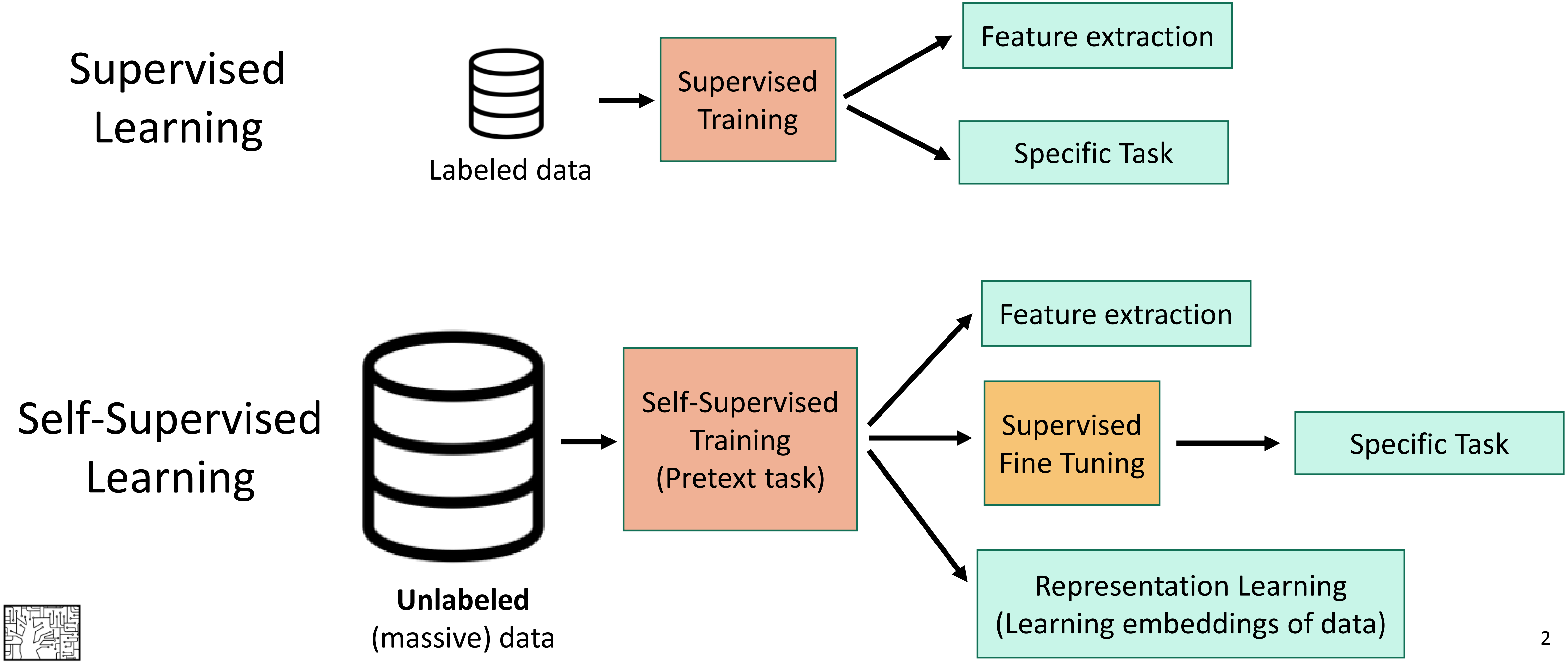
Self-Supervision II

Reminder – Self supervised learning



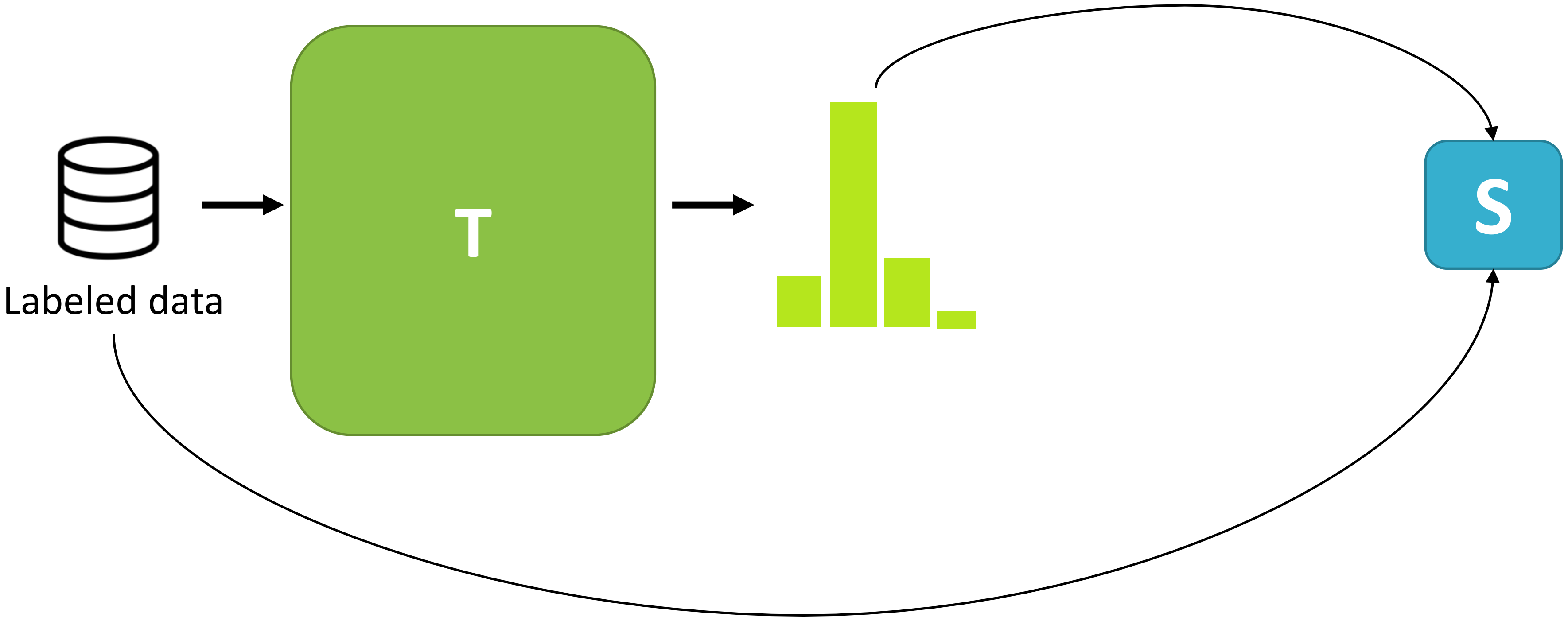
Topics

- self-DIstillation with NO labels
  - **DINO**
- Masked Auto Encoders
  - **MAE**
- Contrastive Language Image Pretraining
  - **CLIP**

WAIC

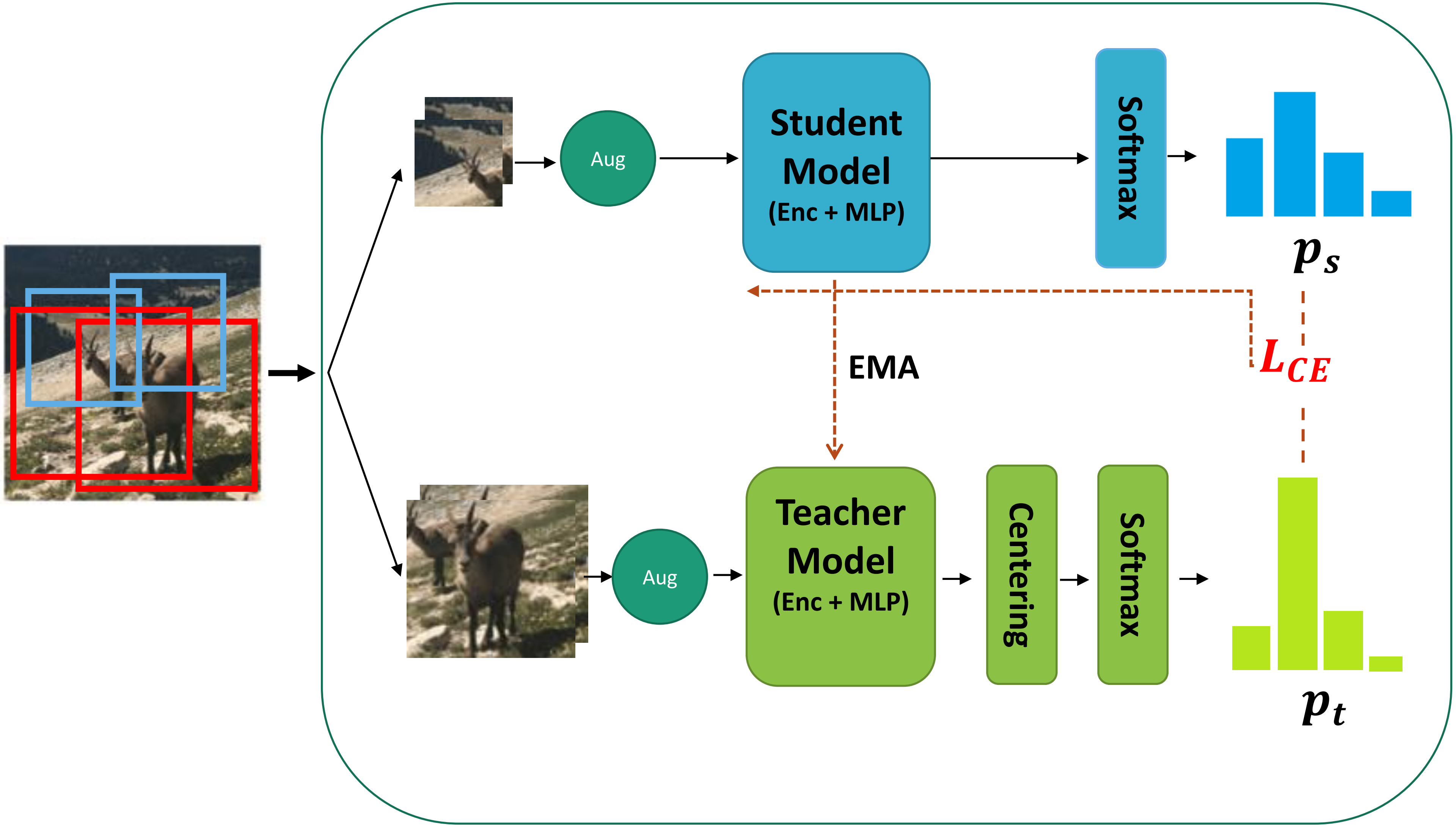
DiNO - Approach

- Self supervised learning as a special case of **knowledge distillation**

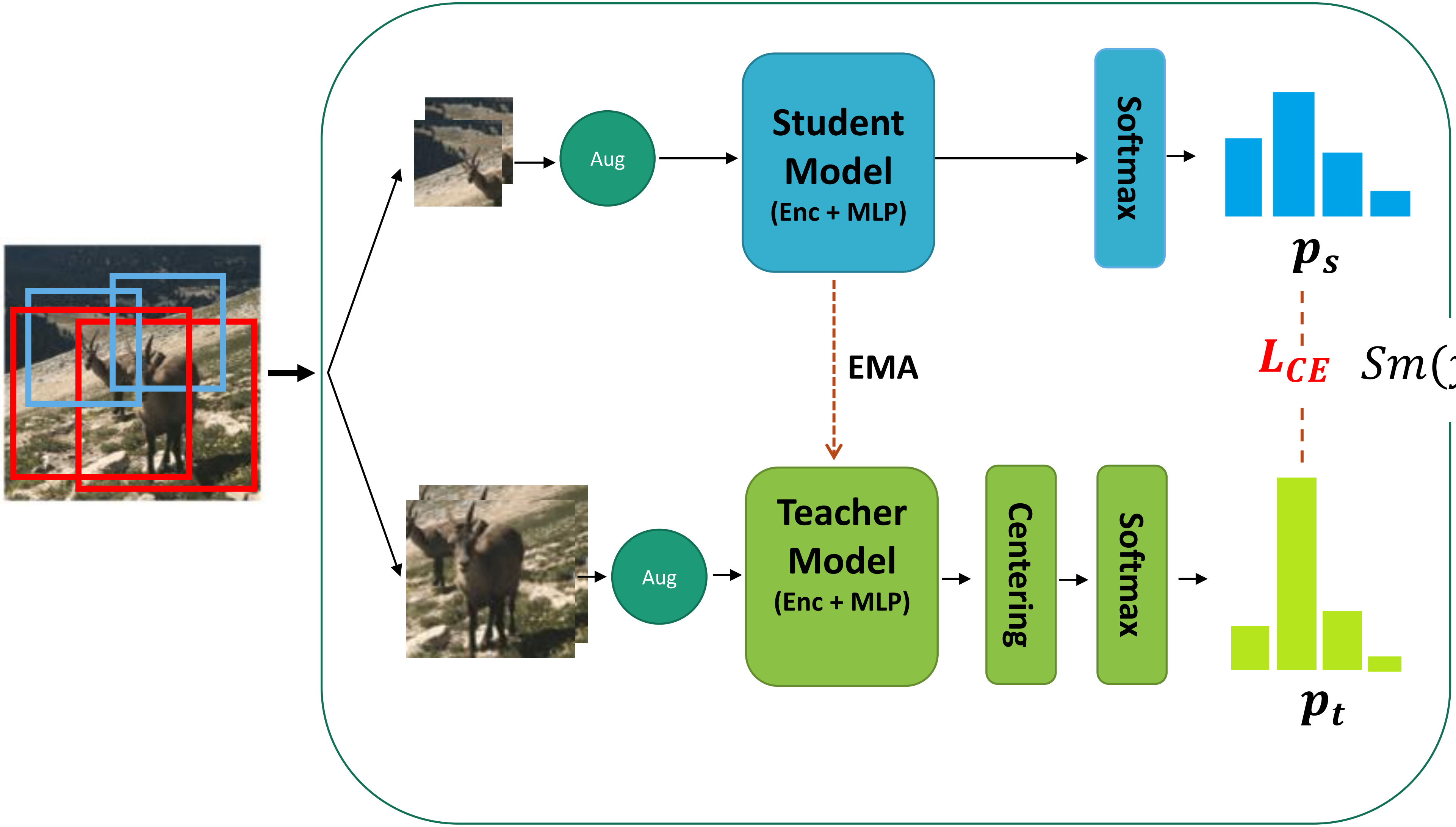


DiNO - Training

DiNO - Training



DiNO - Training



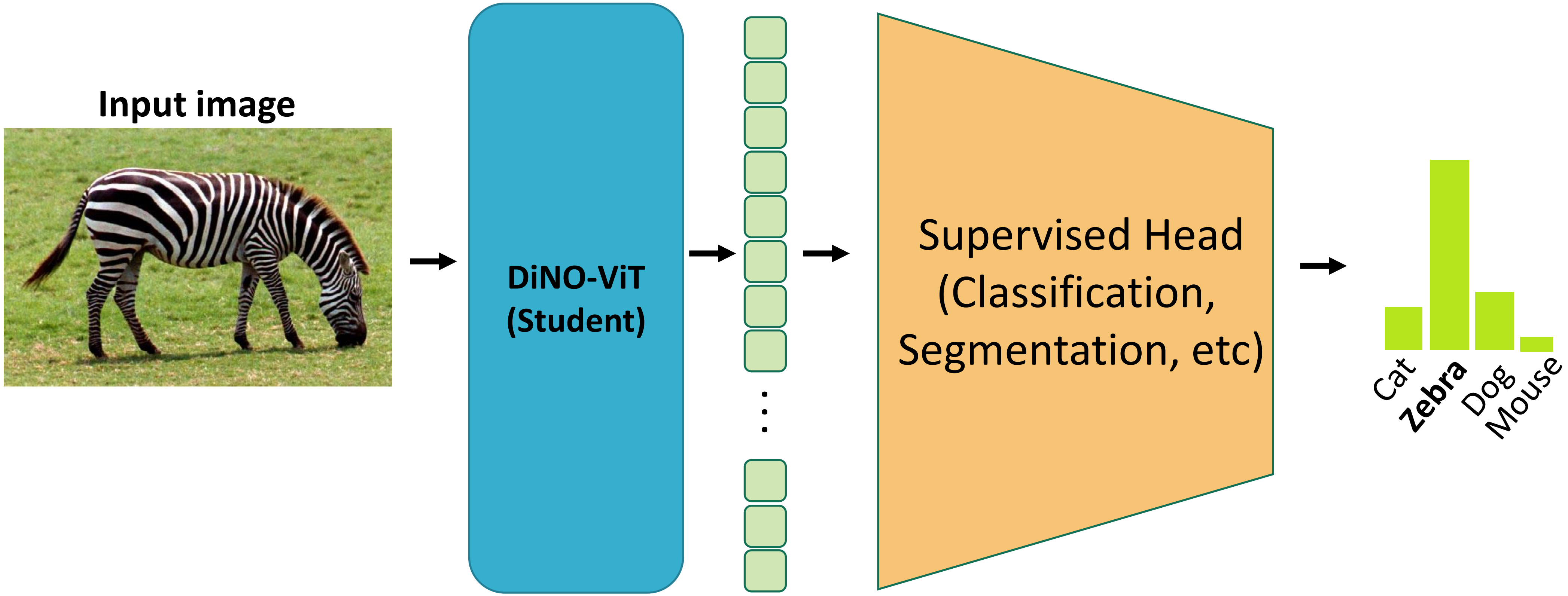
$$Sm(y) = \frac{exp(y_i/\tau)}{\sum_{j=1}^{K} exp(y_j/\tau)}$$

WAIC

DiNO - Explanation

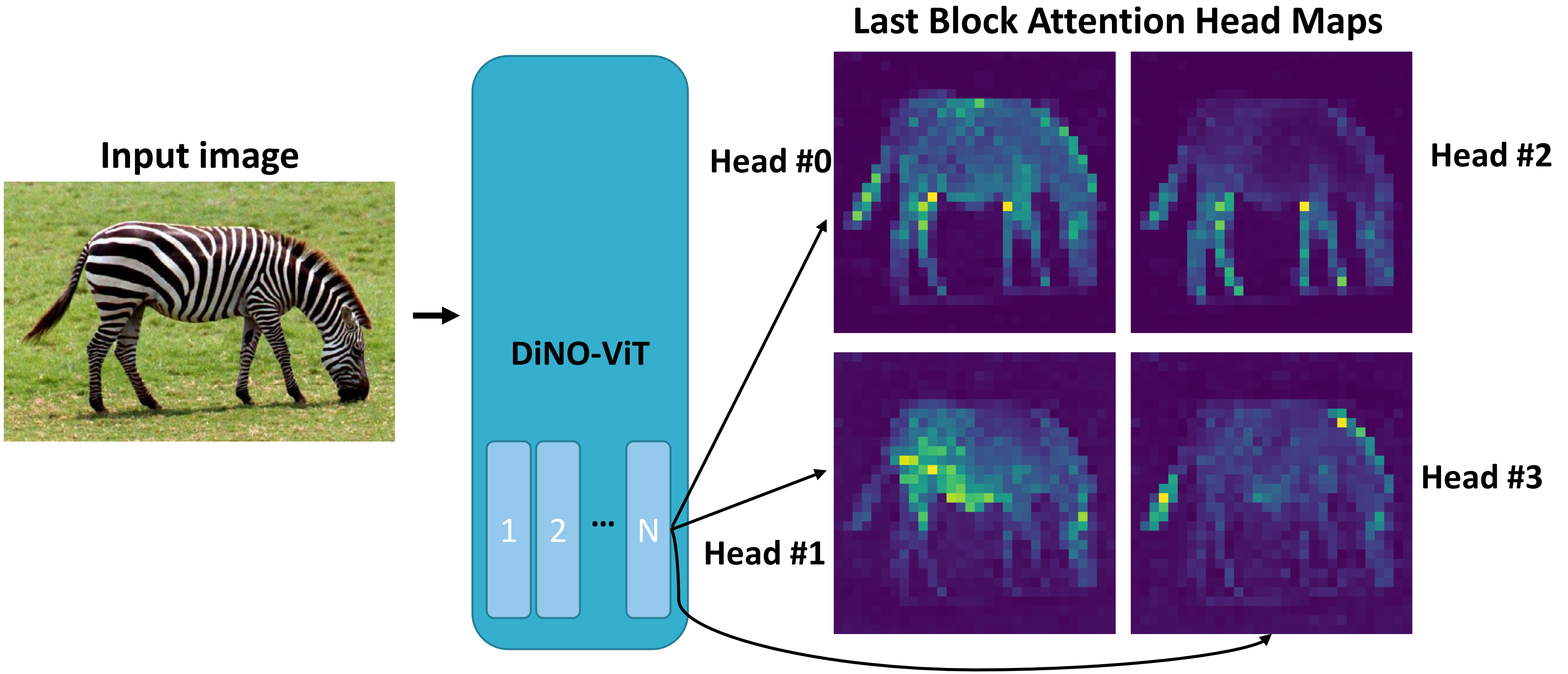
- Augmentations
  - Tells the model what to ignore
    - Collor jitter, Gaussian Blur, Solarize
    - Acts as a data prior
- “Global – local” cropping
- Teacher out-distribution sharpening via centering & Low-temperature in softmax
- The student encoder learns “abstract representations”
  - No awareness of “class labels” or meaning behind logits

DiNO - Inference #1



WAIC

DiNO - Inference #2



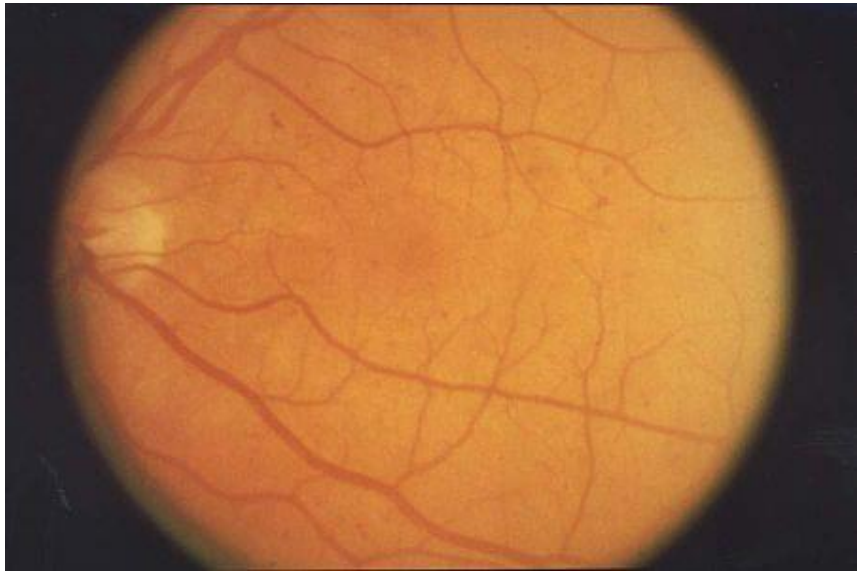
DiNO - Inference #2

Input | **Supervised Segmentation** | **DiNO (SSL) Segmentation**

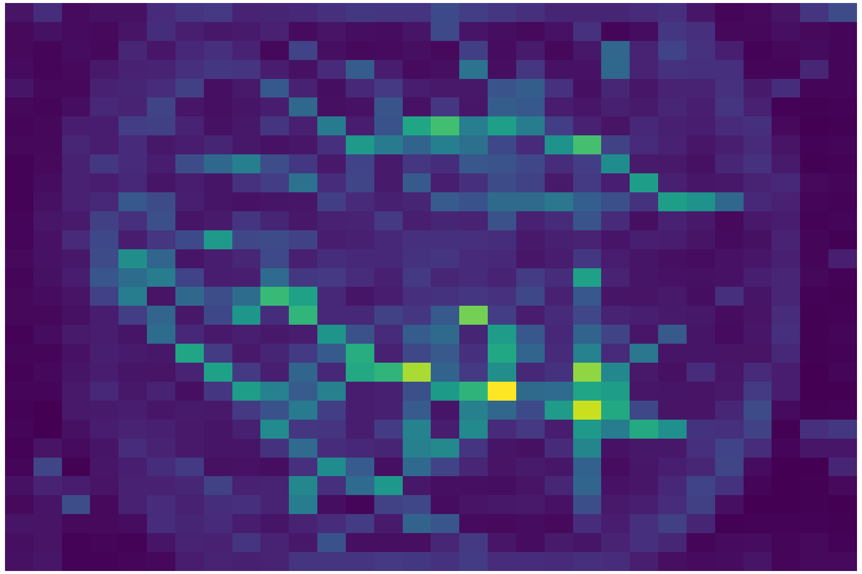
DiNO - Inference #2

- Self supervised learning also makes learned representations applicable to out-of-distribution data

Input image

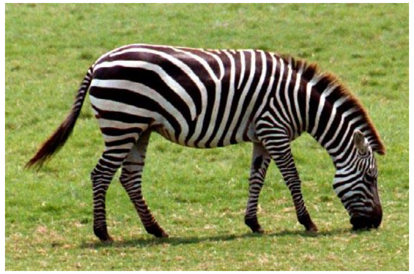


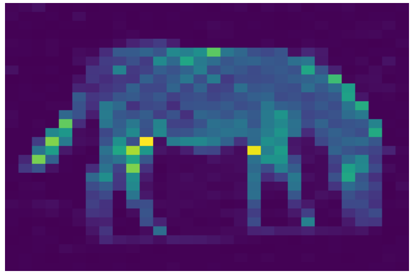
Last block attention map



Usage

```
vit_model = torch.hub.load('facebookresearch/dino:main’,
                           f'dino_vits16', pretrained=True)
img = imread(‘zebra.png’)

x = vit_model.prepare_tokens(img)
for blk in vit_model.blocks[:-1]:
    x = blk(x)
attn_maps = vit_model.blocks[-1](x, return_attention=True)

# Choose head, Get attention map of class token
attn_map = attn_maps[0, HEAD, 0, 1:].reshape((1, 1, H_PATCHES, W_PATCHES))
attn_map = F.interpolate(attn_map, scale_factor=16, mode="nearest")
```

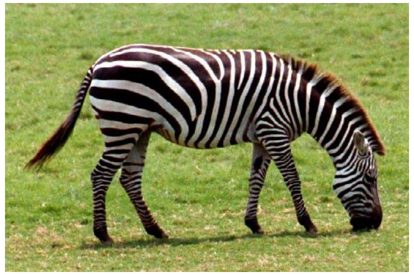


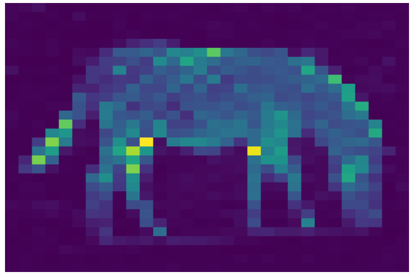


WAIC

Usage

```
vit_model = torch.hub.load('facebookresearch/dino:main’,
                          f'dino_vits16', pretrained=True)

img = imread(‘zebra.png’)

attn_maps = vit_model.get_last_selfattention(img)

# Choose head, Get attention map of class token
attn_map = attn_maps[0, HEAD, 0, 1:].reshape((1, 1, H_PATCHES, W_PATCHES))
attn_map = F.interpolate(attn_map, scale_factor=16, mode="nearest")
```

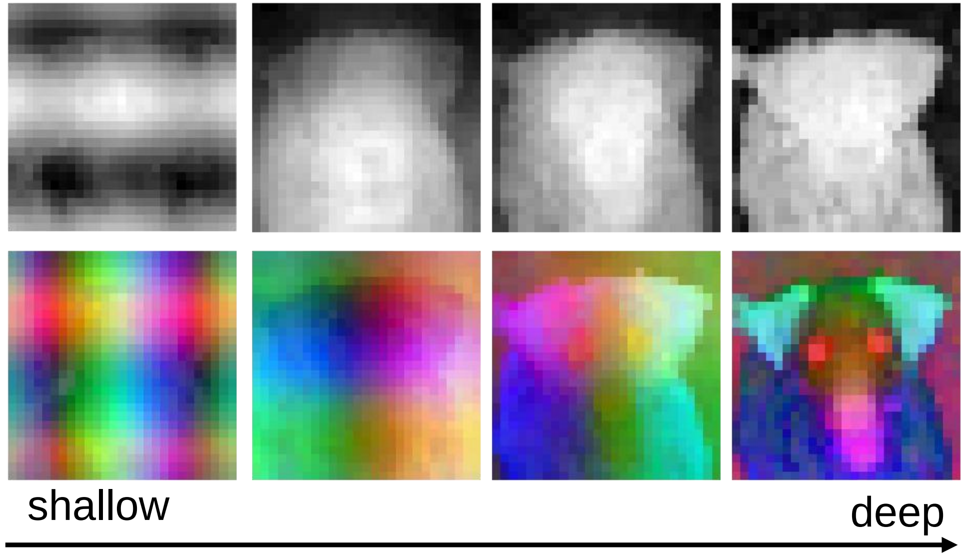




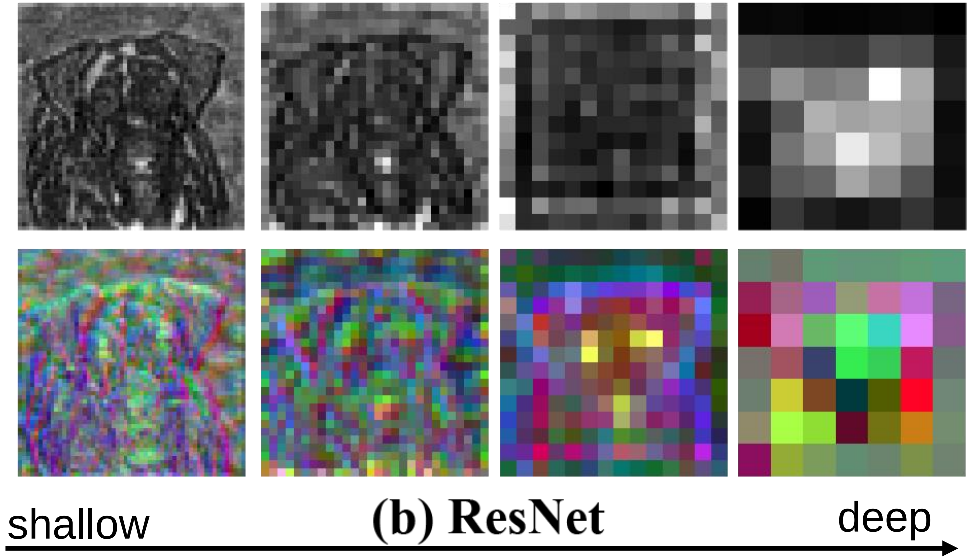
WAIC

PCA (Keys) across layers

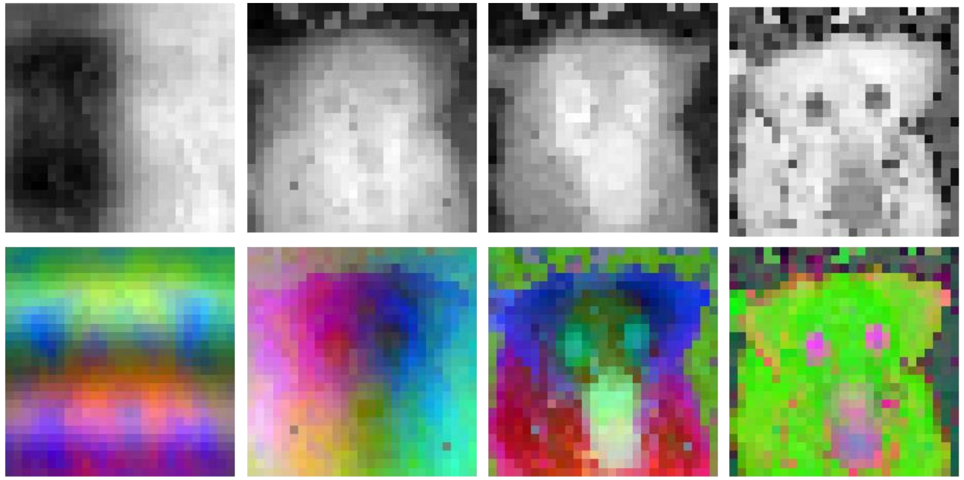
Self-supervised ViT (DINO-ViT)



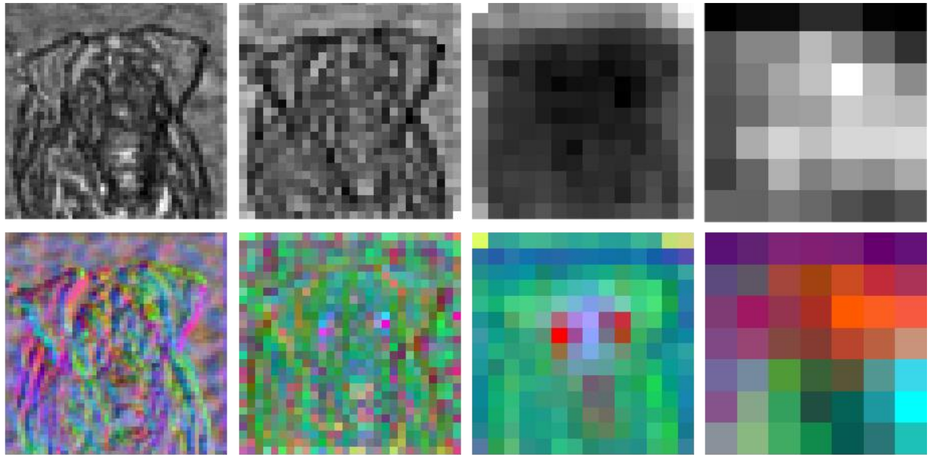
Self-supervised ResNet (DINO-ResNet)



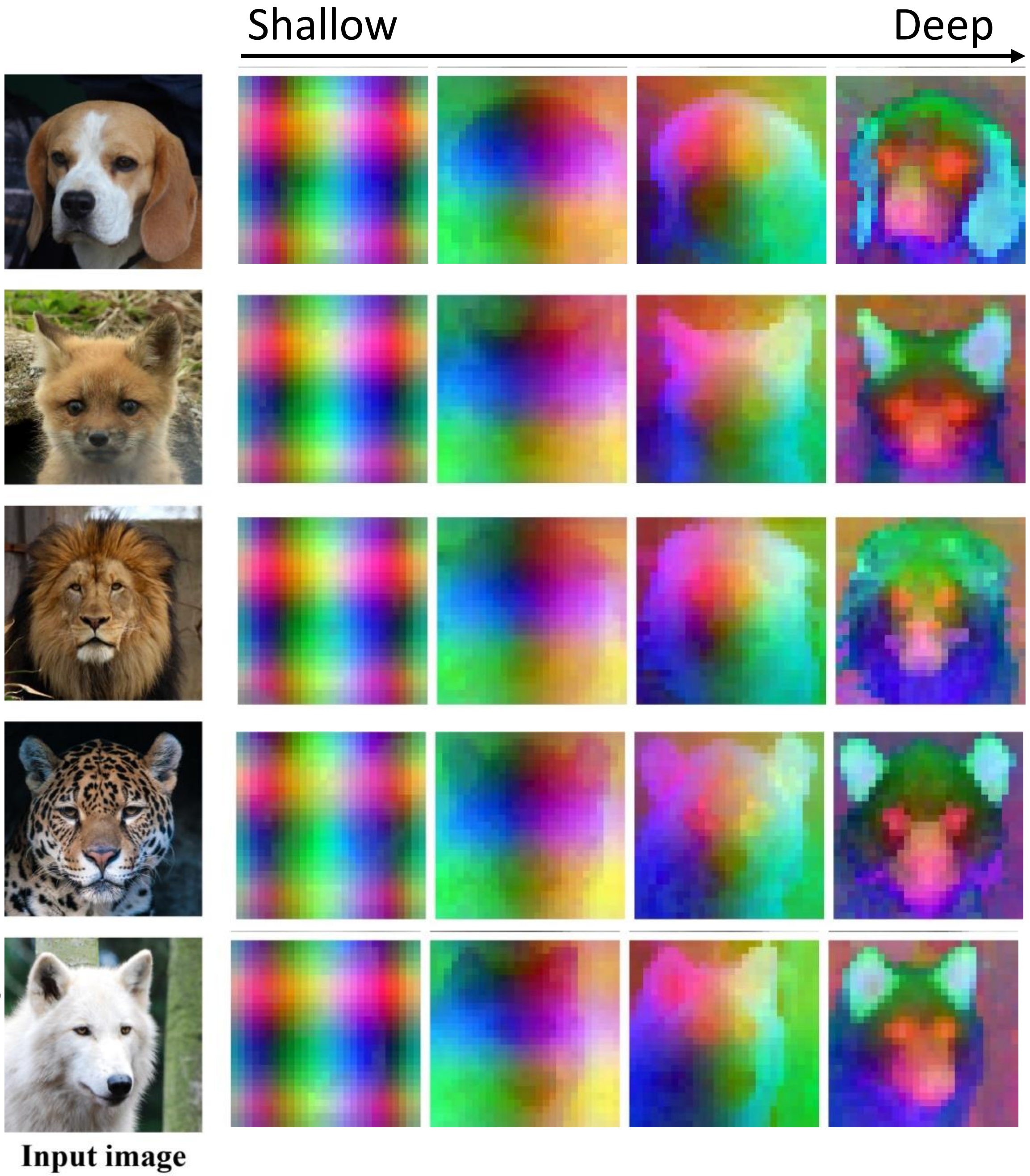
Supervised ViT



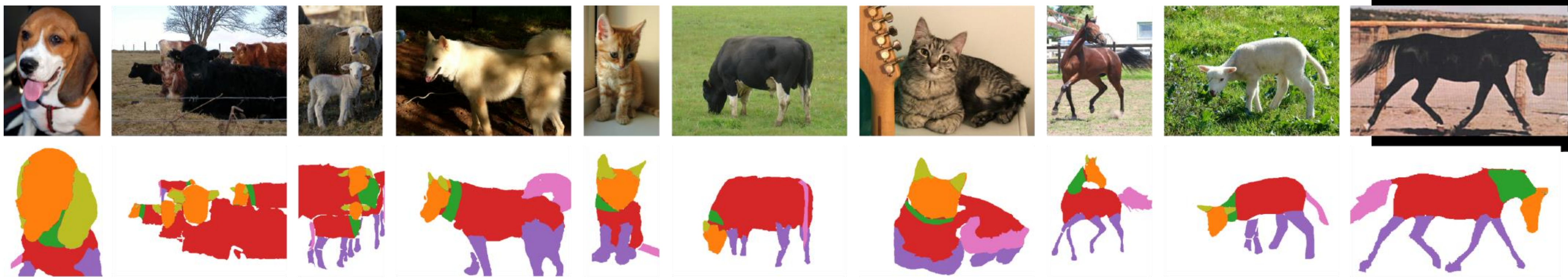
Supervised ResNet



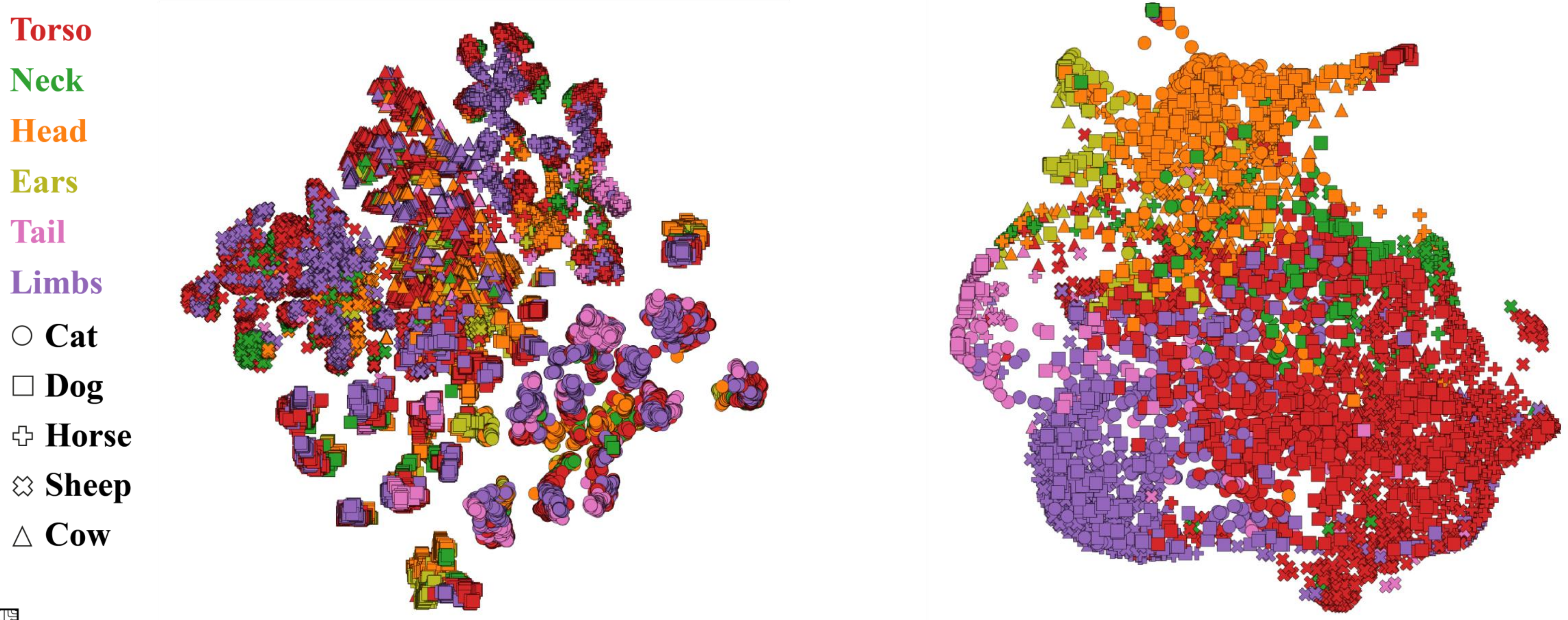
PCA – DiNO ViT



Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.,
“Deep ViT Features as dense visual descriptors”



(a) Sample images and ground truth parts



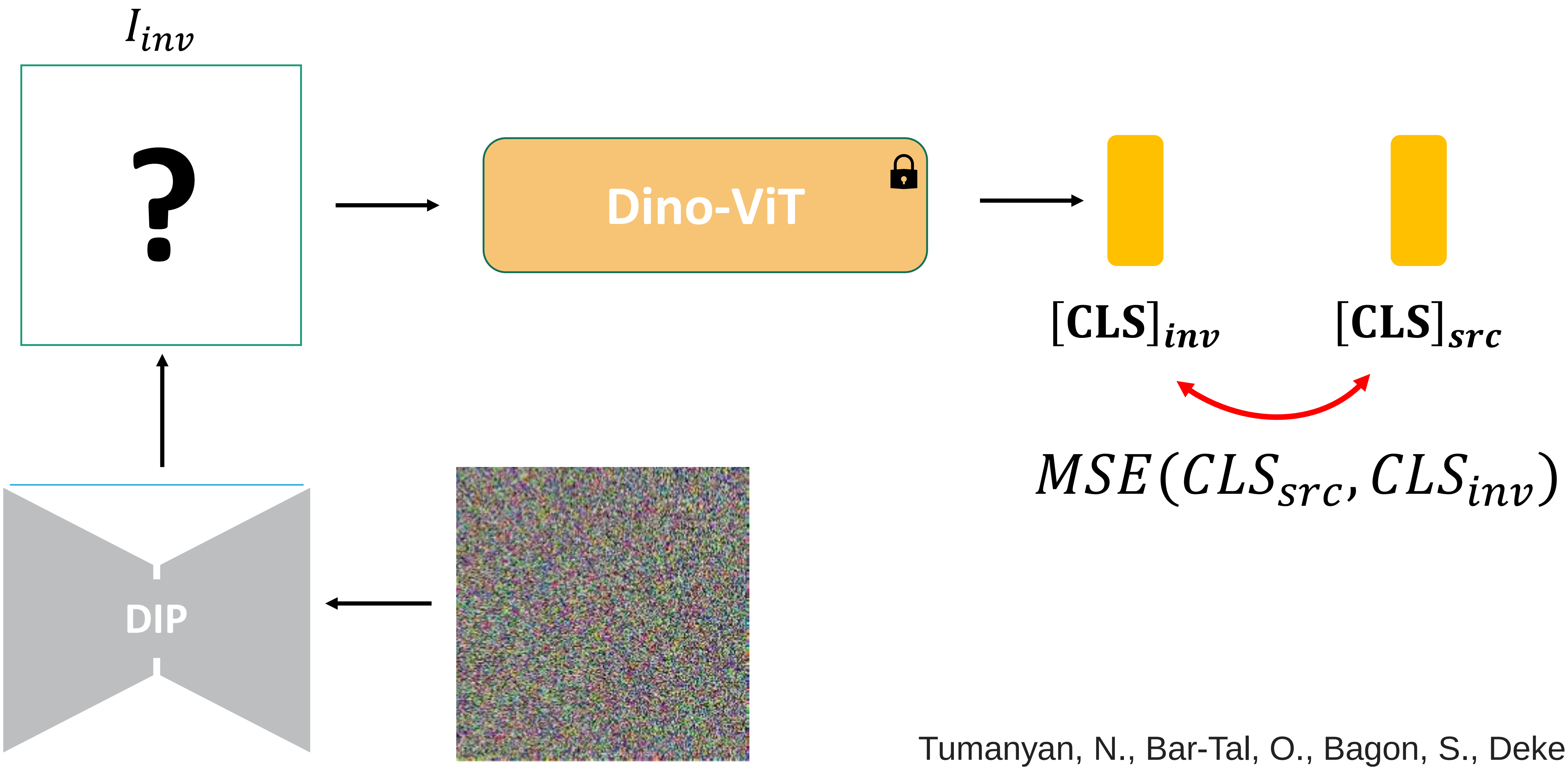
(c) Supervised ViT

(b) Self-Supervised ViT (DINO-ViT)

[CLS] as a **global appearance** representation

Tumanyan, N., Bar-Tal, O., Bagon, S., Dekel, T.
Splicing vit features for semantic appearance transfer (CVPR 2022)

[CLS] as a **global appearance** representation



Tumanyan, N., Bar-Tal, O., Bagon, S., Dekel, T.
Splicing vit features for semantic appearance transfer (CVPR 2022)

WAIC

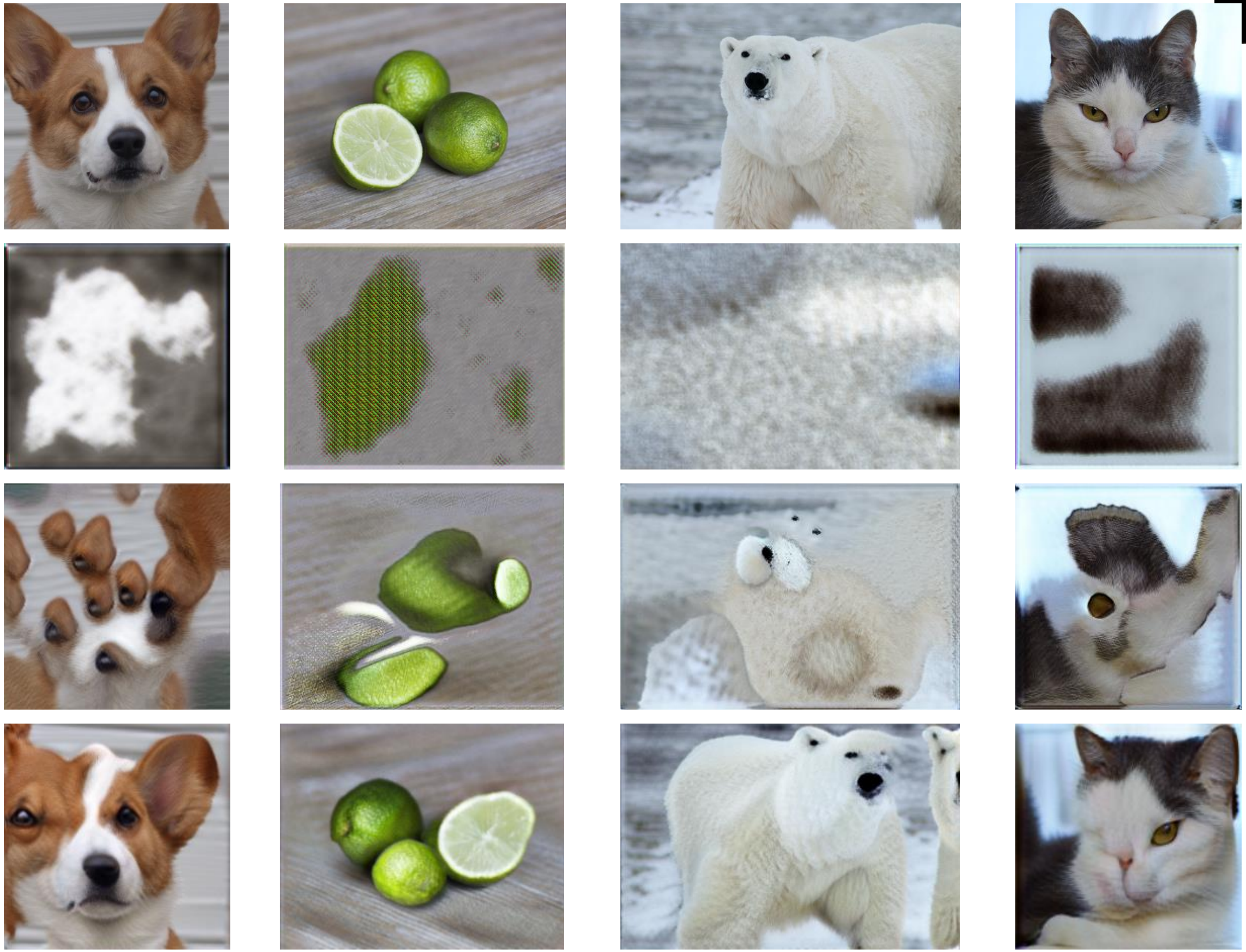
[CLS] as a **global appearance** representation

Input

layer **0**

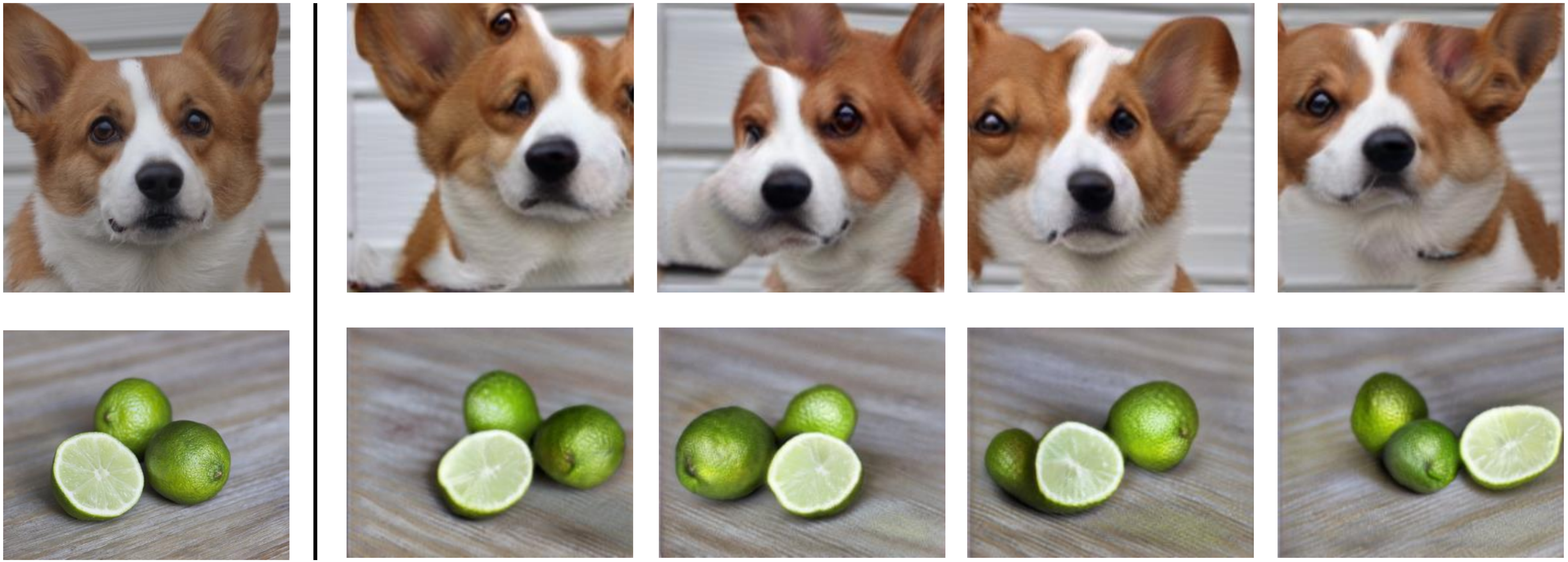
layer **3**

layer **11**



WAIC

[CLS] as a **global appearance** representation



Inversion run 1 | Inversion run 2 | Inversion run 3 | Inversion run 4

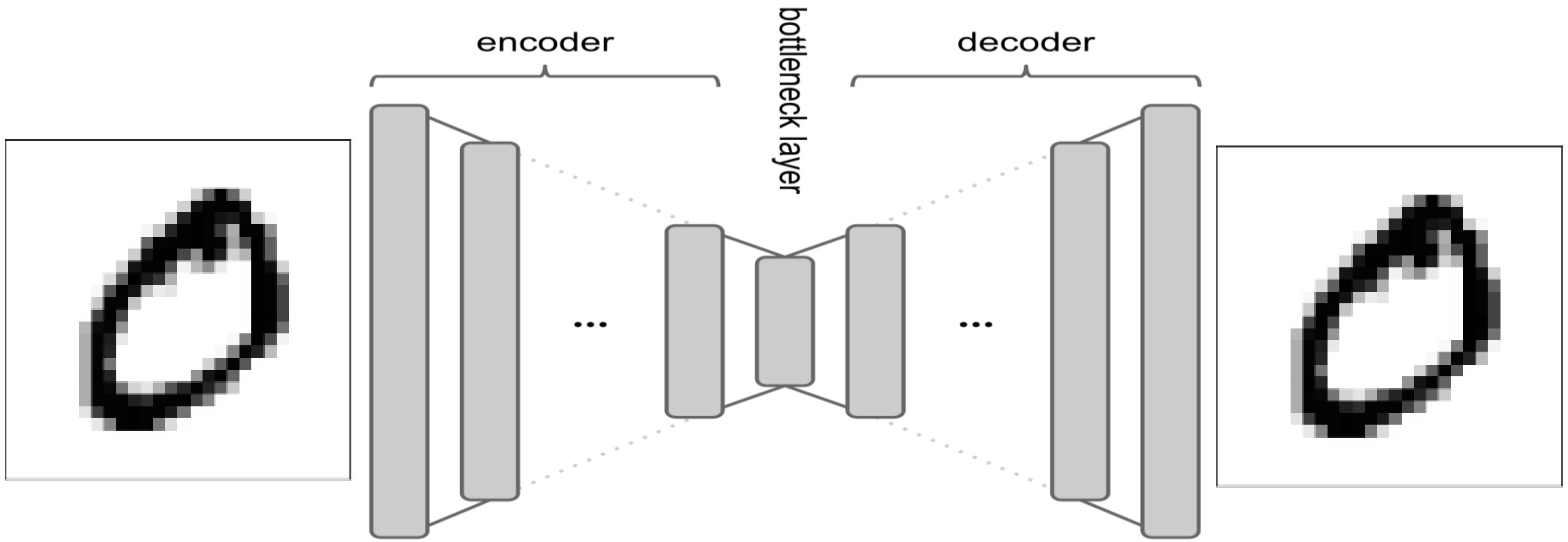
WAIC

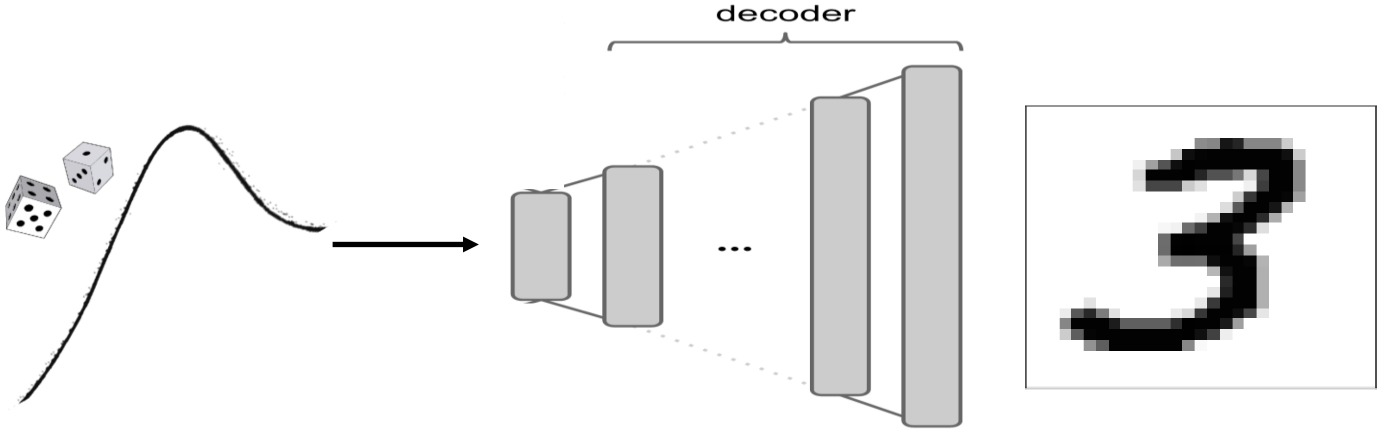
Topics

- self-DIstillation with NO labels
  - DINO
- **Masked Auto Encoders**
  - **MAE**
- Contrastive Language Image Pretraining
  - CLIP

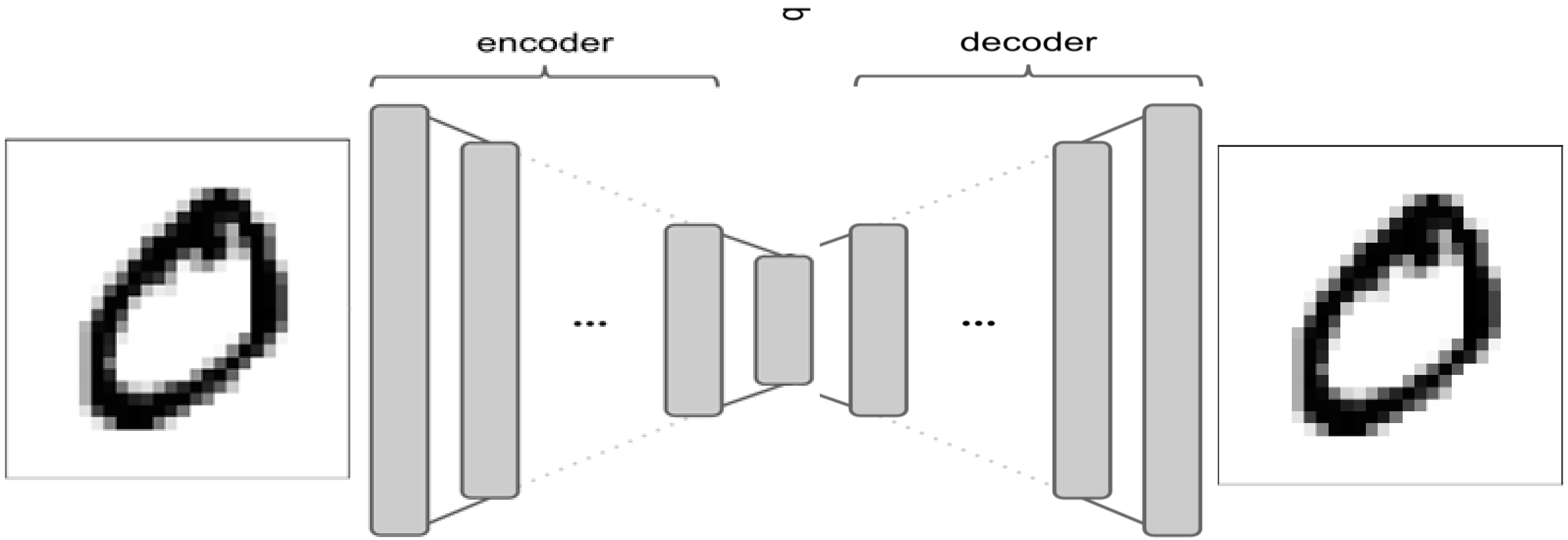
WAIC

Reminder - Auto Encoders





Self-supervised Auto Encoders



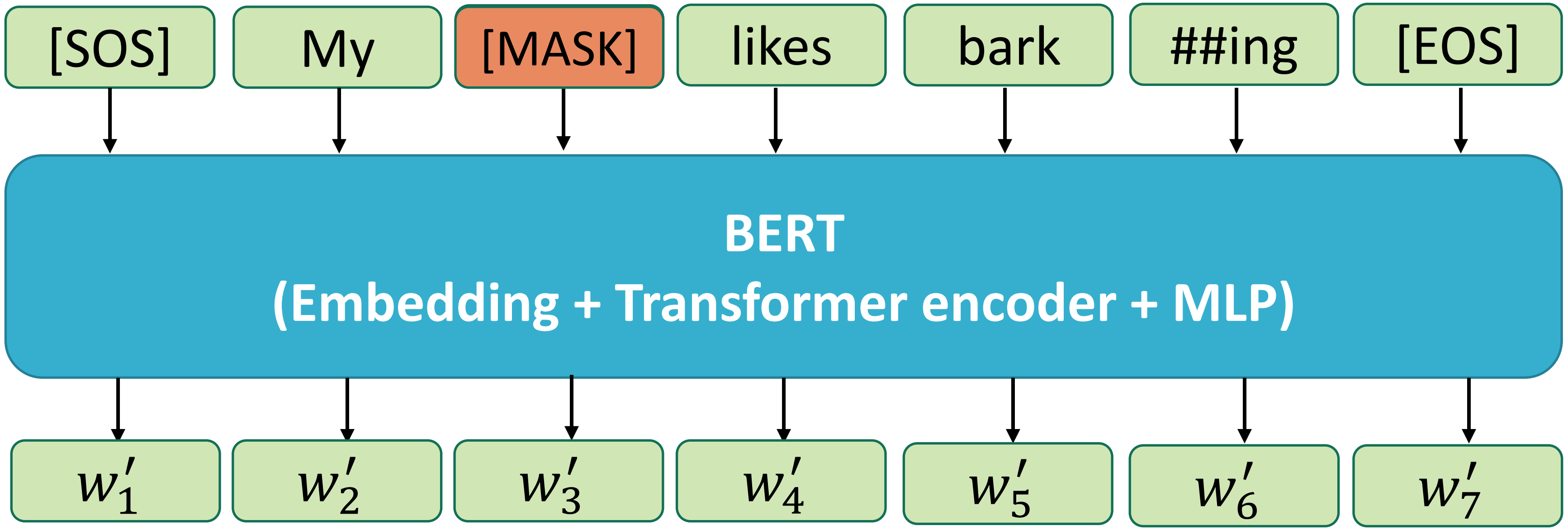
→ **Additional Supervised Layers** → ...

↑

Learned Representation

Masking Approach

- Classical SSL Approach
- Best showcase in NLP
- Input is masked (partially hidden) and then reconstructed



Masking Approach

BeIT

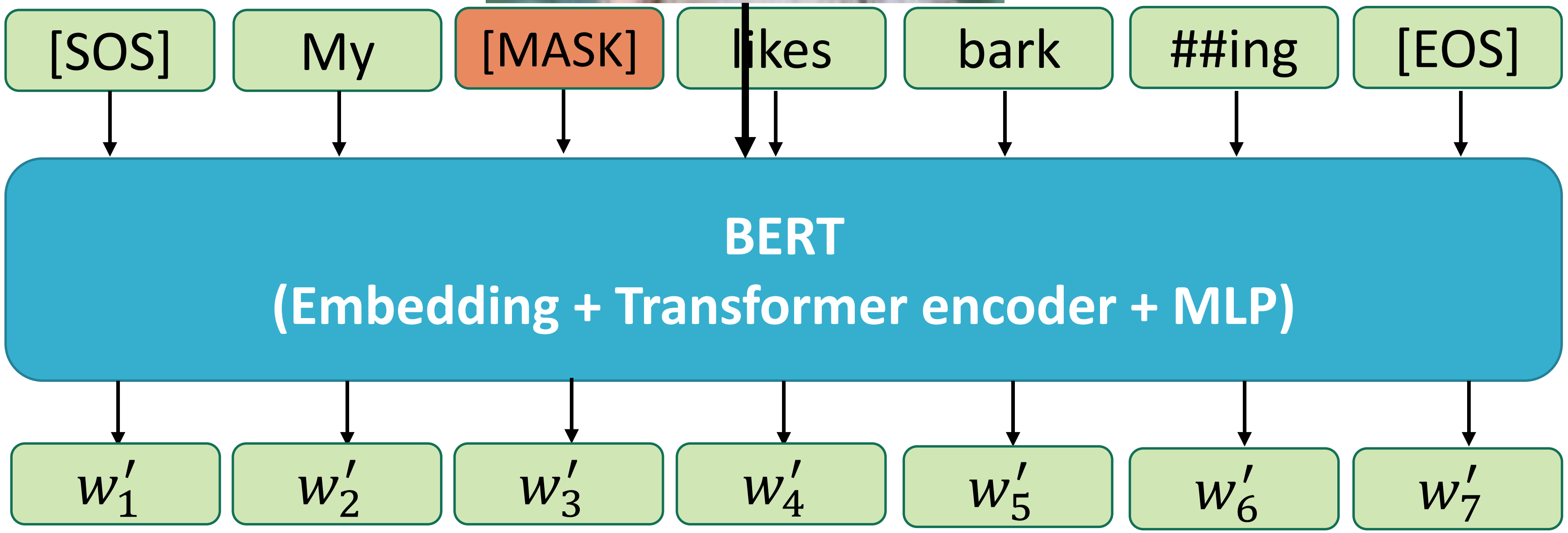
BERT Pre-Training of Image Transformers
(Bao et al., 2021)

SimMIM

A Simple Framework for Masked Image Modeling
(Xie et al., 2021)

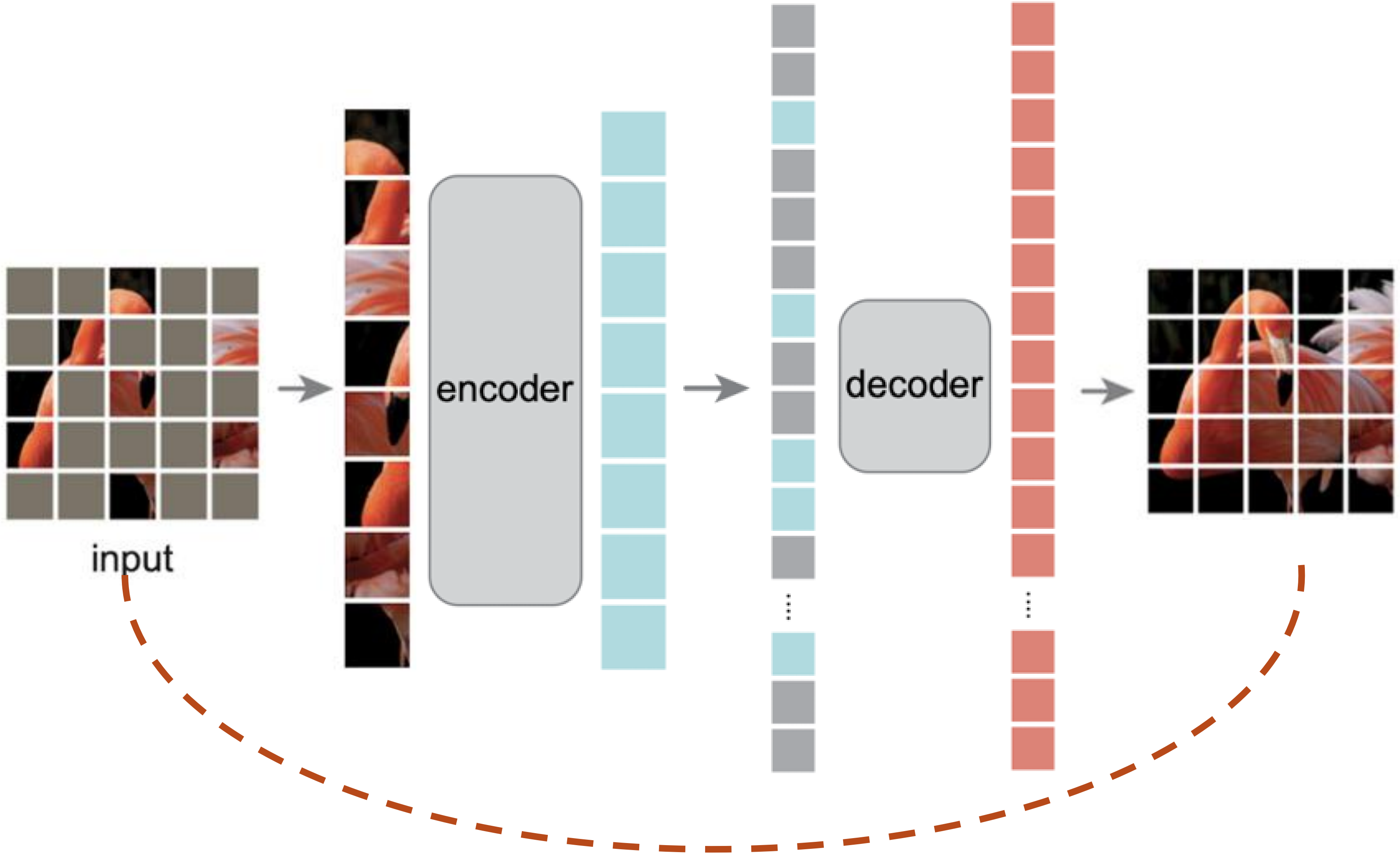
MAE

Masked Autoencoders Are Scalable Vision Learners
(He et al., 2021)



WAIC

MAE - Training

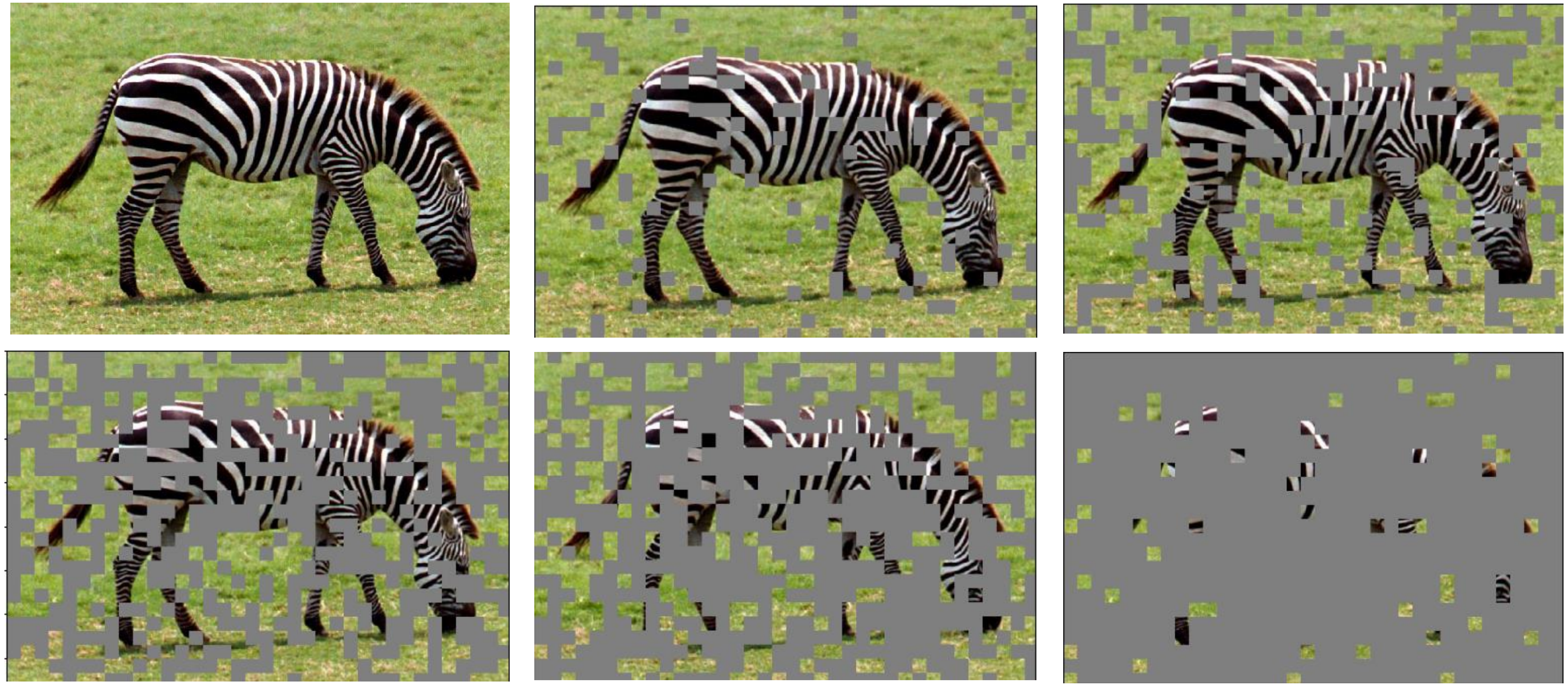


$L_{MSE}(input_{masked}, out_{masked})$

MAE – Masking Factor

- Masking factor is key in this approach
  - Reminder – A good SSL task is neither easy not ambiguous

No masked patches 15% 30%

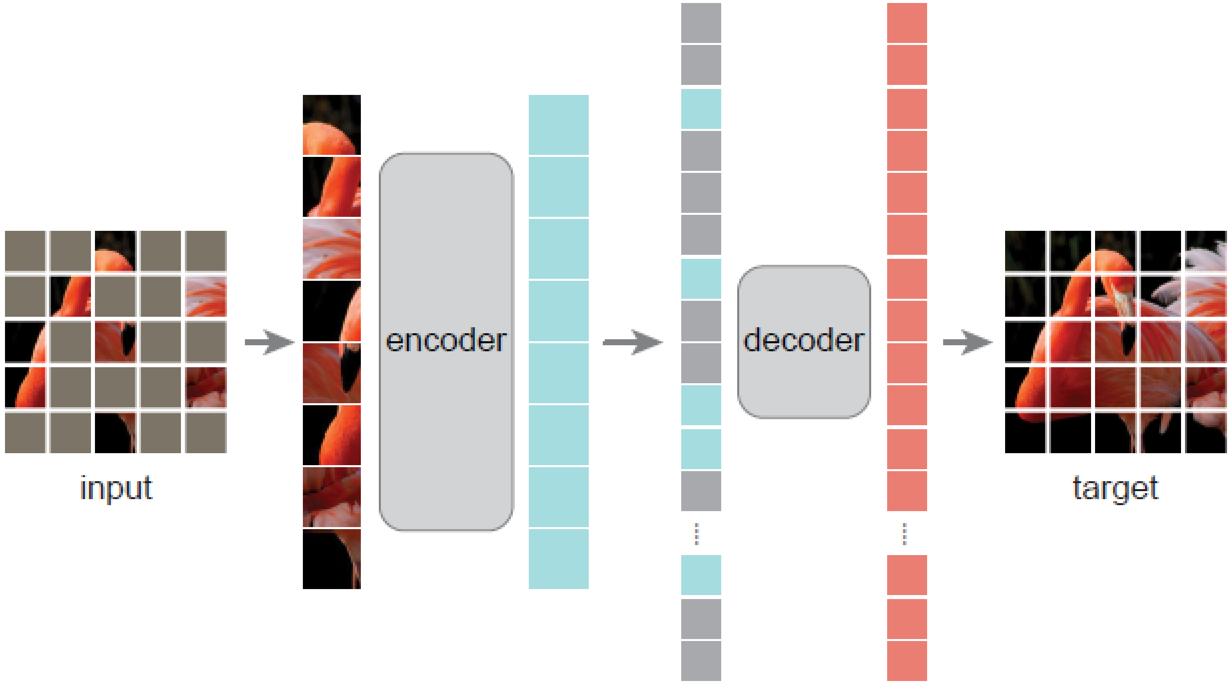


50% 70% 95%

MAE – Masking Factor

- Masking factor is key in this approach
  - Reminder – A good SSL task is neither easy not ambiguous

15% Masked tokens



<u>75% </u>Masked patches

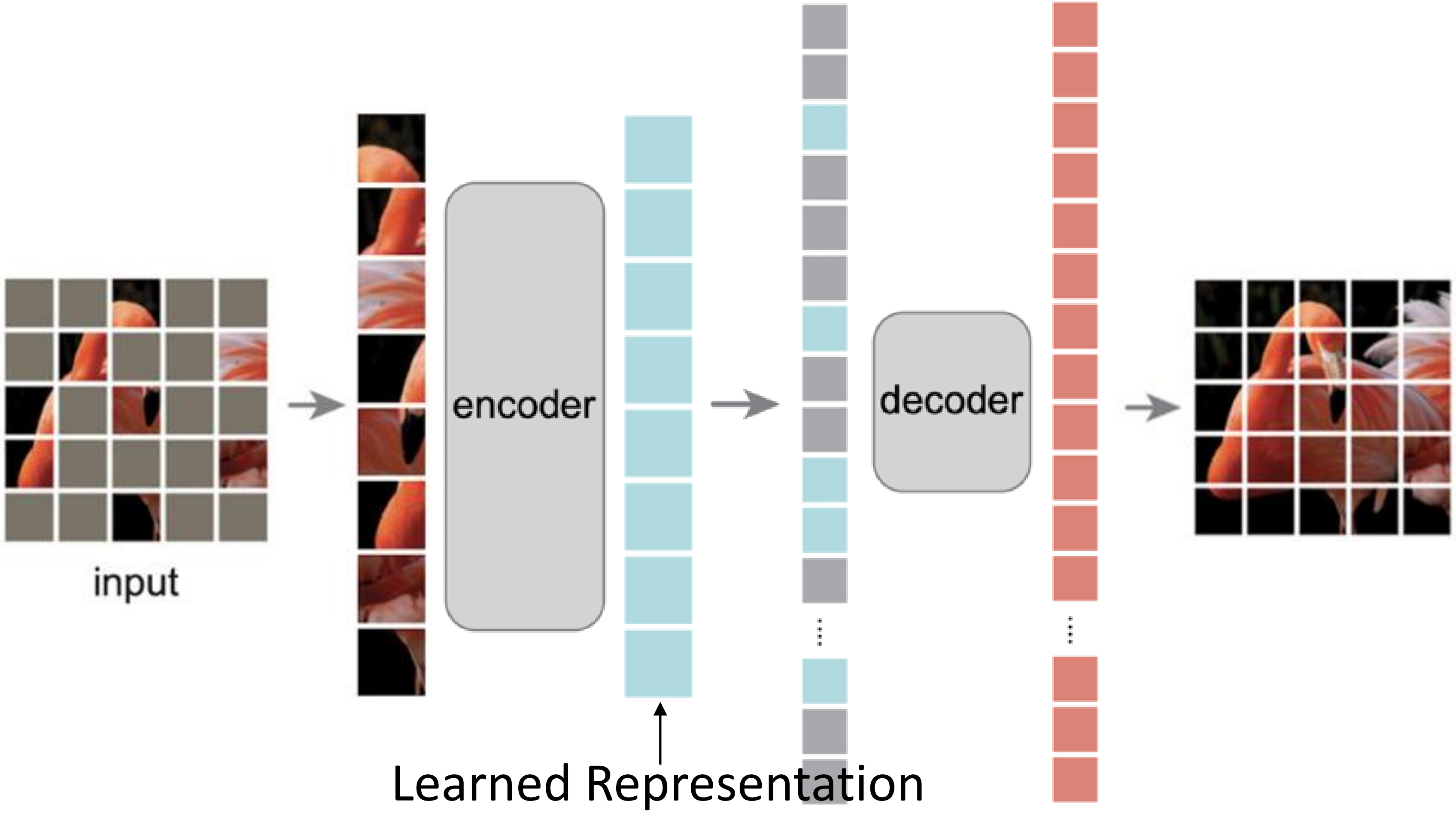
MAE – Reconstruction

Ground Truth Masking 75% Masking 85% Masking 95%

WAIC

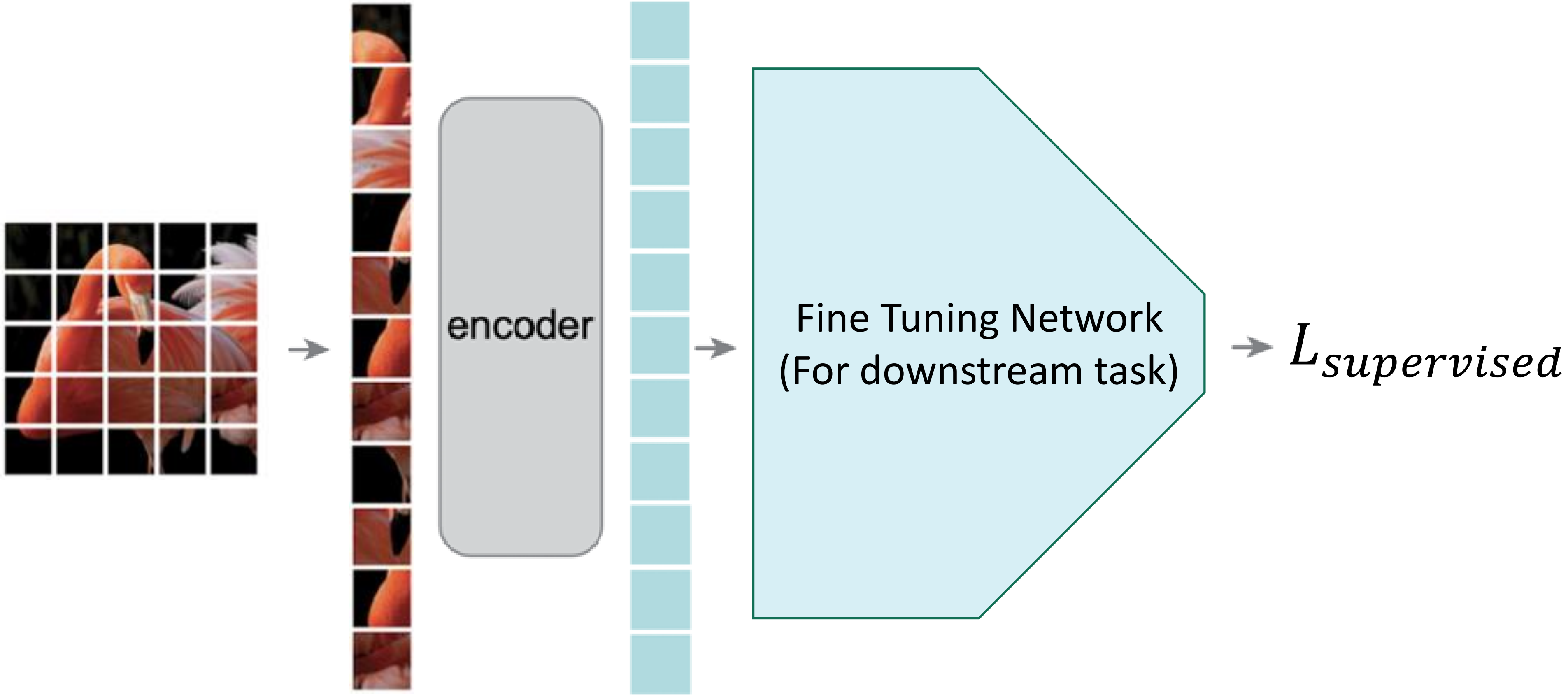
MAE – Fine Tuning

- Learned Representations allow for efficient fine-tuning



MAE – Fine Tuning

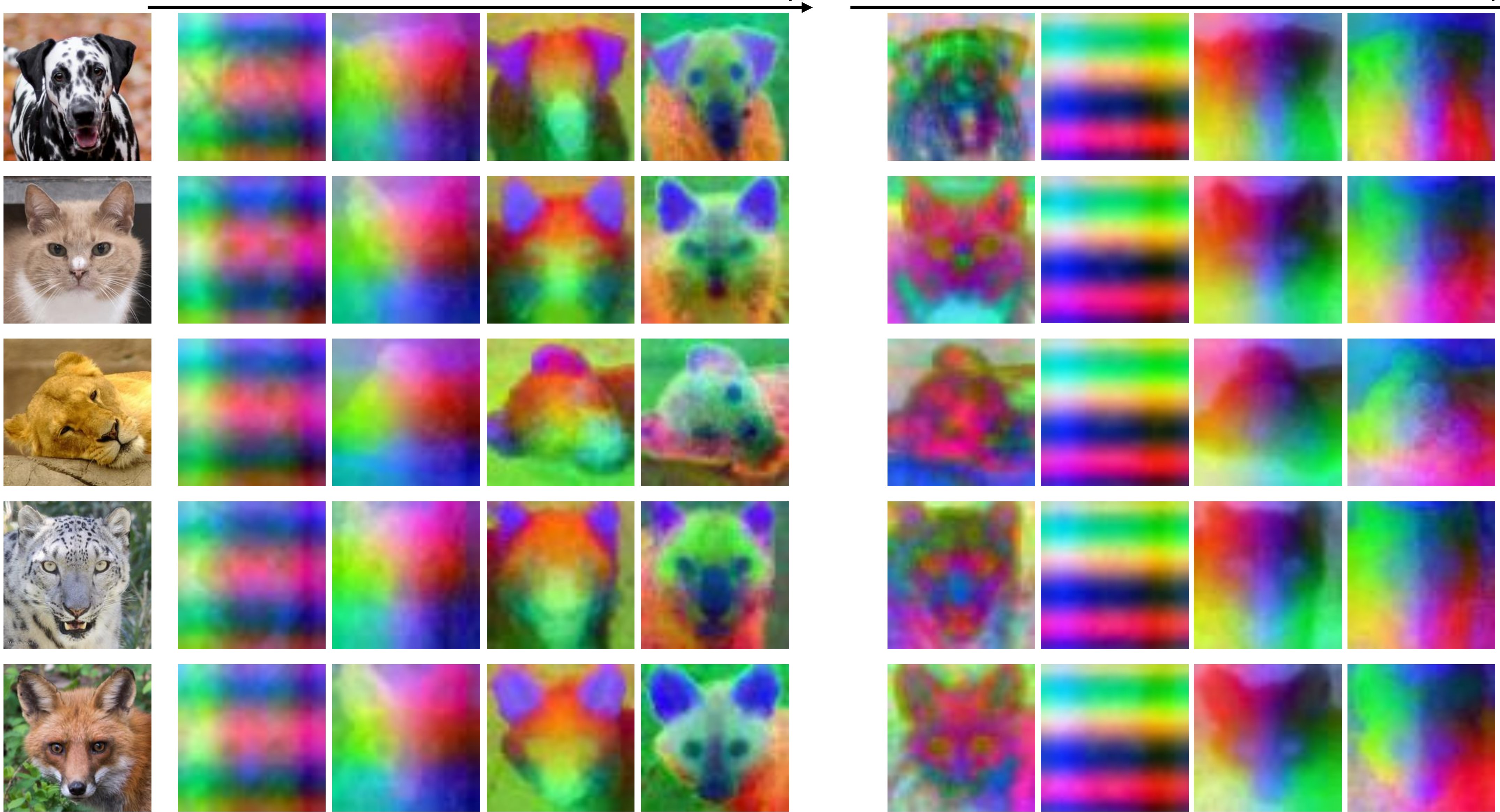
- Learned Representations allow for efficient fine-tuning



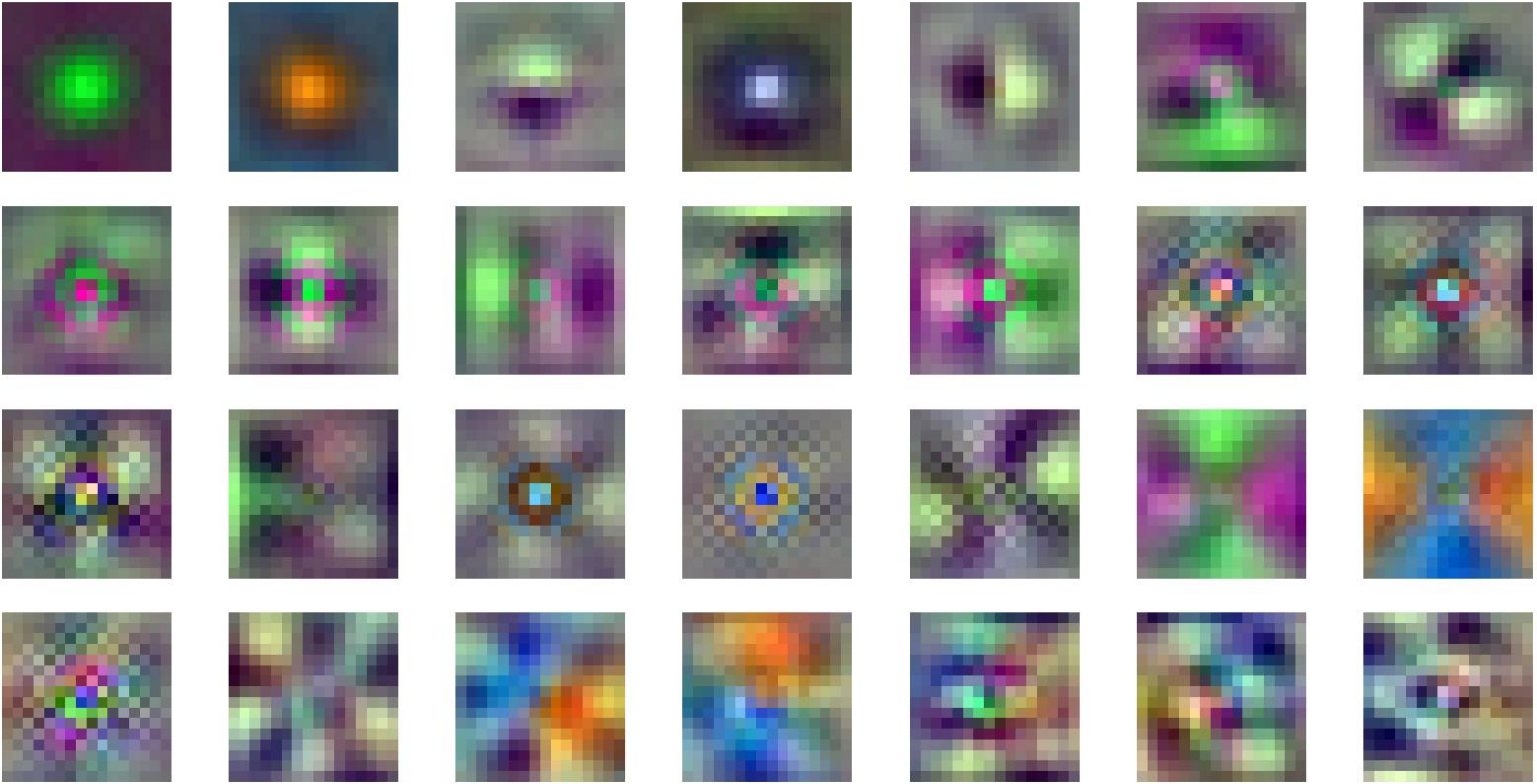
MAE Learned Representation

DINO-ViT | MAE

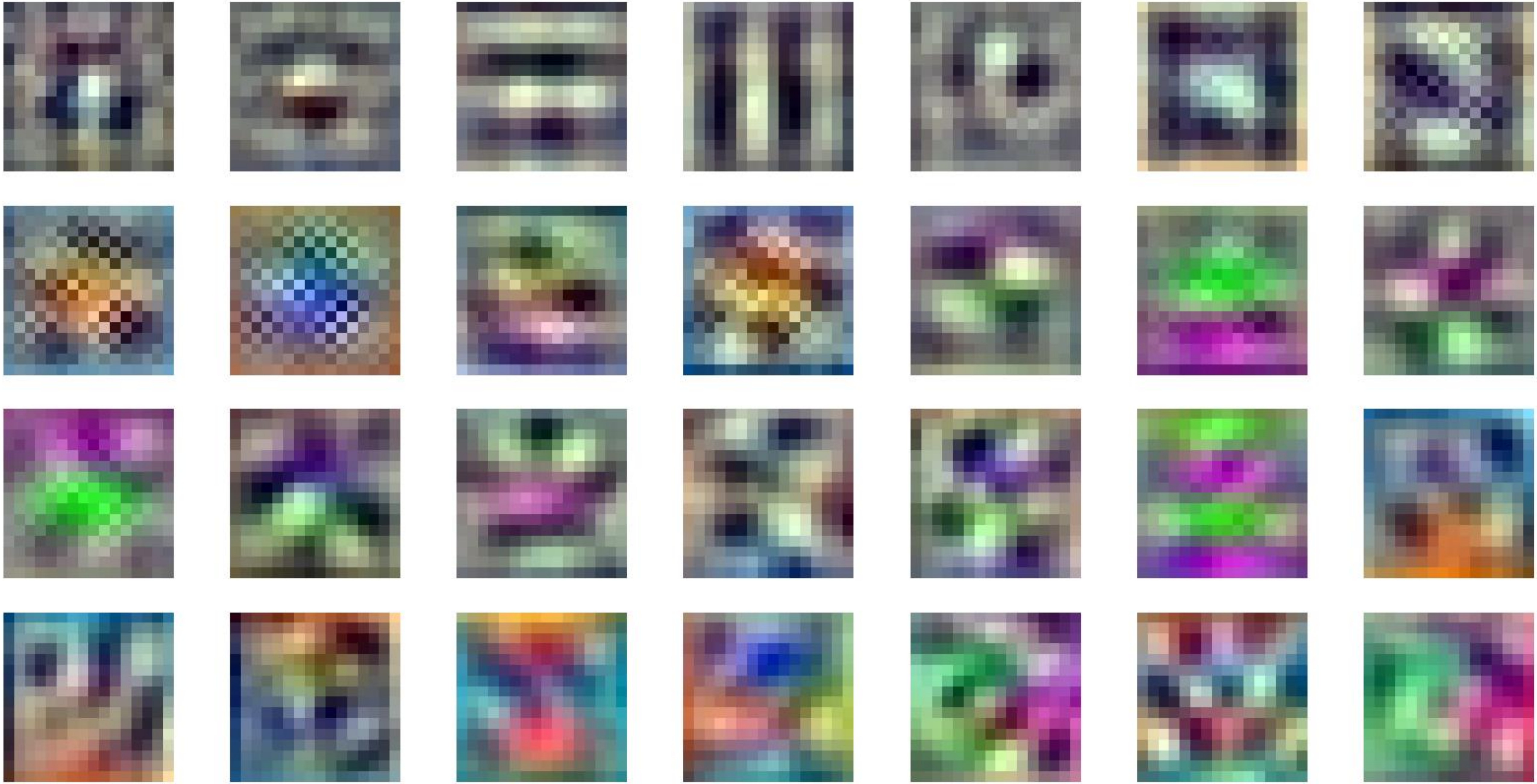
shallow → deep | shallow → deep



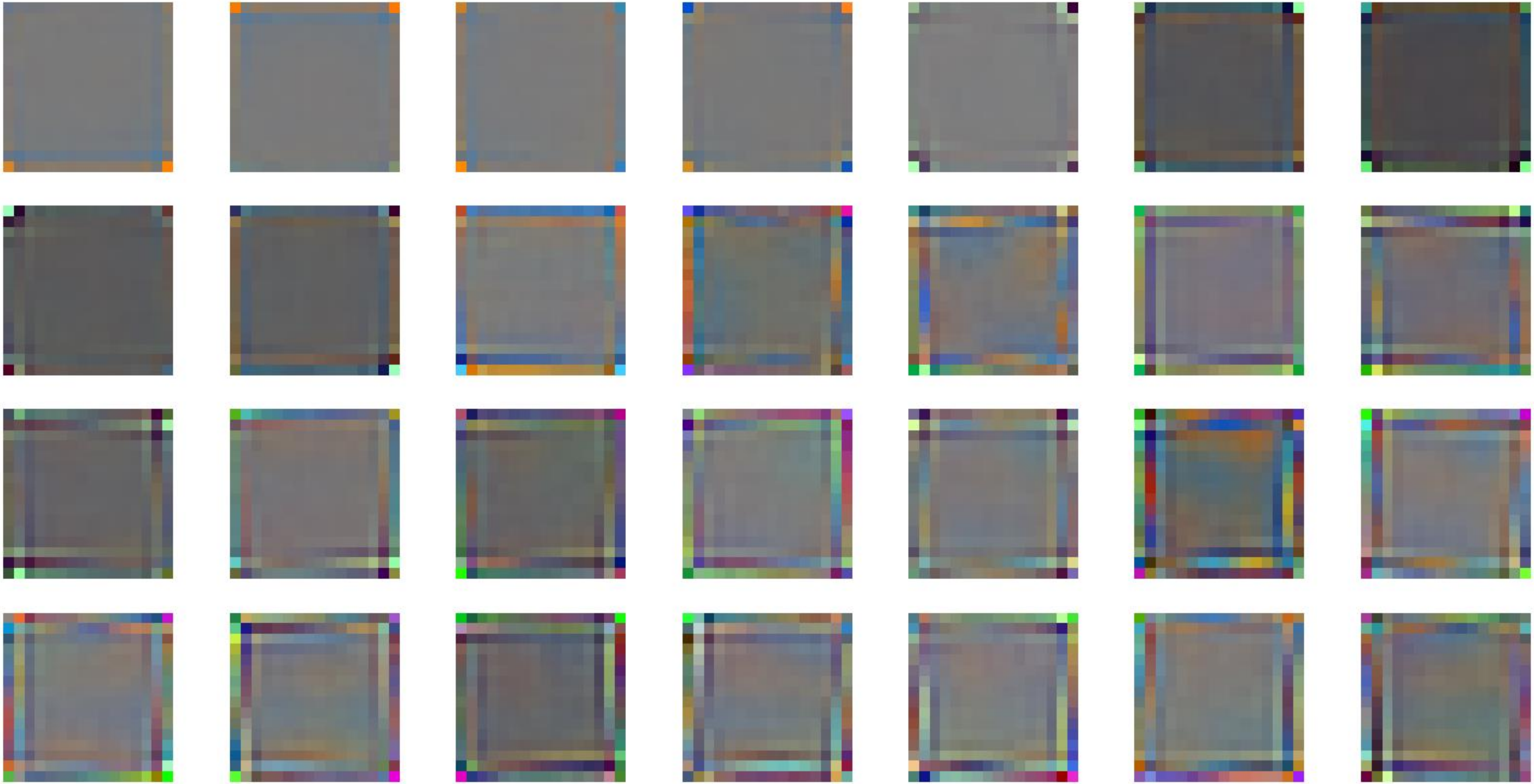
Supervised



DINO



MAE



Topics

- self-DIstillation with NO labels
  - DINO
- Masked Auto Encoders
  - MAE
- **Contrastive Language Image Pretraining**
  - **CLIP**

WAIC

CLIP

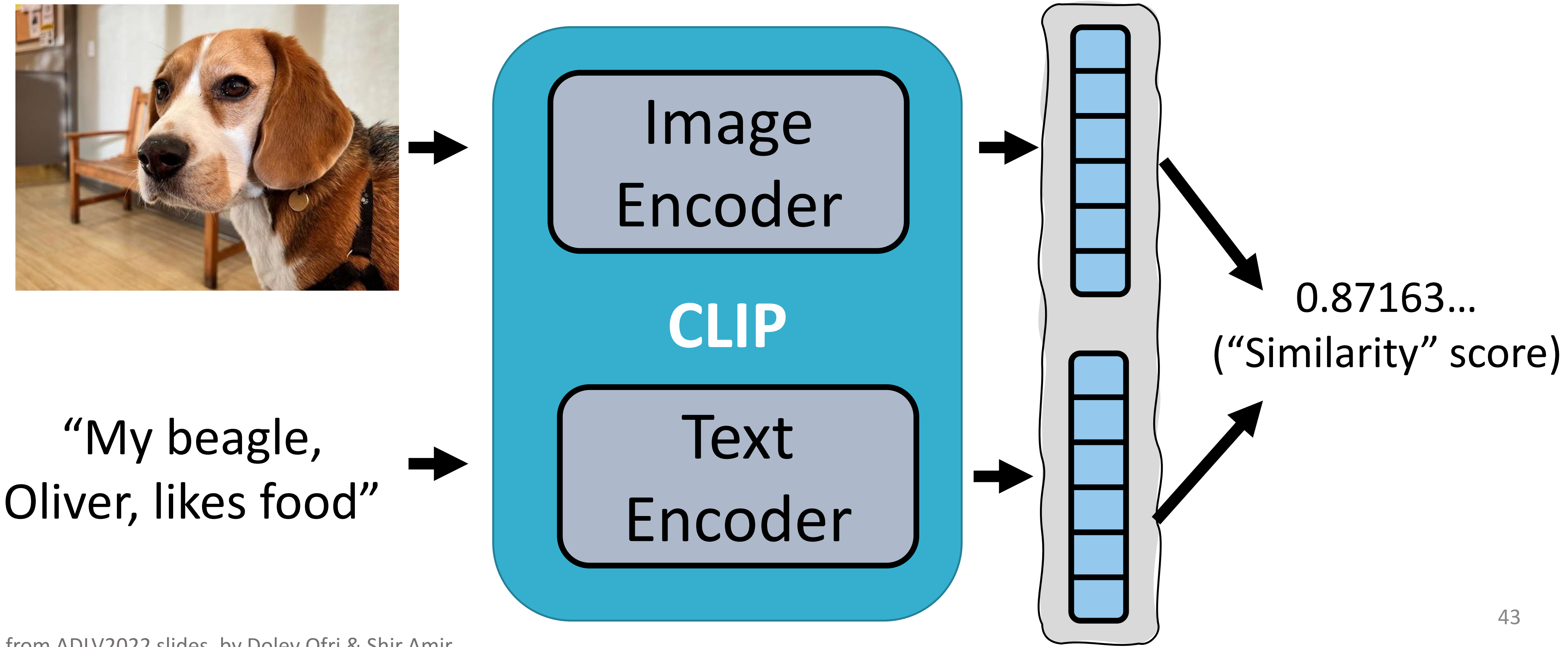
Computer Vision

Natural Language Processing

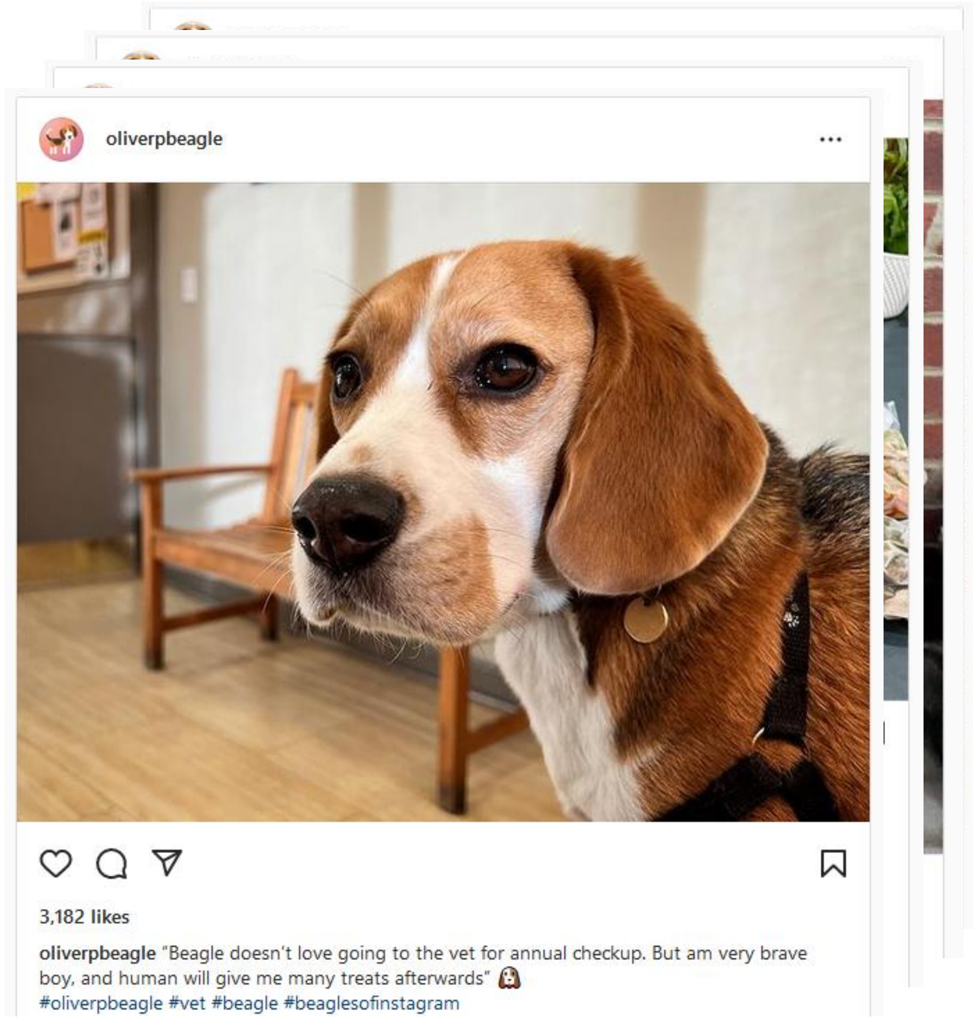
WAIC

CLIP

- **C**ontrastive **L**anguage **I**mage **P**retraining



Dataset

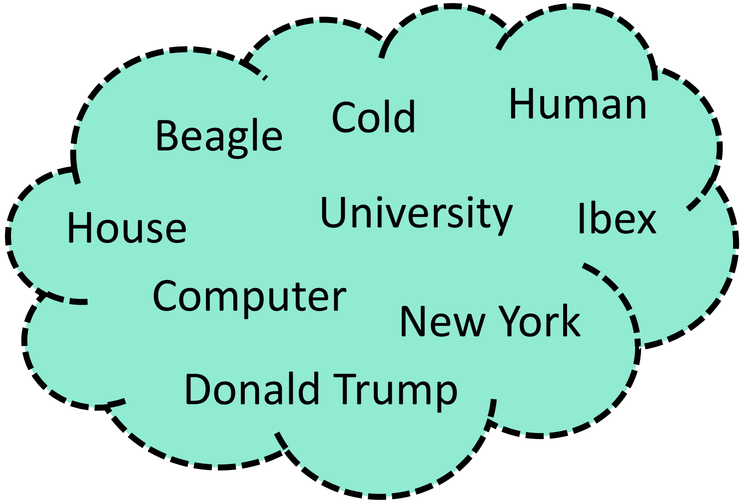

oliverpbeagle
3,182 likes
oliverpbeagle "Beagle doesn't love going to the vet for annual checkup. But am very brave boy, and human will give me many treats afterwards"
#oliverpbeagle #vet #beagle #beaglesofinstagram

MEN.WOMEN.BABIES.ELDERLY.
ARE STILL HELD HOSTAGE BY HAMAS
#BringThemHomeNow

Dataset

WIKIPEDIA
The Free Encyclopedia

→



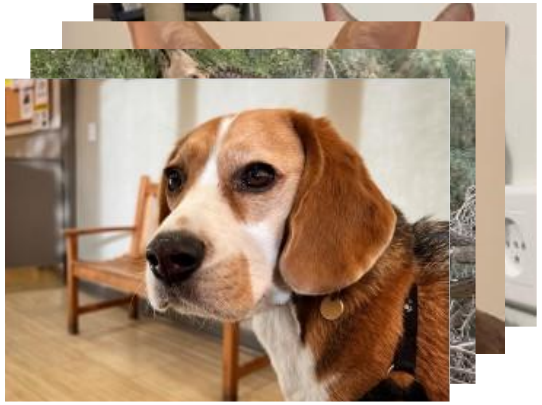
500K Queries

→ $\leq 20K$ pairs per query

( , **oliverpbeagle** “Beagle doesn’t love going to the vet for annual checkup. But am very brave boy, and human will give me many treats afterwards” 🐶 #oliverpbeagle #vet #beagle #beaglesofinstagram) $\times$ 400 Million

WAIC

CLIP - Training

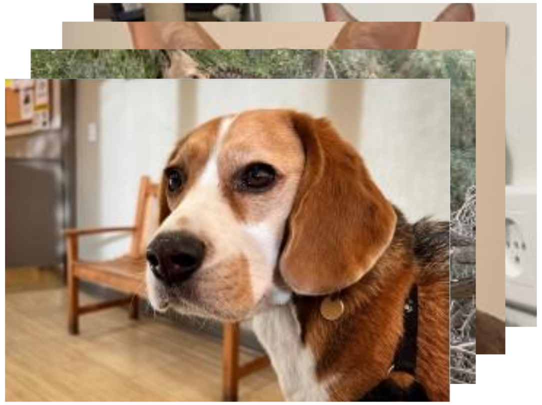


oliverpbeagle
“Beagle doesn’t love going to the vet for annual checkup. But am very brave boy…”

CLIP

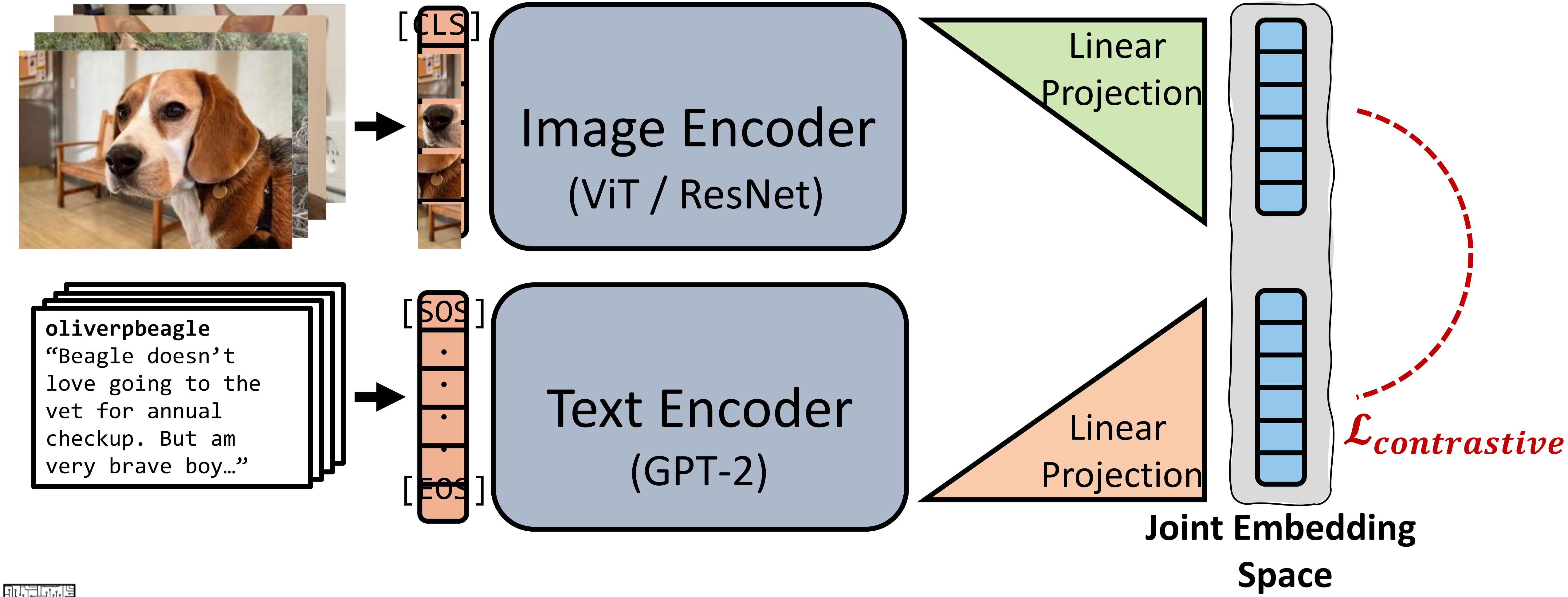
MEN.WOMEN.BABIES.ELDERLY.
ARE STILL HELD HOSTAGE BY HAMAS
#BringThemHomeNow

CLIP - Training



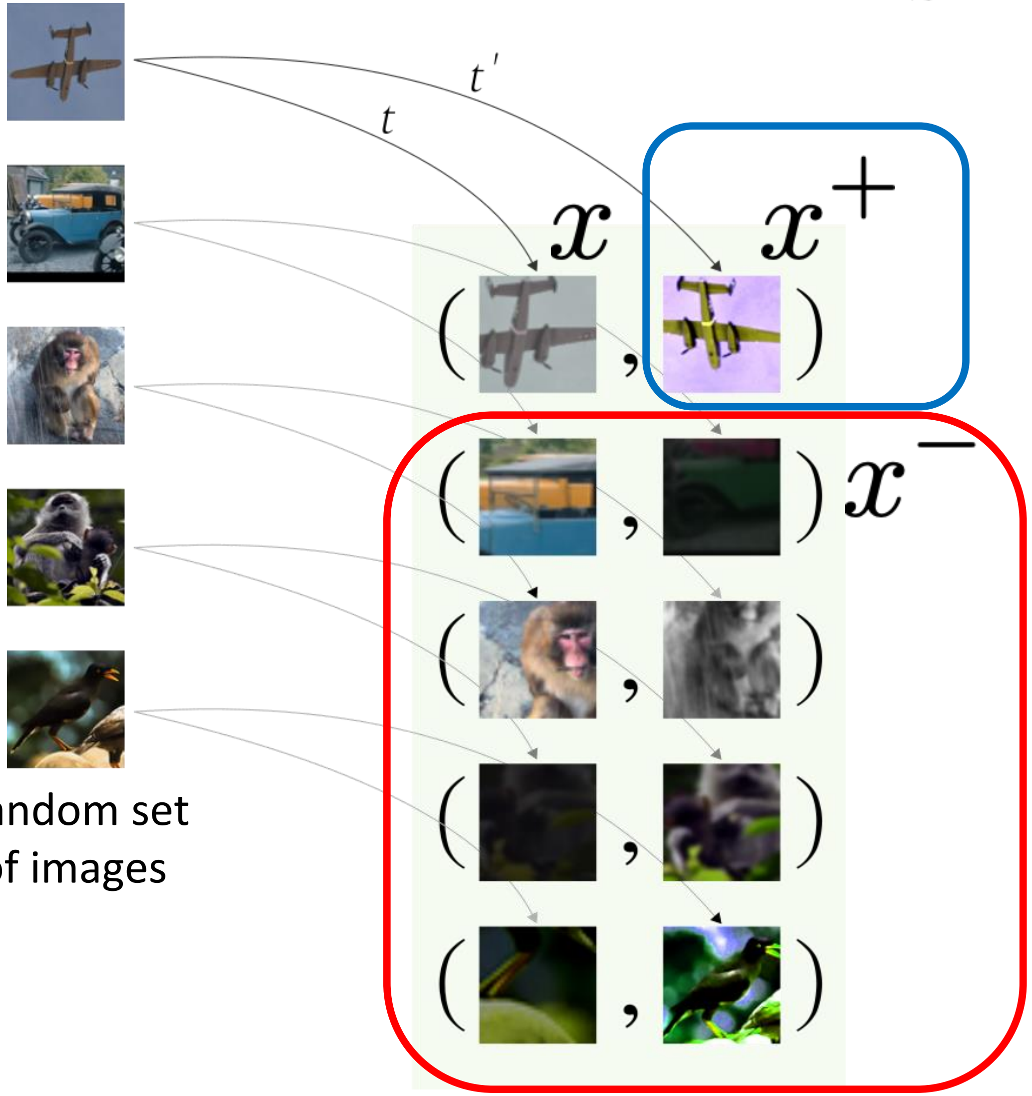
oliverpbeagle
“Beagle doesn’t love going to the vet for annual checkup. But am very brave boy…”

CLIP - Training

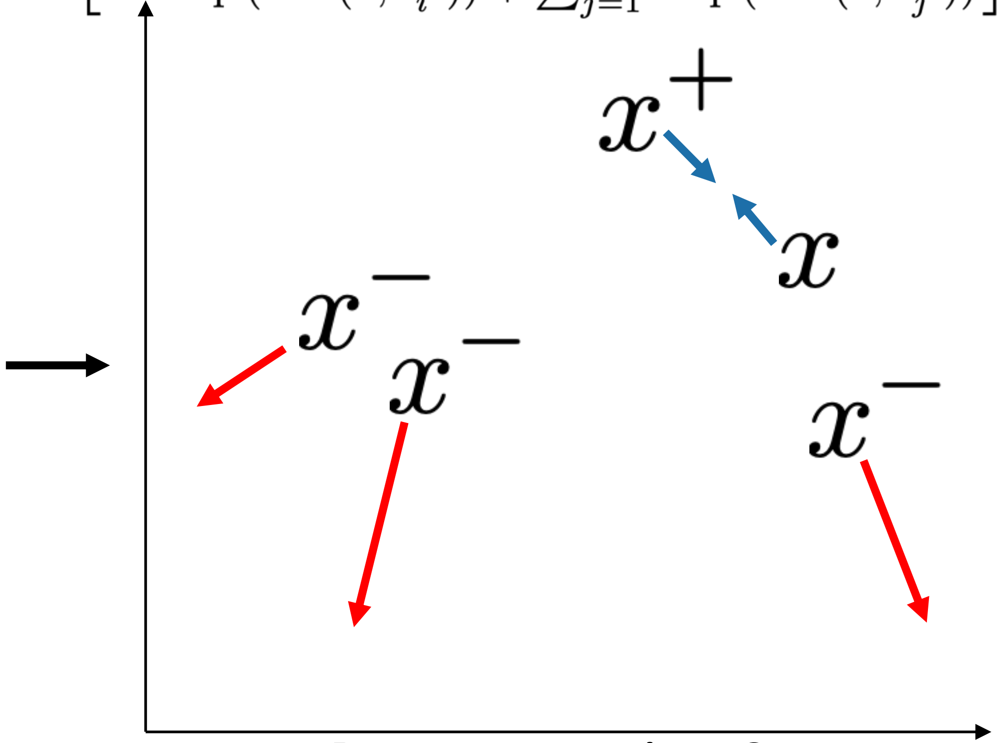


Reminder - Contrastive loss

$$\mathcal{L} = -E_X\left[\log \frac{\exp\left(sim(z, z_i^+)\right)}{\exp\left(sim(z, z_i^+)\right) + \sum_{j=1}^{N-1} \exp\left(sim(z, z_j^-)\right)}\right]$$



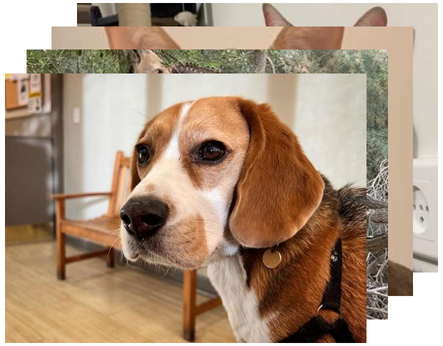
Random set of images



Representation Space

CLIP - Contrastive loss

$$\mathcal{L}_{infoNCE} = \sum_{i=1}^{N} -\log \frac{\exp(I_i T_i)}{\sum_{j=1}^{N} \exp(I_i T_j)}$$



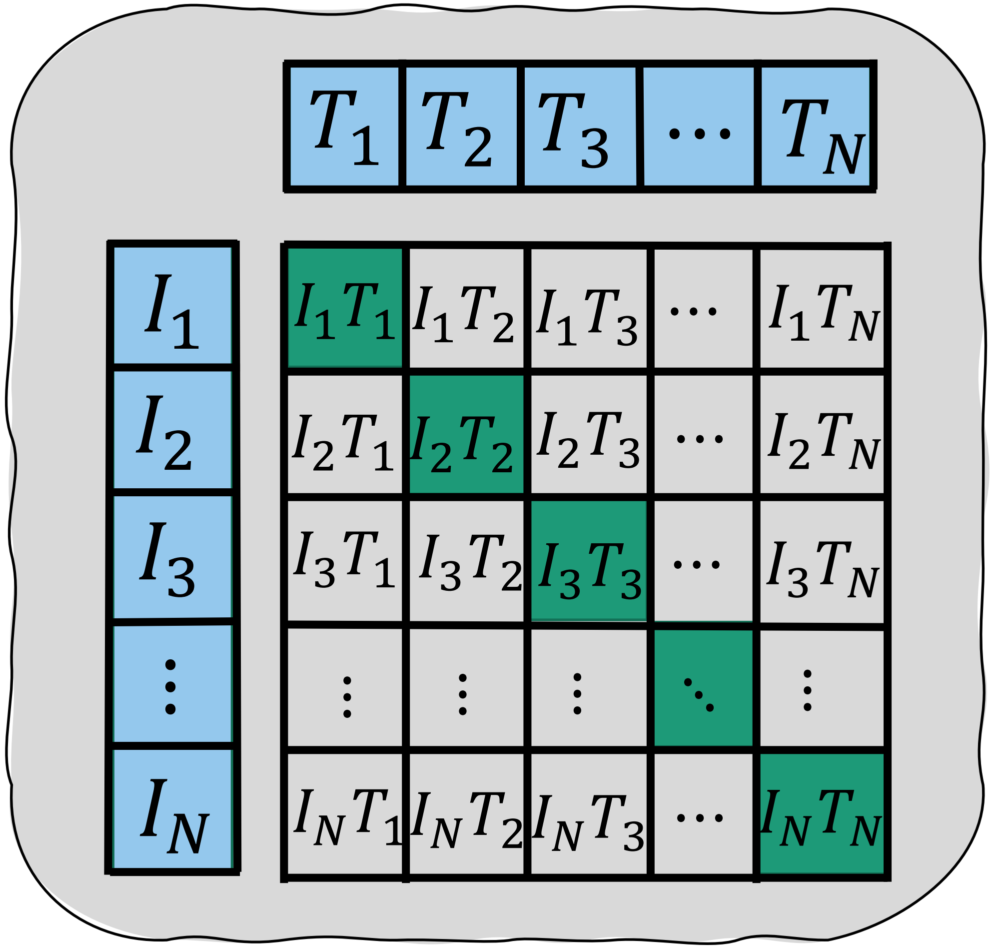
encode images

$I_{1\ldots N}$

oliverpbeagle
“Beagle doesn’t love going to the vet for annual checkup. But am very brave boy…”

encode texts

$T_{1\ldots N}$

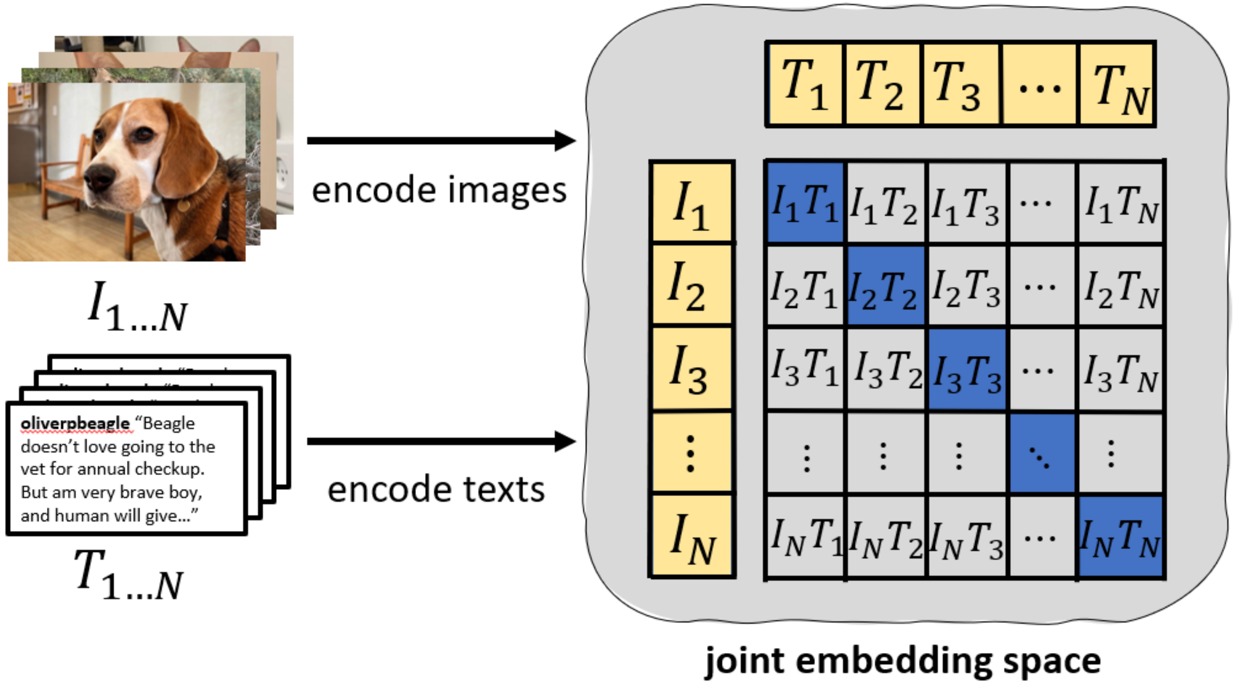


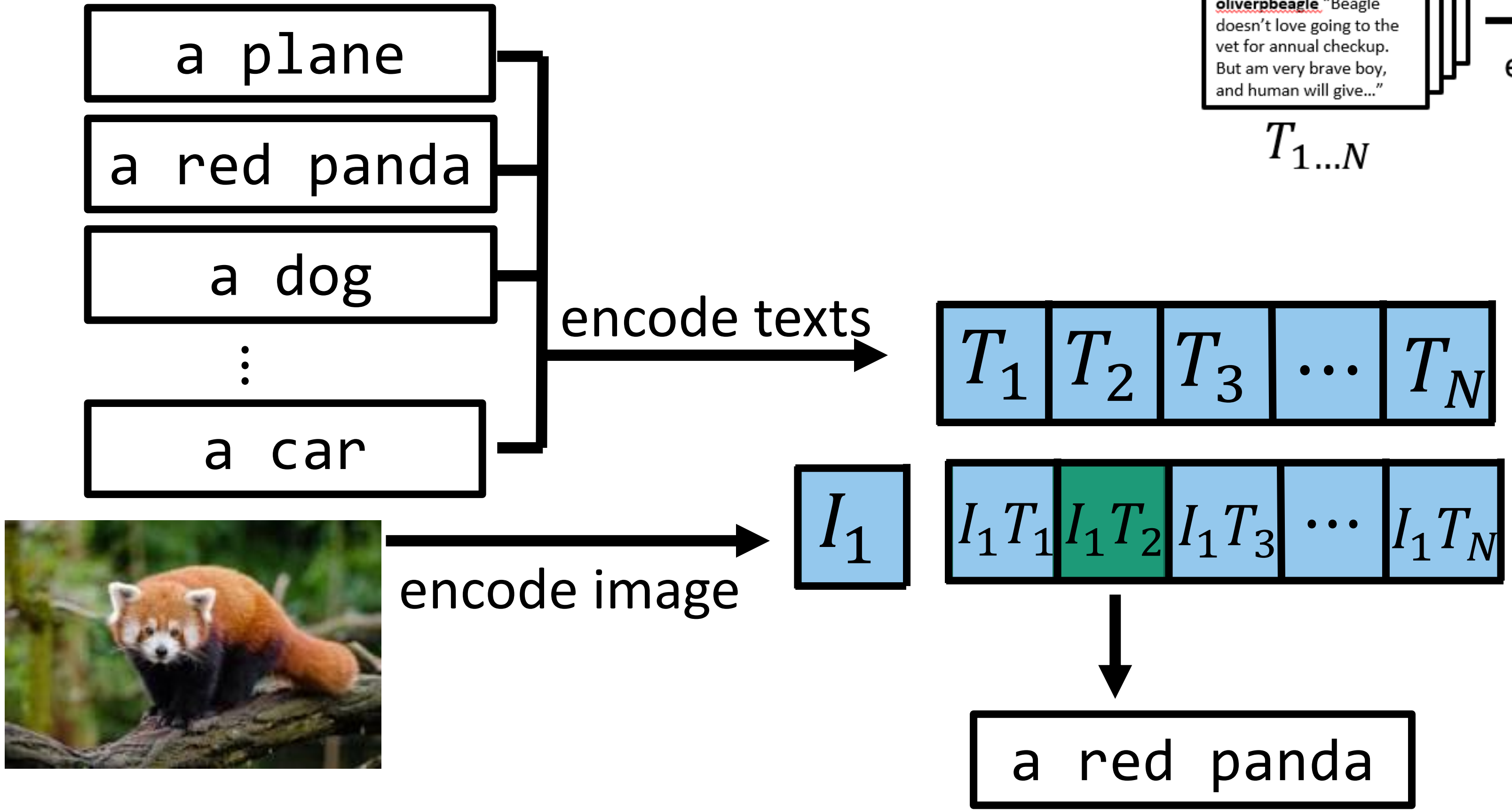
joint embedding space

What is this good for?

- “Zero shot” learning
  - Classification

Classification



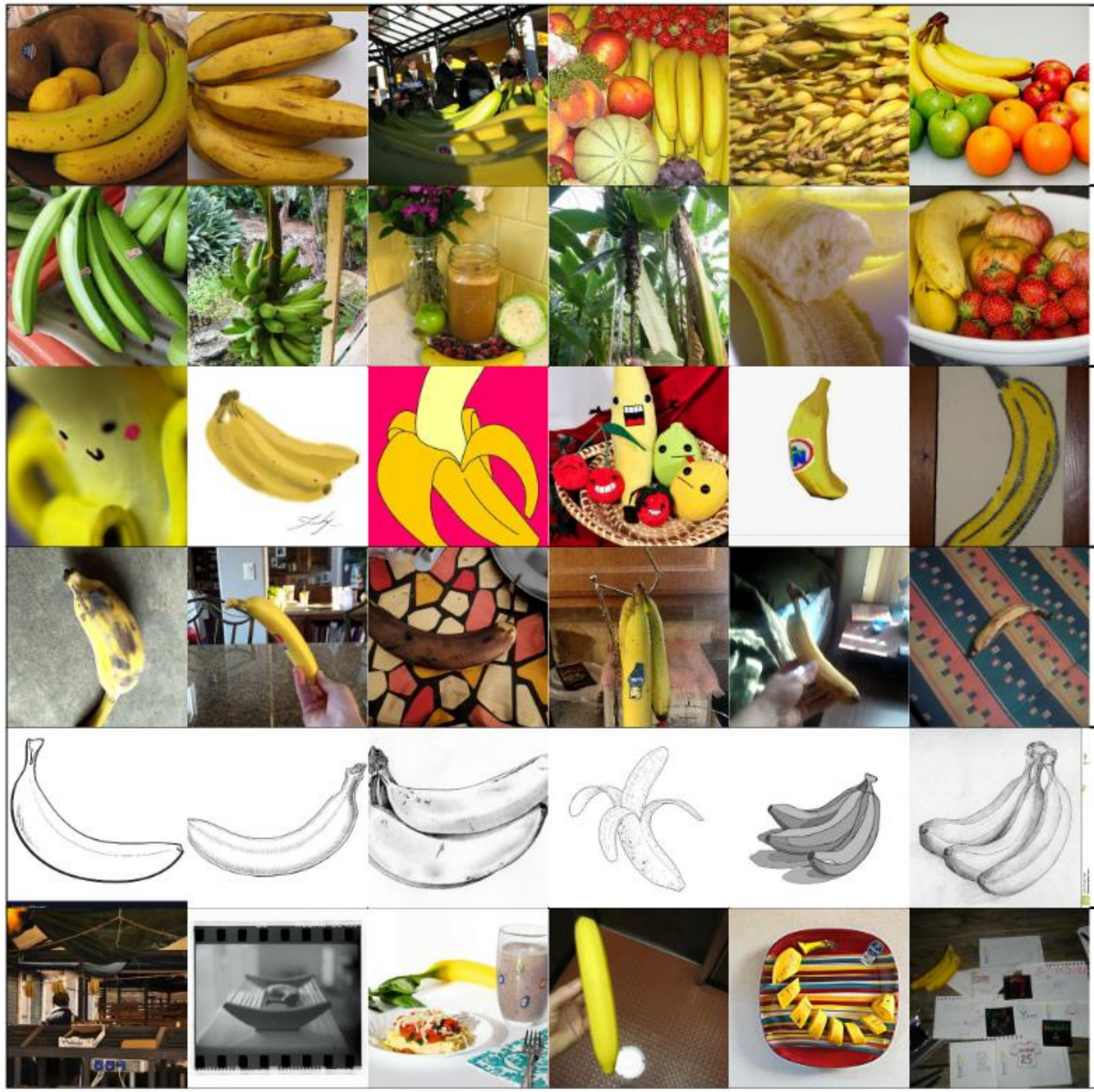


"Zero" Shot Classification

Adapted from ADLV2022 slides, by Dolev Ofri & Shir Amir

Robustness to Different Domains

Dataset Examples



	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	**76.2**	**76.2**	0%
ImageNetV2	64.3	**70.1**	+5.8%
ImageNet-R	37.7	**88.9**	+51.2%
ObjectNet	32.6	**72.3**	+39.7%
ImageNet Sketch	25.2	**60.2**	+35.0%
ImageNet-A	2.7	**77.1**	+74.4%

Adapted from ADLV2022 slides, by Dolev Ofri & Shir Amir

Classification

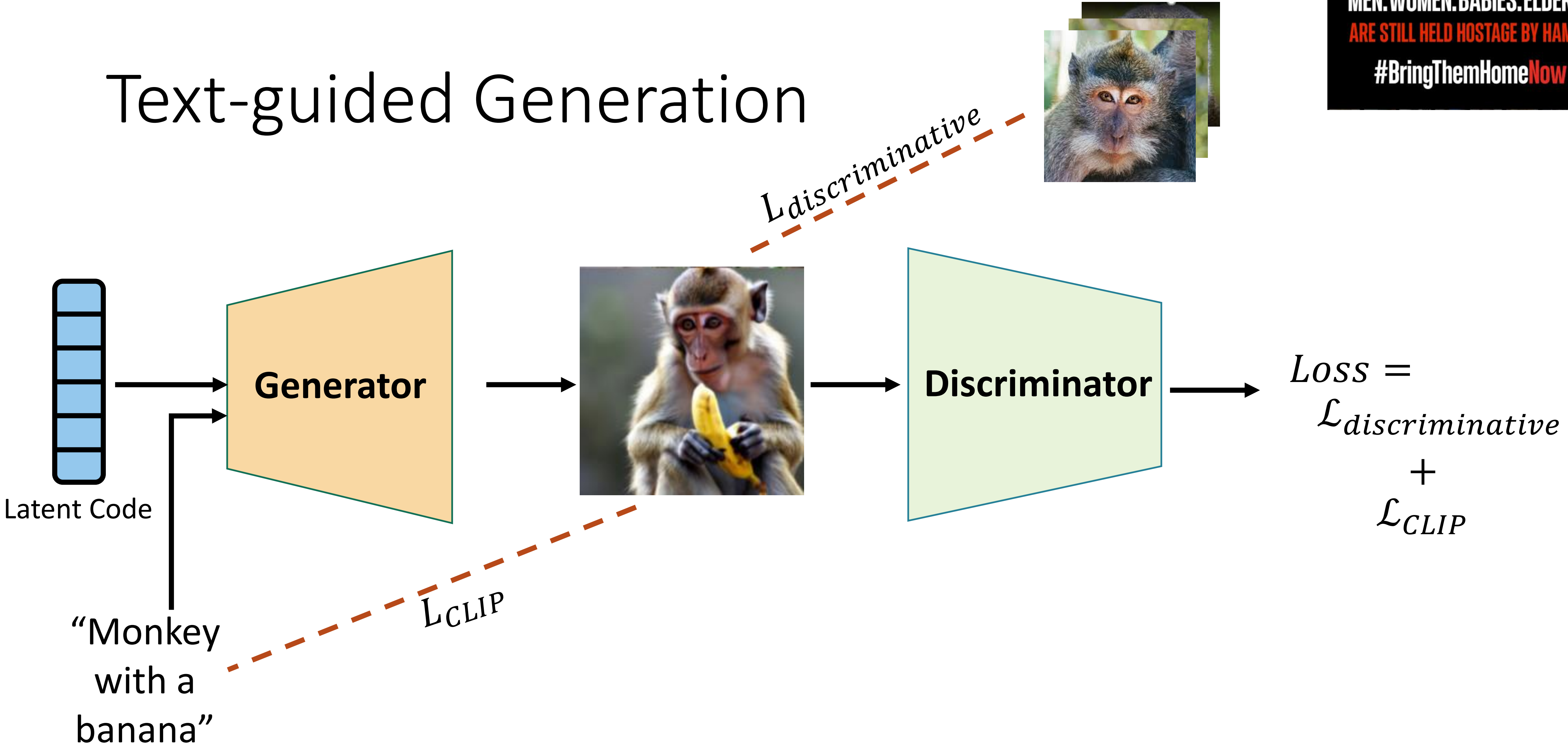
```
from transformers import CLIPModel, CLIPProcessor   # Hugging Face!

model_name = "openai/clip-vit-base-patch32"
processor = CLIPProcessor.from_pretrained(model_name)
model = CLIPModel.from_pretrained(model_name)

inputs = processor(text=[“a red panda”, “a dog“, “a plane”],
                   images=image, return_tensors=“pt”)

model(**inputs).logits_per_image.softmax(dim=1)
# tensor([[0.9815, 0.0110, 0.0075]])
# “A red panda” got the highest score
```

What is this good for?

- “Zero shot” learning
  - Classification
  - Text-guided image generation

WAIC

Text-guided Generation



Weaknesses - Bias

Zero-shot classification of 10,000 faces with additional “bias” categories

	Misclassification rates	
Category	Women	Man
Crime-related Categories	9.8	16.5

WAIC

Weaknesses - Bias

Zero-shot classification of 10,000 faces with additional “bias” categories

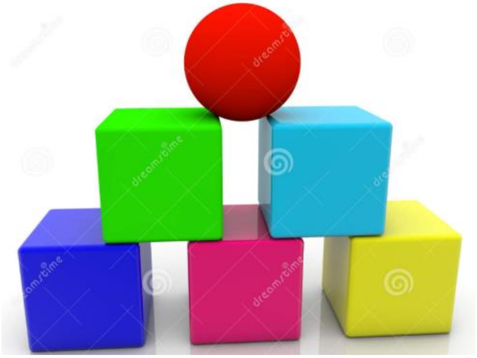
Misclassification rates

Category	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian
Non-human Categories	Total 4.9% misclassified as “non-human”						

WAIC

Weaknesses – Counting and relations

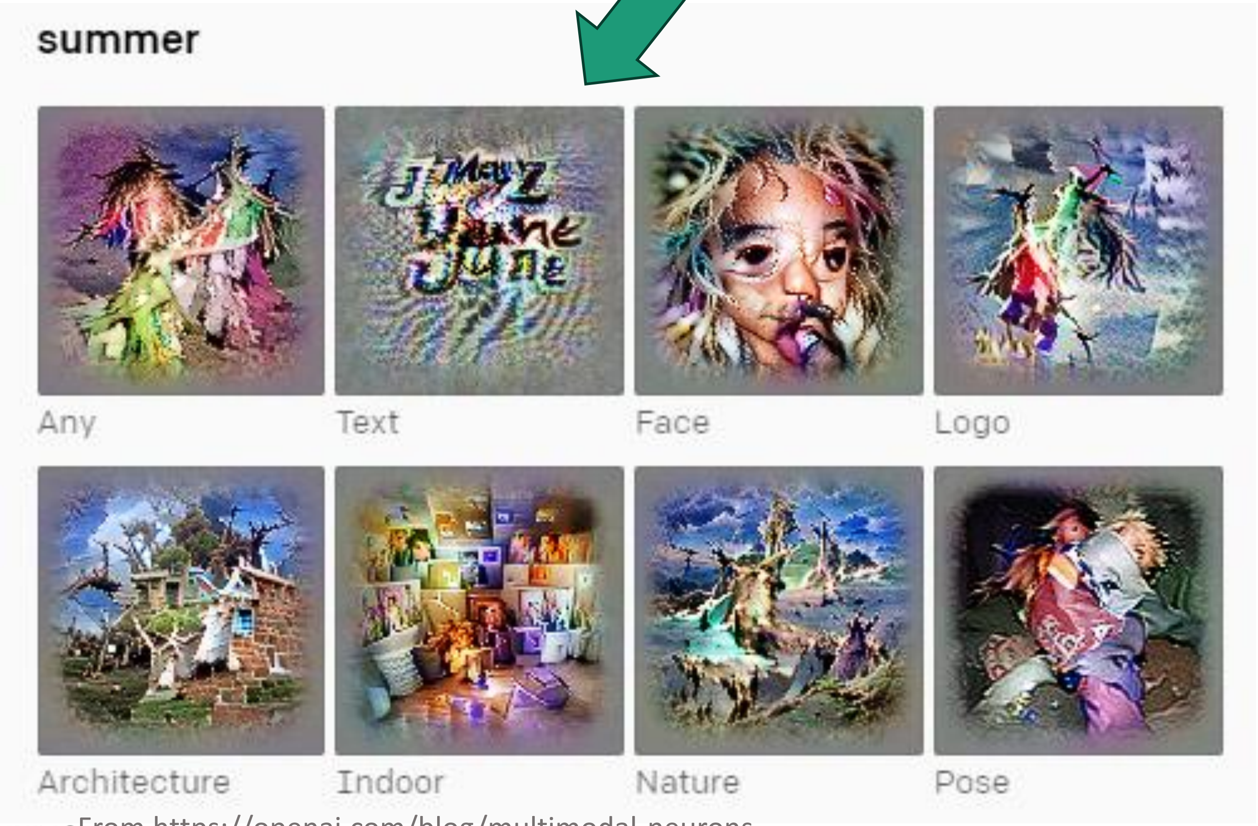
Caption	Probability
Two Balloons	**0.4414**
Three Balloons	0.4054
Four Balloons	0.1531

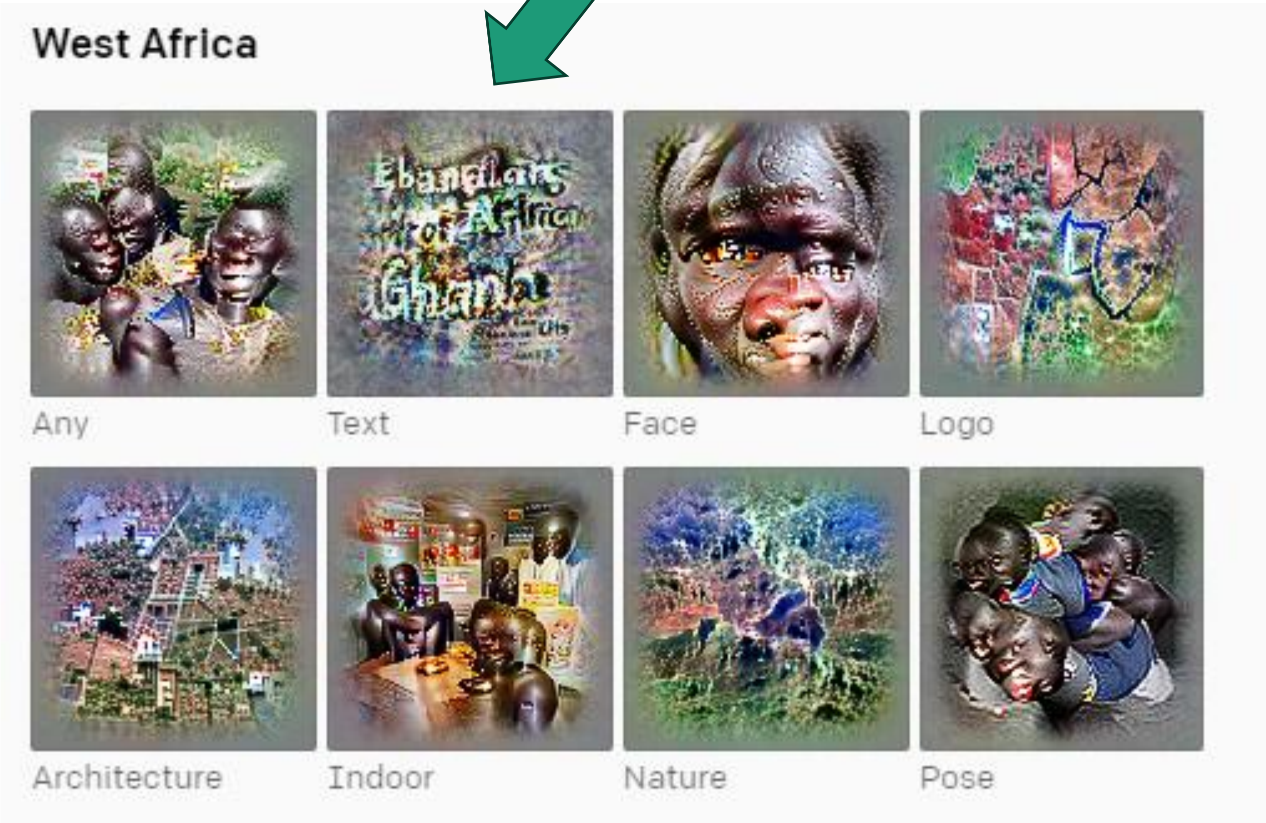


Caption	Probability
A cube next to balls	**0.4743**
A cube over balls	0.3532
A ball over cubes	0.1725

Weaknesses

- Image encoder neurons can be visualized to show concepts





WAIC From https://openai.com/blog/multimodal-neurons

Weaknesses – Typographic Attacks

Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%

Image Standard poodle ⌄

Standard Poodle	39.3%
Angora rabbit	16.0%
Standard Schnauzer	3.6%
Old English Sheepdog	3.3%
Komondor	2.8%
Bedlington Terrier	2.8%

WAIC

Summary

- Self-supervised learning is accelerating as a research field
- Self-supervised foundation models (such as CLIP, DINO, MAE) are highly flexible, generalize well
    - Can learn from given priors (for example, DiNO augmentations)
- Many various approaches to self-supervised learning
    - CLIP – Contrastive learning
    - DINO – Distillation
    - MAE – Masking and reconstruction

Additional Resources

- A cookbook for self-supervised learning
  - https://arxiv.org/abs/2304.12210
- DiNO
  - Paper: Emerging Properties in Self-Supervised Vision Transformers
  - Deep ViT Features as Dense Visual Descriptors
- MAE
  - Paper: Masked Autoencoders Are Scalable Vision Learners
- CLIP
  - Paper: Learning Transferable Visual Models From Natural Language Supervision
  - CLIP Microscope (Neuron concept visualizations)
    - https://microscope.openai.com/models/contrastive_4x/image_block_4_5_Add_6_0
    - https://openai.com/blog/multimodal-neurons

WAIC

MEN.WOMEN.BABIES.ELDERLY.
ARE STILL HELD HOSTAGE BY HAMAS
#BringThemHomeNow

Next time:
“Computer Graphics and Rendering”

WAIC