Swiss Army Knife: Synergizing BiAses in Knowledge from Vision Foundation Models For Multi-Task Learning

Yuxiang Lu*, Shengcao Cao*, Yu-Xiong Wang
Shanghai Jiao Tong University, University of Illinois Urbana-Champaign

Project Code

ICLR

TL;DR

Different VFMs excel at different tasks — what if we could combine their strengths?

Introducing SAK: a "Swiss Army Knife" approach that preserves and exploits the unique representation biases of each model during distillation, optimizing their power for multiple downstream tasks.

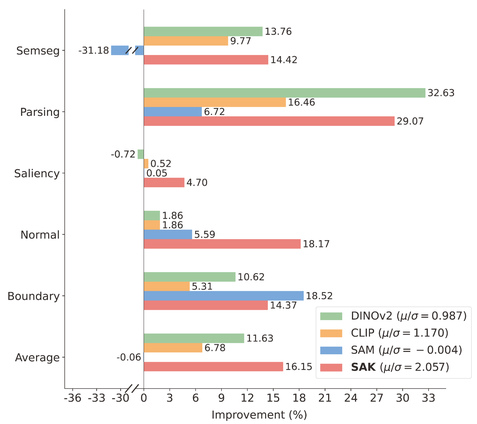
Motivation

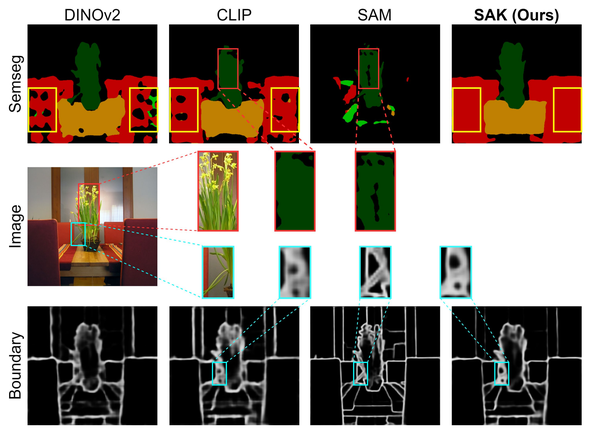
- VFMs are pretrained on diverse datasets, image resolutions, and objectives, introducing **representation biases**
- No single model achieves consistently superior performance across all domains
- Multi-teacher VFM distillation is effective and efficient, but many-to-one distillation risks eliminating the teacher biases and strengths, highlighting the importance of maintaining biases during distillation

Model	Semseg mIoU↑	Parsing mIoU↑	Saliency maxF↑	Normal mErr↓	Boundary odsF↑
Oracle of teachers	81.18 (DINOv2)	74.38 (DINOv2)	81.48 (CLIP)	16.21 (SAM)	75.89 (SAM)
Student w/o biases	80.18 (↓ 1.23%)	69.13 (↓ 7.06%)	82.72 (↑ 1.52%)	16.00 (↑ 1.30%)	71.16(↓ 6.23%)

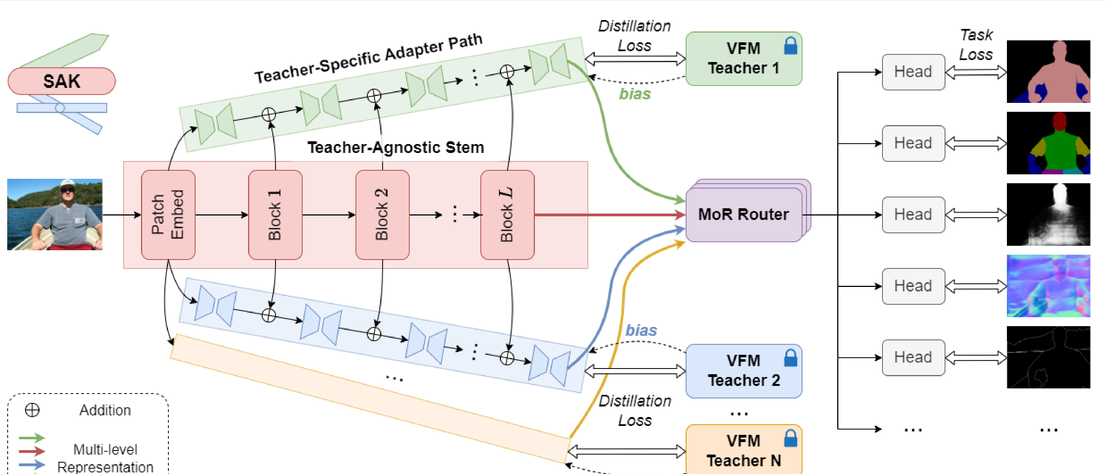
Can we preserve the representation biases of multiple VFMs during distillation to maximize multi-task performance?

Representation Biases in VFM

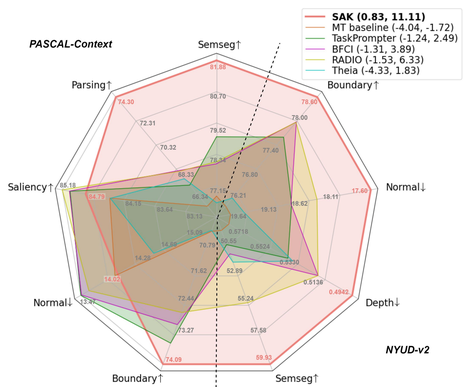


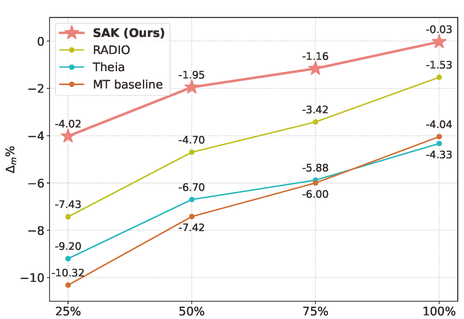


SAK Framework



Experiments





Model	Backbone	#Param	Semseg mIoU↑	Depth RMSE↓	Normal mErr↓	Boundary odsF↑	Δ_m% ↑
Single-task baseline	ViT-L	1259M	54.19	0.5560	19.22	78.09	0.00
Multi-task baseline	ViT-L	346M	52.42	0.5413	19.29	76.50	-0.76
InvPT (Ye & Xu, 2022)	ViT-L	402M	53.56	0.5183	19.04	78.10	1.64
InvPT++ (Ye & Xu, 2024)	ViT-L	~402M	53.85	0.5096	18.67	78.10	2.65
TaskPrompter (Ye & Xu, 2023b)	ViT-L	392M	55.30	0.5152	18.47	78.20	3.36
TaskExpert (Ye & Xu, 2023a)	ViT-L	400M+	55.35	0.5157	18.54	78.40	3.33
BFCI (Zhang et al., 2023b)	ViT-L	400M+	55.51	0.4930	18.47	78.22	4.46
3D-aware (Li et al., 2024a)	ViT-L	409M	54.87	0.5006	18.55	78.30	3.74
TSP (Wang et al., 2024b)	ViT-L	402M	55.39	0.4961	18.44	77.50	4.07
MLoRE (Yang et al., 2024d)	ViT-L	552M	55.96	0.5076	18.33	78.43	4.26
RADIO (Ranzinger et al., 2024b)	ViT-L	362M	59.32	0.4698	17.46	79.41	8.95
SAK (Ours)	ViT-L	394M	**63.18**	**0.4313**	**16.25**	**79.43**	**14.05**