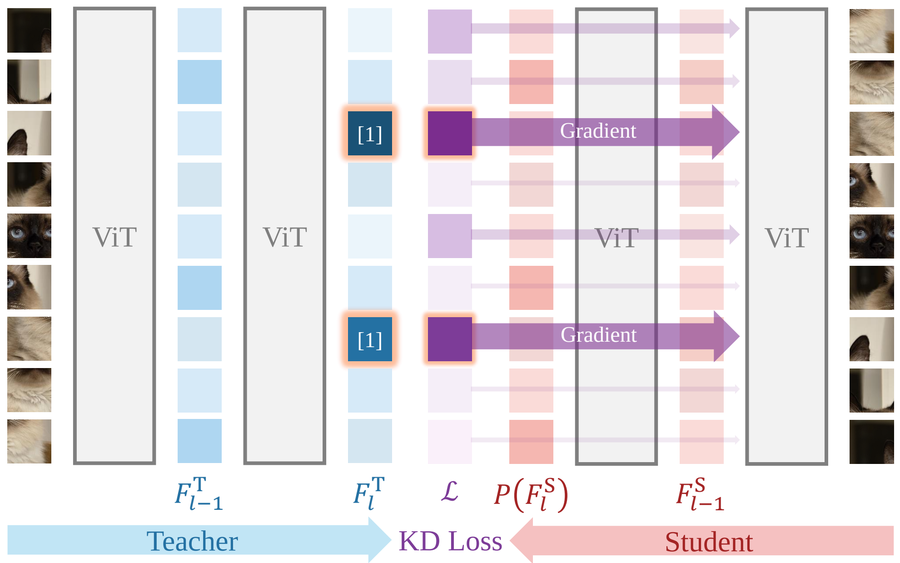
[1] KYUNG HEE UNIVERSITY [2] MOBILTECH

Geunhyeok Yu[1,*] Sunjae Jung[1,2,*] Yoonyoung Choi[1] Jaeseung Kim[2] Hyoseok Hwang[1,†]

SiNGER: A Clearer Voice Distills Vision Transformers Further

ICLR International Conference On Learning Representations

I. Gradient Dominance



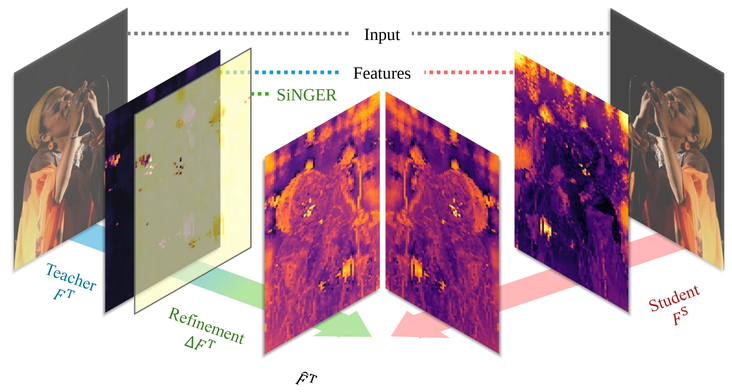
$$\mathcal{L} = \frac{1}{N}\sum_{i}\left\|P_l\left(F^{\mathrm{S}}_{l,i}\right) - F^{\mathrm{T}}_{l,i}\right\|_2^2 \quad \cdots \text{Distance-based loss is proportional to the norm}$$

$$\nabla_{P_l(F^{\mathrm{S}}_l)}\mathcal{L} = \frac{2}{N}\sum_{i} P_l\left(F^{\mathrm{S}}_{l,i}\right) - F^{\mathrm{T}}_{l,i} \quad \cdots \text{thus, the outliers dominate the gradient portion, lead to representation degradation.}$$

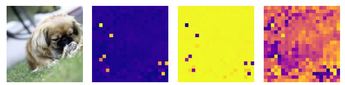
※ Contributions ※

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

II. Training Objectives

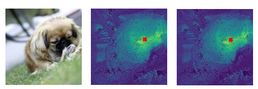


$$\mathcal{L}_{\text{outlier}} = \sum_{l\in\mathcal{D}}\frac{1}{|\mathcal{O}_l|}\sum_{i\in\mathcal{O}_l}\left(\left\|\hat{F}^{\mathrm{T}}_{l,i}\right\|_2 - q_{\alpha,l}\right)\cdot \quad q\text{: quantile}$$

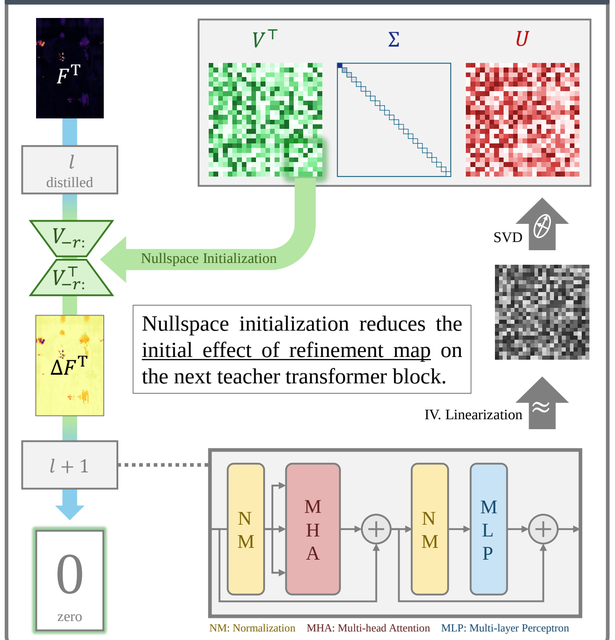


$$\mathcal{L}_{\text{info},l} = \begin{cases}\mathrm{MSE}\left(G(\hat{F}^{\mathrm{T}}_{l+1}), G(F^{\mathrm{T}}_{l+1})\right), & l \in l_{\text{inter}},\\ \mathrm{MSE}\left(G(\hat{F}^{\mathrm{T}}_{l}), G(F^{\mathrm{T}}_{l})\right), & l = l_{\text{final}}.\end{cases}$$

G: Gram matrix

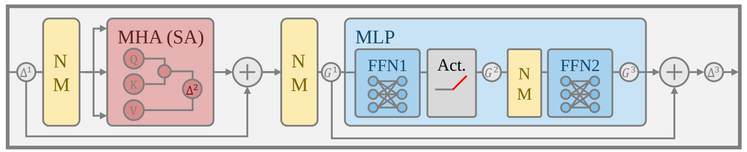


III. Adapter Initialization



IV. Linearization

- To enable SVD on inherently non-linear Transformer blocks, we utilize a linear surrogate $(\widetilde{W} = W_{\text{FFN1}}W_{\text{FFN2}})$ focusing on the FFN sub-layer, which our empirical analysis identifies as the primary driver of norm inflation.
- Verification against the full Jacobian baseline confirms that this FFN-centric proxy is highly robust, preserving the teacher's functional behavior with negligible output deviation.



Metric	SiNGER	Jacobian
L_2 (↓)	**0.169**	0.191
Cosine sim (↑)	**0.9787**	0.9564
CKA (↑)	**0.9975**	0.9947

Table 8. The output deviation of the non-linear block when inputs are perturbed along null directions computed b y ours versus the full Jacobian.

Metric	mean	median	p95
$G_{\text{FFN1}}(G^1, G^2)$	0.0719	0.0714	0.0844
$\boldsymbol{G_{\text{FFN2}}(G^2, G^3)}$	**0.7988**	**0.7945**	**0.8571**
$\Delta_{\text{FFN}}(\Delta^1, \Delta^2)$	0.1871	0.1852	0.2210
$\Delta_{\text{SA}}(\Delta^2, \Delta^3)$	0.1417	0.1415	0.1611

Table 7 in Appendix A. Identification of norm inflation. The p95 metric refers to the norm at the top 5% quantile. G: rhs/lhs, Δ: (rhs − lhs)/lhs.

V. Results

Teacher	Student	Distillation	IN-val top-1 (↑)	ADE-20K mIoU (↑)	NYUd-v2 RMSE (↓)	iNat2019 top-1 (↑)	Domain Shift top-1 (↑)	Fine-Grained top-1 (↑)
ViT-L [2]	ViT-T [2]	FitNet [4]	62.43	18.73	1.0093	40.02	32.32	62.48
		ViTKD [5]	5.07	11.92	1.1903	23.69	2.08	33.52
		SiNGER	**70.59**	**21.76**	**0.9406**	**41.11**	**38.87**	**64.61**
		Δ	+8.16	+3.03	+0.0687	+1.09	+6.55	+2.13
DeiT-III-L [3]	DeiT-III-B [3]	FitNet [4]	60.00	26.79	1.1625	50.04	31.53	74.78
		ViTKD [5]	66.54	19.58	1.2525	30.78	35.77	58.36
		SiNGER	**79.37**	**29.47**	**1.1514**	**55.50**	**49.14**	**57.41**
		Δ	+12.83	+2.68	+.0.0111	+3.46	+13.37	+0.63

Table 1. Multi-task evaluation results. Δ rows indicate the performance gains of SiNGER, computed against thebest-performing baseline among the distilled students (underlined).

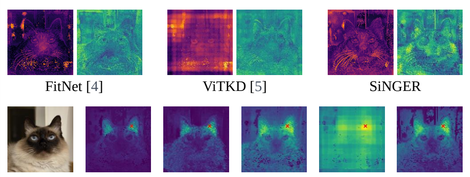


Figure 2. Qualitative analysis. Row 1: KD method comparison. Row 2: Feature map comparison.

Mat	Init	E_{prob} (↓)	E_{safe} (↑)
$\phi_\uparrow$	Rand	0.2565	0.2532
$\phi_\uparrow$	Ours	0.1479	0.8337
$\phi_\downarrow$	Rand	0.3100	0.2494
$\phi_\downarrow$	Ours	0.2847	0.5485

Table 5a. Ablation study on nullspace initialization ($l = 17$).

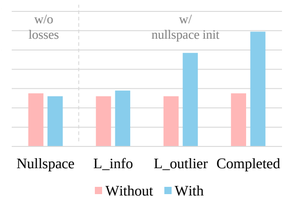


Figure from Table 4. Ablation study on the nullspace initialization and loss.

* Equal contribution. † Corresponding author. [1] Darcet et al., "Vision Transformers Need Registers", ICLR 2024 oral. [2] Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ICLR 2021. [3] Touvron, Cord, and Jégou, "DeiT III: Revenge of the ViT", ECCV 2022. [4] Romero et al., "FitNets: Hints for Thin Deep Nets", ICLR 2015. [5] Yang et al., "ViTKD: Feature-based Knowledge Distillation for Vision Transformers", CVPR 2024 Workshop.