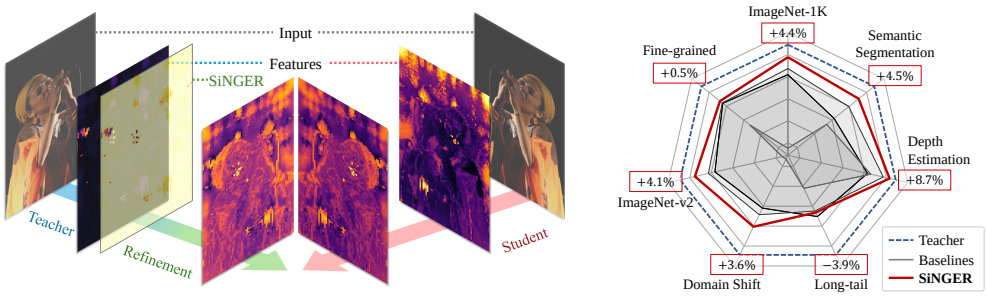
SiNGER: A Clearer Voice Distills Vision Transformers Further

Anonymous authors
Paper under double-blind review

Abstract

Vision Transformers are widely adopted as the backbone of vision foundation models, but they are known to produce high-norm artifacts that degrade representation quality. When knowledge distillation transfers these features to students, high-norm artifacts dominate the objective, so students overfit to artifacts and underweight informative signals, diminishing the gains from larger models. Prior work attempted to remove artifacts but encountered an inherent trade-off between artifact suppression and preserving informative signals from teachers. To address this, we introduce **Si**ngular **N**ullspace-**G**uided **E**nergy **R**eallocation (SiNGER), a novel distillation framework that suppresses artifacts while preserving informative signals. The key idea is principled teacher feature refinement: during refinement, we leverage the nullspace-guided perturbation to preserve information while suppressing artifacts. Then, the refined teacher's features are distilled to a student. We implement this perturbation efficiently with a LoRA-based adapter that requires minimal structural modification. Extensive experiments show that SiNGER consistently improves student models, achieving state-of-the-art performance in multiple downstream tasks and producing clearer and more interpretable representations.



(a) Overview of SiNGER distillation.

(b) Performance gains using SiNGER.

Figure 1: SiNGER suppresses artifacts and enhances transfer. (a) Feature visualizations highlight clearer and more interpretable representations. (b) Radar chart shows consistent multi-task gains.

1 Introduction

Transformers have become the de facto standard architecture in both research and industry due to their scalability and effectiveness (Oquab et al., 2024; Radford et al., 2021). Their token-based self-attention mechanism is broadly applicable with minimal inductive bias (Lu et al., 2022), and has enabled significant advances in computer vision and machine learning (Kim et al., 2024; Sariyildiz et al., 2024). Vision Transformers (ViTs, Dosovitskiy et al. (2021)) extend this paradigm to visual data and form the backbone of Vision Foundation Models (VFMs). Compared to convolutional networks, ViTs rely less on spatial inductive biases and instead exploit scale to achieve high performance. ViTs are also highly scalable since supervised or self-supervised training produces increasingly generalizable representations (Oquab et al., 2024; Touvron et al., 2022). However, the quadratic complexity of self-attention severely limits the practicality of scaling ViTs. This tension between accuracy and efficiency motivates the study of compression. Pruning (Yang et al., 2023) and quantization (Liu et al., 2021) have been explored, but pruning often fails to deliver practical speedup due to structural rigidity (Aghli & Ribeiro, 2021), and quantization can induce numerical

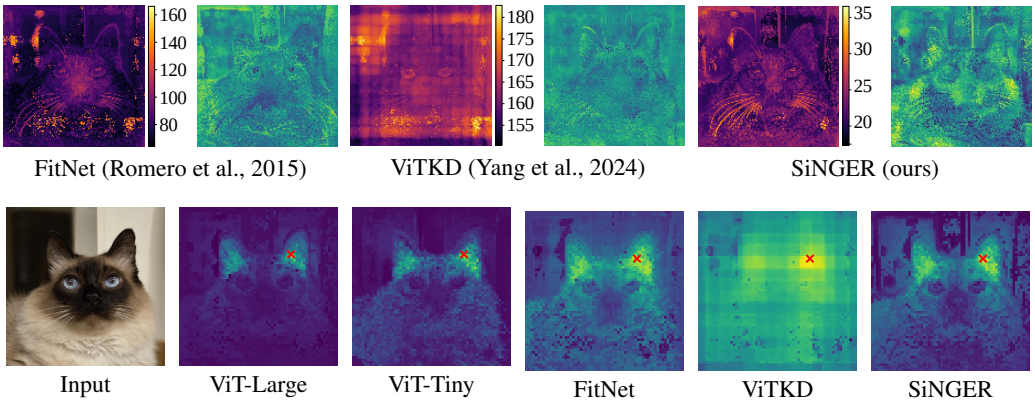


Figure 2: Qualitative analysis. **Row 1**: KD method comparison. Left: distilled feature map colored by patch norm, Right: patch-wise cosine similarity to the teacher. **Row 2**: Input image, two pretrained ViTs, and three ViT-L $\rightarrow$ ViT-T distilled variants. Each panel shows similarity from the $\times$-marked patch. SiNGER most closely preserves teacher semantics, showing the most coherent teacher-consistent similarity patterns.

instability (Jiang et al., 2024; Javed et al., 2024). Knowledge distillation (KD, Hinton et al. (2015)) has emerged as the most reliable solution for transferring knowledge from large ViTs to smaller students. KD methods span diverse targets and frameworks (Sun et al., 2024; Romero et al., 2015; Ranzinger et al., 2024), consistently yielding structurally and numerically stable compact models.

Nevertheless, KD for ViTs suffers from subtle but critical limitations in their representation space. Darcet et al. (2024) revealed that ViT token representations contain high-norm artifacts. Wang et al. (2025) argue that these artifacts are singular defects induced by power-iteration-like accumulation across residual blocks, whereby tokens align with the leading left singular vector of the pre-trained weights. These artifacts interact poorly with the standard feature mean squared error objective in KD: when the teacher and student are matched, gradients concentrate on the few high-norm tokens, producing an outlier-driven optimization bias that obscures informative signals in the inlier structure. Therefore, suppressing outlier norms in teacher features is essential for KD in ViTs as the scale grows. Prior work mitigated this issue via random masking of teacher features (Yang et al., 2024); however, this inevitably removes informative signals. Therefore, a key challenge is to mitigate these artifacts without losing valuable information, a fundamental trade-off that requires a principled approach.

To resolve this trade-off, we introduce a nullspace-guided suppression: we modify only the nullspace component in the teacher features, mathematically, the subspace orthogonal to the downstream space. This yields student-optimal supervision by suppressing artifacts without sacrificing informative signals. Based on this insight, we propose **Si**ngular **N**ullspace-**G**uided **E**nergy **R**eallocation (SiNGER), a framework that addresses this trade-off in ViTs distillation, illustrated in Figure 1a. To minimize the modification to the teacher's signal, we attach a lightweight LoRA-based adapter (Hu et al., 2022) to the KD architecture, which refines the teacher features. The adapter produces a minimal perturbation guided toward the left-nullspace of the next block, suppressing high-norm outliers while leaving the next block output unchanged. Our method achieves superior performance compared to baselines across multiple downstream tasks (Figure 1b). It also produces more structured and interpretable feature maps (Figure 2). Our contributions are summarized as follows:

- We propose a novel distillation framework (SiNGER) that refines teacher signals via the LoRA-based adapter with nullspace initialization to guide effective perturbations.
- We analyze a fundamental limitation of naïve ViT distillation, showing degraded transfer on downstream benchmarks along with qualitative evidence.
- We provide extensive ablation studies to analyze the contribution of each component in SiNGER and validate the robustness of our framework.
- We demonstrate through extensive experiments that our method exceeds baseline performance across tasks and produces more interpretable feature maps.

2 Related Works

Vision Transformers. ViTs (Dosovitskiy et al., 2021) underpin many VFMs and have become the representative architecture for large-scale visual learning. Unlike convolutional networks (He et al., 2016; Howard et al., 2017), ViTs rely on self-attention with fully connected layers. Since ViTs form the architectural core of most VFMs, studying them provides representative insights that generalize broadly. Similarly, parameter-efficient tuning methods such as LoRA (Hu et al., 2022) highlight how minimal perturbations can effectively adapt VFMs, a perspective that motivates our artifact-suppressing perturbations. However, the quadratic complexity of the ViT architecture limits the practicality of large models despite their advanced representation.

Knowledge Distillation. KD compresses models by training a smaller student to mimic a larger teacher (Hinton et al., 2015). Among various approaches, FitNet-style methods (Romero et al., 2015) that align intermediate features are especially influential, as they encourage the student to learn useful representations beyond logits (Sun et al., 2024). Later extensions incorporated relational structures (Park et al., 2019) or multi-teacher settings (Ranzinger et al., 2024), while adaptations for ViTs (Touvron et al., 2022) aimed to respect their architectural characteristics. Despite these advances, distillation applied to VFMs often inherits undesirable properties from teachers, revealing the need for methods that improve not only compression but also the quality of transferred representations.

Artifacts in Transformers. Artifacts are a recurring issue in transformer models, degrading the representation quality. Darcet et al. (2024) demonstrated that ViTs produce high-norm artifacts, particularly in background regions, harming interpretability and dense prediction. Practical suppression strategies include register tokens (Darcet et al., 2024), and recent work argues these artifacts arise from power-method-like accumulation across residual layers, aligning tokens with the leading left singular vector (Wang et al., 2025). In the knowledge distillation domain, ViTKD (Yang et al., 2024) randomly masks teacher features to reduce the mimicking of high-norm artifacts. However, such indiscriminate masking also removes informative inlier signals, motivating artifact-aware KD that suppresses high-norm artifacts while preserving inlier structure.

3 Method

3.1 Problem Formulation

3.1.1 High-Norm Outliers in Vision Transformers

In large ViTs, a non-negligible fraction of patch features in F_l^{T} exhibit high-norm artifacts (outliers). Their prevalence and magnitude increase with model capacity, as Darcet et al. (2024) reported. This is particularly consequential for distillation from a larger teacher to a smaller student: artifact-prone teacher features introduce a systematic imbalance at the feature level and can obscure informative signals.

3.1.2 Knowledge Distillation Objective

Let $F_l^{\mathrm{T}} \in \mathbb{R}^{n\times d^{\mathrm{T}}}$ and $F_l^{\mathrm{S}} \in \mathbb{R}^{n\times d^{\mathrm{S}}}$ denote teacher (T) and student (S) features at layer l. We align dimensions $d^{\mathrm{S}} \to d^{\mathrm{T}}$ with a trainable projection $P_l : \mathbb{R}^{d^{\mathrm{S}}} \to \mathbb{R}^{d^{\mathrm{T}}}$ and define the feature-level KD loss as follow:

$$\mathcal{L}_{\mathrm{KD},l} = \frac{1}{n}\sum_{i=1}^{n} \|F_{l,i}^{\mathrm{T}} - P_l(F_{l,i}^{\mathrm{S}})\|^2, \tag{1}$$

where $F_{l,i}^{\mathrm{T}}$ and $F_{l,i}^{\mathrm{S}}$ denote the i-th patch feature and n is the number of patches and $\|\cdot\|$ means ℓ_2-norm.

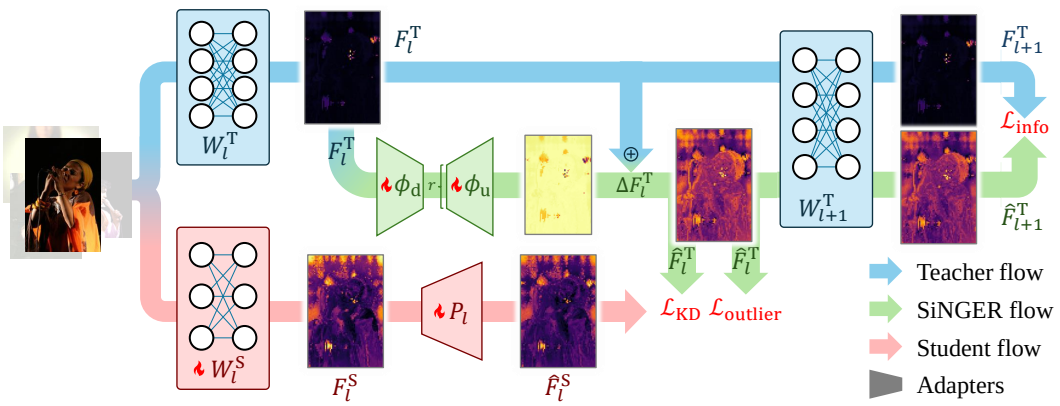


Figure 3: The overall pipeline of knowledge distillation with the SiNGER adapter at lth layer .

3.1.3 Outlier Dominance and Gradient Bias

Partition the patch indices into an outlier set $\mathcal{O}_l$ and an inlier set $\mathcal{I}_l$ to obtain

$$\mathcal{L}_{\mathrm{KD},l} = \frac{1}{n}\left(\underbrace{\sum_{i\in\mathcal{O}_l} \|F_{l,i}^{\mathrm{T}} - P_l(F_{l,i}^{\mathrm{S}})\|^2}_{\text{Outlier Term}} + \underbrace{\sum_{j\in\mathcal{I}_l} \|F_{l,j}^{\mathrm{T}} - P_l(F_{l,j}^{\mathrm{S}})\|^2}_{\text{Inlier Term}}\right). \tag{2}$$

By construction, for $i \in \mathcal{O}_l$ we have $\|F_{l,i}^{\mathrm{T}}\| \gg \|F_{l,j}^{\mathrm{T}}\|$ for $j \in \mathcal{I}_l$. Hence, when the residual magnitudes are of similar order across patches, the outlier term dominates both the objective and its gradients. In particular,

$$\nabla_{P_l(F_{l,i}^{\mathrm{S}})}\,\mathcal{L}_{\mathrm{KD},l} = \frac{2}{n}\big(P_l(F_{l,i}^{\mathrm{S}}) - F_{l,i}^{\mathrm{T}}\big), \tag{3}$$

so outliers induce proportionally larger updates. Optimization is therefore biased toward mimicking a few high-norm outliers, rather than consolidating the majority inlier structure that carries most of the informative signals. This gradient bias disrupts the learning of the dominant inlier representation and leads to suboptimal transfer. We therefore seek to refine the teacher features F_l^{T} at layer l before distillation, so that they are more conducive to transferring informative signals to the student.

3.2 Singular Nullspace-Guided Energy Reallocation

We consider KD at layer l, where high-norm outliers in F_l^{T} induce gradient bias toward a few tokens. To prevent this, the outlier term to the Equation (2) must be weakened, which in practice means reducing the norm of outlier patches in F_l^{T}. However, naïve shrinkage erodes information carried by the larger teacher and can nullify the benefits of distillation.

3.2.1 Perturbation on Nullspace

Let $F_l^{\mathrm{T}} \in \mathbb{R}^{n\times d^{\mathrm{T}}}$ and define a refined feature map $\hat{F}_l^{\mathrm{T}} = F_l^{\mathrm{T}} + \Delta F_l^{\mathrm{T}}$. Our two objectives are:

1. **Suppress Outlier Norms.** Reduce the norm of high-norm patches in F_l^{T} (Figure 4a).
2. **Preserve Information.** Ensure that when the modified features are fed into the next teacher block, the conveyed information is not altered (Figure 4b).

Consider the next block at layer $l+1$ with transformation $W_{l+1} \in \mathbb{R}^{d^{\mathrm{T}}\times d^{\mathrm{T}}}$. Then $\hat{F}_l^{\mathrm{T}}$ preserves the next-block output if and only if

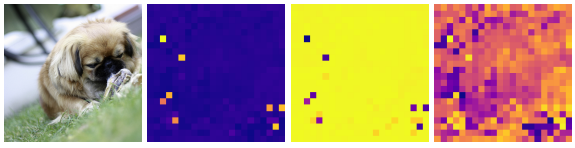
$$(F_l^{\mathrm{T}} + \Delta F_l^{\mathrm{T}})\,W_{l+1} \;=\; \hat{F}_l^{\mathrm{T}}\,W_{l+1} \quad\Longleftrightarrow\quad \Delta F_l^{\mathrm{T}}\,W_{l+1} \;=\; \mathbf{0}. \tag{4}$$

A perturbation ΔF_l^{T} that satisfies the above is obtained by restricting it to the left-nullspace of the next block W_{l+1}. Let us $N_{l+1} := \mathrm{Null}\big((W_{l+1})^{\top}\big)$ denote this left-nullspace. Then the requirement

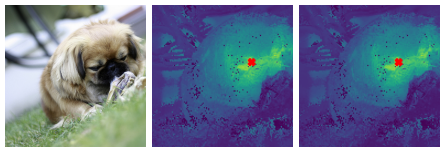
is as follows:

$$\mathrm{row}\big(\Delta F_l^{\mathrm{T}}\big) \subseteq N_{l+1}. \tag{5}$$

Consequently, to allow effective distillation, we refine the features of the teacher F_l^{T} to $\hat{F}_l^{\mathrm{T}}$ by a perturbation guided to the left-nullspace N_{l+1}.



(a) Outlier suppression with the proposed adapter.



(b) Information preservation after the $l+1$th layer.

Figure 4: Two objectives of SiNGER; (a) outlier suppression and (b) information preservation. $\|\Delta F_l^{\mathrm{T}}\|$ in (a) is signed with the cosine-similarity between ΔF_l^{T} and F_l^{T}. In (b), the cosine similarity between $\times$-marked patch and every patch of another feature map is visualized.

3.2.2 Adapter-based Feature Refinement

We refine F_l^{T} by adding a low-rank perturbation produced by a LoRA-based adapter while freezing all teacher weights.

$$\hat{F}_l^{\mathrm{T}} \;=\; F_l^{\mathrm{T}} + \Delta F_l^{\mathrm{T}}, \qquad \Delta F_l^{\mathrm{T}} \;=\; \big(F_l^{\mathrm{T}}\,\phi_{\mathrm{down},l}\big)\,\phi_{\mathrm{up},l}, \tag{6}$$

where $\phi_{\mathrm{down},l} \in \mathbb{R}^{d^{\mathrm{T}} \times r}$, $\phi_{\mathrm{up},l} \in \mathbb{R}^{r \times d^{\mathrm{T}}}$, and $r \ll d^{\mathrm{T}}$.

To bias ΔF_l^{T} toward the left-nullspace N_{l+1} of W_{l+1}, we set the initial weights of adapter,

$$\phi_{\mathrm{down},l} := N_{l+1}, \qquad \phi_{\mathrm{up},l} := N_{l+1}^{\top}. \tag{7}$$

This initialization guides the optimization to remain near N_{l+1} and to find solutions that satisfy the two objectives. Because the next block is nonlinear, its exact nullspace cannot be obtained via SVD. We adopt a practical linearization of the next block, $W_{l+1} \approx \tilde{W}_{l+1}$, and define $\tilde{N}_{l+1}$ as the left singular vectors associated with the r smallest singular values of $\tilde{W}_{l+1}$. By construction, $\tilde{N}_{l+1}$ collects the left singular vectors corresponding to the r smallest singular values of $\tilde{W}_{l+1}$, so $\|\tilde{N}_{l+1}^{\top}\tilde{W}_{l+1}\| = \sigma_{d-r+1}$. Moreover, Appendix A provides a detailed spectral analysis (sublayer perturbations, singular value diagnostics, and ε-null bounds) showing that the same approximate-null relation holds for the nonlinear block. Consequently,

$$\phi_{\mathrm{down},l} := \tilde{N}_{l+1}, \qquad \phi_{\mathrm{up},l} := \tilde{N}_{l+1}^{\top}. \tag{8}$$

3.3 Knowledge Distillation with SiNGER

Figure 3 summarizes the pipeline: SiNGER refines teacher features at selected layers before feature matching. Let $\mathcal{D} = l_{\mathrm{inter}} \cup \{l_{\mathrm{final}}\}$ denote the distillation layers, where l_{inter} is a set of intermediate layers and l_{final} is the final layer. For each $l \in \mathcal{D}$, an adapter ϕ_l transforms F_l^{T} into $\hat{F}_l^{\mathrm{T}}$. Training is guided by three losses, aggregated over $\mathcal{D}$.

Knowledge-Distillation Loss. The student is trained to mimic the refined teacher with $\mathcal{L}_{\mathrm{KD}}$.

$$\mathcal{L}_{\mathrm{KD}} = \sum_{l \in \mathcal{D}} \mathrm{MSE}\left(\hat{F}_l^{\mathrm{T}},\; P_l(F_l^{\mathrm{S}})\right). \tag{9}$$

Outlier Suppression Loss. Adapters are explicitly encouraged to suppress high-norm artifacts. For each l, let $\mathcal{O}_l$ be the indices of patches in $\hat{F}_l^{\mathrm{T}}$ whose norms exceed the α-percentile $q_{\alpha,l}$ as Equation 10.

$$\mathcal{L}_{\mathrm{outlier}} = \sum_{l \in \mathcal{D}} \frac{1}{|\mathcal{O}_l|} \sum_{i \in \mathcal{O}_l} \Big(\|\hat{F}_{l,i}^{\mathrm{T}}\|_2 - q_{\alpha,l} \Big)^2 \tag{10}$$

Information Preservation Loss. To retain informative signals while suppressing norms, we align feature directions via Gram matching. Define

$$\mathcal{L}_{\text{info},l} = \begin{cases} \text{MSE}\left(G(\hat{F}^{\text{T}}_{l+1}),\ G(F^{\text{T}}_{l+1})\right), & l \in l_{\text{inter}}, \\ \text{MSE}\left(G(\hat{F}^{\text{T}}_{l}),\ G(F^{\text{T}}_{l})\right), & l = l_{\text{final}}. \end{cases} \tag{11}$$

where $G(F)$ denotes the Gram matrix of F. This preserves the directional structure passed to the next block for intermediate layers, and preserves the final-layer structure at l_{final}. Consequently, the total information term is $\mathcal{L}_{\text{info}} = \sum_{l\in\mathcal{D}} \mathcal{L}_{\text{info},l}$.

Training Objective. We jointly optimize the student parameters $\boldsymbol{\theta}_{\text{S}}$, projection parameters $\boldsymbol{\theta}_P = \{P_l\}_{l\in\mathcal{D}}$, and SiNGER adapter parameters $\boldsymbol{\theta}_\phi = \{\phi_{\text{down},l}, \phi_{\text{up},l}\}_{l\in\mathcal{D}}$ with a single weighted sum loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{KD}} \ + \ \lambda_{\text{outlier}}\, \mathcal{L}_{\text{outlier}} \ + \ \lambda_{\text{info}}\, \mathcal{L}_{\text{info}}, \tag{12}$$

where λ_{outlier} and λ_{info} balance artifact suppression and information alignment. This objective encourages effective transfer while controlling high-norm artifacts in teacher features.

4 Experiments and Analysis

4.1 Details

Downstream Tasks. To evaluate SiNGER-distilled ViT as a VFM, we adopt the student network to a diverse set of downstream tasks. Specifically, we consider six representative benchmarks: ImageNet-1K validation set for large-scale classification (Deng et al., 2009), ADE-20K for semantic segmentation (Zhou et al., 2019), NYUd-v2 for depth estimation (Ignatov et al., 2024), iNaturalist-2019 for long-tail classification (Van Horn et al., 2018), ImageNet-R and ImageNet-v2 for domain shift robustness (Hendrycks et al., 2021; Recht et al., 2019), and four fine-grained classification datasets (Maji et al., 2013; Parkhi et al., 2012; Bossard et al., 2014; Nilsback & Zisserman, 2006).

Distillation Setup. We assess SiNGER under the canonical ViT (Dosovitskiy et al., 2021), which lies as the core form of ViTs and the variants. For scale, we use a large ViT as the teacher (290.2M parameters, 24 layers) and a tiny ViT as the student (5.5M parameters, 12 layers). Student layers are aligned with every second teacher layer. The official implementation of FitNet and ViTKD employ task-specific loss objectives, such as cross-entropy minimization for classification. In contrast, we target VFM distillation and therefore exclude task-specific losses, distilling the last hidden layer's representation.

Rationale. We aim to probe pre-training agnostic mechanisms of artifact formation and suppression. To this end, we conduct the full ablation suite on canonical ViTs, whose transparent design and widely adopted training recipe allow tighter control, clearer causal attribution, and more reproducible analysis.

4.2 Multi-Task Evaluation

Table 1 summarizes multi-task linear evaluation results across ten benchmarks. The teacher (Large) achieves strong performance, while the Tiny baseline shows significant degradation, particularly on dense prediction tasks; ADE-20K and NYUd-v2. FitNet improves over the Tiny baseline by transferring intermediate features, but still inherits artifacts from the teacher, limiting overall gains. ViTKD performs poorly across all tasks, as its random masking strategy often collapses feature representations and prevents effective learning. Discussion on ViTKD is detailed in Appendix B.

By contrast, SiNGER demonstrates consistent improvements over FitNet and ViTKD on most benchmarks. On IN-val, ADE-20K, NYUd-v2, DS, and FG, SiNGER yields large gains, approaching teacher performance despite the smaller capacity. The only exception is iNat2019, where performance slightly drops, which we attribute to the long-tail nature of the dataset, as Zhang et al. (2023) pointed out. However, SiNGER still outperformed FitNet. We report analysis on iNat2019 in Appendix H. Overall, these results confirm that suppressing artifacts during distillation produces student models that are both more accurate and more generalizable across diverse tasks.

Model	Distillation	IN-val top-1 (↑)	ADE-20K mIoU (↑)	NYUd-v2 RMSE (↓)	iNat2019 top-1 (↑)	DS top-1 (↑)	FG top-1 (↑)
Large	-	79.58	26.57	0.9157	71.42	54.20	82.27
Tiny	-	58.03	14.20	1.1807	43.95	28.75	62.02
Large → Tiny	FitNet	62.43	18.73	1.0093	40.02	32.32	62.48
	ViTKD	5.07	11.92	1.1903	23.69	2.08	33.52
	SiNGER	70.59 8.16	21.76 3.03	0.9406 0.0687	41.11 2.84	38.87 6.55	64.61 2.13

Table 1: Multi-task linear evaluation results. ImageNet-1K validation (IN-val) for large-scale classification, ADE-20K for semantic segmentation, NYUd-v2 for monocular depth estimation, iNaturalist2019 (iNat2019) for long-tail learning, ImageNet-R and ImageNet-v2 for domain shift (DS), and four fine-grained classification (FG) benchmarks: FGVC-Aircraft, Oxford-IIIT Pet, Food-101, and Flowers-102 were tested. Blue represents improvement and red indicates degradation.

4.3 Representation Quality

We assess the quality and interpretability of distilled representations by comparing the feature maps and their Gram matrices. Figure 5 depicts the Gram matrices of the feature maps. Quantitatively, SiNGER's Gram matrix is the most similar one to the teacher's Gram matrix. Gram Distance (GD), defined as ℓ_2 distance between the Gram matrices, confirms this trend Figure 2. This shows that when artifacts are distilled, it disrupts the transfer of patch-wise relation, resulting in degraded student representation. Centered Kernel Analysis (CKA, Kornblith et al. (2019)) measures the linear correlation between two feature maps. FitNet and ViTKD achieve higher similarity by following the teacher too closely, but this reflects replication of artifacts rather than useful knowledge transfer. By contrast, SiNGER learns structurally consistent yet information-preserved representations, balancing similarity with the teacher.

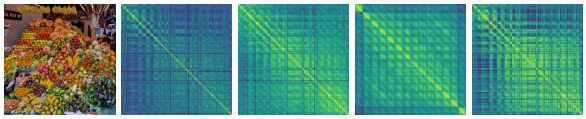


Figure 5: Gram matrices of the patches. The input, the teacher, and distilled features; FitNet, ViTKD, and SiNGER in order.

Layer	FitNet	ViTKD	SiNGER
GD	0.237	0.520	**0.130**
CKA	0.732	0.745	**0.660**

Table 2: The teacher-student representation's similarity in terms of ℓ_2 distance between the Gram Distance (GD) and CKA.

4.4 Adapter Operation

We empirically analyze how the optimized adapter operates on ImageNet-1K. To probe the coupling with the next layer, we evaluate at an intermediate layer $l = 17$.

Patch-Norm Distribution Between F_{l+1} and $\hat{F}_{l+1}$. We visualize the distribution of patch l_2 norms for F_{l+1} and $\hat{F}_{l+1}$ with side-by-side box plots (Figure 6). The teacher produces high-norm artifacts that are distinctly gathered as a group. We observed that SiNGER effectively draws such artifacts into the normal-patch range while preserving informative features. This results in stabilized gradient flow through the normal patches.

Cosine Similarity Between F_{l+1} and $\hat{F}_{l+1}$. To assess information preservation, we compute patch-wise cosine similarities for both F_l vs. $\hat{F}_l$ and F_{l+1} vs. $\hat{F}_{l+1}$, aggregating per image across the dataset (see Figure 3). The $17, 18$-th layers yield cosine similarity of 0.9566 and 0.9731 with negligible variance, respectively, which is clearly considered similar.

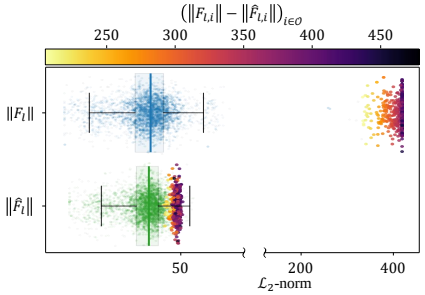


Figure 6: Patch-norm distributions of F_{l+1} and $\hat{F}_{l+1}$. Artifacts are scale-colored separately with `inferno`.

Pair	μ	σ	median
F_{17} vs. $\hat{F}_{17}$	0.9566	0.0038	0.9569
F_{18} vs. $\hat{F}_{18}$	0.9731	0.0021	0.9732

Table 3: Patch-wise cosine similarity (per-image mean over patches) on ImageNet-1K at l=17 and l+1=18.

4.5 Ablation Study

We report four ablations, focusing on initialization, losses, hyperparameters, and distillation layers.

Initialization Method. We validate whether nullspace initialization truly guides the adapter to induce perturbations along the nullspace during optimization by comparing nullspace-biased (SiNGER) and random initializations.

Let the next block at layer $l{+}1$ be linearized to $\tilde{W}_{l+1} \in \mathbb{R}^{D\times D}$. From $\tilde{W}_{l+1}$, define two rank-r bases: the principal basis $P_{l+1} \in \mathbb{R}^{D\times r}$ (largest singular directions) and the null basis $N_{l+1} \in \mathbb{R}^{D\times r}$ (smallest singular directions). We quantify alignment with the normalized Frobenius norm

$$E_{\text{prob}}(\phi) \;=\; \frac{\|\phi\, P_{l+1}\|_f}{\|\phi\|_f}, \qquad E_{\text{safe}}(\phi) \;=\; \frac{\|\phi\, N_{l+1}\|_f}{\|\phi\|_f}, \quad \forall \phi \in \big\{\phi_{\text{up},l},\ \phi_{\text{down},l}^{\top}\big\}. \tag{13}$$

A larger Frobenius norm indicates stronger alignment of the trained adapter matrix($\phi_{\text{up},l}, \phi_{\text{down},l}$) with the corresponding subspace; in our design, the primary goal is to increase E_{safe} (alignment to N_{l+1}).

In Table 4a, initialization markedly increases alignment to N_{l+1}: E_{safe} reaches $0.83/0.76$ for $\phi_{\text{up},l}$ at $l{=}17/23$, and $0.55/0.58$ for ${\phi_{\text{down},l}}^{\top}$. Both are under 0.27 for random initialization. This provides strong evidence that the initialization *guides optimization into the null space*, yielding substantially higher E_{safe} across layers and for both $\phi_{\text{up},l}$ and $\phi_{\text{down},l}^{\top}$, which indicates successful guidance toward the null space directions. Meanwhile, E_{prob} remains lower or comparable under SiNGER, but our objective is not to minimize E_{prob} per se; rather, to ensure that the learned parameters predominantly occupy N_{l+1} so as to suppress high-norm amplification while preserving useful directions.

Layer	Matrix	Init	$E_{\text{prob}}\downarrow$	$E_{\text{safe}}\uparrow$
17	$\phi_{\text{up},l}$	Random	0.2565	0.2532
17	$\phi_{\text{up},l}$	SiNGER	**0.1479**	**0.8337**
23	$\phi_{\text{up},l}$	Random	0.2537	0.2541
23	$\phi_{\text{up},l}$	SiNGER	**0.1833**	**0.7589**
17	${\phi_{\text{down},l}}^{\top}$	Random	0.3100	0.2494
17	${\phi_{\text{down},l}}^{\top}$	SiNGER	**0.2847**	**0.5485**
23	${\phi_{\text{down},l}}^{\top}$	Random	0.3025	0.2641
23	${\phi_{\text{down},l}}^{\top}$	SiNGER	**0.2746**	**0.5774**

(a) Initialization methods.

Pair	$\mathcal{L}_{\text{outlier}}$	$\mathcal{L}_{\text{info}}$	mean $\pm$ std $\downarrow$	median
$F^T \leftrightarrow \hat{F}^T$	✓		14.22 ± 1.45	14.28
$F^T \leftrightarrow \hat{F}^T$	✓	✓	$\mathbf{7.25 \pm 0.84}$	7.19
$F^T \leftrightarrow F^S$	✓		72.36 ± 7.61	71.85
$F^T \leftrightarrow F^S$	✓	✓	$\mathbf{41.71 \pm 7.01}$	40.89

(b) Information preservation term.

Table 4: Ablation studies on the initialization method and the information preservation loss.

Loss Term. We ablate the loss design to verify the role of information preservation. Our full objective uses both outlier suppression $\mathcal{L}_{\text{outlier}}$ and information preservation $\mathcal{L}_{\text{info}}$, whereas the ablated variant uses $\mathcal{L}_{\text{outlier}}$ only. (Using $\mathcal{L}_{\text{info}}$ alone admits the trivial solution $\|\Delta\|{=}0$ and yields no updates.)

To assess preservation of teacher information, we measure the Gram distance between F^T and $\hat{F}^T$. Additionally, to evaluate the final effect on distillation, we measure how well the student features F^S preserve teacher relations by comparing F^S against F^T. Distances are computed per image and summarized over ImageNet-1K. For a feature map F, let $G(X) = FF^{\top}$ and define

$$D_G(F_i, F_j) \;=\; \big\|\, G(F_i) - G(F_j)\,\big\|_f.$$

In Table 4b, lower D_G indicates better preservation of pairwise feature relations. Compared to $\mathcal{L}_{\text{outlier}}$ alone, adding $\mathcal{L}_{\text{info}}$ nearly halves the D_G distance ($14.22 \rightarrow 7.25$) and substantially improves teacher–student alignment ($72.36 \rightarrow 41.71$). Thus, the information preservation term prevents degenerate updates and maintains the relational geometry that is crucial for effective transfer.

Hyperparameter Sensitivity. Two hyperparameters are required in SiNGER: α and r. α determines the strictness of artifact filtering by setting the percentile threshold based on the Gaussian-like distribution of the patch norms $\|F^{\mathrm{T}}_{l,i}\|$ across i. r controls the capacity of the perturbation ΔF^{T} applied to F^{T}. A larger r allows the adapter to explore a wider subspace, but may distort the semantic structure of the features. Conversely, if r is too small, the limited degrees of freedom restrict the adapter from effectively suppressing artifacts, while also risking the loss of informative components.

r	RMSE (↓)	$\delta_{1.25}$ (↑)
8	1.4545	33.97
16	**1.4395**	**33.79**
32	1.4907	33.07
64	1.6485	28.91

(a) Rank sweep with $\alpha = 0.95$.

α	RMSE (↓)	$\delta_{1.25}$ (↑)
0.90	1.5989	29.78
0.95	**1.4395**	**33.79**
0.97	1.4748	33.66
0.99	1.5321	30.80

(b) Quantile threshold sweep with $r = 16$.

Table 5: Rank and quantile threshold sweeps on NYUd-v2. We conduct a grid search over candidate values and select the configuration that yields the best performance.

Table 5 reports the sensitivity of α and r on NYUd-v2. We observe that performance degrades when r is too small or too large, confirming the need for balanced capacity. Similarly, extreme values of α either under-filter artifacts or discard informative signals. The observed trends match our theoretical intuition: performance improves when artifact suppression and information preservation are balanced, but deteriorates when either dominates. At the same time, the results show robustness—performance does not collapse outside the optimal point, indicating stability of the framework. Finally, the chosen hyperparameters ($r = 16$ and $\alpha = 0.95$) generalize well across other tasks and datasets, and we adopt them as the default configuration.

Distillation Layers. Selecting is critical because we aim to distill artifact-prone features. To ensure gradients traverse the entire backbone, we always distill the last layer ($l = 23$ in ViT-L). Beyond this, we select an additional intermediate layer by inspecting teacher feature trends (see Appendix F). Since our method is an artifacts-aware approach, we first pinpointed the location where artifacts occur. For ViT-L, we observed that artifacts appear after $l = 11$. We additionally select the intermediate layer at $l = 17$. Across three variants, the $l = 17, 23$ configuration performs best, as shown in Table 6.

Layers			NYUd-v2
11	17	23	RMSE (↓)
✓		✓	0.9554
	✓	✓	**0.9406**
✓	✓	✓	0.9624

Table 6: Distillation layer selection.

5 Conclusion

In this work, we investigated the challenge of artifact transfer in knowledge distillation for ViTs. We showed that high-norm artifacts in teacher representations degrade interpretability and are naïvely inherited by student models, limiting the effectiveness and benefits from scaling of conventional distillation approaches. To address this issue, we proposed a distillation framework, namely SiNGER, and a nullspace-guided adapter that introduces minimal perturbations to suppress artifacts while preserving informative representations. Our framework demonstrated consistent improvements over existing methods across a diverse set of downstream tasks, yielding both higher accuracy and more interpretable features. We believe this perspective opens new directions for artifact-robust distillation and provides insights into the broader problem of transferring knowledge from over-parameterized models.

Limitations and Future Work. Nevertheless, our method has limitations. It suppresses artifacts rather than fully eliminating their sources. Since the goal is to retain as much teacher information as possible, the root causes of representation degradation remain. As a result, students distilled with our approach cannot achieve the same level of clean representations as models explicitly trained to avoid artifacts, such as SINDER (Wang et al., 2025) and DINOv3 (Siméoni et al., 2025). Future work will extend our approach to a wider range of foundation models and multi-modal settings, exploring whether nullspace-guided perturbations can serve as a general mechanism for reliable model compression and adaptation.

REFERENCES

Nima Aghli and Eraldo Ribeiro. Combining weight pruning and knowledge distillation for cnn compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3191–3198, June 2021.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2dnO3LLiJ1.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL http://arxiv.org/abs/1503.02531.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Xinlei Huang, Jialiang Tang, Xubin Zheng, Jinjia Zhou, Wenxin Yu, and Ning Jiang. Learn from balance: Rectifying knowledge transfer for long-tailed scenarios. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Dmitry Ignatov, Andrey Ignatov, and Radu Timofte. Virtually enriched nyu depth v2 dataset for monocular depth estimation: Do we need artificial augmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6177–6186, 2024.

Saqib Javed, Hieu Le, and Mathieu Salzmann. Qt-dog: Quantization-aware training for domain generalization, 2024. URL https://arxiv.org/abs/2410.06020.

Jiacheng Jiang, Yuan Meng, Chen Tang, Han Yu, Qun Li, Zhi Wang, and Wenwu Zhu. Gaqat: gradient-adaptive quantization-aware training for domain generalization, 2024. URL https://arxiv.org/abs/2412.05551.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model, 2024. URL https://arxiv.org/abs/2406.09246.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28092–28103. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/ec8956637a99787bd197eacd77acce5e-Paper.pdf.

Zhiying Lu, Hongtao Xie, Chuanbin Liu, and Yongdong Zhang. Bridging the gap between vision transformers and convolutional neural networks on small datasets. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

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

Maria-Elena Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 1447–1454, 2006.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12490–12500, June 2024.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

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 Bulent Sariyildiz, Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Unic: Universal classification models via multi-teacher distillation. In *European Conference on Computer Vision (ECCV)*, 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.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15731–15740, June 2024.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 516–533, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20053-3.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Haoqi Wang, Tong Zhang, and Mathieu Salzmann. Sinder: Repairing the singular defects of dinov2. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 20–35, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72667-5.

Huanrui Yang, Hongxu Yin, Maying Shen, Pavlo Molchanov, Hai Li, and Jan Kautz. Global vision transformer pruning with hessian-aware saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18547–18557, June 2023.

Zhendong Yang, Zhe Li, Ailing Zeng, Zexian Li, Chun Yuan, and Yu Li. Vitkd: Feature-based knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1379–1388, June 2024.

Tao Yu, Xu Zhao, Yongqi An, Ming Tang, and Jinqiao Wang. Knowledge distillation dealing with sample-wise long-tail problem. In *Proceedings of the Asian Conference on Computer Vision*, pp. 2354–2370, 2024.

Shaoyu Zhang, Chen Chen, Xiyuan Hu, and Silong Peng. Balanced knowledge distillation for long-tailed learning. *Neurocomputing*, 527:36–46, 2023.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

APPENDIX

SINGER: A CLEARER VOICE DISTILLS VISION TRANSFORMERS FURTHER

A TRICKS FOR CALCULATING NULLSPACE

A.1 WEIGHTS LINEARIZATION

We generally compute a null space via the Singular Value Decomposition (SVD) of a linear operator $M \in \mathbb{R}^{d\times d}$, obtaining a left-nullspace $\mathcal{N} = \mathrm{Null}(M^\top)$. This procedure presumes that the target M is linear. However, a transformer block W_l at layer l is inherently non-linear due to attention, activation, and residual pathways, so the SVD-based nullspace of the block W_l is not directly defined.

Let $x \in \mathbb{R}^{1\times d}$ denote the row-vector feature. A standard Pre-LN transformer block at layer l can be written as

$$y_l = x_l \; + \; \mathrm{MHA}\big(\mathrm{LN}(x_l)\big),$$
$$x_{l+1} = y_l \; + \; \mathrm{FFN}\big(\mathrm{LN}(y_l)\big),$$

where both $\mathrm{MHA}(\cdot)$ and $\mathrm{FFN}(\cdot)$ include non-linear operations (softmax attention, elementwise activations) and the residual additions further couple the sub-layers. Consequently, there is no single linear matrix M that exactly represents W_l for SVD, motivating a linearization that we introduce next.

To compute a nullspace for a non-linear block W_l, we first replace it with a linear surrogate $\tilde{W}_l$. Our key design choice is to linearize only the FFN sub-layer, motivated by an empirical study showing that the FFN induces larger relative feature changes than self-attention (SA).

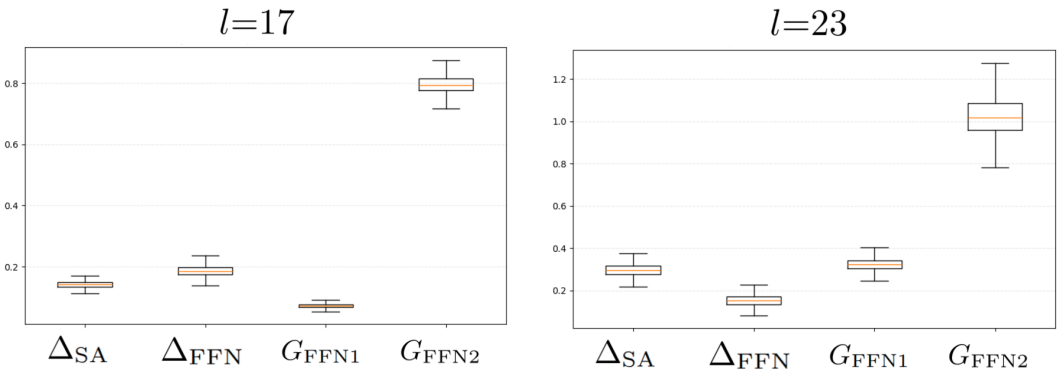
We measure sub-layer-wise changes on a ViT teacher by sampling $N{=}5000$ random ImageNet training images (uniform over class folders), resizing to 224×224, normalizing, and running a forward pass to obtain per-layer tokens. For each layer l and each block, we then reapply the block with instrumented intermediates:

$$x_{\mathrm{in}} \in \mathbb{R}^{B\times(1+P)\times d},$$
$$x_{\mathrm{SA}} = x_{\mathrm{in}} + \mathrm{MHA}\big(\mathrm{LN}(x_{\mathrm{in}})\big),$$
$$h_1 = \mathrm{LN}(x_{\mathrm{SA}}),$$
$$z_1 = h_1 W_1 + b_1,$$
$$a_1 = \mathrm{GELU}(z_1),$$
$$z_2 = a_1 W_2 + b_2,$$
$$x_{\mathrm{out}} = x_{\mathrm{SA}} + z_2,$$

where W_1, W_2 are the FFN weights (expand-then-project), and we ignore the stochastic drop-path in reporting expectations. We exclude the [CLS] token and compute patch-wise ℓ_2-norms.

For each image and layer, we aggregate over patches using the mean and record four quantities:

$$\Delta_{\mathrm{SA}} := \mathrm{mean}_{\text{patch}} \frac{\|x_{\mathrm{SA}} - x_{\mathrm{in}}\|_2}{\|x_{\mathrm{in}}\|_2} \qquad \Delta_{\mathrm{FFN}} := \mathrm{mean}_{\text{patch}} \frac{\|x_{\mathrm{out}} - x_{\mathrm{SA}}\|_2}{\|x_{\mathrm{SA}}\|_2}$$
$$G_{\mathrm{FFN1}} := \mathrm{mean}_{\text{patch}} \frac{\|a_1\|_2}{\|h_1\|_2} \qquad G_{\mathrm{FFN2}} := \mathrm{mean}_{\text{patch}} \frac{\|z_2\|_2}{\|a_1\|_2}$$



Metric	mean	median	p95
G_{FFN1}	0.0719	0.0714	0.0844
G_{FFN2}	**0.7988**	**0.7945**	**0.8571**
Δ_{FFN}	0.1871	0.1852	0.2210
Δ_{SA}	0.1417	0.1415	0.1611

Metric	mean	median	p95
G_{FFN1}	0.3251	0.3231	0.3773
G_{FFN2}	**1.0226**	**1.0194**	**1.1754**
Δ_{FFN}	0.1543	0.1537	0.2001
Δ_{SA}	**0.2960**	**0.2965**	**0.3468**

Figure 7: Sub-layer change analysis at two depths. **Top:** box-plots of relative changes/gains in each layer. **Bottom:** summary statistics in each layer.

From Figure 7, the dominant amplification occurs in the second FFN stage (G_{FFN2} is largest at both depths), and the net FFN residual Δ_{FFN} is comparable to or larger than the SA residual depending on the layer. This indicates that the principal source of norm inflation lies within the FFN pathway, especially its projection stage.

Guided by this analysis, we exclude the non-linear SA pathway when constructing a linear operator for SVD and focus on the FFN inside W_l. Since the non-linearity enters the FFN only via the GELU between two linear maps, removing GELU (and biases) yields a linear surrogate:

$$\mathrm{FFN}(h) \;\approx\; h\,W_{FFN1}W_{FFN2} \;=\; h\,\tilde{W} \tag{14}$$

with row-vector features and right multiplication (the column-vector convention uses $\tilde{W}^\top = W_{FFN2}^\top W_{FFN1}^\top$). We refer to $\tilde{W}_l$ as the linearized weights of block l.

A.2 Nullspace of linearized weights

Now, we can compute the SVD of the linearized FFN matrix $\tilde{W}_l \in \mathbb{R}^{d\times d}$: We compute its SVD

$$\tilde{W}_l \;=\; U_l\,\Sigma_l\,V_l^\top, \qquad \Sigma_l = \mathrm{diag}(\sigma_1 \geq \cdots \geq \sigma_d \geq 0). \tag{15}$$

A left-nullspace basis of dimension r is obtained by selecting the r left singular vectors associated with the r smallest singular values:

$$N_l \;\in\; \mathbb{R}^{d\times r}, \qquad N_l^\top N_l = I_r, \qquad \mathrm{cols}(N_l) \;=\; U_l^{(:,\,d-r+1:d)}. \tag{16}$$

For any row vector $v^\top \in \mathrm{span}(N_l)$ we then have the approximate-null condition

$$v^\top \tilde{W}_l \;=\; v^\top U_l \Sigma_l V_l^\top \;\approx\; 0, \tag{17}$$

since the selected modes correspond to the smallest singular values.

A common concern is that $\tilde{W}_l$ could be numerically full rank, making the exact nullspace trivial. We therefore quantify a practical ε-nullspace using two diagnostics defined below, and visualize them in Figure 8.

Let the singular values of $\tilde{W}_l \in \mathbb{R}^{d\times d}$ be $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0$. Define the cumulative energy

$$E(k) \;:=\; \frac{\sum_{i=1}^{k}\sigma_i^2}{\sum_{i=1}^{d}\sigma_i^2} \;\in [0,1],$$

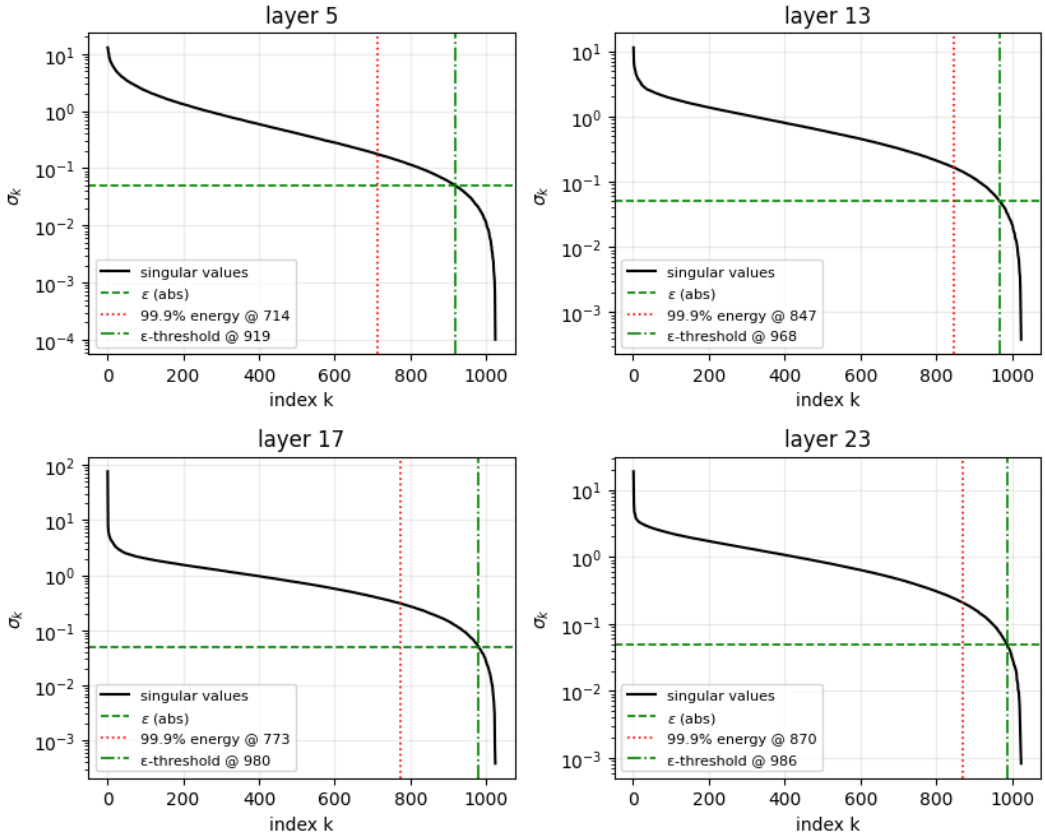


Figure 8: Singular-value spectra of $\tilde{W}_l$ for representative layers ($l = 5, 13, 17, 23$) with a green horizontal line at the absolute threshold $\varepsilon = 0.05$ and red vertical lines marking k_{energy} (99.9% cumulative energy) and k_ε (first index with $\sigma_k \le \varepsilon$).

and for a target level $\rho \in (0, 1)$

$$k_{\text{energy}}(\rho) \;:=\; \min\{\, k \in \{1, \dots, d\} \;:\; E(k) \;\ge\; \rho \,\}. \tag{18}$$

For an absolute tolerance $\varepsilon > 0$, define the first crossing index

$$k_\varepsilon \;:=\; \min\{\, k \in \{1, \dots, d\} \;:\; \sigma_k \le \varepsilon \,\}, \qquad r_\varepsilon \;:=\; d - k_\varepsilon + 1, \tag{19}$$

so that the ε-tail has dimension r_ε and is spanned by the last r_ε left singular vectors.

As summarized by Figure 8, we consider a ViT-L teacher with d=1024 and focus on a intermediate block (l=17). At this layer, the cumulative-energy index is $k_{\text{energy}}(0.999)$=773, so the low-energy tail has size $d - k_{\text{energy}}$=251. With an absolute tolerance ε=0.05, the first-crossing index is k_ε=980, yielding an ε-tail of dimension $r_\varepsilon = d - k_\varepsilon + 1$=45. Consequently, the tail span forms a high-quality *approximate nullspace*: for any $v^\top$ in this subspace,

$$\|v^\top \tilde{W}_l\|_2 \;\le\; \varepsilon \, \|v\|_2,$$

which justifies nullspace-guided updates that suppress outlier energy while preserving informative structure.

B Failure of ViTKD

In this section, we discuss the failure of ViTKD (Yang et al., 2024) in learning teacher representation. The core strategy behind ViTKD is masking and generation. Different from SiNGER, ViTKD does not adaptively detect artifacts and randomly discards patches regardless of their semantic validity,

and generates through convolution, utilizing learnable generative tokens. This ensures students learn artifact-free representation. However, it also makes the whole representation blurry, which is one of the expected trivial solutions to minimize the mean squared error of the generated features, resulting in a significantly degraded representation.

As depicted in Figure 2, ViTKD successfully mimics the teacher's representation in terms of cosine similarity, but fails in building the informatively structured feature map. This structural degradation makes the representation blurry, resulting in poor downstream task adaptation.

C Visualizing Distillation Outputs on a Single Image

We visualize a single sample at token resolution 14×14 (patches only; [CLS] excluded). For the teacher(ViT-large), we show odd-numbered blocks $l \in \{1, 3, \ldots, 23\}$. For the student(ViT-Tiny), we show layers $i \in \{0, \ldots, 11\}$ aligned with the teacher columns. In this sample, the SiNGER adapter is applied only at $l \in \{17, 23\}$.

As shown in Figure 9, the first row F_l^{T} exhibits artifacts: high-norm becomes more pronounced at deeper blocks. After applying SiNGER adapter, the second row $\hat{F}_l^{\mathrm{T}}$ attenuates these artifacts while preserving the informational structure, producing a more transfer-friendly teacher target. The third row visualizes the residual $\Delta F_l^{\mathrm{T}} = \hat{F}_l^{\mathrm{T}} - F_l^{\mathrm{T}}$, confirming that SiNGER removes a small set of outlier's magnitudes. Finally, the fourth row F_l^{S} aligns more closely with $\hat{F}_l^{\mathrm{T}}$ than with F_l^{T}, indicating that the student learns the SiNGER-refined, structure-preserving representation rather than the original outlier-dominated one.

D Visualization of outlier Suppression

This appendix illustrates, on a single sample, how outlier suppression operates numerically. As shown in Figure 10, in the last layer, the patchwise norm map $\|F_l^{\mathrm{T}}\|$ contains an outlier patch with a maximum norm of 638. At the same spatial location, the suppressed map $\|\hat{F}_l^{\mathrm{T}}\|$ drops to 54.1. Finally, the distribution of $\|\hat{F}_l^{\mathrm{T}}\|$ appears much more uniform across patches, indicating that extremely high-norm outliers have been attenuated while the overall scale has been regularized.

E Visualization of Information Preservation

We confirm whether the information is actually preserved right after the $l+1$-th layer. The cross-similarity map between F_{l+1}^{T} and $\hat{F}_{l+1}^{\mathrm{T}}$ are visualized in Figure 11. For each 1×2 cells, the left one shows $F_{l+1}^{\mathrm{T}} \to \hat{F}_{l+1}^{\mathrm{T}}$ similarity map, and the other one shows the contrary. As shown, in both directions, the similarity maps are almost identical This implies the information is actually preserved even after passing the next layer, which is the source of nullspace used for initializing the proposed adapter.

F Student Visualization

We evaluate the quality of our SiNGER-distilled feature by visualizing the similarity map (Figure 12). For each 2×2 cell, the top row is the student, the bottom row is the teacher. The left column is the norm map of the feature map, and the right column is the similarity map to the '$\times$' marked patch. The norms of the teacher's feature maps are artifact-prone, but still produce the similarity map, which semantically makes sense. The student produces artifact-suppressed feature maps while maintaining the semantic relation among the patches. This emphasizes that SiNGER effectively optimizes two objective functions, distills high-quality feature representation.

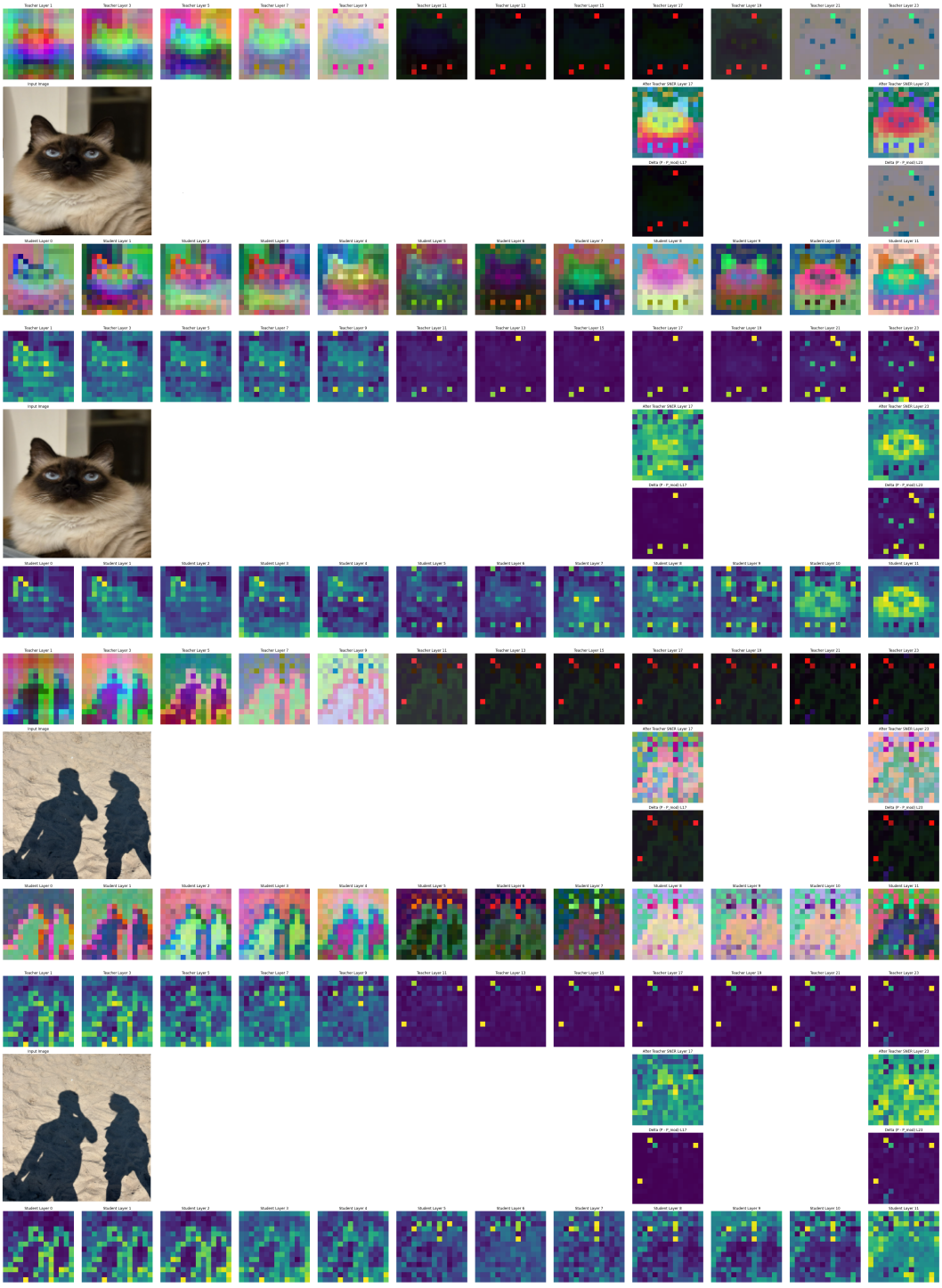


Figure 9: Distillation visualization across stages using complementary views. Each panel renders patch features either as a directional view (PCA with 3 components) or as a magnitude view (patch-wise ℓ_2-norm). Within each panel, rows depict (top to bottom) $F_l^{\mathrm{T}}, \hat{F}_l^{\mathrm{T}}, \Delta F_l^{\mathrm{T}}, F_l^{\mathrm{S}}$.

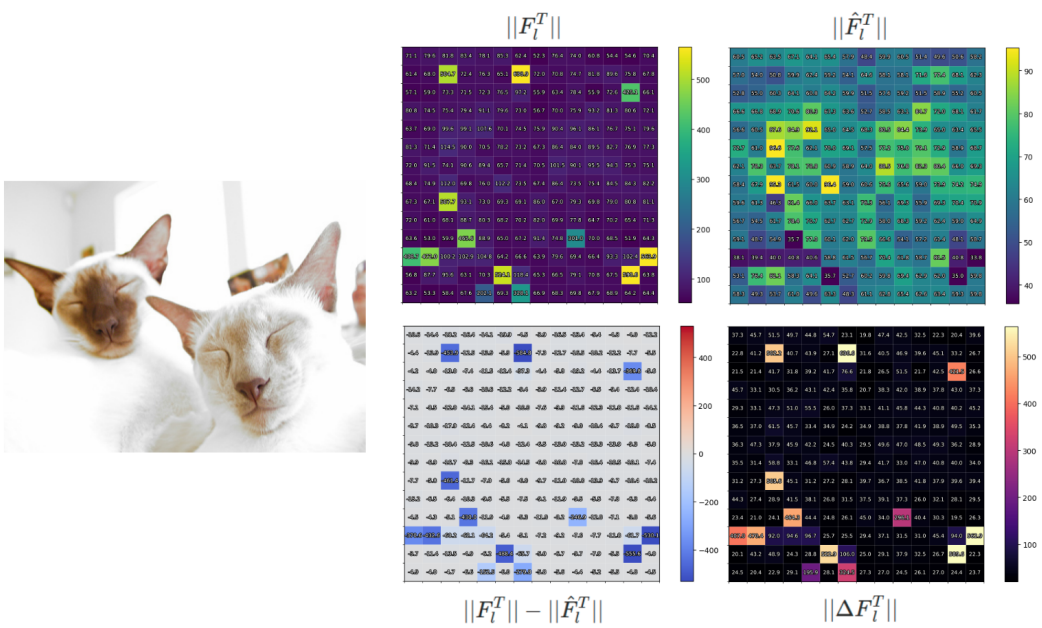


Figure 10: Left: input image. Right: patchwise visualizations. **Row 1:** $\|F_l^{\mathrm{T}}\|$ and $\|\hat{F}_l^{\mathrm{T}}\|$. **Row 2:** $\|F_l^{\mathrm{T}}\|$, $\hat{F}_l^{\mathrm{T}}$ (signed map), and $\|\Delta F_l^{\mathrm{T}}\|$ with $\Delta F_l^{\mathrm{T}} = F_l^{\mathrm{T}} - \hat{F}_l^{\mathrm{T}}$.

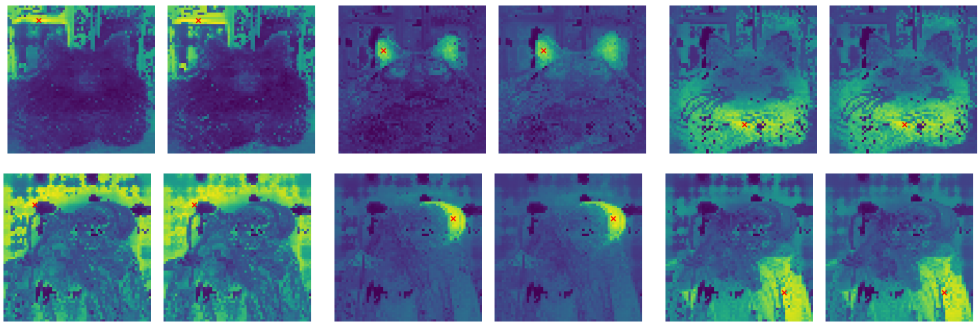


Figure 11: The cross-similarity map btw F_{l+1}^{T} and $\hat{F}_{l+1}^{\mathrm{T}}$ to '×' mark is visualized.

G Correlation of Outlier with Principal and Null Bases

We conduct this experiment to validate the core design choice behind our method: we perturb features along a nullspace to preserve information, but such a perturbation is only meaningful if outliers do not primarily reside in the nullspace. Otherwise, nullspace-directed updates would fail to suppress outliers. To test this, we build on Appendix A.1 and Appendix A.2: from the linearized FFN matrix $\tilde{W}_l$, we take the r left singular vectors with the largest singular values as the **principal basis**, and the r with the smallest singular values as the **null basis**.

Fix a layer l and an image, and let $X \in \mathbb{R}^{(1+P)\times d}$ be the teacher tokens (CLS excluded below). Compute

$$\tilde{W}_l = U_l\,\Sigma_l\,V_l^\top, \quad U_l = [u_1, \ldots, u_d],\ \Sigma_l = \mathrm{diag}(\sigma_1 \geq \cdots \geq \sigma_d \geq 0).$$

Define the two r-dimensional bases

$$U_{\text{prin}} = [u_1, \ldots, u_r], \qquad U_{\text{null}} = [u_{d-r+1}, \ldots, u_d].$$

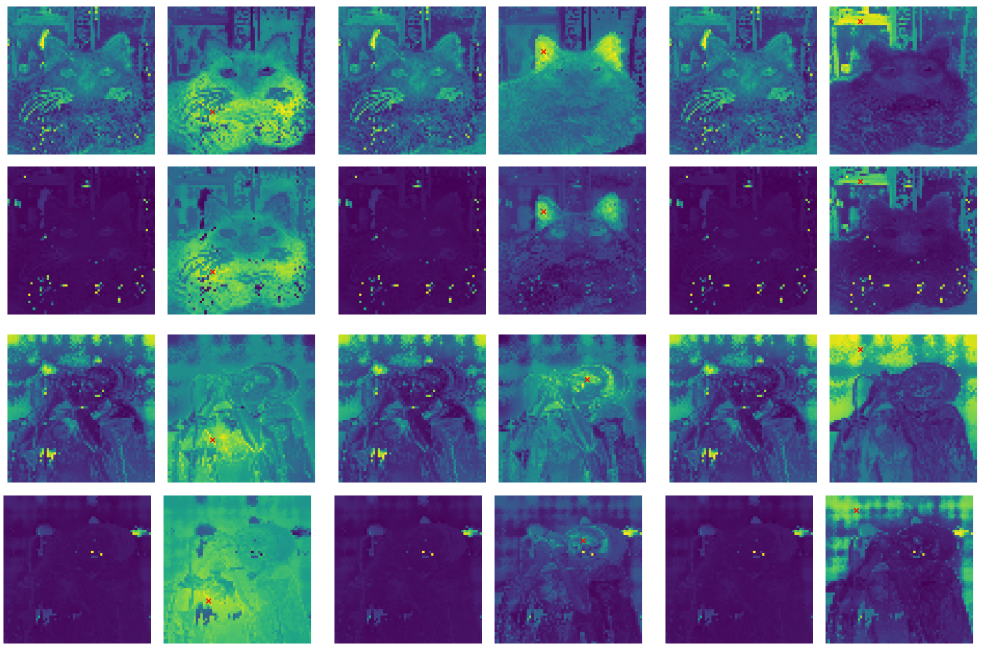


Figure 12: The similarity map to '×' mark is visualized.

Let $x_p \in \mathbb{R}^{1\times d}$ be the p-th patch feature with norm $n_p = \|x_p\|_2$. For a chosen r-dimensional orthonormal basis $U = [u_{i_1}, \ldots, u_{i_r}] \in \mathbb{R}^{d\times r}$ (e.g., columns selected from the left singular vectors of $\tilde{W}_l$), define the normalized subspace energy.

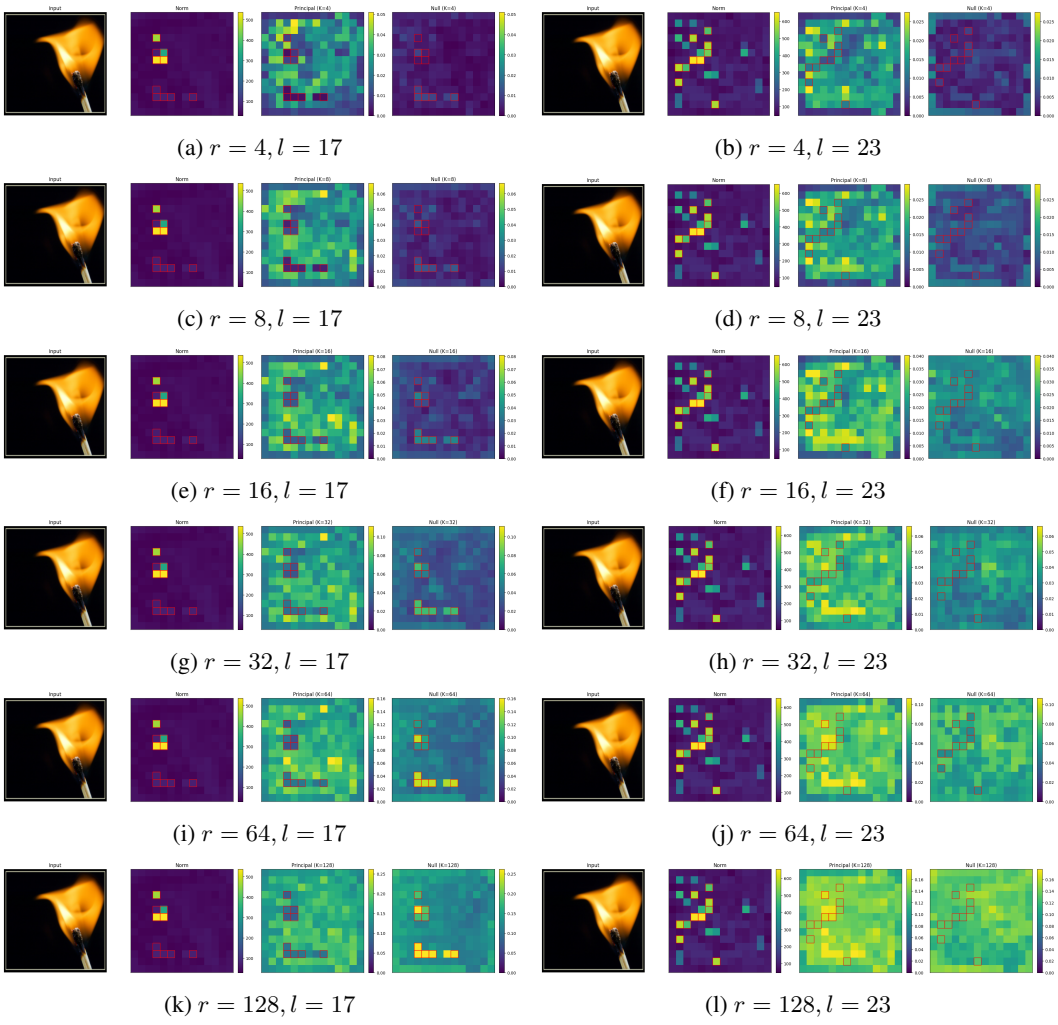
$$E_U(p) = \frac{\sum_{k=1}^{r} \langle x_p, u_{i_k} \rangle^2}{\|x_p\|_2^2 + \varepsilon}$$

In words, $E_U(p)$ is the fraction of the patch's total energy captured by the subspace U—i.e., a norm-invariant measure of how strongly x_p aligns with U. In our analysis, we instantiate U by either of the above bases, i.e.,

$$U \in \{ U_{\text{prin}},\ U_{\text{null}} \}.$$

In Figure 13, across layers we observe distinct behaviors as the rank r increases. At the intermediate layer (l=17), outlier patches (high $\|F_l^{\mathrm{T}}\|_2$) exhibit growing values in $E_{U_{\text{null}}}$ as r increases, while $E_{U_{\text{prin}}}$ over those same patches does not grow accordingly. Conversely, at the last layer (l=23), outlier patches show increasing $E_{U_{\text{prin}}}$ with r, whereas $E_{U_{\text{null}}}$ over outliers does not increase in the same manner. We quantify these patterns in Figure 13 (m),(n). For l=17, the patch norm correlates positively with the null subspace energy ($\text{corr}(\|F_l^{\mathrm{T}}\|_2, E_{U_{\text{null}}}) > 0$) and negatively with the principal subspace energy ($\text{corr}(\|F_l^{\mathrm{T}}\|_2, E_{U_{\text{prin}}}) < 0$). In contrast, for l=23 the signs flip: $\text{corr}(\|F_l^{\mathrm{T}}\|_2, E_{U_{\text{null}}}) < 0$ and $\text{corr}(\|F_l^{\mathrm{T}}\|_2, E_{U_{\text{prin}}}) > 0$.

These results indicate that at intermediate depth (e.g., l=17) the high-norm (outlier) content lies relatively closer to the null subspace, whereas at the final depth (e.g., l=23) it aligns more with the principal subspace. Accordingly, initializing updates along the nullspace at intermediate layers achieves information preservation (by construction) while still enabling outlier suppression, consistent with our design objective.



(a) $r = 4, l = 17$ (b) $r = 4, l = 23$

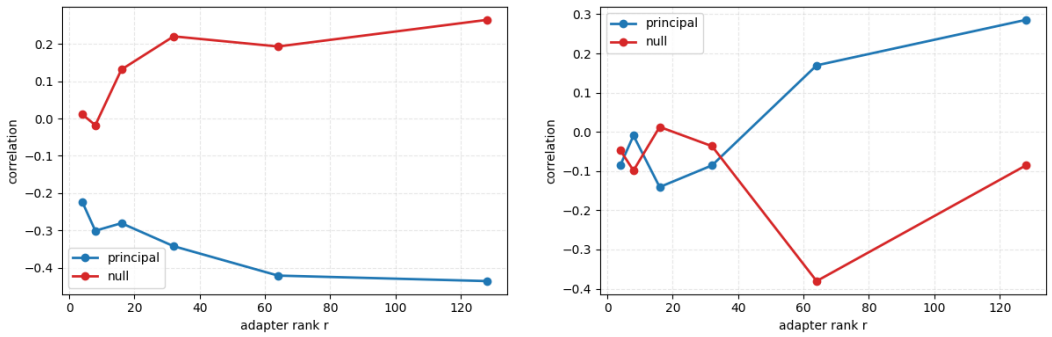
(c) $r = 8, l = 17$ (d) $r = 8, l = 23$

(e) $r = 16, l = 17$ (f) $r = 16, l = 23$

(g) $r = 32, l = 17$ (h) $r = 32, l = 23$

(i) $r = 64, l = 17$ (j) $r = 64, l = 23$

(k) $r = 128, l = 17$ (l) $r = 128, l = 23$



(m) correlation with basis at layer 17 (n) correlation with basis at layer 23

Figure 13: panels (a)–(l) show, for each setting, four views: the input image, the patchwise norm $\|F_l^{\mathrm{T}}\|_2$, and the subspace energies $E_{U_{\mathrm{prin}}}$ and $E_{U_{\mathrm{null}}}$. Panels (m)–(n) show correlation with each basis.

H Distilled Models in Long-Tail Learning

As reported in Table 1 in the main paper, the classification accuracy of the iNat2019 dataset did not overcome the baseline ViT-T. Still, SiNGER achieved the highest accuracy among distillation methods. We consider this borderline as the nature of KD training. Many studies pointed out the performance drop of distilled models in long-tail learning (Zhang et al., 2023; Yu et al., 2024; Huang

et al., 2025). They claim that pure knowledge distillation may harm the accuracy in long-tail learning due to the teacher model's bias transfer and hyperparameter sensitivity.

I Implementation Details

In this section, we report our implementation details for reproducibility. We trained our model on Ubuntu 22.04.5 LTS with CUDA v12.6.85 using eight NVIDIA GeForce RTX 3090 GPUs. Mixed precision training (FP16) was enabled to reduce memory consumption and accelerate computation. All experiments were conducted with a global batch size of 512, distributed evenly across GPUs using PyTorch's DistributedDataParallel (DDP).

For distillation, the 8th layer and 17th layer were selected as the distillation layers for Tiny ViT and Large ViT, respectively. We tuned the weights for the losses equally to 1.0. The model and adapters were optimized using the AdamW optimizer with a cosine annealing learning rate scheduler of a single cycle. The learning rate was initialized at 10^{-4} and decayed to 10^{-8} over 100 epochs. Weight decay was set to 0.05, and gradient clipping with a max norm of 1.0 was applied to stabilize training. Data augmentation followed common practice for ImageNet training, including random resized cropping, horizontal flipping. Input images were resized to 224×224 unless otherwise specified. During evaluation, a center crop was used. All hyperparameters and code will be released upon publication.

J The Use of Large Language Models

As per the ICLR 2026 guidelines on the use of Large Language Models (LLMs), we disclose that an LLM was used for minor grammar corrections and polishing of the text to enhance readability and for searching related research to broaden the scope of the literature review. The LLM did not contribute to the research ideation, methodology, or core findings of the paper.