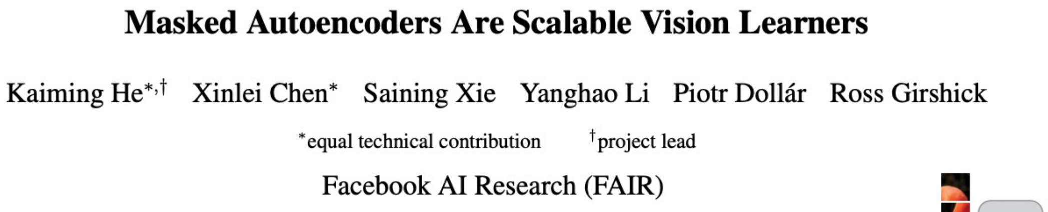
Masked Autoencoders Are Scalable Vision Learners

Facebook AI Research (FAIR)

2021.12.08 이지현



Masked Autoencoders Are Scalable Vision Learners

Kaiming He[*,†] Xinlei Chen[*] Saining Xie Yanghao Li Piotr Dollár Ross Girshick

[*]equal technical contribution [†]project lead

Facebook AI Research (FAIR)

https://arxiv.org/pdf/2111.06377.pdf

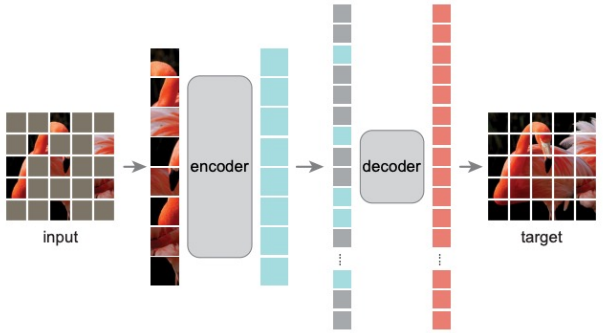


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed 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 to produce representations for recognition tasks.

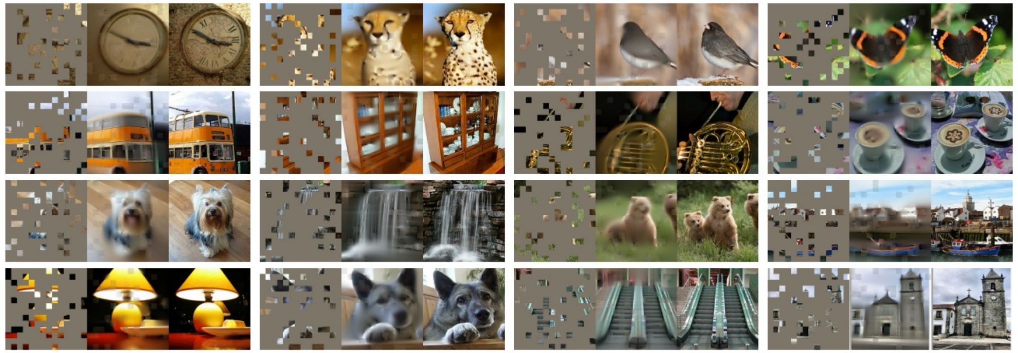


Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction[†] (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix. [†]*As no loss is computed on 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.*

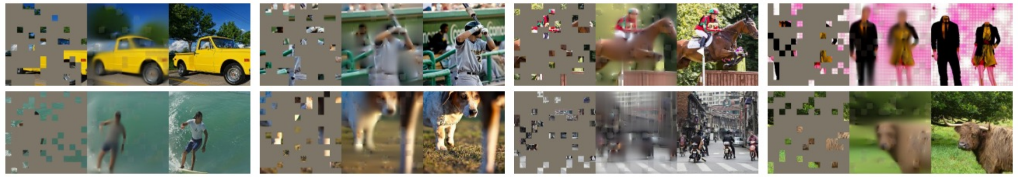


Figure 3. Example results on COCO validation images, using an MAE trained on ImageNet (the same model weights as in Figure 2). Observe the reconstructions on the two right-most examples, which, although different from the ground truth, are semantically plausible.

- FAIR, Kaiming He, Ross Girshick

- **Masked Autoencoders (MAE) 제안**
  - **Autoencoder 활용하는 방법, masked 되지 않은 조각들만 보고 나머지를 reconstructs 할 수 있도록 학습**
    - Encoder: latent representation 학습
    - Decoder: original signal 로 reconstructs

  - **Classical autoencoders 와의 차이점**
    - asymmetric design (encoder 와 decoder 가 asymmetric 하게 생김)
      - Encoder 는 masking token 을 사용하지 않고,
      - Decoder 에서만 masking token 사용

1. Introduction

- **Deep learning 은 폭발적으로 발전을 했고, 성능도 좋아졌지만, 모델 capability 와 capacity 도 함께 커짐.**
- 모델이 커짐으로 인해서, 쉽게 overfit 되고, 더 많은 데이터가 요구됨.
- **NLP 에서는 self-supervised pretraining 으로 해당 문제를 잘 해결함.**
  - **GPT, BERT 모델: they remove a portion of the data and learn to predict the removed content.**
  - GPT: autoregressive 식으로 다음 단어를 예측. (autoregressive language modeling)
  - BERT: 중간 중간 어떤 단어를 masking 한 후, 이를 예측. (masked autoencoding)
  - 모델 사이즈를 키우면서, 많은 데이터를 통해서 self-supervised 학습을 할 수 있었음.
- **그러나, computer vision 분야는 autoencoding 형태의 연구가 부족함.**

- ***What makes masked autoencoding different between vision and language?***

1) 아키텍처가 다름. Architectures were different.
   - Vision: CNN은 일반적으로 regular grids 를 사용하기 때문에,
     masking token, positional embeddings 와 같은 'indicators' 를 CNN 네트워크에 추가하는 것은 어렵다.
   - ViT 가 나오면서 극복

2) Information density 가 다르다! Information density is different between language and vision.
   - Language (human-generated signals) 는 단어 하나하나가 의미가 있음. (highly semantic and information-dense)
   - 이미지는 픽셀 하나가 특별한 의미가 있지 않음. Massing patch 가 recover 하는게 그렇게 어려운 일은 아니다.
   - 이러한 차이를 극복하기 위해서, masking a very high portion of random patches.

3) Autoencoder 의 decoder
   - Vision: 픽셀 예측, 디코더가 픽셀 값 하나하나가 시멘틱한 의미는 없음. (lower semantic level)
   - Languae: missing words , 단어가 많은 semantic information 을 가짐

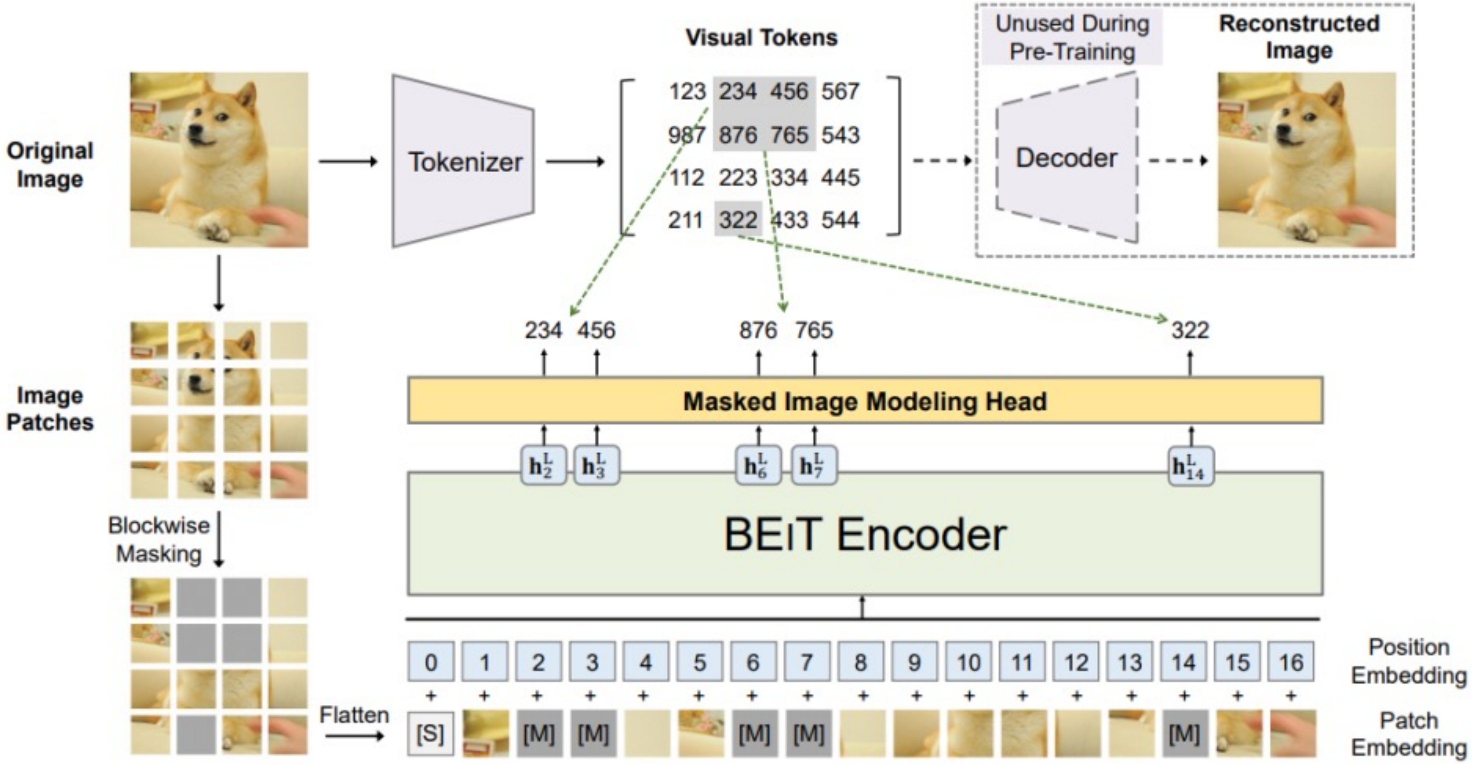
2. Related Work

- **Masked language modeling**
  - BERT, GPT: highly successful methods for pre-training in NLP.
  - 일부분을 없앤 다음에 그것을 예측하는 형태로 학습.
  - scale up 하기가 좋음. Generalize 가 잘돼서 다양한 downstream task 에 적용 가능.

- **Autoencoding**
  - Classicial method for learning representations.
  - Encoder: latent representation 으로 mapping.
  - Decoder: 그걸 다시 input 과 같은 형태로 recontruction.
  - Denoising autoencoders (DAE): input 시그널에 노이즈를 추가한 다음, 노이즈가 없어지는 형태로 복원.

- **Masked image encoding**
  - 이미지를 마스킹 해서 인코딩 하는 방식.
  - Pioneering work: DAE
  - Context encoder : Inpaints large missing regions using cnn. (지워놓고 이미지를 채우는)
  - NLP 에 영감을 받아, iGPT, ViT, DEiT

2. Related Work – BEiT : BERT Pre-Training of Image Transformers



BERT 학습법을 이미지에 사용하여
self-supervised 학습에서 좋은 성능

1) 이미지를 패치로 나누고
2) 일부 마스킹
3) 마스킹 한 것을 BERT 처럼 예측.

- BERT 에서는
  MLM (Masked Language Modeling)
-> MIM (Masked Image Modeling)

- 차이점
  - VAE (Variational AutoEncoder)를 사전학습시키고, freeze -> 이 VAE 의 encoder 부분이 Tokenizer.
  - 이미지를 패치로 나눈 후, 패치를 마스킹한 후 그 마스킹을 맞추는데,
    이 때 패치를 바로 예측하는 것이 아니고, 패치를 tokenizing 하여 tokenizer 를 기준으로 예측. (i.e. label 은 tokenizer 기준).

  - 여러가지 downstream task 에서 적용 가능.

3. Approach

- **Masked autoencoder (MAE)**
  - MAE is a simple autoencoding approach that reconstructs the original signal given its partial observation.
  - **Autoencoder 를 활용.**
    - Encoder: latent representation 학습
    - Decoder: original signal 로 reconstructs

  - **Classical autoencoders 와의 차이점.**
    - asymmetric design (encoder 와 decoder 가 asymmetric 하게 생김, 비대칭적)
      - Encoder 는 masking token 을 사용하지 않고,
      - Decoder 에서만 masking token 사용.

1) Input image 를 패치로 나누고 (overlap 되지 않게)
2) 굉장히 높은 비중으로 masking (default 75%)
3) Decoder 로 갈 때, masking token 추가. -> masking token 예측.

* Decoder 는 pre-training 될 때만 사용.

i.g. 이렇게 모든 학습이 완료된 후,

decoder 는 자르고, layer 몇 개를 붙여서 downstream task 추가 학습.

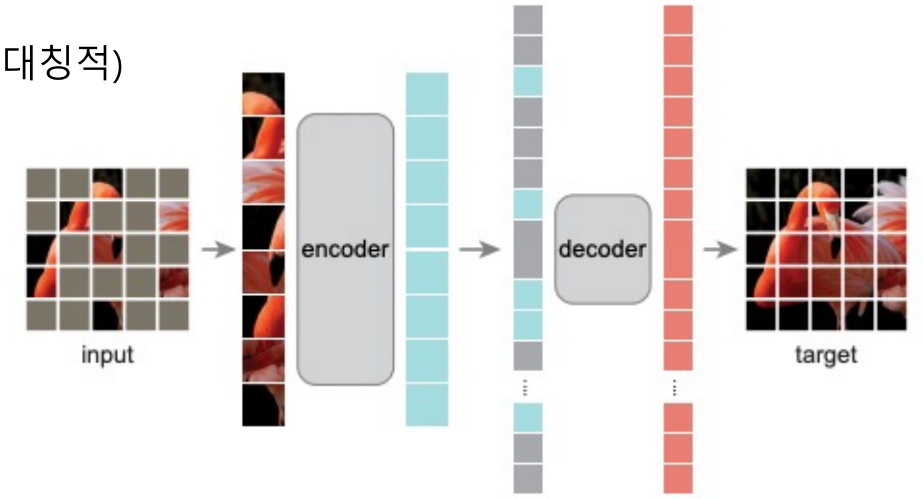


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed 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.

3. Approach - MAE Details

- **Masking**
  - Masking 하는 부분 – overlap 되지 않게
  - Random 하게 masking
  - Replacement X
  - Uniform distribution 에서 sampling (전반적으로 고르게 masking 되도록)
  - High masking ratio 이기 때문에 단순히 extrapolation 방법으로 만들어내기 어렵다.
    - 따라서 이런 방식으로 학습하면 좋은 representation 학습을 야기한다.

- **MAE encoder**
  - MAE encoder is a ViT but applied only on visible, unmasked patches.
  - 일반적인 ViT 와 같이, input patch 를 linear projection 을 통해 positional embeddings.
  - Small subset (e.g. 25%) 정도로 학습하기 때문에, 3배 이상 빠르게 학습한다.
  - 더 큰 모델로 scale up 하기 좋다. (연산량, 메모리 사용량 적음)

- **MAE decoder**
  - The input to the AME decoder is the full set of tokens consisting of (1) encoded visible patches (encoder 에서 만들어준 패치) (2) mask tokens.
  - Mask token 역시 학습하는 토큰. (shared, learned vector)
  - Positional embeddings 사용
  - 실제로 downstream task 에는 사용되지 않기 때문에 decoder architecture 는 flexible designed.
  - MAE's default decoder has < 10% computation per token vs. the encoder. Pre-training time 감소.

3. Approach - MAE Details

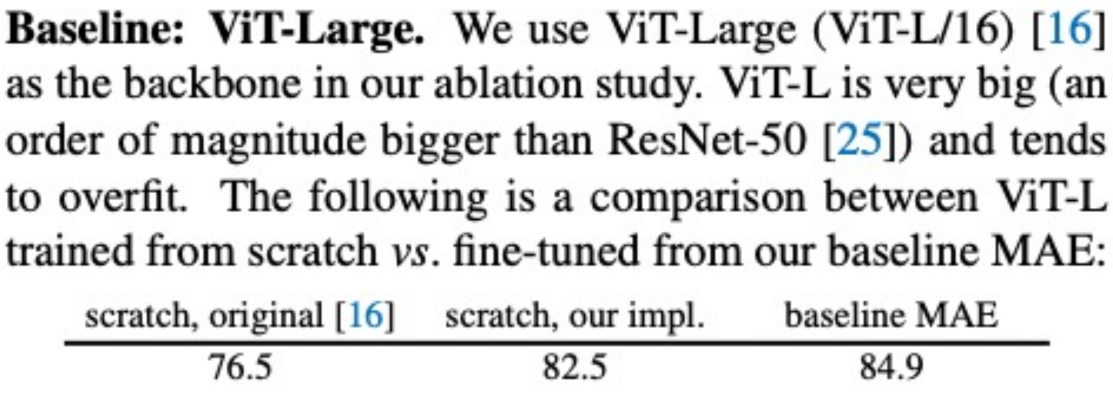
- **Reconstruction target**
  - Original image 그대로 복원.
  - MAE reconstructs the input by predicting the pixel values for each masked patch.
  - The loss function computes the MSE (mean squared error) between the reconstructed and original images in the pixel space. Computing the loss only on masked patches, similar to BERT.

4. Experiments

- **ImageNet**
  - We do self-supervised pre-training on the ImageNet-1K training set.
  - Then we do supervised training to evaluate the representations with (1) end-to-end fine-tuning or (2) linear probing.
    - End-to-end: 모든 layer 학습.
    - Linear probing: self-supervised learning 으로 pre-training 한 네트워크에 layer 1개만 추가하여, classification 과 같은 downstream task 를 학습시켜서 결과를 보는 방법. (pre-training 했던 부분은 freeze)

      Pre-training 이 얼마나 좋은 representation 이 학습되었는지 확인하는 방법.

  - **Baseline: ViT-Large**



Baseline: ViT-Large. We use ViT-Large (ViT-L/16) [16] as the backbone in our ablation study. ViT-L is very big (an order of magnitude bigger than ResNet-50 [25]) and tends to overfit. The following is a comparison between ViT-L trained from scratch *vs*. fine-tuned from our baseline MAE:

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

4. Experiments

- **Masking Ratio**
  - Fine-tuning, linear probing 모두 75% (생각보다 높은 ratio) 에서 성능이 잘 나옴.
  - BERT 는 15% 만 사용하는데, 굉장히 높은 ratio 사용한다.

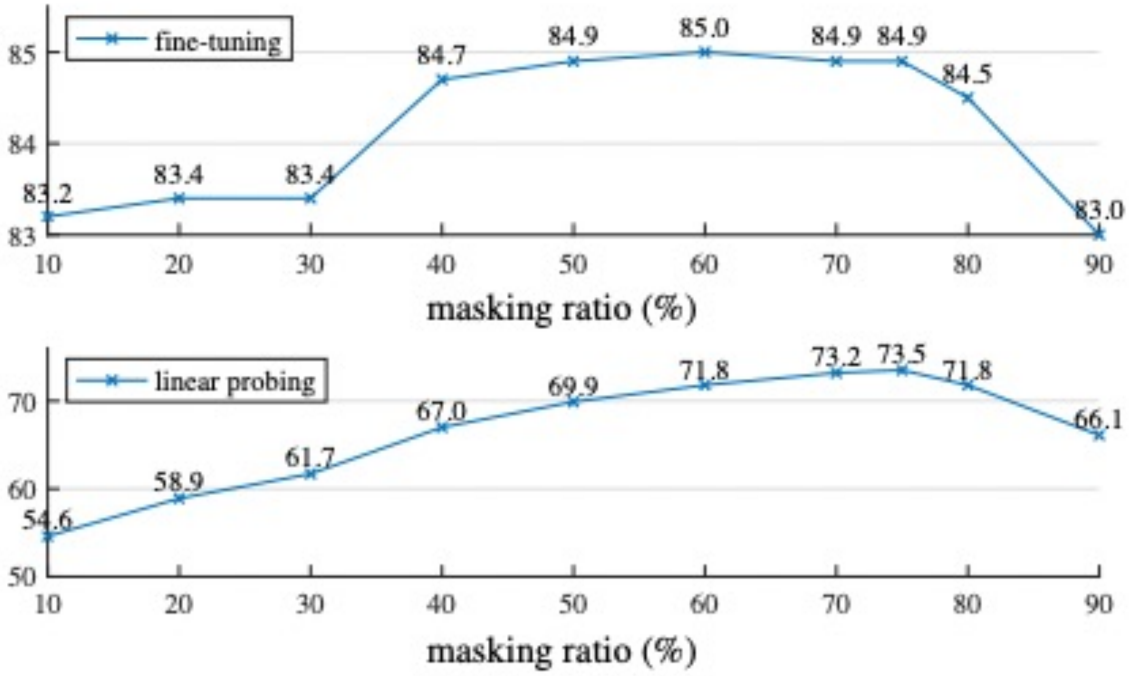
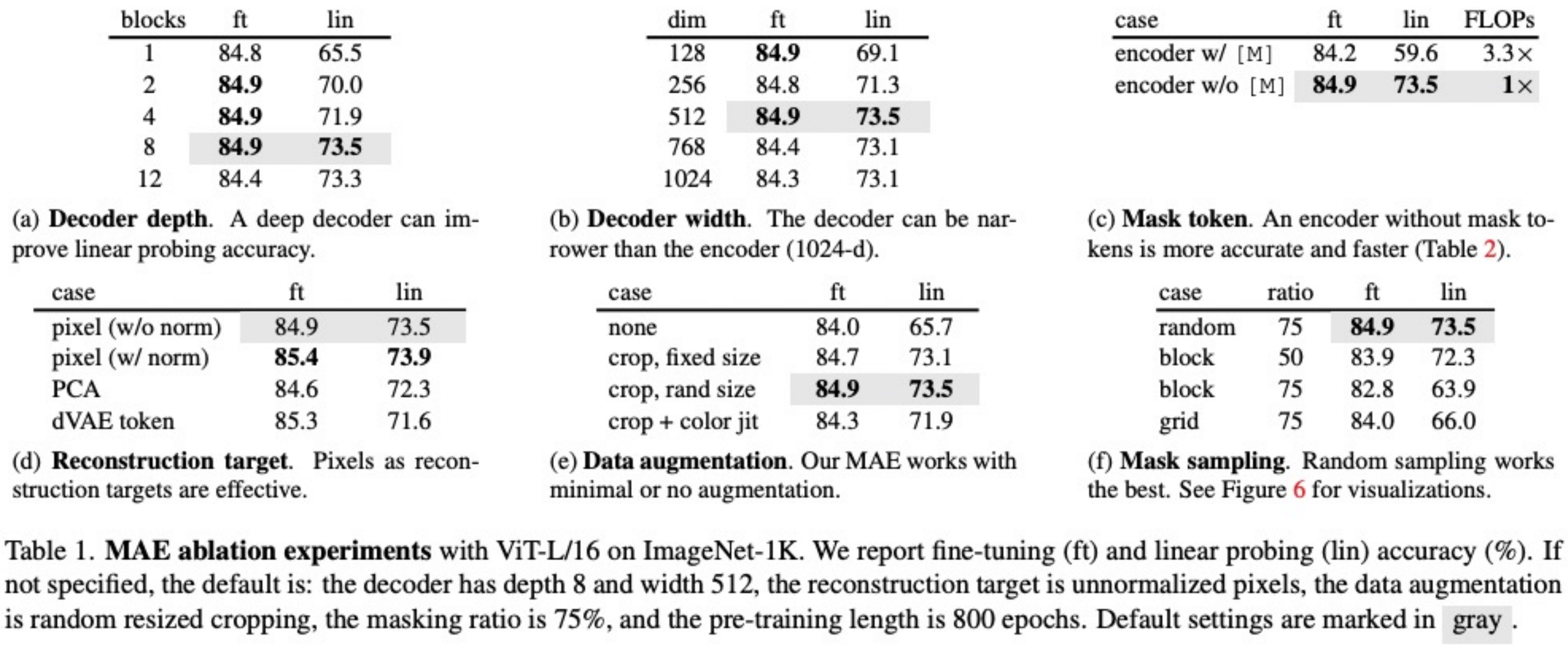


Figure 5. **Masking ratio**. A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

- **Fine-tuning 은** 몇 % 를 마스킹 하던지간에 from the scratch (82.5%) 보다 성능이 좋다.

- **linear probing 은** fine-tuning 에 비해 less sensitive to the ratio.

4. Experiments



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.

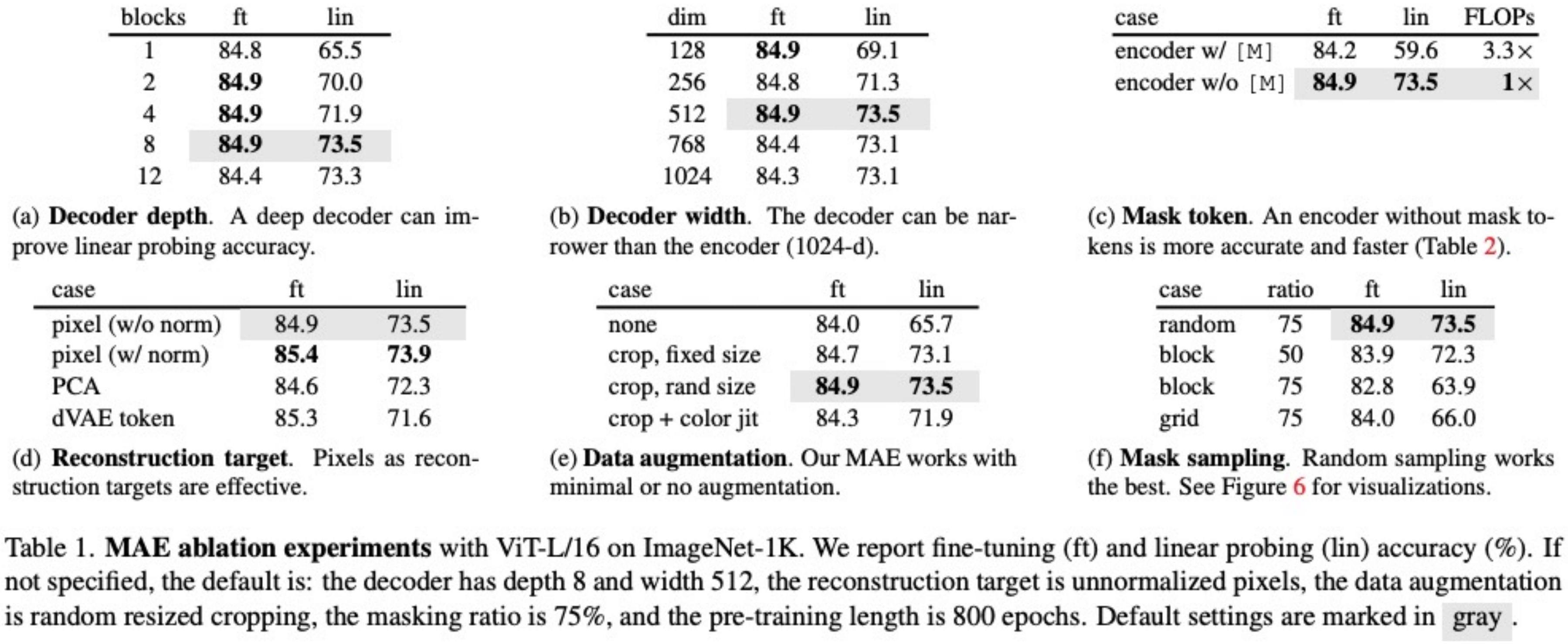
case	ratio	ft	lin
random	75	**84.9**	**73.5**
block	50	83.9	72.3
block	75	82.8	63.9
grid	75	84.0	66.0

(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 .

- **Decoder Design (a), (b)**
  - Depth (transformer block 의 개수), dim 에 따라 실험
  - 충분히 deep 한 decoder 가 필요하다.
  - Default MAE decoder is lightweight. It has 8 blocks and a width of 512-d. It only has 9% FLOPs per token vs. ViT-L (24 blocks, 1024-d)
- **Mask Token (을 사용하느냐 안하느냐) (c)**
  - Encoder 에서 mask token 을 사용하면 오히려 성능이 떨어짐.
  - 실제로 inference 할 때는 masking 안 된 이미지가 들어올텐데, 학습할 때 mask token 을 사용하면 inference 와 차이가 생김.

4. Experiments



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.

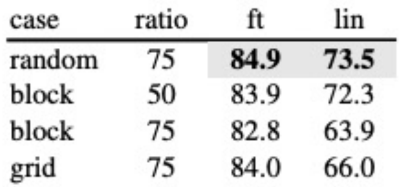
case	ratio	ft	lin
random	75	**84.9**	**73.5**
block	50	83.9	72.3
block	75	82.8	63.9
grid	75	84.0	66.0

(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 .

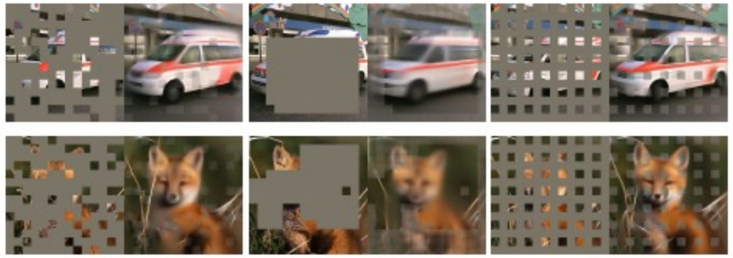
- **Reconstruction target (d)**
  - 픽셀 예측 / PCA 예측 / token 예측 (BEiT)
  - 픽셀 직접 예측하는게 성능이 제일 좋음.
- **Data Augmentation (e)**
  - 아무런 data augmentation 을 사용하지 않아도, 성능 차이가 없음.
  - In MAE, the rold of data augmentation is mainly performed by random masking. The masks are different for each iteration and so they generate new training samples regardless of data augmentation.

4. Experiments



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.



random 75%   block 50%   grid 75%

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.

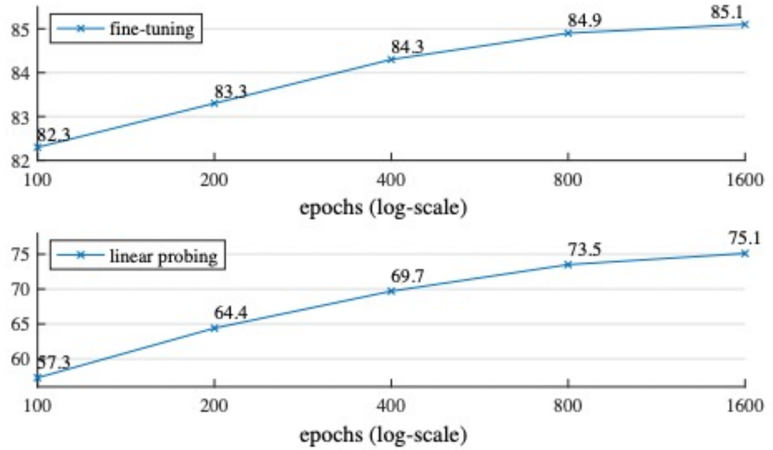


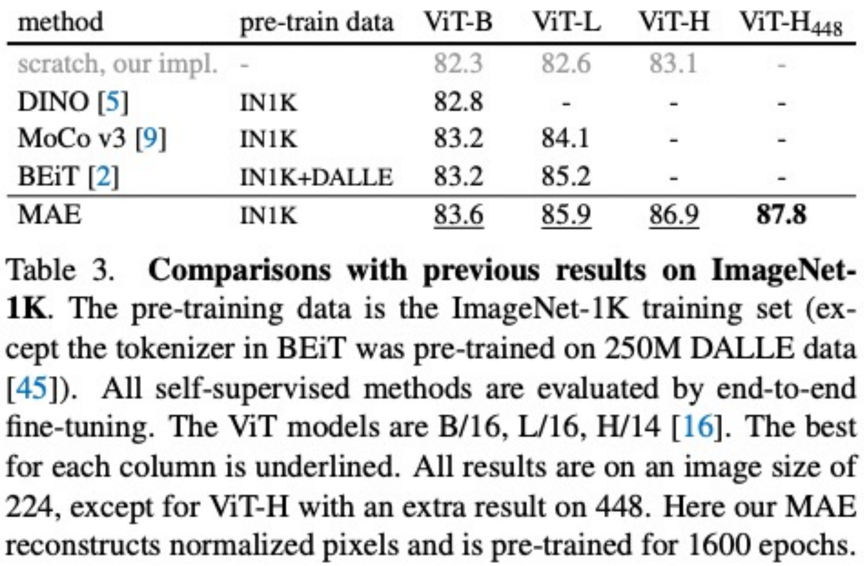
Figure 7. **Training schedules**. A longer training schedule gives a noticeable improvement. Here each point is a full training schedule. The model is ViT-L with the default setting in Table 1.

- **Masking Sampling Strategy (f)**
  - 패치 별로 random / block 형태 / 일정한 grid 형태
  - Block 은 50% 까지는 어느 정도의 성능을 보이고, grid 도 어느 정도를 보이나,
  - Random 이 가장 성능이 좋다.

- **Training schedules (Fig. 7)**
  - Fine-tuning, linear probing 모두 accuracy 가 steadily 증가함. (with longer training)
  - Indeed, saturation of linear probing have not been observed even at 1600 epochs.

4. Experiments

- **Comparisons with Self-supervised Methods.**



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 [45]). 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.

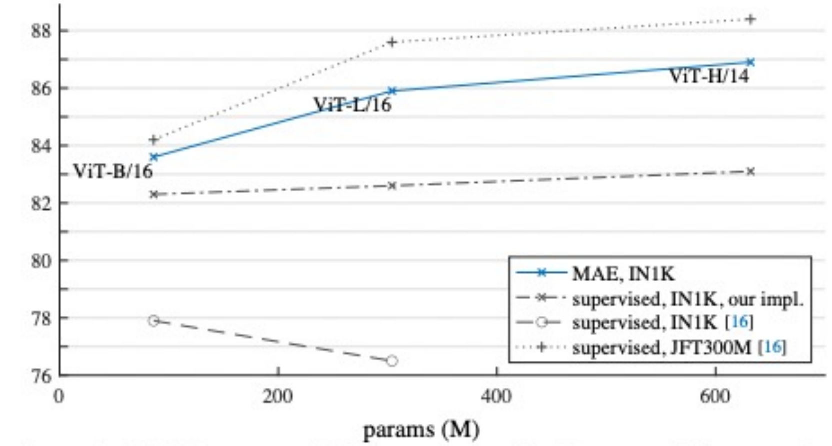
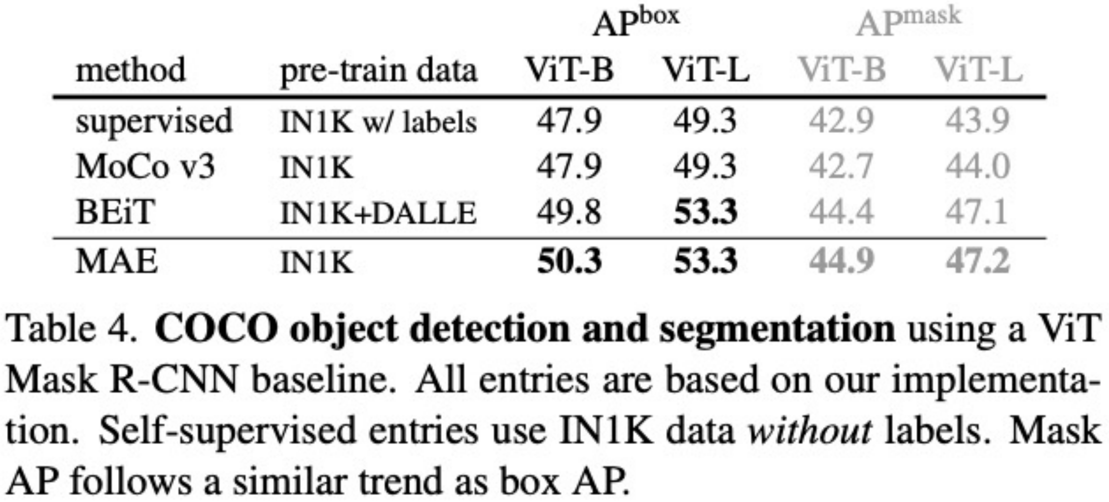


Figure 8. **MAE pre-training *vs*. supervised pre-training**, evaluated by fine-tuning in ImageNet-1K (224 size). We compare with the original ViT results [16] trained in IN1K or JFT300M.

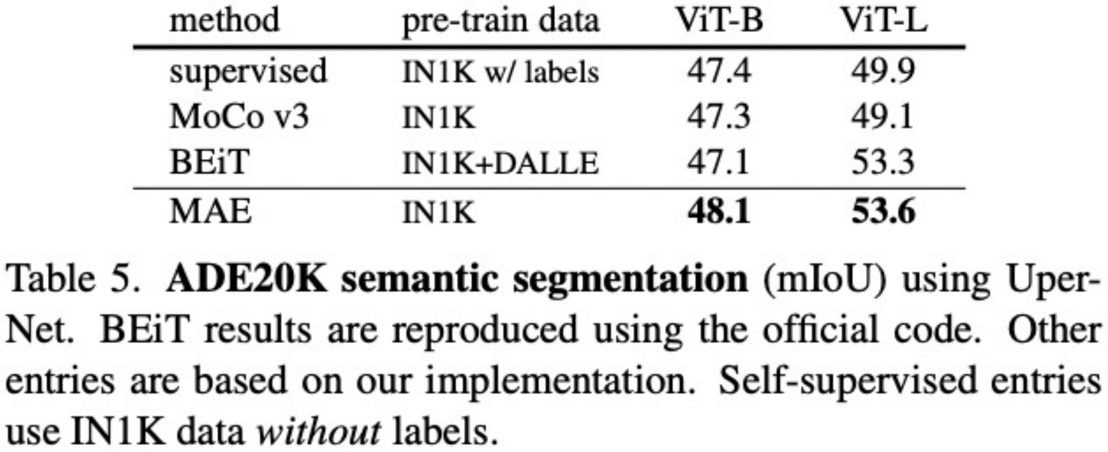
4. Experiments

- **Transfer Learning**



		AP$^{\text{box}}$		AP$^{\text{mask}}$	
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.

- **Object detection and instance segmentation**
  - Mask R-CNN is fine-tuned on COCO.
  - The ViT backbone is adapted for use with FPN.

- **Semantic segmentation**
  - Experiments on ADE20K use UperNet following the code in BEiT.

5. Discussion and Conclusion

- Simple algorithms that scale well are the core of deep learning

- In NLP, simple self-supervised learning methods enable benefits from exponentially scaling models. In computer vision, practical pre-training paradigms are dominantly supervised despite progress in self-supervised learning.

- **Self supervised learning in vision may now be embarking on a similar trajectory as in NLP.**

- 반면에**, 이미지와 language 는 다른 특성을 가진다. 이미지는 빛을 기록한 것이라서 semantic 한 정보를 가지지 않음.**
- **Random 하게 패치를 잘라서 마스킹 한 후, 그걸 reconstruction 해도 semantics 를 잘 학습한다.**

- **The authors hypothesize that this behavior occurs by way of a rich hidden representation inside the MAE.**

감사합니다.