UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

> Reinventa el mundo <

Sesión 5.1

Self-supervised learning II

iBOT. MAE

TRANSFORMATEC

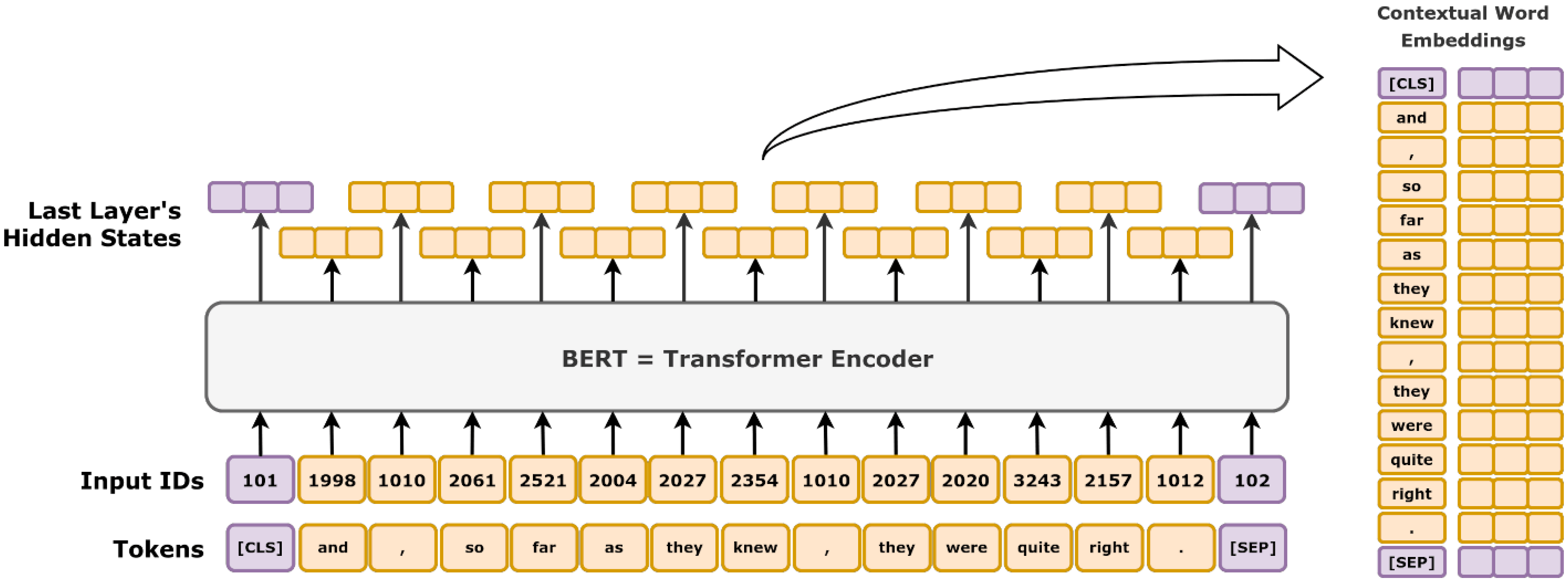
UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

1.

Masked Image *Modeling*

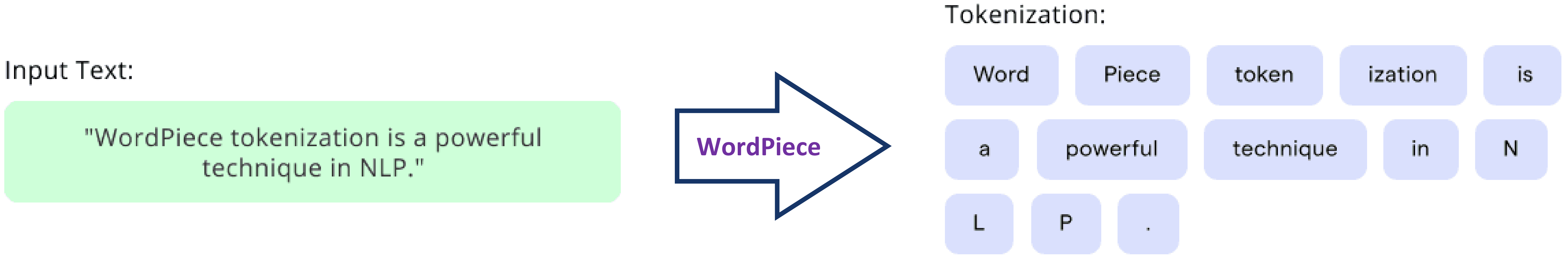
TRANSFORMATEC

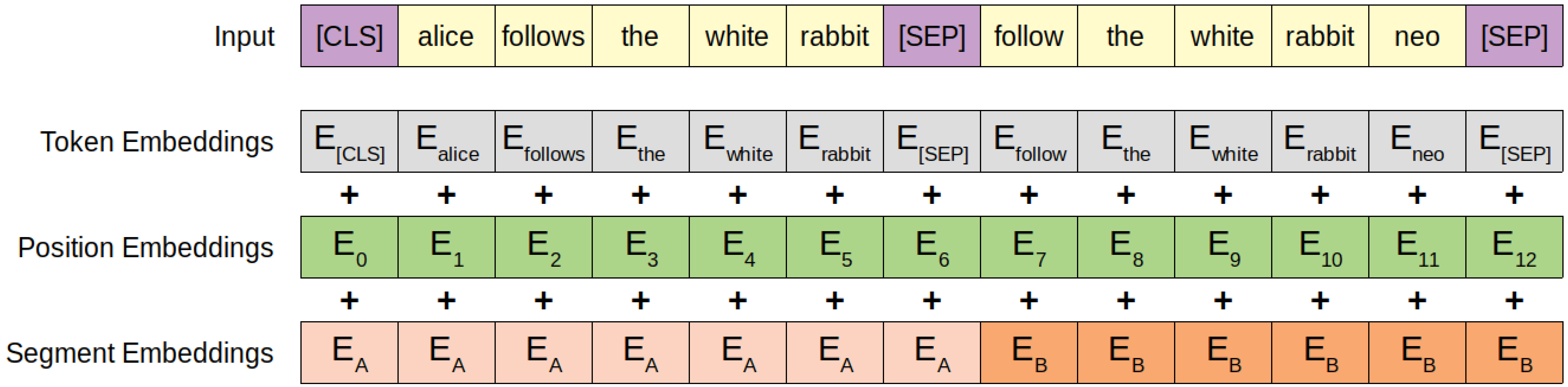
BERT



Jacob Devlin et al. (2018) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". Proceedings of naacL-HLT (Vol. 1, p. 2).

BERT

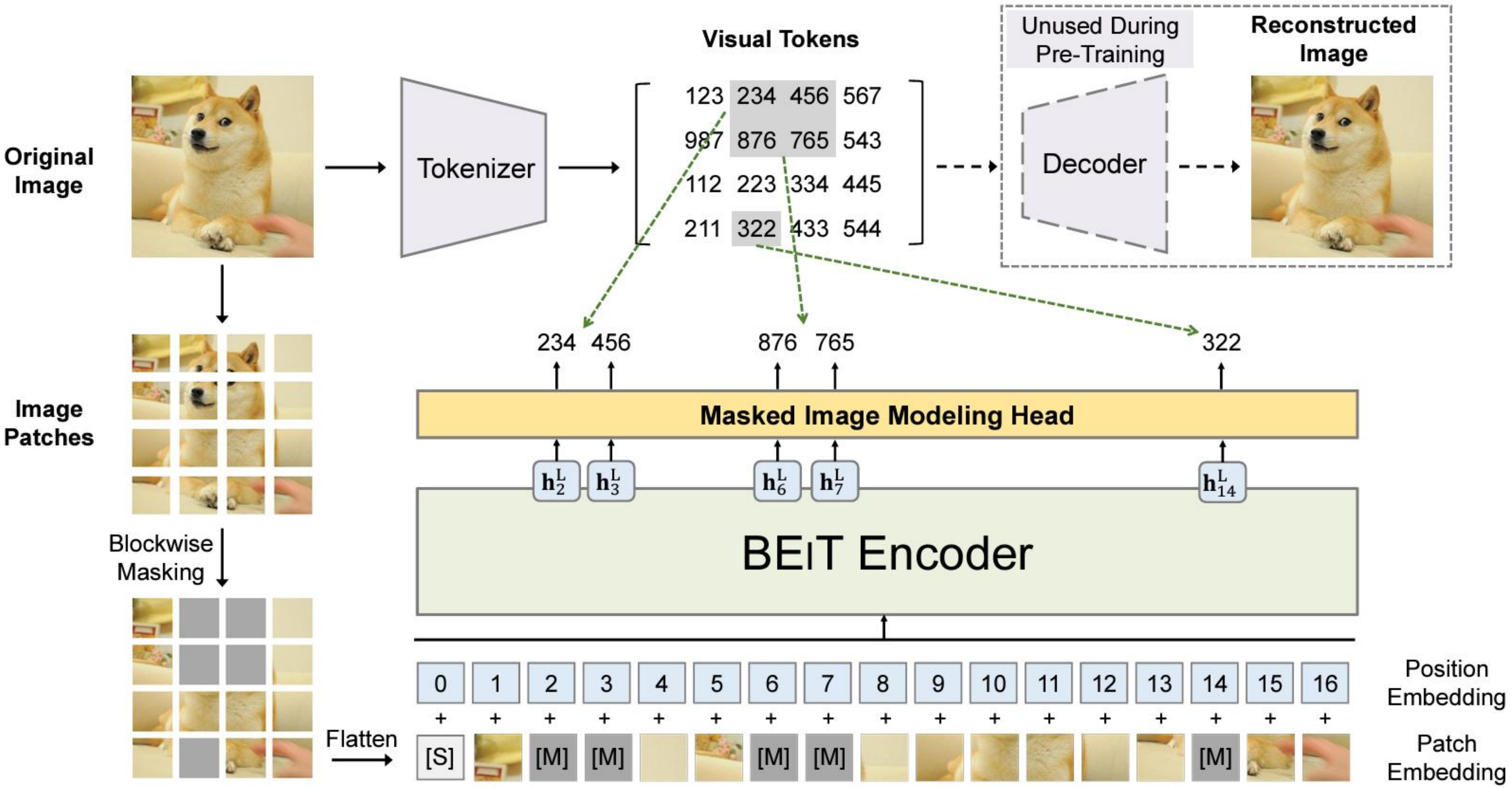




Jacob Devlin et al. (2018) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". Proceedings of naacL-HLT (Vol. 1, p. 2).

BEiT

(Bidirectional Encoder representation for Image Transformers)



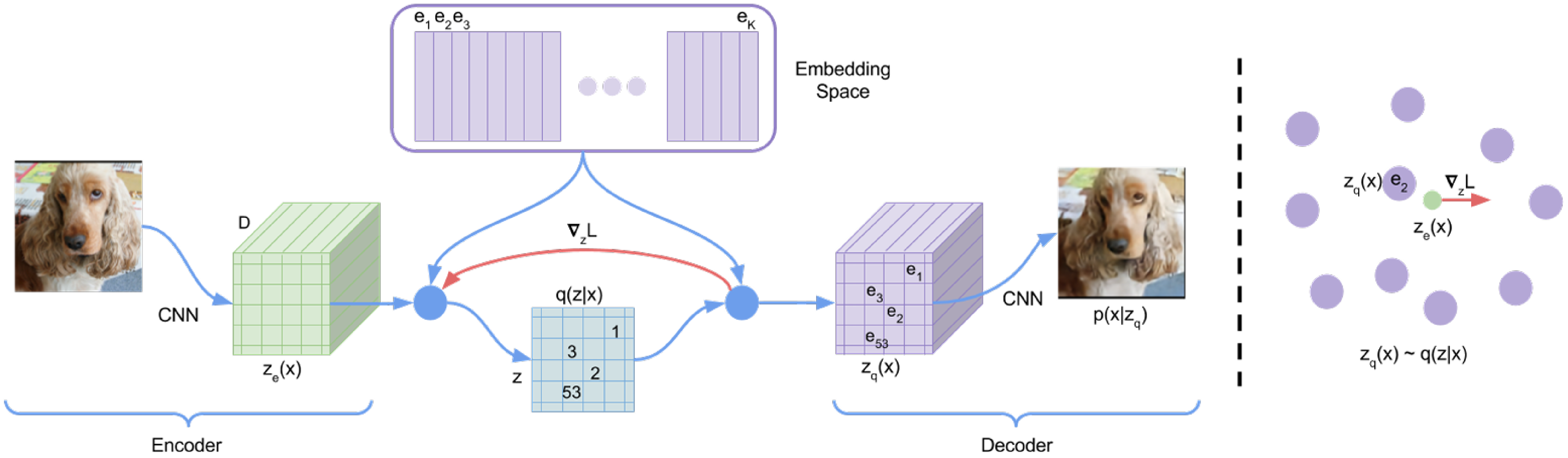
Hangbo Bao et al. (2021) "BEiT: BERT Pre-Training of Image Transformers". arXiv preprint arXiv:2106.08254.

BEiT

(Bidirectional Encoder representation for Image Transformers)

VQ-VAE (Vector Quantised-Variational AutoEncoder)

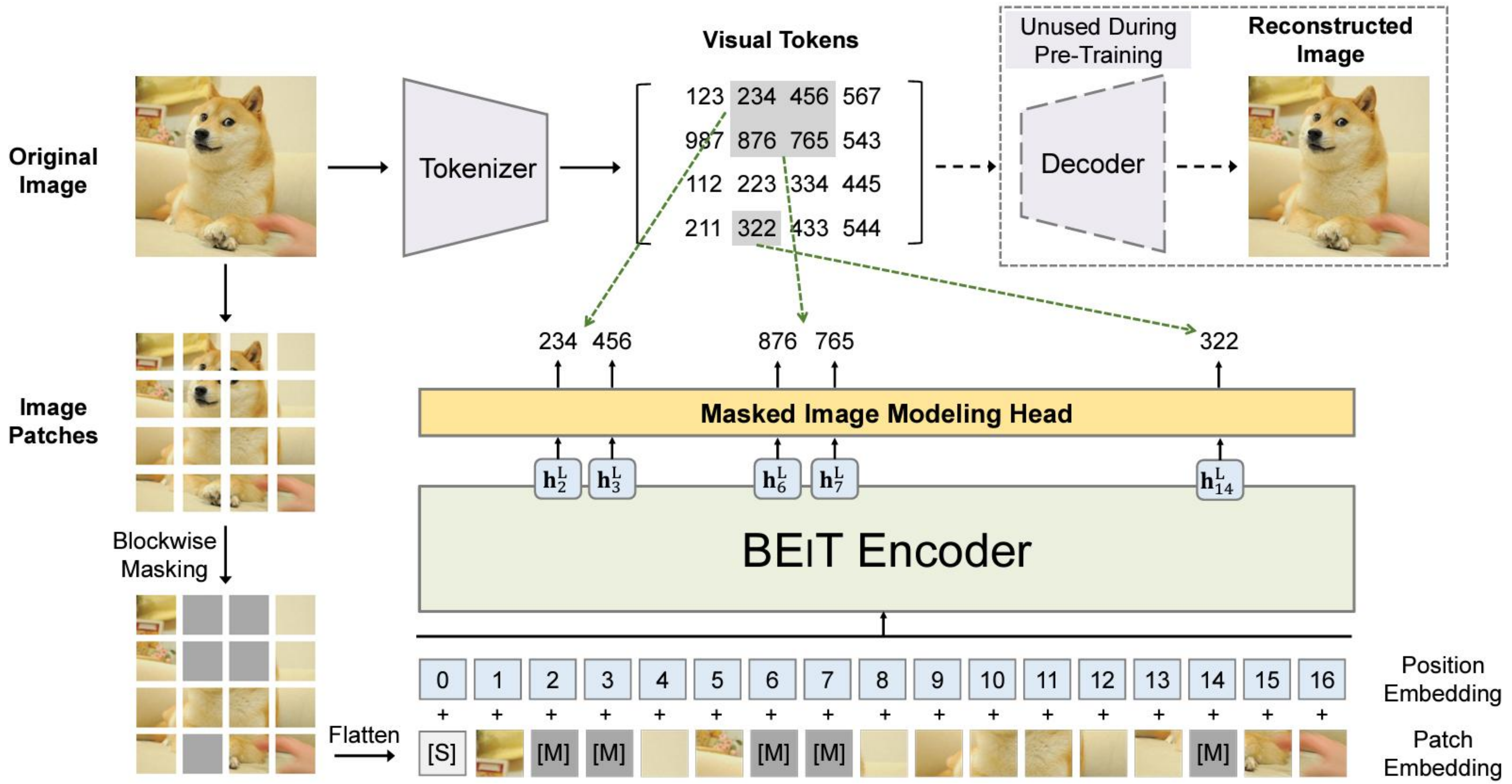
(dVAE es muy similar, pero no había una imagen bonita :D)



Hangbo Bao et al. (2021) “BEiT: BERT Pre-Training of Image Transformers”. arXiv preprint arXiv:2106.08254.

BEiT

(Bidirectional Encoder representation for Image Transformers)



Aplicamos al 40% de la imagen

Maximize the log-likelihood of the correct visual tokens z_i given the corrupted image:

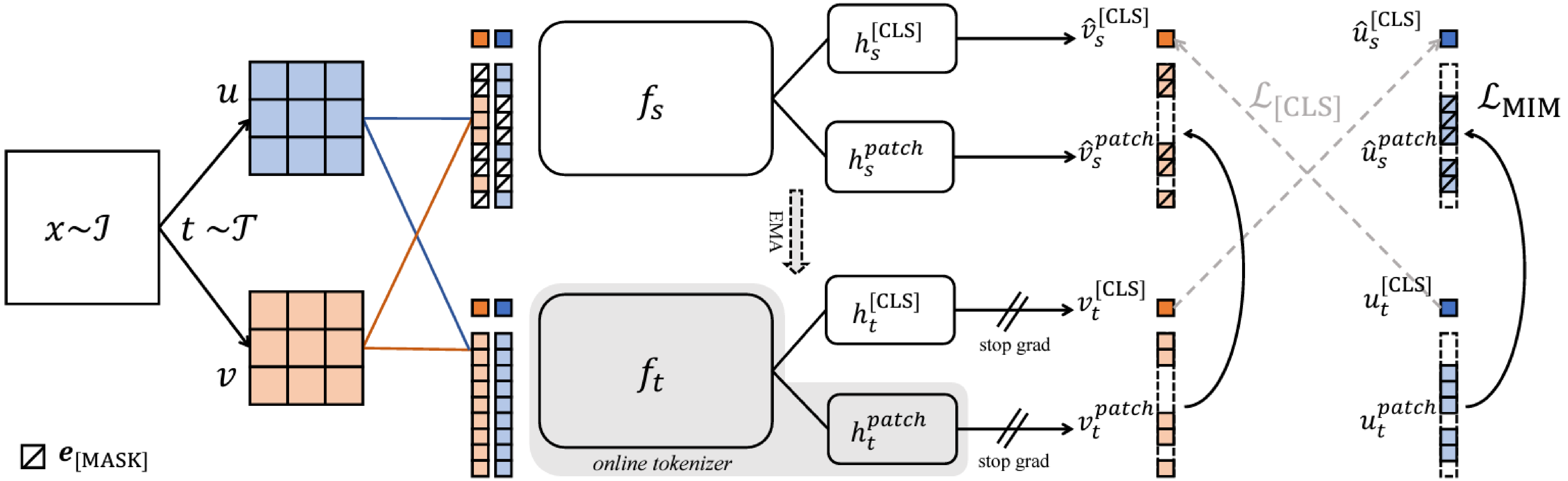
$$\max \sum_{x \in \mathcal{D}} \mathbb{E}_{\mathcal{M}} \left[\sum_{i \in \mathcal{M}} \log p_{\mathrm{MIM}}(z_i | x^{\mathcal{M}}) \right]$$

CrossEntropy

Hangbo Bao et al. (2021) "BEiT: BERT Pre-Training of Image Transformers". arXiv preprint arXiv:2106.08254.

iBOT

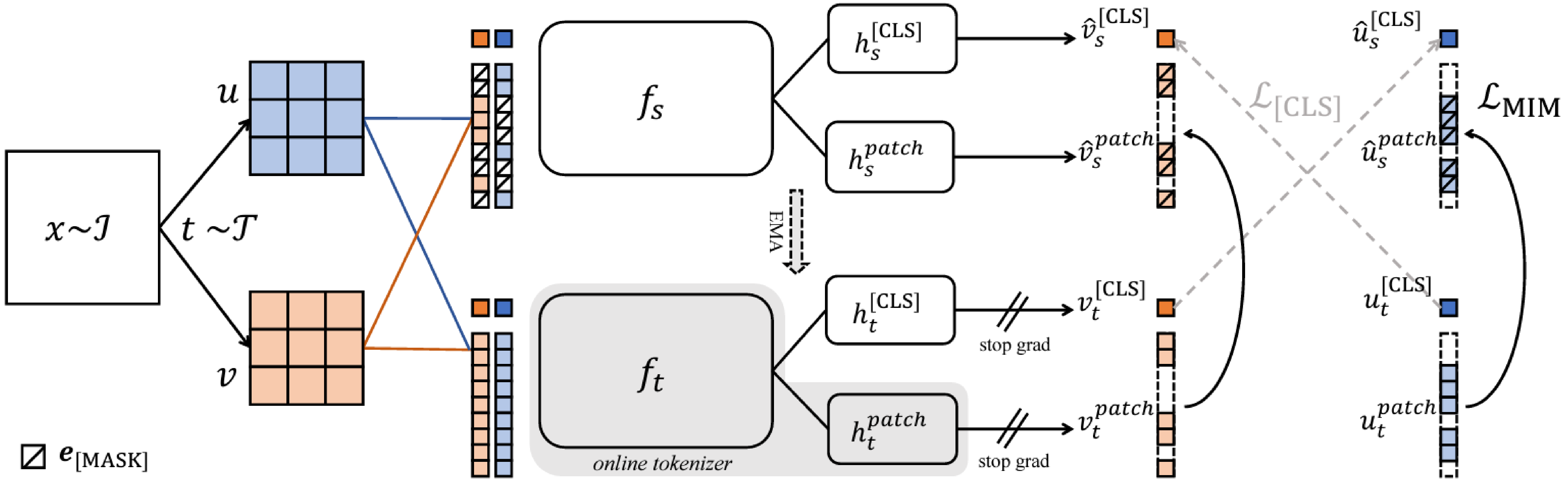
(Image BERT Pretraining with Online Tokenizer)



Jinghao Zhou et al. (2022) "iBOT: Image BERT Pre-Training with Online Tokenizer". arXiv preprint arXiv:2111.07832.

iBOT

(Image BERT Pretraining with Online Tokenizer)



Self-distillation loss:

$$\mathcal{L}_{[CLS]} = -P_{\theta'}^{[CLS]}(\boldsymbol{v})^{T} \log P_{\theta}^{[CLS]}(\boldsymbol{u})$$

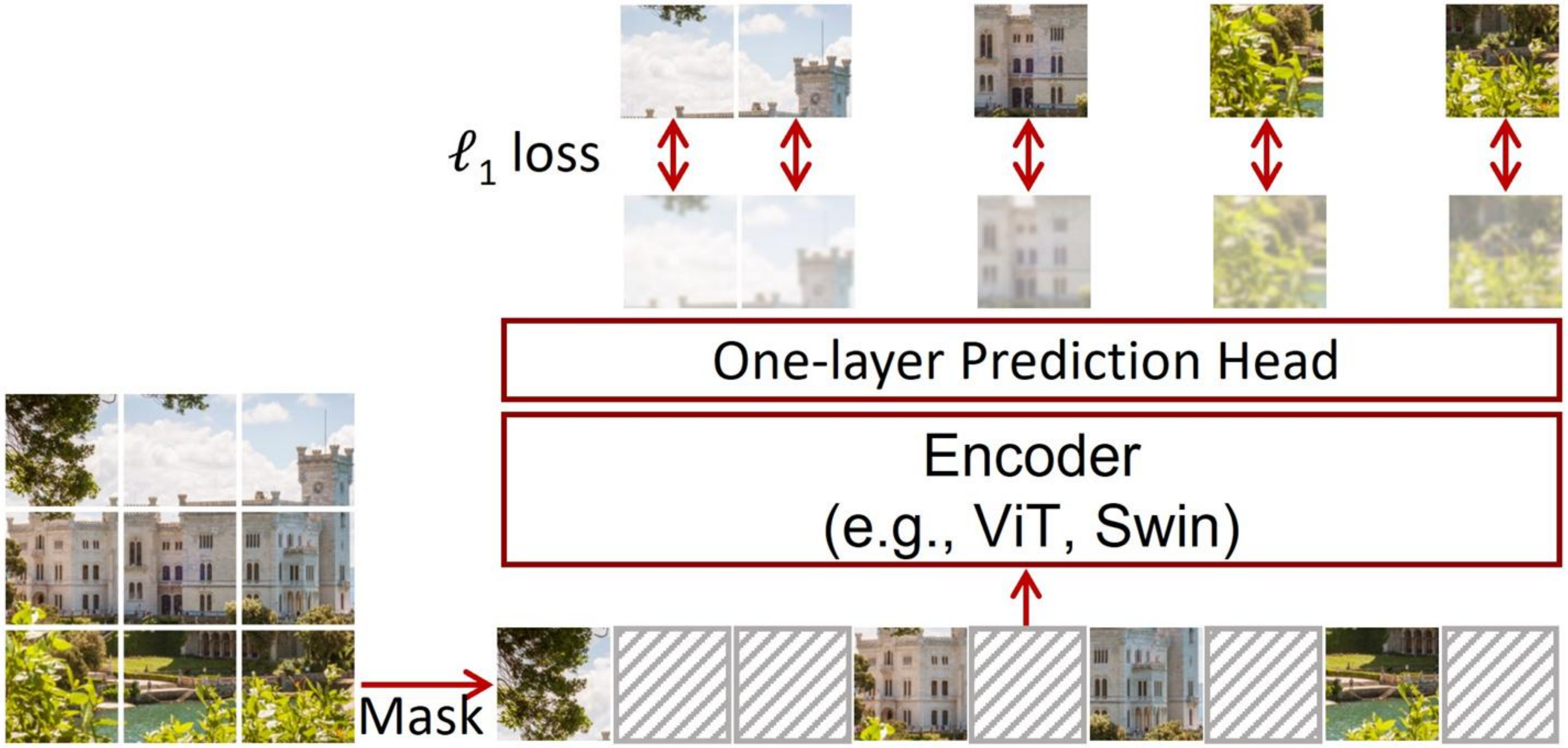
MIM loss:

$$\mathcal{L}_{MIM} = -\sum_{i=1}^{N} m_i \cdot P_{\theta'}^{patch}(\boldsymbol{u}_i)^{T} \log P_{\theta}^{patch}(\hat{\boldsymbol{u}}_i)$$

Jinghao Zhou et al. (2022) "iBOT: Image BERT Pre-Training with Online Tokenizer". arXiv preprint arXiv:2111.07832.

SimMIM

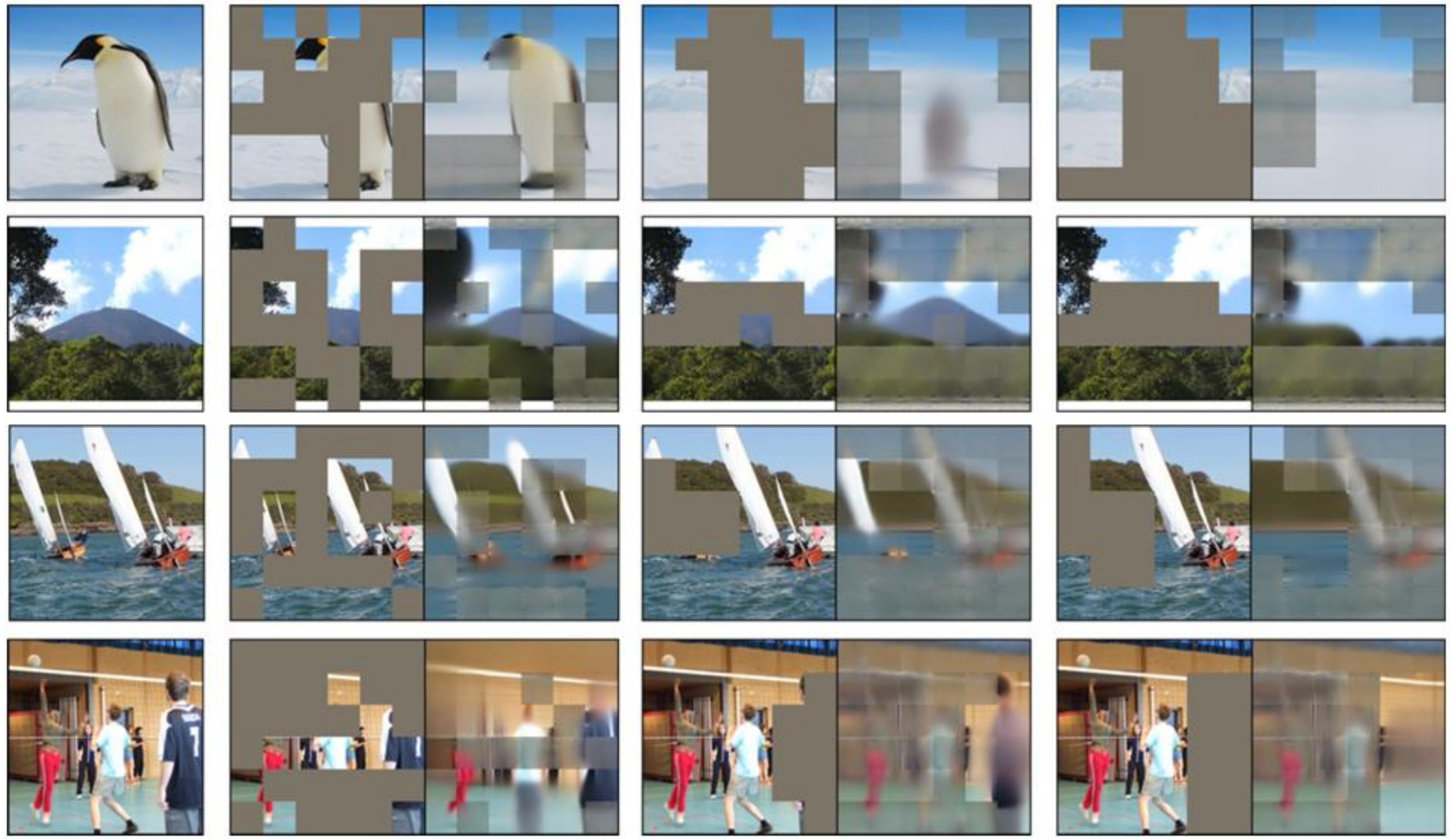
(Simple Masked Image Modeling)



Zhenda Xie et al. (2022) "SimMIM: A Simple Framework for Masked Image Modeling".
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9653-9663).

SimMIM

(Simple Masked Image Modeling)

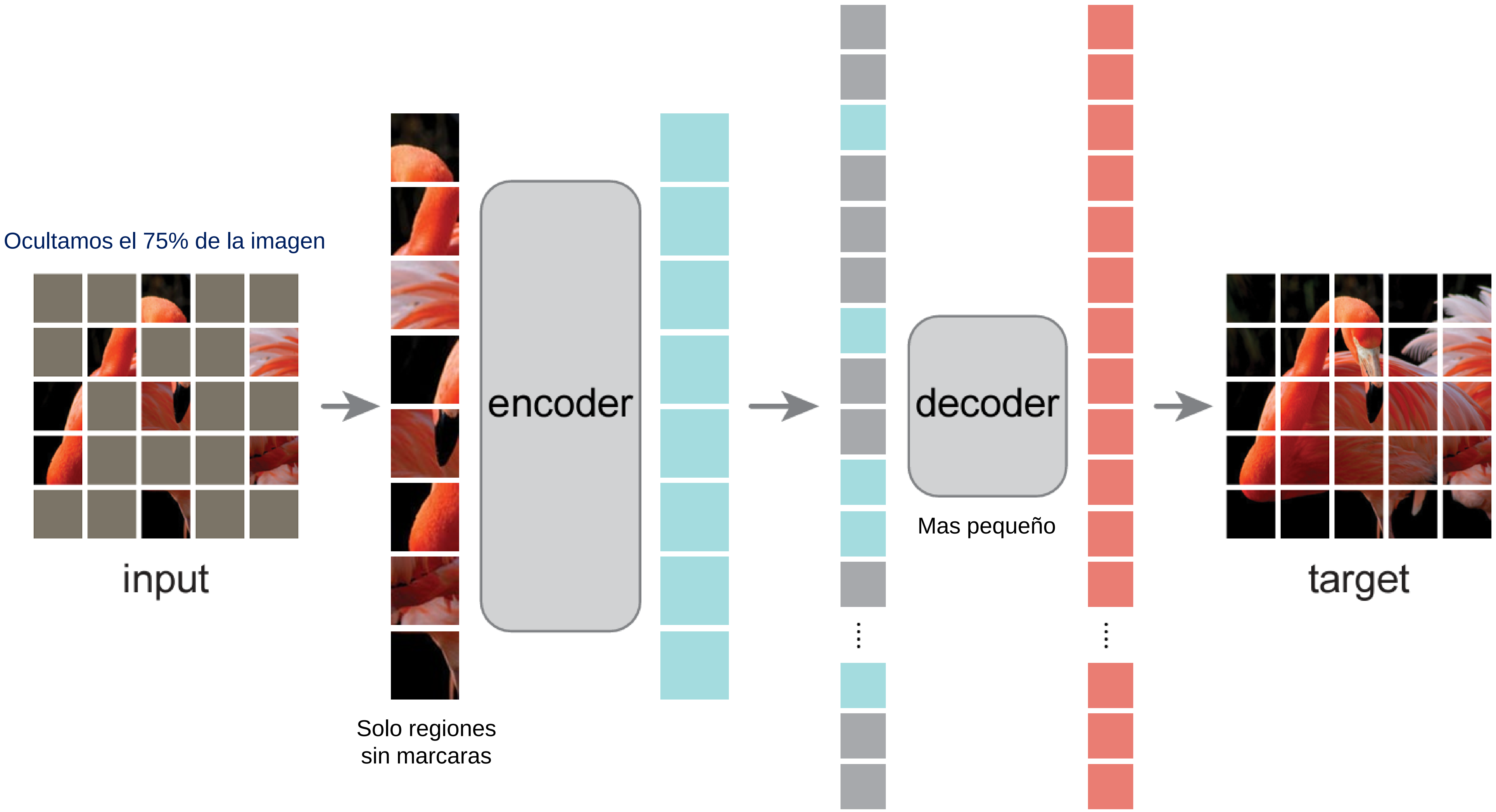


Methods	Input Size	Fine-tuning Top-1 acc (%)	Linear eval Top-1 acc (%)	Pre-training costs
Sup. baseline [44]	224^2	81.8	-	-
DINO [5]	224^2	82.8	78.2	2.0×
MoCo v3 [9]	224^2	83.2	76.7	1.8×
ViT [15]	384^2	79.9	-	∼4.0×
BEiT [1]	224^2	83.2	56.7	1.5×†
Ours	224^2	**83.8**	56.7	1.0×

Zhenda Xie et al. (2022) "SimMIM: A Simple Framework for Masked Image Modeling". Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9653-9663).

MAE

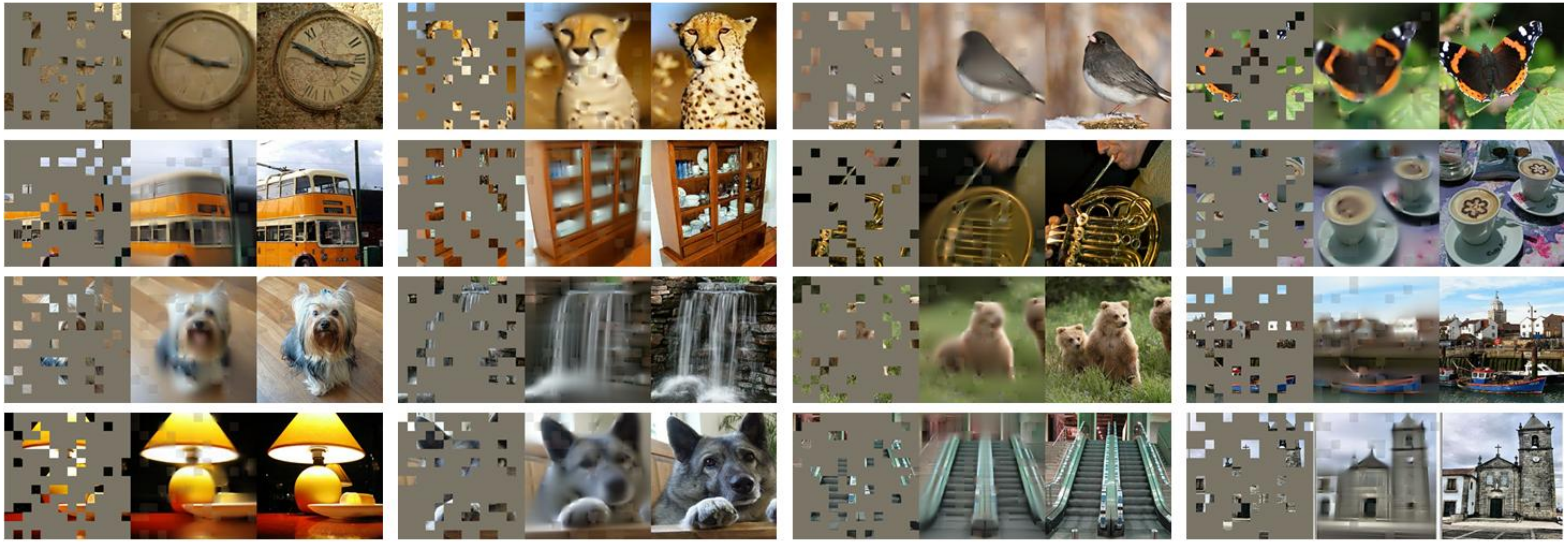
(Masked Autoencoders)



Kaiming He et al. (2022) “Masked Autoencoders Are Scalable Vision Learners”. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16000-16009).

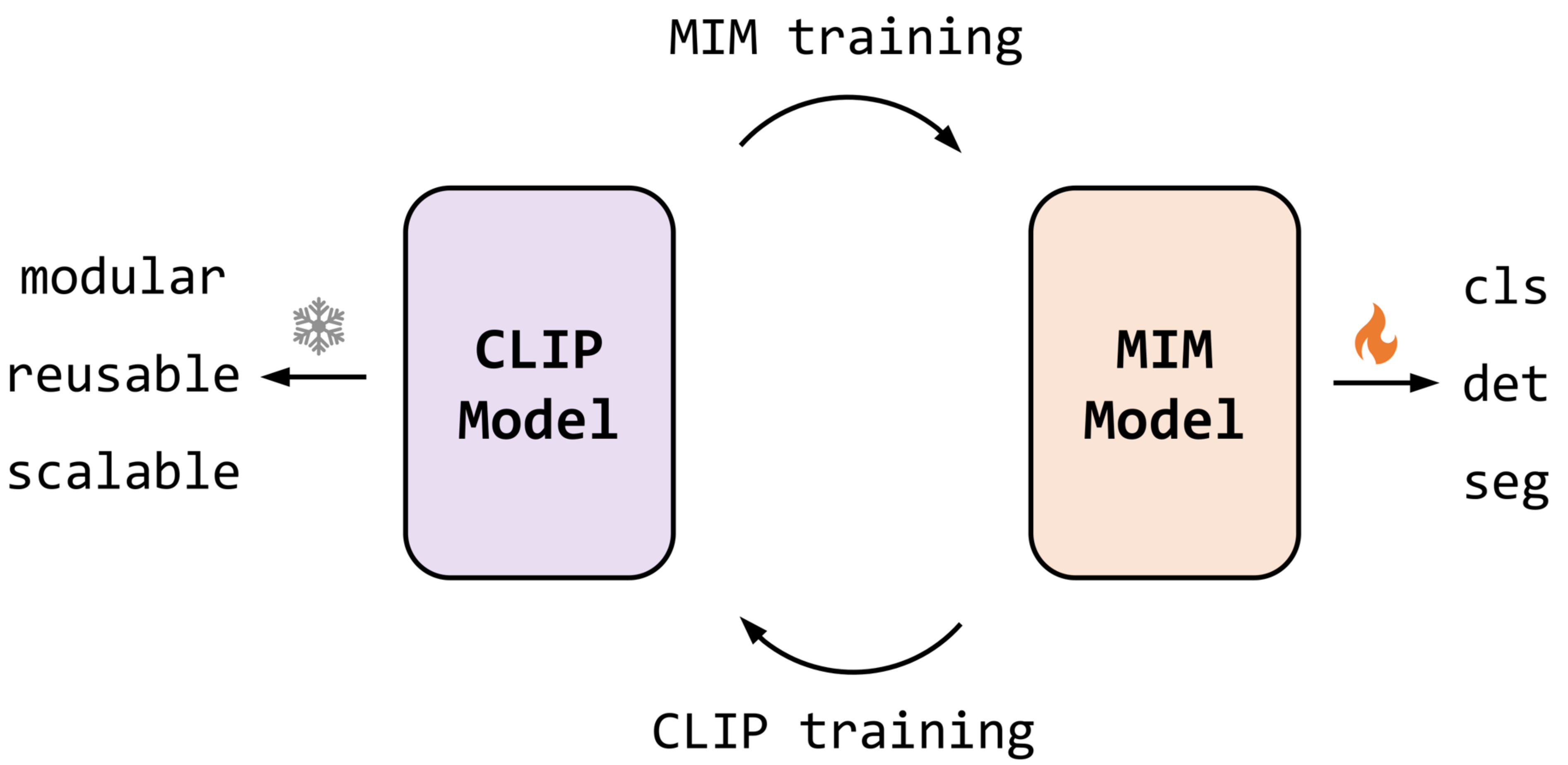
MAE

(Masked Autoencoders)



Kaiming He et al. (2022) "Masked Autoencoders Are Scalable Vision Learners".
Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9653-9663).

EVA



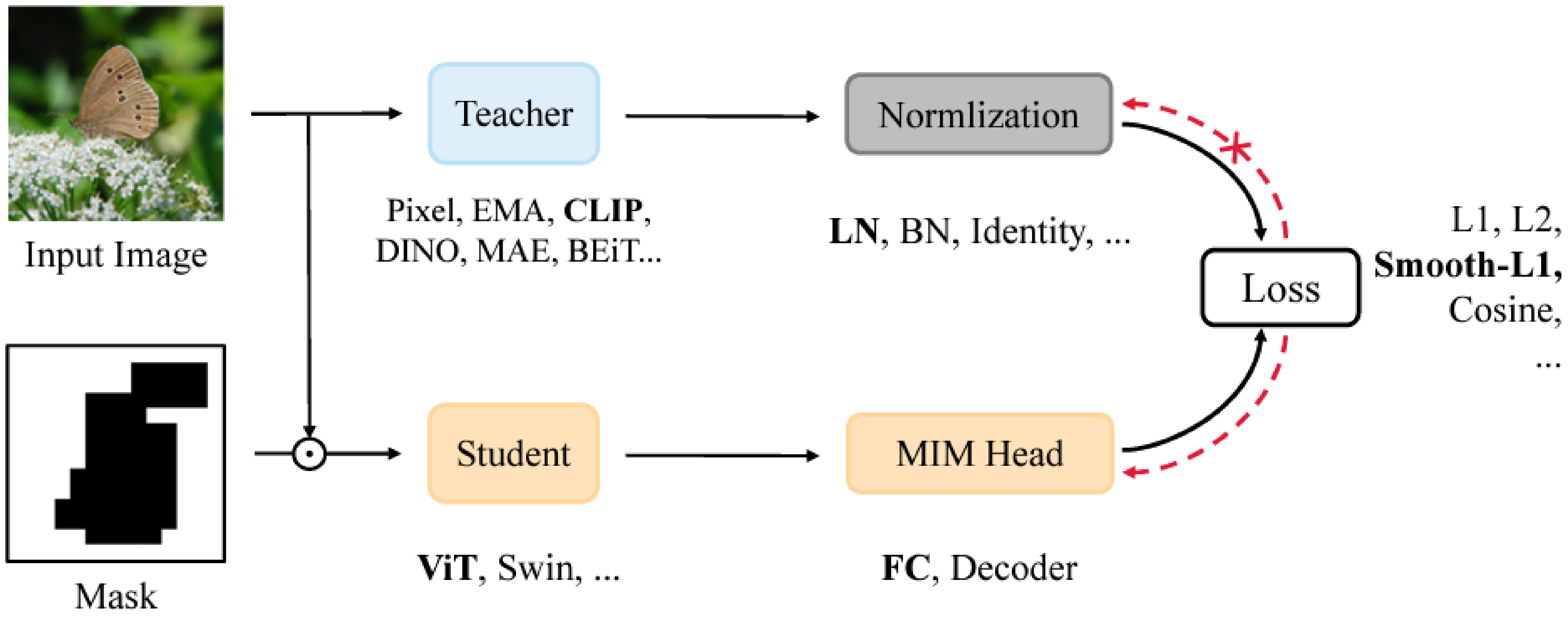
BEiT
Predicción de Tokens Visuales

MAE
Reconstrucción de Píxeles

EVA
Regresión de Features de CLIP

Yuxin Fang et al. (2022) "EVA: Exploring the Limits of Masked Visual Representation Learning at Scale". arXiv preprint arXiv:2211.07636.

EVA



Yuxin Fang et al. (2022) "EVA: Exploring the Limits of Masked Visual Representation Learning at Scale". arXiv preprint arXiv:2211.07636.

UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

2.

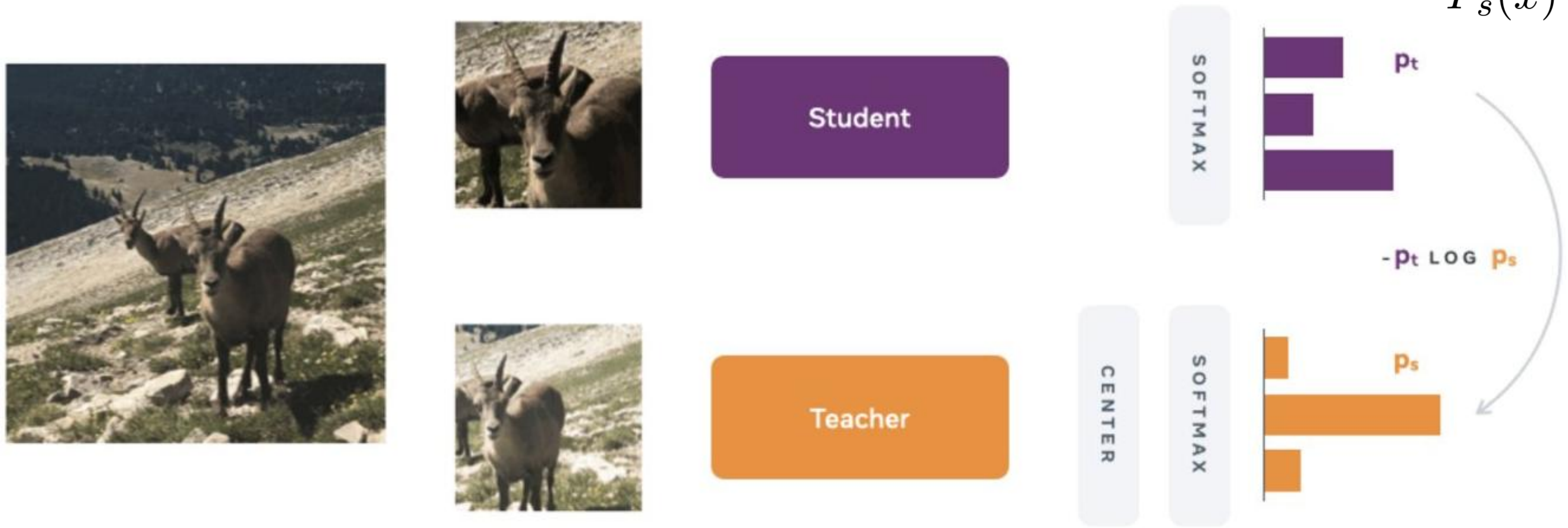
DINO

TRANSFORMATEC

DINO

Mathilde Caron et al. (2021) “Emerging Properties in Self-Supervised Vision Transformers”. Proceedings of the IEEE/CVF international conference on computer vision. p. 9650-9660.

DINO



$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K} \exp(g_{\theta_s}(x)^{(k)}/\tau_s)}$$

$$\min_{\theta_s} \sum_{x \in \{x_1^g, x_2^g\}} \sum_{\substack{x' \in V \\ x' \neq x}} H(P_t(x), P_s(x'))$$

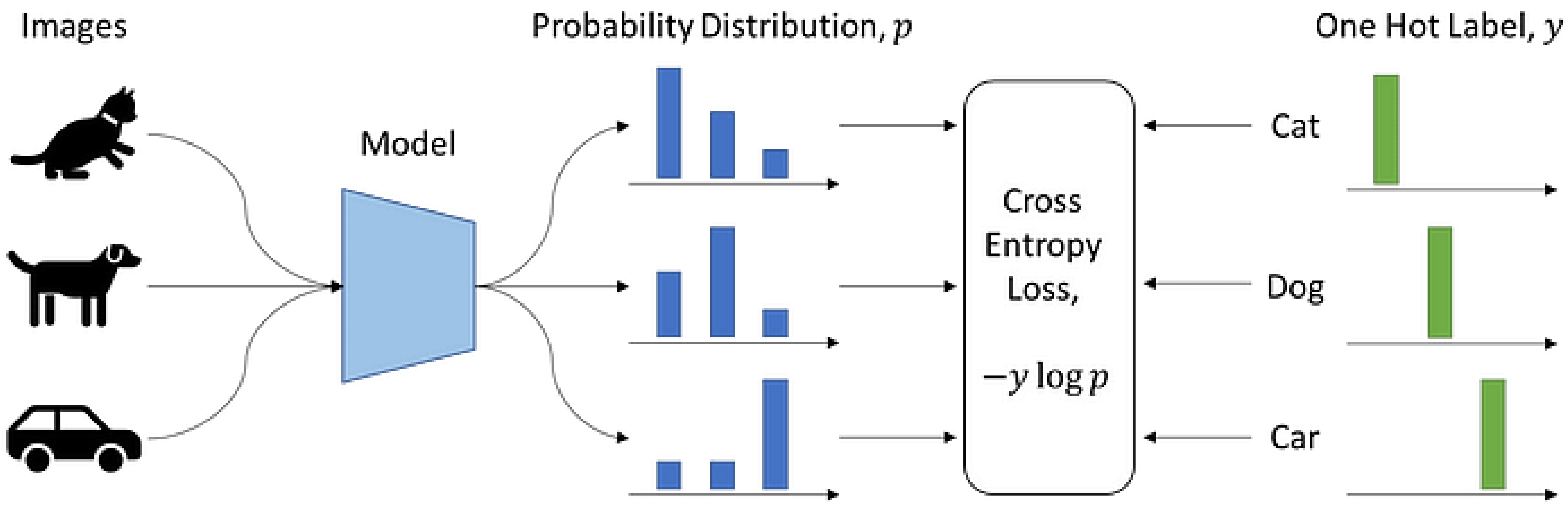
Global views

Local views

$$H(a,b) = -a \log b$$

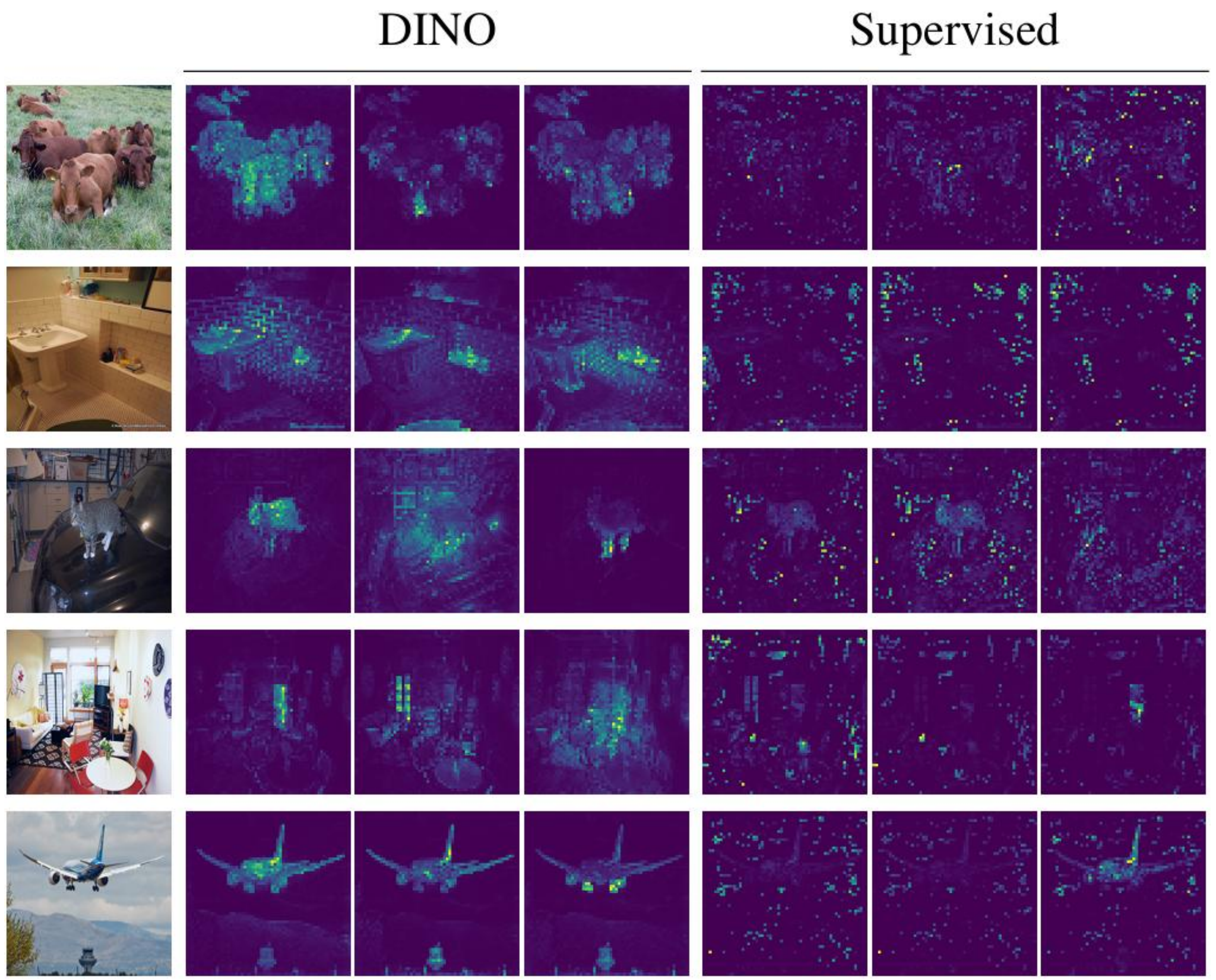
Mathilde Caron et al. (2021) “Emerging Properties in Self-Supervised Vision Transformers”. Proceedings of the IEEE/CVF international conference on computer vision. p. 9650-9660.

DINO



Mathilde Caron et al. (2021) "Emerging Properties in Self-Supervised Vision Transformers". Proceedings of the IEEE/CVF international conference on computer vision. p. 9650-9660.

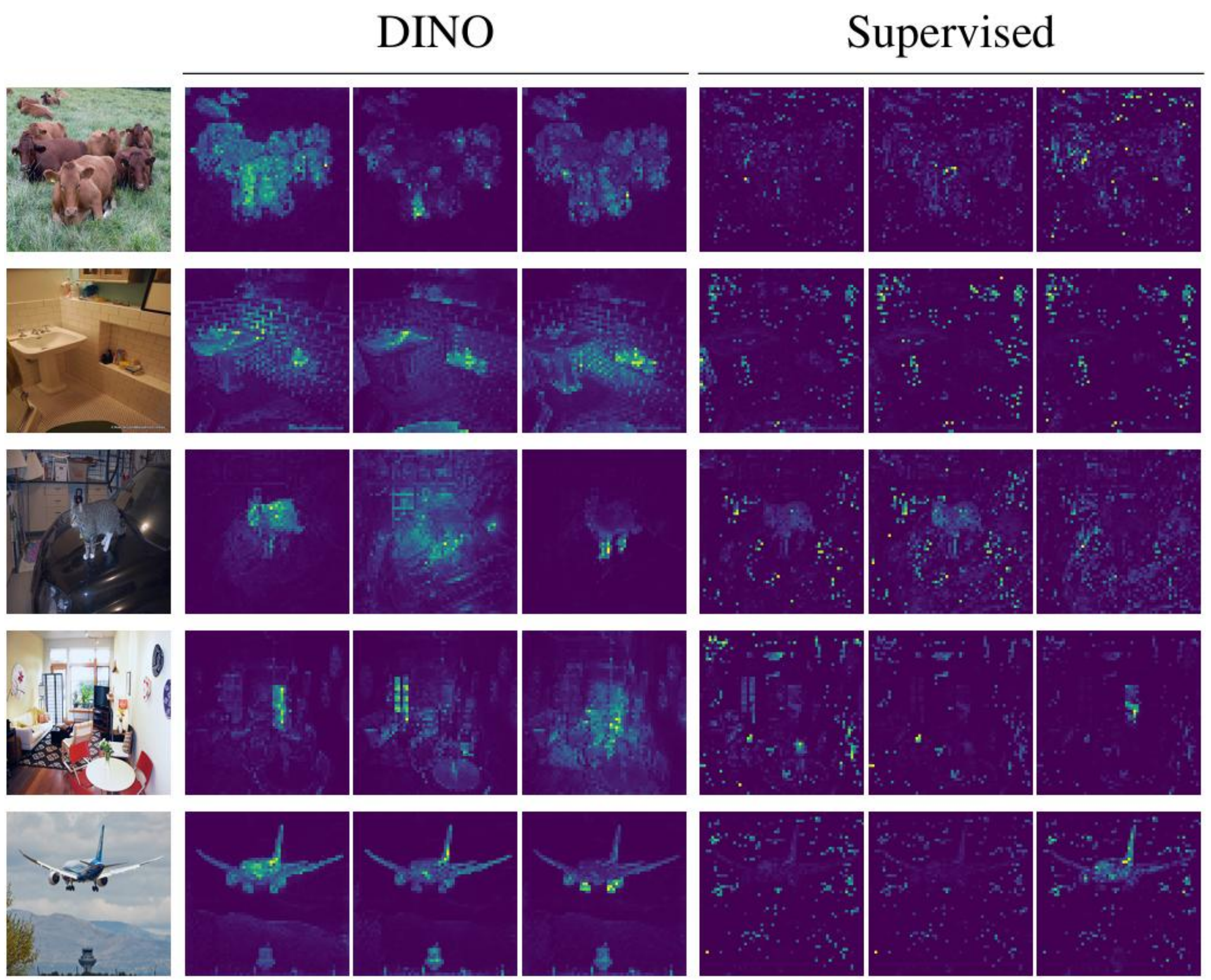
DINO



CODE?

Mathilde Caron et al. (2021) "Emerging Properties in Self-Supervised Vision Transformers". Proceedings of the IEEE/CVF international conference on computer vision. p. 9650-9660.

DINO



Algorithm 1 DINO PyTorch pseudocode w/o multi-crop.

```
# gs, gt: student and teacher networks
# C: center (K)
# tps, tpt: student and teacher temperatures
# l, m: network and center momentum rates
gt.params = gs.params
for x in loader: # load a minibatch x with n samples
    x1, x2 = augment(x), augment(x) # random views

    s1, s2 = gs(x1), gs(x2) # student output n-by-K
    t1, t2 = gt(x1), gt(x2) # teacher output n-by-K

    loss = H(t1, s2)/2 + H(t2, s1)/2
    loss.backward() # back-propagate

    # student, teacher and center updates
    update(gs) # SGD
    gt.params = l*gt.params + (1-l)*gs.params
    C = m*C + (1-m)*cat([t1, t2]).mean(dim=0)

def H(t, s):
    t = t.detach() # stop gradient
    s = softmax(s / tps, dim=1)
    t = softmax((t - C) / tpt, dim=1) # center + sharpen
    return - (t * log(s)).sum(dim=1).mean()
```

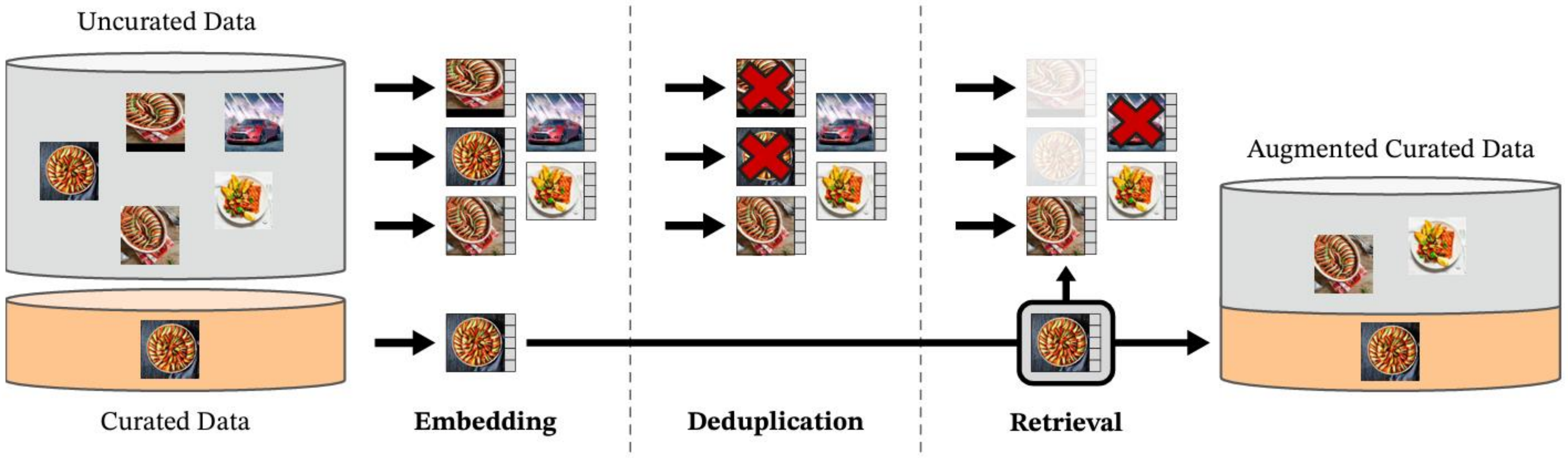

Mathilde Caron et al. (2021) “Emerging Properties in Self-Supervised Vision Transformers”.
Proceedings of the IEEE/CVF international conference on computer vision. p. 9650-9660.

DINO v2

Maxime Oquab et al. (2023) "DINOv2: Learning Robust Visual Features without Supervision". arXiv preprint arXiv:2304.07193.

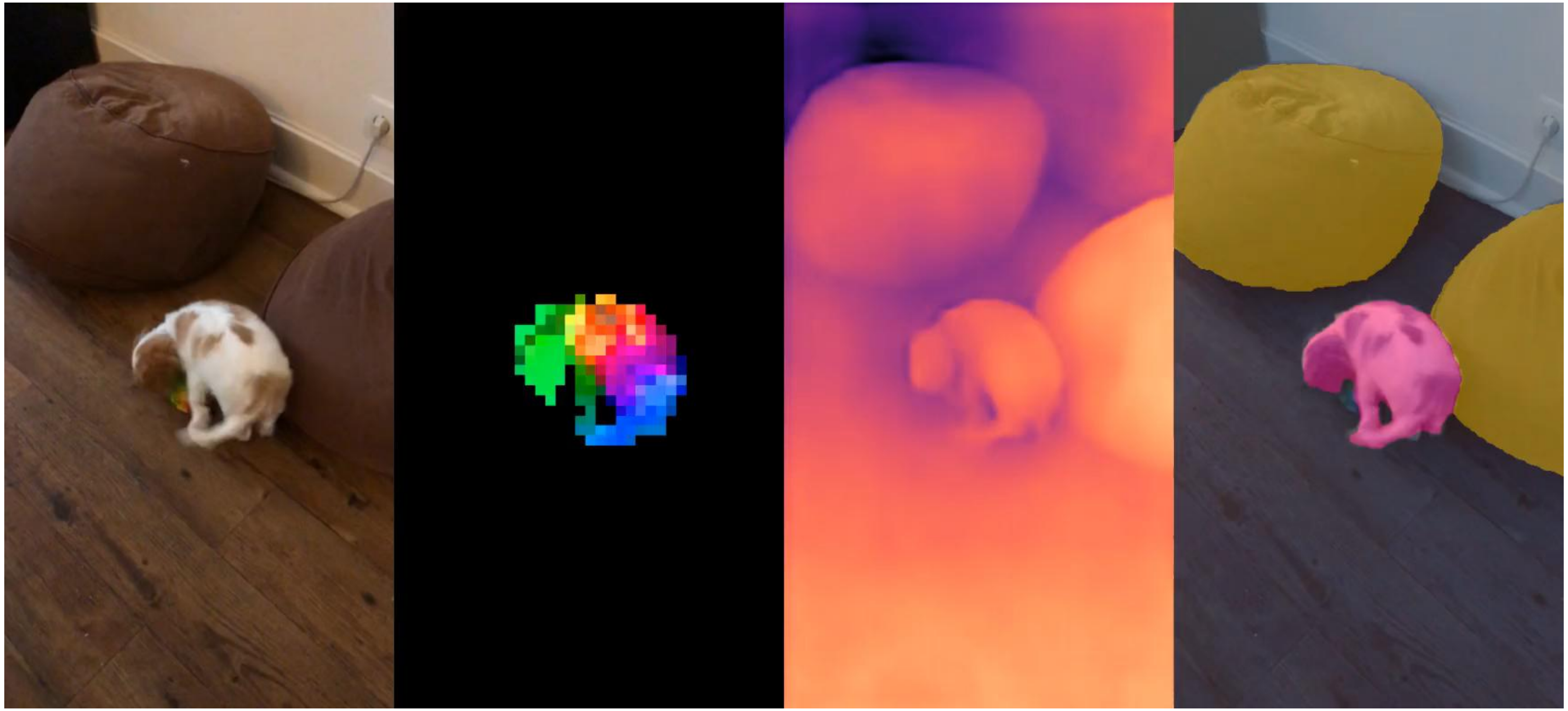
DINO v2

Processing pipeline



Maxime Oquab et al. (2023) "DINOv2: Learning Robust Visual Features without Supervision". arXiv preprint arXiv:2304.07193.

DINO v2



UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

> Reinventa el mundo <

GRACIAS

Victor Flores Benites

TRANSFORMATEC