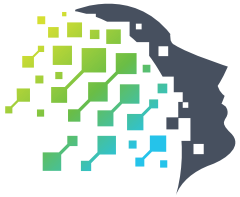


ICLR
International Conference On Learning Representations

SiNGER:
A Clearer Voice Distills Vision Transformers Further

Geunhyeok Yu[1,*] **Sunjae Jung**[1,2,*] **Yoonyoung Choi**[1]

Jaeseung Kim[2] **Hyoseok Hwang**[1,†]

1 AIRLab KYUNG HEE UNIVERSITY 2 MOBILETECH

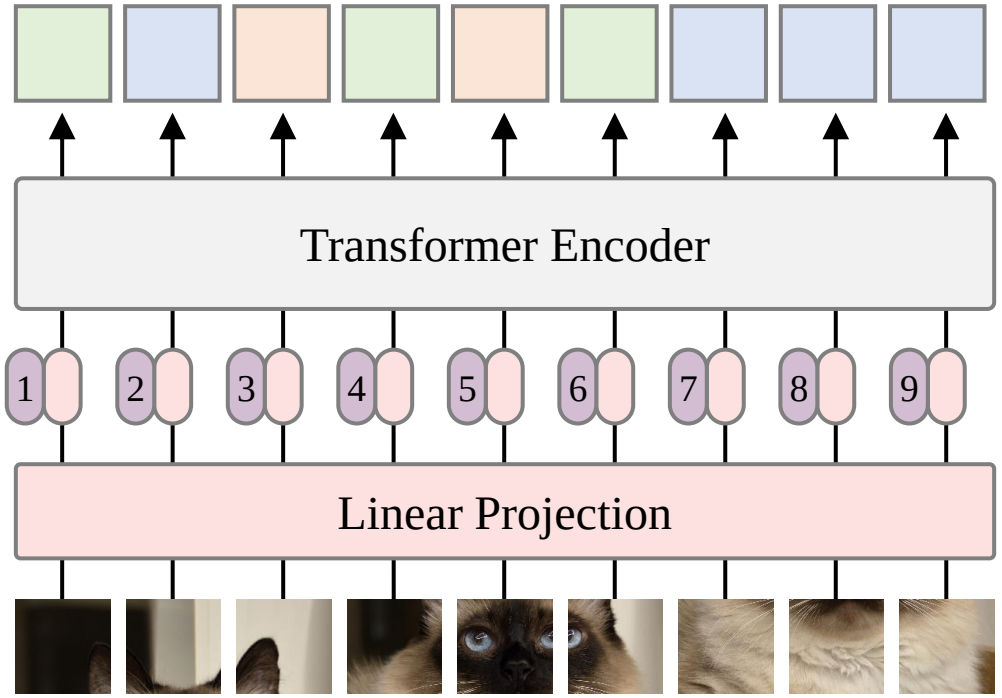
[*]Equal contribution [†]Corresponding author

Vision Transformers

Complexity

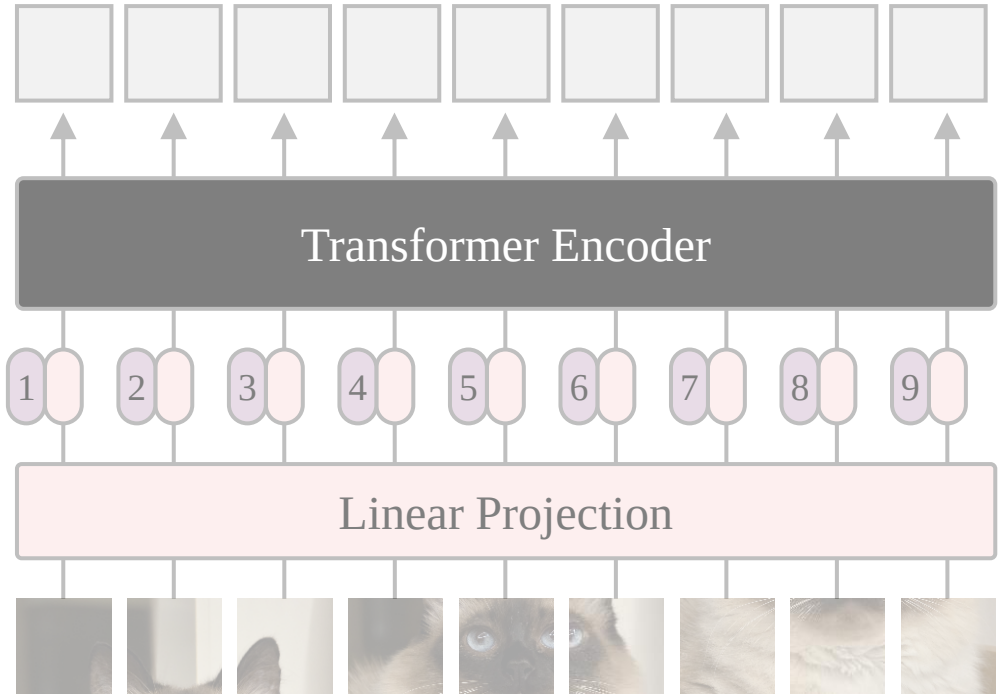
Vision Transformers

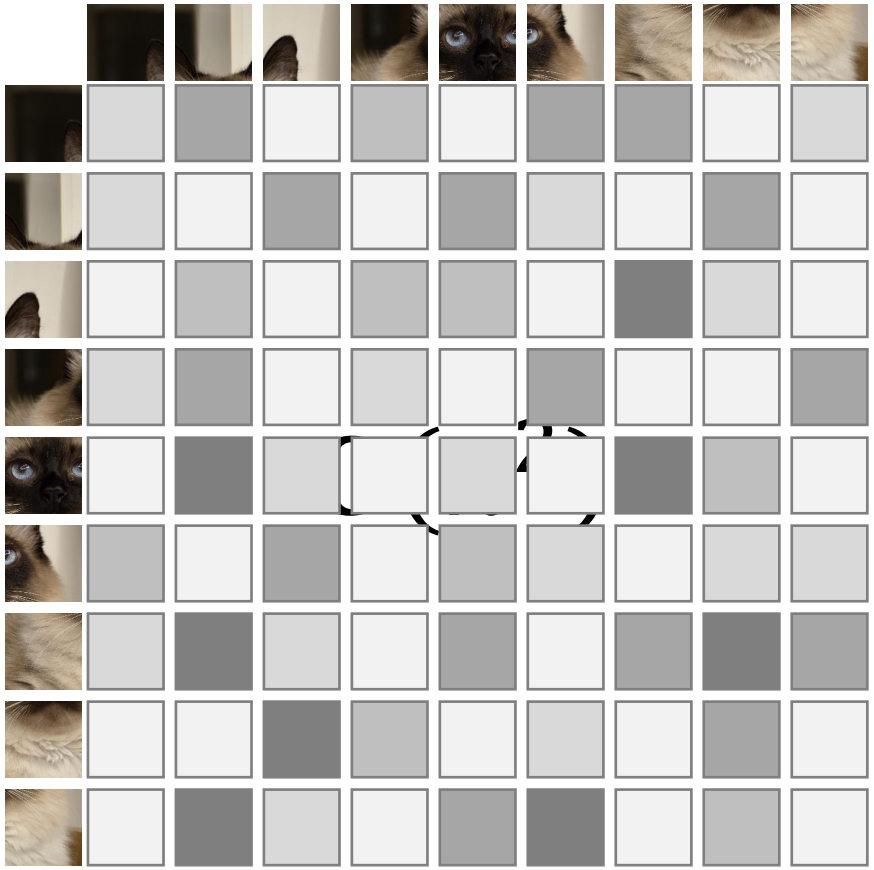
Complexity



Vision Transformers

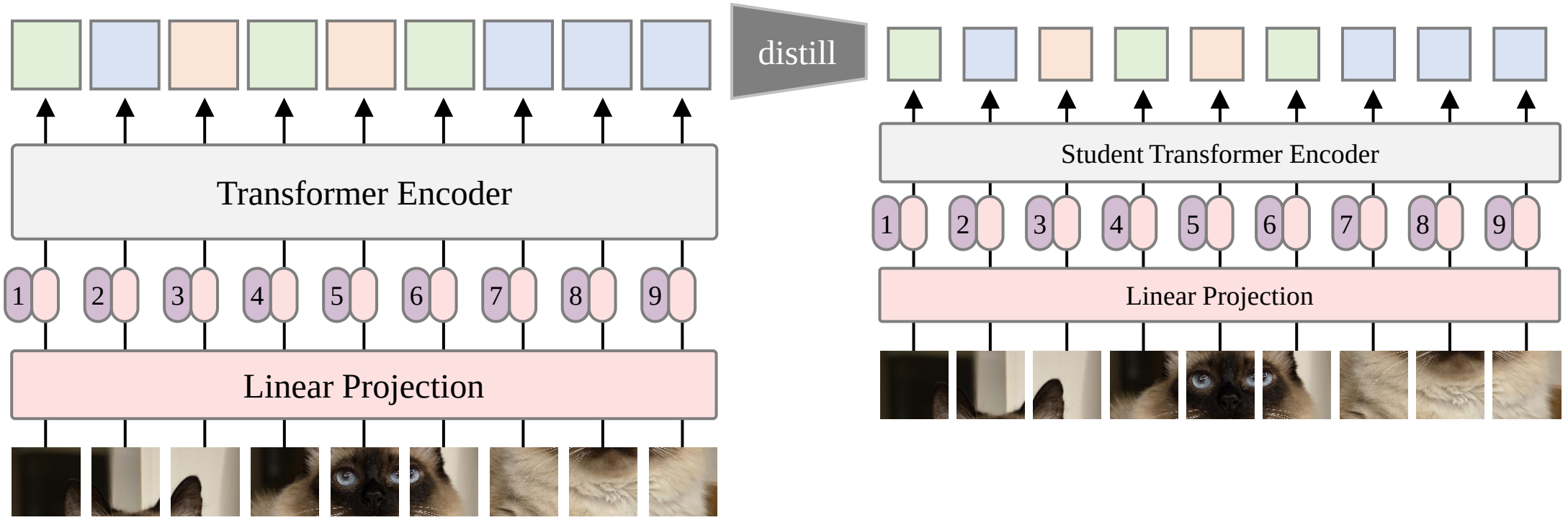
Complexity





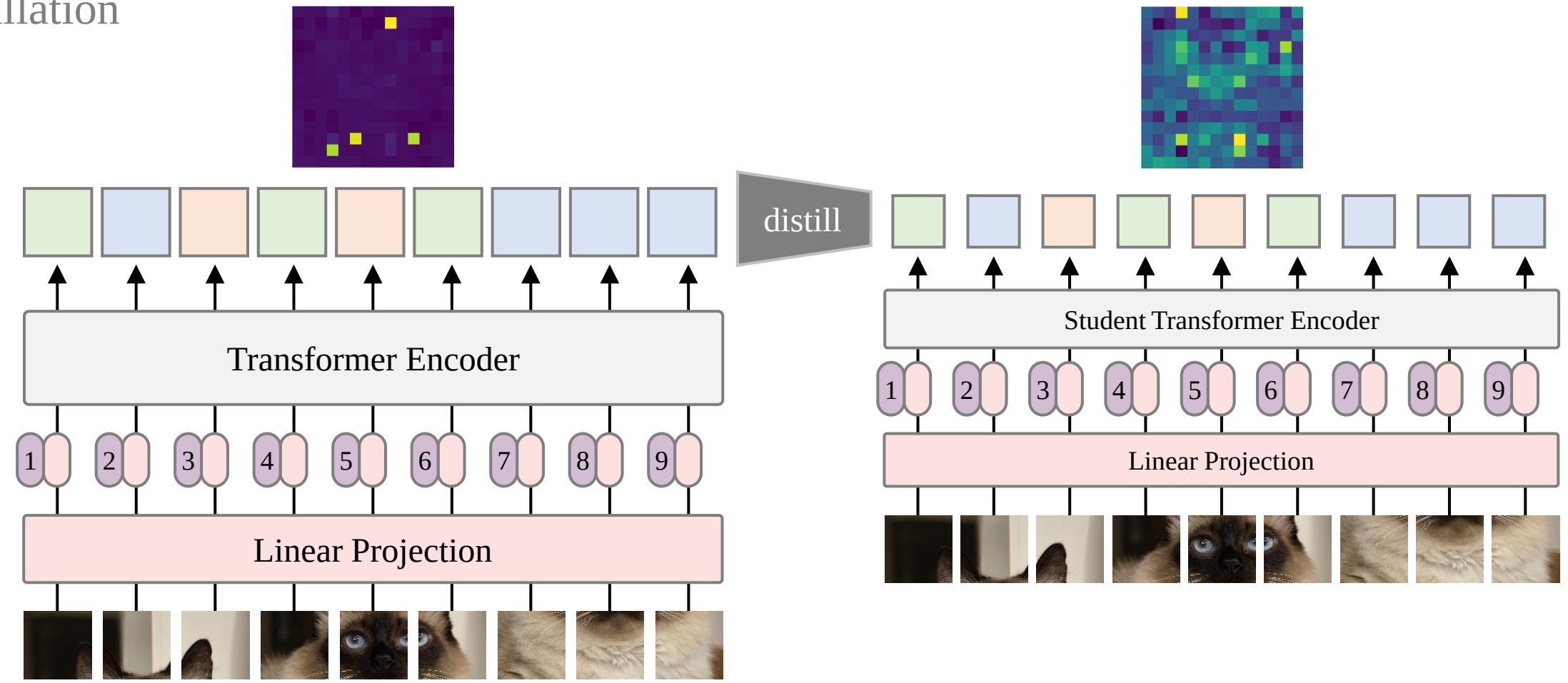
Vision Transformers

Distillation



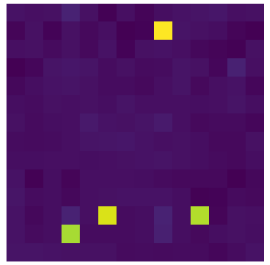
Artifacts

Distillation



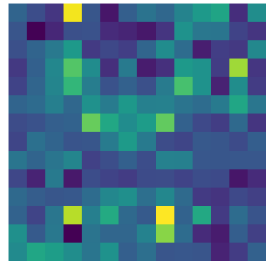
Artifacts

Distillation



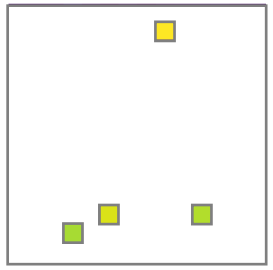
F^{T}

$$\sum \left\| F^{\mathrm{T}} - P\left(F^{\mathrm{S}}\right) \right\|^2$$



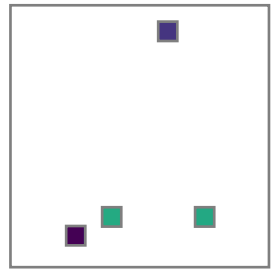
F^{S}

$=$



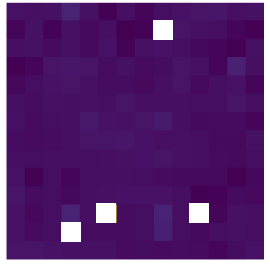
F_i^{T}

$$\sum_{i \in \mathcal{O}} \left\| F_i^{\mathrm{T}} - P\left(F_i^{\mathrm{S}}\right) \right\|^2$$



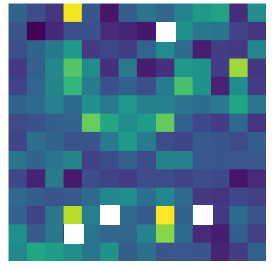
F_i^{S} ⋯ outliers

$+$



F_i^{T}

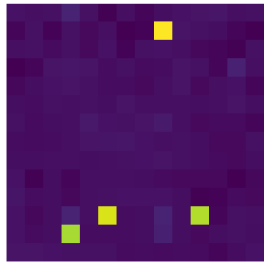
$$\sum_{j \in \mathcal{I}} \left\| F_j^{\mathrm{T}} - P\left(F_j^{\mathrm{S}}\right) \right\|^2$$



F_i^{S} ⋯ inliers

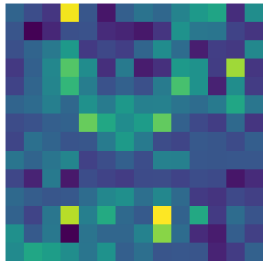
Artifacts

Distillation

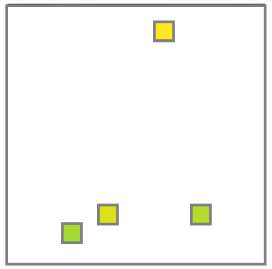


F^{T}

$$\underbrace{\sum \left\| F^{\mathrm{T}} - P\left(F^{\mathrm{S}}\right) \right\|^2}_{\mathcal{L}_{\mathrm{KD}}}$$

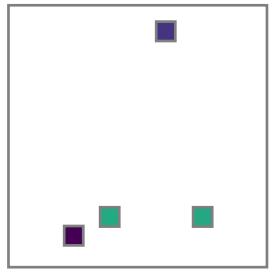


F^{S}



F_i^{T}

$$\sum_{i \in \mathcal{O}} \left\| F_i^{\mathrm{T}} - P\left(F_i^{\mathrm{S}}\right) \right\|^2$$

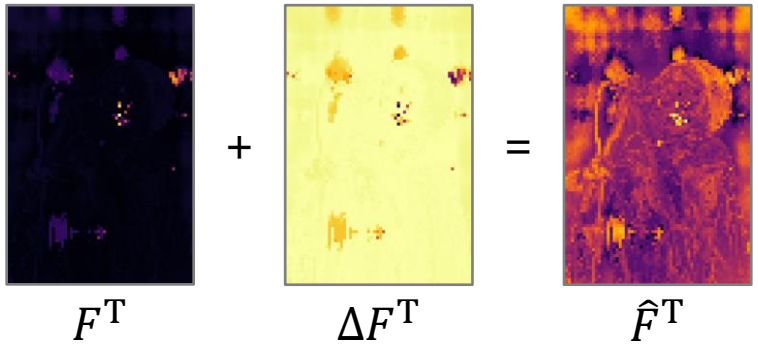


F_i^{S}

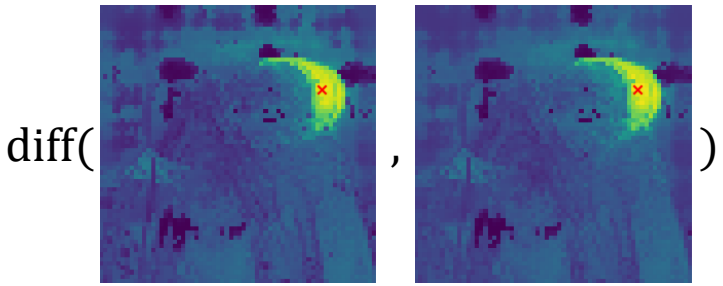
$$\nabla_{P\left(F_i^{\mathrm{S}}\right)} \mathcal{L}_{\mathrm{KD}} = \frac{2}{|\mathcal{O} \cup \mathcal{J}|} \left(P\left(F_i^{\mathrm{S}}\right) - F_i^{\mathrm{T}} \right)$$

SiNGER

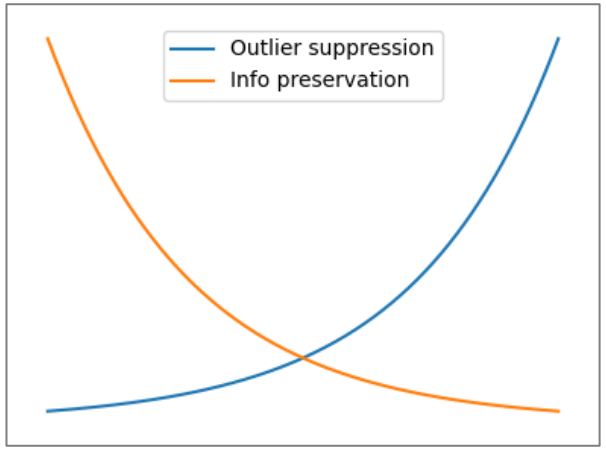
The Trade-Off



(a) Outlier suppression

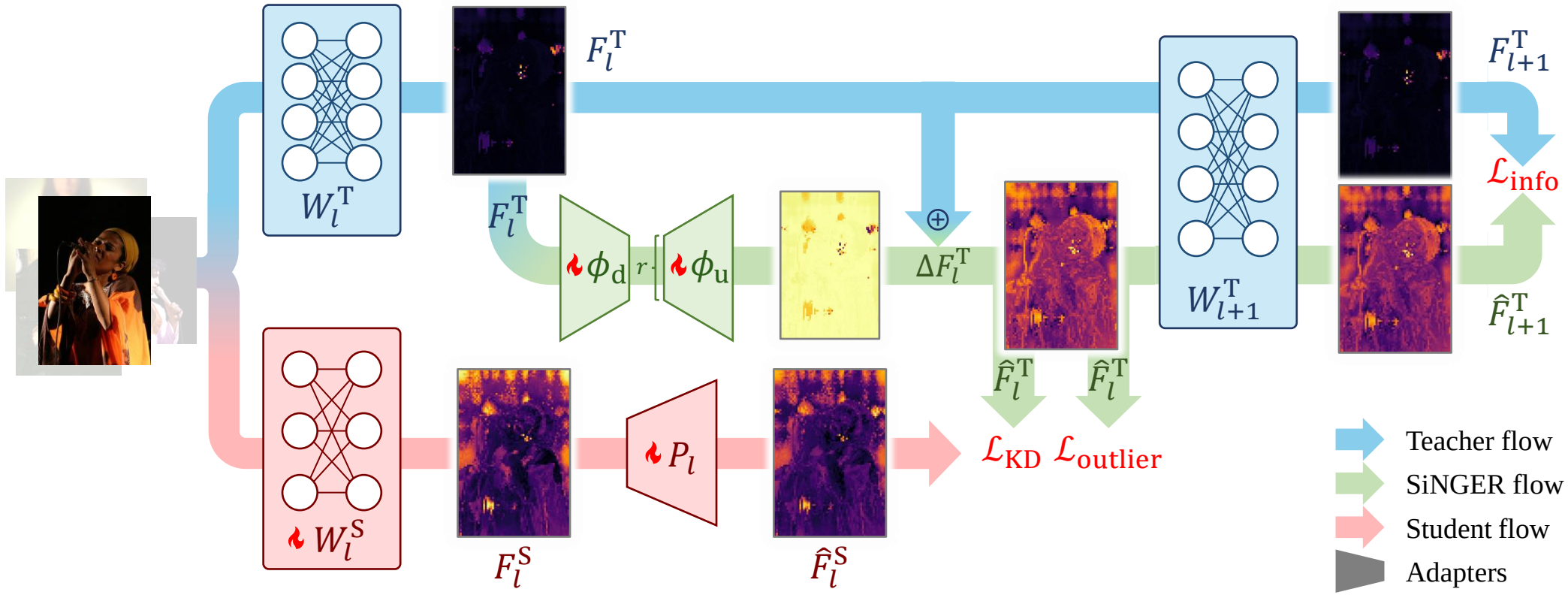


(b) Information preservation



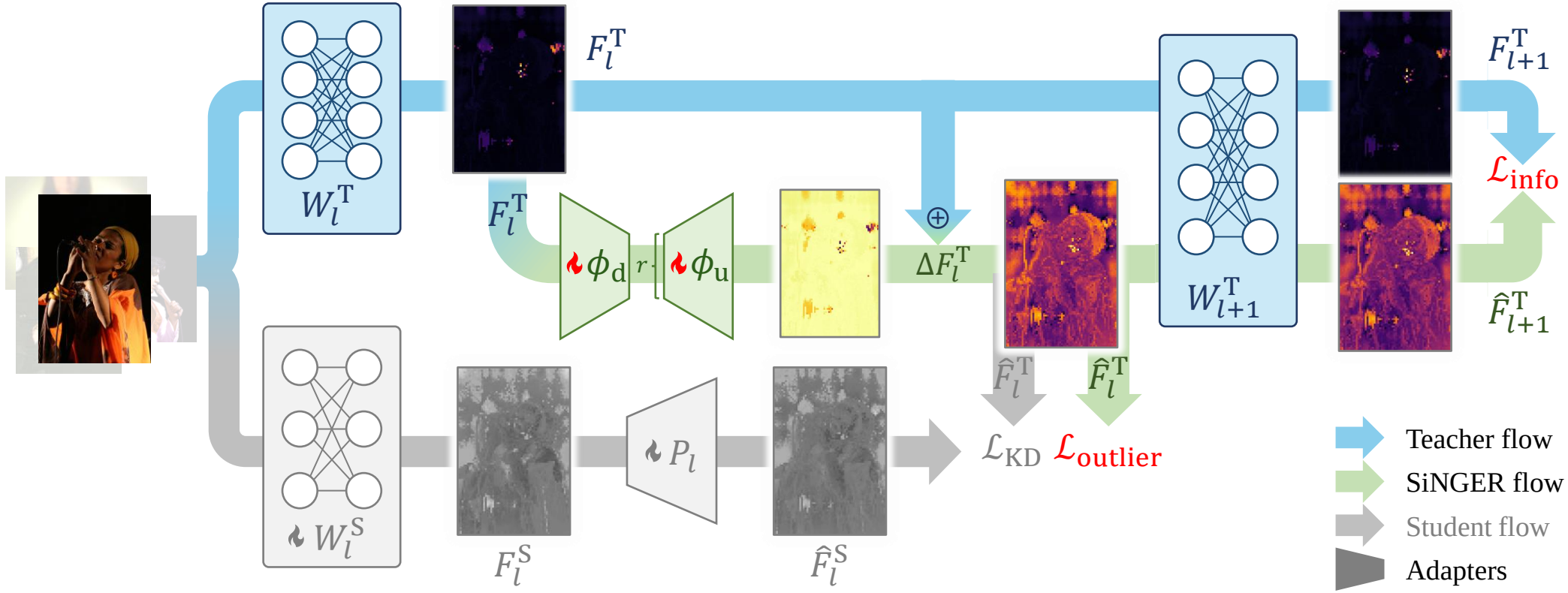
(c) Objective trade-off

SiNGER


W_l^{T}
F_l^{T}
F_l^{T}
ϕ_{d}
r
ϕ_{u}
ΔF_l^{T}
W_{l+1}^{T}
F_{l+1}^{T}
$\mathcal{L}_{\mathrm{info}}$
$\hat{F}_{l+1}^{\mathrm{T}}$
W_l^{S}
F_l^{S}
P_l
$\hat{F}_l^{\mathrm{S}}$
$\hat{F}_l^{\mathrm{T}}$
$\hat{F}_l^{\mathrm{T}}$
$\mathcal{L}_{\mathrm{KD}}$
$\mathcal{L}_{\mathrm{outlier}}$
Teacher flow
SiNGER flow
Student flow
Adapters

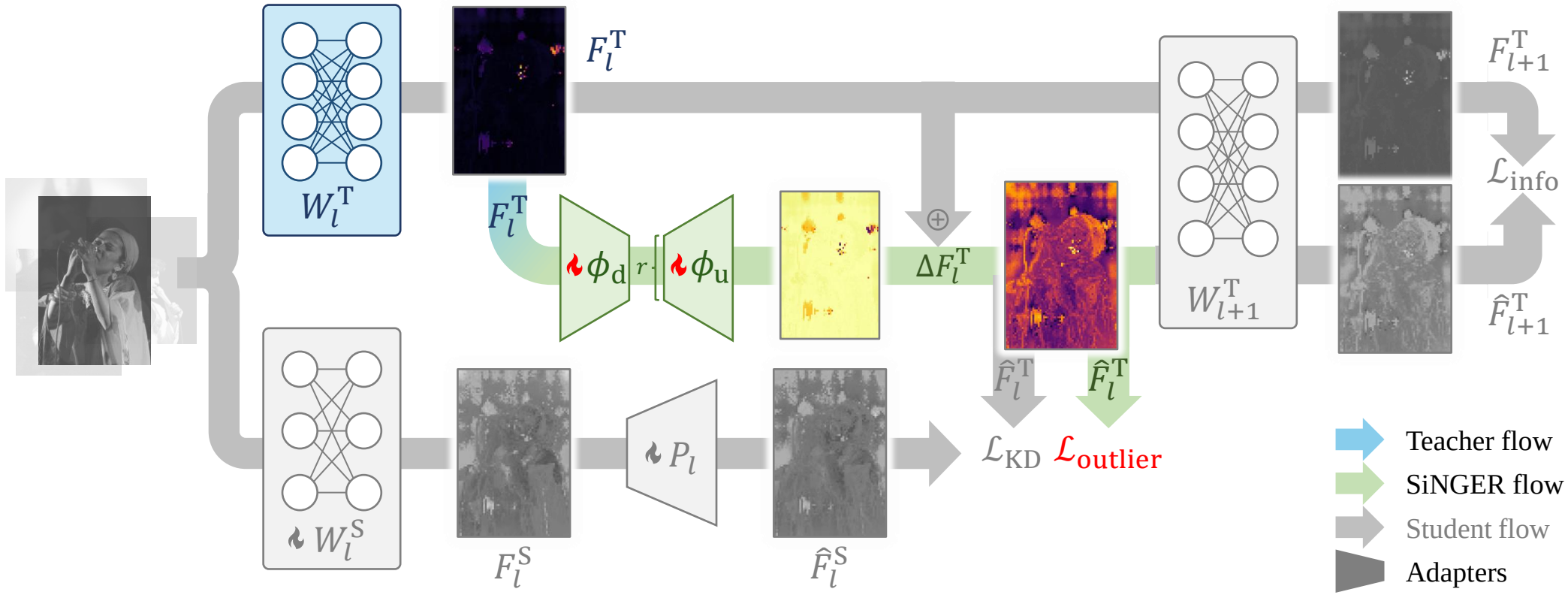
SiNGER

Joint Optimization



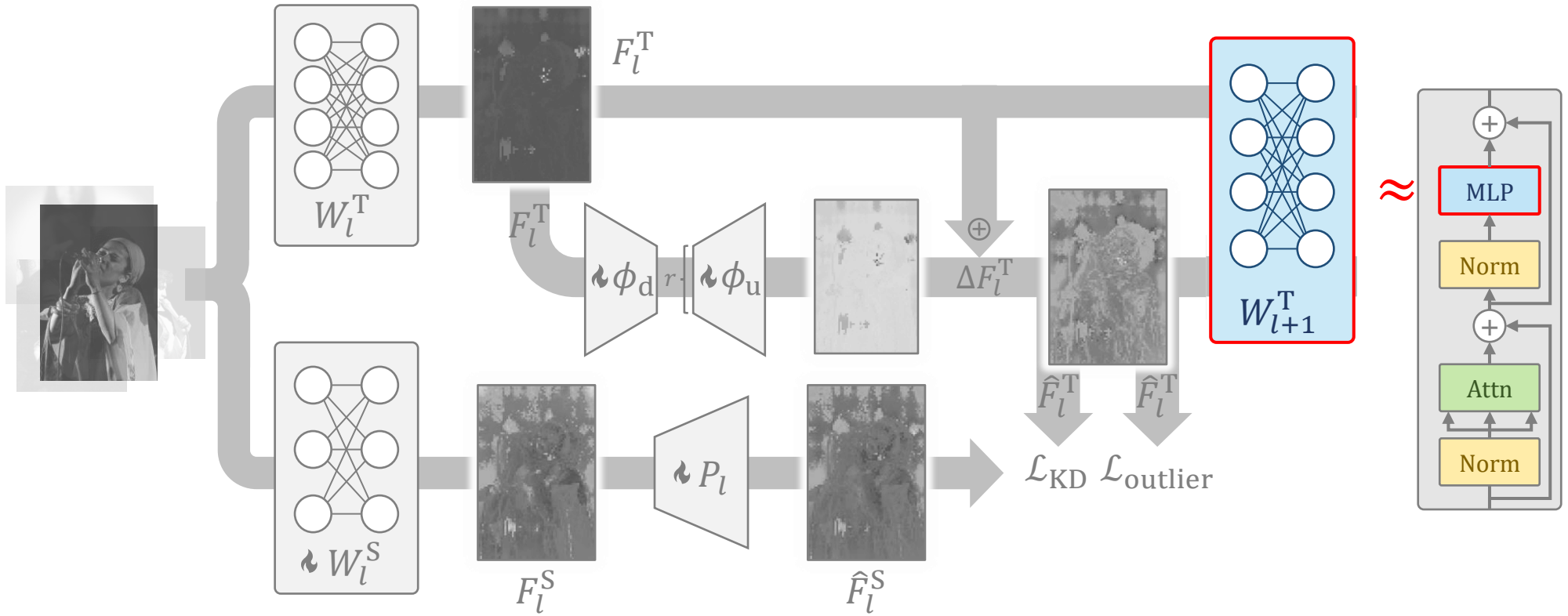
SiNGER

Artifact Suppression



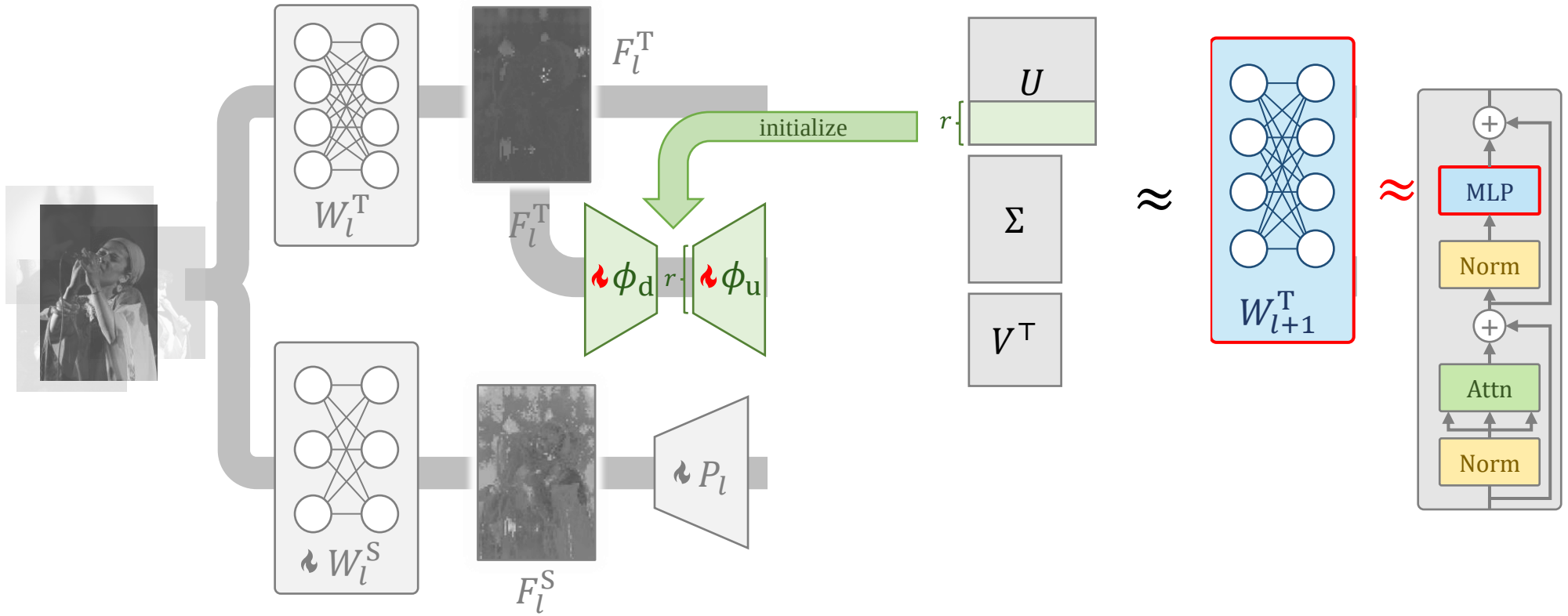
SiNGER

Artifact Suppression



SiNGER

Artifact Suppression



Quantitative Results

Performance Gain

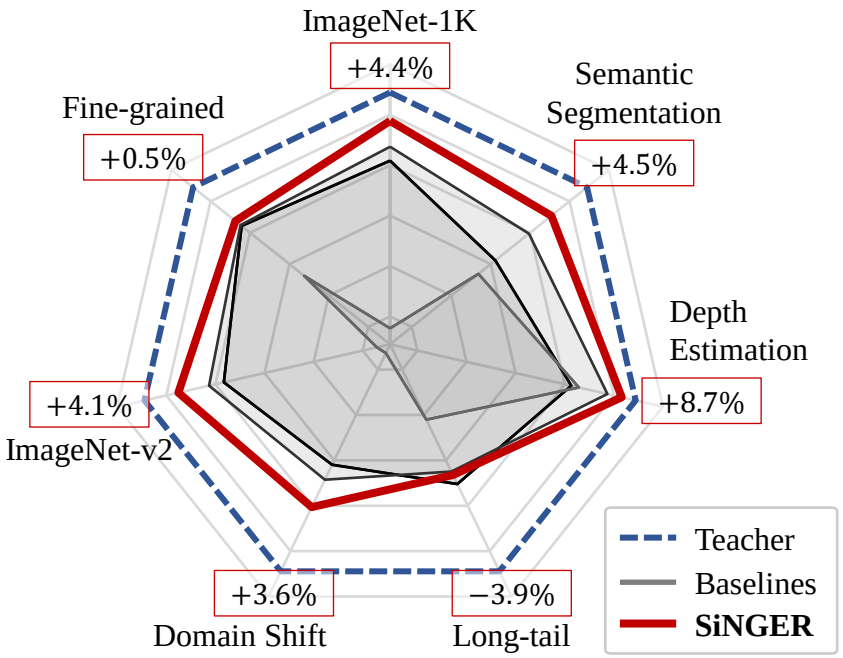


Figure 1b. Performance gains using SiNGER.

Teacher	Student	Method	IN-val top-1 (↑)	ADE-20K mIoU (↑)	NYUd-v2 RMSE (↓)	iNat2019 top-1 (↑)	DS top-1 (↑)	FG top-1 (↑)
ViT-L	ViT-T	FitNet	62.43	18.73	1.0093	40.02	23.23	62.48
		ViTKD	5.08	11.92	1.1903	23.69	2.08	33.52
		SiNGER	**70.59**	**21.76**	**0.9406**	**41.11**	**38.87**	**64.61**
		Δ	+8.19	+3.03	+0.0687	+1.09	+6.55	+2.13
DeiT-L	DeiT-B	FitNet	60.00	26.79	1.1625	50.04	31.53	74.78
		ViTKD	66.54	19.58	1.2525	30.78	35.77	58.36
		SiNGER	**79.37**	**29.47**	**1.1514**	**53.50**	**49.14**	**75.41**
		Δ	+12.83	+2.68	+0.0111	+3.46	+13.37	+0.63

Table1. Multi-task linear evaluation results.

Quantitative Results

Outlier Suppression

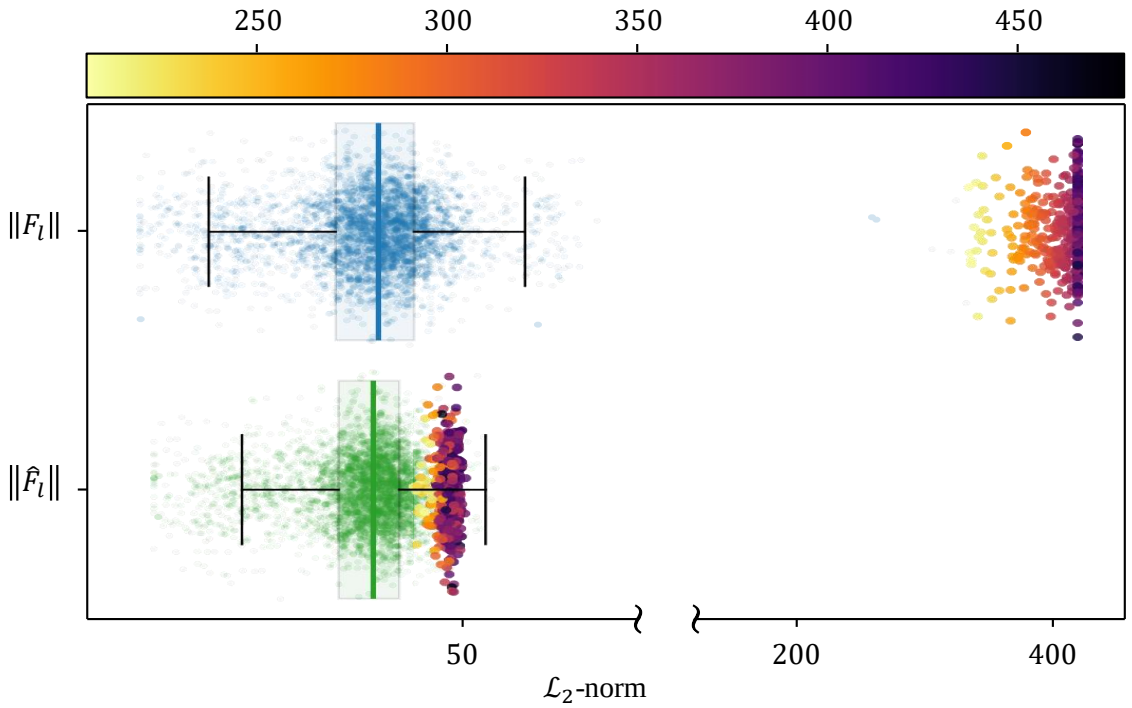


Figure 6. Patch-norm distributions of F_l and $\hat{F}_l$.

Quantitative Results

Outlier Suppression

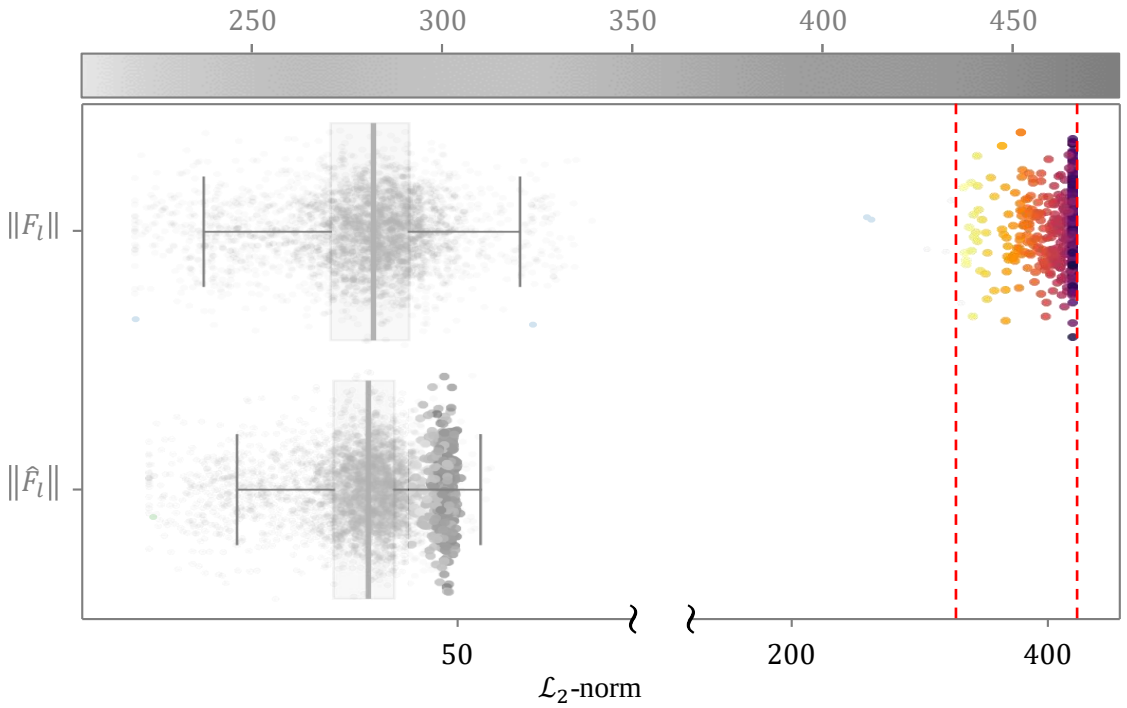


Figure 6. Patch-norm distributions of F_l and $\hat{F}_l$.

Quantitative Results

Outlier Suppression

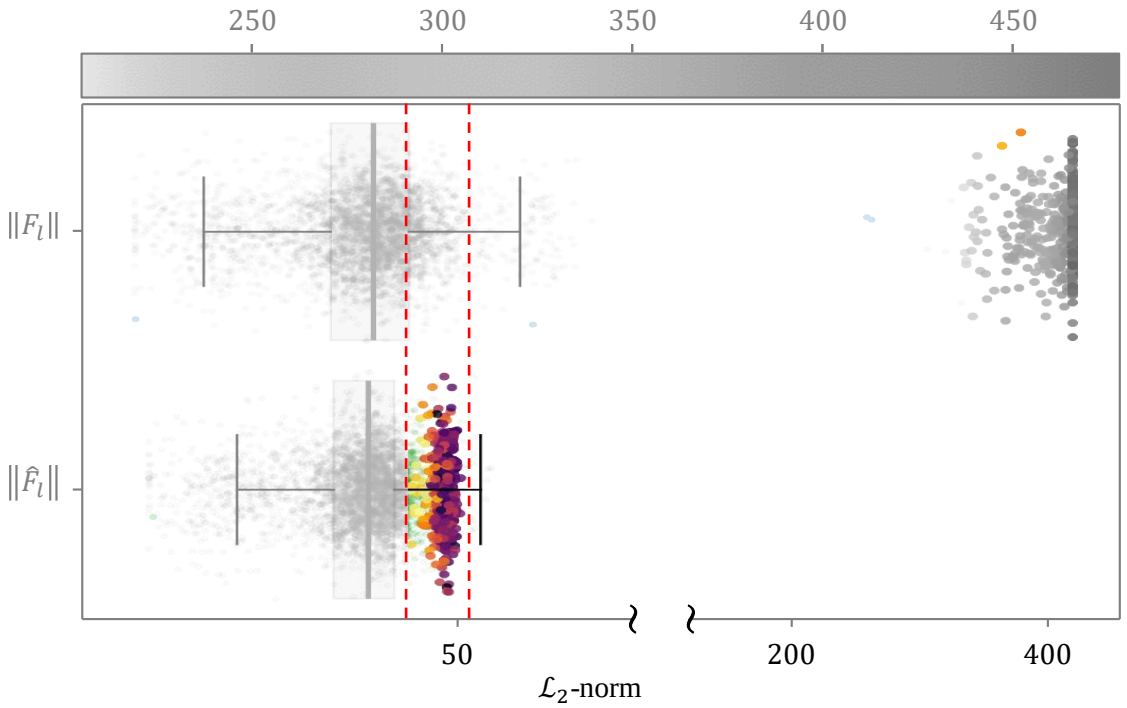


Figure 6. Patch-norm distributions of F_l and $\hat{F}_l$.

Quantitative Results

Outlier Suppression

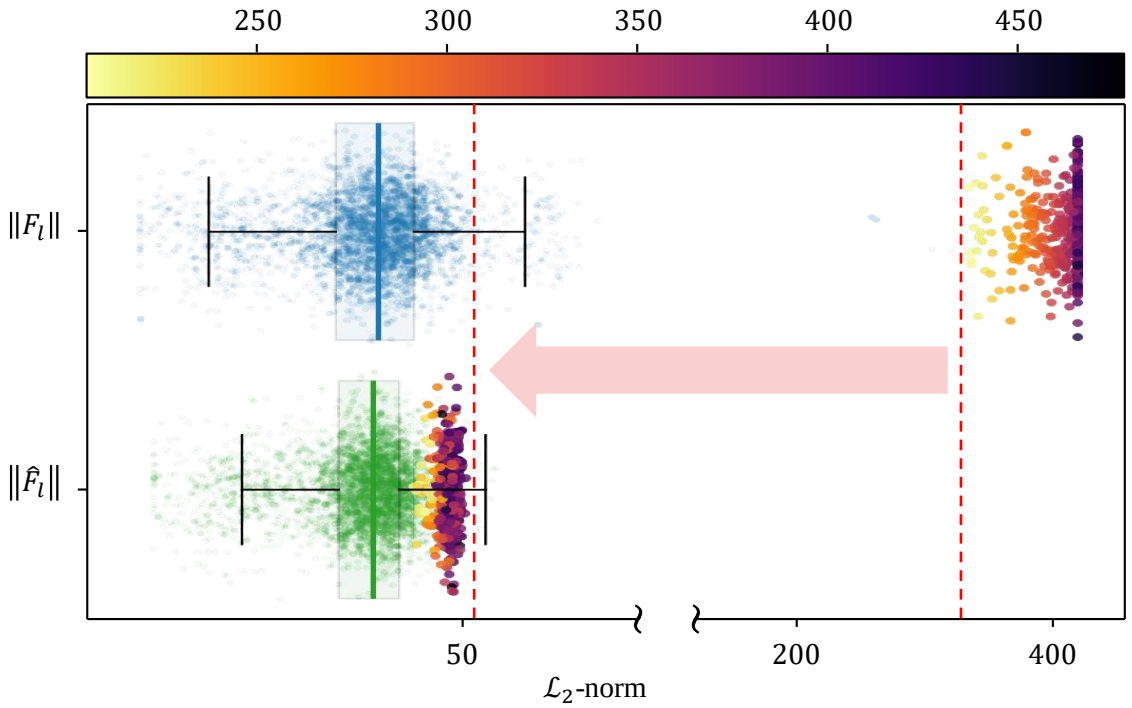


Figure 6. Patch-norm distributions of F_l and $\hat{F}_l$.

Quantitative Results

Information Preservation

Layer	Matrix	Init	E_{prob} (↓)	E_{safe} (↑)
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	$\phi^{\top}_{\text{down},l}$	Random	0.3100	0.2494
17	$\phi^{\top}_{\text{down},l}$	SiNGER	**0.2847**	**0.5485**
23	$\phi^{\top}_{\text{down},l}$	Random	0.3025	0.2641
23	$\phi^{\top}_{\text{down},l}$	SiNGER	**0.2746**	**0.5774**

Table 5a. Initialization methods.

Pair	$\mathcal{L}_{\text{outlier}}$	$\mathcal{L}_{\text{info}}$	mean ± std(↓)	median
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{T}}$	✓		12.22 ± 1.45	14.28
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{T}}$	✓	✓	7.25 ± 0.84	7.19
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{S}}$	✓		73.36 ± 7.67	71.85
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{S}}$	✓	✓	41.71 ± 7.01	40.89

Table 5b. Information preservation term.

Metric	FitNet	ViTKD	SiNGER
Gram Distance	0.237	0.520	**0.130**
CKA	0.732	0.745	**0.660**

Table 2. Teacher-Student similarity.

Quantitative Results

Information Preservation

Layer	Matrix	Init	E_{prob} (↓)	E_{safe} (↑)
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	$\phi^{\top}_{\text{down},l}$	Random	0.3100	0.2494
17	$\phi^{\top}_{\text{down},l}$	SiNGER	**0.2847**	**0.5485**
23	$\phi^{\top}_{\text{down},l}$	Random	0.3025	0.2641
23	$\phi^{\top}_{\text{down},l}$	SiNGER	**0.2746**	**0.5774**

Table 5a. Initialization methods.

Pair	$\mathcal{L}_{\text{outlier}}$	$\mathcal{L}_{\text{info}}$	mean ± std(↓)	median
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{T}}$	✓		12.22 ± 1.45	14.28
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{T}}$	✓	✓	7.25 ± 0.84	7.19
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{S}}$	✓		73.36 ± 7.67	71.85
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{S}}$	✓	✓	41.71 ± 7.01	40.89

Table 5b. Information preservation term.

Metric	FitNet	ViTKD	SiNGER
Gram Distance	0.237	0.520	**0.130**
CKA	0.732	0.745	**0.660**

Table 2. Teacher-Student similarity.

Quantitative Results

Information Preservation

Layer	Matrix	Init	E_{prob} (↓)	E_{safe} (↑)
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	$\phi^{\top}_{\text{down},l}$	Random	0.3100	0.2494
17	$\phi^{\top}_{\text{down},l}$	SiNGER	**0.2847**	**0.5485**
23	$\phi^{\top}_{\text{down},l}$	Random	0.3025	0.2641
23	$\phi^{\top}_{\text{down},l}$	SiNGER	**0.2746**	**0.5774**

Table 5a. Initialization methods.

Pair	$\mathcal{L}_{\text{outlier}}$	$\mathcal{L}_{\text{info}}$	mean ± std(↓)	median
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{T}}$	✓		12.22 ± 1.45	14.28
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{T}}$	✓	✓	7.25 ± 0.84	7.19
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{S}}$	✓		73.36 ± 7.67	71.85
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{S}}$	✓	✓	41.71 ± 7.01	40.89

Table 5b. Information preservation term.

Metric	FitNet	ViTKD	SiNGER
Gram Distance	0.237	0.520	**0.130**
CKA	0.732	0.745	**0.660**

Table 2. Teacher-Student similarity.

Quantitative Results

Information Preservation

Layer	Matrix	Init	E_{prob} (↓)	E_{safe} (↑)
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	$\phi^{\top}_{\text{down},l}$	Random	0.3100	0.2494
17	$\phi^{\top}_{\text{down},l}$	SiNGER	**0.2847**	**0.5485**
23	$\phi^{\top}_{\text{down},l}$	Random	0.3025	0.2641
23	$\phi^{\top}_{\text{down},l}$	SiNGER	**0.2746**	**0.5774**

Table 5a. Initialization methods.

Pair	$\mathcal{L}_{\text{outlier}}$	$\mathcal{L}_{\text{info}}$	mean ± std(↓)	median
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{T}}$	✓		12.22 ± 1.45	14.28
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{T}}$	✓	✓	7.25 ± 0.84	7.19
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{S}}$	✓		73.36 ± 7.67	71.85
$F^{\text{T}} \leftrightarrow \hat{F}^{\text{S}}$	✓	✓	41.71 ± 7.01	40.89

Table 5b. Information preservation term.

Metric	FitNet	ViTKD	SiNGER
Gram Distance	0.237	0.520	**0.130**
CKA	0.732	0.745	**0.660**

Table 2. Teacher-Student similarity.

Qualitative Result

Feature Visualization

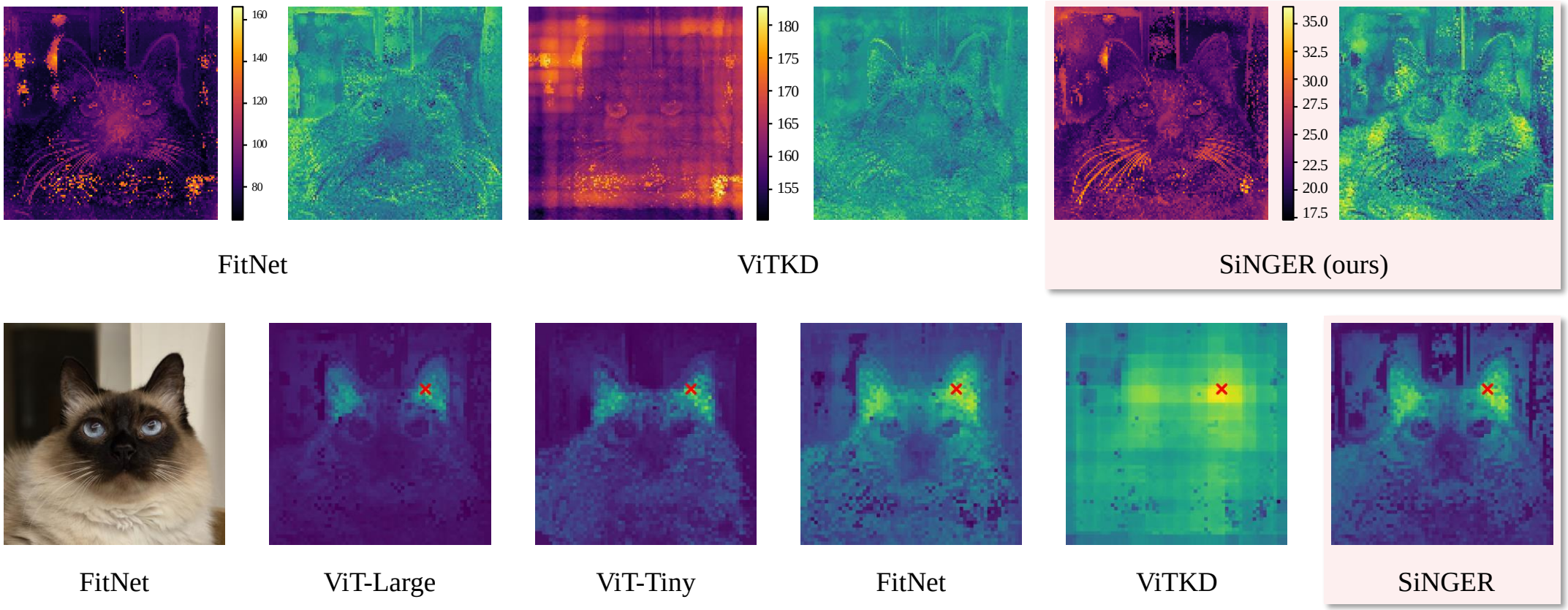


Figure 2. Qualitative analysis. Row 1: KD method comparison. Row 2: Feature map comparison.

Summary

- ✓ 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.

Thank You

Website: airlabkhu.github.io/SiNGER

E-mail: sunj@khu.ac.kr