SLAD : Shared LoRA Adapters for Task Specific Distillation

Supplementary Material

8. Additional Ablation Studies

8.1. Effect of LoRA Rank

We evaluate the sensitivity of SLAD to the rank r of the LoRA adapters on CUB (ViT-B → ViT-S). Results are reported in Table 4.

Rank r	2	4	8	16
Accuracy (%)	90.59	90.63	90.51	90.59

Table 4. Effect of LoRA rank on performance.

We observe that performance remains stable across ranks, indicating that SLAD is not sensitive to the choice of r.

8.2. Effect of Distillation Temperature

We analyze the impact of the temperature T used in the KL divergence term. Results are shown in Table 5 for CUB (ViT-B → ViT-S).

T	0.5	1	2	4	8
Accuracy (%)	89.66	89.99	90.59	**91.11**	90.82

Table 5. Effect of temperature on distillation performance.

Increasing the temperature improves performance up to $T = 4$, suggesting that softer teacher distributions provide richer supervision signals by exposing inter-class similarities. For very large values of T, performance slightly degrades due to over-smoothing of the distribution.

8.3. Effect of Loss Weights

We evaluate different combinations of loss weights $(\alpha_{KL}, \alpha_t, \alpha_s)$, corresponding to the KL divergence loss, teacher cross-entropy (CE) and student CE. Results are reported in Table 6 for CUB.

Weights	(1,1,1)	(2,1,1)	(1,2,2)	(4,1,1)	(2,2,1)	(4,2,1)
ViT-B → ViT-S	90.59	90.84	90.21	**90.89**	90.70	90.63
ViT-L → ViT-S	90.33	91.01	90.06	**91.16**	90.71	91.08

Table 6. Effect of loss weighting on performance.

We observe that increasing the weight of the KL divergence term consistently improves performance, highlighting the importance of strong distillation supervision. However, maintaining task supervision for both teacher and student remains necessary for stable optimization.

9. Additional Comparisons

9.1. Comparison with PEFT Methods

We compare SLAD with alternative parameter-efficient fine-tuning (PEFT) methods, including Singular Value Fine-tuning (SVF) [41] and Visual Prompt Tuning (VPT) [23], on CUB (ViT-L → ViT-S). Results are shown in Table 7.

Method	SVF	VPT	LoRA	SLAD
Accuracy (%)	90.09	90.61	90.92	**91.35**

Table 7. Comparison with alternative PEFT methods.

SLAD consistently outperforms alternative PEFT approaches. While LoRA already provides a good balance between adaptation capacity and feature preservation, SLAD further improves performance by explicitly coupling teacher and student through shared adapters.

10. Additional Results

10.1. Training Efficiency

Standard task-specific distillation requires sequential training of teacher and student, whereas SLAD jointly trains both models in a single stage.

Setting	Probing	LoRA	SLAD
ViT-B → ViT-S	10.36	7.77	**5.08**
ViT-L → ViT-S	17.51	17.49	**10.83**
ViT-L → ViT-B	18.06	19.17	**12.58**

Table 8. Training time (in hours) for CUB dataset.

Although each epoch is slightly more expensive, SLAD reduces overall wall-clock time by approximately 30–40%, making it more efficient in practice.

10.2. VTAB Natural Subset

We report additional results on Oxford Pets and Caltech101 using the same evaluation protocol. Results are presented in Table 9.

On these datasets, performance is already near saturation for all methods, resulting in relatively small performance gaps. Nevertheless, SLAD remains consistently competitive and often improves over both probing and LoRA,

Student	Teacher	Method	Pets	Caltech101
ViT-S	ViT-B	Probing	95.48	98.10
		LoRA	95.23	97.98
		SLAD	**95.64**	**98.39**
ViT-S	ViT-L	Probing	95.23	98.21
		LoRA	95.12	98.44
		SLAD	**95.53**	**98.27**
ViT-B	ViT-L	Probing	95.94	98.16
		LoRA	95.86	98.56
		SLAD	**96.08**	**98.16**

Table 9. Results on VTAB natural datasets (Oxford Pets and Caltech101).

demonstrating its effectiveness even in high-performance regimes.