SOFT PSEUDO SUPERVISED CONTRASTIVE LEARNING FOR TEST-TIME ADAPTATION.

Anonymous authors

Paper under double-blind review

ABSTRACT

Test-time adaptation (TTA) aims to refine pretrained models for shifted test distributions using unlabeled test inputs. Teacher–student self-training shows state-of-the-art performance by minimizing the cross-entropy loss with the pseudo-labels generated by the teacher. However, cross-entropy-based approaches remain vulnerable to noisy pseudo-labels despite the use of various regularization methods. In this work, we propose **Pseudo Supervised Contrastive Learning (PSCL)** as an alternative to cross-entropy that is better suited loss for the noisy labels in TTA. PSCL uses teacher's pseudo-labels to pull together same-class pair representations of test samples while pushing the rest apart. We propose two variants of PSCL. Hard PSCL assigns a binary relation according to whether the teacher predicts two pairs to be the same class, whereas Soft PSCL weights each relation by the teacher-estimated probability that the two samples belong to the same class. Specifically, for each anchor, PSCL aligns the student's similarity distribution over other samples with the teacher-estimated probabilities that those samples belong to the same class as the anchor. Our theory provides a guarantee that, under mild conditions, soft pairwise pseudo-labels better approximate the true pairwise labels than hard pseudo-labels. We show that PSCL, as a sole objective, outperforms state-of-the-art baselines including methods that combine cross-entropy with weak contrastive regularization and other regularization techniques. This result holds across various experiments on the standard continual TTA (CTTA) benchmarks CIFAR-10-C, CIFAR-100-C, and ImageNet-C.

1 INTRODUCTION

Machine learning models often face the challenge of data distribution shift when used in real life applications; a pretrained model may lose accuracy when test data follow a different distribution than the source (pretraining) data distribution. Distribution shifts arise from input corruptions, changes in data collection, or domain change (Hendrycks & Dietterich, 2019; Liang et al., 2020). Test-time adaptation (TTA) addresses this problem by updating a pretrained model using unlabeled test data during inference, and hence improve performance (Wang et al., 2021; 2022).

Most recent gradient-based TTA methods rely on entropy minimization or cross-entropy self-training (Liang et al., 2020; Wang et al., 2021; 2022), often in a student-teacher framework where pseudo-labels generated by an EMA teacher supervise the student on unlabeled target data. Recent approaches stabilize these objectives through reliable sample selection, constrained updates, and replay to reduce error accumulation and forgetting (Niu et al., 2023; Lai et al., 2026; Zhu et al., 2024). However, cross-entropy remains vulnerable to false (noisy) pseudo-labels generated by the teacher.

Graf et al. (2021) shows that supervised contrastive learning is more resistant than cross-entropy to memorizing noisy labels. Whereas cross-entropy fits each label individually, a contrastive loss uses labels only to decide which pairs of samples to pull together, and thus supervises relations between samples rather than individual predictions. We investigate whether this advantage carries over to test-time adaptation, where input corruption also corrupts the pseudo-labels used for training.

Contrastive objectives shape representations through positive and negative relationships van den Oord et al. (2018), were also explored in TTA. AdaContrast (Chen et al., 2022) combines cross-entropy self-training with a contrastive loss that uses augmented views of the same image as posi-

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.

tives and pseudo-labels to exclude likely same-class negatives. This leaves potential positive relationships between distinct images unused and makes the set of competing samples depend on each anchor's pseudo-label. SANTA (Chakrabarty et al., 2023) and RMT (Döbler et al., 2023) instead align features contrastively with source-derived prototypes. Their supervision depends on current feature-to-prototype assignments, which could reinforce errors when distribution shift distorts the representation, and building the prototypes requires source information beyond the pretrained model. None uses pseudo-labels to define positives across distinct images as the sole adaptation signal.

In this work, we study Pseudo-Supervised Contrastive Loss (PSCL) for test-time adaptation. Like relational knowledge distillation (Chen et al., 2018; Park et al., 2019) and relational contrastive methods (Li et al., 2021; Zheng et al., 2021), PSCL uses relations between samples as supervision: the teacher's predictions pull together representations of samples predicted to belong to the same class and push apart those predicted to belong to different classes. We consider two variants. Hard PSCL assigns a binary pair relation according to whether the two hard teacher pseudo-labels agree, following the positive-pair construction of SupCon (Khosla et al., 2020). Since hard assignments discard the uncertainty of the teacher's predictions, Soft PSCL instead weights each pair by the teacher-estimated probability that the two samples belong to the same class.

However, extending these relational objectives to TTA raises an additional challenge: distribution shift degrades calibration even in networks that are well calibrated on the source distribution (Ovadia et al., 2019). The teacher's probabilities may therefore become less reliable, which Soft PSCL uses directly as pair weights.

Since pair weights that recover the true same-class relations exactly would reduce PSCL to SupCon with ground-truth labels. We therefore derive theoretical bounds characterizing the conditions under which soft supervision better approximates the true same-class relations than hard supervision for contrastive learning. We evaluate these conditions with target labels as an analysis. in principle, they can also be verified without target labels; ie. TransCal proposes calibration under covariate shift to be estimated from labeled source and unlabeled target data (Wang et al., 2020).

Using only soft PSCL—without source data, auxiliary regularization, or confidence thresholding—we achieve state-of-the-art performance against recent methods on the standard CTTA benchmarks CIFAR-10C, CIFAR-100C, and ImageNet-C (Hendrycks & Dietterich, 2019).

2 PRELIMINARIES

Supervised contrastive learning (SupCon, Khosla et al. (2020)) . Let $\mathcal{X} = \{x_i\}_{i=1}^N$ denote a batch of clean target samples and $z_i = f(x_i)/\|f(x_i)\|_2$ be the normalized feature vector produced by the model $f(x_i)$ of the i-th sample in a batch of size N and τ being the temperature.
With $\ell_{ij} = \log \frac{\exp(z_i^\top z_j/\tau)}{\sum_{k\neq i}\exp(z_i^\top z_k/\tau)}$, we express the contrastive objective through a pairwise supervision matrix $A \in \{0,1\}^{N\times N}$ with zero diagonal:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}\frac{A_{ij}}{\sum_{k\neq i}A_{ik}}\,\ell_{ij}. \tag{1}$$

with $A_{ij} = \mathbf{1}\{y_i = y_j\}$ and 0 otherwise.

3 METHODS

Current leading TTA methods employ teacher–student self-training: for each unlabeled target sample, the teacher generates a pseudo-label that is used to update the student, typically through cross-entropy. However, an incorrect pseudo-label directly drives the student toward the wrong class. We instead introduce pseudo-supervised contrastive learning (PSCL), which uses the teacher predictions to define pairwise relationships between samples, pulling together representations with the same predicted class and pushing apart those with different predicted classes.

Formally, let the source-pretrained model be $g = h \circ f$, composed of the encoder $f : \mathcal{X} \to \mathcal{Z} \in \mathbb{R}^d$ and the classifier $h(z) = \mathrm{softmax}(Wz)$. At test time, the model receives the unlabeled samples $\mathcal{D}_t = \{x_i\}_{i=1}^N$ drawn from a shifted target distribution. We initialize both the student $g_\theta = h \circ f_\theta$

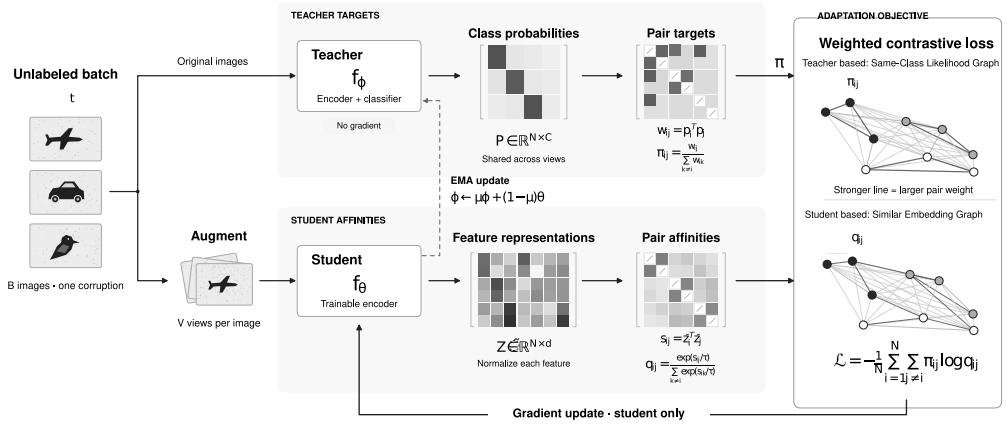


Figure 1: Overview of the proposed pseudo-supervised contrastive test-time adaptation framework.

and the teacher $g_\phi = h \circ f_\phi$ from the source-pretrained model while keeping the head h frozen. For each target sample x_i, the teacher produces the predictive distribution

$$p_i = g_\phi(x_i) \in \Delta^{C-1},$$

with C defined as the number of classes. We rely on the pseudo labels p_i produced by the EMA teacher to define the pairwise relation between two samples, since we have no access to the ground truth labels which define the true positive pairs as in SupCon (Khosla et al., 2020). In the binary pseudo pairing, we define the predicted binary (hard) same-class label to be h_{ij} as an approximation of the true label A_{ij}, computed as:

$$h_{ij} = 1\left[\arg\max_c p_{ic} = \arg\max_c p_{jc}\right].$$

However, using the hard pairing h_{ij} under the noisy nature of the labels induces information loss and discards the teacher's uncertainty on the predicted class. To tackle this issue, we treat all samples as connected (positive pairs) with the probability of each pair being same class pair defined as:

$$\Pr(Y_i = Y_j \mid x_i, x_j) = \sum_{c=1}^{C} p(Y_i = c \mid x_i)\, p(Y_j = c \mid x_j) = \mathbf{p}_i^\top \mathbf{p}_j = w_{ij}. \tag{2}$$

Where we assume conditional independence on the class predictions given two samples.

Analogous to SupCon's adjacency matrix A we use the pseudo soft adjacency $W = [w_{ij}]_{i,j=1}^N \in [0,1]^{N \times N}$ with $w_{ij} = \mathbf{p}_i^\top \mathbf{p}_j$ predicted for each sample by the frozen EMA teacher.

Substituting the adjacency matrix A in hard SupCon (Eq. 1) with the soft weights W, our soft contrastive loss is then defined as:

$$\mathcal{L}_{\text{soft}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j \neq i} \pi_{ij} \log \frac{\exp\left(\mathbf{z}_i^\top \mathbf{z}_j / \tau\right)}{\sum_{k \neq i} \exp\left(\mathbf{z}_i^\top \mathbf{z}_k / \tau\right)}, \tag{3}$$

with the normalized weights $\pi_{ij} = \frac{w_{ij}}{\sum_{k \neq i} w_{ik}}$.

We minimize $\mathcal{L}_{\text{soft}}$ with respect to the student parameters θ, treating the pairwise weights produced by g_ϕ as fixed targets. After each student update, the teacher parameters ϕ are updated as an exponential moving average (EMA) of the student parameters (Figure 1).

Moreover, for each target sample x_b in the batch B, we generate V stochastic views $\tilde{x}_{b,v}$, each of which is mapped by the student to a representation $\mathbf{z}_{b,v} = f_\theta(\tilde{x}_{b,v})$. The teacher prediction of the non-augmented sample $\mathbf{p}_b = g_\phi(x_b)$ is assigned to every view, such that $\mathbf{p}_{b,v} := \mathbf{p}_b$ for $v = 1, \ldots, V$. We then compute the soft pairwise weights from these predictions as in Eq. 2, while setting $w_{ij} := 1$ whenever i and j are two views of the same target sample. The resulting $N = BV$ representations are used as anchors in the loss function $\mathcal{L}_{\text{soft}}$ in Eq. 3.

3.1 When and Why Soft Pseudo-SupCon Outperforms Hard SupCon

Let $A^\star_{ij} = 1[y_i = y_j]$ denote the unknown pairwise class relation between samples i and j. Hard pseudo-SupCon approximates this relation using the binary estimator, whereas soft pseudo-SupCon uses the continuous estimator

$$w_{ij} = \mathbf{p}_i^\top \mathbf{p}_j.$$

We therefore compare the two estimators directly through their expected squared error with respect to the oracle relation $A^\star_{ij}$.

Proposition 1 (Pairwise calibration)**.** *Let $r_{ij} \in [0,1]$ denote the unknown probability that samples i and j belong to the same class, such that $A^\star_{ij} \sim \text{Bernoulli}(r_{ij})$. Given the teacher predictions $\mathbf{p}_i, \mathbf{p}_j \in \Delta^{C-1}$, we let $w_{ij} = \mathbf{p}_i^\top \mathbf{p}_j$ be the soft pairwise estimate, and $h_{ij} \in \{0,1\}$ the corresponding hard estimate.*

Assume perfect pairwise calibration,

$$r_{ij} = w_{ij}.$$

Then

$$\mathbb{E}\big[(A^\star_{ij} - w_{ij})^2\big] \le \mathbb{E}\big[(A^\star_{ij} - h_{ij})^2\big].$$

More precisely (refer to A.2 in Appendix for proof),

$$\mathbb{E}\big[(A^\star_{ij} - h_{ij})^2 - (A^\star_{ij} - w_{ij})^2\big] = \begin{cases} (1 - w_{ij})^2, & h_{ij} = 1, \\ w_{ij}^2, & h_{ij} = 0. \end{cases}$$

The result extends to imperfectly calibrated predictions. Let

$$r_{ij} = w_{ij} + \delta_{ij},$$

where δ_{ij} is the signed pairwise calibration error. Then the soft estimate remains at least as good compared to the hard estimate under condition (refer to A.2 in Appendix for proof):

$$-\frac{w_{ij}}{2} \le \delta_{ij} \le \frac{1 - w_{ij}}{2} \tag{4}$$

3.2 Gradient Function Properties

Unlike cross entropy, the proposed loss has a special property where the gradient update of a sample's representation depends on the entire similarity graph in the representation space, making it more robust to noisy pseudo labels with the gradient function with respect to the representation defined as:

$$\frac{\partial \mathcal{L}_i^{\text{soft}}}{\partial \mathbf{z}_i} = \frac{1}{\tau} \sum_j (q_{ij} - \pi_{ij})\, \mathbf{z}_j, \quad q_{ij} := \frac{\exp\big(\mathbf{z}_i^\top \mathbf{z}_j / \tau\big)}{\sum_a \exp\big(\mathbf{z}_i^\top \mathbf{z}_a / \tau\big)}. \tag{5}$$

Moreover, soft target π_{ij} in Eq. 5 scales the gradient magnitude proportionally to the teacher's confidence, thereby dampening the penalty on false hard positives and mitigating the impact of noisy labels. While, hard pseudo-label h_{ij} (0 or 1) applied to a noisy sample incorrectly forces a hard negative to act as a definitive positive, leading to a larger incorrect gradient update.

3.3 Buffer: an Advantage not only a necessity

The necessity: Given that SupCon Khosla et al. (2020) relies on pairs for learning representations, this poses a challenge in scenarios where the number of possible classes is high relative to the available batch size (e.g., online TTA). We tackle this problem by using a memory buffer instead of assuming that a large batch size is possible, making the proposed solution adequate for each scenario. We use a FIFO buffer with stored raw data $\mathcal{X}_{\text{buffer}} = \{x_i\}_{i=1}^M$. Using the teacher predictions, we retrieve k samples for each predicted class present in the current batch, resulting in $k\mathcal{C}_B$ retrieved buffer samples, where $\mathcal{C}_B$ is the number of predicted unique classes in the batch. These samples are added to the current-batch representations, yielding N samples. in Eq. 3.

4 EXPERIMENTAL SETUP

Evaluation setting. We evaluate our method under online continual test-time adaptation (CTTA), a particularly demanding TTA setting in which the model must adapt to a continual stream of unlabeled target data over different domains (multiple domain shifts). In our setup we assume accessing source data is not possible making the procedure source-free TTA. We stick to the corruption setup defined by CoTTA Wang et al. (2022), where the model is allowed a single update per batch and parameter updates accumulate throughout the stream. This setting therefore evaluates not only adaptation to an individual shift, but also stability under successive sudden distribution changes.

Datasets and source models. Following the standard evaluation adopted by the methods in Table 1 Wang et al. (2021; 2022); Chen et al. (2022); Lai et al. (2026), we evaluate on CIFAR-10-C, CIFAR-100-C, and ImageNet-C Hendrycks & Dietterich (2019). Each benchmark contains 15 corruption types at five severity levels; we follow the standard CTTA protocol and use the highest severity of corruption with level 5. Previous TTA studies employ heterogeneous pretrained architectures, ranging from convolutional networks to Vision Transformers, making direct comparisons between independently reported results difficult. We therefore adopt the setup of GOLD Lai et al. (2026), which reports state-of-the-art results and re-evaluates a broad set of competing methods under consistent source (pretrained) models. Specifically, we use WideResNet-28 for CIFAR-10-C, ResNeXt-29 for CIFAR-100-C, and ResNet-50 for ImageNet-C, with the hyperparameters reported in Table 3 in Appendix A.3. The reported results from GOLD (Lai et al., 2026) use their source-assisted variant with class prototypes similar to methods like SANTA (Chakrabarty et al., 2023). We use the results they report against our source-free TTA setting.

Online adaptation protocol. The model encounters the 15 corruption domains sequentially, without resetting the model, optimizer, teacher, or memory buffer between corruptions. Within each corruption, the samples are processed once in consecutive mini-batches. For each batch $\mathcal{X}_t$, predictions are produced using the current model parameters θ_t and recorded for evaluation. The contrastive loss is then computed on the same unlabeled batch, and exactly one optimizer update is performed to obtain θ_{t+1}. This is done continually for all batches across all corruptions. Ground-truth labels are used only to evaluate and report the error for each batch.

Episodic adaptation protocol. For episodic TTA on CIFAR-10-C, we adapt to each corruption independently for three epochs. Before each corruption, we reset the model, teacher, optimizer, and memory buffer to their source initialization. We use the same protocol for soft PSCL and cross-entropy.

5 RESULTS

Following the standard CTTA benchmarks, we report state of the art results for all datasets as shown in Table 1. The table contains various methods, with results reported using source data and extra regularization as in GOLD (Lai et al., 2026), SANTA (Chakrabarty et al., 2023) and other methods. We furthermore compare the performance between Soft and Hard PSCL in Table 2 on the datasets CIFAR-10-C, CIFAR-100-C, and ImageNet-C.

Table 2 compares Soft PSCL with Hard PSCL, which applies SupCon's positive-pair construction to hard teacher predictions. Soft PSCL reduces mean classification error from 14.15% to 13.87% on CIFAR-10-C, from 28.91% to 28.43% on CIFAR-100-C, and from 67.01% to 58.97% on ImageNet-C. The particularly large ImageNet-C improvement shows that the construction of relational targets can materially affect adaptation performance. These results extend the motivation supplied by Graf et al. by directly testing soft and hard pseudo-label-based relations under continual distribution shift.

We show that Soft PSCL outperforms the hard PSCL with clear gains on CIFAR-100-C and ImageNet-C. Moreover, in Figure 2 we validate the improved performance observed by Soft PSCL by investigating the calibration error bounds defined in Eq. 4. CIFAR-10C has 97.23% of the pairwise weights w_{ij} inside the derived bounds in Eq. 4, and similar for CIFAR-100C with 97.95%
We also compare the performance of soft PSCL against cross entropy in episodic TTA on CIFAR-10-C. We allow three epochs per corruption, where each corruption is learned independently from

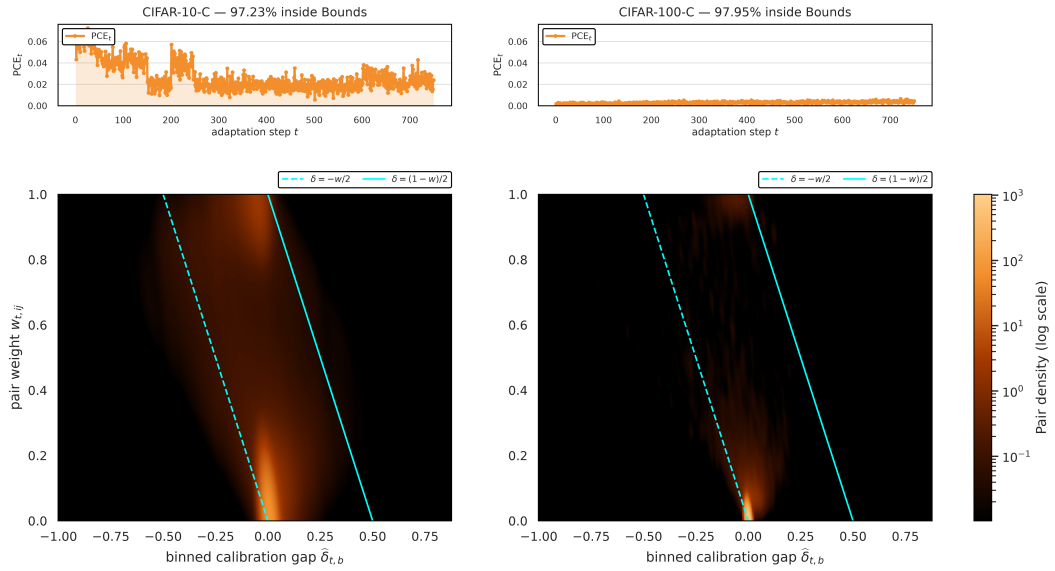


Figure 2: Empirical verification of the sufficient condition in Eq. 4. CIFAR-10-C (left) and CIFAR-100C (right), over all adaption steps. CIFAR-10C has 97.23% of the pairwise weights w_{ij} inside the derived bounds in Eq. 4, and similar for CIFAR-100C with 97.95%. Whereas, CIFAR100-C shows lower pariwise calibration error (PCE) due to sparsity resulting from the number of classes, with $\mathbb{E}[w_{ij}] = 1/C$ under the assumption that the model predictive distribution is uniform over the classes.

the provided source model. Soft PSCL achieves 85.05% compared to 81.11% with cross entropy. We further evaluate the quality of the representations learned after one complete round of continual test-time adaptation on CIFAR-10-C. After sequential adaptation across all corruptions, each final CTTA model (after 15 corruptions) is evaluated over the complete corruption set. The resulting representations are visualized using t-SNE. Our Soft PSCL trained model achieves a 87.69%, where cross-entropy achieves 77.69% accuracy. Moreover our loss produces clearly separated class-wise structures, indicating a qualitative improvement in representation separability and class-consistent neighborhood organization.

6 Conclusion

In this study, we provide empirical evidence showing the robustness of SupCon to noisy pseudo-labels compared with cross-entropy-based methods under data distribution shift. We compare our method against a wide range of methods that use additional regularization and/or source data information, and show that a source-free PSCL objective used as the sole optimization loss outperforms these methods on CIFAR-10-C, CIFAR-100-C, and ImageNet-C. Additionally, We derive sufficient conditions under which soft pairwise pseudo labels better approximate true pairwise labels than hard pairwise pseudo labes.We verify these conditions empirically, and find that soft PSCL outperforms hard PSCL. We conclude that PSCL is a better base loss than cross entropy in the continual and episodic TTA settings studied, with further opportunities to improve its efficiency, hyperparameter robustness, and overall performance through additional regularization.

7 Limitations

Computational Cost We acknowledge the computational complexity of the model due to the quadratic loss scaling with the number of samples compared to cross-entropy or other gradient free methods (Lee et al., 2024; Murphy et al., 2026; Niu et al., 2024; Iwasawa & Matsuo, 2021), especially with the buffer. However this model is not limited to online TTA but should be seen as a

Table 1: Online classification error rates (%) on CIFAR-10-C, CIFAR-100-C, and ImageNet-C under the CTTA setting at severity level 5. Our method uses Soft PSCL, and our results are averaged over five independent runs, with the overall mean reported as mean ± standard deviation. For each dataset and column, the lowest error is highlighted in **bold** and the second-lowest distinct error is underlined; ties receive the same highlighting.

	Method	Gau.	shot	imp.	def.	glass	mot.	zoom	snow	fro.	fog	bri.	con.	ela.	pix.	jpeg	Mean
CIFAR-10-C	TENT Wang et al. (2021)	24.8	20.6	28.6	14.4	31.1	16.5	14.1	19.1	18.6	18.6	12.2	20.3	25.7	20.8	24.9	20.7
	Ada Chen et al. (2022)	29.1	22.5	30.0	14.0	32.7	14.1	12.0	16.6	14.9	14.4	8.1	10.0	21.9	17.7	20.0	18.5
	CoTTA Wang et al. (2022)	24.3	21.3	26.6	11.6	27.6	12.2	10.3	14.8	14.1	12.4	7.6	10.6	18.3	13.4	17.3	16.2
	RMT Döbler et al. (2023)	24.0	20.4	25.6	12.6	25.4	14.2	12.2	15.4	15.1	14.1	10.3	13.7	17.1	13.5	16.0	16.7
	DSS Wang et al. (2024)	24.1	21.3	25.4	11.7	26.9	12.2	10.5	14.5	14.1	12.5	7.8	10.8	18.0	13.1	17.3	16.0
	EATA Niu et al. (2022)	24.2	19.1	27.6	12.5	29.2	14.2	12.0	16.3	15.1	15.1	8.9	13.5	21.2	16.4	21.9	17.9
	SAR Niu et al. (2023)	28.3	25.9	35.7	12.6	34.7	13.8	11.9	17.4	17.5	14.8	8.1	12.9	23.5	19.5	27.2	20.3
	BeCoTTA Lee et al. (2024)	22.9	19.1	26.9	10.2	27.5	12.7	10.4	14.7	14.3	12.4	**7.2**	9.4	20.9	15.2	20.2	16.3
	SANTA Chakrabarty et al. (2023)	23.9	20.1	28.0	11.6	27.4	12.6	10.2	14.1	13.2	12.2	7.6	10.3	19.1	13.3	18.5	16.1
	OBAO Zhu et al. (2024)	22.2	18.3	25.0	10.7	24.7	**11.6**	10.0	12.9	12.7	10.9	7.7	9.6	15.7	11.4	15.2	14.6
	GOLD Lai et al. (2026)	**21.4**	18.1	24.5	**10.0**	24.4	**11.6**	9.9	**12.5**	12.3	**10.5**	7.5	8.9	15.3	10.5	14.9	14.1
	Ours	23.17	**17.42**	**22.28**	11.39	**22.90**	12.14	**9.89**	12.54	**11.05**	10.76	7.92	**8.47**	**14.35**	**10.34**	**13.45**	**13.87**$_{\pm 0.14}$
CIFAR-100-C	TENT Wang et al. (2021)	37.2	35.8	41.7	37.9	51.2	48.3	48.5	58.4	63.7	71.1	70.4	82.3	88.0	88.5	90.4	60.9
	Ada Chen et al. (2022)	42.3	36.8	38.6	27.7	40.1	29.1	27.5	32.9	30.7	38.2	25.9	28.3	33.9	33.3	36.2	33.4
	CoTTA Wang et al. (2022)	40.1	37.7	39.7	26.9	38.0	27.9	26.4	32.8	31.8	40.3	24.7	26.9	32.5	28.3	33.5	32.5
	RMT Döbler et al. (2023)	40.5	36.1	36.3	27.7	33.9	28.5	26.4	29.0	29.0	32.5	25.1	27.4	28.2	26.3	29.3	30.4
	DSS Wang et al. (2024)	39.7	36.0	37.2	26.3	35.6	27.5	25.2	31.4	30.0	37.8	24.2	26.0	30.0	26.3	31.3	30.9
	EATA Niu et al. (2022)	37.2	**33.0**	35.9	27.7	38.1	29.4	27.1	32.8	32.7	36.1	27.0	29.4	33.8	29.5	38.2	32.5
	SAR Niu et al. (2023)	40.5	34.8	37.1	25.6	37.2	28.0	25.5	31.9	30.9	35.8	25.2	27.8	31.8	29.0	37.2	31.9
	BeCoTTA Lee et al. (2024)	42.1	38.0	42.2	30.2	42.9	31.7	29.8	35.1	33.9	38.5	27.9	32.0	36.7	31.6	39.9	35.5
	SANTA Chakrabarty et al. (2023)	**36.6**	33.2	35.1	25.9	34.9	27.7	25.4	29.5	29.9	33.1	23.6	26.7	31.9	27.5	35.2	30.4
	OBAO Zhu et al. (2024)	38.8	35.0	36.6	26.7	33.4	27.4	25.0	28.6	28.0	31.2	24.1	25.1	27.5	25.4	29.1	29.5
	GOLD Lai et al. (2026)	37.1	34.0	35.5	26.1	**32.2**	26.7	24.8	**27.1**	27.0	**29.5**	23.5	25.2	26.8	24.6	**28.3**	28.6
	Ours	38.58	34.65	**33.93**	**25.25**	33.80	**26.08**	**24.15**	27.41	**26.49**	29.61	**22.92**	**23.26**	**26.68**	**24.59**	29.05	**28.43**$_{\pm 0.03}$
ImageNet-C	TENT Wang et al. (2021)	81.6	74.6	72.7	77.6	73.8	65.5	55.3	61.6	63.0	51.7	38.2	72.1	50.8	47.4	53.3	62.6
	Ada Chen et al. (2022)	82.9	80.9	78.4	81.4	78.7	72.9	64.0	63.5	64.5	53.5	38.4	66.7	54.6	49.4	53.0	65.5
	CoTTA Wang et al. (2022)	84.7	82.1	80.6	81.3	79.0	68.6	57.5	60.3	60.5	48.3	36.6	66.1	**47.3**	41.2	46.0	62.7
	RMT Döbler et al. (2023)	80.2	76.4	74.5	77.1	74.4	66.2	57.6	57.0	59.1	48.0	39.1	60.6	**47.3**	42.5	**43.4**	60.2
	DSS Wang et al. (2024)	82.3	78.4	76.7	81.9	77.8	66.9	60.9	**50.8**	60.9	47.7	35.4	69.0	47.5	**40.9**	46.2	62.2
	EATA Niu et al. (2022)	75.7	**66.7**	**65.4**	74.1	**69.2**	64.2	57.5	63.0	62.8	53.4	41.3	64.0	50.5	47.2	50.0	60.4
	SAR Niu et al. (2023)	82.0	74.0	71.9	77.4	73.6	66.1	56.2	61.3	63.1	51.2	37.1	69.0	49.8	46.0	51.2	62.0
	BeCoTTA Lee et al. (2024)	84.1	74.3	72.2	77.4	71.9	63.4	55.1	57.2	61.2	50.7	36.4	66.1	49.2	45.6	48.4	60.9
	SANTA Chakrabarty et al. (2023)	74.1	72.9	71.6	75.7	74.1	64.2	55.5	55.6	62.9	**46.6**	**34.1**	69.9	50.6	44.3	48.5	60.1
	OBAO Zhu et al. (2024)	78.5	75.3	73.0	75.7	73.1	64.5	56.0	55.8	**58.1**	47.6	40.5	**60.1**	47.4	43.9	45.8	59.6
	GOLD Lai et al. (2026)	76.6	72.4	70.9	**73.3**	71.4	**62.2**	54.9	57.7	60.3	48.5	38.7	61.8	48.6	45.3	47.8	59.3
	Ours	**72.65**	68.74	68.74	75.18	73.32	65.00	**54.89**	57.01	59.84	47.21	34.43	62.60	50.26	44.74	49.98	**58.97**$_{\pm 0.13}$

Table 2: Online classification error rates (%) on CIFAR-10-C and CIFAR-100-C under the CTTA setting at severity level 5 comparing Soft and Hard PSCL performance, each using the same hyper-parameters. Our results are averaged over five independent runs and reported as mean ± standard deviation. The best mean is highlighted in **bold** and the second best is underlined.

Dataset	Method	Gau.	Shot.	Imp.	Def.	Glass	Mot.	Zoom	Snow	Frost	Fog	Bri.	Con.	Ela.	Pix.	JPEG	Mean
ImageNet-C	Soft PSCL	**72.65**	**68.74**	**68.74**	**75.18**	**73.32**	**65.00**	**54.89**	**57.01**	**59.84**	**47.21**	**34.43**	**62.60**	**50.26**	**44.74**	**49.98**	**58.97**$_{\pm 0.13}$
	Hard PSCL	78.57	78.27	78.81	82.16	82.91	73.17	62.60	64.27	67.16	54.38	37.92	76.35	57.10	52.13	59.33	67.01$_{\pm 0.12}$
CIFAR-10-C	Soft PSCL	**23.17**	**17.42**	**22.28**	**11.39**	**22.90**	**12.14**	**9.89**	**12.54**	**11.05**	**10.76**	**7.92**	**8.47**	**14.35**	**10.34**	**13.45**	**13.87**$_{\pm 0.14}$
	Hard PSCL	23.37	17.86	22.72	11.65	23.39	12.32	10.19	12.67	11.32	10.80	8.16	8.55	14.84	10.53	13.88	14.15$_{\pm 0.07}$
CIFAR-100-C	Soft PSCL	**38.58**	**34.65**	**33.93**	**25.25**	**33.80**	**26.08**	**24.15**	**27.41**	**26.49**	**29.61**	**22.92**	**23.26**	**26.68**	**24.59**	**29.05**	**28.43**$_{\pm 0.03}$
	Hard PSCL	39.28	35.24	34.01	25.34	34.83	26.47	24.77	27.98	26.67	29.87	23.02	23.67	27.22	25.16	30.05	28.91$_{\pm 0.09}$

general model offering a new framework that can be further optimized for the special case of CTTA. This can be achieved by using more efficient buffer like MoCo He et al. (2020) as well as using efficient sampling in the denominator in q_{ij}, whose computation across all anchors scales as $\mathcal{O}(N^2)$ compared to $\mathcal{O}(NC)$ for cross entropy.

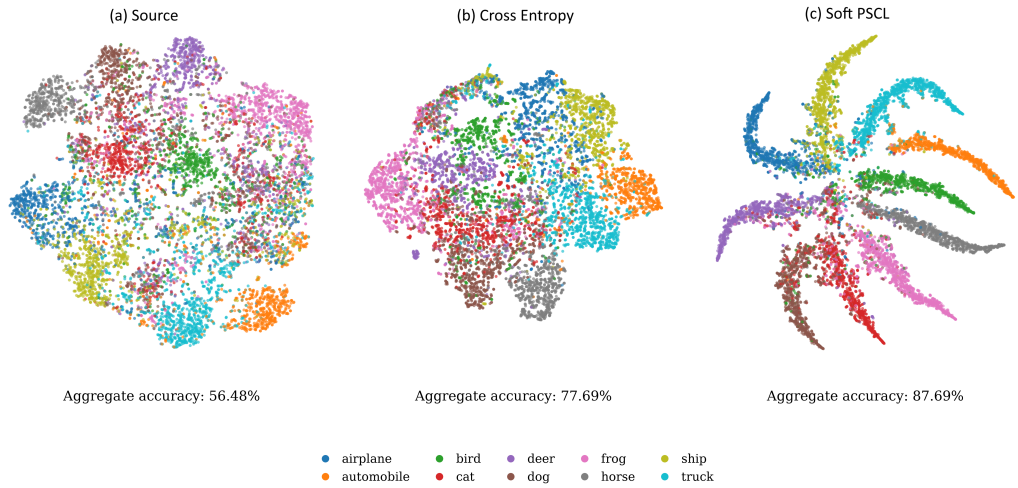


Figure 3: Representations quality learned after a full round of CTTA on CIFAR-10-C. We do offline evaluation for each model. The evaluation is performed over all corruptions together with the final learned model and visualize the TSNE of the representations. a) shows the representations of the pretrained source model with accuracy 56.48%, b) shows the representations learned with cross entropy loss with accuracy 77.69%. C) shows the representations learned with soft PSCL with accuracy of 87.69%, with clear improvement on the representations quality in terms of separability.

The Hyperparameter Optimization Dilemma in TTA In our setting we avoid having systemic hyperparameter optimization with bayesian optimization methods due to the nature of the problem where the true labels should not be available. However we acknowledge the implicit optimization occurring through trial and error. We therefore highlight this as a general problem in the current TTA benchmarking framework where typically methods use explicit hyperparameter optimization although true labels should not be accessible. One solution could be assuming accessibility to limited small labaled set that is only allowed for Hyperparameter Opt. using Bayesian Opt simulating reality where often few labels are available. This does not cancel the implicit optimization, however, but shows how robust a method could be for suboptimal hyperparameter obtained from a small set.

AI USE STATEMENT

We used generative AI tools to brainstorm research approaches search and review related literature, assist with code, and improve the manuscript's wording, notation, and formatting. We reviewed all AI-assisted material and take full responsibility for the methods, claims, results, and final manuscript.

REFERENCES

Goirik Chakrabarty, Manogna Sreenivas, and Soma Biswas. SANTA: Source anchoring network and target alignment for continual test time adaptation. *Transactions on Machine Learning Research*, 2023.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 295–305, 2022.

Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. DarkRank: Accelerating deep metric learning via cross sample similarities transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11783.

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

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3821–3830. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/graf21a.html.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

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

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2427–2440, 2021.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

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

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

Junnan Li, Caiming Xiong, and Steven C. H. Hoi. CoMatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9475–9484, 2021.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6028–6039. PMLR, 2020.

Alexander Murphy, Michal Danilowski, Soumyajit Chatterjee, and Abhirup Ghosh. NEO: No-optimization test-time adaptation through latent re-centering. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

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

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38298–38315. PMLR, 2024.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. doi: 10.1109/CVPR.2019.00409.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

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

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Transferable calibration with lower bias and variance in domain adaptation. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yanshuo Wang, Jie Hong, Ali Cheraghian, Shafin Rahman, David Ahmedt-Aristizabal, Lars Petersson, and Mehrtash Harandi. Continual test-time domain adaptation via dynamic sample selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1701–1710, 2024.

Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. ReSSL: Relational self-supervised learning with weak augmentation. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Zhilin Zhu, Xiaopeng Hong, Zhiheng Ma, Weijun Zhuang, Yaohui Ma, Yong Dai, and Yaowei Wang. Reshaping the online data buffering and organizing mechanism for continual test-time adaptation. In *European Conference on Computer Vision*, pp. 415–433, 2024.

A Appendix

A.1 Gradient Function

$$\frac{\partial \mathcal{L}_i^{\text{soft}}}{\partial \mathbf{z}_i} = -\frac{1}{\tau} \sum_{j \in \mathcal{A}(i)} \pi_{ij} \frac{\partial(\mathbf{z}_i^\top \mathbf{z}_j)}{\partial \mathbf{z}_i} + \frac{\partial}{\partial \mathbf{z}_i} \log \sum_{a \in \mathcal{A}(i)} \exp\left(\frac{\mathbf{z}_i^\top \mathbf{z}_a}{\tau}\right) \tag{1}$$

$$= -\frac{1}{\tau} \sum_{j \in \mathcal{A}(i)} \pi_{ij} \mathbf{z}_j + \frac{\sum_{a \in \mathcal{A}(i)} \exp\left(\mathbf{z}_i^\top \mathbf{z}_a / \tau\right) \frac{\mathbf{z}_a}{\tau}}{\sum_{b \in \mathcal{A}(i)} \exp\left(\mathbf{z}_i^\top \mathbf{z}_b / \tau\right)} \tag{2}$$

$$= -\frac{1}{\tau} \sum_{j \in \mathcal{A}(i)} \pi_{ij} \mathbf{z}_j + \frac{1}{\tau} \sum_{a \in \mathcal{A}(i)} q_{ia} \mathbf{z}_a \tag{3}$$

$$= \frac{1}{\tau} \left(\sum_{j \in \mathcal{A}(i)} q_{ij} \mathbf{z}_j - \sum_{j \in \mathcal{A}(i)} \pi_{ij} \mathbf{z}_j \right) \tag{4}$$

$$= \frac{1}{\tau} \sum_{j \in \mathcal{A}(i)} (q_{ij} - \pi_{ij})\, \mathbf{z}_j, \tag{5}$$

A.2 Inequality proof

Proof: calibrated case. Pairwise probabilities:

$$\mathbb{P}\big(A_{ij}^\star = 1\big) = r_{ij} = w_{ij}, \qquad \mathbb{P}\big(A_{ij}^\star = 0\big) = 1 - w_{ij}.$$

Let the binary expectation be:

$$\mathbb{E}\big[f(A_{ij}^\star)\big] = \sum_{a \in \{0,1\}} \mathbb{P}\big(A_{ij}^\star = a\big)\, f(a) = w_{ij}\, f(1) + (1 - w_{ij})\, f(0). \tag{6}$$

We then compute the soft error as:

$$\mathcal{E}_{\text{soft}} = \mathbb{E}\big[(A_{ij}^\star - w_{ij})^2\big] = w_{ij}(1 - w_{ij})^2 + (1 - w_{ij})\, w_{ij}^2 = w_{ij}(1 - w_{ij}).$$

Similarly for the hard error:

$$\mathcal{E}_{\text{hard}} = \mathbb{E}\big[(A_{ij}^\star - h_{ij})^2\big] = w_{ij}(1 - h_{ij})^2 + (1 - w_{ij})\, h_{ij}^2 = \begin{cases} 1 - w_{ij}, & h_{ij} = 1, \\ w_{ij}, & h_{ij} = 0. \end{cases}$$

yielding the error difference:

$$\mathcal{E}_{\text{hard}} - \mathcal{E}_{\text{soft}} = \begin{cases} (1 - w_{ij})^2, & h_{ij} = 1, \\ w_{ij}^2, & h_{ij} = 0, \end{cases} \geq 0.$$

Proof: miscalibrated case. Calibration error:

$$\delta_{ij} = r_{ij} - w_{ij}, \qquad \mathbb{P}\big(A_{ij}^\star = 1\big) = r_{ij}, \qquad \mathbb{P}\big(A_{ij}^\star = 0\big) = 1 - r_{ij}.$$

Substituting in equation 6 for the soft error:

$$\mathcal{E}_{\text{soft}} = r_{ij}(1 - w_{ij})^2 + (1 - r_{ij})\, w_{ij}^2 = r_{ij}(1 - r_{ij}) + (r_{ij} - w_{ij})^2 = r_{ij}(1 - r_{ij}) + \delta_{ij}^2.$$

Similarly for the hard error:

$$\mathcal{E}_{\text{hard}} = r_{ij}(1 - h_{ij})^2 + (1 - r_{ij})\, h_{ij}^2 = r_{ij}(1 - r_{ij}) + (r_{ij} - h_{ij})^2.$$

taking the difference for both soft and hard:

$$\mathcal{E}_{\text{hard}} - \mathcal{E}_{\text{soft}} = (w_{ij} + \delta_{ij} - h_{ij})^2 - \delta_{ij}^2 = \begin{cases} (1 - w_{ij})(1 - w_{ij} - 2\delta_{ij}), & h_{ij} = 1, \\ w_{ij}(w_{ij} + 2\delta_{ij}), & h_{ij} = 0. \end{cases}$$

This yields the sufficient condition:

$$-\frac{w_{ij}}{2} \leq \delta_{ij} \leq \frac{1 - w_{ij}}{2} \implies \mathcal{E}_{\text{soft}} \leq \mathcal{E}_{\text{hard}}.$$

A.3 HYPERPARAMETERS

Table 3: Hyperparameters for continual test-time adaptation results shown in Table 1.

Hyperparameter	CIFAR-10-C	CIFAR-100-C	ImageNet-C
Learning rate	10^{-4}	10^{-4}	10^{-3}
Batch size	200	200	100
Temperature	0.09	0.04	0.05
Buffer k	0	10	20
EMA momentum	0.99	0.999	0.9995
Number of augmentation views	10	8	6