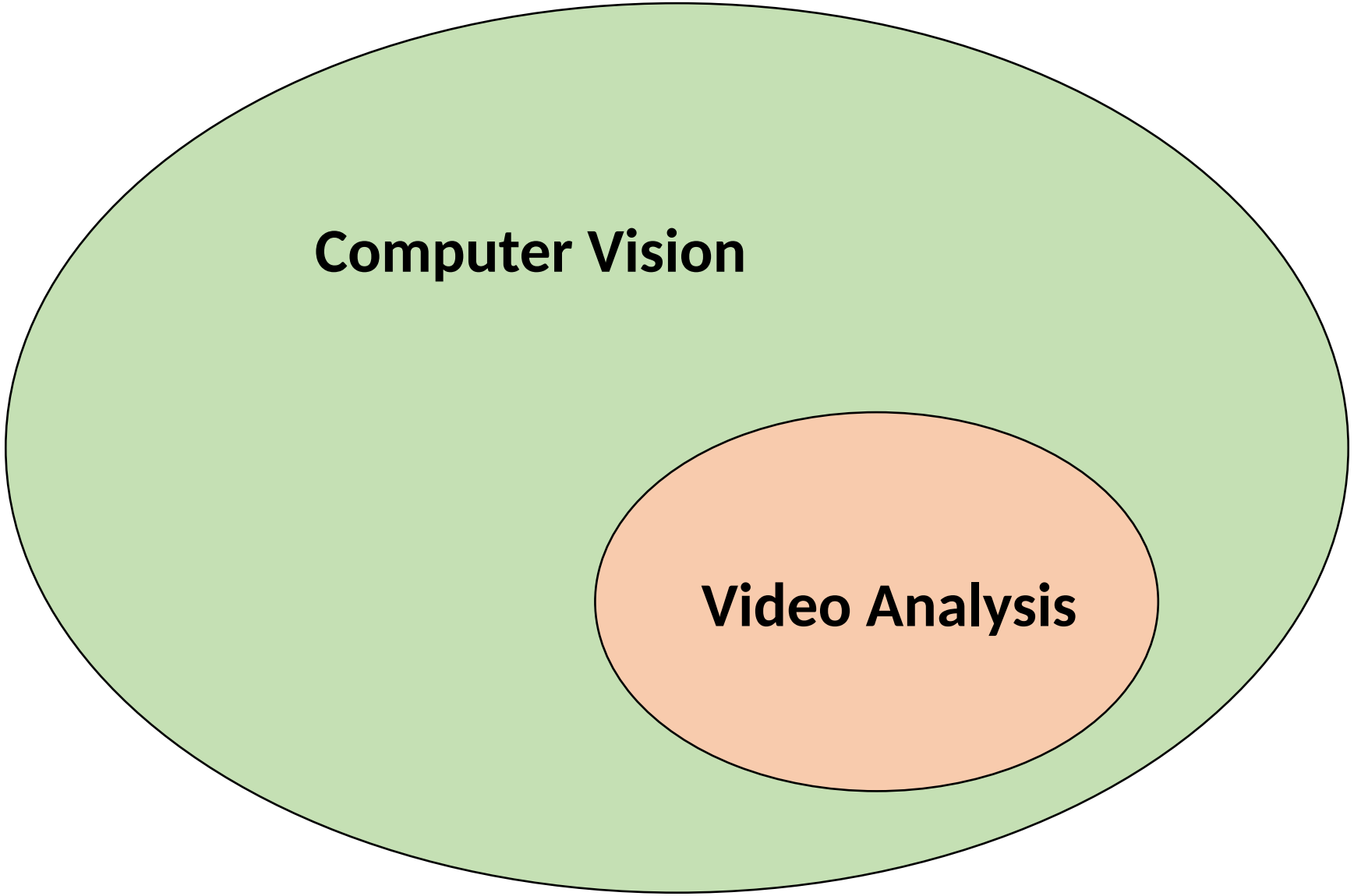
ETH zürich

Siamese Masked Autoencoders

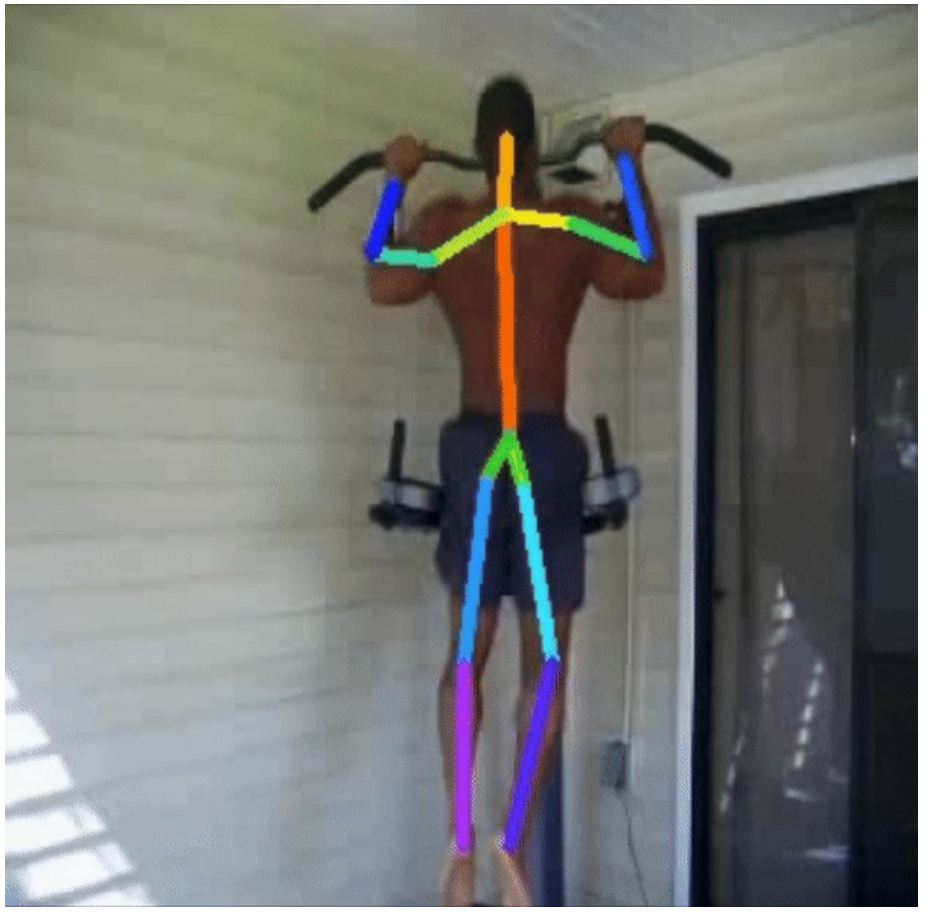
Agrim Gupta, Jiajun Wu, Jia Deng, Li Fei-Fei

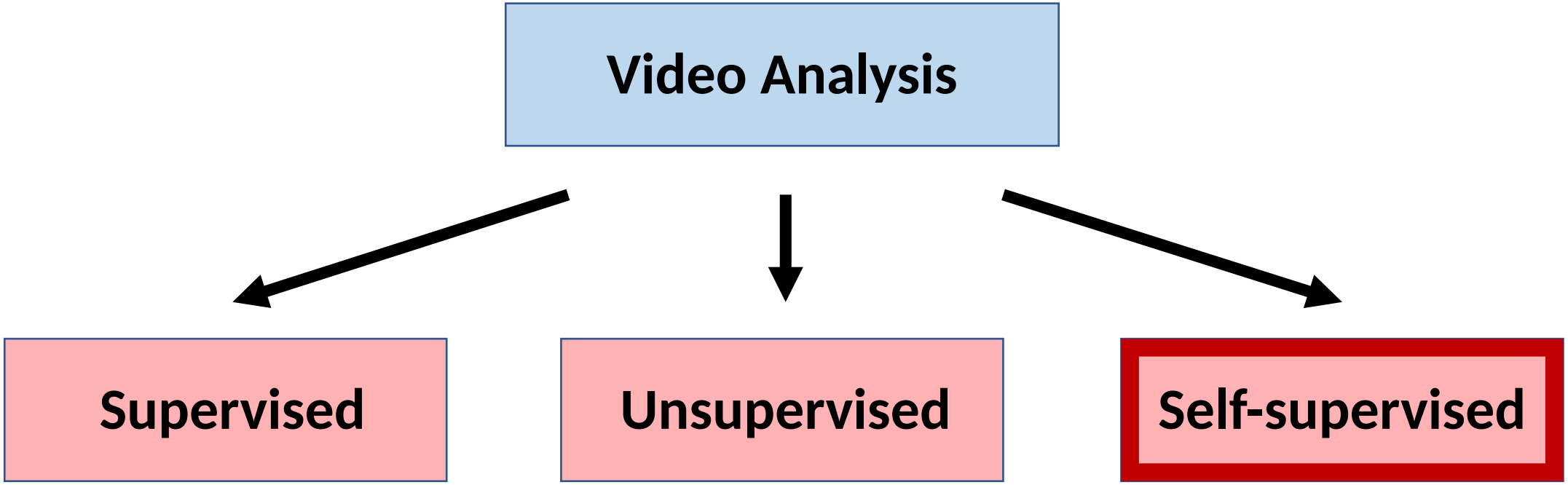
Pyrros Koussios
Seminar in Deep Neural Networks, ETH
26. March 2024


Computer Vision
Video Analysis

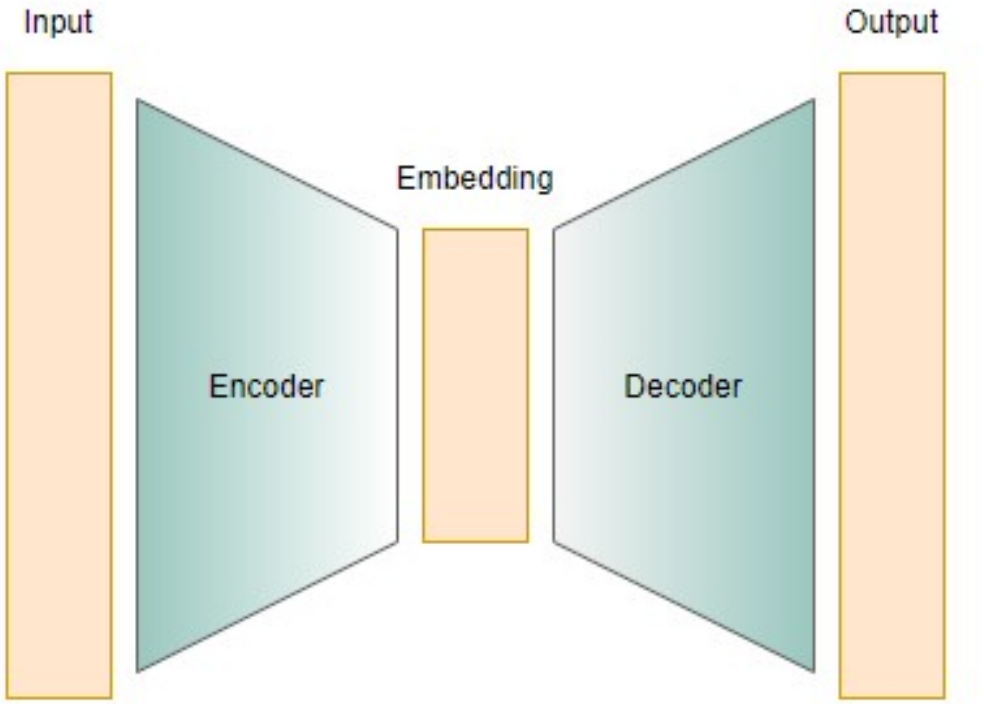
Semantic Segmentation

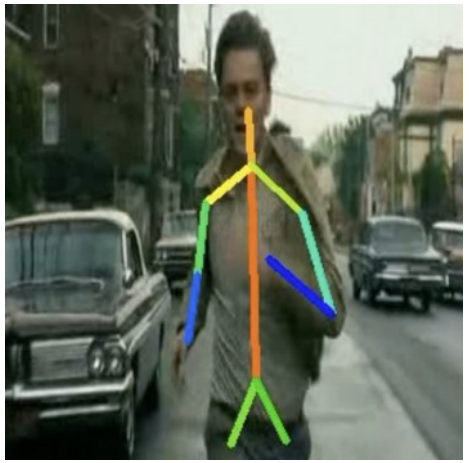
Pose Propagation

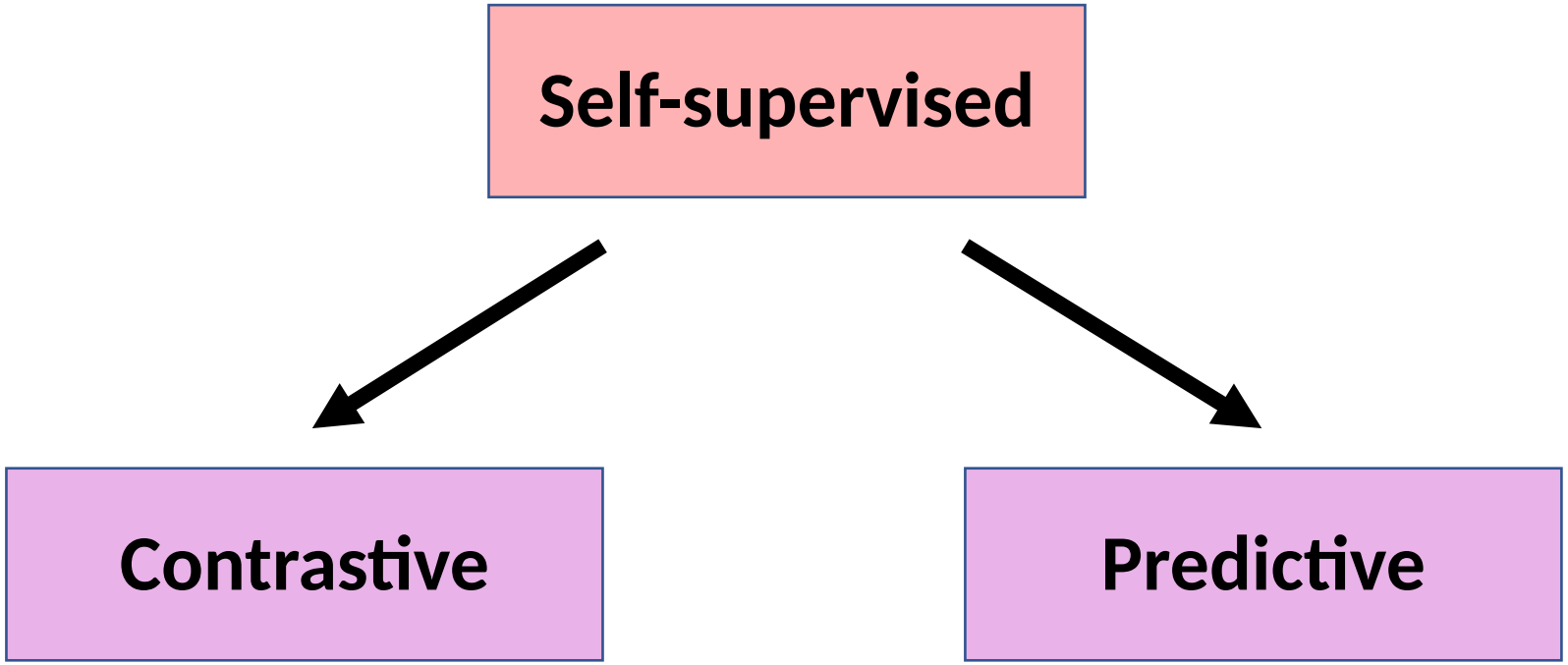



Video Analysis
Supervised
Unsupervised
Self-supervised

Encoder/Autoencoder






Self-supervised
Contrastive
Predictive

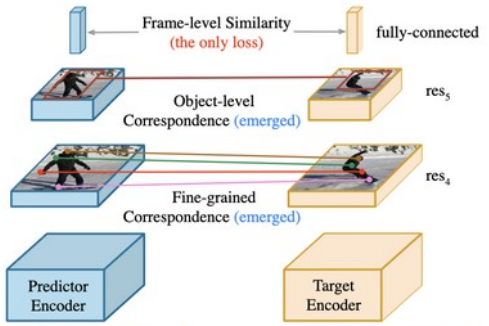
Contrastive

Predictive

Contrastive

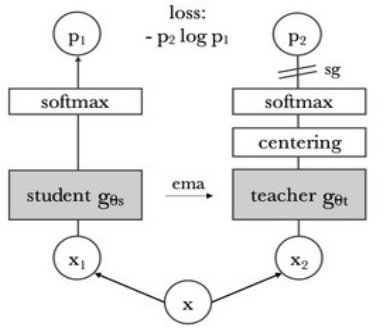
TimeCycle

VFS

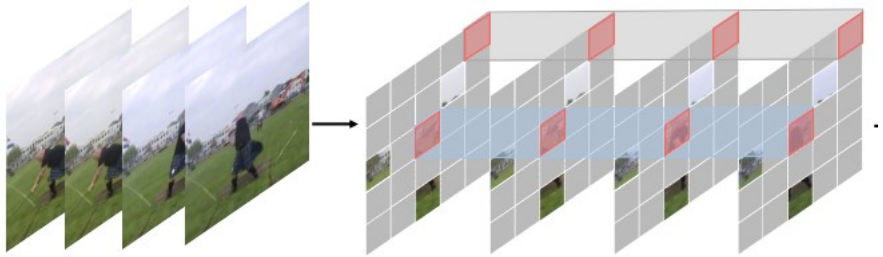


Predictive

DINO



VideoMAE



SiamMAE


MAE
VideoMAE
SiamMAE

Self-supervised training using masking

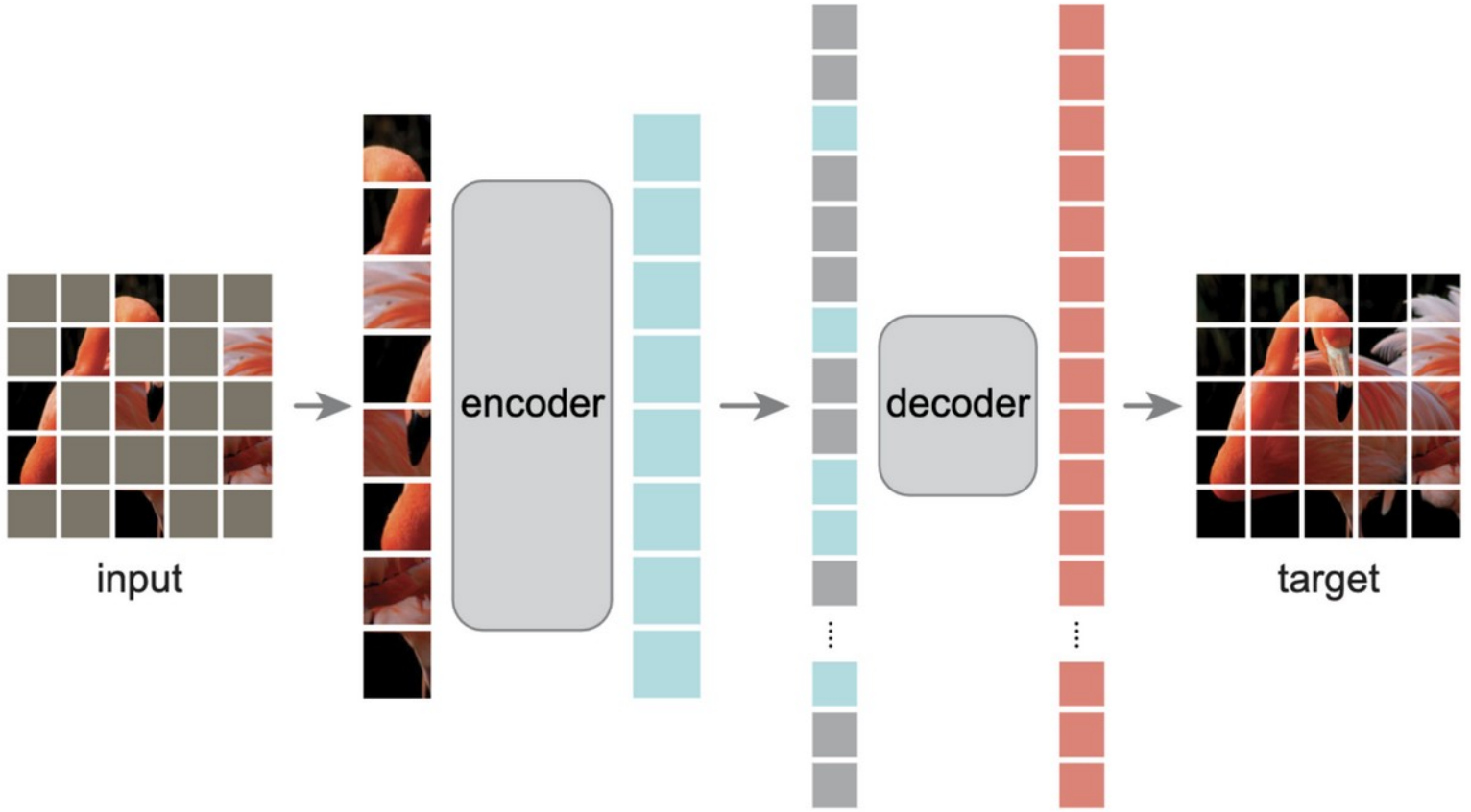
THE BLACK CAT CROSSED THE ROAD

THE BLACK ___ CROSSED THE ROAD

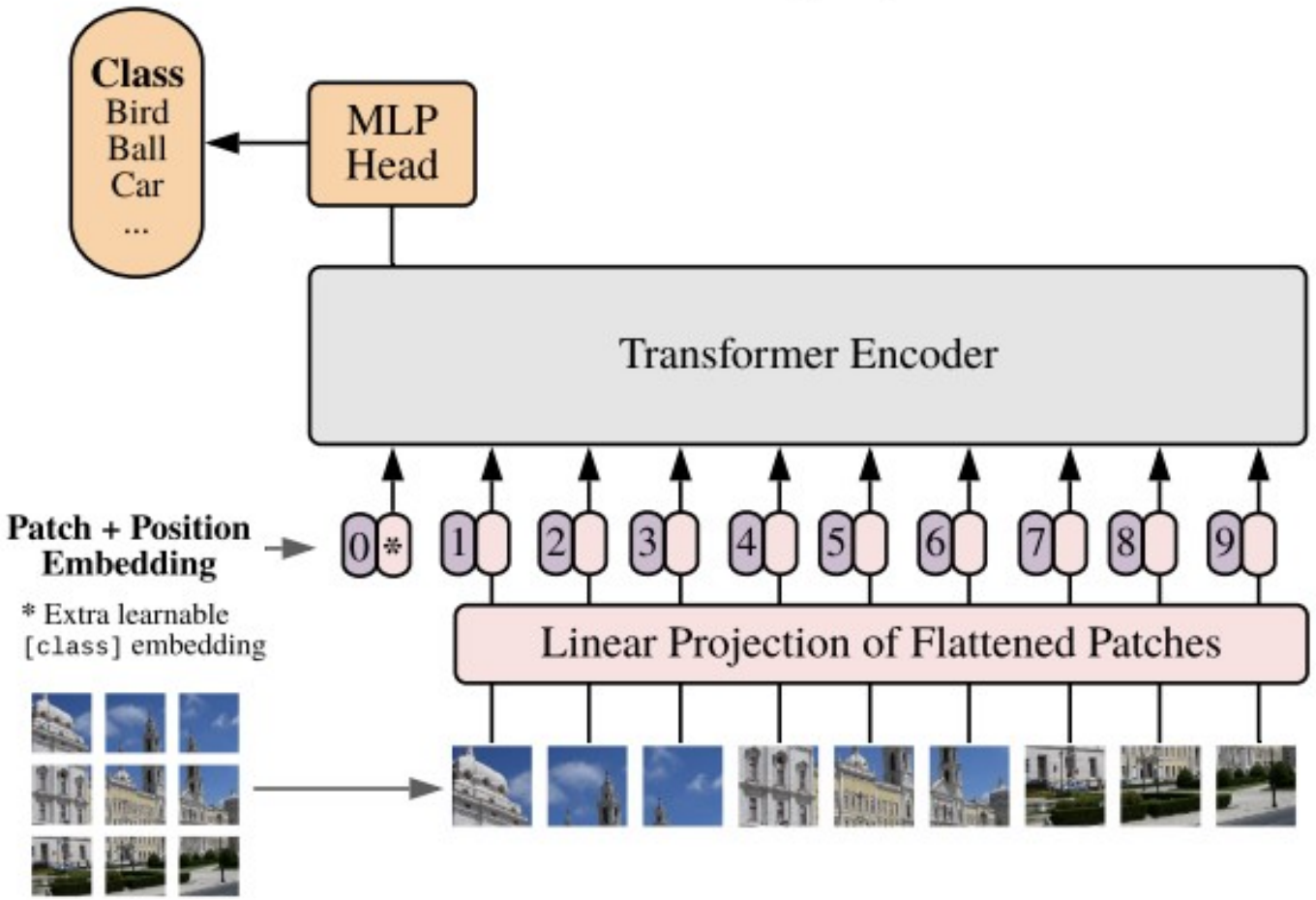
THE _____ ___ CROSSED THE ROAD

Masked AutoEncoders

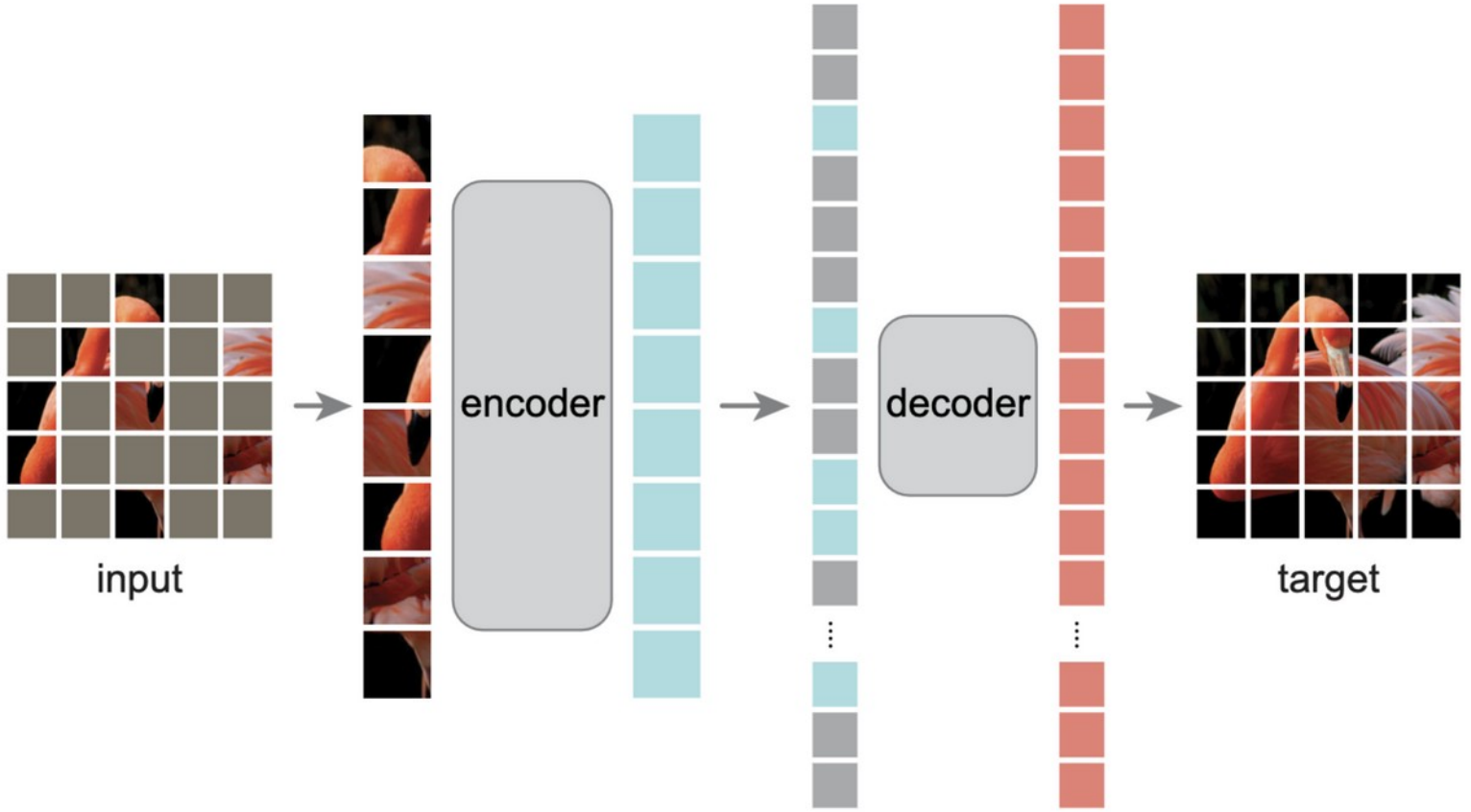
Masked AutoEncoders



Vision Transformer Encoder



Masked AutoEncoders



BERT

Information Reduncancy

MAE → **VideoMAE** → **SiamMAE**

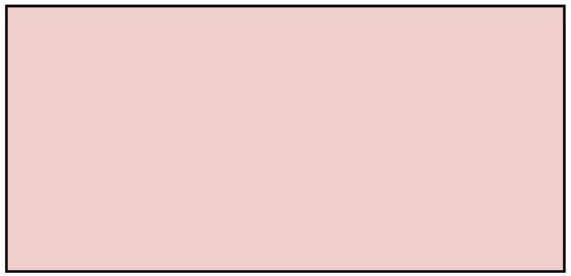
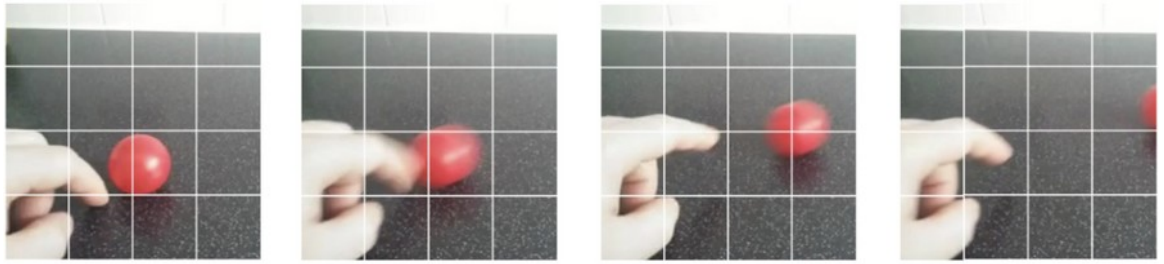


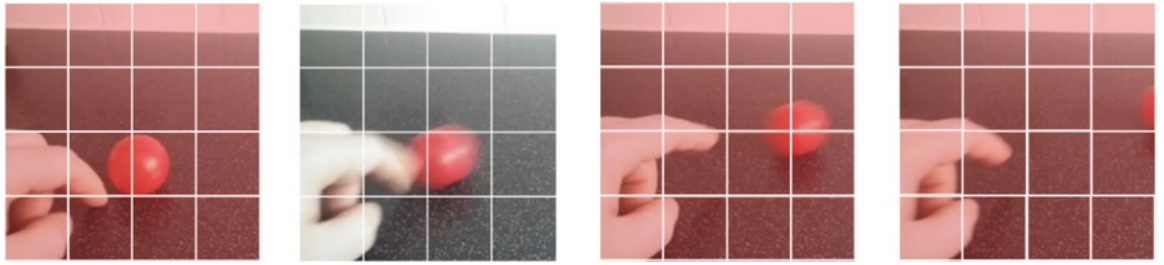
Image
Masking + Reconstruction

Temporal Masking Approaches

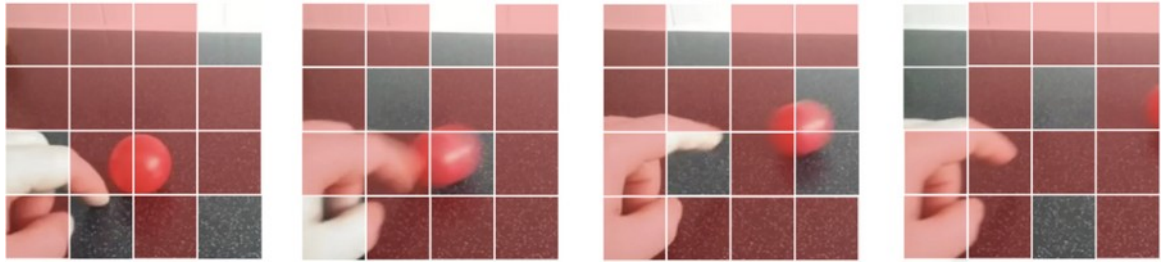
Input Video



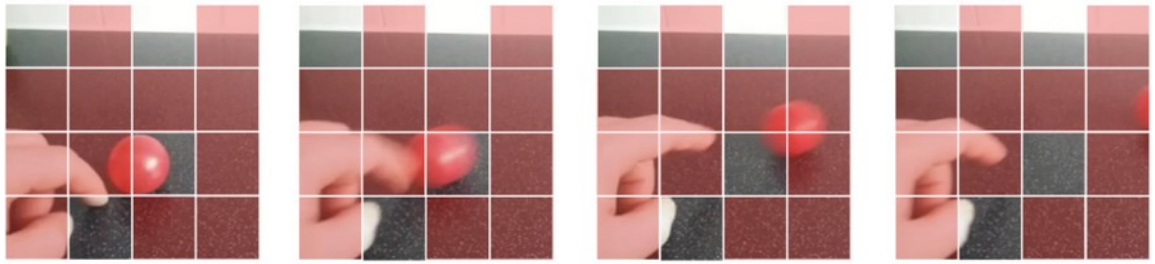
Frame Masking



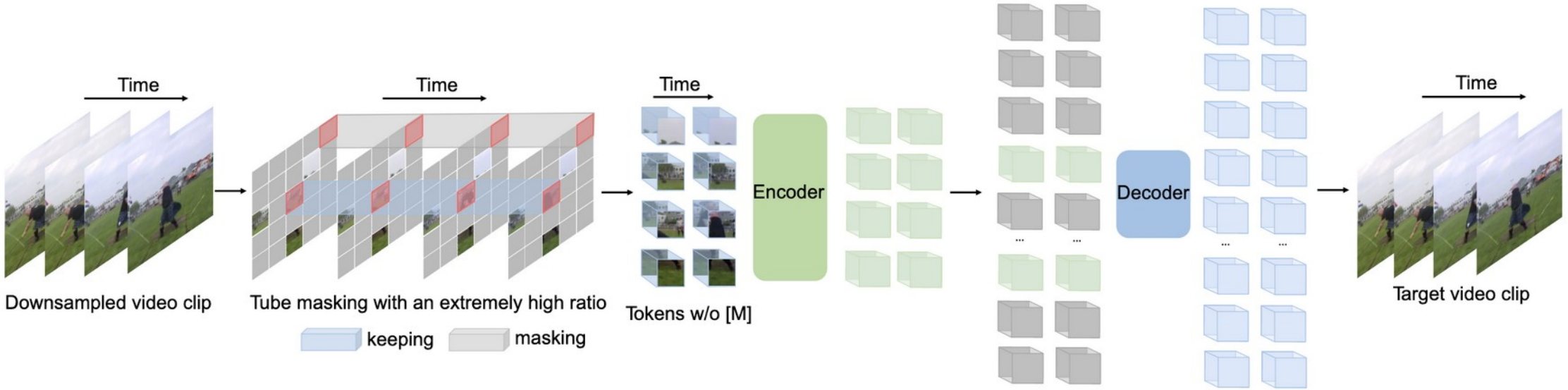
Random Masking

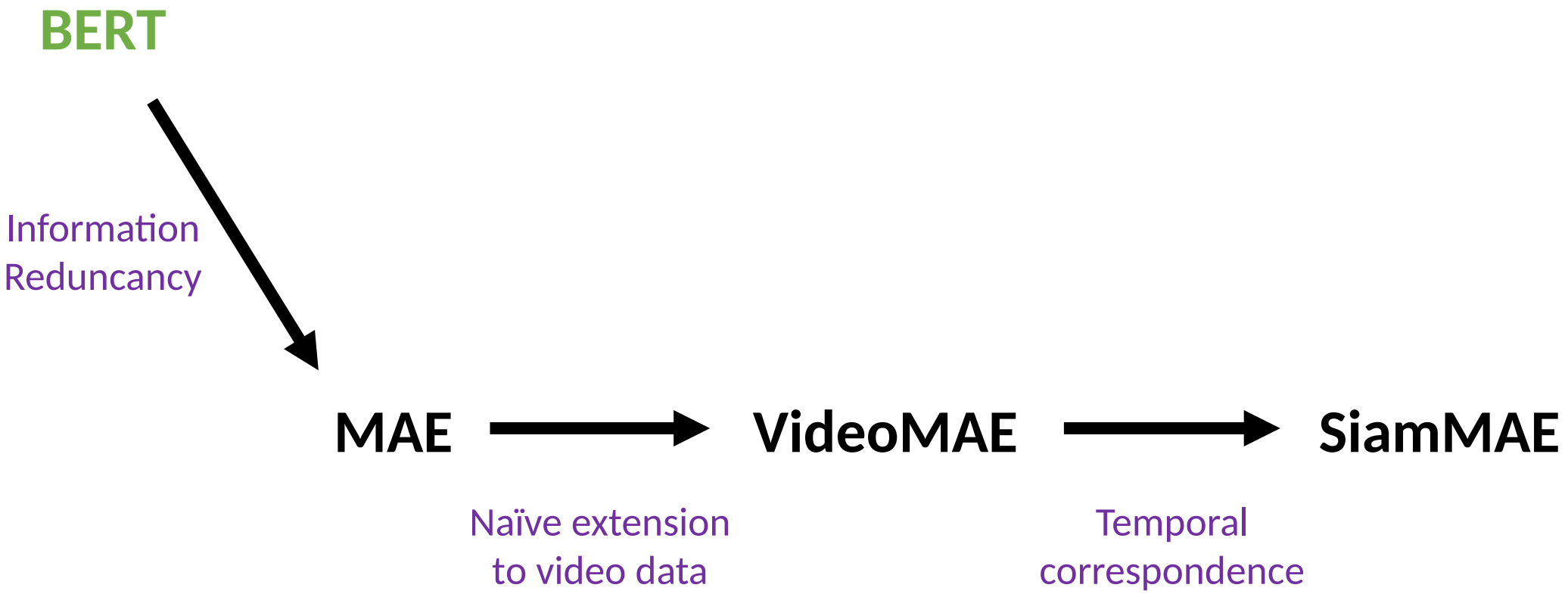


Tube Masking

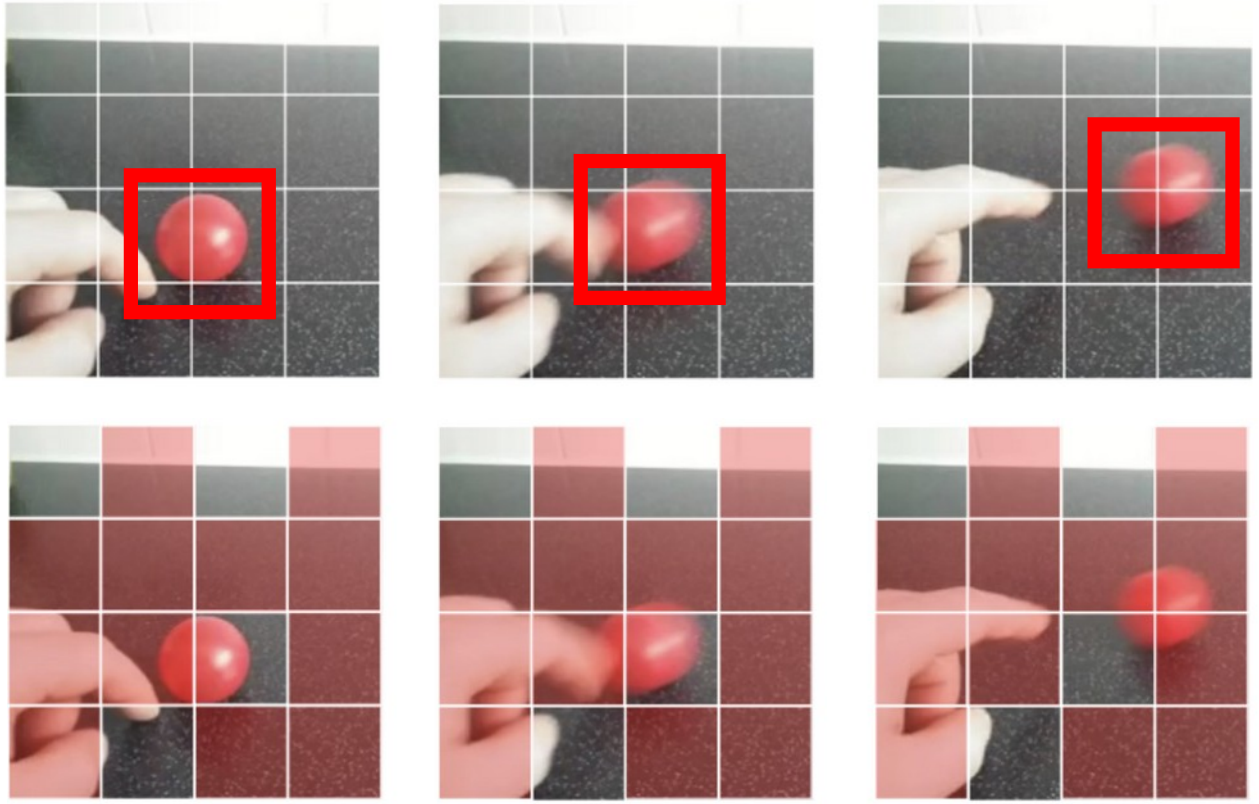


Video Masked AutoEncoders

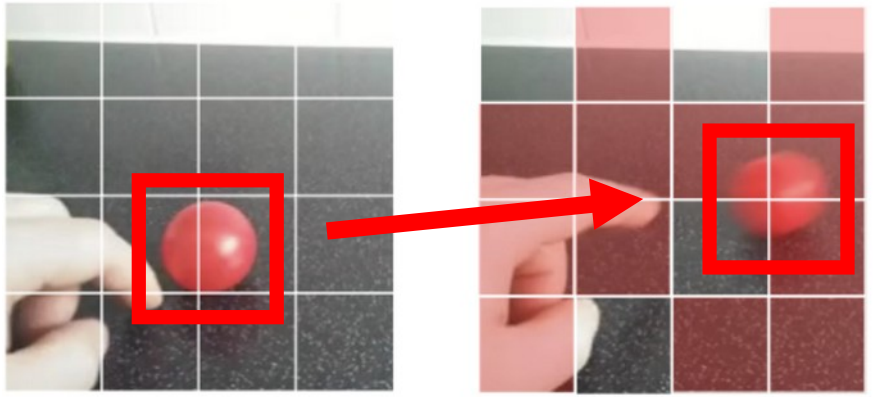



BERT
Information
Reduncancy
MAE
VideoMAE
SiamMAE
Naïve extension
to video data
Temporal
correspondence

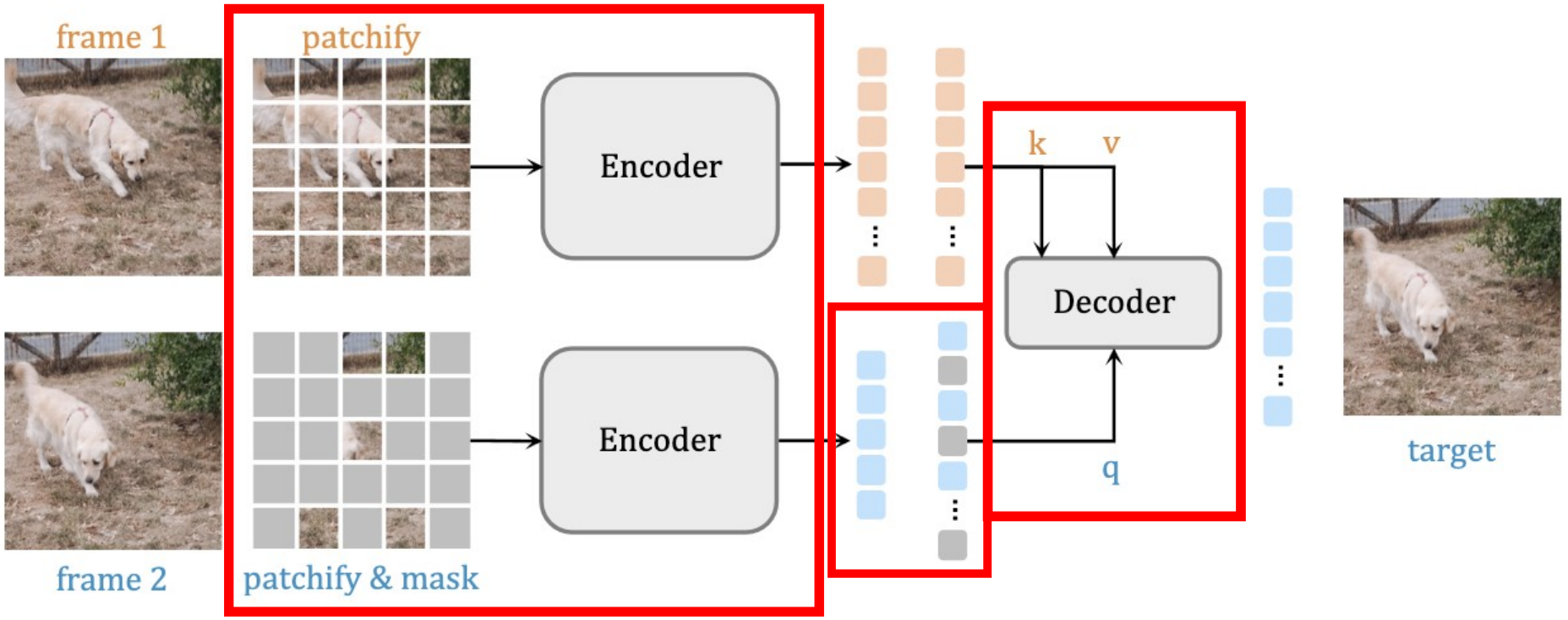
VideoMAE



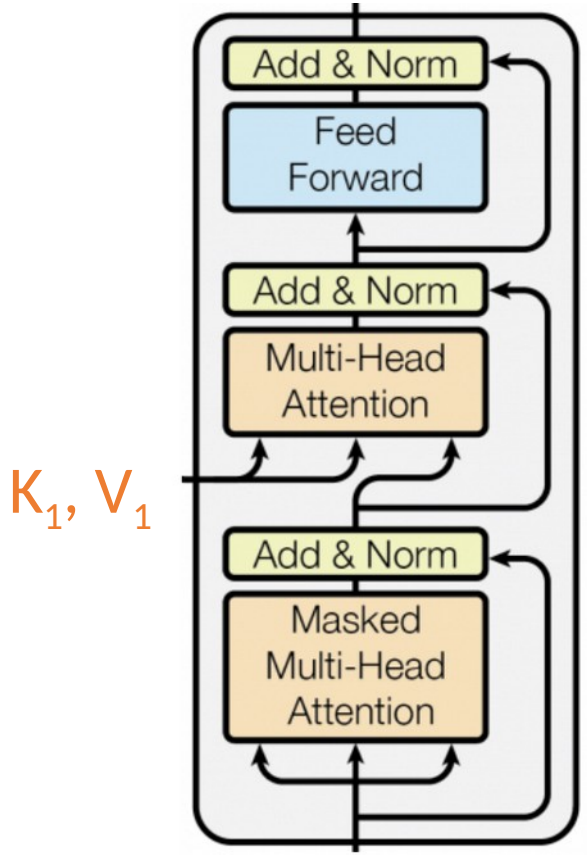
SiamMAE



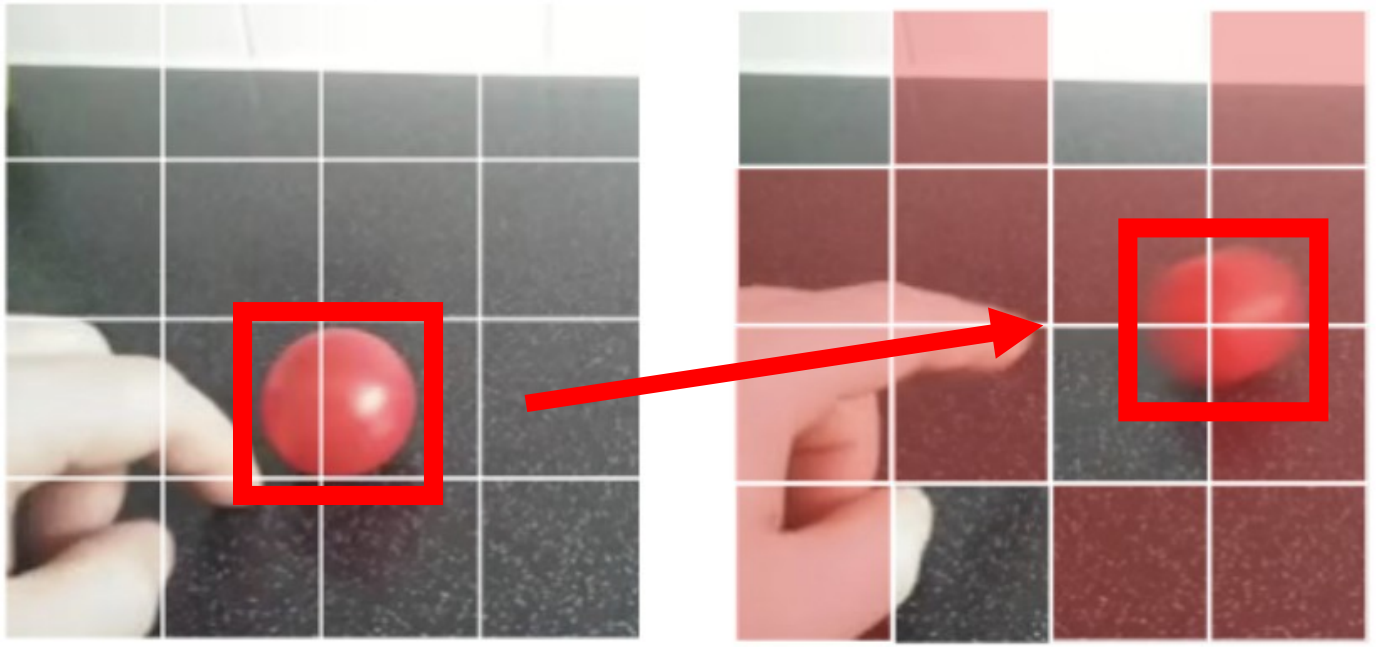
Siamese Masked AutoEncoders



Transformer Decoder



SiamMAE Main Idea:
Learning to propagate patches!



Frame 1

Frame 2

Quantitative Results

Object Propagation

DAVIS

Semantic Part Propagation

VIP

Human Pose Propagation

JHMDB

Inference

Frame 1

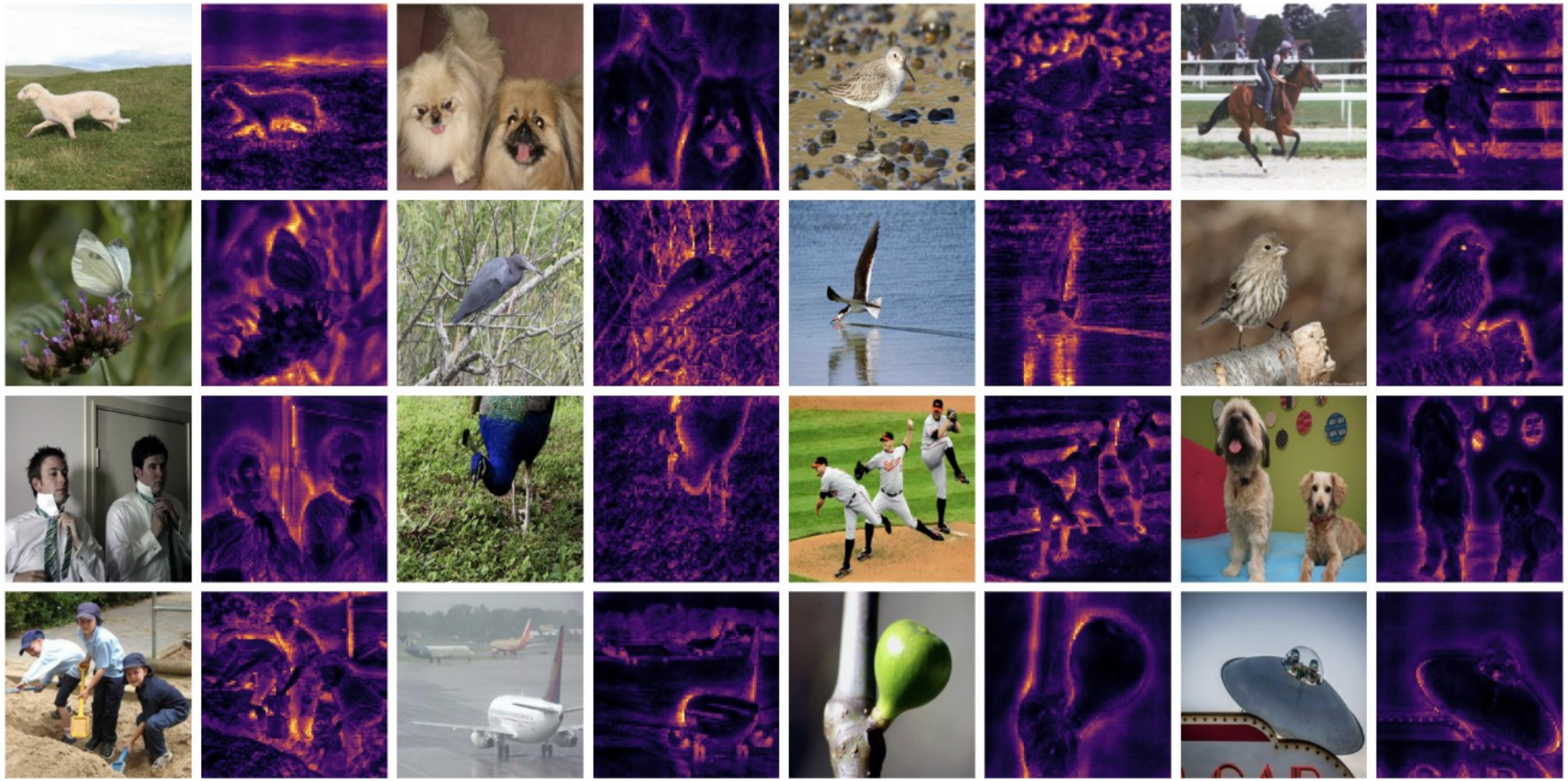
Frame 2

Method	Backbone	Dataset	DAVIS $\mathcal{J}\&\mathcal{F}_m$	DAVIS $\mathcal{J}_m$	DAVIS $\mathcal{F}_m$	VIP mIoU	JHMDB PCK@0.1	JHMDB PCK@0.2
Supervised [98]	ResNet-50	ImageNet	66.0	63.7	68.4	39.5	59.2	78.3
SimSiam [20]	ResNet-50	ImageNet	66.3	64.5	68.2	35.0	58.4	77.5
MoCo [19]	ResNet-50	ImageNet	65.4	63.2	67.6	36.1	60.4	79.3
TimeCycle [14]	ResNet-50	VLOG	40.7	41.9	39.4	28.9	57.7	78.5
UVC [12]	ResNet-50	Kinetics	56.3	54.5	58.1	34.2	56.0	76.6
VFS [16]	ResNet-50	Kinetics	68.9	66.5	71.3	43.2	60.9	80.7
MAE-ST [27]	ViT-L/16	Kinetics	54.6	55.5	53.6	33.2	44.4	72.5
MAE [24]	VIT-B/16	ImageNet	53.5	52.1	55.0	28.1	44.6	73.4
VideoMAE [28]	ViT-S/16	Kinetics	39.3	39.7	38.9	23.3	41.0	67.9
Dino [17]	ViT-S/16	ImageNet	61.8	60.2	63.4	36.2	45.6	75.0
SiamMAE (ours)	ViT-S/16	Kinetics	**62.0**	**60.3**	**63.7**	**37.3**	**47.0**	**76.1**
Dino [17]	ViT-S/8	ImageNet	69.9	66.6	73.1	39.5	56.5	80.3
SiamMAE (ours)	ViT-S/8	Kinetics	**71.4**	**68.4**	**74.5**	**45.9**	**61.9**	**83.8**

SiamMAE quantitative results on three tasks: object segmentation (DAVIS), semantic part propagation (VIP), human pose propagation (JHMDB)

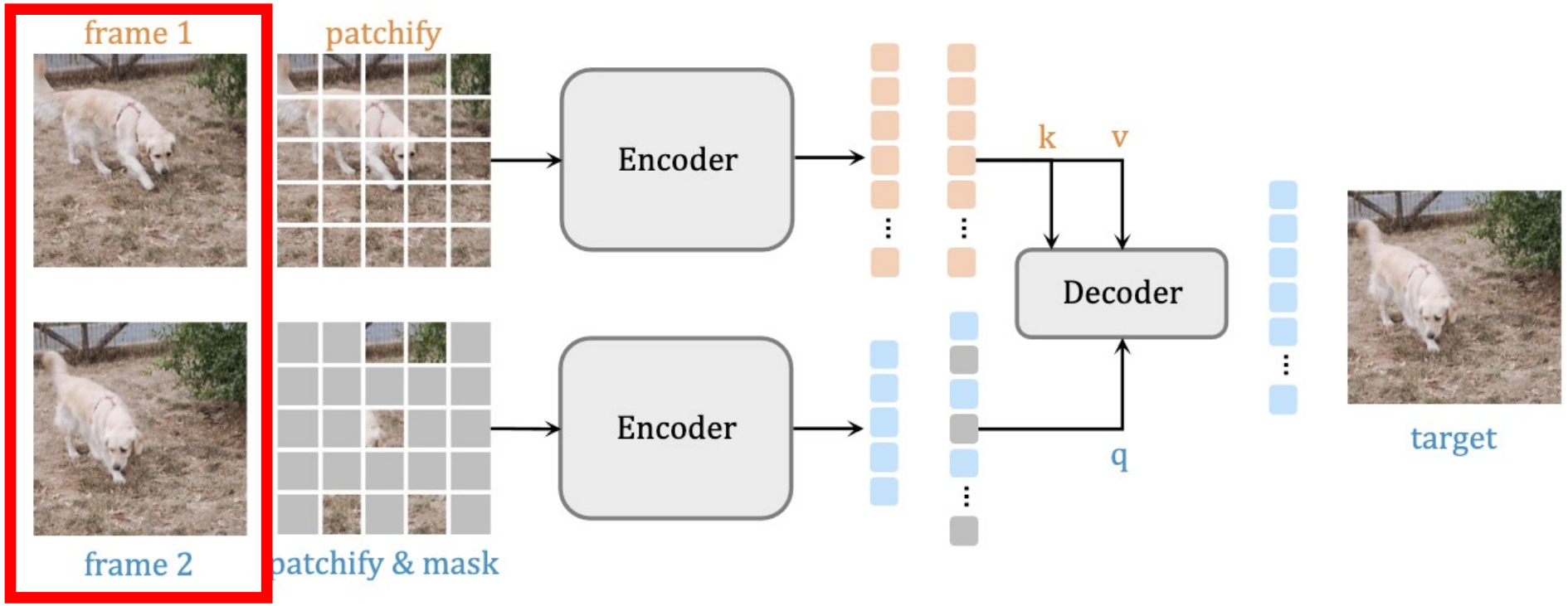
Qualitative Results

SiamMAE decoder self-attention maps



Ablation Studies

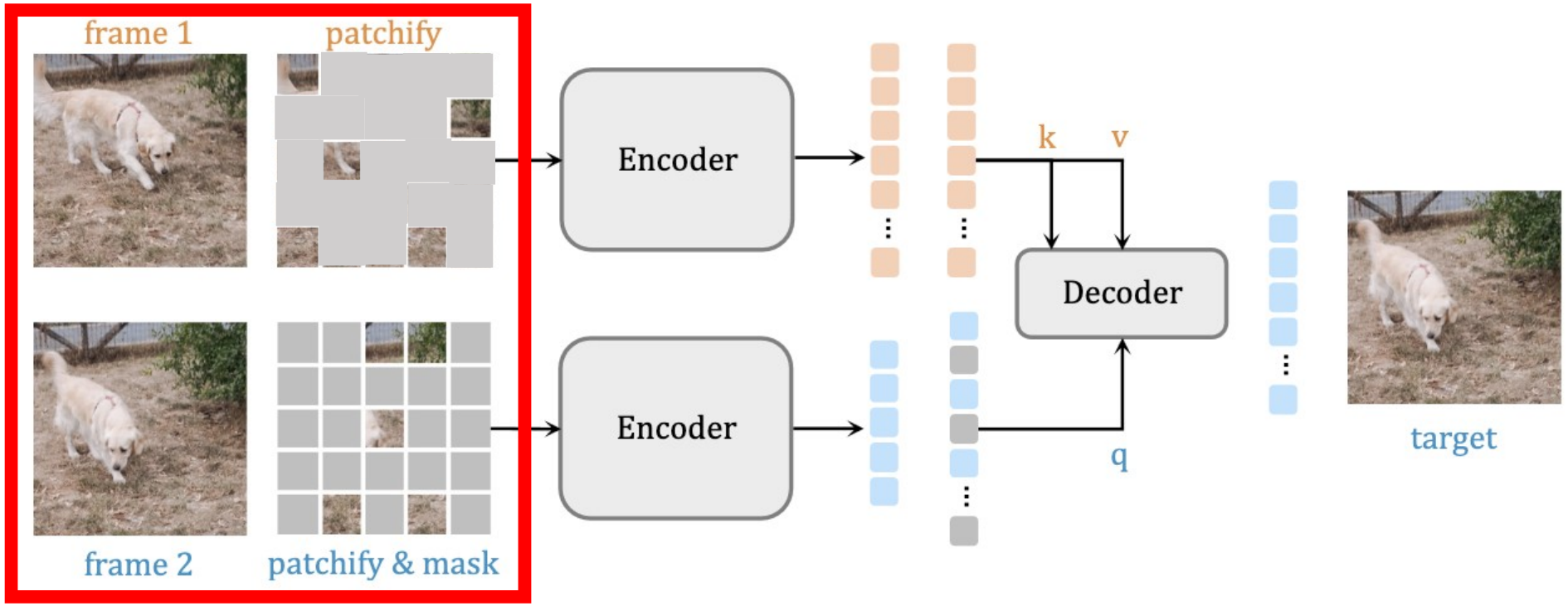
Siamese Masked AutoEncoders



Frame Gap

frame gap	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
4	55.1	53.5	56.7
8	56.4	54.9	57.8
16	58.0	56.7	59.4
32	57.7	56.3	59.1
4-48	**58.1**	**56.6**	**59.6**

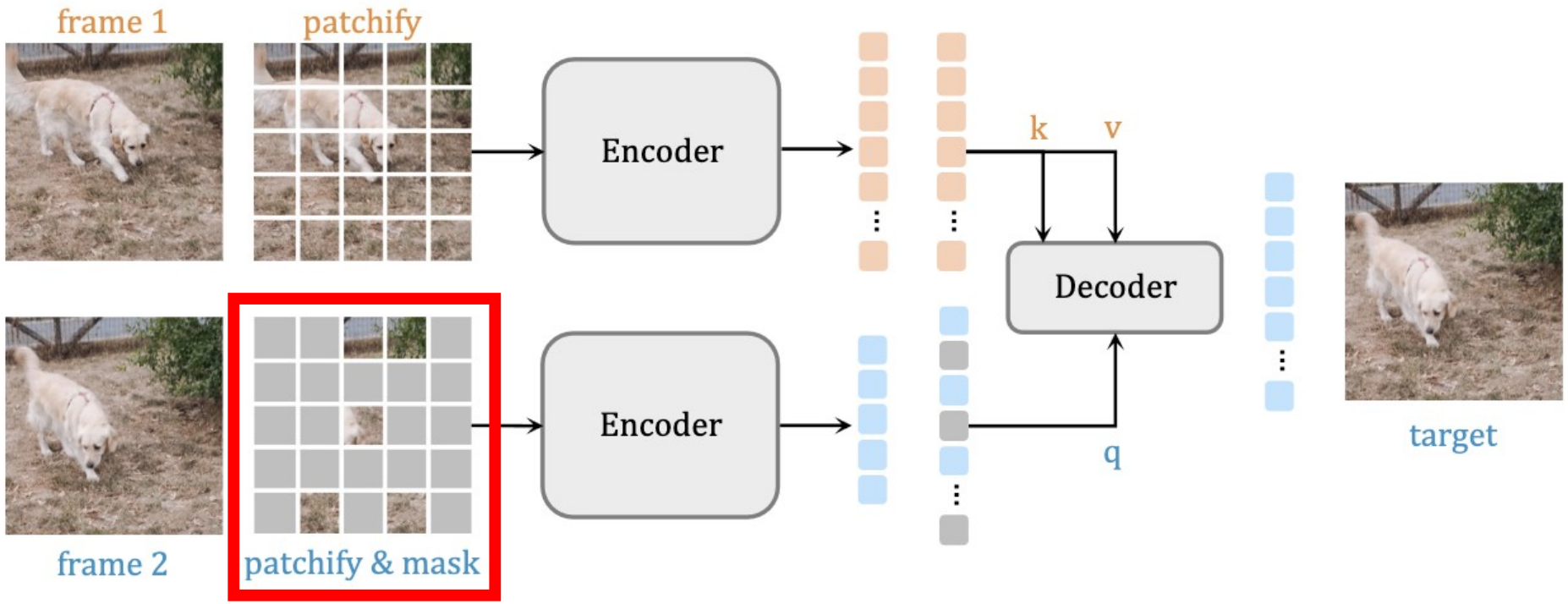
Siamese Masked AutoEncoders



Symmetric Masking

mask ratio	pattern	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
0.50 (s)	random	41.5	40.2	42.7
0.50 (s)	grid	48.2	46.7	49.7
0.75 (s)	random	52.7	51.3	54.1
0.90 (s)	random	51.4	50.0	52.8
0.95 (a)	random	**58.1**	**56.6**	**59.6**

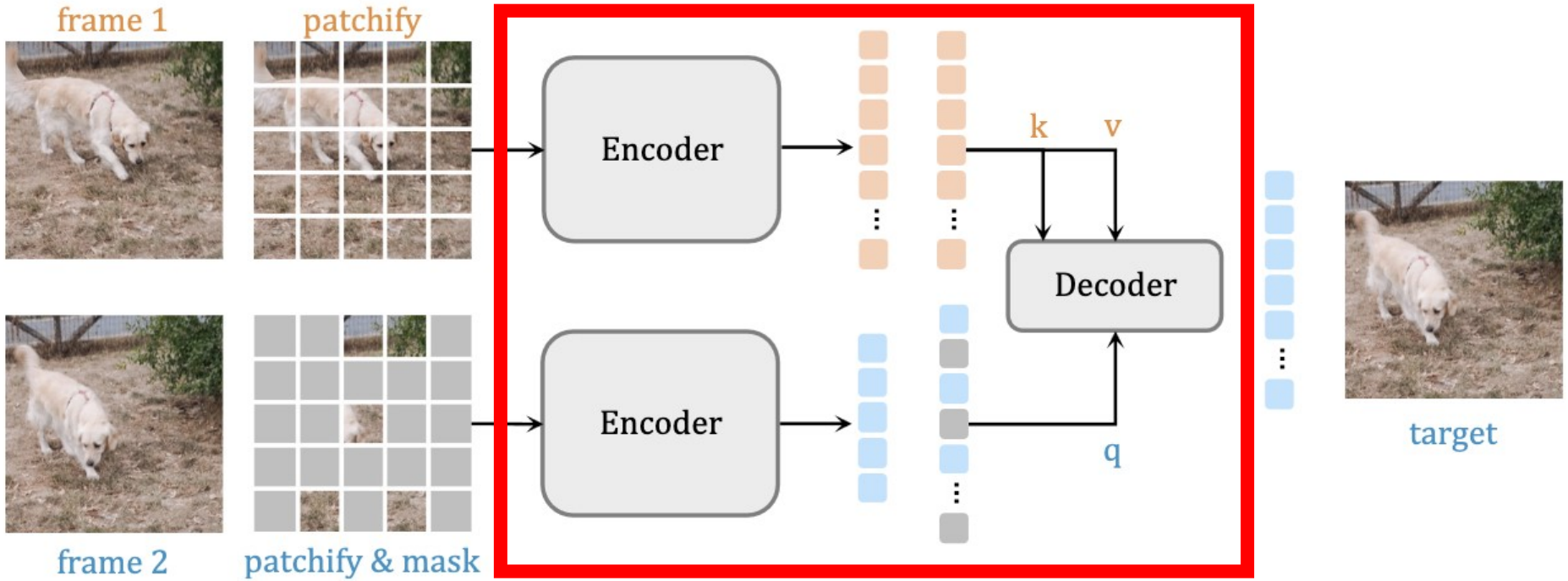
Siamese Masked AutoEncoders



Asymmetric Masking

mask ratio	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
0.50 (a)	49.0	48.4	49.6
0.75 (a)	55.3	54.1	56.4
0.90 (a)	58.4	57.0	59.8
0.95 (a)	**58.1**	**56.6**	**59.6**

Siamese Masked AutoEncoders



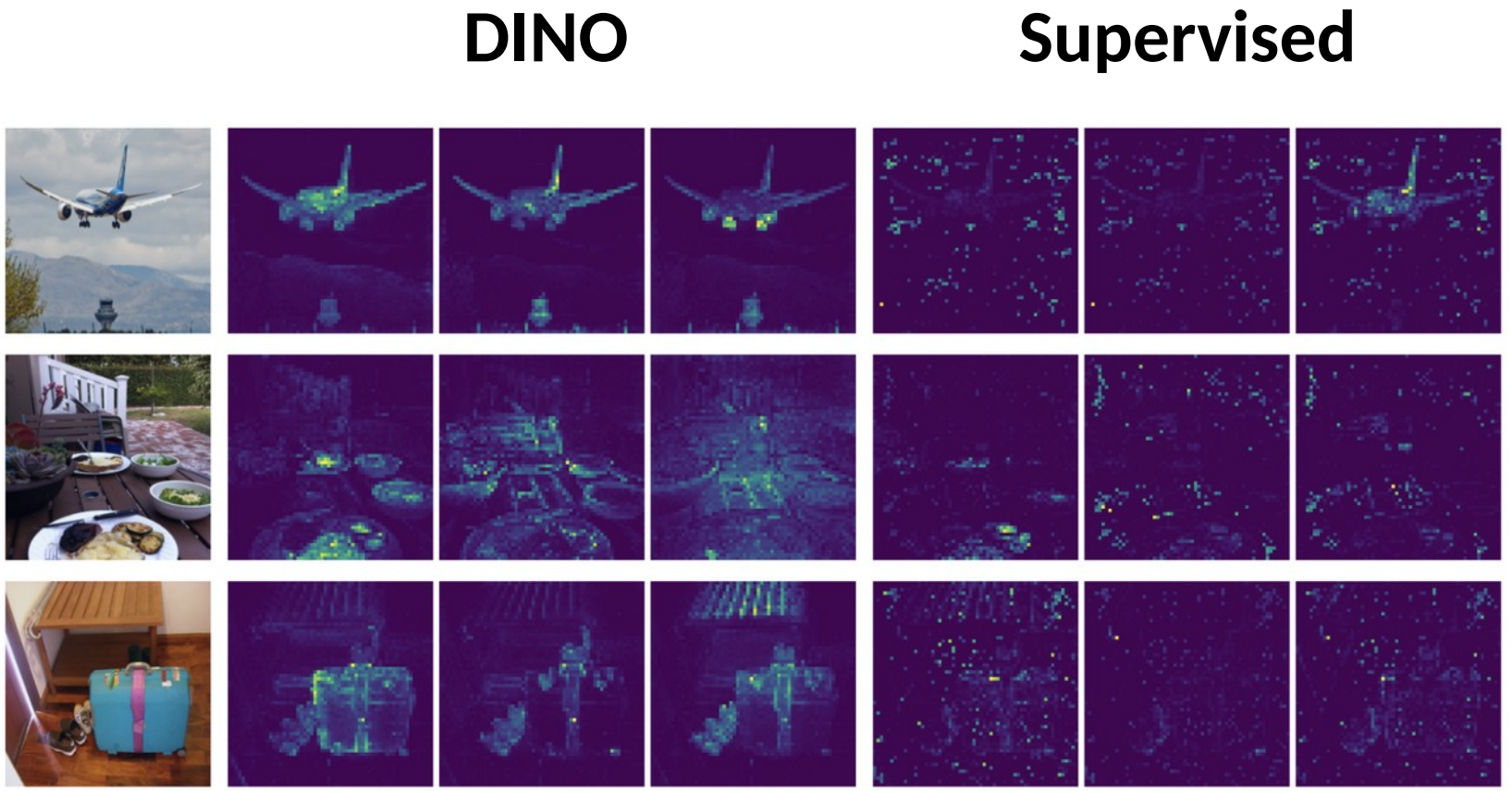
Encoder/Decoder types

encoder	decoder	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
joint	joint	49.7	48.0	51.5
joint	cross	44.6	43.6	45.7
joint	cross-self	41.1	39.6	42.7
siam	joint	56.7	55.4	58.1
siam	cross	52.2	51.2	53.1
siam	cross-self	**58.1**	**56.6**	**59.6**

Thank You!

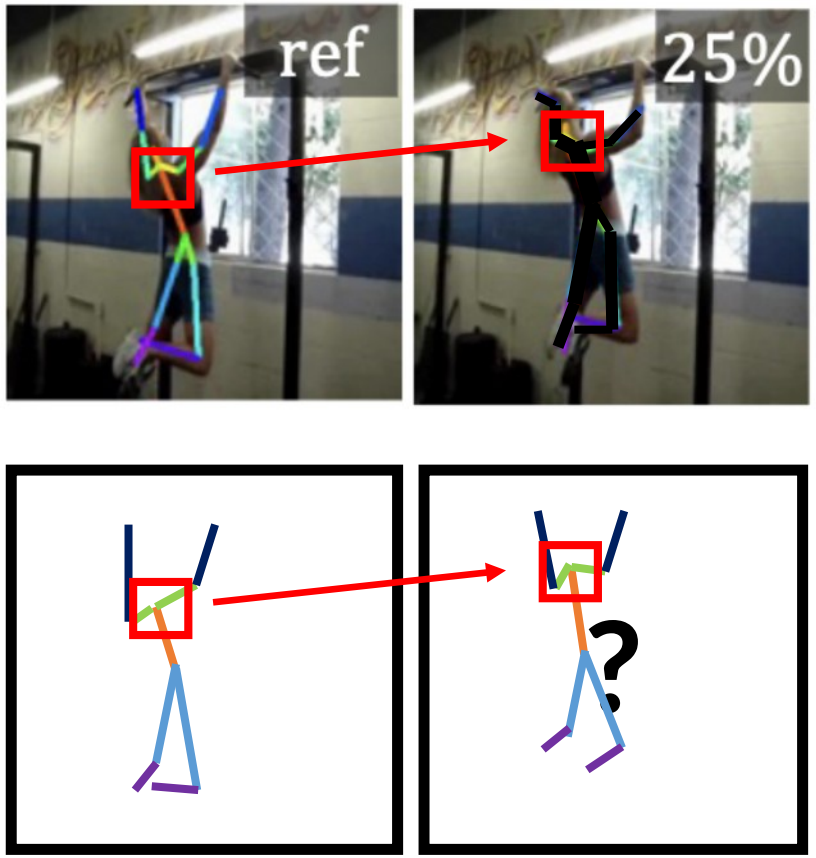
Data Augmentations

spatial	color	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
		56.8	55.5	58.1
✓		**58.1**	**56.6**	**59.6**
	✓	55.8	54.6	57.0
✓	✓	56.7	55.4	57.9


DINO
Supervised

$$A(j,i) = \frac{\exp\left(x^I(j)^\intercal x^p(i)\right)}{\sum_j \exp\left(x^I(j)^\intercal x^p(i)\right)}$$

$$y_i = \sum_j A_{t-1,t}(j,i) y_j$$


ref
25%
?