TEA: Test-time Energy Adaptation

Supplementary Material

6. Appendix Summary

The appendix contains the following sections:

(1) Additional Experiments and Analyses (Sec. 7):
- Detailed Results for Energy Reduction (Sec. 7.1)
- Detailed Results for Image Corruption (Sec. 7.2)
- Hyper-parameters Sensitivity (Sec. 7.3)

(2) Detailed Settings (Sec. 8):
- Datasets(Sec. 8.1)
- Evaluation Metrics(Sec. 8.2)
- Hyper-parameters(Sec. 8.3)
- Computing Resources (Sec. 8.4)

(3) Limitations and Future Explorations (Sec. 9).

7. Additional Experiments

7.1. Detailed Results for Energy Reduction

This section serves as an extension of the energy analysis (Sec. 4.3.1) in the main text, presenting the relationship between TEA's energy reduction and the enhancement of generalizability across all types of corruption. The detailed results are shown in Figs. 7 and 8, where each corruption type is analyzed at five levels of severity, with the analysis examining the correlation between the extent of energy reduction and performance improvements, both before and after adaptation, as severity levels increase.

In our experiments, TEA generally reduced energy and enhanced generalization across various corruptions. Yet, for mild corruptions like "Brightness" at level one, i.e., the mildest in CIFAR-10-C, generalization did not improve and occasionally deteriorated slightly. Correspondingly, energy did not decrease and even increased marginally. These outcomes indicate a strong correlation between generalizability enhancement and energy reduction. However, it is possible that our method may not reduce energy as anticipated for distributions with some less severe corruption types. This may be attributed to these distributions being closely aligned with the original, already at a low energy state. The uniform hyperparameters used in our adaptation may not be optimal for such cases. Addressing this discrepancy will be a priority in future research.

7.2. Detailed Results for Image Corruption

This section serves as an extension of the main adaption results (Sec. 4.2) in the main text, presenting the detailed performance for each corruption type at the most severe corruption level. The detailed results are shown in Tab. 7. In our evaluation, TEA consistently achieves the highest accuracy for every corruption type on CIFAR-10-C and CIFAR-100-C datasets. On TinyImageNet, our model exhibits superior performance on the majority of corruptions. However, it is slightly outperformed by SHOT on a few corruption types. The performance difference might be because the corruptions are mild and similar to the source data, which benefits pseudo-label methods like SHOT that rely on this similarity to produce accurate labels.

7.3. Hyper-parameters Sensitivity

This section provide a new experiments on hyper-parameters sensitivity of our proposed TEA. The main hyper-parameters for TEA are the step and learning rate for Stochastic Gradient Langevin Dynamics (SGLD). Fig. 9 illustrates the variation in model accuracy as the SGLD learning rate is incrementally adjusted from 0.001 to 0.4, while Fig. 10 demonstrates the impact on accuracy when the SGLD step is increased from 1 to 200. The results reveal that the performance of TEA is consistently state-of-the-art under a wide range of hyper-parameters choices, across all types of corruption on CIFAR-10-C.

8. Detailed Settings

8.1. Datasets

We perform experiments on four datasets across two tasks. Image corruption task include CIFAR-10(C), CIFAR-100(C), and TinyImageNet(C) datasets. Domain generalization task include PACS datasets.

Dataset of Clean Distribution Clean distribution of CIFAR-10, CIFAR-100 [29] and TinyImageNet [31] are datasets of clean distribution. CIFAR-10 and CIFAR-100 datasets consist of 60,000 color images, each of size 3x32x32 pixels. CIFAR-10 is categorized into 10 distinct classes with 6000 images per class. CIFAR-100 is more challenging, as these images are distributed across 100 classes, with 600 images per class. TinyImageNet datasets consist of 110,000 color images, each of size 3x64x64 pixels, which are categorized into 200 distinct classes with 550 images per class. Both CIFAR-10 and CIFAR-100 are subdivided into a training set of 50,000 images and a test set of 10,000 images. TinyImageNet is subdivided into a training set of 100,000 images and a test set of 10,000 images.

Dataset of Corrupted Distributions CIFAR-10-C, CIFAR-100-C and TinyImageNet-C[17] are variants of the original CIFAR-10, CIFAR-100 and TinyImageNet datasets that have been artificially corrupted into 19 types

Table 4. Summary of Clean & Corruption Datasets

Dataset	#Train	#Test	#Corr.	#Severity	#Class.
CIFAR-10	50,000	10,000	1	1	10
CIFAR-100	50,000	10,000	1	1	100
TinyImageNet	100,000	10,000	1	1	200
CIFAR-10-C	-	950,000	15	5	10
CIFAR-100-C	-	950,000	15	5	100
TinyImageNet-C	-	750,000	15	5	200

Table 5. Summary of PACS Datasets

Domain	#Sample	#Class	Size
Photo	1,670	7	3x227x227
Art	2,048	7	3x227x227
Cartoon	2,344	7	3x227x227
Sketch	3,929	7	3x227x227

of corruptions at five levels of severity, resulting in 95 corrupted versions of the original test set images. The corruptions include 15 main corruptions: Gaussian noise, shot noise, impulse noise, defocus blur, glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic, pixelation, and JPEG. All these corruptions are simulations of shifted distributions that models might encounter in real-world situations.

Datsset of PACS PACS[35] is an image dataset popular used in transfer learning, which consist of four domains, namely Photo (1,670 images), Art Painting (2,048 images), Cartoon (2,344 images) and Sketch (3,929 images). Each domain contains seven categories.

8.2. Evaluation Metrics

For evaluation on corruption datasets, we employ Average Accuracy and Mean Corruption Error (mCE) [17] as evaluation metrics. For clean and PACS datasets, we employ Accuracy as evaluation metric. These metrics provide a comprehensive evaluation of a model's generalization in handling diverse distributions, thereby offering a multi-faceted perspective on model performance.

Average Accuracy Average Acc is the accuracy averaged over all severity levels and corruptions. Consider there are a total of C corruptions, each with S severities. For a model f, let $\mathcal{E}_{s,c}(f)$ denote the top-1 error rate on the corruption c with severity level s averaged over the whole test set,

$$\mathrm{AverAcc}_f = 1 - \frac{1}{C \cdot S} \sum_{c=1}^{C} \sum_{s=1}^{S} \mathcal{E}_{s,c}(f). \tag{9}$$

Mean Corruption Error mCE is a metric used to measure the performance improvement of model f compared to a baseline model f_0. We use the model without adaptation as the baseline model,

Table 6. Summary of Hyper-parameters

Data	Common				TEA-SGLD		
	Step	LR	BS	Optim	Step	LR	Std
CIFAR-10-C	1	0.001	200	Adam	20	0.1	0.01
CIFAR-100-C	1	0.001	500	Adam	20	0.1	0.01
TinyImageNet-C	1	0.0001	1000	Adam	20	0.1	0.01
PACS-P	10	0.001	full	Adam	20	0.1	0.01
PACS-A	10	0.001	full	Adam	20	0.1	0.01
PACS-C	10	0.002	full	Adam	20	0.1	0.01
PACS-S	20	0.002	full	Adam	20	0.1	0.01

$$\mathrm{mCE}_f = \frac{1}{C} \sum_{c=1}^{C} \frac{\sum_{s=1}^{S} \mathcal{E}_{c,s}(f)}{\sum_{s=1}^{S} \mathcal{E}_{c,s}(f_0)} \tag{10}$$

8.3. Hyper-parameters

This section outlines the hyper-parameters chosen for our experiments. These settings enable the reproducibility of the results presented in our study. For common hyperparameters, we align with those used in Tent [60]. For TEA-specific hyper-parameters, we adjust them following the parameter choices from JEM [13].

8.4. Computing resources

All our experiments are performed on RedHat server (4.8.5-39) with Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz4, 4× NVIDIA Tesla V100 SXM2 (32GB) and 3× NVIDIA Tesla A800 SXM4 (80GB).

9. Limitation and Future Works

Our study has identified key aspects for improvement and future research, which are outlined below: (1) The use of Stochastic Gradient Langevin Dynamics sampling is both time-consuming and unstable. However, ongoing research in energy-based models is addressing these issues through various methods, such as gradient clipping [66], diffusion process [40], additional gradient term [8] and ordinary differential equation based sampling [42]. One of our future directions is to enhance TEA by incorporating these advanced sampling techniques. (2) Overemphasizing the model's sensitivity to the data distribution may significantly impact its discriminative ability. This trade-off between transferability and discriminability is a common theme in TTA research [11, 30]. Another direction for our future work is to explore how to enhance the model's perception of data distribution while maintaining or even improving its discriminative power. We acknowledgethat the limitations identified may present challenges. Nevertheless, we remain confident that our study represents a pioneering effort to integrate energy-based training into test time adaptation. We believe that any future advancements in the training of energy-based models will likely enhance and refine the outcomes we have demonstrated in our research.

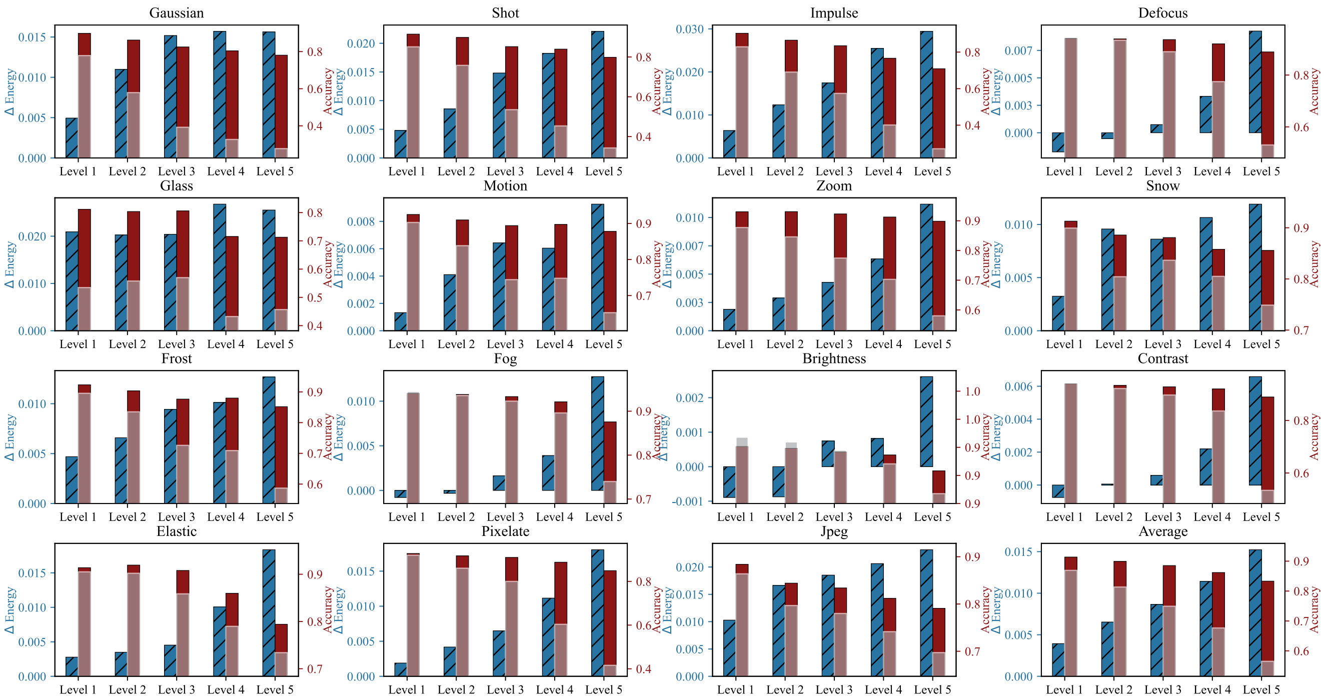

Figure 7. The relationship between TEA's energy reduction and the enhancement of generalizability on CIFAR-10-C, under different types of distribution and different severity level of distribution shifts. Each subfigure plots corruption severity level on the x-axis, energy reduction on the left y-axis, and accuracy on the right y-axis. The accuracy axis contains two bars: the red bar denotes our TEA' accuracy, while the transparent bar denotes baseline's accuracy.

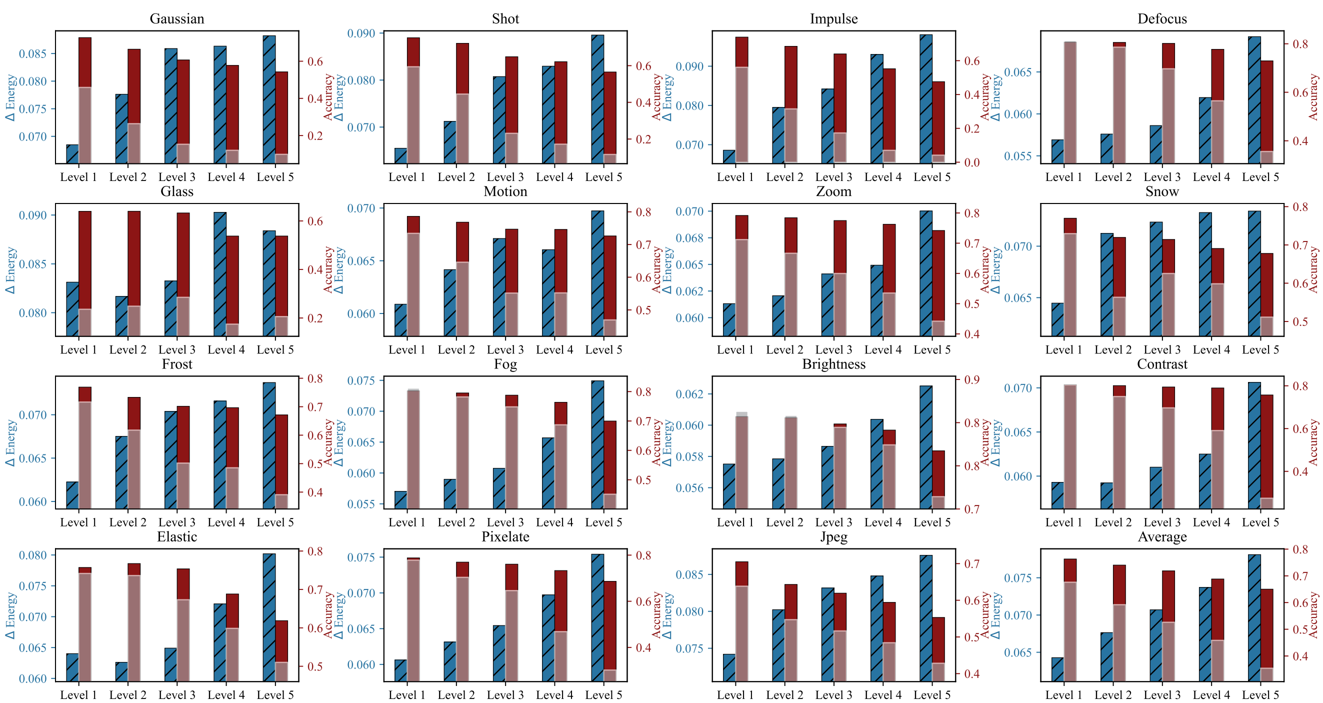

Figure 8. The relationship between TEA's energy reduction and the enhancement of generalizability on CIFAR-100-C, under different types of distribution and different severity level of distribution shifts. Each subfigure plots corruption severity level on the x-axis, energy reduction on the left y-axis, and accuracy on the right y-axis. The accuracy axis contains two bars: the red bar denotes our TEA' accuracy, while the transparent bar denotes baseline's accuracy.

Table 7. Comprehensive comparison of TEA and various baseline models on CIFAR-10-C, CIFAR-100-C and Tiny-ImageNet-C. All evaluated models employ the architecture of WRN-28-10 with BatchNrom. Model weights are sourced from RobustBench for CIFAR-10-C to ensure a fair comparison. Evaluations are based on Accuracy (%) for each individual corruption, as well as Average Accuracy (Acc %), Mean Corruption Error (mCE %) for overall performance. The reported performance of our TEA reflects the average across five runs with varying seeds, with a maximum standard deviation under 0.1%. The most notable results are indicated with **boldface** for the top performance.

	Method	Noise			Blur				Weather			Digital					Avg	
		Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	JPEG	Acc(↑)	mCE(↓)
CIFAR-10(-C)	Source	27.68	34.25	27.07	53.01	45.67	65.24	57.99	74.87	58.68	73.98	90.70	53.37	73.39	41.56	69.71	56.47	100.00
	BN	71.93	73.88	63.74	87.19	64.72	85.83	87.89	82.73	82.61	84.75	91.61	87.35	76.25	80.33	72.70	79.56	52.65
	DUA*	72.60	75.40	64.70	86.90	65.10	85.40	88.40	83.20	82.50	86.90	**92.40**	85.90	77.30	80.70	73.80	80.10	50.78
	PL	14.46	17.77	14.48	44.27	31.29	66.70	60.49	76.46	55.01	79.35	91.05	43.49	75.36	31.61	69.63	51.42	106.98
	SHOT	62.91	66.15	46.41	82.85	61.64	82.39	83.16	79.03	78.24	82.35	89.68	82.20	76.08	75.34	73.19	74.77	63.19
	TENT	75.20	76.52	67.03	88.00	68.22	86.27	89.24	84.06	83.81	86.30	92.16	87.91	78.01	82.77	75.77	81.41	48.13
	ETA	72.21	73.88	63.72	87.21	64.70	85.84	87.89	82.73	82.58	84.77	91.60	87.32	76.22	80.37	72.69	79.58	52.64
	EATA	72.25	73.88	63.74	87.19	64.72	85.83	87.89	82.73	82.61	84.75	91.61	87.35	76.25	80.33	72.70	79.59	52.62
	SAR	71.95	74.14	64.11	87.39	65.20	86.00	88.06	83.08	82.66	85.07	91.90	87.20	76.69	80.41	72.79	79.77	51.94
	TEA	**78.33**	**79.87**	**70.94**	**88.89**	**71.31**	**87.87**	**89.77**	**85.56**	**85.29**	**87.61**	92.37	**88.98**	**79.32**	**84.90**	**78.99**	**83.33**	**43.69**
CIFAR-100(-C)	Source	9.87	11.58	4.15	35.57	20.56	46.92	44.20	51.13	39.07	45.07	71.42	27.43	51.01	30.19	42.82	35.39	100.00
	BN	46.53	48.62	37.15	70.94	47.36	69.04	71.25	63.00	62.96	66.08	75.89	71.31	58.79	64.56	47.46	60.06	63.54
	PL	29.54	34.09	14.99	64.06	40.81	65.36	67.74	60.42	59.47	62.92	75.98	57.29	58.72	59.41	50.34	53.40	72.12
	SHOT	37.50	39.68	21.27	67.90	45.52	67.42	69.86	61.42	60.75	65.01	75.12	63.96	58.96	62.54	51.09	56.53	68.01
	TENT	53.44	54.12	45.53	71.69	51.17	71.54	71.63	64.88	65.30	68.41	75.14	73.59	59.25	66.81	53.93	63.09	59.42
	ETA	48.64	50.95	38.05	69.66	47.52	67.76	70.24	62.51	61.73	66.03	73.40	71.15	56.93	64.40	48.34	59.82	64.52
	EATA	48.76	51.60	39.47	69.29	47.49	68.13	70.68	62.94	62.53	65.14	74.48	71.61	57.53	64.24	49.67	60.24	63.75
	SAR	51.34	54.08	44.62	72.24	50.10	71.06	72.43	64.96	65.35	68.40	76.23	73.95	60.04	67.26	52.29	62.95	59.37
	TEA	**54.29**	**56.55**	**48.59**	**72.96**	**53.78**	**72.63**	**74.20**	**67.78**	**67.14**	**69.98**	**76.74**	**75.71**	**62.18**	**68.65**	**55.32**	**65.10**	**56.07**
TinyImageNet(-C)	Source	12.37	16.07	5.55	6.63	5.67	20.71	19.38	26.79	31.88	14.18	29.75	1.95	32.03	49.00	46.19	21.21	100.00
	BN	24.00	25.82	18.22	28.43	18.46	35.20	33.41	27.59	30.26	25.58	32.58	6.04	34.12	40.81	35.72	27.74	93.42
	PL	20.75	25.99	11.41	17.28	14.42	36.07	33.66	30.19	35.01	22.44	34.87	2.03	**41.31**	51.76	46.75	28.26	91.22
	SHOT	22.25	27.10	14.54	18.74	15.17	36.78	34.63	30.95	**35.68**	23.52	36.01	1.91	41.01	**51.99**	**46.85**	29.14	90.16
	TENT	23.80	25.29	18.68	27.05	17.54	33.21	31.36	26.21	28.56	24.76	30.08	5.68	31.37	37.77	33.33	26.31	95.52
	ETA	24.44	25.76	18.00	28.05	18.25	34.59	32.85	27.11	29.87	25.42	31.86	5.93	32.84	39.55	34.82	27.28	94.12
	EATA	24.48	25.55	17.95	27.68	18.41	34.83	32.84	27.03	29.67	25.39	31.94	5.87	32.71	39.87	35.09	27.28	94.09
	SAR	25.10	26.33	18.87	28.80	18.44	35.59	33.87	28.31	30.62	26.32	33.31	6.30	34.24	41.16	35.97	28.21	92.82
	TEA	**25.50**	**29.35**	**20.58**	**32.76**	**21.14**	**41.72**	**39.06**	**31.43**	34.19	**29.02**	**36.83**	**6.89**	38.94	46.38	41.31	**31.67**	**87.99**

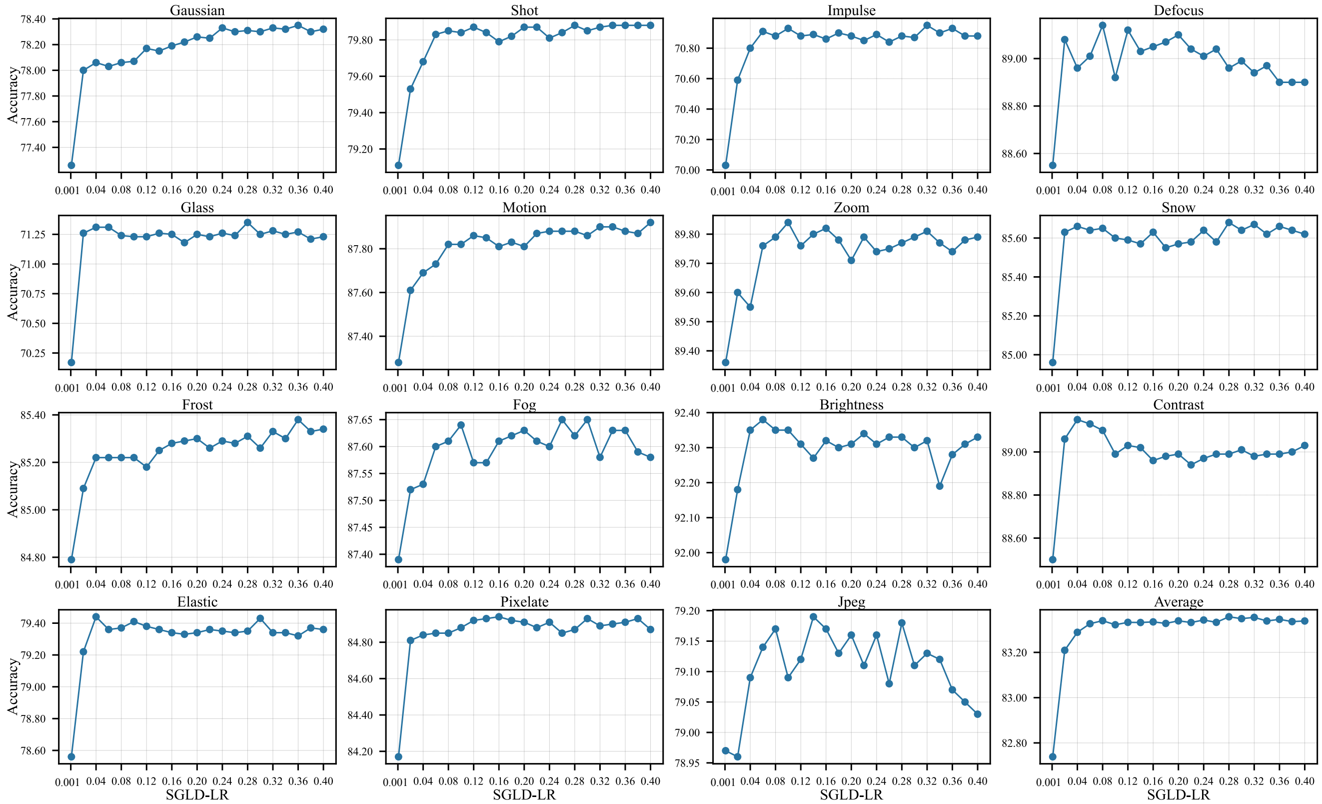

Figure 9. Hyper-parameter stability with respect to the Stochastic Gradient Langevin Dynamics (SGLD) learning rate. The x-axis is the SGLD learning rate varying from 0.001 to 0.4, while the y-axis measures model performance in terms of accuracy.

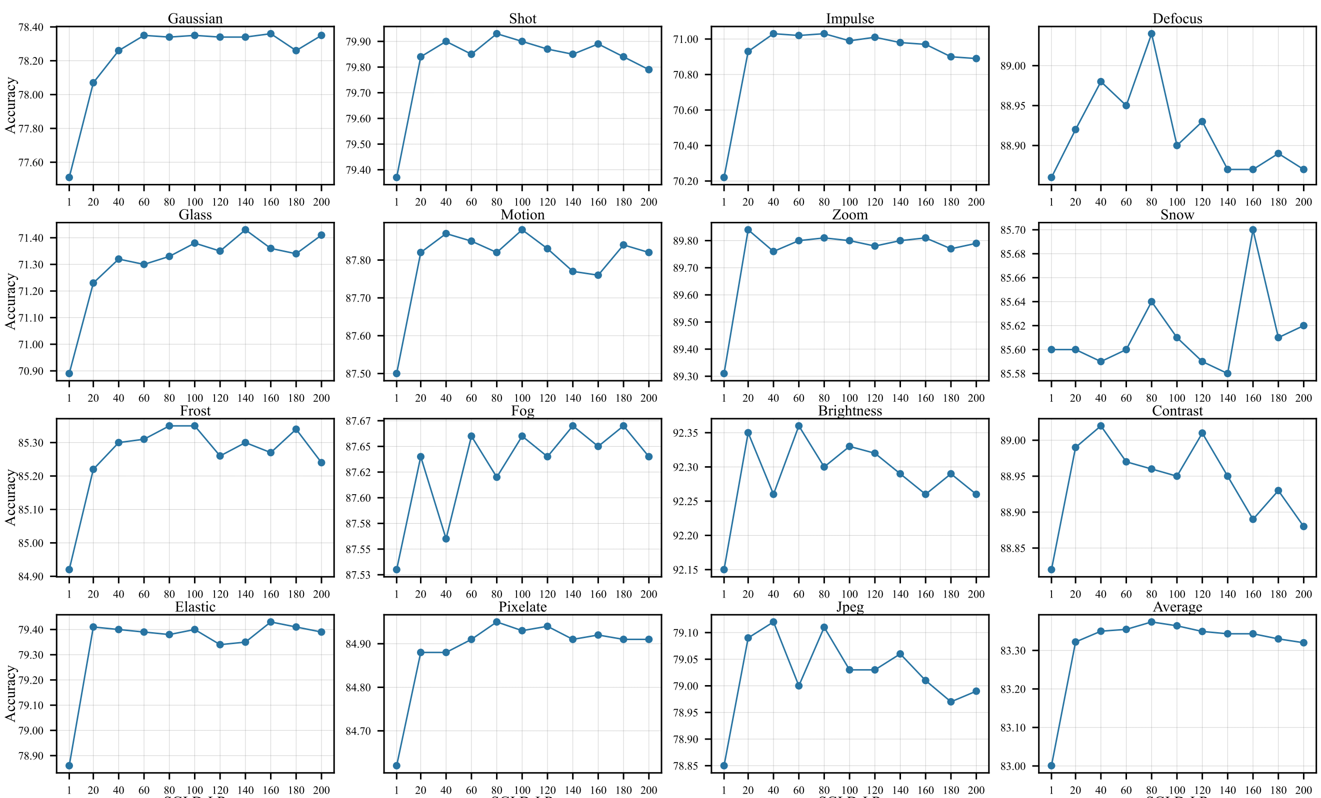

Figure 10. Hyper-parameter stability with respect to the Stochastic Gradient Langevin Dynamics (SGLD) step. The x-axis is the SGLD step varying from 1 to 200, while the y-axis measures model performance in terms of accuracy.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning, 2020. 2

[2] Stephen P Boyd and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004. 5

[3] Miguel A Carreira-Perpinan and Geoffrey Hinton. On contrastive divergence learning. In International workshop on artificial intelligence and statistics, pages 33–40. PMLR, 2005. 4

[4] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 295–305, 2022. 8

[5] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021. 5

[6] Sabrina De Capitani Di Vimercati, Sara Foresti, Giovanni Livraga, and Pierangela Samarati. Data privacy: Definitions and techniques. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 20(06):793–817, 2012. 1

[7] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. arXiv preprint arXiv:1903.08689, 2019. 3, 4

[8] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy-based models. In Proceedings of the 38th International Conference on Machine Learning, pages 2837–2848. PMLR, 2021. 2

[9] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. arXiv preprint arXiv:2103.10360, 2021. 1

[10] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In International Conference on Learning Representations, 2021. 5

[11] Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven adaptation to test-time corruption. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11786–11796, 2023. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020. 2

[13] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In International Conference on Learning Representations, 2020. 3, 4, 2

[14] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In International conference on machine learning, pages 1321–1330. PMLR, 2017. 2, 8

[15] Tian Han, Erik Nijkamp, Xiaolin Fang, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Divergence triangle for joint training of generator model, energy-based model, and inferential model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8670–8679, 2019. 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 5

[17] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In International Conference on Learning Representations, 2019. 5, 1, 2

[18] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. Neural computation, 14(8): 1771–1800, 2002. 2, 3, 4

[19] Liang Hou, Qi Cao, Yige Yuan, Songtao Zhao, Chongyang Ma, Siyuan Pan, Pengfei Wan, Zhongyuan Wang, Huawei Shen, and Xueqi Cheng. Augmentation-aware self-supervision for data-efficient gan training. arXiv preprint arXiv:2205.15677, 2022. 2

[20] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations, 2022. 5

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE international conference on computer vision, pages 1501–1510, 2017. 5

[22] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, pages 448–456. pmlr, 2015. 2, 5

[23] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. Big data privacy: a technological perspective and review. Journal of Big Data, 3:1–25, 2016. 1

[24] Minguk Jang, Sae-Young Chung, and Hye Won Chung. Test-time adaptation via self-training with nearest neighbor information. In The Eleventh International Conference on Learning Representations, 2023. 2

[25] Jing Jiang. A literature survey on domain adaptation of statistical classifiers. URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey, 3(1-12):3, 2008. 3

[26] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. Science, 349(6245): 255–260, 2015. 1

[27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. 5

[28] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In 2nd International Conference on Learning

Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings, 2014. 2

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 1

[30] Jogendra Nath Kundu, Akshay R Kulkarni, Suvaansh Bhambri, Deepesh Mehta, Shreyas Anand Kulkarni, Varun Jampani, and Venkatesh Babu Radhakrishnan. Balancing discriminability and transferability for source-free domain adaptation. In International Conference on Machine Learning, pages 11710–11728. PMLR, 2022. 2

[31] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. CS 231N, 7(7):3, 2015. 5, 1

[32] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. Predicting structured data, 1(0), 2006. 2, 3

[33] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436–444, 2015. 1

[34] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on challenges in representation learning, ICML, page 896. Atlanta, 2013. 1, 2, 5, 6

[35] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In Proceedings of the IEEE international conference on computer vision, pages 5542–5550, 2017. 5, 2

[36] 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, pages 6028–6039. PMLR, 2020. 1, 2, 5, 6

[37] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. arXiv preprint arXiv:2303.15361, 2023. 1, 2

[38] EM Lifshitz and Lev Davidovich Landau. Statistical physics, course of theoretical physics. In Part 2: Theory of the Condensed State. Butterworth-Heinemann Pergamon, London, 1980. 4

[39] Fu Lin, Rohit Mittapalli, Prithvijit Chattopadhyay, Daniel Bolya, and Judy Hoffman. Likelihood landscapes: A unifying principle behind many adversarial defenses. In Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pages 39–54. Springer, 2020. 2

[40] Weijian Luo, Hao Jiang, Tianyang Hu, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Training energy-based models with diffusion contrastive divergences. arXiv preprint arXiv:2307.01668, 2023. 2

[41] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14765–14775, 2022. 1, 2, 5, 6

[42] Weili Nie, Arash Vahdat, and Anima Anandkumar. Controllable and compositional generation with latent-space energy-based models. Advances in Neural Information Processing Systems, 34:13497–13510, 2021. 2

[43] Frank Nielsen and Ke Sun. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. Entropy, 18(12):442, 2016. 5

[44] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. Learning non-convergent non-persistent short-run mcmc toward energy-based model. Advances in Neural Information Processing Systems, 32, 2019. 4

[45] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In International conference on machine learning, pages 16888–16905. PMLR, 2022. 1, 2, 5, 6

[46] 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 The Eleventh International Conference on Learning Representations, 2023. 1, 2, 5, 6

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019. 5

[48] Francesco Pinto, Harry Yang, Ser Nam Lim, Philip Torr, and Puneet Dokania. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. In Advances in Neural Information Processing Systems, pages 14608–14622. Curran Associates, Inc., 2022. 2

[49] Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. Dataset shift in machine learning. Mit Press, 2008. 1

[50] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In International conference on machine learning, pages 1530–1538. PMLR, 2015. 2

[51] Herbert Robbins and Sutton Monro. A stochastic approximation method. The annals of mathematical statistics, pages 400–407, 1951. 5

[52] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. Advances in neural information processing systems, 33:11539–11551, 2020. 1, 2, 5, 6

[53] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference, 90(2):227–244, 2000. 2

[54] Yang Song and Diederik P Kingma. How to train your energy-based models. arXiv preprint arXiv:2101.03288, 2021. 2, 3

[55] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts.

In Proceedings of the 37th International Conference on Machine Learning, pages 9229–9248. PMLR, 2020. 2

[56] Yushun Tang, Ce Zhang, Heng Xu, Shuoshuo Chen, Jie Cheng, Luziwei Leng, Qinghai Guo, and Zhihai He. Neuro-modulated hebbian learning for fully test-time adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3728–3738, 2023. 6

[57] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. Advances in Neural Information Processing Systems, 33: 18583–18599, 2020. 1

[58] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023. 1

[59] Vladimir Vapnik. The nature of statistical learning theory. Springer science & business media, 1999. 1

[60] 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. 1, 2, 5, 6

[61] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. IEEE Transactions on Knowledge and Data Engineering, 2022. 1

[62] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 681–688, 2011. 2, 4

[63] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771, 2019. 1

[64] Yuxin Wu and Kaiming He. Group normalization. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018. 5

[65] Zehao Xiao, Xiantong Zhen, Shengcai Liao, and Cees G. M. Snoek. Energy-based test sample adaptation for domain generalization. In The Eleventh International Conference on Learning Representations, 2023. 2

[66] Xiulong Yang and Shihao Ji. Jem++: Improved techniques for training jem. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6494–6503, 2021. 2

[67] Yige Yuan, Bingbing Xu, Huawei Shen, Qi Cao, Keting Cen, Wen Zheng, and Xueqi Cheng. Towards generalizable graph contrastive learning: An information theory perspective. arXiv preprint arXiv:2211.10929, 2022. 1

[68] Yige Yuan, Bingbing Xu, Bo Lin, Liang Hou, Fei Sun, Huawei Shen, and Xueqi Cheng. Pde+: Enhancing generalization via pde with adaptive distributional diffusion. arXiv preprint arXiv:2305.15835, 2023. 5

[69] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In British Machine Vision Conference 2016. British Machine Vision Association, 2016. 5

[70] Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML, 2023. 1