Deep Learning for Vision & Language

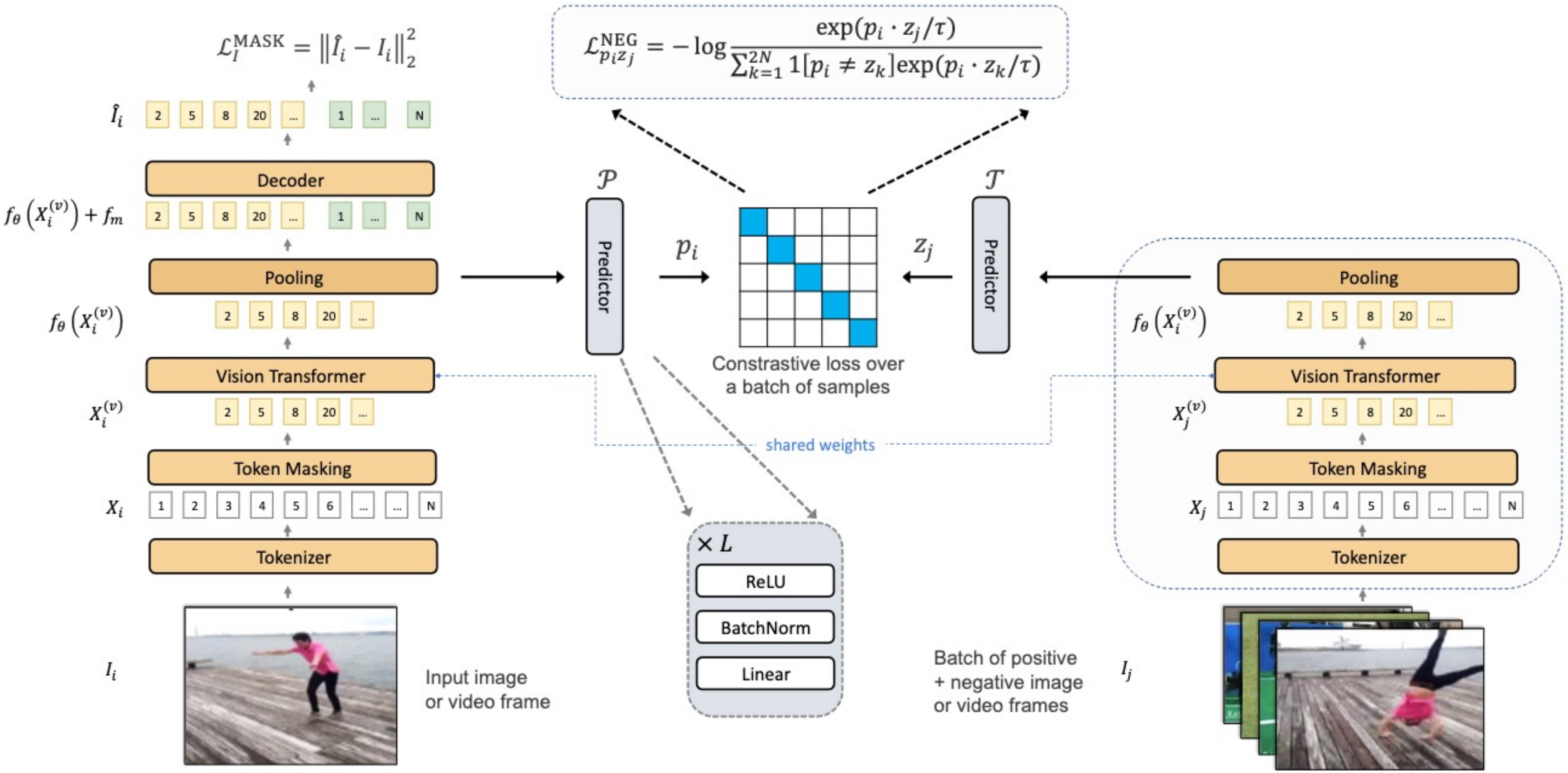
Self-supervised Models for Computer Vision

RICE UNIVERSITY

ViC-MAE: Self-Supervised Representation Learning from Images and Video with Contrastive Masked Autoencoders

Jefferson Hernandez[1], Ruben Villegas[2], Vicente Ordonez[1]
[1]Rice University, [2]Google DeepMind
{jefehern, vicenteor}@rice.edu, rubville@google.com

ViC-MAE



$\mathcal{L}_I^{\mathrm{MASK}} = \|\hat{I}_i - I_i\|_2^2$
$\mathcal{L}_{p_i z_j}^{\mathrm{NEG}} = -\log \frac{\exp(p_i \cdot z_j/\tau)}{\sum_{k=1}^{2N} 1[p_i \neq z_k]\exp(p_i \cdot z_k/\tau)}$
$\hat{I}_i$
Decoder
$f_\theta\left(X_i^{(v)}\right) + f_m$
Pooling
$f_\theta\left(X_i^{(v)}\right)$
Vision Transformer
$X_i^{(v)}$
Token Masking
X_i
Tokenizer
I_i
Input image
or video frame
$\mathcal{P}$
Predictor
p_i
Constrastive loss over
a batch of samples
z_j
$\mathcal{T}$
Predictor
shared weights
$\times L$
ReLU
BatchNorm
Linear
Pooling
$f_\theta\left(X_i^{(v)}\right)$
Vision Transformer
$X_j^{(v)}$
Token Masking
X_j
Tokenizer
Batch of positive
+ negative image
or video frames
I_j

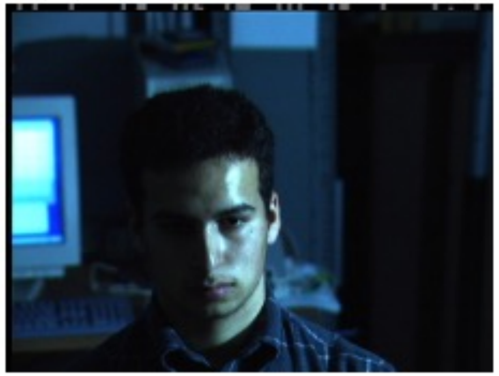
Self-Supervision for Visual Model Learning

- Lots of data but no labels
- Labeling data is expensive

Similarity Learning: Triplet Loss (Supervised)

x_i^a

x_i^p



x_i^n

$$\sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

FaceNet: A Unified Embedding for Face Recognition and Clustering

https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html

Similarity Learning: Triplet Loss (Self Supervised)

x_i^a

x_i^p

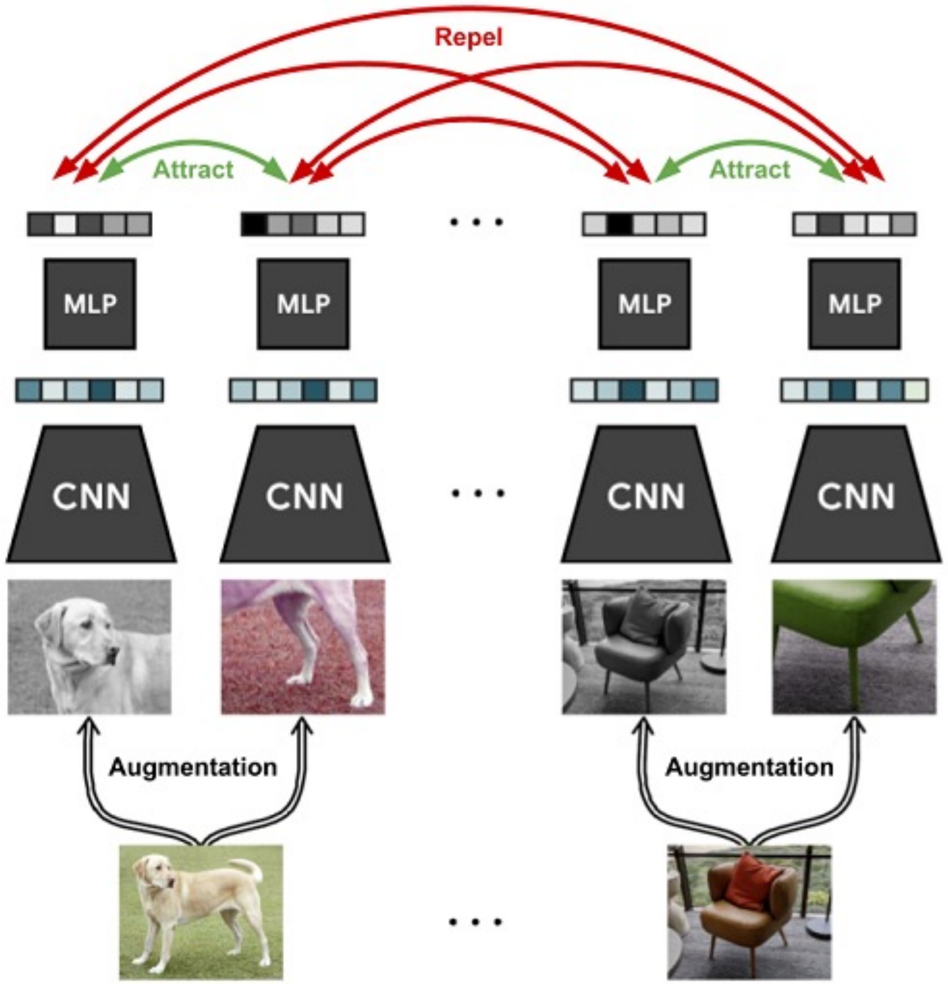
x_i^n

$$\sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

FaceNet: A Unified Embedding for Face Recognition and Clustering

https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Schroff_FaceNet_A_Unified_2015_CVPR_paper.html

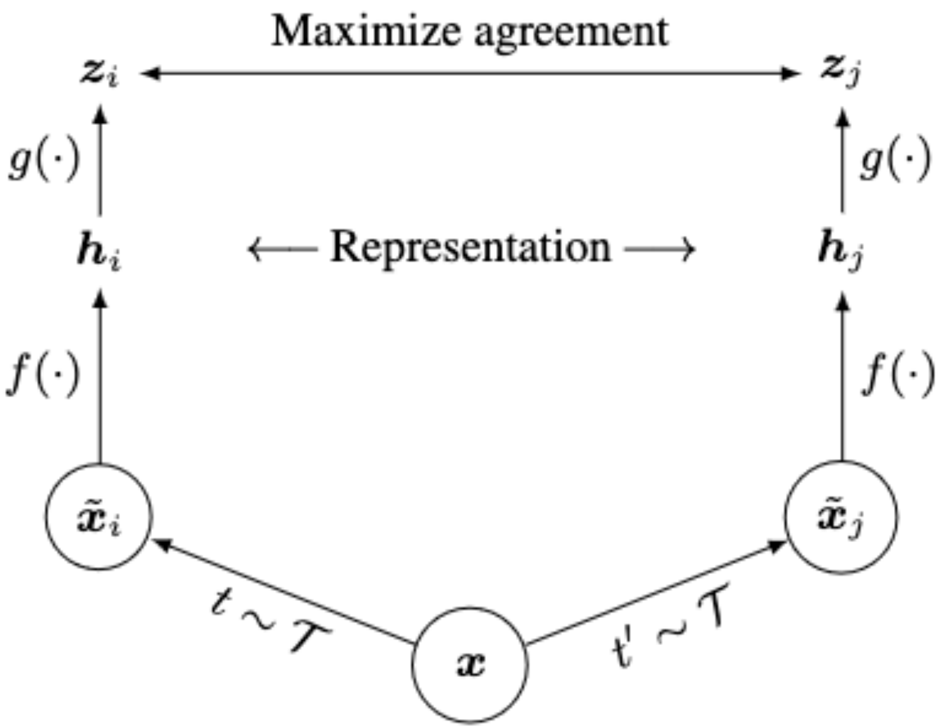
SimCLR: Contrastive Learning



A Simple Framework for Contrastive Learning of Visual Representations

https://arxiv.org/abs/2002.05709

Contrastive Learning



$$\boldsymbol{h}_i = f(\tilde{\boldsymbol{x}}_i) = \mathrm{ResNet}(\tilde{\boldsymbol{x}}_i)$$

$$\boldsymbol{z}_i = g(\boldsymbol{h}_i) = W^{(2)}\sigma(W^{(1)}\boldsymbol{h}_i)$$

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

Positive pair

Negative pairs

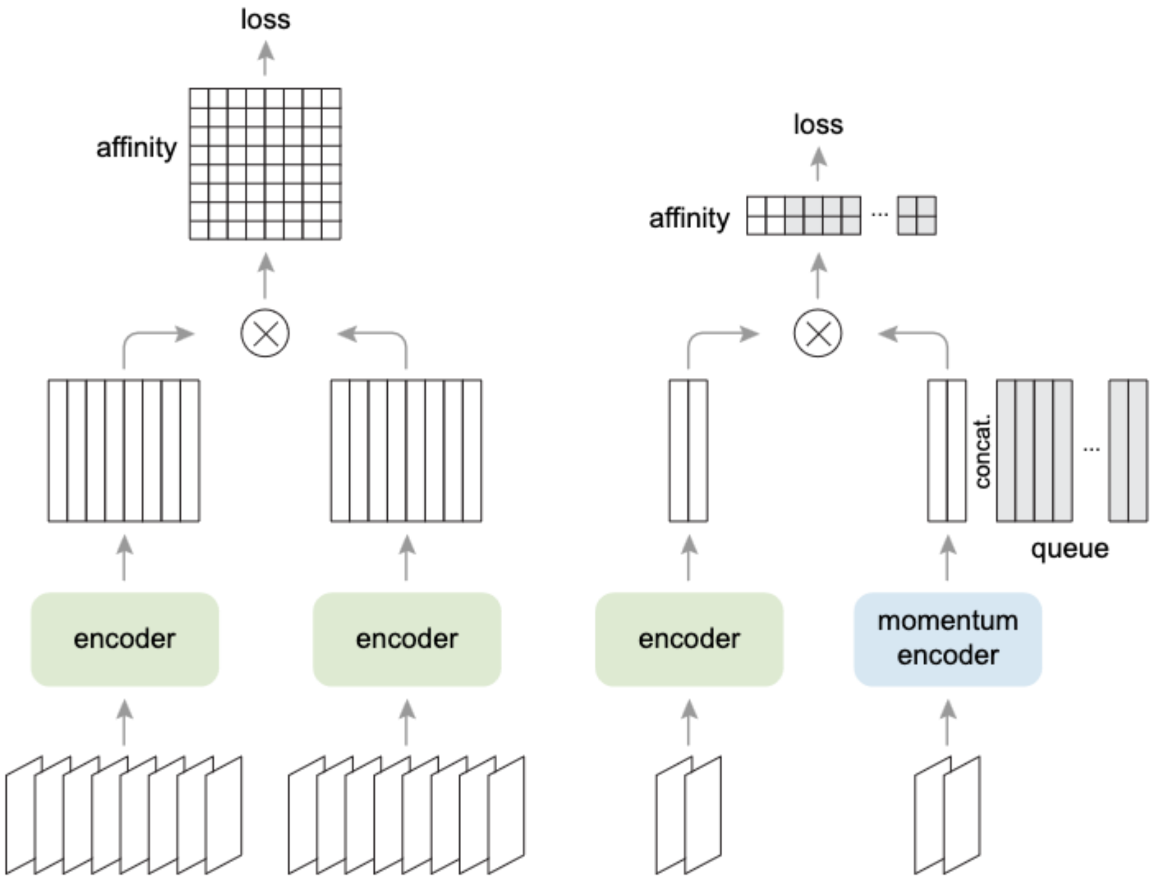
A Simple Framework for Contrastive Learning of Visual Representations

https://arxiv.org/abs/2002.05709

Issue with Contrastive Learning

- Large number of negative examples in the denominator are needed
- In practice this is approximated through batches – all other elements in the batch are negatives.
- Random negative examples are easy
- Large number of negative examples means larger batch sizes which occupy more memory
- GPU memory is expensive
  (What is the max memory a GPU has these days vs your PC?)

Momentum Contrastive Learning (MoCo)

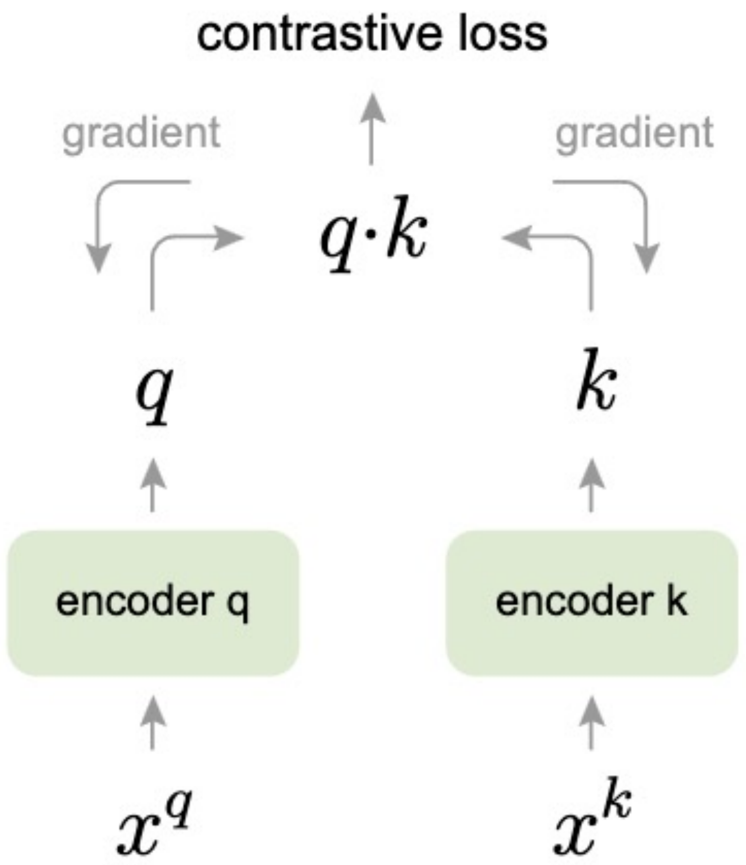


(a) end-to-end (b) Momentum Contrast

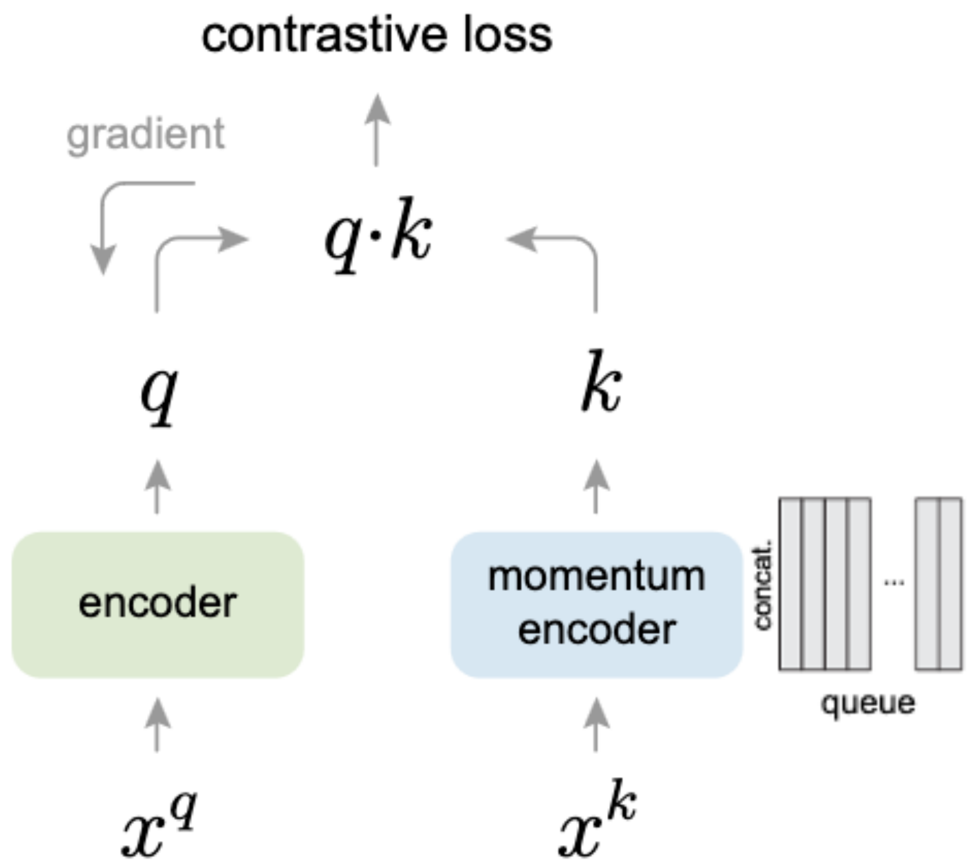
Improved Baselines with Momentum Contrastive Learning

https://arxiv.org/abs/2003.04297 https://arxiv.org/abs/1911.05722

Momentum Contrastive Learning (MoCo)



(a) end-to-end

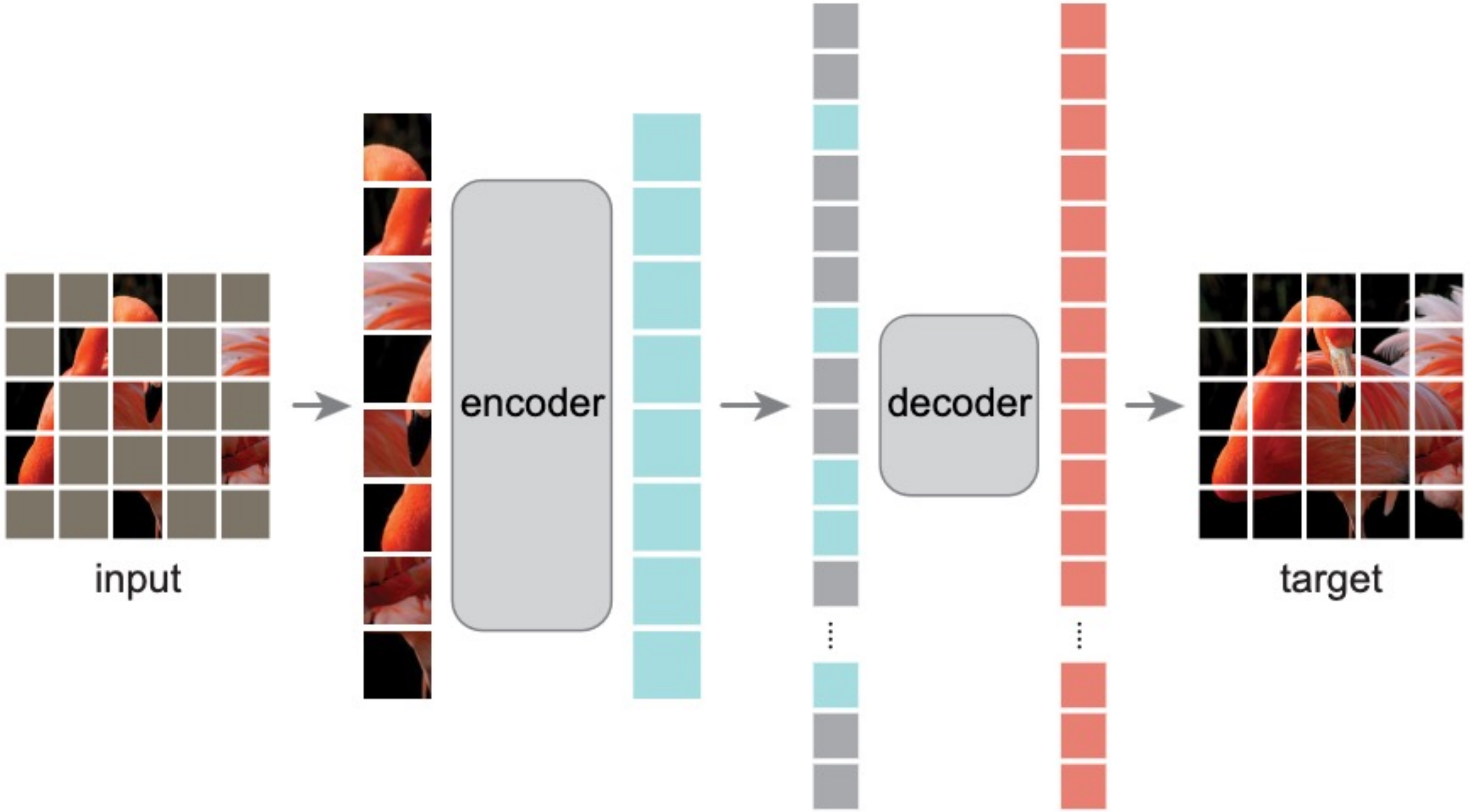


(c) MoCo

Improved Baselines with Momentum Contrastive Learning

https://arxiv.org/abs/2003.04297 https://arxiv.org/abs/1911.05722

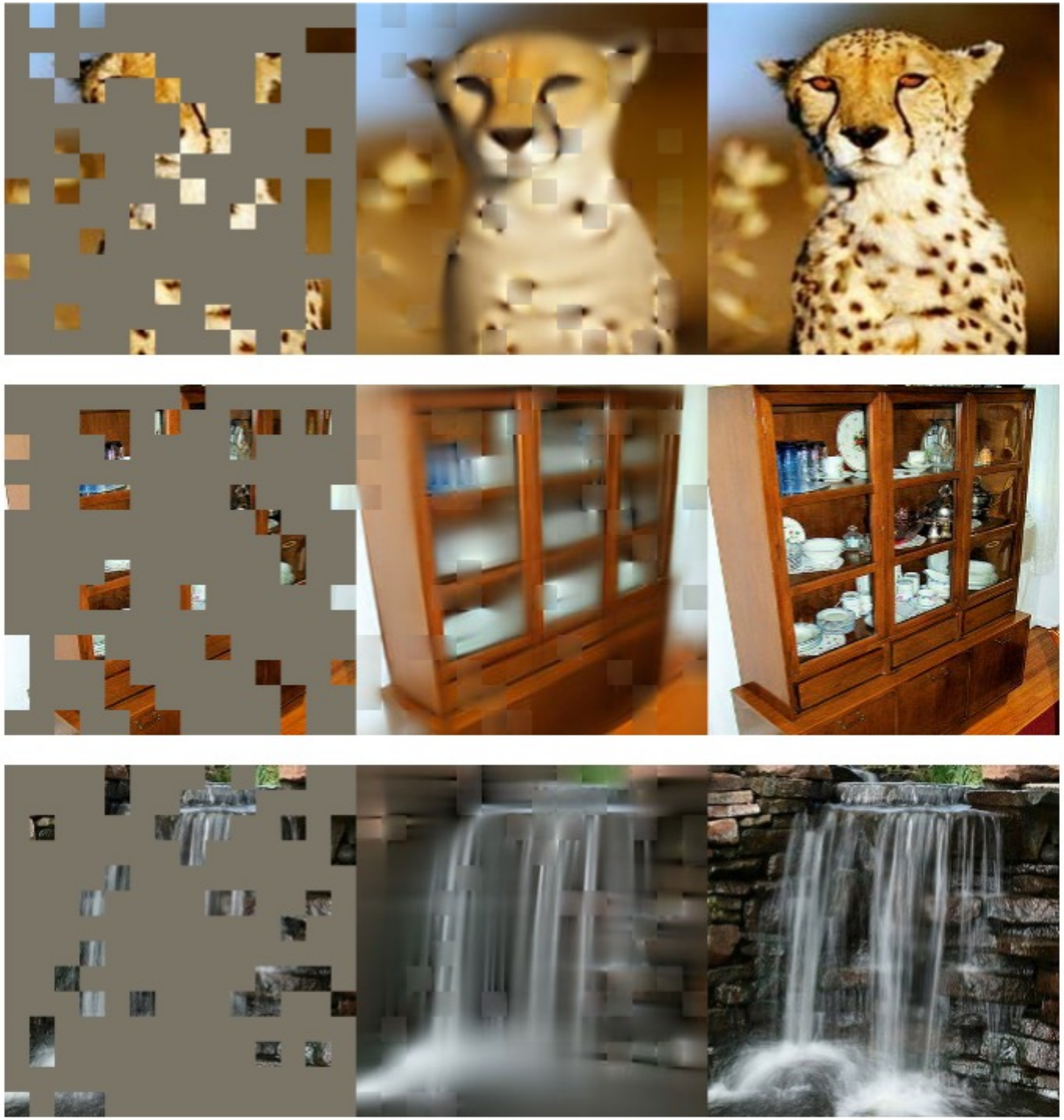
Alternative: Masked AutoEncoders (MAE)



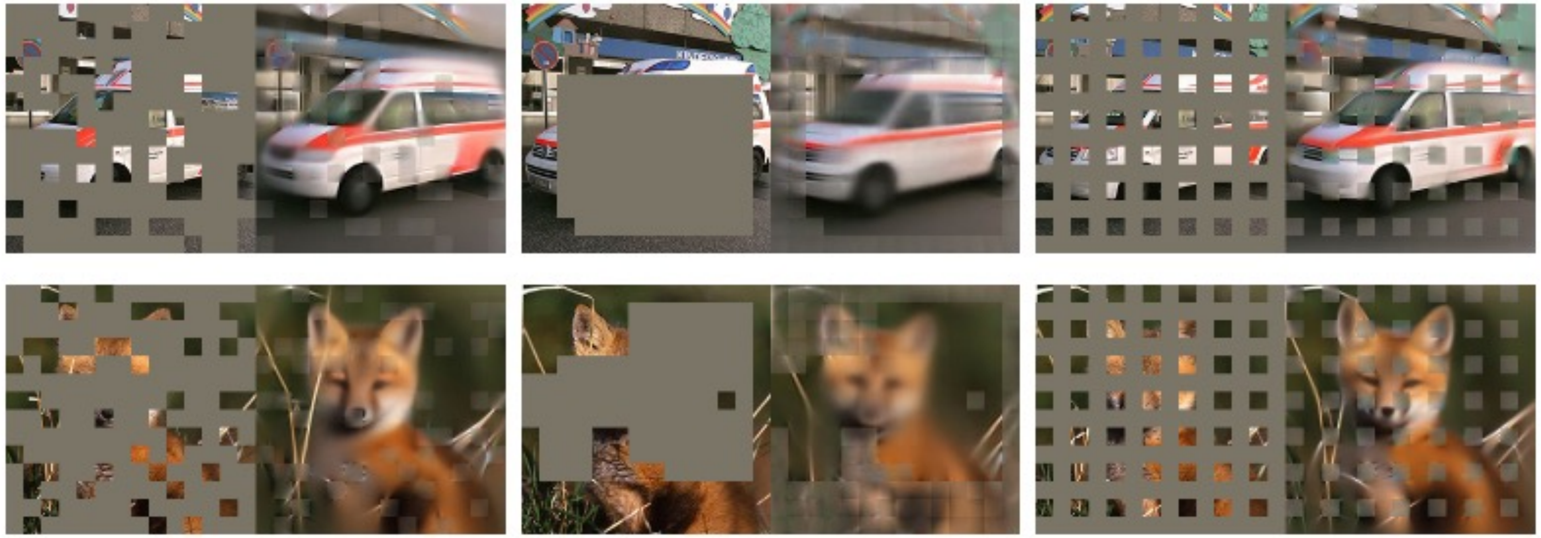
Masked Autoencoders Are Scalable Vision Learners
https://arxiv.org/abs/2111.06377

Examples

- After training, the model learns to fill-in-the-blanks for images.
- Similar to text masked image modeling
- The model can be finetuned for any other task.

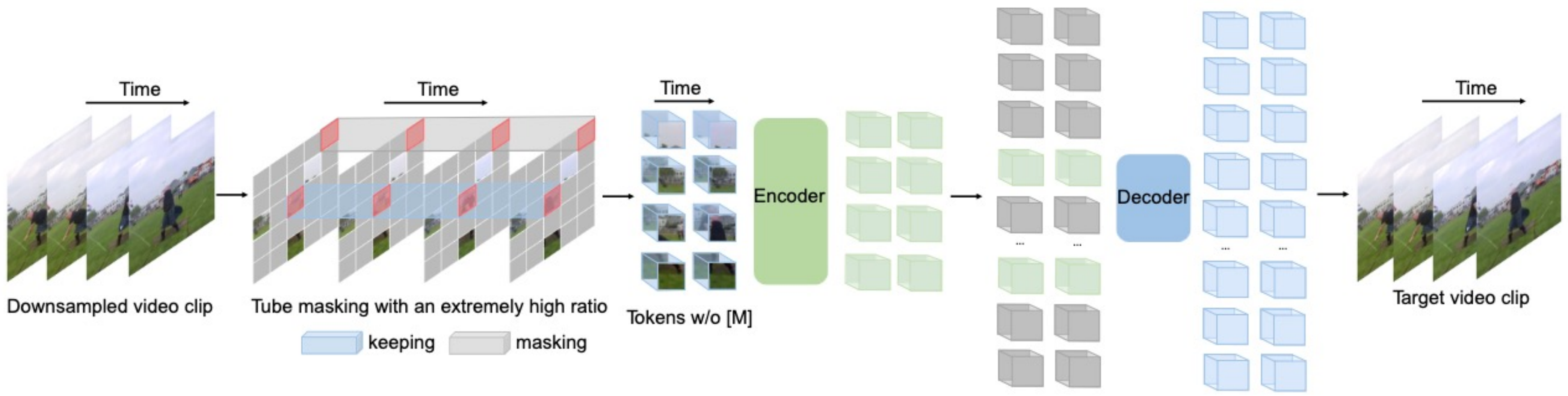


Several Analysis in the Paper



random 75% block 50% grid 75%

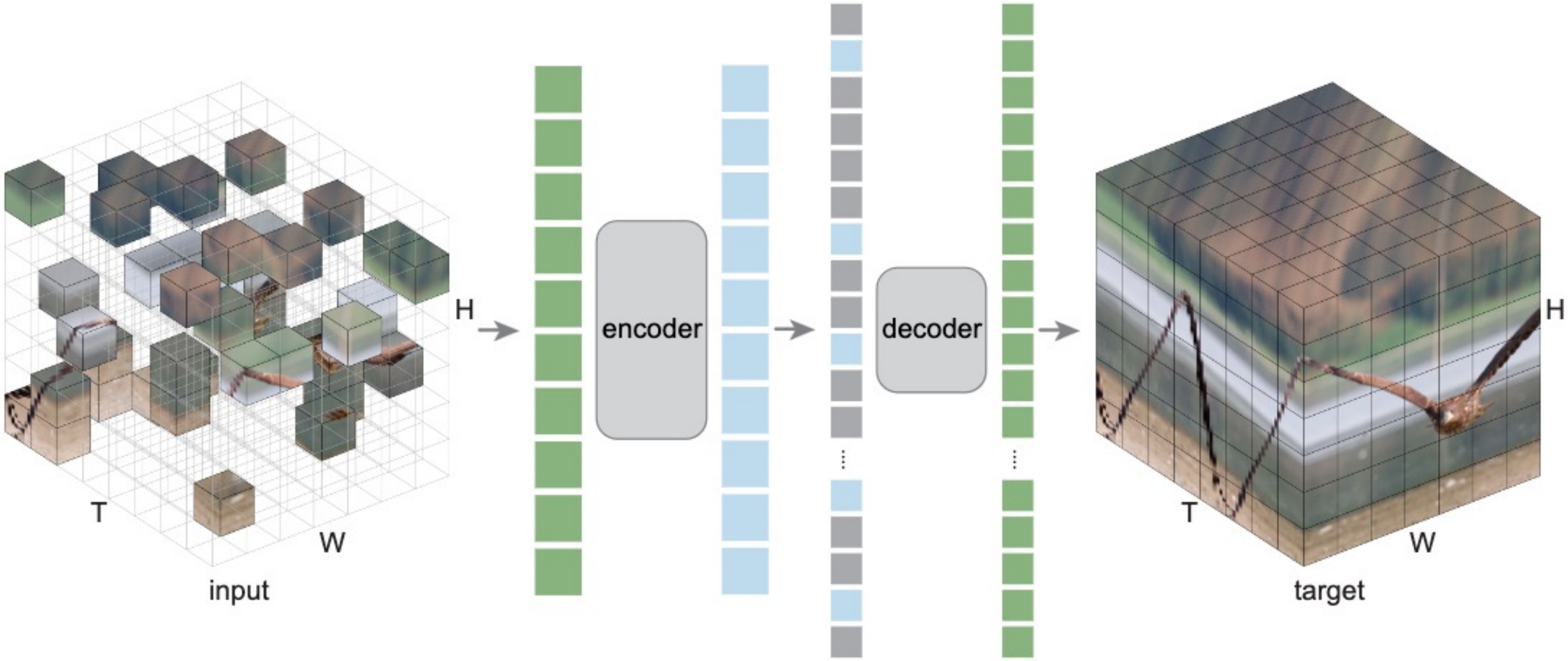
Video MAE



VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training

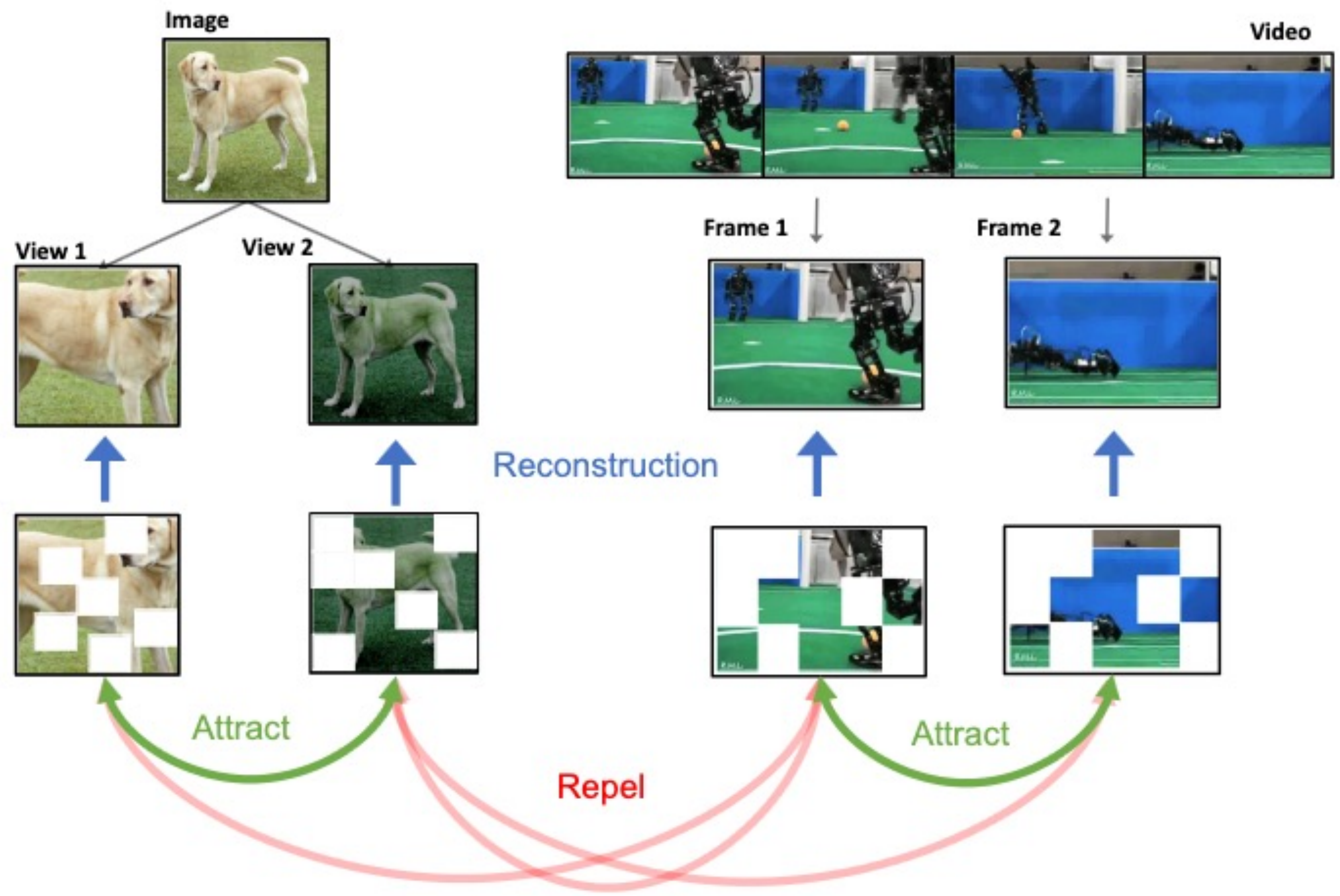
https://arxiv.org/abs/2203.12602

ST MAE



Masked Autoencoders As Spatiotemporal Learners
https://arxiv.org/abs/2205.09113

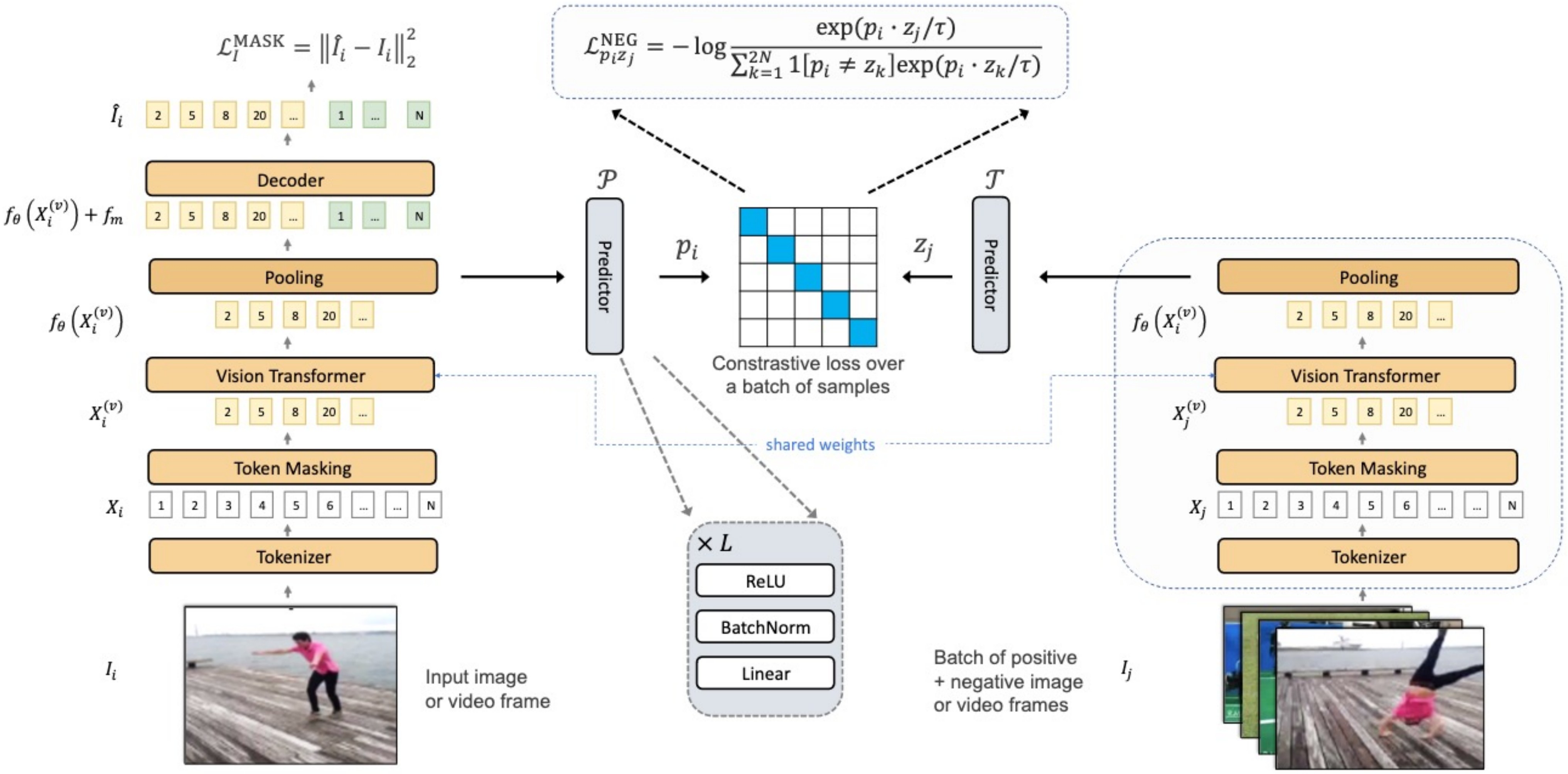
ViC-MAE



ViC-MAE: Self-Supervised Representation Learning from Images and Video with Contrastive Masked Autoencoders

https://arxiv.org/abs/2303.12001

ViC-MAE



$\mathcal{L}_I^{\text{MASK}} = \|\hat{I}_i - I_i\|_2^2$
$\mathcal{L}_{p_i z_j}^{\text{NEG}} = -\log \frac{\exp(p_i \cdot z_j/\tau)}{\sum_{k=1}^{2N} 1[p_i \neq z_k]\exp(p_i \cdot z_k/\tau)}$
$\hat{I}_i$
Decoder
$f_\theta\left(X_i^{(v)}\right) + f_m$
Pooling
$f_\theta\left(X_i^{(v)}\right)$
Vision Transformer
$X_i^{(v)}$
Token Masking
X_i
Tokenizer
I_i
Input image
or video frame
$\mathcal{P}$
Predictor
p_i
Constrastive loss over
a batch of samples
z_j
$\mathcal{T}$
Predictor
shared weights
$\times L$
ReLU
BatchNorm
Linear
Pooling
$f_\theta\left(X_i^{(v)}\right)$
Vision Transformer
$X_j^{(v)}$
Token Masking
X_j
Tokenizer
Batch of positive
+ negative image
or video frames
I_j

Results

	Method	Arch.	Pre-training Data	In-Domain		Out-of-Domain	
				IN1K	K400	Places-365	SSv2
Supervised	ViT [22] *ICML'20*	ViT-B	IN1K	82.3	68.5	57.0	61.8
	ViT [22] *ICML'20*	ViT-L	IN1K	82.6	78.6	58.9	66.2
	OMNIVORE [27] *CVPR'22*	ViT-B	IN1K + K400 + SUN RGB-D	84.0	83.3	**59.2**	68.3
	OMNIVORE [27] *CVPR'22*	ViT-L	IN1K + K400 + SUN RGB-D	**86.0**	84.1	–	–
	TubeViT [63] *CVPR'23*	ViT-B	K400 + IN1K	81.4	88.6	–	–
	TubeViT [63] *CVPR'23*	ViT-L	K400 + IN1K	–	**90.2**	–	**76.1**
Self-Supervised	MAE [35] *CVPR'22*	ViT-B	IN1K	83.4	–	57.9	59.6
	MAE [35] *CVPR'22*	ViT-L	IN1K	85.5	82.3	59.4	57.7
	ST-MAE [26] *NeurIPS'22*	ViT-B	K400	81.3	81.3	57.4	69.3
	ST-MAE [26] *NeurIPS'22*	ViT-L	K400	81.7	84.8	58.1	73.2
	VideoMAE [68] *NeurIPS'22*	ViT-B	K400	81.1	80.0	–	69.6
	VideoMAE [68] *NeurIPS'22*	ViT-L	K400	–	85.2	–	74.3
	OmniMAE [29] *CVPR'23*	ViT-B	K400 + IN1K	82.8	80.8	58.5	69.0
	OmniMAE [29] *CVPR'23*	ViT-L	K400 + IN1K	84.7	84.0	59.4	73.4
	ViC-MAE	ViT-L	K400	85.0	85.1	59.5	73.7
	ViC-MAE	ViT-L	MiT	85.3	84.9	59.7	73.8
	ViC-MAE	ViT-B	K400 + IN1K	83.0	80.8	58.6	69.5
	ViC-MAE	ViT-L	K400 + IN1K	86.0	86.8	60.0	75.0
	ViC-MAE	ViT-B	K400 + K600 + K700 + MiT + IN1K	83.8	80.9	59.1	69.8
	ViC-MAE	ViT-L	K400 + K600 + K700 + MiT + IN1K	**87.1**	**87.8**	**60.7**	**75.9**

Questions