TeamTTA: Efficient Multi-Device Collaboration for Open-Set Test-Time Adaptation via Cloud Integration

ANQI LU*, Harbin Institute of Technology, China
YOUBING HU†, Harbin Institute of Technology, China
YUN CHENG‡, ETH Zurich, Switzerland
DAWEI WEI, Xidian University, China
ZHIQIANG CAO, Harbin Institute of Technology, China
JIE LIU, Harbin Institute of Technology, China
ZHIJUN LI§, Harbin Institute of Technology, China

Deep neural networks (DNNs) deployed on edge devices often suffer from severe performance degradation when exposed to dynamic and continually shifting environments. Test-time adaptation (TTA) has emerged as a promising solution by updating models online with incoming test data. However, edge deployment poses unique challenges: limited computational resources, latency caused by adaptation delays, and knowledge isolation across devices. The situation becomes even more complex in open-world scenarios, where the presence of unknown categories further disrupts adaptation. To overcome these limitations, we propose TeamTTA, a cloud-integrated framework designed for efficient multi-device collaboration open-set test-time adaptation. Specifically, TeamTTA aggregates reliable samples from multiple edge devices through crowdsourcing, uploads them to the cloud, and maintains a memory buffer for continual adaptation. A large vision model (LVM) in the cloud leverages its zero-shot generalization ability to filter out open-set samples and acts as a teacher model, distilling its knowledge into a replicated student edge model stored in the cloud. The adapted model parameters, or alternatively global statistics under poor network conditions, are then transmitted back to the edge devices for efficient inference. Extensive experiments on standard public TTA benchmarks, including corrupted and open-set datasets, show that TeamTTA achieves superior adaptation accuracy, robustness to distribution shifts, and communication efficiency, outperforming state-of-the-art TTA baselines. These results validate the effectiveness of integrating cloud-edge collaboration and LVM-driven knowledge distillation for real-world edge intelligence.

JAIR Associate Editor: Jianwen Xie

*Equal contribution.
†Equal contribution.
‡Corresponding Author.
§Corresponding Author.

Authors' Contact Information: Anqi Lu, ORCID: 0000-0002-8835-6938, luanqi@stu.hit.edu.cn, Harbin Institute of Technology, Harbin, China; Youbing Hu, ORCID: 0000-0003-2181-9659, youbing@stu.hit.edu.cn, Harbin Institute of Technology, Harbin, China; Yun Cheng, ORCID: 0000-0002-0421-1716, yun.cheng@sdsc.ethz.ch, ETH Zurich, Zurich, Switzerland; Dawei Wei, ORCID: 0000-0003-1134-6165, weidawei58@gmail.com, Xidian University, Xi'an, China; Zhiqiang Cao, ORCID: 0000-0002-7351-5887, zhiqiang_cao@stu.hit.edu.cn, Harbin Institute of Technology, Harbin, China; Jie Liu, ORCID: 0000-0001-6209-6886, mjieliu@outlook.com, Harbin Institute of Technology, Harbin, China; Zhijun Li, ORCID: 0000-0001-9129-9957, lizhijun_os@hit.edu.cn, Harbin Institute of Technology, Harbin, China.

DOI: 10.1613/jair.1.21292

JAIR Reference Format:
Anqi Lu, Youbing Hu, Yun Cheng, Dawei Wei, Zhiqiang Cao, Jie Liu, and Zhijun Li. 2026. TeamTTA: Efficient Multi-Device Collaboration for Open-Set Test-Time Adaptation via Cloud Integration. *Journal of Artificial Intelligence Research* 85, Article 43 (April 2026), 28 pages. DOI: 10.1613/jair.1.21292

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success in computer vision (Voulodimos et al. 2018) tasks such as image classification (L. Chen et al. 2021), object detection (Zou et al. 2023), and semantic segmentation (Mo et al. 2022). Despite their outstanding performance in cloud-based and laboratory environments, deploying DNNs on resource-constrained edge devices still faces significant challenges. Specifically, edge devices often operate in complex and dynamically changing environments, such as image blur, illumination variation, and sensor performance degradation (Niu, J. Wu, Y. Zhang, Wen, et al. 2023). These factors lead to distribution shift, where the test data distribution during deployment differs from that in training, thereby severely affecting model inference performance (Murphy 2022). In particular, in edge environments, models must simultaneously cope with two key types of distribution shifts: covariate shift (domain shift) (Stan and Rostami 2024) and semantic shift. Covariate shift refers to changes in the test distribution, such as variations in image characteristics caused by weather changes (Ioffe 2015); semantic shift refers to the occurrence of categories during testing that are not included in the training set, which often happens in open-world environments (Gao et al. 2024). These shifts may lead to severe consequences such as misclassification, thereby significantly undermining the reliability and robustness of the model in real-world tasks. Therefore, improving the adaptability of DNNs in dynamic edge scenarios under distribution shift has become one of the core issues in current edge intelligence research.

To address distribution shift, test-time adaptation (TTA) has emerged as a promising paradigm (D. Wang et al. 2020). The core objective of TTA is to update the model during inference by exploiting incoming unlabeled test data without requiring access to source data. This property is particularly important for resource-constrained edge devices operating in complex and rapidly changing task scenarios. Numerous methods (Liang et al. 2023; Z. Wang, Luo, et al. 2024; Xiao and Snoek 2024; Yu et al. 2023) have been proposed to enhance both the performance (e.g., accuracy) and efficiency (e.g., reduced memory footprint and computational cost) of TTA. For example, CoTTA (Q. Wang et al. 2022) leverages image augmentation to mitigate error accumulation, while RoTTA (Yuan et al. 2023) introduces a robust batch normalization (BN) scheme for more accurate normalization statistics. These approaches significantly improve adaptation accuracy. In contrast, another method emphasizes enhancing the efficiency of TTA—EATA (Niu, J. Wu, Y. Zhang, Y. Chen, et al. 2022) adapts only on reliable samples and LAW (Park et al. 2024) minimizes the number of layers requiring adaptation. Despite these advances, most existing TTA approaches assume closed-set environments and large batch sizes, making them unsuitable for open-set edge deployment.

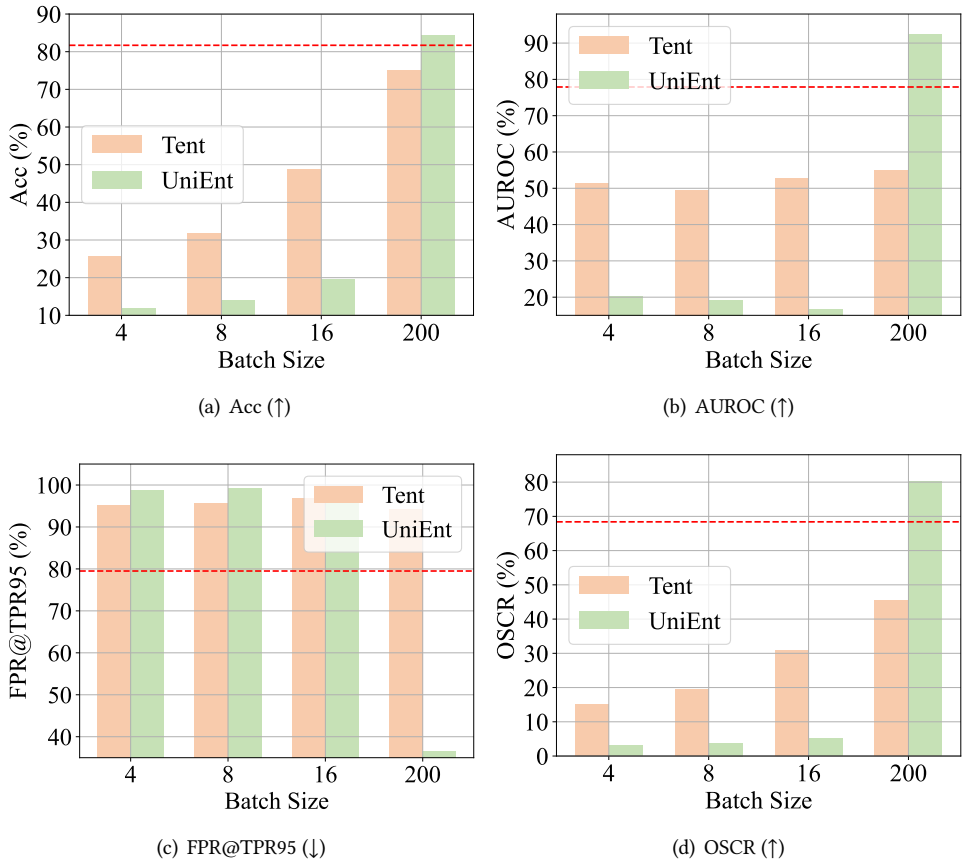
Open-set adaptation (Gao et al. 2024; Lee, Das, et al. 2023) poses a significant challenge, especially for models deployed on resource-constrained edge devices. Unlike the closed-set assumption, where all test categories are present in the training data, open-set scenarios require the model to handle instances from entirely unseen categories during deployment. For example, in autonomous driving scenarios (Wong et al. 2020), a vehicle entering a dimly lit tunnel may encounter novel, unrecognized objects, potentially leading to severe traffic accidents. In such open-set scenarios, models must not only maintain high performance on known categories (closed sets) but also correctly recognize unknown categories (open sets) that are not encountered during training. This dual requirement increases the complexity of TTA, as the model must simultaneously differentiate between known and novel categories when adapting to continuously changing environmental conditions.

Recent studies have explored innovative solutions for open-set adaptation. OSTTA (Lee, Das, et al. 2023) uses crowd wisdom to filter out open-set samples, while UniEnt (Gao et al. 2024) proposes a unified entropy optimization framework. Specifically, the UniEnt framework first employs a distribution-aware filter to distinguish

between covariate-shifted in-distribution (csID) and covariate-shifted out-of-distribution (csOOD) samples. It then minimizes entropy for csID samples while maximizing it for csOOD samples, facilitating more effective adaptation in open-set environments. Although these open-set TTA methods demonstrate strong performance in controlled experimental settings, they exhibit critical limitations that hinder practical deployment in real-world, resource-constrained edge environments: (1) *Heavy reliance on large batch sizes for adaptation.* Model performance degrades sharply as batch size decreases; in extreme cases, adaptation can even cause catastrophic failure, producing results worse than those of the original, unadapted model. As illustrated in Fig.1, when operating with small batch sizes, the state-of-the-art (SOTA) open-set TTA method UniEnt (Gao et al. 2024) performs significantly worse than the closed-set method Tent (D. Wang et al. 2020). For Fig.1, we empirically validate the limitations of existing TTA methods using the CIFAR-10-C dataset (Krizhevsky, Hinton, et al. 2009) to simulate covariate shift, and extend the evaluation to open-set scenarios by introducing the SVHN-C dataset (Netzer et al. 2011) as the out-of-distribution (OOD) domain (See Section 5.1 for specific experimental settings). To better reflect practical deployment conditions—where high-resolution inputs are common—we upsample all images from their original resolutions to 224×224. Experiments are conducted under varying batch sizes of 4, 8, 16, and 200 to investigate performance sensitivity to batch size. The Adam optimizer (Kingma 2014) is employed for model adaptation across all settings. Performance is evaluated using a comprehensive set of metrics: Accuracy (Acc) for in-distribution classification performance, AUROC for open-set separability, FPR@TPR95 to quantify false positive rates at a fixed true positive rate, and OSCR to jointly measure classification and OOD detection capability. Detailed definitions for these metrics are provided in Section 5.1. (2) *High memory consumption with increasing batch sizes.* Even at relatively low image resolutions (e.g., 32×32), memory usage escalates rapidly as batch size grows, as shown in Fig. 2. This issue is exacerbated in real-world edge scenarios, where input images often have substantially higher resolutions, further straining the already limited on-device memory resources. These constraints underscore the pressing need for open-set TTA methods that maintain robustness under small-batch settings, while operating efficiently within the stringent memory and computational limits of edge devices.

In this paper, we identify several key issues that TTA faces in edge scenarios. **Firstly, isolated operation and lack of knowledge sharing.** Edge devices often function as isolated knowledge silos, unable to exchange or leverage useful adaptation information from other devices. This isolation limits their ability to generalize quickly, resulting in suboptimal adaptation performance. **Secondly, severe resource constraints for online adaptation.** Most edge devices possess just enough computational and memory capacity to support forward inference, but lack the additional resources, particularly GPU memory and processing power, required for backpropagation during adaptation. This restricts the application of existing TTA methods. **Thirdly, strict task completion time constraints.** Applications in dynamic edge environments, such as autonomous driving, often have task completion time requirements. These real-time demands make it impractical to deploy sophisticated TTA methods. **Finally, low robustness to open-set conditions.** In practical deployment, models frequently encounter open-set data containing unseen classes. Without robust open-set handling, such conditions can cause severe performance degradation or even lead to model collapse, undermining both reliability and safety. Addressing these issues is essential to enabling TTA methods that are not only accurate but also efficient, robust, and deployable under the stringent constraints of edge intelligence scenarios.

To overcome the aforementioned challenges, we propose TeamTTA, a novel cloud-integrated framework for efficient and open-set TTA in edge environments. Inspired by crowdsourcing (X. Chen et al. 2022; Smirnova et al. 2024; Z. Wang, Zhao, et al. 2021; Zhong et al. 2023), TeamTTA adopts a cloud-edge and multi-device collaborative paradigm, in which the cloud serves as a task publisher that issues sample collection requests, while multiple edge devices act as workers contributing reliable samples. A large vision model (LVM) (Y. Chen, Cai, et al. 2024; Tahmasebi et al. 2024) deployed in the cloud first refines the collected samples by filtering them for open-set instances. Leveraging these refined datasets, the LVM then guides the update of replicated edge models in the cloud, after which the new model parameters are sent back to the corresponding edge devices, completing



(a) Acc (↑) (b) AUROC (↑)

(c) FPR@TPR95 (↓) (d) OSCR (↑)

Fig. 1. Performance comparison of different methods across various batch sizes. The red dashed line represents the original model's performance. ↑ indicates better values, and vice versa.

the adaptation cycle. A central advantage of TeamTTA lies in its ability to offload computationally expensive tasks to the cloud, enabling resource-constrained edge devices to benefit from efficient backpropagation and collaborative learning without exceeding their hardware limitations. In addition, TeamTTA capitalizes on the zero-shot capabilities of LVMs to address open-set challenges, thereby enhancing model robustness in complex real-world scenarios (Table 1). Finally, as a plug-and-play framework, TeamTTA can be seamlessly integrated with existing TTA methods, further improving overall performance and adaptability (Fig. 7).

In summary, our main contributions are as follows:

- We identify the key challenges of edge TTA as knowledge isolation, resource constraints, latency sensitivity, and open-set conditions. To tackle these issues, we propose TeamTTA, a universal and efficient framework, which is the first work to study edge open-set TTA.
- We integrate the idea of crowdsourcing into TeamTTA, enabling multiple edge devices to collaborate and share adaptation knowledge through the cloud, thereby breaking the limitations of single-device isolation

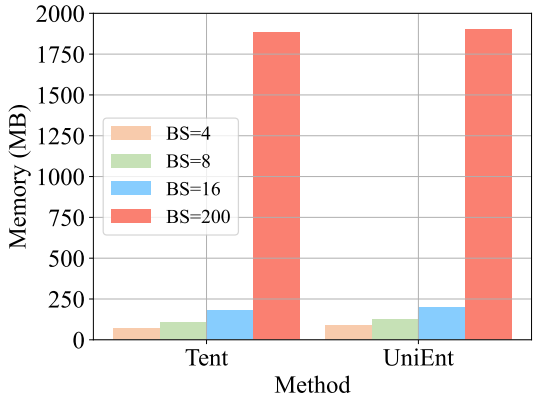


Fig. 2. Memory usage of different methods across various batch sizes. Memory footprint includes the memory consumption of both model parameters and activation storage (Song et al. 2023).

and alleviating resource constraints. Furthermore, we leverage the capabilities of large vision models (LVMs) to guide edge model updates, providing robust open-set recognition.

- We conduct extensive experiments on CIFAR benchmarks, with upsampled high-resolution inputs (224 × 224) to better match real-world application demands, demonstrating that TeamTTA achieves superior performance, robustness, and efficiency compared to state-of-the-art TTA methods, and can be combined with existing TTA methods as a plug-and-play module.

2 Related Work

2.1 Test-Time Adaptation and Optimization

Test-Time Adaptation (TTA) aims to mitigate performance degradation caused by distribution shift by adapting model parameters during inference, without access to the source data (Lee, Das, et al. 2023). A seminal work in this field is Tent (D. Wang et al. 2020), which updates Batch Normalization (BN) parameters by minimizing entropy loss, offering a simple yet highly efficient adaptation strategy. Building on this foundation, CoTTA (Q. Wang et al. 2022) extends TTA into a continual adaptation paradigm, enabling models to handle covariate drift (domain drift) (Sun et al. 2022), which is particularly beneficial for real-world, dynamic scenarios such as autonomous driving.

Research on TTA optimization focuses on performance and efficiency. **Performance Optimization:** Inspired by Tent (D. Wang et al. 2020), many studies have refined entropy minimization to enhance adaptability. For instance, MEMO (M. Zhang et al. 2022) enhances single-sample test-time robustness by leveraging data augmentation and adaptive adjustment of model parameters to minimize the prediction entropy on augmented samples. Domainadaptor (J. Zhang et al. 2023) mitigates inaccurate statistic estimation by dynamically blending training and test statistics. Additionally, it employs a generalized entropy minimization loss with temperature scaling to enhance the model's learning capability for high-confidence samples. SoTTA (Gong, Y. Kim, et al. 2024) enhances robustness through a dual-strategy approach: high-confidence uniform class sampling at the input level to filter out noisy samples and balance class distribution, and entropy sharpness minimization at the parameter level to smooth the loss function, thereby reducing gradient disturbances caused by noisy samples. DeYO (Lee, Jung, et al. 2024) assesses prediction changes before and after data augmentation using pseudo-label probability differences, boosting adaptation accuracy. STAMP (Yu et al. 2025) improves robustness to

distribution shift by constructing a stable memory bank that dynamically stores reliable samples with low entropy and consistent labels, and by combining this with a self-weighted entropy minimization strategy to optimize the model. **Efficiency Optimization:** Efficient adaptation is crucial for resource-constrained devices. For example, EATA (Niu, J. Wu, Y. Zhang, Y. Chen, et al. 2022) introduces an active sample selection strategy that optimizes only reliable and non-redundant test samples, thereby improving computational efficiency. In addition, it incorporates Fisher information-based weight regularization to prevent the model from forgetting knowledge of the original distribution when adapting to a new one. MECTA (Hong et al. 2023) significantly reduces memory overhead by minimizing caching in three dimensions: batch size, channel count, and layer depth. Specifically, it dynamically adjusts the forgetting gates of normalization layers to stabilize small-batch statistics, randomly prunes channels to reduce memory usage for gradient computation, and trains layers on demand to avoid redundant computation. LayerwiseTTA (Park et al. 2024) proposes a layer-wise automatic weighting algorithm based on the Fisher information matrix, which can autonomously identify neural network layers that need to be preserved or intensively adapted. It further incorporates an exponential min–max scaler to amplify the differences in learning weights between layers, thereby effectively handling target distribution shift while reducing computational load. CEMA (Y. Chen, Niu, et al. 2024) innovatively designs a cloud–edge elastic model adaptation paradigm, which offloads computationally intensive tasks to the cloud and optimizes the adaptation process through cloud integration. This work offers new perspectives for deploying TTA on resource-constrained devices.

2.2 Open-Set Test-Time Adaptation

Traditional TTA methods primarily target closed-set scenarios, where the categories in the test data are assumed to fully overlap with those seen during training. This assumption rarely holds in practice, as real-world environments often contain unseen or novel categories that are absent from the training distribution. Such open-set conditions pose a greater challenge, as models must not only adapt to distribution shift but also avoid making overconfident predictions on unknown classes. To tackle this problem, recent studies such as OSTTA (Lee, Das, et al. 2023) and UniEnt (Gao et al. 2024) extend TTA into the open-set setting. These methods employ entropy-based strategies to distinguish between covariate-shifted in-distribution (csID) and out-of-distribution (csOOD) samples, enabling targeted adaptation that improves robustness against open-set noise and unseen categories.

This paper edge TTA in open environments under the Continual Test-Time Adaptation (CTTA) (Nguyen et al. 2024) setting. To strengthen robustness in such challenging conditions, we incorporate the zero-shot generalization capability of large vision models (LVMs) (Chitty-Venkata et al. 2024) into the adaptation process. Unlike conventional TTA methods that primarily focus on covariate shift (changes in input distribution), our proposed framework simultaneously addresses both covariate shift and semantic shift by enabling the model to detect and adapt to previously unseen categories in open-set scenarios (Geng et al. 2020). This dual adaptation strategy not only improves classification accuracy for known categories but also enhances the model's ability to handle novel categories without retraining, thereby broadening the applicability and adaptability of TTA in real-world edge deployments.

3 Problem Statement

Continual Test-time Adaptation. Given a target domain sequence $D = \{D_1, D_2, \ldots, D_t\}$, the target domain at time step $T = k$ is denoted as $D_k = \left\{X_i^k\right\}_{i=1}^{N_k}$. Here, x_i^k represents the i-th unlabeled test sample, X_i^k denotes the set of test samples at step k, and N_k is the number of such samples. The source domain is defined as $D_0 = \left\{(X^S, Y^S)\right\}$, where the model is initially trained. Given a deep learning model $f(\theta_0)$, where θ_0 is the initial parameter pre-trained on the source domain D_0. However, due to storage limitations or privacy concerns, the source domain data D_0 cannot be accessed during the test phase (Y. Wang et al. 2024). The objective of Continual Test-time

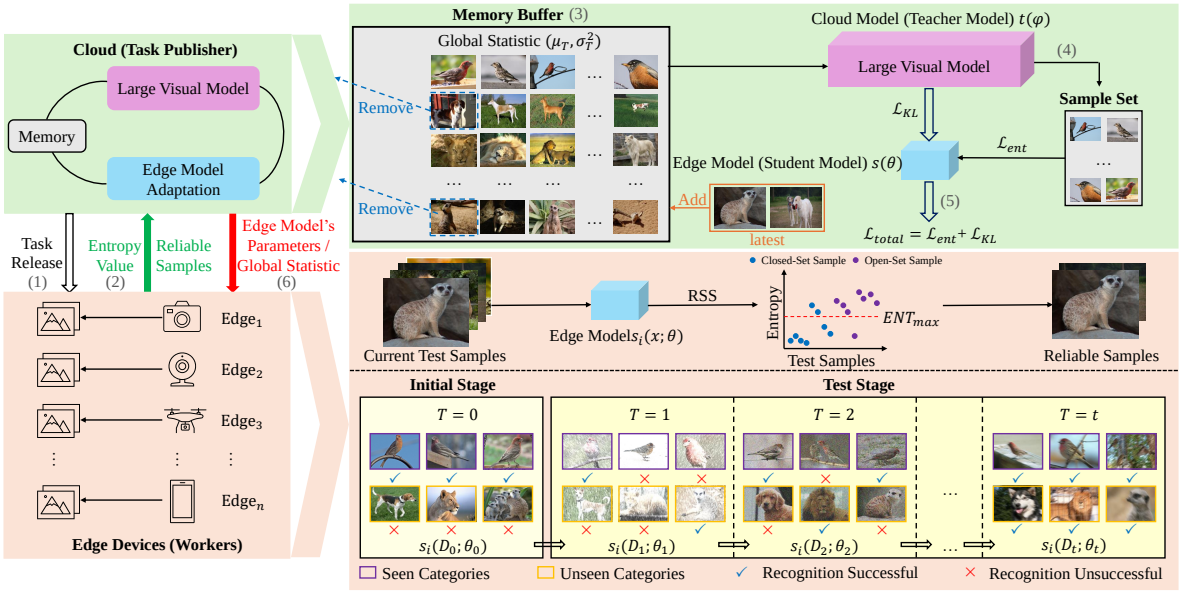


Fig. 3. The overview of TeamTTA framework.

Adaptation (CTTA) is to improve the performance of the model $f(\theta_0)$ in an online manner for a continually changing target domain D_k with $k > 1$. That is, during the inference phase of the target domain D_k, the unlabeled sample data X^k is fed to the model $f(\theta_k)$. The model makes a prediction and adapts its parameters accordingly $\theta_k \longrightarrow \theta_{k+1}$. This continual process enables the model to progressively adapt to domain drift in the target stream without revisiting the source data, ensuring robustness under dynamic environments.

Open-set Continual Test-time Adaptation in TeamTTA.. Building upon the above CTTA setting, our goal of TeamTTA is to further enhance adaptation performance in open-set edge TTA settings by leveraging the collective intelligence of multiple distributed edge devices. We define the set of edge devices as $E = \{Edge_1, Edge_2, \ldots, Edge_n\}$. At the initial stage, all edge devices are deployed with the same model architecture and parameters, capable of classifying only uncorrupted images from the closed-set. Specifically, the initial model on each device $Edge_i$ is denoted as $s_i^0(D_0;\theta_0)$. Besides, we introduce an initial replicated edge model (student model) $s(\theta_0)$ stored in the cloud, which shares the same architecture and initialization as the initial edge models. In parallel, the cloud hosts a large vision model (LVM), denoted as the teacher model $t(\varphi)$. The LVM performs zero-shot inference to guide the updating of student model $s(\theta_k)$ at step k. The adaptation process is formulated as parameter updates $\theta_k \longrightarrow \theta_{k+1}$ based on the teacher model's supervision. As a result, edge devices can progressively improve their ability to both classify corrupted closed-set categories and detect novel open-set categories, achieving reliable performance under real-world data streams.

4 Method

4.1 Overview of TeamTTA

Design. Our proposed TeamTTA framework, illustrated in Fig. 3, comprises a centralized cloud server and n distributed edge devices, such as cameras, drones, and smartphones. The overall process of the TeamTTA framework is as follows. (1) The cloud server, as task publisher, initiates a crowdsourcing task across all participating edge

Algorithm 1 TeamTTA framework at edge device $Edge_i$

Initialization: Initial edge model $s_i(D_0; \theta_0)$

Input: Time step $T = k(k \geq 1)$, target domain $D_k = \left\{X_i^k\right\}_{i=1}^{N_k}$, current edge model $s_i(\theta_k)$

Output: Classification predictions of test samples $\left\{\widehat{Y}_i^{k+1}\right\}_{i=1}^{N_{k+1}}$

for test samples $x \in X_i^k$ in D_k **do**
  Calculate the classification prediction $\widehat{y}$ by $s_i(\theta_k)$.
  Calculate the sample selection score $Score(x)$ by Eq. 1.
  if $Score(x) = 1$ **then**
    Upload the reliable sample x with its entropy value $Ent(x; \theta_k)$ to the cloud.
  end if
end for
Update the edge model parameters $\theta_k \rightarrow \theta_{k+1}$ from the cloud.
Calculate the updated edge model $s_{final}(\theta_{k+1})$ by Eq. 6.
Calculate the classification predictions of test samples $\left\{\widehat{Y}_i^{k+1}\right\}_{i=1}^{N_{k+1}}$ by $s_{final}(\theta_{k+1})$.

devices to collect test samples. (2) Upon receiving the task, each edge device as worker evaluates its incoming data stream, applies the Reliable Sample Selection (RSS) strategy to filter low-entropy samples, and uploads only the selected reliable samples along with their corresponding entropy values to the cloud. (3) The cloud server establishes and incrementally updates a memory buffer, which stores representative samples and continuously maintains their global statistics (i.e., mean and variance). (4) All incoming samples in the memory buffer are processed by the Large Vision Model (LVM), which identifies and filters out open-set samples that do not belong to any known categories for model adaptation. (5) A replicated copy of the edge model is then updated on the cloud side using both the LVM's guidance and the filtered sample set. (6) Finally, the updated model parameters are transmitted back to edge devices when network conditions are good (or global statistics when network conditions fluctuate), completing the adaptation cycle. As shown in Fig. 3, at the initial deployment stage, an edge device $Edge_i$ can only recognize clean and in-distribution categories (e.g., uncorrupted birds), failing to identify unknown categories (e.g., dogs, lions, meerkats). Over time, with the adaptation process enabled by TeamTTA, $Edge_i$ progressively adapts to classify corrupted known categories and detect corrupted unknown categories. For example, at time step $T = t$, the adapted edge model can not only classify heavily fuzzy bird images with high accuracy but also reliably detect previously unseen categories such as fuzzy dogs, lions, and meerkats, even under severe image corruption. This continual improvement is achieved through coordinated adaptation between the cloud and multiple edge devices, where reliable samples are selectively uploaded to the cloud for aggregation, guidance, and feedback. For clarity, the pseudo-code describing the operations of edge devices and the cloud server in our TeamTTA framework is provided in Algorithm 1 and Algorithm 2, respectively.

Edge. Edge devices begin by classifying incoming test samples using their locally stored, current model to obtain preliminary predictions. Next, they apply the RSS method, which leverages prediction entropy to assess the confidence of each sample and filter out unreliable samples. Finally, the selected samples, together with their associated entropy values, are transmitted to the cloud, where they will be aggregated and used for model adaptation.

Cloud. A memory buffer is initialized using the uploaded samples and is continuously updated at each time step to reflect the changing test data stream. Besides, the cloud maintains a group of global statistics that summarizes the aggregated knowledge from all participating edge devices. The LVM then performs zero-shot inference on the

Algorithm 2 TeamTTA framework in the cloud

Initialization: Large visual model $t(\varphi)$, initial copy of edge model $s(\theta_0)$, memory buffer M with its capacity B
Input: Time step $T = k(k \geq 1)$, uploaded test sample set I_k
Output: Updated edge model $s(\theta_{k+1})$, updated global statistics $(\mu_{k+1}, \sigma^2_{k+1})$

```
Add I_k into M.
Calculate the sample importance Imp(x), x ∈ M by Eq. 2.
if |M| > B then
    Remove samples by Imp(x) from smallest to largest.
end if
Calculate the updated group of global statistics (μ_k, σ_k^2) by Eq. 3 and Eq. 4.
Filter a sample set N_k' from M with t(φ).
Update the edge model parameters θ_k → θ_{k+1} via L_total by Eq. 5 with t(φ) and N_k'.
Get the updated edge model s(θ_{k+1}) by Eq. 6.
if Network conditions fluctuate then
    Send the updated global statistics (μ_{k+1}, σ_{k+1}^2) to edge devices.
else
    Send the updated model parameters θ_{k+1} to edge devices.
end if
```

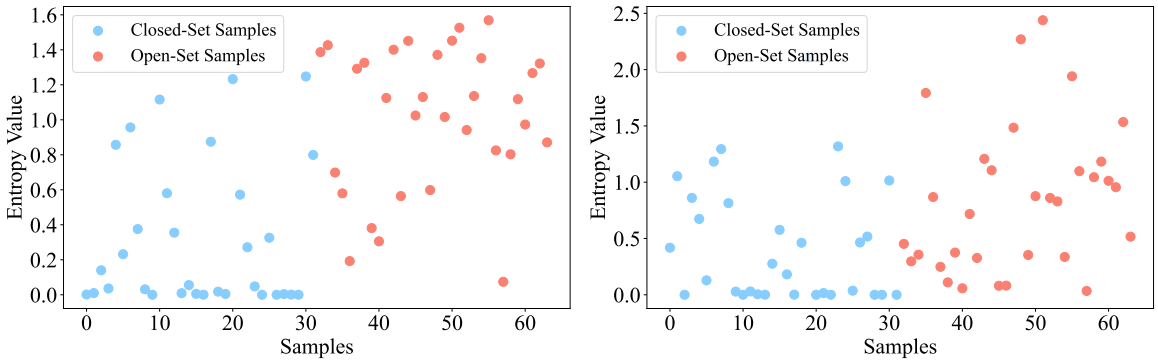
buffered samples to further filter out open-set samples and try to ensure in-distribution data for model adaptation. Beyond filtering, the LVM also distills semantic knowledge into the replicated copy of the edge model, enhancing its ability to handle both covariate shift and semantic shift. Finally, depending on real-time network conditions, the cloud selectively transmits either the global statistics or the updated model parameters back to the edge devices.

4.2 Reliable Sample Selection

We employ an entropy-based Reliable Sample Selection (RSS) method to perform initial filtering of samples on edge devices to ensure that only useful samples are uploaded for model adaptation. Prior work provides strong motivation for this design. EATA (Niu, J. Wu, Y. Zhang, Y. Chen, et al. 2022) points out that high-entropy samples may lead to biased and unreliable gradients, thereby negatively affecting model performance. Moreover, in open-set scenarios, UniEnt (Gao et al. 2024) indicates that unknown-category samples may cause inaccurate estimation of batch normalization (BN) statistics, and minimizing the entropy of samples containing unknown categories may undermine the model's confidence. These issues result in a significant degradation of classification performance on known categories and detection performance on unknown categories. To address these problems, we randomly select a batch of 64 online samples, consisting of 32 closed-set samples and 32 open-set samples, and analyze their entropy values, as shown in Fig.4. The results reveal that most closed-set samples have relatively low entropy values, whereas open-set samples tend to have higher entropy values. Based on this observation, we define a maximum entropy threshold ENT_{max} to filter out unreliable test samples. Additionally, for each test sample, we assign a sample selection score (D. Wang et al. 2020), denoted as $Score(x)$.

$$Score(x) = \mathbb{I}_{\{Ent(x;\theta) < ENT_{max}\}}(x) \tag{1}$$

where $\mathbb{I}_{\{\cdot\}}(\cdot)$ is an indicator function and $Ent(x;\theta)$ is the entropy value of the prediction $s_i(\theta)$ for sample x at edge device $Edge_i$. When $Score(x) = 1$, sample x is deemed reliable. Edge devices transmit reliable samples with entropy values $Ent(x;\theta)$ to the cloud.



(a) CIFAR-10-C dataset (b) CIFAR-100-C dataset

Fig. 4. Entropy values between closed-set and open-set samples.

To further quantify the discriminative behavior of entropy, we additionally compute the Pearson and Spearman correlation coefficients between sample type (closed-set vs. open-set) and entropy values. Specifically, we randomly select 64 online samples (32 closed-set and 32 open-set) consistent with Fig. 4. For CIFAR-10-C/SVHN-C and CIFAR-100-C/SVHN-C datasets, the corresponding Pearson correlations are -0.3744 and -0.2833, and the Spearman correlations are -0.4026 and -0.2927, respectively. Pearson correlation measures linear dependence, whereas Spearman correlation captures monotonic dependence using rank ordering (De Winter et al. 2016). The negative correlation coefficients confirm a consistent statistical trend: closed-set samples generally exhibit lower entropy, while open-set samples tend to have higher entropy. This trend aligns with the entropy distribution patterns observed in Fig. 4 and quantitatively supports the use of entropy as a preliminary screening indicator. Notably, the correlations are slightly stronger for the CIFAR-10-C + SVHN-C setting compared to CIFAR-100-C + SVHN-C. This is expected because the CIFAR-10 dataset contains fewer and more semantically separated categories, making entropy more sensitive to distribution shifts. In contrast, the CIFAR-100 dataset involves a significantly larger number of fine-grained classes, where the semantic distances between closed-set and open-set samples are smaller, naturally reducing the discriminative strength of entropy. The absolute values of the correlation coefficients (approximately 0.2–0.4) indicate low to moderate correlation. This means that entropy alone cannot perfectly distinguish closed-set from open-set samples due to overlapping distributions, yet it does provide statistically meaningful discriminative power. These observations align with our design philosophy—entropy is intentionally used only as a lightweight, low-overhead coarse filtering metric on edge devices, effectively reducing the number of high-uncertainty samples before transmission. The inherent limitations of entropy further justify the necessity of our two-stage architecture, in which the cloud-based CLIP model performs fine-grained filtering to correct potential misclassifications induced by entropy-only selection at the edge.

4.3 Memory Buffer Establishment

As the adaptation process unfolds, the number of samples accumulated in the cloud increases steadily. To manage these samples efficiently, we establish a memory buffer M with a fixed capacity B. Once the number of stored samples exceeds B, unimportant samples must be removed to make room for new ones. Rather than applying a naïve first-in–first-out (FIFO) policy, we selectively discard samples that contribute the least to updating the edge models. The importance of each sample is jointly determined by **timeliness** and **entropy**. Recent samples are typically more representative of the current distribution of the target domain, whereas samples with higher

entropy values often exhibit higher uncertainty. Samples with high uncertainty may generate erroneous gradient information, thereby hindering the model's adaptation (Yuan et al. 2023). To capture this intuition, we introduce a heuristic importance function $Imp(x)$, which normalizes both timeliness and entropy, and computes an aggregate importance score for each sample $x \in M$. Samples with lower $Imp(x)$ values are preferentially removed.

$$Imp(x) = \frac{Max(Age(M)) - Age(x)}{Max(Age(M)) - Min(Age(M))} + \frac{Max(Ent(M)) - Ent(x;\theta)}{Max(Ent(M)) - Min(Ent(M))} \tag{2}$$

where $Age(x)$ represents the duration that sample x has been in the buffer, measured as $Age(x) = T.Min(Age(M))$ and $Max(Age(M))$ denote the shortest and longest durations, respectively. $Ent(x;\theta)$ is the entropy value of sample x uploaded from the corresponding edge device to the cloud.

Besides, we maintain a group of global statistics (μ_T, σ_T^2) to normalize the feature mapping of all the samples stored in the memory buffer. Here, μ_T and σ_T^2 denote the mean and variance of the BN layer at time step T, respectively. Prior to initiating the TTA process, the running mean and variance of $s(\theta)$ are used to initialize the global statistics (μ_0, σ_0^2). To preserve the gradient properties of the calibrated statistics during adaptation, we update the group statistics using the exponential moving average (EMA) (Hong et al. 2023; Yuan et al. 2023) method, as shown in Eq. 3 and Eq. 4.

$$\mu_T = (1-\alpha)\mu_{T-1} + \alpha\mu_g \tag{3}$$

$$\sigma_T^2 = (1-\alpha)\sigma_{T-1}^2 + \alpha\sigma_g^2 \tag{4}$$

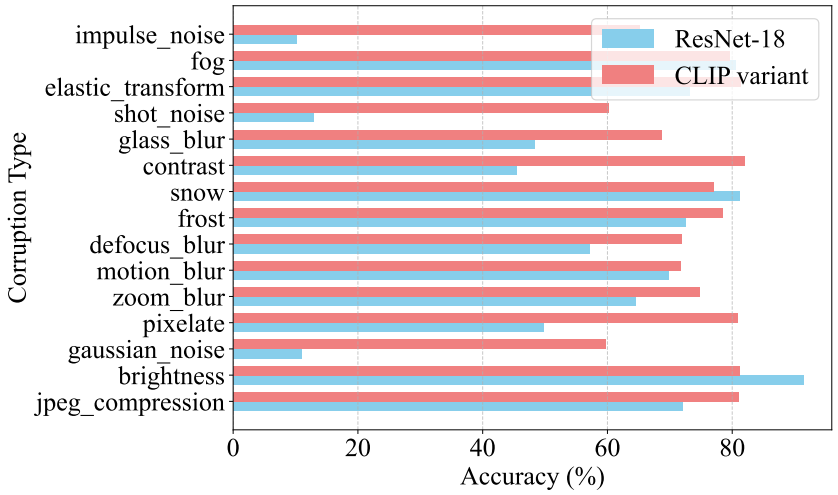
where (μ_g, σ_g^2) represents the global statistic of the current time step's samples, and $\alpha \in [0, 1]$ is a parameter. Smaller α values indicate longer-term memory.

4.4 Model Adaptation

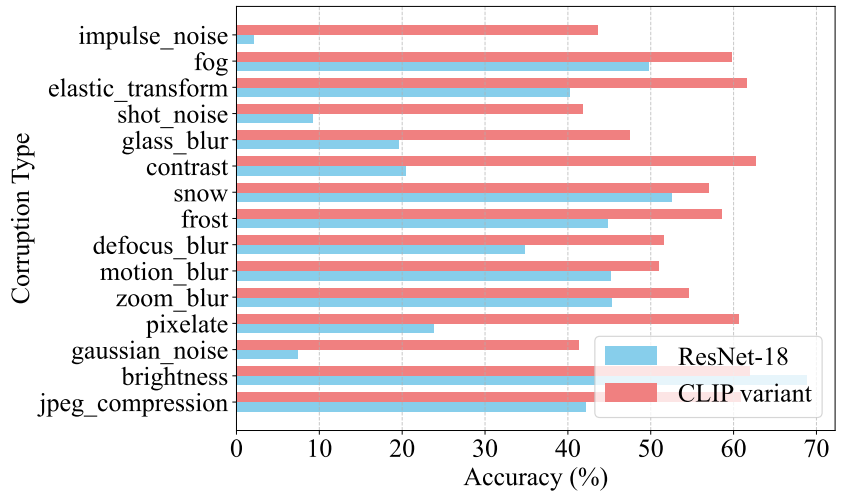
After the initial filtering of test samples at the edge devices, the cloud's memory buffer gradually accumulates a sufficient number of candidate samples. Nevertheless, due to the limited capacity of the lightweight edge models, a portion of open-set samples may remain in the memory buffer. If left unaddressed, these noisy samples may bias the parameter updates of the replicated edge model. To mitigate this, we incorporate Large Vision Models (LVMs) with strong zero-shot inference capability to conduct a secondary filtering step, effectively removing open-set samples and ensuring that only high-quality, noise-free data are retained for model adaptation.

LVM (Bai et al. 2024) utilizes neural networks with a large number of parameters to analyze and understand visual content, thereby capturing complex patterns and relationships in images. For the LVM, we employ CLIP (Contrastive Language–Image Pre-training) (Radford et al. 2021), a powerful vision-language model trained on hundreds of millions of diverse image–text pairs. Unlike conventional supervised models, which require task-specific fine-tuning, CLIP leverages its joint semantic embedding space to generalize across domains and tasks. This property is particularly advantageous in open-world and domain-shift scenarios, where the target distribution contains categories unseen during training. CLIP's zero-shot capabilities enable it to act as a teacher model in the cloud, providing reliable guidance to the lightweight replicated edge models.

To validate CLIP's suitability, we conduct a comparative study on corrupted CIFAR benchmarks, evaluating its zero-shot classification performance against the ResNet-18 (He et al. 2016) edge model. As illustrated in Fig. 5, CLIP consistently achieves substantially higher accuracy across most corruption types, underscoring its robustness and generalization ability under distribution shift. These results strongly support our design choice: CLIP not only serves as a robust filter for open-set noise but also as a knowledge distillation source to enhance the adaptation capacity of edge models.



(a) CIFAR-10-C dataset



(b) CIFAR-100-C dataset

Fig. 5. Performance comparison of ResNet-18 and CLIP accuracy. CLIP functions via zero-shot classification, utilizing an input image paired with a text prompt. For the prompt, we adopt CLIP's default template: 'this is a {category}', where '{category}' represents one of the candidate classes being tested.

Using the LVM, our objective is to update the replicated edge model $s(\theta_k)$ with the filtered sample set $N_k^{'}$ at time step k. To achieve this goal, we introduce knowledge distillation by aligning the prediction distributions of the teacher model (LVM, denoted as $t(\varphi)$) and the student model (replicated copy of the edge model, denoted as $s(\theta_k)$) using Kullback–Leibler (KL) divergence (Y. Chen, Niu, et al. 2024) to optimize the replicated edge model. By minimizing the KL divergence, this alignment process effectively transfers the richer semantic knowledge

encoded in the LVM into the lighter edge model, thereby enhancing its representational capacity and improving generalization on the filtered samples. Consequently, at time step $T = k$, we optimize $s(\theta_k)$ such that $\theta_k \longrightarrow \theta_{k+1}$ by the following equation.

$$\mathcal{L}_{total} = \beta \mathcal{L}_{ent} + (1 - \beta)\mathcal{L}_{KL}(t(\varphi), s(\theta_k)) \tag{5}$$

where β denotes the weight factor that balances the self-supervision loss and the KL divergence. $\mathcal{L}_{ent}$ represents the entropy loss of the replicated edge model, which can be derived from the loss function of existing TTA methods (D. Wang et al. 2020). Besides, we use the KL divergence term to align the prediction distribution between the teacher and student models. The combination of these two losses, through unsupervised learning and knowledge distillation, enables the cloud-side replicated copy of the edge model to learn the knowledge of LVM from the cloud model in a complementary manner.

When updating the replicated copy of the edge model in the cloud, we further adopt the EMA strategy (Hong et al. 2023; Yuan et al. 2023) to mitigate catastrophic forgetting and stabilize the adaptation process. This ensures that the model not only adapts to the changing data distribution but also preserves representations learned previously. Formally, the final adapted model at time step $T = k + 1$ is computed as follows.

$$s_{final}(\theta_{k+1}) = (1 - \alpha)s(\theta_k) + \alpha s(\theta_{k+1}) \tag{6}$$

If network transmission conditions become unstable, the cloud may not be able to deliver the updated model parameters to edge devices in real time. To maintain model adaptability under such constraints, we instead transmit the global statistics (μ_T, σ_T^2) at time step T. These compact statistics require significantly less bandwidth than full parameter updates, enabling efficient synchronization even under limited communication resources. Upon reception, the edge devices leverage these statistics to update their local BN layers' statistics online, thereby aligning the edge model with the latest domain distribution and ensuring more robust inference performance despite fluctuating network conditions.

5 Experiments

5.1 Experimental Setup

Datasets. We follow the common experimental setup in the TTA research field, selecting standard public datasets to ensure fair and reproducible comparisons with existing methods. Specifically, the datasets used in our experiments include the CIFAR (Krizhevsky, Hinton, et al. 2009) benchmarks and the SVHN (Netzer et al. 2011) dataset. The CIFAR benchmark consists of CIFAR-10 and CIFAR-100 datasets, each containing 50,000 training images and 10,000 test images with an original resolution of 32×32. The CIFAR-10 dataset comprises 10 categories: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. In contrast, the CIFAR-100 dataset provides a finer granularity with 100 distinct categories. The SVHN dataset is a real-world street view house number recognition dataset collected and annotated by Google from Google Street View. It also has an original resolution of 32×32 and contains 10 categories corresponding to digits 0–9. Additionally, we use the widely adopted corrupted versions of CIFAR-10 and CIFAR-100 datasets, namely CIFAR-10-C and CIFAR-100-C datasets, to evaluate the proposed TeamTTA framework (D. Wang et al. 2020). These datasets are created by applying various synthetic corruptions to the original CIFAR test sets, aiming to simulate covariate shift encountered in real-world scenarios and comprehensively assess model robustness. Specifically, CIFAR-10-C and CIFAR-100-C datasets both contain 15 types of corruption (e.g., Gaussian noise, motion blur, pixelation) and 5 severity levels (ranging from mild to extreme), providing a standardized benchmark for evaluating model performance under complex perturbations. For the open-set scenario, following prior studies (Gao et al. 2024; Lee, Das, et al. 2023), we introduce the SVHN dataset as unknown category samples into the test environment. To ensure experimental consistency, the SVHN dataset is subjected to the same corruption types and severity levels as the CIFAR-10-C

and CIFAR-100-C datasets. From this, we construct the corresponding SVHN-C dataset using 10,000 test images. This design ensures that open-set samples (SVHN-C) and closed-set samples (CIFAR-10-C and CIFAR-100-C) share identical distribution shift patterns, while their category spaces remain mutually exclusive. That is, SVHN contains digit classes (0–9) that do not overlap with CIFAR's object classes (e.g., airplanes, birds, etc.). Such a setup allows for a rigorous evaluation of the model's ability in two key aspects. (1) Closed-set classification: correctly classifying known categories (from CIFAR-10-C and CIFAR-100-C) under covariate shift; (2) Open-set recognition: effectively rejecting unknown categories (from SVHN-C) and avoiding misclassification as known ones. By jointly using CIFAR-10-C, CIFAR-100-C, and SVHN-C datasets, our experiments provide a unified analysis of both closed-set and open-set performance under identical corruption conditions, thus more faithfully reflecting the model's real-world applicability. The relationship between the CIFAR-10 and SVHN datasets in open-set and closed-set TTA tasks is illustrated in Fig. 6, with a similar relationship holding between the CIFAR-100 and SVHN datasets.

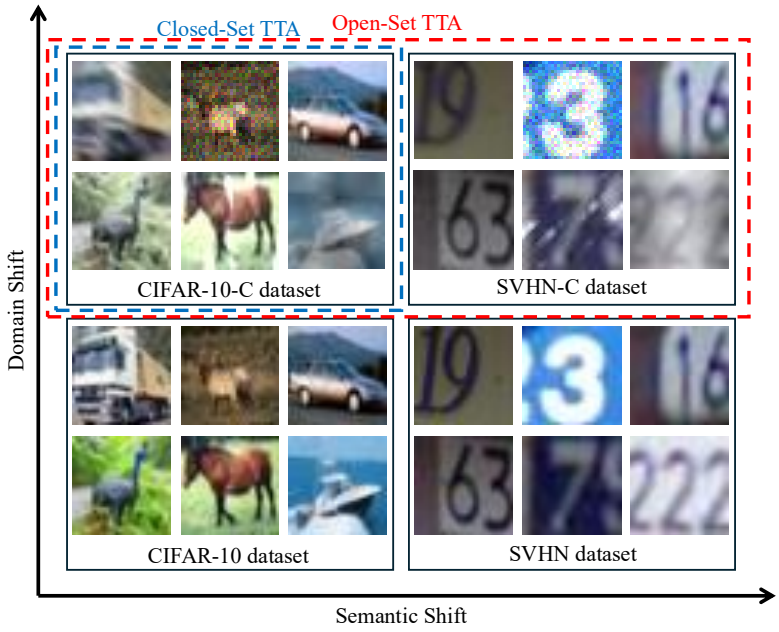


Fig. 6. Relationship between experimental datasets in open-set and closed-set TTA tasks.

Models. We select two representative classification models, one deployed on the cloud and the other on the edge. Specifically, the cloud model utilizes CLIP (Radford et al. 2021), while the edge model employs ResNet-18 (He et al. 2016), which features a simple structure and low computational overhead, making it well-suited for deployment on resource-constrained edge devices. In the training process, the edge model is first trained on a clean, uncorrupted training set under supervised learning to acquire basic image classification capabilities. Subsequently, during the testing phase, the model is evaluated and adapted on a test set containing both corrupted samples and unknown categories.

Metrics. To comprehensively evaluate the performance of the proposed TeamTTA framework, we employ the following set of metrics for quantitative analysis.

(1) **Classification Accuracy (Acc)**: Measures the model's ability to correctly classify samples from known categories, reflecting its recognition performance in closed-set tasks.

(2) **Area Under the Receiver Operating Characteristic curve (AUROC)**: Indicates the model's discriminative ability in distinguishing between known and unknown category samples. Higher values imply better open-set recognition performance.

(3) **False Positive Rate at 95% True Positive Rate (FPR@TPR95)**: Assesses the probability of misclassifying unknown samples as known ones while ensuring the true positive rate reaches 95%. **Lower values indicate stronger robustness in high-recall scenarios.**

(4) **Open-Set Classification Rate (OSCR)** (Dhamija et al. 2018): Provides a joint evaluation of the model's ability to recognize known categories while detecting unknown categories correctly. This metric is particularly suited for open-set scenarios, offering an intuitive measure of the trade-off between recognition and detection performance.

(5) **Number of Uploaded Samples**: Evaluates communication efficiency by measuring the consumption of network resources in edge environments.

Baselines. To validate the effectiveness and advancement of the proposed TeamTTA framework, we conduct comparative experiments against several representative TTA methods, covering approaches from classical techniques to the latest open-set TTA methods, as follows.

(1) **Source Model**: *Source* (Zagoruyko and Komodakis 2016) directly tests the source-domain trained model without any TTA.

(2) **Classical TTA Method**: *Tent* (D. Wang et al. 2020) adapts the model by minimizing the prediction entropy of test samples. This approach requires no access to source data or labels and offers strong generality.

(3) **Efficient TTA Method**: *EATA* (Niu, J. Wu, Y. Zhang, Y. Chen, et al. 2022) selects reliable samples in an active learning–like manner, then updates the model based on these representative samples to reduce computational overhead.

(4) **Classical CTTA Method**: *CoTTA* (Q. Wang et al. 2022) mitigates cumulative errors via weighted averaging and data augmentation. In each iteration, a small portion of neurons is randomly reset to the source pre-trained weights to prevent catastrophic forgetting, enabling long-term adaptation.

(5) **Memory-Enhanced TTA Method**: *RoTTA* (Yuan et al. 2023) proposes a robust batch normalization scheme for estimating normalization statistics, and leverages a memory pool that samples class-balanced data.

(6) **Open-Set TTA Methods**: *OSTTA*(Lee, Das, et al. 2023) filters open-set samples through collective intelligence; *UniEnt* and its improved variant *UniEnt+* (Gao et al. 2024) introduce a unified entropy optimization framework that simultaneously adapts to in-distribution covariate shift and detects out-of-distribution data.

(7) **Latest TTA Method**: *DeYO* (Lee, Jung, et al. 2024) improves adaptation accuracy by evaluating pseudo-label probability differences to assess prediction changes before and after data augmentation.

Implementation Details. To validate the practicality and robustness of the TeamTTA framework in real-world dynamic environments, we evaluate various TTA methods under continuously changing test domains without resetting the model between domains, better simulating realistic deployment scenarios (Q. Wang et al. 2022). Specifically, we set up a multi-device testing environment with three edge devices and randomly assigned the 15 image corruption types from the CIFAR-10-C or CIFAR-100-C datasets to each device, ensuring that the corruption types on each device did not overlap. This simulates the heterogeneity of perceived scenes across different terminals in real-world settings. All methods are evaluated for online adaptation under the most severe corruption level (level 5). During testing, each incoming sample is first used for prediction and then for model adaptation based on the current test sample, following the mainstream TTA process (Y. Wang et al. 2024). Each adaptation batch contains 32 samples, where 16 samples are from the closed-set CIFAR-10-C or CIFAR-100-C dataset and 16 samples are from the open-set SVHN-C dataset. The input resolution for all images is fixed at 224

× 224 to match the precision requirements of high-resolution scenarios. For deployment, the cloud server uses a desktop equipped with an NVIDIA RTX 4090 GPU, while each edge device independently runs its adaptation process as a background thread locally. All experiments are implemented using the PyTorch framework (Paszke et al. 2019). In addition, to assess the model's ability to handle unknown categories, we employ the energy score method (Liu et al. 2020). The energy score measures the log probability density of an input sample to evaluate whether it belongs to the model's known training distribution. Theoretically, energy scores are closely related to the generative probability of a sample: known-category samples typically have lower energy values, while unknown-category samples have higher energy values. Therefore, in open-set recognition, the energy score can serve as an effective criterion—by setting an energy threshold, test samples can be classified as known or unknown, thereby significantly enhancing the model's ability to detect unseen categories. During inference on the edge side, the energy score performs the final open-/closed-set decision for each new sample encountered by the deployed model when computing evaluation metrics. Besides, energy-score threshold follows UniEnt (Gao et al. 2024) and uses the default configuration provided by the official implementation. Code is available at https://github.com/xiaoluludexiaoxiaolu/TeamTTA. The specific parameter configurations are as follows.

(1) Reliable sample selection threshold on the edge side: Maximum entropy threshold ENT_{max} for filtering samples to be uploaded to the cloud.

(2) Cloud-side memory buffer capacity: $B = 128$ to control the scale of historical samples used for adaptation.

(3) Global statistics computation: EMA memory control hyperparameter set to $\alpha = 0.05$ in Eq. 3 and Eq. 4 (Yuan et al. 2023).

(4) Loss function weighting: Weight factor $\beta = 0.9$ for the adaptation loss in Eq. 5.

(5) Optimizer: Adam (Kingma 2014) with a learning rate of 0.001 and momentum of 0.9 to update the cloud-side replicated edge model.

5.2 Experiments Result

Table 1. Performance Comparison of Different Methods across **Three Edge Devices**

Method	CIFAR-10-C				CIFAR-100-C				Average			
	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑
Source	56.0	**86.9**	86.9	52.4	33.8	40.6	99.4	19.2	44.9	63.8	93.2	35.8
Tent	24.7	49.6	96.4	13.9	31.9	62.0	91.5	24.7	28.3	55.8	94.0	19.3
EATA	46.3	51.1	95.3	29.4	**48.7**	70.6	89.1	41.0	47.5	60.9	92.2	35.2
OSTTA	69.7	71.4	87.7	56.5	48.6	**78.2**	80.4	43.5	**59.2**	74.0	83.3	50.9
UniEnt	47.0	51.9	94.2	30.8	33.0	59.2	92.7	24.6	40.0	55.6	93.5	27.7
UniEnt+	61.3	58.7	84.1	36.4	35.7	68.0	91.5	29.6	48.5	63.4	87.8	33.0
TeamTTA (Ours)	**70.6**	77.3	**79.9**	**60.7**	45.7	71.9	**76.7**	**45.6**	58.2	**74.6**	**78.3**	**53.2**

Multi-Device Performance Comparison. We first evaluate the adaptation performance of the proposed TeamTTA framework in a multi-device scenario, and compare it with multiple mainstream TTA baselines. The results are shown in Table 1. The experiments use *a small batch size of 16*, and for both the CIFAR-10-C and CIFAR-100-C datasets, the average values of four key performance metrics are computed over 15 types of image corruptions. To assess multi-device adaptation capability, these metrics are further averaged across the three edge devices. In Table 1, ↑ indicates that larger values are better, and vice versa. From the results, it can be observed that even under small-batch testing conditions, the TeamTTA framework demonstrates excellent adaptation capability, improving classification accuracy for closed-set samples while significantly enhancing recognition performance for open-set samples. Notably, in open-set recognition tasks, TeamTTA shows a clear advantage over existing open-set TTA methods (e.g., OSTTA, UniEnt, and UniEnt+) in identifying unknown categories, indicating

stronger generalization and robustness. Specifically, on both CIFAR-10-C and CIFAR-100-C datasets, TeamTTA outperforms current mainstream methods in most evaluation metrics, fully validating its adaptation performance in multi-device environments. A notable observation is that on CIFAR-100-C, the closed-set accuracy of TeamTTA is slightly lower than OSTTA (45.7% vs. 48.6%). This behavior may stem from a trade-off design in our framework. TeamTTA integrates a KL-based alignment between the edge model and a CLIP-based teacher model, with the total loss Eq. 5. With $\beta = 0.9$, entropy minimization dominates the optimization, while the KL term introduces a smoothing effect due to the inherently softer distribution of CLIP. This alignment significantly enhances open-set robustness—TeamTTA achieves substantially better AUROC, OSCR, and FPR@TPR95 than OSTTA on both datasets—but such smoothing slightly reduces fine-grained discrimination on challenging datasets with many categories. The CIFAR-100-C dataset, containing 100 categories, amplifies this trade-off: the KL-induced regularization mildly attenuates the model's ability to distinguish between visually similar categories, resulting in a small drop in closed-set accuracy. Moreover, Table 1 highlights the impact of category diversity: as the number of categories increases, the closed-set accuracy of all methods decreases due to the increased task difficulty, and TeamTTA exhibits a slight drop—consistent with its design focus on improving open-set performance. Importantly, on the CIFAR-10-C dataset, where the task is less fine-grained, TeamTTA surpasses OSTTA in both closed-set accuracy and open-set metrics. Although its classification accuracy in closed-set TTA tasks is slightly lower than OSTTA (with an average drop of about 1.0%), the improvement in open-set detection is much greater, compensating for the small loss in closed-set accuracy. Overall, these results indicate that the proposed TeamTTA framework can effectively address challenges posed by the combination of covariate shift and semantic shift, providing a more robust and practical TTA solution for real-world edge environments with limited resources.

Table 2. Performance Comparison of Different Methods regarding **Accuracy (%)** on **a Single Edge Device**

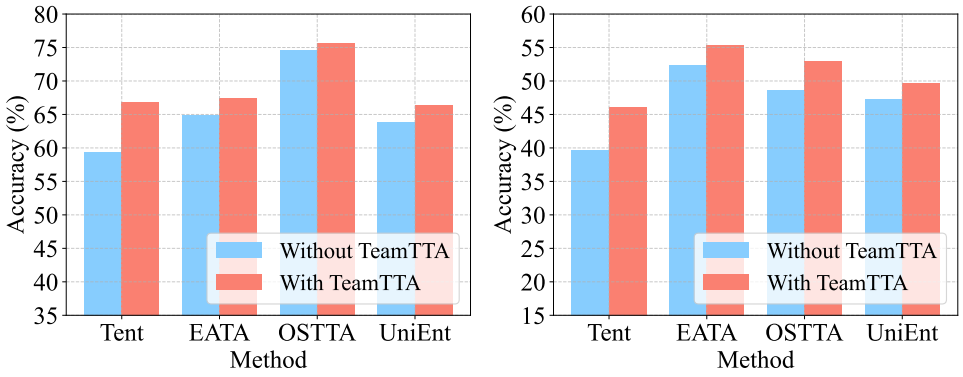
Dataset	Method	Noise			Blur				Weather				Digital				Avg.
		Gauss.	Shot	Impul	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	
CIFAR-10-C	Source	10.9	13.0	10.3	57.1	48.3	69.8	64.6	81.2	72.5	80.6	81.4	45.5	73.2	49.7	72.0	56.0
	Tent	66.1	61.5	33.6	40.7	19.3	17.3	15.4	15.2	14.7	13.1	15.3	15.3	15.0	14.1	14.3	24.7
	CoTTA	51.0	46.1	48.6	46.5	35.3	40.7	40.0	46.6	45.8	31.4	50.4	33.7	34.3	34.8	33.6	41.3
	EATA	66.3	64.4	49.9	62.1	43.8	48.7	45.1	44.3	42.5	42.4	42.5	33.3	36.0	40.0	33.8	46.3
	RoTTA	67.5	**72.8**	**63.6**	**83.1**	61.2	78.3	80.5	79.0	76.7	78.8	86.0	72.0	69.3	71.0	65.5	73.7
	DeYO	**69.3**	72.0	62.7	78.1	59.0	74.9	76.2	74.0	75.5	74.5	72.7	**80.9**	69.7	72.7	65.8	72.5
	TeamTTA (Ours)	66.1	72.7	63.3	82.1	**64.3**	**83.1**	**85.8**	**82.8**	**81.8**	**82.1**	**89.7**	73.5	**73.2**	**74.6**	**70.3**	**76.4**
CIFAR-100-C	Source	7.4	9.2	2.1	34.8	19.7	45.2	45.4	**52.6**	44.9	49.7	**68.8**	20.5	40.3	23.9	42.2	33.8
	Tent	39.6	38.6	31.0	41.8	26.7	35.0	38.6	33.1	32.6	28.7	34.9	23.1	24.8	27.6	22.3	31.9
	CoTTA	19.1	15.0	15.7	16.8	8.8	11.1	10.4	12.0	10.9	5.6	13.0	6.2	7.3	7.8	6.6	11.1
	EATA	**43.2**	**46.4**	**39.5**	56.2	**39.5**	52.2	56.6	50.5	**53.1**	52.7	58.4	53.9	**43.2**	**48.9**	**36.5**	**48.7**
	RoTTA	36.8	38.3	30.0	49.7	26.1	36.6	36.2	29.5	26.9	23.6	28.9	11.1	15.7	14.2	12.6	27.8
	DeYo	41.8	42.4	32.6	45.3	28.9	36.3	42.9	37.8	39.0	32.8	42.3	39.9	25.3	24.8	12.9	35.0
	TeamTTA (Ours)	35.4	39.0	30.9	**57.0**	33.9	**59.3**	**63.0**	51.4	52.5	**54.4**	**64.6**	**54.5**	40.9	44.1	**36.5**	47.8

Single-Device Performance Comparison. Next, we further evaluate the adaptation performance of the proposed TeamTTA framework in a single-device setting on a high-performance edge device with sufficient computational resources, and systematically compare its classification performance in traditional closed-set TTA tasks against various mainstream TTA methods. The results are shown in Table 2. The experiments are conducted on the CIFAR-10-C and CIFAR-100-C datasets, covering 15 common image corruption types, with *a unified batch size of 16* for online adaptation. The results show that on the CIFAR-10-C dataset, TeamTTA achieves the highest classification accuracy for most corruption types, with an average accuracy of 76.4%, significantly surpassing mainstream TTA methods such as Tent, CoTTA, and EATA, demonstrating excellent generalization and adaptability under small-batch test stream conditions. However, on the more challenging CIFAR-100-C dataset, TeamTTA achieves

an average accuracy slightly lower than EATA (by 0.9%), ranking second. This difference may be attributed to EATA's use of source-domain training data during the test phase. In contrast, TeamTTA adapts entirely based on the target-domain test stream without accessing the training set, offering stronger flexibility and better data privacy protection in real-world deployments. It is worth noting that, in combination with the multi-device experimental results in Table 1, we make a structural adjustment to the loss function for the single-device experiments. We remove the KL-divergence regularization term and simplify the original loss function in Eq. 5 to include only the entropy minimization term: $\mathcal{L}_{total} = \mathcal{L}_{ent}$. The motivation behind this adjustment stems from the fact that the LVM in the TeamTTA framework possesses zero-shot open-set recognition capability, enabling it to effectively filter potential unknown-category samples without requiring explicit labels or additional training data. For traditional closed-set TTA tasks (regarding only the metrics of classification accuracy), it is sufficient to perform entropy minimization only on samples identified as belonging to the closed set, without the need for an additional KL-divergence constraint. This results in more efficient optimization and improved classification performance on known categories.

TeamTTA Combined with SOTA TTA Methods. To verify the adaptability and performance gains of TeamTTA when integrated as a plug-and-play module with existing TTA methods, we further compare the classification performance changes of different TTA methods with and without the TeamTTA framework. As illustrated in Fig. 7, *under the batch size setting of 64*, TeamTTA consistently delivers significant and stable performance improvements across multiple baseline TTA methods, demonstrating its strong compatibility and generality. A particularly noteworthy case is the Tent method, which exhibits the most substantial performance improvement after incorporating TeamTTA. On the CIFAR-10-C dataset, the average classification accuracy increases from 59.4% to 66.8%; on the CIFAR-100-C dataset, it rises from 39.6% to 46.1%. These represent gains of 7.4% and 6.5%, respectively. The likely reason for this remarkable improvement is that Tent, as one of the earliest proposed TTA methods, has a relatively simple original design, relying solely on entropy minimization for parameter updates without considering more complex real-world factors such as resource constraints on edge devices or unknown-category recognition. When integrated with our proposed TeamTTA framework, the method benefits from its crowdsourcing mechanism for aggregating reliable samples from multiple devices and its LVM-guided model update process, effectively enhancing Tent's adaptation capability and resulting in a substantial performance boost. In contrast, for more sophisticated TTA methods proposed later, such as OSTTA and UniEnt, which already account for open-set recognition, the performance improvement space for TeamTTA is relatively smaller, though the enhancement trend remains stable. These results indicate that TeamTTA has excellent modularity and can be seamlessly integrated as an enhancement module into various existing TTA methods, improving their ability to handle covariate shift and semantic shift in complex edge scenarios without altering their original design.

Efficiency Comparison. To ensure a fair comparison, we assume that edge devices do not support the model adaptation process, and the computation of all TTA methods is entirely offloaded to the cloud. In Table 3, we compare the cumulative number of uploaded samples when processing 15 types of image corruptions *under a batch size of 128*, to evaluate the practical communication overhead of the proposed TeamTTA framework versus mainstream TTA methods. It is worth noting that Source directly performs inference locally using the source-domain trained model, requiring no test sample uploads and thus incurring zero communication cost. In contrast, methods such as Tent, OSTTA, and UniEnt have no built-in mechanism for selecting samples to upload; they default to sending the entire test set to the cloud for model adaptation, resulting in an upload volume as high as 300k (100%), which leads to extremely high communication costs and makes them unsuitable for deployment in bandwidth-constrained edge environments. Unlike these approaches, our proposed TeamTTA framework incorporates a reliable sample selection mechanism. This mechanism evaluates the prediction entropy of each sample locally on the edge device and uploads only low-entropy, reliable samples. Therefore, it significantly



(a) CIFAR-10-C dataset (b) CIFAR-100-C dataset

Fig. 7. Performance comparison of different methods with or without TeamTTA across **three edge devices**.

Table 3. Number of Uploaded Images for Different Methods

Dataset	Source	Tent	EATA	OSTTA	UniEnt	TeamTTA (Ours)
CIFAR-10-C	-	300k	264k	300k	300k	272k
CIFAR-100-C	-	300k	66.9k	300k	300k	89.7k

reduces the communication burden while preserving model adaptation performance. Specifically, for the CIFAR-10-C dataset, TeamTTA reduces the number of uploaded samples from 300k to 272k, and for the more class-diverse and data-complex CIFAR-100-C dataset, the upload volume drops further to 89.7k, yielding even greater savings. According to the number of uploaded samples, we can get the real transmission volume. Given an image resolution of 224×224 (about 0.15 MB per image), the amount of data uploaded under our reliable sample selection (RSS) strategy is 40.8 MB (CIFAR-10-C dataset) and 13.455 MB (CIFAR-100-C dataset). Compared to a baseline that uploads all 300k test samples (about 45 MB), the CIFAR-10 and CIFAR-100 datasets reduce the transmitted data volume by approximately 9.3% and 70.1%, respectively. The results indicate that TeamTTA is more effective at identifying and filtering out unreliable samples in complex scenarios (such as the CIFAR-100-C dataset), thereby achieving lower communication overhead. It is worth mentioning that, compared with EATA, TeamTTA is slightly more conservative in the number of filtered samples. This may be due to its consideration of open-set samples, which requires its selection strategy to more thoroughly assess the possibility of unknown categories, leading to more cautious upload behavior. While this design sacrifices some transmission efficiency, it trades for stronger generalization and open-set robustness, resulting in a more balanced performance–efficiency trade-off in real-world deployment. Overall, TeamTTA effectively reduces communication costs between edge devices and the cloud while maintaining strong adaptation performance.

Latency Analysis. To evaluate the practical end-to-end runtime of our TeamTTA framework, we construct a real-world network testing environment and use a TP-LINK AC1900 dual-gigabit router to emulate different wireless bandwidth conditions. Specifically, we measure the wall-clock time required to complete a full adaptation cycle under four commonly encountered edge-network bandwidth settings: 10 Mbps, 20 Mbps, 50 Mbps, and 100 Mbps. The results are summarized in Table 4. Across all bandwidth configurations, TeamTTA achieves

Table 4. Wall-clock Adaptation Time under Different Network Conditions

Dataset	10 Mbps	20 Mbps	50 Mbps	100 Mbps
CIFAR-10-C	4.77s	3.03s	1.99s	1.63s
CIFAR-100-C	2.75s	2.14s	1.78s	1.66s

second-level adaptation latency, with total cycle times ranging from approximately 1.6 to 4.8 seconds. These results highlight that TeamTTA introduces minimal adaptation overhead and can operate efficiently even under relatively constrained wireless conditions. This demonstrates that the proposed cloud-integrated TTA framework is well-suited for real-world deployment in bandwidth-limited edge environments.

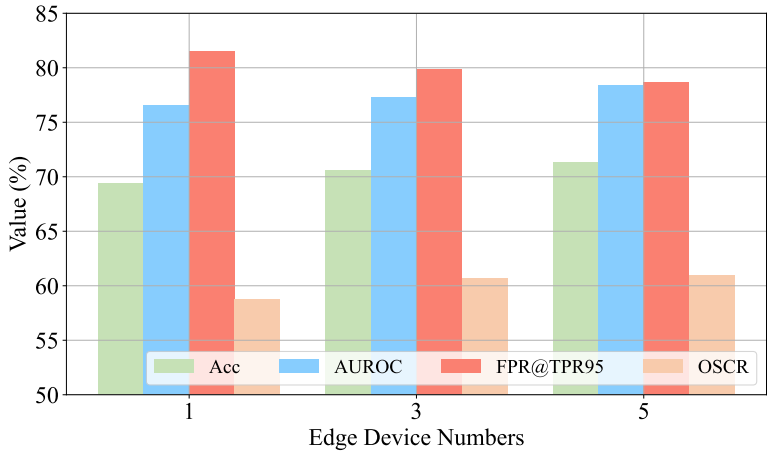
Table 5. Performance of Different Network Conditions

Dataset	Fluctuation	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑
CIFAR-10-C	✗	70.6	77.3	79.9	60.7
	✓	68.5	72.8	84.0	58.9
CIFAR-100-C	✗	45.7	71.9	76.7	45.6
	✓	43.5	67.2	81.1	43.6

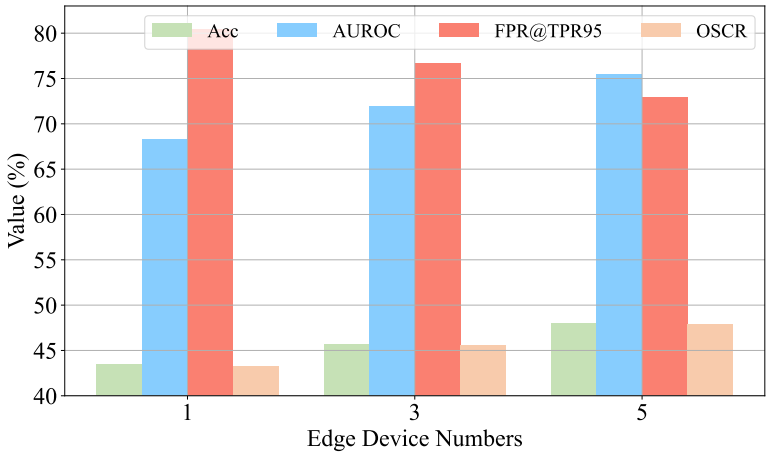
Network Condition. Table 5 presents the performance comparison of the proposed TeamTTA framework under different network transmission conditions *with the batch size of 16*. This experiment aims to evaluate whether the model's adaptation capability is significantly affected during real-world deployment, especially under constrained or unstable network conditions. In practical applications, the updated model parameters on the cloud need to be periodically transmitted to edge devices for synchronized updates. However, due to network bandwidth fluctuations, particularly in remote areas or regions with weak network infrastructure, transmitting the complete set of model parameters is often unsustainable. To address this challenge, the TeamTTA framework introduces a lightweight communication mechanism that, under constrained network conditions, transmits only the globally maintained statistical information from the cloud to the edge devices instead of the complete model parameters. These statistics, including key information such as feature distribution means and variances, account for merely 0.08% of the total model parameter size, significantly reducing the communication burden. As shown in Table 5, even when updating only the statistics without synchronizing the full model parameters, TeamTTA still maintains strong adaptation performance. This result indicates that sharing global statistics plays a crucial role in the adaptation process. The key reason behind this performance retention lies in the fact that these statistics are extracted from reliable samples uploaded by multiple edge devices and crowdsourced on the cloud, effectively integrating diverse environmental information perceived by different devices with high representativeness and stability. This experiment validates TeamTTA's communication robustness and adaptation reliability under bad network conditions.

5.3 Ablation Study

Effect of Edge Device's Number. Fig. 8 illustrates the adaptation performance of the TeamTTA framework under different numbers of edge devices *with the batch size of 16*. Specifically, experiments are conducted with 1, 3, and 5 edge devices, comparing TeamTTA's key performance metrics in both closed-set and open-set TTA tasks. The results show a consistent improvement in overall performance as the number of edge devices increases. For



(a) CIFAR-10-C dataset



(b) CIFAR-100-C dataset

Fig. 8. Effect of edge device's number.

example, on the CIFAR-10-C dataset, the average classification accuracy (Acc) rises from 69.4% (single device) to 70.6% (three devices), and further to 71.3% (five devices). Similarly, in terms of AUROC, the model's ability to distinguish known from unknown categories improves from 76.6% (single device) to 77.3% (three devices), and further to 78.4% (five devices). This trend clearly demonstrates that multi-device crowdsourcing brings stronger generalization and robustness to the model. The core reason for this improvement lies in the fact that as more edge devices participate in the crowdsourcing process, the cloud receives a broader and more semantically diverse range of uploaded samples. When aggregated, these samples can be used to build a more representative memory buffer, effectively enhancing the model's adaptability. Additionally, a multi-device setup facilitates cross-device

collaborative learning, helping edge models overcome the limitations of individual knowledge acquisition and improving their understanding and adaptation to complex perturbation patterns.

Effect of Large Vision Model. Table 6 presents a comparison of the open-set detection performance of the TeamTTA framework with and without the LVM, focusing on three open-set recognition metrics: AUROC, FPR@TPR95, and OSCR. The experimental results clearly demonstrate that incorporating LVM has a significant positive impact on enhancing the open-set robustness of the TeamTTA framework. Due to its strong zero-shot inference capability, LVM can leverage a pretrained semantic space to recognize unseen categories in the target domain. The knowledge it generates can guide the update of edge models, enabling more accurate differentiation between known and unknown samples. Specifically, compared to TeamTTA without LVM, the LVM-enhanced TeamTTA framework achieves comprehensive improvements in open-set recognition performance. For the CIFAR-10-C dataset, AUROC improves by 8.1 percentage points, FPR@TPR95 decreases by 8.1 percentage points, and OSCR increases by 2.8 percentage points. These improvements indicate that, without accessing target labels, LVM can leverage its rich semantic priors to fill the cognitive blind spots of edge models in open-set environments, significantly boosting the overall system's ability to perceive unknown categories. The results validate the effectiveness of integrating LVM as an auxiliary module into the TeamTTA framework, delivering notable performance gains while providing edge devices with a lightweight and efficient open-set awareness enhancement strategy.

Table 6. Effect of the Large Visual Model

Dataset	Large Visual Model	AUROC↑	FPR@TPR95↓	OSCR↑
CIFAR-10-C	✗	69.2	88.0	57.9
	✓	77.3	79.9	60.7
CIFAR-100-C	✗	63.3	85.9	40.1
	✓	71.9	76.7	45.6

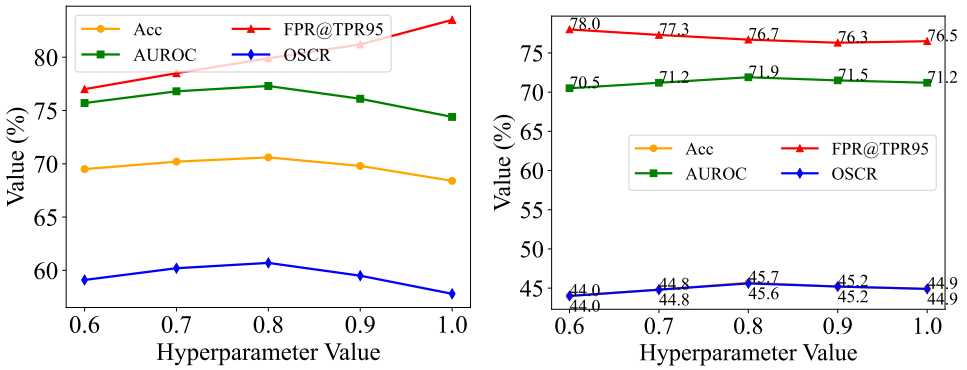
Effect of Memory Buffer. Table 7 presents the performance of the TeamTTA framework under different memory pool capacity settings on the CIFAR-10-C dataset. The experimental results show that the capacity of the memory pool has a certain impact on the model's adaptation performance. As the memory buffer capacity increases, the number of historical test samples accessible to the model grows, leading to continuous improvements in both AUROC and OSCR metrics. However, it is worth noting that this performance gain exhibits a diminishing marginal effect: as the total number of stored samples continues to grow, the improvement rate gradually decreases. Meanwhile, the FPR@TPR95 metric initially decreases with moderate increases in capacity, indicating a reduced false positive rate for unknown samples at high recall, but begins to fluctuate when the memory buffer becomes too large. This fluctuation may be due to excessive historical samples causing the model to overfit to early test streams, thereby interfering with its adaptation to the current distribution. Based on these observations, we ultimately set the memory buffer capacity to $B = 128$ in our experiments. This setting not only effectively enhances the model's robustness to distribution shift and unknown categories but also avoids the computational overhead and adaptation instability caused by redundant information accumulation.

Effect of Memory Buffer. Table 7 presents the performance of the TeamTTA framework under different memory pool capacity settings on the CIFAR-10-C dataset. The experimental results show that the capacity of the memory pool has a certain impact on the model's adaptation performance. As the memory buffer capacity increases, the number of historical test samples accessible to the model grows, leading to continuous improvements in both Acc

and AUROC metrics. However, it is worth noting that this performance gain exhibits a diminishing marginal effect: as the total number of stored samples continues to grow, the improvement rate gradually decreases. For example, the AUROC performance is 4.1% higher when the memory buffer capacity is $B = 128$ than when it is $B = 64$, while the AUROC performance is only 2.4% higher when it is $B = 256$ than when it is $B = 128$. Importantly, the open-set metric, FPR@TPR95, deteriorates when the buffer capacity exceeds $B = 128$. As shown in Table 7, FPR@TPR95 decreases significantly when increasing the buffer capacity B from 64 to 128 (85.2% → 79.9%), indicating improved robustness against misclassifying unknown samples. However, further increasing the buffer capacity B to 256 reverses this trend, raising FPR@TPR95 to 80.1%. Although the numerical increase appears small, it reflects a consistent degradation in open-set reliability: a larger buffer risks admitting more out-of-distribution samples into the adaptation process, thereby increasing the likelihood of falsely accepting unknown categories. This trade-off highlights that simply enlarging the memory buffer does not necessarily benefit open-world adaptation. Excessive historical samples may bias the model toward early test streams, interfere with its responsiveness to the current distribution, and increase memory and latency cost without delivering meaningful accuracy gains. Considering all performance indicators jointly, we set the default memory buffer capacity to $B = 128$. This configuration achieves the best balance between adaptation effectiveness, robustness to unknown categories, and computational efficiency.

Table 7. Effect of the Memory Buffer

Memory Buffer	Acc↑	AUROC↑	FPR@TPR95↓	OSCR↑
64	68.5	73.2	85.2	61.7
128	**70.6**	**77.3**	**79.9**	**60.7**
256	71.5	79.7	80.1	70.6

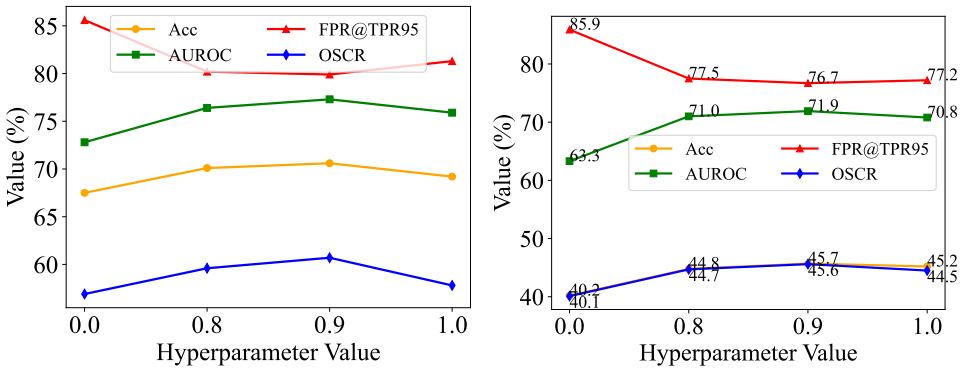


(a) CIFAR-10-C dataset (b) CIFAR-100-C dataset

Fig. 9. Effect of ENT_{max} in RSS method.

Effect of Hyperparameters. To further analyze the tunability of key components in the TeamTTA framework, we evaluate the impact of two core hyperparameters on model adaptation performance: the maximum entropy threshold $ENT_{\max}$ in the Reliable Sample Selection (RSS) method, and the weighting factor β that balances the

two loss terms in the loss function. The experimental results are shown in Fig.9 and Fig.10, respectively. First, in Fig. 9, we observe how varying $ENT_{\max}$ influences the adaptation performance through the RSS method. $ENT_{\max}$ determines the sample-uploading threshold on edge devices, thereby controlling both the quality and quantity of data contributing to model adaptation on the cloud. When $ENT_{\max}$ is set too low, the strict filtering effectively excludes uncertain samples but may also miss potentially valuable adaptation data, leading to performance drops. Conversely, when $ENT_{\max}$ is too high, noisy samples may be included, which can interfere with the LVM's guidance in adaptation. Thus, selecting an appropriate $ENT_{\max}$ strikes a balance between communication cost and adaptation performance. Our experiments find that $ENT_{\max} = 0.9$ yields the best results across most metrics. Second, Fig. 10 illustrates the effect of varying β on the loss function. β is a crucial factor controlling the contribution of the entropy minimization term to the total loss, directly affecting the magnitude of model updates during adaptation. When β is too small, the entropy minimization term is underweighted, limiting the model's adaptation capability. When β is too large, the model may overfit to the current distribution or maladapt to open-set samples, reducing robustness. Experimental results show that $\beta = 0.9$ achieves the best performance in most metrics.



(a) CIFAR-10-C dataset (b) CIFAR-100-C dataset

Fig. 10. Effect of β in loss function.

6 Discussion

6.1 Toward Realistic Non-IID and Open-Set TTA

Although this work follows standard open-set TTA protocols and adopts the SVHN-C dataset as the unseen domain—consistent with prior studies such as OSTTA (Lee, Das, et al. 2023) and UniEnt (Gao et al. 2024)—we acknowledge that the SVHN-C dataset represents a controlled benchmark rather than the full complexity of real-world edge streams. Its use ensures experimental comparability and provides a well-characterized setup for evaluating open-set discrimination, but it does not capture the broader spectrum of semantic, temporal, and contextual variations encountered in practical deployments. At the same time, another important research—non-IID TTA—addresses correlated sampling and temporally dependent test streams, as demonstrated by approaches such as NOTE (Gong, Jeong, et al. 2022) that employ Instance-Aware Batch Normalization and Prediction-Balanced Reservoir Sampling to mitigate batch imbalance and temporal drift. These works, however, focus on closed-set temporal correlation and therefore target a problem orthogonal to the open-set setting considered in our paper. While our experiments do not explicitly evaluate non-IID scenarios, the proposed TeamTTA framework is

naturally extensible: the memory buffer can be adapted into a temporally aware queue, the cloud-side CLIP filtering can be enhanced to detect more complex semantic shifts, and the update mechanism can be generalized to handle batch-dependent drifts. Thus, TeamTTA provides a methodological foundation for future studies that jointly address non-IID sampling and open-set distribution shifts—an emerging and realistic setting for edge intelligence systems.

6.2 Privacy Protection for Image Uploads

Although TeamTTA is designed as an efficient cloud-integrated adaptation framework, deploying it in real-world edge environments also raises important privacy concerns. Uploading image samples to the cloud may expose sensitive visual information in applications such as smart surveillance (Fitwi et al. 2019), medical diagnostics (X. Wang et al. 2022), or industrial inspection (Demertzi et al. 2023). To mitigate this limitation, we explore differentially private noise injection (Brummet et al. 2022) as a practical and theoretically grounded privacy-preserving enhancement. In this approach, each image is perturbed on the edge device with Gaussian or Laplacian noise (Joshi and Fischer 1995) that satisfies a formal differential privacy budget. This mechanism provides mathematically quantifiable privacy guarantees while maintaining full compatibility with the existing cloud-side pipeline, including CLIP-based filtering and knowledge distillation. The primary trade-off lies in noise strength: stronger privacy guarantees (smaller ϵ) introduce more perturbation, which may reduce filtering accuracy and adaptation quality, thereby requiring a balance between privacy and performance. Given its formal guarantees and architectural compatibility, differential privacy–based noise injection represents the most promising extension for reinforcing TeamTTA in privacy-sensitive deployments. In future work, we plan to integrate and systematically evaluate this mechanism to enhance the framework's suitability for real-world cloud-integrated systems.

6.3 Limitations of Entropy-Based RSS

The limitation of the current RSS design lies in the use of a fixed entropy threshold. Although the chosen threshold $ENT_{max} = 0.8$ is empirically validated through ablation experiments on CIFAR benchmarks and performs consistently across corruption types, a static hyperparameter may lack robustness in dynamic, real-world edge environments where data distributions shift over time. In such scenarios, manually tuning the threshold for each new deployment is impractical and undermines the autonomy expected from a TTA system. To address this limitation, we propose adaptive entropy thresholding as an important direction for future work. In particular, a cloud-side sliding-window mechanism can continuously monitor the entropy distribution of uploaded samples and adjust the threshold online to better reflect the current statistics of the incoming data stream. More broadly, while TeamTTA already achieves strong closed-set and open-set performance through its two-stage filtering pipeline, we acknowledge that entropy alone is insufficient when distributions exhibit strong overlap and that a fixed threshold may not generalize well across environments. Future extensions will therefore explore multi-metric sample selection and adaptive threshold adjustment, which we believe will further enhance the robustness and practical deployability of TeamTTA in evolving real-world edge settings.

7 Conclusion

This paper presents TeamTTA, a novel test-time adaptation (TTA) framework designed for edge deployment in open environments. TeamTTA addresses key challenges, including limited resources, adaptation latency, knowledge isolation, and unknown categories, by leveraging crowdsourcing to aggregate reliable test-time samples and offload computation to the cloud. To handle open-set recognition, it further incorporates large vision models (LVMs) to guide the adaptation of edge models. Extensive experiments on CIFAR benchmarks validate the effectiveness and efficiency of TeamTTA. Overall, this work establishes a foundation for open-set TTA at the edge and provides insights to inspire future practical research in this area.

Acknowledgments

Dawei Wei is also affiliated with the Key Laboratory of Cyberspace Security, Ministry of Education, Zhengzhou, China. This work is supposed by the National key R&D Program of China under Grant NO. 2023YFB4503100 and funded by Open Foundation of Key Laboratory of Cyberspace Security, Ministry of Education of China (No. KLCS20240404).

References

Y. Bai, X. Geng, K. Mangalam, A. Bar, A. L. Yuille, T. Darrell, J. Malik, and A. A. Efros. 2024. "Sequential modeling enables scalable learning for large vision models." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22861–22872.

Q. Brummet, E. Mulrow, and K. Wolter. 2022. "The effect of differentially private noise injection on sampling efficiency and funding allocations: Evidence from the 1940 Census." *Harvard Data Science Review*, Special Issue 2.

L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao. 2021. "Review of image classification algorithms based on convolutional neural networks." *Remote Sensing*, 13, 22, 4712.

X. Chen, Y. Zhao, and K. Zheng. 2022. "Task publication time recommendation in spatial crowdsourcing." In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 232–241.

Y. Chen, G. Cai, F. Li, Y. Wang, X. Tan, and X. Li. 2024. "Domain Alignment with Large Vision-language Models for Cross-domain Remote Sensing Image Retrieval." In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 323–333.

Y. Chen, S. Niu, S. Xu, H. Song, Y. Wang, and M. Tan. 2024. "Towards Robust and Efficient Cloud-Edge Elastic Model Adaptation via Selective Entropy Distillation." *arXiv preprint arXiv:2402.17316*.

K. T. Chitty-Venkata, V. K. Sastry, M. Emani, V. Vishwanath, S. Shanmugavelu, and S. Howland. 2024. "WActiGrad: Structured Pruning for Efficient Finetuning and Inference of Large Language Models on AI Accelerators." In: *European Conference on Parallel Processing*. Springer, 317–331.

J. C. De Winter, S. D. Gosling, and J. Potter. 2016. "Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data." *Psychological methods*, 21, 3, 273.

V. Demertzi, S. Demertzis, and K. Demertzis. 2023. "An overview of privacy dimensions on the industrial internet of things (iiot)." *Algorithms*, 16, 8, 378.

A. R. Dhamija, M. Günther, and T. Boult. 2018. "Reducing network agnostophobia." *Advances in Neural Information Processing Systems (NeurIPS)*, 31.

A. Fitwi, Y. Chen, and S. Zhu. 2019. "A lightweight blockchain-based privacy protection for smart surveillance at the edge." In: *2019 IEEE International Conference on Blockchain (Blockchain)*. IEEE, 552–555.

Z. Gao, X.-Y. Zhang, and C.-L. Liu. 2024. "Unified Entropy Optimization for Open-Set Test-Time Adaptation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23975–23984.

C. Geng, S.-j. Huang, and S. Chen. 2020. "Recent advances in open set recognition: A survey." *IEEE transactions on pattern analysis and machine intelligence*, 43, 10, 3614–3631.

T. Gong, J. Jeong, T. Kim, Y. Kim, J. Shin, and S.-J. Lee. 2022. "Note: Robust continual test-time adaptation against temporal correlation." *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 27253–27266.

T. Gong, Y. Kim, T. Lee, S. Chottananurak, and S.-J. Lee. 2024. "SoTTA: Robust Test-Time Adaptation on Noisy Data Streams." *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 770–778.

J. Hong, L. Lyu, J. Zhou, and M. Spranger. 2023. "Mecta: Memory-economic continual test-time model adaptation." In: *2023 International Conference on Learning Representations (ICLR)*.

S. Ioffe. 2015. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167*.

R. L. Joshi and T. R. Fischer. 1995. "Comparison of generalized Gaussian and Laplacian modeling in DCT image coding." *IEEE Signal Processing Letters*, 2, 5, 81–82.

D. P. Kingma. 2014. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*.

A. Krizhevsky, G. Hinton, et al.. 2009. "Learning multiple layers of features from tiny images."

J. Lee, D. Jung, S. Lee, J. Park, J. Shin, U. Hwang, and S. Yoon. 2024. "Entropy is not enough for test-time adaptation: From the perspective of disentangled factors." *arXiv preprint arXiv:2403.07366*.

J. Lee, D. Das, J. Choo, and S. Choi. 2023. "Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 16380–16389.

J. Liang, R. He, and T. Tan. 2023. "A comprehensive survey on test-time adaptation under distribution shifts." *arXiv preprint arXiv:2303.15361*.

W. Liu, X. Wang, J. Owens, and Y. Li. 2020. "Energy-based out-of-distribution detection." *Advances in neural information processing systems*, 33, 21464–21475.

Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao. 2022. "Review the state-of-the-art technologies of semantic segmentation based on deep learning." *Neurocomputing*, 493, 626–646.

K. P. Murphy. 2022. *Probabilistic machine learning: an introduction*. MIT press.

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, et al.. 2011. "Reading digits in natural images with unsupervised feature learning." In: *NIPS workshop on deep learning and unsupervised feature learning* 2. Vol. 2011. Granada, 4.

K. X. Nguyen, F. Qiao, and X. Peng. 2024. "Adaptive Cascading Network for Continual Test-Time Adaptation." In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 1763–1773.

S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan. 2022. "Efficient test-time model adaptation without forgetting." In: *International conference on machine learning (ICML)*. PMLR, 16888–16905.

S. Niu, J. Wu, Y. Zhang, Z. Wen, Y. Chen, P. Zhao, and M. Tan. 2023. "Towards stable test-time adaptation in dynamic wild world." *arXiv preprint arXiv:2302.12400*.

J. Park, J. Kim, H. Kwon, I. Yoon, and K. Sohn. 2024. "Layer-wise auto-weighting for non-stationary test-time adaptation." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1414–1423.

A. Paszke et al.. 2019. "Pytorch: An imperative style, high-performance deep learning library." *Advances in neural information processing systems*, 32.

A. Radford et al.. 2021. "Learning transferable visual models from natural language supervision." In: *International conference on machine learning (ICML)*. PMLR, 8748–8763.

A. Smirnova, J. Yang, and P. Cudre-Mauroux. 2024. "XCrowd: Combining Explainability and Crowdsourcing to Diagnose Models in Relation Extraction." In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2097–2107.

J. Song, J. Lee, I. S. Kweon, and S. Choi. 2023. "Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11920–11929.

S. Stan and M. Rostami. 2024. "Preserving fairness in AI under domain shift." *Journal of Artificial Intelligence Research*, 81, 907–934.

T. Sun, M. Segu, J. Postels, Y. Wang, L. Van Gool, B. Schiele, F. Tombari, and F. Yu. 2022. "SHIFT: a synthetic driving dataset for continuous multi-task domain adaptation." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21371–21382.

S. Tahmasebi, E. Müller-Budack, and R. Ewerth. 2024. "Multimodal misinformation detection using large vision-language models." In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2189–2199.

A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis. 2018. "Deep learning for computer vision: A brief review." *Computational intelligence and neuroscience*, 2018, 1, 7068349.

D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. 2020. "Tent: Fully test-time adaptation by entropy minimization." *arXiv preprint arXiv:2006.10726*.

Q. Wang, O. Fink, L. Van Gool, and D. Dai. 2022. "Continual test-time domain adaptation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7201–7211.

X. Wang, J. Hu, H. Lin, W. Liu, H. Moon, and M. J. Piran. 2022. "Federated learning-empowered disease diagnosis mechanism in the internet of medical things: From the privacy-preservation perspective." *IEEE Transactions on Industrial Informatics*, 19, 7, 7905–7913.

Y. Wang, J. Hong, A. Cheraghian, S. Rahman, D. Ahmedt-Aristizabal, L. Petersson, and M. Harandi. 2024. "Continual test-time domain adaptation via dynamic sample selection." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1701–1710.

Z. Wang, Y. Zhao, X. Chen, and K. Zheng. 2021. "Task assignment with worker churn prediction in spatial crowdsourcing." In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2070–2079.

Z. Wang, Y. Luo, L. Zheng, Z. Chen, S. Wang, and Z. Huang. 2024. "In search of lost online test-time adaptation: A survey." *International Journal of Computer Vision*, 1–34.

K. Wong, S. Wang, M. Ren, M. Liang, and R. Urtasun. 2020. "Identifying unknown instances for autonomous driving." In: *Conference on Robot Learning*. PMLR, 384–393.

Z. Xiao and C. G. Snoek. 2024. "Beyond Model Adaptation at Test Time: A Survey." *arXiv preprint arXiv:2411.03687*.

Y. Yu, L. Sheng, R. He, and J. Liang. 2023. "Benchmarking test-time adaptation against distribution shifts in image classification." *arXiv preprint arXiv:2307.03133*.

Y. Yu, L. Sheng, R. He, and J. Liang. 2025. "STAMP: Outlier-Aware Test-Time Adaptation with Stable Memory Replay." In: *European Conference on Computer Vision (ECCV)*. Springer, 375–392.

L. Yuan, B. Xie, and S. Li. 2023. "Robust test-time adaptation in dynamic scenarios." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15922–15932.

S. Zagoruyko and N. Komodakis. 2016. "Wide residual networks." *arXiv preprint arXiv:1605.07146*.

J. Zhang, L. Qi, Y. Shi, and Y. Gao. 2023. "Domainadaptor: A novel approach to test-time adaptation." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 18971–18981.

M. Zhang, S. Levine, and C. Finn. 2022. "Memo: Test time robustness via adaptation and augmentation." *Advances in neural information processing systems (NeurIPS)*, 35, 38629–38642.

X. Zhong, H. Miao, D. Qiu, Y. Zhao, and K. Zheng. 2023. "Personalized location-preference learning for federated task assignment in spatial crowdsourcing." In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3534–3543.

Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye. 2023. "Object detection in 20 years: A survey." *Proceedings of the IEEE*, 111, 3, 257–276.

Received 10 December 2025; accepted 06 March 2026