THE BATCH SIZE CAN BE ONE: TEST-TIME ADAPTATION FROM CLASSIFIER GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation updates a deployed model on an unlabeled test stream. In practice this stream often arrives one sample at a time, with the classes in no particular order. A single sample then provides limited information for estimating distribution-level statistics, and selecting samples by confidence can admit predictions that are wrong yet confident. A corruption, however, shifts the features while leaving the decision regions of the frozen classifier unchanged. We therefore propose Origin-recentered Test-Time Adaptation (OriTTA), which constructs fixed geometric targets from the source model and adapts selected features toward them. Seen from the origin of the feature space, the classes lean toward a common direction. We choose a new origin to bring class scores closer together while keeping it near the mean feature of one synthetic prototype per class. From this origin, prototype directions spread more evenly, and each class obtains an axis and an ideal point before any test data arrives. During adaptation, OriTTA aligns selected features with their predicted class axis and penalizes excessive radius, keeps one input per predicted class in memory to limit feature changes for classes absent from the stream, and resets the model when re-classifying this memory reveals collapse. Without tuning for each setting, OriTTA achieves the highest mean accuracy at batch size 1 on CIFAR-10-C, CIFAR-100-C and ImageNet-C in all sixteen combinations of four backbones, two protocols and two stream orders. On ImageNet-C fog, recent methods collapse in up to two thirds of one hundred runs, while OriTTA never does.

1 INTRODUCTION

In real-world deployment, a model may encounter test data from distributions that differ from its training data. Test-time adaptation (TTA) addresses this by updating the model on the unlabeled test stream itself, most commonly by minimizing the entropy of the model's own predictions with respect to its normalization parameters (Wang et al., 2021). Later methods refine which samples to learn from and how large a step to take (Niu et al., 2022; 2023; Lee et al., 2024; Hu et al., 2025).

Existing TTA methods face two difficulties in realistic test streams. First, many of them estimate normalization statistics from the current test batch, and these estimates become unreliable when samples arrive one at a time or in a non-shuffled class order (Niu et al., 2023). A single sample then provides limited information for estimating distribution-level statistics, and when a class arrives as one long run, even a large batch is dominated by that class, so its statistics capture the class rather than the distribution shift. Second, methods that update the model with its own predictions risk reinforcing their own mistakes. Confidence-based selection (Niu et al., 2022) filters out uncertain samples but can still admit confidently incorrect predictions, which subsequent updates then reinforce. When adaptation continues across a sequence of shifts (Wang et al., 2022), these errors compound over the stream until predictions collapse onto a few classes (Press et al., 2023).

Since the classifier is frozen during TTA, its decision regions stay unchanged and a corruption only displaces the features. The geometry of these fixed regions then indicates where the features of each class should lie, without depending on the test stream. Using this geometry directly is difficult, however, because of three properties of the classifier and its feature space. First, the classifier rows share a component that shifts every logit equally without changing predictions. Second, some feature directions affect neither predicted labels nor class probabilities. Third, class prototypes point toward a

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.

common direction when measured from the feature-space origin, making their angular arrangement sensitive to the choice of origin.

We therefore propose Origin-recentered Test-Time Adaptation (OriTTA), which places an ideal point for each class using the frozen classifier and synthetic class prototypes, and adapts selected features toward the ideal point of their predicted class. OriTTA centers the classifier rows and projects features onto their span and moves the origin to a point where class scores are close together while staying near the mean feature of one synthetic prototype per class. From this origin, prototype directions spread more evenly, and each class obtains an axis, its ideal point on the axis and a cone around it, all fixed before adaptation begins. On CIFAR-100, the fraction of prototypes inside the cone of their own class rises from 47% to 99%.

During adaptation, OriTTA applies angular and radial losses to the current test feature only if it lies inside the cone of its predicted class and near its ideal point. These losses encourage class-axis alignment and penalize excessive radius. A memory initialized with synthetic prototypes holds one input per predicted class and is updated with low-entropy test inputs. OriTTA periodically pulls stored features toward their ideal points to limit changes while their classes are absent from the stream. Reclassifying this memory detects concentration of predictions on a few classes and triggers restoration of the source model. Without per-setting tuning, OriTTA achieves the highest mean accuracy at batch size 1 on CIFAR-10-C, CIFAR-100-C and ImageNet-C in all sixteen combinations of four backbones, episodic and continual adaptation, and shuffled and class-blocked streams.

Our contributions are the following:

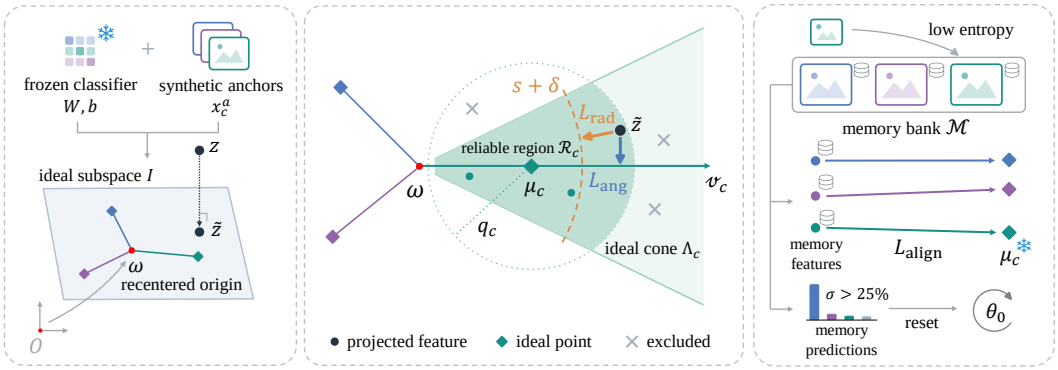
- We propose OriTTA, which takes the target of adaptation from the frozen classifier, measured from a recentered origin computed before any test data arrives.
- We design per-sample losses requiring neither batch statistics nor classifier softmax probabilities, and a collapse detector that evaluates prediction concentration across class-indexed memory slots.
- We analyze classifier and prototype geometry across four source models and demonstrate accuracy gains from projection and recentering in the evaluated ablations.
- We show that OriTTA achieves the highest mean accuracy at batch size 1 in all sixteen settings and never collapses in 100 runs on ImageNet-C fog.

2 Related Work

Single-sample and non-i.i.d. test-time adaptation. TENT (Wang et al., 2021) introduced entropy minimization over normalization affine parameters and EATA (Niu et al., 2022) added sample filtering and anti-forgetting. Single-sample adaptation was addressed by augmentation in MEMO (Zhang et al., 2022) and SITA (Khurana et al., 2021). SAR (Niu et al., 2023) formalized the *wild* setting, which combines batch size 1, online label shift and mixed shifts, and proposed reliable-sample selection with sharpness-aware updates, followed by DeYO (Lee et al., 2024), ReCAP (Hu et al., 2025) and SaTeen (Liu et al., 2026). Temporally correlated streams are handled by adjusting normalization statistics (NOTE (Gong et al., 2022), DELTA (Zhao et al., 2023), UnMIX-TNS (Tomar et al., 2024)), by class-balanced memories (NOTE, RoTTA (Yuan et al., 2023)) or by output correction (LAME (Boudiaf et al., 2022)). OriTTA selects and adapts test features by their position relative to a fixed ideal point for each class, without requiring batch statistics.

Continual adaptation and collapse. CoTTA (Wang et al., 2022), RMT (Döbler et al., 2023) and PETAL (Brahma & Rai, 2023) counter drift with teacher-student schemes and weight restoration. RDumb (Press et al., 2023) uses fixed-interval source resets, while PeTTA (Hoang et al., 2024) detects collapse by measuring divergence from source statistics. OriTTA reclassifies a memory with one input per predicted class and restores the source weights when predictions concentrate on a few classes. This trigger does not directly use the class frequencies of the incoming stream.

Feature geometry. Neural collapse (Papyan et al., 2020) has motivated geometric approaches to adaptation. TCA (Ni et al., 2025) promotes uniformity among pseudo-label centroids, while NEO (Murphy et al., 2026) recenters target embeddings without updating the model. Both estimate class structure from test batches, whereas OriTTA derives its origin and ideal points from the source model before any test data arrive. Further related work is discussed in Appendix B.



(a) Construction (b) Adaptation (c) Stabilization

Figure 1: Overview of OriTTA. (a) The frozen classifier and synthetic anchors define a fixed reference frame. (b) Reliable projected features undergo angular alignment and radial regularization. (c) Stored images are re-encoded for alignment to fixed ideal points, while concentrated current predictions over the memory trigger resets. Snowflakes denote frozen quantities.

3 Method

3.1 Notation and the Recentered Frame

Why recenter. Much of the recent geometric analysis of deep features measures distances and angles from the origin of the feature space: cosine similarity, feature-norm statistics, contrastive objectives on the unit sphere and the simplex-ETF picture of neural collapse all take the origin as the reference point. However, the origin is a property of the parameterization rather than of the data. After a ReLU the features occupy the non-negative orthant, so the origin sits at its corner. After a LayerNorm they are shifted by the affine parameters. In either case the center of the data has no reason to coincide with it. Angles measured from a point that is not the center are dominated by the offset, not by the arrangement of the classes. We therefore recover a center of the source feature distribution and define every quantity of our method in the frame recentered at that point.

The classifier. Let $f_\theta : \mathcal{X} \to \mathbb{R}^D$ be the backbone up to its penultimate layer, producing the feature $z = f_\theta(x)$ of an input x, and let the classifier be linear: for class $c \in \{1, \cdots, K\}$, $w_c \in \mathbb{R}^D$ is the c-th row of W, b_c its bias, and $\ell_c(z) = w_c \cdot z + b_c$ its logit, so that $\ell(z) = Wz + b$ and $\hat{y}(x) = \arg\max_c \ell_c(f_\theta(x))$. Write $\bar{w} = \frac{1}{K}\sum_c w_c$ for the mean row and

$$v_c = \frac{w_c - \bar{w}}{\|w_c - \bar{w}\|}, \qquad \beta_c = \frac{b_c - \bar{b}}{\|w_c - \bar{w}\|}, \qquad c = 1, \ldots, K, \tag{1}$$

collect the rows into $V \in \mathbb{R}^{K\times D}$ and $\beta \in \mathbb{R}^K$. Then the affine hyperplane

$$H_c = \{\, z \in \mathbb{R}^D : \; v_c \cdot z + \beta_c = 0 \,\} = \{\, z : \; \ell_c(z) = \bar{\ell}(z) \,\}, \qquad \bar{\ell} = \frac{1}{K}\sum_c \ell_c \tag{2}$$

is the locus on which class c scores exactly the average logit and $|v_c \cdot z + \beta_c|$ is the Euclidean distance of z from it, since $\ell_c(z) - \bar{\ell}(z) = (w_c - \bar{w}) \cdot z + (b_c - \bar{b})$. The unit normals v_c span the row space of the centered classifier, whose dimension we denote $r = \operatorname{rank} V = \min(K - 1, D)$.

Recentered origin. We define the center of the feature space as the point that is simultaneously closest, in the least-squares sense, to all K hyperplanes H_c. i.e. the point at which the classifier is maximally ambiguous:

$$\mathcal{J}(z) = \sum_{c=1}^{K} \operatorname{dist}(z, H_c)^2 = \|Vz + \beta\|^2. \tag{3}$$

$\mathcal{J}$ always admits a minimizer (not unique in general), but it constrains z only through its position relative to the hyperplanes and says nothing about where the features themselves lie.

A center must also be *where the source features live*, and in test-time adaptation that location has to be recovered without the source data. We obtain it from the classifier itself. For each class c we synthesize a single *anchor image* induced by the cross-entropy,

$$x_c^a = \arg\min_{x\in[0,1]^{C\times H\times W}} \mathrm{CE}\big(\ell(f_{\theta_0}(x)),\, c\big), \tag{4}$$

obtained by gradient descent on the pixels from a random initialization. This is the same cross-entropy the source model was trained with, minimized over the input rather than the weights, so x_c^a is the input the model is most confident belongs to class c, and it depends on the source model alone. Their feature centroid $\bar{z}^a = \frac{1}{K}\sum_{c=1}^{K} f_{\theta_0}(x_c^a)$ is a source-free estimate of where class-prototypical features gather. The *recentered origin* combines the two requirements,

$$\omega = \arg\min_z \ \|Vz+\beta\|^2 + \lambda\|z-\bar{z}^a\|^2 = (V^\top V + \lambda I)^{-1}\big(\lambda\bar{z}^a - V^\top\beta\big), \qquad \lambda = \frac{K}{100}, \tag{5}$$

the point closest to the decision hyperplanes among those near the prototype centroid. The regularizer is weighted at one hundredth of a single hyperplane constraint, so the ridge acts as $K/100$ additional constraints pulling toward $\bar{z}^a$. Appendix D.5 discusses the choice.

Ideal subspace. Let $\mathcal{I} = \omega + \mathrm{span}\{v_1,\dots,v_K\}$ be an ideal subspace which is the r-dimensional affine subspace through the recentered origin spanned by the class normals, and let

$$\tilde{z} := \mathrm{proj}_{\mathcal{I}}(z) = \omega + P\,(z-\omega), \qquad P = V^{+}V, \tag{6}$$

be the orthogonal projection onto it, where V^+ is the Moore-Penrose pseudoinverse of V so that P is the orthogonal projector onto $\mathrm{span}\{v_1,\dots,v_K\}$. Every logit *difference* is preserved by the projection, since $\ell_i(z)-\ell_j(z) = \ell_i(\tilde{z})-\ell_j(\tilde{z})$ for all i,j. Hence $\hat{y}$ and every softmax probability are unchanged, and the component of z orthogonal to $\mathcal{I}$ is invisible to the decision. Thus, we work with $\tilde{z}$ throughout and for a class c decompose $u = \tilde{z}-\omega$ into a radial and a perpendicular part,

$$t_c = u\cdot v_c, \qquad d_{\perp,c} = \sqrt{\|u\|^2 - t_c^2}, \tag{7}$$

so that $\|\tilde{z}-\mu_c\|^2 = (t_c-s)^2 + d_{\perp,c}^2$ splits the distance to the ideal point into a radius error and an angular error. We note that if $K-1 \geq D$, then $\mathcal{I}$ is the whole feature space.

Ideal points. Displacing ω along the normal v_c increases $\ell_c - \bar{\ell}$ faster than any other direction, so the axis $\{\omega + t\,v_c :\ t>0\}$ is the natural home of class c. We place the *ideal point* of class c at a common radius s along its axis,

$$\mu_c = \omega + s\,v_c, \qquad s = \mathrm{med}_{c\in\{1,\cdots,K\}}\ \big\|\,\mathrm{proj}_{\mathcal{I}}(z_c^a) - \omega\big\|, \tag{8}$$

where med denotes the median over classes and z_c^a is the feature vector of x_c^a. Hence, s is the distance at which the model's own prototypical inputs typically land and it is clear that $\mu_c \in \mathcal{I}$. Each anchor depends on its random initialization, but the frame does not. Regenerating the IN-RN50 anchors from ten other seeds moves ω by at most $0.03\,s$ and changes s by at most 0.4% (see Appendix A.2 for more details).

Ideal cone. Within $\mathcal{I}$ we define the region in which the class-c distribution is expected to lie as a cone around the class axis,

$$\Lambda_c = \big\{\tilde{z}\in\mathcal{I} :\ t_c > t_{\min},\ \ d_{\perp,c} < \kappa\, t_c\big\}, \qquad \kappa = \kappa_0\sqrt{r},\ \ \kappa_0 = 0.11,\ \ t_{\min} = 0.5. \tag{9}$$

The half-angle grows with $\sqrt{r}$ because the perpendicular offset of a class-c feature is spread over the $r-1$ directions orthogonal to its axis and concentrates at a radius proportional to $\sqrt{r}$. The ratio κ_0 of that spread to the prototype radius is the same across backbones whose feature scales differ by an order of magnitude, so a single constant serves all of them. The lower bound $t_{\min}$ excludes the immediate neighborhood of ω, where the classifier is ambiguous by construction.

Reliable region. A test feature enters the per-sample losses only if it is *reliable* for its predicted class $c = \hat{y}(x)$: it must lie in the ideal cone Λ_c and, among the cone samples of class c, be among the closest to μ_c,

$$\mathcal{R}_c = \big\{\tilde{z}\in\Lambda_c :\ \|\tilde{z}-\mu_c\| \leq q_c\big\}, \tag{10}$$

where q_c is a running estimate of the 0.2-quantile of $\|\tilde{z}-\mu_c\|$ over the samples that have entered Λ_c as class c so far. A fixed radius would not serve, since the distances depend on the corruption.

3.2 Losses on the Reliable Region

Two losses act on the current feature, and only when it is reliable, $\tilde{z} \in \mathcal{R}_{\hat{y}}$. Since they update normalization parameters shared by every input, moving the reliable features toward their ideal geometry also moves the features of the remaining inputs, on which no loss is computed.

Angular loss. With $c = \hat{y}(x)$, we treat the negative perpendicular distances to the K class axes as logits and apply a cross-entropy toward the predicted class,

$$L_{\text{ang}}(\tilde{z}) = -\log \frac{\exp(-d_{\perp,c}/\tau)}{\sum_{j=1}^{K} \exp(-d_{\perp,j}/\tau)}, \qquad \tau = \sqrt{s}. \tag{11}$$

This is a cross-entropy toward the predicted class, with the logits replaced by negative distances to the class axes. The temperature scales with s because $d_\perp$ does.

Radial loss. A reliable feature that lies far from the recentered origin is pulled inward with a one-sided hinge,

$$L_{\text{rad}}(\tilde{z}) = \frac{1}{r} \max\big(0,\ \|u\| - (s+\delta)\big)^2,\ \delta = \tfrac{1}{3}\, d_{\text{nn}},\ d_{\text{nn}} = \text{med}_{i\in\{1,\cdots,K\}} \min_{j\neq i} \|\mu_i - \mu_j\|. \tag{12}$$

It acts only on the radius and leaves the angle to L_{ang} (see Appendix D.2 for the ablation study).

3.3 Memory Bank and Alignment Loss

The two losses above act on whichever class the current input belongs to. On a class-blocked stream, many classes go unseen for thousands of steps, and nothing prevents the model from drifting away from them. We keep the most recent real input of every class in a memory bank and periodically pull each one back to its ideal point, so that no class is forgotten.

Memory bank. The memory bank $\mathcal{M} = \{m_1, \ldots, m_K\}$ holds one input per class, the most recently admitted one. It is initialized with the anchor images, $m_c \leftarrow x_c^a$, and an anchor is replaced by the first real input admitted for its class. The current input replaces $m_{\hat{y}}$ if its predictive entropy is below $0.4 \ln K$.

Alignment loss. Every $\lceil 30/B \rceil$ steps, we take $\min\{20, K\}$ classes $\mathcal{S}_t$ in round-robin order, pass their bank inputs through the current backbone, and pull each projected feature to its ideal point:

$$L_{\text{align}} = \frac{1}{|\mathcal{S}_t|} \sum_{c\in\mathcal{S}_t} \frac{1}{r} \left\| \text{proj}_{\mathcal{I}} \left(f_\theta(m_c) \right) - \mu_c \right\|^2. \tag{13}$$

where B is the batch size. The bank sample is used rather than the anchor because the anchors drift off μ_c under adaptation as well, and constraining a mixture of anchor and real features would push the real ones to the far side of μ_c. The target is the seed label rather than a fresh prediction, since re-predicting it would let a drifting model relabel its own bank (Appendix D.3 for the discussion).

3.4 Collapse Detection and Reset

Online adaptation can still collapse onto a few classes. We detect this by periodically re-classifying the bank with the current model and measuring the *bank share*

$$\sigma = \frac{1}{K} \max_j \ \big|\{c :\ \hat{y}(m_c) = j\}\big|, \tag{14}$$

the fraction of the bank sent to a single class. Since the bank holds one representative of every class, $\sigma = 1/K$ for a healthy model and rises only under collapse, independently of the label order of the stream. When $\sigma > 0.25$, the model is restored to its source weights and the bank to its anchors (details in Appendix A.6).

The complete objective is

$$L_{\text{total}} = L_{\text{ang}} + L_{\text{align}} + 2\, L_{\text{rad}}. \tag{15}$$

Nothing above is estimated from source data, and nothing is tuned per setting: every quantity is either derived from the source weights, through the classifier and its anchors, or a single constant shared across all datasets, backbones and stream orders.

Table 1: IN-RN50 under the four protocols. Per-corruption standard deviations are in Appendix E. The best result in each column is shown in bold and the second best is underlined.

	episodic	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
shuffled	TENT	12.35	20.22	13.91	13.16	6.46	18.07	18.23	16.00	20.65	1.48	69.80	42.04	6.22	49.38	53.75	24.11±0.40
	EATA	29.49	31.55	30.23	17.93	19.32	30.45	32.62	42.73	41.21	45.55	69.40	43.66	32.98	52.13	54.02	38.22±0.21
	SAR	28.63	31.01	29.40	18.46	15.23	28.82	30.43	44.75	44.39	33.60	71.95	44.61	13.03	47.83	56.24	35.89±0.11
	DeYO	40.87	43.33	42.08	22.07	23.69	38.63	28.12	52.26	50.21	18.88	73.19	51.27	42.97	57.31	58.54	42.89±1.95
	ReCAP	42.07	44.35	43.17	24.69	25.05	41.73	39.64	54.67	52.16	46.54	73.33	53.52	41.54	60.27	60.20	46.86±1.52
	SaTeen	**43.20**	**45.44**	**44.25**	25.80	26.25	**42.44**	26.68	**55.29**	52.69	19.30	**73.34**	**54.00**	48.71	**60.77**	60.13	45.22±2.20
	OriTTA	41.11	43.15	42.22	**33.53**	**31.93**	42.38	**46.33**	54.32	**53.27**	**59.04**	72.80	53.79	**49.64**	59.77	**60.62**	**49.59**±0.05
class-blocked	TENT	15.59	22.05	17.00	13.21	7.87	18.40	18.53	17.01	20.43	1.50	69.84	42.12	6.22	49.45	53.80	24.87±0.80
	EATA	25.27	25.71	25.28	15.68	14.93	21.35	22.70	33.11	32.90	37.47	65.32	34.82	18.24	38.49	50.43	30.78±0.19
	SAR	28.53	31.11	29.44	18.56	15.14	28.83	30.53	44.16	44.36	33.30	71.93	44.67	13.87	47.73	56.22	35.89±0.15
	DeYO	40.89	43.18	42.19	21.84	23.44	38.64	33.34	52.08	50.35	18.67	73.19	51.26	42.93	57.44	58.51	43.20±1.74
	ReCAP	42.13	44.41	43.39	24.72	24.72	41.78	39.92	54.62	52.25	35.69	**73.44**	53.63	42.67	60.32	**60.22**	46.26±2.10
	SaTeen	**43.13**	**45.48**	**44.31**	25.58	25.96	**42.34**	27.82	**55.28**	**52.74**	13.48	73.41	**54.00**	**48.85**	**61.00**	60.14	44.90±1.84
	OriTTA	40.79	42.55	41.81	**33.32**	**31.09**	41.83	**45.80**	53.69	52.67	**57.92**	72.50	53.12	48.79	58.73	60.05	**48.98**±0.10

	continual	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
shuffled	TENT	12.35	0.74	0.22	2.16	0.26	0.22	0.31	0.57	0.18	0.20	1.48	0.31	0.10	0.16	0.79	1.34±0.28
	EATA	29.49	29.30	25.44	12.94	14.77	18.41	23.08	26.36	27.85	31.02	54.08	24.44	21.43	32.12	38.49	27.28±1.91
	SAR	28.63	41.23	44.45	15.66	19.02	9.66	10.66	5.55	3.04	1.21	6.33	1.39	0.48	0.98	1.94	12.68±0.59
	DeYO	40.87	48.80	47.90	10.76	13.74	13.96	6.48	1.53	0.73	0.44	1.76	0.26	0.20	0.19	0.23	12.52±0.55
	ReCAP	42.03	**49.33**	**48.41**	**24.02**	3.50	3.91	2.70	3.55	3.85	5.13	6.98	4.96	1.72	4.31	4.65	13.94±8.28
	SaTeen	**43.20**	48.90	47.61	19.60	1.17	0.19	0.19	0.16	0.12	0.11	0.12	0.11	0.11	0.11	0.11	10.79±0.67
	OriTTA	41.09	43.78	38.40	22.63	**21.81**	**36.39**	**35.26**	**50.72**	**48.64**	**42.97**	**63.05**	**39.22**	**40.62**	**33.09**	**54.44**	**40.81**±1.02
class-blocked	TENT	15.59	1.08	0.27	2.73	0.36	0.27	0.38	0.63	0.19	0.22	1.97	0.41	0.10	0.17	0.97	1.69±0.39
	EATA	25.27	11.04	7.49	6.19	5.24	5.26	6.21	4.71	5.98	5.36	12.14	2.29	4.36	5.29	6.47	7.55±1.19
	SAR	28.53	41.14	44.47	15.76	19.17	11.52	11.18	15.79	18.37	18.04	34.59	20.35	5.68	18.72	20.27	21.57±10.86
	DeYO	40.89	48.73	47.86	11.55	11.50	11.82	6.89	1.88	0.77	0.47	2.02	0.27	0.20	0.20	0.24	12.35±0.84
	ReCAP	42.15	**49.28**	**48.43**	18.17	3.23	0.81	0.31	0.26	0.17	0.15	0.24	0.14	0.13	0.15	0.17	10.92±1.02
	SaTeen	**43.13**	48.78	47.57	24.45	3.17	3.59	2.04	2.06	0.95	0.24	0.16	0.10	0.12	0.10	0.10	11.77±1.78
	OriTTA	40.74	44.45	39.26	**25.80**	**24.91**	**33.45**	**32.44**	**52.02**	**48.77**	**45.97**	**66.77**	**45.32**	**39.96**	**41.61**	**48.27**	**41.98**±0.78

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets and backbones. We evaluate on CIFAR-10-C, CIFAR-100-C and ImageNet-C, on all 15 corruptions at severity 5 (10,000 images per corruption on CIFAR, 50,000 on ImageNet), with WideResNet-28-10 on CIFAR-10, WideResNet-40-2 trained with AugMix on CIFAR-100, and ResNet-50 with GroupNorm and ViT-B/16 on ImageNet, following SAR (Niu et al., 2023). We refer to the four settings as C10, C100, IN-RN50 and IN-ViT.

Protocols. All methods adapt online at batch size 1 with one update per image. In the *episodic* protocol the model is restored to its source weights at each corruption and in the *continual* protocol it adapts through the fifteen corruptions in sequence without any reset. Each protocol is run on a *shuffled* stream and a *class-blocked* one, in which each class arrives as a contiguous block. The stream is redrawn for every seed and shared across methods, and we report accuracy (%) averaged over ten seeds (2021–2030), with the standard deviation as a subscript.

Baselines. We compare against TENT (Wang et al., 2021), EATA (Niu et al., 2022), SAR (Niu et al., 2023), DeYO (Lee et al., 2024), ReCAP (Hu et al., 2025) and SaTeen (Liu et al., 2026), using their official code and published batch-size-1 settings (see Appendix A.10 for more details).

Implementation. OriTTA updates only the normalization affine parameters with Adam, at a learning rate of 2.5×10^{-4} for LayerNorm and $2.5 \times 10^{-4}/8$ for BatchNorm and GroupNorm (see Appendix D.1 for the discussion of Adam and SGD). The anchors and the frame are computed once per backbone, **before** the stream starts, and are never updated. All constants are listed in Appendix A.

4.2 MAIN RESULTS

IN-RN50 results. Table 1 reports the representative setting under all four protocols. In the episodic protocol, OriTTA leads the strongest baseline by about 2.7 on either stream order, and

the one-standard-deviation intervals of the two methods do not overlap. On the three noise corruptions, which come first in the sequence and which every method handles from the source weights, OriTTA trails the best baseline by two to three points. On defocus and glass blur, on zoom blur and on fog, it leads by five to twelve points, and by twenty on fog under the class-blocked stream. The standard deviation of OriTTA is the smallest, at 0.05 and 0.1, which indicates high stability of OriTTA. In the continual protocol the picture changes in kind rather than in degree. Every baseline collapses somewhere in the sequence: TENT within the first two corruptions, DeYO, ReCAP and SaTeen by the fourth or fifth, and the most resilient baseline, EATA on the shuffled stream and SAR on the class-blocked one, retains three quarters and three fifths of its episodic accuracy respectively. OriTTA gives up about nine points relative to its own episodic accuracy on the shuffled stream and seven on the class-blocked one, and ends at 1.5 and 1.9 times the accuracy of the best baseline.

Stability over one hundred seeds. In fact, the large seed-to-seed deviations of DeYO, ReCAP and SaTeen in Table 1 come almost from three corruptions, fog, zoom, and elastic corruption (Table 20). On the other twelve their deviations stay under a point. On these three corruptions the methods behave differently from run to run, and on fog some runs fail outright, with accuracy falling close to zero within the corruption. All three methods minimize prediction entropy, which reinforces the model's own errors once it becomes confidently wrong. 10 seeds are too few to characterize failures of this kind, so we repeat the three corruptions over 100 seeds each (Table 2). The deviations of DeYO, ReCAP and SaTeen remain large on all three. Crucially, on fog they collapse to below 5% accuracy in 66%, 33% and 62% of the seeds respectively. Across the same three hundred runs OriTTA never collapses, its deviation stays at 0.2–0.3.

Table 2: One hundred seeds (2001–2100) on the three corruptions with the least stable baselines for IN-RN50, episodic, shuffled. *Collapse* is the fraction of seeds in which accuracy of fog corruption falls below 5%.

	zoom	elastic	fog	collapse
TENT	18.2±0.6	6.1±0.5	1.5±0.1	100%
EATA	32.9±1.6	33.5±1.9	46.2±2.1	0%
SAR	30.4±0.3	14.1±1.4	33.2±2.2	0%
DeYO	33.1±6.8	43.0±0.5	21.0±25.2	66%
ReCAP	40.4±3.6	45.1±8.2	39.2±25.4	33%
SaTeen	27.2±9.1	48.8±0.4	23.6±27.7	62%
OriTTA	**46.5**±0.2	**49.7**±0.3	**59.0**±0.2	0%

Table 3: Mean accuracy over the 15 corruptions on C10, C100, and IN-ViT. Full results are in Appendix E.

	C10				C100				IN-ViT			
	episodic		continual		episodic		continual		episodic		continual	
	shuffled	class-blk	shuffled	class-blk	shuffled	class-blk	shuffled	class-blk	shuffled	class-blk	shuffled	class-blk
TENT	46.70±2.77	43.52±6.84	15.66±0.77	11.04±0.78	25.83±1.73	18.18±1.18	1.24±0.11	1.17±0.12	50.95±0.13	51.14±0.23	47.85±2.96	48.85±4.09
EATA	56.66±0.08	56.63±0.05	56.96±0.23	56.94±0.17	54.49±0.27	48.98±0.44	29.22±1.37	18.05±0.68	62.18±0.07	45.63±0.10	59.90±0.13	9.06±0.31
SAR	56.33±0.03	56.96±0.10	55.99±0.87	55.49±2.39	44.58±0.24	47.10±0.56	20.36±7.07	31.24±6.53	57.79±0.32	44.58±0.42	59.20±0.10	50.41±0.92
DeYO	51.92±2.94	57.34±6.14	44.42±3.45	37.86±6.62	10.49±1.02	12.48±1.56	1.89±0.55	1.91±0.26	62.91±0.62	62.93±1.09	60.70±0.54	60.60±0.34
ReCAP	48.70±0.98	58.80±0.69	36.58±6.71	43.13±9.33	2.09±0.15	7.48±1.58	1.06±0.03	1.28±0.13	65.00±0.66	64.02±1.56	61.37±0.82	61.13±0.75
SaTeen	37.35±2.55	52.90±8.45	14.27±1.78	27.62±7.73	1.92±0.10	4.94±0.81	1.04±0.01	1.33±0.23	64.68±0.81	64.97±1.20	61.62±0.32	61.64±0.56
OriTTA	**73.38**±0.08	**68.96**±1.51	**73.74**±0.17	**66.04**±1.43	**62.16**±0.09	**61.69**±0.33	**58.77**±0.28	**57.03**±0.29	**65.58**±0.05	**66.23**±0.10	**62.91**±0.96	**62.00**±0.30

Other datasets and backbones. Table 3 summarizes C10, C100 and IN-ViT by their mean accuracy. OriTTA is the most accurate method in each of the twelve columns, though the margin depends on how the baselines fare. On the two CIFAR benchmarks, whose WideResNet normalizes with BatchNorm, several baselines break down even in the episodic protocol, and in the continual one the strongest of them reaches only about half of OriTTA's accuracy on C100. The lead ranges from 8 to 30 points. On IN-ViT the picture is different. LayerNorm makes a transformer considerably more robust to single-sample adaptation, most baselines stay close to their episodic accuracy even in the continual protocol, and OriTTA's lead narrows to 0.4 to 1.3 points.

Continual TTA methods at batch size 1. The failure at batch size 1 extends to methods designed for continual adaptation. CoTTA (Wang et al., 2022), a representative continual method, falls to chance level on every setting when run at batch size 1 (Table 4), and the more recent TCA (Ni et al., 2025) likewise reports low accuracy for continual methods, including its own, at batch size 1.

Table 4: CoTTA at batch size 1 under our protocol, shuffled stream.

	C10	C100	IN-RN50	IN-ViT
episodic	10.31	1.14	0.12	0.12
continual	10.30	0.80	0.35	0.19

Table 5: Geometry of the source model under the three operations. *Rows dev.* is the deviation of the classifier rows' mutual cosines from an equiangular arrangement, $-1/(K-1)$, before and after centering. *pairwise deviation* is the same quantity for the prototypes z_c^a, measured from the origin, after projection, and from ω. *Axis* is the median cosine between prototype c and its own class axis v_c, and *in cone* the fraction of prototypes the ideal cone admits.

	shift			rows dev. ↓		pairwise deviation ↓			axis ↑		in cone ↑	
	$\|\omega\|$	$\|\omega\|/s$	$\|P\omega\|/s$	w_c	v_c	at 0	proj.	at ω	at 0	at ω	at 0	at ω
CIFAR-10-C	4.77	0.96	0.05	0.07	0.07	0.58	0.38	0.36	0.79	0.82	0.00	0.00
CIFAR-100-C	12.11	0.92	0.57	0.10	0.10	0.52	0.36	**0.16**	0.67	**0.81**	0.47	**0.99**
ImageNet-C, RN50	6.92	0.80	0.35	0.65	**0.03**	0.31	0.15	**0.07**	0.54	**0.58**	1.00	1.00
ImageNet-C, ViT	33.62	0.42	0.42	0.06	0.06	0.21	0.21	**0.09**	0.69	**0.77**	1.00	1.00

Table 6: Removing one operation of the frame, episodic. Top: losses computed in the full feature space, IN-RN50 (episodic, shuffled). Bottom: origin moved from ω to 0 and projection to $\mathcal{I}$, with s, μ_c, and δ recomputed, C100 (episodic, class-blocked).

		gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
IN-RN50	OriTTA	41.11	43.15	42.22	33.53	31.93	42.38	46.33	54.32	53.27	59.04	72.80	53.79	49.64	59.77	60.62	**49.59**
	no projection	38.93	40.83	39.50	31.02	27.94	39.57	43.71	53.65	52.58	48.34	72.72	52.62	45.98	57.63	60.35	47.02
	Δ	−2.18	−2.32	−2.72	−2.51	−3.99	−2.81	−2.62	−0.67	−0.69	**−10.70**	−0.08	−1.17	−3.66	−2.14	−0.27	−2.57
C100	OriTTA	48.07	52.29	51.20	70.98	54.74	69.73	70.81	63.13	63.30	58.71	71.93	63.51	62.84	64.83	59.24	**61.69**
	origin at 0	44.78	49.69	48.92	69.94	53.38	68.74	69.25	61.74	61.54	56.09	71.23	60.11	61.74	60.29	58.45	59.73
	Δ	−3.29	−2.60	−2.28	−1.04	−1.36	−0.99	−1.56	−1.39	−1.76	−2.62	−0.70	−3.40	−1.10	**−4.54**	−0.79	−1.96

4.3 What Each Operation of the Frame Contributes

The frame of Section 3.1 is built from three operations on the source model: centering the classifier rows, projecting onto the ideal subspace $\mathcal{I}$, and measuring from the recentered origin ω. Each removes one kind of structure that the classifier does not read, and how much of that structure a backbone carries can be measured before any test data arrives (Table 5). This gives a prediction for each operation: it should matter where the structure it removes is large, and not elsewhere. Table 6 checks the prediction by removing one operation at a time.

Centering. The classifier rows share a component $\bar{w}$ that adds the same amount to every logit and plays no part in any decision. On IN-RN50 it carries 81% of a row's norm, and the raw rows are far from equiangular (deviation 0.65). Once it is removed the axes v_c form the tightest equiangular set of the four settings (0.03). On the other three backbones $\bar{w}$ is negligible and centering changes nothing.

Projection. Logit differences depend on a feature only through its component in $\mathcal{I}$, so the remaining $D-r$ directions carry nothing the classifier reads, while a corruption can still displace a feature far along them. This is large on IN-RN50, where $1{,}049$ of $2{,}048$ dimensions are invisible, and absent on IN-ViT, where $P = I$. Removing the projection on IN-RN50 costs 2.6 points, and the loss falls where the argument places it. Without the projection the cone admits samples on the strength of variation the classifier does not use, and the radial term measures a radius it cannot read.

Recentering. A subspace fixes directions but not a point, and the cone, the ideal points and every distance the method computes need one. We take it from the model's own prototypes, and the effect is that from ω the prototypes lie along their axes: the cosine to the own axis rises in all four settings, and the fraction the ideal cone admits changes where it can, from 47% to 99% on C100. Only the row-space component $P\omega$ acts, and the fraction of the shift it carries grows with r/D. Thus, recentering should matter on C100, where it changes which prototypes the cone admits, and little on IN-RN50, where the cone admits all of them either way. Moving the origin back to 0 costs 2.0 points on the C100 class-blocked stream. On C10, this shift is only $0.05\,s$, and the row space spans just nine of the 640 feature dimensions. The small ridge weight $\lambda = K/100$ limits the pull toward the centroid of ten prototypes, keeping ω close to the classifier hyperplanes (see Appendix D.5).

Table 7: SAR with its own recovery, SAR with OriTTA's detector and reset in its place (SAR+RT), and OriTTA, on IN-RN50, class-blocked. RDumb (Press et al., 2023) restores EATA to its source weights every $64{,}000$ images. The subscript is the mean number of resets per run, shown only where nonzero. The total is given with the average.

episodic	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
SAR	28.53	31.11	29.44	18.56	15.14	28.83	30.53	44.16	44.36	33.30	71.93	44.67	13.87	47.73	56.22	$35.89_{(0)}$
SAR+RT	28.54	31.15	29.42	18.56	15.13	28.83	30.54	44.15	44.32	33.31	71.93	44.67	$20.47_{(1)}$	47.71	56.17	$36.33_{(1)}$
OriTTA	40.79	42.55	41.81	33.32	31.09	41.83	45.80	53.69	52.67	57.92	72.50	53.12	48.79	58.73	60.05	**48.98**$_{(0)}$
continual	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
RDumb	25.44	$22.09_{(1)}$	$22.14_{(1)}$	$15.51_{(1)}$	14.97	$20.73_{(1)}$	$21.44_{(1)}$	$28.41_{(1)}$	$22.80_{(1)}$	37.82	$66.62_{(1)}$	$32.52_{(1)}$	$15.86_{(1)}$	34.16	$48.97_{(1)}$	$28.64_{(11)}$
SAR	28.53	41.14	44.47	15.76	19.17	$11.53_{(8.5)}$	$11.17_{(7.1)}$	$15.79_{(6.1)}$	$18.37_{(36.8)}$	$18.04_{(3564)}$	$34.59_{(8)}$	$20.35_{(144)}$	$5.68_{(3740)}$	$18.72_{(33.1)}$	$20.27_{(1476)}$	$21.57_{(9025)}$
SAR+RT	28.54	41.14	44.46	15.74	19.17	$23.90_{(1)}$	33.83	43.49	46.49	53.29	61.32	46.88	$17.68_{(2)}$	47.43	54.29	$38.51_{(3)}$
OriTTA	40.74	44.45	$39.26_{(0.8)}$	$25.80_{(0.2)}$	$24.91_{(0.8)}$	33.45	$32.44_{(0.9)}$	$52.02_{(0.1)}$	48.77	$45.97_{(0.6)}$	$66.77_{(0.1)}$	$45.32_{(0.1)}$	$39.96_{(1)}$	41.61	48.27	**41.98**$_{(4.6)}$

4.4 Collapse Detection

We transplant the detector into SAR (Niu et al., 2023) in place of its own recovery, which resets the model when the moving average of its entropy falls below 0.2. SAR's loss, filter and optimizer are unchanged (Table 7). The two mechanisms differ in what they see. In the episodic protocol SAR falls to 13.9% on elastic transform, yet its recovery never fires. OriTTA's detector, which re-classifies a class-balanced bank, fires once there and recovers seven points. In the continual protocol SAR's recovery restores the model saved at the start of the current corruption, which is already degraded, and it fires about nine thousand times per run. With OriTTA's detector, SAR resets three times per run and gains seventeen points, from 21.6% to 38.5% with its loss unchanged. This accuracy is higher than the 35.9% that SAR reaches in the episodic protocol. A reset on a fixed schedule is less effective: RDumb (Press et al., 2023) resets eleven times per run and reaches 28.6%, below SAR+RT with three, because it discards adaptation at every interval rather than only when collapse occurs. Moreover, this accuracy is lower than 30.78% that EATA reaches in the episodic protocol. The detector therefore transfers, and the remaining gap to OriTTA is what the frame and the losses contribute. SAR+RT in the continual protocol is also more accurate than SAR in the episodic one: adaptation carries information from one corruption to the next, and a detector that resets only when collapse occurs keeps it, whereas a reset at every corruption boundary would discard it. On OriTTA itself the detector never fires in the episodic protocol and fires 4.6 times per run in the continual one, against fifteen corruption boundaries (see Appendix C for more details).

5 Discussion

Conclusion. We have argued that single-sample test-time adaptation should take its reference from the source model rather than from the test stream, and that the origin of the feature space is the wrong point to take it from. OriTTA recovers a frame from the source classifier and one synthetic prototype per class, adapts inside it with losses that require neither a batch nor the model's confidence, and detects collapse with a signal that does not depend on the label order of the stream. It is the most accurate method in every setting we evaluate, its accuracy varies by a fraction of a point across seeds where recent methods vary by several. Each operation of the frame removes a kind of structure the classifier does not read, and how much a backbone carries of it can be measured before adaptation. The accuracy each operation contributes follows that measurement, which is what allows a single set of constants to serve every dataset, backbone and stream order.

Limitations. The frame works best when the ideal subspace is neither the whole feature space nor a small part of it. When $K - 1 \geq D$, as for ViT-B/16, the projection has nothing to remove and OriTTA's lead over baselines that LayerNorm already keeps stable narrows to about a point. When K is small, as on CIFAR-10, the recentering is nearly invisible to the method and the cone rule, derived for large row spaces, becomes too narrow. The construction also assumes a linear head whose centered rows are close to equiangular, which holds for the models we examine but would not for a head trained on a long-tailed distribution. Finally, our evaluation covers synthetic corruptions only, and natural shifts such as ImageNet-R or ImageNet-Sketch remain to be tested.

Outlook. Restoring the most recent state the detector judged healthy, rather than the source weights, would retain adaptation that a full reset discards. The cone rule can be recalibrated for small row spaces, and the frame extends to any head that is linear in a learned representation, including dense prediction.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing the experimental pipeline and refining sentence-level wording in manuscript drafts written by the authors. The authors developed the theoretical framework, designed the method and experiments, and interpreted the results without generative AI assistance. All AI-assisted code was reviewed and tested, all manuscript edits were checked for accuracy, and all reported results were verified against experimental logs.

ETHICS STATEMENT

This work involves no human subjects, no new data collection and no personally identifiable information. All experiments use the public corruption benchmarks CIFAR-10-C, CIFAR-100-C and ImageNet-C (Hendrycks & Dietterich, 2019) and publicly released source models; ImageNet-derived data are known to carry demographic and geographic biases, which the corruptions inherit and which our method neither measures nor corrects. OriTTA is source-free: it requires no access to the source training data, and its anchors are synthesized from the source model alone. At test time it does, however, keep up to one incoming input per class in memory for the duration of the stream; deployments that process sensitive inputs should treat this buffer accordingly. Finally, online adaptation can degrade a model silently. Our collapse detector reduces this risk but offers no guarantee, and we do not recommend unsupervised test-time adaptation as the sole safeguard in safety-critical settings.

REPRODUCIBILITY STATEMENT

The released code contains the method, the anchor-generation script, the anchors used for all reported results, and the stream construction. The baselines were run from the official implementations released with SaTeen, and the streams they see are produced by the same function. Appendix A lists every constant of the method together with the per-setting quantities it derives from the source model, the learning-rate and batch-size conventions, the anchor-generation procedure, the baseline configurations and the two adjustments we made to run them under our protocol, and the hardware. The test streams are a deterministic function of the seed and are shared by all methods, as described in Appendix A.8. All results in the paper are averages over the seeds 2021–2030 unless stated otherwise.

REFERENCES

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Dhanajit Brahma and Piyush Rai. A probabilistic framework for lifelong test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Xiao Chen, Zhongjing Du, Jiazhen Huang, Xu Jiang, Li Lu, Jingyan Jiang, and Zhi Wang. Neural collapse in test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Zhikang Chen, Min Zhang, Sen Cui, Haoxuan Li, Gang Niu, Mingming Gong, Changshui Zhang, and Kun Zhang. Neural collapse inspired feature alignment for out-of-distribution generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 93671–93689, 2024.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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

Saurabh Garg, Sivaraman Balakrishnan, Zachary C. Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations (ICLR)*, 2022.

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 (NeurIPS)*, 2022.

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

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 (NeurIPS)*, 2024.

Zixuan Hu, Yichun Hu, Xiaotong Li, Shixiang Tang, and Lingyu Duan. Beyond entropy: Region confidence proxy for wild test-time adaptation. In *International Conference on Machine Learning (ICML)*, 2025.

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

Ansh Khurana, Sujoy Paul, Piyush Rai, Soma Biswas, and Gaurav Aggarwal. SITA: Single image test-time adaptation. *arXiv preprint arXiv:2112.02355*, 2021.

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 (ICLR)*, 2024.

Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9641–9650, 2020.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, pp. 6028–6039, 2020.

Chang Liu, Ruotong Zhao, Li Gao, and Yupei Zhang. Sateen: Learning structural alignment for continual test-time adaptation. In *International Conference on Machine Learning (ICML)*, 2026.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Alexander Murphy, Michał Daniłowski, Soumyajit Chatterjee, and Abhirup Ghosh. NEO: No-optimization test-time adaptation through latent re-centering. In *International Conference on Learning Representations (ICLR)*, 2026.

Chenggong Ni, Fan Lyu, Jiayao Tan, Fuyuan Hu, Rui Yao, and Tao Zhou. Maintaining consistent inter-class topology in continual test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15319–15328, 2025.

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

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge. RDumb: A simple approach that questions our progress in continual test-time adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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 (NeurIPS)*, volume 35, pp. 14274–14289, 2022.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR) Workshop*, 2014.

Devavrat Tomar, Guillaume Vray, Jean-Philippe Thiran, and Behzad Bozorgtabar. Un-mixing test-time normalization statistics: Combatting label temporal correlation. In *International Conference on Learning Representations (ICLR)*, 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 (ICLR)*, 2021.

Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. NormFace: L2 hypersphere embedding for face verification. In *ACM International Conference on Multimedia (ACM MM)*, 2017.

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

Yan Wang, Wei-Lun Chao, Kilian Q. Weinberger, and Laurens van der Maaten. SimpleShot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 37991–38002, 2022.

Hongxu Yin, Pavlo Molchanov, Jose M. Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via DeepInversion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bowen Zhao, Chen Chen, and Shu-Tao Xia. DELTA: Degradation-free fully test-time adaptation. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 29820–29834, 2021.

A Implementation Details

A.1 Backbones and Adapted Parameters

Table 8 lists the four settings. On CIFAR we use the RobustBench checkpoints for common corruptions, WideResNet-28-10 on CIFAR-10 and WideResNet-40-2 trained with AugMix on CIFAR-100, which normalize their input internally and receive raw $[0, 1]$ pixels. On ImageNet we use the ResNet-50 with GroupNorm and the ViT-B/16 (AugReg, ImageNet-21k pretrained, ImageNet-1k fine-tuned) checkpoints from `timm`, following SAR (Niu et al., 2023), with ImageNet mean and standard deviation applied externally. For every backbone OriTTA adapts only the affine parameters (γ, β) of the normalization layers. On ViT-B/16 the final LayerNorm before the classifier is kept frozen. The classifier and all convolutional and attention weights are frozen. The feature z is the input of the classifier, taken through a forward hook.

Table 8: Settings. D is the feature dimension, K the number of classes and $r = \min(K - 1, D)$ the dimension of the ideal subspace.

Setting	Dataset	Backbone	Norm	D	K	r
C10	CIFAR-10-C	WideResNet-28-10	BatchNorm	640	10	9
C100	CIFAR-100-C	WideResNet-40-2	BatchNorm	128	100	99
IN-RN50	ImageNet-C	ResNet-50	GroupNorm	2048	1000	999
IN-ViT	ImageNet-C	ViT-B/16	LayerNorm	768	1000	768

A.2 Anchor Generation and Its Algorithm

For each class c one anchor image x_c^a is synthesized from the source model alone by minimizing $\mathrm{CE}(\ell(f_{\theta_0}(x)), c) + 10^{-3}\,\mathrm{TV}(x)$ over the pixels, where TV is the total variation of the image. The image is initialized from Gaussian noise with standard deviation 0.05 in normalized pixel space and optimized with Adam at learning rate 0.1 for $1{,}000$ iterations under cosine decay, with the image clamped to the valid pixel range after every step. Classes are optimized in batches of 50 with seed 0.

Sensitivity to the anchor draw. The anchors depend on their random initialization, so we regenerate the IN-RN50 anchors from ten further seeds and rebuild the frame from each set (Table 9). Individual anchors change substantially. A regenerated anchor lies a median of $0.75\,s$ from the released anchor of its class, and seen from ω the two have a cosine of 0.72. The frame does not follow them. Across the ten sets ω moves by at most $0.032\,s$, s changes by at most 0.4%, and the geometry of Table 5 is unchanged at the reported precision. The frame is built from the centroid and the median of K anchors, which average out the variation of any single one, and ω inherits almost all of the centroid's shift of $0.033\,s$. This is half the shift caused by raising λ from 10 to 100, which leaves accuracy unchanged (Appendix D.5).

Table 9: Frame of IN-RN50 built from the released anchors and from anchors regenerated with ten other seeds (mean and standard deviation). Distances are in units of the released s.

	released	regenerated
$\|\omega\|$	6.93	6.96 ± 0.02
s	8.63	8.62 ± 0.02
δ	3.45	3.44 ± 0.01
shift of ω	–	0.031 ± 0.001
shift of $\bar{z}^a$	–	0.033 ± 0.001
shift of one anchor (median)	–	0.75 ± 0.01
pairwise deviation at ω	0.07	0.07 ± 0.00
axis at ω	0.58	0.58 ± 0.00
in cone	1.00	1.00 ± 0.00

Algorithm 1 Offline: anchors and frame (once per backbone, no data)

Require: source model f_{θ_0} with linear head (W, b), K classes, feature dimension D

$\bar{w} \leftarrow \frac{1}{K}\sum_c w_c, \quad \bar{b} \leftarrow \frac{1}{K}\sum_c b_c$

$v_c \leftarrow (w_c - \bar{w})/\|w_c - \bar{w}\|, \quad \beta_c \leftarrow (b_c - \bar{b})/\|w_c - \bar{w}\|$ ▷ Eq. 1

$V \leftarrow [v_1; \ldots; v_K], \quad P \leftarrow V^{+}V, \quad r \leftarrow \operatorname{rank} V = \min(K-1, D)$

for $c = 1, \ldots, K$ **do** ▷ anchor synthesis, Eq. 4

  $x \leftarrow$ Gaussian noise ($\sigma = 0.05$), clamped to the pixel range

  for 1000 iterations **do**

    $x \leftarrow \mathrm{Adam}\big(x,\ \nabla_x[\mathrm{CE}(\ell(f_{\theta_0}(x)), c) + 10^{-3}\,\mathrm{TV}(x)]\big)$, lr 0.1, cosine decay; clamp x

  end for

  $x_c^a \leftarrow x; \quad z_c^a \leftarrow f_{\theta_0}(x_c^a)$

end for

$\bar{z}^a \leftarrow \frac{1}{K}\sum_c z_c^a, \quad \lambda \leftarrow K/100$

$\omega \leftarrow (V^\top V + \lambda I)^{-1}(\lambda \bar{z}^a - V^\top \beta)$ ▷ Eq. 5

$s \leftarrow \mathrm{med}_c\, \|P(z_c^a - \omega)\|$ ▷ Eq. 8

$\mu_c \leftarrow \omega + s\, v_c$ for all c

$d_{\mathrm{nn}} \leftarrow \mathrm{med}_i \min_{j \neq i} \|\mu_i - \mu_j\|, \quad \delta \leftarrow 0.333\, d_{\mathrm{nn}}, \quad \tau \leftarrow \sqrt{s}, \quad \kappa \leftarrow 0.11\sqrt{r}$

Ensure: $\{x_c^a\}, \omega, P, \{v_c\}, \{\mu_c\}, s, \delta, \tau, \kappa$

A.3 Derived Quantities

Table 10 gives, for each setting, the quantities that Section 3.1 derives from the source classifier and the anchors. All are computed once before the stream starts and never updated.

Table 10: Quantities derived from the source model. s is the prototype radius, d_{nn} the median nearest-neighbor distance between ideal points, $\delta = d_{\mathrm{nn}}/3$ the radial slack, $\tau = \sqrt{s}$ the temperature of L_{ang}, $\kappa = 0.11\sqrt{r}$ the cone slope, and $\lambda = K/100$ the ridge weight in Eq. 5.

Setting	$\|\omega\|$	$\|P\omega\|/s$	s	d_{nn}	δ	τ	κ	λ
C10	4.77	0.05	4.95	7.01	2.33	2.23	0.33	0.1
C100	12.11	0.57	13.21	15.09	5.02	3.63	1.09	1
IN-RN50	6.92	0.35	8.62	10.35	3.45	2.94	3.48	10
IN-ViT	33.62	0.42	79.86	87.26	29.06	8.94	3.05	10

A.4 Constants

Every remaining constant is shared by all settings, protocols and stream orders (Table 11). They were selected on IN-RN50 and are not tuned per setting; the reset threshold $\sigma_{\max}$ was set so that it is not met by two classes merging on a ten-class bank (Appendix C). Learning rates are the only backbone-specific input and are set by normalization type: $2.5 \times 10^{-4}/8$ for BatchNorm and GroupNorm, and 2.5×10^{-4} for LayerNorm.

A.5 Reliable Region and Quantile Tracking

For each class c the reliable region of Eq. 10 uses a radius q_c that tracks the 0.2-quantile of $\|\tilde{z} - \mu_c\|$ over the samples that have entered the cone Λ_c with prediction c. Until 20 such samples have arrived, all of them are treated as reliable and their distances are collected. q_c is then initialized as the 0.2-quantile of these twenty distances (the fourth smallest). Afterwards, each new cone sample of class c updates the radius multiplicatively,

$$q_c \leftarrow \begin{cases} q_c\,(1 - 0.8\,\eta) & \text{if } \|\tilde{z} - \mu_c\| \le q_c, \\ q_c\,(1 + 0.2\,\eta) & \text{otherwise,} \end{cases} \qquad \eta = 0.02, \tag{16}$$

a stochastic approximation of the quantile under which exactly a fraction 0.2 of samples pass in equilibrium. Because the update is driven only by samples of class c, the radius of a class that is absent from the stream does not move. The radii are part of the adaptation state and are cleared on reset.

Table 11: Constants of the method, shared across all settings.

	Value	Role
κ_0	0.11	cone slope per $\sqrt{r}$ (Eq. 9)
$t_{\min}$	0.5	lower bound on the radial coordinate in the cone
trust ratio	0.2	quantile that defines the reliable region (Eq. 10)
warm-up	20	cone samples per class before the quantile engages
η	0.02	rate of the quantile tracker (Appendix A.5)
$w_{\text{ang}}, w_{\text{align}}, w_{\text{rad}}$	$1, 1, 2$	loss weights (Eq. 15)
entropy margin	$0.4 \ln K$	bank admission threshold
snap period, classes	30 frames, 20	schedule of L_{align} (Eq. 13)
probe period	200 frames	how often the bank is re-classified
$\sigma_{\max}$	0.25	reset threshold on the bank share (Eq. 14)
grace	1000 frames	minimum spacing between resets
Adam (β_1, β_2)	$(0.9, 0.999)$	no weight decay

A.6 Memory Bank, Alignment and Reset Schedule

The bank holds one input per class and is initialized with the anchor images. Admission is by predictive entropy below $0.4 \ln K$, evaluated on the input's logits before the current update. The admitted input replaces the bank entry of its predicted class, and when a batch admits several inputs of the same class the one closest to μ_c is kept. The geometric gate is not used for admission because at $K = 1000$ it admits too few samples to fill the bank. Admission by entropy can let a confidently mislabeled input into the bank. The detector does not measure this contamination directly but the drift that produces it: as the model moves toward the classes it favors, entries admitted earlier are re-mapped onto them and σ rises.

L_{align} is evaluated every $\lceil 30/B \rceil$ steps on the next $\min\{20, K\}$ classes in round-robin order, re-passing their bank inputs through the current backbone with gradient enabled so that the loss adjusts the backbone rather than the stored inputs.

The bank is re-classified without gradient every $\lceil 200/B \rceil$ steps, which is also the logging cadence. A reset fires when $\sigma > 0.25$ and at least 1000 frames have passed since initialization or the previous reset. It restores the source affine parameters, re-creates the optimizer (discarding Adam's moment estimates), re-seeds the bank with the anchors, and clears the quantiles q_c and their warm-up counters. No smoothing or persistence is applied to σ. It is recomputed from the whole bank at every probe. Its firing rate in every setting is reported in Appendix C.

A.7 Batch Sizes Above One

All schedules of the method are defined in frames, and a schedule of n frames becomes $\lceil n/B \rceil$ steps at batch size B, so that the snap, the probe and the grace period occur at approximately the same frame intervals of the stream at any batch size. The learning rate scales as $\sqrt{B}$, the rule appropriate for Adam (Malladi et al., 2022). A linear rule over-adapts at $B \geq 16$ and triggers a collapse reset within each corruption, which the detector catches with $\sqrt{B}$ no reset is needed. Admission, the quantile updates and the logging counters are applied per sample within a batch.

A.8 Test Streams

For every corruption the stream is a permutation of its images produced by a single function that both OriTTA and the baseline runner call with the same arguments, so that the two see byte-identical streams for a given seed. The shuffled stream is `torch.randperm` under a generator seeded with the seed. The class-blocked stream permutes the order in which the classes appear and, within each class block, the order of its images, using the same generator, so that a new seed redraws both. All fifteen corruptions of a run use the same seed. The ten seeds 2021–2030 therefore constitute ten independent repetitions of both the stream and the method. We note that in an earlier version of our code the class-blocked stream ignored the seed and used a fixed label-sorted order. All class-blocked results in the paper were produced after this was corrected.

A.9 Main Algorithm

Algorithm 2 Online adaptation (one stream, batch size B)

Require: outputs of Algorithm 1; source affine parameters θ_0

$\theta \leftarrow \theta_0$; $m_c \leftarrow x_c^a$, $q_c \leftarrow \varnothing$, $N_c \leftarrow 0$ for all c; $n \leftarrow 0$, $n_{\mathrm{rst}} \leftarrow 0$
for each batch x of the stream **do**
  $z \leftarrow f_\theta(x)$; $\hat{y} \leftarrow \arg\max_c \ell_c(z)$; $\tilde{z} \leftarrow \omega + P(z - \omega)$ ▷ predict
  $u \leftarrow \tilde{z} - \omega$; $t_c \leftarrow u \cdot v_c$; $d_{\perp,c} \leftarrow \sqrt{\|u\|^2 - t_c^2}$ ▷ Eq. 7
  $\mathcal{L} \leftarrow 0$
  for each sample i in the batch, with $c = \hat{y}_i$ **do**
    if $t_c > t_{\min}$ **and** $d_{\perp,c} < \kappa\, t_c$ **then** ▷ cone, Eq. 9
      if QuantileGate(c, $\|\tilde{z}_i - \mu_c\|$) **then** ▷ Eq. 10
        $\mathcal{L} \leftarrow \mathcal{L} + L_{\mathrm{ang}}(\tilde{z}_i) + 2\, L_{\mathrm{rad}}(\tilde{z}_i)$
      end if
    end if
  end for
  if step $\bmod \lceil 30/B \rceil = 0$ **then** ▷ alignment, Eq. 13
    $\mathcal{L} \leftarrow \mathcal{L} + \frac{1}{|\mathcal{S}_t|} \sum_{c \in \mathcal{S}_t} \frac{1}{r} \|P(f_\theta(m_c) - \omega) - s\, v_c\|^2$, $\mathcal{S}_t$: next $\min\{20, K\}$ classes
  end if
  $\theta \leftarrow \mathrm{Adam}(\theta,\ \nabla_\theta \mathcal{L})$
  if step $\bmod \lceil 200/B \rceil = 0$ **then** ▷ probe, Eq. 14
    $\sigma \leftarrow \frac{1}{K} \max_j |\{c : \hat{y}(m_c) = j\}|$ under the current θ
    if $\sigma > 0.25$ **and** $n - n_{\mathrm{rst}} \geq 1000$ **then** ▷ reset
      $\theta \leftarrow \theta_0$; new optimizer; $m_c \leftarrow x_c^a$; $q_c \leftarrow \varnothing$, $N_c \leftarrow 0$; $n_{\mathrm{rst}} \leftarrow n$
    end if
  end if
  for each sample i with $H(\mathrm{softmax}\, \ell(z_i)) < 0.4 \ln K$ **do** ▷ bank admission
    $m_{\hat{y}_i} \leftarrow x_i$ ▷ per class, the sample closest to $\mu_{\hat{y}_i}$
  end for
  output $\hat{y}$; $n \leftarrow n + B$
end for

function QuantileGate(c, ρ) ▷ Appendix A.5
  $N_c \leftarrow N_c + 1$
  if $N_c \leq 20$ **then**
    collect ρ; **if** $N_c = 20$ **then** $q_c \leftarrow 0.2$-quantile of the collected values
    return true
  else if $\rho \leq q_c$ **then**
    $q_c \leftarrow q_c(1 - 0.8\,\eta)$; **return** true
  else
    $q_c \leftarrow q_c(1 + 0.2\,\eta)$; **return** false
  end if
end function

A.10 Baselines

All baselines are run from the official implementations released with SaTeen, which bundle TENT, EATA, SAR, DeYO, ReCAP and SaTeen, under their batch-size-1 hyperparameters: SGD with momentum 0.9. The learning rate is $(0.00025/64) \cdot B \cdot 2$ for ResNet at $B < 32$ and 0.00025 (episodic) or 0.0001 (continual) at $B \geq 32$, and $(0.001/64) \cdot B$ (episodic) or $(0.0005/64) \cdot B$ (continual) for ViT, with an additional factor of 2 for SAR, DeYO and ReCAP at $B = 1$, entropy margins $0.4 \ln K$ (EATA, SAR), $0.5 \ln K$ and $0.2 \ln K$ (DeYO), $0.2 \ln K$ (SaTeen) and $0.8 \ln K$, $0.7 \ln K$ (ReCAP), ReCAP in its `recap_plpd` variant, and EATA without its Fisher regularizer, the variant its authors call ETA. The regularizer serves to prevent forgetting on in-distribution data, and its Fisher information is estimated from clean in-distribution samples, which our protocol does not provide. The original paper reports that the two variants reach comparable accuracy on corrupted data, with ETA slightly ahead, for example 19.4% against 19.7% error on CIFAR-10-C and 54.5%

against 54.8% on ImageNet-R (Niu et al., 2022, Tables 3 and 4). Since we report accuracy on corrupted data only, omitting the regularizer does not disadvantage EATA at batch size 64. We refer to this variant as EATA throughout. BatchNorm backbones keep their source running statistics and adapt only the affine parameters, as required at batch size 1. The seed is set once and re-set before each corruption, as in the released scripts.

Two adjustments were needed. First, ReCAP requires a per-channel feature-variance prior that is released only for ImageNet. For CIFAR we compute it from the clean CIFAR training features and calibrate its region scale so that the kernel exponent matches the ImageNet setting, without which ReCAP degenerates to no adaptation on CIFAR. Second, in the continual protocol the released scripts re-create the adaptation wrapper at each corruption while the model weights carry over, and we follow them. As a consequence the reference state that SAR's and ReCAP's recovery mechanisms restore is the model at the start of the current corruption rather than the source model. OriTTA receives no corruption-boundary information in either protocol.

RDumb (Press et al., 2023) is ETA restored to its source weights every $64{,}000$ images, the released interval of $1{,}000$ steps at batch size 64, with the continual learning rate of EATA. CoTTA (Wang et al., 2022) is run with its released hyperparameters at batch size 1. SAR+RT uses SAR's loss, filter and optimizer unchanged; its bank is seeded with the anchors of the same backbone, filled by entropy below $0.4 \ln K$, and re-classified every 200 frames, and a reset to the source weights fires when $\sigma > 0.25$ and at least $1{,}000$ frames have passed. We do not include LAME (Boudiaf et al., 2022), whose correction is computed over the samples of a batch and reduces to the source prediction at batch size 1, or NOTE (Gong et al., 2022), which requires retraining the source model with its instance-aware normalization layers.

A.11 Hardware and Runtime

All experiments were run on a single NVIDIA RTX PRO 6000 Blackwell GPU with 96 GB of memory. Table 12 reports the wall-clock time of each method on IN-RN50, episodic, shuffled, at batch size 1. The one-off anchor synthesis, is not included.

Table 12: Wall-clock time on IN-RN50, episodic, shuffled, batch size 1.

	gaus (s)	fog (s)	brit (s)	total (h)
No adaptation	304.0	301.4	300.9	1.3
TENT	715.5	657.3	662.3	2.8
EATA	816.7	1097.6	1328.5	4.5
SAR	1961.1	3438.8	3414.4	12.0
DeYO	710.6	669.4	1139.1	3.5
ReCAP	1065.9	1473.2	1778.3	6.0
SaTeen	1120.9	1882.6	2034.1	7.0
OriTTA	1175.0	1499.0	1215.0	5.4

Per frame, OriTTA performs one forward and one backward pass on the input, plus on average $2/3$ of an additional forward-backward pass for L_{align} (20 bank inputs every 30 frames) and a no-gradient re-classification of the K bank inputs every 200 frames, which is batched and accounts for a small share of the time. OriTTA is faster than SAR, ReCAP and SaTeen, which run a second forward and backward pass for sharpness-aware or augmented updates, and slower than TENT and EATA, which update on the input alone, and DeYO, which adds only a forward pass on a patch-shuffled copy.

B Further Related Work

This section complements Section 2 with the lines of work closest to the frame of OriTTA, which the main text only touches on.

Neural collapse and test-time adaptation. Neural collapse (Papyan et al., 2020) describes the terminal phase of training, in which class means converge to a simplex equiangular tight frame aligned with the classifier rows, and it has recently been carried into adaptation. NCTTA (Chen et al.,

2026) observes that the feature of a single sample aligns with the weight vector of its class, and re-aligns normalized features with the normalized classifier weights during adaptation, combined with entropy minimization and with targets that blend geometric proximity and predictive confidence. TCA (Ni et al., 2025) preserves the inter-class topology by pushing pseudo-label centroids toward uniformity, and neural-collapse-inspired alignment has also been used for out-of-distribution generalization (Chen et al., 2024). NEO (Murphy et al., 2026) assumes, from neural collapse under cross-entropy, that the source features are centered at the origin, and re-centers target features there by subtracting their running mean, without updating the model. OriTTA differs from these methods in two respects. First, it does not measure from the origin. The origin is a property of the parameterization, and Table 5 shows that it lies far from where the classes are arranged. The classifier rows also share a common component, 81% of a row's norm on IN-RN50, that no decision reads, so alignment with the raw rows differs from alignment with the centered axes v_c. Second, unlike NEO and TCA, which estimate a center or class centroids from the stream, OriTTA fixes its frame before any test data. On a class-blocked stream a running mean of the features describes the class that is currently presented rather than the shift, whereas the frame of OriTTA does not change.

Centering and angular geometry of features. That centering matters for angular comparisons is known from few-shot classification, where subtracting the mean of the base-class features before normalization improves nearest-centroid classification (Wang et al., 2019). Cosine classifiers with angular margins (Wang et al., 2017; Deng et al., 2019) and classifiers fixed to a simplex ETF (Zhu et al., 2021; Yang et al., 2022) likewise place the classes on a sphere around the origin. OriTTA uses the same geometric intuition at test time, with two differences. The center ω is recovered from the classifier and synthetic prototypes, since the source data are not available, and the angular loss is measured from ω within the subspace the classifier reads.

Reference structures for adaptation. Several methods adapt toward a reference derived from the source model or from the stream. SHOT (Liang et al., 2020) freezes the source classifier and adapts the feature extractor so that target features fit it, through information maximization and pseudo-labels over the whole target set. T3A (Iwasawa & Matsuo, 2021) replaces the classifier rows by prototypes of pseudo-labeled test features. SaTeen (Liu et al., 2026) maximizes the cross-entropy discrepancy between a sample and a structure-disrupted copy of it, and minimizes the reconstruction error of test features in a subspace spanned by incremental PCA on samples it judges reliable. OriTTA shares the premise that a reference exists, but takes it from the classifier alone. Its subspace is the row space of the centered classifier rather than one estimated from reliable samples. Because this subspace is fixed before the stream, it cannot follow the drift it is meant to correct, and it needs no test samples of any class to be formed.

Synthesizing inputs from the source model. Optimizing an input to maximize a class score is an old tool for visualizing classifiers (Simonyan et al., 2014). For data-free knowledge distillation, synthetic inputs have been made realistic with a generator or by matching the batch-normalization statistics of the network (Chen et al., 2019; Yin et al., 2020), and source-free domain adaptation has trained a class-conditional generator together with the source model to synthesize samples for adaptation (Li et al., 2020). The anchors of OriTTA serve a different purpose. They are needed only for where their features land, not for their appearance, so they are synthesized by cross-entropy with a small total-variation term and no statistical prior, which also makes the procedure identical for BatchNorm, GroupNorm and LayerNorm backbones. Appendix A.2 shows that the frame they define does not depend on the particular draw.

Detecting failure without labels. Estimating the accuracy of a model under shift without target labels is a problem of its own (Garg et al., 2022), usually addressed with confidences calibrated on labeled source data. The bank share of Section 3.4 answers a narrower question, whether the predictions have merged onto a few classes, and needs no source data. Because the bank holds one input per class, it is also insensitive to the label order of the stream, which a confidence statistic computed over the stream is not.

Scope. Test-time adaptation of vision–language models, for example by tuning prompts on a single test input (Shu et al., 2022), adapts a different kind of classifier, whose rows are text embeddings,

and is outside the scope of this work. The frame of Section 3.1 requires only a linear head, and extending it to such models is a direction we leave open.

C Firing Rate Across Settings

Table 13 reports how often the collapse detector of Section 3.4 fires in each of the sixteen settings, as the mean number of resets per run over the ten seeds and per corruption. The detector is armed identically everywhere: it is not disabled in the episodic protocol, and it receives no information about corruption boundaries in either protocol.

Table 13: Resets per run, averaged over ten seeds, per corruption. A dot is zero. *Runs* is the number of the ten seeds in which the detector fired at least once. In the continual protocol the corruptions are presented from left to right.

			gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Total	Runs
IN-RN50	episodic	shuffled	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	0.0	0/10
	episodic	class-blocked	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	0.0	0/10
	continual	shuffled	·	·	1.0	·	1.0	·	0.3	0.7	·	0.1	·	0.8	0.5	0.5	0.3	5.2	10/10
	continual	class-blocked	·	·	0.8	0.2	0.8	·	0.9	0.1	·	0.6	0.1	0.1	1.0	·	·	4.6	10/10
IN-ViT	episodic	shuffled	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	0.0	0/10
	episodic	class-blocked	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	0.0	0/10
	continual	shuffled	·	·	·	0.6	0.3	·	·	0.8	0.1	·	·	0.3	0.9	·	·	3.0	10/10
	continual	class-blocked	·	·	0.3	0.7	·	·	·	1.0	·	·	·	0.7	0.3	·	·	3.0	10/10
C100	episodic	shuffled	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	0.0	0/10
	episodic	class-blocked	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	0.0	0/10
	continual	shuffled	·	·	·	·	·	·	·	·	·	·	·	0.9	·	·	·	0.9	9/10
	continual	class-blocked	·	·	·	·	0.5	·	·	·	0.1	0.4	·	0.1	·	0.1	0.1	1.3	10/10
C10	episodic	shuffled	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	0.0	0/10
	episodic	class-blocked	·	·	·	·	·	0.1	·	·	·	0.1	0.1	0.1	0.1	·	·	0.5	4/10
	continual	shuffled	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	0.0	0/10
	continual	class-blocked	·	0.1	·	0.1	·	·	0.2	0.1	0.4	0.1	·	0.9	0.2	·	0.1	2.2	9/10

Table 14: Bank share σ in runs without and with a reset, pooled over both stream orders and ten seeds. The threshold is 0.25.

		max σ, runs without reset	σ at firing (median)
IN-RN50	episodic	0.037	–
	continual	–	0.258
IN-ViT	episodic	0.022	–
	continual	–	0.293
C100	episodic	0.150	–
	continual	0.030	0.320
C10	episodic	0.200	0.300
	continual	0.200	0.300

Three patterns are visible. In the episodic protocol the detector is silent: over the 10 seeds of all four settings it never fires, except on C10 under the class-blocked stream, where it fires in four of ten runs and 0.5 times per run. Within a single corruption the losses are stable enough that there is nothing to catch. Table 14 gives the largest bank share reached in runs where the detector never fires: on ImageNet it stays below 0.04, a seventh of the threshold, and on C100 it reaches 0.15. On C10 the share moves in steps of $1/K = 0.1$, healthy runs reach 0.2, two classes merged into one, and the detector fires at 0.3, three. The threshold therefore separates healthy runs from collapsing ones in every setting, by a wide margin on ImageNet and by one step of the bank's resolution on C10.

In the continual protocol it fires between 0.9 and 5.2 times per run in every setting but C10 under the shuffled stream, where it never fires. The count grows with the number of classes, from 0.9–2.2 on the CIFAR settings to 3.0 on IN-ViT and 4.6–5.2 on IN-RN50, and it is concentrated on impulse noise, glass blur, contrast and elastic transform account for most of the firings on IN-RN50. No run reaches the repeated firing that SAR's own recovery shows under the same protocol (Subsection 4.4).

Finally, the detector fires in nearly every continual run but never more than a few times: with the exception of C10 under the shuffled stream, at least nine of the ten seeds see one reset, and no

setting exceeds 5.2 per run against the 15 corruption boundaries the stream contains. The detector is therefore neither idle nor hair-triggered. Its rate is similar under the two stream orders on ImageNet and C100, since the bank holds one representative per class regardless of the order; C10 is the exception, where it fires only under the class-blocked stream, whose blocks of one thousand images put the most pressure on the classes absent from it (Appendix D.3).

D Ablation Study

D.1 Optimizer Ablation

Every baseline we compare against optimizes with SGD and momentum 0.9, whereas OriTTA uses Adam. To check that the advantage does not come from this choice, we run OriTTA with SGD at the same momentum on IN-RN50, episodic, class-blocked, over the ten seeds. Table 15 gives the result.

Table 15: Adam against SGD, IN-RN50, episodic, class-blocked, ten seeds. The second line of each method is the standard deviation over seeds.

	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
OriTTA + Adam	40.79	42.55	41.81	33.32	31.09	41.83	45.80	53.69	52.67	57.92	72.50	53.12	48.79	58.73	60.05	48.98
sd	0.17	0.24	0.23	0.29	0.38	0.22	0.13	0.20	0.15	0.36	0.11	0.29	0.35	0.21	0.09	0.10
OriTTA + SGD	43.46	44.45	43.40	34.75	34.41	45.04	48.76	55.86	53.47	55.93	71.82	55.08	54.25	62.63	60.61	50.93
sd	0.20	0.24	0.27	0.27	0.37	0.28	0.18	0.19	0.24	1.03	0.19	0.28	0.39	0.23	0.18	0.32
Δ	$+2.67$	$+1.90$	$+1.59$	$+1.43$	$+3.32$	$+3.21$	$+2.96$	$+2.17$	$+0.80$	-1.99	-0.68	$+1.96$	$+5.46$	$+3.90$	$+0.56$	$+1.95$

SGD is ahead by about two points on average and on thirteen of the fifteen corruptions, most on those where the source model is furthest off and a larger step pays, such as elastic transform and pixelate. The advantage of OriTTA over the baselines is therefore not an artifact of Adam. With the baselines' own optimizer it is larger still. We nonetheless use Adam throughout, for three reasons that the same table exposes.

First, stability. The seed-to-seed deviation of the average triples under SGD, and on fog, the corruption on which the baselines are least stable, it grows from 0.36 to 1.03 while the mean falls by two points. At batch size 1 every update is a single-sample gradient. Adam's normalization by $\sqrt{v}$ bounds the step taken on any one sample, whereas SGD passes its full magnitude through. Three times the selected SGD rate already triggers resets within a single corruption.

Second, scale invariance. Adam's update is unchanged when the objective is multiplied by a constant, so the factors $1/r$ and s^2 in our losses, which differ by orders of magnitude across datasets and backbones, never enter the learning rate. This is what allows one rate per normalization type to serve every setting. Under SGD the rate absorbs the scale of the objective and would have to be chosen again for each dataset and backbone. The value selected on IN-RN50 would have to be rescaled for every other setting, since the gradient magnitude of each loss depends on s and r.

Third, batch size. The $\sqrt{B}$ scaling rule that keeps OriTTA's schedule fixed across batch sizes (Appendix A.7) is the rule appropriate for Adam (Malladi et al., 2022). SGD has no rule that holds down to batch size 1: the released baselines scale their rate linearly with the batch, and SAR, DeYO and ReCAP then double it at batch size 1 as a separate correction, while SaTeen does not (Appendix A.10). Each such correction is one more setting to be validated per method and per batch size.

Since the aim of this work is a single procedure with no per-setting tuning, we regard the two points as the price of that property, and Adam as the appropriate choice.

Table 16: Removing one loss term at a time for IN-RN50, episodic, class-blocked.

	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
OriTTA (full)	40.79	42.55	41.81	33.32	31.09	41.83	45.80	53.69	52.67	57.92	72.50	53.12	48.79	58.73	60.05	**48.98**
w/o L_{ang}	26.13	27.30	25.61	22.63	14.42	28.11	29.11	45.44	48.82	35.14	69.48	39.61	25.09	35.00	53.92	35.05 (-13.92)
w/o L_{rad}	40.00	41.44	40.87	32.06	30.05	40.08	44.61	52.98	51.84	57.48	71.70	52.03	47.44	57.64	58.92	47.94 (-1.03)
w/o L_{align}	40.49	41.36	40.71	33.08	30.23	40.19	44.59	52.62	51.53	58.28	72.35	51.85	47.89	58.28	59.35	48.19 (-0.79)

D.2 Losses Ablation

Table 16 removes one term at a time on IN-RN50. The angular loss does most of the work. Without it accuracy falls by fourteen points, by a third to a half on most corruptions and least on brightness, frost and JPEG compression, where the source model is already accurate. It is the only term that assigns a reliable feature to a class and separates it from the others, and the other two cannot take its place. The radial loss contributes about a point, spread evenly across the corruptions. The alignment loss contributes the least here. Its role is to hold in place the classes the stream is not presenting, and on ImageNet a class is absent only while fifty images of another class pass. Appendix D.3 lengthens that interval.

D.3 Alignment Loss on Long Class Blocks

The alignment loss exists for the classes that the stream is not currently presenting: it pulls a stored representative of every class back to its ideal point, so that the model does not drift away from a class while another one occupies the stream. On IN-RN50 this pressure is weak even under the class-blocked order, because each class arrives as a block of only fifty images, and removing L_{align} there costs 0.8 points (Table 16). The CIFAR benchmarks lengthen the blocks to one hundred images on C100 and one thousand on C10, and we repeat the ablation on both, episodic and class-blocked, over ten seeds (Tables 17 and 18).

Table 17: Removing L_{align}, episodic, class-blocked. *Block* is the number of consecutive images of one class and *resets* is the mean number of detector firings per run and the number of runs, out of ten, in which it fired.

		accuracy			resets per run	
	block	full	w/o L_{align}	Δ	full	w/o L_{align}
IN-RN50	50	48.98	48.19	-0.79	0 (0/10)	0 (0/10)
C100	100	61.69	60.27	-1.41	0 (0/10)	0 (0/10)
C10	1000	68.96	60.94	-8.02	0.5 (4/10)	37.1 (10/10)

Table 18: Per-corruption accuracy without L_{align}, episodic, class-blocked, ten seeds. The line below each result is the standard deviation over seeds.

C10	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
OriTTA	46.96	52.50	39.10	73.22	54.49	78.07	76.05	78.28	72.51	83.49	91.52	75.95	77.83	61.40	73.03	**68.96**
sd	4.01	2.63	3.11	2.79	1.76	2.19	1.99	1.83	1.31	1.30	0.76	1.52	1.48	3.99	2.02	1.51
w/o L_{align}	25.06	31.61	26.24	63.72	44.72	73.25	68.00	80.04	64.90	83.16	93.09	64.91	78.54	43.28	73.55	60.94
sd	1.40	3.66	0.77	1.39	3.19	1.67	2.72	1.10	1.84	1.41	0.83	3.02	1.40	2.56	1.72	0.71
Δ	-21.91	-20.90	-12.86	-9.49	-9.77	-4.82	-8.05	$+1.77$	-7.60	-0.33	$+1.57$	-11.04	$+0.71$	-18.12	$+0.52$	-8.02

C100	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
OriTTA	48.07	52.29	51.20	70.98	54.74	69.73	70.81	63.13	63.30	58.71	71.93	63.51	62.84	64.83	59.24	**61.69**
sd	0.53	0.49	0.33	0.41	0.63	0.55	0.30	0.42	0.40	0.59	0.34	0.72	0.57	0.64	0.37	0.33
w/o L_{align}	45.27	50.12	48.89	70.13	52.93	68.88	70.01	61.94	61.97	57.30	71.12	62.24	61.55	63.70	58.05	60.27
sd	0.79	0.53	0.52	0.44	0.50	0.57	0.34	0.39	0.53	0.52	0.39	0.94	0.59	0.60	0.40	0.37
Δ	-2.80	-2.16	-2.31	-0.85	-1.81	-0.84	-0.80	-1.19	-1.33	-1.42	-0.81	-1.27	-1.29	-1.12	-1.19	-1.41

The cost grows with the block length. On C100, removing L_{align} lowers accuracy on every one of the fifteen corruptions, by 1.4 points on average, and the detector never fires with or without it. On C10 the effect changes in kind. Without L_{align} the model collapses repeatedly: the detector fires 37 times per run instead of 0.5, in every seed rather than four of ten, and accuracy falls by eight points. The losses concentrate where the source model is weakest and adaptation matters most, on the three noise corruptions and on pixelate, by 13 to 22 points, while on snow, brightness, elastic transform and JPEG, where the source model is already accurate, returning to it often costs nothing. Across the three settings the cost of removing L_{align} rises with the length of a class block, from 0.8 points at fifty images to 1.4 at one hundred and 8.0 with repeated collapse at one thousand. On C10 the alignment loss is also applied most often relative to the number of classes: with $K = 10$, every class is pulled back at every alignment step, whereas on ImageNet each class is visited once every $1{,}500$ frames.

The detector does not replace the alignment loss. The standard deviation on C10 falls from 1.51 to 0.71 without L_{align}, but this is not greater stability: with a reset about every 4,000 frames, the model spends most of the stream near its source weights, and its accuracy is consistently lower rather than consistently good. The detector recovers the model after each collapse, yet eight points are lost despite thirty-seven recoveries. The two mechanisms therefore play different roles: L_{align} keeps the classes in place and prevents collapse, and the detector restores the model when collapse occurs nonetheless.

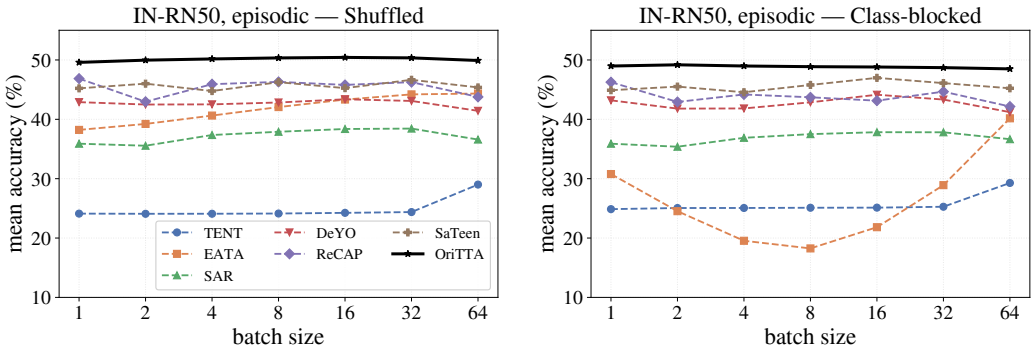


Figure 2: Accuracy against batch size on IN-RN50, episodic, averaged over ten seeds.

D.4 Batch Size Ablation

Figure 2 varies the batch size from 1 to 64 on IN-RN50 in the episodic protocol mean over ten seeds, with every method's learning rate scaled by its own published rule (Appendix A.7). OriTTA is the most accurate method at every batch size on both streams, and its accuracy is nearly flat: it moves within 0.8 points on the shuffled stream and 0.5 on the class-blocked one. A larger batch does not close the gap. DeYO, ReCAP and SaTeen gain little from larger batches, whereas EATA improves by six points on the shuffled stream but falls below 20% at batch sizes 4 to 8 on the class-blocked stream. OriTTA remains stable across the evaluated batch sizes and stream orders. Full per-corruption results at batch size 64 are given in Appendix E.

D.5 Ridge Weight λ

The recentered origin (Eq. 5) minimizes the sum of squared distances to the K hyperplanes H_c together with a ridge toward the prototype centroid $\bar{z}^a$ of weight $\lambda = K/100$. The rule follows from how the reliability of the two terms changes with K.

The hyperplane term is a least-squares system of K equations in D unknowns. When K is small it is far from determined: on C10 it constrains nine directions of a 640-dimensional space, so its solutions form a large affine set that can be satisfied exactly at no cost, and every point of that set is a point at which the classifier is maximally ambiguous. The centroid of ten anchors adds little reliable information beyond selecting a point within that set, and a small λ lets it do exactly that and no more. When K is large the situation is reversed. A thousand constraints pin the least-squares solution to a specific point, and nothing in the system requires that point to lie where features live. On ViT-B/16, where $K-1 > D$ and the system is over-determined, the unregularized solution has norm 1.7 against a feature scale of 80 and is nearly orthogonal to the feature cloud. There the centroid of a thousand prototypes is the more trustworthy of the two, and a large λ lets it dominate. With the per-constraint weight fixed, the effect of the ridge still differs with K, because the hyperplane system does: at small K it is underdetermined and its residual vanishes on a whole affine set, so the ridge only selects a point within that set, whereas at large K no point annihilates the residual and the ridge trades against it directly.

Table 19 tests the small-K end of this argument by replacing the rule on C10 with the ImageNet value $\lambda = 10$, over ten seeds. Trusting the ten-prototype centroid costs ten points on both streams and on every corruption, and on the class-blocked stream the seed-to-seed deviation grows as well. At the large-K end the rule is insensitive: on ImageNet, raising λ from 10 to 100 or dropping the hyperplane term altogether moves ω by less than 0.06 s and leaves accuracy unchanged.

Table 19: C10, episodic, ten seeds, with λ set by the rule ($K/100 = 0.1$) or fixed at the ImageNet value 10. The line below each result is the standard deviation over seeds.

shuffled	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
$\lambda = K/100$	56.14	60.55	41.11	80.25	59.54	82.02	82.34	80.75	76.99	84.44	91.32	80.58	77.84	73.40	73.43	**73.38**
sd	0.30	0.34	0.26	0.40	0.40	0.31	0.19	0.19	0.31	0.24	0.09	0.19	0.08	0.55	0.16	0.08
$\lambda = 10$	43.82	48.71	29.60	67.75	50.51	71.22	70.63	72.90	68.64	76.01	86.84	67.87	68.48	58.96	63.57	63.03
sd	0.28	0.60	0.59	0.58	0.38	0.53	0.58	0.33	0.34	0.29	0.28	0.52	0.40	0.75	0.28	0.21
Δ	−12.32	−11.84	−11.51	−12.50	−9.03	−10.80	−11.71	−7.85	−8.35	−8.43	−4.48	−12.71	−9.36	−14.44	−9.86	−10.35
class-blocked	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
$\lambda = K/100$	46.96	52.50	39.10	73.22	54.49	78.07	76.05	78.28	72.51	83.49	91.52	75.95	77.83	61.40	73.03	**68.96**
sd	4.01	2.63	3.11	2.79	1.76	2.19	1.99	1.83	1.31	1.30	0.76	1.52	1.48	3.99	2.02	1.51
$\lambda = 10$	38.04	41.09	25.39	59.53	47.97	66.38	63.86	71.50	64.49	75.12	87.09	58.99	68.62	46.33	63.38	58.52
sd	5.42	3.99	2.49	2.73	3.68	2.98	2.93	2.12	2.94	2.53	1.51	4.37	3.29	3.72	2.48	2.27
Δ	−8.92	−11.41	−13.71	−13.69	−6.52	−11.69	−12.19	−6.78	−8.02	−8.37	−4.43	−16.96	−9.21	−15.07	−9.65	−10.44

E FULL RESULTS

Tables 20–23 give the accuracy on every corruption for every setting and protocol, as the mean over ten seeds with the standard deviation over seeds as a subscript. Within each block, the best result in each column is shown in bold and the second best is underlined. In the continual protocol the model adapts through the corruptions from left to right without any reset at corruption boundaries. Table 24 gives the full results for IN-RN50 at batch size 64.

Table 20: Full results on ImageNet-C, ResNet-50-GN.

	episodic	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
shuffled	TENT	12.35±2.55	20.22±2.34	13.91±1.47	13.16±0.27	6.46±3.03	18.07±0.97	18.23±0.68	16.00±0.85	20.65±0.70	1.48±0.10	69.80±0.07	42.04±0.25	6.22±0.49	49.38±0.21	53.75±0.09	24.11±0.40
	EATA	29.49±1.04	31.55±1.85	30.23±1.76	17.93±0.80	19.32±0.53	30.45±1.00	32.62±0.88	42.73±1.94	41.21±1.60	45.55±0.56	69.40±0.10	43.66±0.53	32.98±0.55	52.13±0.45	54.02±0.35	38.22±0.21
	SAR	28.63±0.26	31.01±0.22	29.40±0.32	18.46±0.13	15.23±0.28	28.82±0.26	30.43±0.34	44.75±0.60	44.39±0.12	33.60±1.05	71.95±0.09	44.61±0.13	13.03±1.06	47.83±0.21	56.24±0.04	35.89±0.11
	DeYO	40.87±0.30	43.33±0.18	42.08±0.31	22.07±0.26	23.69±0.17	38.63±0.25	28.12±8.19	52.26±0.31	50.21±0.24	18.88±25.72	73.19±0.07	51.27±0.18	42.97±0.54	57.31±0.18	58.54±0.07	42.89±1.95
	ReCAP	42.07±0.26	44.35±0.29	43.17±0.25	24.69±0.27	25.05±0.27	41.73±0.30	39.64±2.44	54.67±0.35	52.16±0.18	46.54±22.42	73.33±0.09	53.52±0.14	41.54±13.41	60.27±0.24	60.20±0.16	46.86±1.52
	SaTeen	**43.20**±0.14	**45.44**±0.17	**44.25**±0.16	25.80±0.45	26.25±0.50	**42.44**±0.33	26.68±10.40	**55.29**±0.12	52.69±0.15	19.30±27.47	**73.34**±0.08	**54.00**±0.13	48.71±0.44	**60.77**±0.22	60.13±0.17	45.22±2.20
	OriTTA	41.11±0.20	43.15±0.12	42.22±0.11	**33.53**±0.16	**31.93**±0.22	42.38±0.15	**46.33**±0.15	54.32±0.16	**53.27**±0.20	**59.04**±0.18	72.80±0.06	53.79±0.13	**49.64**±0.21	59.77±0.17	**60.62**±0.11	**49.59**±0.05
class-blocked	TENT	15.59±3.67	22.05±4.03	17.00±3.45	13.21±0.30	7.87±4.97	18.40±2.31	18.53±0.91	17.01±1.52	20.43±1.09	1.50±0.15	69.84±0.13	42.12±0.28	6.22±0.60	49.45±0.37	53.80±0.21	24.87±0.80
	EATA	25.27±0.50	25.71±1.05	25.28±0.45	15.68±0.78	14.93±0.60	21.35±0.98	22.70±0.51	33.11±0.70	32.90±0.95	37.47±0.51	65.32±0.63	34.82±0.80	18.24±0.76	38.49±1.03	50.43±0.62	30.78±0.19
	SAR	28.53±0.18	31.11±0.31	29.44±0.37	18.56±0.44	15.14±0.47	28.83±0.22	30.53±0.22	44.16±1.40	44.36±0.22	33.30±1.52	71.93±0.05	44.67±0.17	13.87±1.72	47.73±0.30	56.22±0.08	35.89±0.15
	DeYO	40.89±0.16	43.18±0.19	42.19±0.24	21.84±0.30	23.44±0.38	38.64±0.50	33.34±7.50	52.08±0.58	50.35±0.39	18.67±25.79	73.19±0.09	51.26±0.17	42.93±0.74	57.44±0.31	58.51±0.14	43.20±1.74
	ReCAP	42.13±0.24	44.41±0.35	43.39±0.38	24.72±0.47	24.72±0.62	41.78±0.39	39.92±2.35	54.62±0.40	52.25±0.36	35.69±27.54	**73.44**±0.12	53.63±0.27	42.67±11.09	60.32±0.42	**60.22**±0.10	46.26±2.10
	SaTeen	**43.13**±0.15	**45.48**±0.20	**44.31**±0.29	25.58±0.54	25.96±0.56	**42.34**±0.26	27.82±8.78	**55.28**±0.37	**52.74**±0.27	13.48±23.72	73.41±0.12	**54.00**±0.14	**48.85**±0.85	**61.00**±0.36	60.14±0.10	44.90±1.84
	OriTTA	40.79±0.17	42.55±0.24	41.81±0.23	**33.32**±0.29	**31.09**±0.38	41.83±0.22	**45.80**±0.13	53.69±0.20	52.67±0.15	**57.92**±0.36	72.50±0.11	53.12±0.29	48.79±0.35	58.73±0.21	60.05±0.09	**48.98**±0.10

	continual	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
shuffled	TENT	12.35±2.55	0.74±0.23	0.22±0.05	2.16±0.58	0.26±0.09	0.22±0.02	0.31±0.06	0.57±0.05	0.18±0.02	0.20±0.01	1.48±0.49	0.31±0.15	0.10±0.00	0.16±0.01	0.79±0.20	1.34±0.28
	EATA	29.49±1.04	29.30±3.85	25.44±4.01	12.94±2.20	14.77±2.13	18.41±3.53	23.08±3.19	26.36±4.22	27.85±5.39	31.02±6.70	54.08±2.27	24.44±2.40	21.43±2.36	32.12±6.61	38.49±2.63	27.28±1.91
	SAR	28.63±0.25	41.23±0.14	44.45±0.07	15.66±0.25	19.02±0.35	9.66±0.78	10.66±2.84	5.55±6.76	3.04±1.35	1.21±0.36	6.33±0.65	1.39±0.73	0.48±0.38	0.98±0.43	1.94±0.70	12.68±0.59
	DeYO	40.87±0.30	48.80±0.09	47.90±0.12	10.76±1.03	13.74±0.80	13.96±4.12	6.48±3.75	1.53±0.75	0.73±0.18	0.44±0.11	1.76±0.47	0.26±0.05	0.20±0.03	0.19±0.03	0.23±0.03	12.52±0.55
	ReCAP	42.03±0.25	**49.33**±0.16	**48.41**±0.17	**24.02**±7.68	3.50±4.30	3.91±8.72	2.70±6.65	3.55±10.58	3.85±11.75	5.13±15.80	6.98±21.51	4.96±15.30	1.72±5.02	4.31±13.19	4.65±14.27	13.94±8.28
	SaTeen	**43.20**±0.14	48.90±0.19	47.61±0.20	19.60±8.54	1.17±1.70	0.19±0.07	0.19±0.12	0.16±0.07	0.12±0.02	0.11±0.02	0.12±0.03	0.11±0.02	0.11±0.02	0.11±0.02	0.11±0.02	10.79±0.67
	OriTTA	41.09±0.19	43.78±0.74	38.40±0.98	22.63±5.79	**21.81**±3.38	**36.39**±3.28	**35.26**±1.66	**50.72**±2.73	**48.64**±0.59	**42.97**±1.25	**63.05**±3.33	**39.22**±4.98	**40.62**±7.53	**33.09**±13.57	**54.44**±9.66	**40.81**±1.02
class-blocked	TENT	15.59±3.67	1.08±0.43	0.27±0.07	2.73±0.66	0.36±0.11	0.27±0.05	0.38±0.07	0.63±0.10	0.19±0.02	0.22±0.02	1.97±0.52	0.41±0.11	0.10±0.00	0.17±0.02	0.97±0.22	1.69±0.39
	EATA	25.27±0.50	11.04±2.07	7.49±0.99	6.19±1.62	5.24±2.38	5.26±1.45	6.21±2.29	4.71±1.22	5.98±1.98	5.36±2.90	12.14±4.26	2.29±1.25	4.36±2.28	5.29±1.61	6.47±2.75	7.55±1.19
	SAR	28.53±0.18	41.14±0.14	44.47±0.09	15.76±0.30	19.17±0.41	11.52±0.61	11.18±5.27	15.79±15.25	18.37±19.91	18.04±24.14	34.59±28.87	20.35±23.10	5.68±6.21	18.72±22.66	20.27±24.03	21.57±10.86
	DeYO	40.89±0.16	48.73±0.11	47.86±0.12	11.55±4.05	11.50±4.22	11.82±5.74	6.89±4.78	1.88±1.22	0.77±0.39	0.47±0.19	2.02±1.40	0.27±0.09	0.20±0.07	0.20±0.05	0.24±0.10	12.35±0.84
	ReCAP	42.15±0.26	**49.28**±0.22	**48.43**±0.18	18.17±10.74	3.23±5.37	0.81±1.51	0.31±0.22	0.26±0.13	0.17±0.04	0.15±0.04	0.24±0.08	0.14±0.01	0.13±0.03	0.15±0.02	0.17±0.06	10.92±1.02
	SaTeen	**43.13**±0.15	48.78±0.23	47.57±0.18	24.45±5.96	3.17±4.71	3.59±9.03	2.04±5.38	2.06±5.88	0.95±2.65	0.24±0.38	0.16±0.12	0.10±0.01	0.12±0.02	0.10±0.04	0.10±0.02	11.77±1.78
	OriTTA	40.74±0.16	44.45±0.37	39.26±1.70	**25.80**±6.03	**24.91**±5.43	**33.45**±5.97	**32.44**±5.53	**52.02**±2.27	**48.77**±1.37	**45.97**±4.55	**66.77**±5.31	**45.32**±4.59	**39.96**±4.40	**41.61**±4.67	**48.27**±4.20	**41.98**±0.78

Table 21: Full results on ImageNet-C, ViT-B/16.

	episodic	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
shuffled	TENT	44.66±0.22	42.82±0.23	45.42±0.27	52.51±0.17	48.10±0.22	55.55±0.14	50.49±0.28	15.81±0.91	17.72±1.17	66.51±0.27	75.09±0.07	64.82±0.12	53.07±0.34	67.04±0.06	64.57±0.09	50.95±0.13
	EATA	50.24±0.27	49.48±0.34	51.22±0.21	56.69±0.08	55.76±0.18	61.04±0.16	57.94±0.16	64.05±0.21	63.74±0.19	72.08±0.12	77.47±0.07	67.50±0.19	64.61±0.12	71.78±0.09	69.11±0.08	62.18±0.07
	SAR	45.39±0.26	42.95±0.47	46.23±0.33	53.68±0.12	50.47±0.22	57.64±0.17	52.94±0.14	58.78±0.33	56.00±4.06	68.57±0.77	75.28±0.94	65.65±0.08	58.16±0.24	68.67±0.87	66.41±0.11	57.79±0.32
	DeYO	53.01±0.21	52.75±0.18	53.99±0.09	57.66±0.08	57.44±0.15	62.29±0.12	49.17±9.60	65.76±0.16	64.78±0.11	72.63±0.11	77.75±0.09	67.89±0.13	66.19±0.14	72.55±0.08	69.78±0.06	62.91±0.62
	ReCAP	55.04±0.22	55.19±0.34	56.05±0.17	59.02±0.16	59.45±0.14	64.69±0.11	56.49±9.32	68.12±0.16	66.60±0.21	73.83±0.10	**78.43**±0.05	68.07±0.12	68.97±0.10	74.06±0.08	71.04±0.11	65.00±0.66
	SaTeen	55.77±0.20	56.47±0.17	**57.01**±0.21	**59.13**±0.08	**59.84**±0.15	**65.05**±0.13	45.92±12.01	**68.75**±0.09	**66.89**±0.18	**73.99**±0.11	78.38±0.05	68.04±0.15	69.62±0.15	**74.17**±0.12	**71.21**±0.12	64.68±0.81
	OriTTA	**55.81**±0.12	**56.77**±0.18	56.80±0.10	58.64±0.15	59.64±0.13	64.92±0.08	**64.64**±0.15	68.39±0.18	66.53±0.27	73.26±0.09	76.90±0.07	**68.13**±0.12	**70.01**±0.15	73.13±0.11	70.11±0.15	**65.58**±0.05
class-blocked	TENT	44.75±0.23	42.94±0.37	45.57±0.14	52.63±0.15	48.07±0.09	55.66±0.08	50.46±0.19	16.91±2.45	19.09±1.99	66.55±0.05	75.08±0.14	64.90±0.09	52.86±0.85	67.05±0.16	64.61±0.16	51.14±0.23
	EATA	29.56±0.28	27.25±0.97	30.81±0.19	43.45±0.42	36.96±0.45	46.07±0.49	39.27±0.69	44.09±0.78	47.04±0.56	57.73±0.29	69.89±0.23	54.69±1.33	42.08±0.53	58.61±0.50	56.93±0.23	45.63±0.10
	SAR	33.03±2.20	30.51±2.78	34.86±2.01	43.40±0.98	37.71±1.71	46.28±0.81	39.38±0.97	38.64±0.91	42.71±1.89	54.24±0.58	63.50±0.21	55.24±0.41	38.96±0.70	54.65±0.29	55.63±0.38	44.58±0.42
	DeYO	53.27±0.19	53.07±0.18	54.23±0.21	57.83±0.25	57.68±0.10	62.55±0.11	47.21±16.81	66.09±0.13	64.79±0.20	72.68±0.12	77.77±0.09	67.95±0.16	66.36±0.20	72.62±0.21	69.86±0.15	62.93±1.09
	ReCAP	55.38±0.38	55.72±0.34	56.46±0.45	59.47±0.30	59.89±0.33	65.03±0.22	38.03±24.44	68.49±0.23	66.97±0.32	73.99±0.15	**78.53**±0.05	67.29±2.62	69.40±0.18	74.29±0.13	71.37±0.15	64.02±1.56
	SaTeen	56.40±0.23	57.19±0.22	57.42±0.29	**59.73**±0.32	60.34±0.21	**65.42**±0.20	54.36±11.76	**69.07**±0.13	**67.21**±0.30	**74.12**±0.11	78.52±0.04	58.86±21.65	70.02±0.17	**74.38**±0.10	**71.51**±0.17	64.97±1.20
	OriTTA	**56.41**±0.18	**57.36**±0.21	**57.47**±0.13	59.35±0.26	**60.41**±0.15	65.29±0.28	**64.85**±0.33	68.97±0.19	67.15±0.14	74.05±0.28	77.85±0.15	**69.14**±0.24	**70.45**±0.20	73.77±0.11	70.98±0.21	**66.23**±0.10

	continual	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
shuffled	TENT	39.90±0.22	47.44±0.10	52.29±0.09	47.79±0.23	49.56±0.14	54.62±0.10	51.26±0.13	56.18±0.15	59.32±0.15	63.91±0.19	75.06±0.05	61.24±0.15	37.04±7.54	14.46±19.01	7.75±20.28	47.85±2.96
	EATA	45.83±0.22	54.55±0.22	56.06±0.13	52.55±0.14	54.77±0.33	57.62±0.21	55.17±0.27	59.07±0.13	61.47±0.22	68.02±0.34	75.56±0.14	62.37±0.33	60.23±0.49	68.90±0.35	66.26±0.29	59.90±0.13
	SAR	38.79±0.27	50.85±0.15	55.91±0.07	50.59±0.13	54.02±0.08	58.05±0.08	55.18±0.10	59.52±0.07	62.75±0.06	66.93±0.80	75.79±0.21	63.05±0.50	60.78±0.66	67.97±0.91	67.77±0.38	59.20±0.10
	DeYO	49.73±0.16	56.47±0.10	58.31±0.05	53.95±0.13	57.11±0.14	60.30±0.14	36.52±7.43	60.17±0.89	63.06±0.17	69.37±0.16	76.73±0.05	64.74±0.11	64.67±0.19	70.68±0.14	68.73±0.10	60.70±0.54
	ReCAP	52.51±0.33	**59.09**±0.10	**59.76**±0.12	**55.32**±0.30	**59.24**±0.12	**62.61**±0.21	20.24±9.21	62.45±3.96	**65.01**±0.58	**71.22**±0.15	**77.24**±0.09	**65.84**±0.14	**67.68**±0.20	**72.28**±0.09	**70.04**±0.10	61.37±0.82
	SaTeen	46.13±0.23	55.79±0.08	58.15±0.08	53.39±0.21	56.93±0.14	60.53±0.15	51.76±3.19	61.61±1.18	63.87±0.25	69.76±0.13	77.09±0.04	65.17±0.21	64.34±0.16	70.87±0.11	68.90±0.07	61.62±0.32
	OriTTA	**55.78**±0.10	58.03±0.22	57.19±0.48	51.37±5.18	55.14±2.65	60.66±1.89	**60.56**±1.46	**63.17**±5.87	64.10±2.09	70.71±0.83	75.57±0.38	65.13±1.39	66.68±6.03	71.20±0.37	68.42±0.16	**62.91**±0.96
class-blocked	TENT	40.13±0.07	47.48±0.18	52.29±0.13	47.75±0.41	49.75±0.23	54.64±0.11	51.17±0.27	56.29±0.14	59.36±0.05	63.93±0.22	75.06±0.02	61.34±0.19	39.92±7.46	19.51±25.62	14.19±28.78	48.85±4.09
	EATA	26.49±0.12	30.74±0.44	27.36±1.44	16.04±2.10	6.15±1.90	4.16±0.41	3.02±0.60	3.98±0.57	3.46±0.99	1.37±0.44	6.34±0.58	0.49±0.13	1.98±0.96	1.85±0.83	2.47±0.75	9.06±0.31
	SAR	30.92±1.37	38.33±3.45	44.52±1.97	42.10±0.59	41.76±0.44	50.25±1.42	44.51±1.07	47.98±1.47	53.06±1.21	59.97±0.75	74.45±0.70	56.31±0.89	46.65±0.95	62.30±0.69	63.05±0.71	50.41±0.92
	DeYO	49.95±0.30	56.57±0.12	58.28±0.07	53.97±0.19	57.05±0.13	60.25±0.19	34.97±5.42	59.70±0.87	63.04±0.17	69.53±0.13	76.75±0.08	64.73±0.17	64.82±0.24	70.62±0.14	68.72±0.15	60.60±0.34
	ReCAP	52.78±0.49	**59.16**±0.16	**59.82**±0.15	**55.59**±0.33	**59.21**±0.20	**62.57**±0.21	19.77±4.96	59.41±6.83	64.23±1.52	**71.13**±0.35	**77.21**±0.13	**65.74**±0.37	**67.82**±0.24	**72.33**±0.21	**70.15**±0.09	61.13±0.75
	SaTeen	46.37±0.37	55.79±0.14	58.16±0.06	53.42±0.39	56.89±0.12	60.52±0.15	51.97±5.58	61.39±2.09	63.92±0.36	69.90±0.18	77.07±0.02	65.12±0.08	64.40±0.32	70.83±0.18	68.87±0.12	61.64±0.56
	OriTTA	**56.34**±0.23	56.07±0.07	54.79±0.53	55.22±0.46	56.78±0.58	58.58±0.78	**55.88**±0.22	**64.50**±1.49	**64.68**±0.21	68.39±0.49	74.28±0.09	64.03±5.46	65.55±0.39	69.28±1.27	65.61±0.47	**62.00**±0.30

Table 22: Full results on CIFAR-100-C, WRN-28-10.

	episodic	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
shuffled	TENT	4.03±1.15	4.60±0.97	3.90±0.75	53.51±8.67	4.68±2.49	44.72±15.34	58.34±7.00	26.88±10.40	19.21±7.94	6.87±3.52	68.55±0.91	2.53±0.81	27.61±7.16	39.61±17.44	22.44±7.88	25.83±1.73
	EATA	44.73±1.21	47.42±1.21	48.71±0.81	63.27±0.98	46.40±0.80	62.02±1.15	63.46±1.01	54.89±1.08	55.22±0.77	49.52±1.42	66.46±0.65	52.04±1.11	52.99±1.01	59.59±0.91	50.59±0.77	54.49±0.27
	SAR	24.67±0.89	29.86±1.35	29.44±1.00	62.24±1.99	35.97±0.81	60.83±1.04	62.11±1.05	50.20±1.32	46.89±1.85	36.85±1.11	64.84±0.78	28.90±1.29	50.81±1.12	38.01±2.85	47.08±1.20	44.58±0.24
	DeYO	5.46±1.93	5.17±2.84	6.98±2.91	15.89±2.92	2.79±1.47	12.37±3.86	14.44±4.73	13.04±3.39	13.24±3.69	6.98±3.35	25.65±5.01	3.35±1.50	8.06±2.58	15.93±5.48	8.00±3.03	10.49±1.02
	ReCAP	1.44±0.14	1.68±0.24	1.50±0.41	2.88±0.74	1.52±0.36	2.37±0.70	3.10±0.89	2.48±0.58	2.04±0.43	1.30±0.18	3.16±0.81	1.33±0.16	2.16±0.56	2.43±0.70	1.95±0.34	2.09±0.15
	SaTeen	1.38±0.13	1.50±0.13	1.76±0.42	2.16±0.47	1.38±0.24	2.33±0.37	2.16±0.22	2.12±0.27	2.01±0.33	1.37±0.15	3.13±0.32	1.30±0.13	1.98±0.52	2.28±0.52	1.91±0.24	1.92±0.10
	OriTTA	**49.66**±0.44	**53.17**±0.20	**53.11**±0.19	**70.64**±0.15	**55.62**±0.25	**69.51**±0.16	**70.61**±0.20	**63.31**±0.17	**63.29**±0.20	**59.11**±0.22	**71.84**±0.20	**64.81**±0.21	**62.74**±0.22	**65.62**±0.33	**59.42**±0.16	**62.16**±0.09
class-blocked	TENT	3.28±1.82	5.01±2.23	4.78±1.72	36.34±7.30	8.38±3.56	28.48±8.12	32.73±8.61	17.32±4.37	16.53±9.61	6.87±3.00	52.51±9.06	9.24±1.93	14.53±5.00	20.90±6.39	15.81±4.63	18.18±1.18
	EATA	32.80±2.45	38.07±0.80	36.74±2.29	63.44±2.59	39.73±2.76	60.96±1.86	62.19±1.41	52.56±0.52	51.89±1.54	43.65±2.93	63.28±2.34	40.28±5.61	53.51±1.54	45.24±3.53	50.40±1.22	48.98±0.44
	SAR	27.36±1.30	32.52±1.54	33.40±1.84	65.13±1.53	38.63±1.31	62.15±1.94	64.33±1.56	52.60±1.11	48.14±1.51	39.45±1.79	65.72±0.67	33.64±1.20	51.95±2.06	40.97±1.53	50.56±1.14	47.10±0.56
	DeYO	8.10±2.17	7.67±1.88	9.85±3.54	17.42±6.80	4.02±0.99	9.64±2.51	20.26±6.92	10.43±2.71	14.73±2.33	10.04±2.52	29.72±2.99	8.78±2.64	8.17±2.45	16.55±5.99	11.79±3.45	12.48±1.56
	ReCAP	3.52±1.23	4.08±3.63	3.98±2.69	11.36±9.23	2.25±1.36	7.05±1.75	15.41±9.82	7.91±4.67	6.86±4.61	3.83±2.58	24.43±9.66	2.81±0.51	5.40±2.11	7.00±5.92	6.37±2.16	7.48±1.58
	SaTeen	3.54±2.14	3.84±2.82	2.94±1.91	6.81±3.58	3.20±1.71	6.33±1.21	6.36±2.36	7.07±2.24	5.79±1.69	2.17±1.19	7.97±2.71	3.50±2.11	4.23±1.21	5.14±1.07	5.17±1.80	4.94±0.81
	OriTTA	**48.07**±0.53	**52.29**±0.49	**51.20**±0.33	**70.98**±0.41	**54.74**±0.63	**69.73**±0.55	**70.81**±0.30	**63.13**±0.42	**63.30**±0.40	**58.71**±0.59	**71.93**±0.34	**63.51**±0.72	**62.84**±0.57	**64.83**±0.64	**59.24**±0.37	**61.69**±0.33

	continual	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
shuffled	TENT	4.17±1.50	1.12±0.06	1.03±0.01	1.26±0.16	1.00±0.00	1.00±0.00	1.01±0.01	1.00±0.00	1.00±0.00	1.00±0.00	1.01±0.01	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.24±0.11
	EATA	45.03±1.12	43.64±1.45	37.68±2.84	44.59±2.21	29.14±2.39	37.51±3.55	36.11±3.65	27.68±5.59	29.28±3.39	19.86±2.71	28.33±5.78	10.13±4.99	15.79±2.98	19.05±3.49	14.41±1.80	29.22±1.37
	SAR	24.12±0.56	17.85±6.93	14.46±2.88	44.81±8.92	12.82±5.62	32.99±10.34	31.20±13.75	18.67±10.20	15.17±10.78	13.18±8.79	24.97±13.20	13.86±8.59	12.69±10.40	12.98±5.77	15.67±9.95	20.36±7.07
	DeYO	5.84±1.78	4.19±2.74	2.26±0.86	2.53±1.63	1.33±0.37	1.41±0.77	1.40±0.40	1.26±0.30	1.19±0.25	1.02±0.08	1.32±0.27	1.04±0.11	1.25±0.29	1.22±0.23	1.12±0.09	1.89±0.55
	ReCAP	1.46±0.14	1.09±0.16	1.02±0.04	1.25±0.36	1.00±0.01	0.99±0.03	1.00±0.00	1.04±0.08	1.00±0.00	1.00±0.00	1.00±0.01	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.06±0.03
	SaTeen	1.34±0.14	1.07±0.17	1.01±0.02	1.07±0.16	1.05±0.09	1.03±0.04	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.04±0.01
	OriTTA	**49.63**±0.43	**57.69**±0.15	**54.24**±0.24	**68.21**±0.14	**47.86**±2.03	**66.47**±0.97	**68.61**±0.24	**59.51**±0.96	**61.51**±0.30	**56.09**±0.21	**66.12**±0.22	**57.01**±3.22	**54.51**±3.07	**58.49**±1.34	**55.54**±1.10	**58.77**±0.28
class-blocked	TENT	3.28±1.82	1.04±0.02	1.01±0.01	1.27±0.10	1.00±0.00	1.00±0.00	1.01±0.02	1.00±0.00	1.00±0.00	1.00±0.00	1.01±0.01	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.17±0.12
	EATA	32.80±2.45	32.90±1.82	27.20±1.75	40.04±2.30	20.23±2.23	24.50±4.31	20.46±5.43	15.39±3.11	11.82±2.49	9.85±1.35	12.39±2.40	3.64±2.34	7.09±2.01	6.70±3.15	5.75±0.83	18.05±0.68
	SAR	27.36±1.30	20.36±10.95	18.34±8.35	57.65±5.89	20.46±7.74	45.66±10.03	48.01±9.72	29.63±9.02	26.40±12.71	23.65±6.62	45.68±9.42	22.24±6.54	30.45±5.79	24.36±7.63	28.31±5.89	31.24±6.53
	DeYO	8.10±2.17	2.71±0.70	2.71±1.75	1.95±0.47	1.89±0.32	1.14±0.39	1.44±0.15	1.22±0.18	1.42±0.28	1.13±0.31	1.25±0.38	1.00±0.00	0.73±0.37	0.88±0.28	1.07±0.13	1.91±0.26
	ReCAP	3.52±1.23	0.83±0.63	1.28±0.41	2.23±1.33	1.00±0.00	1.00±0.00	1.03±0.06	1.26±0.37	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.01	1.28±0.13
	SaTeen	3.54±2.14	1.14±0.29	1.09±0.20	1.95±1.14	1.18±0.41	1.00±0.01	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.73±1.63	1.20±0.46	1.04±0.09	1.00±0.00	1.00±0.00	1.33±0.23
	OriTTA	**48.07**±0.52	**55.99**±0.22	**50.84**±0.64	**67.25**±0.49	**46.65**±1.90	**66.65**±1.63	**68.11**±1.04	**57.46**±0.51	**57.90**±2.04	**53.88**±2.33	**66.55**±1.83	**55.22**±2.84	**56.51**±0.97	**50.87**±3.63	**53.47**±1.87	**57.03**±0.29

Table 23: Full results on CIFAR-10-C, WRN-28-10.

	episodic	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
shuffled	TENT	11.46±0.21	12.96±0.68	10.94±0.31	42.53±27.44	16.73±0.79	72.56±17.58	66.12±20.77	77.13±1.20	46.58±9.98	79.70±5.12	91.16±0.16	23.50±16.37	70.43±2.82	19.45±2.81	59.32±8.62	46.70±2.77
	EATA	28.07±0.30	34.51±0.16	27.22±0.15	53.38±0.25	45.87±0.17	65.47±0.20	58.17±0.32	74.97±0.08	58.73±0.13	74.11±0.23	90.69±0.01	53.75±0.34	73.47±0.09	41.78±0.26	69.75±0.06	56.66±0.08
	SAR	27.14±0.15	33.53±0.15	25.93±0.27	53.31±0.36	45.46±0.17	65.23±0.10	58.07±0.18	74.95±0.04	58.80±0.10	74.01±0.04	90.69±0.01	53.23±0.27	73.46±0.10	41.45±0.12	69.72±0.08	56.33±0.03
	DeYO	34.65±5.30	47.15±4.56	19.40±5.32	47.17±20.05	15.12±1.74	46.03±12.28	52.38±18.63	75.40±3.05	63.55±3.70	63.39±7.78	90.02±0.43	65.94±7.24	41.57±3.30	67.51±4.30	49.48±8.06	51.92±2.94
	ReCAP	22.01±1.86	28.92±2.35	14.62±1.46	40.20±8.94	33.60±1.56	59.14±1.85	40.88±6.91	73.64±0.61	54.31±1.37	66.44±12.00	90.69±0.04	32.83±6.61	63.78±2.24	47.77±1.52	61.68±3.21	48.70±0.98
	SaTeen	26.22±6.15	37.41±8.06	11.68±3.20	24.33±14.70	13.22±3.27	27.59±11.22	31.10±21.73	63.20±5.28	57.05±4.50	44.53±8.85	83.06±1.23	26.32±19.35	26.12±5.88	56.15±4.18	32.33±11.88	37.35±2.55
	OriTTA	**56.14**±0.30	**60.55**±0.34	**41.11**±0.26	**80.25**±0.40	**59.54**±0.40	**82.02**±0.31	**82.34**±0.19	**80.75**±0.19	**76.99**±0.31	**84.44**±0.24	**91.32**±0.09	**80.58**±0.19	**77.84**±0.08	**73.40**±0.55	**73.43**±0.16	**73.38**±0.08
class-blocked	TENT	12.75±2.11	11.46±1.94	9.12±0.78	42.24±26.23	11.59±1.95	71.42±19.76	61.65±24.86	71.92±15.59	53.95±20.09	58.46±30.29	91.52±0.44	26.27±18.82	67.76±15.44	16.43±7.70	46.34±19.02	43.52±6.84
	EATA	28.18±0.29	34.59±0.14	27.22±0.10	53.36±0.24	45.82±0.20	65.25±0.17	58.17±0.14	74.93±0.05	58.89±0.15	74.04±0.13	90.70±0.02	53.47±0.27	73.42±0.11	41.73±0.14	69.73±0.09	56.63±0.05
	SAR	27.47±0.28	34.13±0.77	27.54±1.03	54.38±0.83	46.11±0.59	65.69±0.40	59.06±0.41	75.19±0.04	59.35±0.22	74.23±0.15	90.71±0.01	54.18±0.52	74.08±0.34	41.67±0.40	70.58±0.17	56.96±0.10
	DeYO	44.70±15.15	43.05±11.26	27.70±11.34	50.02±27.06	17.17±1.96	56.28±13.26	57.90±10.31	**81.65**±5.34	69.83±5.62	64.83±10.60	93.55±0.80	64.57±20.32	58.34±14.66	75.73±8.21	54.79±17.02	57.34±6.14
	ReCAP	40.34±4.99	46.67±4.97	**43.52**±9.94	48.09±3.19	46.84±2.32	63.66±5.85	52.12±2.22	77.23±1.82	62.60±2.49	67.24±3.94	90.91±0.10	46.04±4.00	71.34±3.25	51.54±2.82	**73.84**±3.22	58.80±0.69
	SaTeen	**52.91**±16.63	**67.45**±21.67	24.33±9.99	29.35±22.67	19.76±10.98	41.56±20.37	39.11±27.30	81.02±8.19	**73.45**±17.10	42.46±17.97	**95.29**±1.87	38.16±22.68	55.43±27.35	**77.96**±11.96	55.26±19.99	52.90±8.45
	OriTTA	46.96±4.01	52.50±2.63	39.10±3.11	**73.22**±2.79	**54.49**±1.76	**78.07**±2.19	**76.05**±1.99	78.28±1.83	72.51±1.31	**83.49**±1.30	91.52±0.76	**75.95**±1.52	**77.83**±1.48	61.40±3.99	73.03±2.02	**68.96**±1.51

	continual	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
shuffled	TENT	11.46±0.21	10.29±0.03	10.00±0.00	**77.19**±2.54	12.07±0.33	22.89±8.62	10.95±1.57	10.00±0.00	10.00±0.00	10.00±0.00	10.08±0.02	10.00±0.00	10.00±0.00	10.00±0.00	10.00±0.00	15.66±0.77
	EATA	28.07±0.30	34.94±0.51	27.61±0.21	53.38±0.39	46.38±0.52	65.59±0.32	58.65±0.56	75.29±0.18	59.63±0.42	74.69±0.41	**90.68**±0.06	54.39±0.70	73.78±0.34	41.45±0.43	69.86±0.12	56.96±0.23
	SAR	27.14±0.15	32.70±2.10	25.27±1.34	53.84±2.15	43.80±2.88	65.12±2.38	59.04±3.24	74.62±0.96	57.47±2.09	73.85±1.91	90.53±0.18	53.13±2.23	72.83±1.53	41.20±2.28	69.33±0.51	55.99±0.87
	DeYO	34.65±5.30	54.65±5.92	40.36±3.74	31.97±14.39	22.38±4.29	30.21±13.29	45.23±16.27	52.48±7.69	53.61±5.37	49.02±4.92	75.18±3.31	47.49±10.60	38.11±10.07	47.95±5.02	43.03±8.53	44.42±3.45
	ReCAP	22.01±1.86	27.88±5.76	16.98±1.97	32.74±9.29	24.03±4.21	37.04±14.12	35.13±17.24	54.28±13.00	41.13±15.96	44.45±12.35	74.68±10.99	25.03±10.80	33.37±12.29	46.43±7.22	33.55±11.23	36.58±6.71
	SaTeen	26.22±6.15	35.52±13.22	19.36±7.04	14.52±6.26	11.58±1.49	10.71±0.90	10.30±0.52	12.77±4.96	10.93±2.19	10.90±2.64	11.00±1.74	10.09±0.24	10.14±0.33	10.08±0.29	9.99±0.02	14.27±1.78
	OriTTA	**56.14**±0.29	**69.72**±0.12	**53.20**±0.25	66.23±0.84	**59.95**±0.27	**81.22**±0.30	**85.45**±0.21	**79.06**±0.25	**79.99**±0.34	**84.65**±0.12	90.41±0.20	**81.02**±0.56	**75.06**±0.41	**69.83**±0.70	**74.24**±0.50	**73.74**±0.17
class-blocked	TENT	12.75±2.11	10.35±0.07	10.00±0.00	21.09±9.47	10.54±0.56	10.38±0.32	10.06±0.05	10.29±0.33	10.01±0.02	10.01±0.02	10.14±0.06	10.00±0.00	10.00±0.00	10.01±0.01	10.00±0.00	11.04±0.78
	EATA	28.18±0.29	35.27±0.45	27.65±0.13	53.25±0.36	46.77±0.45	65.30±0.37	58.44±0.52	**75.54**±0.13	60.04±0.42	74.80±0.60	**90.71**±0.07	53.91±1.09	73.54±0.53	41.21±0.66	69.47±0.19	56.94±0.17
	SAR	27.47±0.28	33.00±4.21	26.86±2.98	52.95±2.63	43.22±5.70	63.18±3.49	57.17±2.53	73.70±3.64	56.61±5.95	72.39±3.54	90.30±0.33	51.30±5.91	70.41±5.25	44.48±4.79	69.28±1.89	55.49±2.39
	DeYO	44.70±15.15	56.74±16.70	39.40±8.01	20.65±9.89	18.43±5.21	24.90±9.97	25.23±13.52	39.69±13.64	46.34±9.61	37.16±12.91	65.04±25.78	41.07±24.26	30.72±12.39	48.71±26.66	29.12±16.66	37.86±6.62
	ReCAP	40.34±4.99	45.02±6.64	37.07±9.26	42.88±13.44	26.43±7.18	45.98±20.48	35.15±18.62	55.61±9.83	43.49±9.93	46.50±14.40	74.83±10.42	26.10±18.86	38.38±19.61	50.64±16.78	38.57±27.67	43.13±9.33
	SaTeen	**52.91**±16.63	**66.49**±20.39	30.83±22.13	14.06±11.29	9.45±1.45	13.29±6.42	22.45±19.27	31.24±17.72	26.02±25.60	23.81±17.86	42.07±39.52	15.46±7.86	15.11±9.14	31.13±23.25	19.99±10.02	27.62±7.73
	OriTTA	43.35±12.11	59.98±2.57	**40.76**±2.51	**59.70**±3.12	**53.22**±1.75	**73.91**±1.96	**74.66**±2.46	75.17±1.32	**70.08**±2.86	**79.38**±3.24	87.96±1.48	**72.62**±4.21	**74.60**±2.65	**55.11**±4.31	**70.15**±2.09	**66.04**±1.43

Table 24: Full results on ImageNet-C, ResNet-50-GN, episodic, at batch size 64.

	episodic	gaus	shot	impu	defo	glas	moti	zoom	snow	fros	fog	brit	cont	elas	pixe	jpeg	Average
shuffled	TENT	24.05±1.02	27.90±0.49	25.13±1.07	15.06±0.15	9.76±1.90	21.87±0.38	22.51±0.21	26.56±0.71	33.03±0.52	3.50±0.15	69.88±0.05	42.07±0.10	10.89±0.59	48.41±0.14	54.49±0.05	29.01±0.24
	EATA	36.81±0.25	38.75±0.29	38.15±0.34	28.37±0.47	28.01±0.24	35.97±0.20	38.62±0.24	48.11±0.47	47.90±0.19	54.08±0.30	70.41±0.13	47.68±0.25	40.92±0.47	55.21±0.29	56.16±0.14	44.34±0.10
	SAR	32.05±0.29	34.27±0.29	33.34±0.36	18.62±0.25	18.84±0.38	30.15±0.29	30.69±0.35	42.11±0.60	43.30±0.10	30.49±20.61	70.31±0.04	43.92±0.10	16.37±1.91	49.07±0.17	55.36±0.05	36.59±1.41
	DeYO	38.12±0.20	40.49±0.17	39.57±0.37	20.82±0.23	22.91±0.22	35.52±0.10	33.22±1.75	48.88±0.29	47.09±0.18	26.25±25.36	72.25±0.07	48.15±0.09	37.25±0.31	54.02±0.12	56.90±0.04	41.43±1.69
	ReCAP	39.02±0.26	41.18±0.20	40.24±0.27	22.57±0.43	23.72±0.47	37.67±0.26	38.04±0.45	50.19±0.34	48.10±0.20	42.72±23.04	72.60±0.05	48.93±0.11	39.46±1.35	54.85±0.16	57.31±0.04	43.77±1.51
	SaTeen	**43.10**±0.24	**45.32**±0.13	**44.26**±0.14	24.52±0.24	25.56±0.25	41.53±0.29	27.32±9.30	54.17±0.45	51.90±0.28	33.29±30.06	**73.17**±0.05	52.78±0.19	45.78±0.35	58.86±0.23	59.06±0.16	45.37±1.89
	OriTTA	41.68±0.19	43.83±0.23	42.65±0.26	**34.58**±0.30	**33.38**±0.26	**43.15**±0.16	**46.82**±0.22	**54.67**±0.30	**53.24**±0.20	**57.84**±0.34	72.44±0.11	**53.61**±0.23	**50.68**±0.48	**59.29**±0.08	**60.20**±0.21	**49.87**±0.09
class-blocked	TENT	24.11±3.22	27.84±2.15	25.75±2.08	15.02±0.26	11.00±2.65	22.02±0.90	22.90±0.37	27.18±0.99	33.30±0.50	3.50±0.31	69.85±0.09	42.30±0.19	11.36±0.66	48.52±0.19	54.54±0.15	29.28±0.55
	EATA	33.09±0.97	34.90±0.55	33.67±0.78	22.89±0.32	24.00±0.41	32.57±0.28	34.62±0.63	42.54±1.19	42.63±0.92	49.34±1.06	68.97±0.38	43.62±0.55	34.86±0.71	51.35±0.42	53.52±0.54	40.17±0.21
	SAR	32.62±0.44	34.61±0.32	33.57±0.29	18.48±0.42	18.64±1.63	30.17±0.32	30.44±0.52	39.49±1.18	43.27±0.12	35.62±17.48	70.29±0.04	44.09±0.19	13.74±1.75	49.34±0.32	55.42±0.08	36.65±1.17
	DeYO	38.15±0.29	40.41±0.27	39.51±0.26	20.37±0.31	22.67±0.34	35.42±0.36	32.44±3.68	48.09±0.36	46.91±0.34	26.58±26.21	72.22±0.08	48.17±0.10	36.39±0.73	53.95±0.26	56.87±0.13	41.21±1.79
	ReCAP	39.11±0.23	41.04±0.31	40.35±0.43	21.72±0.38	23.18±0.79	37.38±0.38	35.99±4.20	49.36±0.52	47.88±0.39	27.48±26.94	72.59±0.07	48.98±0.10	35.22±7.66	54.82±0.31	57.29±0.13	42.16±2.08
	SaTeen	**42.60**±0.20	**44.71**±0.16	**43.83**±0.31	23.94±0.43	25.04±0.59	40.97±0.34	25.98±9.97	53.68±0.43	51.65±0.16	36.92±29.59	**73.24**±0.10	**52.51**±0.16	45.44±0.75	**58.80**±0.35	59.14±0.12	45.23±2.07
	OriTTA	40.66±0.26	42.50±0.32	41.64±0.27	**33.24**±0.33	**31.05**±0.43	**41.38**±0.37	**45.15**±0.32	**54.02**±0.16	**52.64**±0.10	**55.77**±0.35	72.31±0.19	52.39±0.30	**48.59**±0.18	57.36±0.22	**59.76**±0.21	**48.56**±0.14