REDUCE, REUSE, AND RECYCLE: NAVIGATING TEST-TIME ADAPTATION WITH OOD-CONTAMINATED STREAMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-Time Adaptation (TTA) aims to quickly adapt a pre-trained Deep Neural Network (DNN) to shifted test data from unseen distributions. Early TTA works only targeted simple and restrictive test scenarios that did not align with the philosophy of TTA that emphasizes practicality. Subsequent research efforts have thus been geared towards exploring more realistic test scenarios. In the same spirit, this work investigates for the first time TTA with data streams contaminated with out-of-distribution (OOD) data. Surprisingly, we observe the existence of benign OOD data that can improve TTA performance. We provide meaningful insights into the causes of benign OOD-contamination by analyzing the feature space of the pre-trained DNN. Inspired by these empirical findings, we propose R3, a novel TTA algorithm that specifically targets OOD-contaminated streams. Our experimental results verify that R3 improves competitive baselines by up to nearly 3%p on OOD-contaminated streams created with CIFAR-10-C and ImageNet-C.

1 INTRODUCTION

Powered by enormous datasets and computational resources, Deep Neural Networks (DNNs) continue to push the boundaries of machine intelligence (LeCun et al., 2015). DNNs, however, are still far from being the omnipotent machine learning model they are mistakenly advertised to be. One key limitation of DNNs is their failure to generalize to corrupted or shifted test data (Pan & Yang, 2010), which is referred to as shifted in-distribution (InD) data from here on. Improving their robustness to diverse distribution shifts thus remains a critical challenge when deploying DNNs in an open world.

Transductive inference aims to address the aforementioned limitation by adapting DNNs to specific distribution shifts (Gammerman et al., 2013). Because it is infeasible to anticipate the myriad of distribution shifts that may occur, transductive inference offers a more realistic solution than its inductive counterpart. Test-Time Adaptation (TTA) (Wang et al., 2020) distinguishes itself from other methods in transduction, such as Unsupervised Domain Adaptation (UDA) (Ganin & Lempitsky, 2015) and Test-Time Training (TTT) (Sun et al., 2020), in two distinct manners: 1) no direct access to training data is allowed, and 2) test data arrive in an online manner and cannot be re-visited. These assumptions make TTA the most practical approach to transduction.

Unlike early works in TTA that dealt with relatively simple and mild test scenarios (Niu et al., 2023), more recent research efforts are being directed towards exploring more challenging yet realistic test scenarios for TTA. For instance, Gong *et al.* (Gong et al., 2022) study temporally correlated, instead of i.i.d. data streams, and Niu *et al.* (Niu et al., 2023) investigate TTA in a "dynamic wild world," where test data have mixed and class-imbalanced distribution shifts. In the same vein, this work pioneers TTA with data streams that are contaminated with irrelevant out-of-distribution (OOD) data. As depicted in Figure 1, once deployed in an open world, the pre-trained DNN will inevitably encounter OOD data. Hence, investigation of the proposed test scenario, dubbed *OOD-contaminated streams*, allows for safer and more flexible deployment of pre-trained DNNs to a wide range of test scenarios without constraints.

Contrary to the popular belief that OOD data undermine the reliability of DNNs (Yang et al., 2021), we observe that performing TTA with OOD-contaminated streams improves the adaptation performance on some shifted InD data. This unusual observation alludes to the existence of "benign

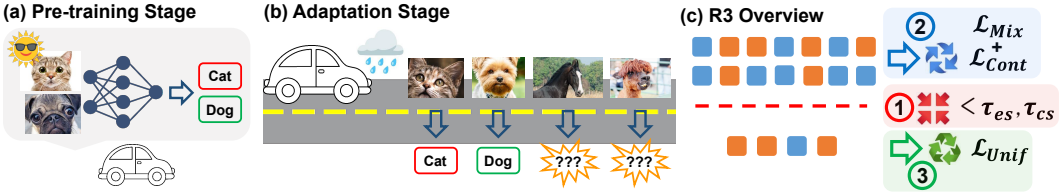

Figure 1: (a) Prior to deployment, the DNN is pre-trained on clean InD data, *e.g.,* data collected on a sunny day. (b) In the real world, the deployed DNN encounters a rainy condition and must be adapted accordingly. Unfortunately, current TTA protocols cannot handle OOD data that exist outside the label set of the pre-training data. (c) The conceptual overview of R3: ① Harmful test instances are reduced with two filtering thresholds. ② Remaining instances are reused via data mixup and contrastive learning. ③ Filtered instances are recycled into an auxiliary loss.

OOD-contamination," whose unrealized potential opens new doors for performance improvement that are unique to OOD-contaminated streams. To understand the cause of benign OOD-contamination, we analyze the feature space of the pre-trained DNN. Our analysis reveals that the feature spaces resided by shifted InD and OOD data overlap significantly, and as a result, OOD data that share the feature space with shifted InD data have the ability to facilitate the domain transfer from clean to shifted InD data. The highly entangled nature of the two data results in the additional side-effect of discarding shifted InD data when filtering out OOD data.

Inspired by these findings, we propose R3 (Reduce, Reuse, and Recycle), a novel TTA algorithm designed specifically for OOD-contaminated streams. To minimize the loss of shifted InD data, R3 conservatively identifies and *reduces* the amount of OOD-contamination with two cost-efficient metrics. The detected OOD instances are *recycled* into an auxiliary loss that drives their predictions closer to the Uniform distribution, effectively preventing the transfer of undesirable features. The remaining instances are *reused* for similarity-based mixup and contrastive learning with class-wise prototypes; these additional signals allow the pre-trained DNN to more robustly fit shifted InD data while preserving the original feature space. We demonstrate that R3 achieves state-of-the-art performance on two benchmark datasets, CIFAR-10-C and ImageNet-C, contaminated with various types of OOD data. Our contributions are largely three-fold:

- This is the first work to explore test-time model adaptation with OOD-contaminated streams that contain both shifted InD and irrelevant OOD data. Because it is impossible for the deployed DNN to evade OOD data in an open world, the proposed test scenario is far more realistic than TTA on curated data streams that explicitly contain targeted distribution shifts.
- We reveal the existence of beneficial OOD data that can assist in improving the adaptation performance on shifted InD data. Further examination of the pre-trained DNN's feature space brings to light that this surprising phenomenon is a result of the highly entangled nature of shifted InD and OOD data.
- We propose R3, a novel TTA algorithm that targets OOD-contaminated streams. Our extensive experimental results on diverse combinations of shifted InD and OOD data demonstrate the superiority of R3 to strong baselines from the TTA literature.

2 Related Works

2.1 Test-Time Adaptation and Its Progression

Test-Time Adaptation (TTA) is a popular branch of transductive inference (Gammerman et al., 2013) that concerns with adapting a pre-trained DNN to specific distribution shifts. The particular appeal of TTA, compared to other variants in transductive inference (*e.g.,* UDA (Ganin & Lempitsky, 2015; Ganin et al., 2016) and SFDA (Li et al., 2020b; Ding et al., 2022; Lee et al., 2022)), lies in its practicality. TTA performs adaptation without direct access to train data and under an online setting, in which test data cannot be revisited. To account for the lack of labels in test data, many of TTA methods perform adaptation by minimizing the entropy of the pre-trained DNN's softmax predictions on test data (Wang et al., 2020). Moreover, fully TTA methods (Boudiaf et al., 2022; Lim et al., 2023; Jang & Chung, 2022; Zhang et al., 2022; Choi et al., 2022) only update affine parameters and statistics of Batch Normalization layers (Ioffe & Szegedy, 2015) for fast and cost-efficient adaptation.

In the beginning, TTA only targeted elementary test scenarios, in which test data could be gathered into a batch to perform batch-wise adaptation, and each individual batch was sampled from the same shifted distribution. More recent works are starting to reflect additional obstacles that may arise in an open world. Continual TTA (Wang et al., 2022b; Song et al., 2023; Döbler et al., 2022) studies TTA in a non-stationary and continually changing test environment. Several works (Zhao et al., 2023; Gong et al., 2022) tackle class-imbalanced or temporally dependent data streams. Single-image TTA (Khurana et al., 2021) explores an extreme setting where a single test instance arrives at a time. Niu *et al.* (Niu et al., 2023) consolidate the above scenarios into a single setting named TTA in a dynamic wild world. To stably perform adaptation in more challenging scenarios, recent methods allow usage of auxiliary signals, such as data augmentation (Khurana et al., 2021) and/or partial information about source data (Döbler et al., 2022; Niu et al., 2022) even at the cost of increased computation and memory consumption. Our work is closely related to Open-set or Universal Domain Adaptation (Panareda Busto & Gall, 2017; Saito et al., 2018; You et al., 2019; Saito et al., 2020), but to the best of our knowledge, this is the first work to study open-world data streams in TTA.

2.2 Detection and Utilization of Out-of-Distribution Data

The presence of OOD data is often believed to significantly degrade the performance and reliability of DNNs. This conventional belief in machine learning has inspired research efforts to successfully detect and exclude OOD data from the inference process (Yang et al., 2021). The most straightforward approach to OOD detection is to utilize the DNN's predictive confidence scores (Hendrycks & Gimpel; Lee et al., a). Unfortunately, DNNs often output miscalibrated and over-confident predictions on OOD data, making confidence scores an unreliable indicator of OOD-ness (Guo et al., 2017). To overcome the drawback of confidence scores, a plethora of new OOD detection metrics, based on the energy-based interpretation of DNNs (Liu et al., 2020), temperature scaling (Liang et al.), rectified activations (Sun et al., 2021), virtual logit matching (Wang et al., 2022a), and various distance measures (Lee et al., 2018) have been suggested. The influence of OOD data is actively studied in various fields of research, including but not limited to: continual learning (Bang et al., 2022) and semi-supervised learning (Huang et al., 2021). Against the long-held belief that OOD data are harmful, some recent studies revealed that OOD data, when leveraged appropriately, can improve the generalization performance of DNNs (Park et al., 2021; Lee et al., b; Wei et al., 2022; Bai et al.).

3 Motivation

3.1 Existence of Benign OOD-contamination

It is easy to assume that performing TTA on shifted InD and OOD data together will lead to sub-optimal performance due to the distributional mismatch between the two. Interestingly enough, we reveal that the presence of OOD data in test streams does not always deteriorate the adaptation performance; we dub the subset of OOD data that can assist, rather than harm, the adaptation performance "benign OOD-contamination." In Figure 2, we compare the adaptation performance of vanilla TENT (Wang et al., 2020), the most widely-adopted TTA approach, on test streams with and without OOD-contamination. We assume that shifted InD and OOD data lie in the same shifted domain. Figure 2 (a) visualizes the results on three different types of CIFAR-10-C streams after inducing OOD-contamination with LSUN (Liang et al., 2017). Surprisingly, executing TENT with LSUN yields higher classification accuracy on CIFAR-10-C. In Figure 2 (b), the same trend is once again observed for ImageNet-C streams before and after OOD-contamination with iNaturalist (Van Horn et al., 2018), demonstrating that the existence of benign OOD-contamination is

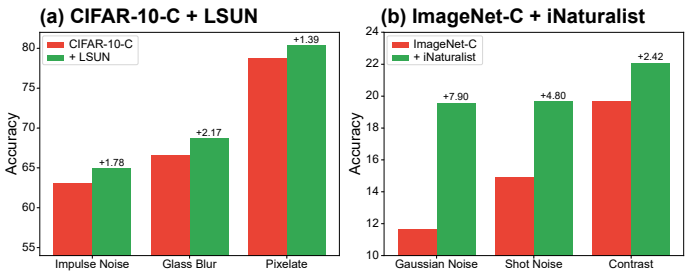

Figure 2: Comparison of TENT adaptation performance with and without OOD-contamination on (a) CIFAR-10-C and (b) ImageNet-C streams. LSUN and iNaturalist are used as OOD-contamination for CIFAR-10-C and ImageNet-C, respectively.

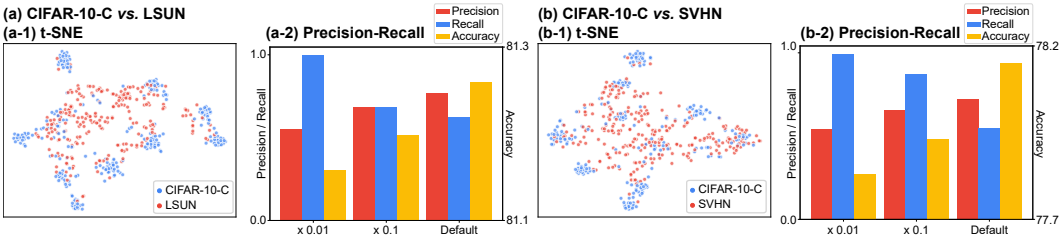

Figure 3: T-SNE and precision-recall analyses of CIFAR-10-C (a) + LSUN and (b) + SVHN. Shifted InD and OOD data are generated from the "Defocus Blur" domain. T-SNE plots show that in the feature space of the pre-trained DNN, shifted InD and OOD data appear nearly inseparable from each other. In precision-recall plots, precision and recall quantify the OOD detection performance on LSUN/SVHN, and accuracy indicates the adaptation performance on CIFAR-10-C. The adaptation accuracy (yellow) exhibits an inverse correlation with recall (blue) in OOD-contaminated streams.

not an isolated incident on smaller datasets. This empirical observation provides a tangible piece of evidence for benign OOD-contamination. In the next section, we analyze the feature space of the pre-trained DNN to offer insights into the phenomenon of benign OOD-contamination.

3.2 Empirical Analysis of the Feature Space

We employ t-SNE (Van der Maaten & Hinton, 2008) to visualize features extracted from penultimate layers of pre-trained DNNs. Figure 3 (a-1) and (b-1) show t-SNE of features of two OOD-contaminated streams: CIFAR-10-C + LSUN and + SVHN. In t-SNE plots, features of CIFAR-10-C and OOD data are colored in blue and red, respectively. Visualization results demonstrate that shifted InD and OOD data appear entangled in the feature space of the pre-trained DNN. This relationship is preserved even in the CIFAR-10-C + SVHN stream even though SVHN is commonly considered to be "far OOD" data that are relatively easier to detect. Such an empirical observation implies that shifted InD and OOD data share domain-specific characteristics, *e.g.,* weather or lighting condition, of the shifted domain. Because Batch Norm layers largely consist of domain-specific information (Li et al., 2017; Schneider et al., 2020), the adaptation protocol of TTA that only updates Batch Norm parameters can be interpreted as performing domain transfer from clean to shifted InD data while preserving the domain-invariant features of InD data. Consequently, during TTA, OOD data that belong in the same domain can provide auxiliary signals about the new shifted domain, thereby contributing to the adaptation performance (Chang et al., 2019; Kang et al., 2019).

A non-negligible side effect of this highly-entangled feature space is that undesirable loss of informative signals inevitably occurs during the filtering process. To corroborate this claim, we perform TENT with three different values of filtering thresholds based on the predictive entropy of a pre-trained DNN; test instances with higher predictive entropy than the threshold are excluded from the adaptation process. The default threshold is set to be $\log(1000) * 0.04$. In Figure 3 (a-2) and (b-2), we show changes in TENT performance according to different threshold values. The left axis shows the OOD detection performance in precision and recall, and the right axis shows the classification accuracy on InD data. TENT performs best in the high precision-low recall region of OOD detection, where OOD data are filtered in a conservative manner, and its performance starts to deteriorate as the precision decreases. Therefore, we deduce that preserving InD data is equally as important as removing harmful OOD-contamination, necessitating a more rigorous form of OOD filtering.

4 Methodology

In this section, we introduce R3, our proposed approach to TTA with OOD-contaminated streams. R3 first reduces the amount of wasteful data by conservatively identifying harmful OOD instances with two cost-efficient metrics for OOD-ness. The identified OOD instances are later recycled and reformulated into an auxiliary loss function that is designed to facilitate a selective transfer of features (Section 4.1). Afterwards, R3 implements similarity-based mixup and contrastive learning with class-wise prototypes by reusing the unfiltered instances (Section 4.2).

Preliminaries and Notations: Let us assume that we have a DNN $f_\theta(x)$ that is parameterized with learnable parameters θ and has been pre-trained on clean InD data $\mathcal{D}_{\text{tr}} = \{(x^i_{\text{tr}}, y^i_{\text{tr}})\}^{L_{\text{tr}}}_{i=1}$,

where $x^i_{\text{tr}} \in \mathcal{X}_{\text{tr}}$ and $y^i_{\text{tr}} \in \mathcal{C}_{\text{tr}}$. TTA aims to adapt $f_\theta(x)$ on arbitrary test data $\mathcal{D}_{\text{te}} = \{(x^i_{\text{te}})\}^{L_{\text{te}}}_{i=1}$. Unless specified otherwise, TTA mitigates the absence of test labels by employing the entropy minimization loss (Wang et al., 2020): $\min -\sum_c y'_{\text{te},c} \log (y'_{\text{te},c})$, where $y'_{\text{te},c} = f_\theta(c|x_{\text{te}})$ denotes the DNN's predictions on a class c. Features extracted from the penultimate layer of the pre-trained DNN are denoted by $z \in \mathcal{Z}$, where $\mathcal{Z}$ corresponds to the feature space. Then, the per-class mean of features computed on the clean train data, *i.e.,* class-wise prototypes, can be denoted as: $r_c = (\sum \mathbb{1}[y^i_{\text{tr}} = c]\ z^i_{\text{tr}}) / \sum \mathbb{1}[y^i_{\text{tr}} = c]$.

In this work, we consider a challenging test scenario in which $\mathcal{D}_{\text{te}}$ consists of both shifted but label-sharing InD data of our interest ($\tilde{\mathcal{D}}_{\text{te}} = \{(\tilde{x}^i_{\text{te}})\}^{\tilde{L}_{\text{te}}}_{i=1}$) and irrelevant OOD data ($\hat{\mathcal{D}}_{\text{te}} = \{(\hat{x}^i_{\text{te}})\}^{\hat{L}_{\text{te}}}_{i=1}$): by definition, $\tilde{\mathcal{D}}_{\text{te}} \cup \hat{\mathcal{D}}_{\text{te}} = \mathcal{D}_{\text{te}}$, and $\tilde{\mathcal{D}}_{\text{te}} \cap \hat{\mathcal{D}}_{\text{te}} = \emptyset$. $\tilde{\mathcal{D}}_{\text{te}}$ shares the same label set as train data but has different data distribution: $\tilde{\mathcal{X}}_{\text{test}} \neq \mathcal{X}_{\text{train}}, \tilde{\mathcal{C}}_{\text{test}} = \mathcal{C}_{\text{train}}$. In the case of $\hat{\mathcal{D}}_{\text{te}}$, however, its data and label sets both diverge away from those of $\mathcal{D}_{\text{train}}$: $\hat{\mathcal{X}}_{\text{test}} \neq \mathcal{X}_{\text{train}}, \hat{\mathcal{C}}_{\text{test}} \neq \mathcal{C}_{\text{train}}$.

4.1 Reduce & Recycle: Conservative OOD Data Filtering

R3 employs two metrics to identify harmful OOD instances prior to performing the adaptation process. For the sake of conciseness, notations for a test instance x_{te} and the associated DNN output y'_{te} is simplified as x and y' from here on. The first measure of OOD-ness in R3 is the energy score (Liu et al., 2020), defined to be: $ES(x; f_\theta) = -T \cdot \log \sum^{\mathcal{C}}_i e^{y'_i/T}$, where T is equivalent to the temperature of the Softmax function. The energy score modifies the naïve confidence score to reduce over-confident predictions on OOD data and increase the separability between InD and OOD data.

However, unlike clean InD data, shifted InD appear intricately entangled with OOD data in the feature space of a pre-trained DNN as shown in Section 3.2. Consequently, the energy score alone remains an insufficient measure for identifying harmful OOD data without sacrificing shifted InD data that we wish to preserve. R3 thus introduces a second measure based on the cosine similarity between the features and class-wise prototypes:

$$CS(x; f_\theta) = \max_c \ \frac{r_c \cdot z}{||r_c||\ ||z||}, \ \text{ for } \ c \in \mathcal{C}_{\text{tr}}. \tag{1}$$

r_c is pre-computed on clean InD data following the pre-training stage and is not updated during the adaptation stage. CS essentially quantifies the similarity between the penultimate feature of a test instance and the closest class-wise prototype. With these two metrics at hand, R3 filters out a test instance only if its negative energy score and cosine similarity both fall below pre-set thresholds:

$$\text{If } \ \mathbb{1}[-ES(x) < \tau_{\text{es}}] \cdot \mathbb{1}[CS(x) < \tau_{\text{cs}}] = \begin{cases} 1, & \text{then}, \quad x^i \in S_{\text{M}} \\ 0, & \text{then}, \quad x^i \in S_{\text{B}}. \end{cases} \tag{2}$$

τ_{es} and τ_{cs} are filtering thresholds for ES and CS, respectively, and are treated as separate hyper-parameters of R3. By introducing a more rigorous set of criteria for eliminating test instances, R3 effectively minimizes the chances of shifted InD instances being unintentionally excluded from the adaptation process. S_{B} and S_{M} refer to a set of InD and OOD instances, identified from the DNN's viewpoint, and an instance in S_{B} and S_{M} is denoted by x_{B} and x_{M}.

Instead of discarding S_{M}, R3 recycles them into an auxiliary loss that minimizes the cross entropy between the Uniform distribution and the DNN's predictions on instances in S_{M}: $\mathcal{L}_{\text{Unif}} = \min -\sum_c \frac{1}{c} \log (y'_{\text{M},c})$. This loss function has been shown to improve the DNN's generalization performance by preventing it from learning irrelevant features (Lee et al., b). Consequently, when utilized for TTA, it can enforce a selective transfer of features that are useful for the shifted domain.

4.2 Reuse: Similarity-based Mixup & Contrastive Learning with Class-wise Prototypes

The absence of test labels creates an inherently noisy learning signal, which is exacerbated by the conservative OOD filtering scheme in R3. To robustify the adaptation process, we employ mixup (Zhang et al., 2017), a popular form of noise-robust learning (Berthelot et al., 2019; Li et al., 2020a). Original mixup randomly mixes two images in a batch with a mixup coefficient sampled from a beta distribution. R3 modifies the data mixing process by adopting the maximum cosine

similarity between the feature and class-wise prototypes as a novel mixup coefficient:

$$x_{\text{mix}} = \alpha \cdot x^i_{\text{B}} + (1-\alpha) \cdot x^j_{\text{B}}, \;\; \text{where} \;\; \alpha = CS(x^i_{\text{B}}; f_\theta). \tag{3}$$

Because $CS(x^i_{\text{B}}; f_\theta)$ is readily available from the previous step, R3's interpretation of mixup does not incur any additional cost compared to the original mixup. With the redefined mixup coefficients, the instances that are closer to class-wise prototypes are weighted more heavily, whereas the opposite holds true for the instances that are farther away from class-wise prototypes. After performing similarity-based mixup, we repeat the filtering process in Eq. (2) to select S_{mix}. Unlike the original mixup, R3 cannot perform mixup in the label space due to the lack of test labels. Instead, R3 defines a separate entropy minimization loss with the instances in S_{mix}: $\mathcal{L}_{\text{mix}} = \min - \sum_c y'_{\text{mix},c} \log (y'_{\text{mix},c})$.

Lastly, R3 performs supervised contrastive learning (Khosla et al., 2020) between penultimate features z of S_{B} and S_{mix} and the class-wise prototypes:

$$\mathcal{L}_{\text{Cont}} = -\sum_i \frac{\exp(\text{sim}(z_i, r_c)) \;/\; T_{\text{con}}}{\sum_{c' \in \mathcal{C}_{\text{tr}}} \exp(\text{sim}(z_i, r_{c'})) \;/\; T_{\text{con}}}, \;\; \text{where} \;\; T_{\text{con}} = 0.1. \tag{4}$$

$\mathcal{L}_{\text{Cont}}$ prevents the feature space from deviating much from that of clean InD data during the adaptation process. Because R3 only updates Batch Norm parameters, it is preferable to retain the feature space of the penultimate layer while alleviating the domain shift, such that the classification layer, which is frozen after the pre-training stage, can map features to correct classes at ease.

Overall Optimization Scheme: The final learning objective of R3 can be expressed as: $\mathcal{L}_{\text{R3}} = \lambda_{\text{Ent}} \mathcal{L}_{\text{Ent}} + \lambda_{\text{Mix}} \mathcal{L}_{\text{Mix}} + \lambda_{\text{Cont}} \mathcal{L}_{\text{Cont}} + \lambda_{\text{Unif}} \mathcal{L}_{\text{Unif}}$, where λ is the coefficient for the associated loss term and is treated as a hyperparameter. The pseudo-code for R3 that encompasses all of the above components and information on how to tune relevant hyperparameters prior to deployment are included in Sections A6 and A7 of Appendix.

5 Experiments

5.1 Experimental Set-up

Baselines: Detailed description of baseline TTA approaches used for comparison can be found in Section A1 of Appendix. These baselines, carefully selected from TTA literature, are a fair reflection of the state-of-the-art in TTA research.

Datasets and Implementation Details: R3 and compared approaches are verified on two types of shifted InD data: CIFAR-10-C (Krizhevsky et al., 2009) and ImageNet-C (Deng et al., 2009). When adapting the DNN on CIFAR-10-C, we consider the following types of OOD-contamination: LSUN (Crop) (Liang et al., 2017), SVHN (Netzer et al., 2011), and Describable Textures Dataset (DTD) (Cimpoi et al., 2014). In the case of ImageNet-C, we induce OOD-contamination with iNaturalist (Van Horn et al., 2018) and DTD. Realistically, OOD data would appear under the same domain or corruption as shifted InD data. Therefore, we apply the same set of corruptions to OOD data following the protocol provided by Hendrycks *et al.* (Hendrycks & Dietterich, 2018). Due to the page constraint, implementation details, including the choice of architectures, pre-training protocols, and test stream configurations, can be found in Section A1 of Appendix.

5.2 Verification under separate and mixed corruption scenarios

We first consider the case where each type of shifted InD data arrives in a separate manner with clear boundaries. We assume that the OOD data are corrupted in the same manner as InD data. The results in terms of the average classification accuracy across all fifteen corruptions types are reported in Table 1. The table also includes standard deviations of adaptation performances computed over five different random seeds. To provide an empirical upper bound performance for reference, we report the performance of TENT on a clean, uncontaminated stream that only contains shifted InD data ("Clean Stream"). On the LSUN-contaminated stream, R3 exhibits further performance improvement from TENT. Furthermore, R3 consistently achieves the best performance on streams contaminated with SVHN and DTD. Note that out of the three OOD datasets, only LSUN, which is "near OOD" (Sastry & Oore, 2020), improves the adaptation performance of TENT. This result supports that near OOD data that bear close resemblance to shifted InD data are more likely to benefit the adaptation process.

Table 1: Comparison against state-of-the-art TTA methods under the **separate** corruption scenario. We report the classification accuracy (%) on CIFAR-10-C averaged over all 15 corruption types. The best result under each OOD-contaminated stream is marked in bold.

Stream	Method	LSUN	SVHN	DTD	LSUN + SVHN	LSUN + DTD	DTD + SVHN
Separate	Test				63.31		
	BN	76.17 ±0.15	69.74 ±0.62	68.54 ±0.13	72.69 ±0.20	73.01 ±0.05	68.92 ±0.09
	TENT	76.39 ±0.17	69.74 ±0.34	69.48 ±0.20	73.39 ±0.10	73.44 ±0.08	69.61 ±0.07
	TENT_f	76.17 ±0.15	67.57 ±0.56	69.31 ±0.32	71.35 ±0.72	73.39 ±0.17	68.23 ±0.30
	EATA	76.12 ±0.12	69.81 ±0.56	68.25 ±0.17	74.09 ±0.54	73.05 ±0.13	68.52 ±0.16
	SAR	76.17 ±0.14	69.75 ±0.63	68.53 ±0.13	72.69 ±0.15	73.01 ±0.07	68.92 ±0.09
	R3	**76.64** ±0.13	**72.08** ±0.20	**72.41** ±0.27	**75.16** ±0.15	**75.10** ±0.30	**71.99** ±0.24
Clean Stream		76.27 ±0.16					

Table 2: Comparison against state-of-the-art TTA methods under the **mixed** corruption scenario. We report the classification accuracy (%) on CIFAR-10-C following the adaptation process. The best result under each OOD-contaminated stream is marked in bold.

Stream	Method	LSUN	SVHN	DTD	LSUN + SVHN	LSUN + DTD	DTD + SVHN
Mixed	Test				64.38		
	BN	65.44 ±0.90	59.00 ±1.07	58.51 ±0.66	64.34 ±1.10	64.58 ±0.92	58.46 ±0.91
	TENT	65.97 ±1.09	59.61 ±0.82	57.52 ±0.96	64.95 ±1.31	64.71 ±1.47	58.21 ±1.03
	TENT_f	65.36 ±1.03	59.85 ±0.89	57.16 ±0.72	64.22 ±1.12	64.83 ±1.70	57.84 ±1.98
	EATA	67.89 ±1.38	60.05 ±1.49	55.02 ±1.71	64.34 ±1.11	64.09 ±0.98	59.31 ±1.30
	SAR	66.09 ±0.88	58.99 ±1.08	58.51 ±0.66	64.34 ±1.10	64.58 ±1.68	58.46 ±1.91
	R3	**67.93** ±0.89	**61.81** ±1.60	**58.87** ±0.83	**66.17** ±0.88	**67.40** ±1.61	**62.25** ±0.92
Clean Stream		63.12 ±0.89					

We additionally validate that LSUN is indeed closer to CIFAR-10-C by quantifying the similarity of OOD datasets to CIFAR-10-C in terms of the 2-Wasserstein distance (Givens & Shortt, 1984). The results and analysis, included in Section A8 of Appendix, uphold that the proximity of LSUN to CIFAR-10-C is what makes it benign OOD-contamination.

We now consider the test scenario in which all types of shifted InD data appear together with no boundary. Likewise, OOD data are corrupted with a mixture of corruption types. We compare the classification accuracy at the end of the adaptation process in Table 2. We observe that LSUN again improves the performance of TENT, demonstrating that it can serve as a benign form of OOD-contamination in different test scenarios. Across all OOD-contamination types, R3 consistently attains the best performance among all the compared approaches.

5.3 Validation on a Larger Dataset

We now verify that the effectiveness of R3 can be scaled to a more complex dataset through experiments on OOD-contaminated ImageNet-C streams. Analogous to the CIFAR-10-C experiments, we compare all approaches under both separate and mixed scenarios. We present the performance of R3 and those of compared approaches in Table 3. R3 exhibits a significant degree of performance improvement under both deployment scenarios and across various OOD-contamination types. These results provide concrete evidence for the scalability and universality of R3.

5.4 Additional Results and Discussion

(1) Compatibility with Various TTA Algorithms: R3 is primarily designed for and implemented in conjunction with the entropy minimization-based loss for TTA. To demonstrate that R3 can be utilized with a broader range of TTA algorithms, we combine R3 with TTA methods based on pseudo-labeling (Goyal et al., 2022) and augmentation invariance (Zhang et al., 2022) and report the results in Table A1 of Appendix. The results clearly show that the performance improvement brought upon by R3 is not exclusive to the entropy minimization-based algorithm for TTA.
(2) Performance on Clean Test Streams: Through comparison against state-of-the-art TTA approaches in Table A2 of Appendix, we validate that R3 can be used for adaptation on clean test

Table 3: Comparison against state-of-the-art TTA approaches on ImageNet-C under separate and mixed corruption scenarios. The best classification accuracy (%) in each column is marked in bold.

Method	Separate			Mixed		
	iNat	DTD	iNat + DTD	iNat	DTD	iNat + DTD
Test	——— 26.65 ———			——— 25.52 ———		
BN	25.62 ±0.14	27.57 ±0.15	27.83 ±0.11	16.37 ±0.85	17.60 ±0.78	19.21 ±0.10
TENT	26.09 ±0.07	27.90 ±0.15	28.57 ±0.21	15.98 ±0.80	15.26 ±1.35	17.89 ±0.57
TENT_f	25.71 ±0.14	27.57 ±0.14	27.92 ±0.09	14.17 ±1.33	15.69 ±1.67	17.36 ±0.97
EATA	25.81 ±0.12	27.47 ±0.15	27.30 ±0.33	17.30 ±0.46	17.28 ±0.71	18.31 ±0.67
SAR	25.61 ±0.14	27.57 ±0.15	27.83 ±0.11	16.38 ±0.72	16.93 ±0.66	19.20 ±0.79
R3	**27.50** ±0.32	**30.83** ±0.22	**30.58** ±0.16	**18.43** ±0.42	**19.50** ±0.78	**20.99** ±0.90
Clean Stream	30.75 ±0.09			18.90 ±0.57		

streams, void of OOD-contamination. R3 visibly outperforms competitive baselines on both clean test streams, showcasing its capability to handle a variety of test scenarios.

(3) OOD Detection Metric: We investigate R3 from the perspective of OOD detection by analyzing H-score, defined as: $H = (2 * ACC_{\text{InD}} * ACC_{\text{OOD}})/(ACC_{\text{InD}} + ACC_{\text{OOD}})$, where ACC_{InD} and ACC_{OOD} refer to the classification accuracy on shifted InD data and OOD detection accuracy, respectively. In Table A3 of Appendix, H-scores of compared approaches measured on one of the CIFAR-10-C + DTD streams are reported. R3 successfully improves the classification accuracy on shifted InD data while maintaining competitive OOD detection accuracy. This result further elucidates that some OOD instances are more beneficial for TTA than others because the compared approaches perform comparably on ACC_{OOD} but still show disparities in ACC_{InD}. Moreover, while R3 does not lead to a noticeable improvement in ACC_{OOD}, it is capable of identifying these benign OOD instances and incorporating them into the adaptation process to effectively boost ACC_{InD}.

6 Ablation Study and Hyperparameter Sensitivity Analysis

6.1 Component-wise Ablation Study

In Figure 4, we visualize the change in the adaptation performance of R3 on CIFAR-10-C + SVHN and ImageNet-C + DTD streams as the filtering scheme and each one of the loss functions are applied incrementally. Each technical component clearly contributes to improving the performance of R3, allowing it to achieve superior performance as demonstrated in earlier sections.

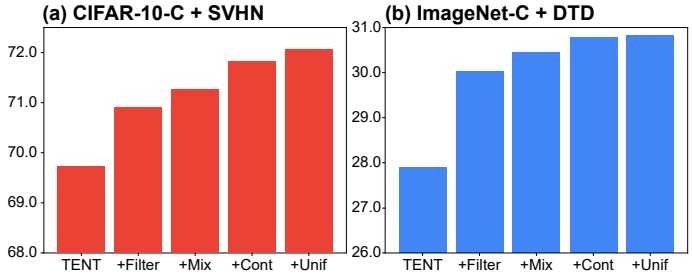

Figure 4: Ablation study on two different OOD-contaminated streams: (a) CIFAR-10-C + SVHN and (b) ImageNet-C + DTD.

Furthermore, we explore different design choices for the filtering scheme and mixup component in R3. First, we study the results of solely using energy-based filtering for compared approaches in Table A4. Two conclusions can be derived from these results. 1) Utilizing energy-based filtering deteriorates the performance of compared approaches in most cases. This result indicates that the use of a more competitive OOD detection method does not always lead to improvement in adaptation performance, confirming the existence of beneficial OOD data. 2) R3 performs better with the proposed dual filtering scheme, indicating that the conservative filtering method in R3 is more effective at conserving beneficial OOD data for TTA than conventional OOD detection scores.

Second, we replace similarity-based mixup in R3 with two more commonly-used forms of mixup - original randomized mixup and CutMix Yun et al. (2019) - and compare their performances in Table A5. Utilizing the proposed similarity-based mixup consistently outperforms the other two, justifying our design of similarity-based mixing coefficients. Collectively, these results support that

Table 4: Results of using different batch sizes for online adaptation. R3 consistently surpasses the two strongest baseline approaches.

BS	Method	C10-C + SVHN	ImgNet-C + DTD
$\times 2$	TENT	70.46	30.25
	SAR	70.44	30.01
	R3	72.05	31.84
$\times \frac{1}{2}$	TENT	66.60	23.69
	SAR	68.55	23.52
	R3	69.52	28.72

Table 5: Results of using different InD-to-OOD data ratios. R3 comes on top of the two other baselines even if the data ratio is changed.

Ratio	Method	C10-C + SVHN	ImgNet-C + DTD
1 : 2	TENT	67.15	27.29
	SAR	66.25	27.07
	R3	68.92	27.68
2 : 1	TENT	73.48	31.72
	SAR	72.45	31.30
	R3	74.98	33.01

although R3 relies on existing ideas, deliberate modifications introduced in R3, *e.g.,* re-defining the mixing coefficient or employing conservative dual filtering, play a critical role in its success.

6.2 Hyperparameter Sensitivity Analysis

Data stream configuration To show that R3 is robust to changes in test scenarios, we report the results of altering two major factors in data stream configuration: the batch size used for online adaptation and the ratio of InD to OOD data. In Table 4, we report the results of doubling and halving default batch sizes used for CIFAR-10-C and ImageNet-C streams. R3 consistently comes on top regardless of the batch size setting. In particular, R3 improves the performance on ImageNet-C + DTD stream by almost 5%p when the batch size is halved. According to Table 5, R3 successfully maintains its competitive performance even when the ratio of InD-to-OOD data changes.

Filtering thresholds & Loss coefficients We perform a hyperparameter sensitivity analysis of two filtering thresholds and four weighting coefficients for loss terms to further confirm the stability of R3. Figure 5 (a) visualizes changes in R3 performance according to different threshold values ($\tau_{\text{es,cs}}$). Similarly, changes in the performance of R3 according to different coefficients for the loss terms ($\lambda_{\text{Ent,Mix,Cont,Unif.}}$) are shown in Figure 5 (b). All other hyperparameters remain unchanged. The grey dotted lines indicate the TENT performance. Within the tested range, R3 steadily outperforms TENT, a strong baseline approach, suggesting that finding the optimal set of hyperparameters is not too difficult.

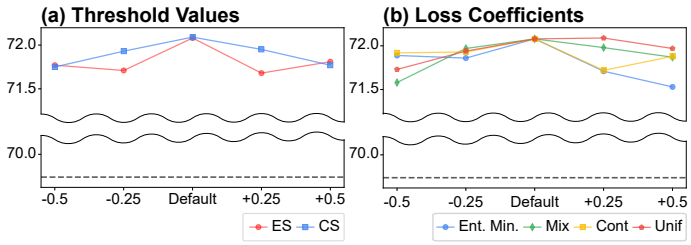

Figure 5: Sensitivity of R3 to (a) different values of the energy score (red) and cosine similarity (blue) thresholds; and (b) to varying coefficients for the four separate loss terms.

7 Concluding Remarks

This paper studied for the first time TTA with OOD-contaminated data streams, a realistic TTA scenario of grave importance. We unearthed the existence of benign OOD data that can improve, rather than harm, the adaptation performance on shifted InD data. To delve into the intriguing phenomenon of benign OOD-contamination, we empirically analyzed the feature space of the pre-trained DNN; our analysis revealed that two types of data share domain-specific characteristics, allowing some OOD data to aid in the domain transfer process in TTA. Motivated by such analytical results, we proposed R3, a novel TTA algorithm designed for OOD-contaminated streams, and showcased its effectiveness and versatility through extensive experiments that span combinations of two shifted InD datasets and four OOD datasets. As a pioneering investigation of its kind, this work will contribute to promoting the safe and robust deployment of pre-trained DNNs in an open world.

REFERENCES

Jianhong Bai, Zuozhu Liu, Hualiang Wang, Jin Hao, Yang Feng, Huanpeng Chu, and Haoji Hu. On the effectiveness of out-of-distribution data in self-supervised long-tail learning. In *The Eleventh International Conference on Learning Representations*.

Jihwan Bang, Hyunseo Koh, Seulki Park, Hwanjun Song, Jung-Woo Ha, and Jonghyun Choi. Online continual learning on a contaminated data stream with blurry task boundaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9275–9284, 2022.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. *arXiv preprint arXiv:2201.05718*, 2022.

Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 7354–7362, 2019.

Sungha Choi, Seunghan Yang, Seokeon Choi, and Sungrack Yun. Improving test-time adaptation via shift-agnostic weight regularization and nearest source prototypes. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pp. 440–458, 2022.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7212–7222, 2022.

Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. *arXiv preprint arXiv:2211.13081*, 2022.

Alex Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. *arXiv preprint arXiv:1301.7375*, 2013.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Clark R Givens and Rae Michael Shortt. A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984.

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

Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and J Zico Kolter. Test time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems*, 35:6204–6218, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Junkai Huang, Chaowei Fang, Weikai Chen, Zhenhua Chai, Xiaolin Wei, Pengxu Wei, Liang Lin, and Guanbin Li. Trash to treasure: harvesting ood data with cross-modal matching for open-set semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8310–8319, 2021.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Minguk Jang and Sae-Young Chung. Test-time adaptation via self-training with nearest neighbor information. *arXiv preprint arXiv:2207.10792*, 2022.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4893–4902, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

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

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 12365–12377. PMLR, 2022.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, a.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Saehyung Lee, Changhwa Park, Hyungyu Lee, Jihun Yi, Jonghyun Lee, and Sungroh Yoon. Removing undesirable feature contributions using out-of-distribution data. In *International Conference on Learning Representations*, b.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020a.

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

Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. *arXiv preprint arXiv:2302.05155*, 2023.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

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*, pp. 16888–16905. PMLR, 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. *arXiv preprint arXiv:2302.12400*, 2023.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, pp. 754–763, 2017.

Dongmin Park, Hwanjun Song, MinSeok Kim, and Jae-Gil Lee. Task-agnostic undesirable feature deactivation using out-of-distribution data. *Advances in Neural Information Processing Systems*, 34:4040–4052, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 153–168, 2018.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, and Kate Saenko. Universal domain adaptation through self supervision. *Advances in neural information processing systems*, 33:16282–16292, 2020.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pp. 8491–8501. PMLR, 2020.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551, 2020.

Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. *arXiv preprint arXiv:2303.01904*, 2023.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

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

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

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*, 2020.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022a.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. *arXiv preprint arXiv:2203.13591*, 2022b.

Hongxin Wei, Lue Tao, Renchunzi Xie, Lei Feng, and Bo An. Open-sampling: Exploring out-of-distribution data for re-balancing long-tailed datasets. In *International Conference on Machine Learning*, pp. 23615–23630. PMLR, 2022.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2720–2729, 2019.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems*, 35:38629–38642, 2022.

Bowen Zhao, Chen Chen, and Shu-Tao Xia. Delta: degradation-free fully test-time adaptation. *arXiv preprint arXiv:2301.13018*, 2023.

A1 BASELINES AND IMPLEMENTATION DETAILS

Following are the baseline methods from TTA literature that R3 is compared to.

- Test: evaluates the pre-trained DNN on new test data without additional modification.
- BN Adapt: replaces the Batch Norm statistics of the pre-trained DNN with those of test data. We use the abbreviation "BN" to refer to this baseline approach.
- TENT (Wang et al., 2020): updates the statistics and affine parameters of Batch Norm layers by minimizing the entropy-based loss. TENT with filtering (TENT_f) performs TENT after removing high-entropy instances. The filtering threshold is set to be $\log(1000) * 0.04$.
- EATA (Niu et al., 2022): utilizes the Fisher regularizer to preserve important parameters and performs instance selection and re-weighting based on the DNN's predictive entropy.
- SAR (Niu et al., 2023): discards unstable test instances with large gradients and replaces the vanilla SGD optimizer with a sharpness-aware minimization optimizer.

All our experiments are implemented using PyTorch (Paszke et al., 2019) and conducted with NVIDIA V100 GPU. We use ResNet-50 (He et al., 2016) with Batch Norm layers for experiments on CIFAR-10-C and ImageNet-C. For CIFAR-10 pre-training, we use the SGD optimizer with the initial learning rate of 0.1, annealed at cosine rate, momentum of 0.9, and weight decay of 0.0005. We use the model provided by the PyTorch timm (Wightman, 2019) as the pre-trained model for ImageNet-C. For the adaptation process, we use the SGD optimizer with a learning rate of 0.0025 and momentum of 0.9. We use the batch size of 32 and 64 for CIFAR-10-C and ImageNet-C, respectively. The default ratio of InD to OOD data is set to be 1:1.

A2 COMPATIBILITY WITH OTHER TTA FRAMEWORKS

Table A1: Compatibility with a broader range of TTA algorithms. PL refers to the pseudo-labeling-based method. MEMO is a method based on augmentation invariance.

Stream	Method	CIFAR-10-C + LSUN	CIFAR-10-C + SVHN	CIFAR-10-C + DTD	ImageNet-C + iNat	ImageNet-C + DTD
Separate	PL	75.78	70.56	70.97	26.21	28.08
	+ R3	76.64	71.79	72.63	27.53	29.32
	MEMO	76.51	74.02	74.54	27.86	27.47
	+ R3	76.83	74.95	75.24	28.67	28.98
Mixed	PL	66.10	60.27	56.62	17.11	16.52
	+ R3	67.65	61.69	58.63	18.26	17.73
	MEMO	67.65	61.69	58.63	18.26	17.41
	+ R3	67.32	65.36	63.85	17.89	19.04

A3 ADAPTATION RESULTS ON CLEAN STREAMS

Table A2: Adaptation results on "clean" streams without OOD data.

Stream	Dataset	EATA	SAR	R3
Sep	CIFAR-10-C	78.19	78.01	79.66
	ImageNet-C	30.14	30.81	31.79
Mix	CIFAR-10-C	63.36	64.57	65.01
	ImageNet-C	17.82	19.37	19.92

A4 OOD DETECTION PERFORMANCE

Table A3: H-score analysis on one of the CIFAR-10-C + DTD streams. H-score allows us to simultaneously consider the classification accuracy on shifted InD data and OOD detection accuracy. R3 again exceeds other strong baselines in terms of H-score.

	ACC_{InD}	ACC_{OOD}	H-score
$\mathrm{TENT_f}$	69.70	53.00	60.21
EATA	68.40	52.80	59.59
SAR	68.39	53.34	59.94
R3	72.50	53.38	61.89

A5 ADDITIONAL DESIGN CHOICE EXPLORATION

Table A4: The filtering component in all compared approaches is replaced with OOD filtering based on the energy score, an advanced metric of OOD-ness.

Stream	Method	CIFAR-10-C + LSUN	CIFAR-10-C + SVHN	CIFAR-10-C + DTD	ImageNet-C + iNat	ImageNet-C + DTD
Sep	TENT	74.82	68.61	68.62	25.73	26.80
	EATA	74.74	68.62	68.24	24.66	27.17
	SAR	74.91	68.57	68.58	24.73	27.21
	R3	76.46	70.82	70.97	26.29	29.07
Mix	TENT	65.85	58.99	56.51	15.70	17.23
	EATA	65.69	60.17	55.02	17.05	18.06
	SAR	66.09	58.99	57.63	17.08	18.14
	R3	66.91	61.27	58.02	17.79	18.88

Table A5: Results of executing R3 with original mixup and CutMix.

Stream	Method	CIFAR-10-C + LSUN	CIFAR-10-C + SVHN	CIFAR-10-C + DTD	ImageNet-C + iNat	ImageNet-C + DTD
Sep	Original	76.61	71.76	71.31	26.07	27.80
	CutMix	76.61	70.93	70.88	25.43	27.39
	Ours	76.64	72.08	72.41	27.50	30.83
Mix	Original	66.74	60.10	58.09	16.98	18.02
	CutMix	66.71	60.27	57.59	16.95	17.51
	Ours	67.93	61.81	58.87	18.43	19.50

Table A6: Results of replacing the filtering component in other algorithms with the proposed filtering scheme, which is denoted by R.F., a shorthand for R3 filtering scheme.

Stream	Method	CIFAR-10-C + LSUN	CIFAR-10-C + SVHN	CIFAR-10-C + DTD	ImageNet-C + iNat	ImageNet-C + DTD
Sep	EATA	76.12	69.81	68.25	25.81	27.47
	+ R.F.	76.30	71.45	71.50	26.87	28.90
	SAR	76.17	69.45	68.53	25.61	27.57
	+ R.F.	76.26	71.08	69.45	25.97	28.91
Mix	EATA	67.89	60.05	55.02	17.30	17.28
	+ R.F.	67.89	61.07	56.95	17.93	17.92
	SAR	66.09	58.99	58.51	16.38	16.93
	+ R.F.	66.27	59.36	58.62	17.08	17.57

A6 OVERALL ALGORITHM FOR R3

Algorithm 1: Test-Time Adaptation with R3

Require: Pre-trained DNN $f_\theta(x)$, Class-wise Prototypes r_c, Test Data Batch $\mathcal{D}_{\text{te}} = \{(x^i_{\text{te}})\}_{i=1}^{B}$
Compute $ES(x^i_{\text{te}}; f_\theta)$ and $CS(x^i_{\text{te}}; f_\theta)$ for $x^i_{\text{te}} \in \mathcal{D}_{\text{te}}$
Determine S_{B} and S_{M} according to Eq. (2)
Compute α and generate x_{mix} for $x_{\text{B}} \in S_{\text{B}}$ according to Eq. (3)
Determine S_{mix} according to Eq. (2)
Compute two entropy minimization losses $\mathcal{L}_{\text{Ent}}(x_{\text{B}})$ and $\mathcal{L}_{\text{Mix}}(x_{\text{mix}})$, where $x_{\text{mix}} \in S_{\text{mix}}$
Compute contrastive loss $\mathcal{L}_{\text{Cont}}$ with $S_{\text{ind}}, S_{\text{mix}}$, and r_c according to Eq. (4)
Compute uniform loss $\mathcal{L}_{\text{Unif}}$ with S_{M}
Update the parameters of Batch Normalization layers using $\mathcal{L}_{\text{R3}}$

A7 R3 HYPERPARAMETER CONFIGURATION

In Table A7, we report a detailed hyperparameter configuration used for R3. The same set of hyperparameters is used for each stream under separate and mixed corruption scenarios. After each update step k, the filtering thresholds and loss coefficients are adjusted at the rate of γ as follows: $\tau_{k+1} = 0.9 * \tau_k + (0.1 * \gamma)\tau_k$, and $\lambda_{k+1} = 0.9 * \lambda_k + (0.1 * \gamma)\lambda_k$.

Table A7: R3 Hyperparameter configurations for different OOD-contaminated streams.

Stream Type	τ_{es}	τ_{cs}	λ_{Ent}	λ_{Mix}	λ_{Cont}	λ_{Unif}	γ
CIFAR-10-C +							
LSUN	6.0	0.8	1.8	1.8	0.1	0.1	0.999
SVHN	6.0	0.8	2.0	2.0	1.5	1.5	0.999
DTD	6.0	0.8	2.0	2.0	1.5	1.5	0.999
LSUN + SVHN	6.0	0.8	2.0	2.0	1.5	1.5	0.999
LSUN + DTD	6.0	0.8	2.0	2.0	1.5	1.5	0.999
DTD + SVHN	6.0	0.8	2.0	2.0	1.5	1.5	0.999
ImageNet-C +							
iNat	0.5	0.45	3.0	3.0	3.0	3.0	0.8
DTD	0.5	0.45	5.0	3.5	2.0	0.5	None
iNat + DTD	0.5	0.45	5.0	3.5	2.0	0.5	None

We elaborate on observations regarding the trends in the hyperparameter configuration that provide meaningful insights into how these values can be tuned prior to deployment. First, the two τ values appear to be closely correlated with the number of classes in the InD dataset; the optimal values of τ_{es} and τ_{cs} for ImageNet-C streams are smaller than those for CIFAR-10-C streams. τ_{es} is based on the model's predictive entropy distribution, and as the number of classes in the InD dataset grows, the model will struggle to output a highly confident prediction on one specific class. Therefore, we would want to use a smaller value for τ_{es} as the number of classes increases. τ_{cs} utilizes the distance between individual features and class-wise prototypes. With a greater number of class-wise prototypes populating the feature space, the distance between an individual feature and its nearest class-wise prototype is likely to decrease.

Second, for λ values, with the exception of streams that exclusively contain near OOD data (e.g., CIFAR-10-C+LSUN and ImageNet-C+iNaturalist), the optimal hyperparameter settings remain mostly consistent across test streams within the same InD dataset. In real deployment scenarios, it is

more likely that some mixture of OOD data types will occur. Therefore, when one cannot explicitly predict the type of OOD data that will be present in test streams in advance, the hyperparameter settings used for the rest of test streams would generally be recommended.

A8 Dataset Analysis with 2-Wasserstein Distance

In this section, we compare how close each OOD dataset is to CIFAR-10-C using 2-Wasserstein distance (Givens & Shortt, 1984) (W) as the measure of similarity. W between CIFAR-10-C and LSUN, SVHN, and DTD are 2.48, 3.62, and 3.53, respectively. We demonstrated in the main paper that LSUN, which is closest to CIFAR-10-C, functions as a benign type of OOD-contamination, while SVHN and DTD, which are relatively farther away, have detrimental effects on the adaptation performance. Therefore, this quantitative analysis further that benign OOD-contamination is induced by the closeness of OOD data to shifted InD data.