THE DEVIL IS IN THE SPECTRUM BIAS: SPECTRUM-BALANCED FEATURE MATCHING FOR ROBUST REPRESENTATION DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large visual foundation models have demonstrated remarkable transferability across a wide range of downstream tasks. To deploy such models efficiently, feature matching has become a popular knowledge distillation approach that transfers teacher representations to smaller student models without requiring labeled data. However, we show that the conventional feature matching objective with L2-distance is inherently biased toward reconstructing dominant spectral directions of the teacher representation, while under-optimizing low-variance directions that often contain task-relevant information. To address this, we propose Spectrum-Balanced Feature Matching, SpecMatch, a simple objective that adaptively emphasizes under-optimized spectral directions while preserving the relative importance of dominant directions. SpecMatch is easy to implement and introduces negligible computational overhead. Extensive experiments on image recognition demonstrate that SpecMatch consistently improves downstream adaptation across diverse tasks, including image classification, anomaly detection, medical image analysis, and domain generalization. In particular, SpecMatch outperforms conventional feature matching in 40 of 42 teacher–student and training-setting combinations, while consistently improving over the original student model in all settings. We further demonstrate that the proposed objective generalizes beyond vision, improving downstream performance across six protein understanding tasks.

1 INTRODUCTION

Recent advances in large-scale representation learning have led to the emergence of large foundation models trained on billions of images (Oquab et al., 2023; Siméoni et al., 2025; Bolya et al., 2026; Chuang et al., 2026; Xu et al., 2024; Radford et al., 2021). Such models have demonstrated remarkable transferability across a wide range of downstream tasks, including object recognition, medical image analysis, and domain generalization. As foundation models continue to grow in scale and capability, they increasingly serve as a source of visual knowledge for downstream applications.

However, directly deploying these large models is often impractical due to computational constraints. Consequently, an important challenge is how to effectively transfer the rich knowledge in large models to smaller and more efficient models. Knowledge distillation has emerged as a promising solution to this problem, enabling student models to inherit useful representations from powerful teachers (Hinton et al., 2015; Jang et al., 2025; Lee et al., 2025; Ranzinger et al., 2024b; Zhang et al., 2025). Feature matching is a popular distillation approach since it transfers teacher representations to a student model and can be applied to diverse scenarios in a label-free manner (Sarıyıldız et al., 2024; Chen et al., 2021; Wang et al., 2019). However, standard feature matching objectives treat the teacher representation as a whole, without explicitly accounting for how information is distributed across different feature directions. This raises an important question: *does feature matching effectively transfer all information encoded in the teacher representation, including information that may be important for downstream tasks?* Our preliminary analysis reveals an important tension: teacher representations learned from large-scale pre-training are highly anisotropic, with a small number of principal directions dominating the feature variance; however, non-dominant directions can also contain information that contributes substantially to downstream tasks. As illustrated in

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.

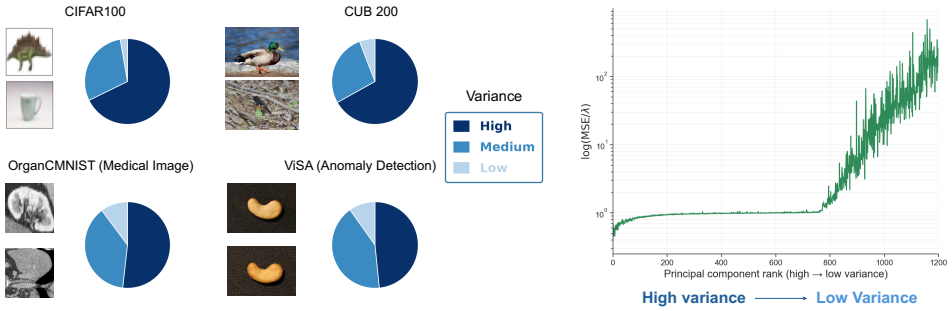


Figure 1: ***(Left)*:** Contribution of principal component (PC) groups to downstream tasks. PCs are sorted by explained variance and divided into three equal-sized groups. Different tasks rely on different spectral regions, motivating the preservation of a broad representation spectrum during distillation. ***(Right)*:** Variance-normalized reconstruction error of vanilla L2 feature matching. Reconstruction deteriorates toward higher-rank, lower-variance PCs, revealing a bias toward high-variance components that may discard information useful for downstream tasks. We further provide evidence in Fig. 7 that lower-variance PCs do not merely represent noise but can substantially improve performance on certain downstream tasks.

Fig. 1, different downstream tasks rely on different spectral regions of the teacher representation, and important task-relevant information can reside in non-dominant directions.

Motivated by this observation, we identify a limitation of conventional feature matching in transferring teacher knowledge to a student. We theoretically and empirically show that feature matching with L2 distance is biased toward reproducing dominant spectral directions: its learning signals scale with teacher feature variance, leading student models to prioritize high-variance directions over low-variance ones. As a result, the student may fail to preserve information that is useful for downstream transfer. Indeed, we empirically observe settings in which vanilla feature matching fails to improve downstream performance. To address this issue, we propose Spectrum-Balanced Feature Matching, *SpecMatch*, a simple objective that adaptively emphasizes under-optimized spectral directions while preserving the relative importance of dominant directions. SpecMatch is easy to implement, introduces negligible computational overhead, and can be incorporated into existing feature matching frameworks with minimal modification.

Extensive experiments on image recognition demonstrate that SpecMatch generally outperforms conventional feature-matching baselines in a downstream adaptation setting, where distillation is performed separately for each target dataset using only its unlabeled images. Across diverse tasks—including image classification, anomaly detection, medical image classification, and domain generalization—our method outperforms vanilla feature matching in 40 out of 42 teacher–student and training-setting combinations, while improving over the original student model in all 42 settings. The method remains effective on large-scale datasets such as ImageNet, as well as on challenging datasets such as iNaturalist, which contains a large number of categories and a highly imbalanced class distribution. Furthermore, the proposed objective generalizes to protein foundation models, achieving consistent improvements across six downstream protein understanding tasks.

2 Related Work

Knowledge distillation is a popular approach to transfer knowledge between different models by training one model to mimic the outputs of another model. Various approaches have been proposed and some employ task logits (Hinton et al., 2015), others utilize intermediate features or embeddings (Heo et al., 2019; Tian et al., 2020), relations between samples (Park et al., 2019). Some of the self-supervised learning approaches include learning mechanisms similar to model distillation (Oquab et al., 2023; Siméoni et al., 2025). Feature distillation approaches are common in distillation literature while they do not account for transferring rich teacher representations (Sarıyıldız et al., 2024; Heo et al., 2019; Chen et al., 2021; Romero et al., 2015). We mainly focus on how to adapt the student model to a downstream task by leveraging a huge teacher model, using unlabeled data from the downstream domain. The downstream tasks can require diverse features for the performance boost, thus transferring diverse features is desirable. Our setting is close to (Vemulapalli

et al., 2024; Jang et al., 2025) in that they assume a label-scarce scenario for cheaper adaptation. While Vemulapalli et al. (2024) focus on obtaining the auxiliary dataset that helps the adaptation to the target domain, our focus is on developing an objective that achieves rich feature transfer. SpecMatch is simple, yet effective in improving the performance of the student model by allowing students to learn rich features.

Normalization in knowledge distillation. Miles et al. (2024); Miles & Mikolajczyk (2024); Lee et al. (2018) employ feature normalization or whitening in knowledge distillation. Our theoretical analysis provides an intuition for why whitening can facilitate feature matching: by equalizing the variance of the teacher representation across spectral directions, it prevents a small number of high-variance directions from dominating the objective. However, our analysis shows that whitening does not consistently improve downstream transfer and can even cause substantial degradation on several datasets. These results suggest that completely equalizing spectral directions can be overly aggressive, since dominant directions may remain important for downstream tasks. In contrast, SpecMatch is designed to reduce the excessive optimization focus on high-variance directions without enforcing equal contributions across the spectrum. Specifically, it moderately emphasizes low-variance directions without fully equalizing their contributions with those of high-variance directions.

Richness of feature representations. The richness of learned representations has been extensively studied in the self-supervised learning (SSL) literature. RankMe (Garrido et al., 2023) showed that the effective rank of learned representations is strongly correlated with downstream performance, suggesting that high-rank representations capture richer semantic information. Motivated by this observation, several SSL methods have been proposed to prevent rank collapse or maintain a high effective rank of learned representations (Zbontar et al., 2021; Jing et al., 2021; He & Ozay, 2022). Our work shares a similar motivation in that we aim to preserve the richness of teacher representations during knowledge transfer. However, rather than designing a self-supervised objective to improve the representation itself, we develop a distillation objective that transfers the rich representations of a powerful teacher model to a student model.

3 Revisiting Feature Matching for Representation Transfer

3.1 Task-Relevant Information is Distributed Across the Spectrum

We analyze the contribution of each principal component to downstream classification. Let $z \in \mathbb{R}^d$ denote the teacher representation, $U = [u_1, \dots, u_d]$ the orthonormal PCA basis computed from the teacher representations (ordered by decreasing variance), and $W = [w_1, \dots, w_C]^\top$ the weight matrix of a linear classifier trained for a downstream task, where C is the number of classes. Since U is orthonormal, each classifier weight can be decomposed as $w_c = \sum_{k=1}^{d} (u_k^\top w_c) u_k$. We quantify the contribution of the k-th principal component by measuring the energy of the classifier weights projected onto that direction: $I_k = \sum_{c=1}^{C} (u_k^\top w_c)^2$. Fig. 1 (left) partitions the principal components into consecutive rank intervals and compute the normalized contribution within each interval: $C_m = \frac{\sum_{k \in \mathcal{B}_m} I_k}{\sum_{k=1}^{d} I_k}$, where $\mathcal{B}_m$ denotes the m-th rank interval. The darker shades of blue indicate principal component groups with larger teacher variance. For the CIFAR-10, the task-relevant directions are strongly aligned with the dominant principal components of the teacher representation, while higher-rank components contribute little to the classifier. In contrast, for anomaly detection and medical image classification, which require recognizing subtle visual patterns, lower-variance principal components also exhibit substantial contributions. These suggest that transferring only the dominant directions of the teacher can be insufficient for downstream tasks. Instead, effective representation transfer should preserve information across a broader spectrum of principal components.

3.2 Insights into Vanilla Feature Matching

We analyze the feature matching with L2-distance from the perspective of the principal components of the teacher representation. Let $z_b^t, z_b^s \in \mathbb{R}^d$ denote the teacher and student representations for the b-th sample, respectively. Applying PCA to the teacher features yields an orthonormal basis $\{u_k \in \mathbb{R}^d\}_{k=1}^d$, where the teacher representation is expressed as $z_b^t = \sum_{k=1}^{d} c_{b,k}^t u_k$, and $\lambda_k = \mathrm{Var}(c_k^t)$ denotes the variance along the k-th principal direction.

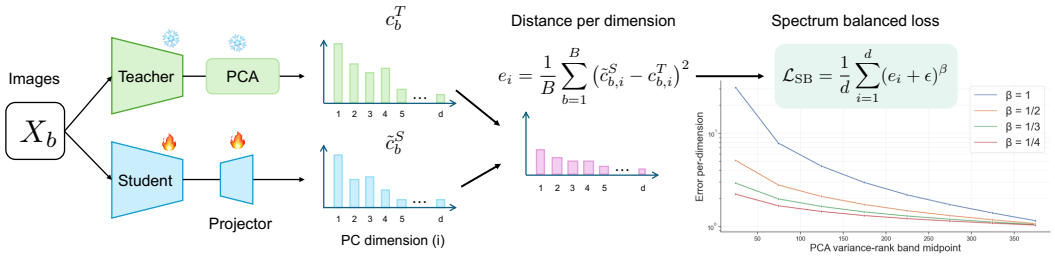


Figure 2: ***(Left)*:** Description of the proposed objective. ***(Right)*:** The illustration of the error (y-axis) on different principal components (x-axis).

Since the PCA basis is orthonormal, feature matching is equivalent to minimizing the reconstruction error in the PCA space:

$$e_k = \frac{1}{B}\sum_{b=1}^{B}(\tilde{c}_{b,k}^s - c_{b,k}^t)^2, \qquad \mathcal{L}_{\text{FM}} = \sum_{k=1}^{d} e_k \tag{1}$$

where $\tilde{c}_{b,k}^s$ is the student prediction along the k-th principal direction. At the early stage of training, the expected reconstruction error satisfies

$$\mathbb{E}[e_k] \approx \lambda_k + C, \tag{2}$$

where C is approximately constant across principal components (See Sec. B for proof). Combining Eq. 2 with the gradient analysis (See Sec. B), we obtain

$$\mathbb{E}\left[\left\|\nabla_{\tilde{\mathbf{c}}_{:,k}}\mathcal{L}_{\text{FM}}\right\|_2^2\right] \propto \lambda_k + C. \tag{3}$$

This shows that the gradient norm along each principal direction k scales with λ_k, i.e., the variance of the teacher features along that direction. This demonstrates that the optimization dynamics of feature matching allocate larger optimization signals to principal components with larger teacher variance, which can hinder the reconstructing target features with higher ranks, low-variance principal directions [1]. In fact, the right of Fig. 1 shows that a model trained with vanilla feature matching struggles to accurately reconstruct the teacher representations along higher-rank directions.

4 SpecMatch: Spectrum-Balanced Feature Matching

To mitigate the spectral bias explained above, we propose Spectrum-Balanced Feature Matching, SpecMatch. Our key observation is that reconstruction errors tend to be larger along dominant principal directions due to their higher variance. Consequently, the normal feature matching allocates disproportionate learning signals to these directions. To address this, we introduce a simple framework that explicitly mitigates the imbalance in the contributions using their reconstruction errors.

Predicting coefficients of the PCA basis. As shown in the left of Fig. 2, the student predicts the coefficients of the teacher representation in the PCA basis, rather than directly regressing the teacher features. This formulation makes the reconstruction error along each principal direction explicit, allowing us to balance the loss across spectral directions. Both the teacher model and the PCA basis are fixed during training. We employ a linear classifier as the prediction head.

Proposed Objective. To quantify the reconstruction error along each principal direction, we first compute the average teacher–student error e_k within each mini-batch, as defined in Eq. 1. As indicated by Eq. 2, e_k scales with the variance λ_k of the corresponding teacher direction. Rather than linearly aggregating these per-direction errors, we employ a sub-linear aggregation:

$$\mathcal{L}_{\text{SB}} = \sum_{i=1}^{d}(e_k + \epsilon)^\beta, \qquad 0 < \beta < 1, \tag{4}$$

where ϵ is a small constant for numerical stability. This transformation reduces the relative influence of large-error directions and increases that of under-optimized directions. β is an important hyperparameter that controls the degree of emphasis placed on low-energy components. As β approaches

[1] Fig. 5 shows that, even after many training iterations, e_k remains approximately proportional to λ_k for high-variance directions. These directions continue to dominate the optimization throughout training.

1, the objective becomes equivalent to the standard L2 feature-matching loss. As β approaches 0, the contributions of low- and high-energy components become more uniform. Importantly, because $(e_k + \epsilon)^\beta$ remains monotonically increasing in e_k, directions with large variance and reconstruction error still receive greater emphasis; the objective only moderates their dominance to achieve a better balance across the spectrum. The right panel of Fig. 2 illustrates $(e_k+\epsilon)^\beta$ for different values of β on real data, with principal components ordered from the highest to the lowest teacher variance. Empirically, high-variance components tend to exhibit larger reconstruction errors, whereas low-variance components incur substantially smaller errors. Our objective compresses this disparity, balancing the contributions of different principal components while preserving the greater importance of dominant directions. The pseudo-code is available in Alg. 1.

Analysis with respect to gradient. The gradient with respect to e_k is

$$\frac{\partial \mathcal{L}_{\mathrm{SB}}}{\partial e_k} = \beta(e_k + \epsilon)^{\beta-1} \tag{5}$$

Since $\beta < 1$, directions with smaller reconstruction errors receive relatively larger gradient coefficients, while directions with large errors are prevented from dominating the optimization. Here, ϵ controls the maximum gradient coefficient. Similarly, the gradient with respect to the student coefficient can be written as

$$\frac{\partial \mathcal{L}_{\mathrm{SB}}}{\partial \tilde{c}^S_{b,k}} = (e_k + \epsilon)^{\beta-1} \frac{\partial \mathcal{L}_{\mathrm{FM}}}{\partial \tilde{c}^S_{b,k}}. \tag{6}$$

This shows that SpecMatch can be interpreted as an adaptive spectral reweighting strategy, where each principal direction receives a weight determined by its reconstruction error. Thus, compared with vanilla feature matching, the proposed objective reduces the variance-dependent learning bias to approximately $\lambda_k^{\beta-1}$. Since $0 < \beta < 1$, the gap between high- and low-variance directions is substantially reduced. In summary, conventional feature matching tends to reproduce dominant spectral directions first, while under-optimizing non-dominant directions that may be important for downstream transfer. SpecMatch counteracts this bias by balancing learning progress across teacher spectral directions, enabling the student to inherit richer representations.

Alternative formulation with static spectral balancing. Our analysis reveals that vanilla feature matching induces an optimization bias across spectral directions: principal directions with larger teacher variance tend to produce larger reconstruction errors and dominate the optimization. This observation suggests that the identified bias can be mitigated by explicitly accounting for the teacher spectrum. One possible formulation is to directly assign a static weight to each principal direction according to its variance and leading to the following objective:

$$\mathcal{L}_{\mathrm{static}} = \sum_{k=1}^{d} (\lambda_k + \epsilon)^{\beta-1} e_k. \tag{7}$$

Since $0 < \beta < 1$, this formulation suppresses the excessive contribution of high-variance directions while relatively emphasizing low-variance directions. Importantly, this objective follows directly from our analysis of the variance-dependent optimization bias and therefore represents another way of exploiting the proposed spectral-balancing principle. We provide the empirical analysis of this objective in Sec. D and confirm the superiority over the vanilla feature matching.

5 Experiments

5.1 Evaluation on Diverse Downstream Tasks

We first evaluate the proposed method across diverse image classification datasets to assess its general applicability. For each dataset, we train a student to match the teacher's representations and evaluate the resulting model on that dataset.

Datasets. We consider 14 benchmark datasets spanning seven categories: general image classification (CIFAR10 and CIFAR100 (Krizhevsky, 2009)), fine-grained recognition (CUB (Wah et al., 2011)), medical image classification ((Yang et al., 2023),(Bandi et al., 2019)), remote sensing ((Cheng et al., 2017), (Helber et al., 2019)), anomaly detection ((Zou et al., 2022), (Bergmann et al., 2019)), and domain generalization ((Beery et al., 2018), (Christie et al., 2018)).

Table 1: Performance comparison across 14 datasets, including general recognition, OCR, fine-grained recognition, medical image classification, and anomaly detection benchmarks. The cell where the distillation model underperforms the student linear probe is highlighted with **red** while the best model is highlighted with **bold**.

		General		OCR		Fine		Remote		Medical		Anomaly		Generalization	
	Average	CIFAR100	CIFAR10	SVHN	GTSRB	NABirds	CUB	Resisc	Eurosat	OrganCMNIST	Camelyon17	VisA	MVTec	iWildCam	FmoW
	# classes	100	10	10	43	555	200	45	10	11	2	2	2	186	62
PE-Core-Large → PE-Core-Tiny															
Teacher	86.5	93.1	99.4	73.6	93.7	87.3	90.6	96.1	97.4	82.9	90.1	81.9	96.3	79.2	50.1
Student	76.0	76.2	92.1	57.7	81.5	59.7	71.1	88.9	94.6	82.9	89.0	75.8	86.3	71.0	36.5
Logits-KD (Hinton et al., 2015)	80.9	79.5	95.1	82.2	96.8	69.9	71.6	93.2	98.4	92.4	**88.8**	**69.2**	**71.9**	74.1	45.3
DINO (Caron et al., 2021)	79.3	**74.2**	94.3	85.5	89.9	69.2	72.7	91.5	97.8	85.9	91.1	**62.9**	89.4	**66.2**	40.4
RKD (Park et al., 2019)	**70.0**	**60.7**	93.7	67.3	90.8	62.3	76.1	**19.2**	**92.7**	84.1	91.5	**66.1**	92.4	74.3	**9.3**
FM	83.0	80.9	97.1	84.1	93.3	70.3	77.5	93.6	98.3	83.8	**94.0**	**74.2**	96.5	71.7	43.6
SpecMatch	**84.6**	**81.9**	**97.3**	**85.8**	**94.8**	**73.7**	**79.6**	**94.8**	**98.4**	**89.2**	93.4	**76.6**	**97.9**	**74.5**	**47.4**
PE-Core-Large → MobileNet-V3															
Teacher	86.5	93.1	99.4	73.6	93.7	87.3	90.6	96.1	97.4	82.9	90.1	81.9	96.3	79.2	50.1
Student	76.0	64.6	86.5	65.0	77.6	55.4	67.6	86.3	94.9	85.2	85.7	72.8	84.6	58.6	26.1
Logits-KD (Hinton et al., 2015)	80.3	76.3	94.2	92.9	96.6	60.8	67.9	93.0	98.1	92.6	88.8	75.1	89.2	**57.8**	40.2
DINO (Caron et al., 2021)	79.3	74.8	94.3	87.4	93.3	**54.7**	**66.6**	91.5	97.8	88.4	93.0	76.5	96.0	59.7	35.7
RKD (Park et al., 2019)	**70.8**	70.0	**92.0**	81.3	89.8	**54.7**	70.0	89.7	97.4	88.4	91.4	76.8	93.7	**56.3**	32.7
FM	80.9	76.0	94.8	87.6	93.9	62.6	71.3	92.6	**98.2**	89.2	93.0	80.7	95.5	**58.5**	39.0
SpecMatch	**82.4**	**77.6**	**95.2**	**89.5**	**94.6**	**64.2**	**72.5**	**93.6**	**98.2**	**90.7**	**93.9**	**81.9**	**96.8**	**59.9**	**42.1**
DINO-V2 → MobileNet-V3															
Teacher	84.0	92.7	99.3	62.7	77.6	87.8	90.3	93.8	96.7	87.6	87.1	82.1	95.2	81.1	41.4
Student	76.0	64.6	86.5	65.0	77.6	55.4	67.6	86.3	94.9	85.2	85.7	72.8	84.6	58.6	26.1
Logits-KD (Hinton et al., 2015)	80.1	76.7	94.3	88.7	95.6	61.0	67.9	92.9	98.3	93.5	92.0	77.0	86.4	**57.7**	39.0
DINO (Caron et al., 2021)	76.9	75.5	94.4	78.2	84.5	**52.6**	**66.3**	91.7	97.5	88.8	86.1	74.4	91.3	**61.2**	35.3
RKD (Park et al., 2019)	**75.8**	69.4	92.0	77.0	83.0	**54.7**	**64.3**	89.9	96.9	87.2	89.5	74.9	91.7	**58.3**	32.1
FM	78.7	77.2	95.2	80.5	87.7	**53.6**	**66.4**	92.6	97.9	89.6	93.2	76.6	93.5	60.2	37.5
SpecMatch	**79.8**	**78.1**	**95.5**	**82.3**	**88.1**	**56.8**	**69.0**	**92.8**	**98.0**	**91.0**	**93.4**	**77.8**	**94.3**	61.1	**39.5**

Models. We use PE-Core-G14 (Bolya et al., 2026) and DINO-V2-ViT-G14 (Oquab et al., 2023) as the teacher model and evaluate two student architectures with different capacities: MobileNetV3-Small-1.0 (Howard et al., 2019) pre-trained on ImageNet (Deng et al., 2009) and PE-Core-T14 (Bolya et al., 2026) pre-trained on image-text data.

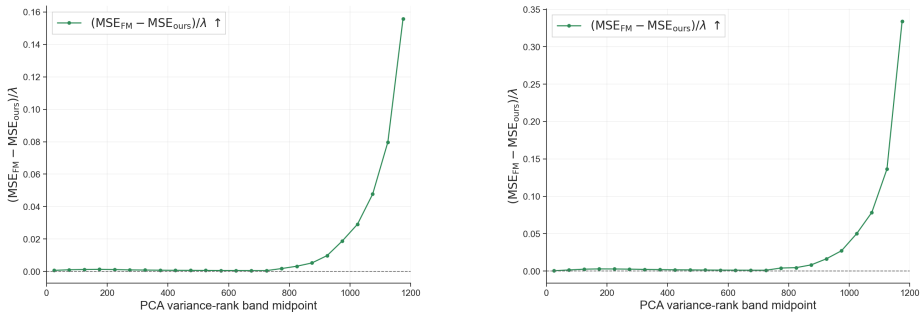
Baselines. Relational knowledge distillation (RKD) (Park et al., 2019), feature matching with L2-distance (FM), and DINO (Caron et al., 2021) are used as the baseline for the model distillation with unlabeled data. The implementation details are available in Sec. C. As a reference to the distillation baseline using labeled data during training, we additionally report Logits-KD (Hinton et al., 2015), where a classification head is trained on the teacher using labeled data, and the resulting logits are used as soft targets for students.

Training. All models are trained with AdamW (Loshchilov & Hutter, 2019). Since FM, DINO and SpecMatch need to train the linear head, we freeze the base model for 500 iterations and tune all parameters. β^{-1} is set as 3 in all datasets. Other details are provided in Sec. C.

Evaluation. After unlabeled feature matching, the learned representations are evaluated by linear-probing and report classification accuracy. For MVTec, we compute the distance to the nearest neighbor to compute AUROC. We report the performance averaged over three runs.

Overview of the results. According to Table 1, SpecMatch outperforms all baselines in almost every setting (40/42 settings), demonstrating the effectiveness of distilling rich feature representations. Notably, the performance gains are consistent across all combinations of teacher pre-training strategies and student architectures, indicating that our method generalizes well across diverse distillation settings. Several observations can be drawn from these results.

SpecMatch consistently improves the pretrained student. First, while several distillation baselines even degrade performance compared to the pretrained student, as highlighted by red numbers, SpecMatch consistently improves performance in all evaluated settings. Transferring rich feature representations is a highly effective strategy for improving downstream task performance.



(a) CIFAR100 (b) NABirds

Figure 3: **Difference in reconstruction errors across the principal components.** We compute the mean reconstruction error between the teacher representation and the student, and report $e_k^{\mathrm{FM}} - e_k^{\mathrm{ours}}$ normalized by the teacher variance λ_k. A higher value shows that SpecMatch reconstructs the teacher representation more accurately than the feature matching baseline.

Table 2: Experiments on distillation with ImageNet. All models are evaluated with classification accuracy (%) after linear-probing on the corresponding labeled subset.

Teacher	Student	Params	Initialization	Method	5-shot	10-shot	Full-shot
PE-Core-G14	ViT-Tiny	5.5M	Scratch	DINO	31.3	37.2	54.6
				FM	48.4	51.6	57.6
				SpecMatch	**50.2**	**52.7**	**58.4**
ConvNEXT-XLarge	ConvNEXT-Tiny	27.8M	Scratch	DINO	36.4	42.9	54.7
				FM	57.6	60.6	67.1
				SpecMatch	**61.2**	**63.9**	**70.0**
PE-Core-G14	PE-Core-T14	6.1M	Pretrained	Base	45.3	52.1	68.1
				DINO	60.0	62.6	68.3
				FM	59.3	63.5	72.1
				SpecMatch	**60.3**	**64.4**	**73.1**

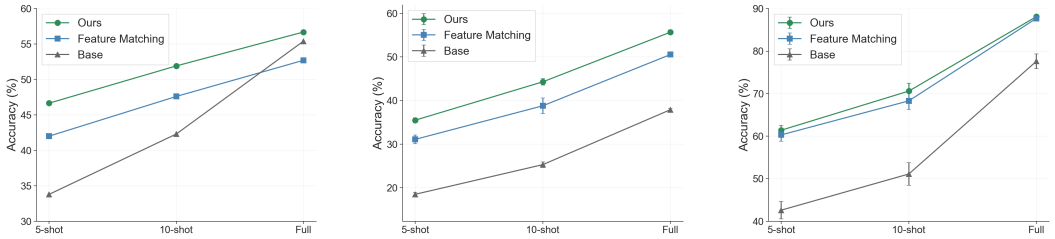
SpecMatch provides larger gains on fine-grained recognition. Second, the improvement over the feature matching (FM) baseline is relatively modest on general recognition tasks such as CIFAR-10 and CIFAR-100, whereas substantially larger gains are observed on fine-grained datasets such as CUB and NABirds. This trend is consistent with the analysis in Figure 1, which shows that general recognition primarily relies on high-variance feature directions, while fine-grained recognition benefits more from preserving lower-variance components.

SpecMatch is effective for domain generalization. SpecMatch yields large improvements on domain generalization benchmarks. We conjecture that preserving richer representations leads to more transferable and robust representations, thereby improving generalization to unseen domains.

The student can even surpass the teacher on out-of-domain tasks. Finally, the student frequently surpasses the teacher on several domains, particularly OCR, medical imaging, and remote sensing, even with FM. These domains are likely underrepresented in the teacher's pretraining data, suggesting that our distillation strategy enables the student to leverage the teacher's rich representations while adapting them more effectively to out-of-domain tasks.

5.2 Analysis.

Reconstruction Error in each basis. Figure 3 compares the reconstruction errors of Feature Matching (FM) and SpecMatch along each principal direction. Let e_k^{FM} and e_k^{ours} denote the mean squared reconstruction errors along the k-th principal direction for FM and SpecMatch, respectively. We visualize their difference, $e_k^{\mathrm{FM}} - e_k^{\mathrm{ours}}$, normalized by the teacher variance λ_k along the corresponding direction. Positive values indicate that SpecMatch reconstructs the teacher representations more accurately than FM. SpecMatch consistently achieves lower reconstruction error across the entire spectrum, demonstrating that the improvement is not limited to a specific subset of principal components. Notably, the gain becomes substantially larger for low-variance components, indicating that our objective effectively addresses the optimization imbalance highlighted in Fig. 1, where conventional feature matching tends to underfit low-variance directions. This observation further suggests that improving the reconstruction of these previously under-learned directions contributes directly to the superior downstream performance. Interestingly, although our objective balances optimiza-



(a) NABirds (b) FGVC (c) GTSRB

Figure 4: Linear probing with different shots. Our approach consistently improves performance over the base student model while vanilla feature matching sometimes fails in improving as in NABirds full-shot case (a).

Table 3: ***(Left)*:** Sensitivity analysis for β^{-1}. $\beta = 1$ corresponds to vanilla feature matching, while $\beta^{-1} = 2$ and 3 consistently improve performance. ***(Right)*:** Comparison among different objectives, including L1, cosine distance, and L2 + whitening. SpecMatch is the best or the second best approach in these results.

β^{-1}	ImageNet	NAB	CIF100	SVHN
1 (FM)	72.1	60.7	76.2	87.2
2	72.7	62.3	77.5	89.4
3	**73.1**	64.2	77.6	89.5
4	72.9	64.9	77.8	89.9
5	72.9	64.9	78.2	89.4

Objective	NAB	CIF100	Cam17	GTSRB	SVHN	FmoW
L1	63.6	76.4	**94.3**	93.9	88.9	38.8
Cosine	61.3	75.8	93.2	92.4	87.1	37.3
L2	62.6	76.0	93.0	93.9	87.6	39.0
L2 + Whitening	63.2	**77.8**	93.1	91.6	86.7	39.4
SpecMatch	**64.8**	77.6	93.9	**94.6**	**89.5**	**42.1**

tion across spectral directions, it also yields a small but consistent improvement in high-variance components. This suggests that balancing the optimization landscape not only prevents the neglect of low-variance directions but also facilitates overall optimization, leading to better reconstruction even in the dominant spectral components.

Experiments on ImageNet. Table 2 presents analysis on ImageNet. We train a student model from scratch or fine-tune a pre-trained student model. In both cases, the models are trained without labels. Our objective shows consistent improvement over vanilla feature matching. These results indicate that the proposed method remains effective even on a large-scale dataset such as ImageNet. Furthermore, the improvements are consistently observed under both 5-shot and 10-shot linear evaluation protocols.

Few-shot probing. Figure 4 presents few-shot linear probing results on fine-grained datasets (NABirds and FGVC Aircraft (Maji et al., 2013)) and GTSRB. Compared with vanilla feature matching, our method consistently improves performance across all datasets and shot settings. The gains are particularly pronounced on the fine-grained datasets, suggesting that balancing feature learning across spectral directions is especially beneficial for tasks requiring subtle visual discrimination. Moreover, while the distilled representations consistently outperform the original student representations, the improvements are larger in the few-shot regime than in the full-shot setting. As more labeled data becomes available, the performance gap gradually narrows.

Analysis of the distance function. Previous feature matching methods have employed L1, L2, or cosine distance. We therefore compare these objectives in the right part of Table 3. SpecMatch achieves either the best or second-best performance on all datasets, yielding the strongest performance on average. In contrast, none of the conventional distance functions consistently performs well across datasets: each can be effective on certain tasks but performs relatively poorly on others. These results demonstrate the robustness of SpecMatch across diverse downstream tasks.

Does whitening target features improve performance? Given the discussion in Sec. 3.2, a natural alternative for balancing the training signal is to whiten the target features, i.e., normalizing each spectral direction by its variance, as studied by Ranzinger et al. (2024a). We therefore train the student to predict whitened teacher features after PCA projection using an MSE loss. The right of Table 3 shows that whitening does not necessarily improve the performance over the L2 distance. We hypothesize that this is because whitening completely equalizes the training signal, disregarding the relative importance of high-variance directions that may capture prominent variations in the data. In contrast, SpecMatch reduces the imbalance across spectral directions while preserving their relative importance, which we find more effective for downstream performance.

Table 4: Study on imbalanced data distribution using PE-Core-G14 and PE-Core-Tiny as teacher and student, respectively. ***(Left)*:** Anomaly detection performance (AUROC) on ViSA with different numbers of anomaly samples used for distillation. ***(Right)*:** Results on iNaturalist2018 across head, medium, and tail classes.

Method	ViSA # Anomaly Samples			
	100	200	500	1000
Teacher	0.898			
Student	0.825			
FM	0.835	0.851	0.863	0.880
SpecMatch	**0.851**	**0.867**	**0.873**	**0.896**

Method	iNaturalist2018			
	Head	Medium	Tail	All
Teacher	86.7	78.0	73.7	77.2
Student	54.2	47.2	44.5	46.8
FM	60.3	51.8	51.0	52.3
SpecMatch	**61.3**	**52.7**	**52.0**	**53.3**

Table 5: Protein downstream performance on six PFMBench tasks. Results are scaled by 100 and reported as mean $\pm$ standard deviation over three runs. Higher is better for all metrics.

Method	Fold: 1195 classes		GO-MF: 489 classes		GO-CC: 320 classes		EC: 585 classes		Cat. Eff. : regression		DeepLoc2: 10 classes	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
Teacher (ESM-2 650M)	$75.0_{\pm 0.1}$	$73.1_{\pm 0.3}$	$55.6_{\pm 0.2}$	$55.5_{\pm 0.4}$	$45.7_{\pm 0.4}$	$47.6_{\pm 0.2}$	$58.8_{\pm 0.2}$	$59.8_{\pm 0.2}$	$7.5_{\pm 0.4}$	$10.6_{\pm 0.5}$	$71.7_{\pm 0.5}$	$70.9_{\pm 0.1}$
Student (ESM-2 6M)	$49.4_{\pm 0.3}$	$51.2_{\pm 0.2}$	$34.5_{\pm 0.4}$	$36.6_{\pm 0.3}$	$30.4_{\pm 0.5}$	$29.7_{\pm 0.2}$	$38.7_{\pm 0.1}$	$40.1_{\pm 0.3}$	$14.3_{\pm 0.3}$	$17.0_{\pm 0.5}$	$64.3_{\pm 0.1}$	$62.9_{\pm 0.1}$
L1	$52.6_{\pm 0.3}$	$53.1_{\pm 0.3}$	$36.5_{\pm 0.1}$	$37.5_{\pm 0.2}$	$35.2_{\pm 0.1}$	$34.8_{\pm 0.1}$	$38.3_{\pm 0.2}$	$40.9_{\pm 0.4}$	$20.7_{\pm 1.3}$	$22.0_{\pm 0.8}$	$\mathbf{70.1}_{\pm 0.1}$	$68.4_{\pm 0.1}$
L2	$51.9_{\pm 0.2}$	$52.2_{\pm 0.3}$	$36.8_{\pm 0.2}$	$38.0_{\pm 0.2}$	$35.2_{\pm 0.2}$	$35.3_{\pm 0.3}$	$39.9_{\pm 0.4}$	$41.0_{\pm 0.4}$	$18.8_{\pm 1.3}$	$27.0_{\pm 0.8}$	$69.6_{\pm 0.1}$	$68.5_{\pm 0.1}$
SpecMatch	$\mathbf{54.4}_{\pm 0.6}$	$\mathbf{55.6}_{\pm 0.5}$	$\mathbf{38.6}_{\pm 0.2}$	$\mathbf{39.6}_{\pm 0.1}$	$\mathbf{35.8}_{\pm 0.2}$	$\mathbf{35.8}_{\pm 0.2}$	$\mathbf{41.7}_{\pm 0.3}$	$\mathbf{42.0}_{\pm 0.6}$	$\mathbf{21.3}_{\pm 0.6}$	$\mathbf{30.7}_{\pm 1.0}$	$69.9_{\pm 0.3}$	$\mathbf{69.2}_{\pm 0.1}$

Hyper-parameter sensitivity. The left part of Table 3 studies the effect of β. Increasing β^{-1} from 1 progressively balances the contributions of different spectral directions. All tested values improve performance over the unweighted baseline, while the differences among them remain relatively small. Thus, the proposed method is not highly sensitive to β within the evaluated range.

Robustness to imbalanced class distribution. Table 4 analyzes the effectiveness of SpecMatch under imbalanced unlabeled data distributions. On the left, we progressively reduce the number of unlabeled anomaly samples used during distillation. As the number of anomaly samples decreases, the overall performance gradually drops, likely because anomaly-specific feature patterns are observed less frequently during training and therefore receive fewer optimization updates. Nevertheless, SpecMatch consistently outperforms vanilla feature matching across all settings, demonstrating its robustness even when anomaly samples are scarce. On the right of Table 4, we assess on iNaturalist2018 (Van Horn et al., 2018), a long-tailed recognition dataset with 8,142 categories. SpecMatch consistently improves over the baseline on all metrics. These indicate that the proposed objective remains effective under highly imbalanced class distributions and learns more discriminative representations for both well-represented and under-represented classes.

Experiments on protein embedding model. SpecMatch also generalizes to protein foundation models. We distill ESM-2 650M into ESM-2 6M (Lin et al., 2023) and evaluate the learned representations via linear probing on six PFMBench tasks (Gao et al., 2025). Fold denotes protein fold classification on the Remote Homology benchmark (Lo Conte et al., 2000). GO-MF and GO-CC are multi-label Gene Ontology prediction tasks for molecular function and cellular component, respectively (Ashburner et al., 2000). EC denotes multi-label enzyme commission number prediction (Bairoch, 2000). Cat. Eff. measures enzyme catalytic efficiency prediction using Spearman correlation (Li et al., 2022), while DeepLoc2 evaluates multi-label protein subcellular localization prediction (Thumuluri et al., 2022). As in the image experiments, distillation is performed separately on the training set of each task, and evaluation follows the official protocol and metrics. The consistent improvements across these diverse tasks suggest that useful information is distributed beyond high-variance components in both image and protein representations. Thus, spectral balancing provides a robust feature-distillation objective that generalizes across modalities.

6 Conclusion

We presented Spectrum-Balanced Feature Matching, SpecMatch, a simple and effective objective for feature distillation that addresses the imbalance of optimization across spectral directions. By dynamically reweighting the reconstruction loss according to the reconstruction error of each principal component, the proposed objective encourages the student to learn richer feature representations while introducing negligible computational overhead. Extensive experiments on diverse downstream tasks demonstrate that SpecMatch consistently improves over conventional feature matching across a wide range of teacher–student combinations and datasets. We hope that this work provides a simple yet effective direction for improving feature distillation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing code and with manuscript writing/editing. All AI-assisted outputs were reviewed by the authors. In particular, AI-assisted code was inspected and verified by the authors, and AI-assisted text was edited and checked for accuracy, consistency with the experiments, and originality. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies knowledge distillation and representation transfer using publicly available datasets and pretrained models. Our experiments do not involve human subjects or the collection of personally identifiable or sensitive information. We follow the licenses and intended research use of the datasets and models used in our experiments. We do not identify any direct ethical concerns specific to the proposed methodology beyond those generally associated with the underlying pretrained models and datasets.

REPRODUCIBILITY STATEMENT

The Pytorch-style code of the proposed objective is shown in Algorithm 1. The overview of the experimental setup is described in Sec. 5.1. More specific details of the experiments, e.g., models and hyper-parameters, are described in the Sec. C. We will also release the code used for our experiments upon acceptance.

REFERENCES

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Amos Bairoch. The enzyme database in 2000. *Nucleic acids research*, 28(1):304–305, 2000.

Peter Bandi, Oscar Geessink, Quirine Manson, Mayer Reiter, Marc Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Brenda Bandettini di Poggio, Iris Rutgers, Geert Litjens, and et al. From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2019. doi: 10.1109/TMI.2018.2867350.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, 2018.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Bangalath, et al. Perception encoder: The best visual embeddings are not at the output of the network. *NeurIPS*, 38:60884–60937, 2026.

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

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. In *CVPR*, 2021.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, Oct 2017. ISSN 1558-2256. doi: 10.1109/jproc.2017.2675998. URL http://dx.doi.org/10.1109/JPROC.2017.2675998.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *NeurIPS*, 38:48009–48036, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Zhangyang Gao, Hao Wang, Cheng Tan, Chenrui Xu, Mengdi Liu, Bozhen Hu, Linlin Chao, Xiaoming Zhang, and Stan Z Li. Pfmbench: Protein foundation model benchmark. *arXiv preprint arXiv:2506.14796*, 2025.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *ICML*. PMLR, 2023.

Bobby He and Mete Ozay. Exploring the gap between collapsed & whitened features in self-supervised learning. In *ICML*. PMLR, 2022.

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*, 2019.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, pp. 1921–1930, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *ICCV*, 2019.

Jinseong Jang, Chunfei Ma, and Byeongwon Lee. Vl2lite: Task-specific knowledge distillation from large vision-language models to lightweight networks. In *CVPR*, 2025.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

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

Jungsoo Lee, Debasmit Das, Munawar Hayat, Sungha Choi, Kyuwoong Hwang, and Fatih Porikli. Customkd: Customizing large vision foundation for edge model improvement via knowledge distillation. In *CVPR*, 2025.

Seung Hyun Lee, Dae Ha Kim, and Byung Cheol Song. Self-supervised knowledge distillation using singular value decomposition. In *ECCV*, 2018.

Feiran Li, Le Yuan, Hongzhong Lu, Gang Li, Yu Chen, Martin KM Engqvist, Eduard J Kerkhoven, and Jens Nielsen. Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nature catalysis*, 5(8):662–672, 2022.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Loredana Lo Conte, Bart Ailey, Tim JP Hubbard, Steven E Brenner, Alexey G Murzin, and Cyrus Chothia. Scop: a structural classification of proteins database. *Nucleic acids research*, 28(1): 257–259, 2000.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

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

Roy Miles and Krystian Mikolajczyk. Understanding the role of the projector in knowledge distillation. In *AAAI*, volume 38, pp. 4233–4241, 2024.

Roy Miles, Ismail Elezi, and Jiankang Deng. $v_\{k\}d$: Improving knowledge distillation using orthogonal projections. In *CVPR*. IEEE, 2024.

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. *arXiv preprint arXiv:2304.07193*, 2023.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Mike Ranzinger, Jon Barker, Greg Heinrich, Pavlo Molchanov, Bryan Catanzaro, and Andrew Tao. Phi-s: Distribution balancing for label-free multi-teacher distillation. *arXiv preprint arXiv:2410.01680*, 2024a.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, 2024b.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets, 2015. URL `https://arxiv.org/abs/1412.6550`.

Mert Bülent Sarıyıldız, Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Unic: Universal classification models via multi-teacher distillation. In *ECCV*. Springer, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Vineet Thumuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, and Ole Winther. Deeploc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic acids research*, 50(W1):W228–W234, 2022.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

Raviteja Vemulapalli, Hadi Pouransari, Fartash Faghri, Sachin Mehta, Mehrdad Farajtabar, Mohammad Rastegari, and Oncel Tuzel. Knowledge transfer from vision foundation models for efficient training of small task-specific models. In *ICML*, 2024.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *CVPR*, 2019.

Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *ICLR*, volume 2024, pp. 47812–47831, 2024.

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.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*. PMLR, 2021.

Yitian Zhang, Xu Ma, Yue Bai, Huan Wang, and Yun Fu. Accessing vision foundation models via imagenet-1k. In *ICLR*, 2025.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*, 2022.

A Appendix

B Theoretical Motivation

Insights into vanilla feature matching. We analyze conventional feature matching from the perspective of the principal components of the teacher representation. Let $z_b^t \in \mathbb{R}^d$ denote the centered teacher representation for the b-th sample, and let $\tilde{z}_b^s \in \mathbb{R}^d$ denote the corresponding student prediction. We apply PCA to the teacher representations and denote the resulting orthonormal principal directions by $\{u_k\}_{k=1}^d$. The teacher representation can then be written as

$$z_b^t = \sum_{k=1}^{d} c_{b,k}^t u_k, \tag{8}$$

where

$$c_{b,k}^t = u_k^\top z_b^t. \tag{9}$$

We similarly denote the student prediction along the k-th principal direction by $\tilde{c}_{b,k}^s$. Since the PCA basis is orthonormal, projecting both representations onto this basis does not change the conventional feature matching objective:

$$\left\|\tilde{\boldsymbol{z}}_b^s - \boldsymbol{z}_b^t\right\|_2^2 = \sum_{k=1}^{d} \left(\tilde{c}_{b,k}^s - c_{b,k}^t\right)^2. \tag{10}$$

Thus, expressing feature matching in the PCA basis is equivalent to conventional feature matching.

The standard feature matching objective is given by

$$\mathcal{L}_{\mathrm{FM}} = \sum_{k=1}^{d} e_k, \qquad e_k = \frac{1}{B}\sum_{b=1}^{B} \left(\tilde{c}_{b,k}^s - c_{b,k}^t\right)^2, \tag{11}$$

where e_k denotes the batch-averaged reconstruction error along the k-th principal direction.

Reconstruction error and teacher variance. Let

$$\lambda_k = \mathrm{Var}(c_k^t), \qquad v_k = \mathrm{Var}(\tilde{c}_k^s), \tag{12}$$

and let ρ_k denote the correlation coefficient between c_k^t and $\tilde{c}_k^s$. Furthermore, define the difference between their means as

$$b_k = \mathbb{E}[\tilde{c}_k^s] - \mathbb{E}[c_k^t]. \tag{13}$$

The expected reconstruction error satisfies

$$\begin{aligned} \mathbb{E}[e_k] &= \mathbb{E}\left[\left(\tilde{c}_k^s - c_k^t\right)^2\right] \\ &= \lambda_k + v_k - 2\rho_k\sqrt{\lambda_k v_k} + b_k^2. \end{aligned} \tag{14}$$

At the early stage of training, the student prediction is largely uninformative about the teacher principal directions. It is therefore reasonable to assume that

$$\rho_k \approx 0, \qquad v_k \approx v, \qquad b_k \approx b, \tag{15}$$

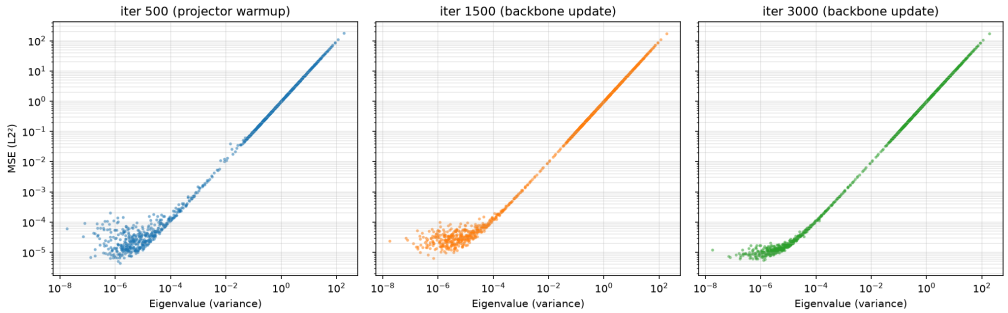


Figure 5: Eigenvalues (x-axis) and reconstruction error.

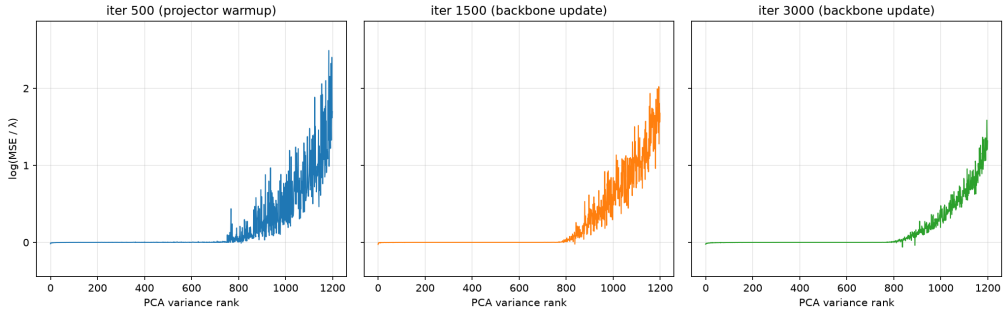


Figure 6: Reconstruction error along principal components.

where v and b are approximately constant across principal components. Under these assumptions,

$$\mathbb{E}[e_k] \approx \lambda_k + C, \qquad C = v + b^2. \tag{16}$$

Therefore, at the early stage of training, the reconstruction error is approximately an affine function of the teacher variance.

Dominance in optimization. Let

$$r_{b,k} = \tilde{c}^s_{b,k} - c^t_{b,k} \tag{17}$$

denote the residual along the k-th principal direction. The gradient of the feature matching objective with respect to the student prediction is

$$\frac{\partial \mathcal{L}_{\mathrm{FM}}}{\partial \tilde{c}^s_{b,k}} = \frac{2}{Bd} r_{b,k}. \tag{18}$$

Consequently, the squared norm of the output-space gradient associated with the k-th principal direction is

$$\begin{aligned} \left\| \nabla_{\tilde{\mathbf{c}}^s_{:,k}} \mathcal{L}_{\mathrm{FM}} \right\|_2^2 &= \sum_{b=1}^{B} \left(\frac{2}{Bd} r_{b,k} \right)^2 \\ &= \frac{4}{Bd^2} e_k. \end{aligned} \tag{19}$$

Thus, the gradient magnitude contributed by each principal direction is directly proportional to its reconstruction error. Combining Equation 16 and 19 gives

$$\mathbb{E}\left[\left\| \nabla_{\tilde{\mathbf{c}}^s_{:,k}} \mathcal{L}_{\mathrm{FM}} \right\|_2^2 \right] \approx \frac{4}{Bd^2} \left(\lambda_k + C \right). \tag{20}$$

Hence, high-variance principal components produce larger reconstruction errors and stronger optimization signals, causing conventional feature matching to preferentially fit the dominant spectral directions of the teacher representation.

Algorithm 1 PyTorch-style pseudocode for SpecMatch.

```
def specmatch_loss(student, teacher, beta_inv=3.0, eps=1e-3):
    """
    student, teacher: PCA coefficients shaped (B, D)
    teacher: teacher features projected onto the PCA basis (B, D)
    """
    # reconstruction error for each principal direction (D,)
    error = (student - teacher).pow(2).mean(dim=0)
    beta = 1.0 / beta_inv
    # mitigate the imbalance across principal directions
    loss = (error + eps).pow(beta).mean()
    return loss
```

Table 6: Dataset statistics.

Dataset	Train	Test	# Classes
CIFAR-100	50,000	10,000	100
CIFAR-10	50,000	10,000	10
SVHN	73,257	26,032	10
GTSRB	26,640	12,630	43
RESISC45	18,900	6,300	45
EuroSAT	21,600	5,400	10
OrganCMNIST	12,975	8,216	11
Camelyon17	302,436	85,054	2
VisA	10,821	962 normal + 1,200 anomaly	2
MVTec	5,354	467 normal + 1,258 anomaly	2
iWildCam	129,809	42,791	182
FMoW	76,863	22,108	62
CUB	5,994	5,794	200
NABirds	23,929	24,633	555

Empirical analysis. Fig. 5 describes the eigenvalues and corresponding error in the axis. Fig. 6 describes the reconstruction error along principal components with L2-distance feature matching loss. Fig. 5 empirically supports the Equation 16, i.e., the error is proportional to the eigenvalues. Also, the trend is consistent across different training iterations.

C Experimetal Details

We provide additional details on the datasets and experimental settings used in our experiments. Table 6 summarizes the datasets used in the experiments reported in Table 1. We report the number of training and test samples and the number of classes for each dataset. Similarly, Table 7 summarizes the statistics of the protein datasets used in the experiments reported in Table 5.

Experimental settings for Table 1. We use AdamW as the optimizer in all experiments. The learning rate for the backbone is set to 1×10^{-4} by default, while we use a smaller learning rate of 1×10^{-5} for MVTec and iWildCam. The learning rate for the projector is set to 1×10^{-3}. We first trained only the linear projector for 500 iterations as a warm-up, and then jointly trained the backbone and projector for an additional 2,500 iterations. These hyperparameters are kept fixed across all three teacher–student combinations to ensure a consistent comparison. We set ϵ as 1e-3 and 1e-4 for the training on PE-Core-Tiny and MobileNet-V3, respectively.

Experiments on ViSA and MvTec. During the distillation training using unlabeled data, we employ both normal and anomaly samples as training data. For the experiments on ViSA, we employ a few labeled normal and anomaly data to train a linear classifier following their evaluation protocol. For the experiments on MvTec, we employ the nearest neighbor distance to the normal samples to compute the anomaly score.

Table 7: Statistics of the protein datasets used in Table 5.

Dataset	Task	Classes	Train	Val	Test	Epochs
Fold prediction	classification	1195	13034	1628	1630	20
GO molecular function	multi labels classification	489	22291	2785	2787	20
GO cellular component	multi labels classification	320	11196	1398	1400	20
Enzyme commission number	multi labels classification	585	12928	1615	1616	60
Enzyme catalytic efficiency	regression	-	10363	1290	1298	60
DeepLoc2Multi	multi labels classification	10	21949	2743	2744	50

Table 8: Overview of teacher and student vision models used in our experiments. Parameter counts correspond to the vision backbone only, excluding classification heads and text encoders.

Role	Model	Source	Pre-training	Input	Params.
Teacher	DINOv2 ViT-g/14	DINOv2	DINOv2	518	1.1B
	PE-Core-G14-448	Meta PE	Perception Encoder	448	1.9B
	CLIP ConvNeXt-XXLarge	OpenCLIP	LAION-2B	256	846.5M
Student	MobileNetV3-Small	timm	IN-1K	224	1.5M
	PE-Core-T16-384	Meta PE	Perception Encoder	384	10M
	ConvNeXt-Tiny	timm	Scratch	224	27.8M
	ViT-Tiny/16	timm	Scratch	224	5.5M

[†]ViT-Tiny/16 is also trained from scratch in experiments where specified. Links in the Source column point to the model repository or the exact checkpoint used.

Experiments on ImageNet. We use a batch size of 512 and train for 200,000 iterations. When initializing the model with PE-Core-Tiny, we use a fixed learning rate of 1×10^{-5} throughout training. For training from scratch, we use an initial learning rate of 1×10^{-3} and decay the learning rate using a cosine schedule.

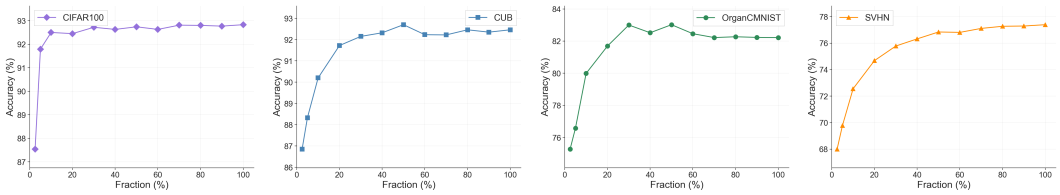
Protein Experiments. For the protein-related experiments in Table 5, we use AdamW with a learning rate of 1×10^{-4}. The learning rate is kept fixed throughout training. Since the number of training samples varies substantially across the protein datasets, we adjust the number of training epochs for each dataset accordingly. The dataset-specific training configurations are summarized in Table 7.

D Additional Results

The number of principal components and linear probe accuracy. In relation to Figure 1, Fig. 7 studies how linear-probing accuracy changes as the number of principal components increases. For tasks such as CIFAR-100, high performance was achieved using only a small number of components. For many other tasks, however, performance continued to improve as up to approximately 50% of the components were included. On SVHN, performance improved further as even more components were used. These results indicate that low-variance principal components can contain information useful for certain downstream tasks, rather than merely representing noise.

Comparison between dynamic and static loss weighting. In Sec 4 (Eq. 7), we mention the alternative formulation to realize the spectral balancing, which aims to mitigate the spectral bias by using the variance in the target representations. Table 11 compares the results by setting $\beta^{-1} = 3$ in both loss. Overall, the dynamic weighting scheme used in the main paper achieves better performance. On ImageNet, however, static weighting performs slightly better, suggesting that the difference between the two schemes is modest. Importantly, both weighting schemes consistently outperform the unweighted L2 loss, demonstrating that variance-based loss weighting improves performance.

Instance recognition. To further evaluate the quality of the representations learned on ImageNet, we assess the ConvNeXt-Tiny model trained in the ImageNet experiments on the Oxford and Paris instance recognition benchmarks. Since this task requires discriminative instance-level representations rather than category-level classification, it provides a complementary evaluation of representation quality. As shown in Table 10, SpecMatch consistently outperforms vanilla feature matching across most metrics, indicating that it better preserves the teacher's rich feature representations.



(a) CIFAR100 (b) CUB (c) OrganCMNIST (d) SVHN

Figure 7: Relationship between linear probe accuracy and the number of used principal components. We measure the linear probe accuracy in using top K% principal components.

Table 9: Comparison with fully supervised fine-tuning. All methods use the same pretrained MobileNetV3-Small student. Δ denotes the performance difference between SpecMatch and supervised fine-tuning.

Method	CIFAR100	CIFAR10	SVHN	GTSRB	NABirds	CUB	RESISC	EuroSAT	OrganCMNIST	Camelyon17	iWildCam	FMoW
Supervised FT	76.5	93.8	**95.2**	**95.4**	50.1	55.8	93.0	**98.4**	**93.4**	90.1	58.5	41.0
SpecMatch (T: PE-Core)	77.6	95.2	89.5	94.6	**64.2**	**72.5**	**93.6**	98.2	90.7	**93.9**	**62.4**	**42.1**
Δ	+1.1	+1.5	-5.7	-0.8	+14.1	+16.7	+0.6	-0.2	-2.7	+3.8	+3.9	+1.1
SpecMatch (T: DINOv2)	78.1	95.5	82.3	88.1	56.8	69.0	92.8	98.0	91.0	93.4	61.1	39.5
Δ	+1.6	+1.8	-12.9	-7.3	+6.7	+13.2	-0.2	-0.4	-2.4	+3.3	+2.6	-1.5

DINO outperforms SpecMatch on many metrics, in contrast to the results on ImageNet (Table 2). This suggests that DINO is more effective at capturing fine-grained, instance-level details, whereas SpecMatch is better suited to transferring class-discriminative features.

Table 10: Instance recognition results on the Oxford and Paris datasets. A ConvNeXt-Large model pre-trained with image-text data is used as the teacher, while ConvNeXt-Tiny serves as the student.

Method	Oxford									Paris								
	Easy			Medium			Hard			Easy			Medium			Hard		
	mAP	P@1	P@10	mAP	P@1	P@10	mAP	P@1	P@10	mAP	P@1	P@10	mAP	P@1	P@10	mAP	P@1	P@10
ConvNEXT-XXlarge → ConvNEXT-Tiny																		
DINO	**21.9**	**41.2**	**27.7**	15.0	**45.7**	**27.3**	3.4	17.1	5.3	**64.8**	**95.7**	**87.7**	**49.9**	**97.1**	**91.3**	**24.7**	**68.6**	**56.0**
FM	17.6	29.4	20.9	13.8	30.0	18.7	3.7	10.0	7.0	53.4	87.1	81.0	40.5	88.6	82.7	17.4	54.3	41.6
SpecMatch	19.7	30.9	22.4	**15.7**	38.6	22.3	**4.9**	**18.6**	**8.6**	55.5	87.1	81.6	43.7	88.6	84.4	21.5	67.1	49.9
PE-Core-G14 → ViT-Tiny																		
DINO	**19.0**	**36.7**	**28.1**	**14.4**	**37.1**	**26.9**	**2.9**	**8.6**	**4.0**	**60.3**	**95.7**	**88.3**	**45.2**	**98.6**	**90.4**	**19.8**	**68.6**	**51.4**
FM	10.0	14.7	14.9	9.1	18.6	16.4	2.2	5.7	3.1	30.3	75.7	64.7	24.7	77.1	66.9	9.1	31.4	19.4
SpecMatch	11.2	29.4	14.3	9.6	31.4	15.3	2.4	5.7	2.7	32.9	77.1	64.9	26.8	78.6	67.6	11.2	35.7	23.4

Table 11: Comparison between static and dynamic weighting.

Method	CIFAR-10	SVHN	NABirds	CUB	RESISC	ImageNet
FM	94.8	87.6	62.6	71.3	92.6	72.1
Static	95.0	88.3	63.7	71.7	93.6	**73.2**
Dynamic	**95.2**	**89.5**	**64.2**	**72.5**	93.6	73.1