Your Dissimilarities Define You: Complementary Learning Exploiting Class Diversities

Supplementary Material

A. Appendix

A1. Gradient Analysis

Let us consider the general cross entropy loss, given by

$$L = -\sum_{i=1}^{N} p_i \log(\hat{p}_i), \tag{1}$$

where $\mathbf{p}$ is an arbitrary target distribution and $\hat{p}_i = \mathrm{softmax}(z)_i$ is the predicted probability for class i. The derivative of the loss with respect to $\hat{p}_i$ is

$$\frac{\partial L}{\partial \hat{p}_i} = -\frac{p_i}{\hat{p}_i}, \tag{2}$$

while the softmax Jacobian has two cases:

$$\frac{\partial \hat{p}_i}{\partial z_j} = \begin{cases} \hat{p}_i(1-\hat{p}_i) & \text{if } i = j, \\ -\hat{p}_i \hat{p}_j & \text{if } i \neq j. \end{cases} \tag{3}$$

Applying the chain rule gives

$$\frac{\partial L}{\partial z_j} = \sum_{i=1}^{N} \frac{\partial L}{\partial \hat{p}_i} \frac{\partial \hat{p}_i}{\partial z_j}, \tag{4}$$

which expands to

$$= \underbrace{\sum_{i \neq j} -\frac{p_i}{\hat{p}_i} \cdot (-\hat{p}_i \hat{p}_j)}_{i \neq j} + \underbrace{\left(-\frac{p_j}{\hat{p}_j} \cdot \hat{p}_j (1-\hat{p}_j) \right)}_{i=j}, \tag{5}$$

leading to the simplified form

$$\frac{\partial L}{\partial z_j} = \sum_{i \neq j} p_i \hat{p}_j \; + \; p_j(\hat{p}_j - 1). \tag{6}$$

We now substitute the two target distributions: the standard one-hot vector $\mathbf{p}$ and the proposed one-cold uniform target vector $\bar{\mathbf{p}}$ used in the definition of the CDL$_{uni}$ loss. For a target class c:

$$p_i = \begin{cases} 1 & i = c, \\ 0 & i \neq c, \end{cases} \qquad \bar{p}_i = \begin{cases} 0 & i = c, \\ \frac{1}{N-1} & i \neq c. \end{cases} \tag{7}$$

For standard cross entropy L_{CE}, substituting $\mathbf{p}$ gives

$$\frac{\partial L_{\mathrm{CE}}}{\partial z_i} = \begin{cases} \hat{p}_i - 1 & i = c, \\ \hat{p}_i & i \neq c. \end{cases} \tag{8}$$

For the proposed uniform Complementary Dissimilarity Loss L_{CD}^{uni}, substituting $\bar{\mathbf{p}}$ results in

$$\frac{\partial L_{\mathrm{CD}}^{uni}}{\partial \bar{z}_i} = \begin{cases} \hat{\bar{p}}_i & i = c, \\ \hat{\bar{p}}_i - \frac{1}{N-1} & i \neq c. \end{cases} \tag{9}$$

We now examine what conditions force $\nabla_{\mathbf{z}} L = \mathbf{0}$ for each loss, assuming no other constraints on the logits.

Cross-entropy (CE) case. The stationary point equations require $\hat{p}_c \to 1$ and $\hat{p}_k \to 0$ $(k \neq c)$. Since the softmax satisfies $\hat{p}_i = \exp(z_i)/\sum_j \exp(z_j)$, we obtain

$$\hat{p}_c \to 1 \implies z_c \to +\infty, \tag{10}$$

$$\hat{p}_k \to 0 \implies z_k \to -\infty \qquad (k \neq c). \tag{11}$$

Thus CE drives the target logit to $+\infty$ and all non-target logits to $-\infty$, diverging in opposite directions.

Proposed (CDL$_{uni}$) case. For the proposed objective, the complementary predicted probabilities satisfy

$$\hat{\bar{p}}_i = \mathrm{softmax}(-z)_i.$$

The gradient conditions require $\nabla_{\bar{\mathbf{z}}} L_{\mathrm{CD}}^{uni} \to 0$, which implies

$$\hat{\bar{p}}_c = \frac{\exp(-z_c)}{\sum_j \exp(-z_j)} \to 0 \implies z_c \to +\infty, \tag{12}$$

$$\hat{\bar{p}}_k = \frac{\exp(-z_k)}{\sum_j \exp(-z_j)} \to \frac{1}{N-1} \implies$$

$$\exp(-z_k) \to \frac{1}{N-1} \sum_{j \neq c} \exp(-z_j). \tag{13}$$

Since the right-hand side is the average over the complementary logits, each non-target logit must converge to the same finite value:

$$z_k \to z_{\mathrm{const}} \qquad \text{for all } k \neq c. \tag{14}$$

Unlike standard CE, which pushes all non-target logits to $-\infty$, CDL$_{uni}$ forces the non-target logits to collapse to a common finite value, while still driving the correct logit z_c to $+\infty$. This symmetric convergence of the non-target logits enables the improved margin and spread behavior analyzed Section 3.3 of the main paper.

A2. Proof of Theorem 1

We fix $K \geq 2$ classes and consider a correctly classified sample with predicted logits $\mathbf{z}$. Let us define the mean of the non-target logits μ_{-c}, the centered deviation of each non-target logit δ_k and the margin between target and non-target mean m as:

$$\mu_{-c} = \frac{1}{K-1} \sum_{k \neq c} z_k, \tag{15}$$

$$\delta_k = z_k - \mu_{-c} \quad (k \neq c), \text{ and} \tag{16}$$

$$m = z_c - \mu_{-c}. \tag{17}$$

The centered deviations satisfy $\sum_{k \neq c} \delta_k = 0$. We collect all non–target centered deviations into the vector $\boldsymbol{\delta}_{-c} = (\delta_k)_{k \neq c}$. The non–target spread is then defined as:

$$\sigma = \sqrt{\frac{1}{K-1} \sum_{k \neq c} \delta_k^2} = \frac{\|\boldsymbol{\delta}_{-c}\|}{\sqrt{K-1}}. \tag{18}$$

We first isolate how a single gradient step affects the centered non-target deviations, which is the quantity controlling the spread σ. A gradient step on logits under SGD, where $\mathbf{z}^+$ denotes the updated logit values, is defined as:

$$\mathbf{z}^+ = \mathbf{z} - \eta \mathbf{g}, \qquad \mathbf{g} = \nabla_{\mathbf{z}} \mathcal{L}(\mathbf{z}; c). \tag{19}$$

Then the following lemma holds.

Lemma 1 (Centering identity). Let $\mathbf{g} = \nabla_{\mathbf{z}} \mathcal{L}(\mathbf{z}; c)$ denote the gradient of the loss with respect to the logits, so that g_k is the gradient entry corresponding to logit z_k. Define the mean gradient over the non-target classes as:

$$\bar{g} = \frac{1}{K-1} \sum_{j \neq c} g_j. \tag{20}$$

After one gradient step $z_k^+ = z_k - \eta g_k$, the updated centered deviation for each non-target logit is given by:

$$\delta_k^+ = z_k^+ - \mu_{-c}^+, \qquad (k \neq c), \tag{21}$$

where μ_{-c}^+ is the updated non-target mean. Then:

$$\delta_k^+ = \delta_k - \eta \, (g_k - \bar{g}), \quad \text{for every } k \neq c. \tag{22}$$

Proof of Lemma 1. The logit update is $z_k^+ = z_k - \eta g_k$. The updated non-target mean is:

$$\mu_{-c}^+ = \frac{1}{K-1} \sum_{j \neq c} z_j^+ = \frac{1}{K-1} \sum_{j \neq c} (z_j - \eta g_j) = \mu_{-c} - \eta \, \bar{g}. \tag{23}$$

The updated centered deviation of non-target logit k is:

$$\delta_k^+ = z_k^+ - \mu_{-c}^+ = (z_k - \eta g_k) - (\mu_{-c} - \eta \bar{g}) = \delta_k - \eta (g_k - \bar{g}). \tag{24}$$

□

To analyze how the spread evolves after one gradient step, we now apply Lemma 1 and derive explicit expressions for the centered gradients $g_k - \bar{g}$ under CE and CDL_{uni}. These expressions determine the linear part of the update for $\boldsymbol{\delta}_{-c}$ and lead directly to the contraction rates claimed in Theorem 1.

Main proof. Lemma 1 yields the centered update

$$\boldsymbol{\delta}_{-c}^+ = \boldsymbol{\delta}_{-c} - \eta \big(\mathbf{g}_{-c} - \bar{g} \, \mathbf{1} \big). \tag{25}$$

We analyze $g_k - \bar{g}$ assuming

$$\|\boldsymbol{\delta}_{-c}\| \leq r, \qquad m \geq m_0, \tag{26}$$

which ensure that all quantities remain in a bounded region, where the following Taylor approximations are valid with uniform constants.

Cross-entropy (CE) case. The gradient of CE with respect to logit z_k, for $k \neq c$, can be written as:

$$g_k = \text{softmax}(\mathbf{z})_k = \frac{\exp(\delta_k)}{\exp(m) + \sum_{j \neq c} \exp(\delta_j)}. \tag{27}$$

We use the fact that each δ_k is small (because $\sum_{k \neq c} \delta_k = 0$ and $\|\boldsymbol{\delta}_{-c}\|$ controls their magnitude). A first–order Taylor expansion of $\exp(x)$ around $x = 0$ then gives the following approximations:

$$\exp(\delta_k) = 1 + \delta_k + O(\|\boldsymbol{\delta}_{-c}\|^2), \tag{28}$$

$$\sum_{j \neq c} \exp(\delta_j) = (K-1) + O(\|\boldsymbol{\delta}_{-c}\|^2). \tag{29}$$

Substituting the expansions (28), (29) into (27) we obtain:

$$g_k = \frac{1 + \delta_k}{\exp(m) + K - 1} + O\left(\frac{\|\boldsymbol{\delta}_{-c}\|^2}{\exp(m) + K - 1} \right). \tag{30}$$

Since $\sum_{k \neq c} \delta_k = 0$, we have:

$$g_k - \bar{g} = \frac{\delta_k}{\exp(m) + K - 1} + r_k^{\text{CE}}, \tag{31}$$

$$\text{with } |r_k^{\text{CE}}| \leq C \, \frac{\|\boldsymbol{\delta}_{-c}\|^2}{\exp(m) + K - 1}. \tag{32}$$

Substituting into (25), we get:

$$\boldsymbol{\delta}_{-c}^+ = \left(1 - \frac{\eta}{\exp(m) + K - 1} \right) \boldsymbol{\delta}_{-c} + \mathbf{r}_{\text{CE}}, \tag{33}$$

$$\text{with } \|\mathbf{r}_{\text{CE}}\| \leq C \, \eta \, \frac{\|\boldsymbol{\delta}_{-c}\|^2}{\exp(m) + K - 1}. \tag{34}$$

Thus:

$$\sigma^+ = \left(1 - \frac{\eta}{\exp(m) + K - 1} \right) \sigma + O(\eta \sigma^2). \tag{35}$$

Proposed (CDL_{uni}) case. The gradient of CDL_{uni} with respect to logit z_k, for $k \neq c$, can be written as:

$$g_k = \frac{1}{K-1} - \mathrm{softmax}(-\mathbf{z})_k, \tag{36}$$

where

$$\mathrm{softmax}(-\mathbf{z})_k = \frac{\exp(-\delta_k)}{\exp(-m) + \sum_{j\neq c}\exp(-\delta_j)}. \tag{37}$$

As in the CE case, each δ_k is small, so a first–order Taylor expansion of $\exp(x)$ around $x = 0$ gives:

$$\exp(-\delta_k) = 1 - \delta_k + O(\|\boldsymbol{\delta}_{-c}\|^2), \tag{38}$$

$$\sum_{j\neq c}\exp(-\delta_j) = (K-1) + O(\|\boldsymbol{\delta}_{-c}\|^2). \tag{39}$$

Substituting (38) and (39) into (37) we obtain:

$$\mathrm{softmax}(-\mathbf{z})_k = \frac{1-\delta_k}{\exp(-m)+(K-1)} + O\left(\frac{\|\boldsymbol{\delta}_{-c}\|^2}{\exp(-m)+(K-1)}\right). \tag{40}$$

Using $\bar{g} = \frac{1}{K-1}\sum_{j\neq c} g_j$ and $\sum_{k\neq c}\delta_k = 0$, we get:

$$g_k - \bar{g} = \frac{\delta_k}{K-1} + r_k^{\mathrm{CDL}}, \tag{41}$$

$$\text{with } |r_k^{\mathrm{CDL}}| \leq C\left(\|\boldsymbol{\delta}_{-c}\|^2 + \exp(-m)\,\|\boldsymbol{\delta}_{-c}\|\right). \tag{42}$$

Substituting (41) into (25), we get:

$$\boldsymbol{\delta}^+_{-c} = \boldsymbol{\delta}_{-c} - \eta\left(\mathbf{g}_{-c} - \bar{g}\,\mathbf{1}\right), \tag{43}$$

which, componentwise for $k \neq c$, becomes

$$\delta_k^+ = \delta_k - \eta\left(\frac{\delta_k}{K-1} + r_k^{\mathrm{CDL}}\right). \tag{44}$$

Thus:

$$\sigma^+ = \left(1 - \frac{\eta}{K-1}\right)\sigma + O(\eta\sigma^2) + \cancelto{0 \text{ for } m \geq m_0}{O(\eta\exp(-m)\,\sigma)} \tag{45}$$

Conclusion. From (35) and (45), for $\sigma \leq r$ and $m \geq m_0$,

$$\sigma^+ = \left(1 - \frac{\eta}{\exp(m)+K-1}\right)\sigma + O(\eta\sigma^2) \text{ for CE,} \tag{46}$$

$$\sigma^+ = \left(1 - \frac{\eta}{K-1}\right)\sigma + O(\eta\sigma^2) \text{ for } \mathrm{CDL}_{uni}. \tag{47}$$

The CE update contracts at a rate proportional to $(\exp(m) + K - 1)^{-1}$, which decays exponentially with the margin m, while CDL_{uni} contracts by the margin–independent factor $(1 - \eta/(K-1))$. □

A3. Neural Collapse Metrics

For completeness, we summarize the metrics used to quantify neural collapse during training. These definitions follow the framework in [39]. The global mean $\mathbf{h}_G$ and class means $\mathbf{h}_k$ of the last-layer features $\{\mathbf{h}_{k,i}\}$ are

$$\mathbf{h}_G = \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbf{h}_{k,i}, \tag{48}$$

$$\mathbf{h}_k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_{k,i} \qquad (1 \leq k \leq K). \tag{49}$$

Within-class and Between-class Variability ($\mathcal{NC}_1$). The within-class and between-class covariance matrices are

$$\boldsymbol{\Sigma}_W = \frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}(\mathbf{h}_{k,i} - \mathbf{h}_k)(\mathbf{h}_{k,i} - \mathbf{h}_k)^\top, \tag{50}$$

$$\boldsymbol{\Sigma}_B = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{h}_k - \mathbf{h}_G)(\mathbf{h}_k - \mathbf{h}_G)^\top. \tag{51}$$

Neural collapse variability is measured by

$$\mathcal{NC}_1 = \frac{1}{K}\,\mathrm{trace}\left(\boldsymbol{\Sigma}_W\boldsymbol{\Sigma}_B^\dagger\right). \tag{52}$$

Simplex ETF Convergence ($\mathcal{NC}_2$). The alignment of the classifier matrix $\mathbf{W}$ with a Simplex ETF is quantified by

$$\mathcal{NC}_2 = \left\|\frac{\mathbf{W}\mathbf{W}^\top}{\|\mathbf{W}\mathbf{W}^\top\|_F} - \frac{1}{\sqrt{K-1}}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\right\|_F. \tag{53}$$

Convergence to Self-duality ($\mathcal{NC}_3$). Let

$$\mathbf{H} = \begin{bmatrix}\mathbf{h}_1 - \mathbf{h}_G & \dots & \mathbf{h}_K - \mathbf{h}_G\end{bmatrix} \in \mathbb{R}^{d\times K}. \tag{54}$$

Self-duality between classifier weights $\mathbf{W}$ and centered class means is measured by

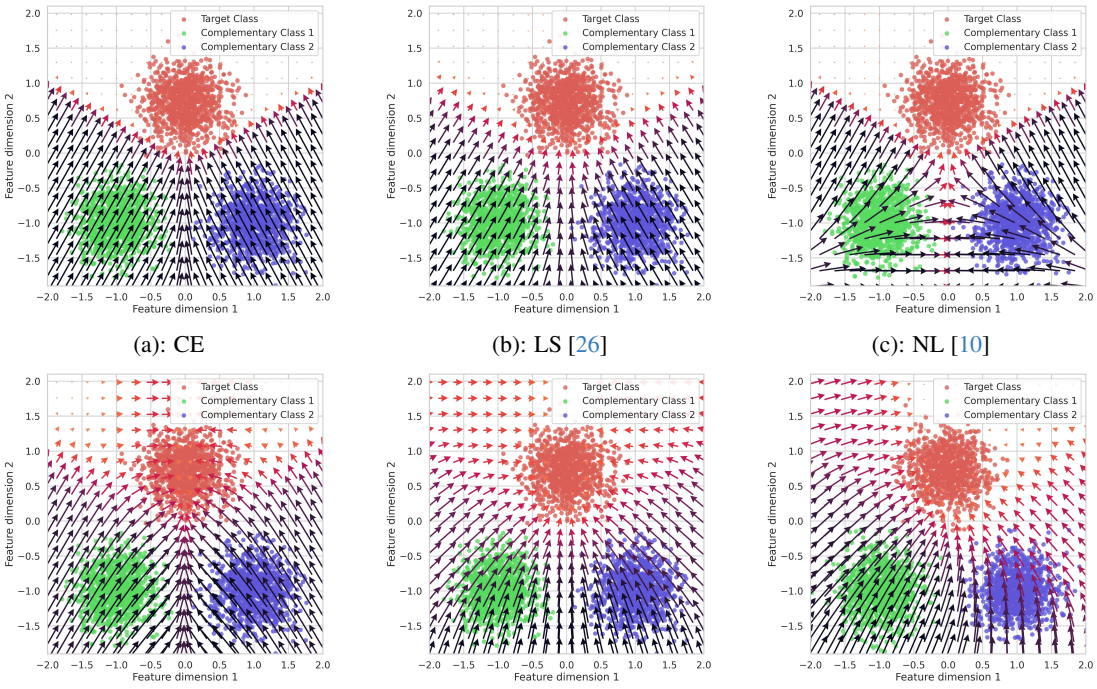
$$\mathcal{NC}_3 = \left\|\frac{\mathbf{W}\mathbf{H}}{\|\mathbf{W}\mathbf{H}\|_F} - \frac{1}{\sqrt{K-1}}\left(\mathbf{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\right\|_F. \tag{55}$$

Collapse of the Bias ($\mathcal{NC}_4$). If the global mean $\mathbf{h}_G$ is nonzero, the bias vector $\mathbf{b}$ often compensates it via

$$\mathbf{W}\mathbf{h}_{k,i} + \mathbf{b} = \mathbf{W}(\mathbf{h}_{k,i} - \mathbf{h}_G) + \underbrace{\mathbf{W}\mathbf{h}_G + \mathbf{b}}_{\approx 0}.$$

Thus, the collapse of bias is captured by the magnitude of

$$\mathcal{NC}_4 = \|\,\mathbf{b} + \mathbf{W}\mathbf{h}_G\,\|_2. \tag{56}$$



(a): CE (b): LS [26] (c): NL [10]

(d): COT [4] (e): Proposed CDL$_{uni}$ (f): Proposed CDL$_{self}$

Figure 1. Comparison of negated gradient vector fields for different loss functions.

A4. Extended Discussion

Figure 1 shows the negated gradient fields of the examined objectives on the same toy example as in Section 3.3, highlighting how each method shapes the feature space and the geometry of non-target classes.

Cross-Entropy (CE). CE increases the target logit while suppressing non-target logits only through the softmax. Gradients on non-target classes vanish rapidly once confidence grows, leaving most of the space without meaningful supervision. CE therefore neither structures non-target classes nor regulates feature geometry, and neural collapse arises only in late training after margin saturation.

Label Smoothing (LS) [26]. LS softens the target distribution but still treats all non-target classes uniformly and solely through their relation to the target. Gradients again decay for highly confident predictions, limiting the influence on non-target geometry. LS reduces overconfidence but also reduces margins and provides no mechanism for shaping inter-class structure.

Negative Learning (NL) [19]. NL penalizes complementary classes but, even when applied to all non-targets, mainly pushes their logits down without encoding their relationships. Its gradients remain dominated by softmax normalization, resulting in behavior close to CE in high-confidence regions and offering little control over the geometry or collapse dynamics of the representation.

Complement Objective Training (COT) [4]. COT encourages uniformity among non-target classes by maximizing their entropy relative to the target probability, yet the softmax suppresses gradients for distant classes and restricts the method to uniform complementary distributions. Consequently, COT flattens some logits but cannot impose richer geometric structure or meaningfully influence neural collapse patterns.

Proposed CDL$_{uni}$. CDL$_{uni}$ assigns a uniform non-target distribution directly to the negated logits, avoiding softmax attenuation and preserving strong, non-vanishing gradients across the entire space. This drives non-target logits toward a shared value, inducing a symmetric arrangement aligned with the Simplex ETF geometry associated with neural collapse. As shown in the main text, the resulting gradient fields contract the non-target spread independently of margin size and accelerate the emergence of $\mathcal{NC}_1$-$\mathcal{NC}_4$.

Proposed CDL$_{self}$. CDL$_{self}$ replaces the uniform prior with a self-distilled dissimilarity prior that adapts to the learned representations. This results in asymmetric gradients across non-target classes, allowing the model to emphasize truly dissimilar classes. CDL$_{self}$ thus preserves low $\mathcal{NC}_1$ and strong separation without enforcing strict ETF symmetry, enabling more flexible and realistic class geometries while retaining the global, non-vanishing gradient benefits of CDL.

B. Experimental Details

B1. Datasets and Network Architectures

Datasets. Table 1 summarizes all datasets used in our main experiments and Appendix. We evaluate our method across a broad spectrum of settings: (i) closed-set classification on CIFAR100, TinyImageNet, and the large-scale ImageNet-1k benchmark, (ii) open-set recognition on SVHN, CIFAR100, and TinyImageNet, (iii) few-shot learning on the standard CUB and MiniImageNet benchmarks, and (iv) domain generalization on the widely used PACS dataset, which includes four distinct test domains. This diversity ensures that our findings hold across small-, medium-, and large-scale datasets as well as across several challenging evaluation protocols.

Table 1. Evaluated datasets details.

Dataset	Dimensions	Classes	Train size
SVHN [20]	$32 \times 32 \times 3$	10	73,257
CIFAR100 [12]	$32 \times 32 \times 3$	100	50,000
TinyImageNet [13]	$64 \times 64 \times 3$	200	100,000
ImageNet-1k [5]	$224 \times 224 \times 3$	1000	1,281,167
CUB [29]	$84 \times 84 \times 3$	200	5900
MiniImageNet [28]	$224 \times 224 \times 3$	200	38,400
PACS [14]	$224 \times 224 \times 3$	7	9,991

Models. Table 2 lists all architectures used in our study. We evaluate our approach on a diverse set of models, ranging from lightweight networks (MobileNetV2), to medium-scale convolutional architectures (ResNet18/34, DenseNet121, WideResNet28-10), and up to large-scale convolutional architectures (ResNet50/101), as well as transformer-based backbones (DeiT-S). This broad coverage demonstrates that our findings hold across varying model capacities, depths, and architectural families.

Table 2. Evaluated models details.

Model	Parameters (M)	Feature dim.
ResNet18 [8]	11.7	512
ResNet34 [8]	21.8	512
ResNet50 [8]	25.6	2048
ResNet101 [8]	44.5	2048
DenseNet121 [9]	8.0	1024
MobileNetV2 [21]	3.4	1280
WideResNet28-10 [35]	36.5	640
DeiT-Small [27]	22.0	384

B2. Classification Experiments

Here we provide all the experimental settings used in our main closed-set classification experiments.

CIFAR-100 and TinyImageNet. We use two distinct hyperparameter configurations for these datasets.

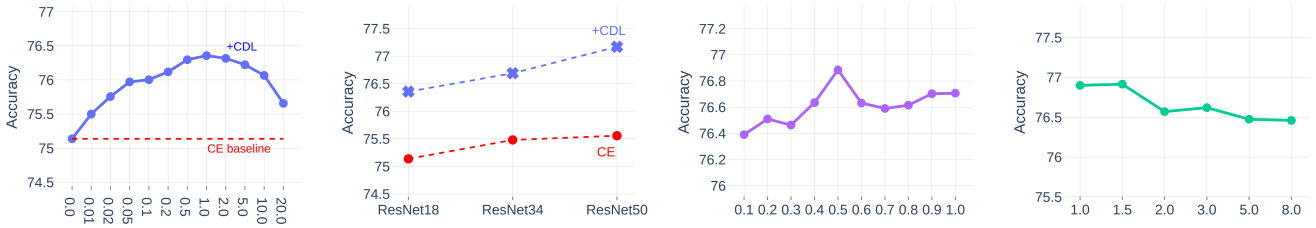
For the comparisons against cross-entropy variants and regularization methods (Table 1), we train for 150 epochs using SGD with momentum 0.9, batch size 128, weight decay 1×10^{-4}, and a base learning rate of 0.1 decayed by a factor of 0.1 at epochs 50 and 100. For the comparisons against Self-KD methods (Table 2), we adopt a longer schedule of 200 epochs with the same optimizer, batch size, and weight decay, and use a base learning rate of 0.1 decayed by 0.2 at epochs 60, 120 and 160.

For the baseline methods, we use official implementations when available and follow their recommended hyperparameters. For Label Smoothing (LS) [26], we set $\epsilon = 0.1$. Negative Learning (NL) [10] is used in combination with CE with weight $w = 1.0$. Complement Objective Training (COT) [4] and Complement Cross Entropy (CCE) [11] use a complement-entropy weight $\gamma = 1.0$, where COT employs two separate optimizers (one for CE and one for complement entropy), whereas CCE combines the losses into a single objective with one optimizer. For Focal Loss (FL) [17], we use $\epsilon = 0.1$ and $\gamma = 2.0$. For its variants Asymmetric Loss (ASL) [1] and Cyclical Focal Loss (CFL) [24], we set $\gamma_{\text{pos}} = 2$ and $\gamma_{\text{neg}} = 2$; for CFL we additionally use $\gamma_{\text{hc}} = 3$ and a cyclical factor of 4. For Maximum Suppression (MaxSup) [38], we follow the proposed linearly increasing schedule for the regularization weight a, starting at 0.1 and increasing to 0.2 over training.

For the Self-KD methods, we use each method's recommended configuration. For Online Label Smoothing (OLS) [36], we use $\alpha = 0.5$. For Zipf's Label Smoothing [16], we set $\lambda = 0.1$ and $\alpha = 1$, where λ controls the regularization strength and α controls the decay shape of the Zipf distribution. For Unified Self-Knowledge Distillation (USKD) [33], we use $\alpha = 1.0$ and $\beta = 0.1$, where α controls the target-class distillation and β the weak non-target distillation.

We also provide a sensitivity analysis of CDL's hyperparameters on CIFAR-100 with 5 seeds, which is prestented in Figure 6.

ImageNet. For ImageNet, we provide all training hyperparameters in the main text. For the compared methods, we follow their recommended hyperparameters as reported in the original papers. For Teacher-free Knowledge Distillation (Tf-KD) [34], we use the official implementation and align the training protocol with the other baselines. Specifically, we adopt their recommended temperature $\tau = 20$ and label-smoothing coefficient $\alpha = 0.1$.



(i) γ weight (ii) architecture scale (iii) α weight (γ=1, τ=1) (iv) τ weight (γ=1, α=0.5)

Figure 6. Sensitivity analysis on CIFAR-100 (5 seeds). **(i)** γ ablation shows that the influence CDL is stable, **(ii)** scaling ablation indicates increasing improvement margins with larger architectures, **(iii)** α ablation shows that self-KD with α=0.5 further improves accuracy, and **(iv)** τ ablation shows that $\tau \approx 1$ is optimal.

B3. Open-set Recognition Experiments

For our Open-set Recognition experiments, we employ the Maximum over Softmax Probabilities (MSP) [22] method to identify unknown classes, using a threshold that distinguishes the lower confidence of unknown instances from the higher confidence of known ones.

Regarding training hyperparameters, following [3], we use the same splits and network architecture for SVHN, where 4 classes are assigned as known and 6 as unknown. For TinyImageNet and CIFAR-100, we use a larger ResNet-18 architecture. In the case of TinyImageNet, we follow the standard split used in prior works (20 known and 180 unknown classes). For CIFAR-100, we define a setup with 50 known and 50 unknown classes.

We compare against Cross Entropy (CE), Binary Cross Entropy (BCE), Generalized Convolutional Prototype Learning (GCPL) [32], Reciprocal Points Learning (RPL) [2], and Adversarial Reciprocal Points Learning (ARPL) [3]. In ARPL, the authors propose the use of Confusing Samples (CS), which involve GAN generated adversarial samples for which the classification network should increase uncertainty by maximizing the entropy of its predictions. In our experiments, this results in improved performance but in an average training overhead of approximately four times longer. We also show how the same augmented confusing samples can be incorporated into our framework by assigning uniform dissimilarities to all known classes.

For our proposed CDL objective weighting factor, we select the value of γ for each dataset based on hold-out validation over the grid $\gamma \in \{0.1, 0.2, 0.5, 1\}$, while for CDL_{self} we set the distillation factor α to 0.5 for all experiments.

B4. Few-Shot Learning Experiments

Few-Shot Classification aims to enable models to quickly adapt to new tasks with limited data. Recent work has shown that the training and adaptation phases can be decoupled and optimized independently [18]. Additionally, neural collapse has been observed to generalize to unseen samples and classes, supporting the use of pretrained feature extractors in few-shot settings [7]. Building on these insights, we study how uniform and class-dependent anti-class distributions affect the emergence of neural collapse during pretraining and its impact on few-shot performance.

We evaluate our method using established meta-learning algorithms, including MatchingNet [28], ProtoNet [25], Finetune [6], and SimpleShot [31] on two widely adopted few-shot learning benchmarks: CUB, which consist of 100 train, 50 validation, and 50 test classes, and miniImageNet, which consist of 64 train, 16 validation, and 20 test classes. Following [6], we use WideResNet28-10 as the backbone and pretrain for 300 epochs with SGD, momentum of 0.9, batch size of 128, a base learning rate of 0.1, decayed at epochs 200 and 250 by 0.1. Evaluation is conducted on 5-way 1-shot and 5-shot tasks. The hyperparameter γ of our CDL objective is selected via a sweep over $\{0.02, 0.05, 0.1, 0.2\}$, using the validation classes for tuning, while the distillation factor α in CDL_{self} is set to 0.5 for all experiments.

B5. Domain Generalization Experiments

For the final set of experiments on Domain Generalization (DG), we consider multiple domains containing the same set of classes. We use the common benchmark dataset PACS [14], which consists of four domains: Photo, Art Painting, Cartoon, and Sketch. In DG, one of the domains is typically held out for testing, while the model is trained on the remaining domains. The goal is to generalize beyond the training domains, thereby evaluating robustness under distributional shifts.

Increasing the degree of neural collapse results in compact, symmetric class representations, where class-conditional features align with a Simplex ETF geometry. As a consequence, decision boundaries become more stable and less sensitive to spurious cues or background variations, promoting better generalization beyond the training distribution and improved robustness to distribution shifts.

For our experiments, we used the DeepDG toolkit [30] and evaluated four commonly used DG methods: Empirical Risk Minimization (ERM), Mixup [37], MMD [15], and DANN [23]. All these methods employ cross-entropy loss, allowing our Complementary Dissimilarity Loss (CDL) to

be seamlessly applied as a complementary objective that encourages class dissimilarities to follow a uniform distribution, thereby suppressing spurious correlations.

We trained a ResNet-18 for 70 epochs with a base learning rate of 10^{-2} using a cosine scheduler, SGD with momentum 0.9, a batch size of 64, and weight decay of 10^{-4}. Following previous works, we split the training data into 90% train and 10% validation, and for each test domain we use the corresponding validation set to select the CDL weighting factor γ via grid search over $\{0.1, 0.2, 0.5, 1.0\}$. For the baseline methods, we use their recommended hyperparameters, specifically $\alpha = 0.2$ for Mixup, $\alpha = 1$ for DANN, and $\gamma_{\text{MMD}} = 1$ for MMD.

To further analyze the robustness of the learned representations under domain shifts, we examine the intra-class structure of the features on the PACS benchmark. Specifically, we evaluate two complementary metrics on the target domain: (i) $\mathcal{NC}_1$: the Neural Collapse within-/between-class variability metric, which measures the ratio of intra-class feature variance to the separation between class means. Lower values indicate more compact and better separated class representations, (ii) Intra-Class Domain Alignment (ICDA): the average cosine similarity between the source-domain and target-domain intra-class feature centroids. Higher values indicate stronger alignment of class representations across domains.

Intuitively, $\mathcal{NC}_1$ captures how tightly samples cluster around their class centers in the target domain, while ICDA measures how well the class structure learned on the source domains transfers to the unseen target domain. Table 3 reports these metrics for standard ERM training with cross-entropy (ERM) training and for CE augmented with the proposed Complementary Dissimilarity Loss (CDL). We observe that CDL substantially reduces $\mathcal{NC}_1$ across all domains, indicating significantly more compact and better separated class representations. At the same time, CDL consistently increases ICDA, demonstrating stronger alignment between source and target domain intra-class representations.

Table 3. Domain generalization intra-class robustness analysis on PACS. Lower $\mathcal{NC}_1$ indicates tighter class clustering, while higher ICDA indicates stronger alignment of class representations across domains.

Metric	Method	P	A	C	S	Mean
$\mathcal{NC}_1$ (↓)	ERM	0.898	1.015	0.618	1.192	0.931
	+CDL	**0.216**	**0.218**	**0.129**	**0.443**	**0.252**
ICDA (↑)	ERM	0.953	0.966	0.947	0.952	0.955
	+CDL	**0.984**	**0.987**	**0.976**	**0.987**	**0.984**

References

[1] Emanuel Ben Baruch, T. Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91, 2021. 5

[2] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. *Learning Open Set Network with Discriminative Reciprocal Points*, page 507–522. Springer International Publishing, 2020. 6

[3] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(11):8065–8081, 2022. 6

[4] Hao-Yun Chen, Pei-Hsin Wang, Chun-Hao Liu, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. Complement objective training. In *International Conference on Learning Representations (ICLR)*, 2019. 4, 5

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 5

[6] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *International Conference on Learning Representations (ICLR)*, 2020. 6

[7] Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations (ICLR)*, 2022. 6

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, pages 630–645, Cham, 2016. 5

[9] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 5

[10] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 101–110, 2019. 4, 5

[11] Yechan Kim, Younkwan Lee, and Moongu Jeon. Imbalanced image classification with complement cross entropy. *Pattern Recognition Letters*, 151:33–40, 2021. 5

[12] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. 5

[13] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015. 5

[14] 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 (ICCV)*, pages 5542–5550, 2017. 5, 6

[15] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2018. 6

[16] Jiajun Liang, Linze Li, Zhaodong Bing, Borui Zhao, Yao Tang, Bo Lin, and Haoqiang Fan. Efficient one pass self-distillation with zipf's label smoothing. In *Computer Vision – ECCV 2022*, pages 104–119, Cham, 2022. Springer Nature Switzerland. 5

[17] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 5

[18] Xu Luo, Hao Wu, Ji Zhang, Lianli Gao, Jing Xu, and Jingkuan Song. A closer look at few-shot classification again. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 23103–23123. PMLR, 2023. 6

[19] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2013. 4

[20] 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. 5

[21] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018. 5

[22] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(7):1757–1772, 2013. 6

[23] Anthony Sicilia, Xingchen Zhao, and Seong Jae Hwang. Domain adversarial neural networks for domain generalization: when it works and how to improve. *Mach. Learn.*, 112(7): 2685–2721, 2023. 6

[24] Leslie N. Smith. Cyclical focal loss, 2022. 5

[25] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4080–4090. Curran Associates Inc., 2017. 6

[26] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. 4, 5

[27] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers amp; distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 5

[28] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 3637–3645. Curran Associates Inc., 2016. 5, 6

[29] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Cub200. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5

[30] Jindong Wang and Wang Lu. Deepdg: Deep domain generalization toolkit. https://github.com/jindongwang/transferlearning/tree/master/code/DeepDG. 6

[31] Yan Wang, Wei-Lun Chao, Kilian Q. Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019. 6

[32] Hong-Ming Yang, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Robust classification with convolutional prototype learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3474–3482, 2018. 6

[33] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17185–17194, 2023. 5

[34] Li Yuan, Francis Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. pages 3902–3910, 2020. 5

[35] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016. 5

[36] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. *Trans. Img. Proc.*, 30:5984–5996, 2021. 5

[37] Hongyi Zhang, Moustapha Cisse, Yann Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. 2017. 6

[38] Yuxuan Zhou, Heng Li, Zhi-Qi Cheng, Xudong Yan, Yifei Dong, Mario Fritz, and Margret Keuper. Maxsup: Overcoming representation collapse in label smoothing, 2025. 5

[39] Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2021. 3