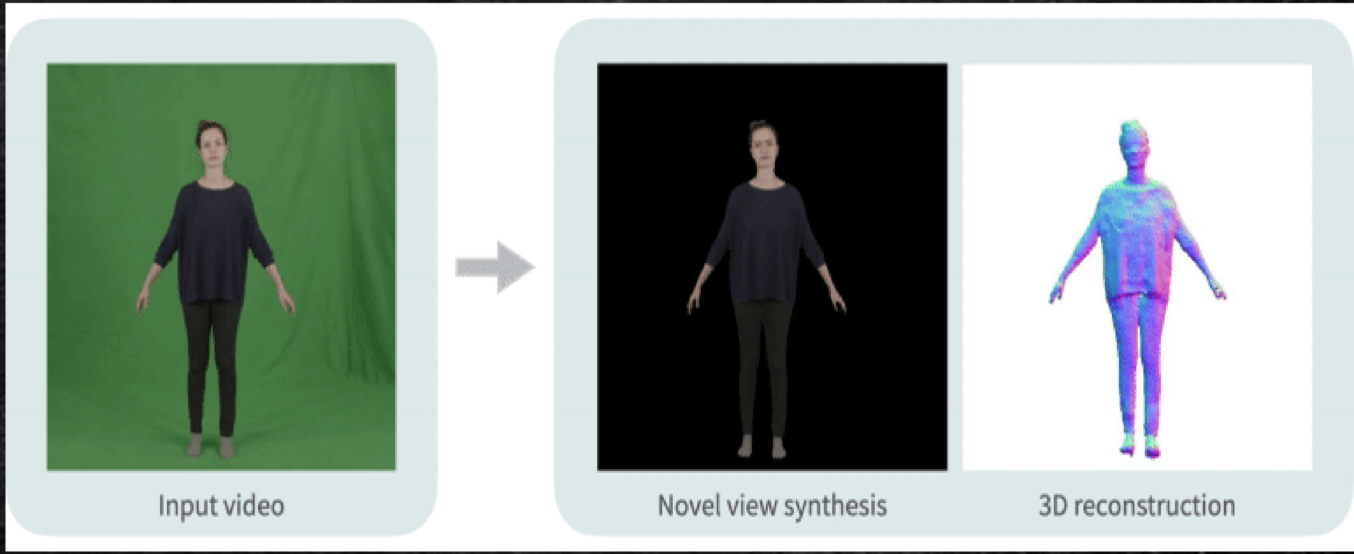
# California State University, Long Beach
# CECS 550 – Pattern Recognition
# Midterm Paper Presentation


- Mentor: Dr. Mahshid Fardadi, Ph.D.

- Student: Ms. Aishwarya Bhavsar (029371509)

# Masked Autoencoders Are Scalable Vision Learners

- Facebook AI Research (FAIR)
- Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, Ross Girshick ´



# Meaning Of Masked Autoencoders

# Why we need Masking?

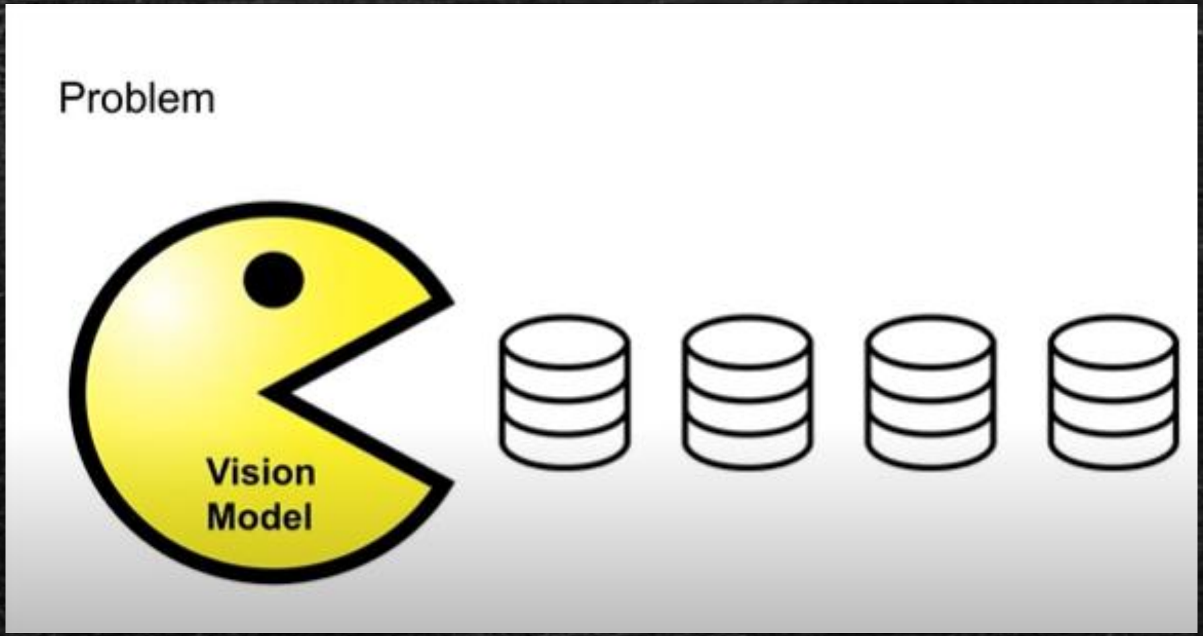
- Data masking is simply a procedure that hides the information of the data by masking.
- We are required to use data masking to train our model more accurately and precisely about the data.
- Reduces the size of training data and models.
- Reduces power, memory and computational utilisation.
- Autoencoders - Encoding and decoding data.
- Advantage of MAE - self-supervision.

# How Vision Transformers (ViT's) Work?

# This paper contains:

- Transformers
- Masking
- Autoencoding
- Self-supervision

# Problem




Problem that the author is trying to tackle here is:

1. Vision Models can easily overfit large amounts of images.
2. Vision Models can demand more data which is obtain publically and is accessible.

# Solution in NLP

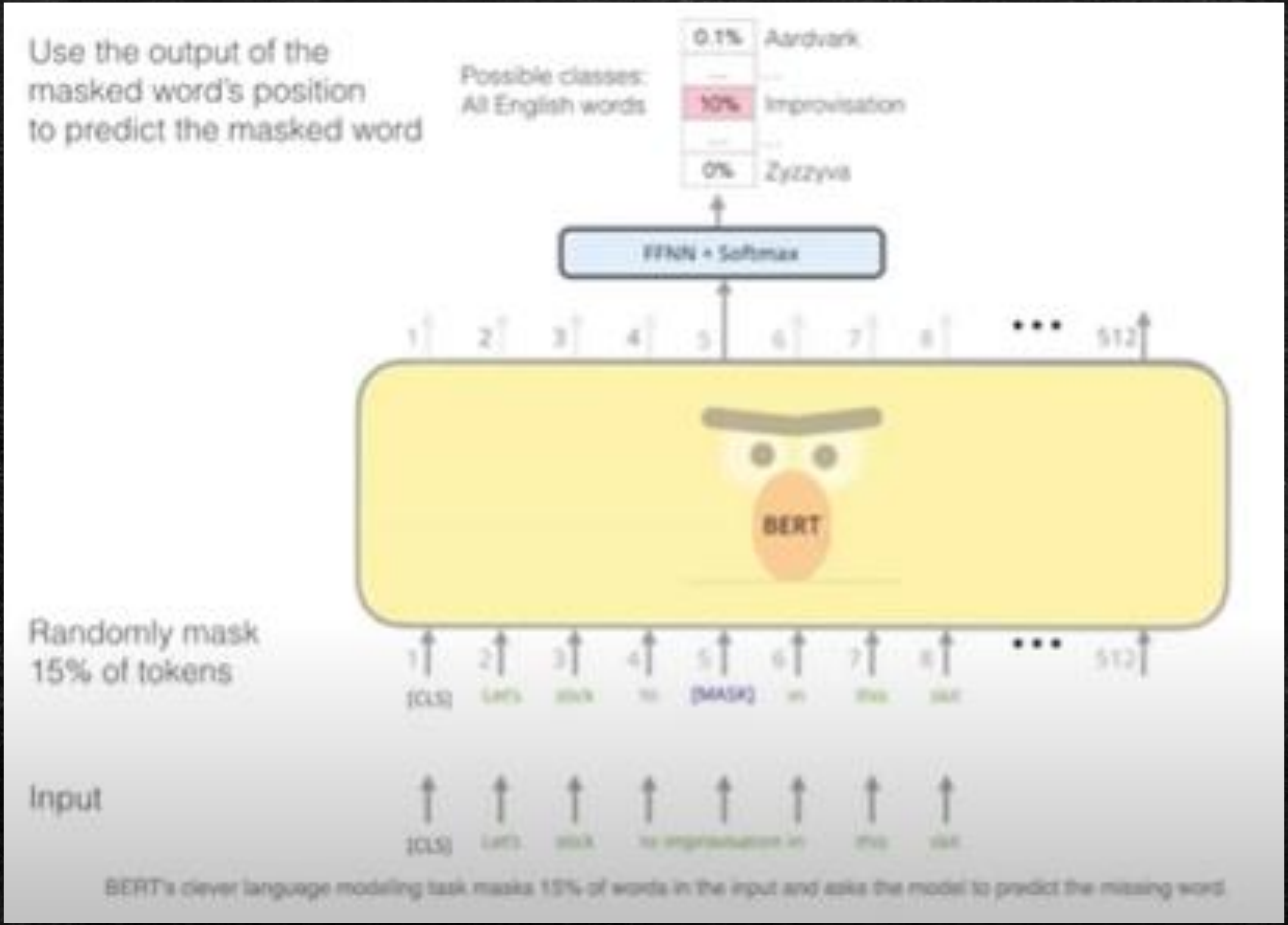
NLP has a solution to this:

Self-supervised pre-training.

# Model

## Model

- Encoder
  - Operates on visible subset of patches
  - No mask tokens
  - ~25% of the full patches
- Decoder
  - Lightweight
  - Reconstructs the input from the latent representation along with mask tokens
  - Outputs a vector of pixel values representing a patch



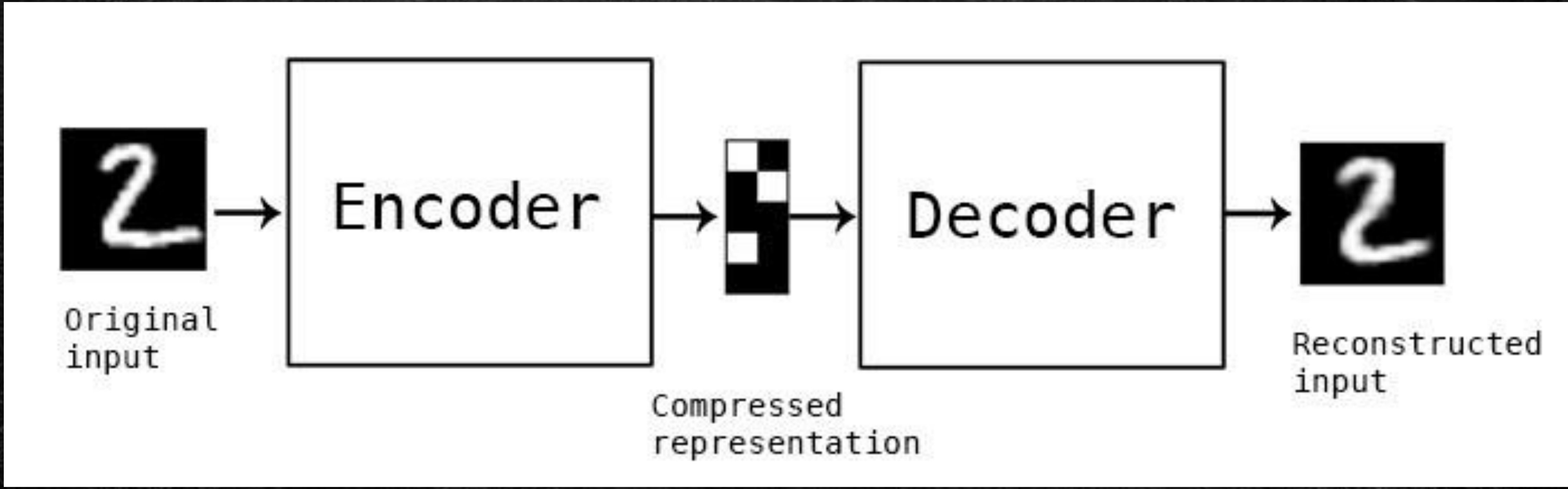

You randomly mask
Certain percentage
Of tokens.

# Research Question

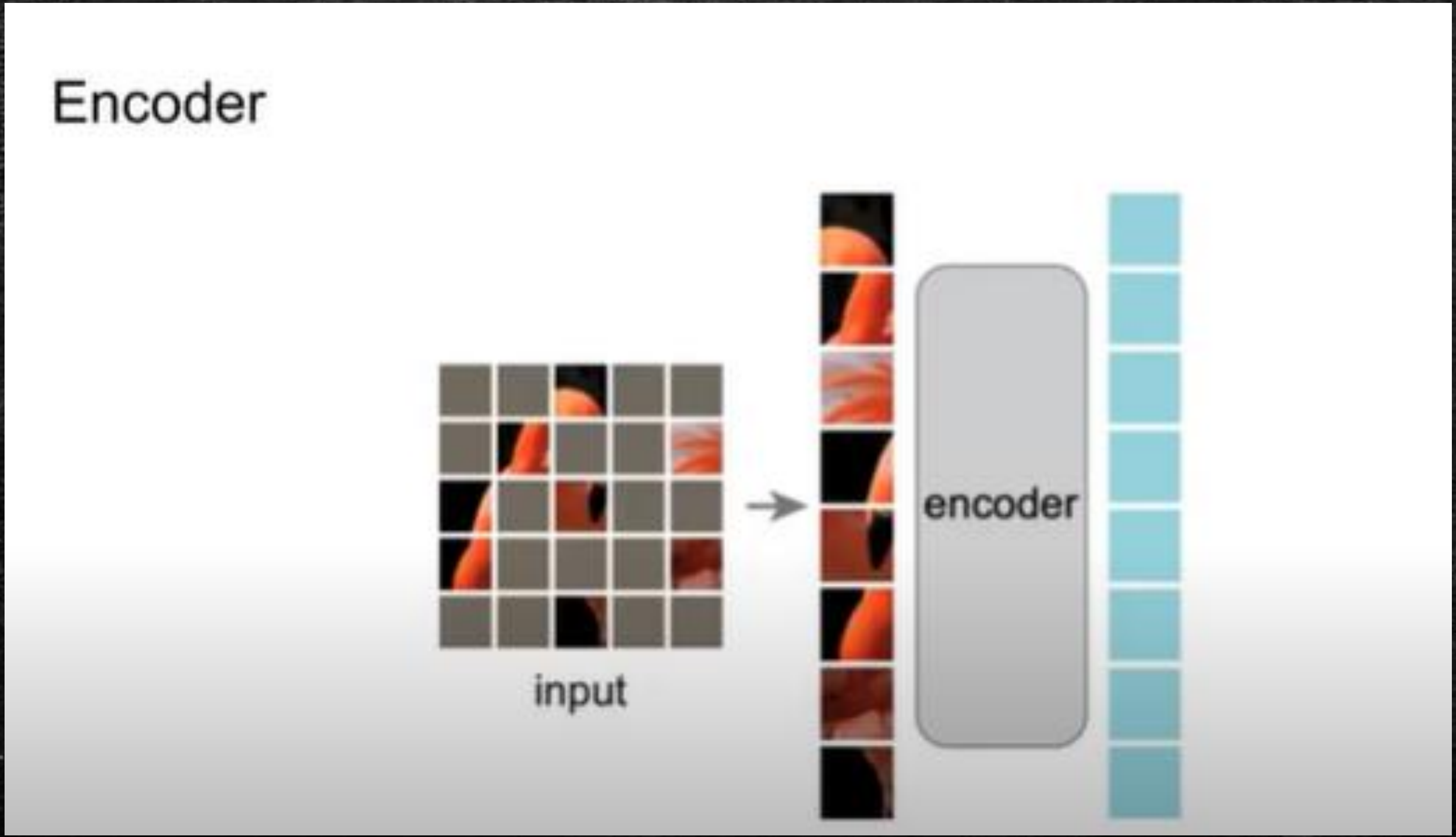
1] How should you design an optimal method for vision autoencoding?

2] What makes masked autoencoding different between vision and language?

# Encoder Decoder Architecture

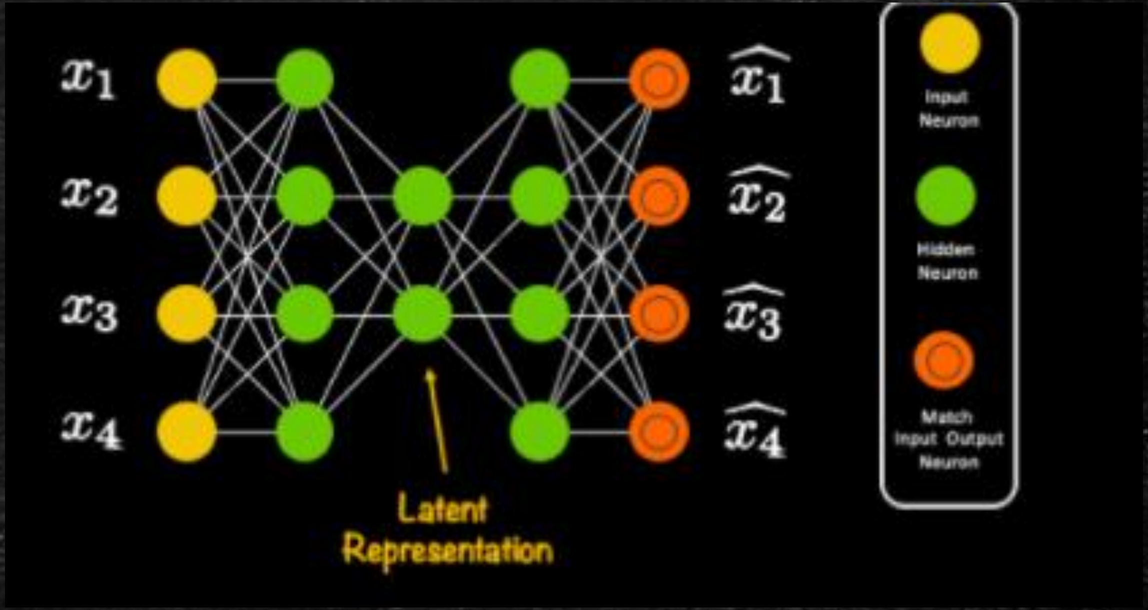


# Encoder – Standard vision transformer which is applied only to visible unmasked patches

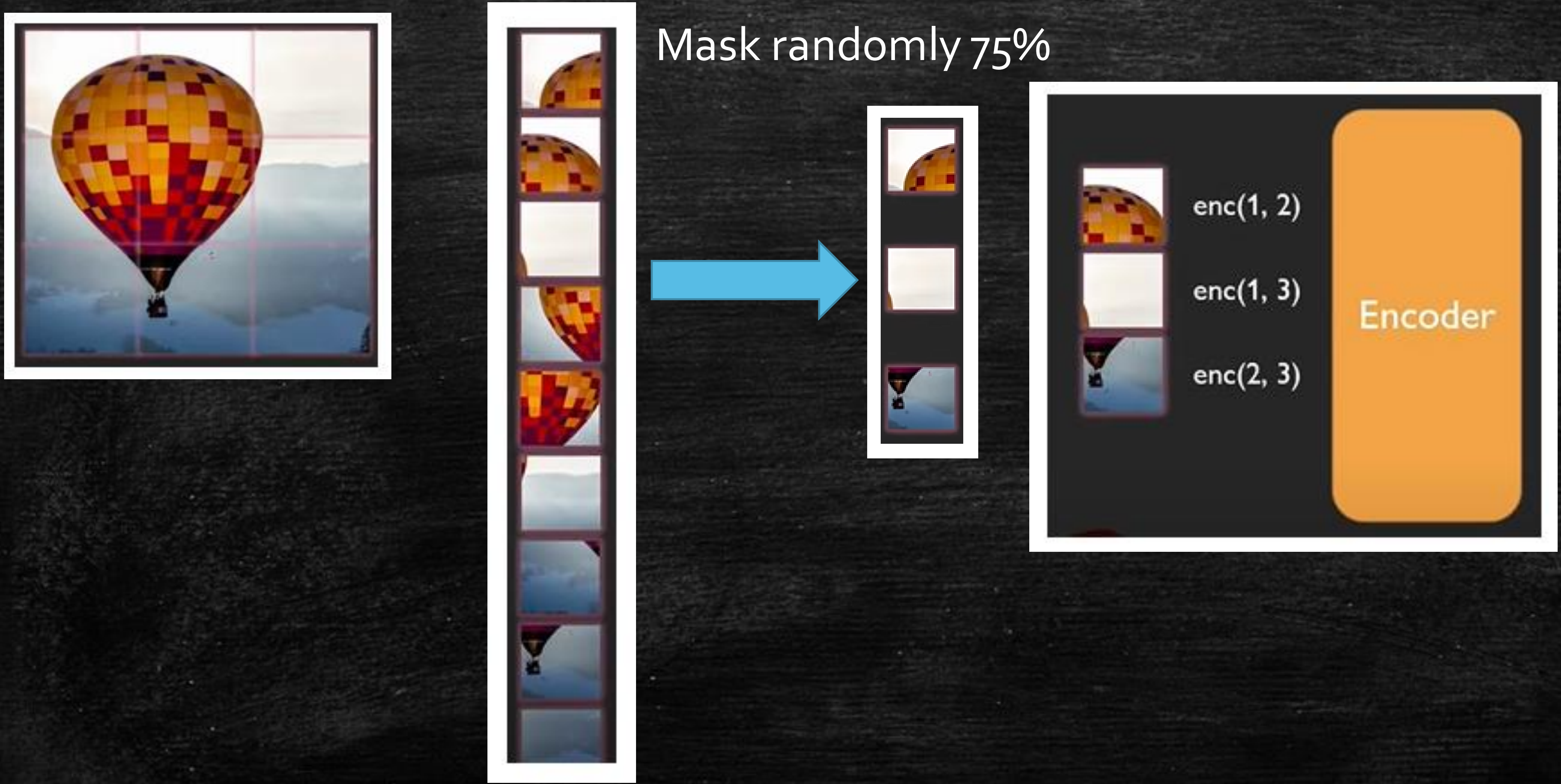


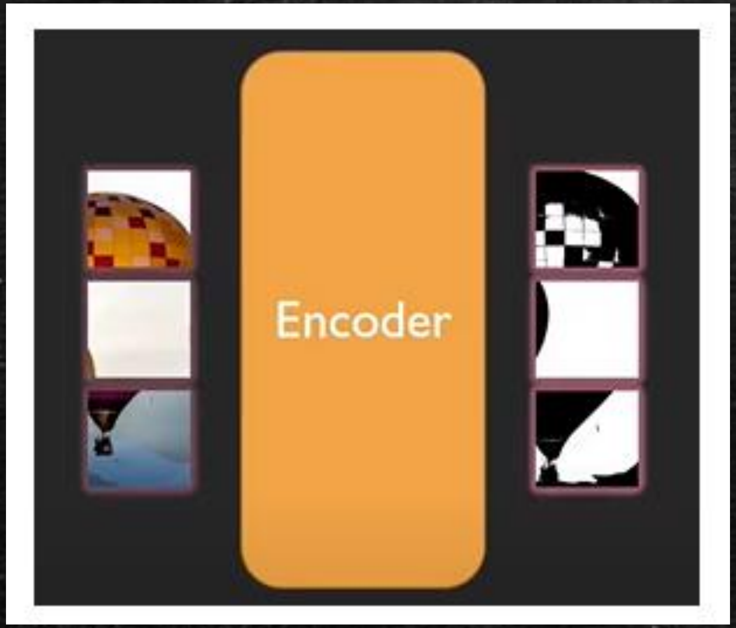
# Working Of Encoders

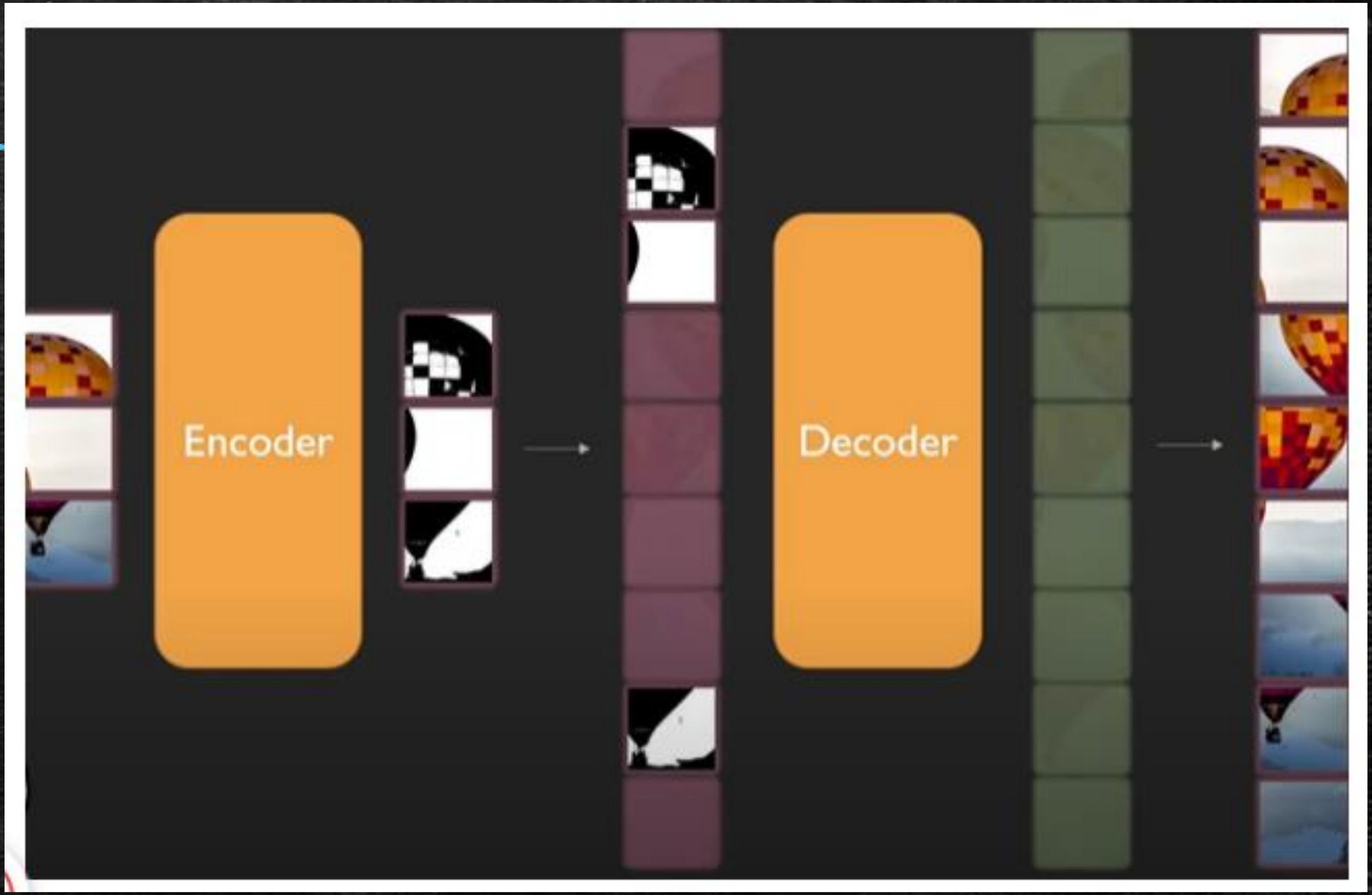
- Encoders embed a patches by using linear projections.
- Uses positional embeddings.



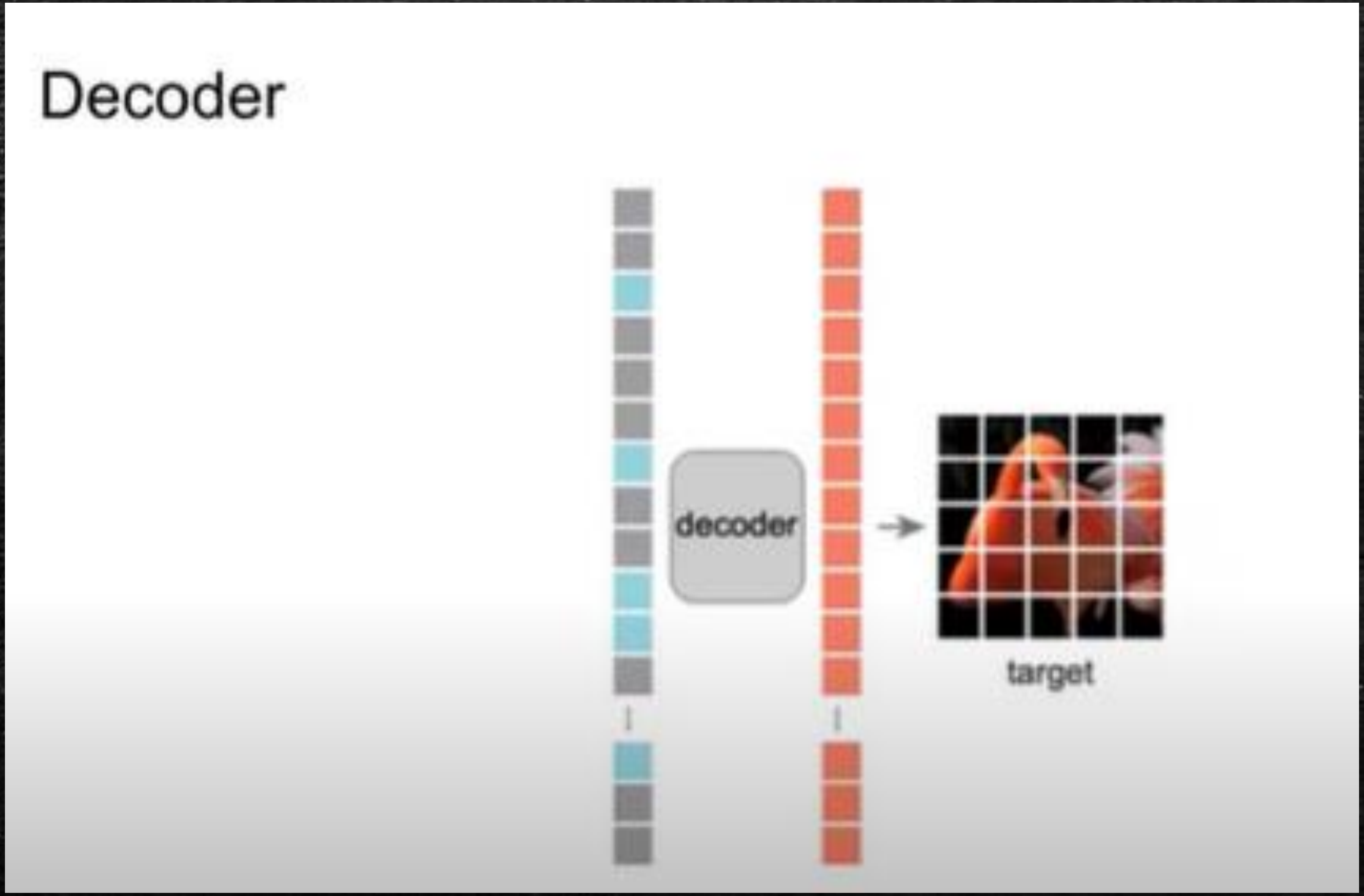
# Working Of Encoders



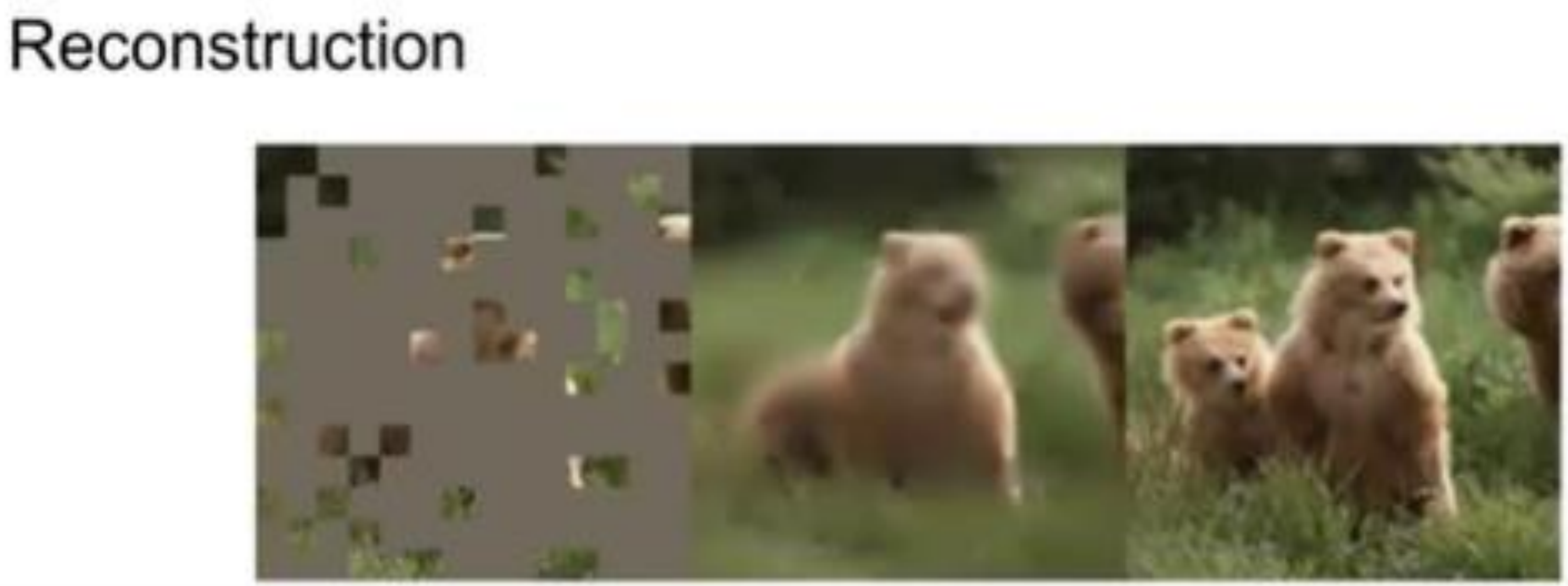

Encoder


Encoder
Decoder

# Decoder – Input to the decoder is full set of tokens which includes encoded visible patches



# Reconstruction – Model reconstructs the input by predicting the individual pixel values for each of the masked patches.




$$\Downarrow$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

# Results:

## Results - ViT-L

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

# Effect Of Masking Ratio:

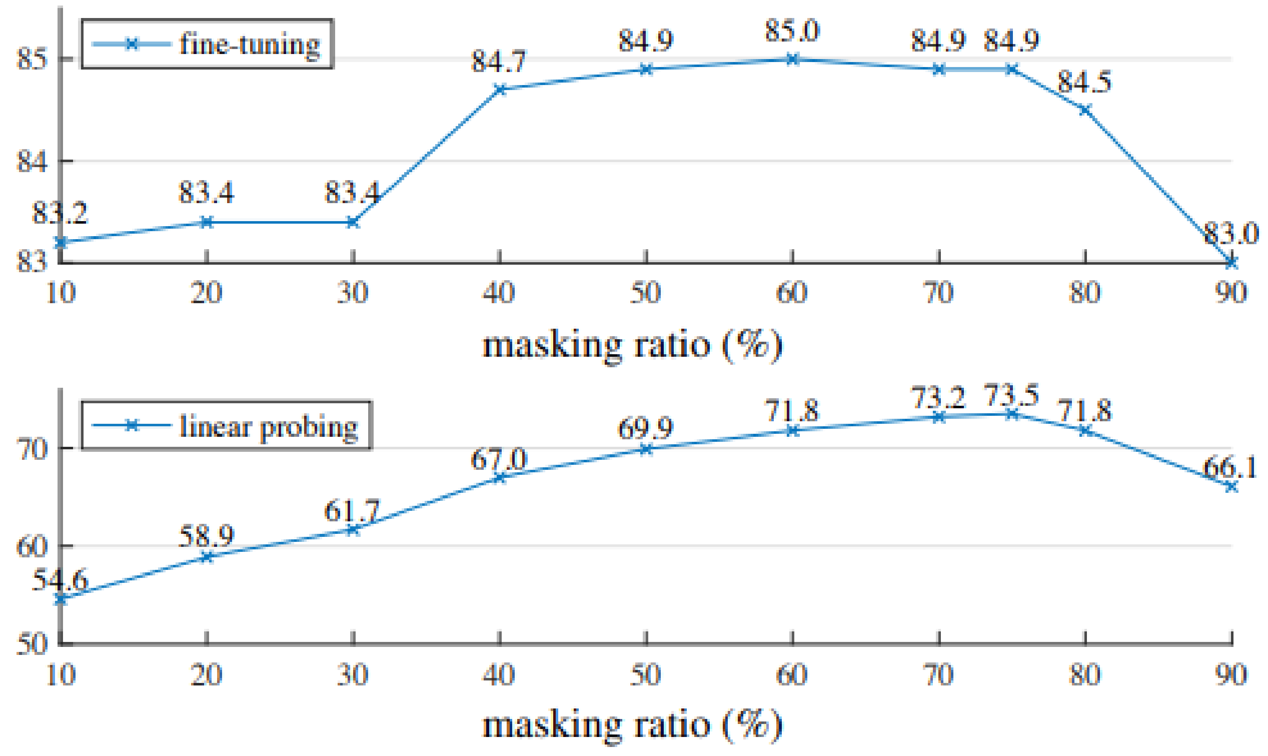



Figure 5. **Masking ratio**. A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

# Ablation:

| 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.

# Mask Sampling Techniques:


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.

# Comparisons:

| 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.

# Comparisons:

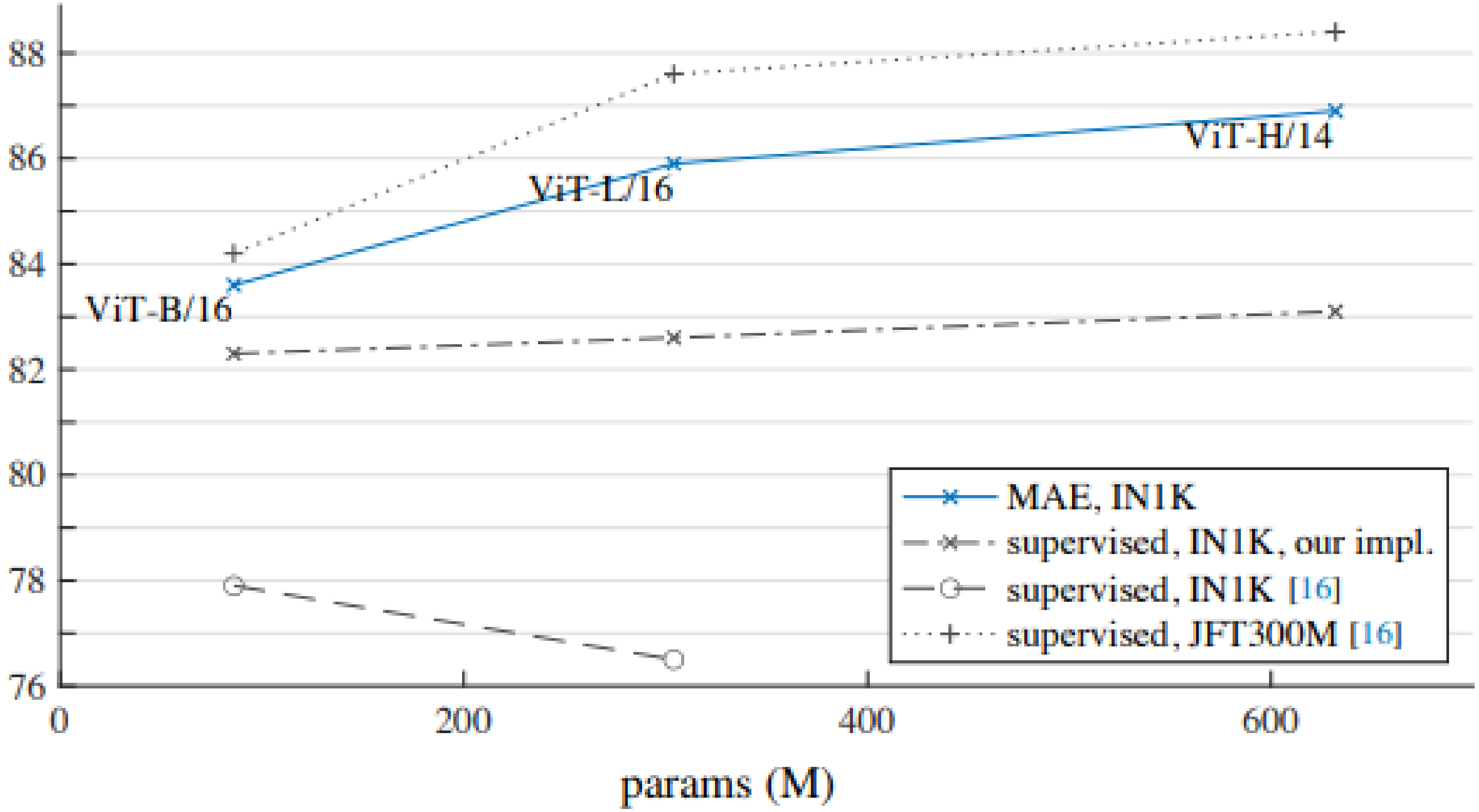



Figure 8. **MAE pre-training *vs.* supervised pre-training**, evaluated by fine-tuning in ImageNet-1K (224 size). We compare with the original ViT results [16] trained in IN1K or JFT300M.

| method | pre-train data | AP^box ViT-B | AP^box ViT-L | AP^mask ViT-B | AP^mask 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.

These observations suggest that linear separability is not the sole metric for evaluating representation quality. It has also been observed (*e.g.*, [8]) that linear probing is not well correlated with transfer learning performance, *e.g.*, for object detection. To our knowledge, linear evaluation is not often used in NLP for benchmarking pre-training.

| 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]$^\dagger$ |
| Places365 | 57.9 | 59.4 | 59.8 | **60.3** | 58.0 [36]$^\ddagger$ |

Table 6. **Transfer learning accuracy on classification datasets,**

# How can we reconstruct the tasks?


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.

# Qualitative Results:

Thank you