Masked Autoencoders Are Scalable Vision Learners

Kaiming He*, Xinlei Chen* et al.

Facebook AI Research (FAIR)

Arxiv

Presenter: Minho Park

Introduction

- Higher model capacity → higher performance.
  - Models today (ViTs, …) can easily overfit. We need more data.
  - E.g., JFT-300M, 3B: Not public.
- Huge success in natural language processing by self-supervised pretraining.
  - Masked language modeling (MLM) tasks.
  - E.g., GPT (autoregressive language modeling), BERT (masked autoencoding).

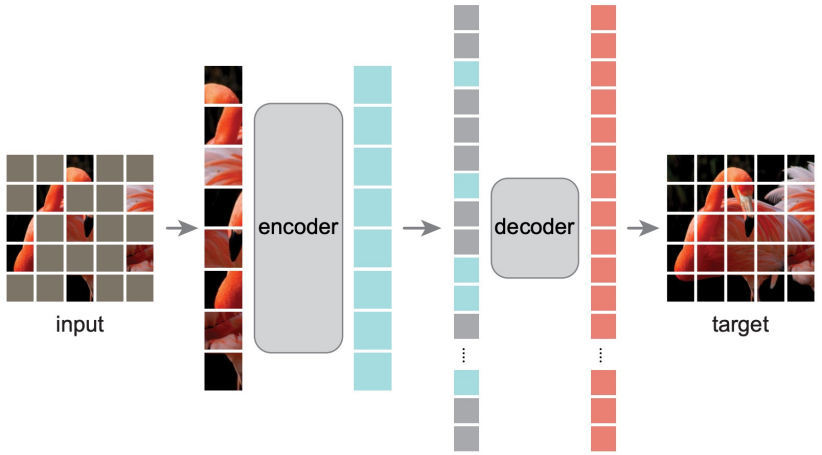

Figure 1. Our MAE architecture

Contribution

- Generalize high-capacity models (ViT-L, ViT-H) with only one million images (ImageNet-1K).
  - They achieve 87.8% accuracy on ImageNet-1K data.
  - They also evaluate transfer learning on object detection, instance segmentation, and semantic segmentation.
- It adapts well to MLM task in computer vision.
  - An asymmetric encoder-decoder architecture
  - Masking high proportion of the image.
- They observe significant gains by scaling up models.

https://paperswithcode.com/sota/image-classification-on-imagenet?tag_filter=171

The difference between vision and language

- What makes masked autoencoding different between vision and language?
- **1) Until recently, architectures were different.**
- **2) Information density is different between language and vision.**
- **3) The autoencoder's decoder.**

The difference between vision and language

- **1) Until recently, architectures were different.**
- Convolutional networks are dominant, and it is not straightforward to integrate ‘indicators’ such as mask tokens **(BEiT overcomes this with ViT).**

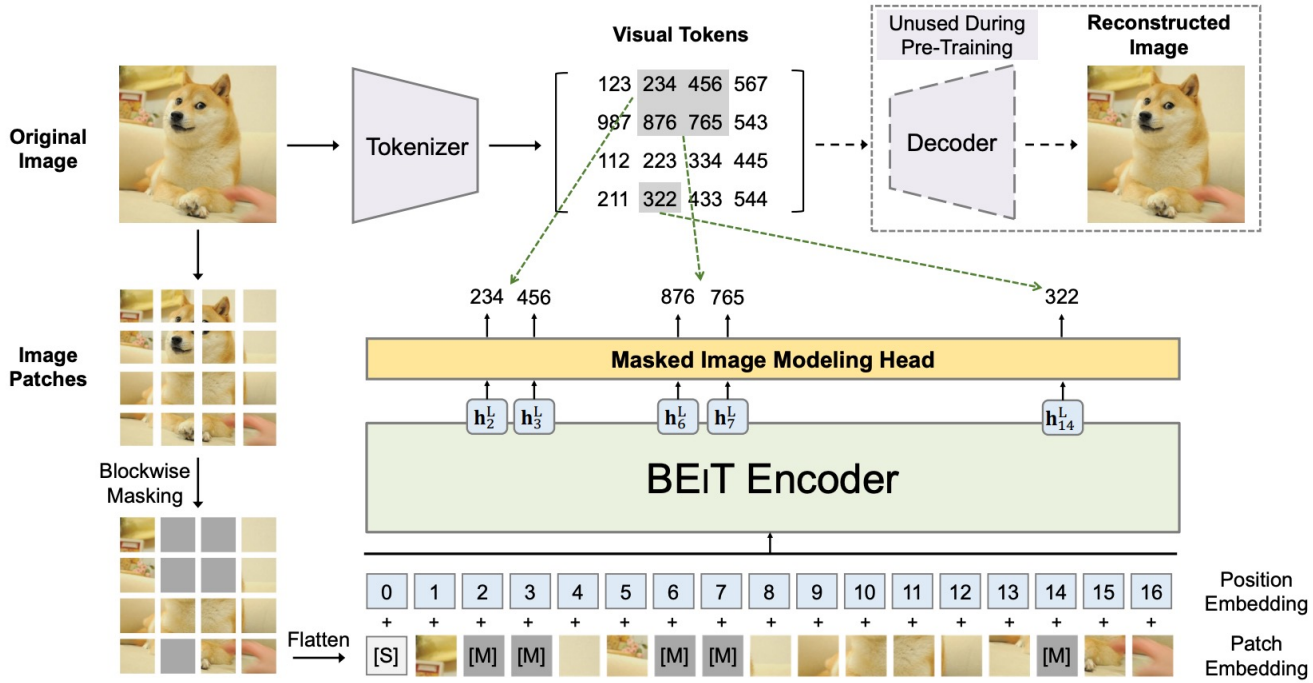

Figure 1: Overview of BEiT pre-training.

Bao, Hangbo, Li Dong, and Furu Wei. "Beit: Bert pre-training of image transformers." *arXiv preprint arXiv:2106.08254* (2021).

The difference between vision and language

- **2) Information density is different between language and vision.**
- Languages are human-generated signals that are highly semantic and information-dense.
- Images are natural signals with heavy spatial redundancy.
  - e.g., a missing patch can be recovered from neighboring patches with little high-level understanding of parts, objects, and scenes.
- ⇒ **Masking a very high portion of random patches (75%).**

Figure 2. Example results on ImageNet validation images.

The difference between vision and language

- **3) The autoencoder's decoder.**
- Words vs. pixels (lower semantic)
- The decoder of BERT can be trivial (an MLP).
- For images, the decoder design plays a key role in determining the semantic level of the learned latent representations.

Approach

- **Masked autoencoder (MAE).**
- Masking.
- MAE encoder.
- MAE decoder.
- Reconstruction target.

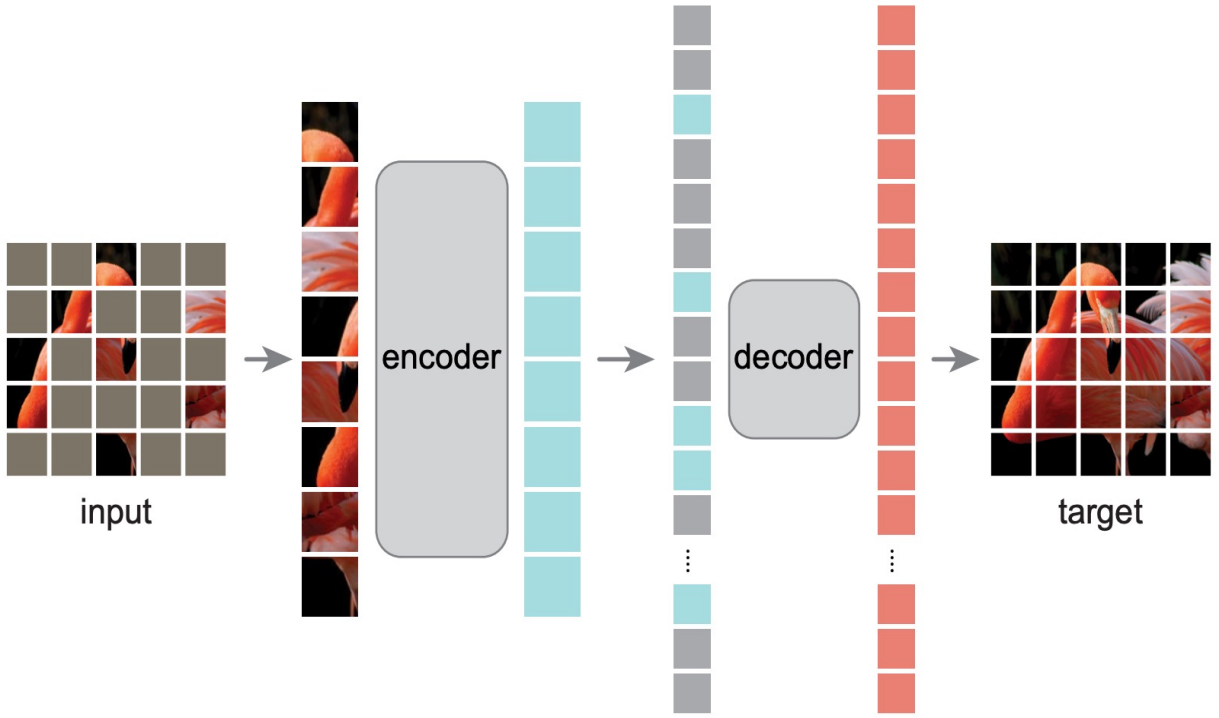

Figure 1. Our MAE architecture

Approach

- **Masking.**
- Following ViT, they divide an image into regular non-overlapping patches.
- Sample random patches following a uniform distribution.
  - Shuffle patches and remove last portion of the list.

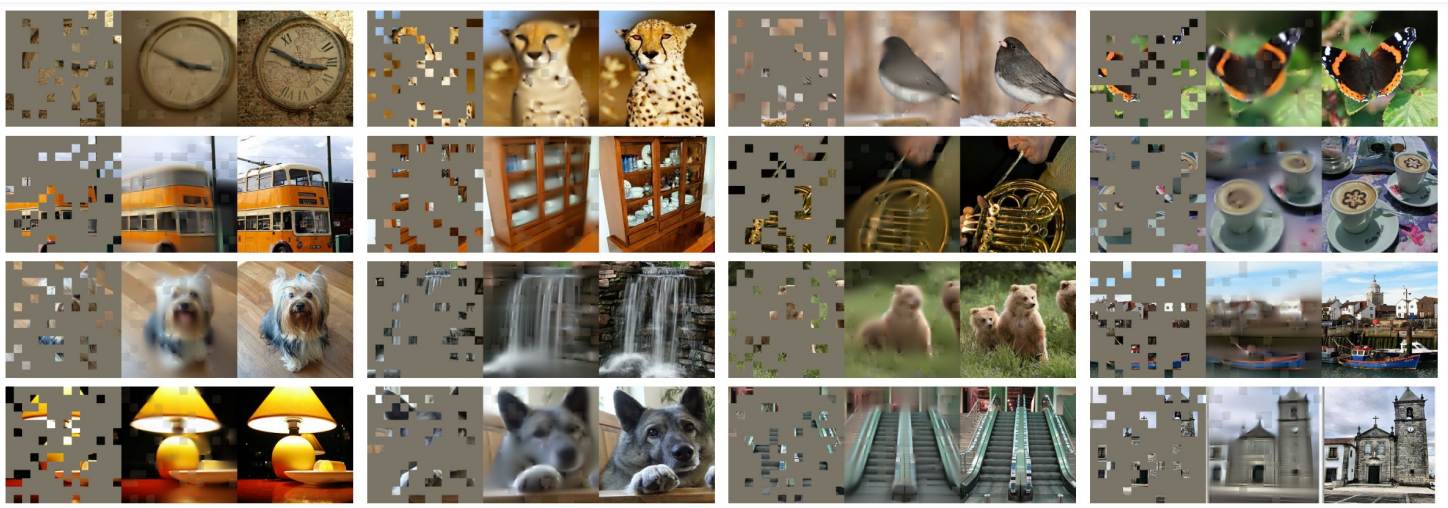

Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction† (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix.
†*As no loss is computed on visible patches, the model output on visible patches is qualitatively worse. One can simply overlay the output with the visible patches to improve visual quality. We intentionally opt not to do this, so we can more comprehensively demonstrate the method's behavior.*

Approach

- **MAE encoder.**
- Encoder is ViT but applied only on visible, unmasked patches (25%).
- This allows them to train very large encoders with only a fraction of compute and memory.

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.

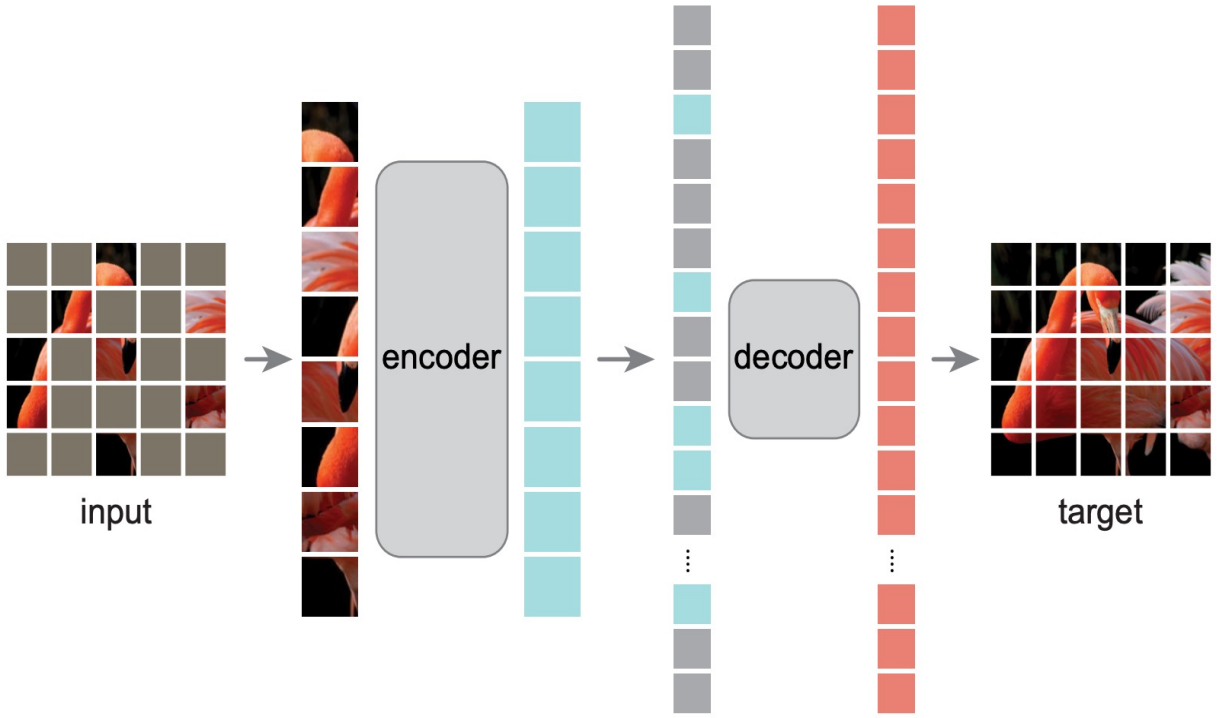

Figure 1. Our MAE architecture

Approach

- **MAE decoder.**
- Another series of Transformer blocks.
- Decoder inputs both of encoded visible patches and mask tokens.
- The MAE decoder is only used during pre-training.
- <10% computation per token vs. the encoder (significantly reduces pre-training time).

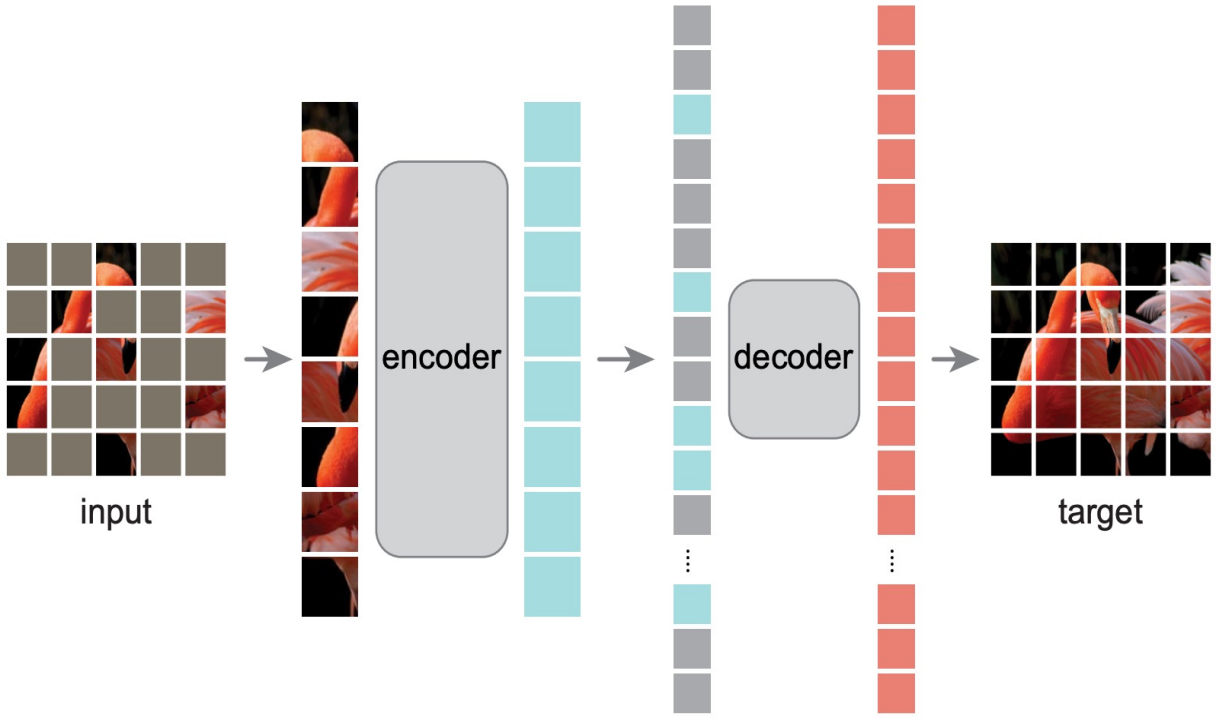

Figure 1. Our MAE architecture

Approach

- **Reconstruction target.**
- The decoder's outputs is reshaped to form a reconstructed image.
- Their loss function computes the mean squared error (MSE) only on masked patches, similar to BERT.
  - This is purely result-driven.

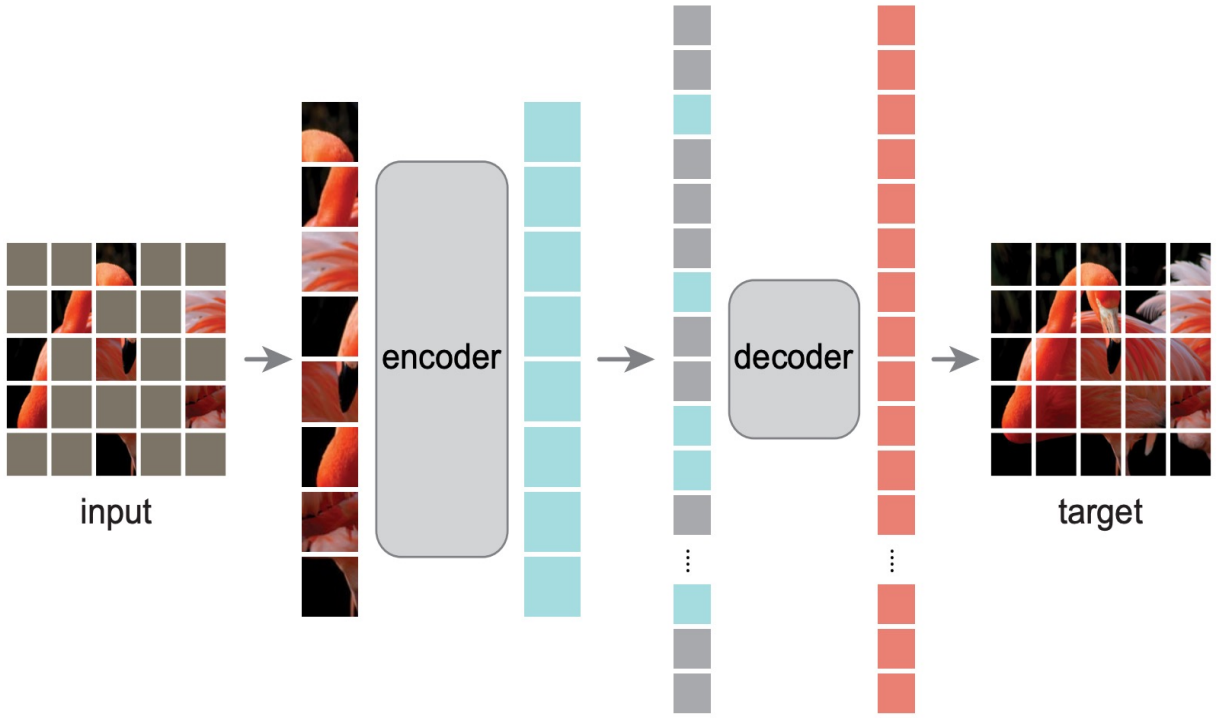

Figure 1. Our MAE architecture

ImageNet Experiments

- Self-supervised pre-training on the ImageNet-1K (IN1K) training set.
- Then they do supervised training to evaluate the representations with end-to-end fine-tuning or linear probing.
- Report top-1 validation accuracy of a single 224×224 crop.
- Baseline: ViT-Large

scratch, original [16]	scratch, our impl.	baseline MAE
76.5	82.5	84.9

config	value
optimizer	AdamW
base learning rate	1e-4
weight decay	0.3
optimizer momentum	β_1, β_2=0.9, 0.95
batch size	4096
learning rate schedule	cosine decay
warmup epochs	20
training epochs	300 (B), 200 (L/H)
augmentation	`RandAug` (9, 0.5) [12]
label smoothing [52]	0.1
mixup [69]	0.8
cutmix [68]	1.0
drop path [30]	0.1 (B), 0.2 (L/H)
exp. moving average (EMA)	0.9999

Table 11. **Supervised training ViT from scratch.**

ImageNet Experiments

- Masking ratio.
  - In contrast with BERT, whose typical masking ratio is 15%

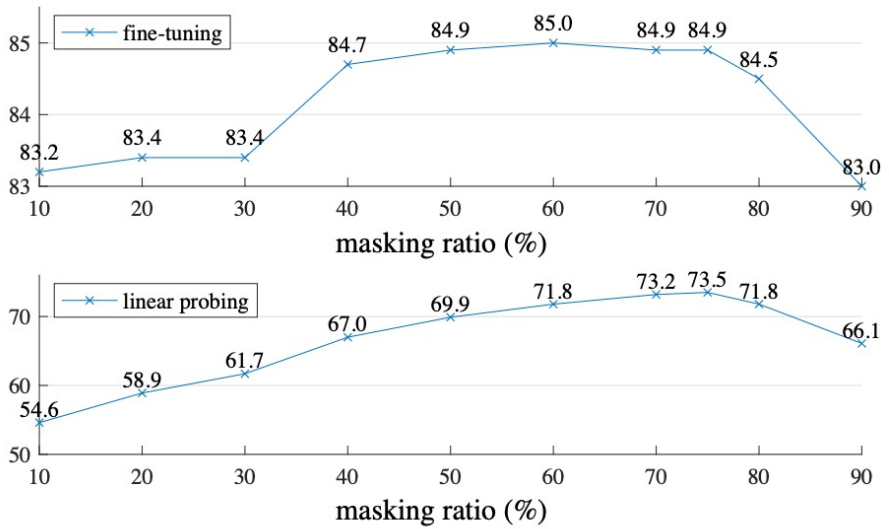

Figure 5. **Masking ratio**. A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

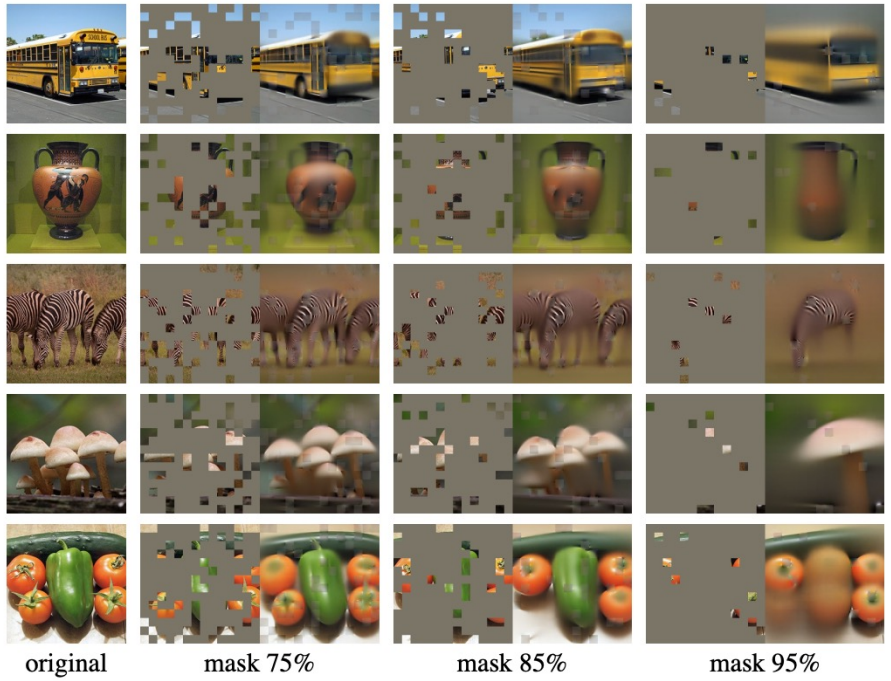

Figure 4. Reconstructions of ImageNet *validation* images using an MAE pre-trained with a masking ratio of 75% but applied on inputs with higher masking ratios. The predictions differ plausibly from the original images, showing that the method can generalize.

ImageNet Experiments

- **Main properties.**

blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

(a) **Decoder depth**. A deep decoder can improve linear probing accuracy.

dim	ft	lin
128	**84.9**	69.1
256	84.8	71.3
512	**84.9**	**73.5**
768	84.4	73.1
1024	84.3	73.1

(b) **Decoder width**. The decoder can be narrower than the encoder (1024-d).

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

(c) **Mask token**. An encoder without mask tokens is more accurate and faster (Table 2).

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

(d) **Reconstruction target**. Pixels as reconstruction targets are effective.

case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

(e) **Data augmentation**. Our MAE works with minimal or no augmentation.

case	ratio	ft	lin
random	75	**84.9**	**73.5**
block	50	83.9	72.3
block	75	82.8	63.9
grid	75	84.0	66.0

(f) **Mask sampling**. Random sampling works the best. See Figure 6 for visualizations.

Table 1. **MAE ablation experiments** with ViT-L/16 on ImageNet-1K. We report fine-tuning (ft) and linear probing (lin) accuracy (%). If not specified, the default is: the decoder has depth 8 and width 512, the reconstruction target is unnormalized pixels, the data augmentation is random resized cropping, the masking ratio is 75%, and the pre-training length is 800 epochs. Default settings are marked in gray.

ImageNet Experiments

- **Decoder design.**
- A sufficiently deep decoder is important for linear probing.
  - This can be explained by the gap between a pixel reconstruction task and a recognition task.
- It only has 9% FLOPs per token vs. ViT-L (24 blocks, 1024-d).

blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

(a) **Decoder depth**. A deep decoder can improve linear probing accuracy.

dim	ft	lin
128	**84.9**	69.1
256	84.8	71.3
512	**84.9**	**73.5**
768	84.4	73.1
1024	84.3	73.1

(b) **Decoder width**. The decoder can be narrower than the encoder (1024-d).

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1: Details of Vision Transformer model variants.

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

ImageNet Experiments

- **Mask token.**
- If the encoder uses mask tokens, it performs worse.
  - Encoder has a large portion of mask tokens in pre-training, which does not exist in uncorrupted images.
- Skipping the mask token in the encoder reduce training computation (3.3× FLOPs, 3~4× speedup).
  - Time and memory efficiency.

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

(c) **Mask token**. An encoder without mask tokens is more accurate and faster (Table 2).

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.

ImageNet Experiments

- **Reconstruction target.**
- Using pixels with normalization improves accuracy.
- BEiT predicts tokens of DALL-E pre-trained dVAE.
  - dVAE tokenizer requires one more pre-training stage, which may depend on extra data (250M images).
  - However, the difference is statistically insignificant.

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

(d) **Reconstruction target**. Pixels as reconstruction targets are effective.

	IN1K			COCO		ADE20K	
	ViT-B	ViT-L	ViT-H	ViT-B	ViT-L	ViT-B	ViT-L
pixel (w/o norm)	83.3	85.1	86.2	49.5	52.8	48.0	51.8
pixel (w/ norm)	83.6	85.9	86.9	50.3	53.3	48.1	53.6
dVAE token	83.6	85.7	86.9	50.3	53.2	48.1	53.4
△	0.0	-0.2	0.0	0.0	-0.1	0.0	-0.2

Table 7. **Pixels *vs*. tokens** as the MAE reconstruction target. △ is the difference between using dVAE tokens and using normalized pixels. The difference is statistically insignificant.

Ramesh, Aditya, et al. "Zero-shot text-to-image generation." *International Conference on Machine Learning*. PMLR, 2021.

ImageNet Experiments

- **Data augmentation and mask sampling.**

case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

(e) **Data augmentation**. Our MAE works with minimal or no augmentation.

case	ratio	ft	lin
random	75	**84.9**	**73.5**
block	50	83.9	72.3
block	75	82.8	63.9
grid	75	84.0	66.0

(f) **Mask sampling**. Random sampling works the best. See Figure 6 for visualizations.

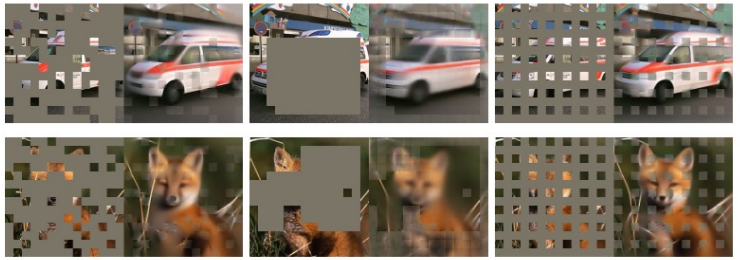

Figure 6. **Mask sampling strategies** determine the pretext task difficulty, influencing reconstruction quality and representations (Table 1f). Here each output is from an MAE trained with the specified masking strategy. Left: random sampling (our default). Middle: block-wise sampling [2] that removes large random blocks. Right: grid-wise sampling that keeps one of every four patches. Images are from the validation set.

Comparison with Previous Results

- By fine-tuning with a 448 size, we achieve 87.8% accuracy, using only IN1K data.
- MAE is more accurate than BEiT while being simpler and faster (3.5× per epoch).

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	83.6	85.9	86.9	**87.8**

Table 3. **Comparisons with previous results on ImageNet-1K**. The pre-training data is the ImageNet-1K training set (except the tokenizer in BEiT was pre-trained on 250M DALLE data [50]). All self-supervised methods are evaluated by end-to-end fine-tuning. The ViT models are B/16, L/16, H/14 [16]. The best for each column is underlined. All results are on an image size of 224, except for ViT-H with an extra result on 448. Here our MAE reconstructs normalized pixels and is pre-trained for 1600 epochs.

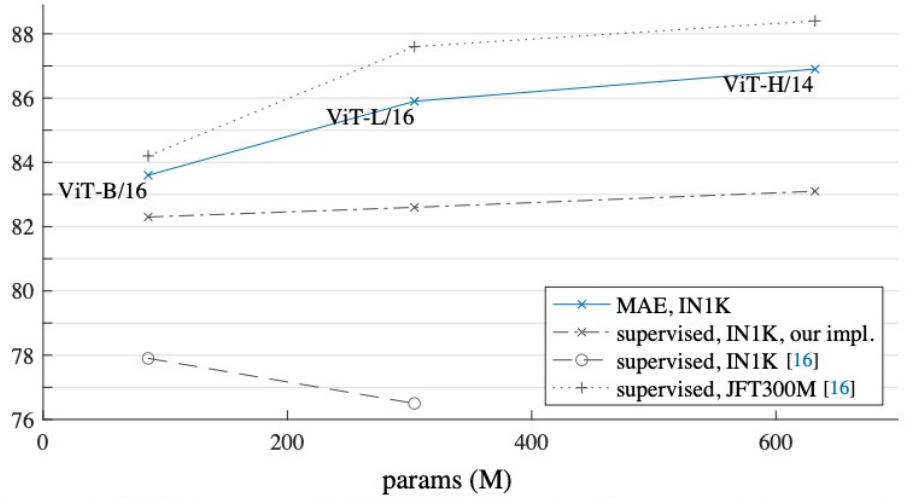

Figure 8. **MAE pre-training *vs*. supervised pre-training**, evaluated by fine-tuning in ImageNet-1K (224 size). We compare with the original ViT results [16] trained in IN1K or JFT300M.

Partial Fine-tuning

- Linear separability is not the sole metric for evaluating representation quality.

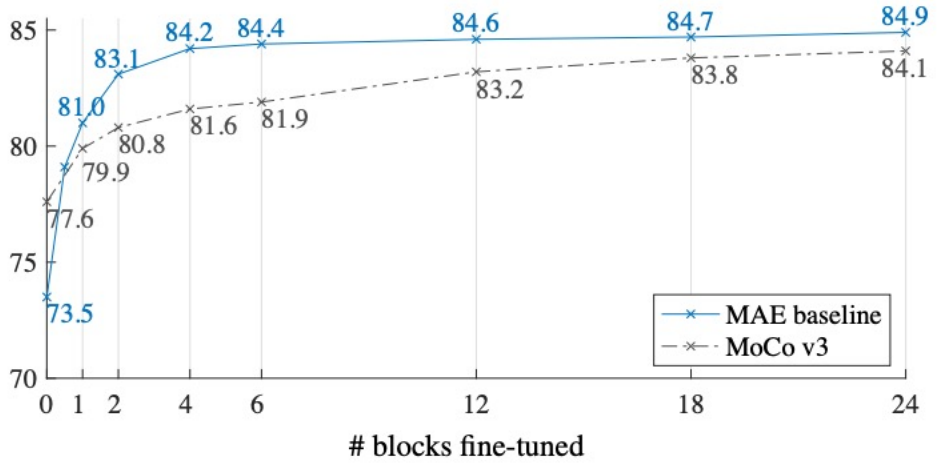

Figure 9. **Partial fine-tuning** results of ViT-L w.r.t. the number of fine-tuned Transformer blocks under the default settings from Table 1. Tuning 0 blocks is linear probing; 24 is full fine-tuning. Our MAE representations are less linearly separable, but are consistently better than MoCo v3 if one or more blocks are tuned.

Transfer Learning Experiments

- Downstream tasks (classification, object detection, instance segmentation, and semantic segmentation)

		AP^{box}		AP^{mask}	
method	pre-train data	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.

Uncurated Random Samples on ImageNet

Uncurated Random Samples on COCO

Thank You