

Self-Supervision

Saurabh Gupta

Is semantic supervision necessary to learn good representations?

- Manual labeling doesn't scale, suffers from biases
- Plenty of unlabeled visual data already, and growing really fast
- *And subject of the Gelato Bet:*
- *If, by the first day of autumn (Sept 23) of 2015, a method will exist that can match or beat the performance of R-CNN on Pascal VOC detection, without the use of any extra, human annotations (e.g. ImageNet) as pre-training, Mr. Malik promises to buy Mr. Efros one (1) gelato (2 scoops: one chocolate, one vanilla).*



Pre-train representations on a pre-text task

E.g. Colorization



After pre-training, use representation for down-stream tasks.

Many other possibilities,

- Spatial relationship between pair of patches

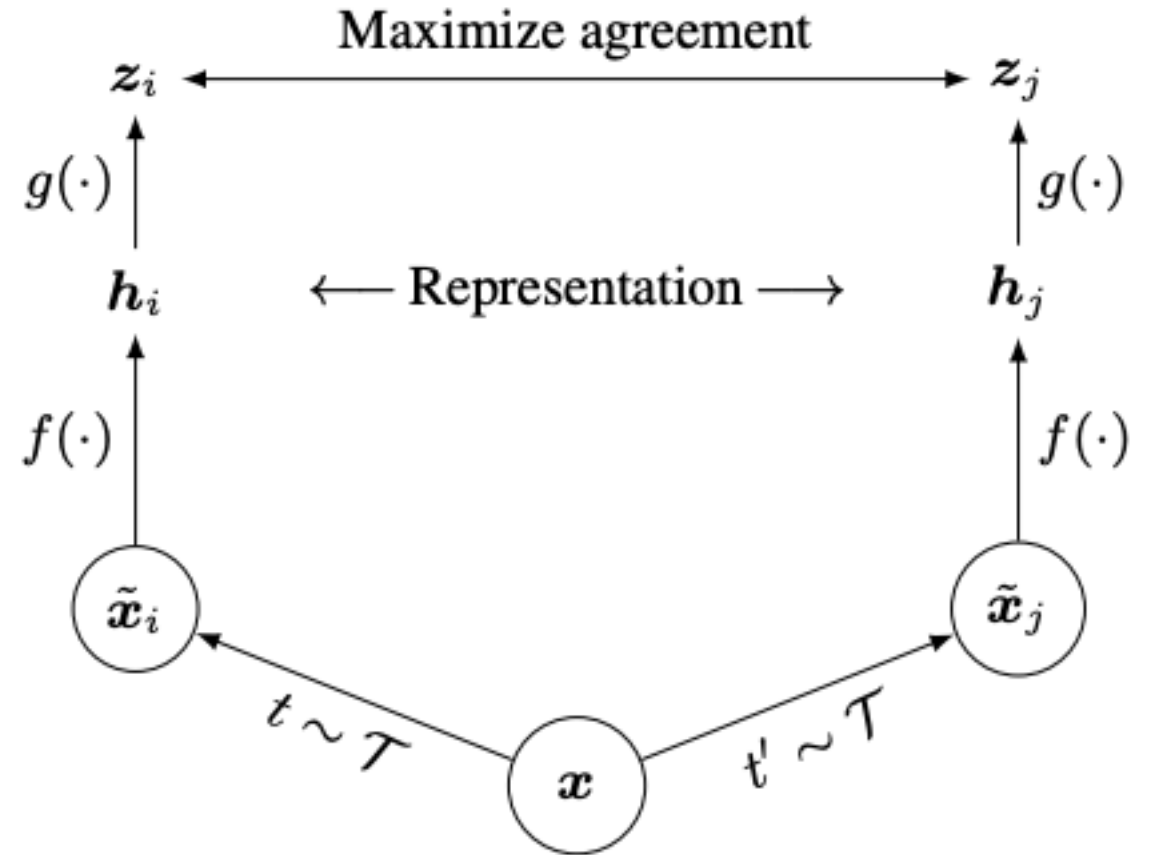


- Predict sound / frame ordering in a video
- Encourage two augmentations of same image to be closer to each other than to another image
- Predict hidden image patches from context

Contrastive Learning

- Encourage two augmentations of an image to be close.
- Using a contrastive loss:

$$l_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$



[A Simple Framework for Contrastive Learning of Visual Representations](#), Chen et al. ICML 2020

See also: [Momentum Contrast for Unsupervised Visual Representation Learning](#), He et al. CVPR 2020

Augmentations



(a) Original



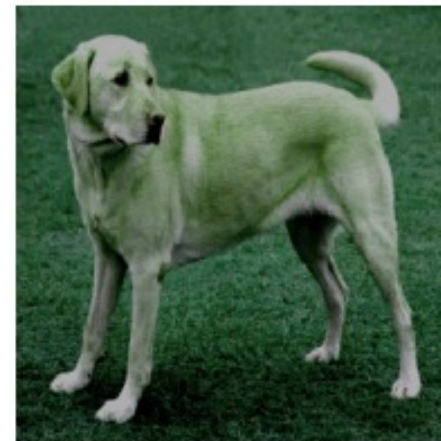
(b) Crop and resize



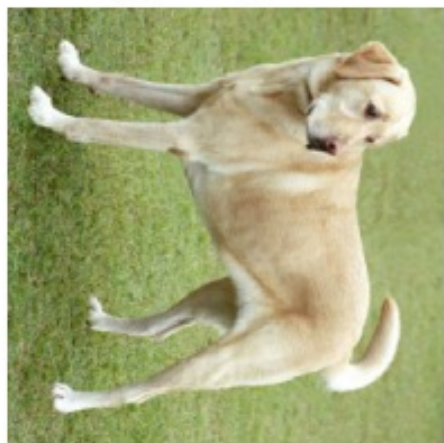
(c) Crop, resize (and flip)



(d) Color distort. (drop)



(e) Color distort. (jitter)



(f) Rotate $\{90^\circ, 180^\circ, 270^\circ\}$



(g) Cutout



(h) Gaussian noise

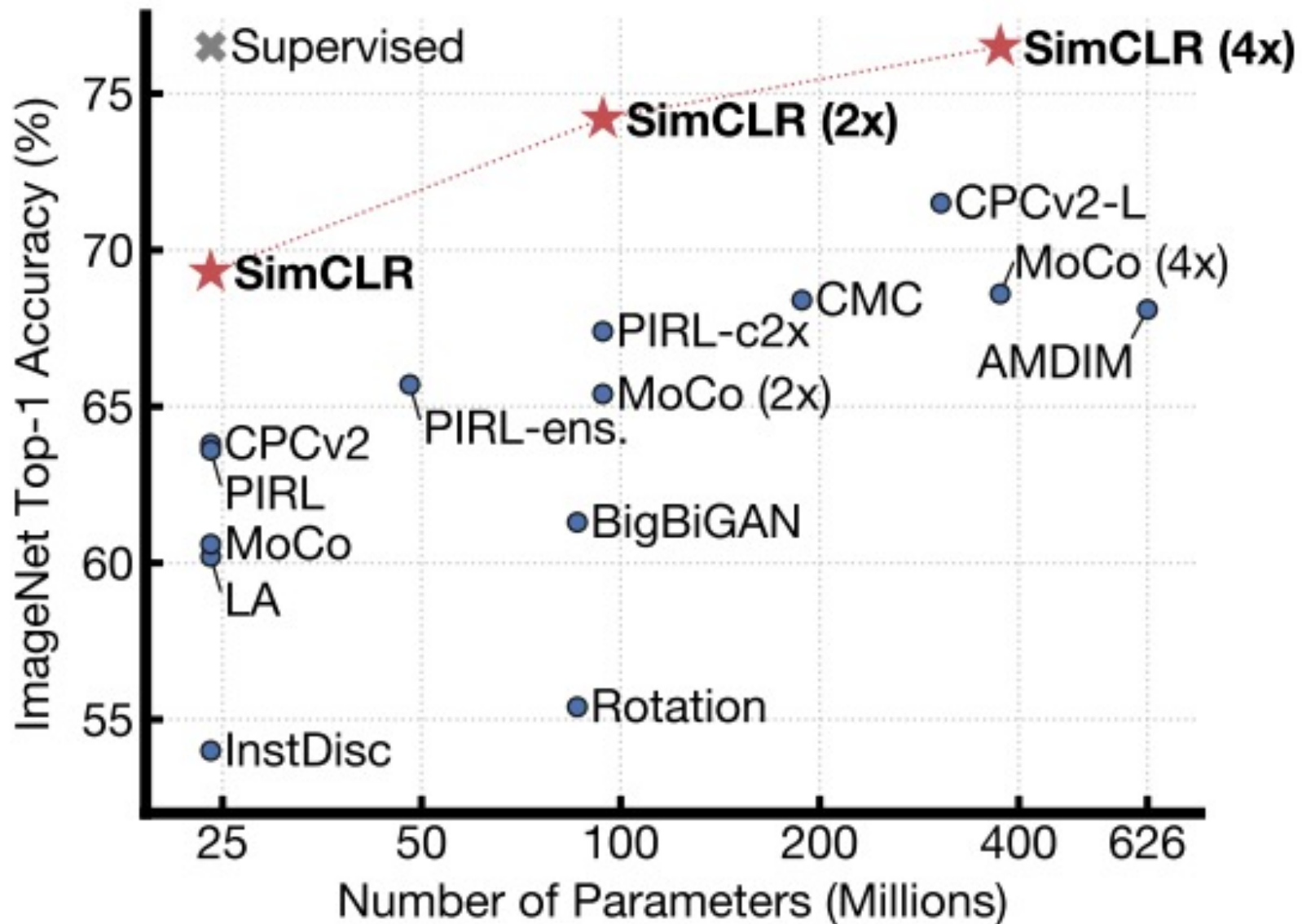


(i) Gaussian blur



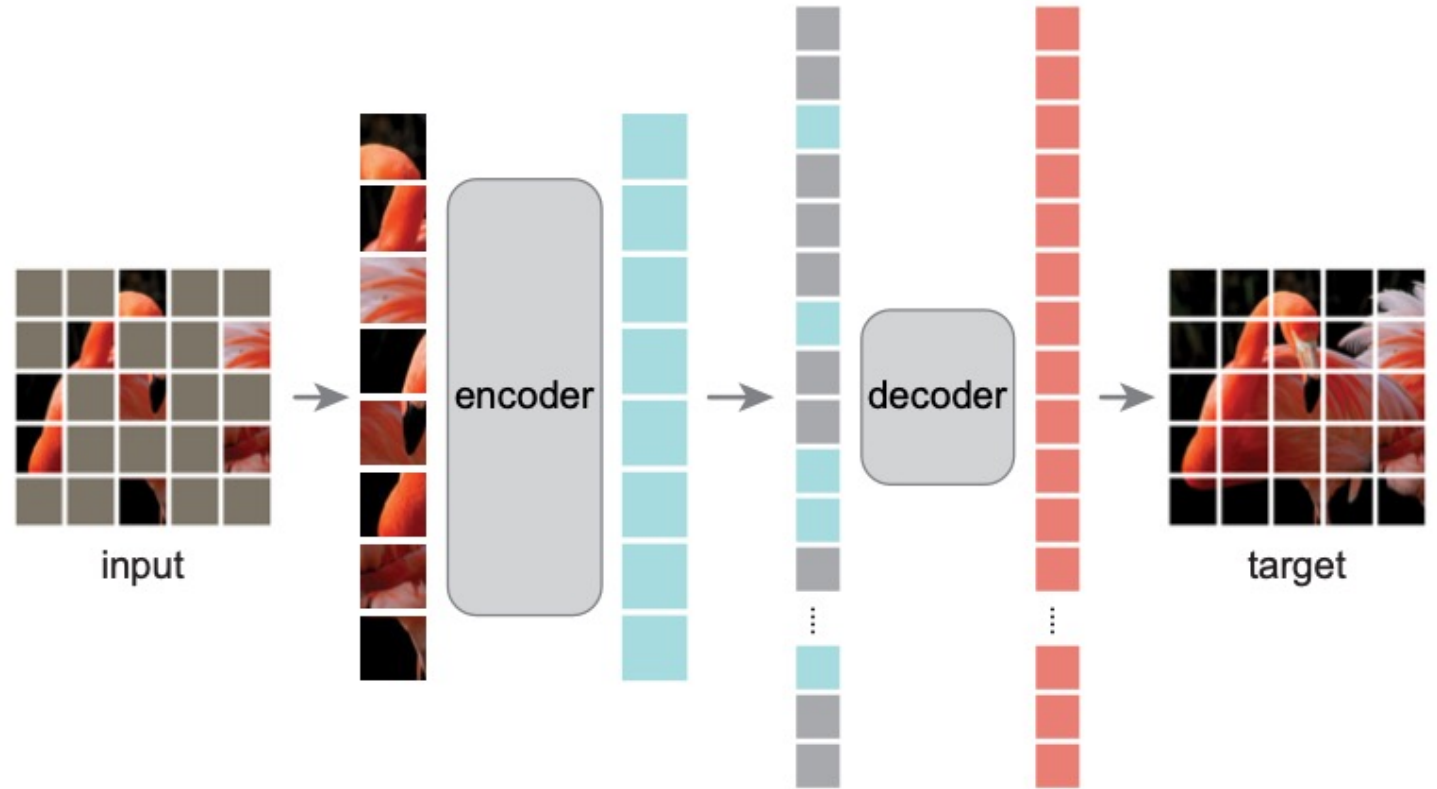
(j) Sobel filtering

Results



Masked Auto-Encoders

- Mask out image patches, predict masked patches from visible patches.
- Pre-train encoder & decoder.
- Use encoder as an image representation.



Better than semantic supervision on ImageNet 1K!

method	pre-train data	AP^{box}		AP^{mask}	
		ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	53.3	44.4	47.1
MAE	IN1K	50.3	53.3	44.9	47.2

Table 4. **COCO object detection and segmentation** using a ViT Mask R-CNN baseline. All entries are based on our implementation. Self-supervised entries use IN1K data *without* labels. Mask AP follows a similar trend as box AP.

method	pre-train data	ViT-B	ViT-L
supervised	IN1K w/ labels	47.4	49.9
MoCo v3	IN1K	47.3	49.1
BEiT	IN1K+DALLE	47.1	53.3
MAE	IN1K	48.1	53.6

Table 5. **ADE20K semantic segmentation** (mIoU) using UperNet. BEiT results are reproduced using the official code. Other entries are based on our implementation. Self-supervised entries use IN1K data *without* labels.

dataset	ViT-B	ViT-L	ViT-H	ViT-H ₄₄₈	prev best
iNat 2017	70.5	75.7	79.3	83.4	75.4 [55]
iNat 2018	75.4	80.1	83.0	86.8	81.2 [54]
iNat 2019	80.5	83.4	85.7	88.3	84.1 [54]
Places205	63.9	65.8	65.9	66.8	66.0 [19] [†]
Places365	57.9	59.4	59.8	60.3	58.0 [40] [‡]

Table 6. **Transfer learning accuracy on classification datasets**, using MAE pre-trained on IN1K and then fine-tuned. We provide system-level comparisons with the previous best results.

[†]: pre-trained on 1 billion images. [‡]: pre-trained on 3.5 billion images.

Improves performance on ImageNet itself

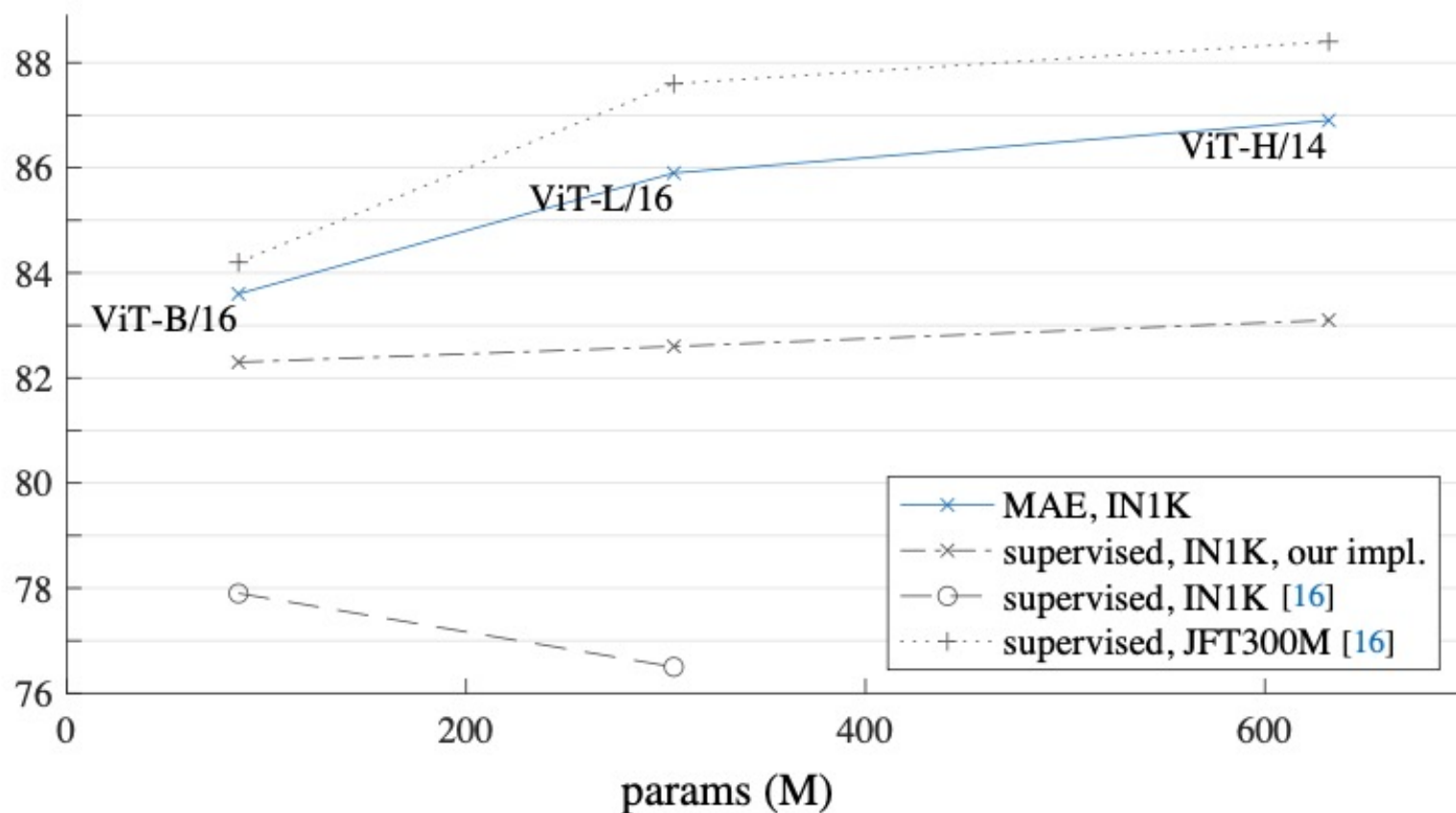
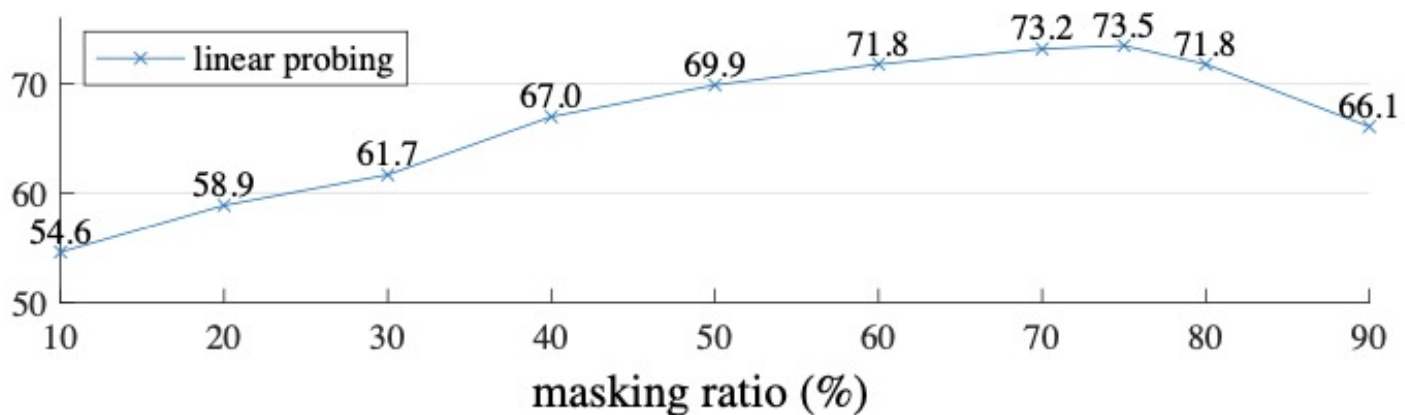
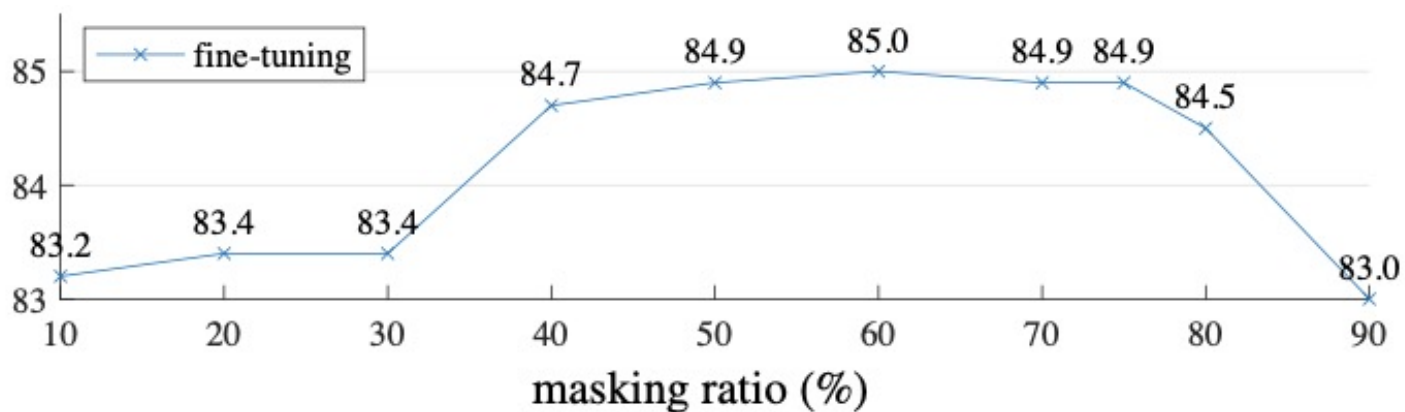


Figure 8. **MAE pre-training vs. supervised pre-training**, evaluated by fine-tuning in ImageNet-1K (224 size). We compare with the original ViT results [16] trained in IN1K or JFT300M.

Better than past self-supervision approaches

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H ₄₄₈
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	<u>83.6</u>	<u>85.9</u>	<u>86.9</u>	87.8

Ablations



Need high masking ratio for good learning.
NLP models use 15-20% masking ratio.

case	ft	lin	FLOPs
encoder w/ [M]	84.2	59.6	3.3×
encoder w/o [M]	84.9	73.5	1×

Faster and better to not input masked out patches to encoder

case	ft	lin
pixel (w/o norm)	84.9	73.5
pixel (w/ norm)	85.4	73.9
PCA	84.6	72.3
dVAE token	85.3	71.6

Normalized pixels are a better target than discrete tokens / PCA coefficients