POST-HOC NORMALISATION: A PREREQUISITE FOR LINEAR EVALUATION OF SELF-SUPERVISED FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear evaluation by fitting a linear classifier on frozen features is the standard yardstick for the quality of self-supervised representations. It is, however, acutely sensitive to a geometric artefact that has nothing to do with representational content: *anisotropy*, the concentration of features in a narrow, off-centre, correlated cone, which is severe in vision transformers and inflates apparent inter-class similarity. We show that the benefit of correcting this geometry *post hoc*, before the linear probe, is not a fixed property of the correction, but scales inversely with how well-conditioned the frozen features already are. We propose a supervised, training-free, representation-space normalization that isotropizes intra-class (noise) covariance while leaving inter-class (signal) structure intact, and position this normalisation precisely to distinguish it from in-network normalization (BatchNorm/LayerNorm), projector-space anti-collapse normalization (DINO centering), and unsupervised post-hoc whitening. On CIFAR-100, the correction lifts a reconstruction-pretrained backbone (BEiT) by $+27.3$ points, a self-distillation backbone trained on ImageNet-1k ($\approx$1.3M images, DINO v1) by $+12.2$ points, an otherwise-identical backbone that adds a patch-level masked-image term at the same scale (iBOT) by $+8.4$ points, and a self-distillation backbone trained on 142M curated images (DINOv2) by $+1.2$ points. This monotone ladder separates two contributing factors of anisotropy: at fixed data scale a stronger objective (DINO v1$\rightarrow$iBOT), and at fixed objective term $\approx 110\times$ more data (iBOT$\rightarrow$DINOv2). Fairness of linear evaluation is clouded by anisotropy induced complexity of optimisation, which the proposed post-hoc normalisation mitigates. Anisotropy diagnostics on 30000 CIFAR-100 photographs and on ten other benchmarks corroborate the mechanism. We also demonstrate the merits of the normalisation on the task of covid classification in Computer Tomography (CT) images.

1 INTRODUCTION

Normalization is ubiquitous in machine learning, and its form and merits have been studied at several distinct points of the deep-network pipeline: on the input data (LeCun et al., 1998; Coates et al., 2011), on internal feature activations during training (Ioffe & Szegedy, 2015; Ba et al., 2016), and on the output of an auxiliary *projector* head in self-supervised learning to prevent representational collapse (Caron et al., 2021; Zbontar et al., 2021). These stages differ not only in *where* they act but in *why*: input and feature normalization improve the conditioning of optimization, and their canonical justification, i.e. the reduction of "internal covariate shift", was later shown to be largely incorrect, the true benefit being a smoothing of the loss landscape (Santurkar et al., 2018); projector normalization instead exists to keep a contrastive or self-distillation objective from collapsing. Section 2 reviews these stages and their historical arguments in full, and uses them to position the present work.

We focus on a fourth, less-studied stage: *post-hoc normalization of the frozen representation*, applicable when a model pretrained by self-supervised learning is to be used, without fine-tuning, for a downstream classification task. This is the setting of linear evaluation, the field's standard measure of representation quality.

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.

We show that the representation of downstream-task data is, in this setting, highly anisotropic, especially for transformer architectures, where a handful of high-norm tokens and dimensions dominate the feature geometry (Timkey & van Schijndel, 2021; Darcet et al., 2024). This anisotropy can partially be neutralized during pretraining, either architecturally, by introducing dedicated *register* tokens that absorb the rogue computation (Darcet et al., 2024), by stronger training objective that explicitly spreads the feature distribution (Zhou et al., 2022), by pretraining at very large scale, as in DINOv2 (Oquab et al., 2024), or by supervised fine-tuning.

However, even large scale foundation models such as DINOv2 or even DINOv3 may be used on data such as medical imaging (MRI, CT, histopathology whole-slide images), remote sensing (multispectral and hyperspectral satellite imagery, synthetic-aperture radar), thermal and infrared imaging, depth maps and LiDAR point clouds, fluorescence and electron microscopy, industrial defect inspection, seismic and geophysical data, astronomical imaging, and audio spectrograms, that is far from natural web images and their domains may amplify anisotropy. Moreover, the use of large-scale models is not always appropriate. Many applications may require a less powerful architecture, e.g. edge based systems, trained on data that is limited in size, private, or proprietary to preclude web-scale pretraining. In such settings, the capacity of a bespoke foundation model will be curtailed and, as we show, it will exhibit stronger residual anisotropy.

Anisotropy introduces correlations that inflate inter-class similarity, and this in turn degrades downstream classification. We demonstrate the effect directly by comparing two backbones from the *same* self-distillation family that differ in training-set size: DINOv2, pretrained on ≈142M curated images, and the original DINO (v1), pretrained only on ImageNet-1k (≈1.28M images, roughly $110\times$ less). On CIFAR-100 their raw linear-probe accuracies are 88.9% and 65.8% respectively, a 23-point gap. As DINOv2 additionally adds a patch-level masked-image (iBOT) term absent from DINO v1, to keep data scale from being confounded with the objective we consider a third, controlled backbone, namely iBOT pretrained on the same ImageNet-1k as well, and show (Section 5) that this term accounts for only about a third of the accuracy gap, leaving the majority attributable to the smaller model's poorer feature conditioning at $110\times$ less data.

To counteract anisotropy we propose a post-hoc normalization applied once, offline, to the frozen features before the linear probe. In essence it subtracts the global mean, estimates the average *within-class* covariance over the training labels, and applies the inverse-square-root of that covariance as a linear whitening map. Unlike whitening the total covariance, which isotropizes every direction including those that separate the class means, within-class whitening isotropizes only the intra-class (noise) covariance and leaves the between-class (signal) structure untouched. This is a supervised, training-free transform closer in spirit to Fisher discriminant analysis than to unsupervised sentence-embedding whitening.

We demonstrate that this correction impacts only marginally on a strong, well-conditioned backbone such as DINOv2 on an easy benchmark, but is very effective on weaker foundation models and on harder datasets where even a strong backbone's raw features are poorly conditioned. Beyond CIFAR-100, performance gains are demonstrated on ten further datasets for both a weak (DINO v1) and a strong (DINOv2) backbone: on the weak backbone the correction improves every dataset, by $+25.6$ points on average, roughly twice its $+13.3$-point effect on the strong one. This validates the method and shows that its benefit is governed by a single principle: it scales with how anisotropic the frozen representation already is, whether that anisotropy comes from a weaker backbone, less effective training or a harder task. To demonstrate the merits of the proposed post-hoc normalisation in scenarios where the data domain changes, we validate it on a task of covid classification in CT scan images.

Our contributions can be summarised as follows:
— We demonstrate the anisotropy of representations produced by self-supervised learning foundation models and its dependence on their training data size and training objective.
— We propose a post-hoc normalisation of the representations (centering and pre-whitening) for the downstream target data as a means to perform efficient linear evaluation and as a tool to gauge whether further training of the backbone is required.
— The merits of the post-hoc normalisation are demonstrated on linear evaluation of the representations carried out on standard benchmarking datasets, as well as on the problem of covid classification in Computer Tomography images of human chest.

The paper is organized as follows. Related literature is presented in Section 2, which concludes with a clear positioning of our contribution with respect to the state of the art. Section 3 introduces the methodology. The experiments are described in Section 4. The findings on benchmarking datasets are discussed in Section 5. The medical data experiment and its results are presented in Section 6. The paper concludes in Section 7.

2 Related work

We organize the normalization literature by the *stage* of the pipeline it acts on, trace the historical arguments that converged each stage to common practice, and thereby locate our contribution.

Input-space normalization. The oldest normalization acts on the input data. LeCun et al. (1998) argued that inputs should be centered, decorrelated, and variance-equalized because this improves the conditioning of the loss Hessian and accelerates gradient descent; taken to its limit this is input *whitening*, standard in early feature-learning pipelines (Krizhevsky, 2009; Coates et al., 2011). The enduring lesson, that decorrelating and equalizing variance improves the conditioning of what a linear model must fit, is exactly the one we re-apply, but to a frozen representation rather than the raw input.

In-network (feature-space) normalization, and the covariate-shift debate. Batch Normalization (Ioffe & Szegedy, 2015) standardizes each feature at every layer. It was introduced with the claim that it reduces *internal covariate shift*. A family of variants followed, each re-choosing the pooling axis: Layer (Ba et al., 2016), Instance (Ulyanov et al., 2016), and Group (Wu & He, 2018) Normalization, and Weight Normalization (Salimans & Kingma, 2016). The instructive part is that the original *justification* did not survive scrutiny but the practice did: Santurkar et al. (2018) showed batch normalization does not reduce covariate shift. Deliberately injecting distributional noise after the normalization layer leaves training fast and stable and its real effect is a smoothing of the optimization landscape, corroborated by Bjorck et al. (2018). Transformer practice converged further, on Pre-LayerNorm placement (Xiong et al., 2020) and RMSNorm (Zhang & Sennrich, 2019). Crucially for us, a more powerful thread was tried and largely rejected: Decorrelated Batch Normalization (Huang et al., 2018) and Iterative Normalization (Huang et al., 2019) generalize standardization to full *whitening* inside the network, but their per-step matrix inverse-square-root proved too costly and unstable to become default. The field settled on cheap per-feature standardization and set full in-network whitening aside.

Projector-space normalization in self-supervised learning. Self-supervised learning introduced a distinct locus of normalization, aimed at preventing representational *collapse*, acting on a projector or prototype head that is discarded at inference. Collapse can be prevented architecturally (BYOL's momentum target (Grill et al., 2020), SimSiam's stop-gradient (Chen & He, 2021)), which is why those methods fall outside a review of normalization; the relevant methods here prevent collapse through an explicit normalization of the projector output. DINO (Caron et al., 2021) centers and sharpens the teacher's output distribution; DINOv2 (Oquab et al., 2024) uses a Sinkhorn–Knopp normalization; Barlow Twins (Zbontar et al., 2021) drives the embedding cross-correlation toward the identity (a whitening-flavoured *objective*); VICReg (Bardes et al., 2022) makes the variance and covariance terms explicit; and W-MSE (Ermolov et al., 2021) whitens the batch as the loss. All operate on a training-only head, so their effect on the frozen backbone geometry is indirect and emergent — a distinction essential to our positioning.

Representation-space (post-hoc) normalization. The stage we contribute to transforms the *frozen* representation after training. In NLP, the anisotropy of contextual embeddings (Gao et al., 2019; Ethayarajh, 2019) prompted training-free corrections: All-but-the-Top removes the mean and top principal directions (Mu & Viswanath, 2018), BERT-flow learns a map to a Gaussian (Li et al., 2020), BERT-whitening applies a single linear whitening transform (Su et al., 2021), and a few high-variance "rogue" dimensions are shown to dominate similarity (Timkey & van Schijndel, 2021). In few-shot vision, SimpleShot centers and L2-normalizes frozen features before a nearest-neighbour classifier (Wang et al., 2019), and neural-collapse theory (Papyan et al., 2020) characterizes the terminal geometry such corrections target. These methods are unsupervised; our within-class whitening differs by using labels to isotropize only the intra-class covariance (closer to Fisher LDA), and by targeting the specific geometry that linear evaluation measures.

Why anisotropy differs across objectives, and positioning. A well-formed self-distillation or contrastive objective explicitly spreads the feature distribution over the hypersphere (alignment and uniformity, Wang & Isola, 2020), an anti-anisotropy pressure absent from purely reconstructive objectives such as BEiT (Bao et al., 2022) and MAE (He et al., 2022) (Yue et al., 2023). This is why raw DINOv2 features are far better conditioned than raw BEiT features. **Our positioning is then precise:** we do not propose another in-network or projector-space normalization competing with BatchNorm or DINO centering. We revive the one operation the in-network settlement rejected, namely full, and *within-class*, whitening, at the single stage where its cost and instability objections vanish. It is applied once, offline, to frozen features, as a deterministic, supervised prerequisite that lets linear evaluation measure representational *content* rather than incidental *conditioning*. Its value, we show, scales inversely with how much conditioning the pretraining objective and data already supplied.

3 METHOD

For representation normalisation we use standard statistical methods for centering and prewhitening. The latter is in two variants, using either an average within-class covariance matrix or the total covariance. It cannot be overemphasised that our contribution lies in applying the proposed post-hoc normalisation for the novel purpose of prewhitening the representations of the target set data generated by a pretrained backbone. The following overview is included just for completeness.

Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be the frozen d-dimensional features of the labelled downstream training set, with $y_i \in \{1, \ldots, C\}$, and let $\boldsymbol{x}$ be a test feature. We seek a single linear map, estimated once on the training set, that conditions the geometry for a linear readout without discarding class information.

Within-class covariance whitening. We first remove the global mean $\boldsymbol{\mu} = \frac{1}{N}\sum_i \boldsymbol{x}_i$. We then estimate the average within-class covariance

$$\boldsymbol{\Sigma}_W \;=\; \frac{1}{C}\sum_{c=1}^{C} \frac{1}{n_c - 1} \sum_{i:y_i=c} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^\top, \qquad \boldsymbol{\mu}_c = \frac{1}{n_c}\sum_{i:y_i=c} \boldsymbol{x}_i, \tag{1}$$

the mean of the C per-class covariance matrices, which captures intra-class (noise) scatter while, by construction, excluding the between-class (signal) scatter of the class means. With the eigendecomposition $\boldsymbol{\Sigma}_W = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, the whitening map is

$$\boldsymbol{W}(\boldsymbol{x}) \;=\; \boldsymbol{\Lambda}^{-1/2}\boldsymbol{U}^\top(\boldsymbol{x} - \boldsymbol{\mu}), \tag{2}$$

which rescales the intra-class variation directions to unit average within-class variance. Finally we remove the residual mean of the whitened class means, $\boldsymbol{r} = \frac{1}{C}\sum_c \frac{1}{n_c}\sum_{i:y_i=c} \boldsymbol{W}(\boldsymbol{x}_i)$, giving the normalized feature $\boldsymbol{z} = \boldsymbol{W}(\boldsymbol{x}) - \boldsymbol{r}$. The map is estimated on the training split and applied unchanged to the test split; the linear probe (or a nearest-class-mean classifier) is then fit on $\boldsymbol{z}$.

Contrast with total-covariance whitening. Whitening the *total* covariance $\boldsymbol{\Sigma}_T = \frac{1}{N-1}\sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$ isotropizes *every* direction, including those along which the class means differ. Since enforcing global isotropy is provably in tension with cluster structure (Mickus et al., 2024), total-covariance whitening flattens signal together with noise, whereas $\boldsymbol{\Sigma}_W$ leaves the between-class directions untouched. Our experiments confirm within-class whitening dominates the total-covariance variant.

Numerical stabilization. With few samples per class relative to d, or in the presence of a degenerate direction induced by a final LayerNorm, $\boldsymbol{\Sigma}_W$ can have near-zero (and, from floating-point noise, slightly negative) eigenvalues, for which $\boldsymbol{\Lambda}^{-1/2}$ in Eq. 2 is ill-defined. We add a small relative floor ε to the eigenvalues before inversion ($\lambda \leftarrow \max(\lambda, 0) + \varepsilon\,\lambda_{\max}$, with $\varepsilon = 10^{-8}$); results are insensitive to ε over several orders of magnitude, and equivalent to simply dropping the degenerate directions.

4 EXPERIMENTS

The aim of our experiments is to demonstrate the dependence of the magnitude of anisotropy on the foundation model strength as defined by training set size and training objective. The vehicle for the

Table 1: Anisotropy of raw *mean-pooled patch-token* features on 30,000 ImageNet-100 photographs. "Eccentricity" is the global-mean norm as a percentage of a typical embedding norm (lower is more centered); "participation" is the covariance participation ratio (higher is whiter; 768 = perfectly white). DINO v1 and iBOT share ImageNet-1k and differ only in the patch-level term: it halves the eccentericity deficit ($89.1\% \to 57.6\%$) while barely moving whiteness.

Backbone	Pretraining data	Eccentricity (mean/typical) ↓	Participation ↑
BEiT-base	ImageNet-22k (≈14M)	93.1%	12.6
DINO v1-base	ImageNet-1k (≈1.3M)	89.1%	80.4
iBOT-base	ImageNet-1k (≈1.3M)	57.6%	87.2
DINOv2-base	LVD-142M (≈142M)	30.1%	95.0

demonstration is the CIFAR-100 dataset. The merits of the proposed post-hoc normalisation are in addition demonstrated on ten other standard benchmarking datasets, as well as on a medical data application.

Setup. We study four ViT-B/768-dimensional backbones that hold model capacity fixed while varying the pretraining objective and data scale: BEiT-base (masked-image reconstruction, ImageNet-22k, ≈14M images), DINO v1 ViT-B/16 (image-level self-distillation, ImageNet-1k, ≈1.28M images), iBOT ViT-B/16 (the same self-distillation *plus* a patch-level masked-image term, at the *same* ImageNet-1k scale, Zhou et al., 2022), and DINOv2-base (self-distillation with the same iBOT patch-level term, LVD-142M, ≈142M images). The iBOT backbone is the controlled middle reference point: it differs from DINO v1 in the objective alone (data held fixed) and from DINOv2 in data scale (the patch-level term held fixed), which lets us separate the two causes the three-backbone ladder confounds. For each image we extract the mean-pooled patch-token representation. Unless stated otherwise the classifier is a linear probe trained for 60 epochs (AdamW, cosine schedule); the normalization map is estimated only on the training split, with the test split held out throughout. We report top-1 accuracy.

4.1 Downstream representations are strongly anisotropic, and objective and scale set the degree

On CIFAR-100, the raw BEiT feature covariance has a maximum-to-minimum eigenvalue ratio of $\approx 485{,}000{:}1$: features occupy a narrow, off-centre cone. Table 1 quantifies this on 30,000 real ImageNet-100 photographs for all four backbones, via the global mean's norm as a fraction of a typical embedding norm (eccentricity) and the covariance *participation ratio* (768 equally contributing dimensions would be perfectly white). Raw BEiT features are severely uncentered (93%) and use only ≈ 13 of 768 effective dimensions. DINOv2, by contrast, is far better centered (30%) and uses ≈ 95. The two ImageNet-1k backbones locate what drives this. DINO v1 is already near-DINOv2 on whiteness (≈ 80), so its image-level objective delivers second-moment conditioning even at small scale, but it is as poorly centered as BEiT (89%). Adding the iBOT patch-level term at the *same* data scale moves eccentricity to 57.6% — roughly half-way to DINOv2 — while barely changing whiteness (≈ 87). First-moment (eccentricity) conditioning is therefore not simply what "degrades with less data": at fixed data scale about half of DINO v1's centering deficit is the missing patch-level objective, the remainder being data scale (Section 5).

4.2 Anisotropy costs downstream accuracy, and the correction scales with it

Table 2 reports the CIFAR-100 linear-probe accuracy of all four backbones under raw features, centering only, total-covariance whitening, and within-class whitening. Two results stand out. First, the raw accuracies themselves track conditioning: DINOv2 (88.9%) > iBOT (72.9%) > DINO v1 (65.8%) ≫ BEiT (31.9%). Second, the benefit of within-class whitening is a clean monotone function of that conditioning: +27.3 points for BEiT, +12.2 for DINO v1, +8.4 for iBOT, and only +1.2 for DINOv2. The correction is marginal exactly where the representation is already well-conditioned, and transformative where it is not. The two ImageNet-1k rows make the reading of this ordering precise. DINO v1 and iBOT are trained on the *same* 1.3M images and differ only in the patch-level term, yet iBOT starts 7.2 points higher raw and needs 3.8 points less correction — so it is

Table 2: CIFAR-100 linear-probe top-1 accuracy under four normalizations, for four capacity-matched ViT-B backbones. "Gain" is within-class whitening minus raw. The correction's benefit grows monotonically as the raw representation becomes less well-conditioned. DINO v1→iBOT isolates the objective (data fixed at ImageNet-1k); iBOT→DINOv2 varies data scale $\approx 110\times$ with the patch-level term held fixed . For calibration, the bottom row shows the results using frozen features of a public ViT-B/16 fine-tuned end-to-end on CIFAR-100 (90.5% test), for which the correction has essentially no headroom left.

Backbone	Objective / data	Raw	Centered	Total-cov	**Within-class**	Gain
BEiT-base	reconstruction / 14M	31.9	42.7	57.4	**59.2**	+27.3
DINO v1-base	self-distill / 1.3M	65.8	72.9	76.9	**78.0**	+12.2
iBOT-base	self-distill+iBOT / 1.3M	72.9	75.7	80.2	**81.3**	+8.4
DINOv2-base	self-distill+iBOT / 142M	88.9	89.1	89.4	**90.1**	+1.2
ViT-B	supervised FT on CIFAR-100	91.2	91.1	91.6	91.6	+0.4

the objective, not data scale, that separates them; iBOT and DINOv2 then share that patch-level term and differ $\approx 110\times$ in data (alongside DINOv2's KoLeo regularizer and minor recipe changes), and it is data scale that chiefly separates *them*. The gain thus tracks how well-conditioned the objective *and* the data scale jointly leave the representation, not either cause alone. Within-class whitening beats the total-covariance variant for every backbone, confirming that preserving the between-class directions by the class mean scatter matters.

4.3 Validation across ten further benchmarks, on both a weak and a strong backbone

To validate the method beyond CIFAR-100, and to test the storyline's core claim directly, we apply the identical pipeline to ten additional datasets for *both* the weak small-data backbone (DINO v1) and the strong large-data backbone (DINOv2), estimating the whitening transform on each dataset's own training labels (Table 3). Two results stand out. First, on the weak DINO v1 backbone the correction is transformative and near-universal: it improves every one of the ten datasets, by $+25.6$ points on average and by as much as $+51.1$ (Oxford Flowers, $34.5 \rightarrow 85.6$) and $+43.9$ (Stanford Cars, $10.6 \rightarrow 54.5$). On the fine-grained sets whose raw DINO v1 features are barely above chance (Stanford Cars 10.6, FGVC Aircraft 8.6), a single deterministic transform multiplies accuracy several-fold. Second, on the strong DINOv2 backbone the same correction averages only $+13.3$ points and is nearly inert on the easy, well-conditioned datasets ($+0.2$ on CIFAR-10, $+1.9$ on Oxford Flowers), its gains surviving only on the fine-grained sets where even DINOv2's raw features are poorly conditioned. The method thus helps roughly twice as much, and far more uniformly, on the weak backbone. The "very effective on weaker foundation models, marginal on strong ones" pattern has now been demonstrated across ten datasets rather than one. Remarkably, after normalization the weak DINO v1 even matches the strong DINOv2 on Caltech-101 (95.5 vs. 95.0).

5 Discussion

A single law: benefit $\propto$ raw anisotropy. Tables 2 and 3 trace the same principle along complementary axes. Whether anisotropy arises from a weaker pretraining objective (BEiT vs. the self-distillation backbones, or DINO v1 vs. iBOT at fixed data), less pretraining data (iBOT vs. DINOv2 at fixed objective term), or a harder downstream task on which even a strong backbone's raw features are poorly conditioned (fine-grained recognition, dofferent domain), the payoff of post-hoc within-class whitening tracks the amount of anisotropy present. Table 3 makes the interaction explicit: the weak DINO v1 backbone, whose raw features are anisotropic on nearly every dataset, benefits everywhere ($+25.6$ mean), whereas the strong DINOv2 backbone benefits only on the hard datasets where its raw features remain poorly conditioned ($+13.3$ mean, near-zero on the easy ones). The method is not a universal accelerant; it is a corrective whose value is exactly the conditioning deficit it removes. This also explains why it belongs before linear evaluation specifically: a fine-tuned or nonlinear readout would learn around the same conditioning, so the geometry the probe is uniquely sensitive to is precisely the geometry the correction repairs.

Table 3: Within-class whitening across ten benchmarks for a weak (DINO v1, ImageNet-1k) and a strong (DINOv2, LVD-142M) backbone, linear-probe top-1. "Δ" is within-class minus raw. The correction improves every dataset for the weak backbone (+25.6 mean), roughly double its effect on the strong one (+13.3 mean), and is near-inert on the strong backbone's already well-conditioned easy datasets.

		DINO v1 (weak, 1.3M)			DINOv2 (strong, 142M)		
Dataset	Cls.	Raw	**Within**	Δ	Raw	**Within**	Δ
Oxford Flowers	102	34.5	**85.6**	+51.1	97.1	**98.9**	+1.9
Stanford Cars	196	10.6	**54.5**	+43.9	35.6	**79.8**	+44.2
CUB-200-2011	200	19.8	**59.8**	+40.1	42.1	**76.4**	+34.3
FGVC Aircraft	100	8.6	**43.7**	+35.0	20.5	**52.7**	+32.3
SUN397	397	27.1	**48.9**	+21.8	66.7	**69.3**	+2.6
Food-101	101	48.0	**66.7**	+18.7	84.8	**88.3**	+3.6
Caltech-101	102	78.0	**95.5**	+17.5	92.7	**95.0**	+2.4
Oxford-IIIT Pets	37	72.8	**86.1**	+13.3	84.6	**91.2**	+6.6
CIFAR-10	10	82.5	**91.6**	+9.1	97.6	**97.8**	+0.2
DTD	47	67.9	**73.8**	+5.9	77.5	**82.6**	+5.1
Mean Δ				**+25.6**			**+13.3**

Objective versus data scale, disentangled directly. The DINO v1→DINOv2 contrast holds the objective family fixed and varies data $\approx 110\times$, but on its own it confounds two changes, since DINOv2 also adds the iBOT patch-level term that DINO v1 lacks. Rather than bound that confound with DINOv2's published ablation (Oquab et al., 2024), which is run at 142M images and measures downstream accuracy, not representation geometry, we run the missing cell of the design: an iBOT backbone at DINO v1's exact ImageNet-1k scale, differing from DINO v1 in the patch-level term alone. The two contrasts then separate cleanly, and they disagree in an informative way. On *geometry*, adding the iBOT term at fixed scale accounts for about half of the full DINO v1→DINOv2 improvement: it closes 53% of the patch-mean centering gap ($89.1\% \to 57.6\%$ of $89.1\% \to 30.1\%$), 47% of the participation-ratio gap, and 49% of an isotropy measure, three statistics on the same $30{,}000$ dataset agreeing on a roughly even split between objective and scale. On *accuracy*, however, the iBOT term closes only $\approx 31\%$ of the raw CIFAR-100 gap ($65.8\% \to 72.9\%$ of $65.8\% \to 88.9\%$) and $\approx 34\%$ of the normalization gain, confining the classification difference chiefly to the data-scale effect, i.e. a direct iBOT-vs-no-iBOT pair at 1.3M rather than 142M images. That the geometric effect is real and not an artifact is confirmed by its specificity to patch tokens: the iBOT term operates on patch-token targets, and accordingly it moves iBOT's *patch-mean* centering by 31.5 points but its *CLS* centering by only 3.1 ($27.8\% \to 24.7\%$). Objective and data scale thus each leave less correctable anisotropy behind, but they act in different proportions on the representation's geometry than on the accuracy that geometry buys.

Limitations. The correction is supervised, requiring downstream labels to estimate $\mathbf{\Sigma}_W$; with very few examples per class this estimate is itself underdetermined (visible as the shrinking headroom on the most data-scarce fine-grained sets), which motivates estimating the transform on a larger disjoint base set.

6 Validation on a small-data medical backbone

Problem and aim. Every backbone in Sections 4 to 5 is a public ImageNet-scale RGB model and every downstream set a natural-image benchmark. The regime that motivates the paper in Section 1, a bespoke model trained on a small non-RGB corpus, is not represented. We therefore test the correction on exactly such a model: a vision transformer self-supervised on roughly 10^4 computed-tomography (CT) scans and linearly evaluated on a clinical classification task. The aim is to check whether the representation of this out-of-domain data shows the same anisotropy behaviour, and whether the same transform recovers the hidden headroom. In addition, we wish to establish to what extent this anisotropy can be rectified by self-supervised pretraining of the architecture using in-domain data, given that its size is relatively small as compared with the usual RGB data pretraining.

Table 4: Anisotropy of frozen CLS features on the 2,196 CC-CCII training scans ($d = 384$) at the ImageNet initialisation and after 200 epochs of iBOT pretraining on CT, raw and within-class whitened. Eccentricity and participation as in Table 1; κ_T, κ_W are $\lambda_{\max}/\lambda_{\min}$ of the total and within-class covariance over non-degenerate directions (one direction is null from the final LayerNorm; the whitened row is measured on the 383 that remain).

Backbone state	Features	Eccentricity ↓	Participation ↑	κ_T ↓	κ_W ↓
iBOT ImageNet init	raw	98.4%	1.3	6.8×10^5	5.6×10^5
iBOT pretrained on CT	raw	96.5%	3.4	1.3×10^6	1.1×10^6

Table 5: CC-CCII (three classes) linear probe accuracy without and with within-class whitening of the frozen CLS features, at the ImageNet initialisation and after CT pretraining. "Selection" is the held-out split used to choose the probe's epoch, the quantity plotted in Figure 1. SOTA is a published reference point on the same dataset, fold and split, obtained with a different backbone.

Backbone state	Features	Selection	Test acc.
SOTA (Pan et al., 2025)	raw	64.1	60.6
iBOT ImageNet init	raw	65.3	59.5
	whitened	74.4	**66.8**
iBOT pretrained on CT	raw	82.1	73.0
	whitened	91.4	**87.9**

We choose iBOT pretrained on ImageNet to guarantee that the backbone has not seen any data from the CT domain.

Backbone, data and protocol. The backbone is iBOT (Zhou et al., 2022) on a ViT-S/16 (d=384) whose patch tokenizer is adapted to CT input. It is initialised from the public ImageNet-1k iBOT checkpoint and pretrained with the iBOT objective for 200 epochs on VoCo-10k (Wu et al., 2024), $\approx$130$\times$ fewer images than ImageNet-1k. Evaluation is on CC-CCII (Zhang et al., 2020), chest CT with three classes (normal, common pneumonia, COVID-19 pneumonia). The 2,928 training scans are split once 75/25 into a norm estimation set (2,196) for every transform. A selection set (732) was used only to choose the probe's epoch. The 741-scan test fold is used for the probe evaluation. Features are the CLS token of the frozen backbone; the probe is one `nn.Linear` trained for 150 epochs (AdamW, learning rate 3×10^{-4}, weight decay 10^{-2}, cosine schedule). The whitening map is fitted on the 2,196 training features and applied unchanged to the selection and test features. Raw and whitened probes share features, split, seed, and hyper-parameters, so the comparison isolates the transform.

Anisotropy before and after CT pretraining. Table 4 gives the diagnostics of Table 1 on the 2,196 training features, for the backbone at the ImageNet initialisation and after the 200 CT epochs. At initialisation, the cloud is almost one-dimensional: 98.4% centering and a participation ratio of 1.3 out of 384. After pretraining, the participation ratio increases to 3.4 and eccentricity drops to 96.5%, but neither of them enough to eliminate anisotropy. The within-class eigenvalue ratio is unchanged in order of magnitude at $\approx10^6$. One direction is exactly null in both states, the LayerNorm-induced degeneracy anticipated in Section 3, and is excluded from the ratios.

Linear evaluation and convergence. Table 5 gives the probe results. On the CT-pretrained backbone, the probe reaches 73.0% test accuracy on raw features and 87.9% on within-class-whitened ones, a gain of $+14.8$ points from a transform with no learned parameters. The ImageNet initialised iBOT gains $+7.3$ points (59.5% to 66.8%). Figure 1 reports convergence for the CT-pretrained backbone: on whitened features the probe reaches 80% selection accuracy at epoch 10 against 65 for raw, exceeds the raw probe's best value over all 150 epochs at epoch 12, and drives the training loss to 0.12 where raw plateaus at 0.51.

Discussion. Set against the rest of the paper, the raw CT features are more anisotropic than any RGB case in Table 1, whose worst entry is BEiT at 93.1% centering and 12.6 effective dimensions, and their within-class eigenvalue ratio is of the order of the $485{,}000$:1 measured for BEiT on

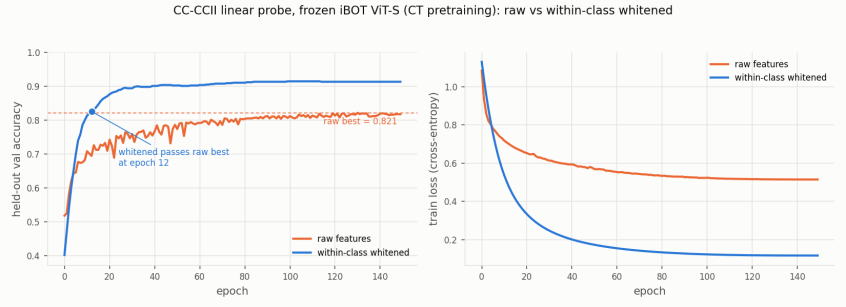


Figure 1: Linear-probe convergence on CC-CCII for the frozen iBOT backbone, raw vs. within-class-whitened CLS features. Left: selection-split accuracy per epoch; the dashed line is the raw probe's best accuracy, which the whitened probe exceeds at epoch 12. Right: training cross-entropy.

CIFAR-100. This demonstrates the impact of out-of-domain applications on the direct usability of foundation models as a basis for their development. The relatively small size ($\approx 10^4$ CT scans) of the in-domain data available for iBOT pretraining is clearly insufficient to mitigate anisotropy. Nevertheless, our backbone model improves the test accuracy for raw features by 13.5%. With external calibration, this compares favourably against a published self-supervised model (Pan et al., 2025) on the same fold and split which reaches 60.6%. Interestingly, the proposed post-hoc whitening transform lifts this enhanced performance by impressive further 21.1%. While admirable per se, the most important message of our post-hoc normalisation is that there is still a capacity to improve performance by further pretraining or fine-tuning. This contrasts with the results presented in Table 2, where after fine-tuning on CIFAR-100, the post-hoc normalisation flagged no headroom for further improvements.

Practical implication. The regime in which this method helps most, namely a bespoke backbone trained on a small, non-web corpus, is exactly the regime forced upon practitioners in domains where large scale domain-relevant pretraining is impossible. We show that a single deterministic transform estimated once on the labelled downstream training set recovers a major part of the accuracy that would be afforded by scale, at negligible cost and with no change to the frozen backbone.

Several observations can be drawn here. First, backbone pretraining or fine-tuning is not an alternative to post-hoc normalisation. These separate processing steps should be used in tandem to achieve the best outcome. Second, the linear probe convergence gap is the practical face of the eigenvalue ratio: with a few directions carrying almost all the variance, gradient descent on raw features spends its budget along them, whereas after whitening every direction has unit variance and one step size serves all of them. The validation uses the CLS token where earlier sections use mean-pooled patch tokens; that the effect transfers across readouts suggests the mechanism is a property of the covariance rather than of a pooling choice. Third, and most importantly, the gains achieved with the post-hoc normalisation are indicative of the capacity of the backbone model to gain from further training. Thus its key practical role is to serve as a gauge for this purpose.

7 Conclusion

We have positioned post-hoc, supervised within-class whitening within the broader landscape of normalization in deep learning, and shown that it fills a specific and useful niche: correcting the anisotropic geometry of frozen self-supervised features, once and offline, as a prerequisite to linear evaluation. Its benefit is governed by a single law: it scales roughly with the anisotropy present, and is therefore marginal on strong, large-data backbones and transformative on the weaker, small-data models that many application domains are obliged to build. As those domains are precisely the ones that cannot directly benefit from a large scale RGB foundation model, the proposed normalization, shown to be most effective exactly where large-scale pre-training is unavailable or its domain is misaligned with the target application, is of direct practical value.

AI USE STATEMENT

In this work, we used generative AI tools for extended literature review, experiment execution and writing. We have not used generative AI tools for creating any synthetic datasets, neither for proving any mathematical claims, as both are absent. More specifically, the ideas, scope, scientific hypothesis, design of the experiments to prove the scientific hypothesis, concept of the paper, as well as a significant part of the writing has been done by the authors. AI was used to execute the experiments and summarise the results. It has also been used to widen the review of the relevant literature and to contribute to writing the paper. The process of writing was i) authors draft detailed paper outline and the introduction to define the problem, aims, and the message of the paper, ii) AI builds on the input in i) to revise the introduction and draft other sections, and iii) authors substantially revise the draft drafted in ii) to produce the final version. No AI generated datasets have been used in the study. The experiments are performed using standard benchmarking datasets. The paper presents no mathematical claims.

ETHICS STATEMENT

The work raises no ethical issues. The medical data used is publicly available, and it is anonymised.

REPRODUCIBILITY STATEMENT

The experiments described in the paper use standard benchmarking datasets, and well known statistical analysis tools which render reproducibility very straightforward. The particulars of foundation models used and other details are provided in the Implementation details section in the Supplementary material section.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Nils Bjorck, Carla P. Gomes, Bart Selman, and Kilian Q. Weinberger. Understanding batch normalization. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Adam Coates, Andrew Y. Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings Machine Learning Research*, 2011.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning (ICML)*, 2021.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations (ICLR)*, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Lei Huang, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. Iterative normalization: Beyond standardization towards efficient whitening. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of Machine Learning Research (PMLR)*, 2015.

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

Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. Efficient BackProp. In *Neural Networks: Tricks of the Trade*. Springer, 1998.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Timothee Mickus, Stig-Arne Grönroos, and Joseph Attieh. Isotropy, clusters, and classifiers. *arXiv preprint arXiv:2402.03191*, 2024.

Jiaqi Mu and Pramod Viswanath. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations (ICLR)*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Tan Pan, Zhaorui Tan, Kaiyu Guo, Dongli Xu, Weidi Xu, Chen Jiang, Xin Guo, Yuan Qi, and Yuan Cheng. Structure-aware semantic discrepancy and consistency for 3d medical image self-supervised learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20257–20267. IEEE, 2025.

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 (PNAS)*, 2020.

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Neural Information Processing Systems (NeurIPS)*, 2016.

Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? In *Neural Information Processing Systems (NeurIPS)*, 2018.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*, 2021.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning (ICML)*, 2020.

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

Linshan Wu, Jiaxin Zhuang, and Hao Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22873–22882. IEEE, 2024.

Yuxin Wu and Kaiming He. Group normalization. In *European Conference on Computer Vision (ECCV)*, 2018.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning (ICML)*, 2020.

Xiaoyu Yue, Lei Bai, Meng Wei, Jiangmiao Pang, Xihui Liu, Luping Zhou, and Wanli Ouyang. Understanding masked autoencoders from a local contrastive perspective. *arXiv preprint arXiv:2310.01994*, 2023.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning (ICML)*, 2021.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Kang Zhang, Xiaohong Liu, Jun Shen, Zhihuan Li, Ye Sang, Xingwang Wu, Yunfei Zha, Wenhua Liang, Chengdi Wang, Ke Wang, et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, 181(6):1423–1433, 2020.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022.

A Implementation details

All features are extracted with the public checkpoints `microsoft/beit-base-patch16-224-pt22k`, `facebook/dino-vitb16`, and `facebook/dinov2-base`, using the mean of the patch tokens of the final layer. The iBOT ViT-B/16 ImageNet-1k teacher backbone has no Hugging Face release; we remap its timm/DeiT-style state dict key-for-key into a `transformers` `ViTModel` — the same translation Hugging Face used to publish `facebook/dino-vitb16`, which makes the two rows architecturally exact counterparts — and validate it two ways before use: the mapped parameter-name set is checked for exact equality (with per-tensor shapes) against the instantiated model and loaded strictly, and the translation is round-tripped on Hugging Face's own DINO v1 weights (HF→timm→HF) requiring bit-exact recovery, which verifies the q/k/v concatenation order a name check cannot see. The

iBOT row uses `facebook/dino-vitb16`'s own image processor so preprocessing matches the DINO v1 row exactly; it is instantiated at LayerNorm $\varepsilon = 10^{-12}$ to match that row (re-extraction at iBOT's native 10^{-6} changes every ladder cell by ≤ 0.07 points and every geometry statistic by ≤ 0.02, leaving all conclusions unchanged). Linear probes are single `nn.Linear` layers trained for 60 epochs with AdamW (learning rate 10^{-3}, weight decay 10^{-4}), a cosine schedule, and batch size 256, on features rescaled by their mean training-set norm. All normalization maps (global mean, total and within-class covariance, residual class-mean) are estimated on the training split only; the test split is never used to fit any transform. The within-class whitening uses the eigenvalue floor of Section 3 with $\varepsilon = 10^{-8}$; on the four backbones studied this floor is inert for BEiT, iBOT, and DINOv2 (their within-class covariances are well conditioned; iBOT has zero negative eigenvalues, and its within-class result is flat to 0.05 points across $\varepsilon \in \{10^{-8}, 10^{-6}, 10^{-4}\}$) and required only for DINO v1, whose within-class covariance has five near-degenerate directions, one slightly negative from floating-point noise. Anisotropy diagnostics (Table 1) are computed on a fixed random sample of $30{,}000$ images from the ungated `clane9/imagenet-100` mirror, identical across backbones. For the ten-benchmark validation (Table 3) we source the standard splits from ungated Hugging Face parquet mirrors, subsampling training sets larger than $20{,}000$ images uniformly at random (SUN397 capped further to $8{,}000/4{,}000$ train/test given its unusually large shards); the identical extraction, normalization, and 60-epoch linear-probe pipeline is applied to DINO v1 and DINOv2 features alike. The mean within-class gain over raw is $+25.6$ points for DINO v1 and $+13.3$ points for DINOv2 across these ten datasets.