SELF-EVOLVING EXPERTS FOR CONTINUAL TEST-TIME ADAPTATION IN AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image detectors are typically trained once and kept static after deployment, making them vulnerable to emerging generators with unseen forensic traces. While continual test-time adaptation (CTTA) is a promising direction, effective self-training requires reliable pseudo-supervision and stable knowledge consolidation across distributions. We propose SEED (Self-Evolving Experts for Detection), a framework that enables pretrained detectors to adapt continually via a dynamically managed pool of residual experts, in the absence of ground-truth labels, generator identities, or domain boundaries. To this end, SEED dynamically manages its expert pool through shift detection for expert creation and decision-orthogonal expert routing. It creates a new expert only when the detector's scores show a persistent change, while using decision-orthogonal features to reuse existing experts when a context recurs. Each expert accumulates feature statistics weighted by the reliability of its pseudo-labels, and updates its residual head through balanced feature replay. Across four public benchmarks, SEED outperforms SOTA methods and improves final accuracy by up to 17.42 percentage points over the frozen detector without CTTA. *All code and data will be publicly available upon acceptance.*

1 INTRODUCTION

Recent AI-generated image detectors generalize across generators by exploiting generalizable representations, local artifacts, and frequency cues (Wang et al., 2020; Ojha et al., 2023; Tan et al., 2024; Durall et al., 2020). Yet most detectors are trained on a dataset and remain static after deployment. As generative models evolve, generated images may exhibit forensic traces absent from the training data (Ojha et al., 2023; Park & Owens, 2025). A static detector cannot keep pace with this evolving environment, motivating continual test-time adaptation (CTTA), in which a detector learns continuously from an unlabeled data stream while retaining previously acquired knowledge.

Most existing CTTA benchmarks consist of test streams with changing image corruptions, while the recognition categories remain unchanged (Wang et al., 2022a; Yuan et al., 2023). In AI-generated image detection, the real/fake label space also remains fixed, but changes in synthesis mechanisms can alter the underlying forensic cues used to distinguish generated images from real ones (Ojha et al., 2023). Learning these unseen forensic signatures through self-training calls for reliable pseudo-supervision to guide adaptation (Nguyen-Le et al., 2025). Continual adaptation further requires stable knowledge consolidation to retain previously acquired forensic knowledge as new generators emerge.

Current literature addresses this problem from isolated angles, yet none fully resolves it. Test-time adaptation (TTA) methods for AI-generated image detection adapt detectors to unseen generators, but primarily study adaptation to the current target distribution rather than retention across evolving data streams (Chen et al., 2022; Nguyen-Le et al., 2025; Yang et al., 2026). Continual deepfake detection preserves knowledge across sequential forgery domains, but relies on labeled training data (Li et al., 2023; Wang et al., 2022b). General CTTA learns from unlabeled streams while reducing forgetting, but remains vulnerable to the high error rates induced by unseen generators (Wang et al., 2022a; Lee et al., 2024; Zhao et al., 2026).

To address these challenges, we propose SEED (**S**elf-**E**volving **E**xperts for **D**etection), a framework for continual test-time adaptation in AI-generated image detection. To support stable continual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

self-training, SEED keeps the base detector frozen and maintains a dynamic pool of lightweight residual experts, each learning residual logit corrections for a context, the current distributional conditions of the stream. Governing this expert lifecycle requires deciding when to create a new expert, entirely without ground-truth labels or domain boundaries. SEED resolves this challenge through *Dual-Space Expert Modeling (DSEM)*, which uses the base-logit space for shift detection and the decision-orthogonal feature space for expert routing. In the base-logit space, DSEM uses persistent shifts in the base detector's score distribution as an unsupervised signal for expert creation. Moreover, DSEM removes components aligned with the base decision direction from the routing features, allowing expert routing to focus on context-specific cues rather than real/fake classification signals. When familiar contexts recur, DSEM routes inputs to suitable existing experts, reusing their residual heads and statistical memories.

To reliably adapt the active expert without ground-truth labels, *Authenticity Uncertainty Estimation (AUE)* derives pseudo-supervision by modeling the expert's base-logit history with a Gaussian mixture model (GMM), assigning lower- and higher-mean component groups to the real and generated pseudo-classes, respectively. Samples whose scores clearly favor one group receive larger reliability weights, whereas samples compatible with both are downweighted as ambiguous. Because the stream shifts over time and individual batches may provide noisy evidence, *Progressive Knowledge Consolidation (PKC)* accumulates reliability-weighted feature statistics in each expert's statistical memory. It then performs balanced feature replay by sampling from the diagonal Gaussian feature models of the two pseudo-classes, helping preserve acquired forensic knowledge without storing raw images. Across four public benchmarks, SEED improves final accuracy over the strongest adaptation baseline by 9.25–15.68 percentage points, with gains of up to 17.42 points over the frozen detector.

The contributions of this paper are fourfold:

- We propose SEED, a framework that continually adapts detectors through a dynamic pool of residual experts, capturing evolving forensic cues while mitigating catastrophic forgetting.
- We introduce Dual-Space Expert Modeling (DSEM), which uses persistent base-logit shifts to trigger expert creation and removes the component along the frozen classifier direction from the features used for expert routing.
- We develop an unsupervised adaptation scheme coupling AUE and PKC: AUE derives pseudo-labels and reliability weights from class-conditional base-logit densities, while PKC stabilizes adaptation through balanced feature replay without storing raw images.
- On four public benchmarks, SEED consistently outperforms competitive CTTA baselines under continual online streams.

2 Related Work

2.1 AI-Generated Image Detection

AI-generated image detection aims to identify forensic evidence that generalizes beyond synthesis models seen during training (Wang et al., 2020; Ojha et al., 2023). Existing efforts primarily enhance transferability through data augmentation (Wang et al., 2020), local artifact extraction (Tan et al., 2024; Zhong et al., 2024), or frequency analysis (Durall et al., 2020; Frank et al., 2020), while reconstruction-based methods exploit discrepancies between the reconstructions of real and generated images (Wang et al., 2023b; Ricker et al., 2024; Chen et al., 2024). More recently, pretrained vision–language models have provided rich semantic priors (Ojha et al., 2023; Cozzolino et al., 2024). These representations are further adapted via forgery-aware tuning (Liu et al., 2024a; Yan et al., 2025), prompt learning and intermediate-feature aggregation (Wang et al., 2026b; Koutlis & Papadopoulos, 2024); large-scale benchmarks also provide linguistic cues for detection (Wang et al., 2026a). These generalization-oriented approaches establish transferable representations for unseen generators. SEED builds upon these representations, extending source-trained detection with continual, feature-dependent residual learning on unlabeled streams.

Deployment-time adaptation has emerged to tackle unseen distribution shifts beyond training. Recent works enhance predictions on incoming target streams through detector parameter updates (Chen et al., 2022), T^2A's uncertainty-aware negative learning (Nguyen-Le et al., 2025), IAPL's image-adaptive prompts (Li et al., 2026), AIGI-Det-Calib's post-hoc calibration (Yang et al., 2026), or offline

curriculum self-training (Zhang et al., 2026). Retention becomes essential as generator distributions change or recur. Complementary continual deepfake detection methods preserve knowledge across labeled forgery domains (Li et al., 2023; Wang et al., 2022b). Building on these adaptation and retention perspectives, SEED pairs residual multilayer perceptron (MLP) updates with the retention of each expert's parameters and statistical memory, supporting real–generated ranking and expert reuse on unlabeled streams.

2.2 Continual Test-Time Adaptation

Test-time adaptation (TTA) adapts models to unlabeled test streams using self-supervision (Sun et al., 2020), TENT's entropy minimization (Wang et al., 2021), EATA's sample selection and SAR's sharpness-aware optimization (Niu et al., 2022; 2023), online pseudo-prototypes (Iwasawa & Matsuo, 2021), or prompt tuning (Shu et al., 2022; Samadh et al., 2023). To handle sequential and recurring distribution shifts, continual TTA (CTTA) seeks to limit error accumulation and catastrophic forgetting through CoTTA's and RMT's teacher–student updates (Wang et al., 2022a; Döbler et al., 2023), RoTTA's memory-based replay (Yuan et al., 2023), diversity-weighted ensembling (Marsden et al., 2024), and persistent adaptation to recurring test distributions (Hoang et al., 2024). LAME (Boudiaf et al., 2022) complements these update-based baselines with parameter-free batch prediction refinement. SEED stabilizes adaptation by keeping the base detector frozen, inferring pseudo-labels and reliability weights from GMMs, and localizing residual learning to the selected expert.

For vision–language adaptation, TDA maintains a dynamic feature cache (Karmanov et al., 2024), DynaPrompt selects and updates an online prompt buffer (Xiao et al., 2025), and CLIPTTA (Lafon et al., 2025) and BATCLIP (Maharana et al., 2025) refine cross-modal alignment through contrastive and bimodal updates, respectively. SEED accumulates reliability-weighted feature statistics for residual MLP updates through balanced feature replay in the frozen feature space.

To mitigate representation interference, isolating stage-specific classifiers in supervised incremental learning (Wang et al., 2023a) provides a rationale for modular test-time adaptation via parameter-efficient fine-tuning. In this regime, lightweight adapters restrict plasticity to localized parameters (Song et al., 2023; Liu et al., 2024b), while modular frameworks route samples across low-rank experts (Lee et al., 2024) or dynamically allocate capacity via shift detection (Zhao et al., 2026). Complementary memory mechanisms retain historical information through generative replay (Shin et al., 2017), feature-distribution modeling (Zhang et al., 2023), or class-conditional calibration (Cui et al., 2025). These works motivate modular adaptation and replay from stored feature statistics. SEED combines shift detection with decision-orthogonal expert routing and preserves each expert's parameters and statistical memory so that expert routing can reuse these states when a context recurs.

3 Method

3.1 Problem Formulation and Overview

Continual test-time adaptation (CTTA) for AI-generated image detection operates on an incoming, unlabeled data stream $\mathcal{X} = (\mathcal{B}_1, \mathcal{B}_2, \dots)$, which arrives in batches $\mathcal{B}$ of n images $\boldsymbol{x}_i$. In this setting, distributional conditions can change over time, and the method receives no generator identities, ground-truth labels, or explicit domain transition boundaries. The system is initialized with a pretrained base detector consisting of a vision encoder f_ψ and the linear binary head θ_0. For each incoming image $\boldsymbol{x}$, the encoder extracts a feature $\boldsymbol{z} = f_\psi(\boldsymbol{x}) \in \mathbb{R}^d$, from which the base head produces a base logit $s(\boldsymbol{z}) = \theta_0(\boldsymbol{z})$ distinguishing real images ($y = 0$) from generated ones ($y = 1$).

To support stable continual self-training on a non-stationary, unlabeled stream, SEED keeps the base detector frozen and introduces a dynamic pool of residual experts. Here, a *context* denotes the current distributional conditions of the stream, which may change with generation mechanisms, image semantics, or other domain properties. Contexts are inferred from observations rather than supplied as generator or domain labels, and our expert need not correspond one-to-one to generators or domains. For each incoming batch, the selected expert e augments the base prediction using a residual MLP θ_e. The residual input is the normalized feature $\bar{\boldsymbol{z}} = \boldsymbol{z}/\|\boldsymbol{z}\|_2$:

$$\widehat{p}_e(\boldsymbol{x}) = \sigma[\theta_0(\boldsymbol{z}) + \theta_e(\bar{\boldsymbol{z}})]\,, \tag{1}$$

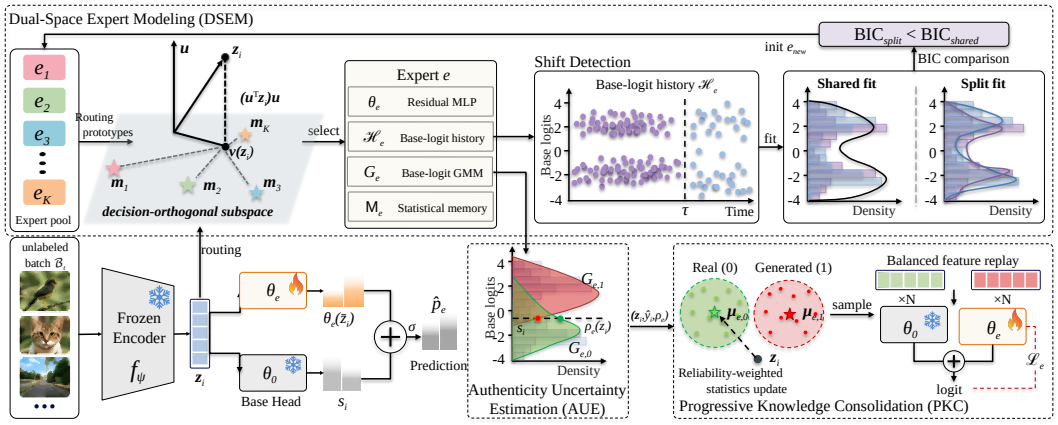

Figure 1: Overview of SEED. DSEM routes each batch using decision-orthogonal features and detects persistent shifts by comparing BIC scores for shared and split fits to an expert's base-logit history. The frozen encoder and base head provide features and logits, while the selected expert adds residual corrections. AUE estimates pseudo-labels and reliability weights, which PKC uses for balanced feature replay. The Real/Generated labels in PKC denote pseudo-classes.

where σ denotes the sigmoid function. Each expert is a self-contained adaptation unit that pairs its MLP head with a routing prototype and statistical memories, enabling expert routing and continued adaptation when its context recurs.

As illustrated in Figure 1, SEED realizes this dynamic routing and updating mechanism through three key components. First, **Dual-Space Expert Modeling (DSEM)** manages the expert pool by monitoring persistent changes in the base detector's logit space as a signal for expert creation, while utilizing decision-orthogonal features to select an appropriate expert when a context recurs. Once a batch is routed to an expert, **Authenticity Uncertainty Estimation (AUE)** mitigates the absence of ground-truth labels by inferring pseudo-labels and instance-level reliability weights. Finally, **Progressive Knowledge Consolidation (PKC)** stabilizes learning by accumulating these reliability-weighted statistics, facilitating balanced feature replay to efficiently update the selected residual θ_e without interfering with other experts.

3.2 Dual-Space Expert Modeling

As discussed in Section 3.1, a single detector suffers from interference under shifting distributions. SEED therefore manages its expert pool online by reusing an existing expert for recurring contexts and creating a new one when encountering persistent distribution shifts. Since explicit transition boundaries and target labels are unavailable, DSEM relies on two complementary mechanisms. First, it uses decision-orthogonal features for expert routing. Second, it relies on the experts' base-logit histories to detect persistent distribution shifts and create a new expert when necessary. Thus, expert routing reuses an existing expert state, while shift detection determines when to create a new state.

With the base detector frozen, each expert e has a lightweight residual MLP θ_e for prediction, a one-dimensional Gaussian mixture model (GMM) G_e fitted to its base logits for pseudo-labeling and shift detection, a routing prototype $\boldsymbol{m}_e$ for expert routing, and a statistical memory M_e for replay.

Firstly, DSEM selects the most suitable expert for each incoming $\mathcal{B}$ by matching decision-orthogonal features. Routing should distinguish distribution contexts despite two confounding factors. First, the raw encoder features contain high-level image semantics, which vary from sample to sample within a context. Second, the feature component along the base decision direction reflects the detector's real–generated judgment, so changes in a batch's class ratio—the proportion of generated samples—can affect its mean independently of the context. For expert routing, we project the features onto the subspace orthogonal to the base decision direction. Let $\boldsymbol{w} \in \mathbb{R}^d$ be the difference between the two weight vectors of the fixed base binary head, and $\boldsymbol{u} = \boldsymbol{w}/\|\boldsymbol{w}\|_2$. With $\mathrm{Normalize}(\boldsymbol{a}) = \boldsymbol{a}/\|\boldsymbol{a}\|_2$, the routing feature is computed as:

$$\boldsymbol{v}(\boldsymbol{z}) = \mathrm{Normalize}\big[(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{z}\big]\,. \tag{2}$$

The batch query $\boldsymbol{v}_{\mathcal{B}}$ averages these projected features: the projection removes the component along the base decision direction to reduce its influence on routing, while averaging reduces per-image semantic variation. Each expert has its routing prototype $\boldsymbol{m}_e = \mathrm{Normalize}\big(\sum_t \sum_{\boldsymbol{z}_i \in \mathcal{B}_t} \boldsymbol{v}(\boldsymbol{z}_i)\big)$ as the normalized mean of routing features. Let $\mathcal{E}$ denote the expert pool, and we route $\mathcal{B}$ to the expert maximizing cosine similarity:

$$e^\star = \arg\max_{e \in \mathcal{E}} \boldsymbol{v}_{\mathcal{B}}^\top \boldsymbol{m}_e. \tag{3}$$

Secondly, DSEM appends each image's base logit to the active expert's base-logit history $\mathcal{H}_e = (s_1, s_2, \ldots, s_{N_e})$, where $s_i = \theta_0(\boldsymbol{z}_i) \in \mathbb{R}$ is the scalar base logit of the i-th image processed by the expert. The base-logit GMM G_e is fitted to this history, with the component count selected by the Bayesian information criterion (BIC) (Schwarz, 1978). Persistent changes in the base-logit distribution serve as an unsupervised signal for expert creation. To distinguish genuine persistent shifts from transient fluctuations, we formulate this as a change-point detection problem. We compare the BIC of two hypotheses: a single GMM representing a shared context versus two independent GMMs split at a candidate time step τ from the evaluated set $\mathcal{T}$. The BIC condition for creating a new expert is:

$$\min_{\tau \in \mathcal{T}} \mathrm{BIC}_{\mathrm{split}}(\mathcal{H}_e; \tau) < \mathrm{BIC}_{\mathrm{shared}}(\mathcal{H}_e). \tag{4}$$

3.3 Uncertainty-Guided Expert Adaptation

Once DSEM routes each $\mathcal{B}$ to an expert e, we update its residual head without ground-truth labels. We achieve this efficiently through two sequential mechanisms: Authenticity Uncertainty Estimation (AUE) infers pseudo-labels and reliability weights, while Progressive Knowledge Consolidation (PKC) accumulates weighted feature moments in statistical memory for balanced feature replay.

To generate pseudo-labels via AUE, we analyze the selected expert's base-logit GMM G_e. We sort its Gaussian components by their means and split them at the maximum adjacent gap. Following the base-logit orientation, we assign the lower- and higher-mean groups to the real and generated pseudo-classes, respectively. Normalizing these two groups separately yields the class-conditional base-logit densities $G_{e,0}$ and $G_{e,1}$. Comparing them with an equal class prior prevents skewed class proportions from directly biasing the decision rule. The pseudo-label $\widehat{y}$ and its reliability weight $\rho_e(\boldsymbol{z})$ are computed below. AUE expresses uncertainty through the relative difference between the two fitted densities: a smaller weight indicates greater ambiguity.

$$\widehat{y} = \mathbf{1}[G_{e,1}(\theta_0(\boldsymbol{z})) \geq G_{e,0}(\theta_0(\boldsymbol{z}))], \qquad \rho_e(\boldsymbol{z}) = \frac{|G_{e,1}(\theta_0(\boldsymbol{z})) - G_{e,0}(\theta_0(\boldsymbol{z}))|}{G_{e,1}(\theta_0(\boldsymbol{z})) + G_{e,0}(\theta_0(\boldsymbol{z}))}. \tag{5}$$

The reliability weight is small when $G_{e,0}(s)$ and $G_{e,1}(s)$ are similar, near the pseudo-label decision boundary defined by their equality, and large when one density dominates.

To consolidate reliability-weighted supervision over time, PKC maintains a statistical memory $\mathsf{M}_e = \{(M_{e,c}, \boldsymbol{\mu}_{e,c}, \boldsymbol{\Sigma}_{e,c})\}_{c \in \{0,1\}}$ in the original frozen feature coordinates $\boldsymbol{z}$: each incoming feature contributes the weight $w_{ic} = \rho_e(\boldsymbol{z}_i)\mathbf{1}[\widehat{y}_i = c]$ to the mass, class mean, and diagonal covariance of its pseudo-class. The stored means and coordinate variances specify a moment-matched diagonal Gaussian feature model that maximizes entropy under these constraints (Jaynes, 1957), and the accumulated weighted moments are its maximum-likelihood parameters (Appendix C.1). Class-conditional Gaussian feature models have been used for deep-feature classification and continual learning (Lee et al., 2018; Zhang et al., 2023). Closed-form merging preserves the accumulated first and diagonal second moments exactly, supporting replay without retaining historical images or individual feature vectors (Appendix C.2). When new data arrive, each class's stored statistics update in closed form. For class c in the current batch, let $m_c = \sum_i w_{ic}$ be the total incoming weight and $\bar{\boldsymbol{z}}_c = m_c^{-1} \sum_i w_{ic} \boldsymbol{z}_i$ the corresponding weighted mean. The stored mass and class mean then merge as

$$M_{e,c}^+ = M_{e,c} + m_c, \quad \boldsymbol{\mu}_{e,c}^+ = \boldsymbol{\mu}_{e,c} + \frac{m_c}{M_{e,c}^+}(\bar{\boldsymbol{z}}_c - \boldsymbol{\mu}_{e,c}). \tag{6}$$

The centered second moments also admit an exact incremental merge. For each coordinate, we maintain the centered second-moment sum $\boldsymbol{S}_{e,c}$ (see Appendix B.3). The covariance used for replay is then $\boldsymbol{\Sigma}_{e,c}^+ = \mathrm{diag}\big(\max\{\boldsymbol{S}_{e,c}^+ / M_{e,c}^+, \nu\}\big)$, with $\nu = 10^{-6}$ guarding against degenerate variances.

Once both pseudo-classes have sufficient support in memory, the expert refines its residual head θ_e by replay: it trains on features sampled from the stored class Gaussians. To balance the class contributions, the replay distribution Q_e mixes the two pseudo-classes with equal probability, and the adaptation objective is the expected logistic loss under Q_e:

$$Q_e(\boldsymbol{z}, y) = \frac{1}{2}\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{e,y}, \boldsymbol{\Sigma}_{e,y}), \tag{7}$$

$$\mathcal{L}_e(\theta_e; \mathsf{M}_e) = \mathbb{E}_{(\boldsymbol{z},y)\sim Q_e} \ell_{\log}(\theta_0(\boldsymbol{z}) + \theta_e(\bar{\boldsymbol{z}}), y)\,. \tag{8}$$

This balanced feature replay samples equally from the two pseudo-class-conditional diagonal Gaussians, preventing the majority pseudo-class from dominating by count. Replay operates entirely in feature space, without storing raw images or training an image generator.

4 Experiments

4.1 Experimental Setup

Benchmarks and Protocols. Our evaluation covers four benchmarks: GenImage (Zhu et al., 2023), AIGCDetectionBenchmark (AIGC Bench.) (Zhong et al., 2024), AIGI-Holmes (Zhou et al., 2025), and OpenSDID (Wang et al., 2025), encompassing 7, 17, 10, and 5 distinct generators, respectively.

Under the **continual protocol**, all methods are initialized from the same source detector. This base detector is built on CLIP ViT-L/14 (Radford et al., 2021) and pre-trained on Stable Diffusion v1.4 from GenImage. The target generator domains are concatenated into a single uninterrupted stream without providing target labels, generator identities, or domain-boundary notifications. Stable Diffusion v1.4 is excluded only from the GenImage target stream; generator sequences in the other benchmarks are retained even when they include the same generator used for source training. Each target domain provides 750 real and 750 generated images for online adaptation and reserves 250 images per class for a holdout set; for each benchmark and seed, the per-domain holdouts are locked and jointly disjoint from the complete online stream across all target domains. Real and generated images undergo identical resizing and JPEG preprocessing. Further implementation details and stream orders are deferred to Appendix B.

For the **episodic protocol** (adaptation state resets between generators), each target generator is processed in an isolated run. Each method is reinitialized to its original pre-adaptation weights before adapting to a new generator, ensuring no parameters or memories are transferred between generators. This protocol enables broader comparisons with standard test-time adaptation (TTA) methods.

Evaluation Metrics. To strictly prevent data leakage, SEED predicts each incoming batch *before* adapting to it. Furthermore, holdout images are utilized exclusively for post-domain evaluation and never update the model or its adaptation state. We report the area under the receiver operating characteristic curve (AUC) and accuracy at a fixed 0.5 probability threshold, using domain-macro online and final holdout evaluation. Domain-macro metrics average the per-domain values equally; subscripts denote sample standard deviations across three random seeds.

For each domain that appears before the final stage, forgetting is the difference between its highest holdout score after arrival and its final holdout score, averaged across those domains. We report forgetting for both AUC and accuracy. AUC and accuracy (Acc.) are reported as percentages; gains and forgetting are reported in percentage points (pp).

Compared Methods. Under the continual protocol, we benchmark against state-of-the-art methods including TENT (Wang et al., 2021), EATA (Niu et al., 2022), SAR (Niu et al., 2023), CoTTA (Wang et al., 2022a), RoTTA-LN (Yuan et al., 2023), LAME (Boudiaf et al., 2022), and T^2A (Nguyen-Le et al., 2025). All methods in Tables 1 and 2, including Frozen, share the same source checkpoint. Under the episodic protocol, we additionally compare against AIGI-Det-Calib, TDA, DynaPrompt, CLIPTTA, BATCLIP, and IAPL using their official implementations; implementation and initialization details are provided in Appendix B.2.

4.2 MAIN RESULTS

Detection on uninterrupted streams. SEED achieves the highest mean AUC and accuracy across all four benchmarks in both online and final holdout evaluation (Tables 1 and 2). Final AUC exceeds the frozen detector by 1.28, 0.78, 1.96, and 1.12 percentage points on GenImage, AIGC Bench., AIGI-Holmes, and OpenSDID, respectively. Final accuracies reach 78.12%, 77.19%, 84.39%, and 89.97%, yielding gains of 9.98–17.42 points over the frozen detector and 9.25–15.68 points over the strongest baseline on each benchmark. Under the shared source initialization, these joint gains demonstrate improved discrimination and fixed-threshold detection through continual adaptation. Figure 2 shows the online accuracy trajectories, and Figure 4 illustrates the expert-pool lifecycle.

Table 1: Online continual detection on four benchmarks. AUC / accuracy (%); subscripts are three-seed standard deviations. Bold marks the best displayed mean.

	GenImage		AIGC Bench.		AIGI-Holmes		OpenSDID	
Method	AUC ↑	Acc. ↑	AUC ↑	Acc. ↑	AUC ↑	Acc. ↑	AUC ↑	Acc. ↑
Frozen	$82.08_{\pm 0.49}$	$67.97_{\pm 0.14}$	$81.96_{\pm 0.11}$	$64.61_{\pm 0.03}$	$90.67_{\pm 0.02}$	$67.10_{\pm 0.10}$	$96.79_{\pm 0.02}$	$80.32_{\pm 0.08}$
TENT	$68.34_{\pm 4.84}$	$51.90_{\pm 3.26}$	$65.20_{\pm 0.88}$	$50.32_{\pm 0.49}$	$75.58_{\pm 0.08}$	$50.02_{\pm 0.00}$	$86.91_{\pm 4.83}$	$67.08_{\pm 9.51}$
EATA	$82.52_{\pm 0.38}$	$67.97_{\pm 0.14}$	$82.30_{\pm 0.12}$	$64.61_{\pm 0.03}$	$90.87_{\pm 0.02}$	$67.10_{\pm 0.10}$	$96.85_{\pm 0.02}$	$80.34_{\pm 0.08}$
SAR	$82.52_{\pm 0.38}$	$67.97_{\pm 0.14}$	$82.30_{\pm 0.12}$	$64.61_{\pm 0.03}$	$90.87_{\pm 0.01}$	$67.10_{\pm 0.10}$	$96.85_{\pm 0.02}$	$80.32_{\pm 0.08}$
CoTTA	$49.30_{\pm 1.94}$	$50.02_{\pm 0.02}$	$50.45_{\pm 1.75}$	$50.08_{\pm 0.02}$	$44.83_{\pm 2.34}$	$50.06_{\pm 0.03}$	$49.95_{\pm 7.71}$	$50.69_{\pm 0.29}$
RoTTA-LN	$82.73_{\pm 0.36}$	$68.00_{\pm 0.12}$	$82.65_{\pm 0.17}$	$64.71_{\pm 0.03}$	$91.20_{\pm 0.05}$	$67.50_{\pm 0.22}$	$96.94_{\pm 0.01}$	$80.80_{\pm 0.09}$
LAME	$81.02_{\pm 0.42}$	$64.04_{\pm 0.05}$	$80.66_{\pm 0.11}$	$58.91_{\pm 0.02}$	$89.71_{\pm 0.04}$	$53.22_{\pm 0.06}$	$96.28_{\pm 0.09}$	$68.93_{\pm 0.37}$
T^2A	$54.30_{\pm 0.54}$	$52.15_{\pm 0.57}$	$53.01_{\pm 0.66}$	$52.02_{\pm 0.36}$	$53.09_{\pm 0.88}$	$51.13_{\pm 0.50}$	$59.62_{\pm 0.63}$	$57.50_{\pm 0.51}$
SEED	$\mathbf{83.51}_{\pm 0.29}$	$\mathbf{78.29}_{\pm 0.59}$	$\mathbf{83.30}_{\pm 0.39}$	$\mathbf{76.70}_{\pm 0.26}$	$\mathbf{92.68}_{\pm 0.35}$	$\mathbf{83.74}_{\pm 0.33}$	$\mathbf{97.27}_{\pm 0.06}$	$\mathbf{90.17}_{\pm 1.55}$

Table 2: Final holdout detection on four benchmarks after the complete stream. AUC / accuracy (%); subscripts are three-seed standard deviations. Bold marks the best displayed mean.

	GenImage		AIGC Bench.		AIGI-Holmes		OpenSDID	
Method	AUC ↑	Acc. ↑	AUC ↑	Acc. ↑	AUC ↑	Acc. ↑	AUC ↑	Acc. ↑
Frozen	$82.14_{\pm 0.68}$	$68.11_{\pm 0.75}$	$82.16_{\pm 0.32}$	$64.62_{\pm 0.10}$	$90.91_{\pm 0.08}$	$66.97_{\pm 0.30}$	$96.50_{\pm 0.05}$	$79.99_{\pm 0.24}$
TENT	$69.82_{\pm 10.87}$	$50.01_{\pm 0.02}$	$60.65_{\pm 3.22}$	$50.00_{\pm 0.00}$	$75.99_{\pm 0.52}$	$50.00_{\pm 0.00}$	$84.51_{\pm 7.98}$	$50.65_{\pm 1.13}$
EATA	$82.52_{\pm 0.64}$	$68.11_{\pm 0.75}$	$82.49_{\pm 0.33}$	$64.61_{\pm 0.09}$	$91.15_{\pm 0.06}$	$66.96_{\pm 0.30}$	$96.55_{\pm 0.05}$	$79.99_{\pm 0.24}$
SAR	$82.52_{\pm 0.65}$	$68.11_{\pm 0.75}$	$82.49_{\pm 0.33}$	$64.62_{\pm 0.10}$	$91.15_{\pm 0.06}$	$66.97_{\pm 0.30}$	$96.55_{\pm 0.05}$	$79.99_{\pm 0.24}$
CoTTA	$48.54_{\pm 3.01}$	$50.00_{\pm 0.00}$	$50.39_{\pm 2.16}$	$50.00_{\pm 0.00}$	$40.41_{\pm 1.18}$	$50.00_{\pm 0.00}$	$48.53_{\pm 9.79}$	$50.00_{\pm 0.00}$
RoTTA-LN	$82.51_{\pm 0.67}$	$68.33_{\pm 0.72}$	$81.88_{\pm 0.17}$	$65.25_{\pm 0.18}$	$92.31_{\pm 0.06}$	$68.71_{\pm 0.57}$	$96.69_{\pm 0.06}$	$80.72_{\pm 0.28}$
LAME	$80.89_{\pm 0.71}$	$63.81_{\pm 0.24}$	$80.80_{\pm 0.34}$	$58.98_{\pm 0.20}$	$89.79_{\pm 0.11}$	$53.27_{\pm 0.39}$	$95.97_{\pm 0.12}$	$68.47_{\pm 0.23}$
T^2A	$48.71_{\pm 8.17}$	$51.26_{\pm 2.18}$	$54.32_{\pm 4.66}$	$51.77_{\pm 2.11}$	$54.23_{\pm 12.14}$	$55.11_{\pm 7.05}$	$48.35_{\pm 13.83}$	$49.72_{\pm 0.97}$
SEED	$\mathbf{83.42}_{\pm 0.64}$	$\mathbf{78.12}_{\pm 0.78}$	$\mathbf{82.94}_{\pm 0.32}$	$\mathbf{77.19}_{\pm 0.18}$	$\mathbf{92.87}_{\pm 0.28}$	$\mathbf{84.39}_{\pm 0.13}$	$\mathbf{97.62}_{\pm 0.13}$	$\mathbf{89.97}_{\pm 0.46}$

Beyond fixed-threshold accuracy, SEED consistently enhances threshold-independent ranking across all benchmarks. Even on strong baselines such as OpenSDID (initial 96.50% AUC) and AIGI-Holmes (initial 90.91% AUC), SEED further elevates final holdout AUC to 97.62% and 92.87%, respectively. These consistent gains across diverse generator streams demonstrate that our residual adaptation genuinely refines the discriminative feature representation, rather than merely shifting the decision threshold to game classification accuracy.

In a residual-form ablation retaining expert creation, routing, pseudo-supervision, and balanced feature replay, the feature-dependent MLP improves final GenImage AUC by 1.47 points and accuracy by 1.43 points over a score-affine residual (Appendix A.4; Table S4).

Episodic adaptation. Under the episodic protocol (Table 3), SEED achieves online accuracies of 78.66%, 77.41%, 84.85%, and 92.03%. This exceeds CLIPTTA—the strongest competing accuracy baseline—by margins ranging from 14.11 to 33.63 percentage points, showing effective per-generator online adaptation with a reset before each generator.

Retention and deployment-time robustness. Figure 3 shows the stage-wise holdout accuracy trajectories across multi-generator streams. Two further stress tests probe deployment-time shift. First, skewing the generated-image prevalence of an entire uninterrupted stream from 10% to 90% moves

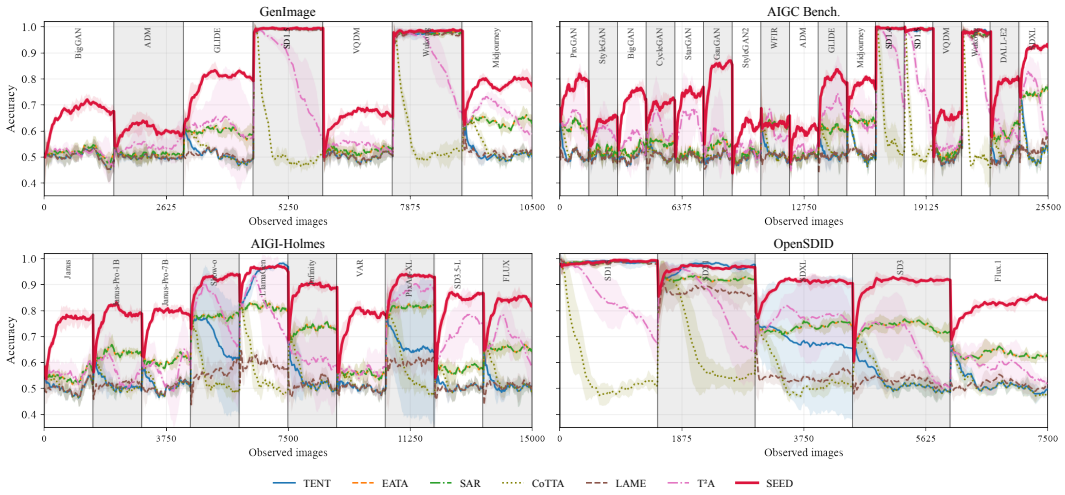

Figure 2: Online accuracy over a 20-batch sliding window for selected methods (mean ± standard deviation over three seeds). Vertical lines mark generator-domain boundaries for visualization only.

Table 3: Episodic online AUC and accuracy (%). Adaptation state resets for every generator.

	GenImage		AIGC Bench.		AIGI-Holmes		OpenSDID	
Method	AUC ↑	Acc. ↑	AUC ↑	Acc. ↑	AUC ↑	Acc. ↑	AUC ↑	Acc. ↑
AIGI-Det-Calib	$64.85_{\pm 0.69}$	$56.11_{\pm 0.47}$	$59.11_{\pm 0.08}$	$53.57_{\pm 0.22}$	$60.24_{\pm 0.47}$	$54.10_{\pm 0.13}$	$63.63_{\pm 0.35}$	$55.30_{\pm 0.81}$
TDA	$65.11_{\pm 0.41}$	$53.69_{\pm 0.15}$	$69.51_{\pm 0.13}$	$53.28_{\pm 0.05}$	$57.74_{\pm 0.30}$	$53.05_{\pm 0.13}$	$56.94_{\pm 0.25}$	$52.78_{\pm 0.14}$
DynaPrompt	$64.21_{\pm 0.58}$	$50.06_{\pm 0.03}$	$71.46_{\pm 0.15}$	$50.09_{\pm 0.01}$	$64.47_{\pm 0.28}$	$50.25_{\pm 0.01}$	$56.14_{\pm 0.42}$	$50.12_{\pm 0.04}$
CLIPTTA	$67.68_{\pm 0.22}$	$62.87_{\pm 0.42}$	$68.63_{\pm 0.59}$	$63.30_{\pm 0.60}$	$70.60_{\pm 0.54}$	$65.00_{\pm 0.34}$	$61.53_{\pm 0.14}$	$58.40_{\pm 0.29}$
BATCLIP	$51.39_{\pm 0.27}$	$50.29_{\pm 0.11}$	$51.37_{\pm 0.19}$	$50.25_{\pm 0.02}$	$51.88_{\pm 0.25}$	$50.90_{\pm 0.26}$	$50.83_{\pm 0.11}$	$50.39_{\pm 0.08}$
IAPL	$64.48_{\pm 0.39}$	$52.57_{\pm 0.34}$	$62.71_{\pm 0.25}$	$53.18_{\pm 0.04}$	$47.20_{\pm 0.10}$	$50.15_{\pm 0.03}$	$49.62_{\pm 0.20}$	$50.27_{\pm 0.02}$
SEED	$\mathbf{83.46}_{\pm 0.16}$	$\mathbf{78.66}_{\pm 0.21}$	$\mathbf{83.07}_{\pm 0.07}$	$\mathbf{77.41}_{\pm 0.18}$	$\mathbf{91.61}_{\pm 0.16}$	$\mathbf{84.85}_{\pm 0.32}$	$\mathbf{96.79}_{\pm 0.08}$	$\mathbf{92.03}_{\pm 0.17}$

balanced holdout accuracy by at most 2.9 points relative to the balanced stream on every benchmark (Table 4; Appendix A.2), and SEED stays at least 7.6 points above the frozen control under every skew, demonstrating robust detection across the tested fixed stream imbalances. Second, under input corruption—JPEG recompression down to quality 30 or Gaussian noise up to σ=0.10, applied to the episodic online streams—the frozen detector falls to 52.63% and 50.16% benchmark-macro online accuracy while SEED retains 63.93% and 65.72%, a paired margin that widens to +20.0 points at σ=0.05 (Table 5; Appendix A.3).

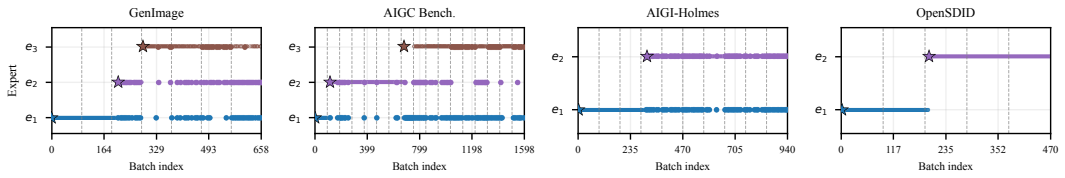

Figure 4: Expert routing and creation (seed 0). Stars mark expert creation; hollow points mark batches served by the newest expert, and filled points mark batches routed to an earlier expert. Dashed lines show generator-domain boundaries, which are unavailable to the method.

4.3 Ablation Studies

We ablate the three modules of SEED on the recurring stream SD1.5 → BigGAN → ADM → SD1.5, where a generator returns after two intervening ones and retention is directly measurable. Table 6 switches each module on (✓) or off (—) independently: disabling DSEM leaves a single continuously updated residual, disabling AUE falls back to frozen-detector hard labels with unit weights, and disabling PKC adapts on current observations only. Row (a) trains a single residual on current observations using frozen-detector hard labels and unit weights; the w/o Expert Routing control retains expert creation and initializes a fresh expert for each returning distribution.

Table 4: Final holdout accuracy (%) of SEED under fixed class-prior imbalance: each row fixes the generated-image fraction of an uninterrupted stream; holdouts stay balanced.

Generated (%)	GenImage	AIGC Bench.	AIGI-Holmes	OpenSDID
10	$75.97_{\pm 0.90}$	$73.80_{\pm 2.02}$	$83.24_{\pm 0.43}$	$89.20_{\pm 1.16}$
25	$75.75_{\pm 0.20}$	$76.22_{\pm 0.21}$	$83.13_{\pm 1.06}$	$90.05_{\pm 0.37}$
50	$78.30_{\pm 0.82}$	$76.67_{\pm 0.27}$	$84.31_{\pm 0.30}$	$90.91_{\pm 0.68}$
75	$77.99_{\pm 0.53}$	$76.79_{\pm 0.46}$	$83.69_{\pm 0.08}$	$91.19_{\pm 1.10}$
90	$77.92_{\pm 0.62}$	$77.19_{\pm 0.28}$	$82.79_{\pm 0.31}$	$88.84_{\pm 1.34}$

Table 5: Online accuracy (%) of SEED under input corruption on the episodic protocol.

Condition	GenImage	AIGC Bench.	AIGI-Holmes	OpenSDID
Clean	$78.66_{\pm 0.21}$	$77.41_{\pm 0.18}$	$84.85_{\pm 0.32}$	$92.03_{\pm 0.17}$
JPEG Q70	$73.88_{\pm 0.13}$	$70.73_{\pm 0.17}$	$78.20_{\pm 0.29}$	$86.84_{\pm 0.22}$
JPEG Q50	$66.40_{\pm 0.62}$	$64.28_{\pm 0.31}$	$68.14_{\pm 0.31}$	$78.63_{\pm 0.16}$
JPEG Q30	$63.66_{\pm 0.41}$	$59.95_{\pm 0.03}$	$60.64_{\pm 0.20}$	$71.47_{\pm 0.38}$
Gaussian noise				
σ=0.02	$74.57_{\pm 0.10}$	$71.48_{\pm 0.21}$	$75.05_{\pm 0.94}$	$82.49_{\pm 0.05}$
σ=0.05	$73.18_{\pm 0.25}$	$67.88_{\pm 0.29}$	$68.68_{\pm 0.43}$	$75.67_{\pm 0.40}$
σ=0.10	$69.00_{\pm 0.52}$	$64.67_{\pm 0.28}$	$62.21_{\pm 0.13}$	$67.01_{\pm 0.76}$

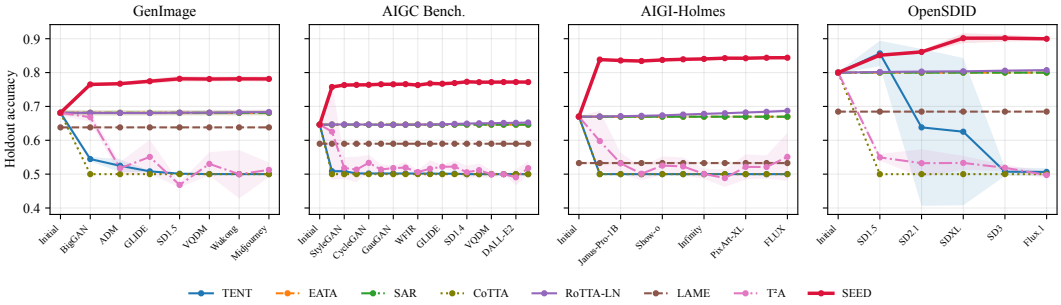

Figure 3: Domain-macro holdout accuracy over all target domains, including unseen ones, after each generator domain (mean ± standard deviation over three seeds). Initial denotes evaluation before adaptation.

Table 6: Component ablations on the recurring stream (SD1.5 → BigGAN → ADM → SD1.5). Detection scores are percentages; Final denotes final holdout evaluation. F_{AUC} is forgetting in AUC percentage points. Subscripts denote sample standard deviations over three seeds. Bold marks the best displayed mean.

Config.	DSEM	AUE	PKC	Online Acc.↑	Final Acc.↑	Final AUC↑	F_{AUC} ↓
(a)	—	—	—	$75.39_{\pm 0.20}$	$74.81_{\pm 0.45}$	$81.81_{\pm 0.97}$	$1.19_{\pm 1.22}$
(b)	✓	—	—	$75.41_{\pm 0.18}$	$74.78_{\pm 0.11}$	$83.34_{\pm 0.34}$	$0.50_{\pm 0.85}$
(c)	—	✓	✓	$75.93_{\pm 0.31}$	$75.41_{\pm 0.67}$	$82.96_{\pm 0.80}$	$0.57_{\pm 0.17}$
(d)	✓	—	✓	$78.11_{\pm 0.55}$	$79.72_{\pm 0.42}$	$85.64_{\pm 1.47}$	$\mathbf{0.05}_{\pm 0.09}$
(e)	✓	✓	—	$79.37_{\pm 0.67}$	$80.36_{\pm 1.91}$	$85.80_{\pm 1.31}$	$0.60_{\pm 0.17}$
w/o Expert Routing	✓	✓	✓	$80.47_{\pm 0.78}$	$75.60_{\pm 0.81}$	$83.84_{\pm 1.08}$	$4.33_{\pm 0.17}$
SEED	✓	✓	✓	$\mathbf{80.78}_{\pm 0.65}$	$\mathbf{82.44}_{\pm 1.04}$	$\mathbf{87.02}_{\pm 1.14}$	$0.09_{\pm 0.08}$

The component ablations in Table 6 validate that all three mechanisms provide complementary benefits on the recurring stream. Compared with the no-retrieval variant, SEED improves final AUC and accuracy by 3.18 and 6.84 percentage points, respectively, while reducing AUC forgetting from 4.33 to 0.09 percentage points. Meanwhile, AUE and PKC provide steady orthogonal improvements—AUE mitigates decision noise by filtering ambiguous pseudo-labels (+1.38% AUC, +2.72% Acc.), and PKC stabilizes residual optimization via class-balanced feature replay (+1.22% AUC, +2.08% Acc.). Operating jointly, the three components establish the best overall performance with near-zero backward interference.

5 Conclusion

In this paper, we introduced SEED for self-evolving AI-generated image detection. The key idea is to preserve each residual expert together with its statistical memory of reliability-weighted feature moments. Base-logit changes create experts, decision-orthogonal features select them, and AUE reliability weights guide statistical memory accumulation and balanced feature replay. This design maintains context-specific residual logit corrections without updating the base detector or storing past images. Experiments on four cross-generator streams show clear online and final gains, while ablations verify the roles of expert creation, expert routing, and residual MLP updates. One limitation is that SEED relies on the frozen base detector retaining useful real–generated separation in its base logits, so severe overlap can undermine pseudo-supervision.

AI USE STATEMENT

Generative AI tools assisted with language editing, manuscript organization, and implementation support. The authors are responsible for checking all mathematical statements, experimental evidence, references, and final claims.

REFERENCES

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

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 7621–7639, 2024.

Liang Chen, Yong Zhang, Yibing Song, Jue Wang, and Lingqiao Liu. OST: Improving generalization of DeepFake detection via one-shot test-time training. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of AI-generated image detection with CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4356–4366, 2024.

Shuang Cui, Jinglin Xu, Yi Li, Xiongxin Tang, Jiangmeng Li, Jiahuan Zhou, Fanjiang Xu, Fuchun Sun, and Hui Xiong. BayesTTA: Continual-temporal test-time adaptation for vision-language models via gaussian discriminant analysis. *arXiv preprint arXiv:2507.08607*, 2025.

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

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7890–7899, 2020.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 3247–3258, 2020.

Trung-Hieu Hoang, Duc Minh Vo, and Minh N. Do. Persistent test-time adaptation in recurring testing scenarios. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957. doi: 10.1103/PhysRev.106.620.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *Proceedings of the European Conference on Computer Vision*, 2024.

Marc Lafon, Gustavo Adolfo Vargas Hakim, Clément Rambour, Christian Desrosier, and Nicolas Thome. Cliptta: Robust contrastive vision-language test-time adaptation. In *Advances in Neural Information Processing Systems*, 2025.

Daeun Lee, Jaehong Yoon, and Sung Ju Hwang. Becotta: Input-dependent online blending of experts for continual test-time adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 27072–27093, 2024.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Chuqiao Li, Zhiwu Huang, Danda Pani Paudel, Yabin Wang, Mohamad Shahbazi, Xiaopeng Hong, and Luc Van Gool. A continual deepfake detection benchmark: Dataset, methods, and essentials. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1339–1349, 2023.

Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable ai-generated image detection via image-adaptive prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024a.

Jiaming Liu, Senqiao Yang, Peidong Jia, Renrui Zhang, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. ViDA: Homeostatic visual domain adapter for continual test time adaptation. In *International Conference on Learning Representations*, 2024b.

Sarthak Kumar Maharana, Baoming Zhang, Leonid Karlinsky, Rogerio Feris, and Yunhui Guo. Batclip: Bimodal online test-time adaptation for clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

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

Hong-Hanh Nguyen-Le, Van-Tuan Tran, Dinh-Thuc Nguyen, and Nhien-An Le-Khac. Think twice before adaptation: Improving adaptability of deepfake detection via online test-time adaptation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 16888–16905, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8245–8257, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9130–9140, 2024.

Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H. Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. doi: 10.1214/aos/1176344136.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. EcoTTA: Memory-efficient continual test-time adaptation via self-distilled regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11920–11929, 2023.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9229–9248, 2020.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024.

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

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-Prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022b. doi: 10.52202/068431-0411.

Yabin Wang, Zhiheng Ma, Zhiwu Huang, Yaowei Wang, Zhou Su, and Xiaopeng Hong. Isolation and impartial aggregation: A paradigm of incremental learning without interference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10209–10217, 2023a. doi: 10.1609/aaai.v37i8.26216.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. Opensdi: Spotting diffusion-generated images in the open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4291–4301, 2025.

Yabin Wang, Xiaopeng Hong, Yaqi Li, Zhiheng Ma, and Zhiwu Huang. Linguistic profiling of deepfakes: An open database for next-generation deepfake detection. *Pattern Recognition*, pp. 113395, 2026a. doi: 10.1016/j.patcog.2026.113395.

Yabin Wang, Zhiwu Huang, Zhou Su, Adam Prugel-Bennett, and Xiaopeng Hong. Penny-wise and pound-foolish in AI-generated image detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026b. doi: 10.1109/TPAMI.2026.3664388.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023b.

Zehao Xiao, Shilin Yan, Jack Hong, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiayi Shen, Qi Wang, and Cees G. M. Snoek. Dynaprompt: Dynamic test-time prompt tuning. In *International Conference on Learning Representations*, 2025.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *International Conference on Learning Representations*, 2025.

Muli Yang, Gabriel James Goenawan, Henan Wang, Huaiyuan Qin, Chenghao Xu, Yanhua Yang, Fen Fang, Ying Sun, Joo-Hwee Lim, and Hongyuan Zhu. Your AI-generated image detector can secretly achieve SOTA accuracy, if calibrated. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11622–11630, 2026. doi: 10.1609/aaai.v40i14.38146.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. SLCA: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19148–19158, 2023.

Yiqian Zhang, Zheyuan Gu, Xiangzhao Hao, Zefeng Zhang, Jingjia Mao, Jiahao Hu, Jiaxu Miao, Jun Yu, Zhenyu Zhang, Shuohuan Wang, and Yu Sun. Test-time curriculum for open-set AIGC detection. *arXiv preprint arXiv:2608.00559*, 2026.

Jianchao Zhao, Chenhao Ding, Songlin Dong, Jiangyang Li, Qiang Wang, Yuhang He, and Yihong Gong. Shared & domain self-adaptive experts with frequency-aware discrimination for continual test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28697–28705, 2026. doi: 10.1609/aaai.v40i34.40102.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2024.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In *Advances in Neural Information Processing Systems*, 2023.