RASA: Reliability-Aware Selective Adaptation via Hybrid Reliability Threshold

Ziqiong Liu, Yijia Wang, Junyang Ji, Yushun Tang, and Zhihai He
Southern University of Science and Technology
Shenzhen, China
Email: {12332162, wangyj2022, 12491002, tangys2022}@mail.sustech.edu.cn, hezh@sustech.edu.cn

***Abstract*—Test-time adaptation (TTA) has emerged as a highly promising paradigm for adapting pre-trained models to target domains. However, in open-world scenarios, standard TTA often suffers from noise overfitting due to indiscriminate entropy minimization, especially when faced with severe semantic shifts and out-of-distribution (OOD) samples. To address these challenges, we propose RASA, a Reliability-Aware Selective Adaptation framework that employs a hybrid reliability threshold for selective test-time adaptation. The core of RASA is a novel History-Aware Thresholding (HAT) module, which integrates hybrid prior information by combining theoretical class priors with running historical entropy statistics, and dynamically adjusts the reliability criteria as the adaptation process progresses. Furthermore, we introduce a dual-stream Transformer architecture that explicitly decouples semantic classification and OOD inference into different feature subspaces, significantly enhancing sensitivity to outliers while maintaining accuracy on in-distribution samples. Extensive experiments demonstrate that RASA consistently outperforms existing state-of-the-art methods, exhibiting superior stability and robustness in open-world environments with severe data perturbations and a high proportion of OOD samples.**

***Index Terms*—Test-time adaptation, Open-world recognition, Out-of-distribution detection, Reliability-aware learning, Noise-robustness, Dual-stream architecture.**

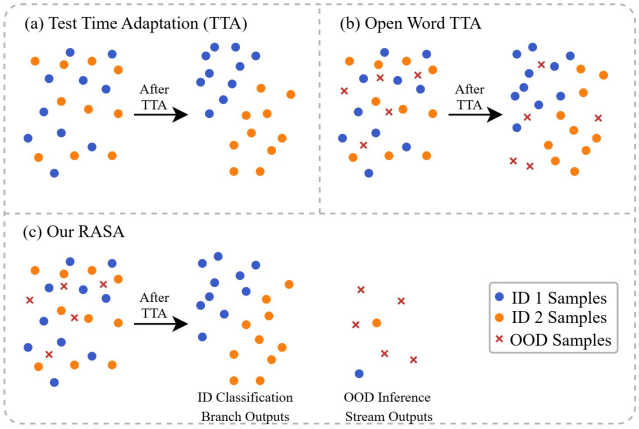

Fig. 1. Comparison of adaptation paradigms in different scenarios. **(a) Standard TTA** works effectively under the closed-set assumption. **(b) Open-World TTA** often leads to decision boundary corruption, as indiscriminate entropy minimization forces OOD samples (red crosses) into In-Distribution (ID) clusters. **(c) Our RASA** effectively decouples OOD detection from semantic classification, rejecting anomalies while maintaining high adaptation accuracy for ID data.

I. Introduction

In real-world deployment environments, pre-trained models inevitably encounter test data streams that deviate significantly from the training distribution (source domain). Excluding confounding factors, domain shifts [1] caused by fluctuating imaging conditions, sensor noise, style drift, and scene changes often lead to sharp degradation in model performance. To mitigate this degradation while adhering to data privacy constraints (i.e., without access to source domain data), Test-Time Adaptation (TTA) [2] has emerged as a vital paradigm. By dynamically fine-tuning the model using the online data stream during the testing phase, TTA enables real-time adaptation to the current distribution, maintaining high utility in complex tasks.

However, challenges in open-world scenarios extend far beyond simple distribution shifts; models frequently face interference from out-of-distribution (OOD) samples [3]. As illustrated in Fig. 1, while standard TTA (Fig. 1(a)) succeeds in closed sets, it fails in open-world settings (Fig. 1(b)) because it indiscriminately forces samples into known categories. These OOD samples, which do not belong to any predefined categories, are often incorrectly classified due to the closed nature of the prediction space. In TTA, this issue is exacerbated by the reliance on entropy minimization [2], [4]. While entropy minimization effectively compresses prediction uncertainty to capture target domain features, it exhibits a "blind certainty" in open-world settings. By indiscriminately forcing high-confidence predictions, the model becomes susceptible to poisoning by noisy pseudo-labels generated from anomalous samples. This results in severe gradient bias [5] and a negative feedback loop of "prediction deviation – erroneous update," ultimately leading to irreversible model collapse [6]. Therefore, accurately identifying and filtering erroneous information in dynamic data streams is crucial for robust TTA deployment.

Although existing methods, such as STAMP [3] and OWTTT [7], attempt to mitigate semantic shift through heuristic filtering, their effectiveness is inherently constrained by the underlying backbone networks. Most TTA research relies on CNN architectures (e.g., ResNet [8]), which possess strong local inductive biases but limited global receptive fields. While adequate for low-resolution benchmarks like CIFAR-10/100-C, these architectures often encounter bottlenecks in large-

scale, fine-grained tasks such as ImageNet-C. In contrast, the Vision Transformer (ViT) [9] models long-range dependencies via global self-attention, providing a holistic feature representation. This global context is highly advantageous in open-world TTA for decoupling stable semantic features from domain-specific noise [10]. Consequently, constructing a Transformer-based TTA framework capable of robustly handling semantic shifts has become a critical research imperative.

To address the severe semantic shift and noise over-fitting problems in open-world TTA, we propose **RASA**, a **Reliability-Aware Selective Adaptation** framework (Fig. 1(c)). RASA aims to mitigate the risks associated with indiscriminate entropy minimization by utilizing a noise-resistant, hybrid reliability threshold setting. Specifically, our main contributions are summarized as follows:

- We propose a noise-robust correction strategy utilizing a hybrid reliability threshold-assisted model update for controlling adaptive processes. Unlike conventional methods, our approach incorporates a security filter to prioritize high-confidence samples, effectively preventing model collapse caused by over-fitting to noisy outliers.
- We design a novel History-Aware Adaptive Threshold (HAT) module. By leveraging hybrid priors that integrate theoretical constraints with historical entropy statistics, this mechanism dynamically calibrates confidence boundaries, ensuring stable adaptation across evolving learning phases.
- We introduce a dual-stream Transformer architecture featuring specialized latent OOD projection layers. By explicitly decoupling semantic classification and out-of-distribution inference into distinct feature subspaces, our model significantly enhances sensitivity to domain shifts while preserving in-distribution accuracy.
- Extensive experiments indicate that RASA consistently outperforms state-of-the-art methods, exhibiting exceptional stability and robustness in open-world environments characterized by heavy data perturbations and high OOD ratios.

II. Related Work

This work is closely related to test-time adaptation (TTA) and out-of-distribution (OOD) detection. We review representative studies in these two areas, with an emphasis on open-world test-time learning settings.

A. Test-time Adaptation

Test-time adaptation (TTA) [2] updates a pretrained model online using unlabeled target data at inference time, typically without access to source data, making it suitable for deployment constrained by annotation cost and privacy [11]–[15]. Existing TTA methods are commonly categorized by adaptation mechanisms: normalization-based approaches recalibrate batch normalization (BN) statistics at test time or blend statistics across domains to improve robustness under covariate shifts with minimal parameter updates [16]–[18]; self-supervised objectives such as entropy minimization [2] and contrastive learning [19] directly refine representations on unlabeled test samples. However, self-training-style TTA can suffer from error accumulation due to unreliable predictions, motivating confidence-/entropy-guided sample selection and update control to suppress noisy gradients [5], [6]. Recent work further extends TTA to non-stationary environments, label shift, and temporally correlated streams, where additional stability mechanisms are required to prevent drift [4], [20]–[22].

Open-world TTA considers test streams containing unknown categories and OOD samples [7], [23]. OWTTT [7] targets adaptation with novel OOD instances, while OSTTA [23] studies jointly corrupted ID/OOD scenarios; AdaODD [24] explores non-parametric online adjustment of the OOD boundary under distribution overlap. Prototype expansion or contrastive objectives have been used to increase ID–OOD separability [3], [7], [25], though some rely on auxiliary source data. Source-free alternatives instead use entropy regularization or related constraints to separate ID from OOD during adaptation, including OSTTA [23], UniEnt [26], and AEO [27]. STAMP [3] combines self-weighted entropy optimization with memory replay, and ATTA [28] studies selective adaptation for dense prediction under domain and semantic shifts.

B. OOD Detection

Out-of-distribution (OOD) detection is critical for reliable deployment, as models that perform well on in-distribution samples can still produce over-confident but incorrect predictions on anomalous inputs [7], [29], [30]. A large body of post-hoc methods detects OOD samples by analyzing model outputs without modifying the underlying architecture or training procedure, including maximum softmax probability (MSP) [31], ODIN [32] and generalized variants [33], energy-based scoring [34], and Mahalanobis distance-based measures [35]. EOOD [36] further improves robustness by selecting network blocks with pronounced information-flow differences and employing conditional entropy as an OOD indicator. Moreover, incorporating auxiliary outliers or additional data beyond the training distribution has been shown to enhance OOD generalization and robustness to unseen inputs in certain settings [37], [38]. Relatedly, STMS [39] explores OOD generalization under model-stealing scenarios via causal inference to handle distribution shifts.

OOD detection has also expanded beyond conventional discriminative scoring to include generative and multimodal paradigms. Early efforts exploit reconstruction-based discrepancies to identify OOD samples [40]. Diffusion-based approaches further leverage differences between original and reconstructed images [41] and trajectory dynamics (e.g., path change rate and curvature) for detecting OOD inputs [42], [43]. In addition, vision-language modeling introduces complementary signals for OOD detection: MCM [44] jointly models image-text representations, while negative prompts can strengthen separability across diverse scenarios [45]–[47]. OLE [48] uses auxiliary outlier class labels as pseudo-OOD prompts to mitigate overconfidence in zero-shot settings. Fi-

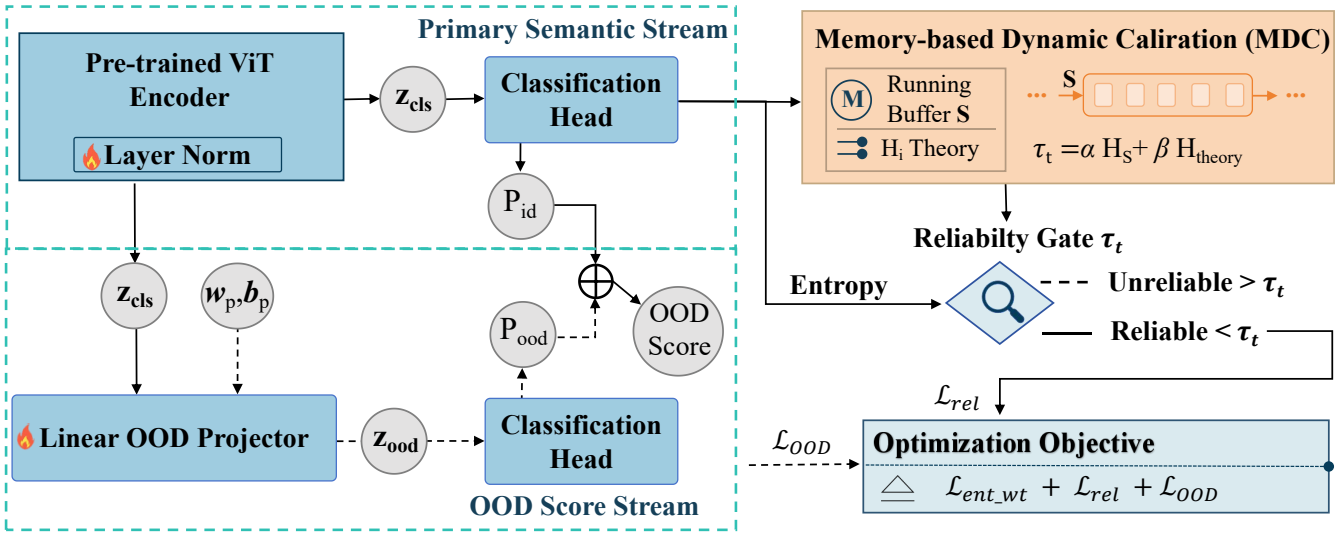

Fig. 2. **Overview of the RASA Framework.** The architecture illustrates the **Decoupled Dual-Stream** mechanism where the input is processed through a shared backbone. (Left) The **Primary Semantic Stream** handles classification using **Dynamic Reliability-Aware Entropy Filtering** to select high-confidence samples. (Right) The **OOD Inference Stream** utilizes a **Latent OOD Projection Layer** to map features into a specialized subspace for anomaly detection. The joint optimization objective ensures that semantic updates do not interfere with distributional sensitivity.

nally, dynamic dictionary-based methods such as OODD [49] maintain representative OOD features online to calibrate OOD scores and improve stability under streaming test conditions.

III. Method

First, we define the problem setup for Test-Time Adaptation (TTA) and Out-of-Distribution (OOD) detection in Sec. III-A. The methodological procedures illustrated in Fig. 2 are detailed across Sec. III-B and Sec. III-C, followed by a summary of our joint optimization and update strategy.

A. Problem Setup: TTA and OOD Detection

In the context of Test-Time Adaptation (TTA), we consider a model f_θ pre-trained on a source domain $\mathcal{D}_s$. At test time, the model encounters a continuous stream of unlabeled data $\mathcal{X}_t = \{x_{t,1}, x_{t,2}, \ldots, x_{t,n}\}$ from a target domain $\mathcal{D}_t$, where the distribution $P_t(x)$ shifts from the source distribution $P_s(x)$. The primary objective of TTA is to adapt the model parameters θ to the target domain $\mathcal{D}_t$ in an online manner.

However, in real-world deployments, the data stream $\mathcal{X}_t$ often contains Out-of-Distribution (OOD) samples—instances that do not belong to any of the C semantic categories defined in $\mathcal{D}_s$. Formally, for any input $x \in \mathcal{X}_t$, the model is required to simultaneously achieve two goals:

1) **Correct Classification:** Assigning the accurate label $y \in \{1, \ldots, C\}$ if x is an In-Distribution (ID) sample.
2) **Reliable Detection:** Identifying whether x belongs to the OOD set $\mathcal{X}_{ood}$, where $x \notin \mathcal{D}_s$.

The absence of ground-truth labels poses a significant challenge: if the model incorrectly treats OOD noise as reliable ID data for self-training, it suffers from *confirmation bias*, leading to semantic collapse and disorganized feature representations. Our proposed framework, **RASA**, addresses this by synergizing structural decoupling with dynamic reliability filtering.

B. Dynamic Reliability-Aware Entropy Filtering

Conventional Test-Time Adaptation (TTA) methods, such as Tent [2], typically rely on a fixed entropy threshold to determine which samples contribute to the model update. However, we observe a significant *"entropy shift"* phenomenon during the continuous adaptation process. In the initial stages, the model's uncertainty is naturally elevated due to the distribution gap between the source and target domains. As the model gradually adapts, the overall entropy distribution tends to shift toward lower values, reflecting increased prediction confidence.

A rigid, fixed threshold fails to capture this temporal evolution, leading to either the inclusion of excessive noise in the initial phase or over-conservative updates in later stages. To address this, we introduce a **Memory-based Dynamic Calibration (MDC)** module. Instead of maintaining a static boundary, RASA maintains a running buffer $\mathcal{S}$ of historical entropy statistics. To maintain computational efficiency while adapting to the current data distribution, we utilize the mean of stored entropy values to extract the characteristics of past samples. Specifically, at each time step t, the MDC module dynamically updates the confidence boundary τ_t as:

$$\tau_t = \alpha \cdot \bar{H}_{\mathcal{S}} + \beta \cdot H_{\text{theory}}, \tag{1}$$

where $\bar{H}_{\mathcal{S}}$ is the running mean entropy in the buffer, and α, β balance empirical statistics and theoretical priors. We define $H_{\text{theory}} = \ln(C)$, the maximum entropy under a uniform distribution over C classes, and set $\beta = 0.2$. This yields a conservative confidence boundary that filters out high-uncertainty

samples, retaining only reliable in-distribution predictions for adaptation.

To ensure numerical stability and prevent the model from drifting into degenerate solutions, such as collapsing to a single class, this hybrid prior effectively fuses empirical historical statistics with theoretical entropy bounds. This mechanism ensures that the selective adaptation process remains both responsive to distribution shifts and grounded in reliable predictions, preventing the boundary from deviating excessively from theoretical expectations.

C. Decoupled Dual-Stream Architecture with Latent OOD Projection

To establish a more precise reliability metric during Test-Time Adaptation (TTA), we propose a decoupled dual-stream Transformer architecture designed to explicitly separate semantic classification and Out-of-Distribution (OOD) inference within the feature subspaces. In unsupervised TTA scenarios, models typically rely on the relative reliability of predictions for self-updating. Traditional single-stream architectures (e.g., STAMP [3]) often utilize a single *class token* to handle both category discrimination and anomaly detection simultaneously. This high degree of coupling inevitably leads to a fundamental conflict between maintaining semantic invariance and enhancing anomaly sensitivity, which may result in a disorganized data distribution after model updates.

Based on our empirical observations of feature visualizations across different Vision Transformer layers in Fig. 3, we identified that the discriminative potential for distinguishing In-Distribution (ID) and OOD samples is significantly pronounced in the deep-layer class tokens. Driven by this insight, we introduce a dedicated **Latent OOD Projection Layer** to construct an independent inference flow. Specifically, the model extracts the global contextual embedding $\mathbf{z}_{cls} \in \mathbb{R}^d$ from the final backbone block and feeds it into bifurcated paths:

(i) Primary Semantic Stream: This path follows the standard discriminative pipeline to ensure precision on ID samples. Its predicted probability distribution P_{ID} is computed via the classification head $h_\omega(\cdot)$ as:

$$P_{ID} = \text{Softmax}(h_\omega(\text{LN}(\mathbf{z}_{cls}))), \tag{2}$$

where $\text{LN}(\cdot)$ denotes Layer Normalization, and h_ω is the fully connected layer mapping to C categories.

(ii) OOD Inference Stream: This stream employs a learnable linear projection operator $\phi(\cdot)$ to re-project the high-dimensional semantic features into a specialized latent distributional space, generating a dedicated OOD token $\mathbf{z}_{OOD}$. This projection process is formulated as:

$$\mathbf{z}_{OOD} = \phi(\mathbf{z}_{cls}) = \mathbf{W}_p\mathbf{z}_{cls} + \mathbf{b}_p, \tag{3}$$

where $\mathbf{W}_p \in \mathbb{R}^{d\times d}$ and $\mathbf{b}_p \in \mathbb{R}^d$ denote the learnable weight and bias of the projection layer, respectively. Subsequently, this token generates an auxiliary distribution P_{OOD} through the same classification head:

$$P_{OOD} = \text{Softmax}(h_\omega(\mathbf{z}_{OOD})). \tag{4}$$

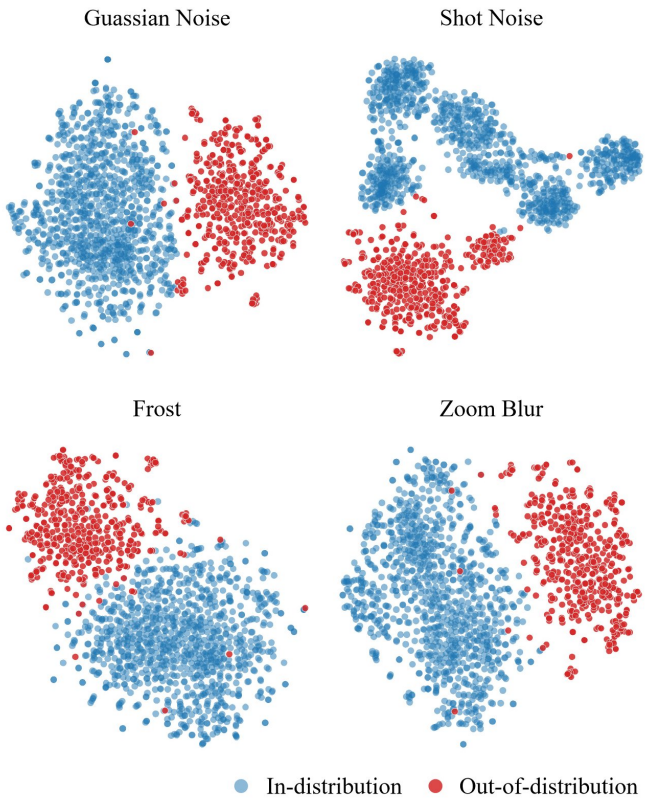

Fig. 3. **t-SNE visualization of deep-layer tokens under ImageNet-C corruptions.** The clear separation between In-Distribution (blue) and Out-of-Distribution (red) features across four typical corruptions validates the significant discriminative potential of deep-layer class tokens, providing the empirical basis for our decoupled architecture.

Final Prediction Fusion and Correction Mechanism: To synergistically leverage the dual-stream features and mitigate gradient bias introduced by anomalous samples, we define the final fused prediction P_{final} as:

$$P_{\text{final}} = \frac{1}{2}\left(P_{\text{ID}} + P_{\text{OOD}}\right). \tag{5}$$

This formulation corresponds to an equal-weight averaging strategy, which balances the discriminative capability of the ID branch and the sensitivity of the OOD branch. Meanwhile, it ensures that the OOD detection process does not adversely affect the semantic classification performance of in-distribution samples.

D. Joint Loss Function

At each time step t, the total optimization objective $\mathcal{L}_{total}$ is designed to simultaneously enhance the model's discriminative capability and suppress noise induced by OOD samples. We formulate the joint loss function as:

$$\mathcal{L}_{total} = \omega_1\mathcal{L}_{ent\_wt} + \omega_2\mathcal{L}_{ood} + \omega_3\mathcal{L}_{rel}, \tag{6}$$

where the first term, $\mathcal{L}_{ent\_weighted}$, follows the strategy-weighting mechanism in STAMP [3], reweighting samples based on their prediction confidence to achieve stable

TABLE I

PERFORMANCE COMPARISON OF CLASSIFICATION ACCURACY (%), AUC (%), AND H-SCORE (%) ON IMAGENET-C (LEVEL 5) WITH FOUR OOD TYPES. BEST RESULTS ARE IN **BOLD**.

Method	Textures			Places			iNaturalist			SUN			Avg.		
	ACC	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*
Source	29.9	64.9	39.6	29.9	34.9	32.1	29.9	54.1	38.3	29.9	38.2	33.4	29.9	48.0	35.9
Tent	40.9	60.7	46.9	41.5	55.2	45.2	40.8	72.9	50.4	41.0	58.4	46.0	41.0	61.8	47.1
CoTTA	54.5	62.5	57.8	53.1	59.9	53.6	51.7	69.7	57.4	53.6	50.1	54.0	53.2	60.5	55.7
SoTTA	56.6	69.0	59.6	56.5	66.8	58.6	59.9	80.3	68.5	56.8	68.6	59.7	57.4	71.2	61.6
SAR	64.0	70.8	67.2	62.7	68.4	65.2	58.8	49.1	52.1	60.4	67.7	63.3	61.5	64.0	62.0
OSTTA	39.4	60.0	45.0	39.0	59.0	44.5	39.1	58.7	44.4	39.1	57.9	44.1	39.2	58.9	43.9
UniEnt	37.9	51.5	43.0	37.3	42.8	39.6	38.1	66.3	47.9	37.3	44.6	40.4	37.7	51.3	42.7
STAMP	59.8	72.6	64.1	62.8	65.4	64.0	63.8	88.9	73.7	62.6	68.8	65.6	62.0	73.9	66.8
Ours	**65.0**	**81.7**	**72.3**	**65.0**	**74.4**	**69.4**	**65.2**	**96.6**	**77.6**	**64.9**	**77.4**	**70.6**	**65.0**	**82.5**	**72.5**
	±0.0	±0.1	±0.0	±0.3	±0.2	±0.1	±0.2	±0.1	±0.2	±0.1	±0.2	±0.1	±0.2	±0.2	±0.1

TABLE II

EXPERIMENTAL RESULTS ON IMAGENET-R AND VISDA-2021 DATASETS, DEMONSTRATING MODEL GENERALIZATION UNDER SEVERE STYLE SHIFTS AND SYNTHETIC-TO-REAL DOMAIN GAPS. BEST RESULTS ARE IN **BOLD**.

Method	Textures			Places			iNaturalist			SUN			Avg.		
	ACC	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*
ImageNet-R															
Source	42.7	58.2	49.2	42.7	52.5	47.1	42.7	60.5	50.1	42.7	55.7	48.3	42.7	46.7	48.7
SAR	60.8	69.3	64.8	61.5	68.8	64.9	55.4	62.8	58.8	54.9	62.1	58.3	58.1	65.8	61.7
STAMP	59.3	73.3	65.5	59.3	66.5	62.7	59.6	73.4	65.8	59.4	71.2	64.8	59.4	71.1	64.7
Ours	**62.0**	**77.7**	**69.0**	**62.1**	**72.1**	**66.7**	**62.1**	**79.5**	**69.7**	**61.9**	**78.2**	**69.1**	**63.1**	**80.5**	**70.7**
	±0.0	±0.3	±0.1	±0.0	±0.1	±0.0	±0.2	±0.3	±0.0	±0.2	±0.5	±0.3	±0.1	±0.3	±0.1
VISDA-2021															
Source	44.3	59.0	50.6	44.3	51.1	47.4	44.3	72.3	54.9	44.3	56.5	49.7	44.3	59.7	50.6
SAR	50.7	63.3	56.3	50.5	55.8	53.1	51.0	78.7	61.9	50.6	61.7	55.4	50.7	64.9	56.6
STAMP	57.2	66.9	61.7	57.4	**66.1**	**61.4**	57.1	83.8	67.9	57.1	59.1	58.1	57.2	69.0	62.3
Ours	**58.6**	**69.6**	**63.6**	**58.6**	58.8	58.7	**58.7**	**86.9**	**70.1**	**58.6**	**67.0**	**62.5**	**58.5**	**70.6**	**63.7**
	±0.3	±0.3	±0.3	±0.2	±0.3	±0.2	±0.1	±0.0	±0.0	±0.3	±0.2	±0.1	±0.1	±0.2	±0.1

global distribution alignment. The second term, $\mathcal{L}_{ood}$, suppresses the uncertainty of detected OOD samples, enhancing discriminability. Finally, the loss for reliable samples obtained via a hybrid reliability threshold, $\mathcal{L}_{reliable}$, is defined as $\bar{H}(\mathcal{B}_{reliable})$, representing the mean entropy of high-confidence ID samples filtered by the dynamic threshold τ_t. By assigning different weights to balance these components, the RASA framework effectively prioritizes learning from reliable samples while maintaining robustness against non-stationary noise in the test-time data stream.

IV. EXPERIMENTS

A. Experimental Setup

Datasets. To comprehensively evaluate the effectiveness of RASA, we conduct experiments under three representative types of domain shift. For in-distribution robustness, we adopt ImageNet-C [1] as the primary benchmark, which contains 15 corruption types with five severity levels. To evaluate out-of-distribution (OOD) detection performance under domain shift, we use four commonly adopted OOD datasets, including Textures [50], Places [51], iNaturalist [52], and SUN [53]. In addition, following the protocol of STAMP [3], we construct Texture-C and Places-C to simulate more challenging open-world scenarios with environmental domain shifts. Furthermore, ImageNet-R [54] is employed to assess semantic robustness and generalization under severe style domain shifts, consisting of approximately 30,000 non-photorealistic images. Finally, we extend the evaluation to the large-scale VisDA-2021 benchmark [55] to validate the robustness of RASA under substantial domain shifts between synthetic and real-world data.

Baselines. We evaluate RASA by comparing it with both a non-adaptive source model and a set of representative online test-time adaptation (TTA) methods. Specifically, the source model without test-time adaptation is used as a reference baseline. In addition, we compare our approach with several recent state-of-the-art TTA methods, including TENT [2], CoTTA [4], SoTTA [56], SAR [6], OSTTA [23], UniEnt [26], and STAMP [3].

Evaluation Metrics. To comprehensively evaluate the model performance, we employ three key metrics. First, classification Accuracy (ACC) is used to measure the recognition capability

TABLE III
DETAILED CLASSIFICATION ACCURACY (%) FOR INDIVIDUAL CORRUPTION TYPES IN IMAGENET-R WITH VIT-L/16 BACKBONE AT SEVERITY LEVEL 5. BEST RESULTS ARE IN **BOLD**.

Method	Textures			Places			iNaturalist			SUN			Avg.		
	ACC	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*
Source	61.5	73.0	66.8	61.5	68.5	64.8	61.5	79.5	69.4	61.5	70.4	65.6	61.5	72.8	66.6
SAR	62.9	74.1	68.0	62.9	69.5	66.0	62.9	71.4	66.9	63.1	**79.5**	70.3	62.9	73.6	67.8
STAMP	65.4	76.5	70.5	65.5	70.5	67.9	65.4	80.3	72.1	65.4	72.1	68.6	65.4	74.8	69.8
Ours	**67.0**	**79.1**	**72.5**	**66.6**	**73.7**	**70.0**	**67.3**	**81.4**	**73.7**	**67.4**	**78.6**	72.6	**67.1**	**78.2**	**72.2**
	±0.7	±0.8	±0.8	±0.8	±0.8	±0.7	±0.4	±0.2	±0.3	±0.1	±0.5	±0.2	±0.5	±0.6	±0.5

TABLE IV
EVALUATION OF ROBUSTNESS UNDER DUAL DISTRIBUTION SHIFTS (ENVIRONMENTAL CORRUPTIONS AND SEMANTIC SHIFTS) USING TEXTURES-C AND PLACES-C DATASETS. BEST RESULTS ARE IN **BOLD**.

Method	Textures-C			Places-C			Avg.		
	ACC	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*	*ACC*	*AUC*	*H-score*
Source	29.9	64.9	39.6	29.9	34.9	32.1	29.9	53.9	37.0
Tent	40.1	67.7	48.9	39.1	62.5	46.6	39.6	65.1	47.7
CoTTA	52.9	67.9	57.9	54.7	65.6	59.4	53.8	66.7	58.6
SoTTA	56.2	69.8	60.7	56.5	68.1	60.7	56.4	69.0	60.7
SAR	57.6	65.3	60.9	57.7	67.6	62.0	57.6	66.4	61.4
OSTTA	39.1	55.8	44.2	37.8	56.5	43.5	38.4	56.2	43.9
UniEnt	38.2	69.4	48.1	37.8	67.8	47.0	38.0	68.6	47.6
STAMP	59.0	79.0	65.8	63.1	76.0	68.8	61.1	77.5	67.3
Ours	**64.5**	**83.4**	**72.6**	**65.2**	**77.5**	**70.8**	**64.8**	**80.5**	**71.7**
	±0.2	±0.4	±0.3	±0.2	±0.1	±0.1	±0.2	±0.2	±0.2

TABLE V
ABLATION STUDY OF THE PROPOSED RASA FRAMEWORK ON IMAGENET-C (ID) AND TEXTURES (OOD) DATASETS. HAT: HISTORY-AWARE THRESHOLDING; DSD: DUAL-STREAM DECOUPLING.

HAT	DSD	*ACC*	*AUC*	*H-score*
–	–	64.1	74.3	68.8
✓	–	64.9	74.3	69.2
–	✓	64.4	81.3	69.2
✓	✓	**65.0**	**81.7**	**72.3**

on in-distribution (ID) samples. Second, the Area Under the Receiver Operating Characteristic curve (AUROC, abbreviated as AUC in our tables) is adopted to assess the efficacy of distinguishing ID samples from out-of-distribution (OOD) instances, providing a threshold-independent evaluation of OOD detection reliability. Finally, to balance the accuracy of ID adaptation and the sensitivity of OOD detection, we incorporate the H-score. As formulated in Eq. (7), the H-score is defined as the harmonic mean of ACC and AUC:

$$Acc_H = 2 \cdot \frac{ACC \cdot AUC}{ACC + AUC}. \tag{7}$$

A high Acc_H indicates that the model achieves a superior trade-off between maintaining classification precision and identifying unknown samples, effectively preventing biased performance toward either task.

B. Implementation Details

We employ ViT-B/16 as the backbone with a batch size of 32. For a fair comparison, all methods utilize the same pre-trained parameters and architecture. Following SAR [6], we use the SGD optimizer and perform two backward passes per iteration to simulate the online adaptation process. The learning rates for LayerNorm parameters are set to 0.01, 0.01, and 0.1 for ImageNet-C, ImageNet-R, and VisDA-2021, respectively. For the larger ViT-L model, a learning rate of 0.001 is adopted. The loss weights in Sec. III-D are configured as $\omega_1 = 1.0$, $\omega_2 = 0.1$, and $\omega_3 = 0.1$. We report the mean and standard deviation across three independent runs with random seeds $\{2024, 2025, 2026\}$. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

C. Main Results

We systematically evaluate the performance of RASA against several state-of-the-art Test-Time Adaptation (TTA) methods across diverse and challenging open-world scenarios. The experimental results validate the superiority of our framework in terms of both in-distribution (ID) accuracy and out-of-distribution (OOD) detection reliability.

a) Performance on ImageNet-C: Table I summarizes the classification performance on ImageNet-C under the highest severity (Level 5) corruptions. Our RASA framework consistently outperforms all competitive TTA baselines across four diverse OOD datasets (Textures, Places, iNaturalist, and SUN). Notably, compared to the second-best method STAMP, RASA achieves a significant improvement in the average H-score, increasing it from 66.8% to 72.5%. Specifically, on the iNaturalist dataset, RASA achieves a remarkable AUC of 96.6%, which is 7.7% higher than STAMP. These results demonstrate that RASA maintains robust feature alignment and superior anomaly sensitivity even under extreme environmental degradations.

b) Robustness against Complex Environmental Perturbations: To further assess the stability in more realistic open-world settings, we evaluate the performance on Textures-C and Places-C, where OOD samples are subjected to severe environmental shifts. As shown in Table IV, RASA consistently achieves the best performance across all evaluation metrics. Compared to STAMP, our method improves the average H-score from 67.3% to 71.7%. On the Textures-C dataset, RASA reaches a leading AUC of 83.4%. These results confirm that by decoupling semantic and OOD inference streams, RASA

TABLE VI
EFFICIENCY COMPARISON OF RASA WITH VIT-B. FLOPS ARE MEASURED FOR A SINGLE FORWARD PASS ON A 224×224 IMAGENET-C IMAGE.

Method	Params (M)	ΔP (%)	FLOPs (G)	ΔF (%)
ViT-B	86.00	0	17.600	0
RASA	**86.59**	**+0.69**	**17.601**	**+0.007**

can effectively mitigate the negative impact of dual distribution shifts (semantic shift and image corruption).

c) Generalization on ImageNet-R and VisDA-2021: We also extend our evaluation to ImageNet-R and the large-scale VisDA-2021 benchmark to verify the generalization capability under severe style shifts and synthetic-to-real domain gaps. The results in Table II and Table III show that RASA maintains a clear performance lead. On ImageNet-R, RASA achieves an average H-score of 66.6%, significantly outperforming SAR and STAMP. Similarly, in the VisDA-2021 scenario, RASA demonstrates exceptional stability, further validating the effectiveness of our Reliability-Aware Selective Adaptation strategy in handling large-scale and high-resolution domain transfers.

D. Ablation Study

To investigate the individual contribution of each component within RASA, we conduct a comprehensive ablation study, with results summarized in Table V. Starting from the baseline (first row), the independent integration of the **HAT** module improves the Accuracy (ACC) from 64.1% to 64.9%. This improvement validates the effectiveness of our history-aware dynamic reliability entropy filtering, confirming that HAT effectively mitigates noise overfitting during the adaptation process. Simultaneously, the **DSD** architecture leads to a significant leap in AUC (from 74.3% to 81.3%), underlining the necessity of decoupling semantic feature subspaces from OOD feature subspaces. Ultimately, by synergizing both modules, the full RASA framework achieves the optimal **H-score of 72.3%**. This demonstrates that RASA successfully balances in-distribution (ID) adaptation with out-of-distribution (OOD) sensitivity.

E. Computational Efficiency Analysis

To evaluate the efficiency of RASA, we compare its computational cost and parameter overhead with the vanilla ViT-B backbone. As shown in Table VI, FLOPs are reported based on a single forward pass for fair comparison. Although the proposed dual-stream architecture and HAT module are introduced for open-world TTA, they incur only marginal overhead, increasing parameters by 0.69% (86.00M to 86.59M) and FLOPs by 0.007% (17.6G to 17.601G). These results demonstrate that RASA achieves strong performance gains with negligible computational cost, making it suitable for resource-efficient test-time adaptation.

V. CONCLUSION

In this work, we proposed the RASA framework to address the critical challenge of noise overfitting in open-world Test-Time Adaptation (TTA). By integrating a History-Aware Adaptive Threshold (HAT) for dynamic reliability filtering and a Decoupled Dual-Stream Architecture that explicitly separates semantic classification from OOD inference, RASA effectively prevents model collapse under severe distribution shifts. Extensive experiments on ImageNet-C, ImageNet-R, and VisDA-2021 benchmarks demonstrate that our method significantly outperforms state-of-the-art approaches, achieving a superior trade-off between in-distribution accuracy and out-of-distribution detection reliability.

VI. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 62331014 and No. 12426312). We acknowledge the computational support of the Center for Computational Science and Engineering at Southern University of Science and Technology.

REFERENCES

[1] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations (ICLR)*, 2019.

[2] D. Wang, E. Shelhamer, J. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *International Conference on Learning Representations*, 2021.

[3] Y. Yu, L. Sheng, R. He, and J. Liang, "Stamp: Outlier-aware test-time adaptation with stable memory replay," in *European Conference on Computer Vision*. Springer, 2024, pp. 375–392.

[4] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7201–7211.

[5] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan, "Efficient test-time model adaptation without forgetting," in *International conference on machine learning*. PMLR, 2022, pp. 16 888–16 905.

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

[7] Y. Li, X. Xu, Y. Su, and K. Jia, "On the robustness of open-world test-time training: Self-training with dynamic prototype expansion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 836–11 846.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[9] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[10] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in neural information processing systems*, vol. 34, pp. 12 116–12 128, 2021.

[11] Y. Ding, J. Liang, B. Jiang, A. Zheng, and R. He, "Maps: A noise-robust progressive learning approach for source-free domain adaptive keypoint detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1376–1387, 2023.

[12] Y. Ding, L. Sheng, J. Liang, A. Zheng, and R. He, "Proxymix: Proxy-based mixup training with label refinery for source-free domain adaptation," *Neural Networks*, vol. 167, pp. 92–103, 2023.

[13] L. Sheng, J. Liang, R. He, Z. Wang, and T. Tan, "Adaptguard: Defending against universal attacks for model adaptation," *arXiv preprint arXiv:2303.10594*, 2023.

[14] Y. Tang, S. Chen, Z. Lu, X. Wang, and Z. He, "Dual-path adversarial lifting for domain shift correction in online test-time adaptation," in *European Conference on Computer Vision*. Springer, 2024, pp. 342–359.

[15] Y. Tang, S. Chen, Z. Kan, Y. Zhang, Q. Guo, and Z. He, "Learning visual conditioning tokens to correct domain shift for fully test-time adaptation," *IEEE Transactions on Multimedia*, 2024.

[16] Z. Nado, S. Padhy, D. Sculley, A. D'Amour, B. Lakshminarayanan, and J. Snoek, "Evaluating prediction-time batch normalization for robustness under covariate shift," *arXiv preprint arXiv:2006.10963*, 2020.

[17] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge, "Improving robustness against common corruptions by covariate shift adaptation," *Advances in neural information processing systems*, vol. 33, pp. 11 539–11 551, 2020.

[18] L. Yuan, B. Xie, and S. Li, "Robust test-time adaptation in dynamic scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 922–15 932.

[19] D. Chen, D. Wang, T. Darrell, and S. Ebrahimi, "Contrastive test-time adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 295–305.

[20] G. Chakrabarty, M. Sreenivas, and S. Biswas, "Santa: Source anchoring network and target alignment for continual test time adaptation," *Transactions on Machine Learning Research*, 2023.

[21] Z. Zhou, L.-Z. Guo, L.-H. Jia, D. Zhang, and Y.-F. Li, "Ods: Test-time adaptation in the presence of open-world data shift," in *International Conference on Machine Learning*. PMLR, 2023, pp. 42 574–42 588.

[22] H. Zhao, Y. Liu, A. Alahi, and T. Lin, "On pitfalls of test-time adaptation," *arXiv preprint arXiv:2306.03536*, 2023.

[23] J. Lee, D. Das, J. Choo, and S. Choi, "Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 380–16 389.

[24] Y. Zhang, X. Wang, T. Zhou, K. Yuan, Z. Zhang, L. Wang, and R. Jin, "Model-free test time adaptation for out-of-distribution detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[25] H. Su, M. Wang, J. Li, B. Wang, D. Liu, and Z. Wang, "Open-world test-time training: Self-training with contrast learning," *arXiv preprint arXiv:2409.09591*, 2024.

[26] Z. Gao, X.-Y. Zhang, and C.-L. Liu, "Unified entropy optimization for open-set test-time adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 975–23 984.

[27] H. Dong, E. Chatzi, and O. Fink, "Towards robust multimodal open-set test-time adaptation via adaptive entropy-aware optimization," *arXiv preprint arXiv:2501.13924*, 2025.

[28] Z. Gao, S. Yan, and X. He, "Atta: Anomaly-aware test-time adaptation for out-of-distribution detection in segmentation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 45 150–45 171, 2023.

[29] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5635–5662, 2024.

[30] A. Filos, P. Tigas, R. McAllister, N. Rhinehart, S. Levine, and Y. Gal, "Can autonomous vehicles identify, recover from, and adapt to distribution shifts?" in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 3145–3153.

[31] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[32] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017.

[33] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 951–10 960.

[34] Z. Lin, S. D. Roy, and Y. Li, "Mood: Multi-level out-of-distribution detection," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 313–15 323.

[35] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in neural information processing systems*, vol. 31, 2018.

[36] G. Yang, C. Hou, W. Peng, X. Fang, Y. Nie, P. Zhu, and K. Tang, "Eood: Entropy-based out-of-distribution detection," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.

[37] S. Park, J. Mok, D. Jung, S. Lee, and S. Yoon, "On the powerfulness of textual outlier exposure for visual ood detection," *Advances in Neural Information Processing Systems*, vol. 36, pp. 51 675–51 687, 2023.

[38] J. Zhang, N. Inkawhich, R. Linderman, Y. Chen, and H. Li, "Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5531–5540.

[39] Y. Yang, X. Chen, Z. Zhao, Y. Xuan, B. Tang, and X. Li, "Stms: An out-of-distribution model stealing method based on causality," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[40] M. S. Graham, W. H. Pinaya, P.-D. Tudosiu, P. Nachev, S. Ourselin, and J. Cardoso, "Denoising diffusion models for out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2948–2957.

[41] Z. Liu, J. P. Zhou, Y. Wang, and K. Q. Weinberger, "Unsupervised out-of-distribution detection with diffusion inpainting," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22 528–22 538.

[42] A. Heng, H. Soh *et al.*, "Out-of-distribution detection with a single unconditional diffusion model," *Advances in Neural Information Processing Systems*, vol. 37, pp. 43 952–43 974, 2024.

[43] R. Gao, C. Zhao, L. Hong, and Q. Xu, "Diffguard: Semantic mismatch-guided out-of-distribution detection using pre-trained diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1579–1589.

[44] Y. Ming, Z. Cai, J. Gu, Y. Sun, W. Li, and Y. Li, "Delving into out-of-distribution detection with vision-language representations," *Advances in neural information processing systems*, vol. 35, pp. 35 087–35 102, 2022.

[45] T. Li, G. Pang, X. Bai, W. Miao, and J. Zheng, "Learning transferable negative prompts for out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 584–17 594.

[46] J. Nie, Y. Zhang, Z. Fang, T. Liu, B. Han, and X. Tian, "Out-of-distribution detection with negative prompts," in *The twelfth international conference on learning representations*, 2024.

[47] H. Wang, Y. Li, H. Yao, and X. Li, "Clipn for zero-shot ood detection: Teaching clip to say no," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1802–1812.

[48] C. Ding and G. Pang, "Zero-shot out-of-distribution detection with outlier label exposure," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[49] Y. Yang, L. Zhu, Z. Sun, H. Liu, Q. Gu, and N. Ye, "Oodd: Test-time out-of-distribution detection with dynamic dictionary," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 30 630–30 639.

[50] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3606–3613.

[51] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[52] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8769–8778.

[53] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3485–3492.

[54] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8340–8349.

[55] D. Bashkirova, D. Hendrycks, D. Kim, H. Liao, S. Mishra, C. Rajagopalan, K. Saenko, K. Saito, B. U. Tayyab, P. Teterwak *et al.*, "Visda-2021 competition: Universal domain adaptation to improve performance on out-of-distribution data," in *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 66–79.

[56] T. Gong, Y. Kim, T. Lee, S. Chottananurak, and S.-J. Lee, "Sotta: Robust test-time adaptation on noisy data streams," *Advances in Neural Information Processing Systems*, vol. 36, pp. 14 070–14 093, 2023.