JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

WED-PM-347

Generic-to-Specific Distillation of Masked Autoencoders

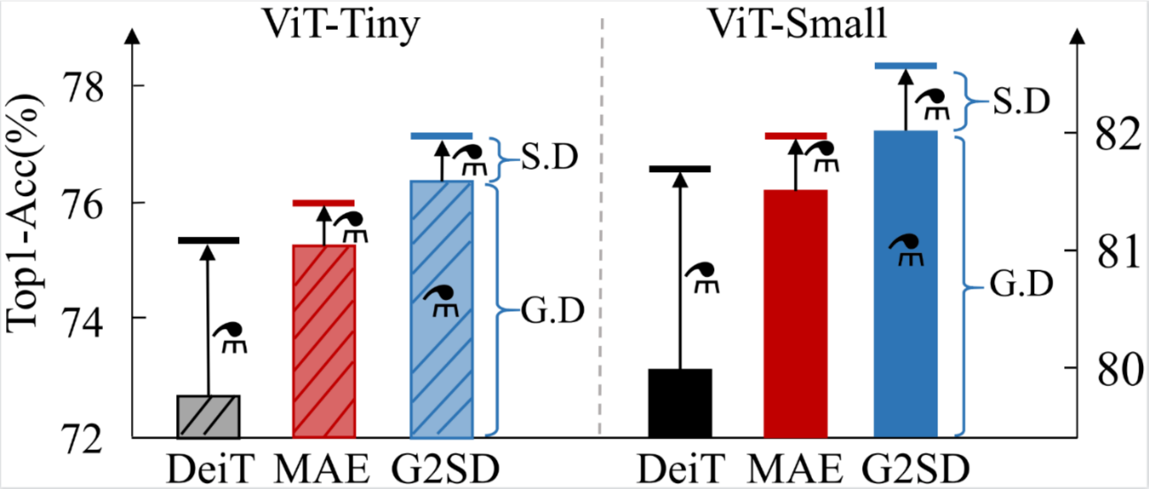
Wei Huang, Zhiliang Peng, Li Dong, Furu Wei, Jianbin Jiao, Qixiang Ye
weihuang19@mails.ucas.ac.cn
University of Chinese Academy of Sciences

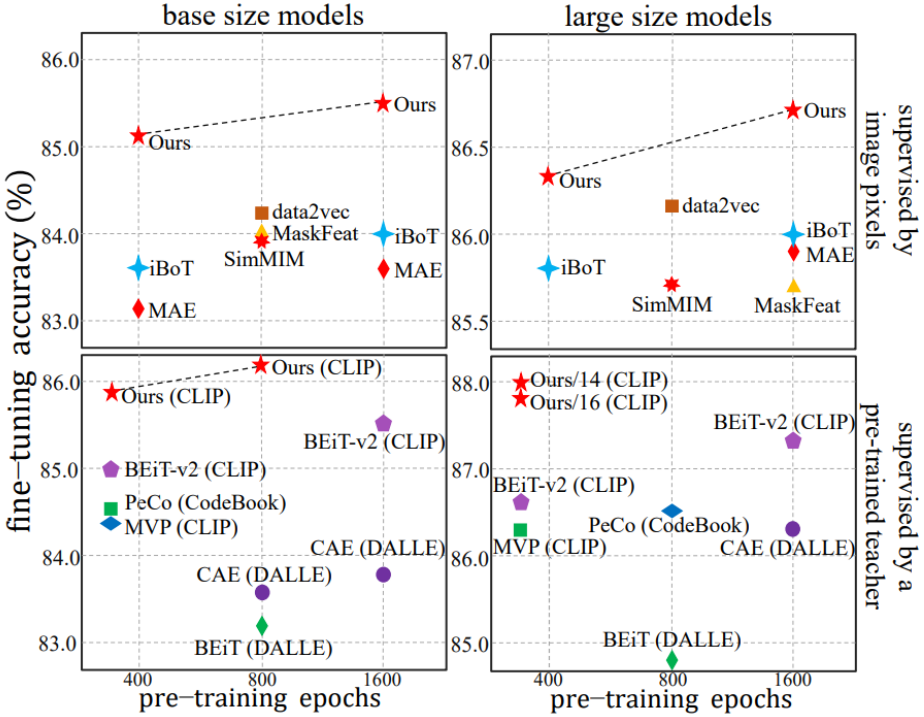
2023.06

中国科学院大学
University of Chinese Academy of Sciences

PriSDL

Motivation





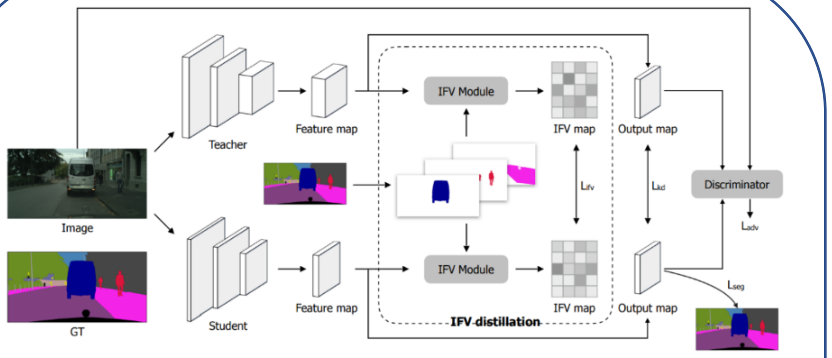
Tian, Yunjie, et al. "Integrally Pre-Trained Transformer Pyramid Networks." *arXiv preprint arXiv:2211.12735* (2022).

Lightweight ViTs :
self-supervised methods fall behind supervised distillation

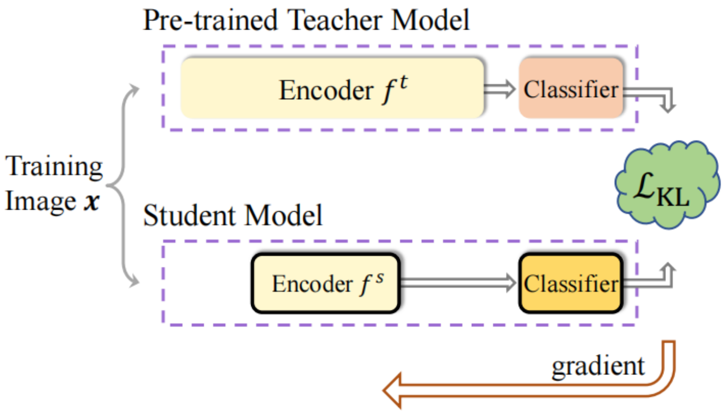
Large ViTs :
benefit a lot from self-supervised pre-training

Motivation

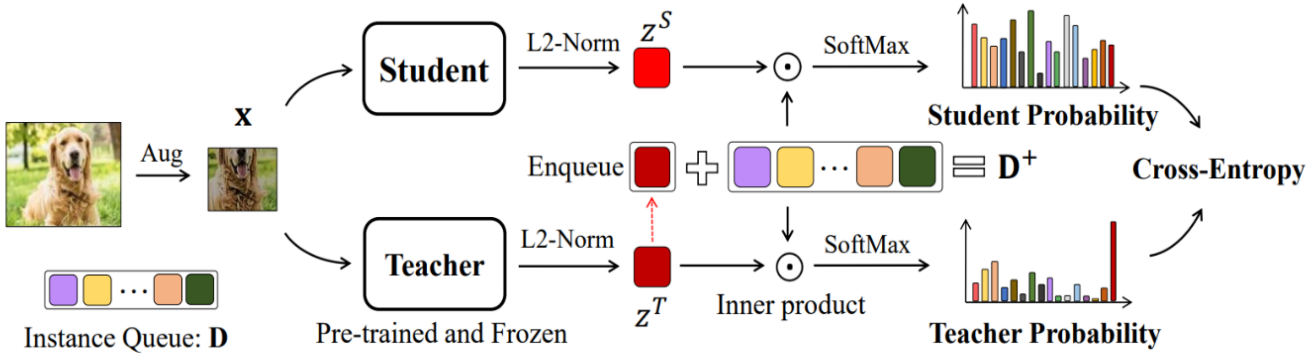
Single-stage distillation



Semantic segmentation



Classification

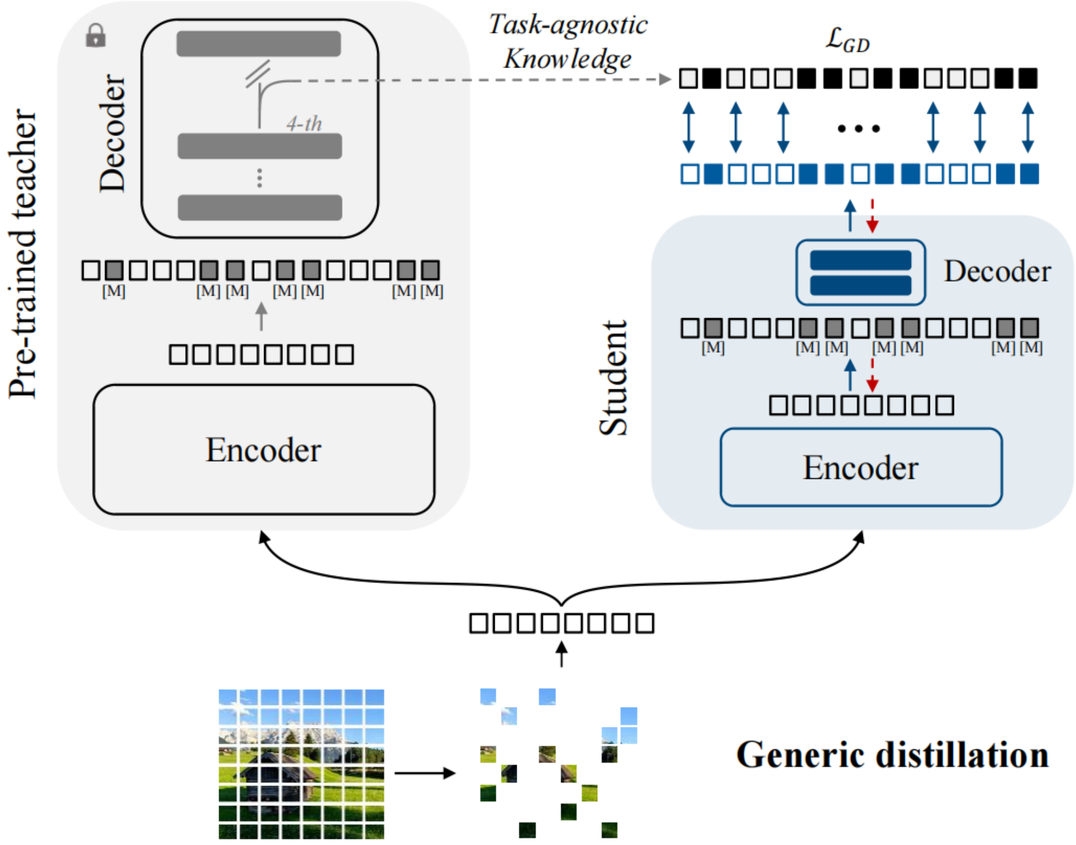


SEED: Self-supervised Distillation For Visual Representation

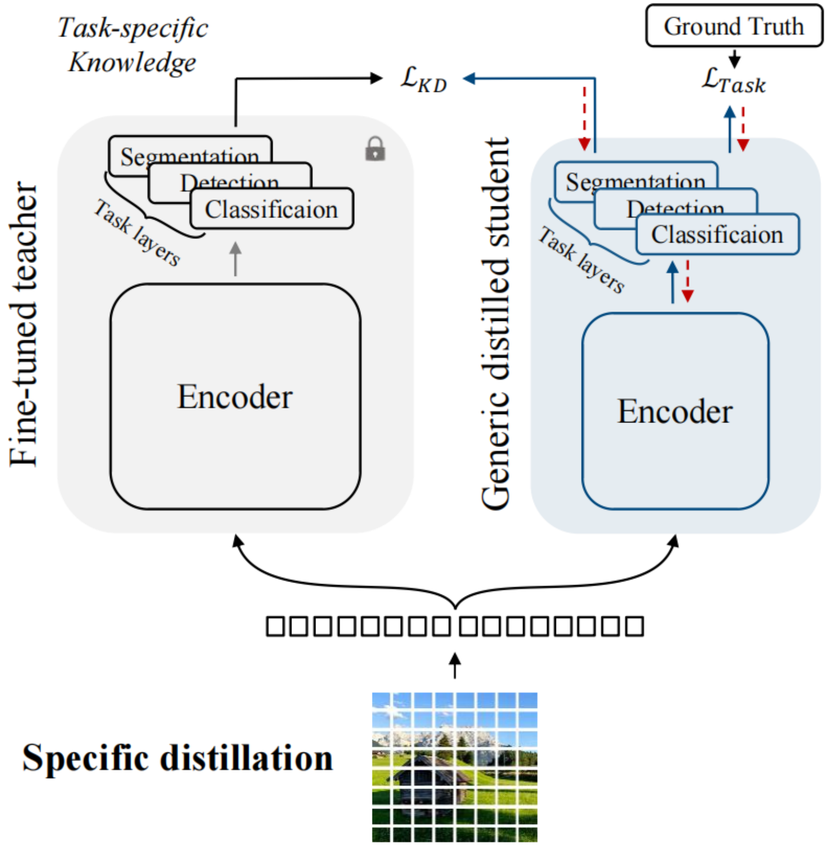
Unsupervised distillation: task-agnostic knowledge

Supervised distillation: task-specific knowledge

Method G2SD: generic-to-specific distillation

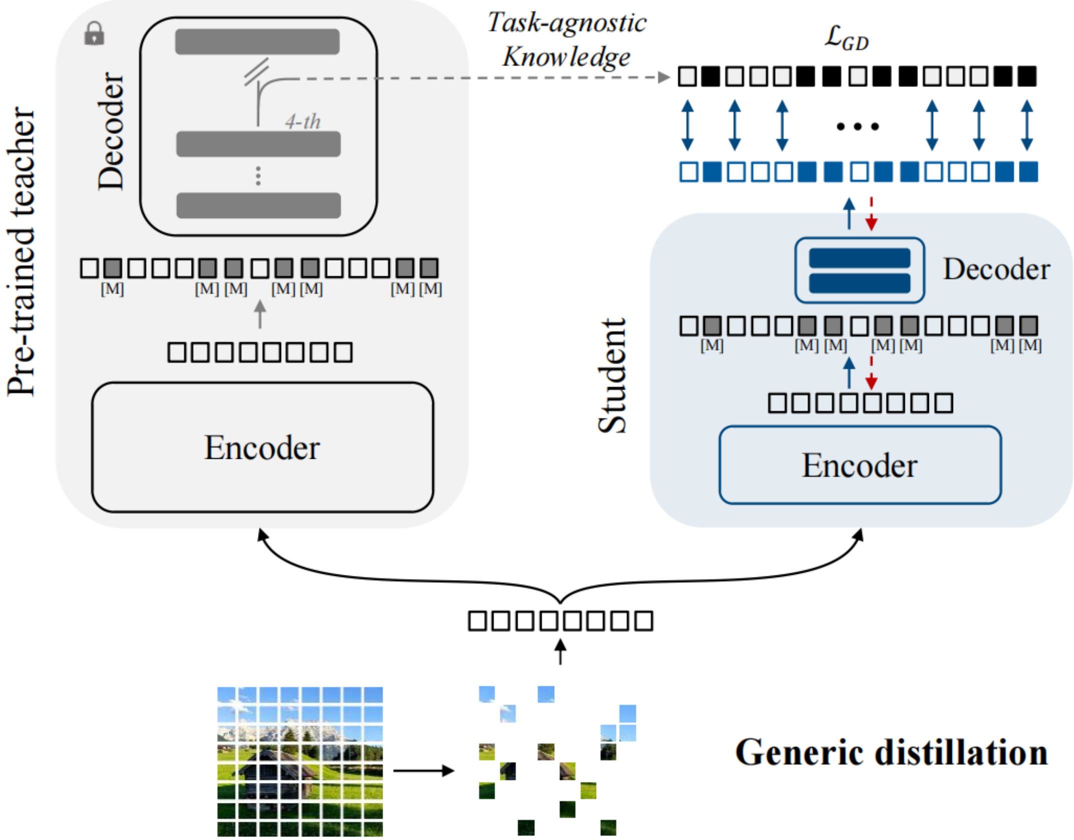


(1) Task-agnostic knowledge Transfer



(2) Task-specific Representation Configuration

Method G2SD: generic-to-specific distillation

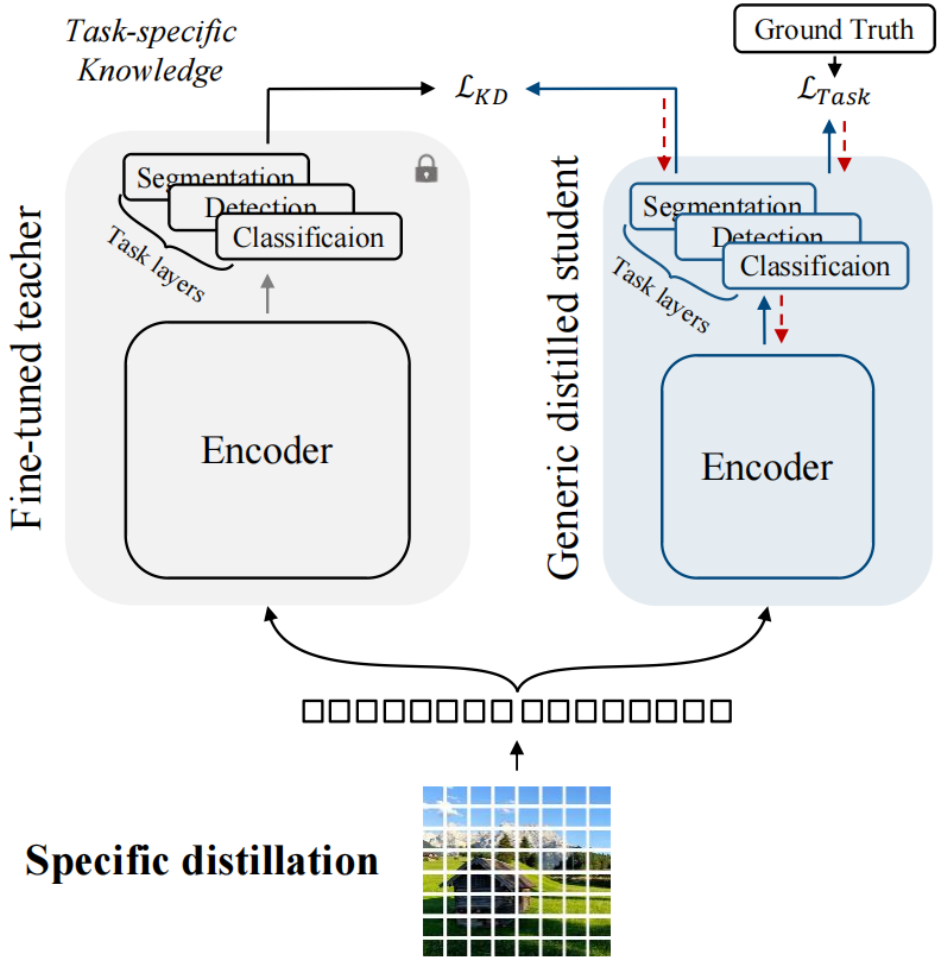


(1) Task-agnostic knowledge Transfer

$$h_i = \boldsymbol{e}_{[\mathrm{M}]} \odot \delta(i \in \mathcal{M}) + \boldsymbol{e}_i \odot (1 - \delta(i \in \mathcal{M})),$$

$$\mathcal{L}_{\mathrm{GD}} = \sum_{i\in\{\mathcal{V}\bigcup\mathcal{M}\}} \text{Smooth-}\ell_1(\mathrm{LN}(\hat{\boldsymbol{z}}_i^t) - \boldsymbol{z}_i^s),$$

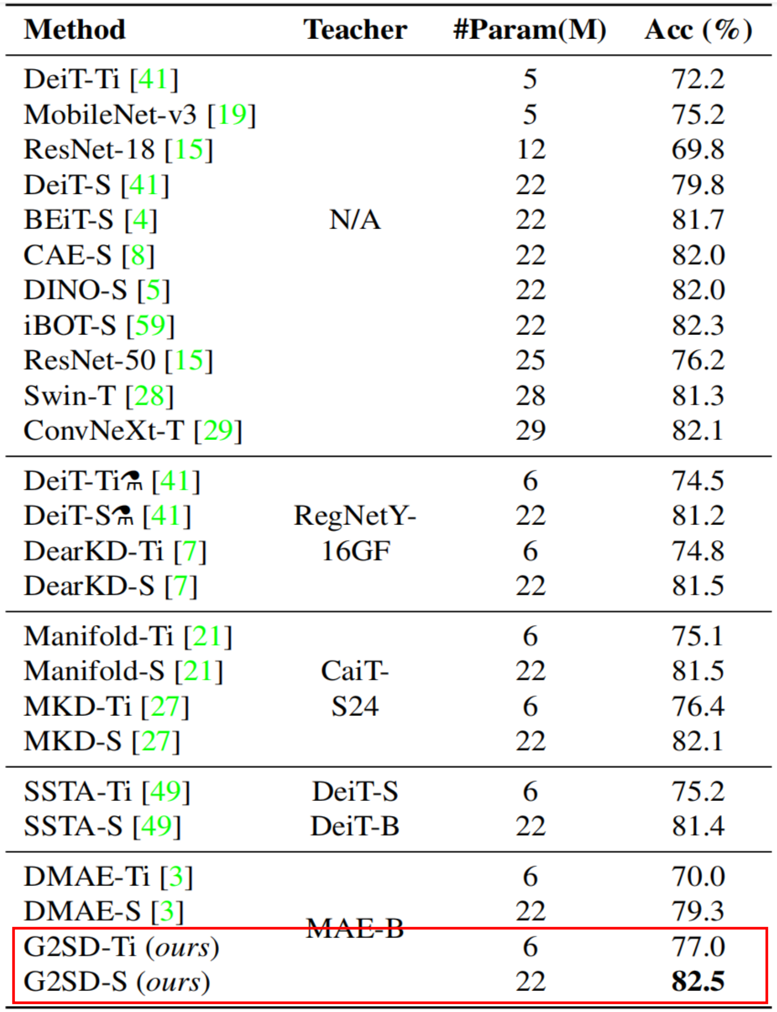
Method G2SD: generic-to-specific distillation



(2) Task-specific Representation Configuration

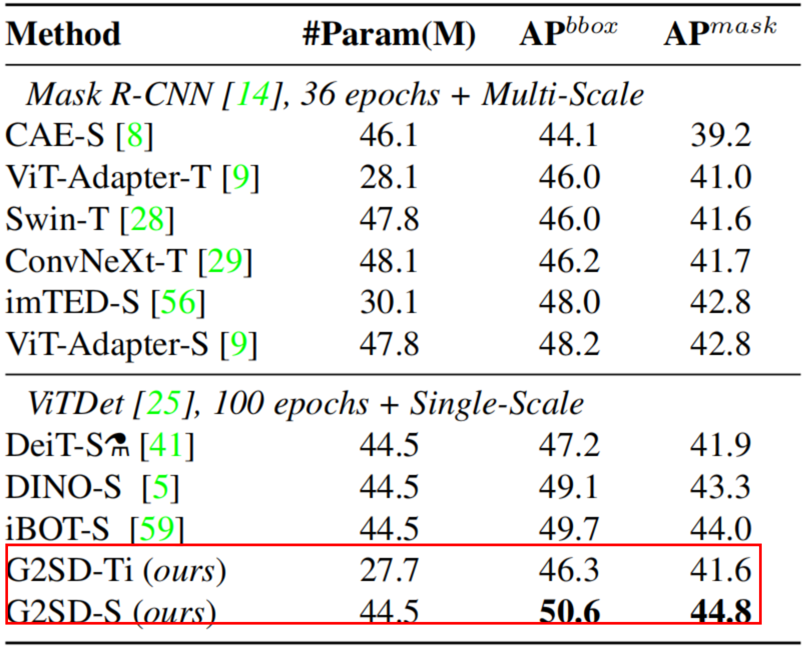
$$\mathcal{L}_{\text{SD}} = \mathcal{L}_{\text{Task}}(f^s(\boldsymbol{x}), Y) + \beta \mathcal{L}_{\text{KD}}(f^s(\boldsymbol{x}), f^t(\boldsymbol{x})),$$

Experiments



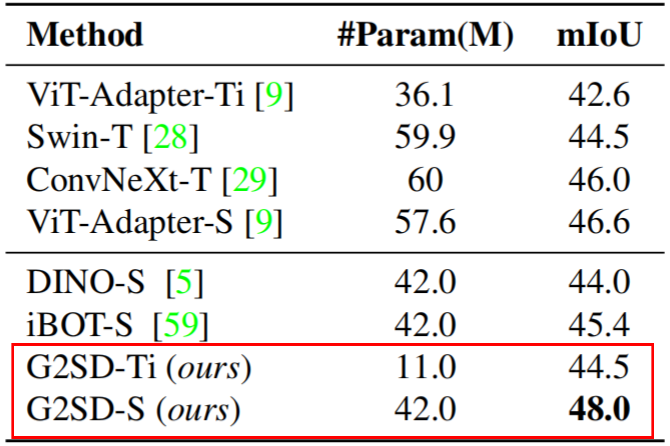
Method	Teacher	#Param(M)	Acc (%)
DeiT-Ti [41]	N/A	5	72.2
MobileNet-v3 [19]		5	75.2
ResNet-18 [15]		12	69.8
DeiT-S [41]		22	79.8
BEiT-S [4]		22	81.7
CAE-S [8]		22	82.0
DINO-S [5]		22	82.0
iBOT-S [59]		22	82.3
ResNet-50 [15]		25	76.2
Swin-T [28]		28	81.3
ConvNeXt-T [29]		29	82.1
DeiT-Ti⚗ [41]	RegNetY-16GF	6	74.5
DeiT-S⚗ [41]		22	81.2
DearKD-Ti [7]		6	74.8
DearKD-S [7]		22	81.5
Manifold-Ti [21]	CaiT-S24	6	75.1
Manifold-S [21]		22	81.5
MKD-Ti [27]		6	76.4
MKD-S [27]		22	82.1
SSTA-Ti [49]	DeiT-S	6	75.2
SSTA-S [49]	DeiT-B	22	81.4
DMAE-Ti [3]	MAE-B	6	70.0
DMAE-S [3]		22	79.3
G2SD-Ti (*ours*)		6	77.0
G2SD-S (*ours*)		22	**82.5**

Classification accuracy on the ImageNet



Method	#Param(M)	AP^{bbox}	AP^{mask}
Mask R-CNN [14], 36 epochs + Multi-Scale			
CAE-S [8]	46.1	44.1	39.2
ViT-Adapter-T [9]	28.1	46.0	41.0
Swin-T [28]	47.8	46.0	41.6
ConvNeXt-T [29]	48.1	46.2	41.7
imTED-S [56]	30.1	48.0	42.8
ViT-Adapter-S [9]	47.8	48.2	42.8
ViTDet [25], 100 epochs + Single-Scale			
DeiT-S⚗ [41]	44.5	47.2	41.9
DINO-S [5]	44.5	49.1	43.3
iBOT-S [59]	44.5	49.7	44.0
G2SD-Ti (*ours*)	27.7	46.3	41.6
G2SD-S (*ours*)	44.5	**50.6**	**44.8**

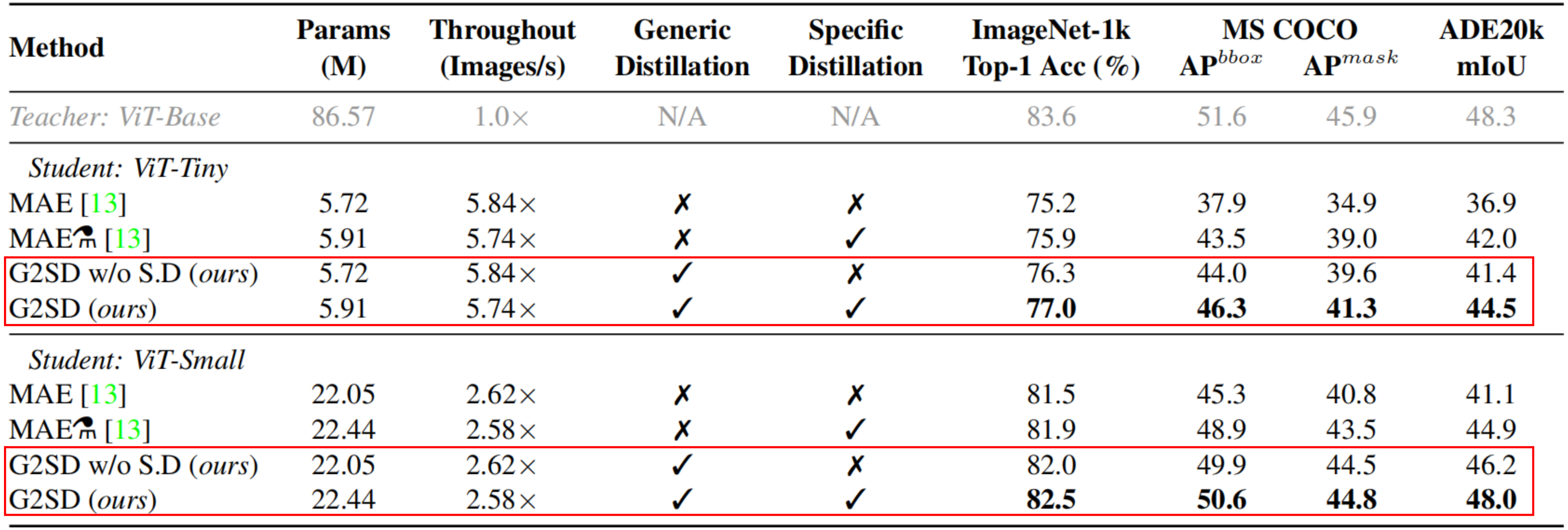
Detection performance on MS COCO



Method	#Param(M)	mIoU
ViT-Adapter-Ti [9]	36.1	42.6
Swin-T [28]	59.9	44.5
ConvNeXt-T [29]	60	46.0
ViT-Adapter-S [9]	57.6	46.6
DINO-S [5]	42.0	44.0
iBOT-S [59]	42.0	45.4
G2SD-Ti (*ours*)	11.0	44.5
G2SD-S (*ours*)	42.0	**48.0**

segmentation performance on ADE20k

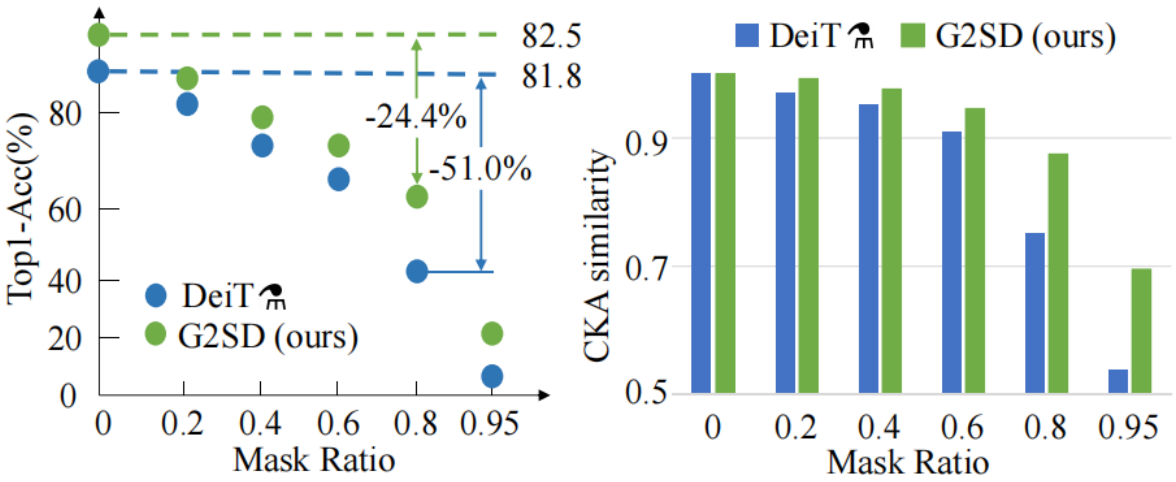
Experiments: Ablation Studies



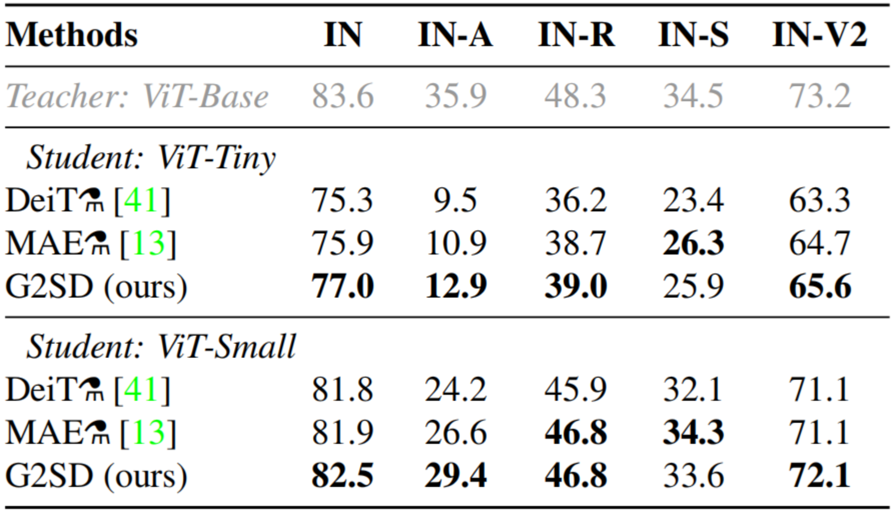
Method	Params	Throughout	Generic	Specific	ImageNet-1k	MS COCO		ADE20k
	(M)	(Images/s)	Distillation	Distillation	Top-1 Acc (%)	AP^{bbox}	AP^{mask}	mIoU
Teacher: ViT-Base	86.57	1.0×	N/A	N/A	83.6	51.6	45.9	48.3
Student: ViT-Tiny								
MAE [13]	5.72	5.84×	✗	✗	75.2	37.9	34.9	36.9
MAE⚗ [13]	5.91	5.74×	✗	✓	75.9	43.5	39.0	42.0
G2SD w/o S.D (*ours*)	5.72	5.84×	✓	✗	76.3	44.0	39.6	41.4
G2SD (*ours*)	5.91	5.74×	✓	✓	**77.0**	**46.3**	**41.3**	**44.5**
Student: ViT-Small								
MAE [13]	22.05	2.62×	✗	✗	81.5	45.3	40.8	41.1
MAE⚗ [13]	22.44	2.58×	✗	✓	81.9	48.9	43.5	44.9
G2SD w/o S.D (*ours*)	22.05	2.62×	✓	✗	82.0	49.9	44.5	46.2
G2SD (*ours*)	22.44	2.58×	✓	✓	**82.5**	**50.6**	**44.8**	**48.0**

Single-stage vs. Two-stage

Experiments: Robustness



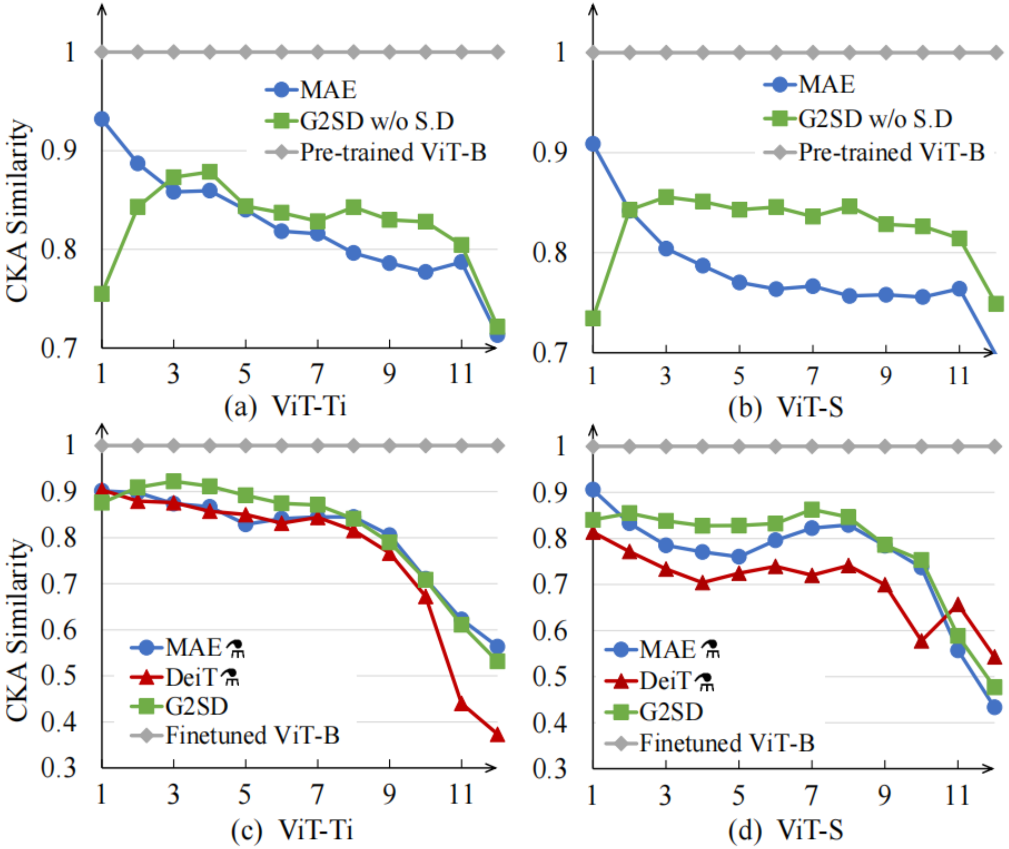
Occlusion Invariance



Methods	IN	IN-A	IN-R	IN-S	IN-V2
Teacher: ViT-Base	83.6	35.9	48.3	34.5	73.2
Student: ViT-Tiny					
DeiT⚗ [41]	75.3	9.5	36.2	23.4	63.3
MAE⚗ [13]	75.9	10.9	38.7	**26.3**	64.7
G2SD (ours)	**77.0**	**12.9**	**39.0**	25.9	**65.6**
Student: ViT-Small					
DeiT⚗ [41]	81.8	24.2	45.9	32.1	71.1
MAE⚗ [13]	81.9	26.6	**46.8**	**34.3**	71.1
G2SD (ours)	**82.5**	**29.4**	**46.8**	33.6	**72.1**

ImageNet variants

Experiments: Analysis



(a) ViT-Ti

(b) ViT-S

(c) ViT-Ti

(d) ViT-S

Representation Similarity with teacher

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thank you for your attention!

Paper

Code

中国科学院大学
University of Chinese Academy of Sciences

PriSDL