Глубинное обучение

Self-supervised learning

Ирина Сапарина

Recap

CLIP: промежуточная задача для обучения - image captioning

40 миллионов пар
картинка + подпись

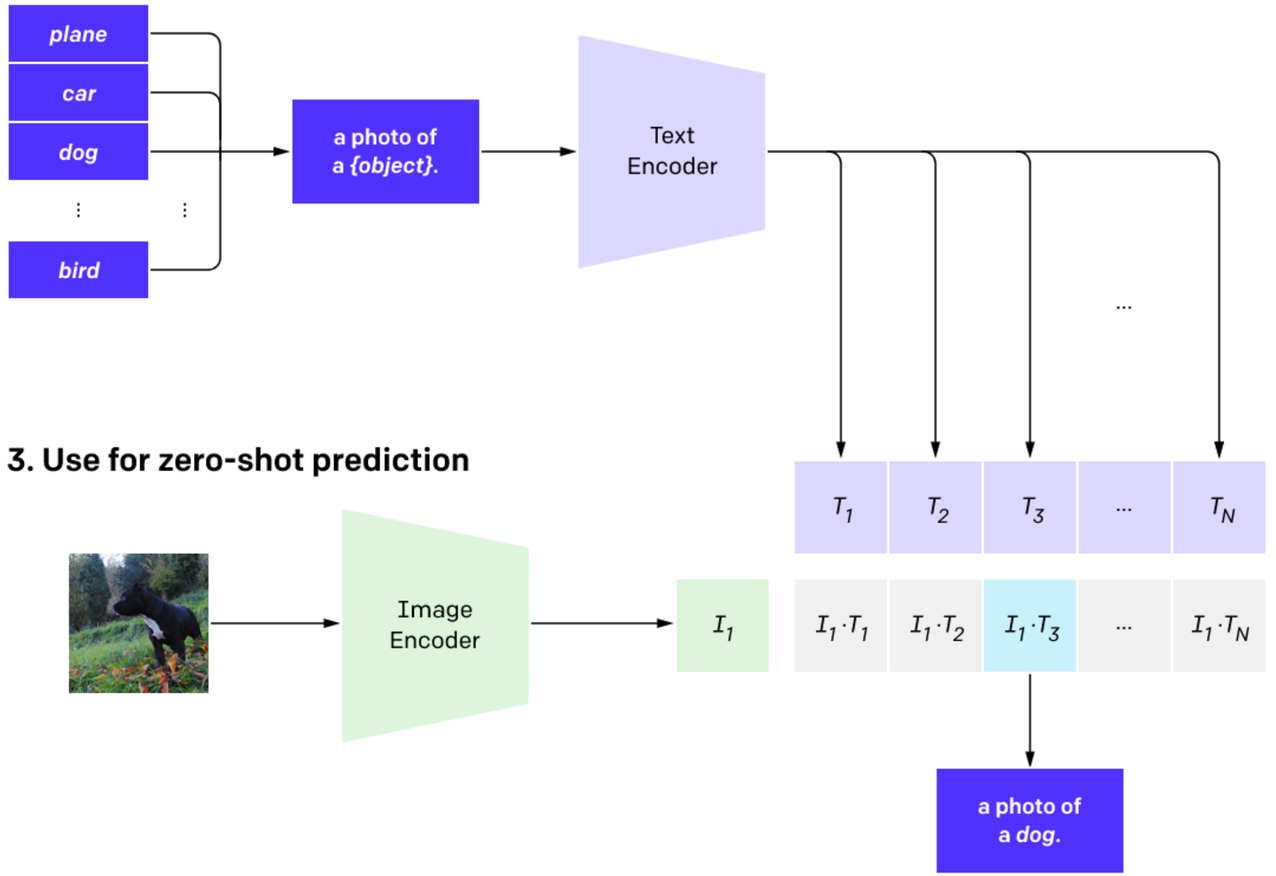


Image credit

Recap

BERT: задачи для обучения MLP, NSP

+

3,300M слов

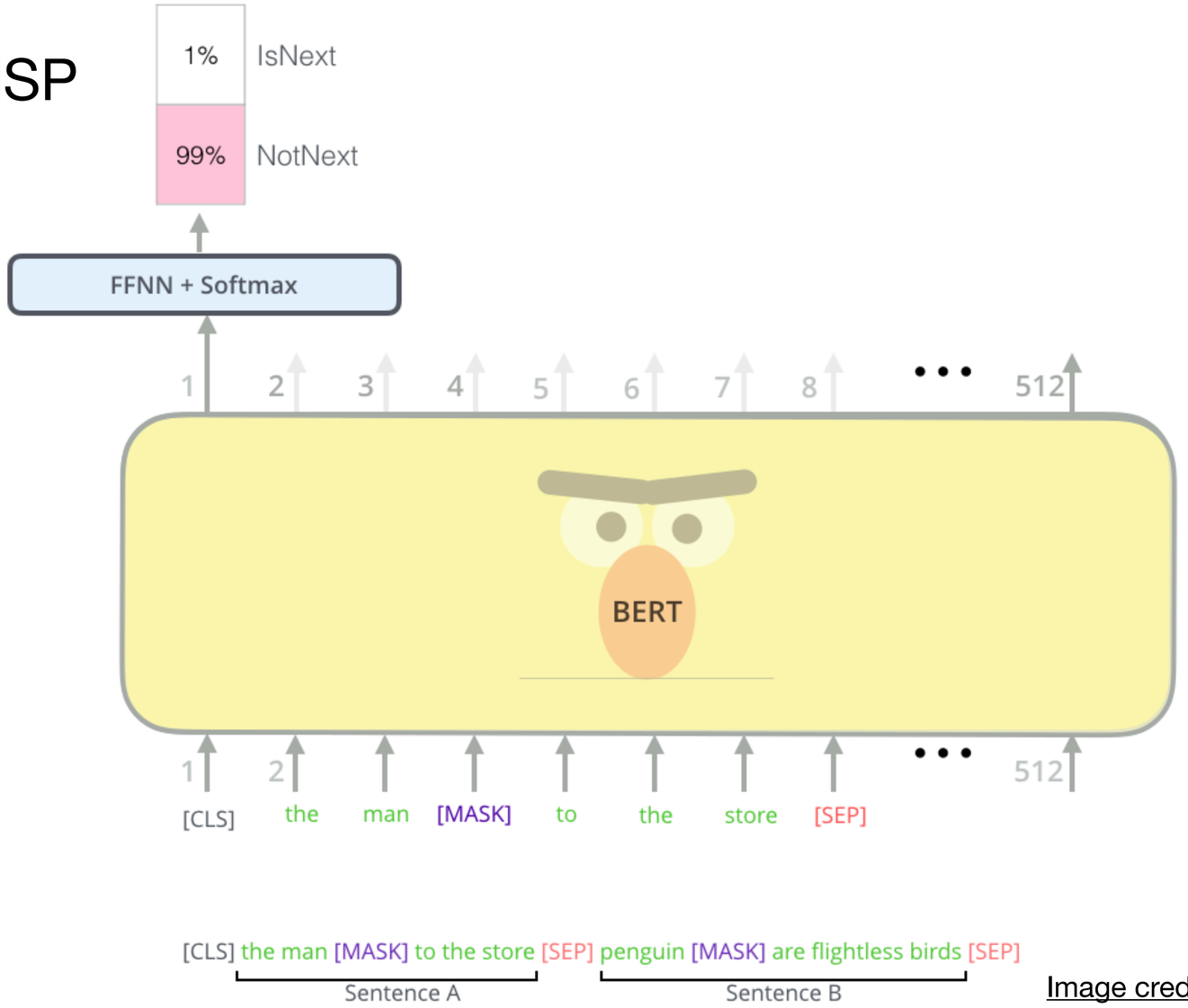


Image credit

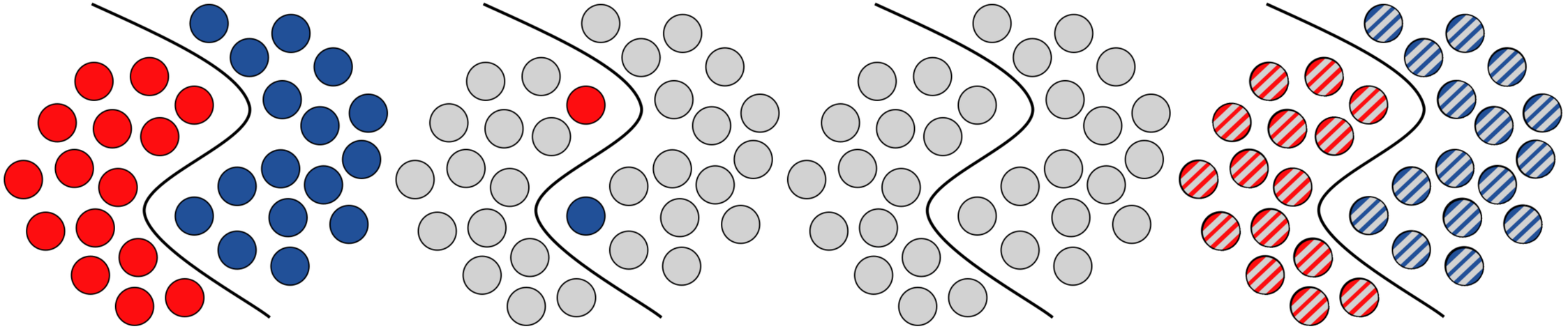
Self-supervised learning

Можно придумать **промежуточную задачу** для обучения модели на **незамеченных данных**

- Задача не должна требовать краудсорсинга для разметки
- Чем больше данных, тем лучше

CLIP, BERT, BART, T5, word2vec, autoencoders, etc.

Self-supervised learning



Supervised | Semi-supervised | Unsupervised | Self-supervised

Image credit

ALBERT: a lite BERT 

ALBERT: a lite BERT

Как уменьшить BERT, не потеряв в качестве?

- факторизация Embedding слоя

Model	Parameter size	Avg test scores (5 NLU tasks[1])
BERT_base	108M	82.3
ALBERT_base	89M	81.7

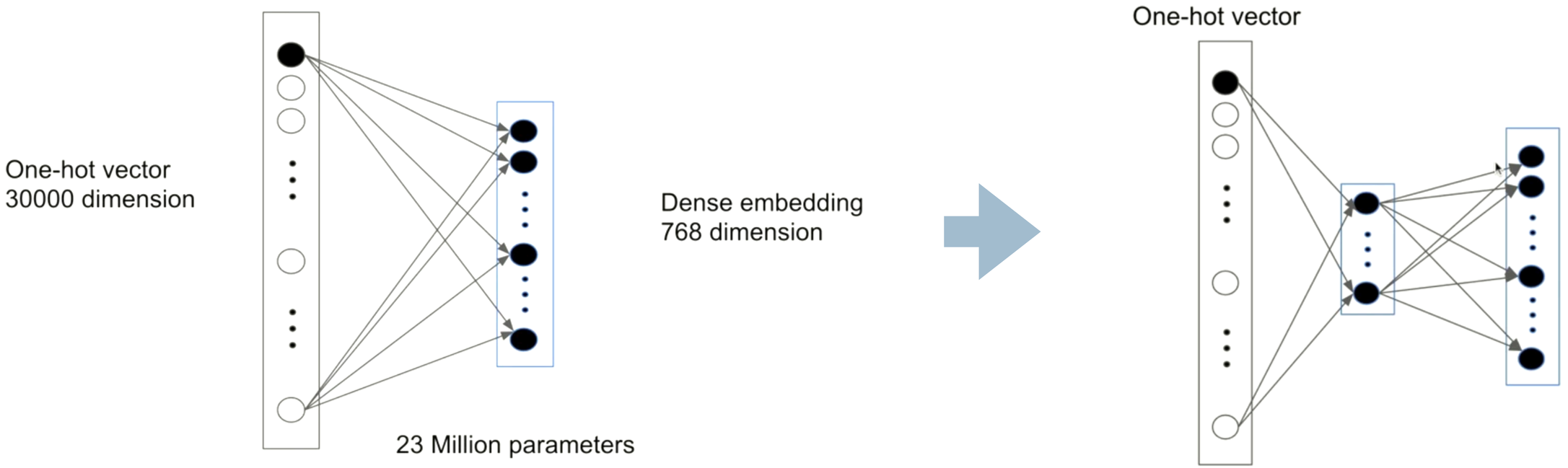


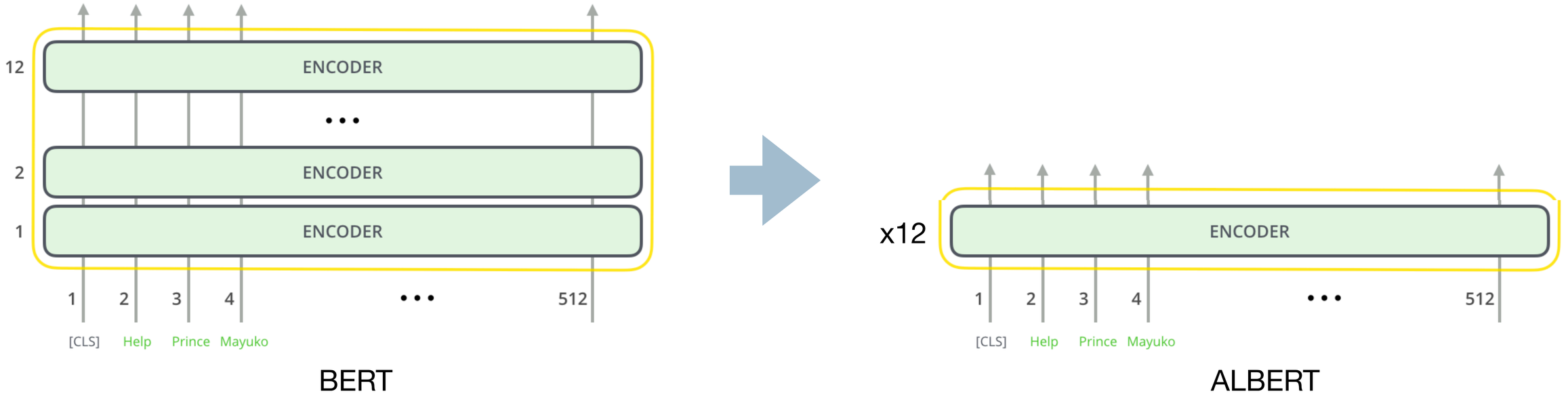
Image credit

ALBERT: a lite BERT

Как уменьшить BERT, не потеряв в качестве?

- sharing весов в Transformer блоках

Model	Parameter size	Avg test scores (5 NLU tasks[1])
BERT_base	108M	82.3
ALBERT_base	31M	79.8



BERT

ALBERT

Image credit

ALBERT: a lite BERT

Model	Avg test scores (5 NLU tasks[1])
Without SOP	79.0
With SOP	80.1

Как улучшить качество?

- вместо NSP используем задачу Sentence Order Prediction

Image credit

ALBERT: a lite BERT

Как улучшить качество?

- больше данных

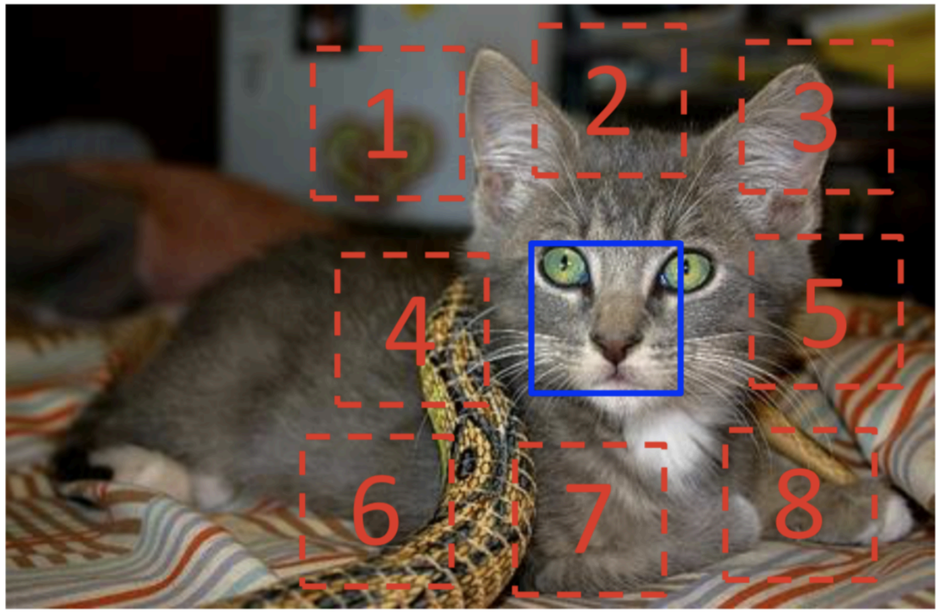
Model	Parameter size	Avg test scores (5 NLU tasks[1])
BERT_Large	334M	83.9
ALBERT_Large	**18M**	85.7
ALBERT_XXLarge	235M	**91.0**

Image credit

Visual representation learning

Visual representation learning: patches

Doersch et al. 2015



X = (,); Y = 3

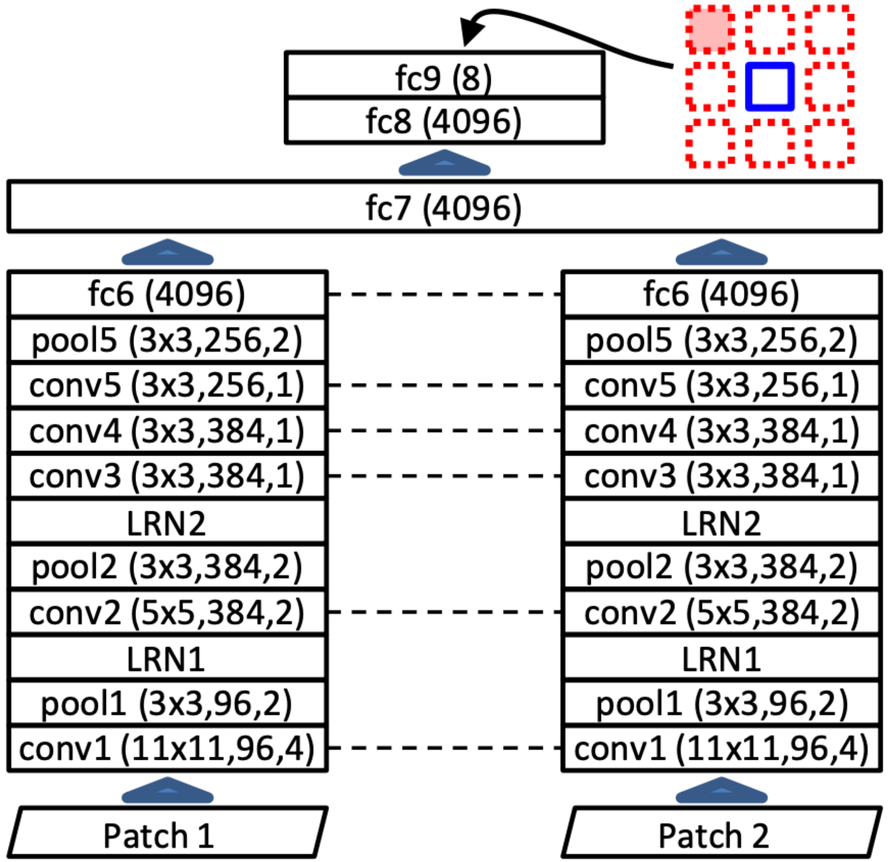
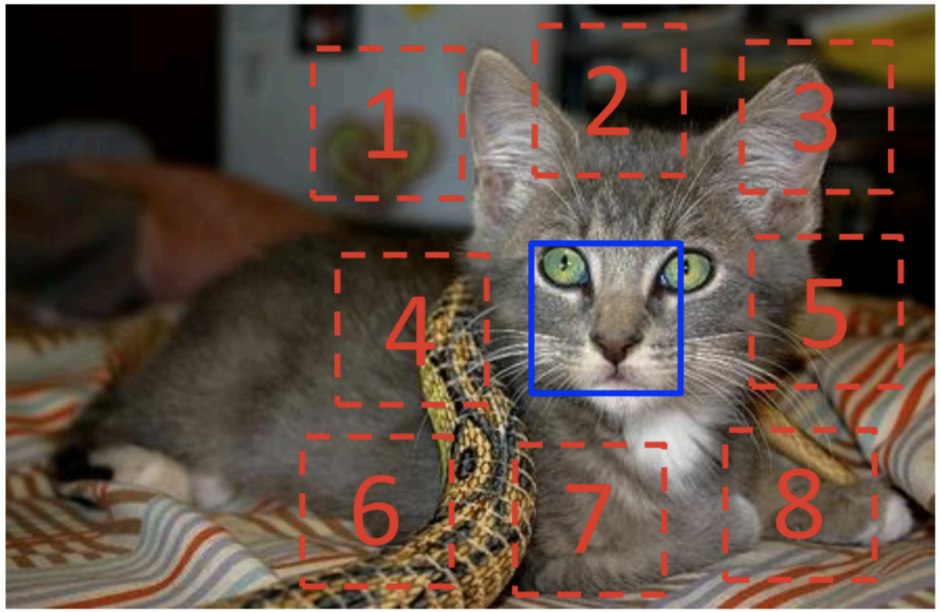
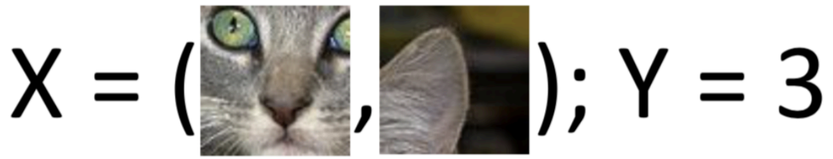


Image credit

Visual representation learning: patches

Doersch et al. 2015





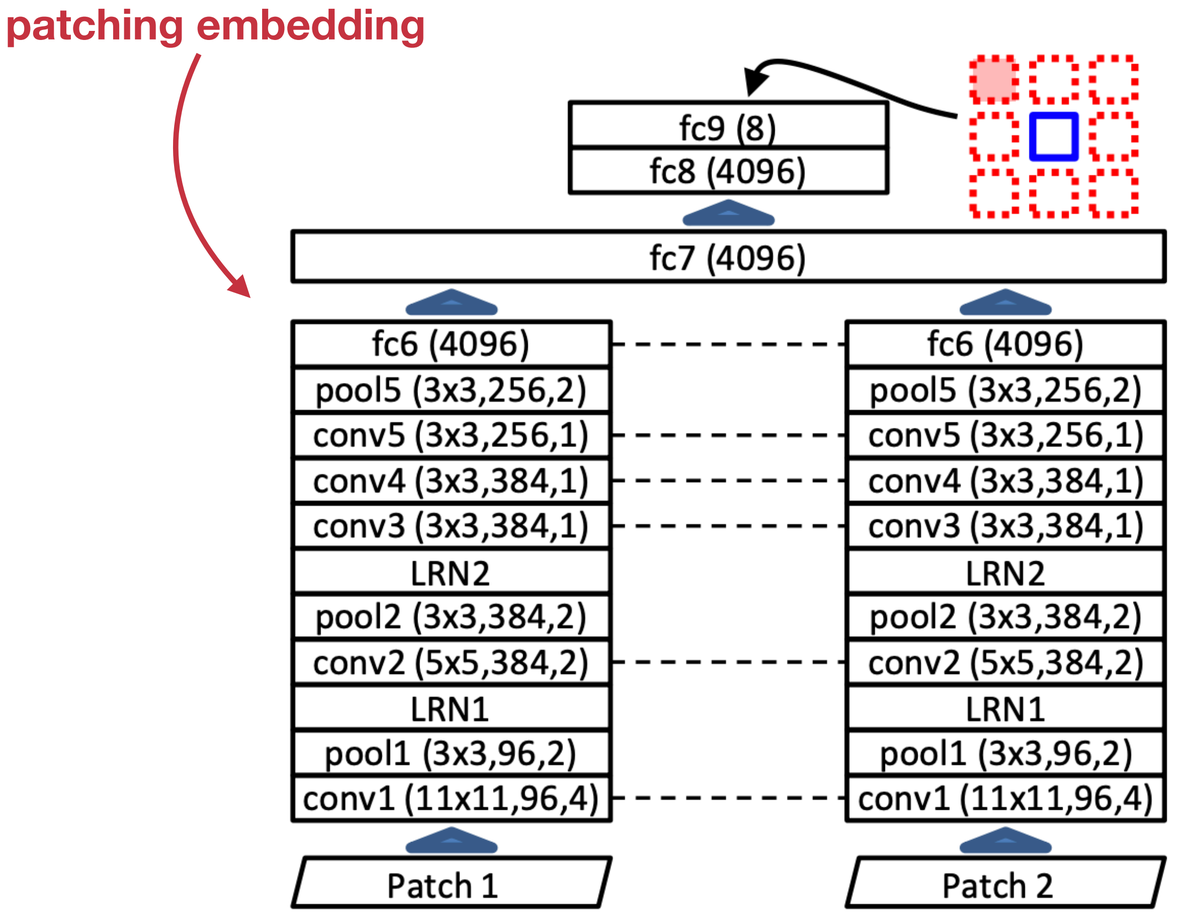


Image credit

Visual representation learning: patches

Doersch et al. 2015

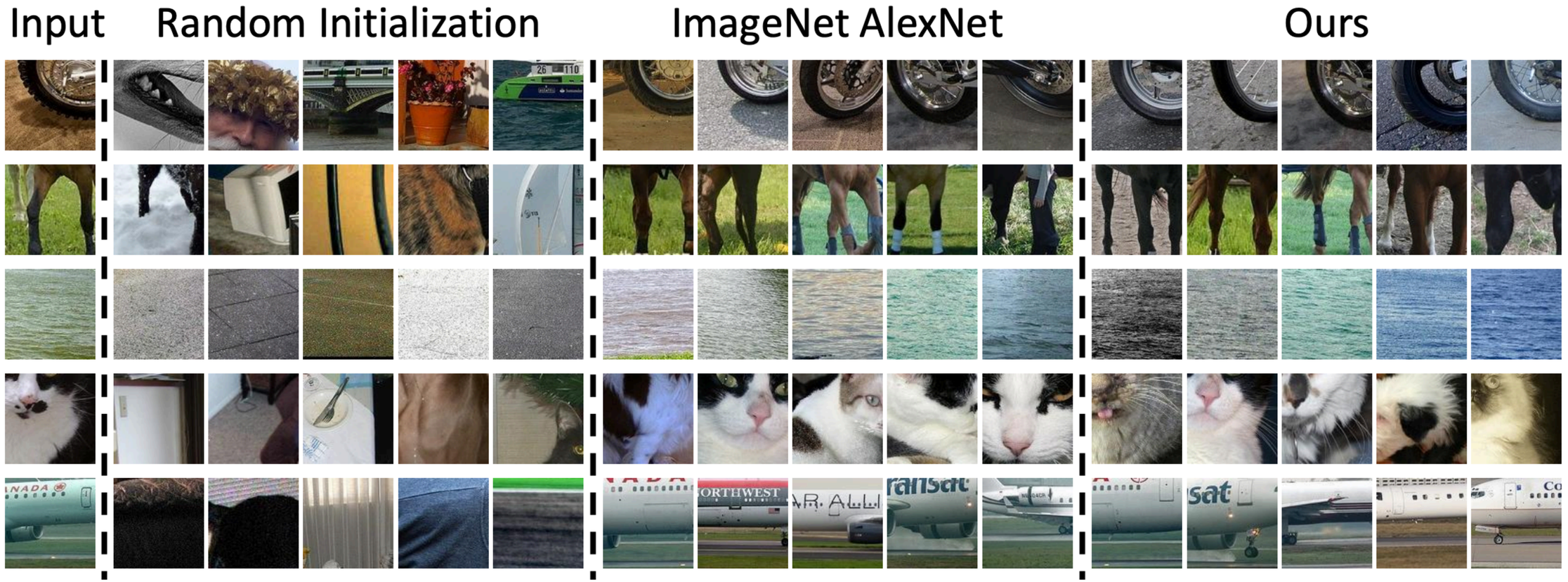
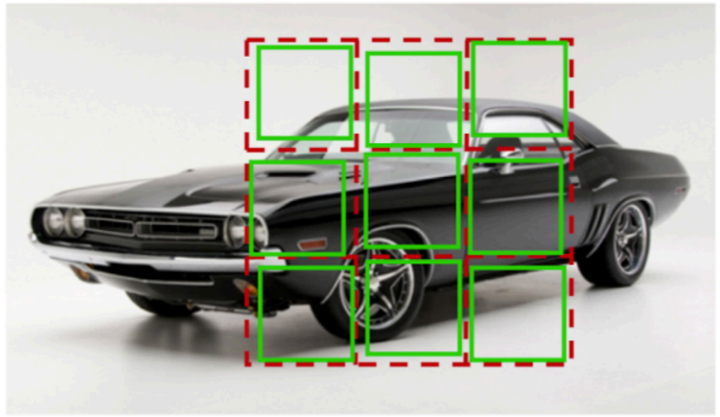


Image credit

Visual representation learning: patches

Noroozi & Favaro, 2016



Permutation Set

index	permutation
64	9,4,6,8,3,2,5,1,7

Reorder patches according to the selected permutation

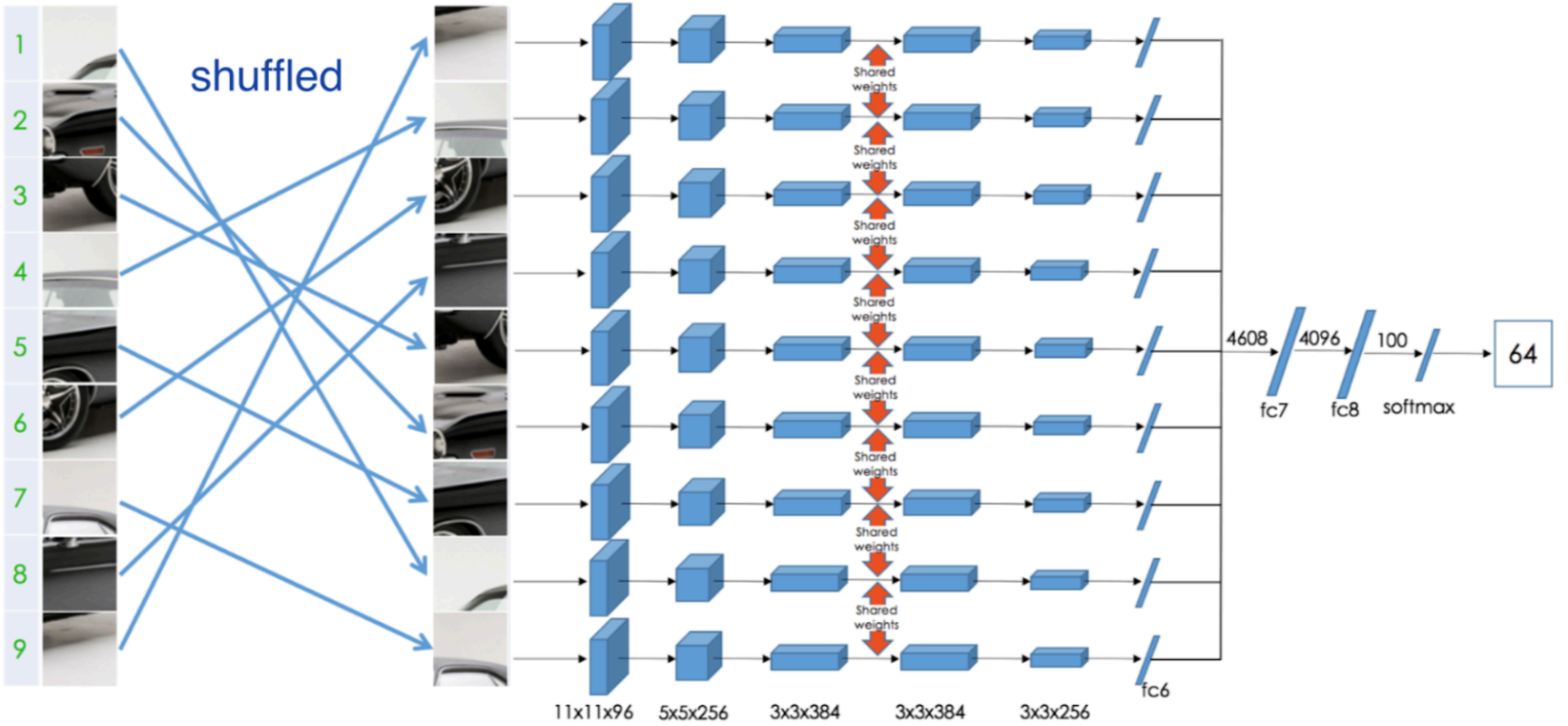
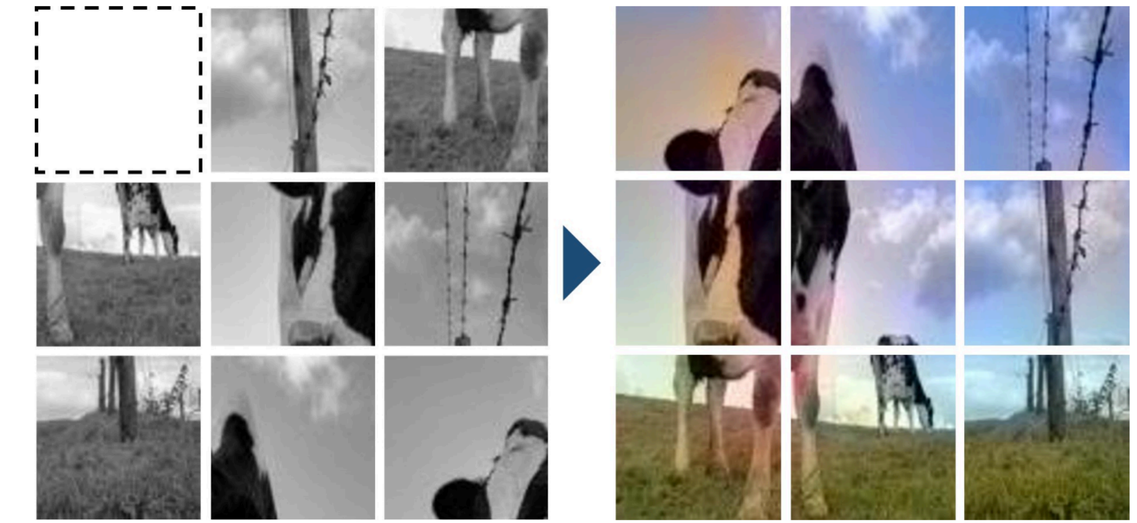


Image credit

Visual representation learning: patches

Kim et al. 2018



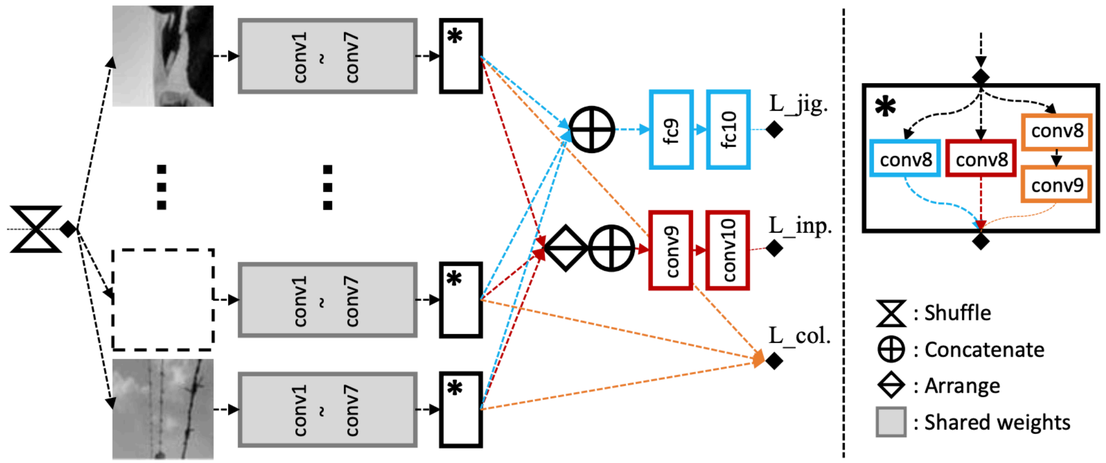
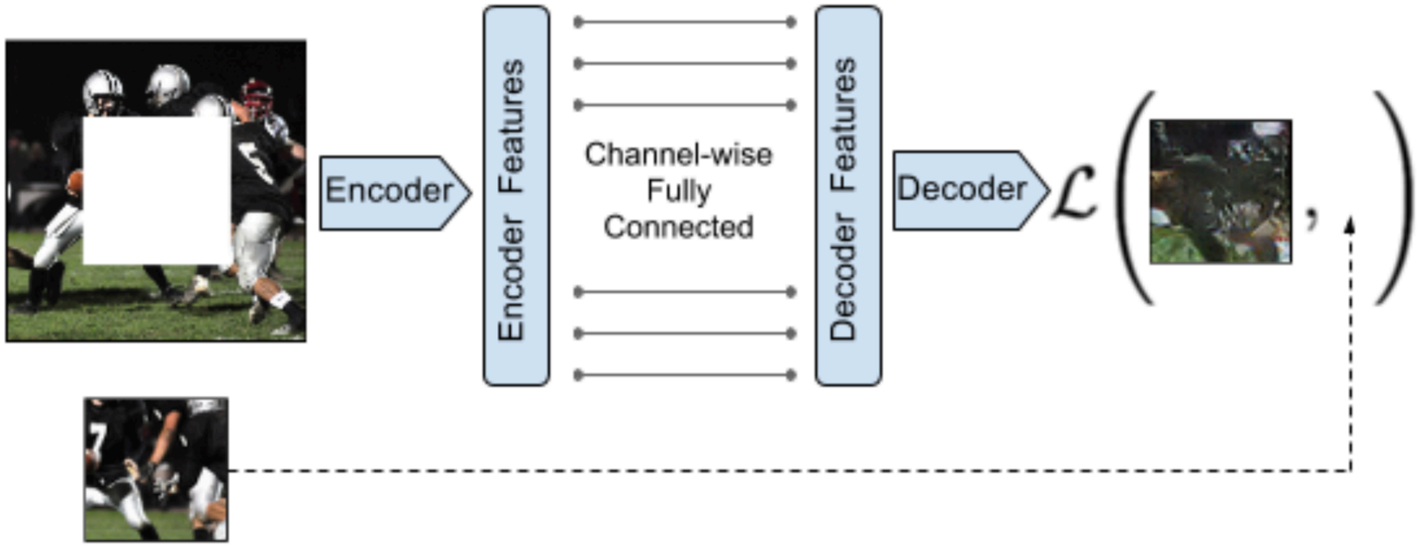


Image credit

Visual representation learning: patches

Pathak et al. 2016

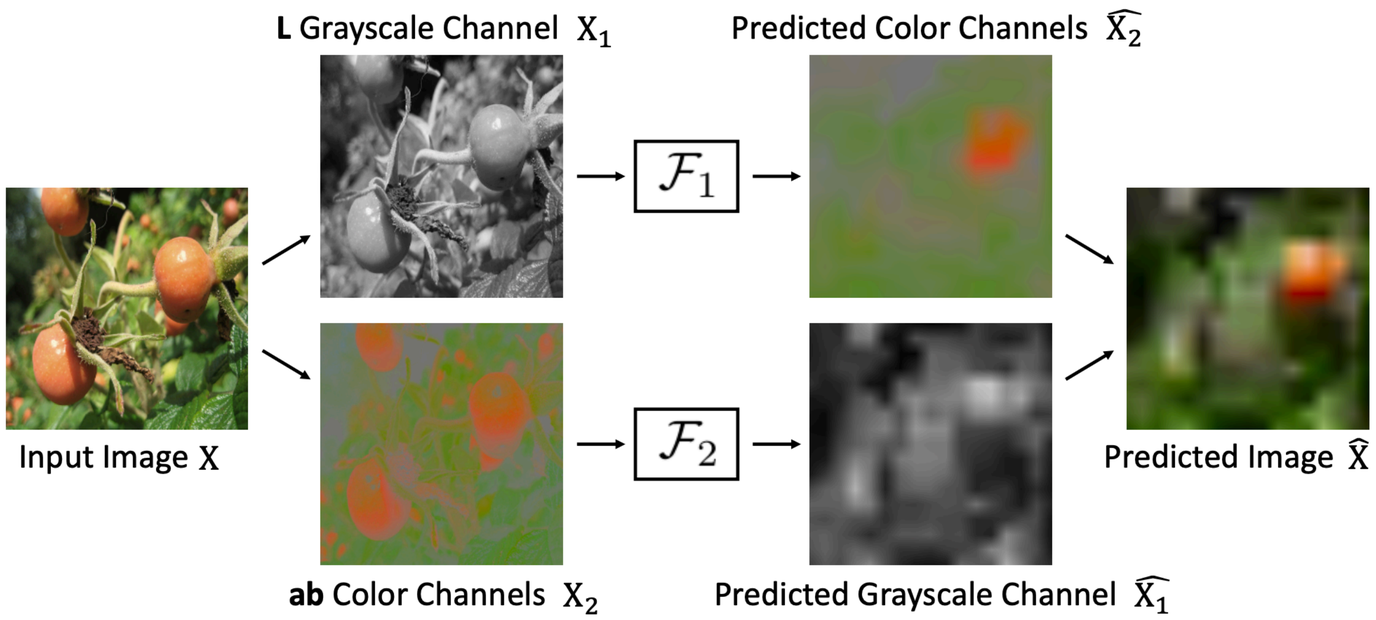


Context Encoder

Image credit

Colorization

Zhang et al. 2016



Spin-Brain Autoencoder

DeOldify project

Modern self-supervised models

SimCLR

Chen et al. 2020

- Given an image find its augmentation among other images.
- Use the rest of batch as negative examples instead of a memory bank (need large batches, N=2-4K)

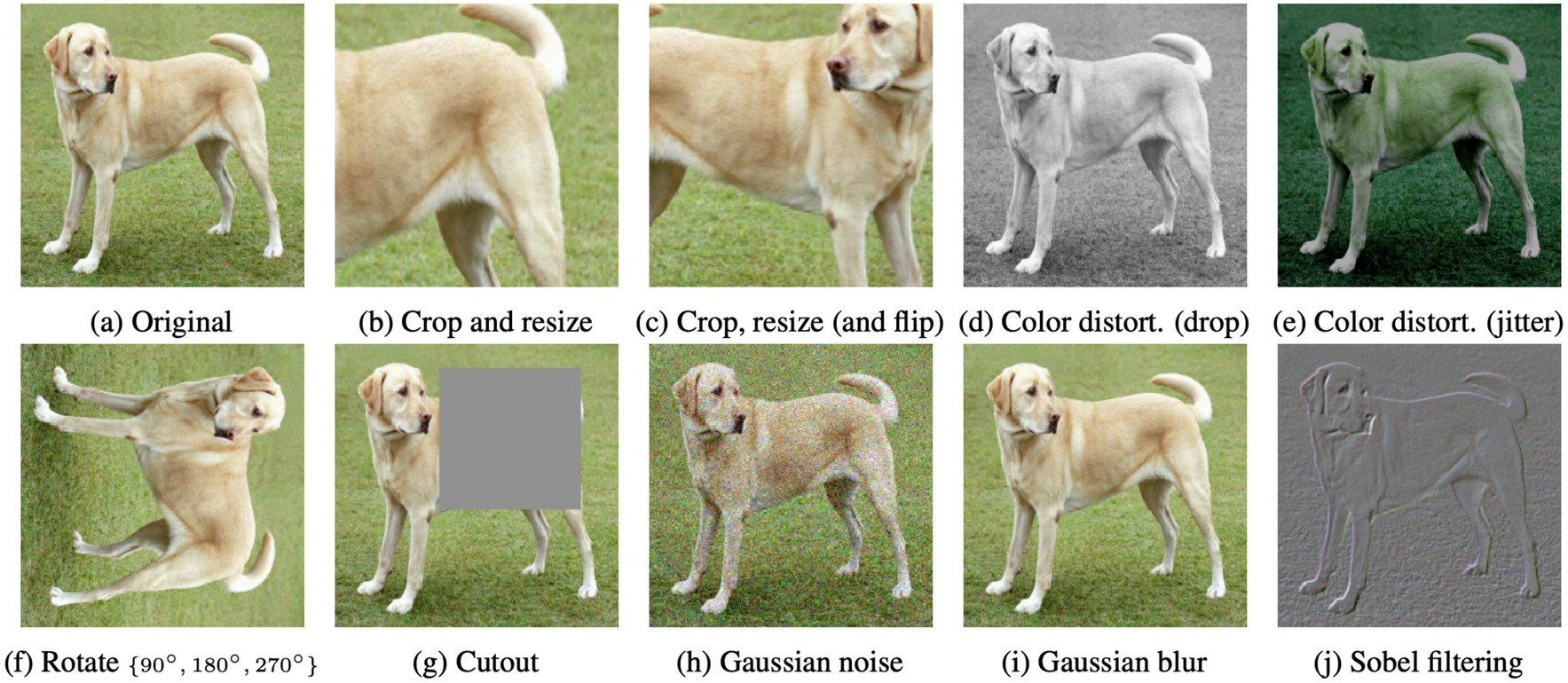
Batch of size N:

2N loss terms:

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

SimCLR

Chen et al. 2020



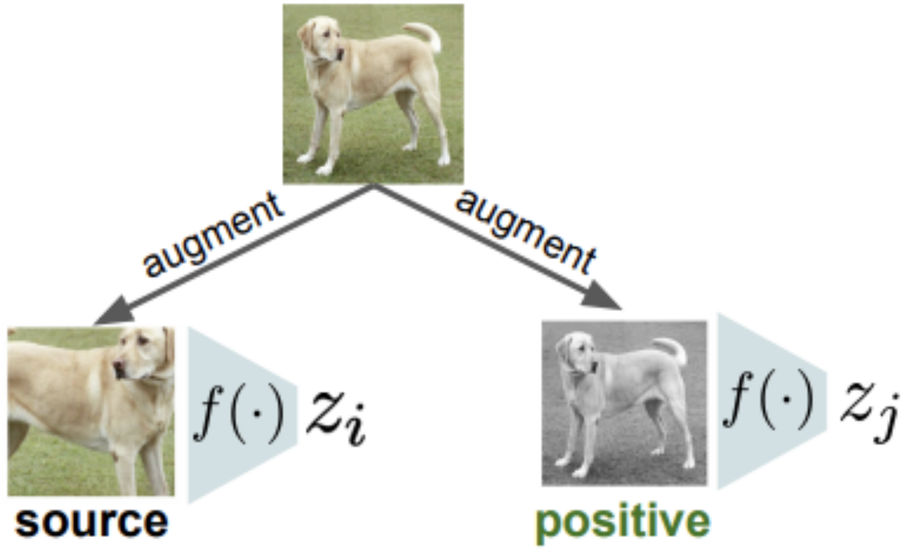
(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

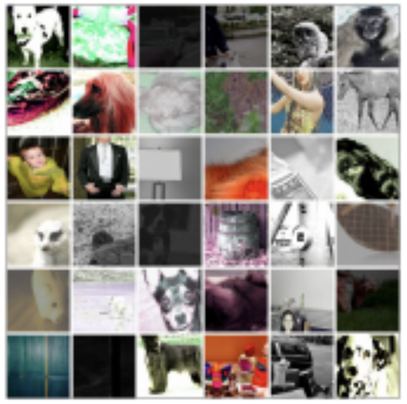
(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

Image credit

SimCLR

Chen et al. 2020



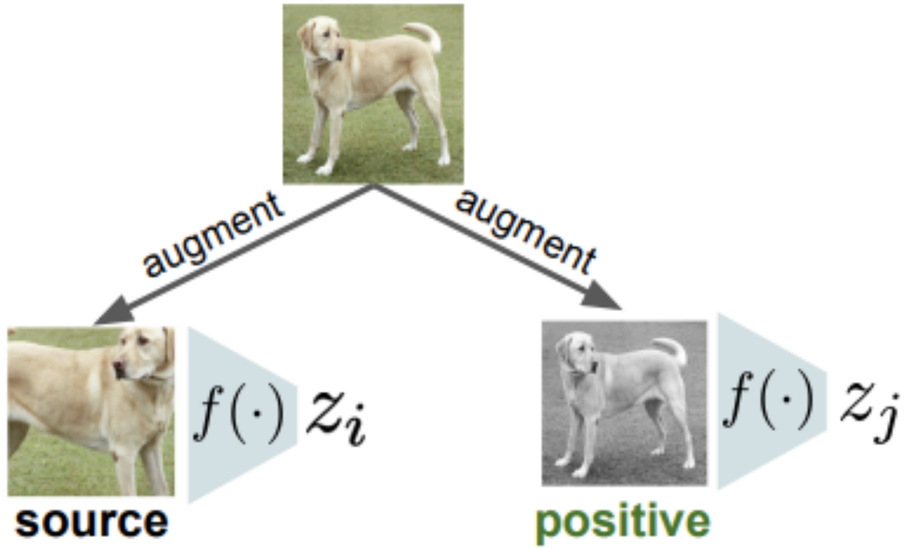


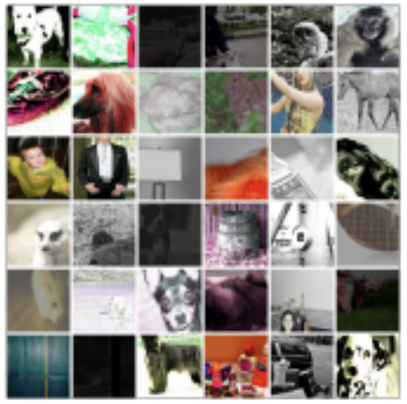
$$L_{i,j} = -\log\left(\frac{\exp(z_i^T z_j)}{\exp(z_i^T z_j) + \sum_l \exp(z_i^T z_l)}\right)$$

-contrastive loss

SimCLR

Chen et al. 2020





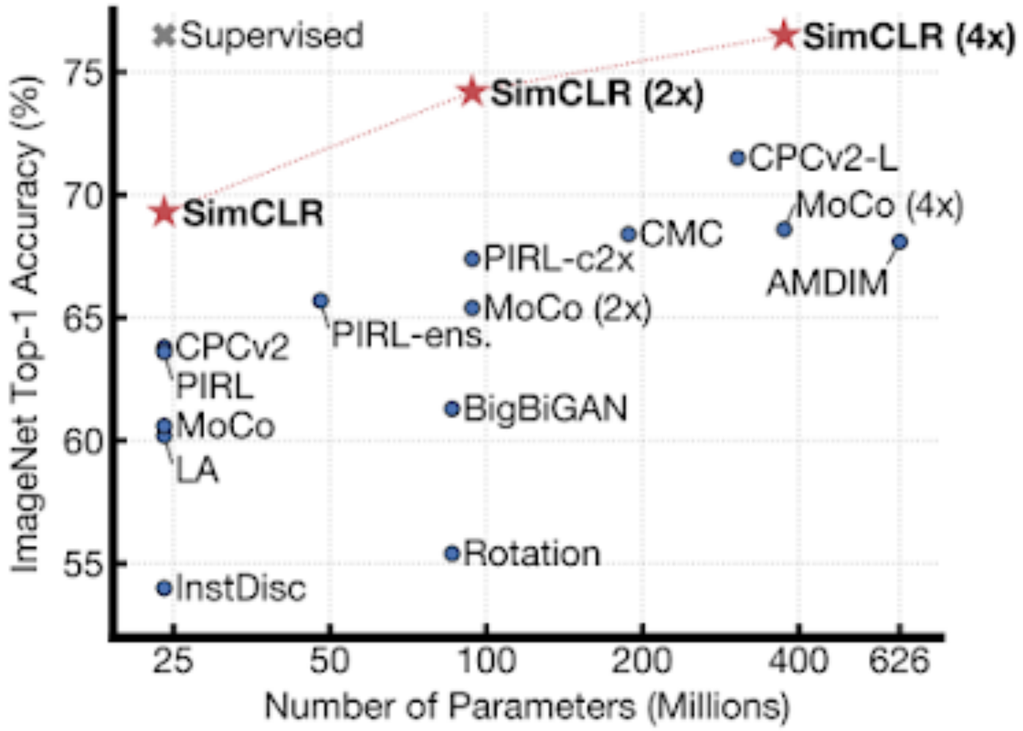
$$L_{i,j} = -\log\left(\frac{\exp(z_i^T z_j)}{\exp(z_i^T z_j) + \sum_l \exp(z_i^T z_l)}\right)$$

-contrastive loss

z - ResNet + projection head

SimCLR

Chen et al. 2020



ImageNet supervised (with labels) vs ImageNet self-supervised

Image credit

BYOL: Bootstrap Your Own Latent

Grill et al. 2020

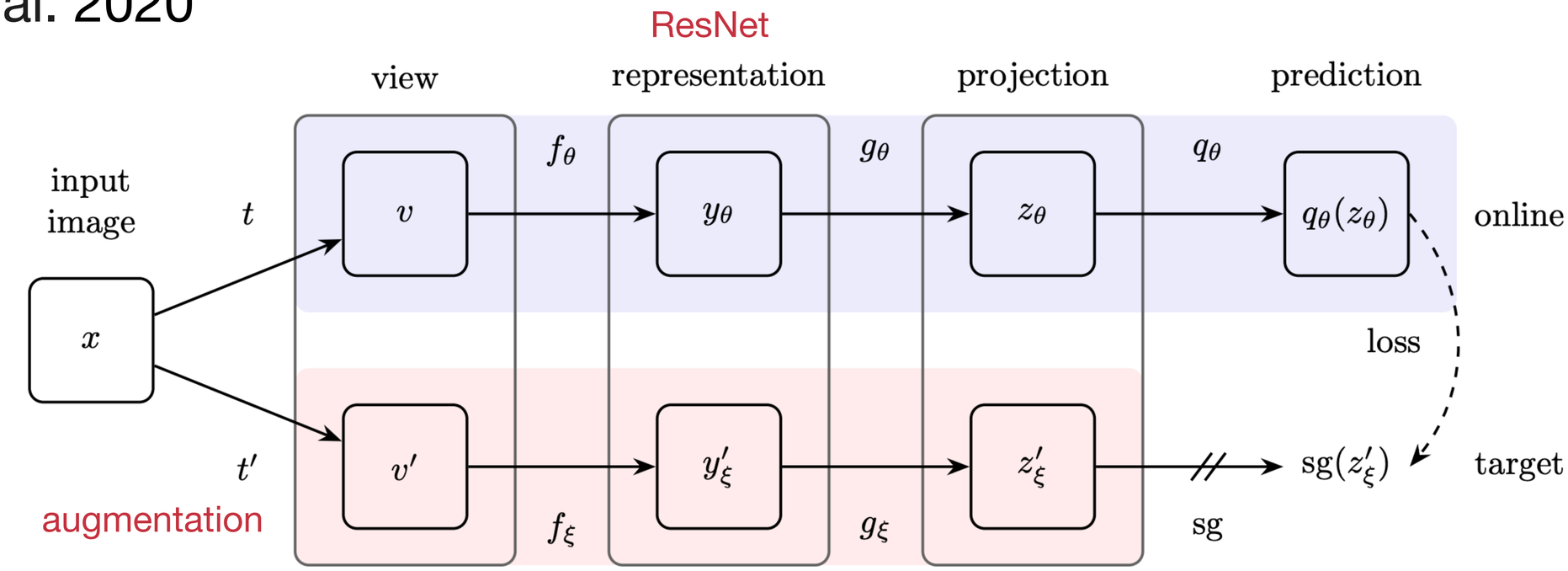
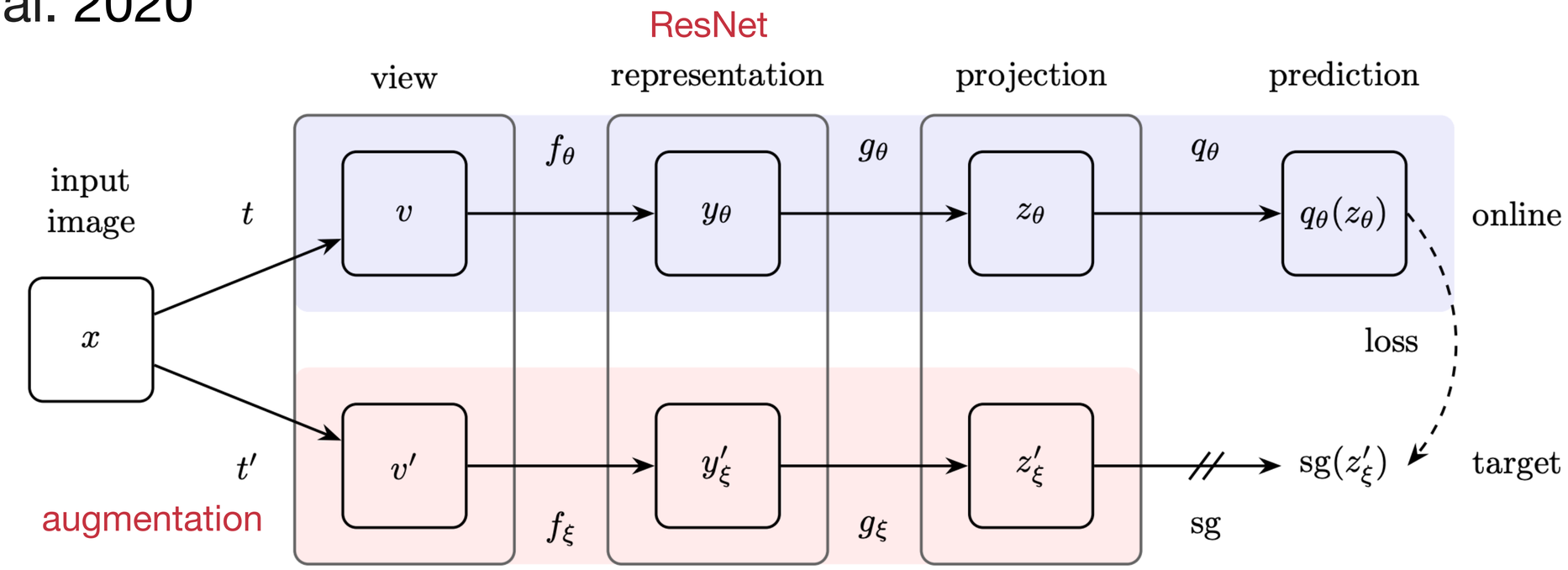


Image credit

BYOL: Bootstrap Your Own Latent

Grill et al. 2020

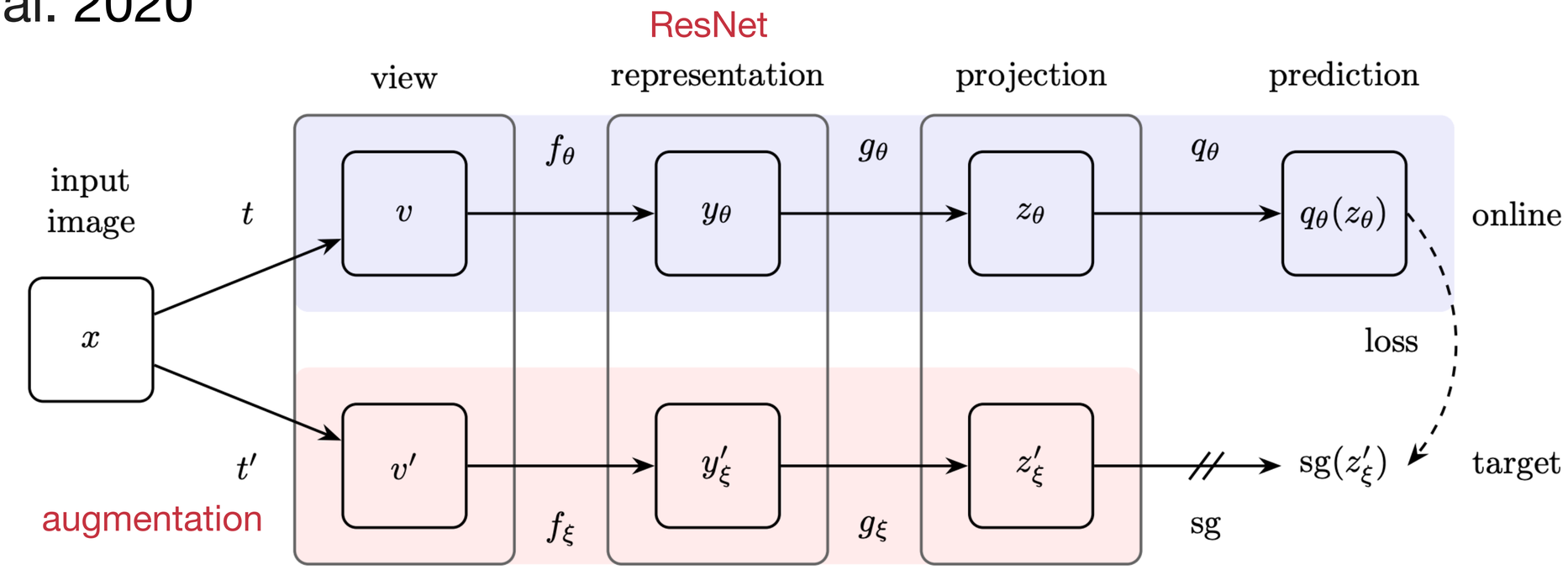


Online NN (θ) - обучается как обычно, L2 loss $\mathcal{L}_{\theta,\xi} \triangleq \left\| \overline{q_\theta}(z_\theta) - \overline{z}'_\xi \right\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\left\| q_\theta(z_\theta) \right\|_2 \cdot \left\| z'_\xi \right\|_2}$

Image credit

BYOL: Bootstrap Your Own Latent

Grill et al. 2020



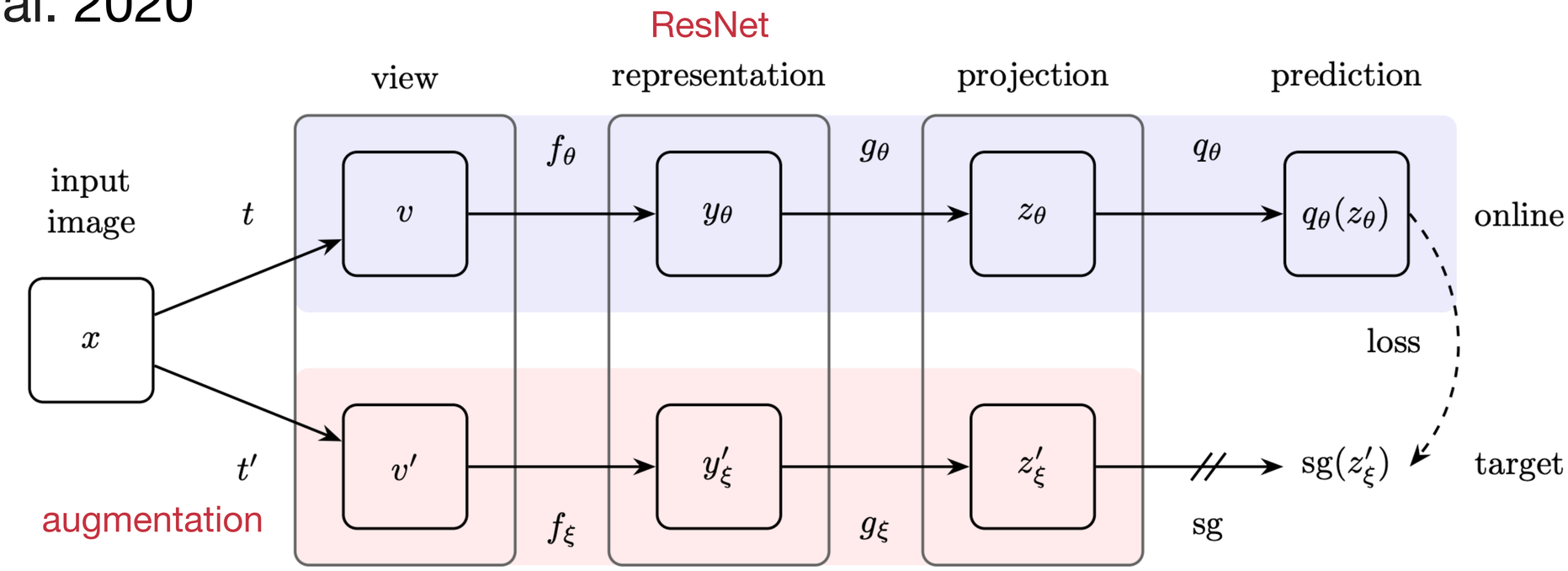
Online NN (θ) - обучается как обычно, L2 loss $\mathcal{L}_{\theta,\xi} \triangleq \left\|\overline{q_\theta}(z_\theta) - \overline{z}'_\xi\right\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\left\|q_\theta(z_\theta)\right\|_2 \cdot \left\|z'_\xi\right\|_2}$

Target NN (ξ) - обновляется через скользящее среднее $\xi \leftarrow \tau\xi + (1-\tau)\theta$

Image credit

BYOL: Bootstrap Your Own Latent

Grill et al. 2020

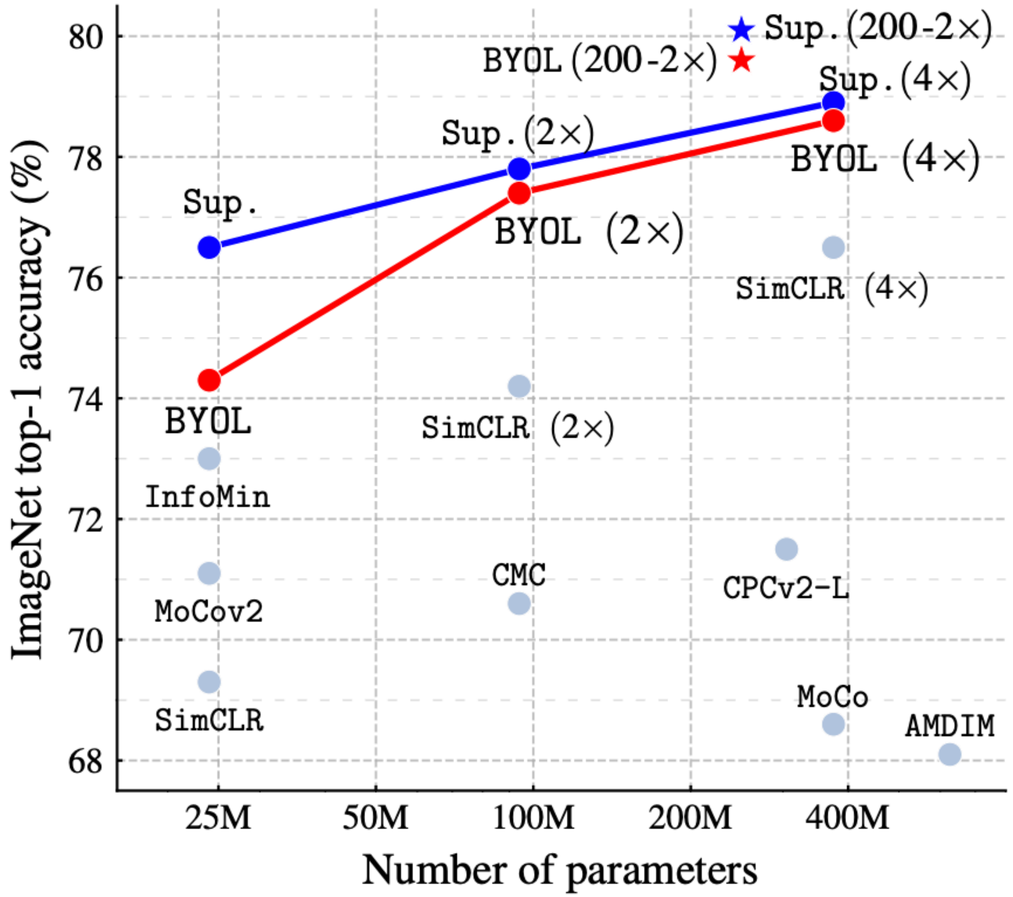


Не нужны negative examples

Image credit

BYOL: Bootstrap Your Own Latent

Grill et al. 2020



ImageNet supervised (with labels) vs ImageNet self-supervised

Image credit

DINO: self-supervised Transformer

Caron et al. 2021

- взяли ViT в качестве NN

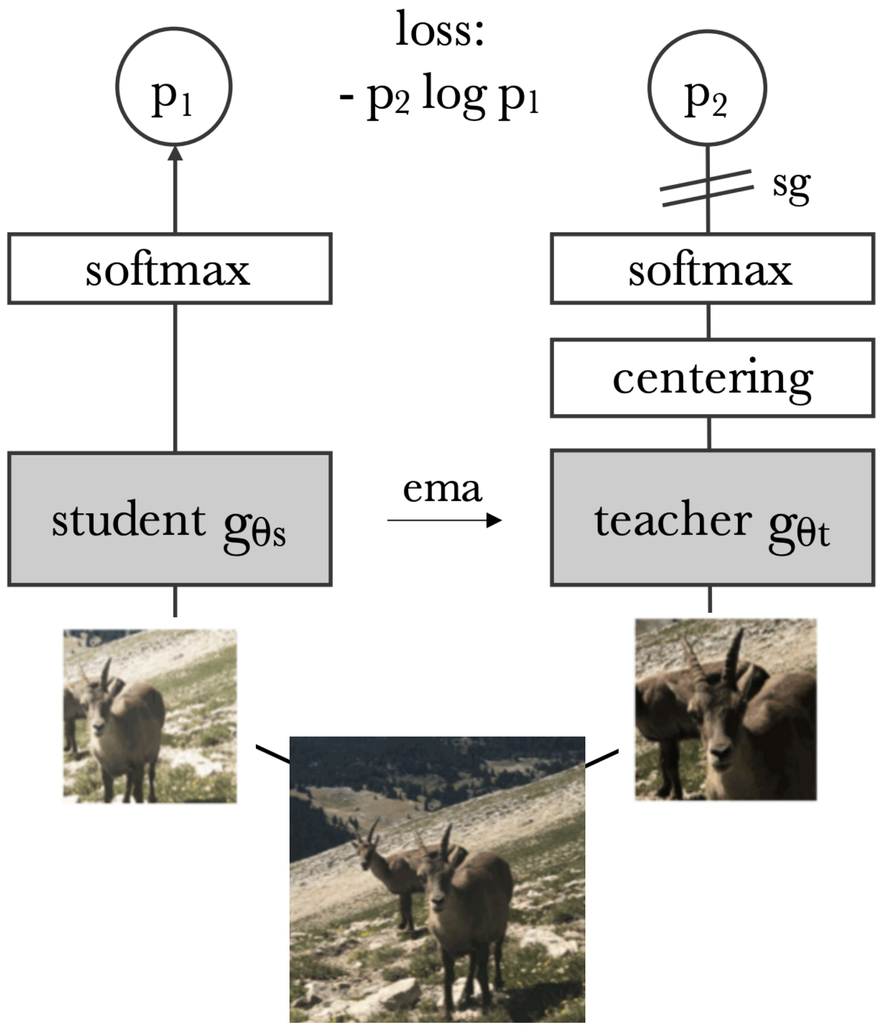


Image credit

DINO: self-supervised Transformer

Caron et al. 2021

Method	Arch.	Param.	im/s	Linear	k-NN
Supervised	RN50	23	1237	79.3	79.3
SCLR [12]	RN50	23	1237	69.1	60.7
MoCov2 [15]	RN50	23	1237	71.1	61.9
InfoMin [67]	RN50	23	1237	73.0	65.3
BarlowT [81]	RN50	23	1237	73.2	66.0
OBoW [27]	RN50	23	1237	73.8	61.9
BYOL [30]	RN50	23	1237	74.4	64.8
DCv2 [10]	RN50	23	1237	75.2	67.1
SwAV [10]	RN50	23	1237	**75.3**	65.7
DINO	RN50	23	1237	**75.3**	**67.5**
Supervised	ViT-S	21	1007	79.8	79.8
BYOL* [30]	ViT-S	21	1007	71.4	66.6
MoCov2* [15]	ViT-S	21	1007	72.7	64.4
SwAV* [10]	ViT-S	21	1007	73.5	66.3
DINO	ViT-S	21	1007	**77.0**	**74.5**

Image credit

DINO: self-supervised Transformer

Caron et al. 2021

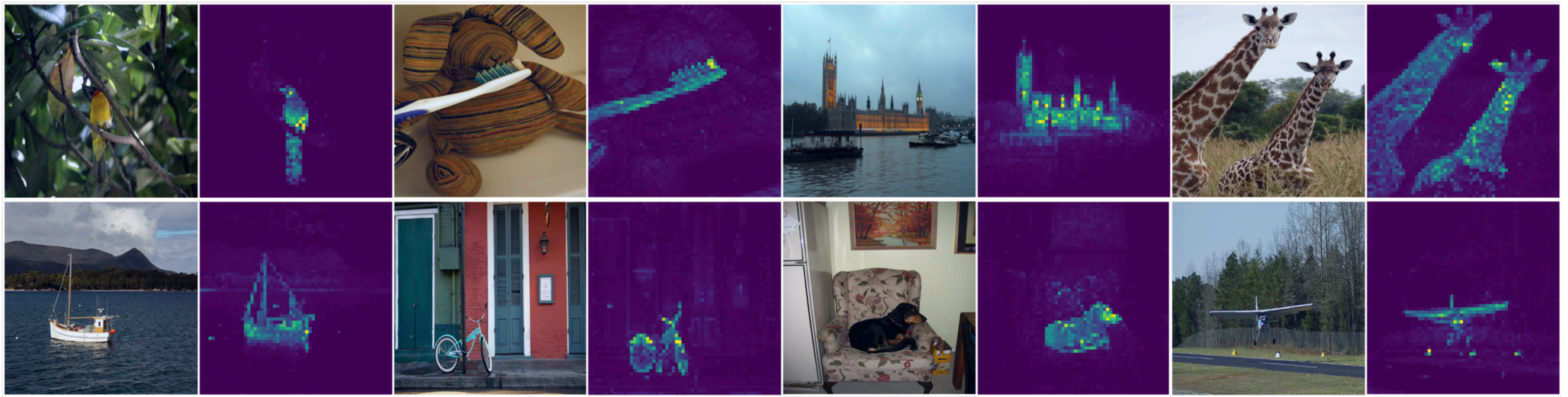
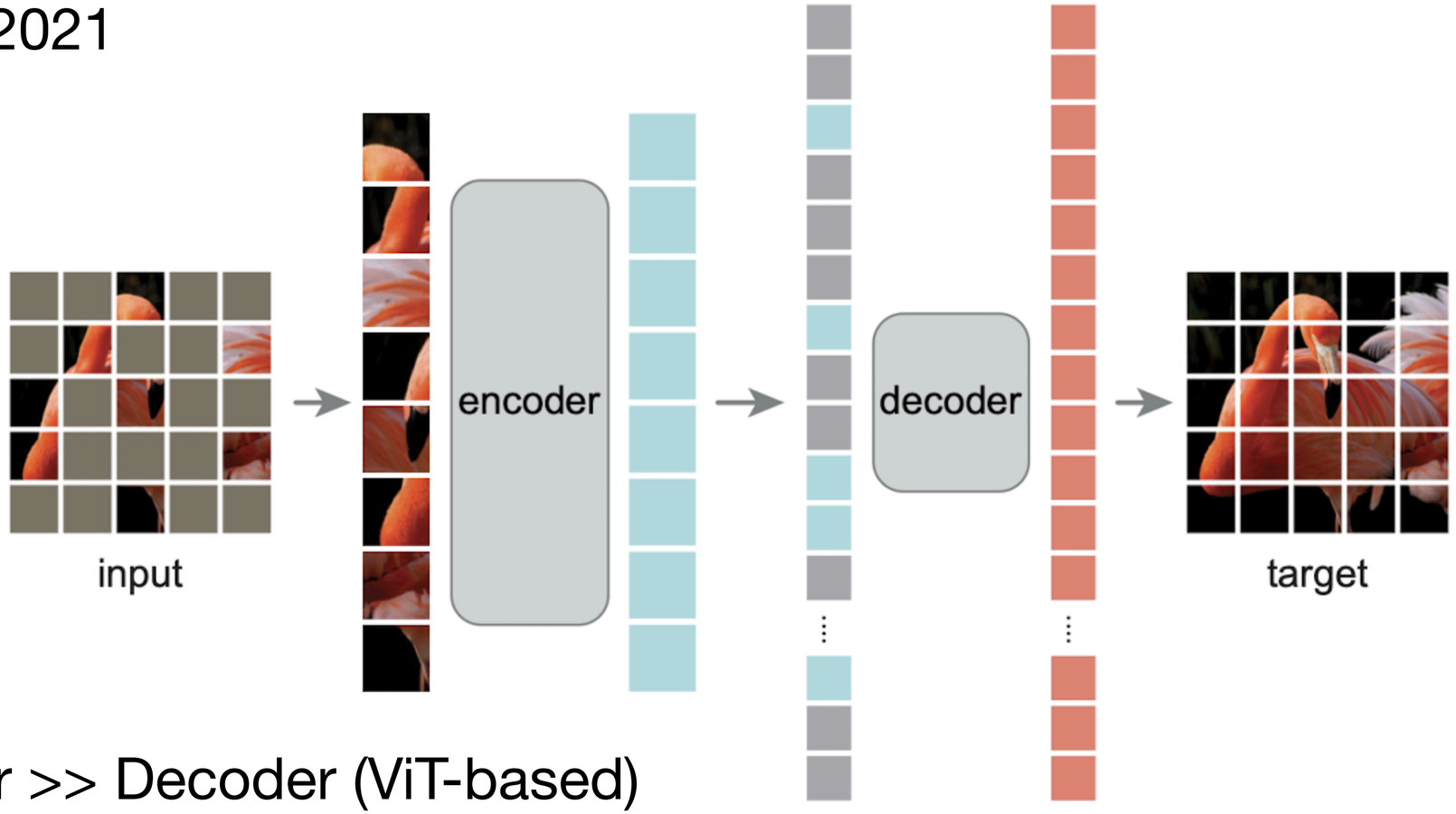


Image credit

Masked Autoencoder Transformer

He et al. 2021



- Encoder >> Decoder (ViT-based)
- Train: MSE loss
- Inference: only Encoder

Image credit

Masked Autoencoder Transformer

He et al. 2021

Image credit

Masked Autoencoder Transformer

He et al. 2021

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**

ImageNet results

Image credit

Masked Autoencoder Transformer

He et al. 2021

		AP^{box}		AP^{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**

Object Detection

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**

Semantic Segmentation

Image credit