THE CONFIDENCE ILLUSION: RETHINKING ENTROPY MINIMIZATION FOR CONTINUAL TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual Test-Time Adaptation (CTTA) aims to adapt a source model to continuously changing target domains using unlabeled test streams. Existing CTTA methods largely rely on entropy minimization, assuming that confident predictions provide reliable adaptation signals. However, we reveal a hidden confidence illusion: confident predictions are not always trustworthy. Under distribution shifts, misclassified samples can also exhibit low entropy and be repeatedly optimized, causing error accumulation. Moreover, such misleading confidence can gradually bias prediction distributions toward a few dominant classes, leading to long-term adaptation instability. To address this issue, we rethink entropy minimization from the perspective of reliable confidence optimization and propose Reliability-Aware Balanced Entropy (RBE). RBE distinguishes trustworthy confidence from misleading confidence through two complementary designs. Specifically, Consistency-aware Entropy Exploration (CAE) estimates prediction reliability by combining weak-augmentation consistency and teacher-student agreement, and selectively emphasizes reliable samples during entropy minimization. Meanwhile, Balanced Prediction Regularization (BPR) softly penalizes persistent concentration in a temporally aggregated prediction distribution, thereby mitigating class-dominant collapse. Extensive experiments on multiple CTTA benchmarks demonstrate that RBE achieves competitive overall performance and stable adaptation under continuously evolving target domains. Our findings highlight the importance of moving beyond confidence maximization toward reliable confidence modeling for continual adaptation.

1 INTRODUCTION

Deep neural networks have achieved strong performance in visual recognition when training and test data follow similar distributions He et al. (2016); Dosovitskiy et al. (2021). However, this assumption is often violated in real-world scenarios. Target data may be affected by weather changes, sensor noise, illumination variation, or other environmental factors. These distribution shifts can degrade the performance of a source model when it is directly deployed to unseen target domains. Test-time adaptation (TTA) aims to adapt a pre-trained source model during inference with unlabeled test data Wang et al. (2021). Continual test-time adaptation (CTTA) further considers a more challenging setting, where target domains change continuously along the test stream Wang et al. (2022); Niu et al. (2022). In this setting, the model must make predictions and update itself online without accessing source data or revisiting previous test samples.

Entropy minimization is one of the most widely used objectives in TTA and CTTA. It encourages the model to produce confident predictions on unlabeled target samples and provides a simple optimization signal when labels are unavailable. Existing CTTA methods improve online adaptation through teacher-student consistency, sample filtering, robust entropy optimization, model restoration, visual prompts, adapters, reconstruction objectives, memory mechanisms, and exploration-exploitation balancing Yang et al. (2025). Although these methods alleviate model degradation from different perspectives, many adaptation objectives still implicitly treat prediction confidence as a proxy for optimization reliability when determining the contribution of target samples. Unlike pre-

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.

vious hard reliability selection strategies, RBE continuously regulates entropy optimization through adaptive reliability weighting rather than discarding samples.

In this paper, we revisit a fundamental assumption behind entropy minimization: whether low-entropy samples necessarily provide reliable optimization signals for CTTA. As shown in Figure 1, we sort target samples from the continual ImageNet-C corruption stream according to prediction entropy and report the mean error of selected lowest-entropy samples. The source model still produces non-negligible errors among low-entropy samples, and the error rate increases as more low-entropy samples are included. This observation suggests that the key issue in CTTA is not only how to further reduce entropy, but also whose entropy should be minimized. Therefore, the reliability of an entropy objective should be evaluated at the sample level rather than directly inferred from prediction confidence. If misclassified low-entropy samples are repeatedly used for online updates, entropy minimization may reinforce incorrect predictions and cause error accumulation over time.

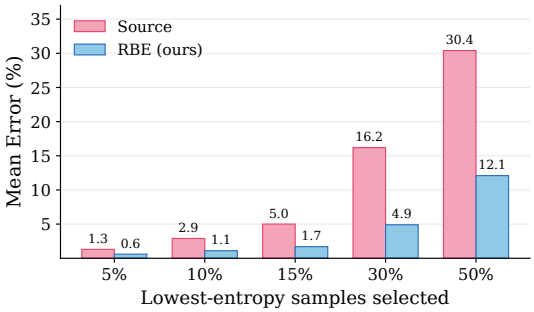

Figure 1: Mean error among the lowest-entropy samples on the continual ImageNet-C stream. Low entropy does not necessarily indicate correct predictions.

Low-entropy error accumulation not only affects individual sample updates, but may also amplify prediction-distribution bias during continual adaptation. Since entropy minimization encourages confident predictions, biased predictions in the current batch can gradually make a few classes dominate the output distribution. Once this class-dominance effect appears, more target samples may be assigned to these dominant classes, reducing prediction diversity and increasing the risk of prediction collapse. Therefore, the key is to make entropy minimization more reliable. On the one hand, the model should reduce the influence of unreliable low-entropy samples on online updates. On the other hand, it should also suppress the prediction-distribution bias that gradually emerges during continual adaptation.

Based on this revisiting, we propose Reliability-Aware Balanced Entropy (RBE), which makes entropy minimization reliability-aware during online adaptation. RBE contains two modules. First, Consistency-aware Entropy Exploration (CAE) estimates whether entropy minimization provides a trustworthy optimization signal for each sample. It jointly considers weak-augmentation consistency and teacher-student agreement, assigning larger entropy weights to samples with stable and consistent predictions while suppressing unreliable low-entropy samples. Second, Balanced Prediction Regularization (BPR) maintains a temporally aggregated prediction distribution using the batch-level predictions observed along the continual test stream. Rather than treating balanced prediction as an adaptation target, BPR uses a uniform reference to define a soft penalty on persistent concentration in this temporal distribution. By reducing sensitivity to short-term batch fluctuations and suppressing persistent class-dominant predictions, BPR improves distribution-level stability during continual adaptation. By combining CAE and BPR, RBE improves sample-level reliability and long-term prediction stability under continuously changing target domains.

The main contributions of this paper are summarized as follows:

- We revisit the fundamental assumption of entropy minimization in CTTA and reveal that low-entropy predictions can still provide unreliable optimization signals, causing error accumulation under continual distribution shifts.
- We propose Consistency-aware Entropy Exploration (CAE), which transforms entropy minimization from confidence-driven optimization into reliability-guided optimization by continuously estimating sample-wise optimization reliability.
- We introduce Balanced Prediction Regularization (BPR), which softly penalizes persistent concentration in a temporally aggregated prediction distribution, alleviating class-dominance effects while reducing sensitivity to individual batch fluctuations.
- Experiments on multiple CTTA benchmarks show that RBE achieves competitive or better overall performance than existing methods, and further analyses support the effectiveness of its reliability-aware design.

2 Related Work

Continual Test-Time Adaptation. Test-Time Adaptation (TTA) aims to adapt a source-trained model during inference using unlabeled test samples, reducing the distribution shift between training and test domains Boudiaf et al. (2022); Liang et al. (2025). Early methods mainly re-estimate batch normalization statistics or update a small set of trainable parameters Nado et al. (2020). Tent further shows that minimizing prediction entropy on test batches and updating normalization parameters can achieve effective online adaptation without source data or target labels. This setting is also related to source-free domain adaptation, where only the source model and unlabeled target data are available during adaptation Kundu et al. (2020).

Continual Test-Time Adaptation (CTTA) extends TTA to continuously changing target-domain streams, requiring the model to update online while avoiding error accumulation and forgetting Wang et al. (2022). Existing CTTA methods improve long-term adaptation from different perspectives, including teacher-student consistency and stochastic restoration, reliable sample selection and regularization Niu et al. (2022), robust mean-teacher learning Döbler et al. (2023), temporal correlation modeling Gong et al. (2022), and memory-efficient adaptation Hong et al. (2023). Unlike RMT, which directly optimizes teacher-student consistency Döbler et al. (2023), CAE uses teacher-student agreement only to estimate sample reliability. Recent methods further introduce contrastive learning, visual prompts, visual adapters, reconstruction objectives, memory mechanisms, and exploration-exploitation balancing to improve continual adaptation under domain shifts Yang et al. (2025). Different from these methods, our work focuses on the reliability of entropy minimization itself and improves CTTA through both sample-level reliability estimation and temporal prediction-distribution regularization.

Entropy Minimization. Entropy minimization encourages models to produce confident predictions on unlabeled samples and has become a fundamental objective in unsupervised adaptation Grandvalet & Bengio (2005). Tent introduces entropy minimization into test-time adaptation by updating normalization parameters during online inference without source data Wang et al. (2021). Subsequent methods improve entropy-based adaptation through different mechanisms, including teacher-student consistency, reliable sample selection, and robust entropy optimization. Specifically, CoTTA combines entropy minimization with augmentation consistency in a teacher-student framework Wang et al. (2022), while EATA reduces ineffective updates through confidence-based sample selection Niu et al. (2022). DeYO measures prediction sensitivity to object-destructive transformations and regards samples with larger prediction changes after destroying object structures as more reliable Lee et al. (2024). In contrast, CAE evaluates invariance under semantics-preserving weak augmentations and combines it with teacher-student agreement to dynamically weight entropy minimization along continual test streams.

However, existing entropy-based methods still implicitly rely on prediction confidence to determine whether target samples provide useful adaptation signals. Under continual distribution shifts, incorrectly classified samples may also produce low-entropy predictions and be repeatedly optimized, leading to error accumulation. Although recent approaches alleviate this issue through confidence filtering, pseudo-label correction, contrastive learning, or robust optimization Chen et al. (2022); Gao et al. (2024), they still depend on entropy, confidence, or limited reliability estimation. Different from these approaches, we propose a reliability-aware entropy optimization strategy that estimates sample-wise reliability from complementary consistency signals and dynamically adjusts the entropy minimization objective.

Prediction Regularization. Minimizing prediction entropy may cause the model to concentrate on a few dominant classes, resulting in class bias or prediction collapse. Prior studies alleviate similar degeneration through consistency regularization, distribution alignment, and diversity constraints Berthelot et al. (2019); Sohn et al. (2020). In CTTA, continuous model updates further amplify this problem, while existing methods mainly rely on teacher-student supervision, sample selection, or knowledge recovery Yu et al. (2024b), without explicitly regulating temporally aggregated predictions. To address this limitation, BPR maintains a temporally aggregated prediction distribution and uses a uniform reference to softly penalize persistent concentration. It targets long-horizon class dominance rather than enforcing balanced predictions within each individual test batch.

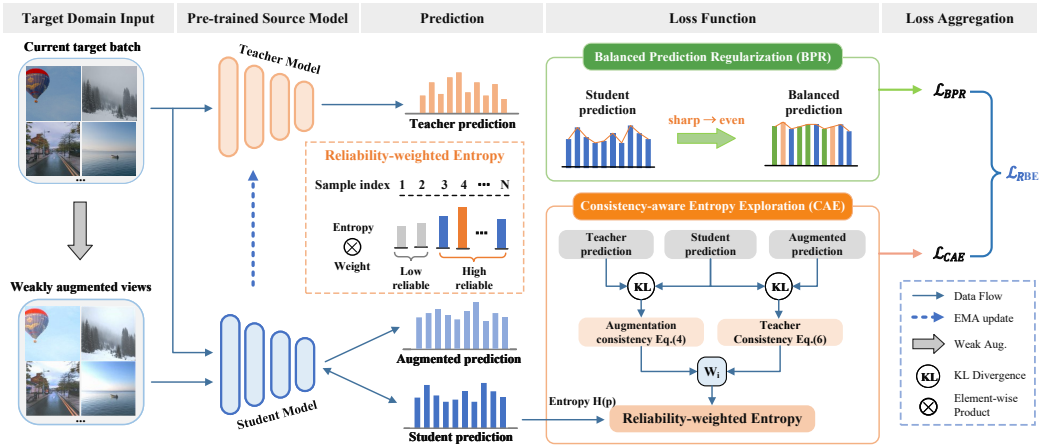

Figure 2: Overview of the proposed Reliability-Aware Balanced Entropy (RBE) framework. CAE uses weak-augmentation consistency and teacher-student agreement to construct reliability-weighted entropy minimization. BPR regularizes a temporally aggregated prediction distribution to alleviate class-dominant collapse while reducing sensitivity to individual batch fluctuations.

3 METHOD

3.1 OVERVIEW

Task Definition. Given a source model f_θ trained on labeled source data $\mathcal{D}_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, CTTA aims to improve its performance during inference on a continually changing target stream $\mathcal{D}_T = \{\mathcal{D}_{T_1}, \ldots, \mathcal{D}_{T_M}\}$, where each $\mathcal{D}_{T_m}$ denotes an unknown unlabeled target domain Wang et al. (2022). Target samples arrive sequentially in mini-batches, and the model must predict each batch before online updating. During testing, source data and target labels are unavailable, and previous predictions cannot be revised.

Overall Framework. As shown in Figure 2, RBE consists of two complementary modules. The original images are processed by both the student and teacher models, while their weakly augmented views are processed by the student model. CAE estimates sample reliability from augmentation consistency and teacher-student agreement, and uses it to weight entropy minimization. BPR softly penalizes persistent concentration in a temporally aggregated prediction distribution. RBE jointly optimizes the CAE and BPR objectives using the unlabeled target stream, while the teacher model is updated by exponential moving average Tarvainen & Valpola (2017).

3.2 CONSISTENCY-AWARE ENTROPY EXPLORATION

Conventional entropy minimization directly optimizes the prediction entropy of all samples in a test batch Wang et al. (2021). However, in CTTA, prediction entropy does not fully indicate sample reliability. Since the target distribution changes continuously, some misclassified samples may also have low entropy; if these samples are repeatedly optimized, the model may further reinforce incorrect predictions. To address this issue, we propose Consistency-aware Entropy Exploration (CAE), which estimates the reliability of each test sample using consistency information and uses the estimated reliability as the weight of entropy minimization.

For an input sample x_i, we first compute the prediction probability of the student model on the original input:

$$p_i = \mathrm{softmax}(f_\theta(x_i)). \tag{1}$$

Then, we apply weak augmentation $\mathcal{A}_w(\cdot)$ to obtain:

$$\tilde{x}_i = \mathcal{A}_w(x_i), \tag{2}$$

and compute the prediction probability of the weakly augmented sample:

$$p_i^w = \mathrm{softmax}(f_\theta(\tilde{x}_i)). \tag{3}$$

If a sample maintains consistent predictions before and after weak perturbation, it is more stable under input perturbations and is therefore more suitable as an online adaptation signal. Based on this intuition, the weak-augmentation consistency weight is defined as:

$$c_i = \exp\left(-D_{\mathrm{KL}}(p_i \| p_i^w)\right). \tag{4}$$

A smaller $D_{\mathrm{KL}}(p_i \| p_i^w)$ corresponds to stronger consistency and a larger c_i. In addition to input-perturbation stability, CAE uses the teacher model as a reliability reference rather than an optimization target. The teacher prediction for the same sample is:

$$p_i^T = \mathrm{softmax}(f_{\bar{\theta}}(x_i)). \tag{5}$$

The consistency weight between the student prediction and the teacher prediction is defined as:

$$t_i = \exp\left(-D_{\mathrm{KL}}(p_i \| p_i^T)\right). \tag{6}$$

A larger t_i indicates that the current student prediction is more consistent with the teacher reference, suggesting a more stable optimization signal. By combining complementary consistency measurements, CAE estimates the reliability of the entropy optimization signal and assigns a continuous weight to each sample:

$$w_i = \alpha c_i + \beta t_i, \tag{7}$$

where α and β control the contributions of the two consistency cues, respectively. We use linear fusion because it keeps the contribution of each reliability cue interpretable and avoids over-suppressing samples when one cue becomes temporarily unstable.

To keep the scale of the entropy objective stable across batches, we rescale the reliability weights by their batch mean:

$$\hat{w}_i = \frac{w_i}{\frac{1}{B}\sum_{j=1}^{B} w_j + \epsilon}, \tag{8}$$

where ϵ is a small constant for numerical stability. This mean normalization preserves relative sample reliability while keeping the average weight close to one. During optimization, the normalized reliability weights $\hat{w}_i$ are detached from the computation graph and serve only as sample-wise coefficients for entropy minimization. Finally, the reliability-aware entropy loss of CAE is defined as:

$$\mathcal{L}_{CAE} = \frac{1}{B}\sum_{i=1}^{B} \hat{w}_i H(p_i). \tag{9}$$

This objective does not apply the same entropy minimization strength to all samples. Instead, it adjusts the contribution of each sample according to consistency-based reliability weights, reducing the negative influence of misclassified low-entropy samples on model updates.

3.3 Balanced Prediction Regularization

CAE modulates entropy minimization at the sample level, but the model may still develop biased prediction distributions during continual adaptation. Specifically, entropy minimization encourages increasingly confident predictions. Once some classes are incorrectly reinforced, the model may gradually assign an excessive number of target samples to these classes, leading to reduced prediction diversity and class-dominant collapse.

Regularizing only the prediction distribution of the current batch may be sensitive to short-term variations, particularly when the incoming batch is small or its class composition is locally imbalanced. To obtain a more stable estimate of the prediction tendency along the continual test stream, BPR aggregates batch-level predictions over time. For the test batch $\mathcal{B}_t$ arriving at time step t, we first compute its mean student prediction as $\bar{p}_t = \frac{1}{B}\sum_{i=1}^{B} p_i$. We then combine $\bar{p}_t$ with the detached running prediction state r_{t-1}:

$$\begin{aligned} q_t &= \rho r_{t-1} + (1-\rho)\bar{p}_t, \\ r_t &= \rho r_{t-1} + (1-\rho)\,\mathrm{sg}(\bar{p}_t), \end{aligned} \tag{10}$$

where $\rho \in [0, 1)$ controls the temporal smoothing strength and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. For the first incoming batch, the running state is initialized with its detached mean prediction. Therefore, q_t remains differentiable with respect to the current student predictions, while historical prediction statistics provide a stable temporal reference.

To quantify and penalize persistent concentration on a small number of classes, we use the parameter-free reference distribution u, where $u_c = 1/C$ for $c = 1, \ldots, C$ and C is the number of classes. This reference is not an estimate of the target class prior; instead, it provides an anchor for defining a soft concentration penalty. The BPR objective is defined as

$$\mathcal{L}_{\mathrm{BPR}} = \frac{1}{1-\rho} D_{\mathrm{KL}}(q_t \| u) = \frac{1}{1-\rho} \sum_{c=1}^{C} q_{t,c} \log \frac{q_{t,c}}{u_c}. \tag{11}$$

The factor $1/(1-\rho)$ compensates for the attenuation of the gradient from q_t to the current batch prediction $\bar{p}_t$ caused by temporal smoothing. Consequently, BPR preserves a stable regularization strength while benefiting from historical prediction statistics.

The uniform reference is used to define a soft long-horizon concentration penalty rather than to estimate or impose the true target class prior. BPR does not require the prediction distribution of each individual test batch to be uniform; instead, it discourages a small number of classes from persistently dominating the temporally aggregated predictions. A more balanced temporal distribution is therefore an effect of mitigating class-dominant collapse rather than a predefined adaptation target. Together with CAE, BPR improves distribution-level stability while retaining confident predictions for reliable individual samples.

3.4 Optimization Objective

The overall objective of RBE combines the reliability-aware entropy loss from CAE and the temporal concentration regularization from BPR:

$$\mathcal{L}_{RBE} = \mathcal{L}_{CAE} + \gamma \mathcal{L}_{BPR}, \tag{12}$$

where γ controls the strength of BPR. For each incoming target batch, RBE estimates sample reliability from augmentation consistency and teacher–student agreement, and then minimizes the reliability-weighted prediction entropy. Meanwhile, BPR softly penalizes persistent concentration in the temporally aggregated prediction distribution to mitigate class-dominant collapse. The student model is optimized using only the current unlabeled target batch, while the teacher model and the running prediction state are updated sequentially along the continual test stream.

4 Experiments

4.1 Experimental Setup

Datasets and Task Setting. We evaluate RBE on three classification CTTA benchmarks, including CIFAR10-to-CIFAR10C, CIFAR100-to-CIFAR100C, and ImageNet-to-ImageNet-C. Following the standard CTTA protocol used in prior works Yang et al. (2025), the target domains are organized as a sequential test stream with the largest corruption severity, i.e., severity level 5. For ViT-Base, experiments are conducted on all three benchmarks. For ResNet-50, experiments are conducted only on ImageNet-C to match the CNN-based comparison setting.

During testing, the model has no access to source-domain data or target-domain labels. Each incoming batch is first evaluated, and the online prediction results are immediately recorded before the model is updated using the same unlabeled batch. For CIFAR10-C and CIFAR100-C, each corruption contains 10,000 test images. For ImageNet-C, each corruption contains 5,000 validation images following the RobustBench protocol Croce et al. (2021).

Implementation Details. We use pre-trained ViT-Base models for CIFAR10-C, CIFAR100-C, and ImageNet-C, and the pre-trained ResNet-50 model from RobustBench for ImageNet-C. The batch size is set to 32 for ViT-Base and 64 for ResNet-50. Following prior CTTA works Liu et al. (2024a), RBE performs one online update for each incoming test batch. For ViT-Base, RBE uses Adam with learning rates of 2×10^{-5} and 9×10^{-6} on CIFAR10-C and CIFAR100-C, respectively, and SGD with a learning rate of 7.5×10^{-5} on ImageNet-C. For ResNet-50 on ImageNet-C, RBE uses SGD with a learning rate of 6×10^{-3}. The hyperparameters are set to $\alpha = 2.0$, $\beta = 0.5$, $\gamma = 0.35$, and $\rho = 0.97$.

Table 1: Classification error rate (%) on ImageNet-C under continual test-time adaptation. Results are reported with ResNet-50 and ViT-Base backbones. The best result within each backbone group is shown in bold.

Time		$t \rightarrow$															
Backbone	**Method**	*Gaussian*	*Shot*	*Impulse*	*Defocus*	*Glass*	*Motion*	*Zoom*	*Snow*	*Frost*	*Fog*	*Brightness*	*Contrast*	*Elastic*	*Pixelate*	*JPEG*	*Mean*
ResNet-50	Source	97.8	97.1	98.2	81.7	89.8	85.2	78.0	83.5	77.1	75.9	41.3	94.5	82.5	79.3	68.6	82.0
	TENT Wang et al. (2021)	81.6	74.6	72.7	77.6	73.8	65.5	55.3	61.6	63.0	51.7	38.2	72.1	50.8	47.4	53.3	62.6
	EATA Niu et al. (2022)	75.9	65.7	**64.2**	74.6	**69.7**	64.7	57.6	62.8	63.0	53.5	40.8	64.5	50.2	47.8	50.0	60.3
	SAR Niu et al. (2023)	80.5	73.2	70.4	77.8	73.8	66.2	56.4	61.5	62.8	51.2	37.7	69.3	50.0	46.9	51.0	61.9
	RMT Döbler et al. (2023)	77.9	73.1	69.9	73.5	71.1	63.1	57.1	57.1	59.2	50.4	42.9	**60.1**	49.0	45.7	**46.9**	59.8
	DPLOT Yu et al. (2024a)	72.8	70.4	69.4	**73.2**	71.1	64.0	58.1	58.5	60.1	51.9	45.4	61.8	50.0	47.2	48.4	60.2
	PALM Maharana et al. (2025)	81.1	73.3	70.8	77.0	71.8	62.3	53.9	56.7	60.8	50.4	36.2	65.9	48.0	45.2	48.0	60.1
	GOLD Lai et al. (2026)	76.6	72.4	70.9	73.3	71.4	62.2	54.9	57.7	60.3	48.5	38.7	61.8	48.6	45.3	47.8	59.3
	RBE (ours)	**72.5**	**65.2**	69.1	76.1	76.5	**62.0**	**52.3**	**54.0**	**58.9**	**44.8**	**33.2**	73.6	**47.1**	**43.4**	49.5	**58.5**
ViT-Base	Source	53.0	51.8	52.1	68.5	78.8	58.5	63.3	49.9	54.2	57.7	26.4	91.4	57.5	38.0	36.2	55.8
	TENT Wang et al. (2021)	52.2	48.9	49.2	65.8	73.0	54.5	58.4	44.0	47.7	50.3	23.9	72.8	55.7	34.4	33.9	51.0
	CoTTA Wang et al. (2022)	52.9	51.6	51.4	68.3	78.1	57.1	62.0	48.2	52.7	55.3	25.9	90.0	56.4	36.4	35.2	54.8
	VDP Gan et al. (2023)	52.7	51.6	50.1	58.1	70.2	56.1	58.1	42.1	46.1	45.8	23.6	70.4	54.9	34.5	36.1	50.0
	ViDA Liu et al. (2024b)	47.7	42.5	42.9	52.2	56.9	45.5	48.9	38.9	42.7	40.7	24.3	52.8	49.1	33.5	33.1	43.4
	DPCore Zhang et al. (2025)	42.2	38.7	39.3	47.2	51.4	47.7	46.9	39.3	36.9	37.4	22.0	44.4	45.1	30.9	29.6	39.9
	DAS Zhu et al. (2026)	**40.0**	38.0	38.3	41.9	52.4	**35.2**	43.8	31.5	34.7	**28.1**	24.6	33.2	37.5	31.6	30.6	36.1
	RBE (ours)	41.8	**37.7**	**37.7**	**41.7**	**44.3**	38.6	**38.3**	**31.3**	**33.9**	29.4	**21.0**	**32.9**	**35.9**	**25.5**	**28.4**	**34.6**

4.2 Main Results

Comparison on ImageNet-C. Table 1 reports the results on ImageNet-C with both ResNet-50 and ViT-Base backbones. For ResNet-50, RBE achieves a mean error of 58.5%, obtaining the best overall result among compared methods. Although the margin over the strongest baseline is moderate and the performance varies across individual corruption types, RBE achieves the best overall mean error among the compared methods. This suggests that RBE can provide reliable online adaptation for CNN backbones without introducing architecture-specific designs.

The improvement becomes more pronounced on ViT-Base. RBE achieves a mean error of 34.6%, outperforming all compared baselines. Compared with the source model and DPCore, RBE reduces the mean error by 21.2 and 5.3 percentage points, respectively. Moreover, RBE surpasses the recent DAS method by 1.5 percentage points. These results show that RBE improves overall adaptation robustness by jointly considering sample reliability and long-term prediction-distribution stability under large-scale continual distribution shifts.

Comparison on CIFAR10-C and CIFAR100-C. Table 2 reports the mean error on CIFAR10-C and CIFAR100-C with ViT-Base. On CIFAR10-C, DAS achieves a lower mean error of 7.0%, compared with 11.6% for RBE. DAS constructs a labeled class-conditional synthetic knowledge base before deployment and uses these synthetic exemplars to provide explicit supervision during adaptation. Such additional class-level supervision is particularly effective on CIFAR10-C, which contains only ten categories.

In contrast, RBE relies solely on the source model and the unlabeled target stream without using external synthetic exemplars. Nevertheless, RBE outperforms the remaining compared methods on CIFAR10-C and achieves the lowest mean error of 19.4% on CIFAR100-C, compared with 22.2% for DAS. This result suggests that RBE is particularly effective in the larger label space, where sample-level reliability and distribution-level anti-collapse regularization become more important.

Table 2: Mean error (%) on CIFAR10-C and CIFAR100-C. The best results are shown in bold.

Method	CIFAR10-C	CIFAR100-C
Source	28.2	35.4
TENT Wang et al. (2021)	23.5	32.1
CoTTA Wang et al. (2022)	24.6	34.8
VDP Gan et al. (2023)	24.1	32.0
ViDA Liu et al. (2024b)	20.7	27.3
DPCore Zhang et al. (2025)	15.4	25.1
DAS Zhu et al. (2026)	**7.0**	22.2
RBE (ours)	11.6	**19.4**

4.3 Ablation Study

Loss Component Analysis. We evaluate the contribution of different loss components in RBE with ViT-Base on CIFAR10-C, CIFAR100-C, and ImageNet-C. As shown in Table 3, all variants keep the basic entropy minimization objective. We first ablate the two reliability cues c_i and t_i, where

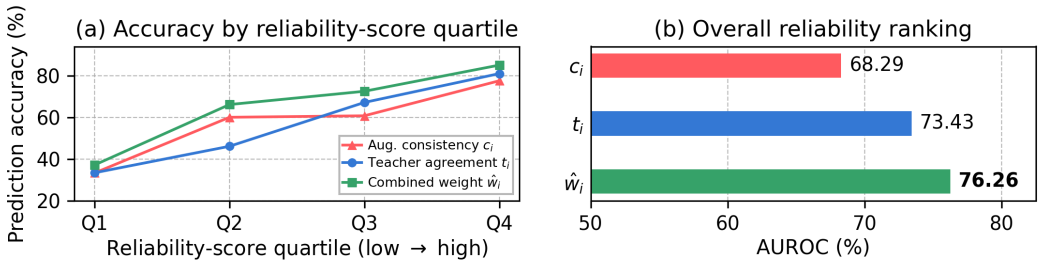

Figure 3: Direct validation of the reliability scores on the Gaussian-noise domain of ImageNet-C.

c_i denotes the weak-augmentation consistency weight and t_i denotes the teacher-student agreement weight. These two cues are combined in Eq. 7 and used by Eq. 9 to construct the reliability-aware entropy loss $\mathcal{L}_{CAE}$. We then evaluate the balanced prediction regularization term $\mathcal{L}_{BPR}$, which is defined in Eq. 11 to regularize the temporally aggregated prediction distribution. This design allows us to examine the effects of sample-level reliability estimation and temporal prediction regularization separately.

Table 3: Loss component ablation with ViT-Base on three benchmarks.

$\mathcal{L}_{CAE}$		$\mathcal{L}_{BPR}$	Mean Error (%)		
c_i	t_i		CIFAR10-C	CIFAR100-C	ImageNet-C
✓			13.4	52.4	66.5
	✓		22.7	38.8	53.9
✓	✓		12.9	25.9	40.6
		✓	70.3	86.6	71.1
✓	✓	✓	**11.6**	**19.4**	**34.6**

From the results, using only c_i performs well on CIFAR10-C but degrades clearly on CIFAR100-C and ImageNet-C, indicating that weak-augmentation consistency can capture local prediction stability but may still treat consistently wrong predictions as reliable under more complex label spaces or severe corruptions. In comparison, using only t_i gives more stable results on CIFAR100-C and ImageNet-C, suggesting that teacher-student agreement provides a more reliable reference constraint. Combining c_i and t_i further reduces the error on all three datasets, validating the complementarity of the two reliability cues. The variant combining unweighted entropy minimization with BPR performs poorly because it lacks reliability-aware sample reweighting. This further demonstrates that BPR primarily serves as a complementary anti-collapse regularizer rather than an independent adaptation objective. The complete RBE achieves the lowest mean error among the evaluated variants on all three datasets, demonstrating that CAE and BPR jointly improve CTTA from sample-level reliability and distribution-level stability.

4.4 Reliability and Stability Analysis

Reliability–Correctness Correlation. To directly examine whether the reliability cues reflect prediction correctness, we rank the samples according to each reliability score and divide them into four groups. As shown in Figure 3, prediction accuracy consistently increases from the low-reliability group to the high-reliability group for all three scores. This provides direct evidence that augmentation consistency c_i and teacher–student agreement t_i are positively correlated with prediction correctness and can therefore serve as effective indicators of sample reliability. Their combination $\hat{w}_i$ further strengthens this correlation and achieves the highest AUROC of 76.26%, confirming that the two consistency cues provide complementary reliability information.

Low-Entropy Sample Reliability. To examine whether low-entropy predictions can be safely used for model adaptation, we evaluate the mean error of samples with the lowest prediction entropy under different selection ratios. The results are shown in Figure 4 (a). RBE consistently achieves lower mean error among low-entropy samples. In particular, when the selected proportion increases, the error of Source and ViDA grows rapidly, indicating that many low-entropy samples are still misclassified. This observation confirms that directly minimizing entropy may repeatedly reinforce incorrect predictions during continual adaptation, while RBE improves the reliability of entropy-selected samples through consistency-aware weighting. DAS reduces this issue to some extent, but still produces higher error than RBE. In contrast, RBE maintains a lower error across different selection ratios, showing that its reliability-aware weighting makes low-entropy samples more trustworthy.

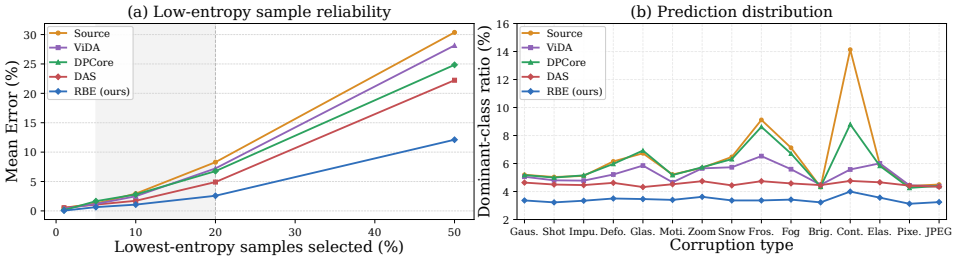

Figure 4: Reliability and prediction-distribution analysis on the continual ImageNet-C stream.

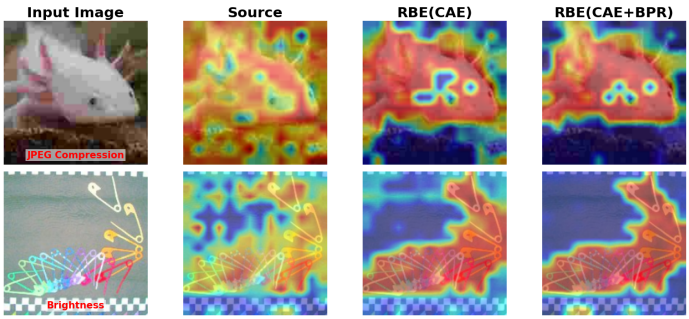

Figure 5: The CAM visualizations on the ImageNet-C dataset.

Prediction Distribution. To analyze whether RBE can alleviate persistent prediction concentration during continual adaptation, we visualize the dominant-class ratio across different corruption types in Figure 4 (b). For each online mini-batch, the dominant-class ratio is computed as $r = \max_c N_c/B$, where N_c denotes the number of samples predicted as class c, and B is the batch size. A larger dominant-class ratio indicates that more samples in a batch are assigned to the same predicted class, reflecting stronger prediction concentration and a higher risk of class-dominant collapse. RBE maintains a lower and more stable dominant-class ratio throughout the corruption stream. This result indicates that BPR mitigates persistent class dominance and improves distribution-level stability. The lower dominant-class ratio is an observed effect of the anti-collapse regularization rather than an objective that explicitly enforces class-balanced predictions.

4.5 Visualization Results

To intuitively analyze the effect of different components on the discriminative regions of the model, we employ CAM visualization on the ImageNet-C dataset. As shown in Figure 5, under severely corrupted inputs, the source model produces relatively scattered activation regions and is easily distracted by local corruption patterns. After introducing CAE, the activation maps become more focused on object-related regions, indicating that reliability-aware entropy minimization helps suppress unreliable adaptation signals. With the further incorporation of BPR, the resulting activation maps exhibit fewer scattered responses and more spatially consistent highlighted regions across different corruptions. These qualitative results suggest that CAE is associated with more object-focused activation patterns, while the complete RBE model exhibits more spatially consistent visual attention under continual distribution shifts.

5 Conclusion

In this paper, we propose Reliability-Aware Balanced Entropy (RBE) for continual test-time adaptation. RBE revisits entropy minimization from the perspective of sample reliability and shows that low-entropy predictions can still be unreliable during online adaptation. To address this issue, CAE reweights entropy minimization using weak-augmentation consistency and teacher-student agreement, while BPR softly penalizes persistent concentration in a temporally aggregated prediction distribution. By combining these two components, RBE improves sample-level reliability and distribution-level stability under continually changing target domains. Experiments on multiple CTTA benchmarks show that RBE achieves competitive overall performance, attaining the lowest mean error on ImageNet-C and CIFAR100-C while remaining competitive on CIFAR10-C.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, literature retrieval, and suggesting candidate experiments for assessing the effectiveness of the proposed method. The authors made all final decisions regarding the selection and design of the experiments, independently implemented the experiments, and verified all analyses and results. All AI-assisted suggestions and text were reviewed by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, pp. 5049–5059, 2019.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *CVPR*, pp. 8344–8353, 2022.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *CVPR*, pp. 295–305, 2022.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: A standardized adversarial robustness benchmark. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Mario Döbler, Robert A. Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *CVPR*, pp. 7704–7714, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Yulu Gan, Xinyuan Ma, Yihang Lou, Yuzhou Bai, Rui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *AAAI*, pp. 7595–7603, 2023.

Zhengqing Gao, Xu-Yao Zhang, and Cheng-Lin Liu. Unified entropy optimization for open-set test-time adaptation. In *CVPR*, pp. 23975–23984, 2024.

Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: Robust continual test-time adaptation against temporal correlation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27253–27266, 2022.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NeurIPS*, pp. 529–536, 2005.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Jinwoo Hong, Lingjuan Lyu, Jie Zhou, and Michael Spranger. MeCOTA: Memory-economic continual test-time model adaptation. In *International Conference on Learning Representations*, 2023.

Jogendra Nath Kundu, Naveen Venkat, and R. Venkatesh Babu. Universal source-free domain adaptation. In *CVPR*, pp. 4544–4553, 2020.

Guannan Lai, Da-Wei Zhou, Zhenguo Li, and Han-Jia Ye. The golden subspace: Where efficiency meets generalization in continual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3866–3875, 2026.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, 2024.

Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

Jiaming Liu, Ran Xu, Senqiao Yang, Renrui Zhang, Qizhe Zhang, Zehui Chen, Yandong Guo, and Shanghang Zhang. Continual-mae: Adaptive distribution masked autoencoders for continual test-time adaptation. In *CVPR*, pp. 28653–28663, 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 *ICLR*, 2024b.

Sarthak Kumar Maharana, Baoming Zhang, and Yunhui Guo. Palm: Pushing adaptive learning rate mechanisms for continual test-time adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19378–19386, 2025. doi: 10.1609/aaai.v39i18.34133.

Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv:2006.10963*, 2020.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, 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 *ICLR*, 2023.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alexey Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, pp. 1195–1204, 2017.

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

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, pp. 7201–7211, 2022.

Pinci Yang, Peisong Wen, Ke Ma, and Qianqian Xu. Learn faster and remember more: Balancing exploration and exploitation for continual test-time adaptation. *Manuscript*, 2025.

Yeonguk Yu, Sungho Shin, Seunghyeok Back, Minhwan Ko, Sangjun Noh, and Kyoobin Lee. Domain-specific block selection and paired-view pseudo-labeling for online test-time adaptation. In *CVPR*, pp. 22723–22732, 2024a.

Yongcan Yu, Lijun Sheng, Ran He, and Jian Liang. Stamp: Outlier-aware test-time adaptation with stable memory replay. In *ECCV*, pp. 375–392, 2024b.

Yunbei Zhang, Akshay Mehra, Shuaicheng Niu, and Jihun Hamm. Dpcore: Dynamic prompt coreset for continual test-time adaptation. In *Forty-second International Conference on Machine Learning*, 2025.

Zhilin Zhu, Yabin Wang, Zhiheng Ma, Yaguang Song, Yaowei Wang, and Xiaopeng Hong. Dance across shifts: Forward-facilitation continual test-time adaptation through dynamic style bridging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.