Test-Time Adaptation Induces Stronger Accuracy and Agreement-on-the-Line

Eungyeup Kim[1] **Mingjie Sun**[1] **Christina Baek**[1] **Aditi Raghunathan**[1] **J. Zico Kolter**[1,2]
[1]Carnegie Mellon University [2]Bosch Center for AI
{eungyeuk, mingjies, kbaek, raditi, zkolter}@cs.cmu.edu

Abstract

Recently, Miller et al. [32] and Baek et al. [3] empirically demonstrated strong linear correlations between in-distribution (ID) versus out-of-distribution (OOD) accuracy and agreement. These trends, coined accuracy-on-the-line (ACL) and agreement-on-the-line (AGL), enable OOD model selection and performance estimation without labeled data. However, these phenomena also break for certain shifts, such as CIFAR10-C Gaussian Noise, posing a critical bottleneck. In this paper, we make a key finding that recent test-time adaptation (TTA) methods not only improve OOD performance, but *drastically strengthen the ACL and AGL trends in models*, even in shifts where models showed very weak correlations before. To analyze this, we revisit the theoretical conditions from Miller et al. [32] that outline the types of distribution shifts needed for perfect ACL in linear models. Surprisingly, these conditions are satisfied after applying TTA to deep models in the penultimate feature embedding space. In particular, TTA causes the data distribution to collapse complex shifts into those can be expressed by a singular "scaling" variable in the feature space. Our results show that by combining TTA with AGL-based estimation methods, we can estimate the OOD performance of models with high precision for a broader set of distribution shifts. This lends us a simple system for selecting the best hyperparameters and adaptation strategy without *any* OOD labeled data. Code is available at `https://github.com/EungyeupKim/TTALine`.

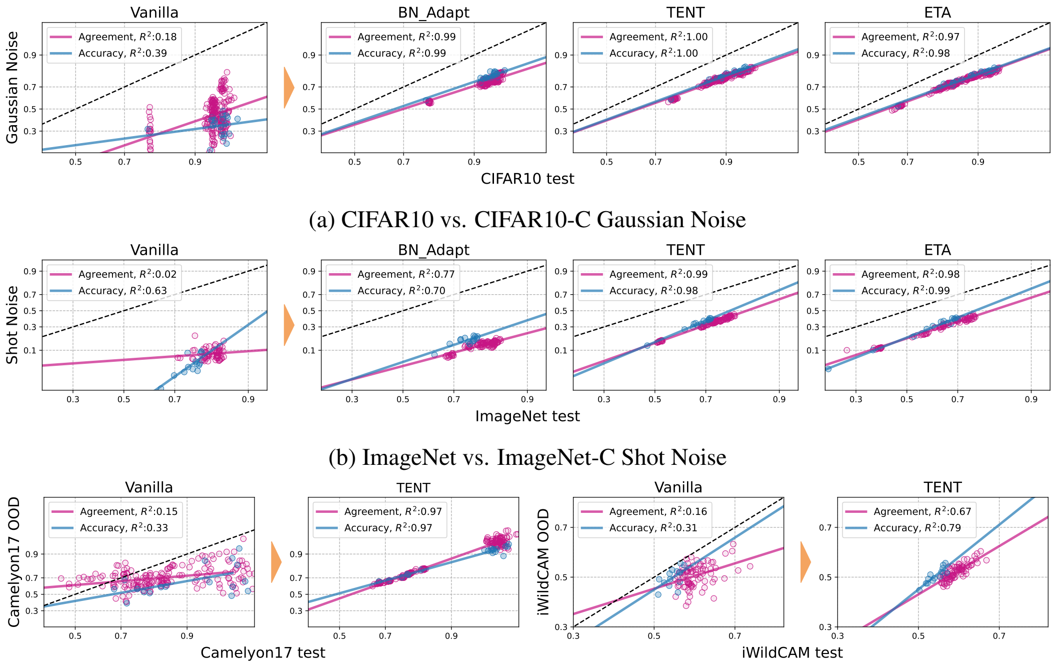
1 Introduction

Neural networks often fail to generalize to out-of-distribution (OOD) data that differs from the in-distribution (ID) data seen at train-time [1, 11, 41]. Thus, characterizing the behaviors of these models under distribution shift becomes crucial for reliable deployment. However, it is often extremely challenging to reliably estimate their performances because in many practical applications, OOD labeled data is scarce. Interestingly, recent studies [32, 3] have found a set of simple empirical laws that describe the behavior of models across many distribution shift benchmarks. In particular, across numerous distribution shift benchmarks, the models' ID versus OOD accuracies, under probit scaling, tend to observe a strong linear correlation across numerous distribution shift benchmarks. Additionally, when accuracy is strongly correlated, the ID and OOD agreement rates between pairs of these models are also strongly correlated with nearly identical slopes and biases. These phenomena, respectively referred to as "accuracy-on-the-line" (ACL) [32] and "agreement-on-the-line" (AGL) [3], can be leveraged for precise OOD accuracy estimation *without access to OOD labels*: one could estimate the slope and bias of the ID vs OOD accuracy trend using agreement rates, then linearly transform ID accuracy using this approximate linear fit[1].

[1] Throughout this paper, we use AGL as comprehensive term for indicate when models show strong linear trends in both accuracy and agreement, and the slopes and biases of these trends are identical.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3820



(a) CIFAR10 vs. CIFAR10-C Gaussian Noise

(b) ImageNet vs. ImageNet-C Shot Noise

(c) Camelyon17 vs. Camelyon17-OOD

(d) iWildCAM vs. iWildCAM-OOD

Figure 1: Linear trends in both accuracy and agreement hold to a substantially stronger degree after applying adaptation methods than before. Each blue and pink dot denotes the accuracy and agreement, followed by the linear fits for each, and R^2 is correlation coefficient.

However, studies [32, 3, 56, 48] have demonstrated that for distribution shifts benchmarks the linear trends breaks down catastrophically. As shown in Fig. 1, such as CIFAR10-C Gaussian Noise [14] or Camelyon17-WILDS [41], models without TTA (*i.e.*, Vanilla) with around $92-95\%$ ID accuracy can have OOD accuracies that vary between $10-50\%$, so ID accuracy alone becomes extremely unreliable for understanding the OOD performance of models. Similarly, the correlation strength of ID and OOD agreement rates also weakens. Ideally, we would like to intervene in models in a way that improves these linear correlations, such that we can reliably predict their performance under distribution shift. While theoretical works [32, 31, 51, 25] provide insights for *when* these linear trends hold, there is little study on how to *strengthen* these trends in models.

In this paper, we empirically demonstrate that recent OOD test-time adaptation (TTA) strategies [44, 27, 46, 53, 9, 35, 55, 36] not only improve OOD performance, but significantly *restore the strong linear trends for a broad range of distribution shifts* where ACL and AGL are not initially observable. For instance, in Fig. 1, the correlation coefficient (R^2) of ID versus OOD agreement and accuracy of Vanilla models is 0.18 and 0.39, respectively, and both of these values improve to 1.0 after applying TENT [53]. We observe such stronger linear trends after TTA throughout our extensive testbed consisting of 9 shifts, 7 adaptation methods, and over 40 network architectures. As test-time adapted models have OOD accuracies that degrade more predictably with respect to the ID accuracy, we are also able to utilize AGL based method ALine [3] to obtain precise OOD performance estimates without any OOD labels. Note that *no OOD labels were utilized throughout this procedure*, neither during TTA or ALine. Our estimates of the models' OOD performances are drastically more precise after adaptation, *e.g.*, estimation error of 11.99% in Vanilla models without TTA vs. 2.34% after applying TENT on CIFAR10-C Gaussian Noise.

Given that TTAs are designed to enhance OOD accuracy, the observation of such strong linear trends is unexpected and non-trivial. While some studies [29, 9] have explored how adaptations lead to improved OOD generalization, these efforts are orthogonal to those investigating the conditions under which linear trends occur [32, 31]. To our knowledge, no studies have attempted to bridge these two areas of research. This naturally raises the question: Why does adapting models at test-time to OOD data lead to stronger linear trends?

To answer this, we revisit the theoretical analysis in Miller et al. [32] of the sufficient conditions for observing highly correlated ID vs OOD accuracy in linear classifiers and Gaussian data. Theoretically, ACL holds exactly under distribution shifts where the direction of the class means and shape of the class covariances are fixed, and only the scale of the mean or covariance changes. Surprisingly, we find that TTA, in practice, seems to enforce exactly this condition: after applying TTA, the cosine similarity between the means and covariances of the penultimate-layer feature embedding of ID and OOD data tend to hover around 1, while their magnitude may change by some scaling constant. This implies that adaptations effectively collapse the complexity of the distribution shift to a singular "scaling" variable in the feature space. In addition to (empirically) justifying the use of TTA for strengthening ACL[2], this discovery casts some insight into the nature of TTA in general, which has previously been a largely heuristic approach.

Furthermore, models adapted with different TTA hyperparameters tend to lie on the *same* linear trend. Fig. 2 illustrates the strong correlation among models of various architectures first trained on ID data for various numbers of epochs, then adapted with different learning rates, batch sizes, adaptation steps. This critically allows us to tune the TTA hyperparameters and choice of TTA method without a held-out OOD labeled set. Notably, correctly hyperparameter tuning without access to labels is an important challenge faced in practice that currently lacks principled approaches [21, 60, 9]. Using ACL and AGL, we are able to select models with accuracy less than 1% away from that of the best model OOD, *e.g.*, in CIFAR10-C. This also allows to select the best TTA methods by comparing their estimated best OOD performances.

To summarize our contributions:

- We observe after TTA, ACL and AGL hold across a wider set of distribution shifts and hyperparameter settings.
- We explain our observation by showing TTA collapses the distribution shift to just a constant scaling of the mean and covariance matrices in the feature space. This satisfies the theoretical conditions studied previously for observing strong linear trends.
- Our findings provide a simple and effective strategy for finding the best TTA hyperparameters and the best TTA strategy without any OOD labels.

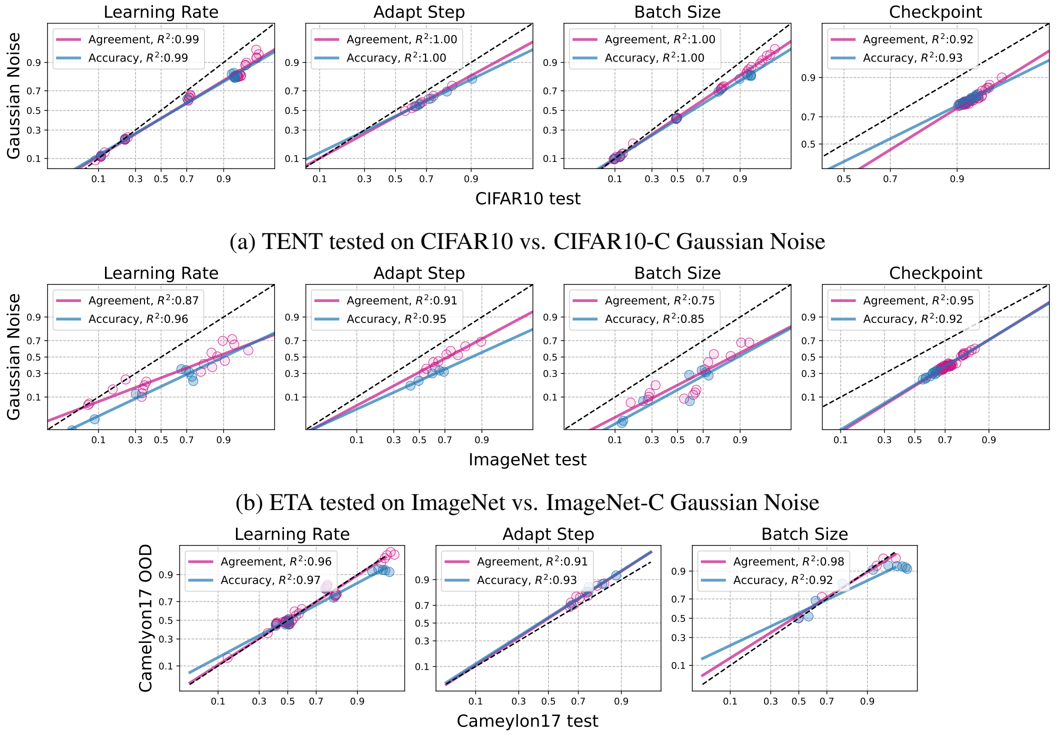
2 Related Work

Understanding accuracy and agreement-on-the-line. Miller et al. [32] and Baek et al. [3] empirically observed a coupled phenomena in deep models when evaluated on many standard distribution shift benchmarks: the ID vs. OOD accuracy and agreement are often strongly correlated and the linear fits match almost exactly. Recent studies [32, 31, 51, 25, 24] attempt to characterize what types of shifts lead to (or break) this phenomena. Miller et al. [32] prove that under a simple Gaussian data setup, ACL does not hold perfectly under distribution shifts that change the direction of the mean or transforms the covariance matrix. For example, they demonstrate that adding isotropic Gaussian noise to CIFAR10 whose covariance is anisotropic, causes the linear trend to break. Mania and Sra [31] provided sufficient conditions directly over the outputs of trained models, in terms of their prediction similarity and distributional closeness. Tripuraneni et al. [51] and LeJeune et al. [25] show that ACL holds asymptotically under certain transformations to the covariance matrix. On the other hand, there has been comparatively little theoretical analysis of AGL and why it appears together with ACL. Lee et al. [24] show that in random feature linear regression, AGL can break break partially, *i.e.*, the slope of the agreement trend matches accuracy's, but the biases may be different. While further theoretical conditions are necessary to guarantee when these phenomena hold jointly, in our empirical findings, we see that the slopes and biases do always match across the wide variety of distribution shifts we test.

Extending upon such theoretical studies, we demonstrate that a simple model intervention by test-time adaption allows these ID versus OOD trends to hold even stronger for a more expansive set of distribution shifts. In fact, we demonstrate that after adaptation, models actually satisfy the theoretical conditions necessary for ACL as described in Miller et al. [32].

[2] AGL has not been similarly shown to hold theoretically for this case of scaled means, but empirically it often seems to hold when ACL holds [3]

https://doi.org/10.52202/079017-3820



(a) TENT tested on CIFAR10 vs. CIFAR10-C Gaussian Noise

(b) ETA tested on ImageNet vs. ImageNet-C Gaussian Noise

(c) TENT tested on Camelyon17 vs. Camelyon17-OOD

Figure 2: Strong AGL by adaptations with varying hyperparameters, including learning rates, adaptation steps, batch sizes, and early-stopped checkpoints of the ID-trained model. Each blue and pink dot denotes the accuracy and agreement, followed by the linear fits for each.

Adaptations under distribution shifts and their pitfalls of reliability. Test-time adaptation aims to enhance model robustness by adapting models to unlabeled OOD test data. Test-time training methods [46, 29, 7] involve learning shift-invariant features via solving self-supervision tasks. Other approaches include computing statistics in Batch Normalization (BN) layers using OOD data [20, 44], instead of using ID statistics stored during training. Subsequent studies further adapt BN parameters by updating them using entropy minimization [53, 35, 36]. Another popular approach is self-training with pseudo-labels [40, 54, 9].

One critical challenge in TTA is that, without OOD labels, it is prohibitively difficult to evaluate how effective the adaptation methods might be. As pointed out in previous studies, adaptation methods may not succeed to address the full spectrum of distribution shifts, such as datasets reproductions [29, 60], domain generalization benchmarks [21, 61], and WILDS [40]. Furthermore, these approaches are known to be sensitive to different hyperparameter choices [4, 61, 36, 23]. Practitioners must take a great care in optimizing such hyperparameters, but their tuning procedures lack clarity. They often follow the settings of the previous studies [35, 36], or rely on some held-out labeled data [21, 60, 9] that is unavailable in practice. There exists a line of studies on unsupervised model validation [33, 42, 34, 52, 17]. Our study leverages *agreement-on-the-line phenomena within TTA* to offer a promising solution for these reliability issues.

3 Adaptations lead to stronger Agreement-on-the-Line

3.1 Experimental setup

Datasets and models. Our testbed includes diverse shifts, including common corruptions (15 failure shifts in CIFAR10-C, CIFAR100-C, and ImageNet-C [14]), dataset reproductions (CIFAR10.1 [38], ImageNetV2 [39]), and real-world shifts (ImageNet-R [16], Camelyon17-WILDS, iWildCAM-WILDS, FMoW-WILDS [41]). Among them, CIFAR10-C Gaussian Noise, Camelyon17-WILDS,

https://doi.org/10.52202/079017-3820

and iWildCAM-WILDS display the weakest correlations in both ID versus OOD accuracy and agreement [32, 3, 56]. We test over 30 different architectures of convolutional neural networks (*e.g.*, VGG [45], ResNet [13, 57, 59], DenseNet [19], MobileNet [43]) and Vision-Transformers (ViTs [5], DeiT [49], SwinT [30]).

TTA methods. We investigate 7 recent state-of-the-art TTA methods, including SHOT [27], BN_Adapt [44], TTT [46], TENT [53], ConjPL [9], ETA [35], and SAR [36]. This testbed includes different training strategies (*e.g.*, self-supervision [46], PolyLoss [26]) and updating certain layer parameters (BN layers [44, 53, 35], LayerNorm (LN) layers [2, 36], entire feature extractors [27]). We test all adaptation baselines, except SAR, on convolutional neural networks with BN layers. We apply SAR specifically to vision transformers [5, 49, 30] because it prevents the model collapse that other adaptation methods exhibit in vision transformers with LN layers [36].

Calculating agreement. Given any pair of models $(h, h') \in \mathcal{H}$ that are tested on distribution $\mathcal{D}$, the expected accuracy and agreement of the models is defined as

$$\text{Accuracy}(h) = \mathbb{E}_{x,y\sim\mathcal{D}}\big[\mathbb{1}\{h(x) = y\}\big], \quad \text{Agreement}(h, h') = \mathbb{E}_{x,y\sim\mathcal{D}}\big[\mathbb{1}\{h(x) = h'(x)\}\big] \quad (1)$$

where $h(x)$ and $h'(x)$ are the normalized logits of models h and h' given datapoint x and y is the class label. Following [32] and [3], we apply probit scaling, which is the inverse of the cumulative density function of the standard Gaussian distribution ($\Phi^{-1} : [0, 1] \to [-\infty, \infty]$), over accuracy and agreement for a better linear fit, specifically in Figs. 1 and 2.

Online test in ID and OOD during TTA. Unlike conventional offline inference at test-time, TTAs involve online learning, which dynamically updates the model's parameters while testing on OOD data. To evaluate this continuously updated model on both ID and OOD data, the data is fed into the model in minibatches and we collect the accuracy and agreement of the model over the i'th minibatch after the dynamic adaptation step i. Details are provided in Algorithm 1 in the Appendix.

3.2 Main observation

We make a striking empirical observation in the models after applying TTA: the correlation strength between ID versus OOD accuracy and agreement increases by a significant margin, especially in distribution shifts where models without TTA show wildly varying trends. In Fig. 1, we demonstrate this on the four distribution shifts that show weakest ACL and AGL trends in Vanilla models: CIFAR10-C Gaussian Noise, ImageNet-C Shot Noise, Camelon17-WILDS, and iWildCAM-WILDS. In particular, the correlation coefficients are at most $R^2 \leq 0.4$ before TTA. These failure shifts have also been noted by previous studies [32, 3, 56]. Surprisingly, after applying TTA methods, such as TENT, the strength of these linear trends as measured by R^2 increase dramatically, (*e.g.*, $0.15 \to 0.97$ and $0.33 \to 0.97$ in agreement and accuracy in Camelyon17-WILDS). Our observations hold consistently across our entire testbed of shifts and adaptation methods, as shown in Figs. 6, 7, 9, 10, 11, and 12. Moreover, in shifts where models without TTA already exhibit strong linear trends, such as ImageNet-V2 or FMoW-WILDS, these linear trends persist even after TTA, irrespective of whether TTA improves or degrades the OOD accuracy (Fig. 4).

Furthermore, models adapted using the same TTA method but with different adaptation hyperparameters lie on the same accuracy and agreement linear trend. Specifically, we vary learning rates, the number of adaption steps, batch sizes, and the number of epochs the model is initially trained on ID data. In Fig. 2 we see that on CIFAR10-C and ImageNet-C Gaussian Noise, and Camelyon17-WILDS, models adapted with different hyperparameters exhibit correlation strength R^2 close to 1 in both ID versus OOD accuracy and agreement. We show that this occurs across distribution shifts in Figs. 8 and 13.

4 Why does adaptations lead to strong linear trends?

In this section, we focus on explaining why TTA leads to restoration of these strong linear trends. Although we do not provide a complete theoretical justification, we identify a key pattern in how TTA modifies models that provides a strong clue as to why these methods substantially strengthen ACL and AGL.

We begin by revisiting the theoretically sufficient conditions for ACL from Miller et al. [32] over a simple Gaussian data and linear classifier setup. In this setting, ACL holds perfectly ($R^2 = 1$) only

https://doi.org/10.52202/079017-3820

when the distribution shifts by a simple scaling factor to the norm of the mean and covariance. On the other hand, if the actual direction of the mean or shape of the covariance matrix changes, the linearity breaks. We then demonstrate that even in CIFAR10-C shifts and deep neural networks, models after applying TTA meet these theoretical conditions better in the penultimate-layer feature space.

4.1 Theoretical conditions for linear trends in Gaussian data

We briefly introduce the theoretical conditions for ACL in Gaussian data and linear models, from Miller et al. [32], restated here for ease of presentation. Consider the binary classification over Gaussian data setup with label $y \in \{-1, 1\}$. The OOD distribution Q differs from the ID distribution P by just some scaling constants $\alpha, \gamma > 0$,

$$P(x \mid y) = \mathcal{N}(y \cdot \mu; \Sigma), \quad Q(x \mid y) = \mathcal{N}\left(y \cdot \alpha\mu; \gamma^2\Sigma\right). \tag{2}$$

Theorem 1 *[Miller et al. [32], simplified] Under the Gaussian data setup in Equation 2, across all linear classifiers $f_\theta : x \mapsto \operatorname{sign}(\theta^\top x)$, the probit-scaled accuracies over P and Q observes perfect linear correlation with a bias of zero and a slope of $\frac{\alpha}{\gamma}$.*

The proof follows immediately from the fact that the test accuracy of a linear classifier on this Gaussian data P is given by $\Phi(\theta^\top\mu/\sqrt{\theta^\top\Sigma\theta})$); applying same result to Q immediately gives the desired linear relationship. The main implication is that if the distribution shifts only in the scale of the mean and covariance, while preserving their direction, ACL is guaranteed to occur across any set of linear classifiers. Note that we focus on the conditions for ACL in particular. Additional theoretical conditions are usually required to observe AGL together with ACL [24], but they have not been well-characterized especially for this linear-Gaussian setup. Thus, we forgo analyzing assumptions for AGL, if any, in our work. Still, we find that AGL is always tightly coupled with ACL across an extensive range of benchmarks and TTA methods, similar to Baek et al. [3].

4.2 Empirical analysis under adaptation to CIFAR10-C

We now investigate how well the above conditions are met before and after TTA for real-world data with deep models. Here, evaluate models trained on CIFAR10 on CIFAR10-C Gaussian Noise. We adapt models using BN_Adapt [44] or TENT [53] at test-time. To match the theoretical setup, we analyze the class-wise feature embeddings from the penultimate-layer of these models that directly precedes the final *linear* classifier. Namely, we can measure the "alignment" or the cosine similarity between the ID and OOD class means and covariances of these embeddings. Similar to our CIFAR10-C Gaussian Noise results in Figs. 1 and 2, we evaluate the embedding alignment of models in each hyperparameter setup: we adapt models at test-time varying a single hyperparameter, e.g., architectures, learning rates, batch sizes, and early-stopped checkpoints, while keeping the remaining hyperparameters fixed.

Shifts become aligned in their mean and covariance after adaptation. Table 1 shows the mean and standard deviation of the cosine similarity measured across different setups. We first notice that, without adaptation (Vanilla), both mean and covariance of CIFAR10 and CIFAR10-C Gaussian Noise features are misaligned with cosine similarity much smaller than 1. Surprisingly, after adaptations, the similarity substantially increases and becomes very close to 1, across every architecture we test (with standard deviation close to 0). In other words, in the feature embedding space, the two distributions have means and covariances that (roughly) point in the same *directions* after adaptation. This also implies that in adapted models, the distribution shift is represented as a simple change in the *scale* of the features alone.

Consider CIFAR10-C Gaussian Noise which corrupts CIFAR10 images by additive isotropic Gaussian Noise. In the input image space, the direction of the covariance matrix clearly changes with the addition of this noise. Yet TTA has the non-trivial ability of being able to "simplify" complex drifts to scale shifts within the features. These observations support the theoretical conditions in Theorem 1 and gives us loose insight into why TTA strengthens ACL and AGL in models in practice.

Theoretical slope matches the empirical slope. Finally, under scale shifts in our theoretical setup, the slope of the ACL trend is exactly $\Delta = \alpha/\gamma$. Assuming the direction of the mean and covariance remains fixed, we may estimate α and γ as simply the average change in magnitude of the class means and covariances, and use this to approximate Δ. In Table 1 we compare how this theoretical

https://doi.org/10.52202/079017-3820

Setup	Cosine Similarity		Slope	
	Mean	Covariance	Theoretical	Empirical
Vanilla (Archs.)	0.691 ± 0.175	0.750 ± 0.109	–	–
BN_Adapt (Archs.)	0.988 ± 0.007	0.972 ± 0.011	0.751 ± 0.075	0.758
TENT (Archs.)	0.990 ± 0.005	0.974 ± 0.011	0.753 ± 0.072	0.778
Learning rates	0.993 ± 0.003	0.977 ± 0.006	0.759 ± 0.041	0.76
Batch Sizes	0.995 ± 0.003	0.982 ± 0.010	0.831 ± 0.101	0.809
Check Points	0.992 ±0.003	0.976 ± 0.008	0.782 ± 0.033	0.838

Table 1: Cosine similarity between mean direction and covariance shape of class-wise penultimate-layer features, followed by the comparison between theoretical and empirical slope. They are evaluated on CIFAR10 vs. CIFAR10-C Gaussian Noise, measured across architectures and hyperparameters. We report their means and standard deviations.

slope compares to the true slope of ID versus OOD accuracy fit by linear regression. These slopes tend to closely match each other, highlighting that the theoretical results of TTA seem to provide correct intuition about why ACL holds in practice, especially after TTA.

TTA improves linear trends irrespective of whether it improves OOD generalization. We make the subtle distiction between TTA's ability to strengthen ACL and AGL phenomena from its original purpose of improving OOD accuracy. One could imagine a trivial reason why ACL happens is because the ID and OOD gap diminishes after TTA, thereby bringing the ID versus OOD accuracy of all models close to the $y = x$ line. However, this explanation does not reflect our empirical findings. While TTA seems to reduce the complexity of the shift to scale shifts, the magnitude of this scaling factor may remain very large, *i.e.*, $\alpha \ll 1$, $\gamma \gg 1$. This results in a near-perfect linear trend that, nonetheless, lies arbitrarily far from the $y = x$. Consider the shift ImageNet-C Shot Noise in Fig. 1. After adaptation, there is still approximately a 40% gap in ID and OOD accuracy, but ACL holds strongly. On the other hand, consider the counterfactual TTA method which improves OOD performance by decreasing the covariance scale γ in the feature embedding space, but the means/covariances remain misaligned. Then the linear trend moves closer to $y = x$, but there is no guarantee about the strength of the linear trend.

5 Experiments

Our finding that TTA induces strong ACL and AGL has two key practical applications: (i) estimation of OOD accuracy, and (ii) unsupervised validation for TTA — all without access to any OOD labels.

5.1 OOD Accuracy estimation after adaptation

Experimental Setup. We employ AGL-based estimation method, namely ALine [3]. This method first estimates the slope and bias of the linear trend in ID versus OOD accuracy, via agreements (which requires no labels), and uses them to linearly transform ID accuracy to get an estimate of OOD accuracy. The full details of ALine are illustrated in Algorithm 2 in Appendix. We apply ALine to models after TTA, and compare these results with (i) those of ALine applied to models before adaptation, and (ii) estimates using existing estimation baselines. These baselines include average thresholded confidence (ATC) [8], difference of confidence (DOC)-feat [10], average confidence [15], and agreement [22]. We evaluate baselines on CIFAR10-C, CIFAR100-C, ImageNet-C [14], and Camelyon17-WILDS and iWildCAM-WILDS [41].

Results. ALine shows accurate estimation under shifts where models demonstrate strong AGL (*e.g.*, in CIFAR10-C snow, mean absolute error (MAE) of estimation is 0.93% in ALine-D), but critically fails when shifts do not have such linear trends (*e.g.*, in CIFAR10-C Gaussian Noise, MAE is 10.76% and in Camelyon17-WILDS, MAE is 12.88%). As a result, as shown in Table 2, estimates using ALine-S/D on Vanilla models are sometimes error-prone. Next, after applying adaptation methods, such as SHOT, BN_Adapt, and TENT, the estimation performance of ALine-S/D improves, showing substantially lower MAE compared to that of Vanilla models (*e.g.*, MAE decreases $10.76\% \rightarrow 2.34\%$ in CIFAR10-C Gaussian Noise and $12.88\% \rightarrow 1.42\%$ in Camelyon17-WILDS). ALine on top of TTA also outperforms other baseline methods on most of the distribution shifts tested.

Dataset	Method	Error	ATC	DOC-feat	AC	Agreement	ALine-S	ALine-D
CIFAR10-C	Vanilla	31.38	8.31	15.03	17.42	5.45	6.02	5.87
	SHOT	15.40	1.63	4.63	7.63	1.78	0.96	**0.77**
	BN_Adapt	16.87	3.69	4.79	7.53	1.93	1.12	**0.91**
	TENT	15.43	4.25	4.65	7.66	1.79	0.97	**0.77**
	ConjPL	16.62	1.80	6.16	11.46	2.02	1.18	**1.01**
	ETA	15.14	4.58	4.50	7.68	1.76	0.92	**0.72**
CIFAR100-C	Vanilla	59.04	5.05	12.82	18.34	6.96	7.49	7.22
	SHOT	40.79	2.21	5.44	14.36	2.52	1.64	**0.90**
	BN_Adapt	42.69	2.89	4.42	11.81	2.33	1.43	**1.13**
	TENT	41.11	6.60	5.59	14.85	2.65	1.64	**0.88**
	ConjPL	42.79	**1.09**	6.55	23.73	2.40	1.67	1.18
	ETA	44.27	7.15	4.92	16.49	4.96	1.44	**0.81**
ImageNet-C	Vanilla	80.41	3.95	13.72	17.34	9.06	6.00	5.95
	BN_Adapt	69.05	7.37	**2.63**	2.86	3.91	6.16	6.09
	TENT	56.58	5.98	6.54	12.70	7.48	4.62	**4.57**
	ETA	56.56	10.21	7.91	34.38	8.02	**3.66**	3.72
	SAR	43.30	5.39	8.61	13.68	5.51	5.19	**4.17**
Camelyon17 -WILDS	Vanilla	34.07	14.91	17.31	21.69	11.95	12.88	13.46
	TENT	14.37	3.00	3.43	6.94	6.49	2.29	**2.27**
	ETA	16.43	3.05	4.38	6.85	5.33	2.24	**1.42**
iWildCAM -WILDS	Vanilla	50.27	7.12	2.73	23.86	3.00	3.53	2.82
	TENT	47.39	5.44	3.20	28.03	3.55	**2.59**	2.96
	ETA	46.49	6.61	3.40	29.34	4.62	**2.14**	2.82

Table 2: The results of OOD accuracy estimation, measured by MAE (%) between estimated and actual OOD accuracy. The gray shades denote the results calculated after applying adaptations, and bold texts indicate the smallest estimation error among estimators. We also report the classification error (%) in both Vanilla and adaptation methods for each dataset.

5.2 Unsupervised validation for TTA

In this section, we demonstrate that strong AGL allows us to validate the best hyperparameter for TTA without any labels. Furthermore, we can use the OOD accuracy estimates to choose the best TTA strategy *overall*.

Experimental setup. First, as we have shown, models adapted using a particular method with varying hyperparameters tend to show strong ACL behavior. When ACL holds, we can simply select the best OOD model by selecting the one with best ID accuracy [32]. Thus, we first collect candidates by systematically sweeping over hyperparameter values, and select the model with the best ID accuracy. We focus on TENT, optimizing its various hyperparameters including learning rates, adaptation steps, architectures, batch sizes, and early-stopped checkpoints of training. See our total pool of models in Appendix Table 6. We compare this naive strategy to existing unsupervised validation methods including MixVal [17], ENT [33], IM [34], Corr-C [52], and SND [42]. We evaluate them on four shifts, CIFAR10-C, ImageNet-C, ImageNet-R, and Camelyon17-WILDS.

Results. Table 3 reports the absolute difference (MAE (%)) between the OOD accuracy of the selected model and the best OOD accuracy, within a set of models where we vary a particular TTA hyperparameter. We find that using ID accuracy for model selection consistently achieves competitive unsupervised validation results, often outperforming other baselines. We noticed that current state-of-the-art unsupervised model selection methods, *i.e.*, MixVal or IM, perform well on CIFAR10-C, ImageNet-C, and ImageNet-R, but they critically fail in Camelyon17-WILDS, *e.g.*, validation error of 7.98% in MixVal and 23.52% in IM. Such failures of existing baselines might come from their assumptions, *e.g.*, low-density separation, that do not generalize to such distribution shifts. In contrast, our method shows low MAE across all distribution shifts, including those where other baselines fail, *e.g.*, 0.62% in Camelyon17-WILDS. We note that there are cases, specifically models adapted with various learning rates for ImageNet-C, where MAE is large (9.70%). We see that models optimized with very small learning rates, *e.g.*, 10^{-5}, deviate from the linear trend. Results on other TTA baselines' unsupervised validation are in Table 7.

Extension to selecting best TTA strategy. After hyperparameter tuning for each TTA strategy using the unsupervised strategy from the previous section, we could further leverage ALine to then select for the best TTA method overall. Different TTA methods exhibit different slopes and biases in their ACL

HyperParameter	CIFAR10-C						ImageNet-C					
	MixVal	ENT	IM	Corr-C	SND	Ours	MixVal	ENT	IM	Corr-C	SND	Ours
Architecture	2.31	1.06	1.06	21.71	2.77	0.03	6.22	0.96	0.47	26.32	20.60	0.75
Learning Rate	6.97	8.88	2.24	11.56	1.87	0.72	12.75	20.49	1.49	20.18	12.61	9.70
Checkpoints	3.21	0.0	0.0	5.53	3.46	0.05	–	–	–	–	–	–
Batch Size	7.85	3.32	0.96	32.37	5.68	0.77	14.29	42.31	0.99	42.31	42.31	5.61
Adapt Step	0.85	0.0	0.0	1.02	0.0	0.23	1.85	1.94	1.25	3.09	2.17	0.30
Average	4.23	2.65	0.85	14.43	2.75	**0.36**	8.77	16.42	**1.05**	14.43	22.97	4.0

HyperParameter	ImageNet-R						Camelyon17-WILDS					
	MixVal	ENT	IM	Corr-C	SND	Ours	MixVal	ENT	IM	Corr-C	SND	Ours
Architecture	1.75	0.62	0.62	22.17	22.17	0.85	28.87	1.03	1.03	28.87	28.87	0.85
Learning Rate	3.12	10.16	4.73	19.16	19.16	2.8	0.91	48.37	46.41	48.37	48.37	1.14
Batch Size	1.83	35.88	0.08	35.88	35.88	1.74	0.0	46.67	46.67	40.45	40.45	1.37
Adapt Step	1.07	1.07	1.07	1.07	1.07	0.0	2.17	33.12	0.0	33.12	33.12	0.0
Average	1.94	14.18	1.62	19.57	19.57	**1.34**	7.98	32.29	23.52	37.70	37.70	**0.62**

Table 3: Results of unsupervised validation, measured by MAE (%) between OOD accuracy of model selected by best ID model and actual best OOD model. Our results are highlighted in shade.

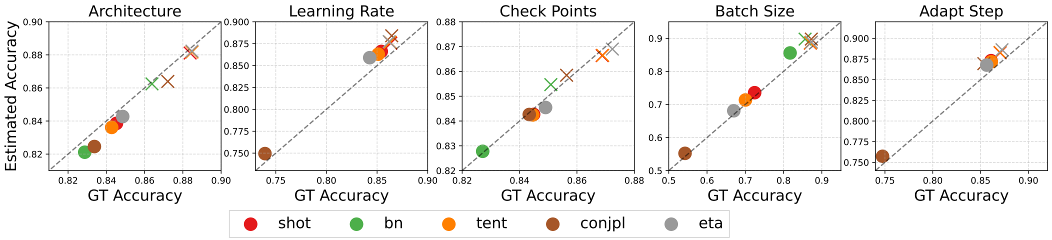


Figure 3: 2-D visualizations of each TTA baseline's ground truth best OOD accuracy (x-axis) and estimated OOD accuracy of best-ID models (y-axis) marked in cross (×). Each color denotes different TTA baselines. Circle dot (○) represents estimates and ground truth accuracies averaged over all hyperparameter values.

trends, so we use AGL-based estimators to estimate the best OOD accuracy across hyperparameters for each method separately, then compare these estimates to choose the best method. Specifically, we select between five TTA baselines — BN_Adapt, TENT, SHOT, ConjPL, and ETA — on CIFAR10-C across all corruptions. For ease of presentation, we consider selecting the best TTA method after tuning a single hyperparameter for each method, with other parameters fixed. In Fig. 3, we plot using "×" the selected models' estimated OOD accuracy for each method versus each method's actual best OOD performances. Our method almost always selects the best OOD model for each method but also accurately estimates their performances, as shown by points lying close to the $y = x$ line. Furthermore, the ranking of TTA methods by OOD accuracy stays the same when using the estimates. We also plot in "○" the average ground truth and estimated accuracy across hyperparameters. This demonstrates that our approach effectively selects the best TTA strategy without OOD labels.

6 Ablation Study: When does TTA not improve linear trends?

So far, we focused on numerous TTA methods that tend to dramatically improve the strength of ACL and AGL phenomena. We then connected this behavior to models representing the distribution shift in its feature space as a scale shift, satisfying the theoretical conditions of ACL from Miller et al. [32]. This behavior was especially prominent in methods that adjust the batch normalization layer statistics, *e.g.* BN_Adapt, or modify the feature extractor, *e.g.* TENT. This section presents an ablation study on the impact of normalization layers (*e.g.*, BN, LN) and the components adapted by TTA (*e.g.*, feature extractor, final classifier) for observing ACL and AGL.

Normalization layers.

As shown in our results for BN_Adapt in Section 4.1, the strength of the ACL and AGL trends tends to be sensitive to the statistics stored in the batch normalization layers. Taking the same set of models from Section 4.1, we test two architectural variants — models with no normalization layers (denoted as Vanilla w/o N) and layer normalization (Vanilla w/ LN) — and report the cosine similarity of the class means and covariance matrices of the ID and OOD feature embeddings.

We also report the correlation coefficient (R^2) of the ACL and AGL trends in Table 4. Interestingly, the mean and covariance alignment as well as correlation strength tends to be larger than that of Vanilla with BN-layers, but still falls short of models with BN-layers adapted with BN_Adapt which is very close to 1.

	Cosine Similarity		Correlations	
Setup	Mean	Covariance	Acc.	Agr.
Vanilla w/o N	0.794 ± 0.117	0.778 ± 0.172	0.29	0.71
Vanilla w/ LN	0.935 ± 0.042	0.854 ± 0.063	0.56	0.86
Vanilla w/ BN	0.691 ± 0.175	0.750 ± 0.109	0.18	0.39
BN_Adapt	0.998 ± 0.007	0.972 ± 0.011	0.99	0.99

Table 4: Cosine similarity of mean direction / covariance shape and correlation coefficients (R^2) in CIFAR10-C Gaussian Noise, measured across architectures. Last two rows are from Table 1.

Methods such as SAR which substitutes BN-layers with LN as a part of their adaptation strategy, also shows stronger linear trends in Fig. 12 than Vanilla counterparts in Fig. 11, but not as consistently as those of BN-layer-adaptation methods in Fig. 11. We conjecture that this stems from how different architectures encode distribution shifts: BN-layers result in severe covariate shift, but most of the shift is encoded in the BN statistics, making it easy to align shifts via adaptation. In contrast, non-BN models encode shifts across the entire model parameters, which may dampen the covariate shift in the representations, but also make it harder to remove by adaptation. Schneider et al. [44] suggests a similar conclusion.

Featurizer or classifier. In this section, we consider the TTA method T3A [21], which leverages a feature extractor without BN layers and only adapts the last linear classifier. This is in contrast to the other TTA methods we study in our work, which largely modifies the feature extractor. We adapt the same set of models in Section 4.1 and observe that the R^2 for accuracy and agreement in T3A are 0.80 and 0.46, which is substantially lower than models adapted using BN_Adapt. This may be because the features remain misaligned after T3A. Overall, our ablation studies clarify what TTA designs ideally lead to strong linear trends. Since non-BN-TTAs are often adopted in numerous practical circumstances, *e.g.*, single batch, it remains an important open question how to develop alternate TTA strategies that achieves strong ACL and AGL trends in such settings, thus improving the reliability of models.

7 Conclusion and Limitations

In this paper, we provide a key observation that recent TTAs lead to stronger AGL across a wide range of distribution shifts, encompassing those with weak correlations before adaptation. We explain this phenomena by the complexity of distribution shifts being substantially reduced to those where the direction/shape of the mean and covariance remains identical, satisfying the theoretical conditions for linear trends. This naturally leads to enhanced estimation of OOD performances across a wider range of shifts than before TTA. We can leverage this to perform unsupervised hyperparameter validation and select for the best TTA method without any labels.

While the method we propose does not require OOD labels, AGL still requires sufficient ID data to accurately estimate agreement rate. As a result, our strategy may not be complementary with TTA methods that do not require any ID data. Moreover, it could make our strategy computationally expensive. In Section D, we conduct an ablation study to minimize the required amount of ID data for observing AGL for accuracy estimation. The results show that even with only 5% of the original ID data, OOD accuracy estimation performance remains nearly the same as with full access, outperforming other estimators. We believe that overcoming dependency on ID data and exploring a fully test-time approach for observing AGL remains a promising direction for future research.

Acknowledgement

We acknowledge the anonymous reviewers for their valuable feedbacks. Eungyeup Kim, Mingjie Sun, and Christina Baek are supported by funding from the Bosch Center for Artificial Intelligence. Aditi Raghunathan gratefully acknowledges support from Open Philanthropy, Google, Apple and Schmidt AI2050 Early Career Fellowship.

References

[1] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization, 2020.

[2] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] C. Baek, Y. Jiang, A. Raghunathan, and J. Z. Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *Advances in Neural Information Processing Systems*, volume 35, pages 19274–19289, 2022.

[4] M. Boudiaf, R. Mueller, I. Ben Ayed, and L. Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8344–8353, June 2022.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[6] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[7] Y. Gandelsaman, Y. Sun, X. Chen, and A. A. Efros. Test-time training with masked autoencoders. In *Advances in Neural Information Processing Systems*, 2022.

[8] S. Garg, S. Balakrishnan, Z. C. Lipton, B. Neyshabur, and H. Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[9] S. Goyal, M. Sun, A. Raghunanthan, and Z. Kolter. Test-time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems*, 2022.

[10] D. Guillory, V. Shankar, S. Ebrahimi, T. Darrell, and L. Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1134–1144, October 2021.

[11] I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 2016. URL `https://api.semanticscholar.org/CorpusID:6447277`.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778. IEEE, June 2016. doi: 10.1109/CVPR.2016.90.

[14] D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

[15] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

[16] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

[17] D. Hu, J. Liang, J. H. Liew, C. Xue, S. Bai, and X. Wang. Mixed samples as probes for unsupervised model selection in domain adaptation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[18] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

https://doi.org/10.52202/079017-3820

[19] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[20] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR.

[21] Y. Iwasawa and Y. Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, pages 2427–2440. Curran Associates, Inc., 2021.

[22] Y. Jiang, V. Nagarajan, C. Baek, and J. Z. Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*, 2022.

[23] A. Khurana, S. Paul, P. Rai, S. Biswas, and G. Aggarwal. Sita: Single image test-time adaptation, 2022.

[24] D. Lee, B. Moniri, X. Huang, E. Dobriban, and H. Hassani. Demystifying disagreement-on-the-line in high dimensions. In *International Conference on Machine Learning (ICML)*, 2023.

[25] D. LeJeune, J. Liu, and R. Heckel. Monotonic risk relationships under distribution shifts for regularized risk minimization. *Journal of Machine Learning Research*, 25(54):1–37, 2024. URL `http://jmlr.org/papers/v25/22-1197.html`.

[26] Z. Leng, M. Tan, C. Liu, E. D. Cubuk, J. Shi, S. Cheng, and D. Anguelov. Polyloss: A polynomial expansion perspective of classification loss functions. In *International Conference on Learning Representations*, 2022.

[27] J. Liang, D. Hu, and J. Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning (ICML)*, pages 6028–6039, 2020.

[28] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L. Li, L. Fei-Fei, A. L. Yuille, J. Huang, and K. Murphy. Progressive neural architecture search. In *European Conference on Computer Vision*, 2018.

[29] Y. Liu, P. Kothari, B. G. van Delft, B. Bellot-Gurlet, T. Mordan, and A. Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[31] H. Mania and S. Sra. Why do classifier accuracies show linear trends under distribution shift?, 2021.

[32] J. P. Miller, R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon, and L. Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7721–7735. PMLR, 18–24 Jul 2021.

[33] P. Morerio, J. Cavazza, and V. Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJWechg0Z`.

[34] K. Musgrave, S. Belongie, and S.-N. Lim. Three new validators and a large-scale benchmark ranking for unsupervised domain adaptation, 2023.

https://doi.org/10.52202/079017-3820

[35] S. Niu, J. Wu, Y. Zhang, Y. Chen, S. Zheng, P. Zhao, and M. Tan. Efficient test-time model adaptation without forgetting. In *The Internetional Conference on Machine Learning*, 2022.

[36] S. Niu, J. Wu, Y. Zhang, Z. Wen, Y. Chen, P. Zhao, and M. Tan. Towards stable test-time adaptation in dynamic wild world. In *Internetional Conference on Learning Representations*, 2023.

[37] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollar. Designing network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[38] B. Recht, R. Roelofs, L. Schimidt, and V. Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv: 1806.00451*, 2018.

[39] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 09–15 Jun 2019.

[40] E. Rusak, S. Schneider, G. Pachitariu, L. Eck, P. Gehler, O. Bringmann, W. Brendel, and M. Bethge. If your data distribution shifts, use self-learning. *Transactions of Machine Learning Research*, 2022.

[41] S. Sagawa, P. W. Koh, T. Lee, I. Gao, S. M. Xie, K. Shen, A. Kumar, W. Hu, M. Yasunaga, H. Marklund, S. Beery, E. David, I. Stavness, W. Guo, J. Leskovec, K. Saenko, T. Hashimoto, S. Levine, C. Finn, and P. Liang. Extending the wilds benchmark for unsupervised adaptation. In *International Conference on Learning Representations (ICLR)*, 2022.

[42] K. Saito, D. Kim, P. Teterwak, S. Sclaroff, T. Darrell, and K. Saenko. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9184–9193, October 2021.

[43] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[44] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, pages 11539–11551. Curran Associates, Inc., 2020.

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[46] Y. Sun, X. Wang, Z. Liu, J. Miller, A. A. Efros, and M. Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, 2020.

[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions, 2014.

[48] D. Teney, Y. Lin, S. J. Oh, and E. Abbasnejad. Id and ood performance are sometimes inversely correlated on real-world datasets, 2023.

[49] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

[50] H. Touvron, M. Cord, and H. J'egou. Deit iii: Revenge of the vit. In *European Conference on Computer Vision*, 2022. URL `https://api.semanticscholar.org/CorpusID:248178188`.

[51] N. Tripuraneni, B. Adlam, and J. Pennington. Overparameterization improves robustness to covariate shift in high dimensions. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=PxMfDdPnTfV`.

[52] W. Tu, W. Deng, T. Gedeon, and L. Zheng. Assessing model out-of-distribution generalization with softmax prediction probability baselines and a correlation method, 2023. URL `https://openreview.net/forum?id=1maXoEyeqx`.

[53] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

[54] J.-K. Wang and A. Wibisono. Towards understanding gd with hard and conjugate pseudo-labels for test-time adaptation. In *Internetional Conference on Learning Representations*, 2023.

[55] Q. Wang, O. Fink, L. Van Gool, and D. Dai. Continual test-time domain adaptation. In *Proceedings of Conference on Computer Vision and Pattern Recognition*, 2022.

[56] F. Wenzel, A. Dittadi, P. V. Gehler, C.-J. Simon-Gabriel, M. Horn, D. Zietlow, D. Kernert, C. Russell, T. Brox, B. Schiele, B. Schölkopf, and F. Locatello. Assaying out-of-distribution generalization in transfer learning. In *Neural Information Processing Systems*, 2022.

[57] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

[58] F. Yu, D. Wang, E. Shelhamer, and T. Darrell. Deep layer aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[59] S. Zagoruyko and N. Komodakis. Wide residual networks, 2017.

[60] M. Zhang, S. Levine, and C. Finn. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, 2022.

[61] H. Zhao, Y. Liu, A. Alahi, and T. Lin. On pitfalls of test-time adaptation. In *International Conference on Machine Learning (ICML)*, 2023.

https://doi.org/10.52202/079017-3820

A Details on Experimental Setup

A.1 Model Architectures

We list the types of architectures for each testbed of dataset.

CIFAR10, 100 Architectures
ResNet18 [13]
ResNet26 [13]
ResNet34 [13]
ResNet50 [13]
ResNet101 [13]
WideResNet28 [59]
PreActResNet18 [12]
PreActResNet34 [12]
PreActResNet50 [12]
PreActResNet101 [12]
RegNet X200 [37]
RegNet X400 [37]
RegNet Y400 [37]
VGG11 [45]
VGG13 [45]
VGG16 [45]
VGG19 [45]
MobileNetV2 [43]
PNASNet-A [28]
PNASNet-B [28]
SENet18 [18]
GoogLeNet [47]

ImageNet Architectures
ResNet18 [13]
ResNet34 [13]
ResNet50 [13]
ResNet101 [13]
ResNet152 [13]
ResNeXT32×4d [57]
DenseNet121 [19]
WideResNet50 [59]
WideResNet101 [59]
DLA34 [58]
DLA46C [58]
DLA60 [58]
DLA102 [58]
DLA169 [58]
ViT-B/16 [5]
ViT-L/16 [5]
ViT-B/32 [5]
SwinT-S [30]
SwinT-B [30]
SwinT-L [30]
DeiT-S/16 [49]
DeiT-B/16 [49]
DeiT3-B/16 [50]
DeiT3-L/16 [50]
DeiT3-H/14 [50]

WILDS Architectures
ResNet18 [13]
ResNet34 [13]
ResNet50 [13]
ResNet101 [13]
ResNet152 [13]
ResNeXT32×4d [57]
DenseNet121 [19]
WideResNet50 [59]
WideResNet101 [59]
VGG11 [45]
VGG13 [45]
VGG16 [45]
VGG19 [45]

Table 5: The list of architecture types for each testbed of datasets, including CIFAR10, CIFAR100, ImageNet, Camelyon17-WILDS, and iWildCAM-WILDS.

For CIFAR10, CIFAR100, and WILDS datasets, we train the models from the scratch, while for ImageNet, we use the pretrained model weights from torchvision and timm package. In addition, since TTT requires specific network composition required for the rotation-prediction task during pretraining, we train them using ResNet-14,26,32,50,104, and 152, which are available in the original implementation[3].

For Camelyon17 dataset, we also test the neural networks with their feature extractor (all parameters before the final linear classifier) randomly initialized, following Miller et al. [32] that tested the low-accuracy models on the dataset. We notice that these models, even with their most of parameters being randomized, still achieve high accuracy in both ID and OOD after training the last linear classifier. They also exhibit improved OOD accuracy after adaptations. We apply this randomization to ResNet18, ResNet34, ResNet50, VGG16, VGG13, and VGG11.

We trained and tested all models and datasets in NVIDIA RTX 6000.

A.2 Distribution Shifts

We test the models on 9 different distribution shifts that include synthetic corruptions and real-world shifts. Here are details of tested benchmarks:

[3] https://github.com/yueatsprograms/ttt_cifar_release

Synthetic corruptions. CIFAR10-C, CIFAR100-C, and ImageNet-C [14] are designed to apply the 15 different types of corruptions, such as Gaussian Noise, on their original dataset counterparts. We use the most severe corruptions, which have severity of 5, in all experiments. These corruptions datasets are most commonly evaluated distribution shifts in a wide range of TTA papers [44, 46, 29, 27, 53, 9, 35, 36, 61].

Dataset reproduction. CIFAR10.1 and ImageNetV2 are reproduced version of their original counterparts, CIFAR10 and ImageNet, respectively. They are collected by carefully following the dataset creation procedure of the original dataset, yet resulting in non-trivial performance degradations to models trained on original counterparts.

Style shifts. ImageNet-R is one variant of ImageNet which contains the images with renditions of various styles. Such styles include paintings, cartoons, sketches and others, which induces significant distribution shifts even when the semantics are invariant.

Real-world shifts. Camelyon17-WILDS [41] contains the medical images of tissue collected from different hospitals, where distribution shifts are designed by different hospitals the model is trained and tested. The task is to predict whether it has tumor tissue or not (binary classification). We followed the same evaluation protocol of Miller et al. [32], collecting 30 whole-slide images (WSI) from 3 hospitals as ID (total 302,436 for train and 33,560 for test), while 10 WSI from a different hospital as OOD (total 85,054). FMoW-WILDS [41] contains the spatio-temporal satellite imagery of 62 different use of land or building categories, where distribution shifts originate from the years that the imagery is taken. Specifically, following Miller et al. [32], we use ID set consists of images taken from 2002 to 2013 (training data of 76,863 and test of 11,327), and OOD set taken between 2013 and 2016 (test data of 22,108). iWildCAM-WILDS [41] contains the images of animals of 182 total species, which are taken with different camera traps. Following Miller et al. [32], we use 129,809 train images and 8,154 test images taken by the same 243 camera traps as ID. OOD data consists of total 42,791 images, which are taken with different camera traps, which involve different camera angle, background, and others.

A.3 Adaptation hyperparameters

Table 6 shows the hyperparameter pools for each setup when used for observing AGL across different hyperparameters. We use SGD optimizer with momentum of 0.9 for all adaptation baselines except for SAR, which uses sharpness-aware minimization (SAM) optimizer [6].

Setup	**CIFAR10, CIFAR100**	**ImageNet**
Learning Rates	$1 \cdot 10^{-5}, 5 \cdot 10^{-5}, 1 \cdot 10^{-4}, 5 \cdot 10^{-4}, 1 \cdot 10^{-3},$ $5 \cdot 10^{-3}, 1 \cdot 10^{-2}, 5 \cdot 10^{-2}, 1 \cdot 10^{-1}, 5 \cdot 10^{-1}$	$1 \cdot 10^{-5}, 2 \cdot 10^{-5}, 5 \cdot 10^{-5}, 1 \cdot 10^{-4}, 2 \cdot 10^{-4},$ $5 \cdot 10^{-4}, 1 \cdot 10^{-3}, 2 \cdot 10^{-3}, 5 \cdot 10^{-3}$
Batch Sizes	$1, 2, 4, 8, 16, 32, 64, 128, 256$	$4, 8, 16, 32, 64, 128, 256, 512$
Early-stopped Checkpoints	$100, 110, 120, 130, 140, 150, 160, 170, 180, 190$ (out of 200 total epochs)	–
Adaptation Steps	$1, 2, 3, 4, 5$	$1, 2, 3, 4, 5$

(a) CIFAR10, CIFAR100, ImageNet

Setup	**Camelyon17-WILDS**	**iWildCAM-WILDS**
Learning Rates	$2 \cdot 10^{-5}, 5 \cdot 10^{-5}, 1 \cdot 10^{-4}, 2 \cdot 10^{-4}, 5 \cdot 10^{-4}, 1 \cdot 10^{-3},$ $2 \cdot 10^{-3}, 5 \cdot 10^{-3}, 1 \cdot 10^{-2}, 2 \cdot 10^{-2}, 5 \cdot 10^{-2}$	$2 \cdot 10^{-5}, 5 \cdot 10^{-5}, 1 \cdot 10^{-4}, 2 \cdot 10^{-4},$ $5 \cdot 10^{-4}, 1 \cdot 10^{-3}, 2 \cdot 10^{-3}, 5 \cdot 10^{-3}$
Batch Sizes	$4, 8, 16, 32, 64, 128, 256, 512$	$1, 2, 4, 8, 16, 32, 64, 128$
Early-stopped Checkpoints	–	–
Adaptation Steps	$1, 2, 3, 4, 5$	$1, 2, 3, 4, 5$

(b) Camelyon17-WILDS, iWildCAM-WILDS

Table 6: The hyperparameter pools utilized for observing AGL across hyperparameters in CIFAR10, CIFAR100, ImageNet, Camelyon17-WILDS, and iWildCAM-WILDS dataset.

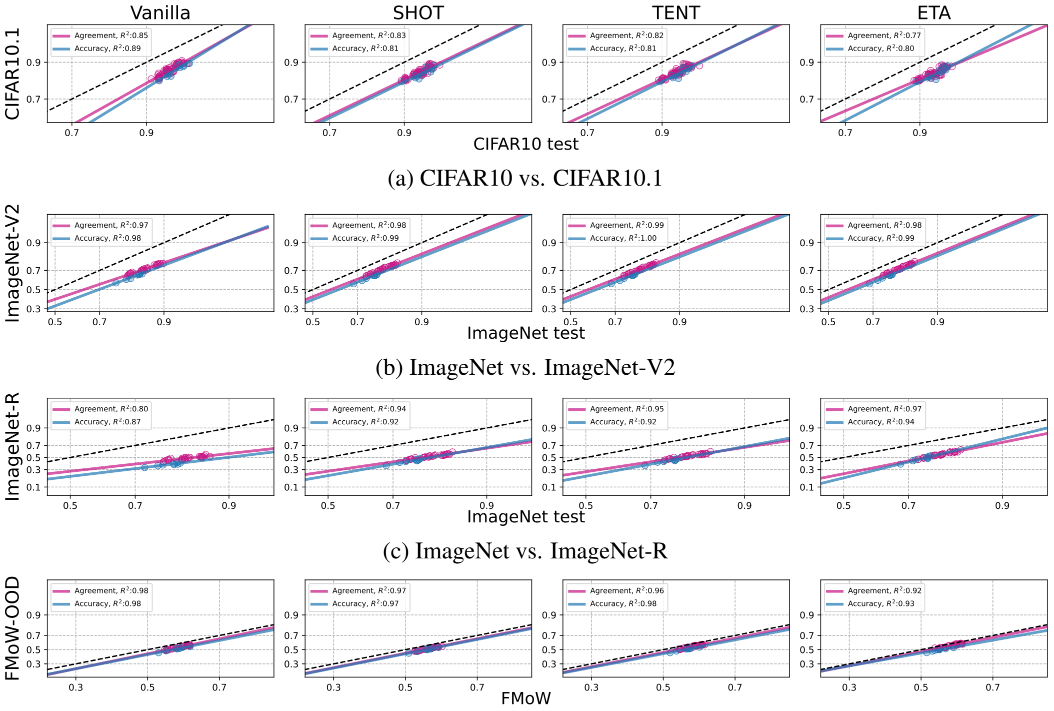
A.4 Online test in ID and OOD during TTA

Algorithm 1 provides the details of how ID and OOD performances are tested during TTA. TTA updates the model parameters at each iteration using the different batch of unlabeled OOD data (x_{OOD}), as described in L6 of Algorithm 1. Note that during TTA, we also provide a batch of ID data (x_{ID}). After adaptation, we inference both ID and OOD data to the updated model to obtain its predictions (L7-8), stored for performance evaluation at the end of test. Stored predictions are used for calculating not just accuracy, but agreements which require additional model.

Such online test is different from traditional testing. Traditional testing tests the model that is fixed in its pre-trained parameters, while the online "test" is testing a model that is continuously updated its parameters during TTA. In other words, the model tested at iteration of $t-1$ is different from that at iteration t. Specifically in Algorithm 1, L6 makes the only difference between traditional and online testing, where L6 updates the model parameters every iteration before testing in L7 and L8.

Algorithm 1 Online Test in ID and OOD during TTA

```
1:  Inputs: A pair of data $\mathcal{X}_{\text{ID}}$, $\mathcal{Y}_{\text{ID}}$ $\mathcal{X}_{\text{OOD}}$, $\mathcal{Y}_{\text{OOD}}$ model $h_\theta$.
2:  Algorithms: Adaptation objective $\mathcal{L}_{\text{TTA}}(\cdot)$.
3:  ____________________________________________
4:  Prediction sets $\mathcal{P}_{\text{ID}} = \emptyset, \mathcal{P}_{\text{OOD}} = \emptyset$
5:  for batch ($x_{\text{ID}}$, $y_{\text{ID}}$), ($x_{\text{OOD}}$, $y_{\text{OOD}}$) in $\mathcal{X}_{\text{ID}}, \mathcal{Y}_{\text{ID}}, \mathcal{X}_{\text{OOD}}, \mathcal{Y}_{\text{OOD}}$ do
6:      $\theta \leftarrow \theta - \eta \nabla \mathcal{L}_{\text{TTA}}(h_\theta(x_{\text{OOD}}))$        ▷ Apply TTA
7:      $\mathcal{P}_{\text{ID}} = \mathcal{P}_{\text{ID}} \cup h_\theta(x_{\text{ID}})$
8:      $\mathcal{P}_{\text{OOD}} = \mathcal{P}_{\text{OOD}} \cup h_\theta(x_{\text{OOD}})$
9:  end for
10: return $\mathsf{Acc}(\mathcal{P}_{\text{ID}}, \mathcal{Y}_{\text{ID}}),\ \mathsf{Acc}(\mathcal{P}_{\text{OOD}}, \mathcal{Y}_{\text{OOD}})$
```



(a) CIFAR10 vs. CIFAR10.1

(b) ImageNet vs. ImageNet-V2

(c) ImageNet vs. ImageNet-R

(d) FMoW vs. FMoW-OOD

Figure 4: The linear trends visualization of shifts that already exhibit strong AGL in Vanilla. Notice that adaptations do not break the linear trends, even when they lead to accuracy drops. We test SHOT, TENT, and ETA on shifts such as CIFAR10.1, ImageNetV2, ImageNet-R, and FMoW-WILDS. Each blue and pink dot denotes the accuracy and agreement, followed by the linear fits for each. The axes are probit scaled.

B Analysis on distribution shifts that have AGL without adaptations

We also examine the distribution shifts that exhibit strong AGL even before TTA, which are dataset reproductions (CIFAR10.1 [38], ImageNetV2 [39]), and other real-world shifts (ImageNet-R [16], FMoW-WILDS [41]), as shown in Fig. 4. The results show that adaptation baselines, such as SHOT, TENT, and ETA, do not improve OOD generalizations or even degrade them, which were also evidenced by previous studies [53, 61]. However, they persist to have the strong linear trends, which implies that TTA does not break the linearity in shifts that show the strong correlations. This suggests that we can reliably predict if TTA may succeed or falter in a wide range of distribution shifts.

C OOD accuracy estimation baselines

C.1 ALine-S and ALine-D

Baek et al. [3] propose ALine-S and ALine-D, which assess the models' OOD accuracy without access to labels by leveraging the agreement-on-the-line among models. We provide the detailed algorithm of ALine-S and ALine-D in Algorithm 2.

C.2 Average thresholded confidence (ATC)

Garg et al. [8] introduce OOD accuracy estimation method, ATC, which learns the confidence threshold and predicts the OOD accuracy by using the fraction of unlabeled OOD samples for which model's negative entropy is less that threshold. Specifically, let $h(x) \in \mathbb{R}^c$ denote the softmax output of model h given data x from $\mathcal{X}_{\text{OOD}}$ for classifying among c classes. The method can be written as

Algorithm 2 ALine-S and ALine-D

```
**Input:** ID predictions $\mathcal{P}_{\text{ID}}$ and labels $\mathcal{Y}_{\text{ID}}$, OOD predictions $\mathcal{P}_{\text{OOD}}$.
**Function:** Probit transform $\Phi^{-1}(\cdot)$, Linear regression $\mathcal{F}(\cdot)$.

$\hat{a}, \hat{b} = \mathcal{F}(\Phi^{-1}(\mathsf{Agr}(\mathcal{P}_{\text{ID}})), \Phi^{-1}(\mathsf{Agr}(\mathcal{P}_{\text{OOD}})))$        ▷ Estimate slope and bias of linear fit
$\widehat{\mathsf{Acc}}^{\text{S}}_{\text{OOD}} = \Phi(\hat{a} \cdot \mathsf{Acc}(\mathcal{P}_{\text{ID}}, y_{\text{ID}}) + \hat{b})$        ▷ ALine-S
Initialize $A \in \mathbb{R}^{\frac{n(n-1)}{2} \times n}, b = \mathbb{R}^{\frac{n(n-1)}{2}}$
i=0
**for** $(p_{j,\text{ID}}, p_{k,\text{ID}}), (p_{j,\text{OOD}}, p_{k,\text{OOD}}) \in \mathcal{P}_{\text{ID}}, \mathcal{P}_{\text{OOD}}$ **do**
    $A_{ij} = \frac{1}{2}, A_{ik} = \frac{1}{2}, A_{il} = 0 \forall l \notin j, k$
    $b_i = \Phi^{-1}(\mathsf{Agr}(p_{j,\text{OOD}}, p_{k,\text{OOD}})) + \hat{a} \cdot \big(\frac{\Phi^{-1}(\mathsf{Acc}(p_{j,\text{ID}}, y_{\text{ID}}) + \Phi^{-1}(\mathsf{Acc}(p_{k,\text{ID}}, y_{\text{ID}}))}{2} - \Phi^{-1}(\mathsf{Agr}(p_{j,\text{ID}}, p_{k,\text{ID}}))\big)$
    i=i+1
**end for**
$w^* = \arg\min_{w \in \mathbb{R}^n} ||Aw - b||_2^2$
$\widehat{\mathsf{Acc}}^{\text{D}}_{\text{OOD}} = \Phi(w^*_i) \forall i \in [n]$        ▷ ALine-D
**return** $\widehat{\mathsf{Acc}}^{\text{S}}_{\text{OOD}}, \widehat{\mathsf{Acc}}^{\text{D}}_{\text{OOD}}$
```

below:

$$\widehat{\mathsf{Acc}}_{\text{OOD}} = \mathbb{E}\big[\mathbb{1}\{s(h(x)) < t\}\big], \tag{3}$$

where s is the negative entropy, *i.e.*, $s(h(x)) = \sum_c h_c(x) \log(h_c(x))$, and t satisfies

$$\mathbb{E}\big[\mathbb{1}\{s(h(x)) < t\}\big] = \mathbb{E}\big[\mathbb{1}\big\{\arg\max_c h_c(x) \neq y\big\}\big]. \tag{4}$$

C.3 Difference of confidence (DOC)-feat

Guillory et al. [10] observe that the shift of distributions is encoded in the difference of model's confidences between them. Based on this observation, they leverage such differences in confidences as the accuracy gap under distribution shifts for calculating the final OOD accuracy. Specifically,

$$\widehat{\mathsf{Acc}}_{\text{OOD}} = \mathsf{Acc}_{\text{ID}} - \Bigg(\mathbb{E}\big[\max_c h_c(x_{\text{ID}})\big] - \mathbb{E}\big[\max_c h_c(x_{\text{OOD}})\big]\Bigg). \tag{5}$$

C.4 Average confidence (AC)

Hendrycks and Gimpel [15] estimate the OOD accuracy based on model's averaged confidence, which can be written as

$$\widehat{\mathsf{Acc}}_{\text{OOD}} = \mathbb{E}\big[\max_c h(x_{\text{OOD}})\big]. \tag{6}$$

C.5 Agreement

Jiang et al. [22] observe that disagreement between the models that are trained with different setups closely tracks the error of models in ID. We adopt this as the baseline for assessing generalization under distribution shifts, where we can estimate $\widehat{\mathsf{Acc}}_{\text{OOD}} = \mathsf{Agr}(\mathcal{P}_{\text{OOD}})$, where $\mathcal{P}_{\text{OOD}}$ denotes the set of predictions of the models on OOD data $\mathcal{X}_{\text{OOD}}$.

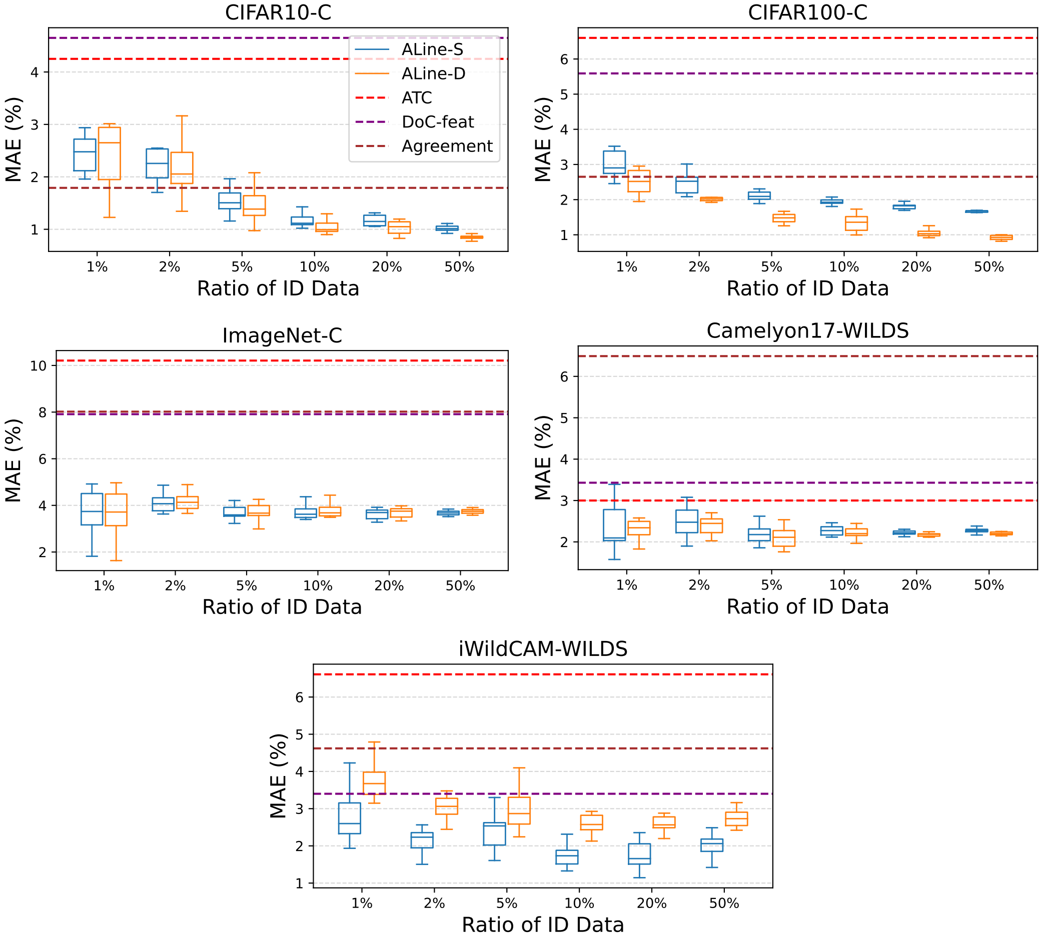


Figure 5: OOD accuracy estimation results with limited amount of ID data, decreasing from 50% to 1% of entire data pool. We randomly sampled 10 different subsets for each ratios, and visualize the distribution of MAE (%) results. The results of baseline including ATC, DoC-feat, and Agreement are included for comparison.

D Ablation study on the number of ID data

Even if access to labeled ID data is often easily available in practice more than obtaining OOD data, ID data could be limited in certain applications due to privacy issue or computational complexity. Therefore, this section further explores how sensitive our accuracy estimations are under conditions where the number of ID samples is critically limited.

Specifically, for each dataset, we gradually decrease the ratio of ID data for calculating both accuracy and agreement in ID, ranging from 50% to 1% of entire data. To report the confidence interval of different random subsets, we iterate 10 independent sampling of data for each ratio. Box plots in Fig. 5 show that for CIFAR10-C and CIFAR100-C, with even 5% begins to become close to precision of using 10× more, *i.e.*, 50%, achieving the state-of-the-art estimation performances outperforming other baselines. Also, in more complex dataset such as ImageNet-C, Camelyon17-WILDS, and iWildCAM-WILDS, even using 1% of entire ID data achieves the best estimation performances among baselines. These results indicate that our framework is practically feasible in such practical applications where access to ID is highly limited.

HyperParameter	CIFAR10-C						CIFAR100-C				
	BN_Adapt	TTT	SHOT	TENT	ETA	ConjPL	BN_Adapt	SHOT	TENT	ETA	ConjPL
Learning Rate	–	3.71	0.65	0.72	0.72	0.42	–	0.90	1.05	0.67	0.68
Adapt Step	–	0.04	0.23	0.23	0.24	0.12	–	0.63	0.0	0.55	0.63
Architecture	0.21	0.49	0.03	0.03	0.01	0.04	0.61	1.13	0.69	0.84	0.38
Batch Size	0.0	–	0.73	0.77	0.77	0.18	0.02	0.57	0.50	0.62	0.75
Checkpoints	0.0	0.48	0.07	0.05	0.01	0.11	0.01	0.00	0.24	0.0	0.0
Average	0.07	1.18	0.34	0.36	0.35	0.17	0.21	0.48	0.49	0.53	0.48

HyperParameter	ImageNet-C				ImageNet-R			Camelyon17		iWildCAM	
	BN_Adapt	TENT	ETA	SAR	BN_Adapt	TENT	ETA	BN_Adapt	TENT	BN_Adapt	TENT
Learning Rate	–	9.70	6.67	6.21	–	2.80	5.50	–	1.14	–	3.17
Adapt Step	–	0.30	0.0	3.07	–	0.0	0.0	–	0.0	–	1.97
Architecture	2.23	0.39	0.24	0.75	0.06	0.85	1.83	1.22	0.0	0.0	4.08
Batch Size	0.00	5.61	5.70	7.21	0.0	1.74	4.02	0.0	1.37	0.15	2.04
Average	1.15	4.0	3.15	4.31	0.03	1.34	2.83	0.61	1.25	0.07	2.81

Table 7: Results of unsupervised validation of TTA methods other than TENT, measured by MAE (%) between OOD accuracy of model selected by best ID model and actual best OOD model. Since BN_Adapt do not involve parameter updates, we exclude learning rate and the adaptation steps. In addition, TTT uses a single batch, and we exclude the batch size.

E Unsupervised validation results on TTA baselines

This section supplements Table 3 by the unsupervised validation results of various TTA methods other than TENT, which include those of BN_Adapt, TTT, SHOT, ETA, ConjPL, and SAR. Table 7 reports the difference in OOD accuracy (MAE (%)) between the model with best ID accuracy and the actual best OOD model, tested across different adaptation hyperparameters. The OOD model selection via best ID model achieves less than 1% MAE over almost every hyperparameter setup and adaptations, in CIFAR10-C and CIFAR100-C every corruptions in average. It also demonstrates that the hyperparameter optimization in ImageNet-C, R and WILDS benchmark datasets have reasonably low MAE across diverse setups and adaptation baselines. These indicate that the strong linear trends across hyperparameters enable the selection of near-optimal OOD adaptations across various shifts and TTA strategies.

https://doi.org/10.52202/079017-3820

F Additional results on CIFAR10-C, CIFAR100-C, and ImageNet-C

https://doi.org/10.52202/079017-3820

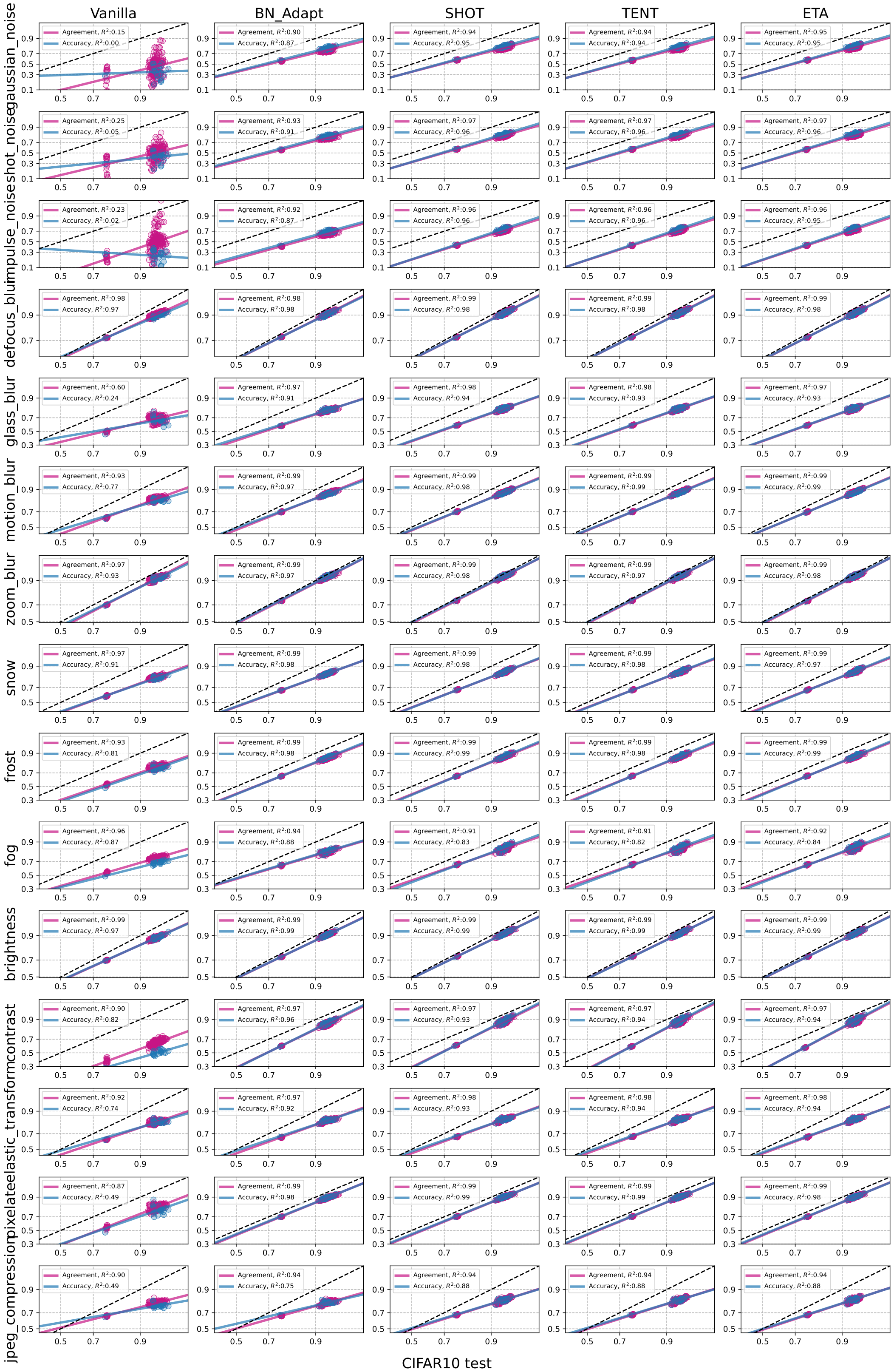


Figure 6: Results on CIFAR10-C corruptions (different architectures), BN_Adapt [44], SHOT [27], TENT [53], and ETA [35].

https://doi.org/10.52202/079017-3820

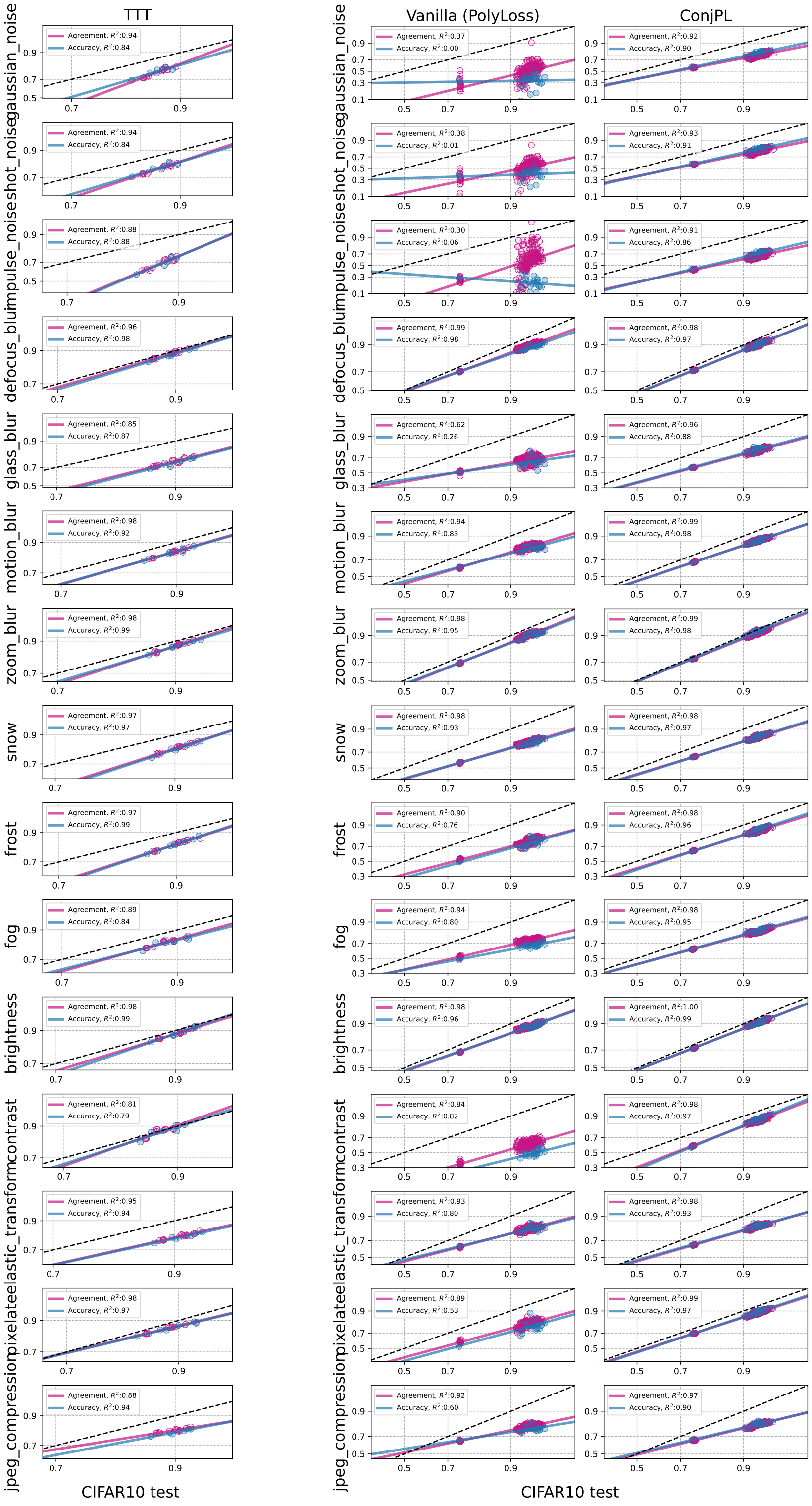


Figure 7: Results on CIFAR10-C corruptions (different architectures), TTT [46], ConjPL [9]

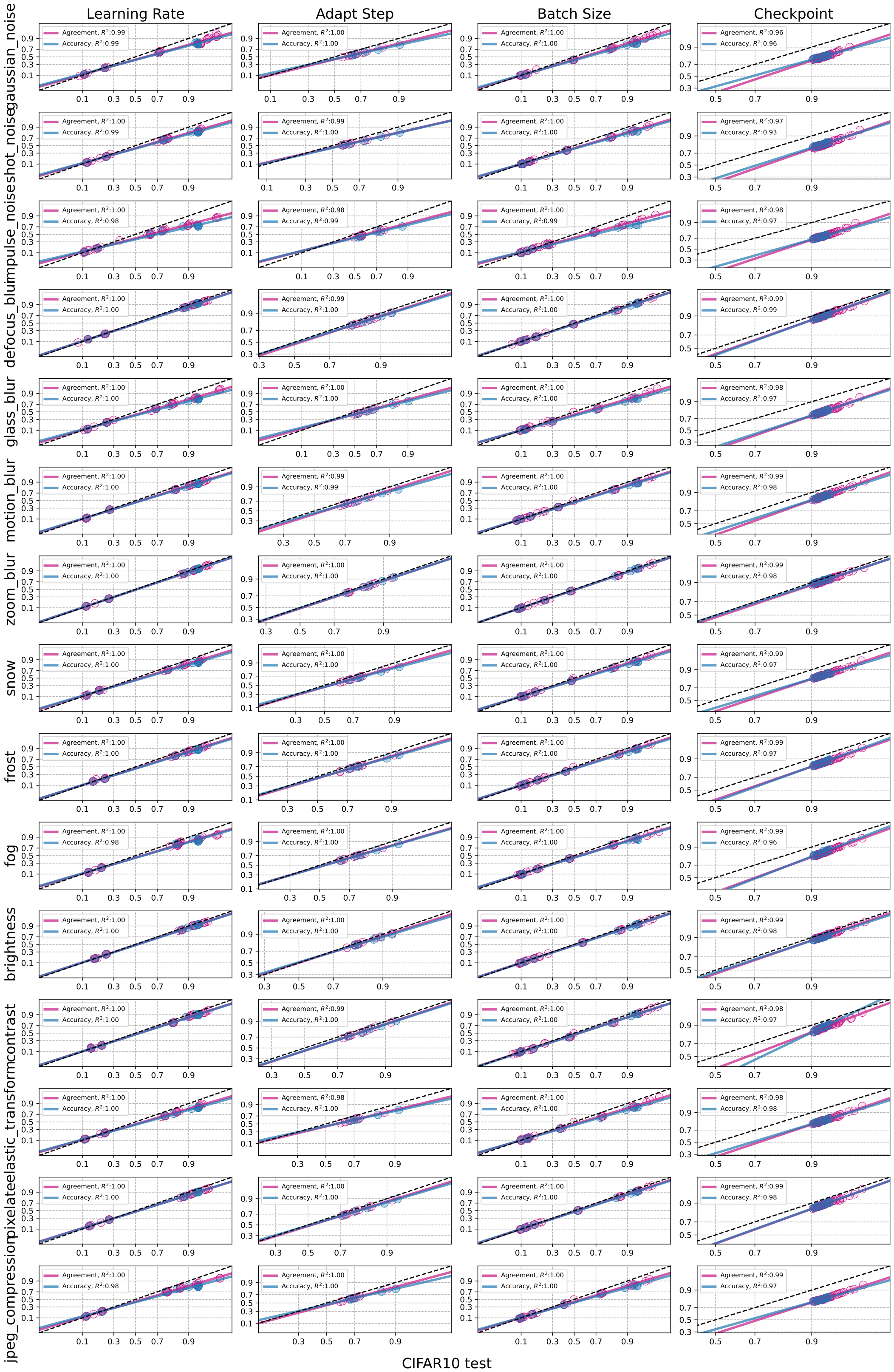


Figure 8: Results on CIFAR10-C corruptions (adaptation hyperparameters of TENT [53]).

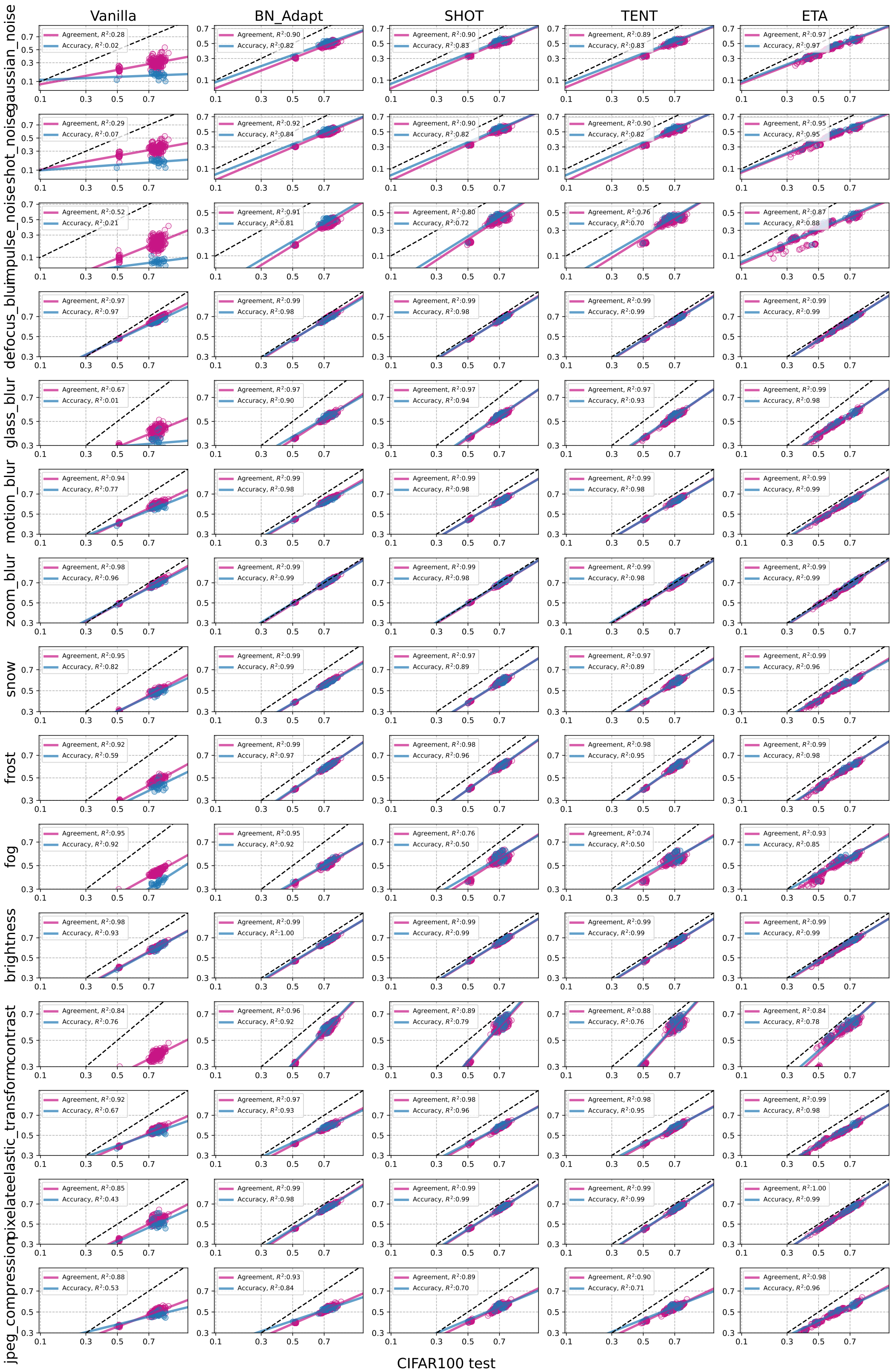


Figure 9: Results on CIFAR100-C corruptions (different architectures), BN_Adapt [44], SHOT [27], TENT [53], and ETA [35].

https://doi.org/10.52202/079017-3820

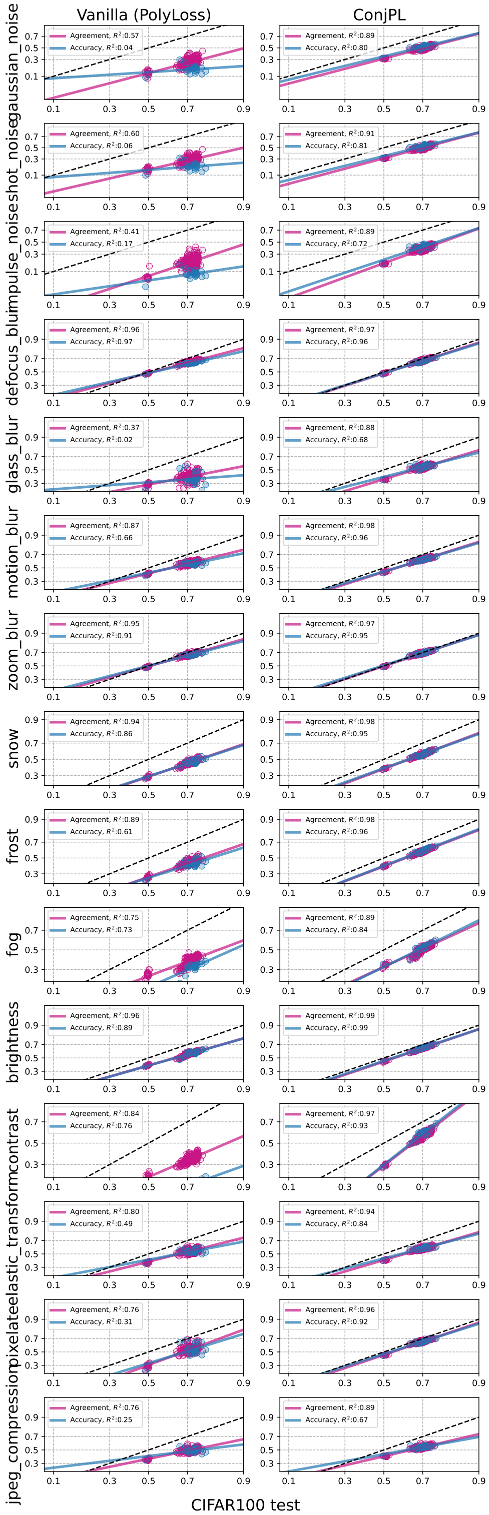


Figure 10: Results on CIFAR100-C corruptions (different architectures), ConjPL [9]

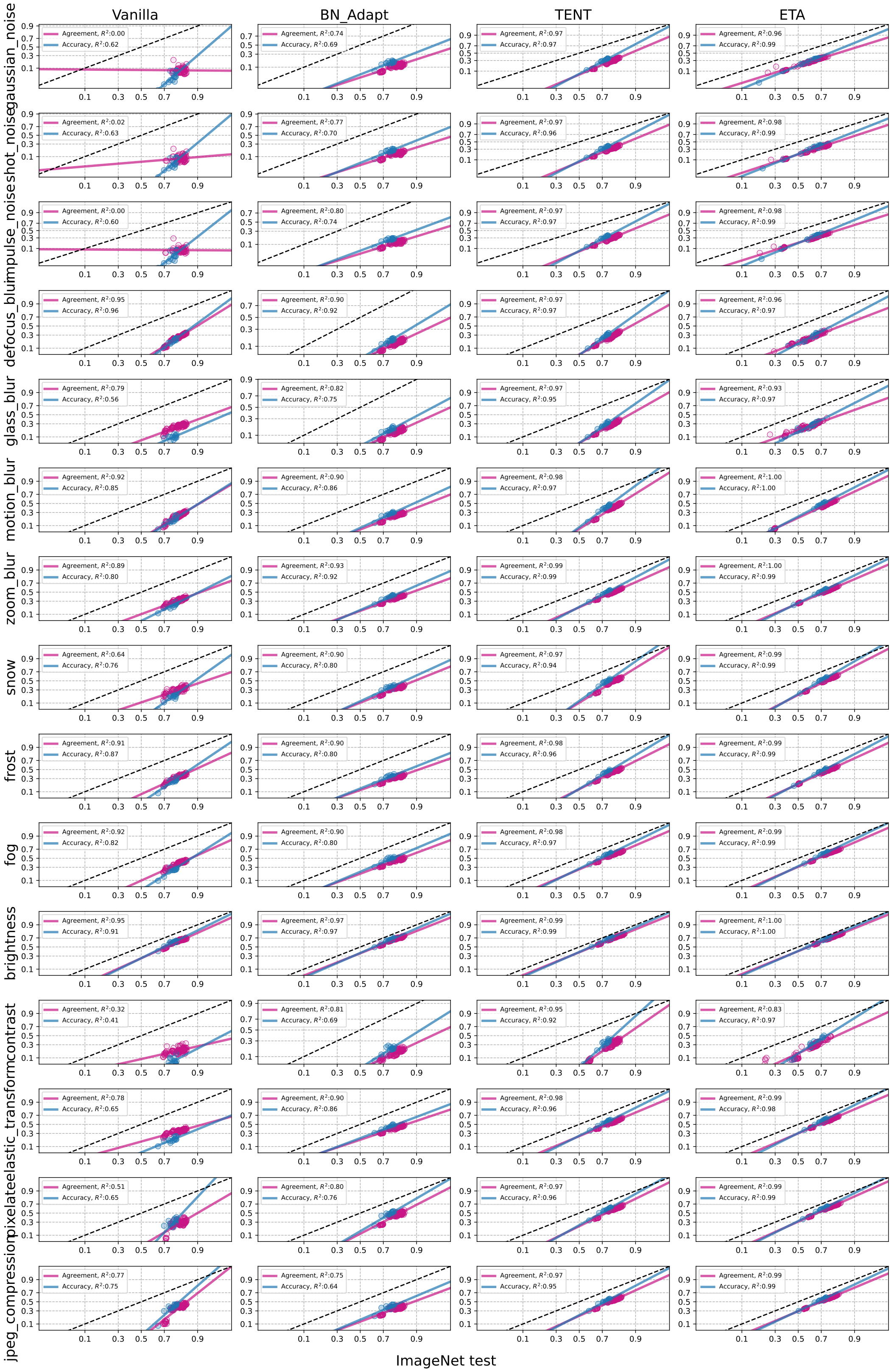


Figure 11: Results on ImageNet-C corruptions (different architectures), BN_Adapt [44], TENT [53], and ETA [35].

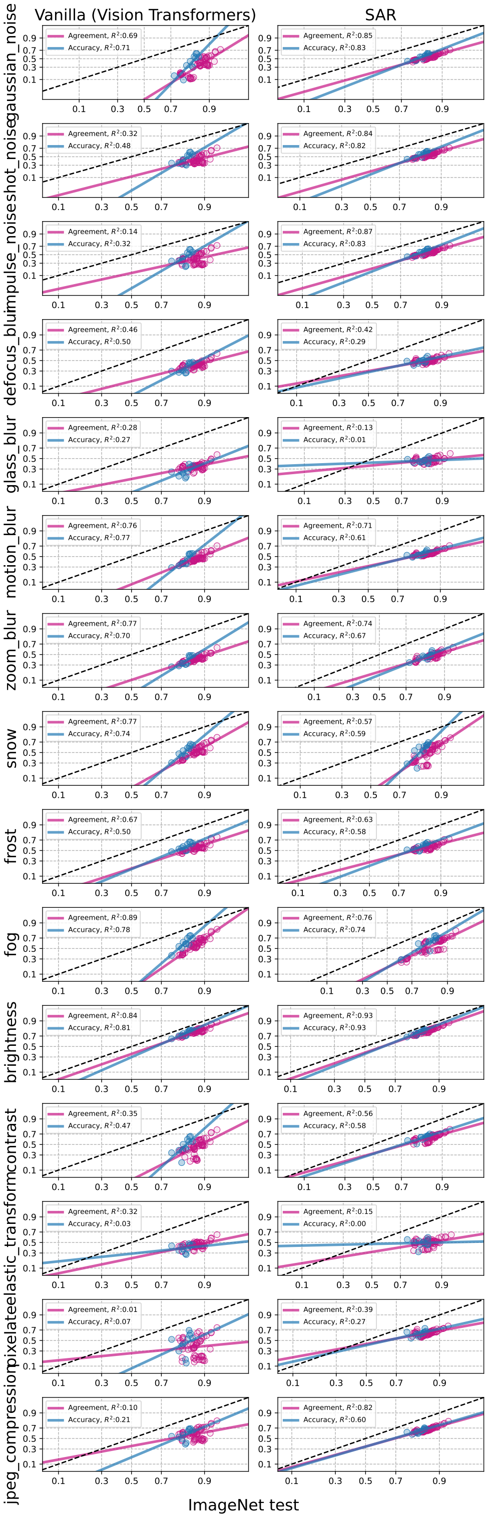


Figure 12: Results on ImageNet-C corruptions (different architectures), SAR [36]

https://doi.org/10.52202/079017-3820

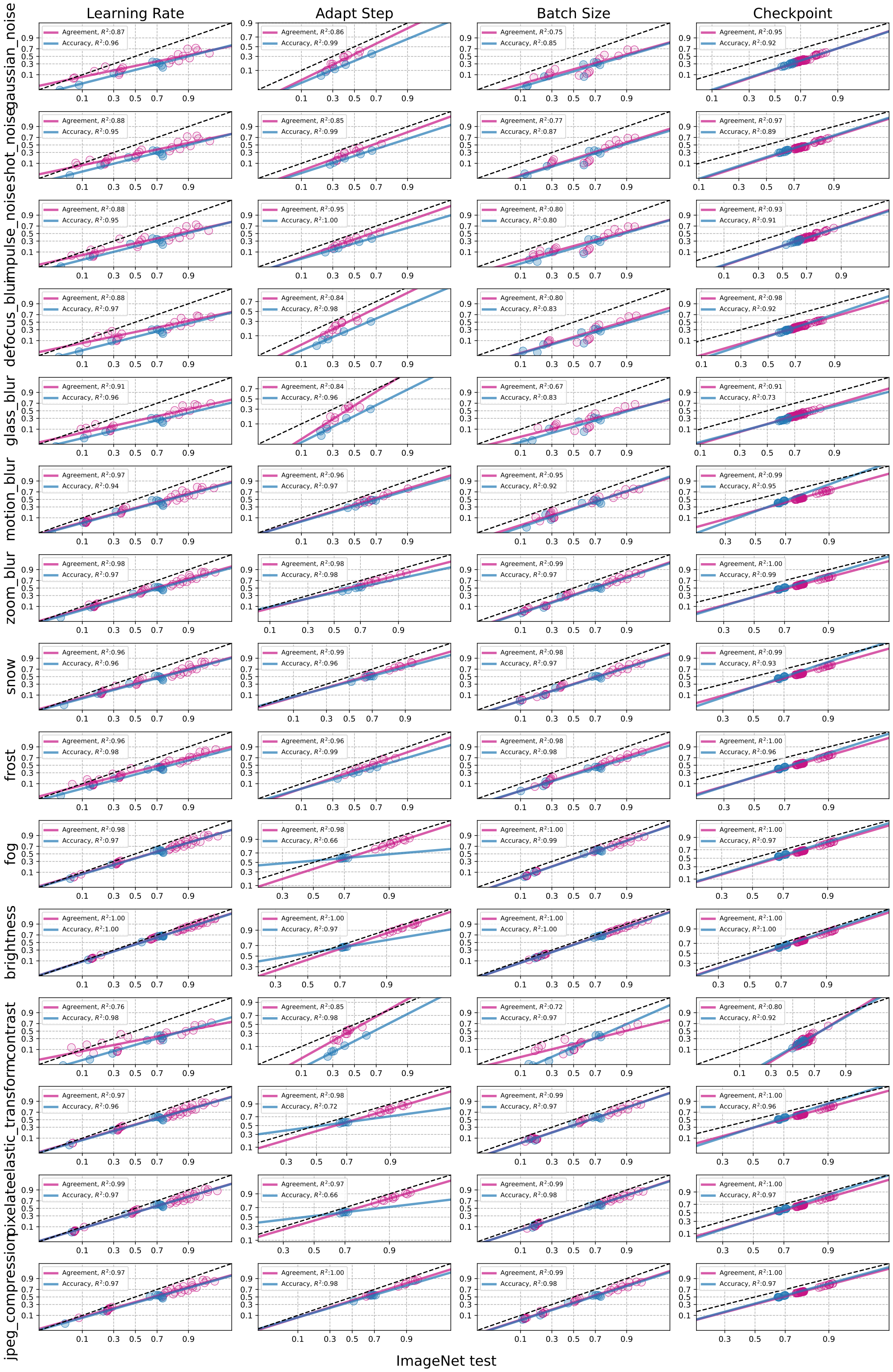


Figure 13: Results on ImageNet-C corruptions (adaptation hyperparameters of ETA [35]).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Both abstract and introduction includes our main observations, analysis on them, followed by experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 7, we discuss about our paper's limitation that AGL requires ID data during test-time, and we conduct the ablation study on limited number of ID data in Section D.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-3820

Answer: [Yes]

Justification: In Section 4.1, we provide exact notations, setups, theorems and proofs with proper numbers and references.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide the detailed algorithm for online test in ID and OOD in Algorithm 1, which is very easy to follow. Also, the estimation algorithm we employ in other paper already has the public code, but we provide detailed algorithm in Algorithm 2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

https://doi.org/10.52202/079017-3820

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our work mainly focuses on existing test-time adaptation methods as well as publicly available datasets, which are already highly accessible.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed experimental setups for our observation in Section 3.1 as well as our experiments in Sections 5.1, 5.2, and Section A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We conducted iterative experiments on measuring mean and covariance similarity in Table 1, as well as subsampling ID data in Fig. 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

https://doi.org/10.52202/079017-3820

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the details about the type of GPU machines for our entire experiments in Section A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics, and confirmed that our paper confirms them.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Section 7, we discuss the positive societal impacts of our work, the significant improvements of reliability of neural networks under distribution shift.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This work does not include such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly cite all the test-time adaptation baselines and datasets we used in our paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

https://doi.org/10.52202/079017-3820

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This work does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.