LINEAR CKA CONFLATES CLASS-MEAN AND WITHIN-CLASS AGREEMENT: AN EXACT COVARIANCE DECOMPOSITION ON LABELED BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear centered kernel alignment (CKA) is a widely used measure of similarity between neural network representations, and it is commonly evaluated on labeled benchmarks. Its first step, global centering, removes the dataset-wide mean but not the offset of each class mean from it. A high CKA score can come from two networks placing the same classes in the same regions of feature space, even when they arrange the images inside each class in unrelated ways. Splitting each representation into orthogonal between-class and within-class components decomposes the cross-covariance matrix in the CKA numerator, and the two self-covariance matrices in its denominator, exactly into a between-class and a within-class term, whose squared Frobenius norms combine with a signed interference term. A constructive example makes ordinary CKA arbitrarily close to 1 while within-class CKA remains exactly 0, so a high score alone does not establish within-class agreement. Across four backbones and four labeled datasets, the final-depth gap between ordinary and within-class CKA is positive in all 24 (model pair, dataset) combinations, ranging from 0.022 to 0.273, with a positive interference term throughout. After projecting both representations to a common 64 dimensions, the gap exceeds every value of a 100-permutation label-shuffling null, at the floor empirical $p = 1/101$ and Benjamini-Hochberg adjusted $q = 0.0099$, while a random balanced partition of matching granularity reduces it to near zero. The overlap between the class subspace and each backbone's dominant principal-component subspace exceeds a random-subspace null by factors of six to five hundred. On ImageNet-1K, replacing ordinary CKA with within-class CKA moves the global maximum of the cross-depth heatmap for four of six model pairs, so class-mean structure can change the layer correspondence such a heatmap reports.

1 INTRODUCTION

Representation similarity analysis asks whether separately trained neural networks learn similar internal representations, and linear centered kernel alignment (CKA) (Kornblith et al., 2019) is a widely used measure for comparing two networks, or two layers within a network. A CKA value close to 1 is commonly interpreted as evidence that two representations encode overlapping geometric structure. This reading supports claims that independently trained architectures converge to similar solutions (Li et al., 2015; Mehrer et al., 2020) and that a given layer of one network corresponds to a given layer of another (Nguyen et al., 2021; Raghu et al., 2021).

These comparisons are made on labeled evaluation sets. ImageNet (Deng et al., 2009) validation images, CIFAR (Krizhevsky, 2009) test images, and other standard classification benchmarks are commonly used to compute a CKA heatmap, and the labels attached to these images play no part in the similarity computation. Ordinary linear CKA begins with global centering, which removes the component along the all-ones sample direction from each representation. It does not remove the offset of each class mean from the dataset mean. On a labeled dataset, a large share of the total feature variance can come from between-class differences, and two representations that place the same classes in the same coarse regions of feature space can then score high CKA even when the two models arrange the images within a class in unrelated ways. Ordinary CKA therefore mixes class-level agreement with residual within-class agreement. The subspace that survives global centering

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

has one dimension fewer than the number of classes and so grows with the granularity of the label set, so the ambiguity follows from the evaluation protocol rather than from any particular pair of models.

We ask how much ordinary linear CKA changes when shared class-mean structure is removed from a labeled dataset, and how much residual similarity remains. Using a centered class-indicator projector P, we split each centered feature matrix into a between-class component and a within-class residual component. The cross-covariance matrix driving CKA's numerator, together with the two self-covariance matrices that enter its denominator, then splits exactly into a between-class and a within-class term, and their squared Frobenius norms combine with a signed interference term. The decomposition defines a within-class CKA that removes class-mean structure from each side before the ratio is computed.

When ordinary CKA is interpreted as reflecting fine-grained representational agreement, class-mean structure introduces an ambiguity. We show constructively that this interpretation does not follow in general, since ordinary CKA can be driven arbitrarily close to 1 by a shared class-mean component while the corresponding within-class CKA is exactly zero. We then assess to what extent this difference depends on the class labels, using label permutations, random balanced partitions, a class-level bootstrap, and a comparison between the class subspace and the dominant-variance subspace. We apply these analyses to four backbones (ResNet-18, ConvNeXt-Tiny, EfficientNet-B0, and a self-supervised vision transformer, ViT-S, trained with DINO) on four labeled datasets (CIFAR-10, CIFAR-100, Food-101, and the ImageNet-1K validation split).

Our contributions are:

- An exact class-conditional decomposition of the cross- and self-covariance matrices underlying linear CKA, whose between-class term is a cross product of the two class-mean configurations and has no single-representation analogue, with a signed interference term in the squared Frobenius norms. The resulting within-class CKA inherits CKA's invariance to orthogonal transformations and isotropic scaling, and we prove constructively that a high ordinary CKA score does not in general imply a high within-class score, together with sufficient conditions under which the two coincide.
- A statistical protocol, using 100 label permutations with Benjamini-Hochberg correction, random balanced partitions at three granularities, and a 500-resample class-level bootstrap, showing that the change in CKA induced by class-mean removal is reproduced neither by label permutations nor by random partitions of matching granularity.
- A quantitative account of how the class subspace relates to each backbone's dominant-variance directions, using principal-angle overlap between the class projector and a top-k principal component analysis (PCA) subspace with $k = \min(c-1, d-1)$, against a rank-k random-subspace null.
- An analysis showing that, on ImageNet-1K, removing class-mean structure moves the globally best-matching depth pair for four of six model pairs.

2 Related Work

Measures of representation similarity. Representations of two networks are commonly compared with canonical correlation analysis (CCA) (Hotelling, 1936) and its neural variants, singular vector CCA (SVCCA) (Raghu et al., 2017) and projection-weighted CCA (Morcos et al., 2018), with representational similarity analysis (RSA) (Kriegeskorte et al., 2008), with Procrustes-type shape metrics (Schönemann, 1966; Williams et al., 2021), and with linear CKA (Kornblith et al., 2019). CKA descends from kernel-target alignment (Cristianini et al., 2001) and centered alignment (Cortes et al., 2012), both built on the Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005), and the unbiased HSIC estimator (Song et al., 2012) yields the debiased CKA. Recent work establishes close connections among RSA, CKA, and CCA under appropriate centering (Williams, 2024) and surveys the wider family (Klabunde et al., 2025), in each case without conditioning on any label structure of the evaluation set.

Reading layer correspondence from CKA. CKA heatmaps are widely used to identify which layer of one network matches which layer of another, for example across width and depth (Nguyen et al.,

2021) and between vision transformers and convolutional networks (Raghu et al., 2021), and our depth-matching analysis tests this interpretation directly.

Reliability and confounds. Several studies examine when CKA can mislead. Similarity measures disagree with one another and can be insensitive to changes that matter functionally (Ding et al., 2021), CKA can move under simple transformations and outliers while behavior is nearly unchanged (Davari et al., 2023), and the block structure in CKA heatmaps has been traced to a few dominant principal components and datapoints (Nguyen et al., 2022). Most closely related, Cui et al. (2022) show that population structure in the input data confounds CKA and RSA, and remove it by covariate adjustment. Functional alternatives such as model stitching (Bansal et al., 2021) and intervention-based tests (Hayne et al., 2024) question whether geometric similarity captures the information a network uses.

Class-mean geometry in learned representations. The split of feature variation into between-class and within-class parts goes back to linear discriminant analysis (Fisher, 1936). The neural collapse literature shows that this split becomes extreme in trained classifiers, with within-class variability shrinking and class means approaching a simplex equiangular tight frame (Papyan et al., 2020), a configuration that arises under several training objectives and architectures (Zhu et al., 2021; Han et al., 2022; Dang et al., 2023), with class frequencies shaping the resulting geometry (Fang et al., 2021). This literature establishes that class-mean structure can occupy a large share of a representation, but it does not ask what that structure does to a similarity score computed between two representations. A scatter decomposition splits the variation of one representation; the object decomposed here is the cross-covariance of two, whose between-class term $B = X_b^\top Z_b$ measures the alignment of two class-mean configurations and has no single-representation counterpart.

Relation to prior work. Kornblith et al. (2019) report a within-class CKA in their Appendix F.7, computed separately inside each CIFAR-10 class and averaged over classes. We instead remove the class means from both representations jointly and then compute a single CKA, which keeps all samples in one comparison and admits the exact decomposition of Proposition 1, including the signed interference term that a per-class average does not expose. A per-class score is also estimated from the samples of a single class, of which the ImageNet-1K validation split provides 50, while the pooled residual retains all n samples in one estimate. Relative to the deconfounding of Cui et al. (2022), the structure we remove is defined by the evaluation labels rather than by input-space population structure. To our knowledge, no prior work provides an exact label-conditional decomposition of the matrices entering the CKA numerator and denominator. The controls of Section 4 then measure how far a CKA value on a labeled vision benchmark responds to the removal of shared class-mean geometry.

3 Background and Problem Setting

3.1 Linear CKA

Let n be the number of paired samples and let $X \in \mathbb{R}^{n \times d_x}$ and $Z \in \mathbb{R}^{n \times d_z}$ be feature matrices from two representations, with row i of X and row i of Z corresponding to the same input. We assume X and Z are globally centered, so $\mathbf{1}^\top X = 0$ and $\mathbf{1}^\top Z = 0$, where $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector and $J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix. Linear CKA is

$$\mathrm{CKA}(X, Z) = \frac{\|X^\top Z\|_F^2}{\|X^\top X\|_F \, \|Z^\top Z\|_F}, \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm (Kornblith et al., 2019). $\mathrm{CKA}(X, Z) \in [0, 1]$, with 1 attained when the sample-space Gram matrices $XX^\top$ and $ZZ^\top$ are proportional. Our decomposition acts on the cross-covariance matrix $X^\top Z$ and the self-covariance matrices $X^\top X$ and $Z^\top Z$ that enter this ratio.

3.2 The Centered Class Projector

Definition 1 (Centered class projector)**.** *Let c be the number of classes and let $H \in \{0, 1\}^{n \times c}$ be the class-indicator matrix, $H_{i,k} = 1$ if sample i belongs to class k. Let $Y = JH$. Define*

$$P = Y(Y^\top Y)^+ Y^\top \in \mathbb{R}^{n \times n}, \tag{2}$$

where $(\cdot)^+$ is the Moore-Penrose pseudoinverse, needed because Y loses one rank to centering, $\mathrm{rank}(Y) \leq c - 1$.

P is symmetric, idempotent ($P^2 = P$), and satisfies $P\mathbf{1} = 0$, so $\mathrm{range}(P)$ lies in $\mathbf{1}^\perp$ and global centering and class projection commute. Since $\mathbf{1} \in \mathrm{range}(H)$ and $\mathrm{range}(P) = \mathrm{range}(H) \cap \mathbf{1}^\perp$, the span of the raw class indicators decomposes as $\mathrm{span}\{\mathbf{1}\} \oplus \mathrm{range}(P)$, of dimension c. Global centering removes one of these c directions and leaves the other $c - 1$ in the representation that ordinary CKA compares. For any globally centered X, define the between-class and within-class components, in the sense of linear discriminant analysis (Fisher, 1936),

$$X_b = PX, \qquad X_w = (I - P)X, \tag{3}$$

and analogously Z_b, Z_w. X_b contains only class-constant structure (every row of X_b for samples in the same class is identical), while X_w holds the deviation of each sample from its class mean, including the full within-class covariance.

Proposition 1 (Exact class-conditional decomposition)**.** *For globally centered X, Z and the projector P above,*

$$X^\top Z = X_b^\top Z_b + X_w^\top Z_w = B + W, \tag{4}$$

with $B := X_b^\top Z_b$ and $W := X_w^\top Z_w$, and

$$\|X^\top Z\|_F^2 = \|B\|_F^2 + \|W\|_F^2 + 2\langle B, W\rangle, \tag{5}$$

where $\langle B, W\rangle = \mathrm{tr}(B^\top W)$ is a signed cross term, which we call the interference term. The same expansion applies to $\|X^\top X\|_F^2$ and $\|Z^\top Z\|_F^2$.

A full proof is given in Appendix A.1; it follows from $P(I - P) = 0$, which makes the two cross terms $X_b^\top Z_w$ and $X_w^\top Z_b$ vanish exactly. The covariance matrices therefore decompose exactly, while the scalar CKA ratio, whose numerator and two denominator factors each carry an interference term of their own, is not additive.

Definition 2 (Between-class and within-class CKA)**.**

$$\mathrm{CKA}_b(X, Z) = \frac{\|B\|_F^2}{\|X_b^\top X_b\|_F \, \|Z_b^\top Z_b\|_F}, \tag{6}$$

$$\mathrm{CKA}_w(X, Z) = \frac{\|W\|_F^2}{\|X_w^\top X_w\|_F \, \|Z_w^\top Z_w\|_F}. \tag{7}$$

We define the paired difference

$$\Delta = \mathrm{CKA}(X, Z) - \mathrm{CKA}_w(X, Z), \tag{8}$$

which measures how much the CKA score changes once class-mean structure is removed from both representations. Since $\mathrm{CKA}_w(X, Z) = \mathrm{CKA}(X_w, Z_w)$ and $I - P$ acts on samples while a change of feature basis acts on columns, within-class CKA inherits CKA's invariance to orthogonal transformations and isotropic scaling.

CKA_b and CKA are normalized by different denominators, the between-class self-covariance norms in one case and the full self-covariance norms in the other, so their values are not directly comparable. We therefore report the normalized interference term $2\langle B, W\rangle / \|B + W\|_F^2$ alongside them. A positive value means that the between-class and within-class cross-covariance terms have a positive Frobenius inner product, which raises the squared norm of their sum above $\|B\|_F^2 + \|W\|_F^2$; a negative value has the opposite effect. Since X_b and X_w are orthogonal, their squared Frobenius norms add, and we also report the feature-variance fraction $\rho(X) = \|X_b\|_F^2 / \|X\|_F^2$, which separates the amount of class-mean variance from the change in similarity that residualization produces.

Proposition 2 (Existence of an arbitrarily large gap)**.** *There exist centered X, Z with a shared nonzero class component in* $\mathrm{range}(P)$*, and residuals in* $\mathrm{range}(I - P)$ *with nonzero self-covariance, such that $X_w^\top Z_w = 0$ identically. Scaling the shared class component by a drives* $\mathrm{CKA}(X, Z) \to 1$ *as $a \to \infty$, while* $\mathrm{CKA}_w(X, Z) \equiv 0$ *for every a.*

This is proved constructively in Appendix A.1 and instantiated numerically in Appendix A.2.

Proposition 3 (Sufficient conditions for equality). $\mathrm{CKA}(X, Z) = \mathrm{CKA}_w(X, Z)$ *holds whenever either (a) both class components vanish,* $X_b = 0$ *and* $Z_b = 0$*, or (b) there exists a single scalar* $s > 0$*, applied consistently to the numerator and to both denominator factors, such that*

$$\|B + W\|_F^2 = s^2 \|W\|_F^2,$$
$$\|X_b^\top X_b + X_w^\top X_w\|_F = s\|X_w^\top X_w\|_F, \qquad \|Z_b^\top Z_b + Z_w^\top Z_w\|_F = s\|Z_w^\top Z_w\|_F. \tag{9}$$

Condition (b) constrains the numerator interference term and both denominator interference terms by the same scalar s, and a condition on the numerator alone is not sufficient. Read in the other direction, $\Delta = 0$ *requires these three terms to be in exact proportion, so a small* Δ *does not by itself place the class components at zero.*

Condition (a) covers the single-class case: with one class, $Y = J\mathbf{1} = 0$ and P vanishes, so nothing is removed and within-class CKA reduces to ordinary CKA. Ordinary CKA is the special case of this family with no label-conditioned residualization. Propositions 2 and 3 bracket the two ends of the range: Δ can be pushed arbitrarily close to its maximum, and it is exactly zero under the stated conditions. Neither end is decided by the definition of CKA, so where a given pair of representations falls between them is an empirical question, which the rest of the paper measures.

4 Method

4.1 Models and Datasets

We evaluate four ImageNet-pretrained backbones that differ in architecture and training objective. These are ResNet-18 (He et al., 2016) (RN18; four residual stages, feature widths $64, 128, 256, 512$), ConvNeXt-Tiny (Liu et al., 2022) (CNXT; four stages, widths $96, 192, 384, 768$), EfficientNet-B0 (Tan & Le, 2019) (EffB0; four feature depths, widths $24, 80, 192, 1280$), and the self-supervised ViT-S (Dosovitskiy et al., 2021) trained with DINO (Caron et al., 2021) (twelve transformer blocks, hidden width 384). For the three convolutional backbones we take the output of each of the four architectural stages and apply global average pooling over spatial locations to obtain one feature vector per image per depth. For the DINO ViT-S backbone we average the patch tokens, excluding the classification (CLS) token, at four evenly spaced blocks, the third, sixth, ninth, and twelfth, giving a four-depth representation; CLS-token pooling is compared in Appendix A.10, where Δ stays positive under both choices. Pooling to one vector per image lets a single sample-space projector P, built from the labels alone, act on all four backbones. We index these depths 1 to 4 for every backbone and also report a normalized relative depth in $\{0.25, 0.5, 0.75, 1.0\}$.

We use four labeled evaluation sets, the CIFAR-10 and CIFAR-100 test splits (Krizhevsky, 2009) (10 and 100 classes), the Food-101 test split (Bossard et al., 2014) (101 classes), and the complete ImageNet-1K validation split (Deng et al., 2009) (1000 classes, $50{,}000$ images). CIFAR-10, CIFAR-100, and Food-101 carry label sets different from the one the supervised backbones were pretrained on, while the ImageNet-1K validation split carries that same label set.

All images pass through the same deterministic preprocessing, namely a shorter-side resize, a center crop to 224×224, and standard ImageNet channel normalization, with no data augmentation at any stage. Further implementation details are given in Appendix A.11.

4.2 Decomposition and Statistical Controls

For each pair of backbones, dataset, and pair of depths, we form X_b, X_w, Z_b, Z_w from the per-class means and compute CKA, CKA_b, CKA_w, and Δ, with ρ for each representation.

To determine how far the measured Δ depends on the class labels, and how the class subspace relates to the dominant directions of variance, we run three null comparisons and one uncertainty analysis. Each is applied at the final depth pair and at the depth pair that maximizes ordinary CKA, after a Gaussian random projection of the centered features to 64 dimensions, so that every control and its observed value are evaluated at the same dimension.

- **Label permutation.** We draw 100 permutations of the label vector that preserve the per-class sample count, recompute the class means from each permuted labeling, and recompute Δ. Holding the class sizes fixed holds $\mathrm{rank}(P)$ and the number of samples entering

each class mean fixed as well, so a permuted labeling removes a subspace of the same dimension as the true labeling, and the two differ only in which samples are grouped together. The two-sided empirical p-value is $p = \frac{1+\#\{\text{permutations with } |\Delta_{\text{perm}}| \geq |\Delta_{\text{obs}}|\}}{101}$, with floor $1/101$.

- **Balanced partition.** We assign samples to 10, 100, or 1000 near-equally sized groups uniformly at random, independent of any label, recompute the projector for this partition, and recompute Δ. Varying the number of groups varies the rank of the removed subspace over two orders of magnitude, which separates the effect of removing a subspace of a given size from the effect of removing the one the labels define.
- **Two-stage class bootstrap** (Efron, 1979)**.** We build 500 resamples, each drawing c classes with replacement and then drawing images with replacement within every drawn class while keeping the original class sizes, and recompute Δ and ρ on each resample to obtain 95% percentile intervals. Because each resample redraws the classes themselves, the resulting interval describes variation over evaluation sets of the same design and need not be centred on the value computed from the original sample. Resampling at the class level matches the level at which the projector is defined, since classes and not individual images determine P, and the intervals therefore describe variation in the class composition of the evaluation set as well as in its image composition.
- **Subspace comparison.** The class projector is defined by the labels and the principal-component projector by the variance spectrum, so comparing the two subspaces asks whether the label-defined subspace aligns with the directions that account for most feature variance. We compute the top-k principal-component projector Q_k of the sample-space Gram matrix, with $k = \min(c-1, d-1)$ for feature dimension d, and measure the overlap between $\mathrm{range}(P)$ and $\mathrm{range}(Q_k)$ as $\frac{1}{k}\|U_P^\top U_{Q_k}\|_F^2 = \frac{1}{k}\sum_{i=1}^{k} \cos^2\theta_i$, where U_P and U_{Q_k} hold orthonormal bases of the two subspaces and θ_i are their principal angles. The two subspaces are rank-matched, $k = \mathrm{rank}(P) = c-1$, whenever $d \geq c$, which holds for every backbone on the three datasets with fewer than 1000 classes and, on ImageNet-1K, for EfficientNet-B0 alone; otherwise the feature dimension limits the rank to $k = d-1$. We compare the overlap against a null distribution built from 100 independently drawn random rank-k orthonormal (Haar) subspaces of the sample space, whose expected overlap with any fixed subspace is $\mathrm{rank}(P)/n$.

All p-values from the permutation test, across every model pair, dataset, and both depth pairs, are corrected jointly using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). At the same two depth pairs we compute a bias-corrected estimator of CKA (debiased CKA) (Song et al., 2012), and at the final depth we compute SVCCA (Raghu et al., 2017) with a per-representation rank chosen to retain 99% of variance. These correct the finite-sample bias of the CKA estimator and replace the CKA ratio by a canonical-correlation statistic.

4.3 Depth-Matching Analysis

For every model pair and dataset we identify the depth pair that maximizes ordinary CKA and the depth pair that maximizes within-class CKA over the full 4×4 depth grid, and record whether these two best-matching depth pairs coincide, which tests whether residualization changes the strongest cross-model depth correspondence.

5 Results

5.1 Class-Mean Structure at the Final Depth

Scaling the shared class component of the construction in Proposition 2 drives ordinary CKA from 0 to 0.969 while within-class CKA stays exactly at 0 (Appendix A.2). On real features, ordinary CKA exceeds within-class CKA in all 24 combinations. All values here are computed in the full feature space, the projection entering only the controls. Table 1 gives the full decomposition for eight combinations, and Appendix Tables 5 and 4 give further diagnostics and Δ for all 24. Across all six model pairs on CIFAR-10, the difference Δ ranges from 0.176 to 0.273, and across all six pairs on ImageNet-1K from 0.022 to 0.211. Although the two are normalized differently, between-class CKA exceeds ordinary CKA in all 24 combinations, reaching 0.961 against 0.679 for ResNet-18 versus

ConvNeXt-Tiny on CIFAR-10, so the two class-mean configurations agree closely even where the full representations agree much less. The normalized interference term $2\langle B, W\rangle/\|B + W\|_F^2$ is positive in all 24 combinations and equals $+0.24$ for this pair, so the two terms reinforce each other and raise the ordinary numerator above $\|B\|_F^2 + \|W\|_F^2$, while the within-class numerator is $\|W\|_F^2$ alone. The between-class variance fraction ρ stays between 0.152 and 0.587 across settings while Δ reaches 0.273, so a moderate share of the total variance can coincide with a large change in the score once class-mean structure is removed. Debiased CKA stays close to ordinary CKA, and SVCCA on the residualized features is lower than on the raw features in every combination.

Table 1: Final-depth decomposition for eight of the 24 combinations. CKA is ordinary linear CKA, CKA_b and CKA_w are the between-class and within-class components, $\Delta = \mathrm{CKA} - \mathrm{CKA}_w$, and ρ_X, ρ_Z are the between-class variance fractions of each side.

Dataset	Pair	CKA	CKA_b	CKA_w	Δ	ρ_X	ρ_Z
CIFAR-10	RN18 vs CNXT	0.679	0.961	0.406	+0.273	0.177	0.216
CIFAR-10	CNXT vs EffB0	0.689	0.963	0.461	+0.228	0.216	0.152
CIFAR-100	RN18 vs CNXT	0.587	0.881	0.399	+0.188	0.247	0.266
CIFAR-100	EffB0 vs DINO	0.506	0.843	0.311	+0.195	0.219	0.296
Food-101	RN18 vs CNXT	0.527	0.840	0.333	+0.194	0.193	0.319
ImageNet-1K	RN18 vs CNXT	0.162	0.297	0.088	+0.074	0.383	0.587
ImageNet-1K	RN18 vs DINO	0.417	0.715	0.206	+0.211	0.383	0.334
ImageNet-1K	CNXT vs EffB0	0.366	0.451	0.344	+0.022	0.587	0.283

5.2 The Difference Grows with Depth

Figure 1 shows the full 4×4 depth grid of ordinary CKA, within-class CKA, and their difference for the ResNet-18/ConvNeXt-Tiny pair on all four datasets. On CIFAR-10 the largest ordinary CKA in the grid, 0.78, lies off the matched diagonal, at source depth 2 and target depth 1. Within-class CKA is lower at all 16 depth pairs on CIFAR-10, CIFAR-100, and Food-101, and at 14 of the 16 on ImageNet-1K, dropping sharply off the diagonal. The two exceptions, at shallow source and deeper target depths, exceed ordinary CKA by at most 0.016. The difference Δ (bottom row) is small at shallow depths and reaches its grid maximum at the deepest matched depth on every dataset, between 0.188 and 0.273 on CIFAR-10, CIFAR-100, and Food-101, where the matched-depth diagonal peaks at the final depth for all eighteen model pairs (Appendix A.9). On ImageNet-1K that maximum is 0.074. As a cross product of the two between-class components, $B = X_b^\top Z_b$ reflects the alignment of the two class-mean configurations, while the normalization terms depend on each representation separately.

5.3 The Difference Is Not Reproduced by Random Labelings

Figure 2 shows the permutation null for the ResNet-18/ConvNeXt-Tiny pair at the final depth on all four datasets. Every permuted value stays below 0.001 in absolute value while the observed differences range from 0.044 to 0.190, so the Δ computed with the class labels exceeds every permuted value by a factor of at least seventy, at the floor empirical value $p = 1/101$. The same value is obtained at all 48 tested (model pair, dataset, depth pair) combinations, where the Benjamini-Hochberg adjusted value is $q = 0.0099$. Since a class-size-preserving permutation removes a subspace of the same rank as the true labeling, this comparison isolates which samples share a label from how many dimensions the projector takes out, and only the former moves the score. In particular, residualizing a rank-$(c-1)$ subspace does not by itself raise ordinary CKA above within-class CKA in finite samples.

Table 2 reports the balanced-partition check alongside the permutation result. At 10 and at 1000 groups, the recomputed Δ for the ResNet-18/ConvNeXt-Tiny final depth pair is at most 0.0006 in absolute value, smaller than the value obtained with the class labels by factors between 220 and 950. Across the other final depth pairs the partition value stays at least an order of magnitude below the corresponding difference, and the bootstrap intervals exclude zero on all four datasets.

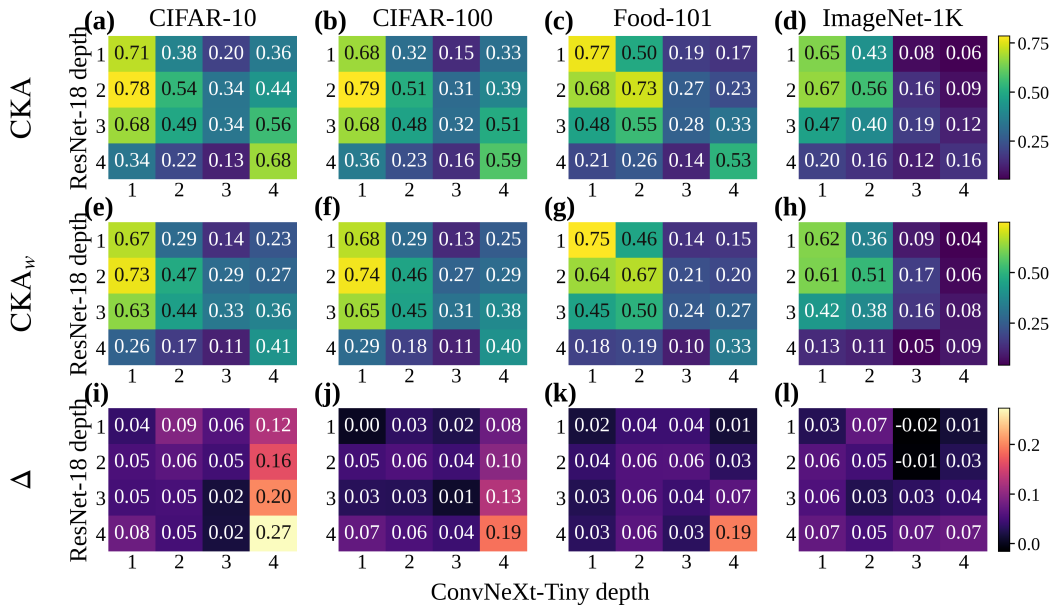

Figure 1: Cross-depth grids of ordinary CKA (top row), within-class CKA_w (middle row), and $\Delta = CKA - CKA_w$ (bottom row) for the ResNet-18/ConvNeXt-Tiny pair. (a),(e),(i) CIFAR-10; (b),(f),(j) CIFAR-100; (c),(g),(k) Food-101; (d),(h),(l) ImageNet-1K. The color scale is shared across datasets within each row.

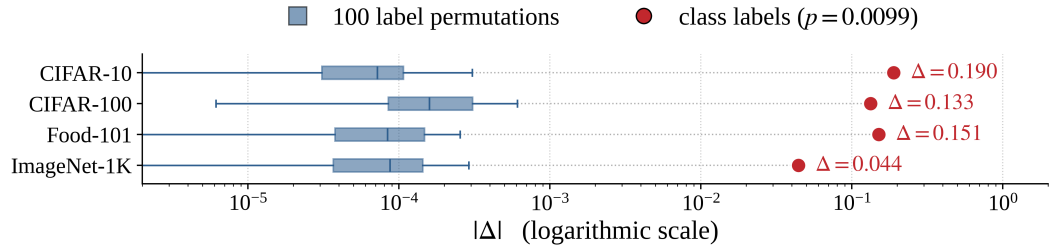

Figure 2: Permutation null of Δ under 100 class-size-preserving label permutations (final depth, ResNet-18/ConvNeXt-Tiny, 64-dimensional Gaussian projection). Boxes give the median, quartiles, and full range of $|\Delta|$; circles give $|\Delta|$ from the class labels. The horizontal axis is logarithmic.

A second question is whether the difference follows the class labels or the dominant directions of variance, to which CKA is known to be sensitive (Nguyen et al., 2022) and with which the class subspace can overlap. The principal-angle overlap between the class projector and the top-k PCA projector, with $k = \min(c-1, d-1)$, lies between 0.098 and 0.475 on every dataset and backbone, while a random rank-k subspace null has a mean between 0.0009 and 0.020. This overlap links sensitivity to dominant directions with sensitivity to class structure, and it is partial, so removing a rank-matched PCA subspace is not interchangeable with removing P. Per-backbone values are given in Appendix A.3.

5.4 Effect on Depth Matching

For every model pair on ImageNet-1K we compare the depth pairs maximizing ordinary and within-class CKA (Appendix A.7). For four of the six pairs the best-matching depth pair changes, with ResNet-18/ConvNeXt-Tiny shifting from $(2, 1)$ to $(1, 1)$ and ResNet-18/DINO from $(3, 2)$ to $(1, 1)$. Ordinary CKA selects five different depth pairs across the six model pairs, while within-class CKA selects depth $(1, 1)$ for all six, so removing class-mean structure turns a scattered set of layer cor-

Table 2: Statistical controls for the final depth pair of ResNet-18/ConvNeXt-Tiny, following the definitions in Section 4. Proj. Δ is the observed value after the 64-dimensional projection, Perm. mean is the mean over 100 label permutations, and the partition columns give one random balanced partition at 10 and at 1000 groups. Perm. mean and the two partition columns are in units of 10^{-4}, and < 0.5 denotes a magnitude below the displayed resolution. The last column is the 95% percentile interval over the 500 bootstrap resamples.

Dataset	Proj. Δ	Perm. mean	p	Part.$_{10}$	Part.$_{1000}$	Boot. 95% CI
CIFAR-10	$+0.190$	$+0.06$	0.0099	-2.0	$+1.0$	$[+0.127, +0.289]$
CIFAR-100	$+0.133$	-0.13	0.0099	< 0.5	$+6.0$	$[+0.121, +0.166]$
Food-101	$+0.151$	-0.02	0.0099	-1.0	$+2.0$	$[+0.127, +0.199]$
ImageNet-1K	$+0.044$	-0.16	0.0099	< 0.5	$+1.0$	$[+0.046, +0.058]$

respondences into a single one. On CIFAR-10, CIFAR-100, and Food-101, whose label sets differ from the pretraining one, the best-matching depth pair is unchanged for every pair shown in Appendix Table 6. The effect on depth matching is therefore most pronounced on ImageNet-1K, where the between-class variance fraction also reaches its largest values (Appendix A.4). Large class-mean structure inside one representation is expected there (Papyan et al., 2020); the grid adds that it changes which depth pair a comparison between two representations selects.

6 Discussion

Ordinary linear CKA computed on a labeled evaluation set responds to the removal of class-mean structure: the difference is positive at the final depth in every combination we measure, and on ImageNet-1K residualization moves the globally best-matching depth pair for four of six model pairs. Neither the 100 label permutations nor the random balanced partitions of matching granularity reproduce it, and the principal-angle analysis places the class subspace inside the dominant variance directions that ordinary CKA is known to be sensitive to (Nguyen et al., 2022).

For CKA analyses on labeled benchmarks, within-class CKA reported alongside ordinary CKA distinguishes class-level from residual within-class agreement. A claim about where the classes sit rests on the between-class part and a claim about how samples are arranged inside a class on the within-class part, and ordinary CKA alone does not say which of the two a given value reflects. Proposition 3 shows that changes in the numerator and in the two normalization terms can offset one another, so Δ is most informative when read together with the between-class variance fraction and the interference term, both of which the decomposition provides.

The size of this effect varies across architecture, depth, and the label set of the evaluation dataset. The between-class variance fraction trajectories do not separate the three supervised backbones from the DINO backbone (Appendices A.4 and A.8), and on ImageNet-1K we observe the largest differences at deeper and matched depths for supervised-backbone pairs, with the peak depth varying by pair (Appendix A.9).

Limitations. Proposition 1 holds for any features, while the magnitude of the effect is measured on four ImageNet-pretrained backbones and four datasets. On ImageNet-1K, exact rank matching between the class and PCA subspaces is available only for EfficientNet-B0.

7 Conclusion

Ordinary linear CKA on a labeled dataset mixes agreement on the placement of classes with agreement on the arrangement of images inside each class. An exact decomposition of the cross- and self-covariance matrices behind CKA separates the two, and a construction makes ordinary CKA approach 1 while within-class CKA remains exactly zero. Across 24 combinations of four backbones and four datasets the gap is positive at the final depth, falls to near zero under label permutation and under random partitions, and on ImageNet-1K moves the best-matching depth pair for four of six pairs. Built from the labels alone, it applies to any CKA comparison on a labeled set.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX formatting, experiment code, and figure plotting. They were not used to produce experimental results, numerical values, or scientific claims. The authors take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

Proposition 1 is proved in full in Appendix A.1. The models, datasets, preprocessing, pooling conventions, and all hyperparameters of the statistical controls are given in Section 4 and Appendix A.11.

REFERENCES

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236, 2021.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101: Mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision*, 2014.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13(28):795–828, 2012.

Nello Cristianini, John Shawe-Taylor, André Elisseeff, and Jaz S. Kandola. On kernel-target alignment. In *Advances in Neural Information Processing Systems*, volume 14, pp. 367–373, 2001.

Tianyu Cui, Yogesh Kumar, Pekka Marttinen, and Samuel Kaski. Deconfounded representation similarity for comparison of neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19138–19151, 2022.

Hien Dang, Tho Tran Huu, Stanley Osher, Hung The Tran, Nhat Ho, and Tan Minh Nguyen. Neural collapse in deep linear networks: From balanced to imbalanced data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6873–6947. PMLR, 2023.

MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. Reliability of CKA as a similarity measure in deep learning. In *International Conference on Learning Representations*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1556–1568, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2): 179–188, 1936.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*, volume 3734 of *Lecture Notes in Computer Science*, pp. 63–77. Springer, 2005.

X. Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022.

Lucas Hayne, Heejung Jung, and R. McKell Carter. Does representation similarity capture function similarity? *Transactions on Machine Learning Research*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3–4):321–377, 1936.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, volume 44 of *Proceedings of Machine Learning Research*, pp. 196–212. PMLR, 2015.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and Tim C. Kietzmann. Individual differences among deep neural network models. *Nature Communications*, 11(1):5725, 2020.

Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5727–5736, 2018.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. In *International Conference on Learning Representations*, 2021.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. On the origins of the block structure phenomenon in neural network representations. *Transactions on Machine Learning Research*, 2022.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6076–6085, 2017.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128, 2021.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(47):1393–1434, 2012.

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*, 2019.

Ross Wightman. PyTorch image models. `https://github.com/huggingface/pytorch-image-models`, 2019.

Alex H. Williams. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. In *Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models*, volume 285 of *Proceedings of Machine Learning Research*, pp. 10–23. PMLR, 2024.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750, 2021.

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 *Advances in Neural Information Processing Systems*, volume 34, pp. 29820–29834, 2021.

A Appendix

A.1 Proofs

Proof of Proposition 1. Write $X = X_b + X_w$ and $Z = Z_b + Z_w$ with $X_b = PX$, $X_w = (I-P)X$, and analogously for Z. Then

$$X^\top Z = (X_b + X_w)^\top (Z_b + Z_w) = X_b^\top Z_b + X_b^\top Z_w + X_w^\top Z_b + X_w^\top Z_w. \tag{10}$$

Since P is symmetric and idempotent, $P(I-P) = P - P^2 = 0$. Hence

$$X_b^\top Z_w = X^\top P^\top (I-P)Z = X^\top P(I-P)Z = 0, \tag{11}$$

and by the same argument with the roles of X and Z exchanged, $X_w^\top Z_b = (X^\top (I-P)^\top PZ)^\top = 0$ since $(I-P)P = 0$ as well. This leaves $X^\top Z = X_b^\top Z_b + X_w^\top Z_w = B + W$ exactly, with no remaining cross terms. Squaring the Frobenius norm of a sum of two matrices gives $\|B+W\|_F^2 = \|B\|_F^2 + \|W\|_F^2 + 2\langle B, W\rangle$ by definition of the Frobenius inner product, where $\langle B, W\rangle = \mathrm{tr}(B^\top W)$. The same argument with Z replaced by X gives $X^\top X = X_b^\top X_b + X_w^\top X_w$ and the corresponding expansion of $\|X^\top X\|_F^2$, and likewise for $Z^\top Z$. ■

Proof of Proposition 2. Let $c \geq 2$ classes each have m samples, $n = cm$. Let $v \in \mathrm{range}(P)$ be any nonzero centered class-constant vector (constant within each class, and averaging to zero overall), and let $u_x, u_z \in \mathrm{range}(I-P) \cap \mathbf{1}^\perp$ be any two vectors chosen so that $u_x^\top u_z = 0$. Taking the residuals in $\mathbf{1}^\perp$ keeps X and Z globally centered, since $\mathbf{1} \in \mathrm{range}(I-P)$; this subspace has dimension $n - c$, so the choice is possible once $n \geq c + 2$. Set $X = av\mathbf{e}_1^\top + u_x\mathbf{e}_1^\top$ and $Z = av\mathbf{e}_1^\top + u_z\mathbf{e}_1^\top$ for a scalar a, restricting to one feature dimension for clarity (the argument extends coordinatewise to higher dimensions). Then $X_b = av\mathbf{e}_1^\top$, $X_w = u_x\mathbf{e}_1^\top$, and analogously for Z, so $X_w^\top Z_w = u_x^\top u_z = 0$ by construction for every value of a, giving $\mathrm{CKA}_w(X, Z) \equiv 0$. Meanwhile $X^\top Z = a^2 v^\top v + av^\top u_z + au_x^\top v + u_x^\top u_z$. Since $v \in \mathrm{range}(P)$ and $u_x, u_z \in \mathrm{range}(I-P)$, the cross terms $v^\top u_z$ and $u_x^\top v$ vanish, and $u_x^\top u_z = 0$, so $X^\top Z = a^2\|v\|^2$, while $X^\top X = a^2\|v\|^2 + \|u_x\|^2$ and $Z^\top Z = a^2\|v\|^2 + \|u_z\|^2$ grow at the same rate in a. Therefore $\mathrm{CKA}(X, Z) = \frac{(a^2\|v\|^2)^2}{(a^2\|v\|^2+\|u_x\|^2)(a^2\|v\|^2+\|u_z\|^2)} \to 1$ as $a \to \infty$. ■

Proof of Proposition 3. If $X_b = 0$ and $Z_b = 0$, then trivially $X = X_w$ and $Z = Z_w$, so $\mathrm{CKA}(X, Z) = \mathrm{CKA}_w(X, Z)$ by definition. For condition (b), suppose there exists a scalar $s > 0$ such that

$$\begin{gathered} \|B+W\|_F^2 = s^2\|W\|_F^2, \\ \|X_b^\top X_b + X_w^\top X_w\|_F = s\|X_w^\top X_w\|_F, \qquad \|Z_b^\top Z_b + Z_w^\top Z_w\|_F = s\|Z_w^\top Z_w\|_F. \end{gathered} \tag{12}$$

By Proposition 1, $X^\top Z = B + W$, $X^\top X = X_b^\top X_b + X_w^\top X_w$, and $Z^\top Z = Z_b^\top Z_b + Z_w^\top Z_w$, so the left-hand sides above are exactly $\|X^\top Z\|_F^2$, $\|X^\top X\|_F$, and $\|Z^\top Z\|_F$. Substituting,

$$\mathrm{CKA}(X, Z) = \frac{s^2\|W\|_F^2}{s\|X_w^\top X_w\|_F \cdot s\|Z_w^\top Z_w\|_F} = \frac{\|W\|_F^2}{\|X_w^\top X_w\|_F\|Z_w^\top Z_w\|_F} = \mathrm{CKA}_w(X, Z). \tag{13}$$

The scalar s must be shared across the numerator condition and both denominator conditions; imposing a proportionality condition on the numerator alone, without matching conditions on both denominator factors, is not sufficient to conclude $\mathrm{CKA} = \mathrm{CKA}_w$, since the denominator interference terms $2\langle X_b^\top X_b, X_w^\top X_w\rangle$ and $2\langle Z_b^\top Z_b, Z_w^\top Z_w\rangle$ would otherwise remain unconstrained. The conditions are sufficient and not necessary, since equal CKA with unequal proportionality constants is not excluded. ■

A.2 Numerical Instantiation of the Counterexample

Table 3 instantiates Proposition 2 on synthetic centered features with an explicit orthogonal within-class residual. As the shared class-constant component is scaled up, ordinary CKA rises from 0 to 0.969 while the within-class CKA remains exactly 0 at every scale.

Table 3: The constructive counterexample of Proposition 2. Ordinary CKA and within-class CKA on synthetic centered features, with the shared class-constant component scaled by the factor in the first column.

Class scale	Ordinary CKA	Within-class CKA
0	0	0
1	0.250	0
2	0.640	0
4	0.886	0
8	0.969	0

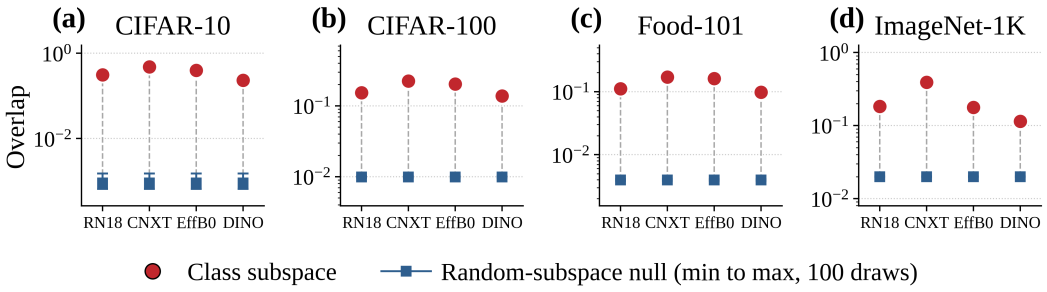

Figure 3: Overlap between the class projector and each model's top-k PCA subspace, with $k = \min(c-1, d-1)$ (circles), against the same overlap for 100 random rank-k orthonormal subspaces of the sample space (squares with min-to-max bars), on (a) CIFAR-10, (b) CIFAR-100, (c) Food-101, (d) ImageNet-1K. The vertical axis is logarithmic. On ImageNet-1K, EfficientNet-B0 is the only backbone with k exactly matched to $\mathrm{rank}(P)$.

A.3 Class Subspace and Dominant Variance Directions

Figure 3 compares the subspace overlap between the class projector and the top-k PCA projector, with $k = \min(c-1, d-1)$ as described in Section 4, against a null distribution built from 100 independently drawn random rank-k orthonormal subspaces of the sample space. On every dataset and backbone, the observed overlap substantially exceeds the random-subspace null, whose mean matches the predicted $\mathrm{rank}(P)/n$ to four decimal places on all four datasets. The null mean runs from 0.0009 on CIFAR-10 to 0.020 on ImageNet-1K, with a standard deviation of at most 0.0002, while the overlap between the class and PCA subspaces ranges from 0.098 to 0.475. On CIFAR-10, CIFAR-100, and Food-101, where $k = c - 1$ matches $\mathrm{rank}(P)$ exactly for every backbone, this overlap is one to two orders of magnitude above the null maximum. On ImageNet-1K, EfficientNet-B0's final feature dimension of 1280 gives $k = 999 = \mathrm{rank}(P)$ exactly, the one setting where the two subspaces are rank-matched at the full 999-dimensional class rank, and there EfficientNet-B0's overlap of 0.178 is $8.9\times$ the null maximum of 0.020. For the remaining three backbones on ImageNet-1K, feature dimensions smaller than 999 cap k below $\mathrm{rank}(P)$, and the margin above the null narrows, with DINO's overlap of 0.115 at about $5.7\times$ the null maximum.

The class subspace is neither contained in the dominant PCA subspace nor orthogonal to it. Proposition 1 quantifies the class-aligned component of the numerator and of both denominator factors without requiring a choice between a class-specific and a dominant-direction account.

A.4 Between-Class Variance Across Depth

Figure 4 shows the between-class variance fraction ρ as a function of normalized relative depth for every backbone on every dataset, a summary of the between- and within-class geometry that the neural collapse literature (Papyan et al., 2020) studies at the terminal phase of supervised training. On ImageNet-1K, ResNet-18 and ConvNeXt-Tiny show ρ increasing monotonically with depth, most strongly for ConvNeXt-Tiny, which reaches $\rho = 0.587$ at the final depth. EfficientNet-B0's

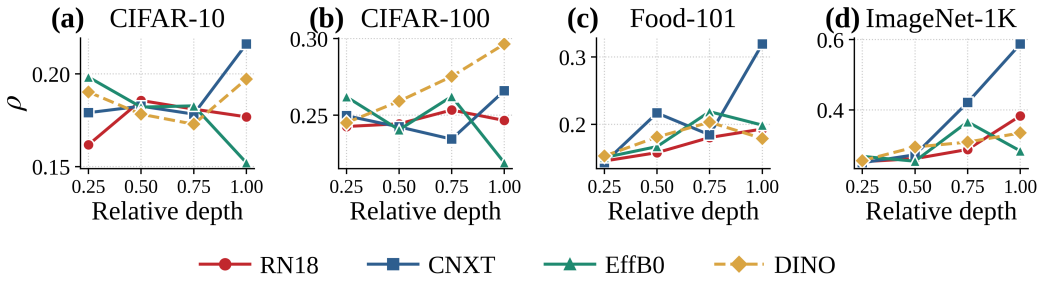

Figure 4: Between-class variance fraction $\rho = \|X_b\|_F^2 / \|X\|_F^2$ as a function of normalized relative depth for each backbone, on (a) CIFAR-10, (b) CIFAR-100, (c) Food-101, (d) ImageNet-1K.

trajectory on ImageNet-1K is non-monotonic (0.267, 0.254, 0.366, 0.283 across the four depths), peaking at the third depth instead of the last. The self-supervised DINO ViT-S also increases with depth on this dataset and reaches a final-depth value of 0.334, above EfficientNet-B0's 0.283. On CIFAR-10 and CIFAR-100, all four backbones show flat or non-monotonic trajectories, and the DINO backbone again exceeds ResNet-18 and ConvNeXt-Tiny at several depths on CIFAR-100.

A.5 Final-Depth Difference for All Combinations

Table 4 gives Δ at the final depth for every model pair on every dataset, covering all 24 combinations summarized in Section 5.

Table 4: Final-depth $\Delta = \mathrm{CKA} - \mathrm{CKA}_w$ for all six model pairs on all four datasets.

Pair	CIFAR-10	CIFAR-100	Food-101	ImageNet-1K
RN18 vs CNXT	+0.273	+0.188	+0.194	+0.074
RN18 vs EffB0	+0.211	+0.179	+0.178	+0.111
RN18 vs DINO	+0.180	+0.182	+0.078	+0.211
CNXT vs EffB0	+0.228	+0.175	+0.189	+0.022
CNXT vs DINO	+0.185	+0.181	+0.080	+0.055
EffB0 vs DINO	+0.176	+0.195	+0.092	+0.046

A.6 Extended Numerical Results

Table 5 extends Table 1 with reconstruction error, debiased CKA, PCA-removed CKA, and residual effective rank for further model pairs and datasets at the final depth. The debiased-CKA and PCA-removed-CKA values shown here are computed at the final depth pair.

Table 5: Extended final-depth decomposition table (further combinations, complementing Table 1) with reconstruction error, debiased CKA, PCA-removed CKA, and effective rank (EffRank) of each side's within-class Gram matrix, defined as $(\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ for its eigenvalues λ_i.

Dataset	Pair	CKA	CKA_w	Δ	Recon. err. (10^{-15})	Debiased	PCA-removed	$\mathrm{EffRank}_X$	$\mathrm{EffRank}_Z$
CIFAR-10	RN18 vs CNXT	0.679	0.406	+0.273	9.4	0.678	0.314	76.0	98.5
CIFAR-10	RN18 vs EffB0	0.677	0.466	+0.211	9.3	0.676	0.398	76.0	134.7
CIFAR-10	RN18 vs DINO	0.522	0.342	+0.180	7.9	0.521	0.230	76.0	25.1
CIFAR-100	RN18 vs CNXT	0.587	0.399	+0.188	3.1	0.585	0.116	80.6	114.5
CIFAR-100	EffB0 vs DINO	0.506	0.311	+0.195	3.1	0.505	0.093	160.4	26.7
Food-101	RN18 vs CNXT	0.527	0.333	+0.194	5.2	0.526	0.076	90.0	65.1
Food-101	CNXT vs DINO	0.299	0.220	+0.080	5.6	0.299	0.076	65.1	15.4
ImageNet-1K	RN18 vs CNXT	0.162	0.088	+0.074	7.5	0.160	< 0.001	178.9	80.9
ImageNet-1K	RN18 vs DINO	0.417	0.206	+0.211	6.9	0.416	< 0.001	178.9	34.4
ImageNet-1K	CNXT vs EffB0	0.366	0.344	+0.022	7.0	0.365	0.001	80.9	250.1

PCA-removed CKA, obtained by removing the top-k sample-space PCA subspace (in place of the class projector P) before recomputing CKA, is comparable in magnitude to CKA_w on the three datasets with fewer than 1000 classes, where k is rank-matched to $\mathrm{rank}(P)$ for every backbone, and is near zero on ImageNet-1K, where $k = d - 1$ for ResNet-18, ConvNeXt-Tiny, and DINO ViT-S, so the removed subspace leaves a single direction of the feature space and the column carries no information for those three backbones. The comparison is therefore read on the three datasets with fewer than 1000 classes.

A.7 Full Consequence Table

For every source depth we also record whether the nearest target depth under ordinary CKA differs from the nearest target depth under within-class CKA. Among the four ImageNet-1K pairs whose global best-matching depth pair changes, the nearest-target assignment at each of the four source depths changes for one pair only, ConvNeXt-Tiny versus EfficientNet-B0, and at one of its four source depths; the other three pairs keep all four per-depth nearest-target assignments while their single global best match moves.

Table 6 reports the depth-matching comparison across all four datasets; the ImageNet-1K rows cover all six model pairs. On the three datasets whose label sets differ from the pretraining one, the best-matching depth pair is unchanged for every pair shown; a per-depth nearest-layer assignment changes at one source depth for CIFAR-100 CNXT vs DINO and at one source depth for Food-101 RN18 vs DINO, without the global best-matching depth pair changing.

Table 6: Depth-matching comparison across all datasets (selected pairs; ImageNet-1K rows cover all six pairs). The two columns give the (source depth, target depth) pair maximizing ordinary and within-class CKA over the full 4×4 grid, with depths indexed 1 to 4.

Dataset	Pair	Best (CKA)	Best (CKA_w)	Differs
CIFAR-10	RN18 vs CNXT	$(2, 1)$	$(2, 1)$	No
CIFAR-10	RN18 vs EffB0	$(1, 1)$	$(1, 1)$	No
CIFAR-10	RN18 vs DINO	$(2, 1)$	$(2, 1)$	No
CIFAR-100	RN18 vs CNXT	$(2, 1)$	$(2, 1)$	No
CIFAR-100	CNXT vs DINO	$(1, 1)$	$(1, 1)$	No
Food-101	RN18 vs DINO	$(3, 2)$	$(3, 2)$	No
ImageNet-1K	RN18 vs CNXT	$(2, 1)$	$(1, 1)$	Yes
ImageNet-1K	RN18 vs EffB0	$(1, 1)$	$(1, 1)$	No
ImageNet-1K	RN18 vs DINO	$(3, 2)$	$(1, 1)$	Yes
ImageNet-1K	CNXT vs EffB0	$(1, 2)$	$(1, 1)$	Yes
ImageNet-1K	CNXT vs DINO	$(1, 1)$	$(1, 1)$	No
ImageNet-1K	EffB0 vs DINO	$(2, 2)$	$(1, 1)$	Yes

A.8 Pairs With and Without the DINO Backbone

We compare the mean final-depth Δ for supervised-supervised model pairs against pairs that include the self-supervised DINO backbone, per dataset. Supervised-supervised pairs give a larger mean Δ on CIFAR-10 (0.237 versus 0.180) and Food-101 (0.187 versus 0.083), a comparable mean on CIFAR-100 (0.181 versus 0.186), and a smaller mean on ImageNet-1K (0.069 versus 0.104). The reversal across datasets shows that the pattern is not consistently associated with whether a pair includes the DINO backbone.

A.9 Depth-Diagonal Sensitivity

The difference Δ measured along the matched-depth diagonal, at normalized relative depths $0.25, 0.5, 0.75, 1.0$, increases with depth for most model pairs on CIFAR-10, CIFAR-100, and Food-101, with the largest value at relative depth 1.0 for every pair on these three datasets. On ImageNet-1K the pattern is less uniform. Several pairs (for example ResNet-18 versus ConvNeXt-Tiny, and ResNet-18 versus DINO) peak at relative depth 1.0, while the ResNet-18/EfficientNet-B0 diagonal

peaks at relative depth 0.75 ($\Delta = 0.123$, versus 0.111 at relative depth 1.0), and the ConvNeXt-Tiny/EfficientNet-B0 diagonal peaks at relative depth 0.5 ($\Delta = 0.045$, versus 0.022 at relative depth 1.0). The depth at which Δ peaks on ImageNet-1K therefore depends on the pair of backbones.

A.10 CLS-Token Pooling Ablation for the DINO Backbone

Table 7 compares the final-depth patch-mean pooling used throughout the main text against CLS-token pooling for every pair involving the DINO ViT-S backbone. Ordinary CKA and within-class CKA are higher under CLS pooling for every pair and dataset, and Δ stays positive under both pooling choices, at $+0.078$ to $+0.239$ under CLS pooling against $+0.046$ to $+0.211$ under patch-mean pooling.

Table 7: Final-depth CKA / CKA$_w$ / Δ for every DINO ViT-S pair under patch-mean (PM) pooling, used throughout the main text, versus CLS-token pooling.

Dataset	Pair	PM: CKA/CKA$_w$/Δ	CLS: CKA/CKA$_w$/Δ
CIFAR-10	RN18 vs DINO	$0.522/0.342/+0.180$	$0.670/0.485/+0.185$
CIFAR-10	CNXT vs DINO	$0.491/0.306/+0.185$	$0.661/0.435/+0.225$
CIFAR-10	EffB0 vs DINO	$0.478/0.303/+0.176$	$0.686/0.511/+0.175$
CIFAR-100	RN18 vs DINO	$0.517/0.335/+0.182$	$0.629/0.453/+0.176$
CIFAR-100	CNXT vs DINO	$0.485/0.304/+0.181$	$0.590/0.402/+0.188$
CIFAR-100	EffB0 vs DINO	$0.506/0.311/+0.195$	$0.623/0.446/+0.176$
Food-101	RN18 vs DINO	$0.413/0.335/+0.078$	$0.664/0.491/+0.174$
Food-101	CNXT vs DINO	$0.299/0.220/+0.080$	$0.597/0.390/+0.207$
Food-101	EffB0 vs DINO	$0.335/0.243/+0.092$	$0.661/0.494/+0.168$
ImageNet-1K	RN18 vs DINO	$0.417/0.206/+0.211$	$0.582/0.343/+0.239$
ImageNet-1K	CNXT vs DINO	$0.126/0.071/+0.055$	$0.204/0.127/+0.078$
ImageNet-1K	EffB0 vs DINO	$0.135/0.089/+0.046$	$0.324/0.216/+0.107$

A.11 Implementation Details

Final-depth feature dimensions are 512 for ResNet-18, 768 for ConvNeXt-Tiny, 1280 for EfficientNet-B0, and 384 for DINO ViT-S, which sets the PCA rank $k = \min(c-1, d-1)$ used in the subspace control of Section 4. The DINO ViT-S blocks are taken at zero-based indices 2, 5, 8, and 11. The relative reconstruction error $\|X^\top Z - (B+W)\|_F / \|X^\top Z\|_F$ of Proposition 1 stays below 10^{-14} for every depth pair. Feature extraction uses PyTorch (Paszke et al., 2019) with automatic mixed precision at bfloat16 for the forward pass on all four backbones, with pooled features stored in float32 and all decomposition cross-products accumulated in float64. The shorter side is resized to 256 pixels before the center crop, and normalization uses channel mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. The CIFAR-10, CIFAR-100, and Food-101 datasets are loaded through their standard Hugging Face `datasets` splits (test splits for all three); ImageNet-1K uses the complete $50{,}000$-image validation split. The three convolutional backbones use torchvision `IMAGENET1K_V1` pretrained weights (`resnet18`, `convnext_tiny`, `efficientnet_b0`); the DINO ViT-S backbone is the `timm` (Wightman, 2019) checkpoint `vit_small_patch16_224.dino`, from which the patch tokens, excluding the CLS token, are averaged at each of the four blocks; CLS-token pooling is compared in Appendix A.10. Statistical controls use 100 label permutations, 500 two-stage class bootstrap resamples, random balanced partitions at 10, 100, and 1000 groups, and 100 Haar-random-subspace draws for the PCA-overlap null; the controls that require dimension matching operate on a Gaussian random projection of the centered features to 64 dimensions, shared by every control applied to a given (model pair, dataset, depth pair) combination.