CLIPArTT: Adaptation of CLIP to New Domains at Test Time - Supplementary Material

1. Dataset Details

We evaluate CLIPArTT's performance across diverse TTA datasets using established methodologies. These datasets simulate challenging scenarios, offering insights into our approach's efficacy. Additionally, we explore CLIPArTT's adaptability on other datasets through zero-shot test-time adaptation.

Our evaluation framework encompasses *natural images*, *varied styles and textures images*, *common corruptions*, *simulated images*, and *video* providing a comprehensive assessment of the model's performance across diverse challenges.

In our evaluation of *natural images* (also known as zero-shot scenario), we utilize CIFAR-10, CIFAR-10.1, and CIFAR-100 datasets, each comprising 10,000 images and featuring 10 and 100 classes, respectively. These datasets represent natural imagery and are novel to the model under scrutiny. Notably, CIFAR-10.1 introduces a natural domain shift from CIFAR-10, thereby enriching our assessment with varied and nuanced data distributions. We also evaluate our method on ImageNet and extend our investigation on two datasets mostly used in the field of domain generalization: PACS [2] and OfficeHome [5] datasets, instrumental in understanding *texture and style variations*. The PACS dataset consists of 9,991 images across four domains (Art, Cartoons, Photos, Sketches) and seven classes. Lastly, the OfficeHome dataset includes 15,588 images across four domains (Art, Clipart, Product, Real) and 65 classes. Evaluating across these distinct scenarios showcases the generalizability of our method.

Transitioning to our investigation of *common corruptions*, we turn to the CIFAR-10-C and CIFAR-100-C [1]. These datasets offer a diverse range of 15 distinct corruptions, including elastic transform and impulse noise, among others. Each corruption is characterized by 5 severity levels, yielding a total of 75 unique testing scenarios per dataset. Within each severity level, there are 10,000 images, contributing to a comprehensive evaluation of the model's robustness against a variety of corruption types and intensities. Finally, we test our method on ImageNet-C to evaluate its performance on a larger dataset with 1,000 classes.

Finally, we examine the VisDA-C dataset's [4] two domain shifts: *simulated* (3D) and *video* (YT). The former comprises a set of 152,397 images rendered in 3D across 12 different classes. The latter includes 72,372 YouTube video frames spanning the same categories. This dataset presents an important challenge, as it bridges the gap of the type of imagery that a model can be applied on.

2. Unsupervised clustering

In Fig. 1, tSNE visualizations of data points are shown. We show how the distribution of data points change after adaptation, which improves the accuracy of class predictions and facilitates the assignment of ground truth labels.

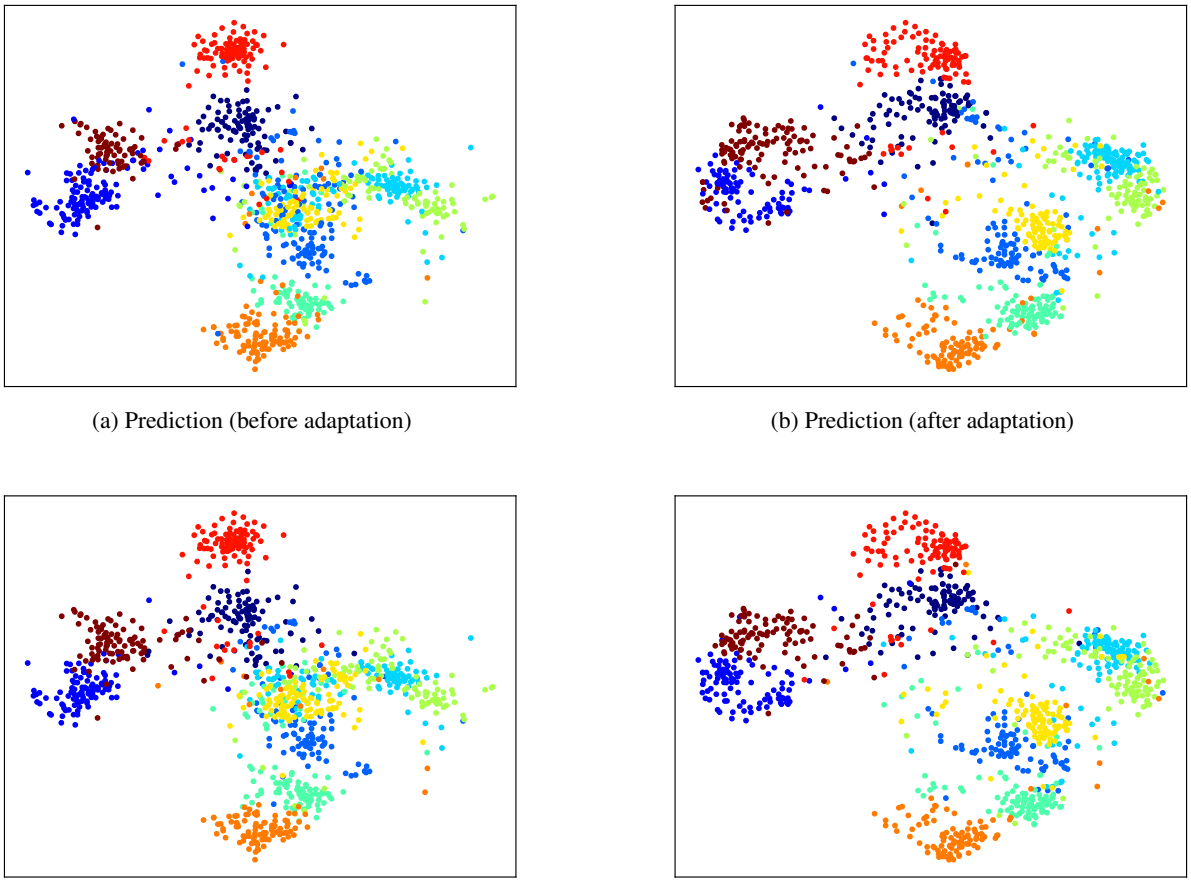
3. Additional settings

We explore additional experimental settings to further explore the strengths and weaknesses of CLIPArTT. Although these scenarios deviate from the standard practices in the TTA literature, they are useful to evaluate a method's performance in potentially challenging real-world applications. For most of the following experiments, we utilize the CIFAR-10 dataset's variants, unless otherwise is mentioned.

3.1. Imbalanced batch instances

CLIPArTT adapts to batch data in a trandsductive manner by computing the image-to-image similarity. The diversity of the batches cannot be ensured due to their finite size. The opposite case naturally arises in large scale classification datasets such as ImageNet, where including at least one image per class in a batch of size 128 is impossible. On the other hand, adaptation to highly imbalanced batches is understudied.

In this experiment, we force extreme imbalances in the batches, by allowing only C randomly chosen classes to be present in the batch. As CLIPArTT is dependant also on the most probable predictions, it is expected that the missclassification due to the imbalance would drive the model to directions opposite to the actual correct classes. In this experiment, we evaluate with $C = \{2, 3, 4, 5\}$ as the possible number of classes to be present in the batches, and compare CLIPArTT against TENT. Final results are shown



(a) Prediction (before adaptation)

(b) Prediction (after adaptation)

(c) Ground truth (before adaptation)

(d) Ground truth (after adaptation)

Figure 1. The t-SNE visualizations exhibit discernible attributes of brightness within the visual features derived from CLIPArTT. Panels (a) and (b) present the model's predictions before and after 10 iterations of adaptation, respectively. Panels (c) and (d) demonstrate the actual labels in the absence of adaptation and following adaptation of the representations, respectively.

in Table 1).

It can be observed that TENT shows a higher robustness to class imbalance, as for each prediction, it depends less on the other images' predictions (i.e., induction). Interestingly, we observe a trend where TENT's performance decrease when more classes are present (e.g., $C=5$ instead of $C=2$), whereas CLIPArTT increase its performance as C grows.

3.2. Open-Set classification

Another interesting scenario is open-set classification, where images from classes that are not considered are present in the batch. These images could potentially affect performance, specially in transductive methods and particularly in CLIPArTT, as the image-to-image and text-to-text similarities are combined. The only available text prompts are wrongly assigned to the out-of-distribution (OOD) samples.

To measure the effectiveness of our method in this scenario, we utilize SVHN-C, a corrupted version of the SVHN [3] dataset, analogous to CIFAR-10.C. For each batch of 128 image, 128 additional OOD images are included. Both parts are used at the same time for adaptation, and accuracy is only measured on the first half.

As seen in Table 2, CLIPArTT defeats TENT's open-set accuracy by a significant margin. This encouraging result suggests that our method is robust to semantic out-of-distribution perturbations.

3.3. Generalization after adaptation

In TTA, the classification performance is evaluated directly on the adapted set of images in an episodic manner. However, a high accuracy on this set does not necessarily guarantee a good performance in a separate set of images unknown to the model.

	CLIPArTT				TENT			
	2	3	4	5	2	3	4	5
Original	76.99	81.12	84.37	86.92	97.07	96.14	95.27	94.09
CIFAR-10.1	74.21	78.96	82.57	85.14	96.47	94.33	92.56	90.68
Gaussian Noise	47.12	50.83	54.18	58.63	63.76	49.49	49.92	47.15
Shot Noise	48.67	53.30	56.37	60.24	69.89	55.83	56.68	53.55
Impulse Noise	45.98	49.82	53.15	57.48	66.87	54.47	54.77	52.88
Defocus Blur	60.80	67.78	71.24	74.85	89.19	85.39	83.05	81.44
Glass Blur	47.82	53.07	56.45	60.12	77.11	65.34	62.72	58.78
Motion Blur	60.85	66.82	70.67	73.43	83.33	79.88	79.28	75.89
Zoom Blur	56.78	66.52	70.43	72.94	86.88	84.15	81.89	81.35
Snow	59.20	66.38	71.29	73.97	91.84	87.60	85.32	82.45
Frost	60.02	68.08	72.12	75.15	92.50	88.62	86.18	84.25
Fog	58.57	66.11	70.54	73.12	91.26	87.01	84.07	83.12
Brightness	72.29	77.62	80.75	83.55	95.45	93.82	92.29	91.15
Contrast	62.18	71.68	75.08	77.64	89.75	87.11	85.72	83.25
Elastic Transform	50.85	58.91	62.12	65.12	88.17	82.05	78.13	76.05
Pixelate	51.75	57.20	60.42	63.08	77.35	71.38	69.22	66.85
JPEG Compression	49.63	56.08	59.32	62.05	79.98	72.34	69.85	68.17

Table 1. Accuracy on imbalance datasets with different numbers of available classes at each batch.

	CLIPArTT	TENT
Gaussian Noise	59.72 $\pm$0.05	41.27 $\pm$0.03
Shot Noise	62.11 $\pm$0.07	48.14 $\pm$0.05
Impulse Noise	55.23 $\pm$0.10	47.86 $\pm$0.07
Defocus Blur	75.44 $\pm$0.07	72.03 $\pm$0.05
Glass Blur	60.12 $\pm$0.07	42.08 $\pm$0.08
Motion Blur	74.60 $\pm$0.05	65.71 $\pm$0.05
Zoom Blur	76.15 $\pm$0.05	71.45 $\pm$0.06
Snow	74.69 $\pm$0.05	73.99 $\pm$0.08
Frost	77.72 $\pm$0.05	74.49 $\pm$0.09
Fog	72.87 $\pm$0.03	70.57 $\pm$0.04
Brightness	85.70 $\pm$0.09	82.34 $\pm$0.05
Contrast	76.49 $\pm$0.04	71.67 $\pm$0.04
Elastic Transform	67.42 $\pm$0.01	65.99 $\pm$0.12
Pixelate	62.02 $\pm$0.11	54.25 $\pm$0.08
JPEG Compression	60.81 $\pm$0.07	54.09 $\pm$0.10
Average	69.41 $\pm$8.86	62.40 $\pm$13.20

Table 2. Open-set accuracy on CIFAR-10-C when using SVHN-C as the OOD dataset.

Using a batch size of 160 images, we measure CLIPArTT's generalization by separating 25% of each batch as a test-time validation split. The remaining 75% is used for adaptation prior to testing on the validation split. The accuracies on the adaptation split and the validation split are both reported and contrasted against TENT. Results are shown in Table 3.

While TENT obtains a higher accuracy on the adaptation splits of natural images (i.e., CIFAR-10 and CIFAR-10.1), following a similar trend as in the main experiments' results, the difference in the generalization accuracy is smaller with respect to CLIPArTT's. Moreover, our method demonstrates a consistently better generalization and adaptation accuracy on corrupted samples, representing stronger domain shifts.

	CLIPArTT		TENT	
	Acc.	Acc. Gen.	Acc.	Acc. Gen.
Original	95.43 ±0.01	71.23 ±0.0008	97.02 ±0.02	72.68 ±0.0023
CIFAR-10.1	88.75 ±0.85	65.87 ±0.8083	90.49 ±0.38	67.07 ±1.1
Gaussian Noise	63.79 ±0.20	46.87 ±0.0008	45.30 ±0.05	30.83 ±0.0030
Shot Noise	66.38 ±0.10	79.80 ±0.0038	50.65 ±0.18	35.85 ±0.0025
Impulse Noise	59.31 ±0.19	44.56 ±0.0023	51.91 ±0.20	37.99 ±0.0037
Defocus Blur	81.26 ±0.03	60.13 ±0.0016	84.48 ±0.07	60.32 ±0.0028
Glass Blur	65.38 ±0.04	49.47 ±0.0020	56.68 ±0.26	41.32 ±0.0035
Motion Blur	80.41 ±0.12	60.08 ±0.0022	75.70 ±0.13	55.72 ±0.0013
Zoom Blur	81.97 ±0.05	60.63 ±0.0009	81.14 ±0.13	59.33 ±0.0007
Snow	82.06 ±0.14	60.83 ±0.0014	83.08 ±0.13	61.08 ±0.0008
Frost	83.85 ±0.19	61.73 ±0.0009	84.70 ±0.11	66.75 ±0.0006
Fog	90.04 ±0.27	59.48 ±0.0027	81.56 ±0.09	60.69 ±0.0033
Brightness	91.73 ±0.04	68.01 ±0.0017	92.93 ±0.03	68.80 ±0.0018
Contrast	82.63 ±0.07	61.28 ±0.0025	83.53 ±0.06	62.97 ±0.0009
Elastic Transform	74.04 ±0.22	54.87 ±0.0011	74.08 ±0.10	54.92 ±0.0009
Pixelate	70.58 ±0.02	51.53 ±0.0042	67.35 ±0.18	50.19 ±0.0004
JPEG Compression	67.32 ±0.06	49.56 ±0.0023	66.25 ±0.14	48.85 ±0.0021
Average	**75.38** ±9.04	55.92 ±6.62	71.78 ±14.14	52.77 ±11.10

Table 3. Generalization results on CIFAR-10 variants. Each batch is divided into an adaptation split and a validation split. Accuracy (Acc.) is measured on the former after adaptation, whilst the generalization accuracy (Acc. Gen.) is measured on the later.

4. Computational Cost

In this section, we compare the computational cost of CLIPArTT with other TTA methods through a thorough evaluation under consistent conditions, using an NVIDIA A6000 GPU within the same Python environment. The provided table 4 compares adaptation time, memory usage, and the number of learnable parameters across various TTA methods, including our proposed CLIPArTT. The results demonstrate that CLIPArTT maintains competitive adaptation time and memory usage relative to other approaches, such as TENT and TPT.

5. Comprehensive experimental results

We present comprehensive tables containing all the detailed information about results that was summarized in the main paper.

References

[1] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. 2019.

[2] 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.

[3] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

[4] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[5] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

Method	Adaptation Time	Memory	Pct. of Learnable Parameters
TENT	0.28 s	1.5 GB	0.026%
TPT	0.26 s	1.7 GB	0.001%
CLIPArTT	0.55 s	1.7 GB	0.026%

Table 4. Comparison of Computational Cost.

	CIFAR10		CIFAR100	
	Top 1	Top 3	Top 1	Top 3
Original	88.74	100.00	61.68	97.34
Gaussian Noise	35.27	99.87	14.8	63.66
Shot noise	39.67	99.99	16.03	67.02
Impulse Noise	42.61	100.00	13.85	64.4
Defocus blur	69.76	100.00	36.74	90.14
Glass blur	42.40	100.00	14.19	61.66
Motion blur	63.97	100.00	36.14	90.36
Zoom blur	69.83	100.00	40.24	91.27
Snow	71.78	100.00	38.95	91.40
Frost	72.86	100.00	40.56	92.23
Fog	67.04	99.98	38.00	91.51
Brightness	81.87	100.00	48.18	93.10
Contrast	64.37	100.00	29.53	84.67
Elastic transform	60.83	100.00	26.33	78.96
Pixelate	50.53	100.00	21.98	75.65
JPEG compression	55.48	100.00	25.91	80.81
Average	59.22	99.99	29.43	81.12

Table 5. Accuracy (%) on CIFAR-10/100 and CIFAR-10/100-C datasets with Level 5 corruption for the top 1 or the top 3 predicted classes.

	K = 1	K = 3	K = 4
Original	89.8 $\pm$0.05	90.04 $\pm$0.13	90.41 $\pm$0.07
CIFAR 10.1	85.37 $\pm$0.17	86.35 $\pm$0.27	86.07 $\pm$0.21
Gaussian Noise	60.20 $\pm$0.24	59.90 $\pm$0.36	59.71 $\pm$0.15
Shot noise	62.08 $\pm$0.11	62.77 $\pm$0.07	62.17 $\pm$0.16
Impulse Noise	54.33 $\pm$0.07	56.02 $\pm$0.16	56.27 $\pm$0.15
Defocus blur	77.16 $\pm$0.02	76.74 $\pm$0.05	76.79 $\pm$0.11
Glass blur	61.91 $\pm$0.15	61.77 $\pm$0.16	61.72 $\pm$0.23
Motion blur	74.94 $\pm$0.15	76.01 $\pm$0.19	76.33 $\pm$0.10
Zoom blur	76.84 $\pm$0.13	77.40 $\pm$0.20	77.15 $\pm$0.04
Snow	76.87 $\pm$0.05	77.29 $\pm$0.16	76.56 $\pm$0.16
Frost	77.81 $\pm$0.04	79.20 $\pm$0.08	78.42 $\pm$0.04
Fog	75.83 $\pm$0.28	75.74 $\pm$0.14	75.65 $\pm$0.06
Brightness	85.55 $\pm$0.12	86.59 $\pm$0.16	86.83 $\pm$0.10
Contrast	78.02 $\pm$0.18	77.82 $\pm$0.14	78.27 $\pm$0.14
Elastic transform	69.42 $\pm$0.07	70.20 $\pm$0.01	69.81 $\pm$0.20
Pixelate	66.07 $\pm$0.09	66.52 $\pm$0.13	66.45 $\pm$0.08
JPEG compression	64.82 $\pm$0.26	63.51 $\pm$0.14	62.72 $\pm$0.25
Average	70.79	71.17	70.99

Table 6. Accuracy (%) on CIFAR-10, CIFAR-10.1 and CIFAR-10-C datasets with Level 5 corruption for different number of K selected classes to create pseudo-label.

	$K=1$	$K=3$	$K=5$	$K=7$	$K=10$	$K=20$
Original	69.00 ±0.22	69.79 ±0.04	69.68 ±0.07	69.56 ±0.02	69.78 ±0.02	**69.93 ±0.08**
Gaussian Noise	**26.05 ±0.11**	25.32 ±0.14	24.69 ±0.03	24.60 ±0.03	24.70 ±0.15	23.73 ±0.07
Shot noise	**28.88 ±0.11**	27.90 ±0.05	27.25 ±0.26	26.75 ±0.07	26.83 ±0.26	25.73 ±0.13
Impulse Noise	24.04 ±0.09	**25.62 ±0.09**	25.12 ±0.14	25.25 ±0.14	24.95 ±0.23	24.57 ±0.12
Defocus blur	49.03 ±0.15	**49.88 ±0.23**	49.75 ±0.11	49.74 ±0.25	49.62 ±0.10	49.49 ±0.07
Glass blur	26.77 ±0.14	**27.89 ±0.03**	27.76 ±0.23	27.28 ±0.07	26.57 ±0.07	25.52 ±0.09
Motion blur	46.50 ±0.09	**47.93 ±0.14**	47.48 ±0.21	47.57 ±0.10	47.53 ±0.09	47.36 ±0.09
Zoom blur	52.08 ±0.12	52.70 ±0.06	52.22 ±0.10	52.10 ±0.24	52.62 ±0.24	**52.82 ±0.09**
Snow	49.24 ±0.07	**49.72 ±0.01**	48.98 ±0.08	48.87 ±0.08	49.13 ±0.08	49.54 ±0.13
Frost	**49.91 ±0.07**	49.63 ±0.12	48.43 ±0.17	48.11 ±0.06	48.72 ±0.08	49.13 ±0.10
Fog	47.15 ±0.04	48.77 ±0.04	48.95 ±0.18	48.78 ±0.22	**49.06 ±0.05**	48.74 ±0.36
Brightness	60.01 ±0.08	**61.27 ±0.08**	60.77 ±0.16	60.89 ±0.19	60.98 ±0.18	61.03 ±0.19
Contrast	46.90 ±0.21	**48.55 ±0.24**	49.01 ±0.14	49.07 ±0.03	**49.27 ±0.09**	49.08 ±0.12
Elastic transform	36.32 ±0.10	37.45 ±0.08	**37.63 ±0.12**	37.31 ±0.09	37.13 ±0.16	36.94 ±0.11
Pixelate	32.52 ±0.17	33.88 ±0.14	34.40 ±0.15	34.38 ±0.16	**34.65 ±0.09**	34.32 ±0.02
JPEG compression	35.81 ±0.11	**36.07 ±0.32**	35.77 ±0.01	35.60 ±0.10	35.63 ±0.10	35.29 ±0.14
Average	40.75	**41.51**	41.21	41.09	41.16	40.89

Table 7. Accuracy (%) on CIFAR-100 and CIFAR-100-C datasets with Level 5 corruption for different number of K selected classes to create pseudo-labels.

	Iter = 1	Iter = 5	Iter = 10	Iter = 20
Original	89.59 ±0.01	**90.54 ±0.09**	90.04 ±0.13	88.32 ±0.12
CIFAR 10.1	84.78 ±0.02	**86.67 ±0.06**	86.35 ±0.27	84.33 ±0.31
Gaussian Noise	39.75 ±0.04	53.79 ±0.08	**59.90 ±0.36**	59.33 ±0.20
Shot noise	43.80 ±0.04	57.16 ±0.24	62.77 ±0.07	**63.08 ±0.43**
Impulse Noise	45.19 ±0.07	52.44 ±0.14	56.02 ±0.16	**56.73 ±0.01**
Defocus blur	72.93 ±0.07	76.36 ±0.10	**76.74 ±0.05**	75.33 ±0.13
Glass blur	46.61 ±0.06	57.45 ±0.13	61.77 ±0.16	**62.01 ±0.27**
Motion blur	67.89 ±0.03	74.34 ±0.14	**76.01 ±0.19**	75.94 ±0.28
Zoom blur	73.24 ±0.05	77.03 ±0.07	**77.40 ±0.20**	75.42 ±0.13
Snow	73.81 ±0.07	76.51 ±0.08	**77.29 ±0.16**	76.18 ±0.21
Frost	74.80 ±0.04	78.13 ±0.12	**79.20 ±0.08**	77.44 ±0.30
Fog	69.81 ±0.06	74.16 ±0.09	**75.74 ±0.14**	74.66 ±0.02
Brightness	84.16 ±0.01	**86.61 ±0.10**	86.59 ±0.16	85.14 ±0.42
Contrast	67.75 ±0.03	74.85 ±0.04	**77.82 ±0.14**	77.75 ±0.11
Elastic transform	63.15 ±0.08	68.53 ±0.19	**70.20 ±0.01**	68.48 ±0.24
Pixelate	54.20 ±0.02	61.87 ±0.04	66.52 ±0.13	**67.13 ±0.15**
JPEG compression	57.46 ±0.09	62.00 ±0.13	63.51 ±0.14	**63.64 ±0.20**
Average	62.30	68.75	**71.17**	70.55

Table 8. Accuracy (%) on CIFAR-10, CIFAR-10.1 and CIFAR-10-C datasets with Level 5 corruption for different number of iterations to update the model at test-time.

	Image	Text	Image + Text
Original	**90.18** ±**0.02**	89.05 ±0.14	90.04 ±0.13
CIFAR 10.1	86.25 ±0.37	84.85 ±0.40	**86.35** ±**0.27**
Gaussian Noise	59.05 ±0.30	59.29 ±0.27	**59.90** ±**0.36**
Shot noise	62.11 ±0.18	61.89 ±0.31	**62.77** ±**0.07**
Impulse Noise	55.43 ±0.12	55.48 ±0.10	**56.02** ±**0.16**
Defocus blur	76.88 ±0.09	76.25 ±0.10	**76.74** ±**0.05**
Glass blur	61.56 ±0.03	61.28 ±0.33	**61.77** ±**0.16**
Motion blur	**76.32** ±**0.15**	75.37 ±0.27	76.01 ±0.19
Zoom blur	**77.66** ±**0.09**	76.29 ±0.11	77.40 ±0.20
Snow	77.28 ±0.08	76.03 ±0.16	**77.29** ±**0.16**
Frost	78.80 ±0.18	78.05 ±0.21	**79.20** ±**0.08**
Fog	75.69 ±0.17	73.70 ±0.15	**75.74** ±**0.14**
Brightness	**86.62** ±**0.20**	84.69 ±0.04	86.59 ±0.16
Contrast	77.43 ±0.12	74.52 ±0.02	**77.82** ±**0.14**
Elastic transform	69.63 ±0.02	69.33 ±0.20	**70.20** ±**0.01**
Pixelate	66.33 ±0.16	64.86 ±0.29	**66.52** ±**0.13**
JPEG compression	**63.92** ±**0.13**	63.44 ±0.20	63.51 ±0.14
Average	70.98	70.03	**71.17**

Table 9. Accuracy (%) on CIFAR-10, CIFAR-10.1 and CIFAR-10-C datasets with different targets.

	CLIP	BS = 8	BS = 16	BS = 32	BS = 64	BS = 128
Original	88.74	82.50 ±0.13	85.89 ±0.19	88.25 ±0.15	89.48 ±0.15	90.04 ±0.13
CIFAR 10.1	83.25	77.2 ±0.92	81.55 ±0.53	84.00 ±0.31	85.40 ±0.08	86.35 ±0.27
Gaussian Noise	35.27	47.30 ±0.37	50.91 ±0.35	54.23 ±0.28	57.89 ±0.13	59.90 ±0.36
Shot noise	39.67	49.62 ±0.26	53.1 ±0.27	56.88 ±0.23	60.56 ±0.12	62.77 ±0.07
Impulse Noise	42.61	47.24 ±0.22	50.24 ±0.48	52.7 ±0.21	54.88 ±0.17	56.02 ±0.16
Defocus blur	69.76	68.24 ±0.35	72.22 ±0.04	75.09 ±0.16	75.97 ±0.27	76.74 ±0.05
Glass blur	42.40	49.49 ±0.30	53.27 ±0.04	57.18 ±0.24	60.12 ±0.14	61.77 ±0.16
Motion blur	63.97	65.22 ±0.06	69.02 ±0.30	72.54 ±0.27	74.71 ±0.18	76.01 ±0.19
Zoom blur	69.83	67.69 ±0.20	71.33 ±0.11	74.53 ±0.11	76.35 ±0.07	77.40 ±0.20
Snow	71.78	68.68 ±0.42	72.37 ±0.11	74.93 ±0.18	76.53 ±0.41	77.29 ±0.16
Frost	72.86	70.35 ±0.25	73.93 ±0.34	76.81 ±0.23	78.22 ±0.13	79.20 ±0.08
Fog	67.04	66.25 ±0.31	69.71 ±0.24	72.36 ±0.23	73.96 ±0.21	75.74 ±0.14
Brightness	81.87	77.36 ±0.17	81.20 ±0.20	84.07 ±0.08	85.58 ±0.25	86.59 ±0.16
Contrast	64.37	65.12 ±0.07	69.02 ±0.12	72.60 ±0.46	75.79 ±0.24	77.82 ±0.14
Elastic transform	60.83	59.61 ±0.11	63.67 ±0.13	66.36 ±0.26	68.74 ±0.07	70.20 ±0.01
Pixelate	50.53	56.78 ±0.24	60.01 ±0.06	62.57 ±0.19	64.64 ±0.03	66.52 ±0.13
JPEG compression	55.48	57.59 ±0.26	60.78 ±0.12	62.63 ±0.06	63.43 ±0.16	63.51 ±0.14
Average	59.22	61.10	64.72	67.70	69.82	71.17

Table 10. Accuracy (%) on CIFAR-10, CIFAR-10.1 and CIFAR-10-C datasets with Level 5 corruption for different batch sizes.

	CLIP	TENT	TPT (BS=32)	CLIPArTT
Original	89.25	**92.75** ±**0.17**	89.80 ±0.05	92.61 ±0.05
CIFAR 10.1	84.00	88.52 ±0.33	83.75.0.21 ±	**88.72** ±**0.33**
Gaussian Noise	37.75	31.04 ±0.38	35.35 ±0.15	**60.89** ±**0.26**
Shot noise	41.10	40.54 ±0.41	41.03 ±0.19	**65.19** ±**0.21**
Impulse Noise	51.71	58.03 ±0.16	54.86 ±0.07	**67.55** ±**0.09**
Defocus blur	70.07	77.57 ±0.03	70.29 ±0.02	**78.92** ±**0.12**
Glass blur	42.24	47.16 ±0.05	37.86 ±0.17	**57.18** ±**0.20**
Motion blur	65.81	76.16 ±0.05	67.43 ±0.11	**76.59** ±**0.06**
Zoom blur	72.50	**79.64** ±**0.12**	72.91 ±0.02	79.62 ±0.11
Snow	73.23	**81.68** ±**0.03**	72.98 ±0.32	81.13 ±0.29
Frost	76.52	**83.22** ±**0.05**	75.87 ±0.16	81.24 ±0.08
Fog	68.35	**80.78** ±**0.15**	69.13 ±0.27	78.47 ±0.19
Brightness	83.36	**89.85** ±**0.11**	83.67 ±0.14	88.66 ±0.15
Contrast	61.90	**79.24** ±**0.19**	62.16 ±0.06	75.15 ±0.07
Elastic transform	53.16	62.54 ±0.08	51.26 ±0.23	**69.49** ±**0.08**
Pixelate	48.48	67.08 ±0.24	44.65 ±0.21	**71.80** ±**0.16**
JPEG compression	56.05	65.42 ±0.05	56.73 ±0.07	**66.42** ±**0.25**
Average	60.15	68.00	59.75	**73.22**

Table 11. Accuracy (%) on CIFAR-10, CIFAR-10.1 and CIFAR-10-C datasets with ViT-B/16 as visual encoder.

	CLIP	TENT	TPT (BS=32)	CLIPArTT
Original	95.36	**96.13** ±**0.06**	95.18 ±0.02	95.16 ±0.03
CIFAR 10.1	91.20	**92.22** ±**0.25**	91.32 ±0.12	91.02 ±0.02
Gaussian Noise	64.64	68.87 ±0.20	64.44 ±0.11	**70.04** ±**0.31**
Shot noise	67.82	**71.95** ±**0.06**	66.81 ±0.19	71.44 ±0.16
Impulse Noise	78.21	**80.22** ±**0.19**	76.46 ±0.17	79.42 ±0.15
Defocus blur	80.73	**83.10** ±**0.03**	79.01 ±0.23	81.75 ±0.19
Glass blur	50.29	57.12 ±0.07	49.64 ±0.23	**58.13** ±**0.23**
Motion blur	80.75	**82.69** ±**0.11**	78.85 ±0.04	80.76 ±0.12
Zoom blur	82.75	**84.91** ±**0.08**	82.32 ±0.13	83.39 ±0.05
Snow	83.01	**85.99** ±**0.11**	82.69 ±0.10	84.48 ±0.07
Frost	84.90	**87.15** ±**0.12**	84.63 ±0.08	85.21 ±0.06
Fog	78.44	**81.30** ±**0.07**	77.56 ±0.17	79.27 ±0.07
Brightness	91.67	**93.07** ±**0.04**	90.94 ±0.04	91.87 ±0.09
Contrast	84.20	**87.93** ±**0.04**	82.88 ±0.09	86.19 ±0.06
Elastic transform	65.45	**69.96** ±**0.12**	64.81 ±0.14	67.43 ±0.24
Pixelate	75.10	**77.89** ±**0.05**	72.92 ±0.12	77.11 ±0.10
JPEG compression	72.58	**75.49** ±**0.07**	71.18 ±0.19	74.46 ±0.11
Average	76.04	**79.18**	75.01	78.06

Table 12. Accuracy (%) on CIFAR-10, CIFAR-10.1 and CIFAR-10-C datasets with ViT-L/14 as visual encoder.

	CLIP	TENT	TPT (BS=32)	CLIPArTT
Original	64.76	**71.73 ±0.14**	67.15 ±0.23	71.34 ±0.07
Gaussian Noise	15.88	12.28 ±0.20	15.43 ±0.03	**19.01 ±0.24**
Shot noise	17.49	15.07 ±0.21	16.88 ±0.07	**20.27 ±0.21**
Impulse Noise	21.43	13.13 ±0.16	**22.12 ±0.15**	17.66 ±0.10
Defocus blur	40.10	**50.35 ±0.03**	41.08 ±0.22	49.86 ±0.13
Glass blur	13.48	4.84 ±0.14	**18.43 ±0.15**	18.34 ±0.31
Motion blur	39.82	49.85 ±0.37	40.85 ±0.26	**50.00 ±0.09**
Zoom blur	45.45	**54.76 ±0.04**	46.77 ±0.06	54.13 ±0.08
Snow	42.77	52.38 ±0.18	47.24 ±0.18	**52.80 ±0.27**
Frost	45.39	**51.66 ±0.04**	48.61 ±0.14	49.56 ±0.08
Fog	38.98	**50.74 ±0.14**	39.92 ±0.16	49.92 ±0.11
Brightness	52.55	**64.26 ±0.09**	55.83 ±0.10	63.76 ±0.13
Contrast	33.32	**48.69 ±0.08**	33.13 ±0.16	47.86 ±0.02
Elastic transform	24.39	**33.56 ±0.28**	27.36 ±0.10	32.93 ±0.23
Pixelate	21.89	36.20 ±0.28	21.26 ±0.10	**39.49 ±0.21**
JPEG compression	27.21	30.80 ±0.05	30.97 ±0.10	**35.56 ±0.23**
Average	32.01	37.90	33.73	**40.08**

Table 13. Accuracy (%) on CIFAR-100 and CIFAR-100-C datasets with ViT-B/16 as visual encoder.

	CLIP	TENT	TPT (BS=16)	CLIPArTT
Original	73.28	78.03 ±0.08	76.85 ±0.06	**79.42 ±0.08**
Gaussian Noise	30.55	36.93 ±0.03	36.10 ±0.11	**41.46 ±0.15**
Shot noise	34.58	40.96 ±0.16	38.23 ±0.13	**44.27 ±0.09**
Impulse Noise	44.89	49.09 ±0.14	49.69 ±0.21	**51.44 ±0.23**
Defocus blur	48.88	55.23 ±0.07	50.43 ±0.19	**56.55 ±0.22**
Glass blur	23.46	27.02 ±0.23	24.35 ±0.22	**30.47 ±0.14**
Motion blur	50.83	56.03 ±0.20	51.94 ±0.04	**56.98 ±0.18**
Zoom blur	56.02	61.19 ±0.10	56.96 ±0.16	**62.56 ±0.04**
Snow	49.03	55.60 ±0.09	54.89 ±0.11	**58.81 ±0.11**
Frost	53.27	58.21 ±0.15	58.15 ±0.33	**60.38 ±0.23**
Fog	48.51	53.37 ±0.25	49.26 ±0.13	**54.38 ±0.04**
Brightness	60.53	67.34 ±0.17	66.60 ±0.10	**69.63 ±0.14**
Contrast	50.24	59.91 ±0.13	53.64 ±0.24	**63.39 ±0.13**
Elastic transform	35.07	38.49 ±0.12	35.72 ±0.09	**39.57 ±0.39**
Pixelate	43.86	48.37 ±0.17	44.32 ±0.10	**50.45 ±0.16**
JPEG compression	39.11	44.42 ±0.09	43.44 ±0.11	**47.45 ±0.14**
Average	44.59	50.14	47.58	**52.52**

Table 14. Accuracy (%) on CIFAR-100 and CIFAR-100-C datasets with ViT-L/14 as visual encoder.

	CLIP	LAME	TENT	TPT (BS=32)	CLIPArTT
Original	88.74	89.36 ±0.06	**91.69 ±0.10**	88.06 ±0.06	90.04 ±0.13
CIFAR 10.1	83.25	81.22 ±0.33	**87.60 ±0.45**	81.80 ±0.27	86.35 ±0.27
Gaussian Noise	35.27	24.71 ±0.11	41.27 ±0.27	33.90 ±0.08	**59.90 ±0.36**
Shot noise	39.67	27.44 ±0.09	47.20 ±0.23	38.20 ±0.02	**62.77 ±0.07**
Impulse Noise	42.61	31.38 ±0.15	48.58 ±0.31	37.66 ±0.20	**56.02 ±0.16**
Defocus blur	69.76	62.45 ±0.44	**77.12 ±0.16**	67.83 ±0.28	76.74 ±0.05
Glass blur	42.40	29.96 ±0.06	52.65 ±0.30	38.81 ±0.12	**61.77 ±0.16**
Motion blur	63.97	54.00 ±0.36	71.25 ±0.09	63.39 ±0.13	**76.01 ±0.19**
Zoom blur	69.83	61.97 ±0.36	76.20 ±0.19	68.95 ±0.16	**77.40 ±0.20**
Snow	71.78	64.61 ±0.48	**78.29 ±0.20**	70.16 ±0.10	77.29 ±0.16
Frost	72.86	65.17 ±0.17	**79.84 ±0.09**	72.39 ±0.22	79.20 ±0.08
Fog	67.04	59.13 ±0.49	**77.39 ±0.01**	64.31 ±0.28	75.74 ±0.14
Brightness	81.87	80.05 ±0.23	**87.78 ±0.03**	81.30 ±0.18	86.59 ±0.16
Contrast	64.37	56.91 ±0.37	**79.47 ±0.11**	62.26 ±0.31	77.82 ±0.14
Elastic transform	60.83	53.89 ±0.20	70.00 ±0.25	56.43 ±0.27	**70.20 ±0.01**
Pixelate	50.53	39.67 ±0.34	63.74 ±0.18	42.80 ±0.40	**66.52 ±0.13**
JPEG compression	55.48	47.24 ±0.14	62.64 ±0.14	53.67 ±0.25	**63.51 ±0.14**
Average	59.22	50.57	67.56	56.80	**71.17**

Table 15. Accuracy (%) on CIFAR-10, CIFAR-10.1 and CIFAR-10-C datasets with ViT-B/32 as visual encoder.

	CLIP	LAME	TENT	TPT (BS=32)	CLIPArTT
Original	61.68	58.27 ±0.17	69.74 ±0.16	63.78 ±0.28	**69.79 ±0.04**
Gaussian Noise	14.80	12.72 ±0.04	14.38 ±0.14	14.03 ±0.10	**25.32 ±0.14**
Shot noise	16.03	13.78 ±0.08	17.34 ±0.27	15.25 ±0.17	**27.90 ±0.05**
Impulse Noise	13.85	7.82 ±0.14	10.03 ±0.13	13.01 ±0.13	**25.62 ±0.09**
Defocus blur	36.74	33.38 ±0.11	49.05 ±0.07	37.60 ±0.17	**49.88 ±0.23**
Glass blur	14.19	9.00 ±0.05	3.71 ±0.07	16.41 ±0.02	**27.89 ±0.03**
Motion blur	36.14	32.79 ±0.13	46.62 ±0.27	37.52 ±0.23	**47.93 ±0.14**
Zoom blur	40.24	37.57 ±0.15	51.84 ±0.15	42.99 ±0.11	**52.70 ±0.06**
Snow	38.95	35.49 ±0.18	46.71 ±0.21	42.35 ±0.13	**49.72 ±0.01**
Frost	40.56	37.22 ±0.21	44.90 ±0.27	43.31 ±0.14	**49.63 ±0.12**
Fog	38.00	35.94 ±0.09	47.31 ±0.04	38.81 ±0.17	**48.77 ±0.04**
Brightness	48.18	44.93 ±0.08	60.58 ±0.18	50.23 ±0.11	**61.27 ±0.08**
Contrast	29.53	27.52 ±0.06	45.90 ±0.11	28.09 ±0.09	**48.55 ±0.24**
Elastic transform	26.33	24.01 ±0.02	33.09 ±0.08	28.12 ±0.15	**37.45 ±0.08**
Pixelate	21.98	19.55 ±0.13	26.47 ±0.09	20.43 ±0.14	**33.88 ±0.14**
JPEG compression	25.91	21.77 ±0.14	29.89 ±0.07	28.82 ±0.09	**36.07 ±0.32**
Average	29.43	26.23	35.19	30.46	**41.51**

Table 16. Accuracy (%) on CIFAR-100 and CIFAR-100-C datasets with ViT-B/32 as visual encoder.