Neural Collapse Inspired Knowledge Distillation

Shuoxi Zhang, Zijian Song, Kun He*

School of Computer Science and Technology, Huazhong University of Science and Technology
{zhangshuoxi,songzijian88,brooklet60}@hust.edu.cn,

Abstract

Existing knowledge distillation (KD) methods have demonstrated their ability to achieve student network performance on par with their teachers. However, the knowledge gap between the teacher and student remains significant and may hinder the effectiveness of the distillation process. In this work, we introduce the structure of Neural Collapse (NC) into the KD framework. NC typically occurs in the final phase of training, resulting in a graceful geometric structure where the last-layer features form a simplex equiangular tight frame. We hypothesize that NC can alleviate the knowledge gap in distillation, thereby enhancing student performance. This paper begins with an empirical analysis to bridge the connection between KD and NC. Through this analysis, we establish that transferring the teacher's NC structure to the student benefits the distillation process. Therefore, instead of merely transferring instance-level logits or features, as done by existing distillation methods, we encourage students to learn the teacher's NC structure. We propose the new distillation paradigm termed Neural Collapse-inspired Knowledge Distillation (NCKD). Comprehensive experiments demonstrate that NCKD is simple yet effective, improving the generalization of all distilled student models and achieving state-of-the-art accuracy performance.

Introduction

In recent decades, deep learning has made remarkable strides in the field of computer vision, resulting in significant advancements in performance and generalization across various downstream tasks, including image classification (He et al. 2016; Hu, Shen, and Sun 2018; Dosovitskiy et al. 2020), object recognition (Lin et al. 2017; Chen et al. 2019), and semantic segmentation (Poudel, Liwicki, and Cipolla 2019), *etc*.

These remarkable achievements have been largely attributed to the effectiveness of over-parameterized networks. However, the cumbersome deep models typically require substantial computation and memory resources during training and inference stages, making it challenging to deploy on mobile devices with limited resources. To address this issue, knowledge distillation (Hinton, Vinyals, and Dean 2015) (KD) has emerged as a crucial technique for model compression and performance improvement. By transferring knowledge encapsulated in a large, well-trained teacher model to a smaller student model, KD may achieve comparable performance in a resource-efficient manner. This process is particularly beneficial in scenarios where deploying large models is impractical due to computational constraints. Despite its widespread adoption, the efficacy of KD is often limited by a persistent knowledge gap between the teacher and student models, resulting in suboptimal student performance.

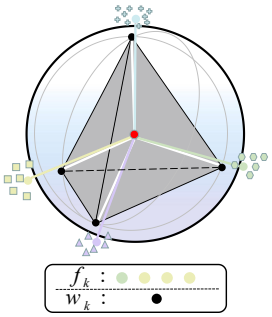


Figure 1: Description of the structure of Neural Collapse. All class features collapse toward their centroids, forming an equiangular structure. Also, classifier $\boldsymbol{w}$ will align with its corresponding last-layer normalized centroid $\tilde{\boldsymbol{f}}$.

Meanwhile, a parallel line of research has uncovered the phenomenon of `Neural Collapse` (Papyan, Han, and Donoho 2020) (NC), where the final layer representations of a deep neural network exhibit a surprisingly symmetric and structured geometry as training progresses. Neural collapse is characterized by the alignment of within-class feature vectors, which converge to their respective class means, forming a simplex equiangular tight frame (ETF) (see Figure 1). The occurrence and prevalence of NC have been empirically verified through experiments with various datasets and network architectures(Zhou et al. 2022).

The study of NC arguably provides a better understanding on the properties of deep features. Nonetheless, existing research has not addressed the following questions: *Are desirable KD methods the result of inducing a better simplex ETF structure? Can we improve the distillation process by*

*Corresponding Author.

encouraging the student to learn the teacher's NC structure?

Given the geometric elegance and generalization benefits of neural collapse, we hypothesize that integrating these properties into the student model can bridge the knowledge gap more effectively. Thus, we strive to investigate whether existing KD techniques enable the student model to obtain the NC structure of the teacher and leverage this phenomenon to enhance KD performance.

In this paper, we first conduct an empirical analysis to explore the relationship between the student's NC structure and its impact on the distillation process. Through this analysis, we establish that a well-aligned NC structure plays a crucial role in bridging the knowledge gap and improving performance within the KD paradigm. Accordingly, we exploit the properties of $\mathcal{NC}_1$, where the features of a well-trained network collapse towards their respective centroids. We design a contrastive loss that encourages the student's feature space to align with the teacher's centroids. Next, we extend this approach by transferring the teacher's ETF structure to the student, ensuring that the student's class not only aligns with the corresponding teacher centroids but also forms a consistent ETF structure relative to other classes. Finally, considering that the primary goal of KD is to reduce computational costs, we capitalize on the properties of $\mathcal{NC}_3$ by using normalized prototypes as the classifier, thereby reducing computational overhead. The above three key components form the foundation of our **N**eural **C**ollapse-inspired **K**nowledge **D**istillation (NCKD) framework.

We conduct extensive experiments to evaluate the effectiveness of NCKD across various benchmarks. Our method not only outperforms state-of-the-art distillation techniques on multiple vision tasks but also demonstrates its versatility as a plug-and-play loss that can be integrated into other popular distillation methods to enhance their performance.

Our main contributions can be summarized as follows:

- We explore the intersection of two intensively studied fields, knowledge distillation and neural collapse, and attempt to establish a connection. To the best of our knowledge, we are the first to apply the principles of NC within the KD framework.
- We distill the teacher's NC structure into the student model. Our approach goes beyond merely distilling class semantics; more critically, we also distill the ETF structure formed by the classes, thereby encouraging the student to construct a similarly elegant structure as that of the teacher.
- Our approach consistently outperforms state-of-the-art baselines in extensive experiments, encompassing various network architectures and diverse tasks including classification and detection.

Related Work

In this section, we first provide a brief overview of the related studies on knowledge distillation. Following that, we review the research literature on neural collapse and discuss its applications in various specific domains.

Knowledge Distillation

Knowledge distillation (Hinton, Vinyals, and Dean 2015) was first introduced by Hinton *et al*., who utilized dark knowledge hidden within the well-trained teacher network to improve the performance of the student. They employed the probabilistic relationships from the negative logits to provide additional supervision and better regularization (Yun et al. 2020). Building on this, logit-based distillation has demonstrated its potential in improving student model performance and generalization. Subsequent works have further refined logit-based KD through structural information (Park et al. 2019) or graph-level knowledge (Zhang, Liu, and He 2024). However, a significant knowledge gap persists between teacher and student models, prompting researchers to explore more effective knowledge transfer methods. For example, Zhao *et al*. (Zhao et al. 2022) decoupled traditional KD loss to achieve more efficient and adaptable distillation.

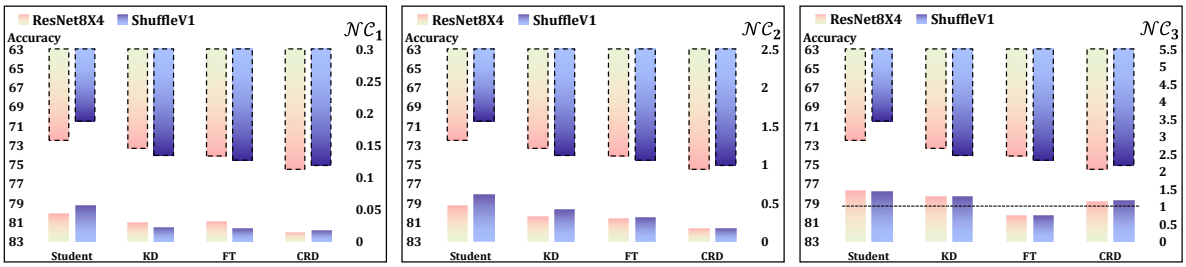
Another line of KD research leverages information concealed in intermediate features, attempting to align the feature maps between the teacher and student. FitNet (Romero et al. 2014) initiated this line by mimicking the teacher's intermediate features, setting the stage for feature-based distillation. Subsequent methods have refined the alignment and knowledge transfer from teacher features, incorporating attention mechanisms (Zagoruyko and Komodakis 2016; Guo et al. 2023), neural selectivity (Huang and Wang 2017), and specifically designed alignment modules (Kim, Park, and Kwak 2018; Chen et al. 2021a,b; Zheng and Yang 2024).

Neural Collapse

Neural collapse (NC) refers to a phenomenon where the features and classifiers of a neural network's final layer progressively converge to form a simplex *equiangular tight frame* (ETF), an elegant geometric structure. Empirical evidence of NC has been observed with both cross-entropy loss (Papyan, Han, and Donoho 2020; Lu and Steinerberger 2022) and mean squared error (MSE) loss (Zhou et al. 2022). This phenomenon is pervasive in deep training, arising unbiased to disparate datasets or architectures. Conceptually, NC represents the network's goal to maximize inter-class distances, thereby enhancing both generalization and adversarial robustness (Papyan, Han, and Donoho 2020). Consequently, NC has been effectively employed to improve performance in areas such as class incremental learning (Yang et al. 2023; Seo et al. 2024) and out-of-distribution detection (Ammar et al. 2023). However, the manifestation of NC in knowledge distillation, and its potential integration into distillation strategies, remain largely unexplored.

Problem Formulation

In this section, we first introduce several fundamental KD methods for subsequent analysis and provide necessary notations to facilitate the ensuing illustrations. We then review the concept of neural collapse, outlining its core properties and the metrics used to characterize this phenomenon. Finally, we empirically examine the impact of neural collapse



(a) $\mathcal{NC}_1$ (b) $\mathcal{NC}_2$ (c) $\mathcal{NC}_3$

Figure 2: Comparison of NC metrics and prediction performance across different methods. Both networks were distilled from ResNet32x4 on CIFAR-100. The ideal NC results are characterized by $\mathcal{NC}_{1,2}$ approaching 0, and $\mathcal{NC}_3$ approaching 1.

on the generalization of networks trained with various representative KD methods.

Knowledge Distillation

Consider the K-class classification problem $\mathcal{D} = \{(\boldsymbol{x}_k^{(n)}, \boldsymbol{y}_k)\}_{k\in[K],n\in[N_k]}$. Here N_k is the number of samples in the k-th class. For simplicity, our distillation framework assumes a balanced dataset, meaning $N_k \equiv N$, resulting in a total dataset size of $N*K$. Each sample consists of a data point $\boldsymbol{x}_k^{(n)}$ and the one-hot label $\boldsymbol{y}_k \in \mathbb{R}^K$. In addition, we utilize $\boldsymbol{f}$ and $\boldsymbol{z}$ to denote the feature function of the last layer and the corresponding logits, respectively.

In the basic KD paradigm, knowledge from the teacher is encapsulated and transferred through prediction logits or intermediate features. The total loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha\mathcal{L}_{\text{distill}}, \tag{1}$$

where $\mathcal{L}_{\text{cls}}$ is the classification loss with ground-truth labels, and the term $\mathcal{L}_{\text{distill}}$ indicates the distillation loss.

In Hinton's vanilla KD, it uses $\mathcal{L}_{KD}$ as $\mathcal{L}_{\text{distill}}$ to measure the KL divergence (Joyce 2011) of softened logit predictions $(\boldsymbol{z}_S, \boldsymbol{z}_T)$ between the teacher and the student:

$$\mathcal{L}_{KD} = \tau^2 KL\big(\sigma(\boldsymbol{z}_S/\tau), \sigma(\boldsymbol{z}_T/\tau)\big), \tag{2}$$

where σ denotes the Softmax operation, and the temperature τ is used to soften the logits.

Beyond distilling knowledge from logits, researchers also leverage the knowledge contained in intermediate features. Feature-based methods (*e.g.*, FT (Kim, Park, and Kwak 2018)) leverage the intermediate features from the teacher to guide the student's training. Accordingly, the distillation loss $\mathcal{L}_{FT}$ for $\mathcal{L}_{\text{distill}}$ is given by:

$$\mathcal{L}_{FT} = \mathcal{D}\left(\Phi(f_L^T), F_L^S\right), \tag{3}$$

where F_L^T and F_L^S are the L-th intermediate features of teacher and student, respectively. $\mathcal{D}(\cdot)$ denotes the distance function, utilized to measure the discrepancy of the selected features and thereby guide the distillation process. Extra transformation layer Φ aligns the feature sizes between teacher and student.

Neural Collapse

Neural collapse constructs an elegant geometric structure on the last-layer feature and the classifier in the final training phase. For simplicity, we denote the last-layer feature of the sample $\boldsymbol{x}_k^{(n)}$ by $\boldsymbol{f}_k^{(n)}$. Then, the k-th *class means* and *global mean* of the features are calculated by:

$$\boldsymbol{f}_k := \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{f}_k^{(n)}, \qquad \boldsymbol{f}_G := \frac{1}{K}\sum_{k=1}^{K}\boldsymbol{f}_k.$$

The NC phenomenon includes the following properties:

1. **NC1: Within-class variability collapse.** $\mathcal{NC}_1$ depicts the relative magnitude of within-class variability $\boldsymbol{\Sigma_W} = \frac{1}{NK}\sum_{k=1}^{K}\sum_{n=1}^{N}(\boldsymbol{f}_k^{(n)} - \boldsymbol{f}_k)(\boldsymbol{f}_k^{(n)} - \boldsymbol{f}_k)^\top$ in relation to the total variability. We compute $\mathcal{NC}_1$ by using within-class covariance $\boldsymbol{\Sigma_W}$ and between-class covariance $\boldsymbol{\Sigma_B} = \frac{1}{K}\sum_{k=1}^{K}(\boldsymbol{f}_k - \boldsymbol{f}_G)(\boldsymbol{f}_k - \boldsymbol{f}_G)^\top$. Thus, we can measure the $\mathcal{NC}_1$ collapse by measuring the magnitude of the between-class covariance $\boldsymbol{\Sigma_B} \in \mathbb{R}^{d\times d}$ compared to the within-class covariance $\boldsymbol{\Sigma_W} \in \mathbb{R}^{d\times d}$ of the learned features via:

$$\mathcal{NC}_1 := \frac{1}{K}\text{Trace}\left(\boldsymbol{\Sigma_W}\boldsymbol{\Sigma_B}^\dagger\right), \tag{4}$$

where $\boldsymbol{\Sigma_B}^\dagger$ denotes the pseudo inverse of $\boldsymbol{\Sigma_B}$.

2. **NC2: Convergence to Simplex ETF.** The penultimate feature centroids exhibit a simplex ETF structure with the following property: if we define the normalized class means as $\tilde{\boldsymbol{f}}_k = \frac{\boldsymbol{f}_k - \boldsymbol{f}_G}{\|\boldsymbol{f}_k - \boldsymbol{f}_G\|_2}$, then $\langle\tilde{\boldsymbol{f}}_k, \tilde{\boldsymbol{f}}_{k'}\rangle = -\frac{1}{K-1}$ for $k \neq k'$, indicating that the centered means are equiangular. Then we define the $\mathcal{NC}_2$ as:

$$\mathcal{NC}_2 = \text{avg}_{k\neq k'}\left(\left|\langle\tilde{\boldsymbol{f}}_k, \tilde{\boldsymbol{f}}_{k'}\rangle + \frac{1}{K-1}\right|\right). \tag{5}$$

3. **NC3: Convergence to self-duality.** The within-class means centered by the global mean will be aligned with their corresponding classifier weights, which means the classifier weights will converge to the same simplex ETF:

$$\mathcal{NC}_3 = \text{avg}\left\|\frac{\langle\tilde{\boldsymbol{f}}_k, \boldsymbol{w}_k\rangle}{\|\tilde{\boldsymbol{f}}_k\| \cdot \|\boldsymbol{w}_k\|}\right\|_F. \tag{6}$$

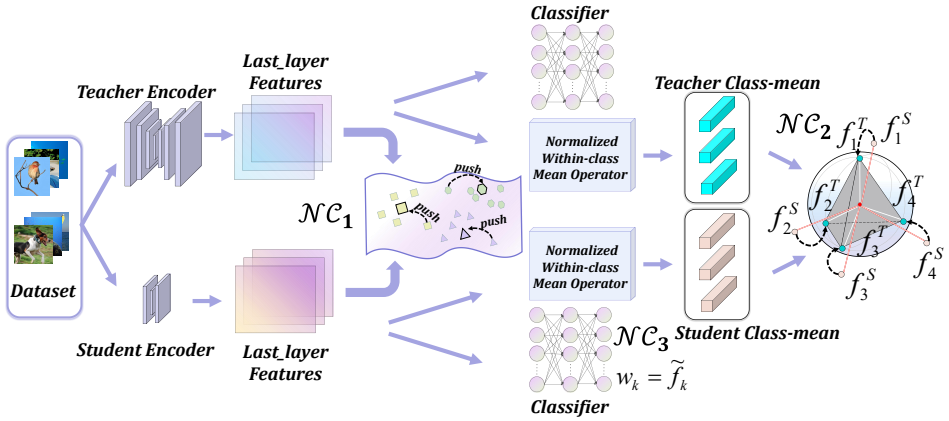


Figure 3: The overall framework of our NCKD. We distill the $\mathcal{NC}_{1,2}$ from the teacher to the student. We normalize within-class mean $\boldsymbol{f}$ to $\tilde{\boldsymbol{f}}$ to construct the ETF structure. Then the teacher's ETF structure information can be transerred to the student via $\mathcal{NC}_2$ distillation. $\mathcal{NC}_3$ classifier is leveraged to maintain the NC self-dual structure.

The Relationship between KD and NC Properties

We evaluate the student model's last-layer feature and classifier under different training conditions — namely, standalone student training, KD, FT, and CRD (Tian, Krishnan, and Isola 2019) — and compare the resulting NC metrics with their respective distillation performance (as shown in Figure 2). In both distillation pairs, a strong correlation between NC and distillation outcomes is evident. Improved distillation often corresponds with decreases in $\mathcal{NC}_1$ and $\mathcal{NC}_2$. Among the methods, CRD achieves the best distillation results, with $\mathcal{NC}_1$ and $\mathcal{NC}_2$ values closest to zero and $\mathcal{NC}_3$ closest to one. This indicates that the distillation process may implicitly steer the student toward an optimal NC structure. Thus, directly leveraging NC properties in distillation would be a highly effective strategy.

The Proposed Method

Building on the relationship between NC and KD, we propose our NC-inspired distillation method to promote the NC-like behavior of the student model. Our approach comprises three key components: 1) a contrastive learning module that aligns the student with the teacher's prototypes; 2) a mechanism to distill the teacher's neural ETF structure into the student; and 3) a $\mathcal{NC}_3$ classifier designed to reduce computation. The overall framework is shown in Figure 3.

$\mathcal{NC}_1$ Distillation

In the above analysis, we have established the $\mathcal{NC}_1$ property of a well-trained network, indicating that the last-layer features exhibit reduced within-class variance, effectively collapsing to their respective class centroids. This naturally leads to the idea of directly aligning the student features with the teacher's corresponding prototypes. To achieve this alignment, we leverage the paradigm of contrastive learning, which has already demonstrated its ability to preserve the NC phenomenon (Kini et al. 2023). We introduce the prototype alignment loss as follows:

$$\mathcal{L}_{\mathcal{NC}_1} = -\frac{1}{NK}\sum_{n,k}^{NK}\log\frac{\exp\left(\mathrm{sim}\left(g^S(x_k^{(n)}), \boldsymbol{f}_k^T\right)/\tau\right)}{\sum_{k=1}^{K}\exp\left(\mathrm{sim}\left(g^S(x_k^{(n)}, \boldsymbol{f}_k^T\right)/\tau\right)}. \tag{7}$$

Here, τ is the temperature that controls the feature space structure, and $\mathrm{sim}(\cdot)$ denotes the similarity measure. To address the norm gap between teacher and student, as discussed in (Wang et al. 2023), we use standard cosine similarity $\cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a}\cdot\boldsymbol{b}}{|\boldsymbol{a}|\cdot|\boldsymbol{b}|}$, to quantify the disparity between student features and their corresponding teacher centers.

The CRD loss (Tian, Krishnan, and Isola 2019) is most closely related to our approach, as it also employs a contrastive framework to enhance distillation matching. However, the key difference lies in the alignment strategy: while CRD aligns teacher and student features on an instance-wise basis, our method directly aligns student features with the teacher's prototypes. This design choice is driven by the observation that a well-trained teacher's features naturally collapse toward class centers, reflecting the $\mathcal{NC}_1$ property. In the experimental section, we will compare the effects of the two loss functions.

$\mathcal{NC}_2$ Distillation

To fully leverage the structured feature space of a well-trained teacher model, it is essential to distill the simplex ETF structure into the student model. As described earlier, the modified within-class feature means $\tilde{\boldsymbol{f}}_k$ collectively form an equiangular fabric. For simplicity, we organize all **prototype**s of the student and teacher into matrices $\tilde{\boldsymbol{F}}^S, \tilde{\boldsymbol{F}}^T \in \mathbb{R}^{K\times D}$, where each row represents the corresponding class mean. We aim to ensure that each student's normalized centroid $\tilde{\boldsymbol{f}}_k^S$ mimics the ETF structure of the teacher, thereby preserving the inter-class relationships. To achieve this, we propose the following loss function:

$$\mathcal{L}_{\mathcal{NC}_2} = \left\| \tilde{\boldsymbol{F}}^S(\tilde{\boldsymbol{F}}^T)^\top - \frac{K}{K-1}\left(\boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K{}^\top\right)\right\|_2^2. \tag{8}$$

Method	Homogeneous architecture			Heterogeneous architecture			Average
	ResNet-56	WRN-40-2	ResNet-32×4	ResNet-50	ResNet-32×4	ResNet-32×4	
	ResNet-20	WRN-40-1	ResNet-8×4	MobileNet-V2	ShuffleNet-V1	ShuffleNet-V2	
`teacher` (T)	72.34	75.61	79.42	79.34	79.42	79.42	77.59
`student` (S)	69.06	71.98	72.50	64.60	70.50	71.82	70.08
Logit-based Method							
KD	70.66	73.54	73.33	67.65	74.07	74.45	72.28
DKD	71.97	74.81	75.44	70.35	76.45	77.07	74.34
DIST	71.78	74.42	75.79	69.17	75.23	76.08	73.75
MLKD	72.19	75.35	76.98	69.58	77.18	**77.92**	74.87
Feature-based Method							
FitNet	69.21	72.24	73.50	63.16	73.59	73.54	70.87
RKD	69.61	72.22	71.90	64.43	72.28	73.21	70.61
CRD	71.16	74.14	75.51	69.11	75.11	75.65	73.45
ReviewKD	71.89	75.09	75.63	69.89	77.45	77.78	74.62
NORM	71.55	74.82	76.49	70.56	77.42	77.87	74.79
SimKD	71.68	75.56	77.22	70.32	77.11	75.42	74.55
TTM	71.83	74.32	76.17	69.24	74.18	76.52	73.71
NCKD	**72.63**	75.71	77.23	70.12	77.48	77.42	75.10
CRD+NCKD	72.26(↑1.10)	75.16(↑1.02)	76.88(↑2.74)	69.88(↑0.77)	76.32(↑1.21)	76.68(↑1.03)	**75.53(↑2.08)**
SimKD+NCKD	72.47(↑0.79)	**75.81(↑0.25)**	**78.18(↑0.94)**	**70.67(↑0.35)**	**77.71(↑0.60)**	76.98(↑1.56)	75.30(↑0.75)

Table 1: Benchmarking results (mean of three repeats) on the CIFAR-100. Methods are reported with top-1 accuracy (%). ↑ indicates the improvement of our approach when incorporated into others. The best results are highlighted with **bold**.

Here, I_K represents the identity matrix of dimension K, and $\mathbf{1}_K$ denotes a vector of ones with K elements. Notably, the product $\mathbf{1}_K\mathbf{1}_K^\top$ yields a $K \times K$ matrix where all elements are equal to 1. Ideally, when the $\mathcal{L}_{\mathcal{NC}_2}$ loss is optimized to 0, each normalized centroid $\tilde{\boldsymbol{f}}_k^S$ of the student model will have a similarity score of 1 with the corresponding teacher's centroid $\tilde{\boldsymbol{f}}_k^T$, while displaying an inner product of $-\frac{1}{K-1}$ with the centroids of other classes, thus elegantly matching the teacher's simplex ETF structure. Consequently, optimizing this loss allows us to effectively distill the $\mathcal{NC}_2$ structural knowledge from teacher to student, ensuring that the student model accurately mimics the geometric configuration of the teacher's class centroids. The total loss in our framework can be formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_1\mathcal{L}_{\mathcal{NC}_1} + \lambda_2\mathcal{L}_{\mathcal{NC}_2}, \tag{9}$$

where λ_1, λ_2 are the balancing coefficients.

$\mathcal{NC}_3$-inspired Classifier

The primary goal of KD is to minimize computational costs in practical applications while maintaining the model performance. Given the previously discussed $\mathcal{NC}_3$ property, where the final-layer features tend to form a self-dual space towards the end of the training phase — meaning that the features of each class align closely with their corresponding functional (i.e., the classifier). Therefore, a natural idea is to eliminate the classifier computation. Instead, we utilize the normalized centroid to represent the corresponding classifier weight $\boldsymbol{w}$ in the following form:

$$\boldsymbol{w}_k = \tilde{\boldsymbol{f}}_k.$$

This approach leverages the $\mathcal{NC}_3$ property to reduce computational overhead by eliminating the need for a separate linear classification layer. Notably, several existing distillation methods, such as SimKD (Chen et al. 2022), also implicitly utilize the $\mathcal{NC}_3$ property, though this is not always explicitly recognized in their design. We will explore this aspect further in our experimental section through a case study, providing additional insights.

Experiments

Baselines

We compare our approach with two main kinds of KD baselines[1] (*i.e.*, logit-based and feature-based distillation):

- **Logit-based** methods include the vanilla KD (Hinton, Vinyals, and Dean 2015), DKD (Zhao et al. 2022), DIST (Huang et al. 2022) and MLKD (Jin, Wang, and Lin 2023).
- **Feature-based** methods include FitNet (Romero et al. 2014), RKD (Park et al. 2019), CRD (Tian, Krishnan, and Isola 2019), ReviewKD (Chen et al. 2021b), FGFI (Wang et al. 2019), NORM (Liu et al. 2023), SimKD (Chen et al. 2022) and TTM (Zheng and Yang 2024).

Main Results

CIFAR-100. To validate the effectiveness of our approach, we compared NCKD against a range of state-of-the-art distillation methods. Our experiments included both similar-architecture and cross-architecture distillation to demonstrate the universality of our method. As shown in Table 1, NCKD outperformed all existing baselines, achieving an average accuracy of 75.10%. Additionally, when we integrated our NC-inspired losses as a plug-in module into two mainstream methods, CRD and SimKD, we observed a significant improvement in distillation performance. These results confirm the effectiveness of our approach in enhancing distillation generalization and highlight its versatility as a plug-and-play module suitable for various distillation frameworks and real-world applications.

[1]The detailed implementation of the experiments is provided in the appendix, available at https://arxiv.org/abs/2412.11788.

Student (Teacher)	Metric	Teacher	Student	FT	KD	SP	CRD	ReviewKD	DIST	TTM	DisWOT	NCKD
ResNet18 (ResNet34)	Top-1	73.31	69.75	70.70	70.66	70.62	71.17	71.61	71.88	72.09	72.08	**72.44**
	Top-5	91.42	89.07	90.00	89.88	89.80	90.13	90.51	90.42	90.48	90.38	**91.12**
MobileNet (ResNet50)	Top-1	76.16	70.13	70.78	70.68	70.99	71.37	72.56	72.94	73.09	73.22	**73.61**
	Top-5	92.86	89.49	90.50	90.30	90.61	90.41	91.00	91.12	90.77	90.22	**91.56**

Table 2: Evaluation results of baseline settings on ImageNet. We use ResNet34 and ResNet50 as our teacher network.

		mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
Method	Teacher	ResNet101			ResNet101			ResNet50		
		42.04	62.48	45.88	42.04	62.48	45.88	40.22	61.02	43.81
	Student	ResNet18			ResNet50			MobileNetV2		
		33.26	53.61	35.26	37.93	58.84	41.05	29.47	48.87	30.90
Feature	FitNet	34.13	54.16	36.71	38.76	59.62	41.80	30.20	49.80	31.69
	FGFI	35.44	55.51	38.17	39.44	60.27	43.04	31.16	50.68	32.92
	ReviewKD	36.75	56.72	34.00	40.36	60.97	44.08	33.71	53.15	36.13
Logits	KD	33.97	54.66	36.62	38.35	59.41	41.71	30.13	50.28	31.35
	DIST	34.89	56.32	37.68	39.24	60.82	42.77	31.98	52.33	34.02
	DKD	35.05	56.60	37.54	39.25	60.90	42.73	32.34	53.77	34.01
	NCKD (Ours)	**37.36**	**57.96**	**37.94**	**40.68**	**62.12**	**44.89**	**33.97**	**54.32**	**35.41**

Table 3: Comparison results on MS-COCO. We take Faster-RCNN (Ren et al. 2015) with FPN (Xie et al. 2017) as the backbones, and AP, AP_{50}, and AP_{75} as the evaluation metrics. The original accuracy results of the teacher and student models are also reported.

ImageNet-1k. To validate the effectiveness of our method on large-scale vision tasks, we conducted experiments on the ImageNet-1k dataset, using both similar-architecture (ResNet34/ResNet18) and cross-architecture (ResNet50/MobileNet) network pairs. As presented in Table 2, our method consistently outperforms the baselines, aligning with our findings on CIFAR-100. Remarkably, our approach even surpasses the advanced KD search method, DisWOT, by a substantial margin for the respective student-teacher pairs. These results highlight the effectiveness of our method in large-scale learning.

MS-COCO. We verify the efficacy of the proposed NC-inspired loss in knowledge distillation tasks for object detection on the COCO dataset, as shown in Table 3. All methods are evaluated under uniform training conditions to ensure comparability. Specifically, NCKD yields a significant improvement in performance, demonstrating their effectiveness and efficiency in knowledge distillation for dense prediction tasks.

Extensions

Visualization We employ t-SNE to evaluate the efficacy of our distillation method in enhancing the feature representation, as shown in Figure 4. KD, CRD, and DIST serve as our primary baselines. While the baseline models exhibit considerable class overlap, indicating poor feature separation, our method produces distinct clusters, demonstrating improved discriminative power. These results empirically validate the effectiveness of our approach and highlight its potential to enhance model generalization.

Ablation Study

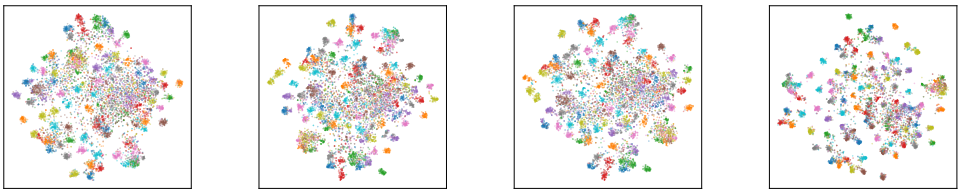
Distillation from Bigger Models. In principle, effective knowledge distillation should lead to GREAT TEACHERS PRODUCING OUTSTANDING STUDENTS, meaning that a superior teacher should guide the student to better distillation. However, in practice, such ideal case is not always achieved. We do evaluation using ResNet and Swin models of varying scales, as shown in Table 4. One can observe that existing methods do not consistently guarantee steady improvements in student performance as the teacher model's size increases. In contrast, our approach effectively addresses this issue, likely because better models establish a refined NC structure, which facilitates the student's consistent enhancement.

Teacher	Student	Teacher	Student	KD	DIST	DKD	**NCKD**
ResNet-34	ResNet-18	73.31	69.76	71.21	71.88	71.68	**72.44**
ResNet-50		76.13		71.35	72.04	71.91	**72.56**
ResNet-101		77.37		71.09	72.01	72.05	**72.71**
ResNet-152		78.31		71.12	72.06	72.03	**72.77**
Swin-T	ResNet-34	81.70	73.31	74.56	74.78	74.92	**74.95**
Swin-S		83.00		74.68	74.69	74.82	**75.01**
Swin-B		83.48		74.59	74.75	74.84	**75.05**

Table 4: Performance of ResNet-18/34 on ImageNet distilled from different large teachers.

Does $\mathcal{NC}$ impact KD? Yes! We evaluate the contribution of each $\mathcal{NC}$ property to the distillation process through ablation study, as shown in Table 5. The results show that removing any NC property would reduce the student prediction accuracy, with $\mathcal{NC}_{\mathbf{2}}$ having the most significant impact. This underscores the critical role of each module in our framework, especially the importance of preserving the teacher's ETF structure for effective knowledge transfer. Additionally, when combined with standard KD, our method further improves the distillation performance.

Does $\mathcal{NC}_{\mathbf{3}}$-classifier trade performance for efficiency? No! We conduct ablation study on the $\mathcal{NC}_{\mathbf{3}}$ classifier, with



(a) KD (b) CRD (c) DIST (d) NCKD

Figure 4: t-SNE of features learned by several KD methods. We use ResNet-32×4/ResNet-8×4 as the teacher/student pair.

Module	KD	Distillation			ResNet-8×4	ShuffleV1
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	$\mathcal{NC}_3$		
Baseline	-	-	-	-	72.51	70.50
KD	✓	-	-	-	74.12	74.00
CRD	✓	-	-	-	75.51	75.11
CRD+$\mathcal{NC}$	-	-	✓	✓	76.88	76.32
w/o $\mathcal{NC}_2$	-	✓	-	✓	75.98	76.48
w/o $\mathcal{NC}_3$	-	✓	✓	-	77.00	77.24
Ours	-	✓	✓	✓	77.23	77.48
Ours+KD	✓	✓	✓	✓	**77.41**	**77.55**

Table 5: Ablation study on the $\mathcal{NC}$-inspired distillation components on CIFAR-100. The baseline denotes the student's plain training. In other cases, the knowledge from pre-trained ResNet-32×4 is used for distillation.

Method	top-1	$\mathcal{NC}_1$	top-1	$\mathcal{NC}_3$	Method
KD	70.66	2.7e-2	70.66	1.47	KD
CRD	71.17	1.4e-2	72.01	1.11	SimKD
NCKD	**72.44**	8.1e-3	**72.44**	1.07	NCKD

Table 6: We use ResNet34/ResNet18 pair training on ImageNet to test the implicit $\mathcal{NC}$ properties of some existing approaches.

results presented in Figure 5. Notably, the $\mathcal{NC}_3$ classifier either outperforms or matches the standard classifier's results. Additionally, considering that we reduce the computational cost of training the classifier, this suggests that the design effectively balances performance and efficiency.

Case Study While we are the first to explicitly integrate NC into the KD framework, we recognize that some existing methods have implicitly leveraged $\mathcal{NC}$ to enhance distillation, albeit without explicit acknowledgment. Here, we investigate the role of $\mathcal{NC}$ in the effective distillation results of two representative methods, CRD and SimKD.

Case 1: CRD uses contrastive learning at the instance level to align teacher and student features, implicitly encouraging feature convergence toward class centroids (Khosla et al. 2020). This is reflected in the significant reduction of $\mathcal{NC}_1$ in CRD compared to KD (see Table 6), indicating its implicit use of $\mathcal{NC}_1$. Our approach, however, better preserves the $\mathcal{NC}_1$ property, resulting in improved performance.

Case 2: SimKD replaces the student's classifier with the teacher's, focusing solely on feature matching. We hypothesize that this implicitly leverages the teacher's $\mathcal{NC}_3$ property — where the reused classifier weights $\boldsymbol{w}$ preserve the teacher's normalized centroids. Our calculations, shown in Table 6, indicate that SimKD achieves $\mathcal{NC}_3$ values closer to 1 compared to standard KD. This suggests that SimKD gets benefit from this alignment, resulting in improved feature semantics and, consequently, better distillation outcomes.

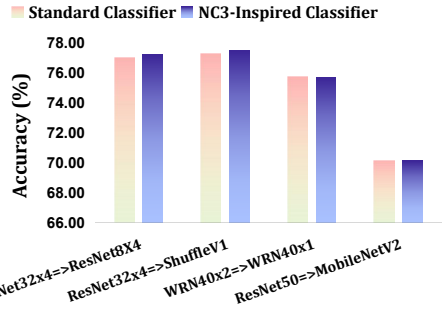


Figure 5: Distillation results with standard and our classifiers on CIFAR-100.

Conclusion

In this work, we introduced a novel approach to knowledge distillation by incorporating the structure of Neural Collapse (NC) into the distillation process. Our method, Neural Collapse-inspired Knowledge Distillation (NCKD), enables student models to learn the geometrically NC structure present in the teacher's final-layer representations. This strategy effectively bridges the knowledge gap between teacher and student models, resulting in superior student performance. Comprehensive experiments across diverse tasks and network architectures consistently demonstrated that our method outperforms state-of-the-art techniques, affirming its efficacy in enhancing performance. These findings highlight the robustness and adaptability of our NCKD, marking it a significant advancement in the field of KD.

While our study focused on distillation with a pre-trained teacher model, an unresolved area in the field is mutual distillation, where the student model also transfers knowledge back to the teacher during the distillation process. In future work, we will investigate whether NC can similarly benefit mutual distillation. Additionally, we aim to design NC-based criteria for selecting the most appropriate teacher model for a given student within the distillation framework.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. U22B2017).

References

Ammar, M. B.; Belkhir, N.; Popescu, S.; Manzanera, A.; and Franchi, G. 2023. NECO: NEural Collapse Based Out-of-distribution detection. *arXiv preprint arXiv:2310.06823*.

Chen, D.; Mei, J.-P.; Zhang, H.; Wang, C.; Feng, Y.; and Chen, C. 2022. Knowledge Distillation with the Reused Teacher Classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11933–11942.

Chen, D.; Mei, J.-P.; Zhang, Y.; Wang, C.; Wang, Z.; Feng, Y.; and Chen, C. 2021a. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7028–7036.

Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. 2019. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021b. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5008–5017.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Guo, Z.; Yan, H.; Li, H.; and Lin, X. 2023. Class attention transfer based knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11868–11877.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv:1503.02531*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35: 33716–33727.

Huang, Z.; and Wang, N. 2017. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv:1707.01219*.

Jin, Y.; Wang, J.; and Lin, D. 2023. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24276–24285.

Joyce, J. M. 2011. Kullback-leibler divergence. In *International Encyclopedia of Statistical Science*, 720–722. Springer.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33: 18661–18673.

Kim, J.; Park, S.; and Kwak, N. 2018. Paraphrasing complex network: Network compression via factor transfer. In *Advances in Neural Information Processing Systems*.

Kini, G. R.; Vakilian, V.; Behnia, T.; Gill, J.; and Thrampoulidis, C. 2023. Supervised-contrastive loss learns orthogonal frames and batching matters. *arXiv preprint arXiv:2306.07960*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, X.; Li, L.; Li, C.; and Yao, A. 2023. Norm: Knowledge distillation via n-to-one representation matching. *arXiv preprint arXiv:2305.13803*.

Lu, J.; and Steinerberger, S. 2022. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59: 224–241.

Papyan, V.; Han, X.; and Donoho, D. L. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663.

Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3967–3976.

Poudel, R. P.; Liwicki, S.; and Cipolla, R. 2019. Fast-scnn: Fast semantic segmentation network. *arXiv preprint arXiv:1902.04502*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv:1412.6550*.

Seo, M.; Koh, H.; Jeung, W.; Lee, M.; Kim, S.; Lee, H.; Cho, S.; Choi, S.; Kim, H.; and Choi, J. 2024. Learning Equiangular Representations for Online Continual Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23933–23942.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive representation distillation. *arXiv:1910.10699*.

Wang, T.; Yuan, L.; Zhang, X.; and Feng, J. 2019. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4933–4942.

Wang, Y.; Cheng, L.; Duan, M.; Wang, Y.; Feng, Z.; and Kong, S. 2023. Improving knowledge distillation via regularizing feature norm and direction. *arXiv preprint arXiv:2305.17007*.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks.

In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 1492–1500.

Yang, Y.; Yuan, H.; Li, X.; Lin, Z.; Torr, P.; and Tao, D. 2023. Neural collapse inspired feature-classifier alignment for few-shot class incremental learning. *arXiv preprint arXiv:2302.03004*.

Yun, S.; Park, J.; Lee, K.; and Shin, J. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 13876–13885.

Zagoruyko, S.; and Komodakis, N. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv:1612.03928*.

Zhang, S.; Liu, H.; and He, K. 2024. Knowledge Distillation via Token-Level Relationship Graph Based on the Big Data Technologies. *Big Data Research*, 36: 100438.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11953–11962.

Zheng, K.; and Yang, E.-H. 2024. Knowledge distillation based on transformed teacher matching. *arXiv preprint arXiv:2402.11148*.

Zhou, J.; Li, X.; Ding, T.; You, C.; Qu, Q.; and Zhu, Z. 2022. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In *International Conference on Machine Learning*, 27179–27202. PMLR.