Neural Collapse in Test-Time Adaptation

Supplementary Material

A. Details about Neural Collapse

A.1. Additional Neural Collapse Properties

In Section 3.2, we presented the mathematical formalization of NC3. In this appendix, we offer a concise yet comprehensive overview of the remaining three hallmark phenomena that collectively manifest in deep classifiers during the TPT. We begin by introducing the necessary notation and terminology, proceed with a detailed explanation of each NC property, and conclude by elucidating their intrinsic interconnections.

Notation. Let $\mathbf{h}_i^c \in \mathbb{R}^L$ be the feature embedding of the i-th sample in class c. We define the class mean and global mean as

$$\mu_c \;=\; \frac{1}{n_c}\sum_{i=1}^{n_c}\mathbf{h}_i^c, \qquad \mu_G \;=\; \frac{1}{K}\sum_{c=1}^{K}\mu_c, \tag{1}$$

where n_c is the number of samples in class c and K is the total number of classes. The within-class covariance matrix is

$$\Sigma_W \;=\; \frac{1}{N}\sum_{c=1}^{K}\sum_{i=1}^{n_c}\big(\mathbf{h}_i^c-\mu_c\big)\big(\mathbf{h}_i^c-\mu_c\big)^\top, \quad N=\sum_{c=1}^{K} n_c. \tag{2}$$

Empirically, as the training loss approaches zero, the following four properties arise in concert.

Variability Collapse (NC1). All within-class feature scatter vanishes, so that

$$\Sigma_W \;\longrightarrow\; 0. \tag{3}$$

In other words, features from the same class concentrate precisely at their class mean.

Simplex Equiangular Tight Frame (NC2). After centering, the class means become equidistant and maximally separated, forming a regular simplex:

$$\big\|\mu_c-\mu_G\big\|_2^2 \;-\; \big\|\mu_{c'}-\mu_G\big\|_2^2 \;\longrightarrow\; 0, \tag{4}$$

$$\big\langle \tilde{\mu}_c, \tilde{\mu}_{c'}\big\rangle \;\longrightarrow\; \frac{K}{K-1}\,\delta_{c,c'} \;-\; \frac{1}{K-1}, \tag{5}$$

where $\tilde{\mu}_c = \mu_c - \mu_G$ and $\delta_{c,c'}$ is the Kronecker delta. This configuration maximizes the pairwise angular separation between class centroids.

Convergence to Self-Duality (NC3). Remarkably, the learned classifier weights $\{w_c\}_{c=1}^K$ and the class means $\{\mu_c\}_{c=1}^K$ align up to a global rotation and scaling, see Section 3.2 for details.

Simplification to Nearest Class-Center (NC4). As a consequence of NC1–NC3, the softmax decision rule reduces to a nearest-centroid classifier in feature space:

$$\arg\max_{c'}\langle w_{c'}, \mathbf{h}_i\rangle \;\longrightarrow\; \arg\min_{c'}\big\|\mathbf{h}_i-\mu_{c'}\big\|_2. \tag{6}$$

Thus, inference is geometrically interpretable as assigning each feature to its closest class mean.

Interplay of NC Phenomena. These properties collectively describe a systematic progression during training: 1) the intra-class variability collapses, leading to tightly clustered feature representations for each class (NC1); 2) the class means organize into the vertices of a maximally symmetric simplex (NC2); 3) the classifier weights align with these simplex vertices, mirroring their geometric structure (NC3); and 4) the network's decision rule simplifies to a nearest-center classification scheme (NC4).

This unified framework provides a compelling explanation for why, in the overparameterized regime, deep classifiers converge to solutions that are not only highly symmetric but also remarkably simple, despite the inherent complexity of the training process.

A.2. Empirical validation of NC3+

In Section 3.2, we propose Sample-wise Alignment Collapse (NC3+), with a theoretical proof provided in Appendix A.3. Experiments are conducted on CIFAR-10, CIFAR-100, and ImageNet100. The backbones for CIFAR-10 and CIFAR-100 are ResNet50 and ViT-S/16, while the backbones for ImageNet100 are ResNet50 and ViT-B/16. All models are trained for 200 epochs, and for each epoch, we compute:

$$\bar{d}_{iy_i} = \frac{1}{N_{\text{train}}}\sum_{i=1}^{N_{\text{train}}} d_{iy_i}, \tag{7}$$

$$\bar{d}_{ij} = \frac{1}{N_{\text{train}}\times(K-1)}\sum_{i=1}^{N_{\text{train}}}\sum_{j=0,j\neq y_i}^{K-1} d_{ij}, \tag{8}$$

and plot the line charts accordingly, as shown in Figure 1.

Table 1. Comparisons with baselines on CIFAR-10-C at severity level 5 regarding accuracy (%). The **bold** value signifies the top-performing result and the underline is the second-best.

	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg.
no_adapt	33.69	40.07	35.74	53.11	49.86	67.04	56.85	73.89	62.59	64.66	89.00	43.57	74.37	41.96	74.43	57.39
BN_adapt	62.43$_{\pm 0.02}$	64.33$_{\pm 0.02}$	57.30$_{\pm 0.08}$	84.48$_{\pm 0.00}$	59.46$_{\pm 0.03}$	82.02$_{\pm 0.04}$	83.25$_{\pm 0.03}$	76.87$_{\pm 0.01}$	74.30$_{\pm 0.01}$	79.29$_{\pm 0.00}$	87.24$_{\pm 0.02}$	83.33$_{\pm 0.02}$	71.72$_{\pm 0.01}$	73.41$_{\pm 0.01}$	71.10$_{\pm 0.02}$	74.03
Tent	64.71$_{\pm 0.03}$	66.97$_{\pm 0.03}$	59.61$_{\pm 0.01}$	84.97$_{\pm 0.03}$	60.96$_{\pm 0.04}$	82.48$_{\pm 0.04}$	83.93$_{\pm 0.01}$	77.78$_{\pm 0.01}$	75.42$_{\pm 0.00}$	80.00$_{\pm 0.04}$	87.78$_{\pm 0.01}$	83.54$_{\pm 0.00}$	72.62$_{\pm 0.05}$	74.73$_{\pm 0.01}$	72.32$_{\pm 0.03}$	75.19
EATA	62.43$_{\pm 0.02}$	64.33$_{\pm 0.02}$	57.31$_{\pm 0.08}$	84.48$_{\pm 0.00}$	59.46$_{\pm 0.03}$	82.02$_{\pm 0.04}$	83.25$_{\pm 0.02}$	76.88$_{\pm 0.01}$	74.31$_{\pm 0.01}$	79.28$_{\pm 0.00}$	87.24$_{\pm 0.02}$	83.33$_{\pm 0.02}$	71.72$_{\pm 0.01}$	73.42$_{\pm 0.01}$	71.13$_{\pm 0.02}$	74.04
SAR	63.74$_{\pm 0.07}$	65.87$_{\pm 0.06}$	58.47$_{\pm 0.02}$	84.70$_{\pm 0.01}$	60.51$_{\pm 0.02}$	82.22$_{\pm 0.02}$	83.62$_{\pm 0.02}$	77.26$_{\pm 0.03}$	74.74$_{\pm 0.02}$	79.74$_{\pm 0.02}$	87.40$_{\pm 0.01}$	83.29$_{\pm 0.02}$	72.24$_{\pm 0.03}$	74.15$_{\pm 0.00}$	71.87$_{\pm 0.02}$	74.67
NOTE	55.17$_{\pm 0.03}$	60.13$_{\pm 0.15}$	55.13$_{\pm 0.06}$	74.53$_{\pm 0.00}$	59.43$_{\pm 0.08}$	78.61$_{\pm 0.04}$	75.72$_{\pm 0.08}$	79.90$_{\pm 0.00}$	76.46$_{\pm 0.01}$	78.18$_{\pm 0.05}$	**90.91**$_{\pm 0.01}$	70.43$_{\pm 0.04}$	76.18$_{\pm 0.03}$	58.91$_{\pm 0.12}$	75.83$_{\pm 0.00}$	71.03
MEMO	52.22$_{\pm 0.00}$	58.09$_{\pm 0.00}$	51.40$_{\pm 0.00}$	73.17$_{\pm 0.00}$	58.01$_{\pm 0.00}$	76.50$_{\pm 0.01}$	74.61$_{\pm 0.00}$	77.74$_{\pm 0.00}$	73.01$_{\pm 0.00}$	75.47$_{\pm 0.00}$	90.27$_{\pm 0.00}$	63.89$_{\pm 0.00}$	**76.95**$_{\pm 0.01}$	55.86$_{\pm 0.00}$	75.62$_{\pm 0.00}$	68.85
EATA-C	62.43$_{\pm 0.17}$	64.33$_{\pm 0.19}$	57.30$_{\pm 0.35}$	84.48$_{\pm 0.01}$	59.46$_{\pm 0.21}$	82.02$_{\pm 0.25}$	83.25$_{\pm 0.20}$	76.88$_{\pm 0.12}$	74.30$_{\pm 0.13}$	79.28$_{\pm 0.06}$	87.24$_{\pm 0.19}$	83.33$_{\pm 0.18}$	71.72$_{\pm 0.15}$	73.41$_{\pm 0.14}$	71.10$_{\pm 0.18}$	74.04
SAR^2	64.40$_{\pm 0.34}$	66.56$_{\pm 0.22}$	59.26$_{\pm 0.37}$	84.83$_{\pm 0.11}$	60.91$_{\pm 0.11}$	82.38$_{\pm 0.22}$	83.81$_{\pm 0.20}$	77.44$_{\pm 0.11}$	74.95$_{\pm 0.12}$	79.84$_{\pm 0.19}$	87.51$_{\pm 0.11}$	83.35$_{\pm 0.15}$	72.53$_{\pm 0.16}$	74.42$_{\pm 0.10}$	72.26$_{\pm 0.15}$	74.96
AdaDEM	64.06$_{\pm 0.18}$	66.29$_{\pm 0.23}$	58.92$_{\pm 0.20}$	84.80$_{\pm 0.08}$	60.73$_{\pm 0.11}$	82.30$_{\pm 0.20}$	83.72$_{\pm 0.20}$	77.41$_{\pm 0.12}$	74.89$_{\pm 0.17}$	79.77$_{\pm 0.15}$	87.53$_{\pm 0.10}$	83.35$_{\pm 0.12}$	72.37$_{\pm 0.17}$	74.34$_{\pm 0.04}$	72.03$_{\pm 0.25}$	74.83
COME	62.80$_{\pm 0.22}$	65.01$_{\pm 0.04}$	57.85$_{\pm 0.22}$	84.68$_{\pm 0.11}$	59.95$_{\pm 0.17}$	82.17$_{\pm 0.18}$	83.49$_{\pm 0.12}$	77.17$_{\pm 0.07}$	74.67$_{\pm 0.13}$	79.53$_{\pm 0.09}$	87.38$_{\pm 0.17}$	83.55$_{\pm 0.13}$	71.90$_{\pm 0.15}$	73.89$_{\pm 0.17}$	71.50$_{\pm 0.10}$	74.37
DeYO	67.61$_{\pm 0.27}$	70.15$_{\pm 0.02}$	63.15$_{\pm 0.15}$	85.59$_{\pm 0.00}$	62.98$_{\pm 0.03}$	83.09$_{\pm 0.13}$	84.57$_{\pm 0.00}$	79.05$_{\pm 0.00}$	76.35$_{\pm 0.02}$	80.68$_{\pm 0.03}$	88.17$_{\pm 0.00}$	83.57$_{\pm 0.00}$	74.16$_{\pm 0.04}$	76.45$_{\pm 0.02}$	74.13$_{\pm 0.08}$	76.65
NCTTA (ours)	**70.94**$_{\pm 0.14}$	**72.88**$_{\pm 0.28}$	**66.18**$_{\pm 0.60}$	**86.25**$_{\pm 0.11}$	**64.69**$_{\pm 0.30}$	**83.88**$_{\pm 0.30}$	**85.54**$_{\pm 0.43}$	**80.16**$_{\pm 0.28}$	**77.73**$_{\pm 0.22}$	**81.55**$_{\pm 0.24}$	88.82$_{\pm 0.10}$	**83.81**$_{\pm 0.13}$	75.54$_{\pm 0.09}$	**78.57**$_{\pm 0.31}$	**75.92**$_{\pm 0.44}$	**78.16**

Table 2. Comparisons with baselines on CIFAR-100-C at severity level 5 regarding accuracy (%). The **bold** value signifies the top-performing result and the underline is the second-best.

	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg.
no_adapt	10.06	11.67	6.67	31.87	17.52	37.58	36.61	39.01	27.27	29.71	60.98	16.87	41.92	19.26	41.79	28.59
BN_adapt	32.18$_{\pm 0.01}$	32.98$_{\pm 0.01}$	28.95$_{\pm 0.05}$	59.25$_{\pm 0.08}$	35.05$_{\pm 0.05}$	56.45$_{\pm 0.08}$	59.45$_{\pm 0.01}$	46.86$_{\pm 0.05}$	46.12$_{\pm 0.07}$	49.09$_{\pm 0.09}$	62.41$_{\pm 0.00}$	58.32$_{\pm 0.03}$	45.86$_{\pm 0.04}$	47.84$_{\pm 0.01}$	40.73$_{\pm 0.07}$	46.77
Tent	33.39$_{\pm 0.03}$	34.17$_{\pm 0.11}$	29.80$_{\pm 0.06}$	59.98$_{\pm 0.09}$	36.04$_{\pm 0.05}$	57.06$_{\pm 0.13}$	60.05$_{\pm 0.03}$	47.54$_{\pm 0.02}$	46.73$_{\pm 0.04}$	50.04$_{\pm 0.19}$	62.93$_{\pm 0.00}$	58.61$_{\pm 0.02}$	46.48$_{\pm 0.02}$	49.09$_{\pm 0.03}$	41.81$_{\pm 0.11}$	47.58
EATA	34.61$_{\pm 0.13}$	35.72$_{\pm 0.24}$	30.82$_{\pm 0.04}$	60.09$_{\pm 0.04}$	36.31$_{\pm 0.03}$	57.00$_{\pm 0.03}$	60.10$_{\pm 0.03}$	47.84$_{\pm 0.05}$	47.06$_{\pm 0.05}$	50.84$_{\pm 0.20}$	62.90$_{\pm 0.01}$	58.55$_{\pm 0.05}$	46.65$_{\pm 0.05}$	50.19$_{\pm 0.25}$	42.75$_{\pm 0.14}$	48.10
SAR	33.64$_{\pm 0.06}$	34.60$_{\pm 0.11}$	29.77$_{\pm 0.21}$	59.82$_{\pm 0.06}$	36.36$_{\pm 0.03}$	56.96$_{\pm 0.08}$	60.00$_{\pm 0.04}$	47.43$_{\pm 0.01}$	46.68$_{\pm 0.03}$	49.99$_{\pm 0.14}$	62.77$_{\pm 0.03}$	58.52$_{\pm 0.03}$	46.57$_{\pm 0.01}$	49.14$_{\pm 0.00}$	41.85$_{\pm 0.10}$	47.61
NOTE	23.44$_{\pm 0.05}$	25.80$_{\pm 0.03}$	22.00$_{\pm 0.08}$	46.70$_{\pm 0.06}$	29.17$_{\pm 0.06}$	49.87$_{\pm 0.07}$	51.82$_{\pm 0.02}$	48.54$_{\pm 0.02}$	44.09$_{\pm 0.24}$	44.78$_{\pm 0.11}$	**66.13**$_{\pm 0.02}$	35.03$_{\pm 0.08}$	46.81$_{\pm 0.03}$	34.70$_{\pm 0.07}$	44.46$_{\pm 0.05}$	40.89
MEMO	18.60$_{\pm 0.00}$	21.30$_{\pm 0.00}$	16.47$_{\pm 0.01}$	44.62$_{\pm 0.00}$	26.13$_{\pm 0.01}$	46.67$_{\pm 0.00}$	50.27$_{\pm 0.01}$	45.80$_{\pm 0.01}$	40.45$_{\pm 0.00}$	40.09$_{\pm 0.01}$	64.70$_{\pm 0.00}$	27.42$_{\pm 0.01}$	46.04$_{\pm 0.00}$	32.24$_{\pm 0.01}$	44.59$_{\pm 0.01}$	37.69
EATA-C	32.48$_{\pm 0.17}$	33.26$_{\pm 0.16}$	29.13$_{\pm 0.35}$	59.55$_{\pm 0.36}$	35.29$_{\pm 0.32}$	56.64$_{\pm 0.36}$	59.72$_{\pm 0.13}$	47.10$_{\pm 0.20}$	46.38$_{\pm 0.30}$	49.31$_{\pm 0.47}$	62.70$_{\pm 0.14}$	58.52$_{\pm 0.16}$	46.15$_{\pm 0.18}$	48.29$_{\pm 0.11}$	41.13$_{\pm 0.50}$	47.04
SAR^2	34.04$_{\pm 0.28}$	34.84$_{\pm 0.43}$	30.30$_{\pm 0.24}$	60.00$_{\pm 0.21}$	36.47$_{\pm 0.25}$	57.05$_{\pm 0.31}$	60.05$_{\pm 0.27}$	47.58$_{\pm 0.18}$	46.80$_{\pm 0.11}$	50.23$_{\pm 0.46}$	62.86$_{\pm 0.13}$	58.54$_{\pm 0.20}$	46.62$_{\pm 0.14}$	49.35$_{\pm 0.06}$	42.15$_{\pm 0.45}$	47.79
AdaDEM	33.05$_{\pm 0.20}$	33.74$_{\pm 0.29}$	29.60$_{\pm 0.27}$	59.71$_{\pm 0.35}$	35.75$_{\pm 0.32}$	56.89$_{\pm 0.49}$	59.91$_{\pm 0.13}$	47.24$_{\pm 0.23}$	46.57$_{\pm 0.34}$	49.73$_{\pm 0.56}$	62.72$_{\pm 0.23}$	58.53$_{\pm 0.20}$	46.30$_{\pm 0.10}$	48.71$_{\pm 0.21}$	41.48$_{\pm 0.40}$	47.33
COME	32.57$_{\pm 0.22}$	33.35$_{\pm 0.16}$	29.28$_{\pm 0.27}$	59.64$_{\pm 0.40}$	35.35$_{\pm 0.25}$	56.77$_{\pm 0.37}$	59.72$_{\pm 0.08}$	47.23$_{\pm 0.25}$	46.43$_{\pm 0.35}$	49.40$_{\pm 0.49}$	62.65$_{\pm 0.19}$	58.60$_{\pm 0.21}$	46.02$_{\pm 0.13}$	48.37$_{\pm 0.15}$	41.17$_{\pm 0.40}$	47.10
DeYO	36.92$_{\pm 0.07}$	38.44$_{\pm 0.01}$	32.46$_{\pm 0.03}$	60.81$_{\pm 0.03}$	37.89$_{\pm 0.00}$	57.75$_{\pm 0.03}$	60.68$_{\pm 0.01}$	48.65$_{\pm 0.06}$	47.88$_{\pm 0.01}$	51.82$_{\pm 0.08}$	63.29$_{\pm 0.04}$	**59.19**$_{\pm 0.00}$	47.61$_{\pm 0.04}$	51.73$_{\pm 0.00}$	44.32$_{\pm 0.19}$	49.30
NCTTA (ours)	**38.22**$_{\pm 0.46}$	**40.05**$_{\pm 0.43}$	**33.35**$_{\pm 0.40}$	**61.20**$_{\pm 0.13}$	**38.43**$_{\pm 0.28}$	**58.17**$_{\pm 0.43}$	**61.09**$_{\pm 0.12}$	**49.52**$_{\pm 0.29}$	**48.69**$_{\pm 0.29}$	**52.79**$_{\pm 0.30}$	63.76$_{\pm 0.27}$	59.17$_{\pm 0.14}$	**47.85**$_{\pm 0.05}$	**52.94**$_{\pm 0.14}$	**45.46**$_{\pm 0.65}$	**50.05**

A.3. Proof of NC3+

We extend the NC framework by introducing Sample-wise Alignment Collapse (NC3+), which asserts that, in the TPT, each feature embedding $\mathbf{h}_i$ and its corresponding classifier weight w_{y_i} become asymptotically co-linear. Empirical evidence (Section 3.2) shows that d_{ij} satisfies $d_{iy_i} \longrightarrow 0$ and d_{ij} $(j \neq y_i)$ remains essentially constant. We now present a concise derivation, preserving the original gradient argument.

Cross-Entropy Setup. For a given sample (x_i, y_i), the per-sample cross-entropy loss is defined as:

$$\mathcal{L}_{\mathrm{CE}}(x_i, y_i) = -\log p_{iy_i}, \tag{9}$$

where $p_{ij} = \frac{\exp(\mathbf{z}_{ij})}{\sum_{k=1}^{K} \exp(\mathbf{z}_{ik})}$ and $\mathbf{z}_{ij} = w_j \mathbf{h}_i$.

Logit Gradients. By the chain rule,

$$\frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial \mathbf{z}_{ij}} = \frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial p_{ij}} \frac{\partial p_{ij}}{\partial \mathbf{z}_{ij}} = \begin{cases} p_{ij} - 1, & j = y_i, \\ p_{ij}, & j \neq y_i. \end{cases} \tag{10}$$

During the TPT, $p_{iy_i} < 1$, so

$$\frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial \mathbf{z}_{iy_i}} = p_{iy_i} - 1 < 0, \quad \frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial \mathbf{z}_{ij}} = p_{ij} > 0 \ (j \neq y_i). \tag{11}$$

Gradient descent thus increases target logit $\mathbf{z}_{iy_i}$ and suppresses $\mathbf{z}_{ij}$ for $j \neq y_i$.

Angle and Distance Reduction. We begin by expressing the logit as

$$\mathbf{z}_{ij} = \|\mathbf{h}_i\|_2 \, \|w_j\|_2 \, \cos \alpha_{ij}, \tag{12}$$

where

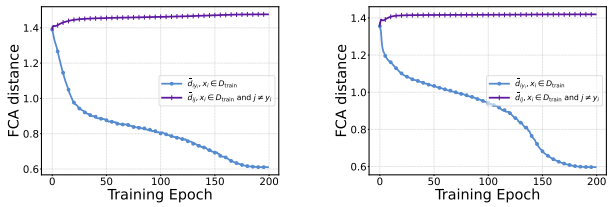
$$\cos \alpha_{ij} = \frac{\langle \mathbf{h}_i, w_j \rangle}{\|\mathbf{h}_i\|_2 \, \|w_j\|_2}$$

denotes the cosine of the angle between $\mathbf{h}_i$ and w_j.

Our empirical results (Figure 2) show that the growth of the ground-truth logit $\mathbf{z}_{iy_i}$ during training is predominantly attributed to increased alignment, i.e., an increase in $\cos \alpha_{iy_i}$. For example, when training ResNet-50 on CIFAR-10, $\mathbf{z}_{iy_i}$ increases by a factor of $37.3\times$, which is mainly driven by a $32.3\times$ increase in alignment $\cos \alpha_{iy_i}$, while the norm product $\|\mathbf{h}_i\|_2 \|w_{y_i}\|_2$ grows by only $1.2\times$.

Table 3. Comparisons with baselines on ImageNet-C at severity level 5 regarding accuracy (%). The **bold** value signifies the top-performing result and the underline is the second-best.

	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg.
no_adapt	34.93	33.67	36.41	32.38	22.22	36.61	31.10	22.15	26.73	53.16	61.17	50.27	31.96	54.52	55.87	38.88
Tent	51.93$_{\pm 0.01}$	52.77$_{\pm 0.00}$	53.05$_{\pm 0.01}$	52.59$_{\pm 0.03}$	45.53$_{\pm 0.97}$	54.89$_{\pm 0.01}$	48.07$_{\pm 0.07}$	17.33$_{\pm 0.87}$	19.79$_{\pm 1.10}$	66.34$_{\pm 0.09}$	73.19$_{\pm 0.16}$	65.06$_{\pm 0.01}$	46.44$_{\pm 12.61}$	66.57$_{\pm 0.03}$	64.53$_{\pm 0.01}$	51.87
EATA	55.72$_{\pm 0.01}$	56.94$_{\pm 0.00}$	56.76$_{\pm 0.03}$	57.71$_{\pm 0.02}$	55.81$_{\pm 0.05}$	62.07$_{\pm 0.00}$	60.45$_{\pm 0.01}$	64.98$_{\pm 0.48}$	63.74$_{\pm 0.04}$	71.90$_{\pm 0.01}$	76.71$_{\pm 0.05}$	67.81$_{\pm 0.24}$	66.29$_{\pm 0.00}$	72.48$_{\pm 0.01}$	69.25$_{\pm 0.00}$	63.91
SAR	51.93$_{\pm 0.01}$	52.90$_{\pm 0.01}$	53.06$_{\pm 0.00}$	52.66$_{\pm 0.00}$	47.08$_{\pm 0.09}$	54.88$_{\pm 0.05}$	48.42$_{\pm 0.01}$	26.60$_{\pm 10.31}$	34.87$_{\pm 4.86}$	66.28$_{\pm 0.13}$	72.79$_{\pm 0.01}$	64.67$_{\pm 0.00}$	50.80$_{\pm 0.58}$	66.40$_{\pm 0.00}$	64.36$_{\pm 0.00}$	53.97
NOTE	35.11$_{\pm 0.00}$	33.86$_{\pm 0.00}$	36.58$_{\pm 0.00}$	32.71$_{\pm 0.00}$	22.39$_{\pm 0.00}$	36.92$_{\pm 0.00}$	31.30$_{\pm 0.00}$	22.33$_{\pm 0.00}$	26.90$_{\pm 0.00}$	53.14$_{\pm 0.00}$	62.25$_{\pm 2.02}$	51.18$_{\pm 0.00}$	32.04$_{\pm 0.00}$	54.64$_{\pm 0.00}$	55.95$_{\pm 0.00}$	39.15
MEMO	43.78$_{\pm 0.00}$	43.25$_{\pm 0.01}$	45.83$_{\pm 0.00}$	32.90$_{\pm 0.03}$	28.46$_{\pm 0.03}$	45.67$_{\pm 0.07}$	38.25$_{\pm 0.12}$	30.08$_{\pm 0.06}$	34.63$_{\pm 0.15}$	54.80$_{\pm 0.20}$	69.99$_{\pm 0.37}$	56.43$_{\pm 0.16}$	34.47$_{\pm 0.01}$	62.56$_{\pm 0.03}$	59.61$_{\pm 0.04}$	45.38
SAR^2	52.15$_{\pm 0.19}$	53.15$_{\pm 0.07}$	53.28$_{\pm 0.05}$	53.08$_{\pm 0.08}$	47.35$_{\pm 0.38}$	55.19$_{\pm 0.18}$	48.84$_{\pm 0.22}$	26.93$_{\pm 4.49}$	33.61$_{\pm 4.85}$	66.53$_{\pm 0.32}$	72.97$_{\pm 0.09}$	65.07$_{\pm 0.04}$	50.55$_{\pm 1.34}$	66.53$_{\pm 0.06}$	64.60$_{\pm 0.03}$	53.99
AdaDEM	50.29$_{\pm 0.07}$	50.96$_{\pm 0.04}$	51.54$_{\pm 0.05}$	50.20$_{\pm 0.05}$	43.80$_{\pm 0.10}$	52.71$_{\pm 0.05}$	47.22$_{\pm 0.12}$	43.10$_{\pm 0.73}$	40.74$_{\pm 0.24}$	64.47$_{\pm 0.19}$	71.55$_{\pm 0.10}$	63.31$_{\pm 0.09}$	48.91$_{\pm 0.81}$	62.66$_{\pm 0.07}$	61.41$_{\pm 0.28}$	53.53
COME	52.65$_{\pm 0.08}$	53.85$_{\pm 0.07}$	53.93$_{\pm 0.11}$	54.90$_{\pm 0.03}$	48.45$_{\pm 0.21}$	57.81$_{\pm 0.03}$	51.82$_{\pm 0.07}$	60.01$_{\pm 0.16}$	58.84$_{\pm 0.15}$	70.06$_{\pm 0.19}$	77.32$_{\pm 0.06}$	66.74$_{\pm 0.06}$	54.97$_{\pm 0.53}$	70.30$_{\pm 0.06}$	67.19$_{\pm 0.05}$	59.92
DeYO	55.68$_{\pm 0.01}$	56.92$_{\pm 0.02}$	56.42$_{\pm 0.03}$	57.85$_{\pm 0.01}$	55.71$_{\pm 0.33}$	62.68$_{\pm 0.10}$	46.53$_{\pm 0.38}$	66.42$_{\pm 0.01}$	65.46$_{\pm 0.01}$	72.45$_{\pm 0.03}$	78.38$_{\pm 0.01}$	66.52$_{\pm 0.05}$	67.48$_{\pm 0.04}$	73.63$_{\pm 0.00}$	70.18$_{\pm 0.01}$	63.49
NCTTA (ours)	**57.58**$_{\pm 0.05}$	**59.03**$_{\pm 0.23}$	**58.98**$_{\pm 0.16}$	**60.14**$_{\pm 0.04}$	**58.90**$_{\pm 0.35}$	**65.14**$_{\pm 0.04}$	**63.21**$_{\pm 0.19}$	**68.73**$_{\pm 0.29}$	**67.80**$_{\pm 0.02}$	**74.24**$_{\pm 0.16}$	**78.75**$_{\pm 0.15}$	**69.19**$_{\pm 0.15}$	**69.25**$_{\pm 0.16}$	**74.57**$_{\pm 0.06}$	**71.39**$_{\pm 0.01}$	**66.46**



(a) CIFAR-10 (ResNet50) (b) CIFAR-100 (ResNet50)

(c) CIFAR-10 (ViT-S/16) (d) CIFAR-100 (ViT-S/16)

Figure 1. **Supplementary Validation of Sample-wise Alignment Collapse (NC3+).** This figure provides additional evidence supporting the NC3+ property by showing the behavior of FCA distances d_{iy_i} and $d_{ij}, j \neq y_i$ during training. Results are evaluated on CIFAR-10, CIFAR-100, and ImageNet-100 using different backbones (ResNet50 and ViT). These supplementary results further confirm the convergence of d_{iy_i} to zero during the TPT.

Therefore, the increase of $\mathbf{z}_{iy_i}$ is primarily associated with an increase in $\cos\alpha_{iy_i}$, implying a decrease in the angle α_{iy_i}. Consequently, the normalized feature–classifier distance satisfies

$$\begin{aligned} d_{iy_i} &= \left\| \frac{\mathbf{h}_i}{\|\mathbf{h}_i\|_2} - \frac{w_{y_i}}{\|w_{y_i}\|_2} \right\|_2 \\ &= \sqrt{2 - 2\,\frac{\langle \mathbf{h}_i, w_{y_i} \rangle}{\|\mathbf{h}_i\|_2\,\|w_{y_i}\|_2}} \\ &= \sqrt{2 - 2\cos\alpha_{iy_i}} \;\longrightarrow\; 0. \end{aligned} \tag{13}$$

In contrast, d_{ij} for $j \neq y_i$ does not exhibit the same systematic decrease. Over repeated optimization steps, this

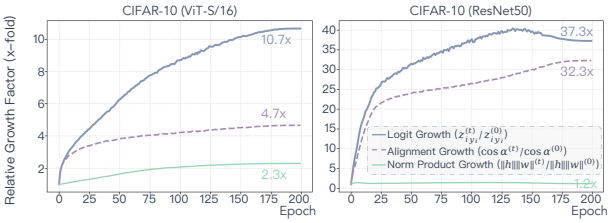


Figure 2. **Growth dynamics of the ground-truth logit during training.** We decompose the growth of the ground-truth logit $\mathbf{z}_{iy_i}$ into the alignment term $\cos\alpha_{iy_i}$ and the norm product $\|\mathbf{h}_i\|_2\|w_{y_i}\|_2$. Empirically, the increase of $\mathbf{z}_{iy_i}$ is predominantly driven by improved alignment rather than norm growth.

mechanism drives a sample-wise collapse of $\mathbf{h}_i$ toward its corresponding classifier weight w_{y_i}, thereby establishing NC3+.

B. Details of Datasets, Baselines and Hyperparameters

B.1. Datasets and Licenses

In this paper, we primarily evaluate the performance of all methods on OOD data using widely recognized and large-scale benchmarks, including CIFAR-10-C, CIFAR-100-C, ImageNet-C, Waterbirds, and PACS. These datasets are well-established for assessing model robustness under domain shifts and cover diverse types of domain corruptions, spurious correlations, and domain generalization tasks. Below, we provide an overview of each dataset.

CIFAR-10-C, CIFAR-100-C, and ImageNet-C. CIFAR-10-C, CIFAR-100-C, and ImageNet-C are OOD evaluation datasets derived from their respective base datasets—CIFAR-10, CIFAR-100, and ImageNet—by applying systematic corruptions. These datasets include 15 distinct corruption types, such as Gaussian noise, shot noise, impulse noise, defocus blur, glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic

(a) Gaussian Noise (b) Shot Noise (c) Impulse Noise

(d) Defocus Blur (e) Glass Blur (f) Motion Blur

(g) Zoom Blur (h) Snow (i) Frost

(j) Fog (k) Brightness (l) Contrast

(m) Elastic (n) Pixelate (o) JPEG

Figure 3. **Illustration of ImageNet-C under five levels of severity.** The dataset showcases 15 types of algorithmically generated corruptions across four categories: noise, blur, weather, and digital. Each corruption type is illustrated at five increasing levels of severity, demonstrating the progressive impact of these corruptions.

transformation, pixelation, and JPEG compression. Each corruption type is further categorized into five severity levels, with higher levels representing more extreme domain shifts. An example of these corruptions is illustrated in Figure 3. These datasets serve as a standard benchmark for evaluating model robustness to common corruptions and noise.

Waterbirds. The Waterbirds dataset is designed to assess robustness to spurious correlations. It is constructed by combining images from the CUB dataset (a fine-grained bird classification dataset) and background scenes from the Places dataset. Specifically, bird species are overlaid onto either water or land backgrounds, creating spurious correlations between the bird class and the background type. For example, most "waterbird" images are paired with water backgrounds, while most "landbird" images are paired with land backgrounds. The test set, however, includes examples where these spurious correlations do not hold, making it a challenging benchmark for evaluating a model's ability to generalize beyond such biases.

PACS. PACS is a domain generalization benchmark consisting of images from four distinct domains: Paintings, Artistic images, Cartoons, and Sketches. It contains seven object categories common across all domains, such as "dog," "guitar," and "person." Each domain exhibits unique visual characteristics, with substantial domain shifts between them. The primary task involves training on three domains and testing on the held-out domain, assessing a model's ability to generalize to unseen domain shifts. PACS is widely used to evaluate domain generalization methods due to its diverse and realistic domain variations.

Source code. We use the implementation of existing baseline methods for reporting their results in this paper, conducted on the TTAB benchmark. Source code used in this paper is under the MIT License.

B.2. Baselines

In this section, we provide an overview of the baseline methods evaluated in our study. All experiments are conducted on the TTAB benchmark, with algorithmic parameters set to their default values unless otherwise specified.

no_adapt refers to directly using a pre-trained model for inference on the test data without any adaptation. The accuracy is calculated based on this straightforward approach.

BN_adapt proposes replacing the Batch Normalization (BN) layer statistics, such as mean and variance, with the statistics computed from the current test batch, which is referred to as Target Batch Normalization (TBN).

Tent introduces entropy minimization as a self-supervised loss for TTA. By minimizing the entropy of the predictions, the model is encouraged to adapt to the target domain effectively.

EATA addresses the issue of noisy gradients caused by high-entropy samples. To mitigate this, it introduces an entropy-based filtering mechanism and a weighting strategy. In our experiments, the entropy filtering threshold E_0 is set to $\texttt{math.log}(K) \times 0.4$, where K is the number of

Table 4. Ablation study of proposed components on CIFAR-10-C. Average classification accuracy (%) under corruption severity level 5. The **bold** entries indicate the highest performance.

	$\mathcal{L}_{ENT}$	$\mathcal{L}_{NC}$	S_{ENT}	λ	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg.
(1) Tent	✓				64.88	66.91	59.60	84.89	61.19	82.42	83.77	77.67	75.43	80.27	87.67	83.63	72.72	74.62	72.18	75.19
(2) SAR	✓		✓		64.32	66.29	58.93	84.71	60.68	82.28	83.54	77.21	74.77	79.93	87.34	83.51	72.39	74.24	71.95	74.81
(3)		✓			65.07	67.18	59.88	84.98	61.19	82.55	83.71	77.86	75.60	80.29	87.64	83.93	72.78	74.74	72.19	75.31
(4)		✓	✓		64.16	66.40	59.31	84.89	60.87	82.41	83.61	77.46	75.08	80.08	87.51	83.63	72.49	74.36	72.00	74.95
(5)	✓	✓			67.05	69.23	62.09	85.53	62.39	82.93	84.33	78.54	76.19	80.68	88.14	84.00	73.75	75.60	73.29	76.25
(6)	✓	✓	✓		65.69	68.10	60.68	85.13	61.93	82.70	84.16	78.06	75.54	80.36	87.76	83.54	73.00	75.07	72.86	75.64
(7)		✓	✓	✓	67.97	70.57	63.44	85.70	62.87	83.13	84.85	79.55	76.81	81.01	88.35	83.87	74.40	76.45	74.12	76.87
(8)	✓		✓	✓	68.00	70.10	62.73	85.57	63.07	83.16	84.70	78.85	75.81	80.70	87.95	83.41	74.12	76.24	73.85	76.55
(9) NCTTA(ours)	✓	✓	✓	✓	**71.08**	**72.53**	**65.97**	**86.20**	**64.79**	**83.86**	**85.84**	**80.42**	**77.55**	**81.59**	**88.69**	**83.79**	**75.46**	**78.21**	**75.55**	**78.10**

Table 5. Ablation study of proposed components on ImageNet-C. Average classification accuracy (%) under corruption severity level 5. The **bold** entries indicate the highest performance.

	$\mathcal{L}_{ENT}$	$\mathcal{L}_{NC}$	S_{ENT}	λ	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg.
(1) Tent	✓				51.80	52.74	52.94	52.31	45.14	54.86	47.52	13.68	18.86	66.22	73.20	64.99	41.40	66.72	64.60	51.13
(2) SAR	✓		✓		51.90	52.86	52.93	52.92	47.06	54.82	48.46	18.32	19.88	66.60	72.70	64.95	49.75	66.47	64.41	52.27
(3)		✓			47.44	47.71	48.87	48.57	38.48	50.97	45.22	48.83	48.55	64.47	73.27	63.40	45.25	64.02	62.44	53.17
(4)		✓	✓		42.54	43.13	45.46	46.59	38.50	49.93	44.76	48.01	46.09	63.49	72.15	60.66	44.88	63.16	61.55	51.39
(5)	✓	✓			53.51	54.63	54.67	54.82	48.74	57.72	51.10	24.18	22.46	69.12	76.21	67.06	53.20	68.94	66.26	54.84
(6)	✓	✓	✓		53.77	54.94	54.88	55.49	50.44	58.07	52.66	37.51	39.08	69.39	76.03	67.10	55.63	68.85	66.26	57.34
(7)		✓	✓	✓	44.16	43.96	46.54	48.92	43.69	52.71	47.90	58.09	50.07	65.72	74.07	53.63	51.81	66.19	62.16	53.97
(8)	✓		✓	✓	51.35	32.02	45.18	57.72	56.32	62.52	37.20	4.01	9.96	72.60	77.73	67.59	66.01	72.91	69.85	52.20
(9) NCTTA(ours)	✓	✓	✓	✓	**57.68**	**59.07**	**59.11**	**60.07**	**59.27**	**65.17**	**63.01**	**68.61**	**67.65**	**74.46**	**78.79**	**69.50**	**69.36**	**74.52**	**71.30**	**66.50**

classes in the dataset. The learning rate ϵ is set to 0.05, and the trade-off parameter β is set to $1/2000$.

SAR investigates the impact of noisy gradients in more challenging scenarios, such as small batch sizes and mixed data distributions. Similar to EATA, the entropy threshold E_0 is set to $\texttt{math.log}(K) \times 0.4$. Additionally, the threshold e_m for the model recovery scheme is set to 0.2.

NOTE introduces Instance-Aware Batch Normalization (IABN) to correct normalization for OOD samples. The memory size is set to 64, with the following parameter configurations: `bn_momentum` = 0.01, `temperature` = 1.0, `iabn_k` = 4, and `threshold_note` = 1 (used to discard adjustments based on the skip threshold).

MEMO improves model robustness by averaging probabilities across multiple augmented views of the same input. After adapting to each batch, the model is reset, making the adaptation process episodic.

EATA-C extends EATA by replacing the naive entropy minimization objective with a consistency-based uncertainty reduction loss and a conditional min-max entropy regularizer for confidence calibration, while retaining the Fisher regularization to mitigate catastrophic forgetting. The entropy threshold E_0 is set to $0.4 \log K$, and the cosine similarity threshold d_0 is set to 0.05 for redundant sample filtering. The stochastic depth ratio is set to 0.2. The smoothing parameter p in the consistency loss is 0.5, and the balancing coefficient α for the entropy regularizer is 0.1.

SAR^2 extends SAR by incorporating centroid-based feature regularization with a momentum-updated feature bank. The entropy filtering threshold E_0 is set to $0.4 \log K$, and the model recovery threshold e_m is set to 0.2. A feature redundancy loss weighted by α (set to 10^3) and a feature inequity loss weighted by β (set to 50).

AdaDEM builds upon Tent by replacing the standard entropy minimization loss with the proposed Decoupled Entropy Minimization objective. The marginal update coefficient is set to $\pi = 0.1$.

COME replaces the conventional entropy loss with Dirichlet entropy. In our implementation, we substitute Tent's $\mathcal{L}_{\text{Ent}}$ with the COME loss, while keeping all other components identical to Tent.

DeYO proposes a novel sample selection strategy called Pseudo-Label Probability Difference (PLPD), which is combined with entropy-based filtering to ensure higher-quality sample selection. In our experiments, the entropy threshold τ_{Ent}, the PLPD threshold τ_{PLPD}, and Ent_0 is set to $\texttt{math.log}(K) \times 0.4$, $\texttt{math.log}(K) \times 0.5$, and 0.3, respectively.

NCTTA (Ours) incorporates parameters α, k, γ_{Ent}, τ_{Ent}, η and ν. Specifically, η is set to 5.0, ν to 1, and γ_{Ent} and τ_{Ent} are consistent with prior methods, both set to $\texttt{math.log}(K) \times 0.4$. To align the gradient scales of $\mathcal{L}_{\text{ENT}}$ and $\mathcal{L}_{\text{NC}}$, the latter is multiplied by a factor of 5. α and k are set to values that work well across all situations. Specifically, α is fixed at 0.3, and k is set to 3 for CIFAR-10-C, 5 for CIFAR-100-C, 10 for ImageNet-C, 2 for PACS, and 1

for Waterbirds.

B.3. Hyperparameters

To ensure a fair comparison across all methods, we adopt consistent experimental settings. Below, we detail the hyperparameters used for each dataset and scenario.

For CIFAR-10-C and CIFAR-100-C, we employ ResNet50 with Batch Normalization as the backbone. The learning rate is set to 1×10^{-4}, and the batch size is 64.

For ImageNet-C, we utilize the ViT-B/16 architecture as the backbone. The learning rate is set to 1×10^{-3}, with a batch size of 64. In scenarios with a batch size of 1, the learning rate is scaled down by factors of 16 and 32 for CIFAR-10-C/100-C and ImageNet-C, respectively, to ensure stable optimization.

For Waterbirds and PACS, ResNet50 is also used as the backbone. The learning rate is configured as 2.5×10^{-4}, and the batch size is 32.

For experiments on ImageNet-C with batch size = 1 and for CTTA experiments, the domain sampling ratio is set to 0.1; for all other experiments, the domain sampling ratio is set to 1. The random seeds are set to 2022, 2023, and 2024. The reported results are the mean and variance over three runs. For ablation studies, the random seed is set to 2022.

All experiments were conducted on a single NVIDIA Tesla V100 GPU, using SGD as the optimizer. These parameter configurations are carefully selected to ensure consistency and reliability across all experiments, facilitating a robust comparison of the proposed and baseline methods.

C. Additional Experiments and Analysis

C.1. Comparing Experiments on CIFAR-100-C

Table 2 presents the classification accuracy of TTA methods on CIFAR-100-C under corruption severity level 5. As CIFAR-100-C contains 100 classes, it provides a more challenging benchmark compared to CIFAR-10-C. Among all methods, NCTTA achieves the highest average accuracy of 50.15%, outperforming the second-best method, DeYO (49.30%), and Tent (47.58%). These results demonstrate the capability of NCTTA in maintaining robust performance under severe domain shifts. Additionally, NCTTA consistently outperforms other baselines across a variety of corruption types, such as Gaussian Noise (37.98%), Glass Blur (38.50%), and Zoom Blur (61.16%), showcasing its superior adaptability.

Compared to DeYO, NCTTA achieves higher accuracy on several challenging corruption types, including Frost (48.82% vs. 47.88%) and Glass Blur (38.53% vs. 37.89%). While DeYO performs competitively on some corruptions, such as Defocus Blur (60.81%), NCTTA demonstrates a stronger ability to generalize across all corruptions, delivering more consistent performance. Tent, a popular entropy-minimization method, falls behind NCTTA in most scenarios, particularly under extreme noise conditions like Gaussian Noise (33.39%) and Impulse Noise (29.80%), where NCTTA provides significant improvements. Similarly, methods like EATA and SAR, while incorporating stability mechanisms, are unable to match the performance of NCTTA, especially in scenarios requiring better feature-classifier alignment.

Overall, the results highlight the robustness and generalization capabilities of NCTTA on CIFAR-100-C. The method's ability to maintain high accuracy across diverse corruptions, combined with its feature-classifier alignment strategy, ensures strong performance even in highly challenging scenarios. These findings solidify NCTTA as the state-of-the-art approach for TTA on CIFAR-100-C, outperforming existing baselines by a clear margin.

C.2. Ablation Experiments

We conduct an ablation study to evaluate the contribution of each component in NCTTA across three datasets: CIFAR-10-C, CIFAR-100-C and ImageNet-C (Table 5). The components analyzed include the entropy minimization loss (L_{ENT}), the alignment loss (L_{NC}), sample filtering (S_{ENT}), and weighting (λ). The full NCTTA model, which incorporates all these components, consistently achieves the highest accuracy across all datasets, demonstrating the importance of synergizing these mechanisms to enhance robustness under domain shifts.

On ImageNet-C (Table 5), ablation experiments demonstrate the effectiveness of each component in NCTTA. The full model achieves an average accuracy of 64.76%, highlighting the synergy between its design elements. The analysis also reveals that entropy loss, while beneficial in some cases, can harm performance under certain corruptions. For instance, as shown in Table 5, applying entropy loss alone under the Snow and Frost corruption results in degraded accuracy. In contrast, the alignment loss consistently enhances performance, including under Snow and Frost, by preserving feature consistency and guiding the model toward stable adaptation. This demonstrates that alignment loss is more robust in scenarios where entropy minimization may inadvertently weaken the model's predictions.

The results confirm that the full NCTTA model, which integrates all components, achieves the best performance across all corruption types. This validates the importance of alignment-based weighting and careful component integration in achieving state-of-the-art results for TTA.

In summary, the ablation experiments demonstrate that all components of NCTTA—entropy minimization, alignment loss, sample filtering, and adaptive weighting—are essential for achieving state-of-the-art performance. The synergy between these mechanisms ensures robust adaptation under severe distributional shifts across datasets of varying

complexity.

C.3. Cost Analysis

Consistent with TTA protocols, our approach relies on backpropagation for online optimization. As shown in Table 6, the computational overhead of the extra $\mathcal{L}_{\text{NC}}$ term is negligible. NCTTA preserves Tent's memory footprint while introducing negligible latency overhead ($1.05\times$ slower).

Table 6. Efficiency Statistics on ImageNet-C (ViT-B/16) under Brightness corruption severity level 5.

Metric	Tent	DeYO	NCTTA (ours)
Avg. Backward Time (s) ↓	0.242	0.238	0.242
Avg. Forward Time (s) ↓	0.190	0.384	0.192
GPU Memory (GiB) ↓	4.67	4.85	4.67