Reliable Test-Time Adaptation via Agreement-on-the-Line

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

Abstract

Test-time adaptation (TTA) methods aim to improve robustness to distribution shifts by adapting models using unlabeled data from the shifted test distribution. However, there remain unresolved challenges that undermine the reliability of TTA, which include difficulties in evaluating TTA performance, miscalibration after TTA, and unreliable hyperparameter tuning for adaptation. In this work, we make a notable and surprising observation that TTAed models strongly show the agreement-on-the-line phenomenon [1] across a wide range of distribution shifts. We find such linear trends occur consistently in a wide range of models adapted with various hyperparameters, and persist in distributions where the phenomenon fails to hold in vanilla models (*i.e.*, before adaptation). We leverage these observations to make TTA methods more reliable in three perspectives: (i) estimating OOD accuracy (without labeled data) to determine when TTA helps and when it hurts, (ii) calibrating TTAed models without label information, and (iii) reliably determining hyperparameters for TTA without any labeled validation data. Through extensive experiments, we demonstrate that various TTA methods can be precisely evaluated, both in terms of their improvements and degradations. Moreover, our proposed methods on unsupervised calibration and hyperparameters tuning for TTA achieve results close to the ones assuming access to ground-truth labels, in terms of both OOD accuracy and calibration error.

1 Introduction

Machine learning models often fail to generalize to new distributions [2, 3, 4] – so-called out-of-distribution (OOD) data — which differ from the one they were trained on, referred to as in-distribution (ID) data. This can lead to a significant degradation in their performance during test time. Recently, there has been a surge in research on test-time adaptation (TTA), a technique that adapts models to the target distribution using only unlabeled test data. These involve adaptation strategies including estimating test-time feature statistics [5], self-supervision [6, 7, 8], entropy minimization [9, 10, 11, 12, 13], and self-training with pseudo-labels [9, 14, 15]. These efforts have aimed to enhance model robustness in the face of distribution shifts where labeled data is unavailable.

Despite the progress, several critical bottlenecks persist, undermining the reliable applications of TTA methods in practice. Firstly, TTA is not universally effective for all distribution shifts and can sometimes lead to performance degradation [10, 7, 16]. Moreover, the absence of labeled test data hinders the evaluation of model performance, thereby making it unclear in advance whether these methods will work or not. Secondly, TTA methods often result in poorly calibrated models [17, 18, 14], posing potential risks in safety-critical applications. Thirdly, TTA methods are often extremely sensitive to their hyperparameters during adaptation [19, 16, 20], and their tuning procedures lack clarity. Most of them often follow the same settings of the previous studies [12, 13], or rely on some held-out labeled data [21, 11, 15], which might be unavailable in practice. To our knowledge, there is little work addressing these shortcomings given no access to the labels during test time.

In a separate line of work, [1] show that the ID and OOD *agreement* between classifiers (*i.e.*, the average extent to which two classifiers make the same prediction on an unlabeled datasets)

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

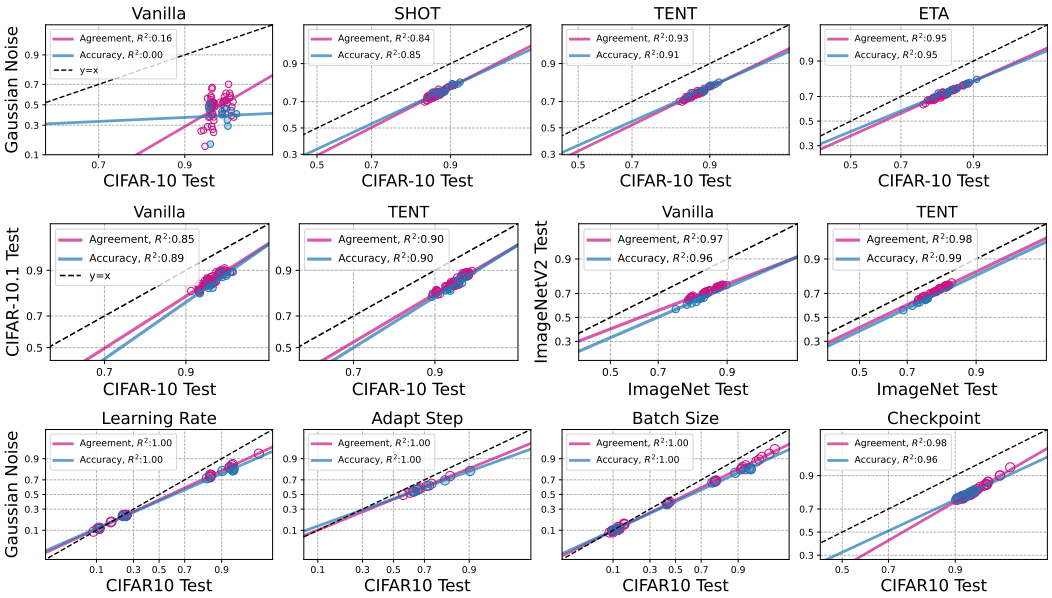

Figure 1: TTA makes the AGL and ACL notably stronger than their base counterparts (first row), or persist (second row), across synthetic and real-world shifts. Furthermore, adaptations with varying hyperparameters also show strong linear trends (third row). Each blue point indicates the accuracy of the models, while pink denotes the agreement between a pair of them. The axes are probit scaled.

show a strong linear correlation, akin to the same phenomenon showed for ID vs. OOD *accuracy*, demonstrated by [22]. Taken together, these phenomena, the so-called agreement-on-the-line (AGL) and accuracy-on-the-line (ACL) present a method for assessing OOD accuracy without labeled data.

In this paper, we observe a noteworthy phenomenon: after applying TTA, AGL and ACL *persist* or even hold to a *stronger* degree than in their base counterparts. In other words, when we assess the accuracies and agreements of the models adapted to OOD data, the strong correlations in ID vs. OOD consistently hold across distribution shifts, including those where vanilla models do not exhibit such trends. Interestingly, these correlations occur not only when TTA improves OOD accuracy, but also when it fails to enhance or even negatively affects OOD accuracy, especially under real-world shifts. We observe such trends not just among the models with different architectures, but within those with same architecture but adapted with varying values of their adaptation hyperparameters, including learning rates, the number of adaptation steps, and others.

These findings, with strong AGL and ACL after TTA, lead to the enhancement of TTA methods for improved reliability. We can first *predict the effectiveness of TTA methods*, *i.e.*, whether they succeed or fail and to what extent, across distribution shifts. Specifically, our approach uses the ALine-S and ALine-D techniques from [1], and applies them to test-time adapted models. The result is that, without any labeled data at all, we can estimate the accuracy of TTAed models better than we can for vanilla models, especially for shifts where vanilla falls short . Such estimation results also enable the identification of shifts where TTA methods might potentially struggle to improve accuracy. Second, we introduce a novel variant of the temperature-scaling method, which achieves *model calibration solely through estimated accuracy*, representing an unsupervised approach that eliminates the need for labeled data as required by the original temperature scaling [23]. We observe that it effectively reduces the expected calibration error (ECE) [23] close to the best achievable lower-bound using ground-truth labels. Finally, we introduce the *reliable hyperparameter optimization* strategy for adaptations without access to labels: selecting model with the highest ID accuracy. Across all TTA baselines we employ, the majority of models chosen through our approach exhibit performance comparable to those selected using ground-truth labels .

2 Strong agreement-on-the-line after TTA

2.1 Experimental setup

Datasets and models. We evaluate on both synthetic corruptions (CIFAR10-C, CIFAR100-C, ImageNet-C [24]) with highest severity, datasets reproductions (CIFAR10.1 [25], CIFAR10.2 [26],

ImageNetV2 [27]), and real-world shifts (ImageNet-R [28], FMoW-WILDS [29, 4]). We leverage a variety of different network architectures, which encompass ResNet [30, 31], ResNext [32], VGG [33], GoogLeNet [34], DenseNet [35], and MobileNet [36] with differing depths and widths. For evaluation on ImageNet and its shifts, we use pretrained weights publicly accessible from torchvision[1], except for ConjPL, which require training with PolyLoss [37]. We train models for the other datasets. Additional details are provided in Section 4.1 in appendix.

Test-time adaptation baselines. To gain generality, we test TTA methods that involve different update parameters (*e.g.*, batch normalization (BN) layers, entire encoder parameters), objectives (*e.g.*, entropy minimization, self-supervision task), and source-training objectives (*e.g.*, cross-entropy loss, PolyLoss). These include BN_Adapt [5], SHOT [9], TTT [6], TENT [10], ConjPL [15], ETA [12], and SAR [13]. We examine their key adaptation hyperparameters that are shared among all baselines, including learning rates, number of adapt steps, and batch size. We also test different checkpoints of the source-trained model as another possible hyperparameter to select.

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

$$\text{Agreement}(h, h') = \mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\mathbb{1}\{h(x) = h'(x)\} \right], \tag{1}$$

where $h(x)$ and $h'(x)$ are final predictions of models h and h'. Following [22, 1], we apply the inverse of the cumulative density function of the standard Gaussian distribution, namely probit-transformation ($\Phi^{-1} : [0, 1] \to [-\infty, \infty]$), on the axes of accuracy and agreement, for a better linear fit.

2.2 Main observation

Agreement (and accuracy) on-the-line persists or holds stronger after TTA. We find that after TTA methods, there are strong correlations in agreement (and accuracy) between ID and OOD across both synthetic and real-world shifts. Most importantly, we notice that such trends become more persistent after adaptations, exhibiting strong correlations even across datasets where the models before adaptations have failed to have, as noted in [22] and [1]. We show such phenomenon in Figure 1 (first row): when tested on CIFAR10 test vs. Gaussian-Noise corruption, models with different architectures have weak correlations with their coefficients of determination (R^2) values being significantly low. After applying various TTA methods, SHOT, TENT, and ETA, these models consistently have much stronger AGL and ACL (*e.g.*, R^2 improves $0.16 \to 0.95, 0.00 \to 0.95$ after ETA), as well as the alignments between the lines. Furthermore, these trends also occur in real-world shift datasets, where TTA sometimes even fails to improve OOD accuracy, as shown in Figure 1 (second row). Here, we examine on CIFAR10.1 and ImageNetV2, where models before and after TENT have similar linear trends and maintain high R^2 values. In particular, as seen in the plots of CIFAR10.1 and ImageNetV2, TENT does not improve generalization, or even results in degradation [10, 16]. This highlights the consistent AGL and ACL trends across TTA methods, irrespective of TTA's actual performance improvement across diverse distribution shifts where it may succeed or falter.

Such linear trends are also observed when varying TTA hyperparameters. We find that these trends in TTA can also be obtained by leveraging models adapted with varying values of TTA setups. As mentioned in Section 4.1, we examine learning rates, the number of adaptation steps, batch size, and the checkpoints of the source-trained model. Here we fix the model architecture while systematically varying specific hyperparameters within defined ranges. Figure 1 (third row) shows that models adapted with different hyperparameter values exhibit strong AGL trends among them, with their R^2 values close to 1, when tested on CIFAR10 vs. CIFAR10-C Gaussian-Noise corruption. The resulting TTA performances vary according to the different hyperparameter values, and they seem sensitive particularly in terms of learning rates and batch sizes. Still, these results, improved or degraded, lie on the same positive correlation line in both agreement and accuracy. Such trends are of significant value for TTA, where leveraging models with different architectures may necessitate training them separately in advance. By using different hyperparameters on a single model, we can eliminate the need to train multiple models instead. See Section 4.2 in appendix for additional results.

[1] https://pytorch.org/vision/stable/models.html

Dataset	Method	ATC	DOC-feat	AC	Agreement	ALine-S	ALine-D
CIFAR10-C	Vanilla	9.26	14.44	16.74	7.70	5.50	5.17
	SHOT	2.91	5.32	9.94	1.48	0.73	**0.56**
	TENT	9.21	5.32	10.08	1.51	0.73	**0.53**
	ETA	10.18	5.10	11.02	1.50	0.71	**0.56**
	SAR	1.14	4.71	7.76	1.79	0.89	**0.77**
ImageNet-C	Vanilla	4.27	13.47	17.76	22.28	5.87	5.87
	SHOT	5.83	4.46	9.84	14.74	4.17	**4.13**
	TENT	10.92	6.30	20.25	13.74	**4.16**	4.18
	ETA	6.34	5.94	23.56	13.20	**4.00**	4.21
	SAR	7.04	**3.86**	11.00	12.27	4.31	5.25

Table 1: Mean absolute error (MAE) (%) of the accuracy estimation on TTAed models with different architectures. ALine-S/D on TTAed models leads to substantially lower errors in accuracy estimation, compared to ALine-S/D applied on the vanilla models as well as other estimation baselines.

3 Reliable test-time adaptation

Based on the observations, we investigate the enhancement of existing TTA methods in three ways: (i) accurate estimation of OOD accuracy, (ii) calibration after TTA, and (iii) reliable hyperparameter optimization – all performed without assuming access to the labels.

3.1 Accuracy estimation

We illustrate how the strong AGL trend shown among TTAed models enables the precise accuracy estimation on target OOD data. We employ the estimation method called ALine-S and ALine-D, presented in [1], but utilize them within the context of TTA models. Specifically, we leverage the stronger correlations and alignments in agreement and accuracy exhibited by TTAed models to further optimize such estimation method. The details of our utilization are described in Algorithms 1. This approach yields improved accuracy estimation compared to vanilla models and other estimation baselines. Additionally, in Table 5, we investigate the forecastability of TTAed models in the face of various distribution shifts.

Algorithm 1 Accurate Estimation of TTA

Inputs: Labeled ID data $\mathcal{X}_{\text{ID}}, \mathcal{Y}_{\text{ID}}$, unlabeled OOD data $\mathcal{X}_{\text{OOD}}$, a set of ID-trained n models $\mathcal{H} = \{h_\theta, ..., h_\theta\}$, sets $\mathcal{P}_{\text{ID}} = \emptyset$, $\mathcal{P}_{\text{OOD}} = \emptyset$.
Algorithms: TTA objective $\mathcal{L}_{\text{TTA}}(\cdot)$, ALine-S/D$(\cdot)$.

for batch $x_{\text{ID}}, x_{\text{OOD}}$ in $\mathcal{X}_{\text{ID}}, \mathcal{X}_{\text{OOD}}$ **do**
  for $h_\theta \in \mathcal{H}$ **do**
    $\theta \leftarrow \theta - \eta \nabla \mathcal{L}_{\text{TTA}}(h_\theta(x_{\text{OOD}}))$ ▷ Apply TTA
    $\mathcal{P}_{\text{ID}} = \mathcal{P}_{\text{ID}} \cup h_\theta(x_{\text{ID}})$
    $\mathcal{P}_{\text{OOD}} = \mathcal{P}_{\text{OOD}} \cup h_\theta(x_{\text{OOD}})$
  end for
end for
return ALine-S/D$\big(\mathcal{P}_{\text{ID}}, \mathcal{Y}_{\text{ID}}, \mathcal{P}_{\text{OOD}}\big)$

Strong ACL and AGL in TTA lead to precise accuracy estimation. Table 1 reports the mean absolute error (MAE) between actual and the estimated accuracy for both vanilla and the adapted models. We evaluate them on widely used benchmarking distribution shifts in TTA literature. In particular, for CIFAR10-C, applying ALine-S/D on TTAed models achieve substantially lower MAE compared to that of vanilla models (*e.g.*, $5.17\% \rightarrow 0.53\%$ of ALine-D after TENT). This suggests that ALine-S/D can be substantially more *effective* in estimating accuracy of the TTAed models, primarily attributed to their *stronger AGL* than vanilla models. We also compare with other estimation baselines and find that across different datasets, ALine-S and ALine-D consistently outperform existing baselines on estimating TTA performance. See Table 4 for additional results.

3.2 Unsupervised calibration

Proposed method. In this section, we introduce a variant of temperature-scaling [23] that calibrates models only *with the estimated accuracy* of the model, not the labeled data. Specifically, let X the random variable for data and $f(X)$ be the logit output of the neural network f given input X for classifying among c categories. We define a simple root-finding problem that finds an optimal temperature value τ that scales the model's averaged confidence to match to the estimated accuracy Acc_{est}. This can be written as

$$\text{Find } \tau \text{ such that } \mathbb{E}\left[\max_c \text{softmax}\left(\frac{f(X)}{\tau}\right)\right] = \text{Acc}_{\text{est}} \tag{2}$$

We use Newton's method to find the optimal τ via root-finding. Once the optimal τ is found, we then temperature-scale the prediction using this value.

Method	CIFAR10-C			CIFAR100-C			ImageNet-C		
	Uncalib.	Ours	Oracle	Uncalib.	Ours	Oracle	Uncalib.	Ours	Oracle
Vanilla	17.48	9.71	3.42	15.25	15.82	3.09	14.97	7.92	1.79
BN_Adapt	7.99	2.73	2.49	9.59	2.85	2.37	3.21	1.98	1.67
TENT	7.76	3.11	2.77	13.40	2.10	2.06	7.38	4.56	2.94
ETA	7.72	3.13	2.74	15.66	4.93	4.48	12.80	8.23	7.40
Vanilla*	21.80	12.99	3.96	25.70	15.66	2.90	30.32	11.63	2.03
ConjPL	11.89	4.43	3.56	25.48	3.01	3.00	17.13	5.50	3.84

Table 2: For various TTA methods, our unsupervised calibration method significantly reduces the ECE compared to that of vanilla, while also manifesting negligible gap to oracle-bound results.

Experimental results. Table 2 presents a comparison of ECE between vanilla and adapted models using various TTA methods. "Vanilla*" denotes the vanilla model pretrained with PolyLoss [37], as used in ConjPL. The "Oracle" approach represents the lower-bound of the best achievable ECE through temperature scaling, where we sweep over the grids of temperature candidates and find τ that minimizes ECE using ground-truth labels. The results in "Uncalib." first demonstrate that TENT, ETA, and ConjPL result in worse calibration compared to BN_Adapt [38]. Such results are mainly attributed to entropy minimization employed in these methods, applied across all samples regardless of their correctness [39, 18]. After applying our calibration method, however, their calibration errors substantially decrease to levels close to the lower-bound achieved by the oracle in every dataset. Interestingly, when we apply our method to vanilla models, where the presence of ACL and AGL is less pronounced, a substantial disparity still remains in the calibration results between our method and the lower bound represented by the oracle. This emphasizes the effectiveness of our method specifically in the setting of TTA where we observe prominent AGL.

HyperParameter	TTA Method						
	BN_Adapt	SHOT	TENT	ETA	ConjPL	SAR	TTT
Learning Rate	–	0.65	0.72	0.72	0.42	0.24	3.71
Adapt Step	–	0.23	0.23	0.24	0.12	0.43	0.04
Architecture	0.21	0.03	0.03	0.01	0.04	0.20	0.49
Batch Size	0.0	0.73	0.77	0.77	0.18	0.06	–
Checkpoints	0.0	0.07	0.05	0.01	0.11	0.01	0.48

Table 3: Mean Absolute Error (MAE) (%) between accuracy of models selected by our method vs. ground-truth. Across TTA methods, our method consistently identifies hyperparameters that lead to TTAed models achieving OOD accuracy close to those selected using ground-truth labels.

3.3 Reliable hyperparameter optimization

Given the ACL phenomena of TTAed models across varying hyperparameter values (see Figure 1 third row), *selecting the best-performing model on ID data* emerges as a straightforward and effective strategy for model selection in OOD shifts. Hyperparameter tuning has remained a significant challenge in the context of TTA [19, 16, 20], owing to the absence of labelled OOD test data. We propose to find the best hyperparameters, by first exploring a wide range of hyperparameter candidates and then selecting the TTAed models that best performs in ID. Table 3 reports the OOD accuracy gap between the models selected by our approach vs. best-performing in OOD using ground-truth labels, tested on CIFAR10-C. The results show that across different corruptions as well as TTA baselines, selecting the best-performing hyperparameters in ID consistently results in negligibly high OOD accuracy, with less than 1% MAE compared to those selected with labeled data. For sensitive hyperparameters, e.g. learning rates and batch sizes, where incorrect selections can lead to significant performance degradation, our approach consistently identifies near-optimal hyperparameters.

4 Conclusion and Discussions

We observe that TTA maintains or reinforces the strong linear ID vs. OOD correlations in the accuracy and agreement, leading to enhancement of TTA reliability. This naturally raises questions about how to theoretically characterize the conditions under which adaptations enhance these linear trends. This is a promising future direction that can ascertain the reliable observations of AGL, leading to reliable TTA across *any* types of distribution shifts. Additionally, observing AGL and ACL requires access to ID test data, which might raise privacy concerns and also demand additional computational resources. Overcoming such dependencies on ID data and exploring "fully" test-time approach of observing AGL and ACL remain promising directions for future research.

Acknowledgments. Eungyeup Kim and Mingjie Sun 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] Christina Baek, Yiding Jiang, Aditi Raghunathan, and J. Zico 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.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

[3] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[4] Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, Sara Beery, Etienne David, Ian Stavness, Wei Guo, Jure Leskovec, Kate Saenko, Tatsunori Hashimoto, Sergey Levine, Chelsea Finn, and Percy Liang. Extending the wilds benchmark for unsupervised adaptation. In *International Conference on Learning Representations (ICLR)*, 2022.

[5] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias 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.

[6] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, 2020.

[7] Yuejiang Liu, Parth Kothari, Bastien Germain van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[8] Yossi Gandelsaman, Yu Sun, Xinlei Chen, and Alexei A. Efros. Test-time training with masked autoencoders. In *Advances in Neural Information Processing Systems*, 2022.

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

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

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

[12] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *The Internetional Conference on Machine Learning*, 2022.

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

[14] Evgenia. Rusak, Steffen Schneider, George Pachitariu, Luisa Eck, Peter Gehler, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. If your data distribution shifts, use self-learning. *Transactions of Machine Learning Research*, 2022.

[15] Sachin Goyal, Mingjie Sun, Aditi Raghunanthan, and Zico Kolter. Test-time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems*, 2022.

[16] Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *International Conference on Machine Learning (ICML)*, 2023.

[17] Cian Eastwood, Ian Mason, Christopher K. I. Williams, and Bernhard Schölkopf. Source-free adaptation to measurement shift via bottom-up feature restoration. In *International Conference on Learning Representations*, 2022.

[18] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *CVPR*, 2022.

[19] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca 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.

[20] Linus Ericsson, Da Li, and Timothy Hospedales. Better practices for domain adaptation. In *AutoML Conference 2023*, 2023.

[21] Yusuke Iwasawa and Yutaka 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] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig 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.

[23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.

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

[25] Benjamin Recht, Rebecca Roelofs, Ludwig Schimidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv: 1806.00451*, 2018.

[26] Shangyun Lu, Bradley Nott, Aaron Olson, Alberto Todeschini, Hossein Vahabi, Yair Carmon, and Ludwig Schmidt. Harder or different? a closer look at distribution shift in dataset reproduction. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

[27] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal 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.

[28] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

[29] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian 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.

[31] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks, 2017.

[32] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.

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

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

[35] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[37] Zhaoqi Leng, Mingxing Tan, Chenxi Liu, Ekin Dogus Cubuk, Jay Shi, Shuyang Cheng, and Dragomir Anguelov. Polyloss: A polynomial expansion perspective of classification loss functions. In *International Conference on Learning Representations*, 2022.

[38] Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift, 2021.

[39] Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Sentry: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8558–8567, 2021.

[40] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

[41] Saurabh Garg, Sivaraman Balakrishnan, Zachary Chase Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

[42] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1134–1144, October 2021.

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

[44] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*, 2022.

[45] Jun-Kun Wang and Andre Wibisono. Towards understanding gd with hard and conjugate pseudo-labels for test-time adaptation. In *Internetional Conference on Learning Representations*, 2023.

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

[47] Ansh Khurana, Sujoy Paul, Piyush Rai, Soma Biswas, and Gaurav Aggarwal. Sita: Single image test-time adaptation, 2022.

[48] Florian Wenzel, Andrea Dittadi, Peter V. Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, Bernhard Schölkopf, and Francesco Locatello. Assaying out-of-distribution generalization in transfer learning. In *Neural Information Processing Systems*, 2022.

[49] Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. Id and ood performance are sometimes inversely correlated on real-world datasets, 2023.

[50] Weixin Liang, Yining Mao, Yongchan Kwon, Xinyu Yang, and James Zou. Accuracy on the curve: On the nonlinear correlation of ml performance between data subpopulations, 2023.

Supplementary Material

4.1 Experimental setup

This section provides the experimental details used throughout the paper, including distribution shifts, network architectures, and adaptation setups.

Distribution shifts. We test the models on 8 different distribution shifts that include synthetic corruptions and real-world shifts. Synthetic corruptions datasets, CIFAR10-C, CIFAR100-C, and ImageNet-C [24] 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 [5, 6, 7, 9, 10, 15, 12, 13, 16].

We also test on real-world shifts, which include CIFAR10.1 [25], CIFAR10.2 [26], ImageNetV2 [27], ImageNet-R [28], and FMoW-WILDS [4]. CIFAR10.1, CIFAR10.2, and ImageNetV2 are the reproduced datasets of their base counterparts by following the original dataset creation procedures. ImageNet-R is the variant of ImageNet which contains the images with renditions of various styles, such as paintings or cartoons. FMoW-WILDS [4] 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 [22], we use ID set consists of images taken from 2002 to 2013, and OOD set taken between 2013 and 2016.

Network architectures. We use a different set of the network architectures for specific datasets and their shifts in Sections 2 and 3.1. Specifically, for CIFAR10 and CIFAR100 vs. their OOD shifts, we use ResNet-18,26,34,50,101 [30], WideResNet-28-10 [31], MobileNetV2 [36], VGG11,13,16,19 [33]. For ImageNet and FMoW-WILDS vs. their OOD shifts, we leverage ResNet-18,34,50,101,152, WideResNet-50,101, DenseNet121 [35], ResNeXt50-32x4d [32]. As mentioned in Section 2.1, we use the pretrained model weights from torchvision, except for TTT [6] and ConjPL [15]. In addition, since TTT requires the rotation-prediction task during pretraining on source data, we train them ourselves using ResNet-14,26,32,50,104, and 152, which are available in their original implementation[2] and use them for Section 3.3. For Figures 1(third row), 4, and Tables 5, 2, and 3, we use the default network architecture, which is ResNet-26 for CIFAR10 and CIFAR100, and ResNet-50 for ImageNet and FMoW.

Optimizer and learning rates. We use SGD optimizer with momentum of 0.9 for all TTA baselines except for SAR, which uses sharpness-aware minimization (SAM) optimizer [40]. For Tables 1, 3, and Figures 1(third row) 4, we adapt the models adapted with different learning rates. Specifically, for Table 1, we evaluate the models with different architectures but adapted with the same learning rates, and then we average their estimation results across learning rates. For Table 3 and Figures 1(third row) 4, we evaluate the models (with same architecture) adapted with different learning rates, and observe the linear trends. For these experiments, we sweep over in the grid of $\{10^{-4}, 2 \cdot 10^{-4}, 5 \cdot 10^{-4}, 10^{-3}, 2 \cdot 10^{-3}, 5 \cdot 10^{-3}, 10^{-2}\}$ for CIFAR10, CIFAR100, and their OOD shifts, while using $\{5 \cdot 10^{-5}, 10^{-4}, 2 \cdot 10^{-4}, 5 \cdot 10^{-4}, 10^{-3}, 2 \cdot 10^{-3}, 5 \cdot 10^{-3}\}$ for ImageNet and their OOD counterparts. For Figures 1 2, 3 and Tables 5, 2, we use the fixed learning rates of 10^{-3} for CIFAR10, CIFAR100 and their OOD counterparts, and $2.5 \cdot 10^{-4}$ for ImageNet, FMoW and their counterparts.

Batch sizes. In Figures 1(third row) 4 and Table 3, we test a wide spectrum of different values of batch sizes, by sweeping over the grid of $\{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$. For the rest of the experiments, we fix the batch size of 128 for TTA on CIFAR10, CIFAR100 and their OOD, and 64 on ImageNet, FMoW for their OOD. Note that for TTT, we use a single batch following the original paper setup.

Number of adapt steps. In Figures 1(third row) 4 and Table 3, we sweep through one to five adaptation steps for every TTA baseline we use. For the rest of the experiments, we use a single step of the adaptations. We utilize the online adaptation (*i.e.*, no initialization for each batch) strategy for all baselines.

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

4.2 Additional results on strong AGL and ACL trends after TTA

In this section, we supplement the main text by providing additional results of TTAed models exhibiting strong AGL and ACL linear trends across datasets and TTA methods. Specifically, in Figures 2 we provide the results of SHOT, TENT, and ETA comparing with that of vanilla, across CIFAR10 and ImageNet against their Gaussian Noise corruptions with highest severity. The axes are probit-scaled, and each blue and pink dot represent the accuracy and agreement of the model, respectively.

In Figure 3, we add the results on AGL and ACL trends of TTA methods' results on the real-world shifts. We consistently observe that AGL and ACL consistently persist or even become stronger than vanilla models, when applying various TTA methods across distribution shifts, evidenced by the high R^2 values in all cases.

In Figures 1(third row) 4, we observe strong AGL and ACL among models adapted with varying values of TTA setups, including learning rates, the number of adapt steps, batch sizes, and checkpoints of the source-trained model. We plot the results of TENT-adapted models on CIFAR10-C Gaussian Noise in first row, and ETA-adapted models on ImageNet-C Gaussian noise in second row. Each point with different colors denote models using different values in each hyperparameter.

We also provide the results of SHOT, TENT, ETA, and SAR comparing with that of vanilla, across every corruption types of CIFAR10-C with highest severity, in Figures 5 and 6. In Figure 7, we add the results on CIFAR10.2 along with other TTA methods' results on the real-world shifts. We consistently observe that AGL and ACL consistently persist or even become stronger than vanilla models, when applying various TTA methods across distribution shifts, evidenced by the high R^2 values in all cases.

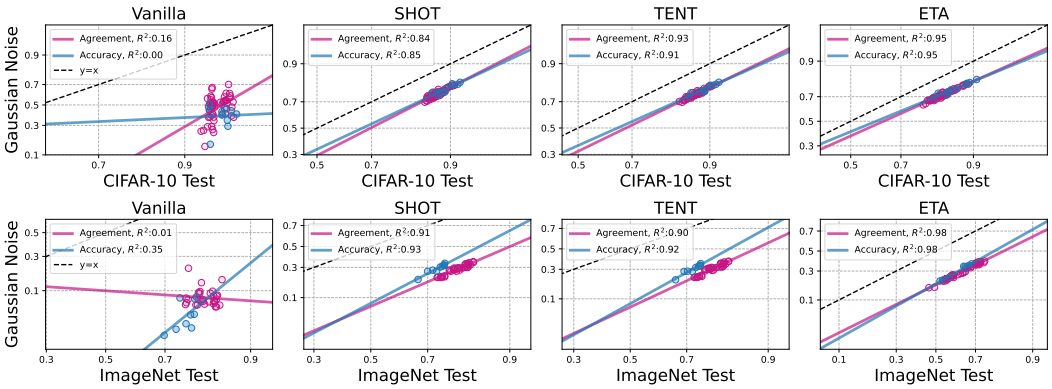

Figure 2: Various TTA methods make the correlations in agreement and accuracy notably stronger than their base counterparts, with their R^2 values substantially increased. We test SHOT, TENT, and ETA on CIFAR10 (first row) and ImageNet (second row) against their Gaussian Noise corruption.

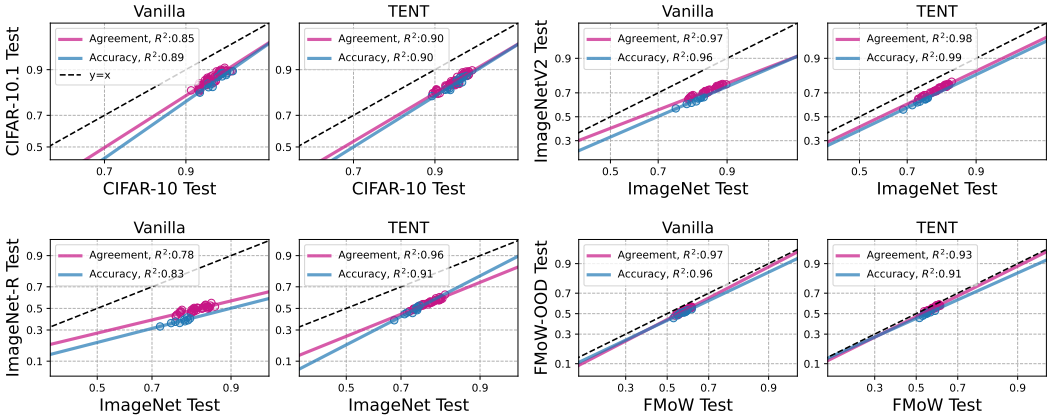

Figure 3: Across real-world shifts, such as CIFAR10.1, ImageNetV2, ImageNet-R, and FMoW-WILDS, the models after TTA maintain their strong linear trends. Notably, this observation also holds true despite possible accuracy degradation, e.g. on CIFAR10.1 and ImageNetV2 (See Table 5).

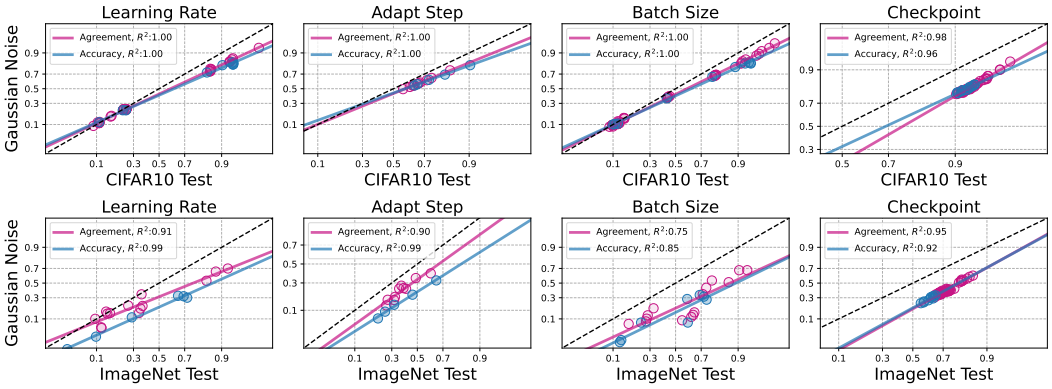

Figure 4: We observe strong AGL and ACL among models adapted with varying values of TTA setups, including learning rates, the number of adapt steps, batch sizes, and checkpoints of the source-trained model. We plot the results of TENT-adapted models on CIFAR10-C Gaussian Noise in first row, and ETA-adapted models on ImageNet-C Gaussian noise in second row. Each point with different colors denote models using different values in each hyperparameter.

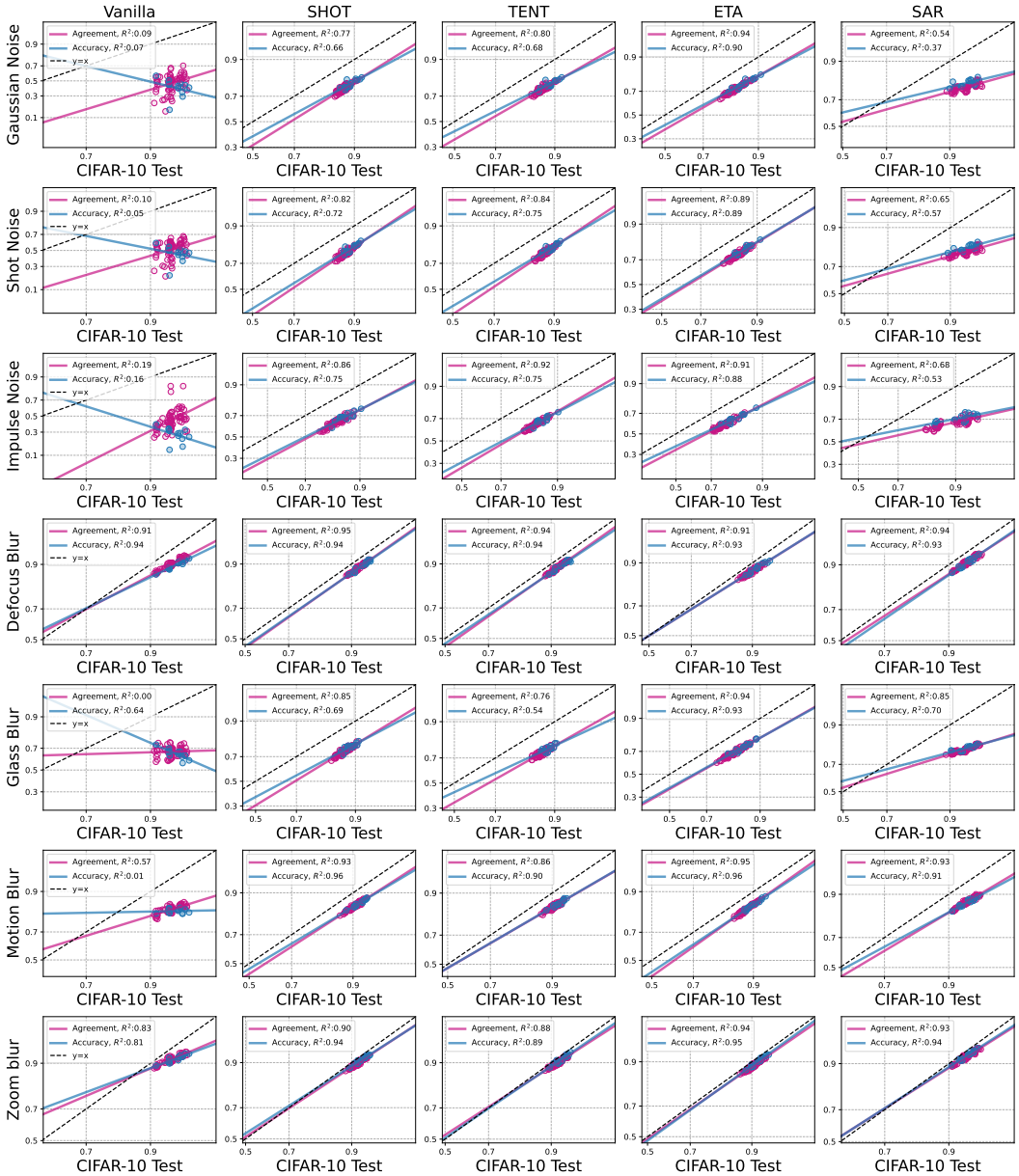

Figure 5: Additional results across every corruption types of CIFAR10-C, where models after TTA show strong AGL and ACL than vanilla models.

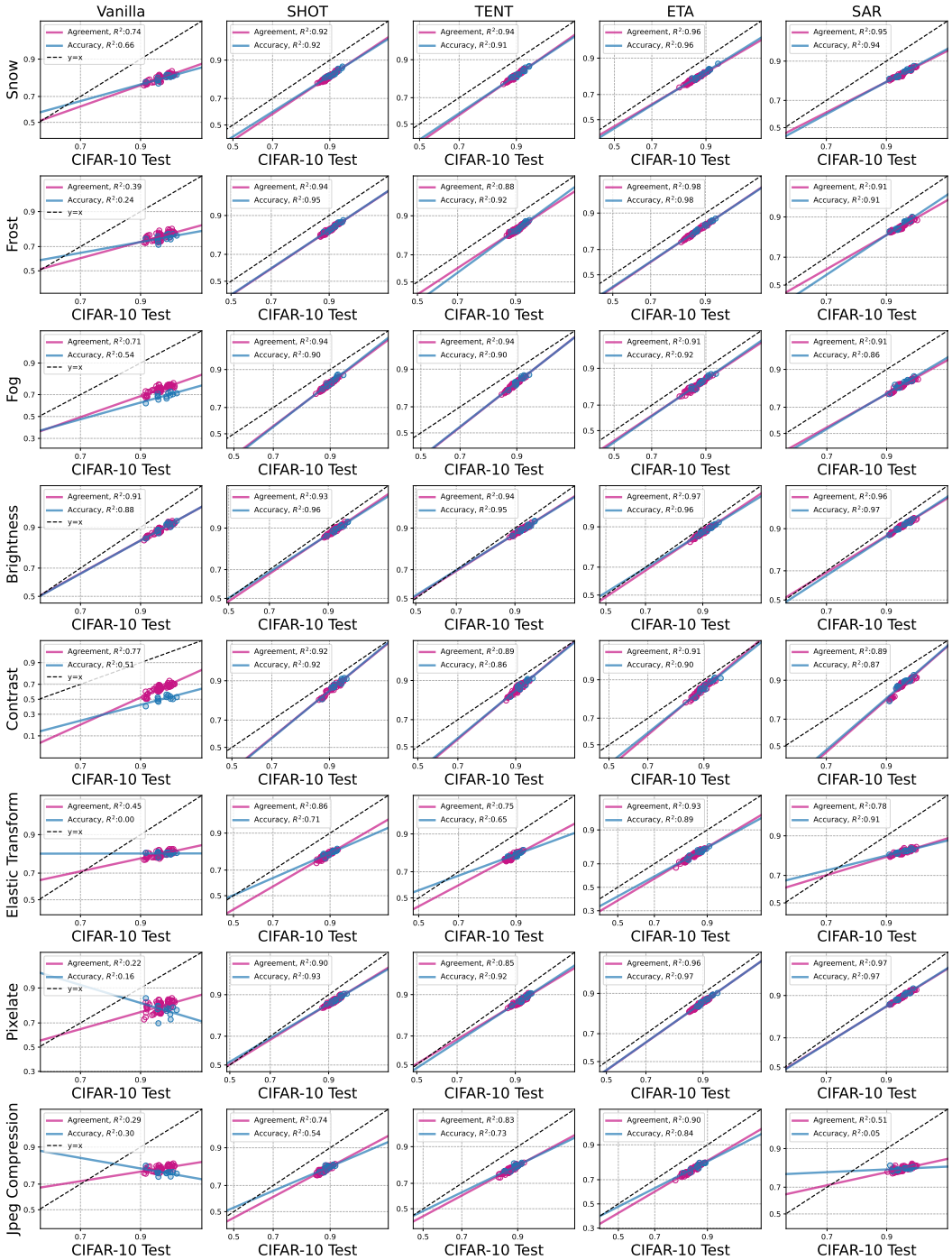

Figure 6: Additional results across every corruption types of CIFAR10-C, where models after TTA show strong AGL and ACL than vanilla models.

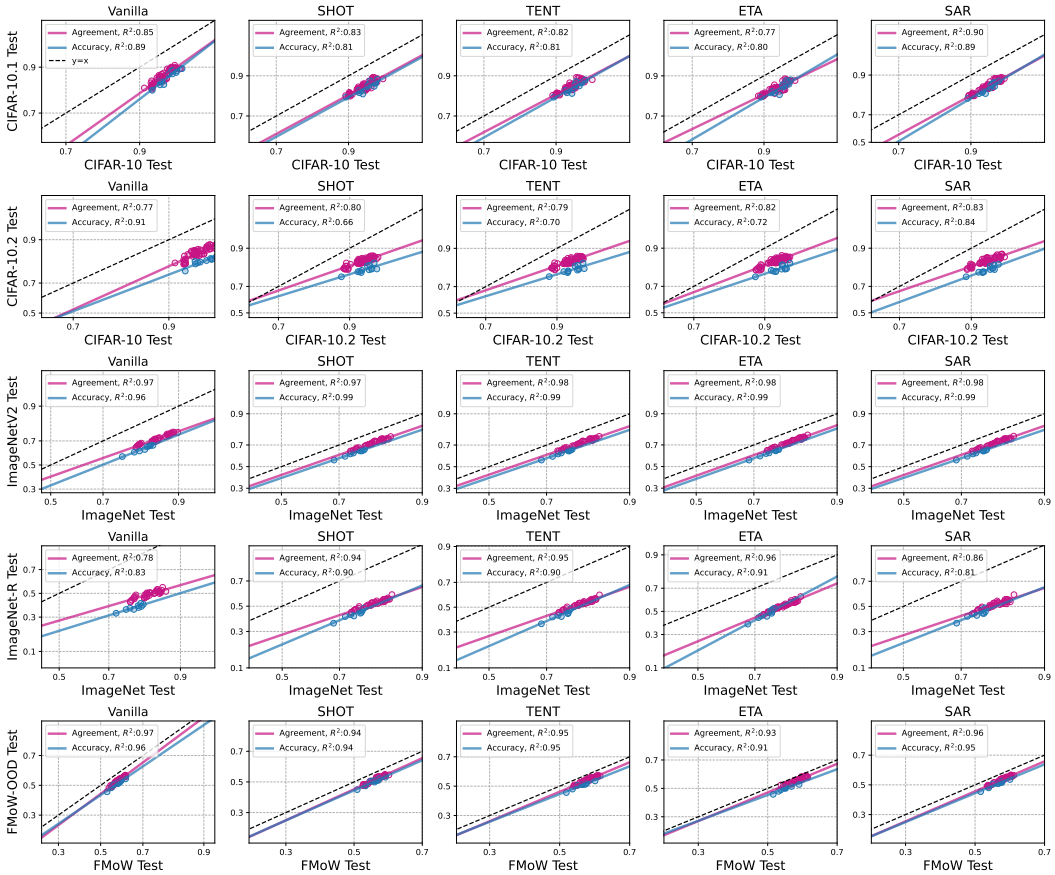

Figure 7: Additional TTA baselines' results under real-world shifts, CIFAR10.1, CIFAR10.2, ImageNetV2, ImageNet-R, and FMoW-WILDS, where models after TTA persist strong ACL and AGL.

Dataset	Method	ATC	DOC-feat	AC	Agreement	ALine-S	ALine-D
CIFAR100-C	Vanilla	5.20	10.42	14.35	15.07	8.37	8.20
	SHOT	4.49	6.57	18.34	4.87	0.84	**0.69**
	TENT	19.17	6.53	21.96	5.16	0.82	**0.67**
	ETA	25.40	4.10	28.32	6.18	**0.70**	1.07
	SAR	1.85	3.48	16.68	4.54	0.90	**0.77**
ImageNet-R	Vanilla	1.72	14.31	17.74	21.56	7.62	7.62
	SHOT	5.29	12.91	17.85	17.92	3.22	**3.22**
	TENT	9.14	14.94	24.10	18.30	3.44	**3.44**
	ETA	11.71	12.61	34.12	17.49	2.33	**2.17**
	SAR	3.78	9.44	13.49	16.11	2.82	**2.44**

Table 4: Mean absolute error (MAE) (%) of the accuracy estimation on TTAed models with different architectures, tested on CIFAR100-C and ImageNet-R. ALine-S/D on TTAed models leads to substantially lower errors in accuracy estimation, compared to ALine-S/D applied on the vanilla models as well as other estimation baselines.

Dataset	SHOT		TENT		ETA		SAR	
	GT	Est.	GT	Est.	GT	Est.	GT	Est.
CIFAR10C-Snow	+3.94	+4.32	+4.09	+4.45	+4.51	+4.69	+2.54	+2.80
CIFAR100C-Bright	+6.59	+8.28	+6.79	+8.24	7.55	+8.72	+7.34	+8.29
ImageNetC-Gauss	+24.53	+13.45	+25.79	+14.64	+30.25	+22.82	+31.28	+22.19
ImageNet-R	+6.94	+2.70	+6.19	+2.52	+10.35	+5.30	+8.9	+2.44
CIFAR10.1	-2.30	-1.51	-2.10	-1.68	-2.25	-1.33	-2.30	-1.56
CIFAR10.2	-1.70	-1.64	-1.80	-1.65	-1.40	-0.66	-1.90	-1.44
ImageNetV2	-0.27	-2.50	-1.18	-3.42	-10.34	-12.60	-0.21	-2.40
FMoW-WILDS	-0.22	-0.73	+0.37	+1.03	+0.29	+0.70	+0.62	+0.85

Table 5: Actual (GT) and estimated (Est.) improvement/degradation (%) in OOD accuracy after applying each TTA method w.r.t their base counterparts. The values with green indicate the improvement, whiled red is the degradation. Our estimations consistently have the same predictions (*i.e.*, colors) on whether TTA methods enhance or diminish accuracy across distribution shifts, enabling to forecast their generalizations without labeled data.

4.3 Additional results on accuracy estimation of TTA

To supplement the accuracy estimation results in the main text, we additionally provide the estimation results of TTA on CIFAR100-C and ImageNet-R in Table 4. Furthermore, we extend to various datasets that include common corruptions in CIFAR10-C, CIFAR100-C, ImageNet-C, and real-world shifts, such as ImageNet-R, CIFAR10.1, CIFAR10.2, ImageNetV2, and FMoW-WILDS. In Table 5, we present actual and estimated OOD accuracy improvement or degradation by each TTA baseline, including SHOT, TENT, ETA, and SAR, with respect to their base counterparts. Specifically, we first obtain the actual (by using ground-truth labels) and estimated OOD accuracy of both TTAed models and vanilla models, and calculate the differences between them.

We initially observe that TTA methods sometimes fail to enhance generalization under real-world shifts, *e.g.*, CIFAR10.1, CIFAR10.2 and ImageNetV2, as shown as red in columns of "GT", constrated with their effectiveness in synthetic datasets, denoted as green. Interestingly, the overall trend of such relative accuracy changes after TTA, represented by red or green, in estimated OOD accuracy in the "Est." columns closely mirrors that of GT columns. For instance, the generalization trends of different TTA methods on FMoW-WILDS precisely align with those revealed in the estimated results, *i.e.*, all TTA baselines except SHOT improve accuracy. This underscores the potential of our accuracy estimation on TTA, which offers practical guidance for selecting or determining the suitability of TTA methods in the face of distribution shifts from the wild, with no labeled data.

Given the strong AGL and ACL trends that persist under real-world shifts (Figure 3), we demonstrate the estimation results of the TTAed models' OOD accuracy by comparing them with the ground-truth accuracy on such shifts. In Figure 8, we compare the estimated and ground-truth OOD accuracy of vanilla and the adapted models using TENT, ETA, and SAR. It shows that their accuracies can be closely estimated to the actual accuracies (*i.e.*, closer to $y = x$) on CIFAR10.1, ImageNetV2, ImageNet-R, and FMoW-WILDS. In particular, as shown in ImageNetV2 and ImageNet-R, the estimation results of TTAed models are located closer to $y = x$ than those of vanilla, indicating TTAed models' accurate estimation performances compared to those of vanilla models.

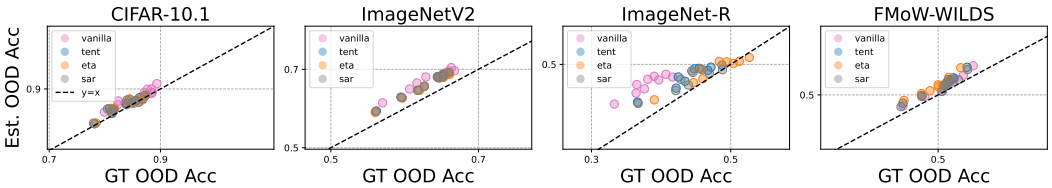

Figure 8: Comparison of GT vs. estimated OOD accuracy of vanilla and TTAed models under real-world shifts. Each pink dot represents vanilla, blue the TENT, orange the ETA, and gray the SAR results. The dotted $y = x$ line denotes the *perfect* estimation line, where the closer dots are located to the line, the more accurate the estimations are.

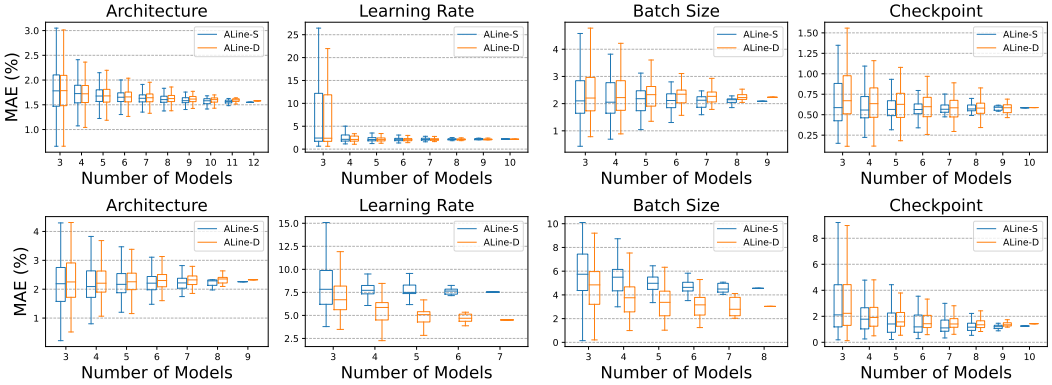

Figure 9: Distribution of MAE of accuracy estimation over different number of models. We test models adapted with different TTA setups. The results on first row are those adapted by TENT on CIFAR10-C Gaussian Noise, and those on second row are adapted by ETA on ImageNet-C Gaussian Noise.

4.4 Analysis on the number of TTAed models for accurate estimation.

In this section, we investigate how the accuracy estimation of test-time adapted model's performances change as we vary the number of the models used in estimation. To this end, we examine the models with different architectures as well as the hyperparameters examined in Section 3.3, which include learning rates, batch sizes, and the checkpoints of the source-trained models. We vary the size of models, n, and for each size, we calculate all possible sets of models' estimation error.

Figure 9 illustrates the changes of the distributions of the estimation error (MAE) with respect to the number of models used for implementing AGL and ACL. We test TENT on CIFAR10 test vs. CIFAR10-C Gaussian Noise, and ETA on ImageNet test vs. ImageNet-C Gaussian Noise. We observe that even using the minimum number of models for estimation, which is three, the estimation results achieve the low MAE, particularly when using different architectures or the checkpoints of the source-trained models. In addition, we note the rapid decrease in MAE is easily attainable by adding only a small number of additional models, which is pronounced when using learning rates or batch size (in ImageNet-C). These results indicate that during TTA, accurate estimation of OOD accuracy can be achieved with a reasonable number of models, each with different hyperparameter values, thereby offering practical feasibility during testing.

4.5 OOD Accuracy estimation methods

ALine-S and ALine-D Baek *et al.* [1] 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.

Average thresholded confidence (ATC) Garg *et al.* [41] 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

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 $\boldsymbol{A} \in \mathbb{R}^{\frac{n(n-1)}{2} \times n}, \boldsymbol{b} = \mathbb{R}^{\frac{n(n-1)}{2}}$
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**
  $\boldsymbol{A}_{ij} = \frac{1}{2}, \boldsymbol{A}_{ik} = \frac{1}{2}, \boldsymbol{A}_{il} = 0 \forall l \notin j, k$
  $\boldsymbol{b}_i = \Phi^{-1}(\mathsf{Agr}(p_{j,\text{OOD}}, p_{k,\text{OOD}})) + \hat{a} \cdot \left(\frac{\Phi^{-1}(\mathsf{Acc}(p_{j,\text{ID}}, y_{\text{ID}}) + \Phi^{-1}(\mathsf{Acc}(p_{k,\text{ID}}, y_{\text{ID}}))}{2} - \Phi^{-1}(\mathsf{Agr}(p_{j,\text{ID}}, p_{k,\text{ID}})) \right)$
  i=i+1
end for
$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w} \in \mathbb{R}^n} ||\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}||_2^2$
$\widehat{\mathsf{Acc}}^{\text{D}}_{\text{OOD}} = \Phi(\boldsymbol{w}_i^*) \forall i \in [n]$ ▷ ALine-D
return $\widehat{\mathsf{Acc}}^{\text{S}}_{\text{OOD}}, \widehat{\mathsf{Acc}}^{\text{D}}_{\text{OOD}}$

---

classifying among c classes. The method can be written as below:

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

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

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

Difference of confidence (DOC)-feat Guillory *et al.* [42] 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,

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

Average confidence (AC) Hendrycks *et al.* [43] estimate the OOD accuracy based on model's averaged confidence, which can be written as

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

Agreement Jiang *et al.* [44] 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 $\mathcal{X}_{\text{OOD}}$.

5 Related Work

Test-time adaptation and its pitfalls of reliability Test-time adaptation (TTA) enhances model robustness by adapting models to unlabeled test data. One research direction uses self-supervision tasks during both training and testing [6, 7, 8], while another explores "fully" test-time adaptations that require no specific pretraining procedures, relying on objectives such as entropy minimization [10, 12, 13], data augmentation invariance [11, 18], and self-training with pseudo-labels [14, 15, 45].

Some studies have extended their evaluation beyond corruptions to include more challenging shifts, such as datasets reproductions [7, 11], domain generalization benchmarks [21, 16], and WILDS [14]. Yet, as pointed out in [16], TTA methods may not effectively address the full spectrum of distribution shifts in the wild. Another persistent issue is that TTA, especially those based on entropy minimization, can lead to overconfident predictions. Specifically, while [38] showed that using test-time batch statistics enhances calibration under distribution shifts [46], several work [14, 18] observed that entropy minimization diminishes this effect. Moreover, the performance of TTA methods can be severely impacted by the negligent selection of hyperparameters such as learning rates [19, 16],

adaptation steps [16], or batch sizes [13, 47]. A recent study [19] addresses this issue by adapting only the model's outputs instead of its parameters. Our study leverages the remarkable observation of the *agreement-on-the-line phenomena within TTA*, offering promising solutions to these reliability issues.

Accuracy and agreement-on-the-line As mentioned above, the basic accuracy-on-the-line and agreement-on-the-line phenomena were first identified in [22] and [1] respectively. However, the underlying reason for such phenomenon remain unclear, and certain datasets reveal weak accuracy and agreement-on-the-line trends, as observed in CIFAR10 test vs. Gaussian Noise corruption in CIFAR10-C. Recent work [48, 49, 50] have investigated the broader types of distribution shifts, identifying shifts that show different trends beyond the linear correlations. In this study, we investigate the impact of model adaptation on reinforcing (or maintaining) linear correlations, and identify novel conditions that lead to such trends–adapting with varying TTA hyperparameters.

5.1 Limitations

Even though we consistently observe the linear trends across various models and distribution shifts during TTA, we also find an exceptional case where such trends do not manifest. As shown in Figure 10 right, we empirically observe that varying learning rates in TTT [6] exhibits that the ID and OOD accuracies are negatively correlated, leading to a complete misalignment with the agreement line. While in other experimental setups, such as using different architectures, we still observe the strong AGL and ACL among the models, as shown in Figure 10 left. Such negative correlation results in the accuracy of the models selected using our method, *i.e.*, best-performing ID accuracy, suffer significant deviation from the best-performing OOD accuracy, as shown in Table 3.

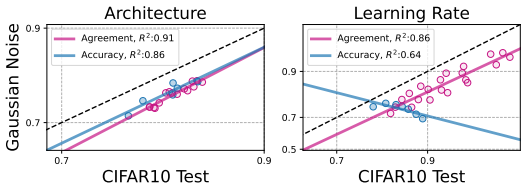

Figure 10: Compared to the adapted models with varying architectures, using learning rates in TTT [6] shows negative correlations in accuracies, resulting in misalignment between agreement and accuracy lines.