NPU
Northwestern
Polytechnical University

SaTeen: Learning Structural Alignment for Continual Test-Time Adaptation

Chang Liu, Ruotong Zhao, Li Gao, Yupei Zhang
Northwestern Polytechnical University | MIIT Lab of Big Data Storage and Management | ICML 2026

Code / Paper
github.com/scienceliuc/SaTeen
QR can be added after the final project URL is fixed.

Entropy-only adaptation collapses under imbalance

Test-time adaptation (TTA) updates a source model online from unlabeled target samples, with each sample used once. Most methods select low-uncertainty samples and minimize entropy, but shifted streams can become prediction-imbalanced.

Failure mechanism

noisy features → imbalanced prediction → collapse

Key diagnosis

If a dominant predicted class grows, entropy minimization can keep pushing updates along a collapse direction instead of recovering source structure.

SaTeen uses structure, not uncertainty alone

- Intra-sample alignment compares each test image with a patch-shuffled counterpart.
- It encourages attention to object structure rather than corruption artifacts.
- Inter-sample alignment builds an online IPCA subspace from reliable target features.
- Incoming samples are aligned to this stable stream manifold.

L_SaTeen(x) = alpha(x) [I(x in Gamma) L_1(x) + lambda_2 L_2(x)]

Biased shortcuts: strong worst-group gains

ColoredMNIST	Avg	Worst-Group
ResNet18-BN	$65.3_{\pm0.1}$	$20.1_{\pm0.1}$
Tent (Wang et al., 2021)	$59.8_{\pm0.2}$	$13.0_{\pm0.7}$
SENTRY (Prabhu et al., 2021)	$63.2_{\pm0.4}$	$13.0_{\pm0.8}$
EATA (Niu et al., 2022)	$63.4_{\pm0.2}$	$20.8_{\pm0.2}$
SAR (Niu et al., 2023)	$61.3_{\pm0.2}$	$15.6_{\pm0.5}$
DeYO (Lee et al., 2024)	$\underline{76.4}_{\pm0.3}$	$\underline{52.7}_{\pm0.4}$
SaTeen (Ours)	$\mathbf{79.3}_{\pm0.3}$	$\mathbf{67.2}_{\pm0.3}$
Waterbirds	**Avg**	**Worst-Group**
ResNet50-BN	$84.3_{\pm0.1}$	$67.0_{\pm0.1}$
Tent (Wang et al., 2021)	$83.2_{\pm0.6}$	$54.5_{\pm2.0}$
SENTRY (Prabhu et al., 2021)	$85.7_{\pm0.6}$	$61.0_{\pm1.4}$
EATA (Niu et al., 2022)	$83.0_{\pm0.4}$	$54.0_{\pm1.3}$
SAR (Niu et al., 2023)	$83.1_{\pm0.5}$	$53.8_{\pm1.4}$
DeYO (Lee et al., 2024)	$\underline{88.2}_{\pm0.7}$	$\underline{74.6}_{\pm1.3}$
SaTeen (Ours)	$\mathbf{89.0}_{\pm0.6}$	$\mathbf{76.3}_{\pm1.5}$

Mixed shifts	Lev.5	Lev.3	Avg.
ResNet50-GN	30.6	54.0	42.3
DeYO	38.6	59.2	48.9
ReCAP	41.7	59.0	50.4
SaTeen	**43.7**	**60.5**	**52.1**
ViT DeYO	59.4	72.1	65.8
ViT SaTeen	**60.3**	**72.3**	**66.3**

Mixed ImageNet-C shifts, Table 2. SaTeen gives the best average on both ResNet50-GN and ViT-Base-LN.

+14.5 worst-group on ColoredMNIST over DeYO

+1.7 worst-group on Waterbirds over DeYO

Visible mechanism: attention and manifolds align

Learning Structural Alignment for Continual Test-Time Adaptation

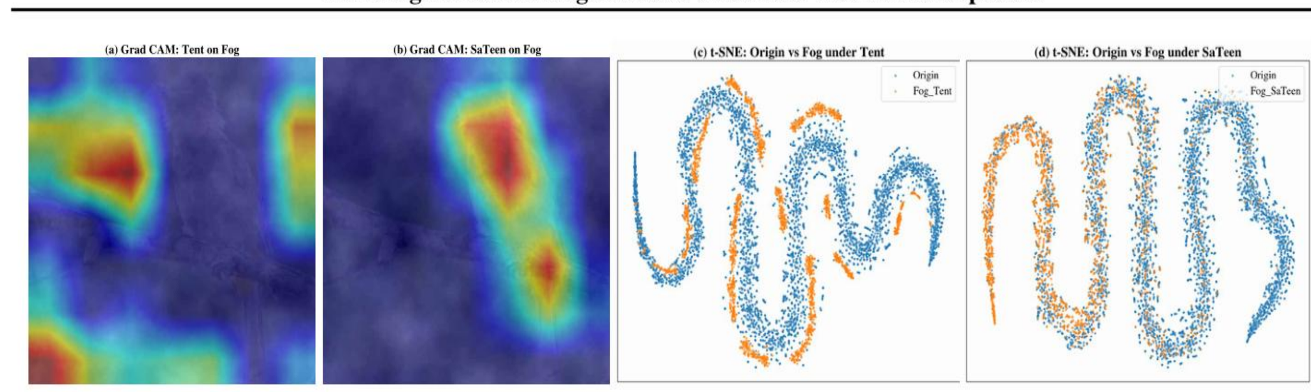
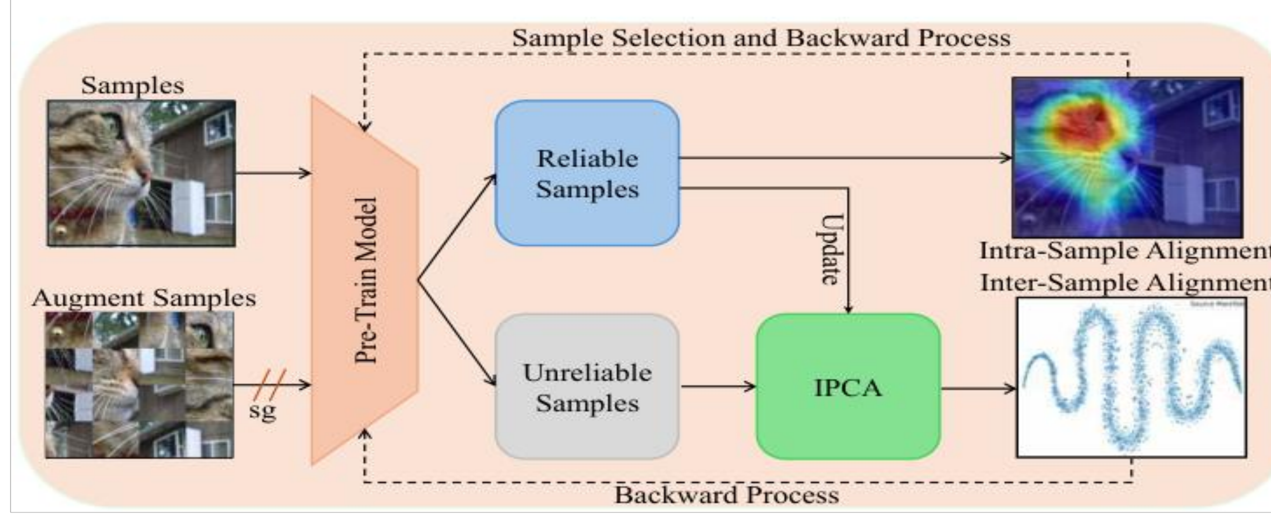


Figure 4. Visualizations of structure alignment. (a) and (b) show the Grad-CAM visualizations of a 'fog' image from ImageNet-C after

SaTeen shifts Grad-CAM toward object structure and makes source/target t-SNE manifolds less separable.

SaTeen framework: reliable samples update two alignments



Reliable samples: intra-sample + inter-sample alignment. Unreliable samples: inter-sample alignment only.

ImageNet-C severity 5: single-domain TTA

Table 3. Accuracy (%) under single TTA on ImageNet-C across corruption types at severity level 5.

Single TTA	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impl.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	
ResNet50-GN	15	19.8	17.9	19.7	11.3	21.4	24.9	40.4	47.3	33.6	69.3	36.3	18.6	28.4	52.2	30.6
Tent (Wang et al., 2021)	5.4	6.1	5.8	14.8	6.5	21.8	22.2	25.8	33.6	3.7	70.5	42.4	11.7	48.2	54.3	33.5
EATA (Niu et al., 2022)	28.9	30.9	31.2	19.7	19.0	31.0	32.1	41.6	42.4	20.3	70.1	43.7	16.9	49.7	56.2	36.9
SAR (Niu et al., 2023)	28.4	31.6	29.9	18.5	18.1	30.8	30.4	42.0	43.6	5.2	70.7	44.0	17.0	48.7	55.2	34.3
DeYO (Lee et al., 2024)	39.0	41.9	40.9	21.9	23.7	38.7	<u>37.6</u>	50.8	49.3	3.5	73.1	50.1	42.0	56.1	57.7	42.4
ReCAP (Hu et al., 2025)	40.5	42.7	41.8	21.4	2.7	40.1	**41.1**	16.8	48.6	56.2	72.9	50.5	3.5	56.3	57.8	39.5
SaTeen (Intra-only)	<u>41.1</u>	<u>44.0</u>	<u>42.9</u>	<u>25.3</u>	<u>25.7</u>	<u>41.2</u>	37.5	<u>53.5</u>	<u>51.6</u>	<u>57.7</u>	<u>73.2</u>	<u>52.3</u>	<u>46.1</u>	<u>58.6</u>	<u>58.8</u>	$\underline{47.1}_{\pm0.2}$
SaTeen (Intra and Inter)	**42.2**	**45.4**	**44.3**	**32.4**	**29.9**	**44.0**	**41.1**	**56.6**	**53.9**	**60.3**	**73.7**	**54.0**	**49.9**	**60.4**	**59.8**	$\mathbf{49.9}_{\pm0.2}$
VitBase-LN	9.4	6.7	8.3	29.1	23.4	34.0	27.1	15.8	26.4	47.4	54.7	44.0	30.5	44.5	47.6	29.9
Tent (Wang et al., 2021)	42.3	1.7	43.4	51.9	47.9	55.1	50.1	19.3	20.5	66.4	74.7	64.7	52.1	66.7	64.1	48.1
EATA (Niu et al., 2022)	44.1	15.0	46.1	53.2	49.7	54.9	52.6	55.3	55.4	67.1	73.7	64.6	55.4	66.9	64.3	54.5
SAR (Niu et al., 2023)	44.3	13.0	45.5	52.9	50.0	55.7	51.4	56.4	55.2	66.4	74.7	64.4	55.1	66.7	64.0	54.4
DeYO (Lee et al., 2024)	53.0	53.2	53.8	58.1	58.5	63.2	58.5	67.0	65.8	<u>73.4</u>	<u>78.2</u>	<u>68.1</u>	67.7	73.2	70.0	64.1
ReCAP (Hu et al., 2025)	47.6	45.7	46.0	56.2	55.3	60.8	57.0	63.7	63.5	72.0	77.3	67.3	63.8	71.6	68.6	61.5
SaTeen (Intra-only)	<u>54.3</u>	<u>55.0</u>	<u>55.5</u>	<u>58.8</u>	<u>59.4</u>	<u>64.6</u>	<u>60.3</u>	<u>68.3</u>	<u>66.4</u>	**73.6**	<u>78.2</u>	**68.3**	<u>68.9</u>	<u>73.9</u>	<u>70.6</u>	$\underline{65.1}_{\pm0.1}$
SaTeen (Intra and Inter)	**54.7**	**55.4**	**55.6**	**59.2**	**59.9**	**64.9**	**61.9**	**68.5**	**66.6**	**73.6**	**78.4**	<u>68.1</u>	**69.1**	**74.1**	**70.7**	$\mathbf{65.4}_{\pm0.1}$

ResNet50-GN: 49.9 vs 42.4 DeYO

ViT-Base-LN: 65.4 vs 64.1 DeYO

Continual TTA: gains grow when shifts arrive as a stream

EATA (Niu et al., 2022)	22.1	32.7	36.9	18.9	20.9	23.1	26.3	39.9	44.6	45.7	67.6	42.3	24.1	42.6	54.9	36.2
SAR (Niu et al., 2023)	23.1	33.5	37.3	18.6	19.9	22.9	26.0	39.4	43.6	44.7	67.1	41.5	23.7	42.0	53.6	35.8
DeYO (Lee et al., 2024)	28.0	<u>41.1</u>	<u>43.6</u>	17.2	21.2	24.7	31.7	42.0	47.1	51.0	70.0	48.2	27.2	51.6	56.8	40.1
ReCAP (Hu et al., 2025)	19.1	24.7	27.3	20.4	15.9	26.5	29.6	42.8	44.3	48.5	70.0	44.3	22.0	45.0	55.9	36.4
SaTeen (Intra-only)	**37.1**	39.8	39.0	<u>20.6</u>	<u>22.6</u>	<u>36.8</u>	<u>35.4</u>	<u>47.3</u>	47.1	<u>53.9</u>	**72.5**	<u>48.8</u>	<u>36.6</u>	<u>54.1</u>	<u>56.9</u>	$\underline{43.2}_{\pm1.2}$
SaTeen (Intra and Inter)	<u>35.0</u>	**47.8**	**47.9**	**27.1**	**30.1**	**38.8**	**38.8**	**48.9**	**53.5**	**57.9**	<u>70.1</u>	**51.7**	**41.1**	**58.3**	**59.4**	$\mathbf{47.1}_{\pm1.0}$
VitBase-LN	47.0	48.2	47.9	31.5	21.2	41.5	36.7	50.1	45.8	42.3	73.6	8.6	42.5	62.0	63.8	44.2
Tent (Wang et al., 2021)	47.8	51.1	50.8	34.2	27.0	45.5	41.6	56.0	52.3	49.7	76.1	27.2	45.3	65.6	66.1	49.0
CoTTA (Wang et al., 2022)	47.1	48.4	48.6	31.7	21.9	42.9	38.0	51.8	47.3	44.7	74.1	10.0	43.6	63.6	64.8	45.2
EATA (Niu et al., 2022)	50.3	55.8	55.3	42.5	40.2	54.6	49.2	59.1	56.2	59.0	75.9	42.9	51.1	67.5	67.9	55.1
SAR (Niu et al., 2023)	50.7	56.2	55.1	41.8	39.1	53.9	48.2	58.7	55.9	58.2	76.2	42.8	50.1	67.1	67.3	54.8
DeYO (Lee et al., 2024)	53.0	<u>57.7</u>	<u>58.4</u>	54.1	58.1	<u>61.6</u>	35.5	62.4	63.2	70.4	76.8	64.7	67.2	<u>71.7</u>	69.3	61.6
ViDA (Liu et al., 2024b)	52.3	57.5	57.1	47.8	43.1	54.5	51.1	61.1	57.3	59.3	75.7	47.2	50.9	66.5	66.9	56.6
CMAE (Liu et al., 2024a)	<u>53.7</u>	58.1	57.5	48.6	45.1	56.7	<u>59.3</u>	<u>65.8</u>	64.2	35.7	76.6	39.7	62.5	70.8	68.6	57.5
ReCAP (Hu et al., 2025)	37.9	47.8	52.9	49.1	50.9	56.3	53.1	58.4	61.5	65.8	76.3	62.8	57.9	67.3	67.2	57.7
REM (Han et al., 2025)	**56.5**	**61.9**	**60.8**	46.8	51.0	56.5	57.2	62.5	<u>64.8</u>	64.6	76.8	53.2	58.4	71.1	<u>69.8</u>	60.8
SaTeen (Intra-only)	51.1	56.8	58.0	<u>54.5</u>	<u>58.1</u>	61.3	54.1	62.7	64.2	<u>70.5</u>	<u>76.9</u>	<u>65.4</u>	<u>66.0</u>	71.4	69.2	$\underline{62.7}_{\pm0.8}$
SaTeen (Intra and Inter)	53.6	53.8	54.4	**58.7**	**59.2**	**63.9**	**60.5**	**67.5**	**66.1**	**73.0**	**78.1**	**68.6**	**67.9**	**73.5**	**70.1**	$\mathbf{64.6}_{\pm0.5}$

two benchmarks with extreme spurious correlation shifts. We conduct single TTA and continual TTA experiments on ImageNet-C under mild and wild (e.g., imbalance la-

trainable parameters of SaTeen during TTA, we adapt the affine parameters of batch/group/layer normalization layers in ResNet18-BN/ResNet50-GN/VitBase-LN, similar to the

ResNet50-GN: 47.1 vs 40.1 DeYO, 8.2 Tent

ViT-Base-LN: 64.6 average

Why the two alignments matter

Intra-sample

Prevents update directions that overfit corruption/noisy features.

Full single TTA: 49.9

Inter-sample

Uses reliable target features to track a stable online subspace.

Full continual TTA: 47.1

Hyperparameters are stable around the chosen setting

Learning Structural Alignment for Continual Test-Time Adaptation

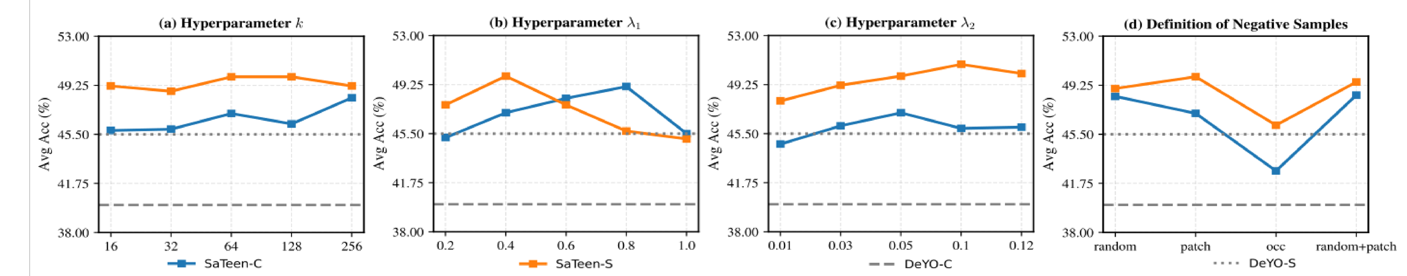


Figure 6. Hyperparameter experiments of SaTeen on ImageNet-C under single and continual TTA at severity 5.

Figure 6: SaTeen remains strong across k, lambda1, lambda2, and negative-sample definitions.

Ablation: all components are needed

Variant	L1	L2	tau	alpha	Avg.
Tent	-	-	-	-	8.2
DeYO	-	-	-	-	40.1
L1+tau	Y	-	Y	-	38.5
L1+tau+alpha	Y	-	Y	Y	43.2
SaTeen	**Y**	**Y**	**Y**	**Y**	**47.1**

Main ablation claim

Tent collapses to 8.2 in continual TTA at ImageNet-C severity 5; full SaTeen reaches 47.1 by combining structural losses, thresholding, and weighting.

8.2 -> 47.1 continual TTA average accuracy

Efficiency: strong accuracy without extra heavy adapters

Method	Params	Time	Acc.
ViTBase-LN	0	-	44.2
DeYO	0.04M	10m17s	61.6
REM	0.03M	17m21s	60.8
ReCAP	0.04M	13m45s	57.7
SaTeen	**0.04M**	**12m21s**	**64.6**

Table 6: SaTeen keeps trainable parameters low while reaching the best accuracy.

Takeaways and limitation

What to remember

- Entropy-only TTA is brittle when predictions become imbalanced.
- Patch-shuffled negatives encourage structure-aware updates.
- Online IPCA reduces error accumulation in continual streams.

Limitation

On Places365-C, prediction imbalance is weak and SaTeen is not the top continual-TTA method, suggesting negative-sample suppression is most useful when collapse pressure is present.

Built from the ICML-style PPTX template and grounded in the SaTeen paper figures/tables. Replace the code URL/QR placeholder after the final project link is fixed.