Effortless Active Labeling for Long-Term Test-Time Adaptation

Supplementary Material

This supplementary material is organized into five sections: Section A discusses the importance of border samples in TTA. Section B presents additional experimental results, including the substitution of human annotators with large models (Table A), and the robustness of our method across various backbones (Tables B and C). Section C offers further ablation studies, such as the effects of hyper-parameter variations (Table D). Section D provides detailed descriptions of the datasets used, including ImageNet-C, -A, -R, -K, and PACS. Finally, Section E explains the pre-training protocol and implementation details of TTA and ATTA baselines.

A. Importance of Border Samples in TTA

TTA adapts a pre-trained model in real-time based on online target data with significant domain shifts, which presents challenges from both noisy labels and efficiency requirements. ATTA mitigates the adverse effect of noisy labels by introducing human/large model annotations of a small set of samples. However, as we demonstrate in Figure 3, these samples could be difficult for the model to learn, thereby hindering adequate adaptation. To address this problem, we propose to select samples that are both informative and feasible to learn from a single-step optimization perspective. Specifically, we prefer samples that are distributed at the border between the source and target domains, as they provide learnable target domain knowledge and help the model gradually adapt to the target domain [5, 20].

Furthermore, we provide a toy example to compare two sampling strategies, where the source model is fine-tuned with data points (blue stars) that are far from and close to the source domain, respectively. To this end, we first train a toy model on red data and test it on green data. We then select two data points per class in the green data to fine-tune the pre-trained toy model and visualize the decision boundary in Figure A.

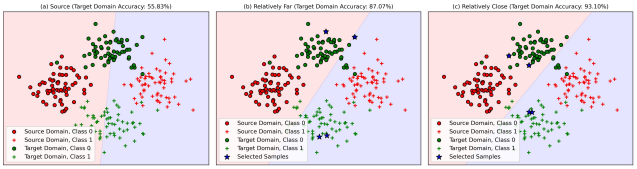

Figure A. Performance comparisons with two sampling strategies for model adaptation on a toy example. (a) The initial decision boundary of the source model. (b) The updated decision boundary using data points that are relatively far from the source domain. (c) The updated decision boundary using data points that are relatively close to the source domain.

As shown in Figure A, the model performs better when fine-tuned on data points close to the source domain, achieving 93.10%, compared to 87.07% when fine-tuned on points far from the source domain. This demonstrate border samples are the optimal ones to facilitate robust adaptation.

B. More Experimental Results

Replacing Human Annotator with Large Models. This experiment is an extension of Table 8 of the main paper. We perform experiments in the fully test-time adaptation setting and present the results in Table A. SimATTA [10] shows modest improvements over the baseline where a large model provides the annotations. However, it underperforms compared to the baseline where human experts provide annotations. In contrast, our method consistently outperforms the baseline, regardless of whether the annotations come from human experts or large models. This highlights the potential of our method to deploy in real-world applications.
Performance across Various Backbones. We provide experimental results on ImageNet-C for continual and fully test-time adaptation settings using ResNet-50 with GroupNorm (Table B) and ViT-B-16 (Table C), respectively. As shown in Table B, the basic version of our method (*i.e.*, Ours*) surpasses the baseline by 1.4% and 1.1% for CTTA and FTTA, respectively. Moreover, under the same annotation and buffer setting, our method outperforms SimATTA [10] by 6.3% and 4.4% for CTTA and FTTA, respectively. As shown in Table C, the basic version of our method surpasses the baseline by 1.9% and 0.4% for CTTA and FTTA, respectively. Moreover, under the same annotation and buffer setting, our method outperforms SimATTA by 3.1% and 3.6% for CTTA and FTTA, respectively. This demonstrates the robustness of our method across various backbones.

C. More Ablation Studies

All experiments in this section use the ResNet-50 with BatchNorm as the backbone.
Hyper-Parameters. We discuss the effect of several hyper-parameter variations and summarize them in Table D: ① the trade-off parameter α in Eq. 5 of the main paper; ② the standard deviation level σ in Gaussian noise perturbation.

As shown in Table D ①, omitting this trade-off parameter α for adjusting the two dynamic weights results in suboptimal performance (*e.g.*, $\alpha = 0$), highlighting the effectiveness of the proposed gradient norm-based debiasing. Furthermore, a larger α leads to a lower average error rate, highlighting its effectiveness in refining these weights and

Table A. Performance comparisons between different ATTA methods on ImageNet-C for fully test-time adaptation setting (*i.e.*, '**F**'). Except for Baseline*(GT), all methods adopt the ViT-L-16 model to annotate the selected samples. The Baseline*(GT) adopts ground-truth labels. "*" and "†" indicate 1 and 3 samples are annotated per batch, respectively. $\mathcal{BFS}$ is the buffer size. The **best** and second-best performances are highlighted.

		Noise			Blur				Weather				Digital				
F	Methods	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg. Err.
Active	• Baseline*(GT)	69.3	67.5	68.1	72.6	70.2	57.7	51.2	52.3	57.7	43.0	33.3	63.0	45.6	42.1	47.7	56.1
	• Baseline*	73.8	71.4	71.6	74.3	75.3	64.1	59.0	57.1	59.0	46.4	33.8	65.6	51.4	54.0	51.4	60.6
	• SimATTA† [10] ($\mathcal{BFS}=300$)	71.1	68.8	69.9	72.7	75.6	62.4	59.5	56.1	57.4	46.6	33.2	81.7	52.5	47.4	49.3	60.3
	• Ours*	67.3	65.4	66.1	69.5	70.1	56.8	**50.6**	50.8	56.0	42.4	32.7	63.7	**44.3**	**40.9**	46.2	54.9
	• Ours†	**66.7**	65.1	66.2	69.0	69.5	56.8	50.8	**50.7**	56.0	42.4	32.7	61.1	**44.5**	**41.3**	**46.2**	54.6
	• Ours† ($\mathcal{BFS}=300$)	66.9	**64.9**	**65.8**	**68.9**	**69.0**	**56.1**	51.6	**50.7**	**55.3**	**42.1**	**32.4**	**59.4**	44.6	41.8	46.3	**54.4**

Table B. Performance comparisons on ImageNet-C for continual (*i.e.*, **C**) and fully (*i.e.*, **F**) test-time adaptation settings. The backbone is ResNet-50 with GroupNorm. $\mathcal{BFS}$ is the buffer size. The **best** and second-best performances are highlighted.

		Noise			Blur				Weather				Digital				
C	Methods	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg. Err.
Non-Active	• Source	82.0	80.2	82.1	80.3	88.7	78.6	75.1	59.6	52.7	66.4	30.7	63.7	81.4	71.6	47.7	69.4
	• TENT [43]	95.1	99.6	99.8	95.6	99.8	99.7	99.7	99.2	99.8	99.8	99.0	99.9	99.9	99.9	99.4	99.1
	• CoTTA [47]	89.8	73.8	82.1	87.9	82.9	80.8	76.3	82.1	74.5	73.4	55.0	75.6	78.5	56.1	60.8	76.0
	• SAR [32]	71.8	58.2	56.1	83.6	80.0	86.3	96.8	98.9	70.4	50.0	29.6	53.4	86.2	93.7	98.1	74.2
	• ETA [31]	64.1	59.3	60.9	77.8	73.7	71.5	62.7	60.5	56.3	52.4	38.1	53.5	60.2	54.4	47.1	59.5
Active	• Baseline*	62.4	53.5	53.9	71.6	64.6	59.4	54.4	51.5	45.4	41.8	29.9	46.7	53.2	43.0	39.1	51.3
	• SimATTA† [10] ($\mathcal{BFS}=300$)	63.0	53.1	52.8	75.1	69.7	64.0	56.9	54.9	45.6	45.6	31.2	48.8	61.1	45.8	39.1	53.8
	• Ours*	61.7	51.9	52.4	71.8	63.6	58.3	52.2	48.8	43.8	40.2	28.3	45.1	50.8	40.8	37.9	49.9
	• Ours†	61.1	50.7	50.6	69.8	63.1	56.6	51.1	47.9	43.1	39.4	27.8	**44.2**	50.0	40.6	37.0	48.9
	• Ours† ($\mathcal{BFS}=300$)	**58.6**	**49.5**	**49.7**	**67.8**	**60.5**	**54.9**	**49.0**	**46.5**	**42.2**	**38.2**	**27.5**	44.3	**47.6**	**39.3**	**36.6**	**47.5**

		Noise			Blur				Weather				Digital				
F	Methods	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg. Err.
Non-Active	• Source	82.0	80.2	82.1	80.3	88.7	78.6	75.1	59.6	52.7	66.4	30.7	63.7	81.4	71.6	47.7	69.4
	• TENT [43]	95.1	93.9	94.2	85.2	89.7	77.7	77.8	73.0	65.7	96.8	29.7	57.8	88.6	51.8	45.6	74.8
	• CoTTA [47]	97.5	69.0	68.6	84.6	83.7	79.2	72.7	70.9	52.9	96.1	32.5	77.9	83.9	55.6	47.2	71.5
	• SAR [32]	71.7	68.8	70.1	81.4	81.3	69.4	69.7	59.0	56.8	95.1	29.3	56.3	82.7	51.3	44.8	65.8
	• ETA [31]	63.7	61.7	62.9	72.1	71.7	63.9	61.2	52.1	51.9	45.8	29.4	52.6	58.6	45.0	43.9	55.8
Active	• Baseline*	62.3	60.7	61.5	69.0	71.0	62.1	58.5	48.3	47.2	45.0	27.8	49.7	56.3	43.9	42.0	53.7
	• SimATTA†($\mathcal{BFS}=300$) [10]	62.2	60.7	61.4	70.5	71.7	63.2	59.2	48.1	46.1	45.0	27.9	51.4	59.0	47.2	42.1	54.4
	• Ours*	62.2	59.3	61.1	67.8	69.6	60.7	57.7	47.1	46.9	42.1	27.1	48.7	54.4	43.1	41.0	52.6
	• Ours†	61.1	58.9	60.0	67.3	69.1	60.1	57.3	46.3	46.5	42.0	27.1	48.4	54.0	42.7	41.0	52.1
	• Ours†($\mathcal{BFS}=300$)	**58.8**	**56.3**	**57.2**	**66.3**	**65.7**	**58.1**	**54.4**	**44.8**	**44.7**	**40.7**	**26.4**	**46.7**	**50.6**	**40.8**	**39.3**	**50.0**

Table C. Performance comparisons on ImageNet-C for continual (*i.e.*, **C**) and fully (*i.e.*, **F**) test-time adaptation settings. The backbone is ViT-B-16. $\mathcal{BFS}$ is the buffer size. The **best** and second-best performances are highlighted.

		Noise			Blur				Weather				Digital				
C	Methods	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg. Err.
Non-Active	• Source	66.0	66.8	65.0	68.5	74.7	64.0	66.9	57.3	45.0	49.4	28.7	81.8	57.8	60.8	49.9	60.2
	• TENT [43]	58.6	54.0	57.8	59.7	77.4	99.7	99.8	99.7	99.9	99.9	99.8	99.9	99.9	99.9	99.9	87.1
	• CoTTA [47]	63.5	64.2	70.4	93.8	84.5	85.8	78.6	99.6	99.9	99.9	99.8	99.9	99.9	99.9	99.9	89.3
	• SAR [32]	54.9	50.3	52.2	57.7	60.3	60.9	49.6	60.0	48.1	35.0	26.4	43.5	51.7	44.5	35.9	48.7
	• ETA [31]	54.2	50.3	52.3	57.5	53.7	50.2	50.8	50.7	47.2	44.4	31.4	57.8	44.6	42.4	41.7	48.6
Active	• Baseline*	55.0	49.4	49.9	54.7	50.3	46.4	46.1	46.9	41.0	39.4	28.2	48.6	41.1	38.3	37.5	44.9
	• SimATTA†($\mathcal{BFS}=300$) [10]	55.2	49.0	48.9	54.2	50.6	45.4	46.0	46.4	39.1	36.9	26.7	43.7	42.0	38.3	36.6	43.9
	• Ours*	54.5	48.7	48.7	53.9	49.5	44.1	43.4	46.3	38.8	36.1	26.4	46.0	38.2	34.7	35.2	43.0
	• Ours†	53.9	47.5	47.4	52.1	47.8	42.6	42.2	44.0	38.0	34.6	**25.6**	41.7	37.1	34.1	34.5	41.5
	• Ours†($\mathcal{BFS}=300$)	**53.0**	**46.7**	**46.9**	**50.9**	**46.8**	**42.0**	**40.6**	**43.0**	**37.7**	**33.7**	**25.6**	**41.4**	**36.0**	**33.6**	**33.6**	**40.8**

		Noise			Blur				Weather				Digital				
F	Methods	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg. Err.
Non-Active	• Source	66.0	66.8	65.0	68.5	74.7	64.0	66.9	57.3	45.0	49.4	28.7	81.8	57.8	60.8	49.9	60.2
	• TENT [43]	58.6	56.3	56.8	58.1	61.7	52.3	56.6	72.2	43.5	93.2	26.7	54.5	50.0	41.8	41.7	54.9
	• CoTTA [47]	63.0	78.6	71.4	96.3	95.3	97.8	88.5	97.9	94.8	45.9	94.0	98.0	78.1	94.6	50.8	83.0
	• SAR [32]	54.9	52.9	53.8	53.4	53.9	46.0	50.4	54.3	40.5	38.3	25.7	55.0	41.4	35.9	37.0	46.2
	• ETA [31]	54.2	52.4	53.3	52.6	51.6	45.1	46.1	41.9	40.1	36.0	25.5	46.5	38.7	35.0	36.2	43.7
Active	• Baseline*	55.0	53.0	53.8	53.1	53.1	46.8	48.6	43.7	39.3	35.6	26.0	46.5	40.9	36.8	38.1	44.7
	• SimATTA†($\mathcal{BFS}=300$) [10]	55.2	53.2	54.0	55.1	54.8	49.1	50.9	44.9	39.6	36.5	26.5	48.1	44.0	39.6	39.1	46.0
	• Ours*	54.5	52.6	53.6	53.4	53.2	46.2	47.8	43.5	39.4	36.5	25.8	46.4	39.5	35.8	36.9	44.3
	• Ours†	53.9	52.0	52.9	52.1	52.4	45.4	47.3	42.3	38.7	35.8	25.4	44.1	39.3	35.5	36.7	43.6
	• Ours† ($\mathcal{BFS}=300$)	**53.0**	**51.1**	**52.2**	**51.2**	**50.2**	**44.2**	**45.3**	**41.1**	**37.7**	**34.3**	**25.0**	**43.1**	**37.5**	**34.5**	**35.3**	**42.4**

enabling stable model adaptation during long-term distribution shifts.

As shown in Table D ②, our method yields similar results when a small standard deviation is applied, indicating that slight perturbation is sufficient for selecting optimal samples. Conversely, a large standard deviation could significantly alter the model's predictions for samples that do not borders between the source- and target-domain data distributions, thereby weakening the discriminative power of our method.

Table D. Performance comparisons on hyper-parameter variations, including ① the trade-off parameter α, and ② the standard deviation level σ. The **best** and second-best performances are highlighted.

	Variants	ImageNet-C	ImageNet-R	ImageNet-K	ImageNet-A	Avg. Err.
①	0.0	57.3	52.6	65.5	99.0	68.6
	0.2	54.1	51.7	65.1	98.2	67.3
	0.4	54.3	51.6	64.5	98.4	67.2
	0.6	54.5	**51.1**	64.7	**98.1**	67.1
	0.8	**53.8**	51.5	**64.3**	**98.1**	**66.9**
②	0.01	53.8	51.5	**64.3**	**98.1**	**66.9**
	0.02	**53.7**	**51.4**	**64.3**	98.3	**66.9**
	0.03	54.1	51.8	64.9	98.5	**66.9**
	0.1	53.8	51.6	64.7	**98.1**	67.1
	1.0	54.6	52.9	66.8	98.0	68.1

D. Dataset Details

ImageNet-C. ImageNet-C [12] is a dataset derived from the validation set of the original ImageNet with common corruptions and perturbations, such as 'Gaussian Noise', 'Shot Noise', 'Impulse Noise', 'Defocus Blur', 'Glass Blur', 'Motion Blur', 'Zoom Blur', 'Snow', 'Frost', 'Fog', 'Brightness', 'Contrast', 'Elastic Transform', 'Pixelate', and 'JPEG Compression'. Each corruption type is applied at five levels of severity, resulting in 50,000 images per corruption type. Overall, the dataset comprises 750,000 images across 1,000 classes.

ImageNet-Rendition (R). ImageNet-R [13] is a dataset with diverse artistic renditions, such as cartoons, paintings, origami, embroidery, toys, sculptures, and so on. It features renditions of 200 ImageNet classes, comprising a total of 30,000 images.

ImageNet-Sketch (K). ImageNet-K [45] is a dataset designed to provide sketch-based representations of objects belonging to the ImageNet database. It consists of hand-drawn sketches corresponding to 50,000 images from 1,000 different categories in ImageNet.

ImageNet-A. ImageNet-A [14] provides natural adversarial examples that are challenging for models to recognize correctly while still being visually similar to the original classes. It contains 7,500 images across 200 categories. Each category corresponds to a class from the original ImageNet dataset.

PACS. PACS [23] includes a total of 9,991 images across four domains, such as 'Photo', 'Art Painting', 'Cartoon' and 'Sketch'. Each domain contains seven categories.

E. More Experimental Details

E.1. Pre-training Protocol on PACS

We employ the ResNet-18 with pre-trained weights, specifically 'ResNet18_Weights.DEFAULT' from PyTorch. Following [10] and [11], we fix the statistics in the batch normalization layers in the pre-trained model. We set the batch size to 32 and train the model for 40 epochs using the Adam optimizer, with a learning rate of 0.0001 and a weight decay of 5e-5.

E.2. Implementation Details of TTA and ATTA Baselines

TENT. For TENT [43], we use the SGD optimizer with a learning rate of 0.00025 and a momentum of 0.9 on ImageNet-C, -R, -K, and -A. We use the Adam optimizer with a learning rate of 0.005 on PACS. The implementation follows the official code[1].

CoTTA. For CoTTA [47], we use the SGD optimizer with a learning rate of 0.01 and a momentum of 0.9 on ImageNet-C, -R, -K, and -A. And the restoration factor, exponential moving average factor, the average probability threshold, and the augmentation number are set to 0.01, 0.999, 0.1, 32, respectively. Moreover, we use the Adam optimizer with a learning rate of 0.01 on PACS. And the restoration factor, exponential moving average factor, the average probability threshold, and the augmentation number are set to 0.01, 0.999, 0.72, 32, respectively. The implementation follows the official code[2].

ETA. For ETA [31], we use the SGD optimizer with a learning rate of 0.00025 and a momentum of 0.9 on ImageNet-C, -R, -K, and -A. Moreover, we use the Adam optimizer with a learning rate of 0.001 on PACS. We set the exponential moving average factor, the cosine similarity threshold, and the entropy threshold to 0.9, 0.05, and $0.4 \times \ln(\mathcal{C})$, respectively. Here, $\mathcal{C}$ is the number of classes. The implementation follows the official code[3].

SAR. For SAR [32], we use the SAM optimizer with a learning rate of 0.001 and a momentum of 0.9 on ImageNet-C, -R, -K, and -A. Moreover, we use the Adam optimizer with a learning rate of 0.001 on PACS. We set the reset factor, the entropy threshold, and the exponential moving average factor to 0.2, $0.4 \times \ln(\mathcal{C})$, and 0.9, respectively, for all datasets. The implementation follows the official code[4].

[1] https://github.com/DequanWang/tent
[2] https://github.com/qinenergy/cotta
[3] https://github.com/mr-eggplant/EATA
[4] https://github.com/mr-eggplant/SAR

SimATTA. For SimATTA [10], we employ the SGD optimizer with a learning rate of 0.00025 and a momentum of 0.9 on ImageNet-C, -R, -K, and -A databases. For PACS, we use the Adam optimizer with a learning rate of 0.005. The maximum length of anchors is set to 50, and the entropy threshold is set to $0.4 \times \ln(\mathcal{C})$. This adjustment to the entropy threshold is necessary because the original threshold is not appropriate for ImageNet-C, leading to suboptimal performance. The buffer size is fixed to 300 for fair comparison. The implementation follows the official code[5].

CEMA. For CEMA [6], we employ the SGD optimizer with a learning rate of 0.00025 and a momentum of 0.9 on ImageNet-C, -R, -K, and -A databases. And the maximum entropy threshold, the minimum entropy threshold, and the decreasing factor are set to $0.4 \times \ln(\mathcal{C})$, $0.02 \times \ln(\mathcal{C})$, and 1.0, respectively. The buffer size is set to 300 for fair comparison. The implementation follows the official code[6].

HILTTA. For HILTTA [26], we use the experimental results reported in the original paper.

Baseline. The Baseline method, which builds on TENT, randomly selects a specified number of samples from each online batch for manual annotation and then performs ATTA using Eq. 2 of the main paper. We employ the SGD optimizer with a learning rate of 0.00025 and a momentum of 0.9 on ImageNet-C, -R, -K, and -A databases. We use the Adam optimizer with a learning rate of 0.005 on PACS.

[5] https://github.com/divelab/ATTA

[6] https://github.com/chenyaofo/CEMA