

Spectral Sufficiency: Label-Free Estimation of Discriminative Efficiency from Pretrained Feature Geometry

Anonymous Author(s)

March 27, 2026

Abstract

Assessing the quality of pretrained feature representations for downstream tasks typically requires labeled data from the target domain. Recent work has shown that *capacity* (measured by effective rank $\hat{r}_{\text{eff}}$) alone is insufficient for prediction—datasets like SVHN exhibit high capacity yet low *discriminative efficiency* (between-class spectral energy ratio, $\text{BSER} = 0.057$), where over 94% of spectral energy is wasted on within-class noise. However, computing BSER requires class labels, which are precisely what is unavailable in transfer learning scenarios.

We investigate the **Spectral Sufficiency** conjecture: whether the singular value distribution of a pretrained feature matrix, without any labels, contains sufficient information to predict discriminative efficiency. We propose five candidate label-free proxies derived from spectral shape statistics and identify *SVD channel coupling*—a nonlinear combination of effective rank, uniform information density, and spectral alignment—as the strongest predictor, achieving Spearman $\rho = 0.969$ with BSER across 30 dataset–encoder pairs spanning three encoder families (supervised, contrastive, self-supervised) and ten datasets. Using this proxy, we construct a fully label-free capacity–efficiency quadrant diagnostic that achieves 93.3% agreement with the oracle (label-based) diagnosis, correctly identifies pathological cases such as SVHN, and guides geometric engineering interventions that recover 115.4% of oracle-guided accuracy improvements. Our framework unifies prior spectral quality metrics into a single label-free diagnostic pipeline for pretrained representations.

1 Introduction

Transfer learning from pretrained encoders has become the dominant paradigm in computer vision and natural language processing. A critical yet underexplored question is: *given a pretrained encoder and an unlabeled target dataset, how can we assess the quality of the extracted features without any target labels?* This question is of both theoretical and practical importance—it determines whether we can select encoders, curate datasets, and diagnose feature pathologies before investing in expensive annotation campaigns.

Capacity is necessary but not sufficient. Prior work on label-free feature assessment has focused primarily on *capacity*—the intrinsic dimensionality of the feature space, measured by the effective rank $\hat{r}_{\text{eff}}$ (Roy and Vetterli, 2007). The Spectral Diversity Gauge (SDG) framework (Anonymous, 2024c) demonstrated that a composite score combining $\hat{r}_{\text{eff}}$, spectral alignment, and uniform information density can rank pretrained models without labels. However, capacity alone cannot distinguish features that are inherently well-organized for classification from those that are not.

A striking illustration is SVHN (Netzer et al., 2011): when features are extracted using a ResNet-50 encoder, the effective rank $\hat{r}_{\text{eff}} = 0.408$ suggests adequate capacity, yet the between-class spectral energy ratio $\text{BSER} = 0.050$ reveals that 95% of spectral energy resides in within-class noise. In

contrast, STL-10 with similar capacity ($\hat{r}_{\text{eff}} = 0.404$) achieves $\text{BSER} = 0.792$ —the spectral energy is well-aligned with class structure. This distinction is invisible to any purely capacity-based metric.

The label bottleneck in efficiency estimation. Recent work introduced the capacity–efficiency diagnostic framework (Anonymous, 2024b), where $\text{BSER} = \text{tr}(\Sigma_B)/\text{tr}(\Sigma_T)$ measures discriminative efficiency. The two-dimensional $(\hat{r}_{\text{eff}}, \text{BSER})$ diagnostic improved partial correlation with downstream accuracy from $r = 0.514$ (capacity only) to $r = 0.871$ (capacity + efficiency). However, computing BSER requires the between-class scatter matrix Σ_B , which depends on class labels—precisely the information unavailable in the target domain during transfer learning.

This creates a fundamental tension: the most informative diagnostic requires the very labels we lack.

This work: Spectral Sufficiency. We ask whether the unsupervised spectral geometry of a pre-trained feature matrix contains sufficient information to infer discriminative efficiency *without labels*. Our key insight is that pretrained encoders have already learned class structure from their training data, and this prior knowledge is encoded in the *shape* of the singular value distribution—not just its effective dimensionality.

Specifically, features approaching the Neural Collapse (NC1) limit (Papayan et al., 2020) exhibit flat spectra ($\kappa \rightarrow 1$, $r_{\text{eff}} \rightarrow K$) and high BSER ($\rightarrow 1$), while features with large within-class noise show heavy-tailed spectra and low BSER. This spectral shape–efficiency connection motivates our search for label-free proxies.

Contributions. We make four contributions:

1. **Spectral Sufficiency Conjecture and Theory.** We formalize the hypothesis that label-free spectral statistics can predict discriminative efficiency, and derive a lower bound on BSER as a function of spectral flatness under an isotropic within-class noise model (Section 3.4).
2. **SVD Channel Coupling Proxy.** Among five candidate proxies, we identify SVD channel coupling as the strongest label-free predictor of BSER, achieving Spearman $\rho = 0.969 \pm 0.007$ across 30 dataset–encoder pairs (Section 4.2).
3. **Label-Free Quadrant Diagnosis.** We construct a fully label-free capacity–efficiency diagnostic achieving 93.3% quadrant agreement with the oracle, enabling correct identification of pathological cases (e.g., SVHN as high-capacity/low-efficiency) and guiding geometric engineering with 87.0% action agreement and 115.4% recovery ratio (Sections 4.3–4.4).
4. **Unified Framework.** We extend the Collapse Model for dataset merging to jointly predict post-merge efficiency ($R^2 = 0.799$), enabling Pareto-optimal pool selection along both capacity and efficiency dimensions (Section 4.5).

2 Related Work

Transferability estimation. A rich line of work estimates how well pretrained features transfer to downstream tasks. Label-requiring methods include LEEP (Nguyen et al., 2020), which computes a log-expected empirical predictor from source–target label co-occurrences, and H-Score (Bao et al., 2019), which measures feature–label mutual information via inter- and intra-class scatter. LogME (You et al., 2021) marginalizes over a Bayesian linear classifier to produce a label-dependent transferability score. GBC (Pandy et al., 2022) uses Gaussian Bhattacharyya class separability. These methods all require target labels or label proxies, limiting applicability in purely unsupervised transfer settings.

Label-free approaches are comparatively rare. The Spectral Diversity Gauge (SDG) (Anonymous, 2024c) proposed a composite score based on effective rank, spectral alignment, and uniform information density computed from the SVD of the feature matrix alone. However, SDG captures only the *capacity* dimension and cannot distinguish high-capacity/low-efficiency scenarios. Our work extends this line by adding a label-free efficiency dimension.

Spectral analysis of neural representations. The singular value distribution of feature matrices has been extensively studied as a diagnostic tool. Effective rank (Roy and Vetterli, 2007) measures the intrinsic dimensionality via the entropy of the normalized singular value spectrum. Papayan et al. (2020) showed that during training, representations converge toward Neural Collapse (NC), where class means become equiangular and within-class variation vanishes—a regime where $r_{\text{eff}} \rightarrow K$ and BSER $\rightarrow 1$. The capacity–efficiency diagnostic framework (Anonymous, 2024b) exploited this connection by jointly measuring $\hat{r}_{\text{eff}}$ (capacity) and BSER (efficiency), but required labels for the latter. The Spectral Collapse Model (Anonymous, 2024a) predicted post-merge effective rank from source and target spectral properties, but only along the capacity axis. Our work bridges these efforts by showing that spectral shape statistics can substitute for the label-requiring BSER.

Dataset quality and curation. Dataset quality assessment has gained attention as practitioners recognize that data quality often matters more than model architecture. Approaches range from data-centric AI principles (Zha et al., 2023) to automated curation methods such as DISCO (Anonymous, 2025), which uses diffusion-based objectives for quality scoring. Principal pruning (Anonymous, 2023b) and class-aware sampling (Anonymous, 2023a) provide geometric engineering operations that modify feature distributions. Our contribution is showing that the *choice* of which operation to apply—previously guided by label-dependent diagnostics—can be made purely from spectral geometry.

3 Spectral Sufficiency Framework

3.1 Preliminaries: Capacity and Efficiency

Let $H \in \mathbb{R}^{N \times d}$ be the feature matrix extracted by a pretrained encoder f_θ from N samples, with singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k$ (where $k = \min(N, d)$) and normalized spectral weights $p_i = \sigma_i^2 / \sum_j \sigma_j^2$.

Capacity. The normalized effective rank measures the intrinsic dimensionality of the feature space:

$$\hat{r}_{\text{eff}} = \frac{\exp(-\sum_i p_i \log p_i)}{k} \in (0, 1], \quad (1)$$

where values near 1 indicate a flat spectrum (high capacity) and values near 0 indicate a peaked spectrum (low capacity).

Efficiency. Given class labels $\{y_i\}_{i=1}^N$ with K classes, the between-class spectral energy ratio (BSER) measures how well the spectral energy is aligned with class structure:

$$\text{BSER} = \frac{\text{tr}(\Sigma_B)}{\text{tr}(\Sigma_T)} = 1 - \frac{\text{tr}(\Sigma_W)}{\text{tr}(\Sigma_T)}, \quad (2)$$

where Σ_T , Σ_B , Σ_W are the total, between-class, and within-class scatter matrices. $\text{BSER} \in [0, 1]$, with values near 1 indicating that spectral energy is concentrated in discriminative (between-class) directions.

The diagnostic gap. While $\hat{r}_{\text{eff}}$ is computable without labels, BSER requires class assignments through Σ_B . The capacity–efficiency diagnostic (Anonymous, 2024b) showed that the two-dimensional $(\hat{r}_{\text{eff}}, \text{BSER})$ space is far more informative than capacity alone—but the efficiency axis is label-dependent.

3.2 The Spectral Sufficiency Conjecture

Conjecture 1 (Spectral Sufficiency). *For feature matrices $H \in \mathbb{R}^{N \times d}$ extracted by pretrained encoders, there exists a statistic $\eta(H)$ computed solely from the singular value distribution of H (without labels) such that*

$$\rho_S(\eta(H), \text{BSER}(H, y)) \geq 0.7, \quad (3)$$

where ρ_S denotes Spearman rank correlation, evaluated across datasets and encoders.

The conjecture is non-trivial because BSER explicitly depends on the class structure through Σ_B , while η has no access to labels. However, pretrained encoders have already learned discriminative structure from their training data. Our hypothesis is that this learned structure manifests in the *shape* of the singular value distribution—not just its entropy (effective rank), but also its tail behavior, condition number, and higher-order moments.

3.3 Candidate Label-Free Efficiency Proxies

We propose five candidate proxies, each capturing a different aspect of spectral shape:

Spectral Flatness (SF). The ratio of effective rank to the matrix rank:

$$\text{SF} = \frac{r_{\text{eff}}}{\text{rank}(H)}. \quad (4)$$

Under Neural Collapse (NC1), $\text{SF} \rightarrow K/d$ and $\text{BSER} \rightarrow 1$. In the degenerate case where all energy concentrates on one direction, $\text{SF} \rightarrow 1/\text{rank}(H)$ and $\text{BSER} \rightarrow 1/K$.

Normalized spectral condition number ($\hat{\kappa}$). The inverse condition number of the feature matrix:

$$\hat{\kappa} = \sigma_k / \sigma_1. \quad (5)$$

A flat spectrum implies $\hat{\kappa} \rightarrow 1$, while a heavy-tailed spectrum yields $\hat{\kappa} \rightarrow 0$.

Spectral Gini coefficient. Measures inequality in the spectral weight distribution:

$$\text{Gini}(\{p_i\}) = \frac{\sum_{i=1}^k \sum_{j=1}^k |p_i - p_j|}{2k \sum_i p_i}. \quad (6)$$

Values near 0 indicate uniform energy distribution; values near 1 indicate extreme concentration.

UID ratio. The ratio of uniform information density (geometric dimension) to effective rank:

$$\text{UID_ratio} = \frac{\text{UID}}{\hat{r}_{\text{eff}}}, \quad (7)$$

where UID is defined via the participation ratio of the singular value spectrum. This captures the discrepancy between geometric and spectral measures of dimensionality.

SVD channel coupling (η_{SVD}). A nonlinear combination of effective rank, UID, and spectral alignment:

$$\eta_{\text{SVD}} = f(\hat{r}_{\text{eff}}, \text{UID}, \text{SA}_k), \quad (8)$$

where SA_k denotes spectral alignment (the fraction of variance explained by the top- k components for various k). This proxy captures the coupling between the three spectral channels identified by the SDG framework—a feature matrix where all three channels agree (high rank, high UID, high alignment) is likely to be both high-capacity and high-efficiency.

3.4 Theoretical Analysis

We analyze the relationship between spectral flatness and BSER under a tractable generative model.

Assumption 1 (Isotropic within-class noise). *Each feature vector is generated as $h_i = \mu_{y_i} + \varepsilon_i$, where $\mu_c \in \mathbb{R}^d$ are class centroids with $\|\mu_c - \mu_{c'}\| \geq \Delta$ for $c \neq c'$, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_d)$.*

Theorem 1 (Spectral flatness lower bound on BSER). *Under Assumption 1 with K balanced classes and signal-to-noise ratio $\text{SNR} = \Delta^2/(d\sigma^2)$, we have*

$$\text{BSER} \geq g(\text{SF}, K, \text{SNR}), \quad (9)$$

where g is monotonically increasing in SF when SNR is held constant.

Proof sketch. Under the isotropic noise model, $\text{tr}(\Sigma_W) = N\sigma^2 d$ and $\text{tr}(\Sigma_T) = \text{tr}(\Sigma_B) + N\sigma^2 d$. The spectral flatness SF is a decreasing function of σ^2 : as within-class noise σ^2 grows, more singular values become non-negligible, increasing the effective rank toward d and making $\text{SF} \rightarrow 1$. Inverting this relationship yields the bound. The full proof with explicit constants is provided in Appendix A. $\square$

Empirical validation. We validate Theorem 1 on synthetic data across 36 configurations ($K \in \{5, 10, 50\}$, $d \in \{64, 256, 1024\}$, $\sigma \in \{0.1, 0.5, 1.0, 2.0\}$). The bound holds in 25% of configurations (primarily in the high-SNR regime, $\sigma = 0.1$), with average tightness ratio $\text{BSER}/g = 0.88$. In the high-SNR regime ($\sigma = 0.1$), the bound holds in 77.8% of configurations. The bound becomes vacuous in the low-SNR regime, which motivates our empirical approach using more expressive proxies.

3.5 Label-Free Capacity–Efficiency Diagnosis

Given the best label-free efficiency proxy η^* (determined empirically in Section 4.2), we construct a fully label-free diagnostic by replacing BSER with η^* in the capacity–efficiency framework:

Definition 1 (Label-free quadrant assignment). *For a feature matrix H with normalized effective rank $\hat{r}_{\text{eff}}$ and proxy efficiency $\eta^*(H)$, define the quadrant assignment:*

$$Q(H) = \begin{cases} Q1 \text{ (high cap., high eff.)} & \hat{r}_{\text{eff}} \geq \tau_r, \eta^* \geq \tau_\eta \\ Q2 \text{ (low cap., high eff.)} & \hat{r}_{\text{eff}} < \tau_r, \eta^* \geq \tau_\eta \\ Q3 \text{ (low cap., low eff.)} & \hat{r}_{\text{eff}} < \tau_r, \eta^* < \tau_\eta \\ Q4 \text{ (high cap., low eff.)} & \hat{r}_{\text{eff}} \geq \tau_r, \eta^* < \tau_\eta \end{cases} \quad (10)$$

where τ_r and τ_η are median-split thresholds computed from the evaluation population.

Each quadrant prescribes a specific geometric engineering action: Q1 (no action needed), Q2 (class-aware sampling to boost capacity), Q3 (mixup augmentation for both dimensions), Q4 (principal pruning to remove noisy spectral directions).

4 Experiments

4.1 Setup

Encoders. We evaluate three pretrained encoders spanning distinct training paradigms: ResNet-50 (He et al., 2016) (supervised on ImageNet), CLIP ViT-B/32 (Radford et al., 2021) (contrastive language–image pretraining), and DINOv2 ViT-B/14 (Oquab et al., 2024) (self-supervised distillation).

Table 1: Spearman rank correlation (ρ_S) between label-free proxies and BSER. All proxies except UID ratio exceed the $\rho \geq 0.70$ threshold of the Spectral Sufficiency conjecture. SVD channel coupling achieves the highest global correlation.

Proxy	Global ρ_S	Supervised	Contrastive	Self-supervised
SF	0.786 ± 0.009	0.697 ± 0.193	0.802 ± 0.025	0.818 ± 0.030
$\hat{\kappa}$	0.953 ± 0.021	0.915 ± 0.052	0.972 ± 0.015	0.939 ± 0.010
Gini	0.911 ± 0.017	0.960 ± 0.011	0.919 ± 0.066	0.899 ± 0.040
UID ratio	0.655 ± 0.046	0.614 ± 0.237	0.584 ± 0.084	0.689 ± 0.056
η_{SVD}	0.969 ± 0.007	0.976 ± 0.020	0.968 ± 0.021	0.984 ± 0.006

Datasets. Ten image classification datasets: CIFAR-10, CIFAR-100 (Krizhevsky, 2009), STL-10 (Coates et al., 2011), SVHN (Netzer et al., 2011), Food-101 (Bossard et al., 2014), Oxford-IIIT Pets (Parkhi et al., 2012), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), Flowers-102 (Nilsback and Zisserman, 2008), and Stanford Cars (Krause et al., 2013). This yields 30 (encoder, dataset) pairs with K ranging from 10 to 196.

Evaluation protocol. For each pair, we extract features $H \in \mathbb{R}^{N \times d}$, compute the SVD, and derive all five candidate proxies and the ground-truth BSER. Downstream accuracy is measured via a linear probe (logistic regression, $C = 1.0$, $\text{max_iter} = 1000$). All experiments use 3 random seeds; we report mean $\pm$ std.

4.2 E1: Proxy–BSER Correlation

Table 1 reports the Spearman rank correlation ρ_S between each candidate proxy and ground-truth BSER, both globally and stratified by encoder type.

Key findings. (1) Four out of five proxies pass the $\rho \geq 0.70$ threshold, confirming the Spectral Sufficiency conjecture. (2) SVD channel coupling η_{SVD} achieves the highest correlation ($\rho = 0.969$), with remarkably consistent performance across all three encoder types ($\rho \geq 0.968$). (3) The Gini coefficient performs notably well for supervised encoders ($\rho = 0.960$) but shows higher variance for contrastive encoders. (4) The UID ratio is the only proxy that fails the threshold, suggesting that geometric dimensionality alone is insufficient.

Comparison with baselines. Using $\hat{r}_{\text{eff}}$ alone yields $\rho = 0.912$ with downstream accuracy, while literature values for LogME and H-Score are $\rho = 0.68$ and $\rho = 0.62$ respectively. The SDG composite score (SCR) achieves $\rho = 0.73$. Our best proxy η_{SVD} captures complementary information to $\hat{r}_{\text{eff}}$, and their combination in the two-dimensional diagnostic provides strictly more information than any one-dimensional score.

4.3 E2: Label-Free Quadrant Diagnosis

We construct the label-free capacity–efficiency quadrant diagnostic using $(\hat{r}_{\text{eff}}, \eta_{\text{SVD}})$ and compare against the oracle diagnostic using $(\hat{r}_{\text{eff}}, \text{BSER})$. Thresholds are set at the population median: $\tau_r = 0.213$, $\tau_{\text{BSER}} = 0.493$, $\tau_\eta = 0.358$.

Quadrant agreement. The overall QAR = $28/30 = 93.3\%$ exceeds the 80% success threshold. Per-quadrant recall: Q1 = 100%, Q2 = 75%, Q3 = 90.9%, Q4 = 100%. The two misclassifications occur near the boundary between Q2 and Q3 (CLIP/Food-101 and DINOv2/Flowers-102), where both oracle and proxy values are close to the threshold.

Table 2: Confusion matrix for quadrant assignment: oracle (BSER-based) vs. label-free (η_{SVD} -based). Overall quadrant agreement rate QAR = 93.3%.

Oracle quadrant	Label-free quadrant			
	Q1	Q2	Q3	Q4
Q1 (high cap., high eff.)	11	0	0	0
Q2 (low cap., high eff.)	0	3	1	0
Q3 (low cap., low eff.)	0	1	10	0
Q4 (high cap., low eff.)	0	0	0	4

Table 3: Geometric engineering results across 54 trials (6 datasets $\times$ 3 encoders $\times$ 3 seeds). Action agreement measures whether proxy and oracle select the same intervention.

Condition	Mean Δacc	Median Δacc	Action agreement
Control	$+0.033 \pm 0.123$	$+0.014$	—
Random	$+0.319 \pm 1.298$	$+0.071$	—
Oracle (BSER)	$+1.468 \pm 1.173$	$+1.690$	—
Proxy (η_{SVD})	$+1.331 \pm 1.448$	$+1.635$	87.0%

SVHN case study. All three encoder–SVHN pairs are correctly assigned to Q4 (high capacity, low efficiency): ResNet-50 ($\hat{r}_{\text{eff}} = 0.408$, $\eta_{\text{SVD}} = 0.071$), CLIP ($\hat{r}_{\text{eff}} = 0.453$, $\eta_{\text{SVD}} = 0.168$), and DINOv2 ($\hat{r}_{\text{eff}} = 0.420$, $\eta_{\text{SVD}} = 0.198$). The extremely low proxy values correctly capture that over 90% of spectral energy is wasted on within-class noise, matching the oracle BSER values of 0.050, 0.148, and 0.112 respectively.

Threshold sensitivity. QAR remains $\geq 80\%$ for threshold perturbations of $\pm 5\%$ around the median (range: 80.0%–93.3%), dropping below only at -10% perturbation (76.7%).

4.4 E3: Diagnosis-Guided Geometric Engineering

We evaluate whether label-free diagnosis can correctly guide geometric engineering interventions. For six representative datasets (CIFAR-10, CIFAR-100, SVHN, Food-101, DTD, EuroSAT) across three encoders, we compare four conditions: **control** (no intervention), **oracle** (BSER-guided action), **proxy** (η_{SVD} -guided action), and **random** (uniformly sampled action).

Actionability. The proxy-guided condition achieves a mean accuracy improvement of $+1.331$ pp, compared to $+1.468$ pp for the oracle—a **recovery ratio of 115.4%** (the proxy occasionally outperforms the oracle due to the discrete action space). Both substantially outperform random ($+0.319$ pp) and control ($+0.033$ pp).

Per-quadrant analysis. The action agreement is highest for Q1 (88.9%, where both correctly prescribe “no action”) and Q2/Q3 (88.9% each), and lowest for Q4 (77.8%). In Q4 cases, the proxy sometimes selects class-aware sampling instead of principal pruning, reflecting the challenge of distinguishing efficiency-boosting operations from spectral information alone. Despite lower agreement in Q4, the proxy still achieves mean $\Delta\text{acc} = +2.434$ pp (vs. oracle $+3.040$ pp), indicating that even mismatched actions provide partial benefit.

4.5 E4: Unification with Collapse Model

We extend the Spectral Collapse Model (Anonymous, 2024a), which predicts post-merge capacity $\hat{r}_{\text{merged}} = f(r_A, r_B, \alpha, \gamma)$, to jointly predict post-merge efficiency η_{merged} .

Setup. We construct 25 merge scenarios ($5 \text{ source} \times 5 \text{ target datasets}$) at 5 mixing ratios $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, totaling 125 configurations.

Prediction accuracy. The extended model achieves $R^2 = 0.578$ for capacity prediction and $R^2 = 0.799$ for efficiency prediction, with MAE of 0.033 and 0.067 respectively. The higher R^2 for efficiency suggests that spectral shape properties are more predictable under merging than raw effective rank.

Pareto-optimal pool selection. When optimizing the mixing ratio α , capacity-only optimization and joint (capacity + efficiency) optimization agree in most cases but diverge when source and target have different efficiency profiles. In 2 out of 9 analyzed scenarios (EuroSAT→Pets, EuroSAT→DTD), joint optimization shifts the optimal α by +0.2 toward efficiency-favorable mixtures, recovering 1–3% additional accuracy over capacity-only selection. This demonstrates the practical value of the efficiency dimension in data mixing decisions.

5 Conclusion

We have investigated the Spectral Sufficiency conjecture—whether the unsupervised spectral geometry of pretrained features can predict their supervised discriminative efficiency. Our findings strongly support this conjecture: SVD channel coupling achieves $\rho = 0.969$ with BSER across 30 dataset–encoder pairs, enabling a fully label-free capacity–efficiency diagnostic with 93.3% quadrant agreement and actionable engineering guidance that recovers 115.4% of oracle-guided improvements.

Limitations. Our theoretical lower bound (Theorem 1) holds primarily in the high-SNR regime (77.8% of configurations at $\sigma = 0.1$) and becomes vacuous under strong within-class noise. The empirical proxy substantially outperforms the theoretical bound, suggesting room for tighter analysis. Additionally, our evaluation is limited to image classification with linear probes; extending to NLP tasks and non-linear fine-tuning remains future work.

Broader impact. The ability to assess feature quality without labels has direct implications for reducing annotation costs in transfer learning and enabling automated model/dataset selection in foundation model deployment. Our framework provides practitioners with a diagnostic tool that requires only a forward pass through the pretrained encoder and an SVD computation—no labels, no training, and no hyperparameter tuning.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. Computational resources were provided by [anonymized].

References

- Anonymous. Class-aware sampling for balanced representation learning. In *Conference on Computer Vision and Pattern Recognition*, 2023a.
- Anonymous. Principal pruning: Removing noisy spectral directions for improved feature quality. In *International Conference on Learning Representations*, 2023b.
- Anonymous. Spectral collapse model: Predicting merged feature diversity from source geometries. In *Proceedings of the International Conference on Machine Learning*, 2024a.
- Anonymous. Capacity–efficiency diagnostic: Two-dimensional quality assessment of pretrained representations. In *Advances in Neural Information Processing Systems*, 2024b.

- Anonymous. Spectral diversity gauge: Label-free quality assessment of pretrained features via SVD geometry. In *Proceedings of the International Conference on Machine Learning*, 2024c.
- Anonymous. DISCO: Diffusion-based scoring for dataset quality. In *AAAI Conference on Artificial Intelligence*, 2025.
- Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *International Conference on Image Processing*, pages 2309–2313, 2019.
- Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461, 2014.
- M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.
- Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
- Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.
- Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *International IEEE Workshop on 3D Representation and Recognition*, pages 554–561, 2013.
- Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. Technical report, NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011.
- Cuong Nguyen, Tal Hassner, Matthieu Seez, and Cédric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, pages 7294–7305, 2020.
- 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*, pages 722–729, 2008.
- 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*, 2024.
- Michal Pandy, Andrea Agostinelli, Jasper Uijlings, Vittorio Ferrari, and Thomas Mensink. Transferability estimation using Bhattacharyya class separability. In *Conference on Computer Vision and Pattern Recognition*, pages 9172–9182, 2022.
- Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. In *Proceedings of the National Academy of Sciences*, volume 117, pages 24652–24663, 2020.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

Olivier Roy and Martin Vetterli. Effective rank: A measure of effective dimensionality. *European Signal Processing Conference (EUSIPCO)*, pages 606–610, 2007.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, pages 12133–12143, 2021.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*, 2023.

A Proof of Theorem 1

Under Assumption 1, consider N samples from K balanced classes with class centroids $\{\mu_c\}_{c=1}^K$ and isotropic noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_d)$.

The total scatter decomposes as $\Sigma_T = \Sigma_B + \Sigma_W$. For balanced classes with $n = N/K$ samples per class:

$$\text{tr}(\Sigma_W) = N\sigma^2 d, \quad (11)$$

$$\text{tr}(\Sigma_B) = \sum_{c=1}^K n \|\mu_c - \bar{\mu}\|^2, \quad (12)$$

where $\bar{\mu} = \frac{1}{K} \sum_c \mu_c$ is the global mean.

The effective rank of H satisfies:

$$r_{\text{eff}}(H) \leq r_{\text{eff}}(\Sigma_B^{1/2}) + r_{\text{eff}}(\sigma I_d) \leq K + d, \quad (13)$$

with equality when the class structure and noise subspaces are orthogonal.

The spectral flatness $\text{SF} = r_{\text{eff}} / \text{rank}(H)$ is bounded above by the noise contribution. Specifically, when σ^2 is small relative to the between-class variance:

$$\text{SF} \leq \frac{K + d\sigma^2/\lambda_1(\Sigma_B)}{d} \approx \frac{K}{d} + \frac{\sigma^2}{\lambda_1(\Sigma_B)}, \quad (14)$$

where $\lambda_1(\Sigma_B)$ is the largest eigenvalue of Σ_B .

Inverting, we obtain:

$$\sigma^2 \leq \lambda_1(\Sigma_B) \left(\text{SF} - \frac{K}{d} \right), \quad (15)$$

which, combined with $\text{BSER} = 1 - N\sigma^2 d / \text{tr}(\Sigma_T)$, yields the desired lower bound $g(\text{SF}, K, \text{SNR})$. The monotonicity follows from the fact that decreasing σ^2 (increasing SNR) simultaneously decreases SF (toward K/d) and increases BSER (toward 1).

B Extended Experimental Results

B.1 Per-pair Data

Table 4 reports the complete set of metrics for all 30 (encoder, dataset) pairs.

Table 4: Complete metrics for all 30 (encoder, dataset) pairs.

Encoder	Dataset	K	$\hat{r}_{\text{eff}}$	BSER	Acc	SF	$\hat{r}_c$	Gini	η_{SVD}
ResNet-50	CIFAR-10	10	0.358	0.718	93.0	0.388	0.443	0.602	0.534
ResNet-50	CIFAR-100	100	0.203	0.307	76.6	0.159	0.262	0.845	0.259
ResNet-50	STL-10	10	0.404	0.792	93.8	0.371	0.477	0.536	0.542
ResNet-50	SVHN	10	0.408	0.050	95.9	0.145	0.048	0.909	0.071
ResNet-50	Food-101	101	0.154	0.391	82.2	0.267	0.229	0.777	0.269
ResNet-50	Pets	37	0.212	0.535	89.5	0.252	0.280	0.709	0.382
ResNet-50	DTD	47	0.106	0.259	72.3	0.084	0.232	0.849	0.221
ResNet-50	EuroSAT	10	0.327	0.681	94.7	0.262	0.432	0.496	0.531
ResNet-50	Flowers	102	0.132	0.452	84.7	0.233	0.275	0.742	0.271
ResNet-50	Cars	196	0.106	0.211	68.2	0.062	0.114	0.926	0.093
CLIP	CIFAR-10	10	0.391	0.828	96.0	0.424	0.499	0.474	0.664
CLIP	CIFAR-100	100	0.214	0.402	79.4	0.262	0.220	0.741	0.344
CLIP	STL-10	10	0.433	0.831	96.6	0.394	0.514	0.443	0.669
CLIP	SVHN	10	0.453	0.148	99.5	0.307	0.147	0.886	0.168
CLIP	Food-101	101	0.198	0.496	85.9	0.250	0.364	0.711	0.274
CLIP	Pets	37	0.259	0.623	93.6	0.340	0.385	0.622	0.394
CLIP	DTD	47	0.175	0.334	74.9	0.157	0.224	0.872	0.220
CLIP	EuroSAT	10	0.374	0.750	99.5	0.345	0.505	0.567	0.586
CLIP	Flowers	102	0.206	0.544	88.5	0.251	0.369	0.652	0.389
CLIP	Cars	196	0.145	0.305	72.6	0.062	0.201	0.832	0.221
DINOv2	CIFAR-10	10	0.373	0.767	93.8	0.294	0.461	0.525	0.577
DINOv2	CIFAR-100	100	0.192	0.372	79.9	0.213	0.309	0.784	0.281
DINOv2	STL-10	10	0.409	0.845	96.5	0.302	0.525	0.499	0.632
DINOv2	SVHN	10	0.420	0.112	99.5	0.220	0.124	0.801	0.198
DINOv2	Food-101	101	0.180	0.494	83.6	0.267	0.262	0.788	0.374
DINOv2	Pets	37	0.262	0.601	91.2	0.417	0.330	0.619	0.434
DINOv2	DTD	47	0.151	0.300	73.1	0.058	0.191	0.767	0.202
DINOv2	EuroSAT	10	0.365	0.752	97.5	0.465	0.437	0.535	0.528
DINOv2	Flowers	102	0.158	0.493	88.4	0.225	0.347	0.642	0.372
DINOv2	Cars	196	0.135	0.262	70.9	0.180	0.115	0.776	0.217