Generic-to-Specific Distillation of Masked Autoencoders

Wei Huang[1,*], Zhiliang Peng[1,§,*], Li Dong[2], Furu Wei[2], Jianbin Jiao[1,†], Qixiang Ye[1,†]
University of Chinese Academy of Sciences[1]
Microsoft Research[2]

model	image size	#param.	FLOPs	throughput (image / s)	IN-1K top-1 acc.
ImageNet-1K trained models					
• RegNetY-16G [54]	224^2	84M	16.0G	334.7	82.9
• EffNet-B7 [71]	600^2	66M	37.0G	55.1	84.3
• EffNetV2-L [72]	480^2	120M	53.0G	83.7	85.7
○ DeiT-S [73]	224^2	22M	4.6G	978.5	79.8
○ DeiT-B [73]	224^2	87M	17.6G	302.1	81.8
○ Swin-T	224^2	28M	4.5G	757.9	81.3
• ConvNeXt-T	224^2	29M	4.5G	774.7	**82.1**
○ Swin-S	224^2	50M	8.7G	436.7	83.0
• ConvNeXt-S	224^2	50M	8.7G	447.1	**83.1**
○ Swin-B	224^2	88M	15.4G	286.6	83.5
• ConvNeXt-B	224^2	89M	15.4G	292.1	**83.8**
○ Swin-B	384^2	88M	47.1G	85.1	84.5
• ConvNeXt-B	384^2	89M	45.0G	95.7	**85.1**
• ConvNeXt-L	224^2	198M	34.4G	146.8	**84.3**
• ConvNeXt-L	384^2	198M	101.0G	50.4	**85.5**

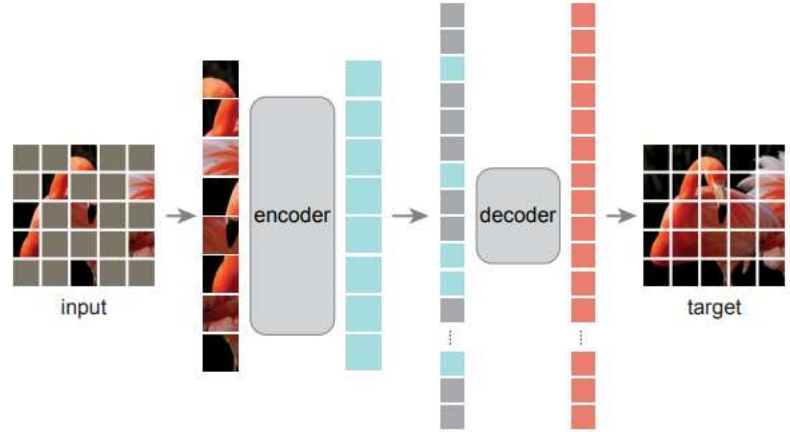


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.

encoder	dec. depth	ft acc	hours	speedup
ViT-L, w/ [M]	8	84.2	42.4	-
ViT-L	8	84.9	15.4	2.8×
ViT-L	1	84.8	11.6	**3.7**×
ViT-H, w/ [M]	8	-	119.6†	-
ViT-H	8	85.8	34.5	3.5×
ViT-H	1	85.9	29.3	**4.1**×

Table 2. **Wall-clock time** of our MAE training (800 epochs), benchmarked in 128 TPU-v3 cores with TensorFlow. The speedup is relative to the entry whose encoder has mask tokens (gray). The decoder width is 512, and the mask ratio is 75%. †: This entry is estimated by training ten epochs.

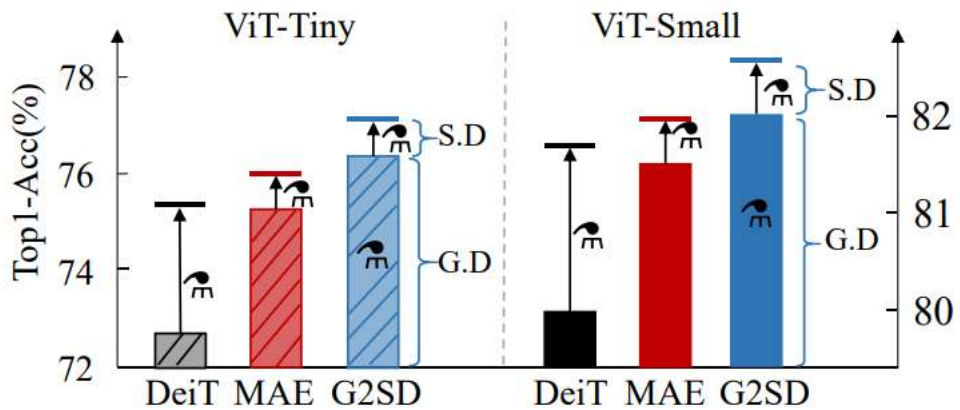
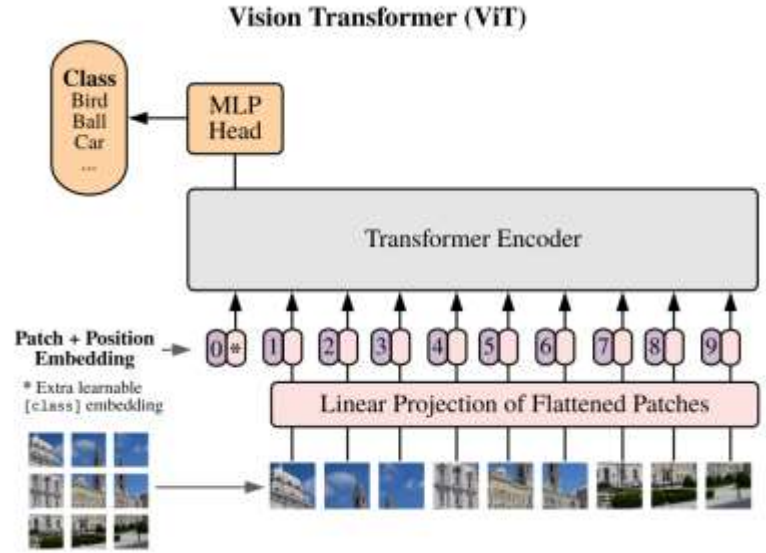


Figure 1. Comparison of single-stage distillation models (from scratch [41] and pre-trained by the self-supervised method MAE [13]) with the two-stage distillation counterparts (G2SD) using the same teacher model. G.D and S.D respectively denote generic and specific distillation. ⚗ is the symbol of distillation.

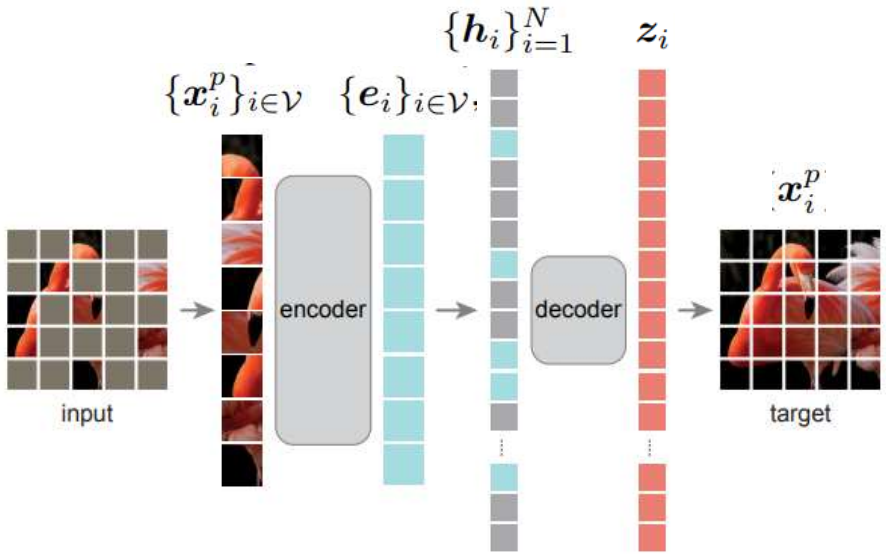
- We propose general-to-specific distillation (G2SD) to transfer task-agnostic and task-specific knowledge from masked autoencoders to lightweight ViTs, setting a solid baseline for two-stage vision model distillation.

- We design a simple-yet-effective generic distillation strategy by aligning the student's predictions with hidden features of the pre-trained masked autoencoder at visible and masked patches.

- Experiments show that the lightweight student model with G2SD achieves competitive results across vision tasks, improving the performance of lightweight ViT models to a new height.



将一张224 × 224 ×3大小的图像重构为一张14 × 14网格的图像小块，每个小块大小为16×16×3。



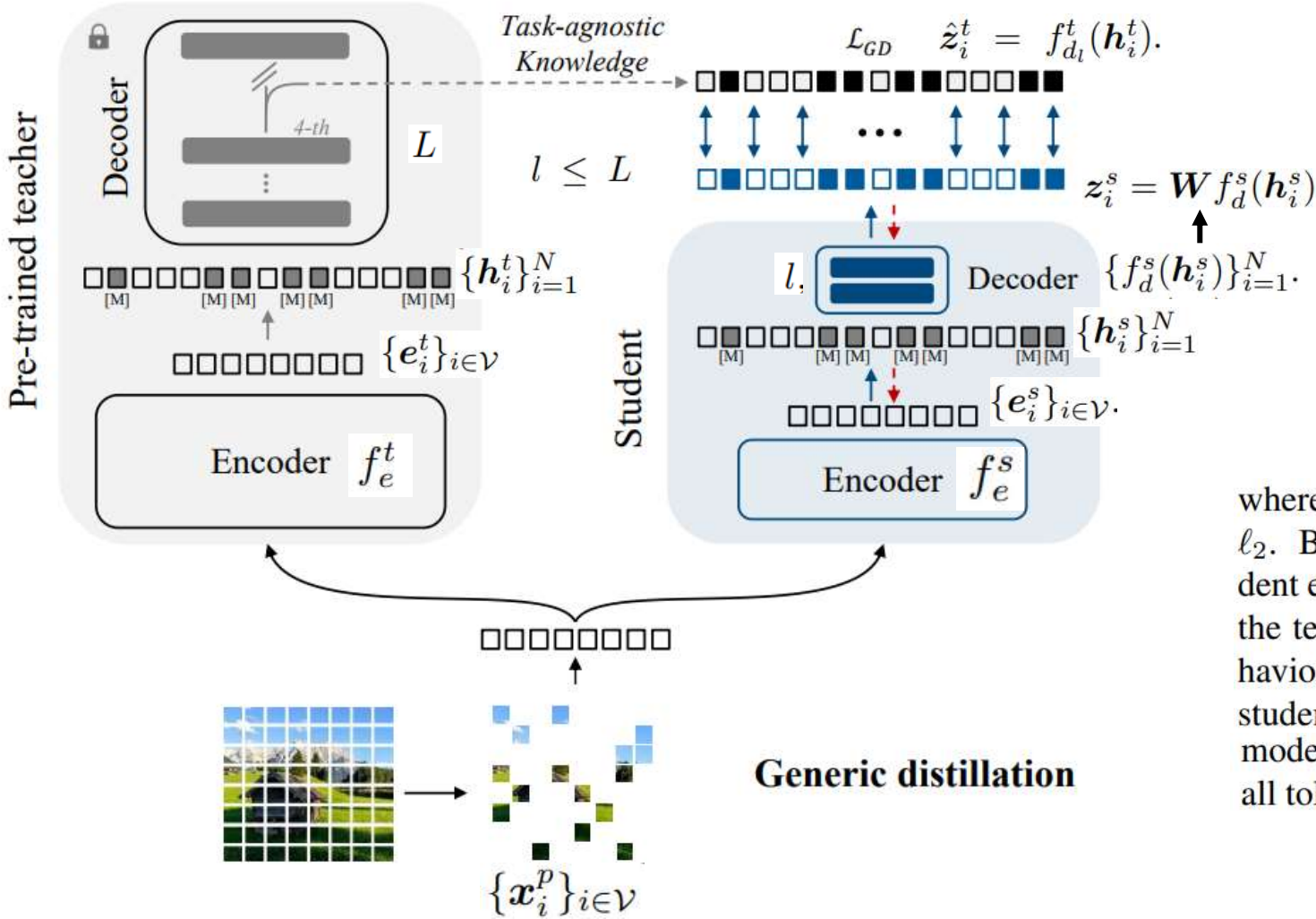
Visible token $\{x_i^p\}_{i\in\mathcal{V}}$

masked token $\{x_i^p\}_{i\in\mathcal{M}}$

$$h_i = e_{[\mathrm{M}]} \odot \delta(i \in \mathcal{M}) + e_i \odot (1 - \delta(i \in \mathcal{M})),$$

$$\mathcal{L}_{\mathrm{MAE}} = \sum_{i\in\mathcal{M}} ||\mathrm{LN}(x_i^p) - z_i||_2,$$

Generic Distillation: Task-agnostic knowledge Transfer

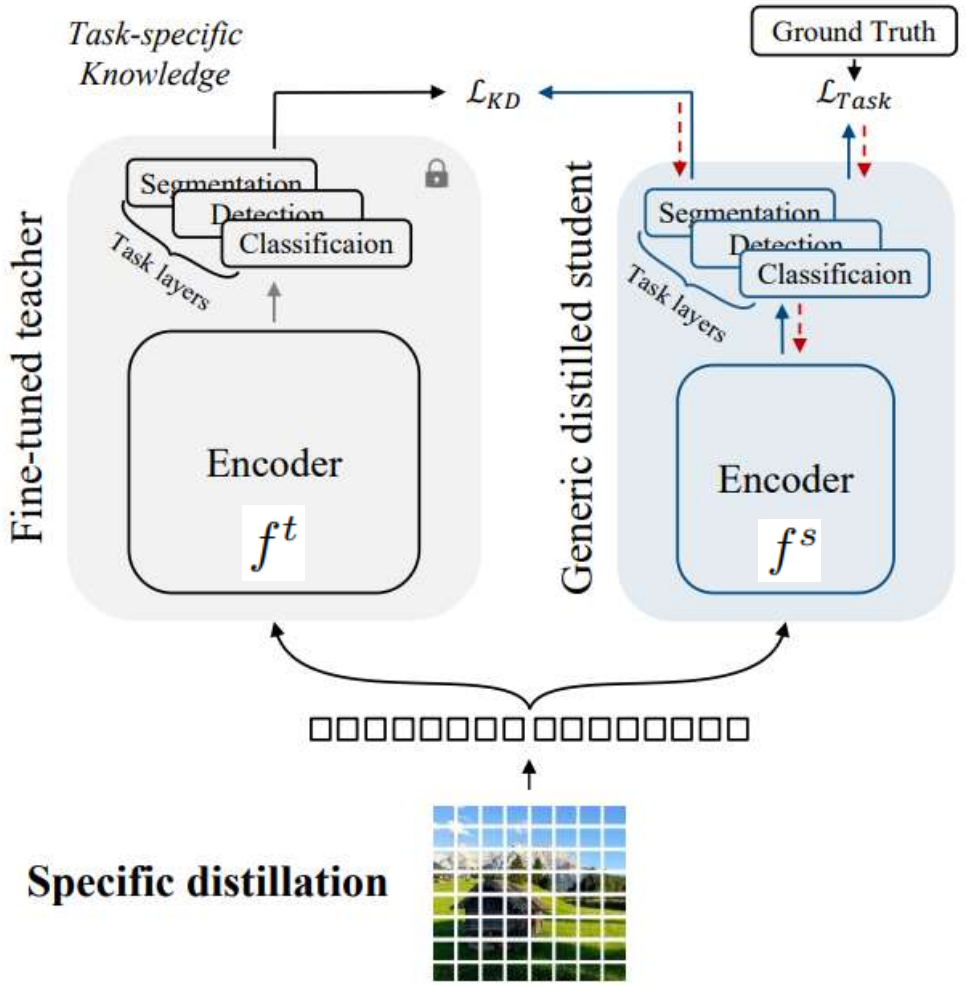


$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & , \ |x| \leq 1 \\ |x| - 0.5 & , \ |x| > 1 \end{cases}$$

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

where Smooth-$\ell_1(\cdot)$ is a trade-off function between ℓ_1 and ℓ_2. By minimizing $\mathcal{L}_{\text{GD}}$ on the visible tokens $\mathcal{V}$, the student encoder is optimized to extract features in the way like the teacher encoder, *i.e.*, mimicking feature extraction behavior. By minimizing $\mathcal{L}_{\text{GD}}$ on the masked tokens $\mathcal{M}$, the student encoder and decoder are optimized to learn context modeling ability from teacher models. Optimizing $\mathcal{L}_{\text{GD}}$ on all tokens transfers task-agnostic knowledge.

Specific Distillation: Task-specific Representation Configuration



$\mathcal{L}_{\text{Task}}$ as the task loss function

$\mathcal{L}_{\text{KD}}$ as the task-specific distillation loss function

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

where Y is the ground truth and β is the regularization factor (Refer to Appendix A for details).

Analysis

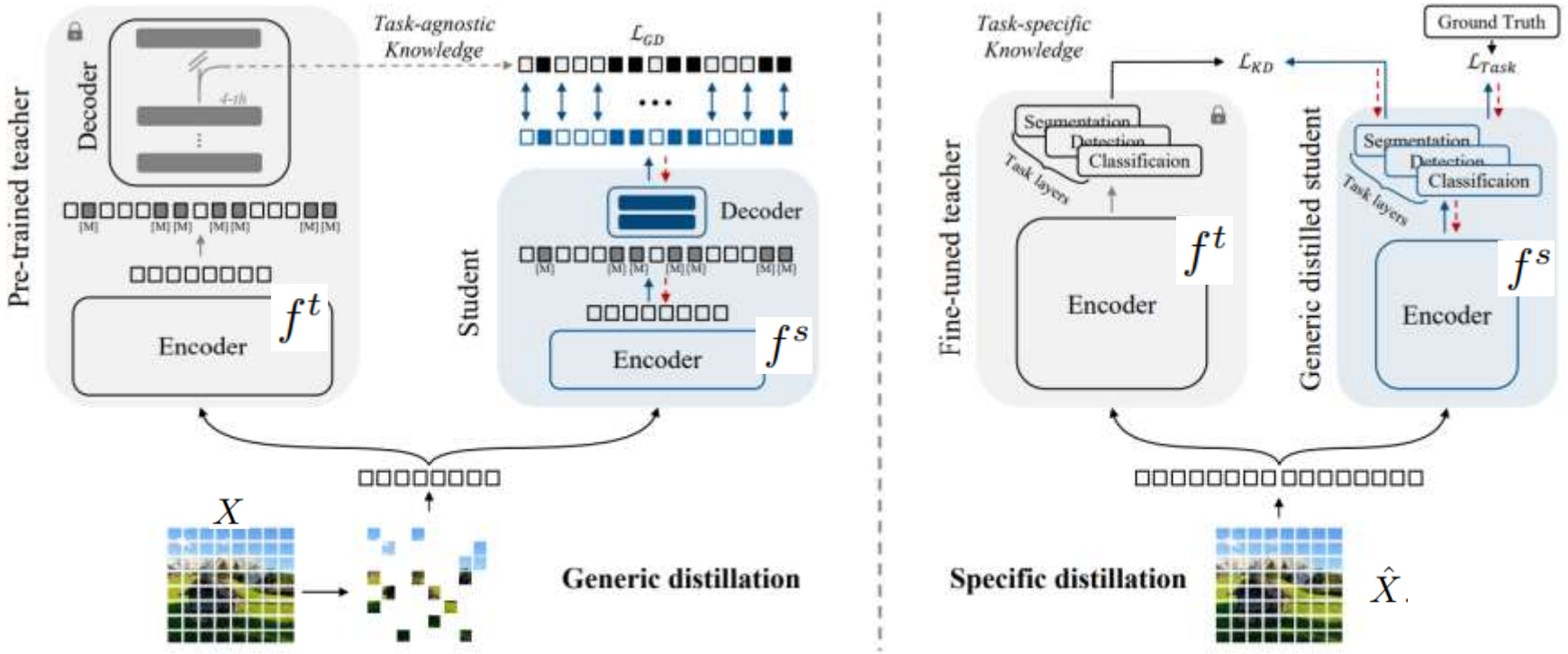


Figure 2. Diagram of the proposed generic-to-specific distillation (G2SD). [M] denotes mask token. In the generic distillation stage (*left*), masked images are converted to patches and fed to both the teacher and student encoders for feature extraction. Feature predictions of the student decoder are aligned with those of the teacher at both visible and predicted patches. In the specific distillation stage (*right*), student models are trained to have consistent predictions with teacher models fine-tuned on the specific task.

The single-stage task-specified distillation is interpreted as

$$\arg\max_{\theta^s} \mathcal{I}_{\theta^s,\theta^t}(f^t, f^s|\hat{X}),$$

The proposed G2SD is interpreted as

$$\arg\max_{\theta^s} \mathcal{I}_{\theta^s,\theta^t}(f^t, f^s|X) + \mathcal{I}_{\theta^s,\theta^t}(f^t, f^s|\hat{X}) - \mathcal{I}_{\theta^s,\theta^t}(f^t, f^s|(X, \hat{X})),$$

Obviously, the mutual information defined by G2SD is larger than that by single-stage, which implies more information can be transferred by our G2SD approach.

Experiments

Datasets. The generic distillation is conducted on ImageNet-1k [39] training set with 1.2M images. Following self-supervised recipes [13], we do not use the label information, so that lightweight models focus on soaking up the task-agnostic representations. In specific distillation, the models are fine-tuned from the previous stage on ImageNet-1k [39], MS COCO [26] and ADE20K [58] datasets.

For image classification, we take a fine-tuned ViT-base model as the teacher, which is officially released by MAE [13] and achieves 83.6% top-1 accuracy. Following DeiT [41] distillation recipe, we append a distillation token to the student model for token-based distillation and use the hard decision of the teacher as the distillation label. The student model is trained for 200 epochs.
achieves 82.5% top-1 accuracy, which outperforms CNN-based ConvNeXt by 0.4%, by using fewer parameters (22M *vs*. 29M). G2SD consistently outperforms self-supervised methods, BEiT and CAE, by 0.8% and 0.5%, respectively. Compared with those distillation methods, G2SD shows the superiority. Remarkably, with the limited parameters (∼6M), G2SD reports a substantial gain compared to DeiT-Ti⚗ and carefully designed MobileNet-v3.

Table 1. Top-1 accuracy on ImageNet-1k.

Method	**Teacher**	**#Param(M)**	**Acc (%)**
DeiT-Ti [41]		5	72.2
MobileNet-v3 [19]		5	75.2
ResNet-18 [15]		12	69.8
DeiT-S [41]		22	79.8
BEiT-S [4]	N/A	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]		6	74.5
DeiT-S⚗ [41]	RegNetY-	22	81.2
DearKD-Ti [7]	16GF	6	74.8
DearKD-S [7]		22	81.5
Manifold-Ti [21]		6	75.1
Manifold-S [21]	CaiT-	22	81.5
MKD-Ti [27]	S24	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]		6	70.0
DMAE-S [3]	MAE-B	22	79.3
G2SD-Ti (*ours*)		6	77.0
G2SD-S (*ours*)		22	**82.5**

For object detection and instance segmentation tasks, we follow the ViTDet [25] framework, where the official ViTDet-Base [25] model are used as the teacher. The Feature-Richness Score method [12] is adopted to stress important features that are distilled from the teacher to the student model. Student models are trained with batch size 64 for 100 epochs. The input image resolution is 1024×1024.

Table 2. Object detection and instance segmentation results on the MS COCO dataset.

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**

For semantic segmentation, we use UperNet [51] task layers and distill the model for 160K iterations. Due to the absence of officially released model weights, we fine-tune the MAE pre-trained ViT-Base model on ADE20k by using the BEiT [4] semantic segmentation codebase to get teacher model, which achieves 48.3 mIoU, is comparable to MAE official report. During specific distillation, besides the supervision from the ground-truth, activation maps from the student and the teacher are aligned $w.r.t.$ the channel dimension [40].

Table 3. ADE20K validation results using UperNet [51]. The input image resolution is 512×512.

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**

Ablation studies

Table 4. Ablation study on single-stage and two-stage distillation methods, where G2SD w/o S.D denotes **only** performing generic distillation (*i.e.*, without specific distillation) and MAE⚗ means performing task-specific distillation during fine-tuning phase of MAE [13].

Method	Params (M)	Throughout (Images/s)	Generic Distillation	Specific Distillation	ImageNet-1k Top-1 Acc (%)	MS COCO AP^{bbox}	MS COCO AP^{mask}	ADE20k 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**

Table 5. Ablation study on generic distillation targets. $\boldsymbol{e}_i^t$, $\hat{\boldsymbol{z}}_i^t$ and $\mathcal{R}(\boldsymbol{e}_i^t)$ respectively denote teacher encoder features, teacher decoder features, the relation among teacher encoder features. #5 is the default setting.

Target	$\boldsymbol{e}_i^t$ $i \in \mathcal{V}$	$\mathcal{R}(\boldsymbol{e}_i^t)$ $i \in \mathcal{V}$	$\hat{\boldsymbol{z}}_i^t$ $i \in \mathcal{V}$	$\hat{\boldsymbol{z}}_i^t$ $i \in \mathcal{M}$	Accuracy (%)	mIoU (%)
#1	✓				81.60	43.69
#2		✓			81.45	43.64
#3				✓	81.96	45.20
#4	✓			✓	81.85	44.12
#5			✓	✓	**81.99**	**46.19**

Table 6. Ablation on the mask ratio (*top*) and target layer of the teacher model used for distillation (*bottom*).

Mask ratio	0.05	0.25	0.55	0.75	0.9
Top-1 Acc(%)	81.7	81.7	81.6	**82.0**	81.8

Layer Index	1	2	4	6	8
Top-1 Acc(%)	81.6	81.8	**82.0**	81.8	81.7

Table 7. Ablation study on the width and depth (D) of the student decoder. The depth and width of the teacher's decoder are 8 and 512, respectively.

Width	D	Acc(%)	D	Acc(%)	D	Acc(%)
128		81.9		81.8		81.7
256	2	81.7	4	**82.0**	8	81.7
512		81.8		81.7		80.3

Width: embedding dimension
Depth: number of Transformer blocks

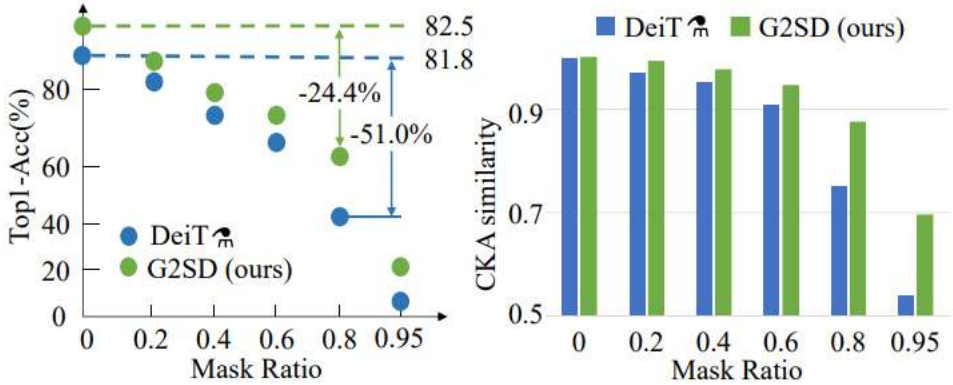


Figure 3. Performance degradation (*left*) and CKA similarity (*right*) between the representations generated by the complete image and the corrupted image with various mask ratios.

Table 8. Robustness evaluation. "IN" is short for ImageNet.

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**

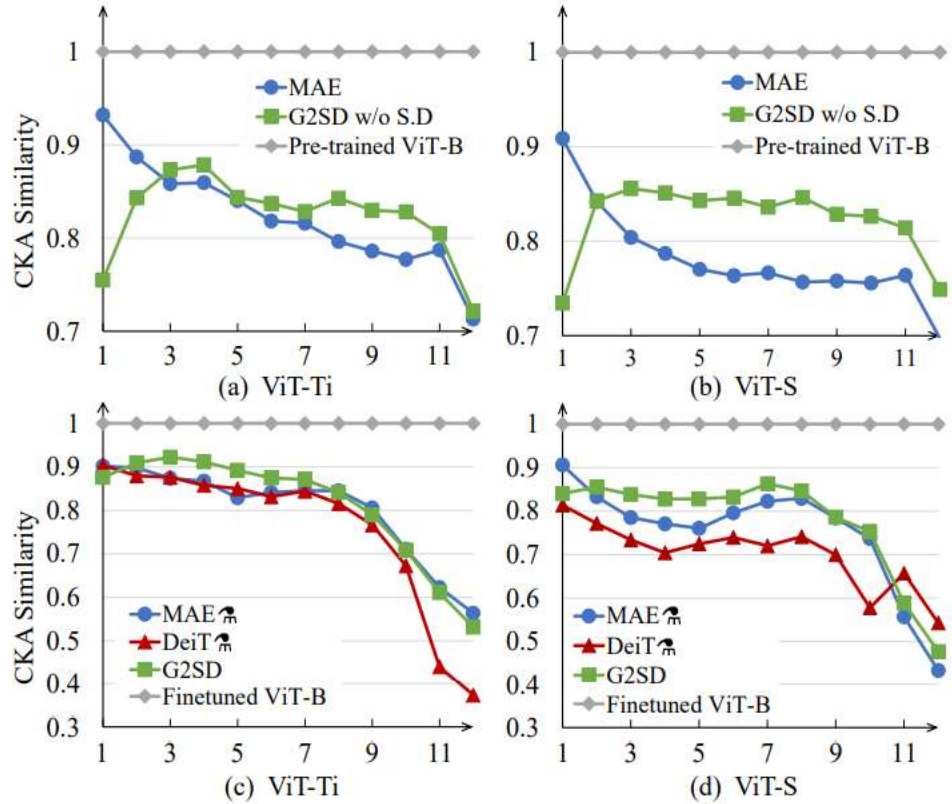


Figure 4. CKA similarity between representations generated by pre-trained MAE-B with: (a) pre-trained MAE-Ti and G2SD-Ti w/o S.D, and (b) MAE-S and G2SD-S w/o S.D. CKA similarity between representations generated by fine-tuned MAE-B with: (c) fine-tuned MAE-Ti⚗, DeiT-Ti⚗, and G2SD-Ti, and (d) fine-tuned MAE-S⚗, DeiT-S⚗, and G2SD-S. *x-axis* denotes network depth.

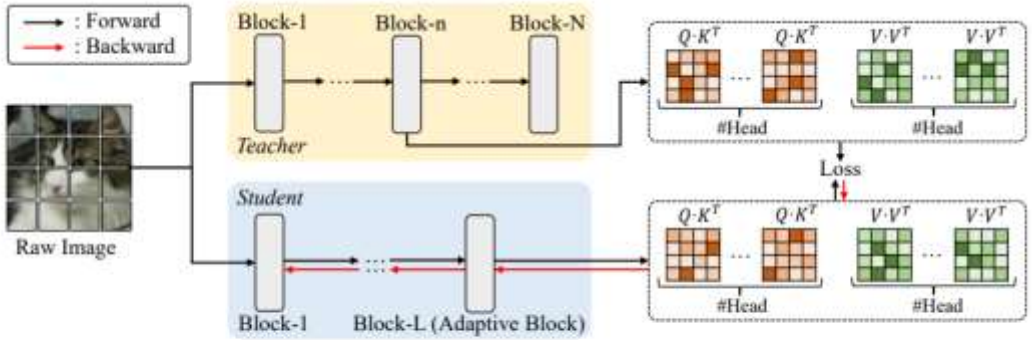


Figure 3. The default knowledge distillation strategy of TinyMIM. The student (*e.g.* ViT-B) is optimized to mimic the relations generated by the intermediate block of a MIM pre-trained teacher (*e.g.* ViT-L) with raw image as input. We replace the last block of the student with an adaptive block to match teacher's head number (no extra computational cost). After pre-training (knowledge distillation), the student model can be transferred to various downstream tasks.

Method	Pretraining Epochs	Tokenizer/ Teacher	Tokenizer/Teacher Data	Classification Top-1 Acc (%)	Segmentation mIoU
Tiny-size models (ViT-T/16)					
Scratch [44]	300	Label	IN1K	72.2	38.0
MAE† [18]	1600	Pixel	IN1K	71.6	37.6
MoCo [5]	1600	EMA	IN1K	73.3	39.3
TinyMIM (Ours)	300	TinyMIM-ViT-S	IN1K	**75.8**	**44.0/44.6**‡
TinyMIM* (Ours)	300	TinyMIM-ViT-S	IN1K	**79.6**	**45.0**‡
Small-size models (ViT-S/16)					
Scratch [44]	300	Label	IN1K	79.9	43.1
MAE† [18]	1600	Pixel	IN1K	80.6	42.8
MoCo [5]	1600	EMA	IN1K	81.4	43.9
DINO [3]	1600	EMA	IN1K	81.5	45.3
CIM [13]	1600	Pixel	IN1K	81.6	-
TinyMIM (Ours)	300	TinyMIM-ViT-B	IN1K	**83.0**	**48.4/48.9**‡
Base-size models (ViT-B/16)					
Scratch [44]	300	Label	IN1K	81.2	47.2
BeiT [2]	800	DALL-E	DALLE250M+IN22K+IN1K	83.2	45.6
MAE [18]	1600	Pixel	IN1K	83.6	48.1
SIM [43]	1600	EMA	IN1K	83.8	-
CAE [4]	1600	DALL-E	DALLE250M+IN22K+IN1K	83.9	50.2
MaskFeat [48]	1600	HOG	IN1K	84.0	-
SdAE [7]	300	EMA	IN1K	84.1	48.6
data2vec [1]	800	EMA	IN1K	84.2	-
PeCo [11]	300	VQGAN	IN1K	84.1	46.7
PeCo [11]	800	VQGAN	IN1K	84.5	48.5
TinyMIM (Ours)	300	MAE-ViT-L	IN1K	**85.0**	**52.2/52.6**‡

Table 3. Fine-tuning results on ImageNet-1K and ADE20K. All models are pre-trained on ImageNet-1K. "Tokenizer/Teacher Data": training data of teacher and tokenizer. †: reproduced result using official code. *: the model is fine-tuned for 1000 epochs with DeiT-style [44] knowledge distillation. ‡: the model adopts an intermediate fine-tuning on ImageNet-1K classification before ADE20K segmentation fine-tuning.

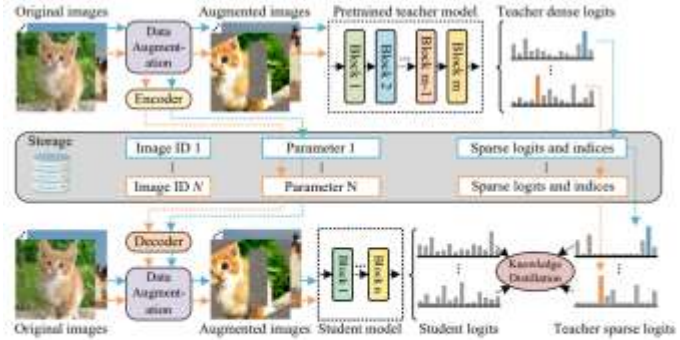


Fig. 2: Our fast pretraining distillation framework. **Top:** the branch for saving teacher logits. Encoded data augmentation and sparsified teacher logits are saved. **Middle:** the disk for storing information. **Bottom:** the branch for training the student. The decoder reconstructs the data augmentation, and distillation is conducted between the teacher logits and student outputs. Note that the two branches are independent and asynchronous.

Table 4: TinyViT performance on IN-1k [18] with comparisons to state-of-the-art models. MACs (multiply–accumulate operations) and Throughput are measured using the GitHub repository of [1,24] and a V100 GPU. ⚗: pretrain on IN-21k with the proposed fast distillation; ↑: finetune with higher resolution.

	Model	Top-1 (%)	Top-5 (%)	#Params (M)	MACs (G)	Throughput (images/s)	Input	Arch.
5-10M #Params	MoblieViT-S [46]	78.4	-	6	1.8	2,661	256	Hybrid
	ViTAS-DeiT-A [60]	75.5	92.4	6	1.3	3,504	224	Trans
	GLiT-Tiny [9]	76.3	-	7	1.5	3,262	224	Trans
	Mobile-Former-214M [14]	76.7	-	9	0.2	3,105	224	Hybrid
	CrossViT-9 [10]	77.1	-	9	2.0	2,659	224	Trans
	TinyViT-5M (ours)	**79.1**	94.8	5.4	1.3	3,060	224	Hybrid
	TinyViT-5M⚗ (ours)	**80.7**	95.6	5.4	1.3	3,060	224	Hybrid
11-20M	ResNet-18 [28]	70.3	86.7	12	1.8	8,714	224	CNN
	PVT-Tiny [66]	75.1	-	13	1.9	2,791	224	Trans
	ResT-Small [81]	79.6	94.9	14	2.1	2,037	224	Trans
	LeViT-256 [24]	81.6	-	19	1.1	7,386	224	Hybrid
	CoaT-Lite Small [71]	81.9	95.6	20	4.0	1,138	224	Trans
	TinyViT-11M (ours)	**81.5**	95.8	11	2.0	2,468	224	Hybrid
	TinyViT-11M⚗ (ours)	**83.2**	96.5	11	2.0	2,468	224	Hybrid
>20M	DeiT-S [64]	79.9	95.0	22	4.6	2,276	224	Trans
	T2T-ViT-14 [74]	81.5	95.7	21	4.8	1,557	224	Trans
	AutoFormer-S [11]	81.7	95.7	23	5.1	1,341	224	Trans
	Swin-T [43]	81.2	95.5	28	4.5	1,393	224	Trans
	CrossViT-15 [10]	82.3	-	28	6.1	1,306	224	Trans
	EffNet-B5 [62]	83.6	96.7	30	9.9	330	456	CNN
	TinyViT-21M (ours)	**83.1**	96.5	21	4.3	1,571	224	Hybrid
	TinyViT-21M⚗ (ours)	**84.8**	97.3	21	4.3	1,571	224	Hybrid
	TinyViT-21M⚗ ↑384 (ours)	**86.2**	97.8	21	13.8	394	384	Hybrid
	TinyViT-21M⚗ ↑512 (ours)	**86.5**	97.9	21	27.0	167	512	Hybrid

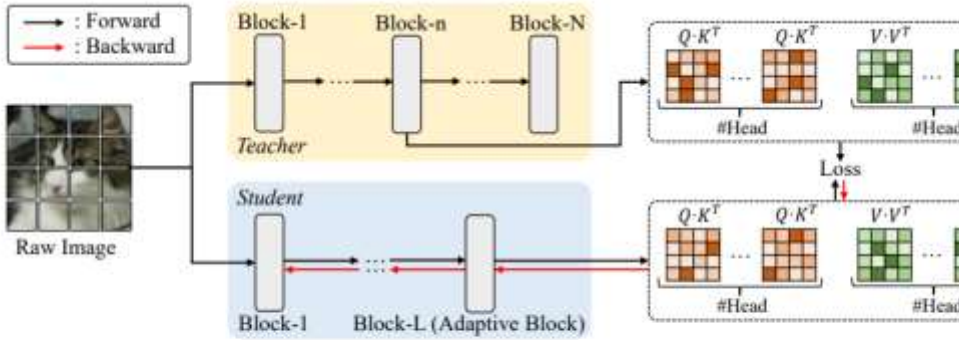


Figure 3. The default knowledge distillation strategy of TinyMIM. The student (*e.g.* ViT-B) is optimized to mimic the relatio the intermediate block of a MIM pre-trained teacher (*e.g.* ViT-L) with raw image as input. We replace the last block of the adaptive block to match teacher's head number (no extra computational cost). After pre-training (knowledge distillation), th can be transferred to various downstream tasks.

Method	Pretraining Epochs	Tokenizer/ Teacher	Tokenizer/Teacher Data	Classification Top-1 Acc (%)	
Tiny-size models (ViT-T/16)					
Scratch [44]	300	Label	IN1K	72.2	
MAE† [18]	1600	Pixel	IN1K	71.6	
MoCo [5]	1600	EMA	IN1K	73.3	
TinyMIM (Ours)	300	TinyMIM-ViT-S	IN1K	**75.8**	**44.0/44.6**‡
TinyMIM* (Ours)	300	TinyMIM-ViT-S	IN1K	**79.6**	**45.0**‡
Small-size models (ViT-S/16)					
Scratch [44]	300	Label	IN1K	79.9	43.1
MAE† [18]	1600	Pixel	IN1K	80.6	42.8
MoCo [5]	1600	EMA	IN1K	81.4	43.9
DINO [3]	1600	EMA	IN1K	81.5	45.3
CIM [13]	1600	Pixel	IN1K	81.6	-
TinyMIM (Ours)	300	TinyMIM-ViT-B	IN1K	**83.0**	**48.4/48.9**‡
Base-size models (ViT-B/16)					
Scratch [44]	300	Label	IN1K	81.2	47.2
BeiT [2]	800	DALL-E	DALLE250M+IN22K+IN1K	83.2	45.6
MAE [18]	1600	Pixel	IN1K	83.6	48.1
SIM [43]	1600	EMA	IN1K	83.8	-
CAE [4]	1600	DALL-E	DALLE250M+IN22K+IN1K	83.9	50.2
MaskFeat [48]	1600	HOG	IN1K	84.0	-
SdAE [7]	300	EMA	IN1K	84.1	48.6
data2vec [1]	800	EMA	IN1K	84.2	-
PeCo [11]	300	VQGAN	IN1K	84.1	46.7
PeCo [11]	800	VQGAN	IN1K	84.5	48.5
TinyMIM (Ours)	300	MAE-ViT-L	IN1K	**85.0**	**52.2/52.6**‡

Table 3. Fine-tuning results on ImageNet-1K and ADE20K. All models are pre-trained on ImageNet-1K. "Tokenizer/Teacher Data": training data of teacher and tokenizer. †: reproduced result using official code. *: the model is fine-tuned for 1000 epochs with DeiT-style [44] knowledge distillation. ‡: the model adopts an intermediate fine-tuning on ImageNet-1K classification before ADE20K segmentation fine-tuning.

Contraction factors. We consider the following factors to form a model:

- $\gamma_{D_{1-4}}$: embeded dimension of four stages respectively. Decreasing them results in a thinner network with fewer heads in multi-head self-attention.
- $\gamma_{N_{1-4}}$: the number of blocks in four stages respectively. The depth of the model is decreased by reducing these values.
- $\gamma_{W_{2-4}}$: window size in the last three stages respectively. As these values become smaller, the model has fewer parameters and higher throughput.
- γ_R: channel expansion ratio of the MBConv block. We can obtain a smaller model size by reducing this factor.
- γ_M: expansion ratio of MLP for all transformer blocks. The hidden dimension of MLP will be smaller if scaling down this value.
- γ_E: the dimension of each head in multi-head attention. The number of heads will be increased when scaling it down, bringing lower computation cost.

	Model	Top-1 (%)	Top-5 (%)	#Params (M)	MACs (G)	Throughput (images/s)	Input	Arch.
5-10M #Params	MoblieViT-S [46]	78.4	-	6	1.8	2,661	256	Hybrid
	ViTAS-DeiT-A [60]	75.5	92.4	6	1.3	3,504	224	Trans
	GLiT-Tiny [9]	76.3	-	7	1.5	3,262	224	Trans
	Mobile-Former-214M [14]	76.7	-	9	0.2	3,105	224	Hybrid
	CrossViT-9 [10]	77.1	-	9	2.0	2,659	224	Trans
	TinyViT-5M (ours)	**79.1**	94.8	5.4	1.3	3,060	224	Hybrid
	TinyViT-5M⚗ (ours)	**80.7**	95.6	5.4	1.3	3,060	224	Hybrid
11-20M	ResNet-18 [28]	70.3	86.7	12	1.8	8,714	224	CNN
	PVT-Tiny [66]	75.1	-	13	1.9	2,791	224	Trans
	ResT-Small [81]	79.6	94.9	14	2.1	2,037	224	Trans
	LeViT-256 [24]	81.6	-	19	1.1	7,386	224	Hybrid
	CoaT-Lite Small [71]	81.9	95.6	20	4.0	1,138	224	Trans
	TinyViT-11M (ours)	**81.5**	95.8	11	2.0	2,468	224	Hybrid
	TinyViT-11M⚗ (ours)	**83.2**	96.5	11	2.0	2,468	224	Hybrid
>20M	DeiT-S [64]	79.9	95.0	22	4.6	2,276	224	Trans
	T2T-ViT-14 [74]	81.5	95.7	21	4.8	1,557	224	Trans
	AutoFormer-S [11]	81.7	95.7	23	5.1	1,341	224	Trans
	Swin-T [43]	81.2	95.5	28	4.5	1,393	224	Trans
	CrossViT-15 [10]	82.3	-	28	6.1	1,306	224	Trans
	EffNet-B5 [62]	83.6	96.7	30	9.9	330	456	CNN
	TinyViT-21M (ours)	**83.1**	96.5	21	4.3	1,571	224	Hybrid
	TinyViT-21M⚗ (ours)	**84.8**	97.3	21	4.3	1,571	224	Hybrid
	TinyViT-21M⚗ ↑384 (ours)	**86.2**	97.8	21	13.8	394	384	Hybrid
	TinyViT-21M⚗ ↑512 (ours)	**86.5**	97.9	21	27.0	167	512	Hybrid