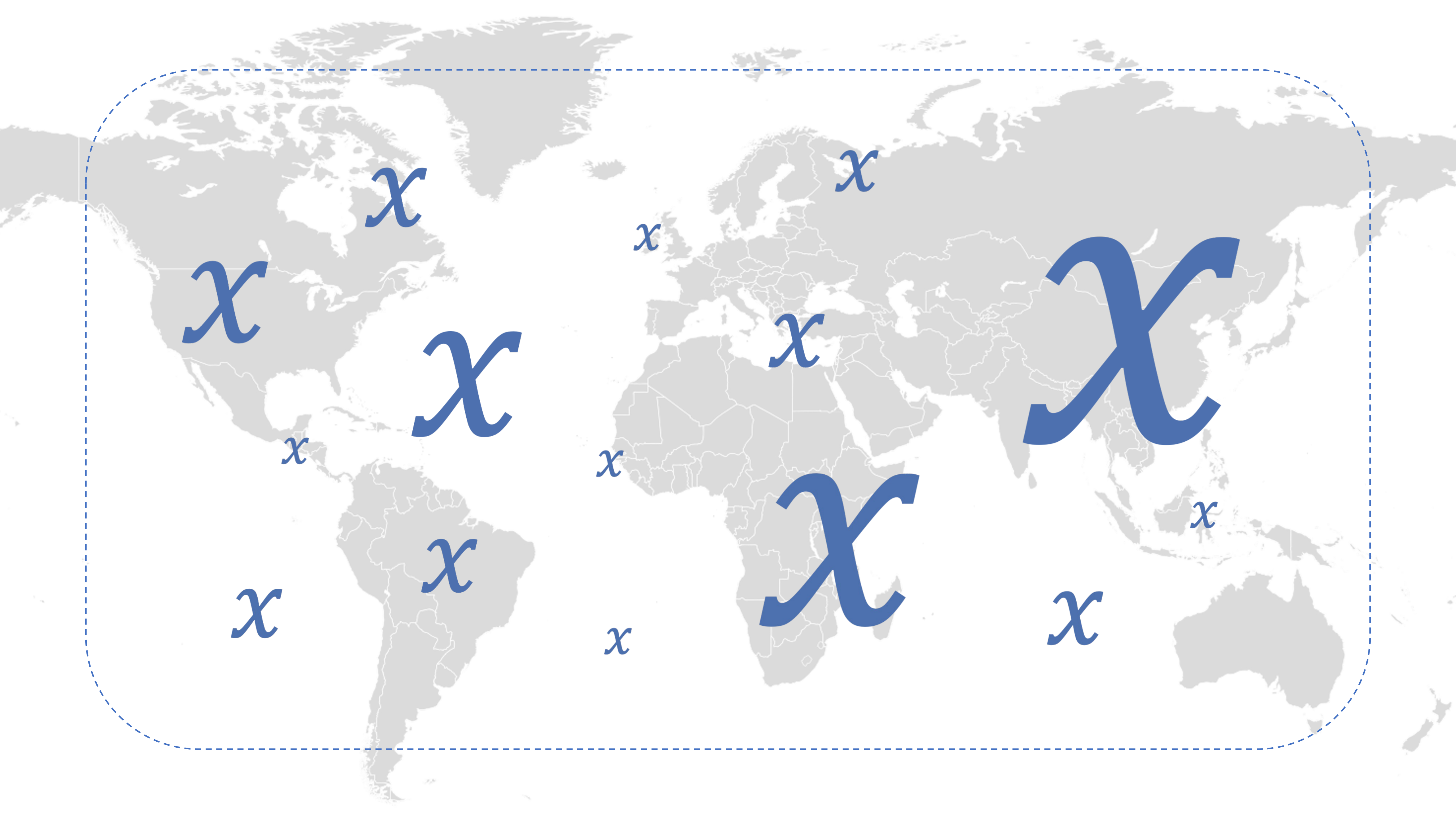
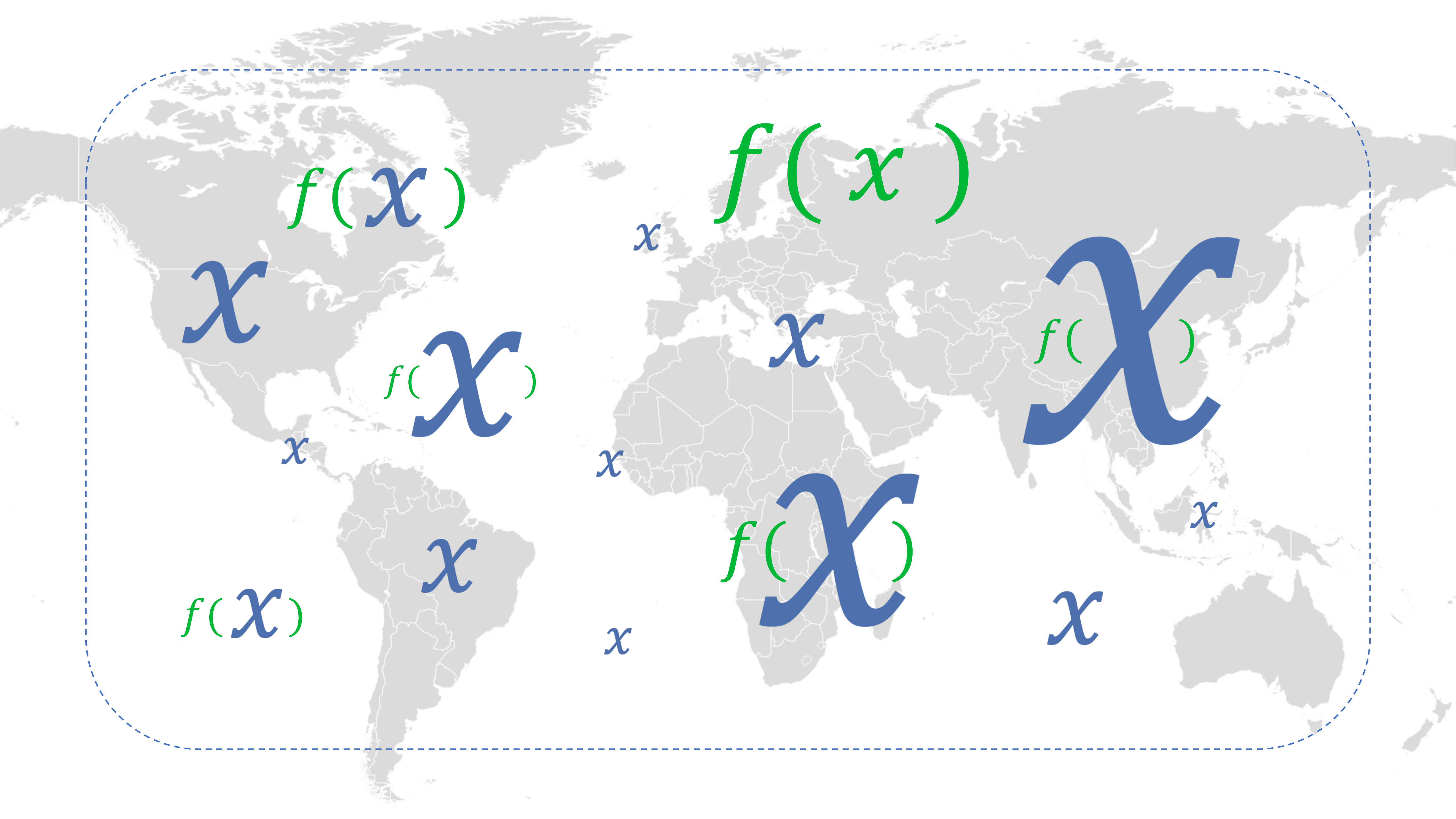
Masked Auto-Encoders as Scalable Vision Learners

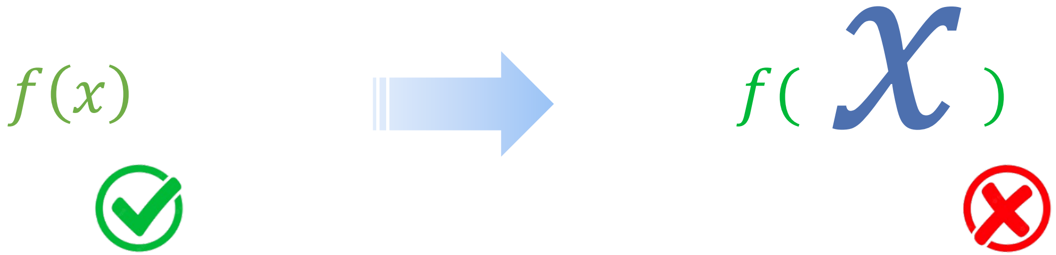
Xinlei Chen

ECCV 2022 tutorial on self-supervised representation learning in computer vision

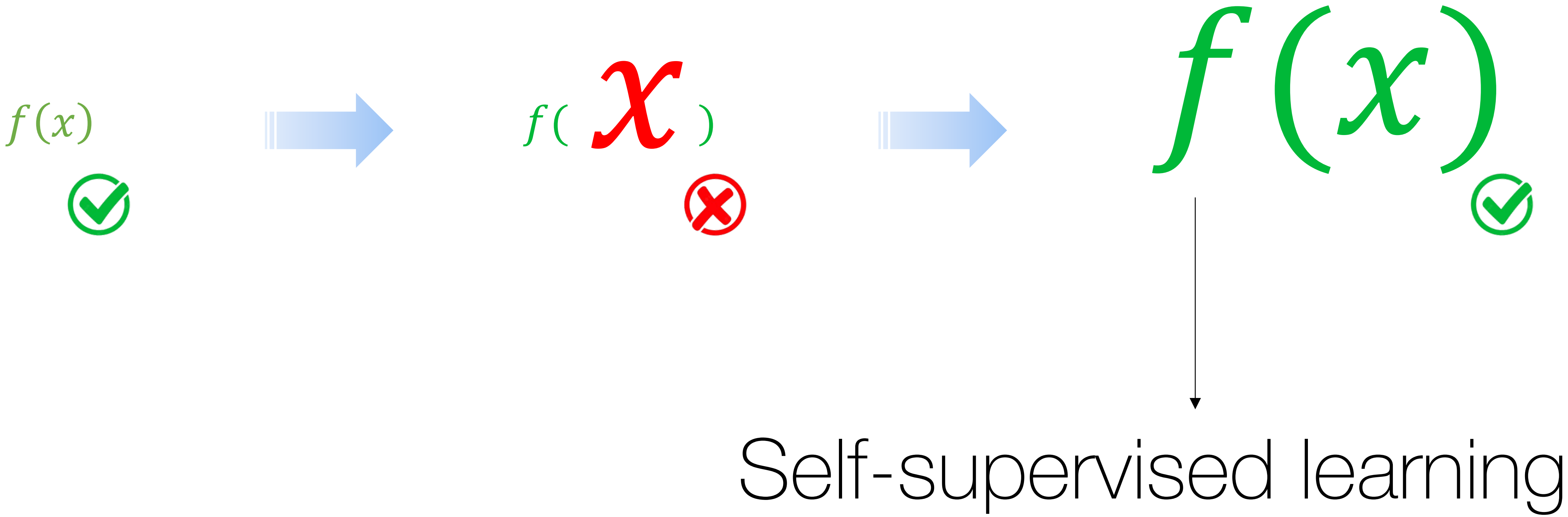

x
x
x
x
x
x
x
x
x
x
x
x
x
x
x



f(x)
x
f(x)
x
f(x)
x
f(x)
x
x
x
x
f(x)
f(x)
x
x


$f(x)$
$f(X)$

$f(x)$
$f(x)$
$f(x)$


f(x)
f(x)
f(x)
Self-supervised learning

Self-Supervised Learning

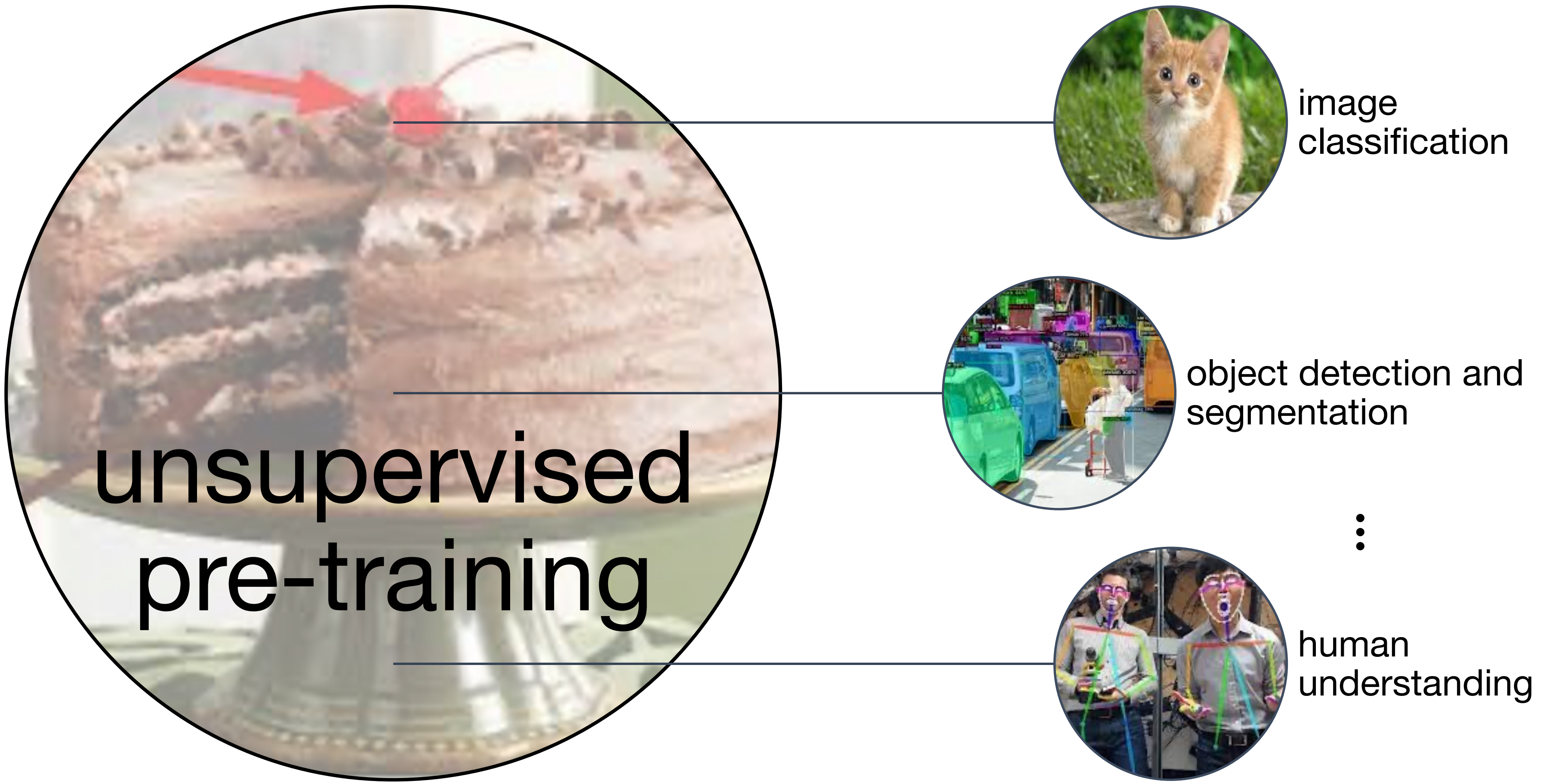
- Pre-train representations <u>without labels</u> for downstream tasks

Self-Supervised Learning

- Pre-train representations without labels for downstream tasks

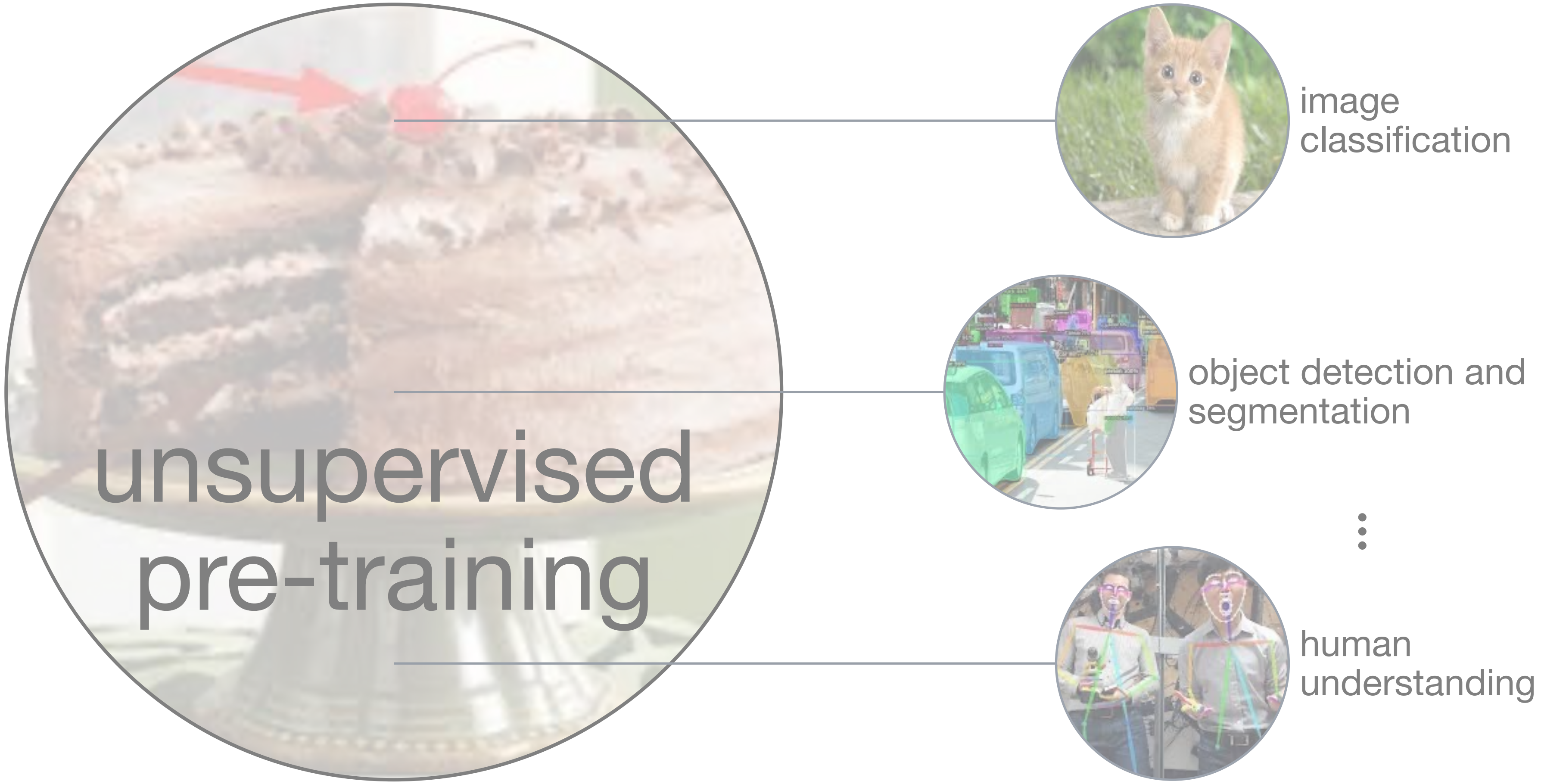
Self-Supervised Learning

- Pre-train representations without labels for downstream tasks



Self-Supervised Representation Learning

- Pre-train representations without labels for downstream tasks



Self-Supervised Representation Learning

- Scalable: use unlimited data to train unlimited-sized models

[Devlin et al, NAACL 2019] [Brown et al, NeurIPS 2020]

Self-Supervised Representation Learning

- Scalable: use unlimited data to train unlimited-sized models
- Tremendously successful in NLP

Language

[Devlin et al, NAACL 2019] [Brown et al, NeurIPS 2020]

Self-Supervised Representation Learning

- Scalable: use unlimited data to train unlimited-sized models
- Tremendously successful in NLP

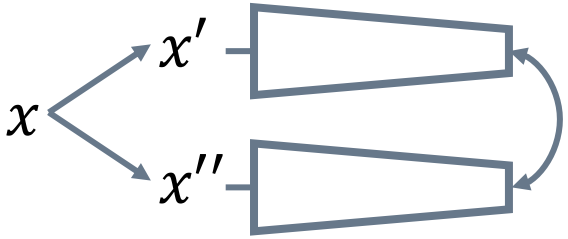
Language

Vision

[Chen et al, ICML 2020] [He et al, CVPR 2022]

Self-Supervised Paradigms Covered

- Contrastive / Siamese



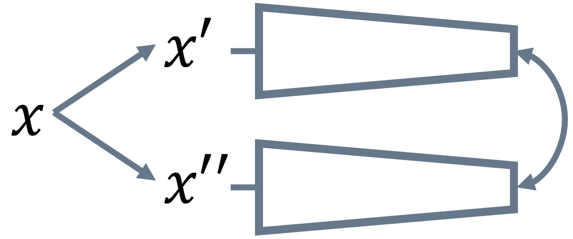
→ Tutorial from Ting Chen

SimCLR 1st author, Google

5:30 pm – 6:15pm

Self-Supervised Paradigms Covered

- Contrastive / Siamese

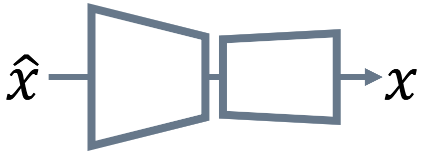


→ Tutorial from Ting Chen

SimCLR 1st author, Google

5:30 pm – 6:15pm

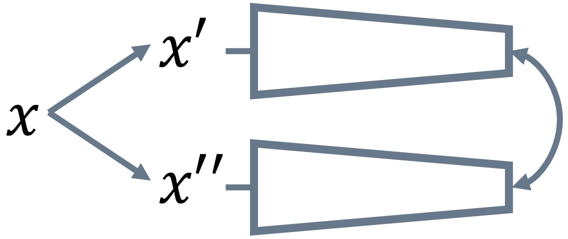
- Reconstructive / Auto-Encoding



[Chen et al, ICML 2020] [He et al, CVPR 2022]

Self-Supervised Paradigms Covered

- Contrastive / Siamese

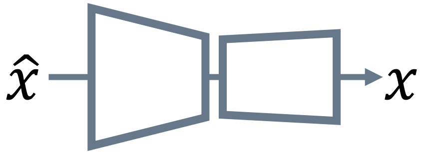


→ Tutorial from Ting Chen

SimCLR 1st author, Google

5:30 pm – 6:15pm

- Reconstructive / Auto-Encoding



Masked **A**uto-**E**ncoders Are Scalable Vision Learners:

Kaiming, Xinlei, Saining, Yanghao, Piotr, Ross

CVPR 2022

[He et al, CVPR 2022]

What is MAE?

- Very simple method, but highly effective

[Devlin et al, NAACL 2019] [He et al, CVPR 2022]

What is MAE?

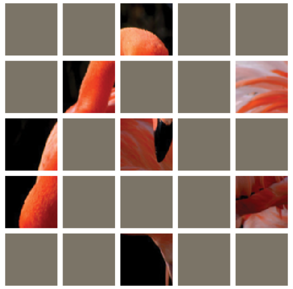
- Very simple method, but highly effective
- BERT-like algorithm, but with crucial design changes for vision

[Devlin et al, NAACL 2019] [He et al, CVPR 2022]

What is MAE?

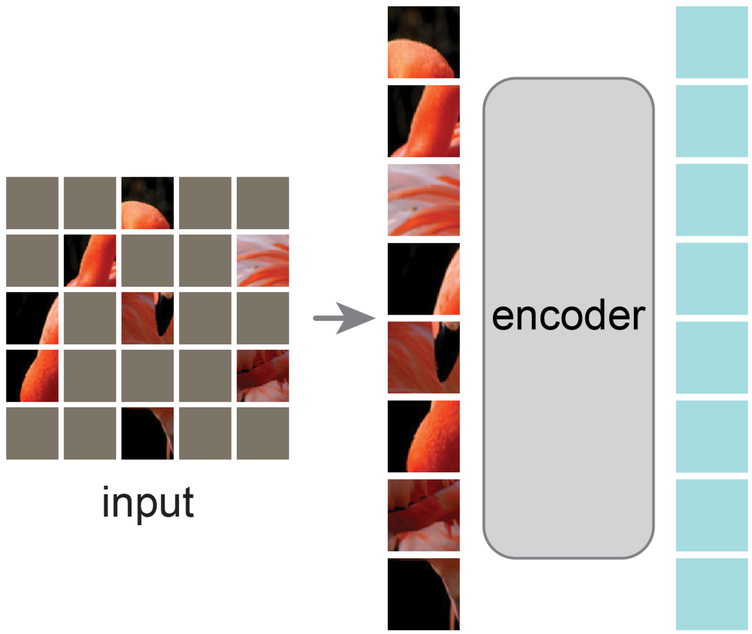
- Very simple method, but highly effective
- BERT-like algorithm, but with crucial design changes for vision
- Intriguing properties – better scalability and more from analysis

How MAE Works?



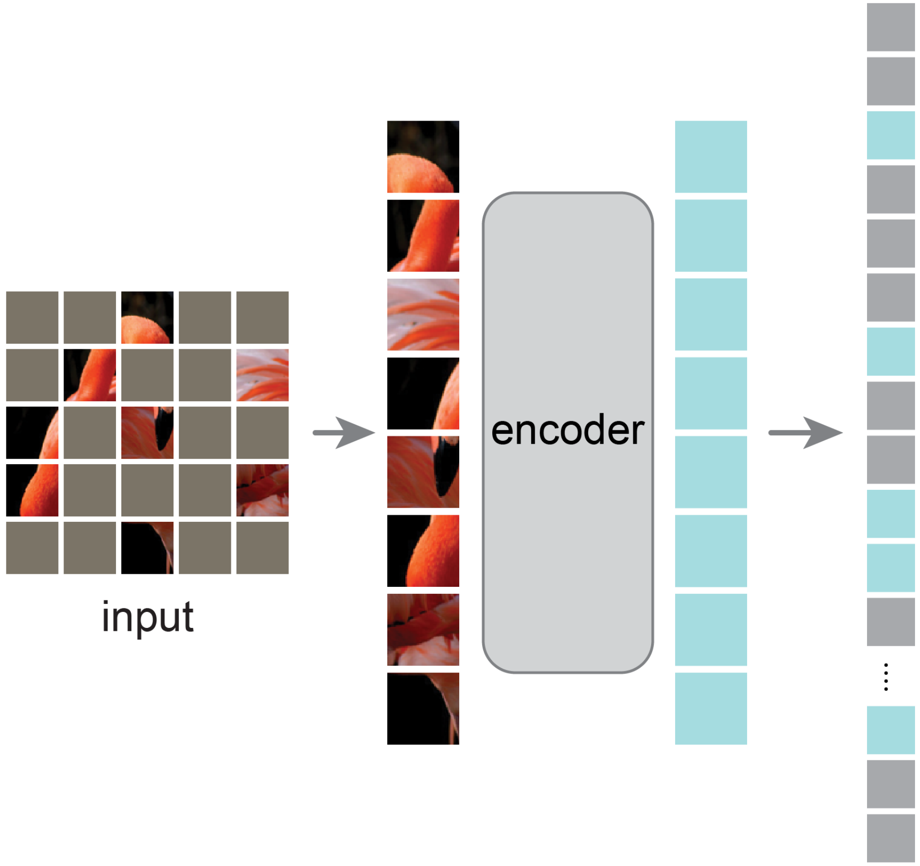
Random masking

How MAE Works?



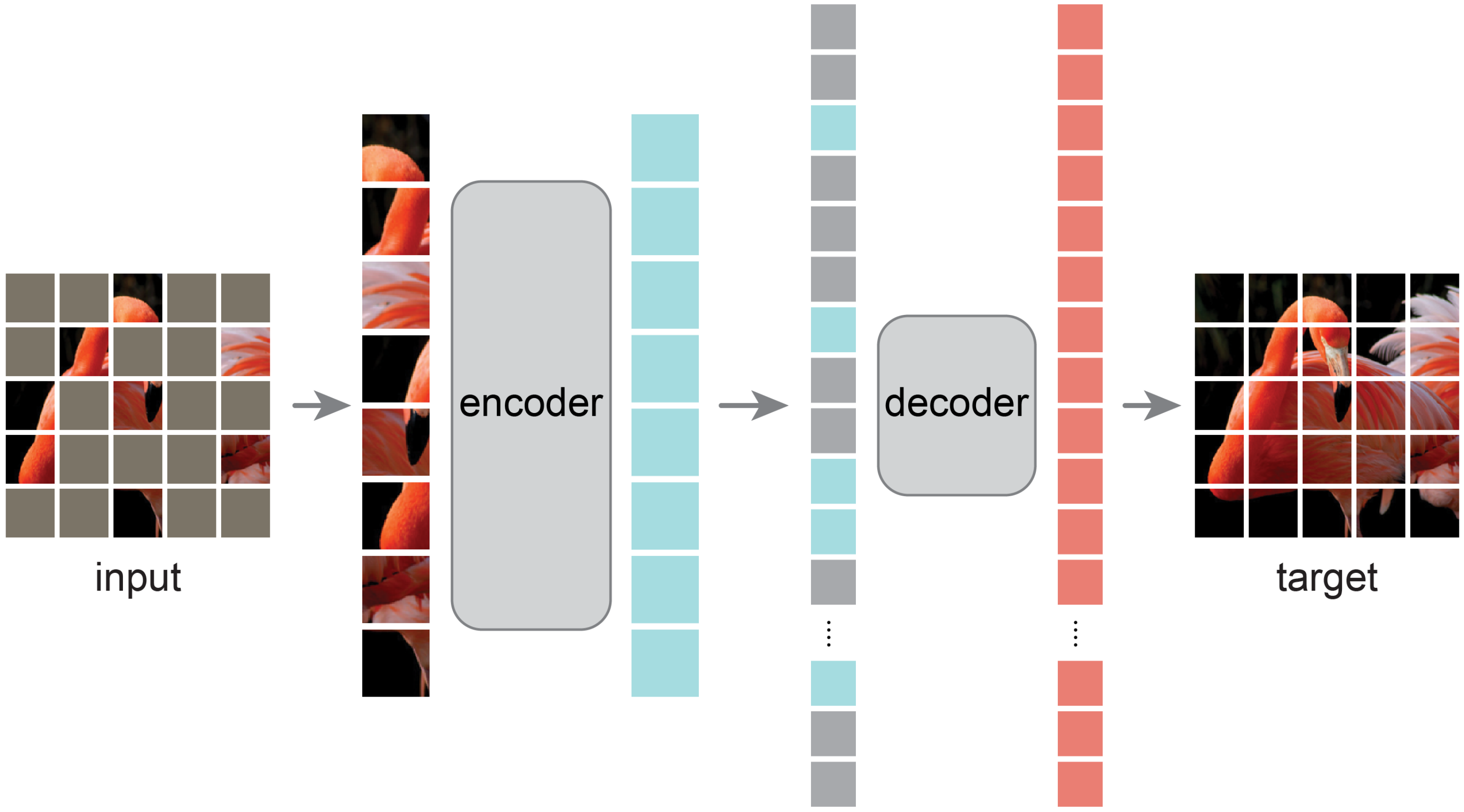
Encode visible patches

How MAE Works?



Add mask tokens

How MAE Works?



Reconstruct

MAE Reconstruction Example

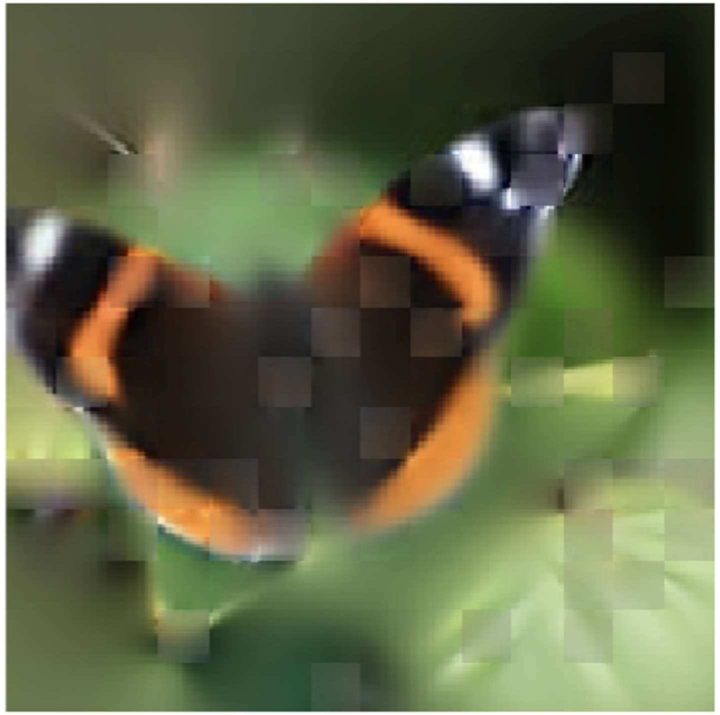
?

Masked input: 80%

You guess?

MAE Reconstruction Example

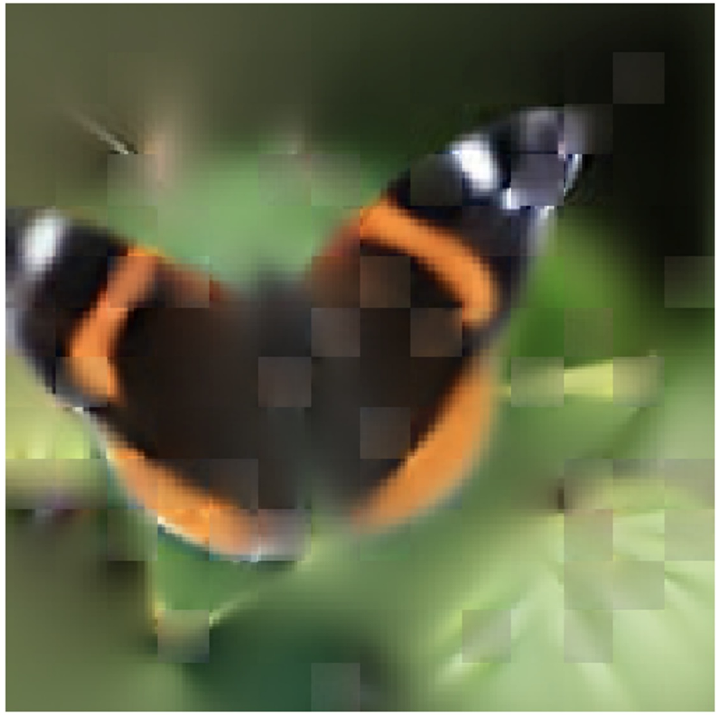
Masked input: 80%



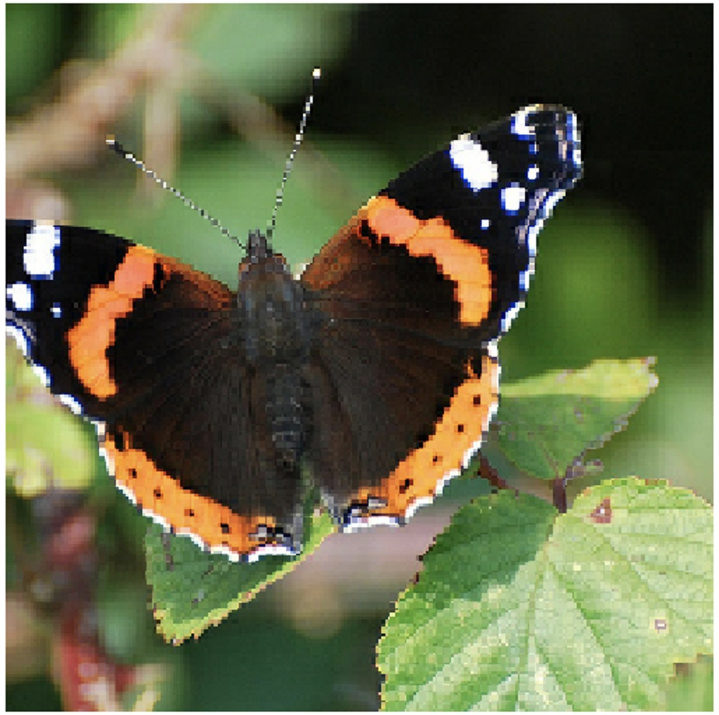
MAE’s guess

MAE Reconstruction Example

Masked input: 80%

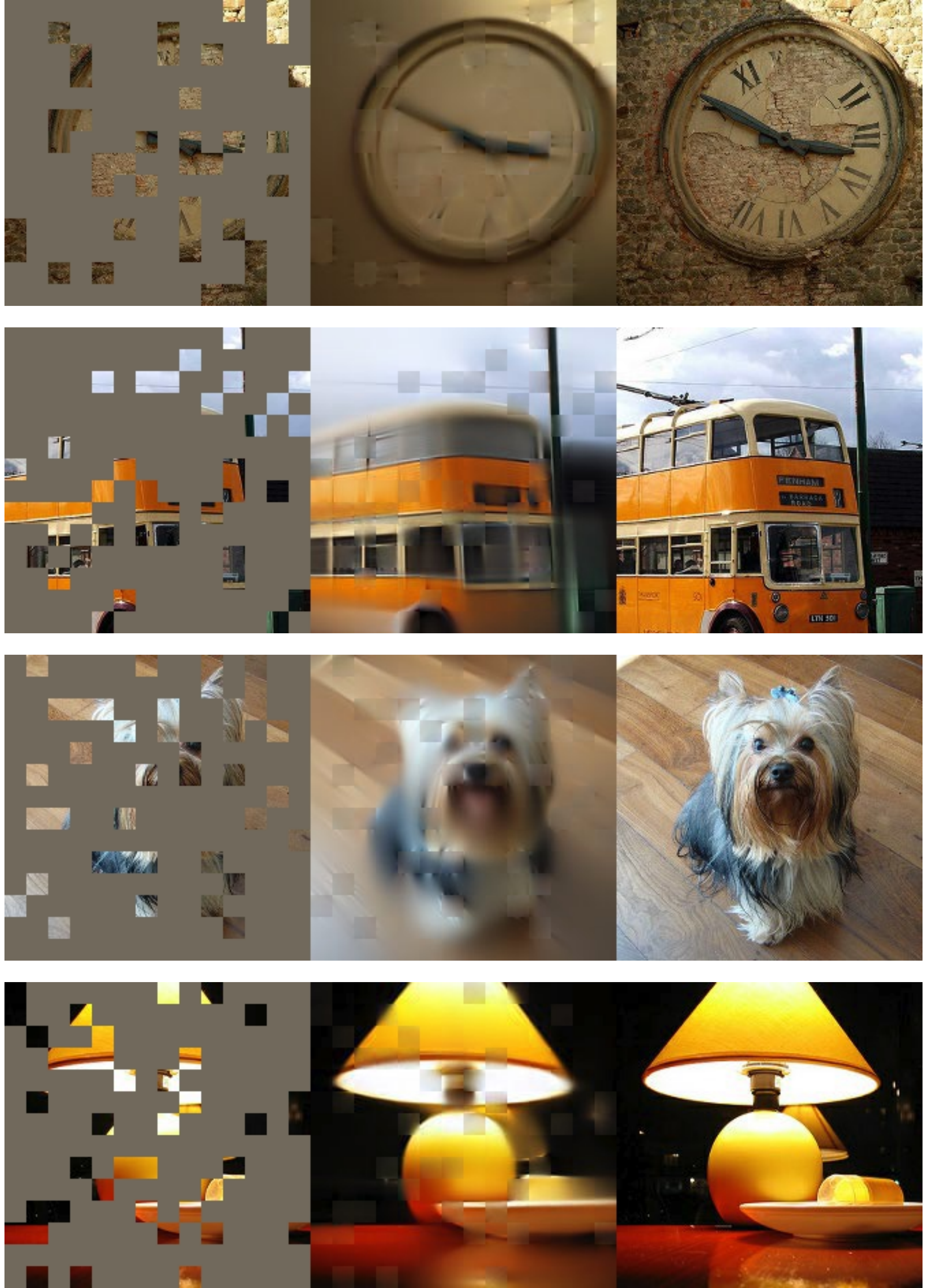


MAE's guess

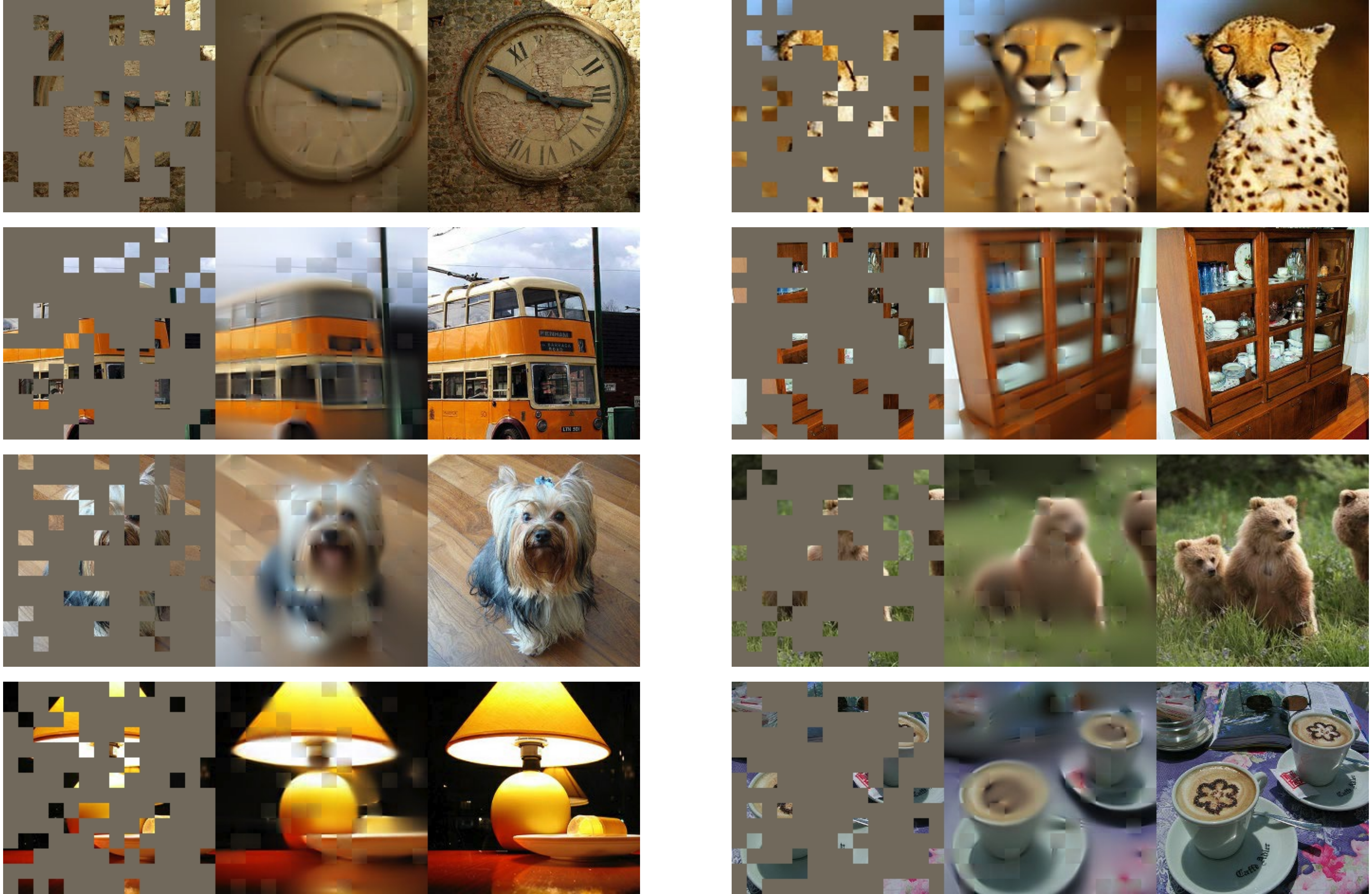


Ground truth

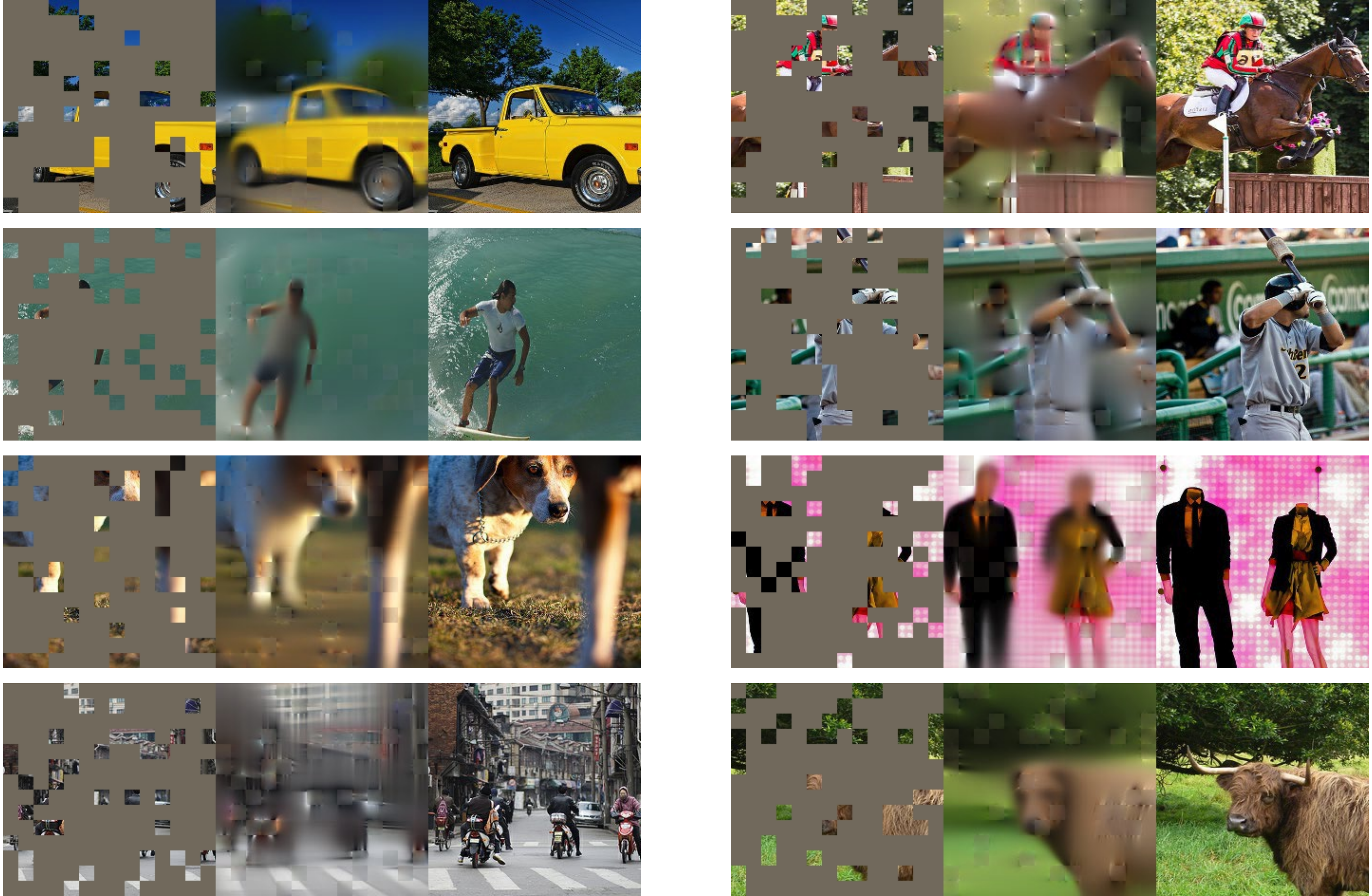
ImageNet val set (unseen)



ImageNet val set (unseen)



COCO val set (unseen)



75% mask

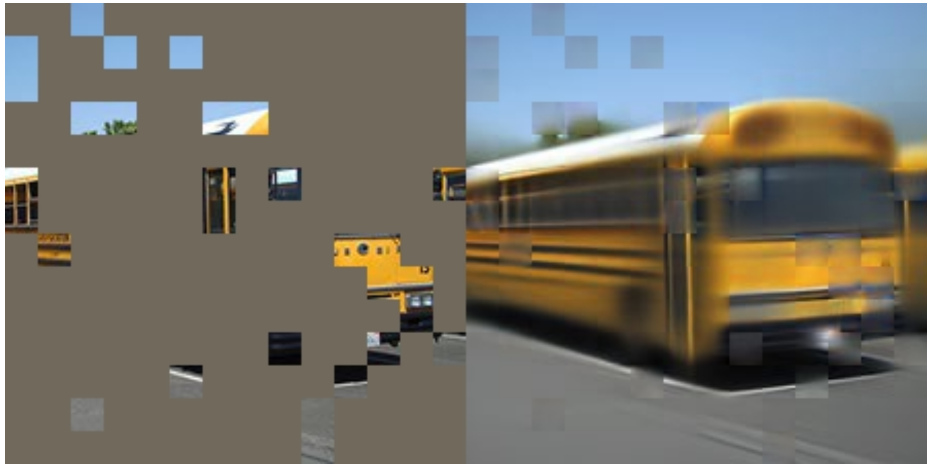
original

MAE Can Generalize

MAE Can Generalize

original

75% mask

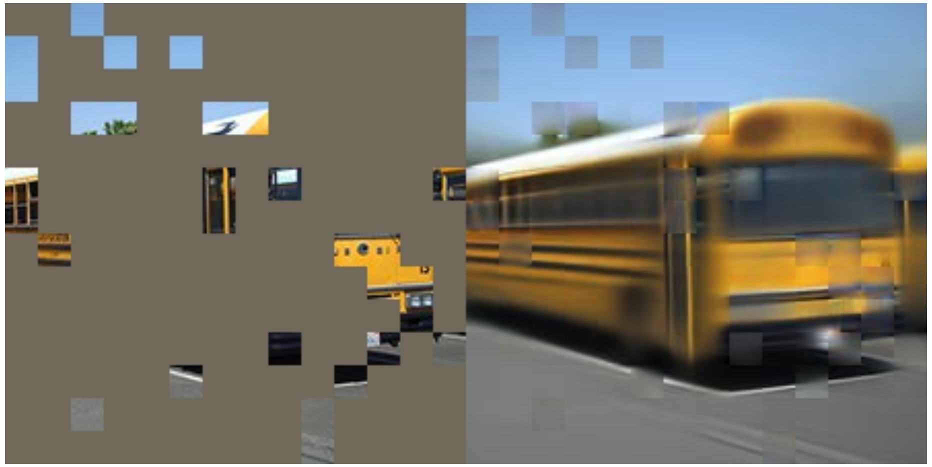


85% mask

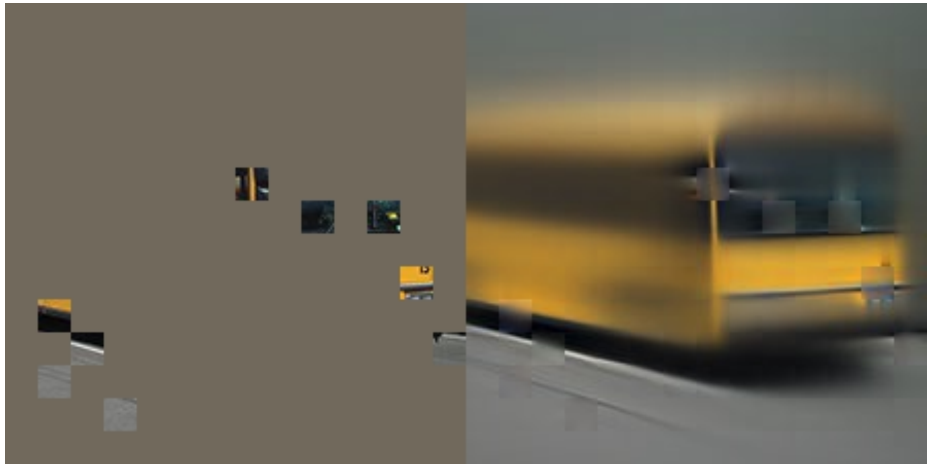
MAE Can Generalize

original

75% mask



85% mask



95% mask

original

75% mask

MAE Can Generalize

original

75% mask

85% mask

MAE Can Generalize

MAE Can Generalize

original

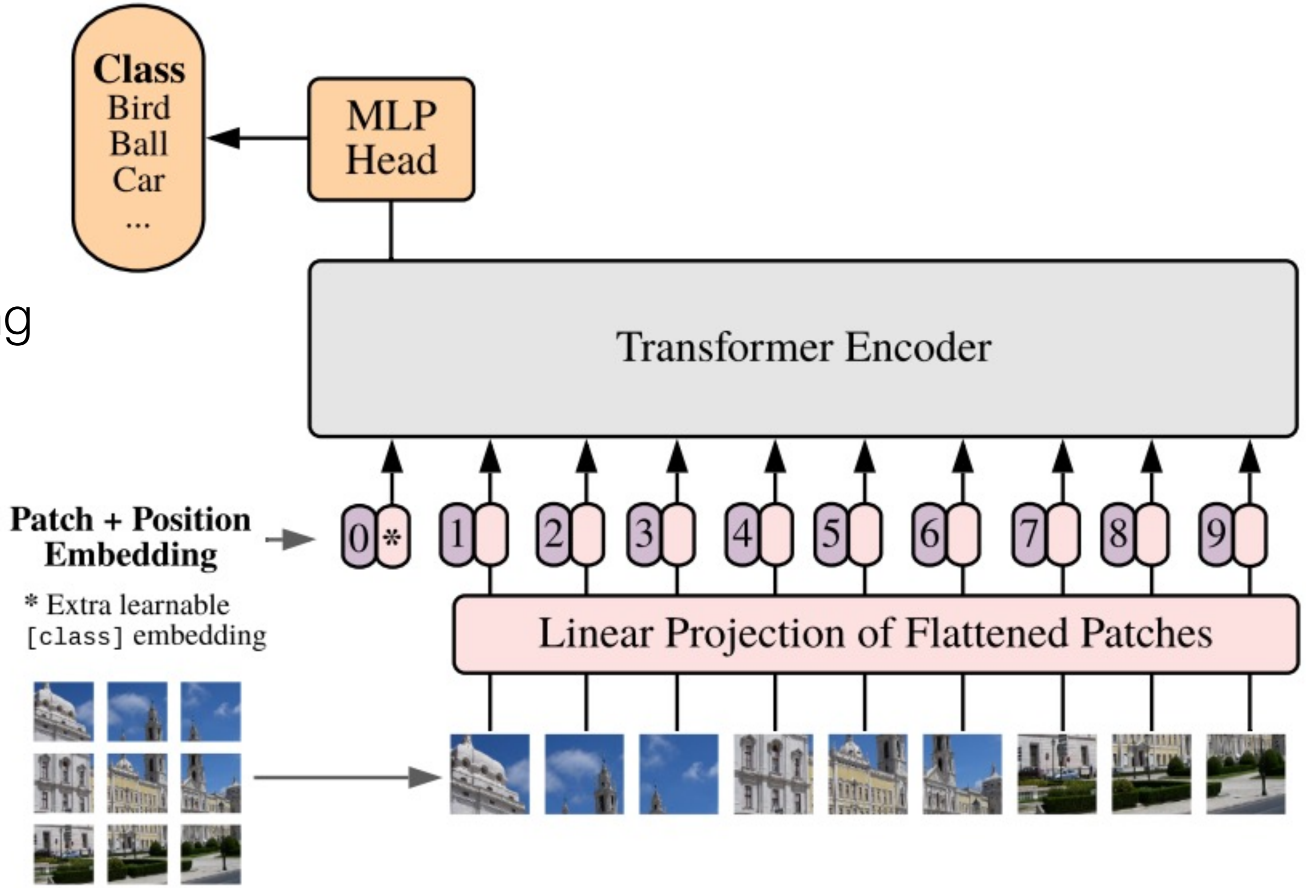
75% mask

85% mask

95% mask

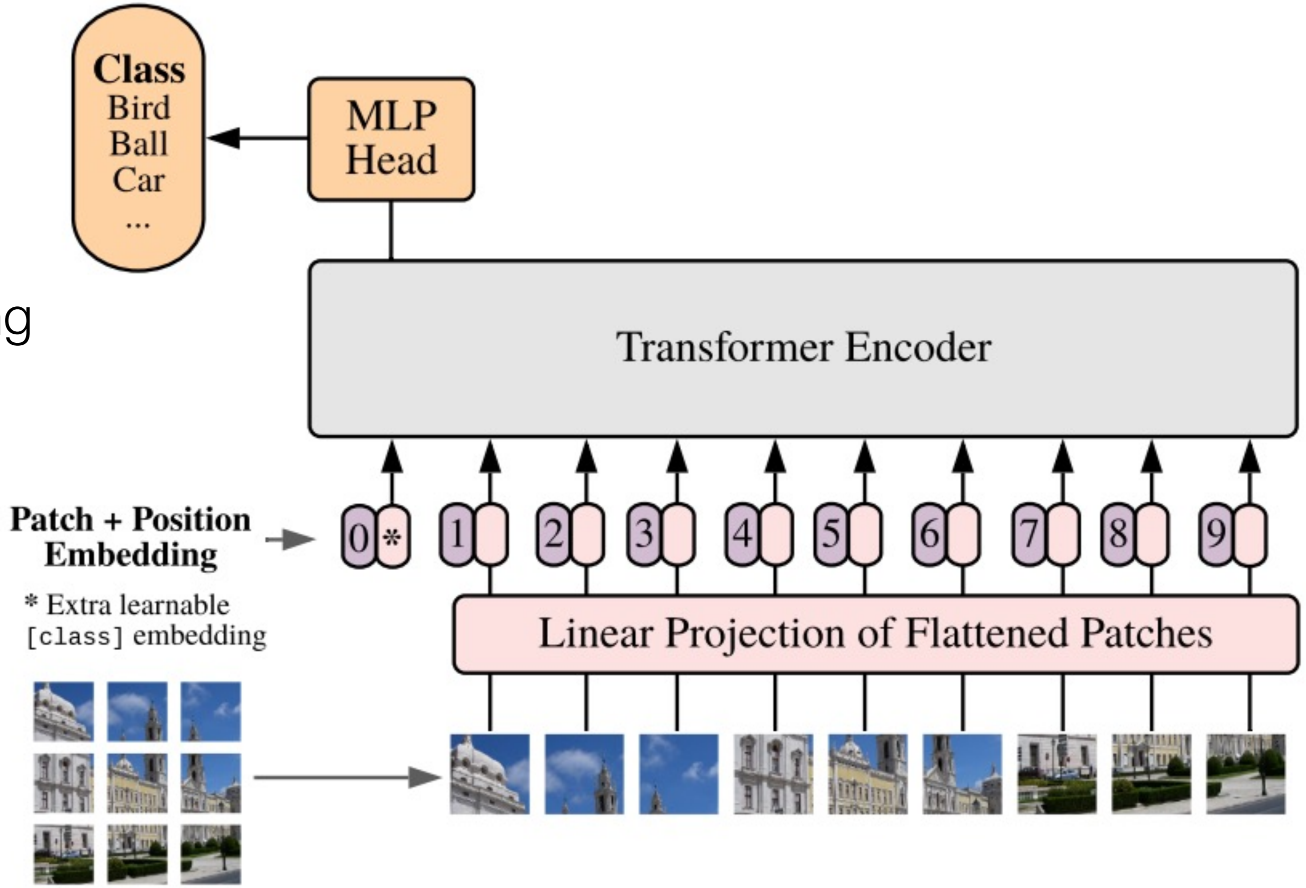
BERT-like: Transformers

- Vision Transformer (ViT)
  - Less inductive bias
  - Non-overlapping tokenization
    - Easier for masked auto-encoding



BERT-like: Transformers

- Vision Transformer (ViT)
  - Less inductive bias
  - Non-overlapping tokenization
    - Easier for masked auto-encoding
- *Scalable*
  - with larger models
  - on larger datasets

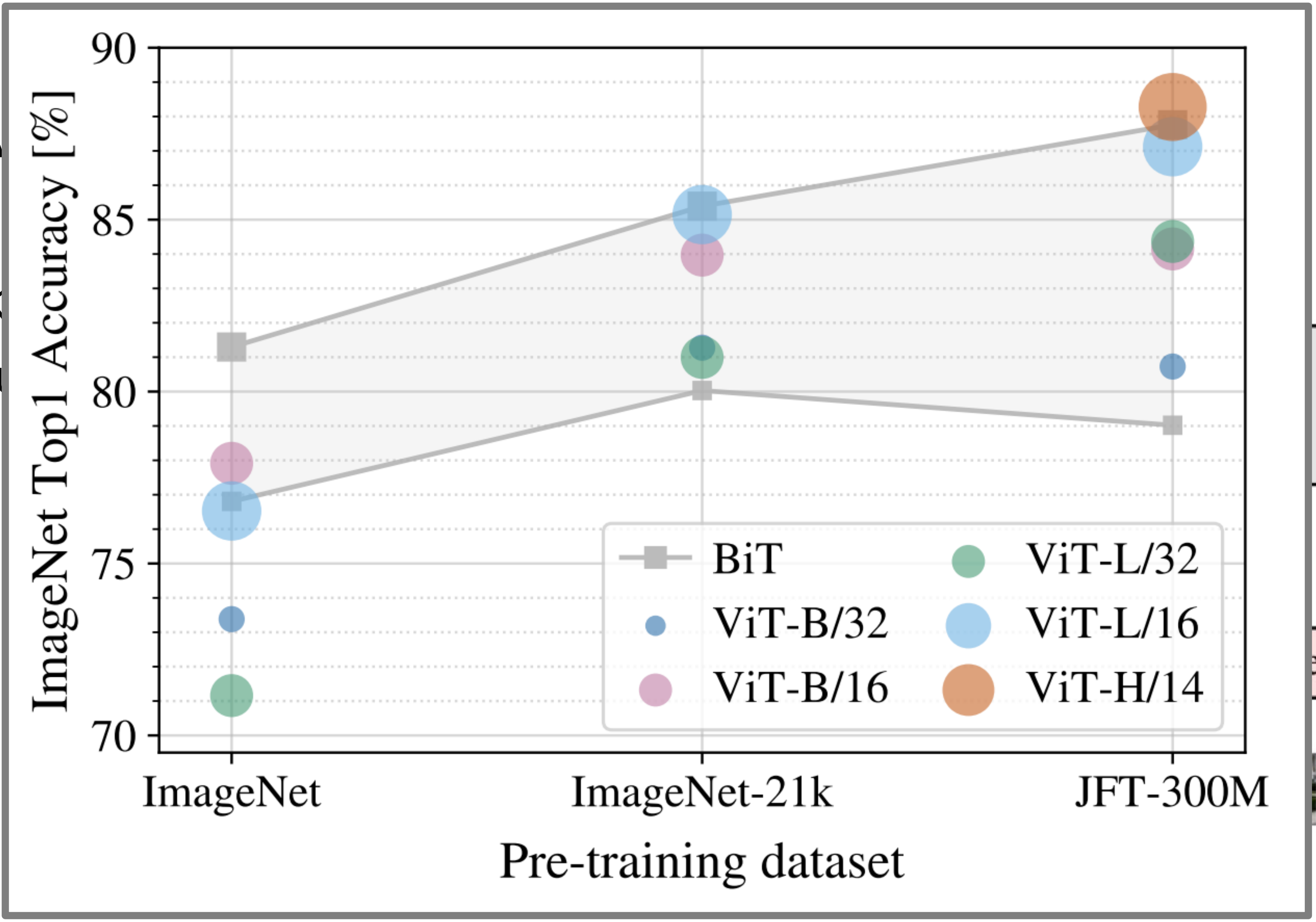


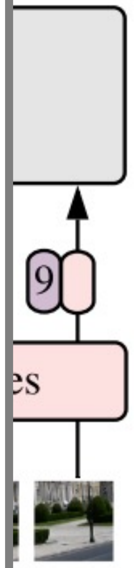
[Dosovitskiy et al, ICLR 2021]

BERT-like: Transformers

- Vision Transformer (V
  - Less inductive bias
  - Non-overlapping tok
    - Easier for masked a

- *Scalable*
  - with larger models
  - on larger datasets





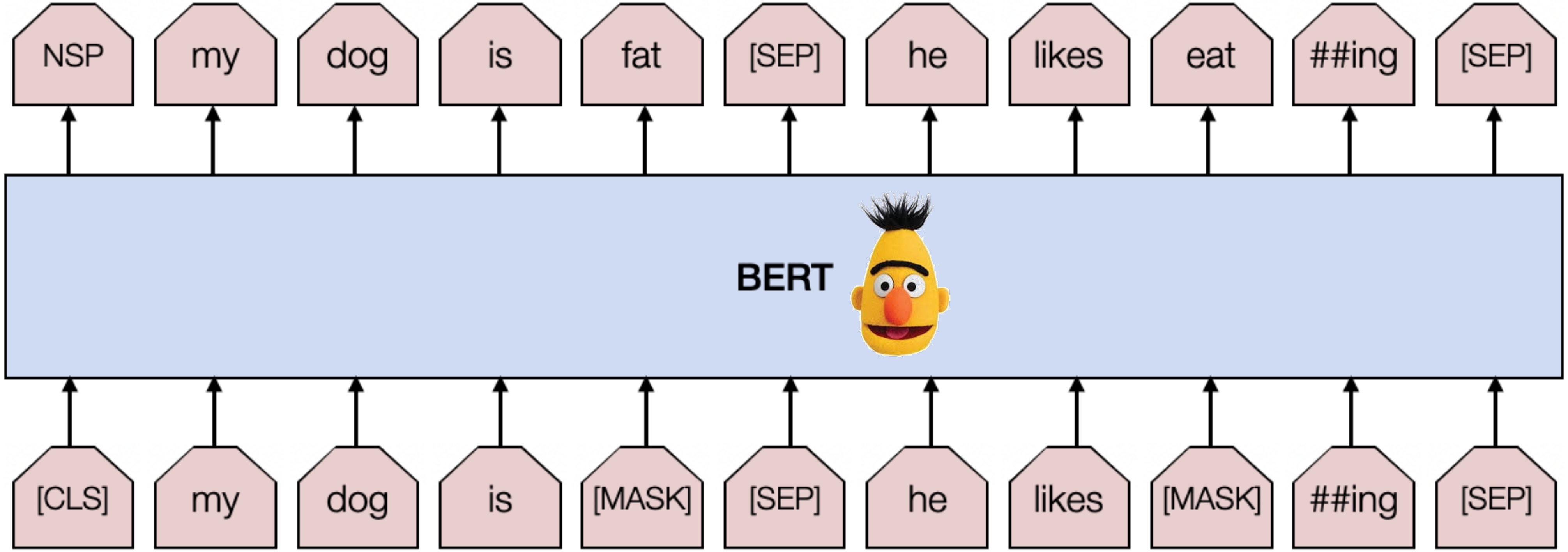
[Devlin et al, NAACL 2019]

BERT-*unlike*: Mask Ratio

- BERT: 15% is enough to create a challenging task
- MAE: a high ratio of 75% - 80% is about optimal

BERT-*unlike*: Encoder-Decoder

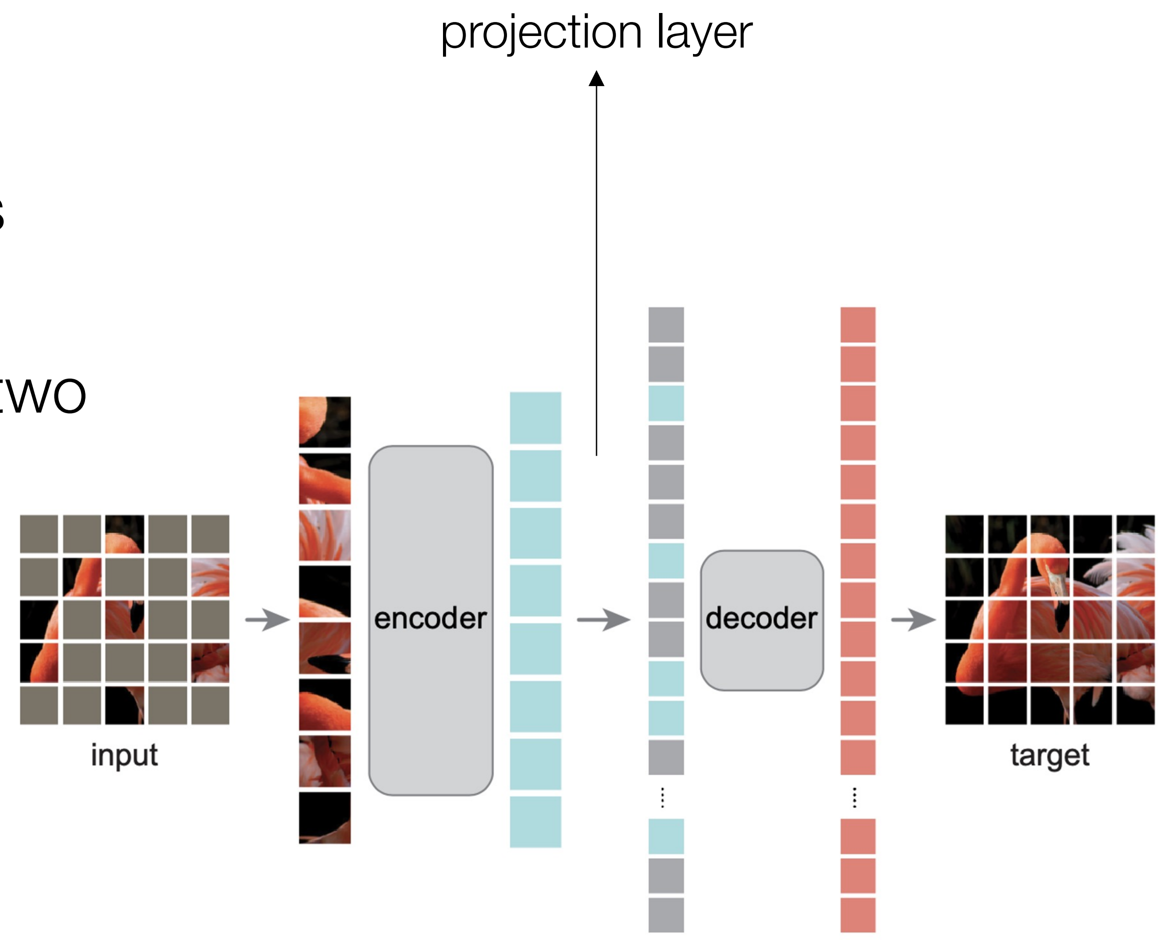
- BERT: encoder-*only* pre-training



[Devlin et al, NAACL 2019]

BERT-*unlike*: Encoder-Decoder

- MAE:
  - *Large* encoder on *visible* tokens
  - Small decoder on *all* tokens
  - *Projection* layer to connect the two

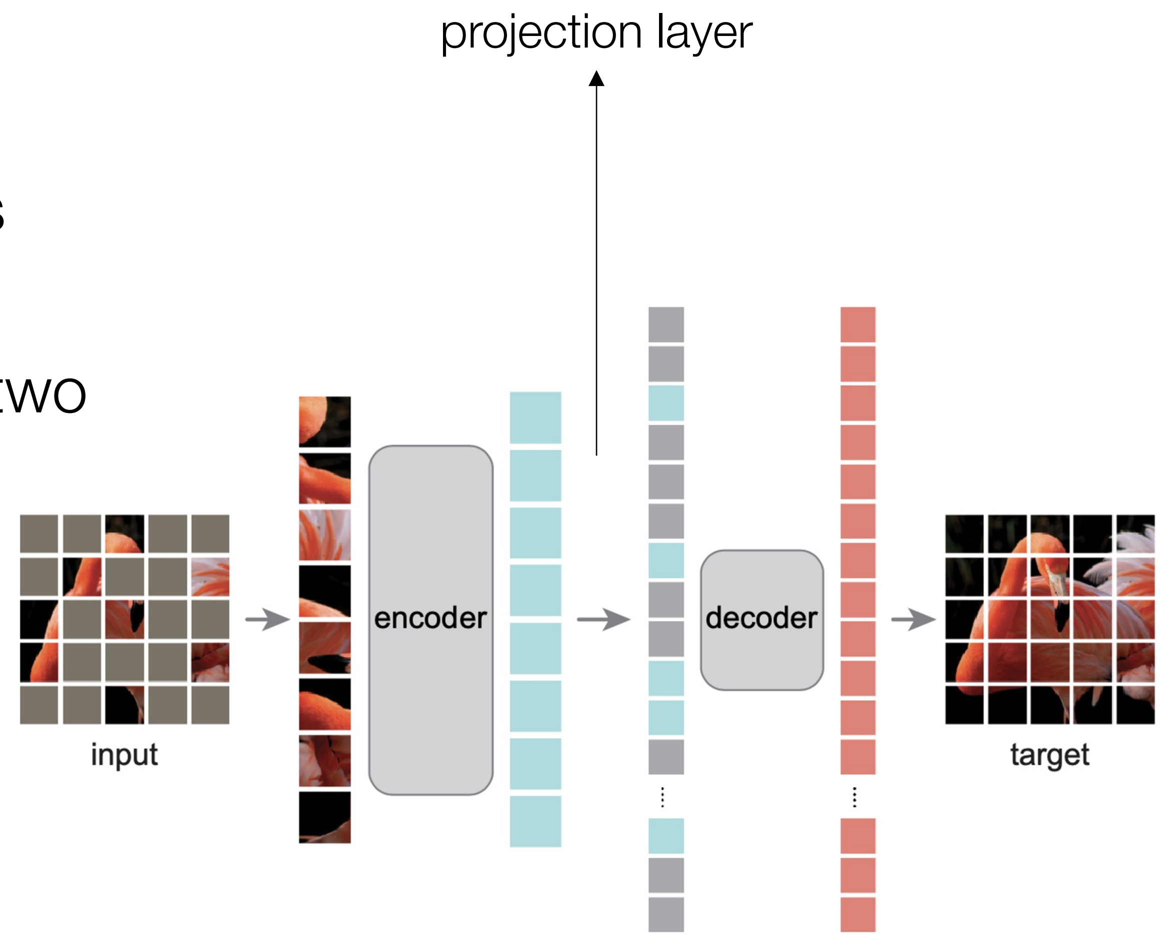


[Devlin et al, NAACL 2019]

BERT-*unlike*: Encoder-Decoder

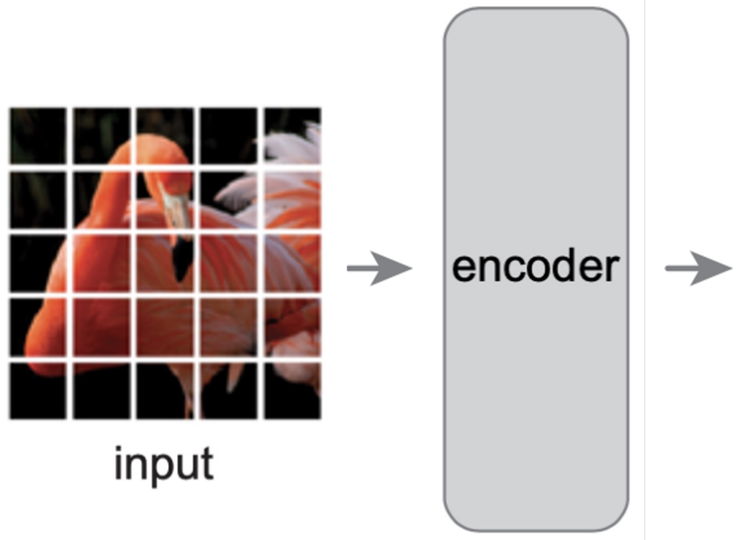
- MAE:
  - *Large* encoder on *visible* tokens
  - Small decoder on *all* tokens
  - *Projection* layer to connect the two

- Very efficient when coupled with <u>high</u> mask ratio (75%)



MAE for Downstream Tasks: *Encoder Only*

- After MAE pre-training, just *throw away* the decoder
- Encoder is used for representations with *full-sequence* input



Experimental Protocols

- Pre-training dataset: ImageNet-1K
- Architecture: ViT-*Large* encoder, 512-dim decoder

Experimental Protocols

- Pre-training dataset: ImageNet-1K
- Architecture: ViT-*Large* encoder, 512-dim decoder
- Transfer task: ImageNet-1K classification
  - "*ft*": end-to-end tuning with MAE as an initialization
  - "*lin*": linear probing, a single classifier on top of frozen encoder features

Analysis: Decoder Size

- Encoder has 24-blocks, 1024-dimensional

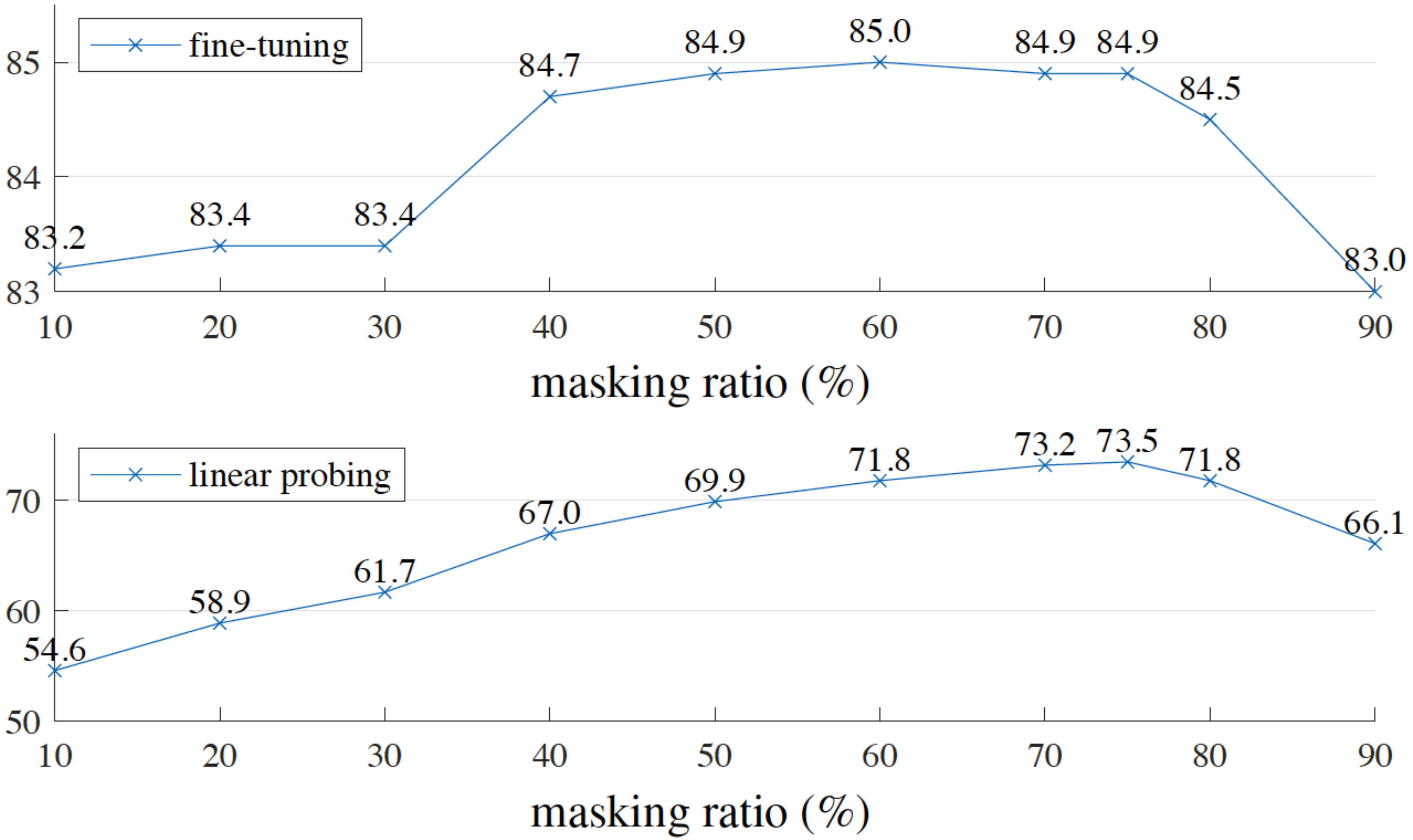
blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

Decoder depth

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

Decoder width

Analysis: Mask Ratio



Analysis: Mask Token [M] in Encoder

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

- Encoder w/ [M] is default in BERT
- Big domain gap for linear probing
  - Pre-train sees 25% of the images only, while evaluation sees 100%

Analysis: Reconstruction Target

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

- Pixels with normalization: per-patch -- minus *mean*, divide by *std*
- PCA: only low-frequency component is retained
- dVAE token: from DALLE, expensive to compute

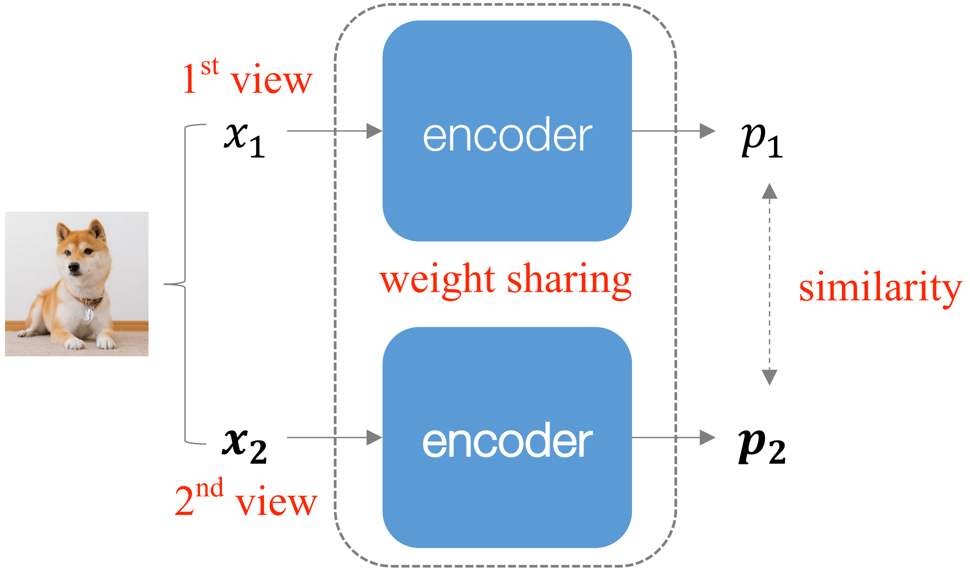
Analysis: Augmentations

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

- MAE can work with minimal data augmentation

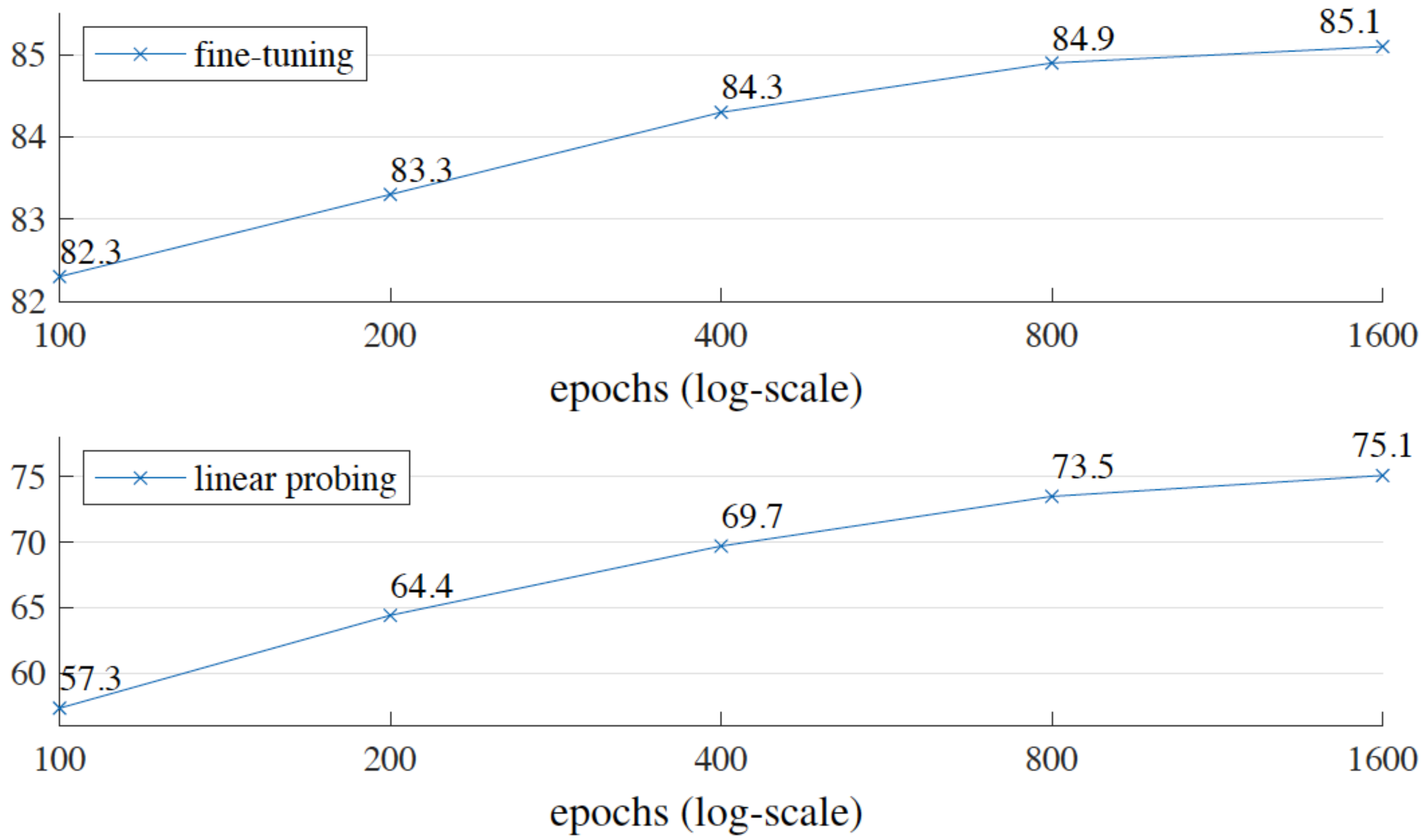
Analysis: Augmentations

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

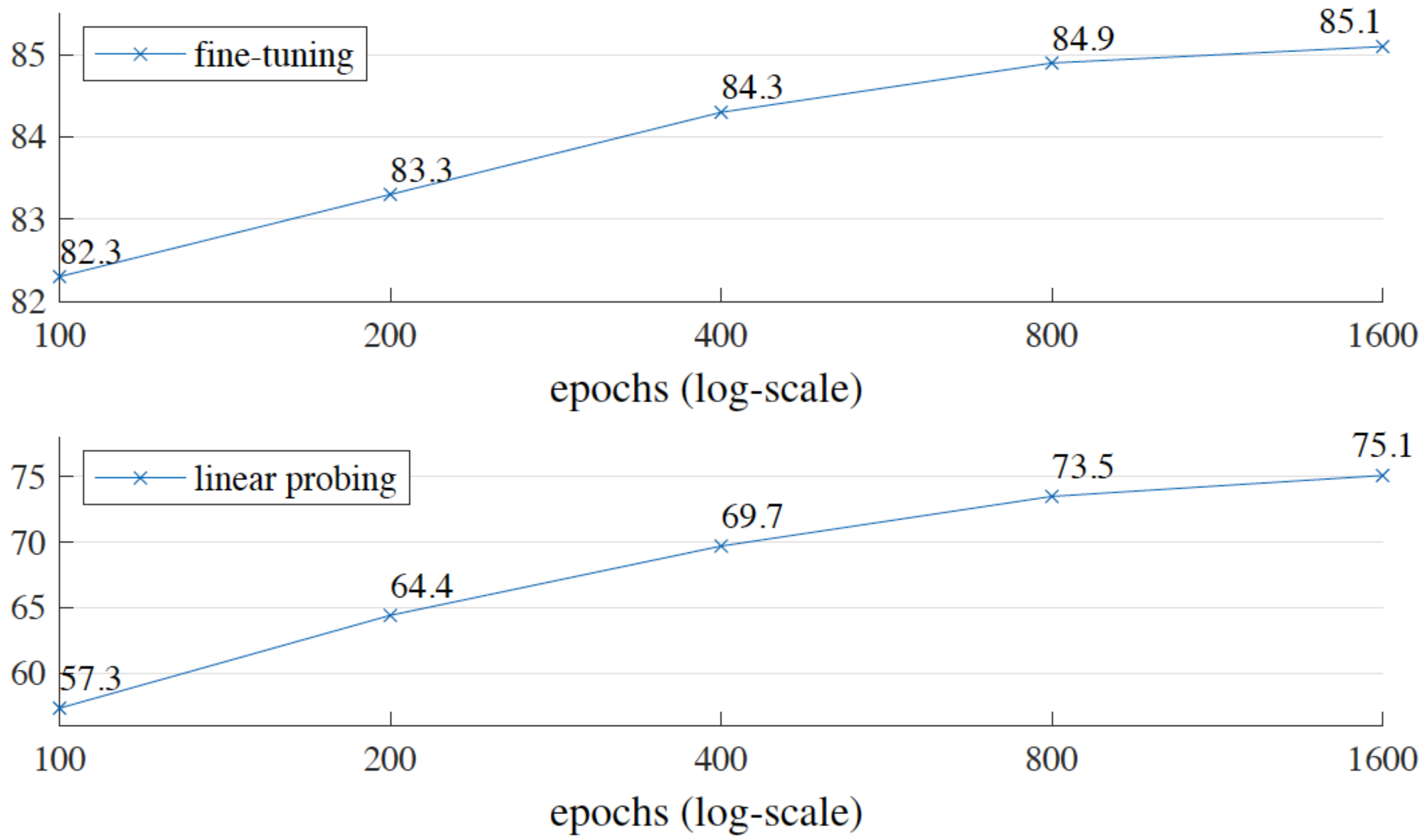


- MAE can work with minimal data augmentation
- For Siamese learning, augmentation is crucial

Scalability: Longer Training

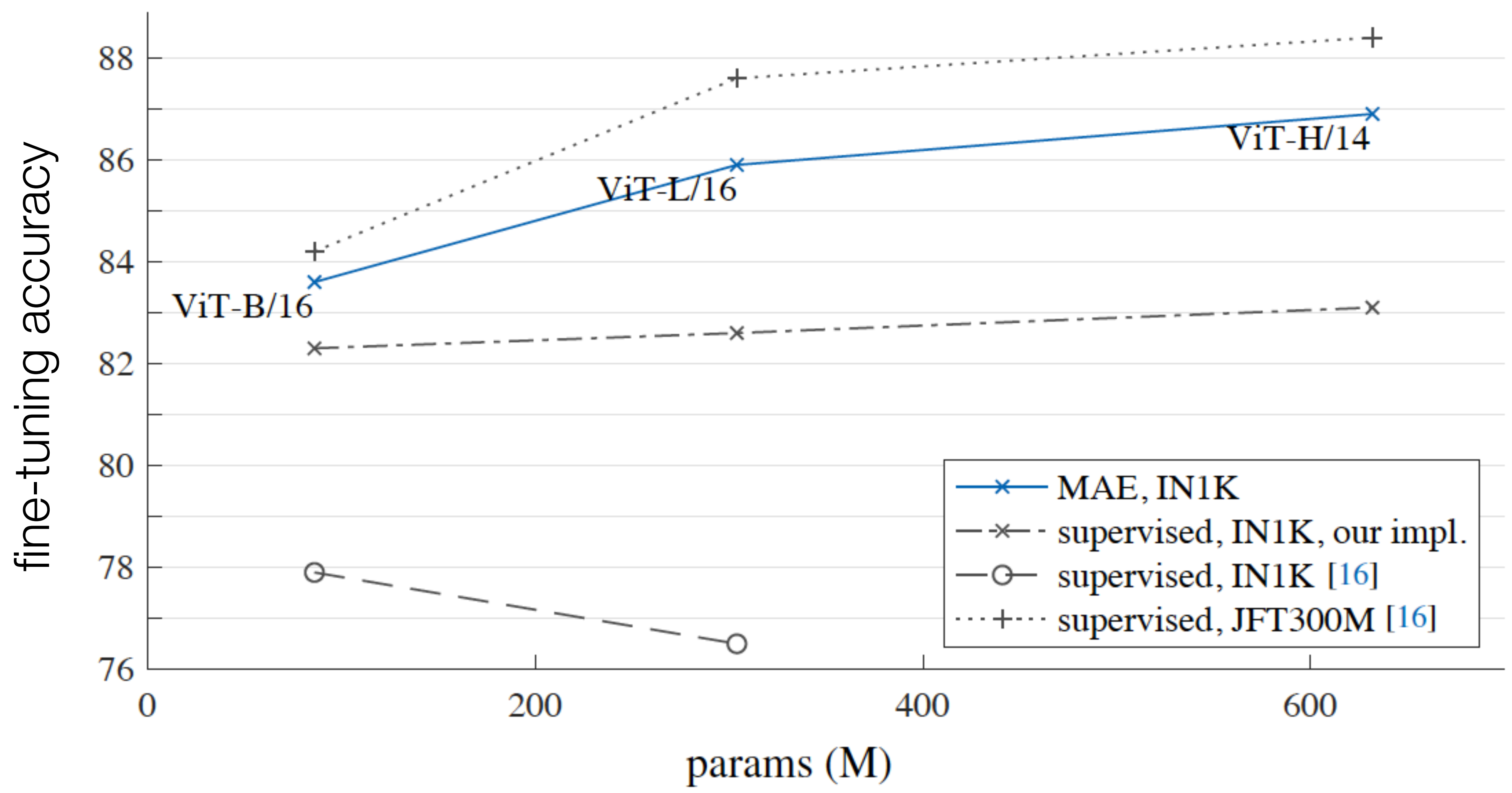


Scalability: Longer Training

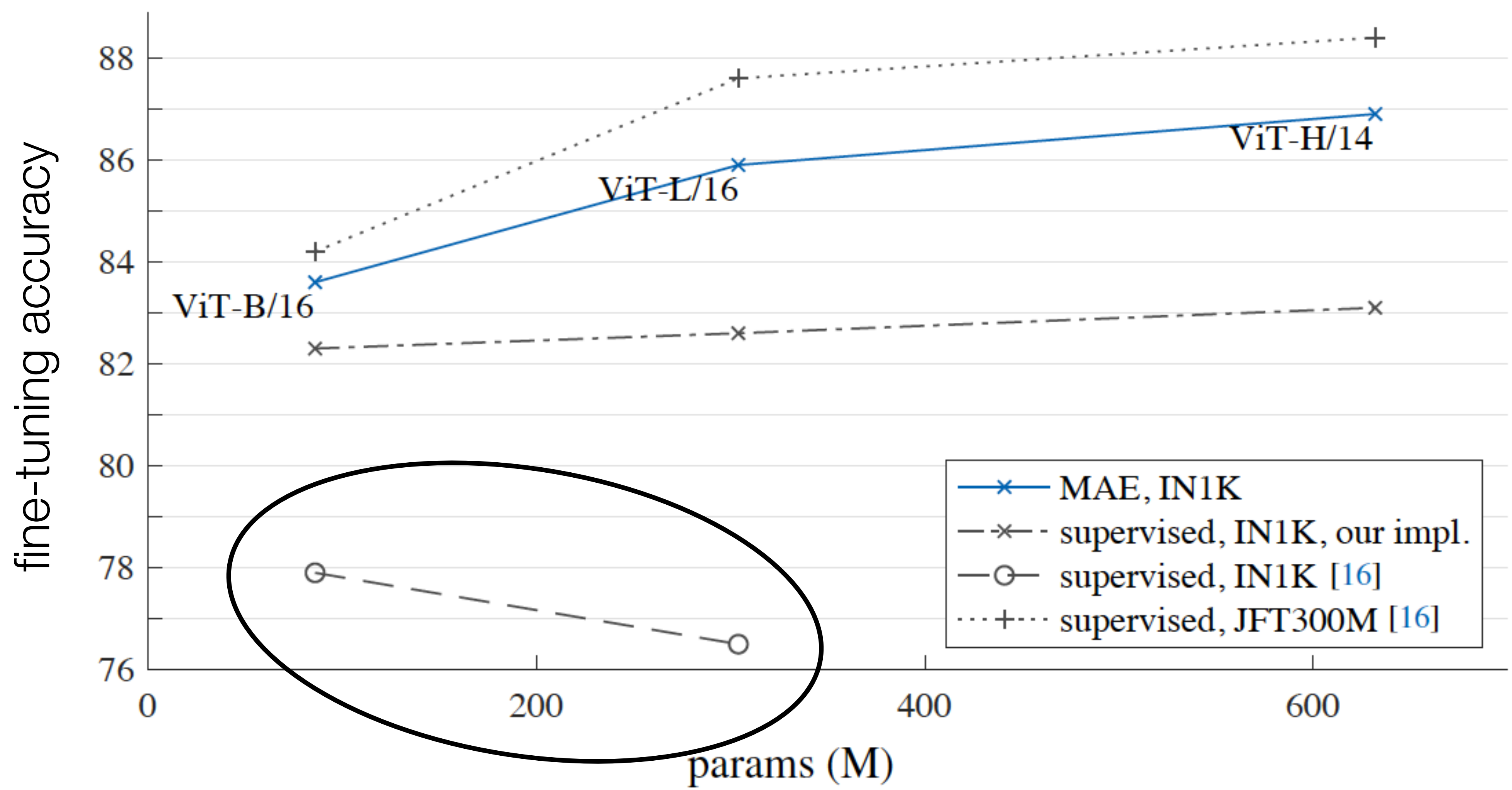


Wall-clock speed still efficient thanks to MAE design

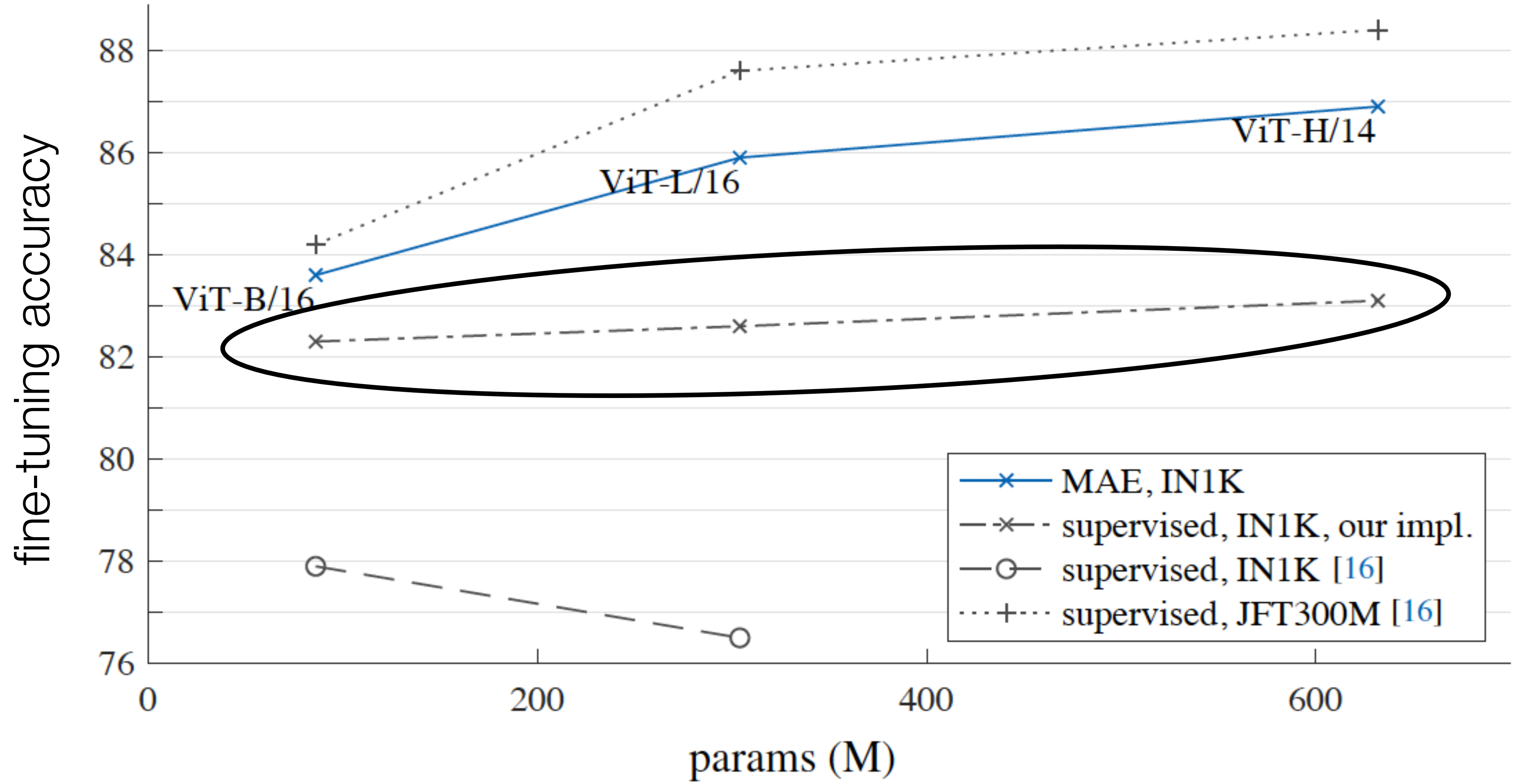
Scalability: Larger Models



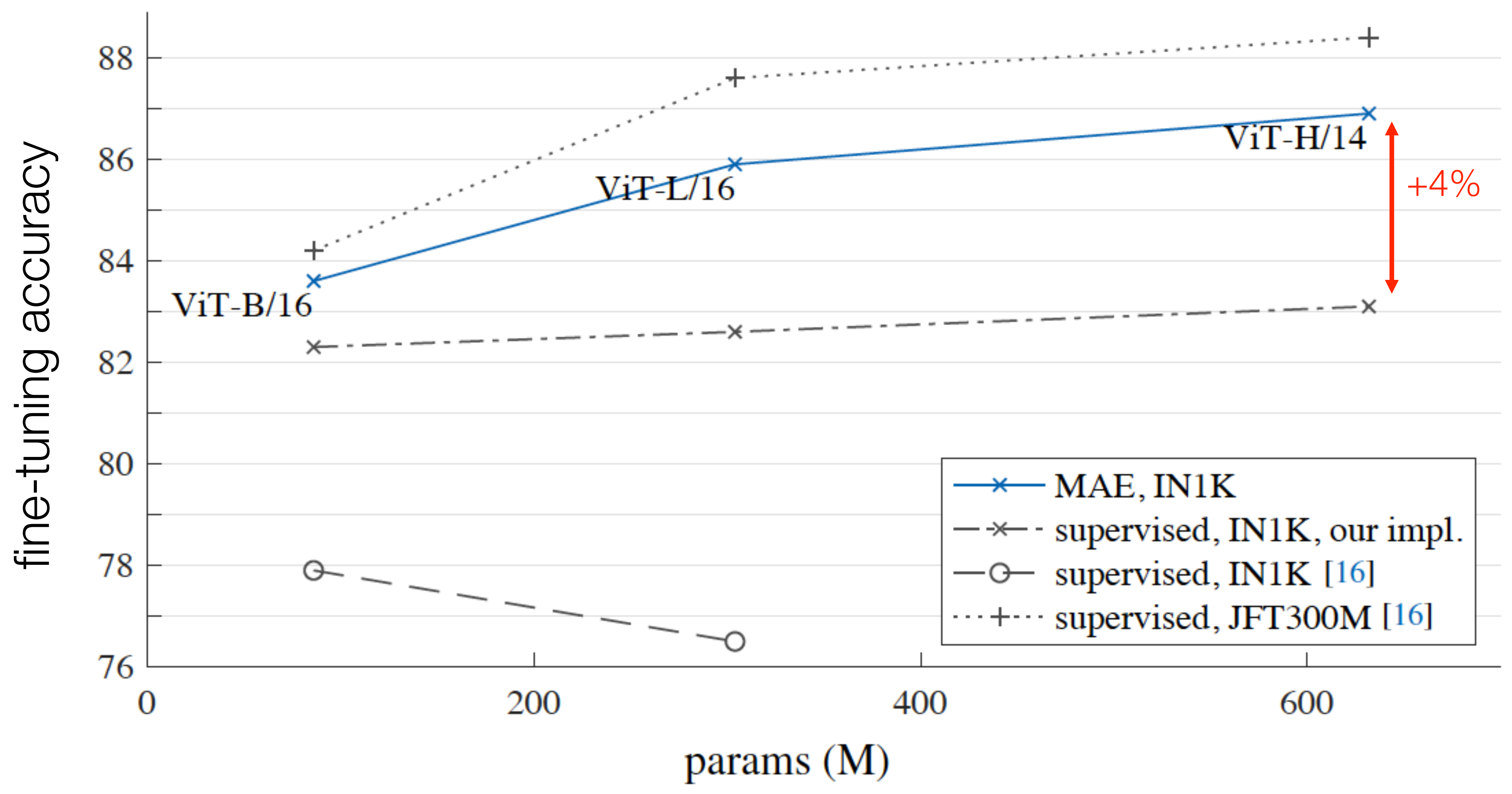
Scalability: Larger Models



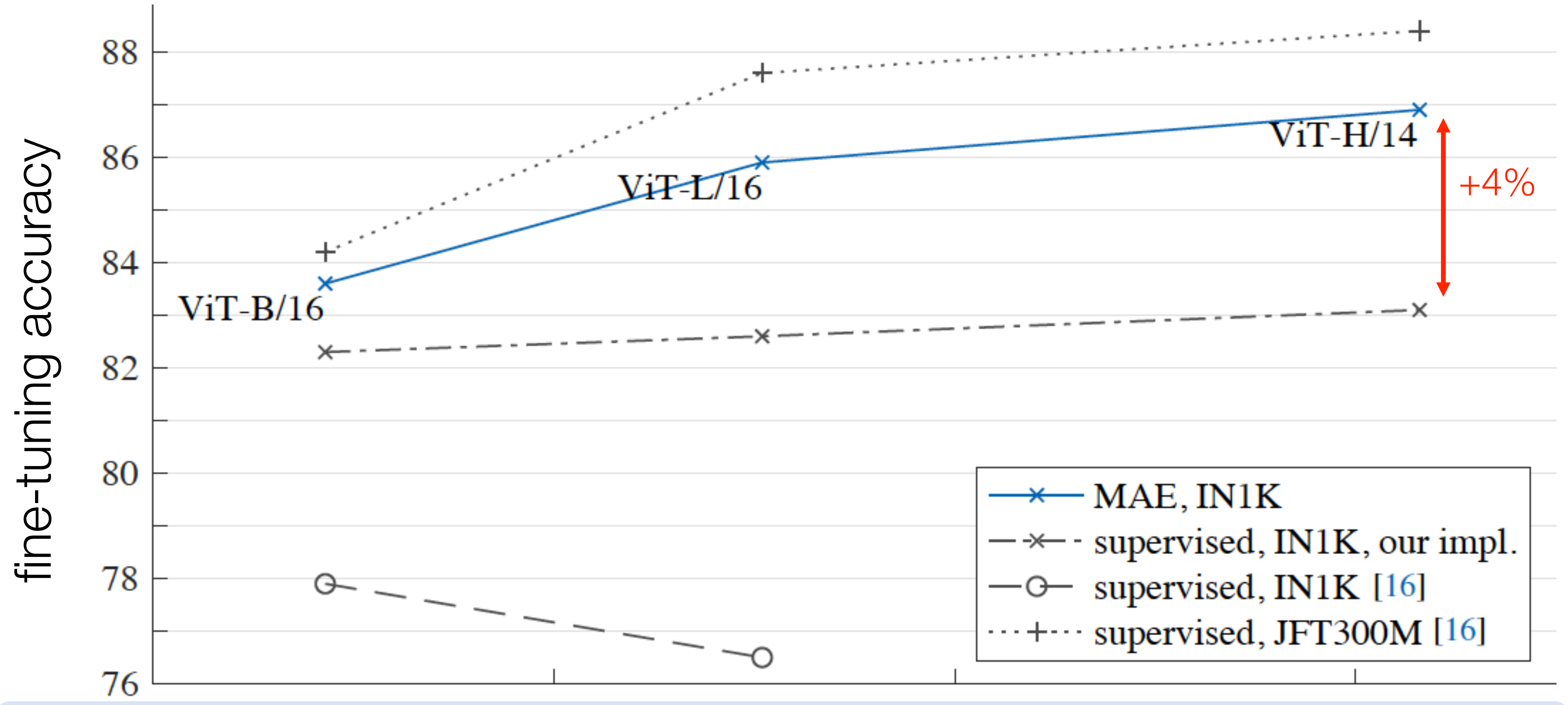
Scalability: Larger Models



Scalability: Larger Models



Scalability: Larger Models



new SOTA on ImageNet-1K (no extra data): **87.8%**

Scalability: Larger Models

dataset	ViT-B	ViT-L	ViT-H	ViT-H$_{448}$	prev best
iNat 2017	70.5	75.7	79.3	**83.4**	75.4 [50]
iNat 2018	75.4	80.1	83.0	**86.8**	81.2 [49]
iNat 2019	80.5	83.4	85.7	**88.3**	84.1 [49]
Places205	63.9	65.8	65.9	**66.8**	66.0 [19]†
Places365	57.9	59.4	59.8	**60.3**	58.0 [36]‡

new SOTA on 5 large-scale classification datasets

dataset	ViT-B	ViT-L	ViT-H	ViT-H$_{448}$	prev best
IN-Corruption ↓ [27]	51.7	41.8	**33.8**	36.8	42.5 [32]
IN-Adversarial [28]	35.9	57.1	68.2	**76.7**	35.8 [41]
IN-Rendition [26]	48.3	59.9	64.4	**66.5**	48.7 [41]
IN-Sketch [60]	34.5	45.3	49.6	**50.9**	36.0 [41]

new SOTA on 4 ImageNet robust evaluations

Scalability: Larger Models

method	pre-train data	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3
MoCo v3	IN1K	47.9	49.3
BEiT	IN1K+DALLE	49.8	**53.3**
MAE	IN1K	**50.3**	**53.3**

COCO detection: **+4.0%**

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**

ADE20K segmentation: **+3.7%**

[Hu et al, arXiv 2202.04639]

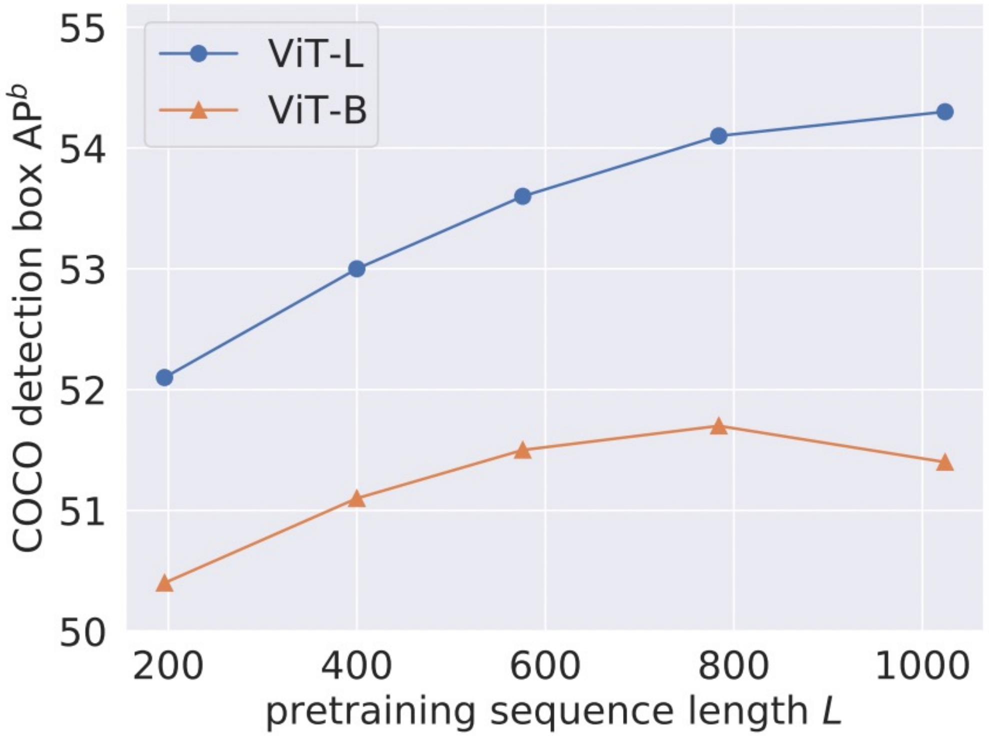
Scalability: Sequence Length

- Longer sequence length during pre-training, but fixed length during downstream transfers

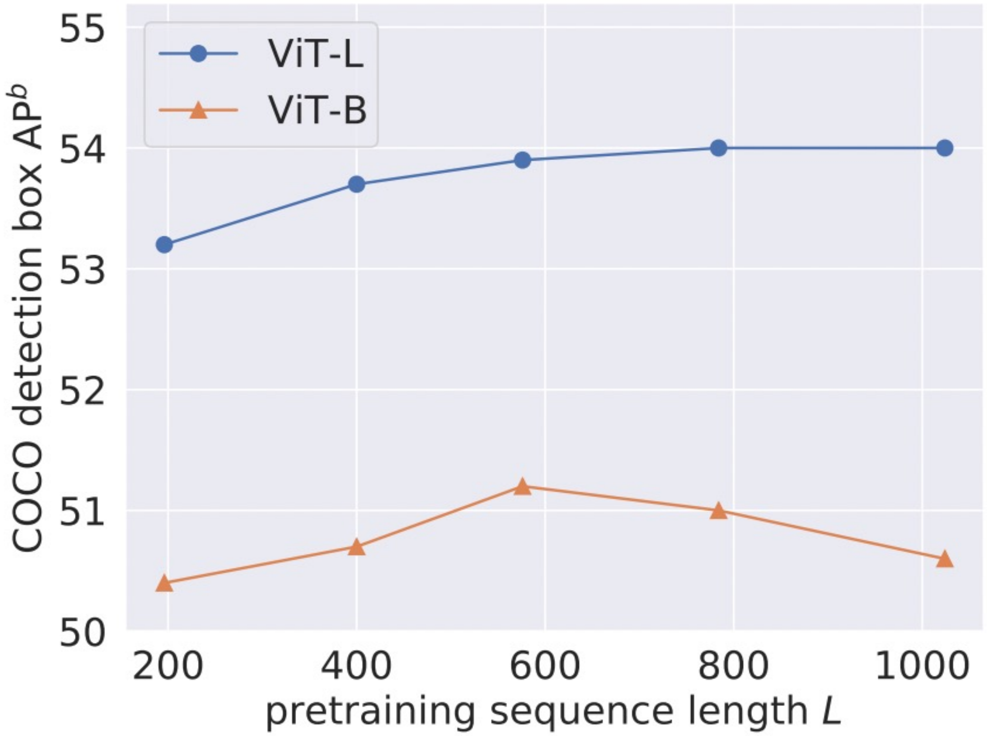
[Hu et al, arXiv 2202.04639]

Scalability: Sequence Length

- Longer sequence length during pre-training, but fixed length during downstream transfers



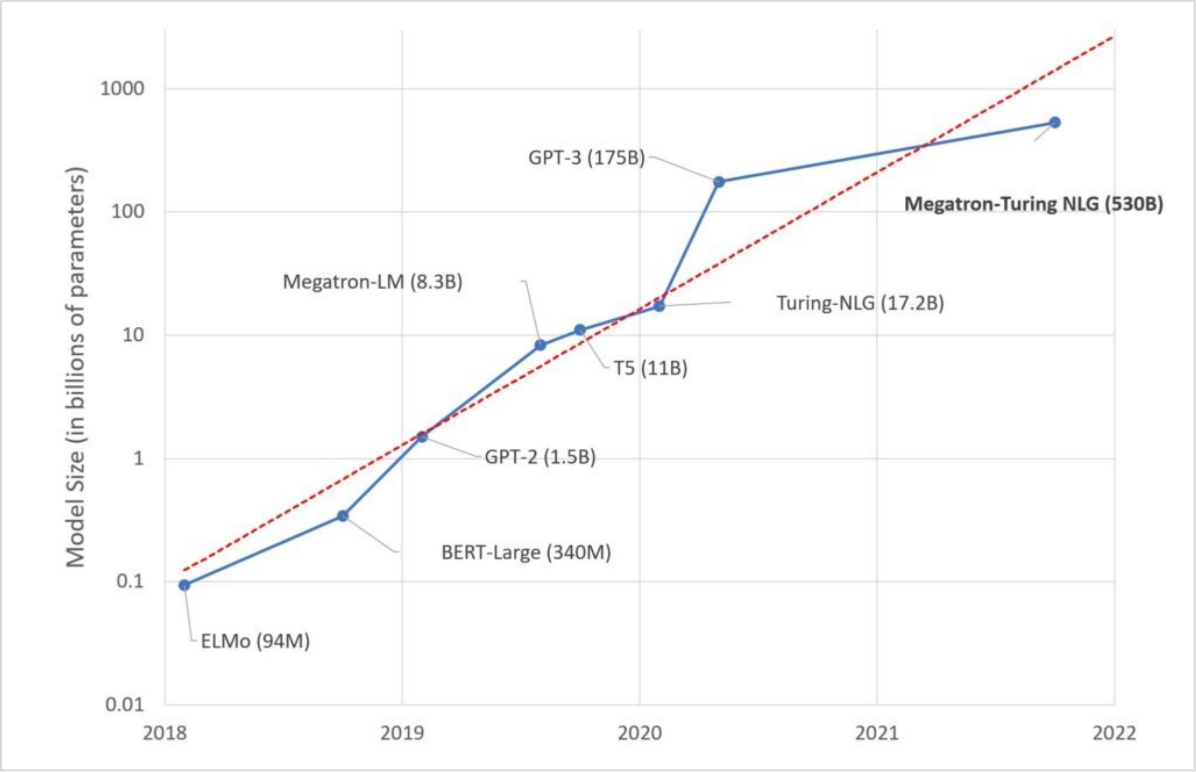
COCO pre-training



ImageNet-1K pre-training

Is the Journey 99% Done?

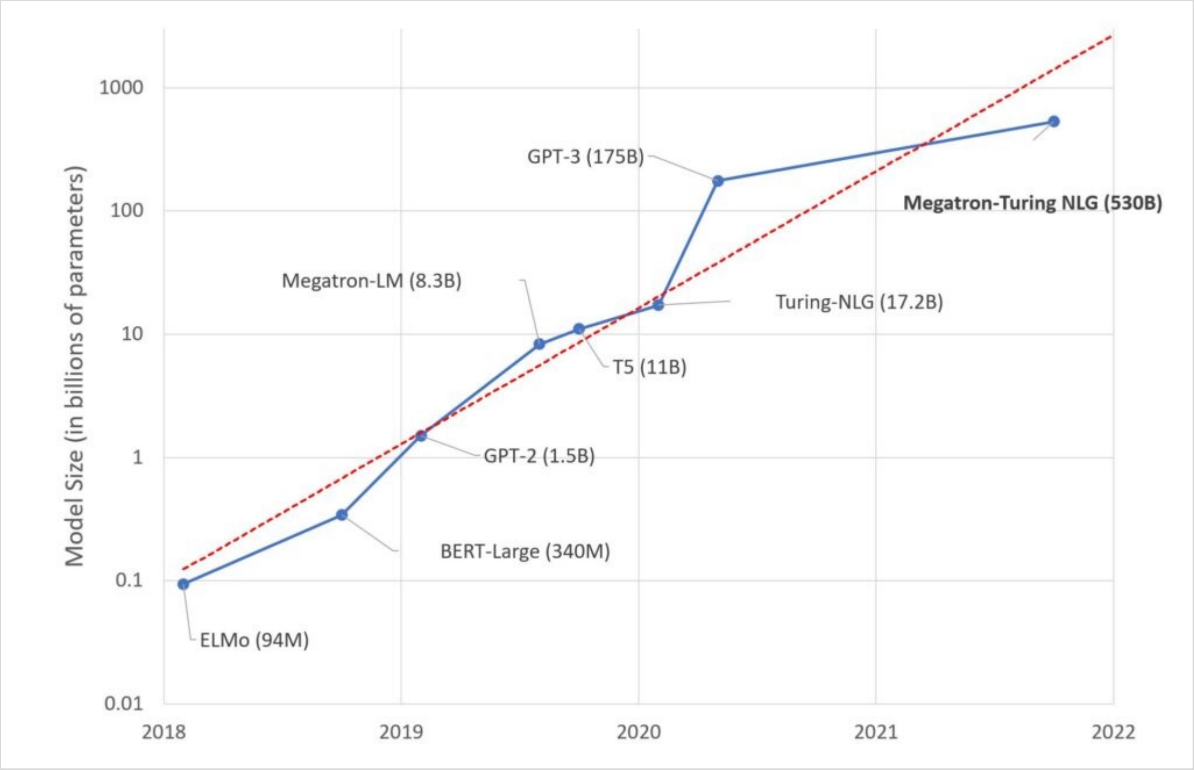
- NLP has witnessed amazing progress in scaling since BERT



Is the Journey 99% Done?

- NLP has witnessed amazing progress in scaling since BERT

- It's just *starting* in vision:
  - Temporal data – Christoph
  - Architectures – ConvNets?
  - Other modalities? 3D?
  - Other downstream tasks?
  - Other axes to scale?
  - [`Your exploration`] here!



code (GPU): https://github.com/facebookresearch/mae

code (TPU): https://github.com/facebookresearch/long_seq_mae

Take-aways

- Self-supervised learning aims at *scalable* representation learning

- Masked auto-encoders can serve as scalable vision learners

- Exciting years ahead in this direction!