A MATTER OF DIRECTION: EIGENFRAME DISTRIBUTION MATCHING FOR SELF-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised objectives seek representations that preserve informative variation while avoiding collapse. Yet a representation can distinguish images while losing variation along some directions, a partial form of collapse. Why might such an objective respond weakly to this loss? We investigate this question in methods that match feature distributions through random one-dimensional projections. Each projection mixes many directions, allowing variation in healthy directions to mask a deficiency in another. In a local Gaussian model, we show that this effect becomes stronger as the representation dimension grows. Averaging more projections reduces sampling noise but does not strengthen the expected response. We call this effect *spectral dilution*. This observation motivates VARTE, a self-supervised objective that replaces random directions with the covariance eigenvectors of features averaged across augmented views. These directions separate the representation's modes of variation, allowing the matching objective to examine each one directly. We analyze this choice theoretically and test its consequences through controlled interventions and image representation learning. On ImageNet-1K, VARTE reaches 74.7% linear-probe accuracy with a ViT-B/16 and transfers competitively across recognition and specialized-domain tasks. Our work highlights a simple principle: preventing collapse depends not only on the distribution we ask features to match, but also on the directions along which we measure them.

1 INTRODUCTION

Self-supervised representation learning often relies on objectives built around multiple augmented views of the same image. These objectives encourage their representations to agree across changes in appearance without collapsing all images to a single feature vector. Yet avoiding this constant solution does not prevent a subtler form of collapse, where a representation can still distinguish images while losing variation along some directions (Jing et al., 2022; Weng et al., 2024). This raises a more specific question: how strongly does such an objective respond when one direction begins to lose variation?

Several approaches address this problem by constraining feature statistics (Ermolov et al., 2021; Zbontar et al., 2021; Bardes et al., 2022). Distribution matching instead asks the features to follow a desired distribution (Balestriero & LeCun, 2025; Zimmermann et al., 2025; Wu et al., 2026). However, because comparing distributions directly in high dimensions can be difficult, a common alternative is to compare their one-dimensional projections (Peyré & Cuturi, 2019; Nadjahi et al., 2020). This makes the comparison tractable, but leaves an important choice of which directions to use when comparing the feature distributions.

To see why, consider the current feature distribution in Figure 1(a), which matches the target variance in every covariance mode except the weak mode u. The matching objective compares current and target distributions through one-dimensional projections along a direction a. When a is random, as in panel (b), its overlap with u can be small, so the weak mode is mixed with healthy variation and the projected distributions remain close. When $a = u$, as in panel (c), this mixing disappears and the full variance mismatch becomes visible.

One might expect that using enough random projections would resolve the problem. However, our analysis of small variance errors in a local Gaussian model shows otherwise. When the covariance

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.

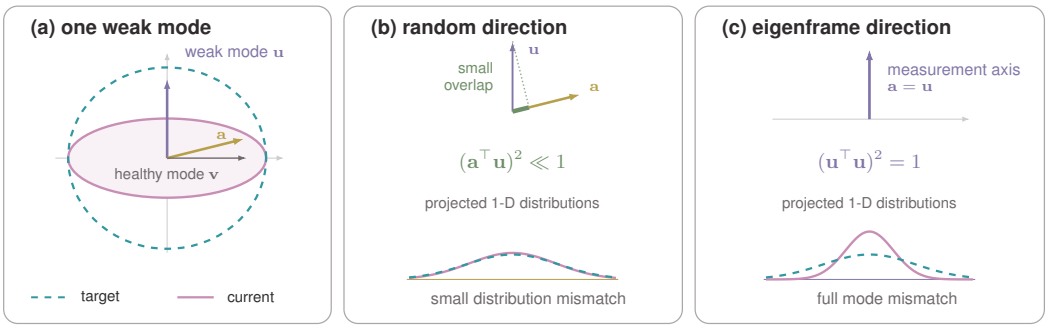

Figure 1: **The intuition behind spectral dilution.** A random direction $\mathbf{a}$ reveals only a fraction of the variance error along the weak mode $\mathbf{u}$, set by their squared alignment $(\mathbf{a}^\top \mathbf{u})^2$. Observing along $\mathbf{u}$ itself reveals the full error.

modes are observed directly, each variance error is measured on its own, whereas random projections mix that error with variation from other modes. Consequently, as dimension grows, this mixing weakens the expected response to the error. Moreover, averaging more random projections reduces sampling noise but does not strengthen that expected response. We call this effect *spectral dilution*.

This suggests a simple alternative: observe each covariance mode separately. The covariance eigenvectors form such an *eigenframe*. We put this idea into practice with View-Averaged Representation Transport in Eigenframes (VARTE), a teacher-free objective that matches view-averaged features to a Gaussian along their covariance eigenvectors and keeps each image's views close to one another. The central change is thus not the target distribution, but the directions along which it is compared with the features. We test whether these directions help detect and correct weakened modes, whether they produce useful representations on ImageNet-1K and downstream tasks, and their time and memory costs.

On ImageNet-1K, VARTE reaches 74.69% linear-probe accuracy with ViT-B/16 after 400 epochs and 74.82% with ViT-L/14 after 100 epochs. The ViT-B model transfers competitively with non-heuristic methods across recognition, specialized domains, and dense prediction. Moreover, matched H100 measurements show similar training-step times at the selected configuration, while using less allocated memory.

In summary, our contributions are:

- A theoretical analysis of *spectral dilution*: why random projections can respond weakly to a loss of variation in one mode, and why averaging more projections does not remove this attenuation.
- VARTE, a teacher-free objective that matches feature distributions along every estimated covariance mode, allowing each mode's variation to be compared directly with its target.
- Empirical evidence that observation directions affect both the response to weakened modes and learned representation quality, supported by controlled experiments and ImageNet-1K evaluations.

2 Related Work

Preventing collapse. Self-supervised methods take several routes to preserving differences between images while making their augmented views agree. Contrastive methods use negative examples to keep representations apart (Chen et al., 2020; Wang & Isola, 2020), whereas BYOL and SimSiam rely on asymmetry between the branches that produce and predict a representation (Grill et al., 2020; Chen & He, 2021). Another route is to constrain the features directly: Barlow Twins controls cross-correlations, VICReg penalizes low variance and correlated coordinates, and W-MSE compares whitened features (Zbontar et al., 2021; Bardes et al., 2022; Ermolov et al., 2021). Yet avoiding a constant output does not by itself ensure that variation is preserved in every direction: features can still concentrate in only a few directions (Balestriero & LeCun, 2022; Weng et al., 2022; He & Ozay, 2022; Jing et al., 2022).

This has motivated objectives that act on the covariance spectrum itself. CorInfoMax discourages vanishing eigenvalues through a log-determinant term, VNE encourages a more even spectrum through entropy, and spectrum modulation controls how eigenvalues decay (Ozsoy et al., 2022; Kim et al., 2023; Weng et al., 2024). Other work investigates how different covariance modes evolve during learning, with IsoLoss addressing unequal convergence rates across eigenmodes (Tian et al., 2021; Halvagal et al., 2023). Our work asks a complementary question: before a weak mode can be corrected, how much of its variance error is visible in the directions used to measure the features?

Matching feature distributions. Rather than constrain individual statistics, distribution-matching methods ask features to follow a chosen reference distribution. Distribution Matching uses the Mallows distance to combine this goal with agreement between augmented views (Jiao et al., 2025). LeJEPA motivates an isotropic Gaussian target and introduces SIGReg, which matches it through random one-dimensional projections, while KerJEPA develops matching objectives based on kernel discrepancies (Balestriero & LeCun, 2025; Zimmermann et al., 2025). VISReg combines coordinate-wise variance regularization with random projections that match the shape of standardized features (Wu et al., 2026). Other methods normalize features onto a sphere and compare them with spherical reference distributions through sliced Wasserstein distances (Bonet et al., 2023; Tran et al., 2024). Within this family, VARTE matches view-averaged features to a Gaussian target along their covariance eigenvectors, jointly measuring differences in center, scale, and shape.

Choosing projection directions. The choice of directions has also received attention in optimal transport. Here, sliced distances make distribution comparison tractable by averaging one-dimensional transport costs over random projections (Peyré & Cuturi, 2019; Nadjahi et al., 2020). Orthogonal sampling can make this estimate less noisy (Rowland et al., 2019), while other approaches seek directions that reveal larger differences. For example, max-sliced and subspace-robust distances optimize a direction or subspace (Deshpande et al., 2019; Paty & Cuturi, 2019); distributional, amortized, adaptive, and random-path methods instead learn or condition how directions are sampled (Nguyen et al., 2021; Nguyen & Ho, 2022; Ohana et al., 2023; Nguyen et al., 2024). Rather than seeking projections with large discrepancies, VARTE uses a complete covariance eigenframe so that modes with little variance are examined alongside those with substantial variance.

3 VARTE: EIGENFRAME DISTRIBUTION MATCHING

VARTE encourages agreement between augmented views while matching the one-dimensional distributions of their average features to a Gaussian reference. For a minibatch of B images, a shared encoder and projector produce D-dimensional features from $V \geq 2$ augmented views. We write $Z = (Z_1, \ldots, Z_V)$, with $Z_v \in \mathbb{R}^{B\times D}$ for $v = 1, \ldots, V$, and denote the feature vector of image b in view v by $z_{vb} \in \mathbb{R}^D$. Both distribution matching and view agreement operate directly on these features, without a teacher network, a separate predictor, or negative pairs.

3.1 CONSTRUCTING THE EIGENFRAME

We first average each image's view features. Let m_b denote the view-mean feature vector of image b and $\bar{m}$ its minibatch average:

$$m_b = \frac{1}{V}\sum_{v=1}^{V} z_{vb}, \qquad \bar{m} = \frac{1}{B}\sum_{b=1}^{B} m_b. \tag{1}$$

To find their directions of variation, we compute their sample covariance and its eigenvectors:

$$\widehat{C}_M = \frac{1}{B-1}\sum_{b=1}^{B}(m_b - \bar{m})(m_b - \bar{m})^\top = U\Lambda U^\top. \tag{2}$$

The columns u_i of U form the covariance eigenframe. We retain all D eigenvectors, including those with small estimated variance, so that each covariance mode is measured separately. The frame is recomputed at each update but held fixed when differentiating the loss. Gradients therefore pass through the features, but not through the eigendecomposition.

3.2 Matching distributions along its directions

We express each image's view mean in the eigenframe, giving feature $x_{bi} = m_b^\top u_i$ along direction u_i. Although these directions are determined from centered features, the original features are the ones projected onto them. This preserves the location of the feature distribution, allowing the matching loss to penalize a nonzero mean as well as differences in scale and shape.

Along each direction, we match the projected features to B fixed reference values approximating a standard Gaussian. Where the values correspond to evenly spaced probability levels:

$$q_b = \Phi^{-1}\left(\frac{b}{B+1}\right), \qquad b = 1, \dots, B. \tag{3}$$

Here Φ is the standard Gaussian CDF, so $\Phi^{-1}(p)$ gives the value with cumulative probability p. Using these fixed quantiles avoids the noise introduced by random target samples.

We sort the projected features and pair them with the reference values in order, that is, smallest with smallest, second smallest with second smallest, and so on. This pairing minimizes the average squared difference, yielding the squared Wasserstein distance between the two empirical distributions (Peyré & Cuturi, 2019). Now let $x_{(b),i}$ denote the b-th smallest feature along u_i. Averaging the resulting squared differences over images and directions, and multiplying by V, gives

$$\mathcal{L}_{\text{shared}} = \frac{V}{BD} \sum_{i=1}^{D} \sum_{b=1}^{B} \left(x_{(b),i} - q_b\right)^2. \tag{4}$$

Sorting makes the comparison independent of sample order, matching corresponding probability levels to capture differences in mean, scale (standard deviation), and distribution shape.

3.3 View agreement and the complete objective

Matching each image's view mean does not ensure agreement among its individual views. We therefore also penalize their squared distance from each image's mean:

$$\mathcal{L}_{\text{residual}} = \frac{1}{(V-1)BD} \sum_{v=1}^{V} \sum_{b=1}^{B} \|z_{vb} - m_b\|_2^2. \tag{5}$$

Combining distribution matching with view agreement gives the complete objective:

$$\boxed{\mathcal{L}_{\text{VARTE}} = \alpha \mathcal{L}_{\text{shared}} + (1-\alpha)\mathcal{L}_{\text{residual}}.} \tag{6}$$

where $\alpha \in (0, 1)$ balances the two terms. Section B.3 explains the normalization of these two terms through a decomposition into view averages and differences.

Given that a minibatch covariance has rank at most $B-1$, a zero empirical eigenvalue need not indicate a missing population mode. Finite-batch and quantile-grid effects are discussed in Section B.1 and fixed-frame gradients are treated in Section B.2.

Computationally, the $D \times D$ eigendecomposition adds small-batch overhead, but its size is independent of B. In matched ViT-B/16 training on 16 H100s at $B = 528$, $D = 512$, VARTE takes 232 ms per step, close to VISReg's 228 ms and LeJEPA's 234 ms, while using 39% less peak allocated memory than LeJEPA (Figure 2). The complete training-step sweeps and measurement protocols are in Section E.7.

4 Spectral dilution and covariance directions

We now ask how the matching loss changes with the choice of observation directions. To separate this question from minibatch sampling, we consider population Wasserstein matching to a standard Gaussian. For a projected distribution P with mean μ, standard deviation σ, and standardized distribution $\widetilde{P}$, the loss decomposes as

$$W_2^2(P, \mathcal{N}(0,1)) = \underbrace{\mu^2}_{\text{mean}} + \underbrace{(\sigma-1)^2}_{\text{scale}} + \underbrace{\sigma W_2^2(\widetilde{P}, \mathcal{N}(0,1))}_{\text{shape}}. \tag{7}$$

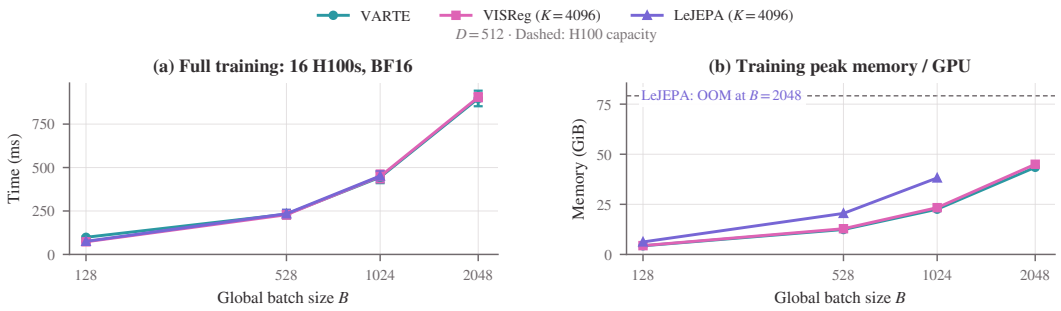

Figure 2: **Matched training cost.** ViT-B/16, 16 H100s, ten views, BF16; $D = 512$, global batch B, and $K = 4096$ for VISReg and LeJEPA. (a) Step time: mean $\pm$ standard deviation of three 100-step window averages. (b) Peak allocated memory per GPU; dashed: capacity of GPU. LeJEPA is OOM at $B = 2048$. Full sweeps and protocols: Section E.7.

Mean, scale, and shape are therefore components of the same matching loss rather than separate penalties. The derivation is given in Section A.1. Because dimensional collapse concerns a loss of variation, we focus on the scale component. Writing the projected variance as $s = \sigma^2$, this contribution is $\phi(s) = (\sqrt{s} - 1)^2$.

To compare observation frames, let $C \in \mathbb{R}^{D\times D}$ be the population covariance of the view means across images and U its eigenframe. For any orthogonal frame $R = [r_1, \ldots, r_D]$, with $R^\top R = I_D$, the variance along r_j is $r_j^\top C r_j$. Averaging the scale component over the frame gives

$$\mathsf{S}_R(C) = \frac{1}{D}\sum_{j=1}^{D} \phi(r_j^\top C r_j). \tag{8}$$

For $a \in \mathbb{R}^D$ uniform on the unit sphere, the random-direction counterpart is $\mathsf{S}_{\text{slice}}(C) = \mathbb{E}_a \phi(a^\top C a)$.

4.1 Spectral dilution of a weak mode

Suppose one direction u has too little variance, while every orthogonal direction has the target variance of one. This is the setting illustrated in Figure 1. Writing δ for the variance error along the unit vector u, the covariance is

$$C_\delta = I + \delta u u^\top, \qquad \|u\|_2 = 1, \quad \delta > -1. \tag{9}$$

Projecting variation along u onto another unit direction a multiplies distances from the mean by their alignment, $a^\top u$. Because variance averages squared differences, the projected variance error is multiplied by $(a^\top u)^2$. Thus the variance along a is $1 + \delta(a^\top u)^2$, while observing along u retains the full error δ. Near the target variance, their average scale losses satisfy:

Theorem 1 (Local spectral dilution). *For fixed $D \geq 2$,*

$$\lim_{\delta\to 0} \frac{\mathsf{S}_{\text{slice}}(C_\delta)}{\mathsf{S}_U(C_\delta)} = \frac{3}{D+2}. \tag{10}$$

Random projections retain a smaller fraction of the eigenframe's scale loss as dimension grows. We call this effect *spectral dilution*, since mixing a weak mode with healthy modes weakens its contribution to the objective. Increasing the number of independent and identically distributed (i.i.d.) projections reduces sampling variance without changing the expected loss, and a uniformly random orthogonal frame has the same expectation. Thus, neither more samples nor orthogonality alone removes the attenuation. The proof and expectation arguments are given in Section A.2.

4.2 Eigenframes expose each mode directly

The advantage of covariance directions extends beyond a small error in one mode. For any covariance matrix, the eigenframe maximizes the average scale loss among complete orthogonal frames:

Theorem 2 (Eigenframes maximize the scale-matching loss). *Let $C \succeq 0$ have eigenvalues $\lambda_1, \ldots, \lambda_D$ and eigenframe U. For every orthogonal frame R,*

$$\mathsf{S}_R(C) \leq \mathsf{S}_U(C) = \frac{1}{D}\sum_{i=1}^{D}(\sqrt{\lambda_i} - 1)^2. \tag{11}$$

A direction that mixes covariance modes observes a weighted average of their variances, allowing deviations below and above the target to partially offset before the loss is evaluated. Since ϕ is convex, applying it after this averaging cannot exceed the corresponding weighted average of the mode-wise losses. Across an orthogonal frame, each eigenmode receives the same total weight, so the eigenframe attains the maximum by measuring the modes separately (proof in Section A.3).

The same result gives $\mathsf{S}_{\text{slice}}(C) \leq \mathsf{S}_U(C)$ for any covariance matrix. Thus, uniformly sampled directions have a weaker expected scale response than observing the covariance modes directly. Increasing the number of directions estimates this response more precisely without changing its expectation. The argument is given in Section A.3.

This difference is particularly relevant to dimensional collapse, where a representation loses variation along one or more modes. Observing a weakened mode's eigenvector exposes the deficiency directly, while random directions mix it with healthy variation and can attenuate its contribution to the loss. When the population covariance becomes rank-deficient, this direct exposure yields a rank-dependent lower bound on the matching loss, derived in Section A.4. Increasing the loss weight cannot generally recover the difference because random projections respond differently to different patterns of variance error, as analyzed in Section A.5. In Section 5.1, we test these theoretical statements through loss attenuation, gradient rescaling, and controlled repair of weakened modes.

5 Experimental evaluation

We first test the theoretical predictions on controlled variance defects, then ask whether covariance-aligned directions improve representation learning. Experiments cover ImageNette controls, ImageNet-1K pretraining with ViT-B/16 and ViT-L/14, frozen probing, transfer learning, and fine-tuning. ImageNette runs use ViT-S/8 for 800 epochs with four global views, while ImageNet-1K uses four global and six local views. Training and evaluation protocols are in Table S8 and Section E.3.

5.1 Eigenframe directions expose weak modes

Before asking whether observation directions produce better corrections, we first ask how strongly they expose a variance error. We therefore test whether the attenuation predicted by Theorem 1 persists beyond the local regime as it is depicted in Figure 3(a). Consistent with the theory, the random-direction/eigenframe scale-loss ratio decreases with dimension across all tested contractions. However, one might argue that this weaker response could simply reflect a difference in magnitude, in which case a larger loss weight might recover the eigenframe response. To test this, we fit one shared multiplier to best match the eigenframe gradients across five variance-error patterns, including contractions and expansions of individual modes or groups of modes. Because a scalar rescales every gradient component equally, it cannot change their relative strengths. Even after this adjustment, the residual mismatch remains approximately 97% of the eigenframe gradient norm at $D = 128$ (Figure 3(b)). Thus no single loss weight recovers the eigenframe's correction pattern across these errors. Moreover, increasing the number of random directions reduces sampling noise without recovering the mode-specific response.

To test whether this difference persists in learned representations, we first rescale frozen VARTE and VISReg projector features so that each covariance mode has near-target variance. We then insert a 50% contraction into five eigenmodes. During repair, we subtract the pre-defect gradient and use normalized gradients with a common maximum step length. VARTE reaches 90% repair in 19 updates in all ten source–mode cases (Figure 3(c)), showing that the more targeted correction persists across feature sources. We then test whether this advantage carries over to representation learning in Section 5.2. Further analysis of the mechanism is in Section C.

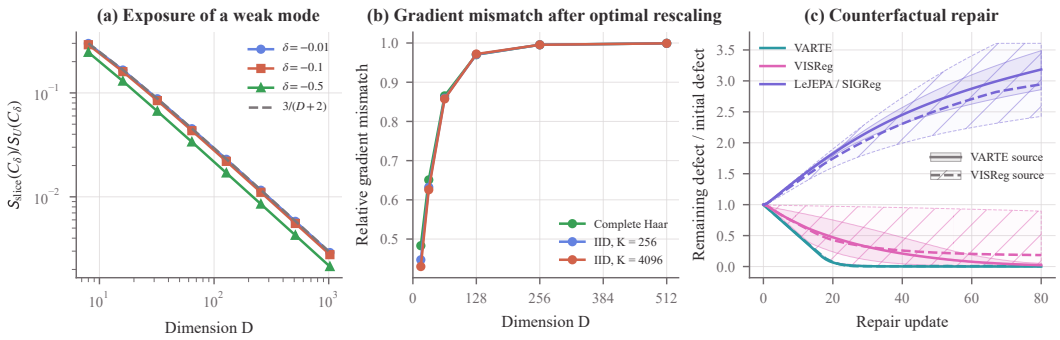

Figure 3: **Eigenframe directions expose and repair variance defects.** (a) Expected random-direction/eigenframe scale-loss ratio for variance $1 + \delta$. The ratio falls with dimension, including for finite contractions beyond the local regime of Theorem 1. (b) Relative gradient mismatch after fitting a single shared multiplier jointly across five variance-error patterns. Zero denotes exact recovery. Gradients are w.r.t. Gaussian log standard deviations, not encoder weights. (c) Remaining defect under normalized-gradient updates with a common step-length cap after subtracting each objective's pre-defect gradient. Solid/dashed curves denote VARTE/VISReg feature sources. Shading gives min/max over five modes. VISReg and SIGReg use $K = 4096$.

5.2 Observation directions during self-supervised learning

We now ask whether observation directions affect representation learning. Keeping VARTE's objective fixed, we replace only its directions. The i.i.d. variant samples $K = 4096$ independent uniform directions, while the Haar variant samples a fresh uniformly random orthonormal frame at each batch. The native variant instead uses the projector's original coordinate axes. As shown in Figure 4(a), the eigenframe directions outperform all three alternatives. Thus neither more random directions nor orthogonality alone closes the gap.

We then test the dimensional trend predicted by spectral dilution. At fixed batch size, the accuracy gap between the eigenframe and the other directions grows, in general, with projector dimension (Figure 4(b)), consistent with weaker random-direction exposure in higher dimensions. Fixed Haar and native-coordinate frames are also substantially worse, showing that orthogonality alone does not recover adaptive spectral exposure. Finally, because our theory isolates the scale component, we also ask how much learning depends on matching distributional shape. Replacing full matching with its exact mean-and-scale component yields 90.34% frozen accuracy, compared with 91.15% for VARTE. . Further design ablations are reported in Section D.2.

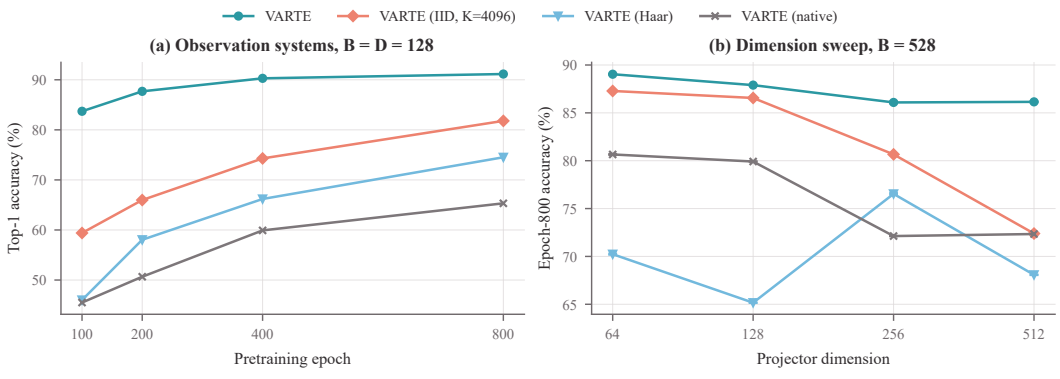

Figure 4: **Eigenframe directions support representation learning.** ImageNette frozen accuracy. (a) Trajectories at $B = D = 128$. (b) Dimension sweep at $B = 528$. The i.i.d. $K = 4096$, Haar, and native-coordinate variants change only VARTE's observation directions. Eigenframe and Haar use D directions.

5.3 IMAGENET-1K AND TRANSFER

We next scale VARTE from the ImageNette controls to ImageNet-1K classification and other downstream transfer tasks. We ablate the loss weight, learning rate, batch size, and projector dimension (Tables S5 and S6), selecting batch size 528, projector dimension 512, loss weight 0.05, and peak learning rate 3×10^{-4}. Complete training and evaluation settings are given in Section E.

At the selected projector dimension of 512, the ViT-L/14 runs provide the most direct large-scale comparison with VISReg. Both train for 100 epochs, with VISReg using 512 random directions and VARTE using the 512 covariance eigenvectors. Under this matched setting, VARTE outperforms VISReg on ImageNet-1K and recognition transfer while maintaining comparable specialized-domain performance (Tables 1 and 2). Beyond this matched comparison, VARTE with ViT-B/16 remains close to VISReg across both transfer suites, although DINO and iBOT retain higher recognition averages.

Table 1: **ImageNet-1K and downstream linear probing.** Top-1 Accuracy (%); Avg.: seven downstream tasks. Downstream reruns use 10-epoch probes; LeJEPA uses published 100-epoch probes. †: published (Wu et al., 2026; Balestriero & LeCun, 2025); **bold**/underline: best/second best.

Methods	Backbone	Epochs	Aircraft	Cars	CIFAR-10	CIFAR-100	Flowers	Food-101	Pets	Avg.	ImageNet-1K
MoCo v3	ViT-B/16	300	57.79	67.63	96.94	85.21	91.54	81.81	89.75	81.52	76.7†
DINO	ViT-B/16	400	63.64	73.76	96.52	84.95	**94.58**	83.07	93.57	84.30	78.2†
iBOT	ViT-B/16	400	63.55	72.43	97.06	85.88	93.87	84.13	93.43	84.34	79.8†
VISReg (K=4096)	ViT-B/16	400	57.07	64.78	94.64	78.84	90.42	82.84	88.20	79.54	75.7†
VARTE	ViT-B/16	400	55.42	63.23	94.75	80.86	88.55	81.35	88.28	78.92	74.69
MAE	ViT-L/16	1600	62.26	61.35	93.40	78.01	85.41	78.61	91.31	78.62	75.1†
iBOT	ViT-L/16	250	**66.07**	**76.16**	**97.56**	**87.26**	94.05	**86.19**	**93.98**	**85.89**	**81.0**†
LeJEPA†	ViT-L/14	100	57.0	57.3	96.5	83.7	91.2	82.1	89.7	79.6	75.6†
VISReg (K=4096)	ViT-L/14	400	56.65	66.17	94.13	71.85	90.13	83.35	89.15	78.78	77.0†
VISReg (K=512)	ViT-L/14	100	50.50	55.53	93.48	80.22	87.74	79.50	83.54	75.79	72.90
VARTE	ViT-L/14	100	52.87	59.13	94.80	81.34	86.14	81.18	87.54	77.57	74.82
I-JEPA	ViT-H/14	300	55.33	59.12	97.36	85.82	86.78	83.31	92.75	80.07	79.3†

Table 2: **Specialized-domain linear probing.** Top-1 Accuracy (%) except ChestXRay, which reports exact-match multilabel accuracy; Avg.: six tasks with the same probe protocol of Table 1.

Methods	Backbone	Epochs	DTD	Galaxy10	AID	ChestXRay	Retina.	OrganA.	Avg.
MoCo v3	ViT-B/16	300	73.70	73.07	90.19	37.21	64.49	91.37	71.67
DINO	ViT-B/16	400	74.15	72.66	**91.51**	37.71	62.75	91.64	71.74
iBOT	ViT-B/16	400	74.57	72.44	91.40	37.35	62.75	91.98	71.75
VISReg (K=4096)	ViT-B/16	400	75.69	74.07	90.90	37.50	62.25	**93.42**	72.30
VARTE	ViT-B/16	400	73.67	70.80	89.07	37.58	64.75	91.51	71.23
MAE	ViT-L/16	1600	72.77	71.70	86.43	37.80	63.00	90.03	70.29
iBOT	ViT-L/16	250	75.21	72.66	91.24	37.78	63.25	90.68	71.80
VISReg (K=4096)	ViT-L/14	400	**76.38**	**75.70**	90.07	**38.05**	64.25	92.99	**72.91**
VISReg (K=512)	ViT-L/14	100	73.35	72.83	88.30	37.57	61.50	92.22	70.96
VARTE	ViT-L/14	100	72.45	71.03	88.03	37.86	64.75	91.51	70.94
I-JEPA	ViT-H/14	300	70.00	70.91	88.74	37.60	**67.00**	92.72	71.16

Beyond classification transfer, the frozen backbone retains spatial information that supports linear segmentation on ADE20K, with performance close to published VISReg (Table 4). To test whether these representations also provide a useful starting point for adaptation, we then fine-tune the encoder. VARTE remains competitive across downstream tasks, including 82.0% on ImageNet-1K (Table 3). Together, these results support both frozen reuse and task-specific adaptation.

Spatial organization without dense supervision. The patch maps provide a qualitative view of the spatial relationships retained in these representations (Figure 5). The hand query highlights the other hand and face, while the keyboard query responds to characters on other keys. These examples complement the segmentation results by illustrating spatial structure learned without dense supervision. Additional examples and the visualization protocol are in Section E.6.

Table 3: **Full fine-tuning.** ViT-B/16 top-1 accuracy (%) with different pretrainings. †: published in (Wu et al., 2026). Bold: best, including ties.

Methods	CIFAR-10	CIFAR-100	Flowers	Galaxy10	ImageNet-1K
Supervised†	99.0	89.5	98.5	—	81.5
DINO†	99.1	91.7	98.8	86.6	82.8
VISReg†	99.2	**91.8**	**99.0**	**87.0**	**83.0**
VARTE	**99.3**	**91.8**	98.9	86.7	82.0

Table 4: **ADE20K linear segmentation.** ImageNet-pretrained ViT-B/16; mIoU on 40 decoder epochs.

Methods	Epochs	mIoU
MoCo v3†	300	**31.69**
DINO†	400	29.40
MAE†	1600	23.60
VISReg†	400	30.16
VARTE	400	29.92

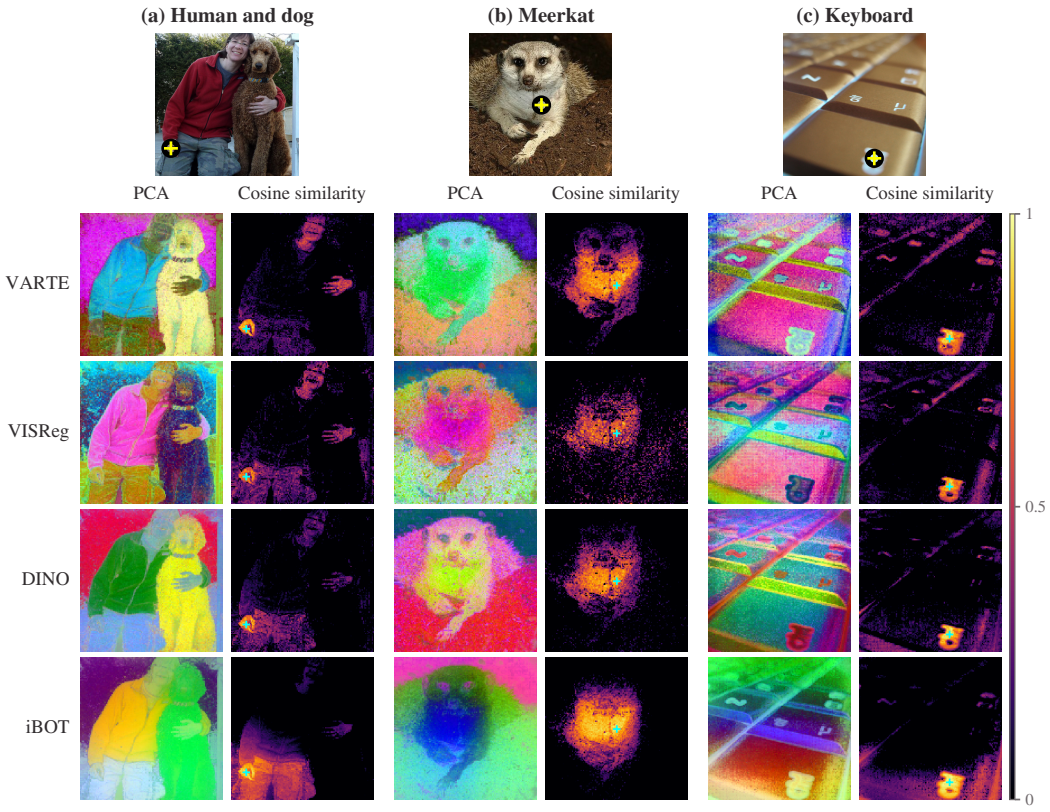

Figure 5: **Spatially structured patch representations.** Frozen 400-epoch ViT-B/16 backbones use identical 2048-pixel inputs and 128×128 grids. Rows show PCA and shared-query cosine maps; PCA is fitted independently and cosine maps use Equation (34) normalization.

6 Conclusion

We studied dimensional collapse through the directions used to observe feature distributions. Our analysis identifies spectral dilution, where random projections can obscure a weakening covariance mode. Increasing their number reduces sampling noise but does not strengthen the expected response. VARTE addresses this weakness by matching distributions along covariance eigenvectors, yielding more targeted correction in controlled experiments and competitive ImageNet-1K and downstream transfer. At the selected configuration, it also uses less peak allocated memory than VISReg and LeJEPA with similar training-step times. Our experiments also show that preserving variation alone does not ensure that the retained variation is useful downstream. Future work should therefore combine observation geometry with learning signals that organize this preserved variation into transferable information. Overall, VARTE shows that the directions used to measure a feature distribution matter for how its variation is exposed and preserved, while the learning task determines how useful that variation becomes.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery, implementation and mathematical proof development. It was also used to polish the writing. The authors reviewed all AI-assisted work and take responsibility for the final content of this work, including any text, claims, code produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The method is specified in Section 3, with theoretical assumptions and claims in Section 4 and proofs in Section A. Appendices B–C detail implementation qualifications and controlled diagnostics. Appendices D–E document hyperparameter ablations, training settings, data processing, evaluation protocols, and computational measurements. We plan to release the code and checkpoints publicly.

REFERENCES

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2205.11508.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. URL https://arxiv.org/abs/2511.08544.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2105.04906.

Clément Bonet, Paul Berg, Nicolas Courty, François Septier, Lucas Drumetz, and Minh-Tan Pham. Spherical sliced-wasserstein. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jXQ0ipgMdU.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. URL https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021. URL https://arxiv.org/abs/2011.10566.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. URL https://www.robots.ox.ac.uk/~vgg/data/dtd/.

Harald Cramér and Herman Wold. Some theorems on distribution functions. *Journal of the London Mathematical Society*, s1-11(4):290–294, 1936. doi: 10.1112/jlms/s1-11.4.290. URL https://doi.org/10.1112/jlms/s1-11.4.290.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. URL https://image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G. Schwing. Max-sliced wasserstein distance and its use for gans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/

Deshpande_Max-Sliced_Wasserstein_Distance_and_Its_Use_for_GANs_CVPR_2019_paper.html.

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. URL https://arxiv.org/abs/2010.11929.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, 2021. URL https://arxiv.org/abs/2007.06346.

fastai. ImageNette, 2019. URL https://github.com/fastai/imagenette.

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, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.07733.

Manu Srinath Halvagal, Axel Laborieux, and Friedemann Zenke. Implicit variance regularization in non-contrastive self-supervised learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c837ab3eebe77bffac634939f22ac458-Abstract-Conference.html.

Bobby He and Mete Ozay. Exploring the gap between collapsed and whitened features in self-supervised learning. In *International Conference on Machine Learning*, pp. 8613–8634, 2022. URL https://proceedings.mlr.press/v162/he22c.html.

Yuling Jiao, Wensen Ma, Defeng Sun, Hansheng Wang, and Yang Wang. Distribution matching for self-supervised transfer learning. *arXiv preprint arXiv:2502.14424*, 2025. URL https://arxiv.org/abs/2502.14424.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Jaeill Kim, Suhyun Kang, Duhun Hwang, Jungwook Shin, and Wonjong Rhee. VNE: An effective method for improving deep representation by manipulating eigenvalue distribution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3799–3810, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Kim_VNE_An_Effective_Method_for_Improving_Deep_Representation_by_Manipulating_CVPR_2023_paper.html.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. URL https://openaccess.thecvf.com/content_iccv_workshops_2013/W19/html/Krause_3D_Object_Representations_2013_ICCV_paper.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

W. Henry Leung and Jo Bovy. Galaxy10 DECaLS. Zenodo dataset, 2024. URL https://doi.org/10.5281/zenodo.10845026.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. URL https://arxiv.org/abs/1306.5151.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/eefc9e10ebdc4a2333b42b2dbb8f27b6-Abstract.html.

Khai Nguyen and Nhat Ho. Amortized projection optimization for sliced wasserstein generative models. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=2dxsDFaESK.

Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QYjO70ACDK.

Khai Nguyen, Shujian Zhang, Tam Le, and Nhat Ho. Sliced wasserstein with random-path projecting directions. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2401.15889.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. URL https://www.robots.ox.ac.uk/~vgg/data/flowers/102/.

Ruben Ohana, Kimia Nadjahi, Alain Rakotomamonjy, and Liva Ralaivola. Shedding a pac-bayesian light on adaptive sliced-wasserstein distances. In *International Conference on Machine Learning*, 2023. URL https://openreview.net/forum?id=bKxkWVTrxx.

Serdar Ozsoy, Shadi Hamdan, Sercan O. Arik, Deniz Yuret, and Alper T. Erdogan. Self-supervised learning with an information maximization criterion. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/e4cd50120b6d7e8daff1749d6bbaa889-Abstract-Conference.html.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. URL https://www.robots.ox.ac.uk/~vgg/data/pets/.

François-Pierre Paty and Marco Cuturi. Subspace robust wasserstein distances. In *International Conference on Machine Learning*, pp. 5072–5081, 2019. URL https://proceedings.mlr.press/v97/paty19a.html.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. Foundations and Trends in Machine Learning. Now Publishers, 2019. URL https://arxiv.org/abs/1803.00567.

Mark Rowland, Jiri Hron, Yunhao Tang, Krzysztof Choromanski, Tamas Sarlos, and Adrian Weller. Orthogonal estimation of wasserstein distances. In *International Conference on Artificial Intelligence and Statistics*, pp. 186–195, 2019. URL https://proceedings.mlr.press/v89/rowland19a.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. URL https://arxiv.org/abs/1409.0575.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, 2021. URL https://proceedings.mlr.press/v139/tian21a.html.

Huy Tran, Yikun Bai, Abihith Kothapalli, Ashkan Shahbazi, Xinran Liu, Rocio Diaz Martin, and Soheil Kolouri. Stereographic spherical sliced wasserstein distances. In *International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=vLtVGtEz5h.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/2005.10242.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://arxiv.org/abs/1705.02315.

Xi Weng, Lei Huang, Lei Zhao, Rao Anwer, Salman H. Khan, and Fahad Shahbaz Khan. An investigation into whitening loss for self-supervised learning. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c057cb81b8d3c67093427bf1c16a4e9f-Abstract-Conference.html.

Xi Weng, Yunhao Ni, Tengwei Song, Jie Luo, Rao Anwer, Salman Khan, Fahad Khan, and Lei Huang. Modulate your spectrum in self-supervised learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/03474669b759f6d38cdca6fb4eb905f4-Abstract-Conference.html.

Haiyu Wu, Randall Balestriero, and Morgan Levine. Visreg: Variance-invariance-sketching regularization for jepa training. *arXiv preprint arXiv:2606.02572*, 2026. URL https://arxiv.org/abs/2606.02572.

Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. doi: 10.1109/TGRS.2017.2685945. URL https://arxiv.org/abs/1608.05167.

Jiancheng Yang, Rui Shi, and Bingbing Ni. MedMNIST classification decathlon: A lightweight AutoML benchmark for medical image analysis. In *IEEE International Symposium on Biomedical Imaging*, pp. 191–195, 2021. URL https://medmnist.com/.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2—a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1):41, 2023. URL https://arxiv.org/abs/2110.14795.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, 2021. URL https://arxiv.org/abs/2103.03230.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. URL https://ade20k.csail.mit.edu/.

Eric Zimmermann, Harley Wiltzer, Justin Szeto, David Alvarez-Melis, and Lester Mackey. Kerjepa: Kernel discrepancies for euclidean self-supervised learning. *arXiv preprint arXiv:2512.19605*, 2025. URL https://arxiv.org/abs/2512.19605.