北京大学
PEKING UNIVERSITY

Siamese Masked Autoencoders

NeurIPS 2023 Oral

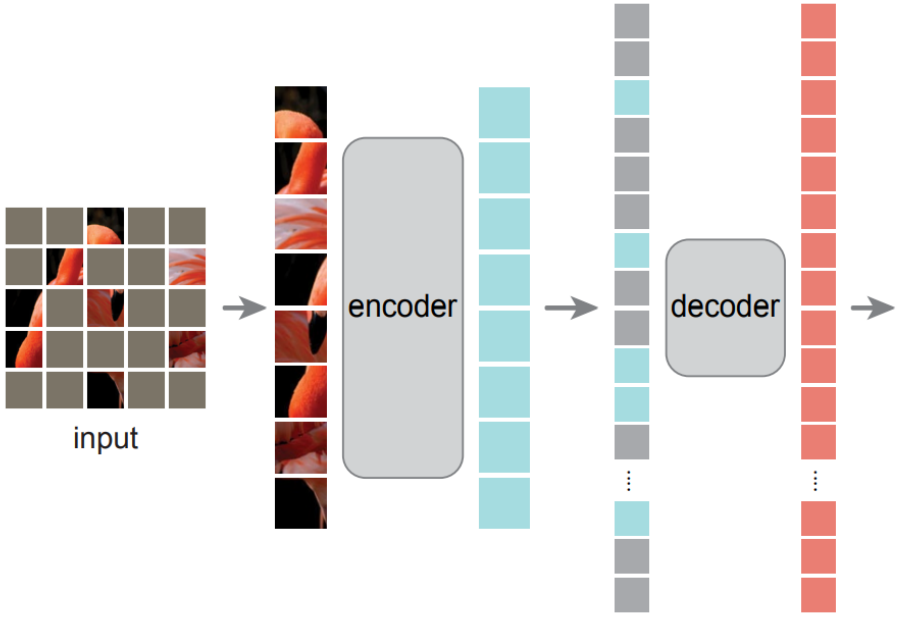
Presenter: Shaofan Sun

2024.1.14

Directory

北京大学 PEKING UNIVERSITY

Background: masked autoencoders (MAE)

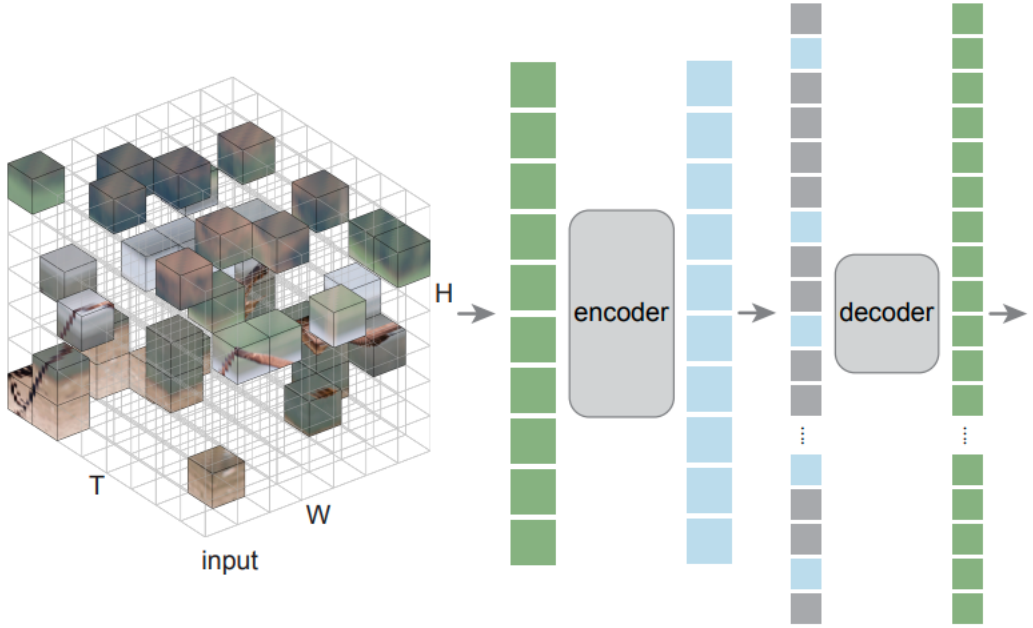


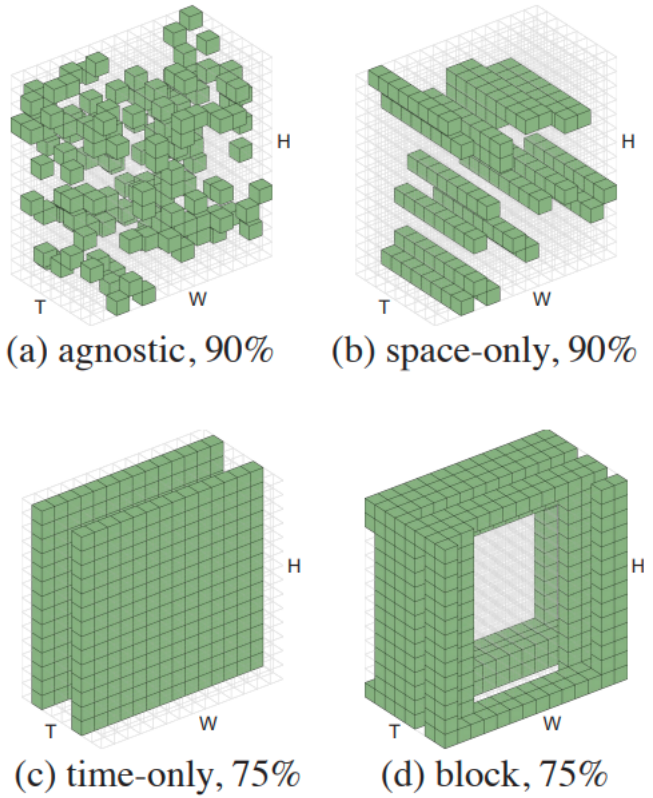
- Encoder: Vision Transformer
- Decoder: can be flexibly designed
- Masking ratio: 75%
  - BERT: 15%
  - More redundant information

Masked autoencoding: inspired by BERT

Masked Autoencoders Are Scalable Vision Learners, Kaiming He, et al., CVPR 2021

Background: MAE for videos

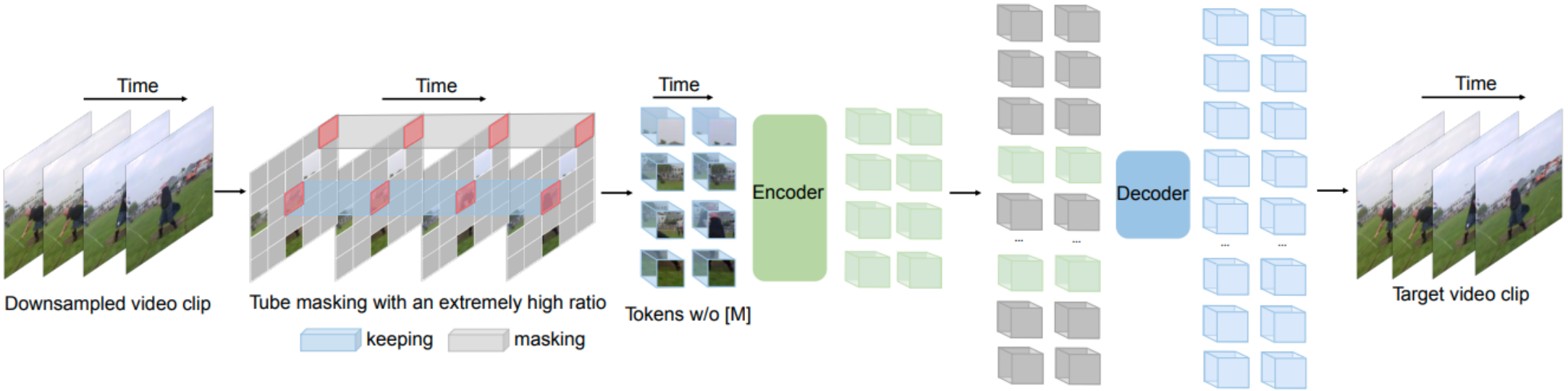




- Masking strategy: agnostic masking
- Masking ratio: 90%
  - More redundant information than images

Masked Autoencoders As Spatiotemporal Learners, Christoph Feichtenhofer, et al., NeurIPS 2022

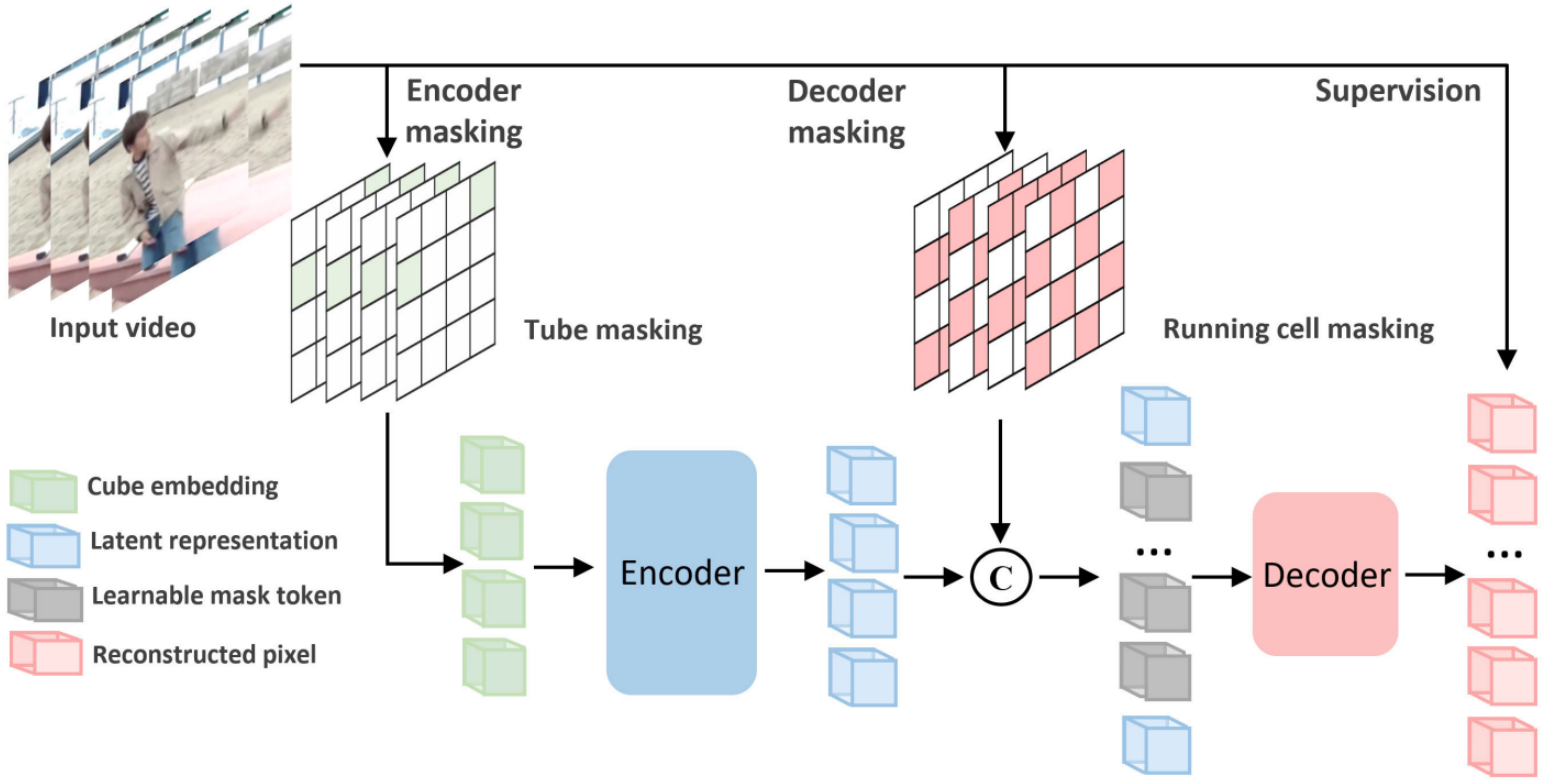
Background: VideoMAE



- Masking strategy: tube masking
- Masking ratio: 90%

VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, Zhan Tong, et al., NeurIPS 2022

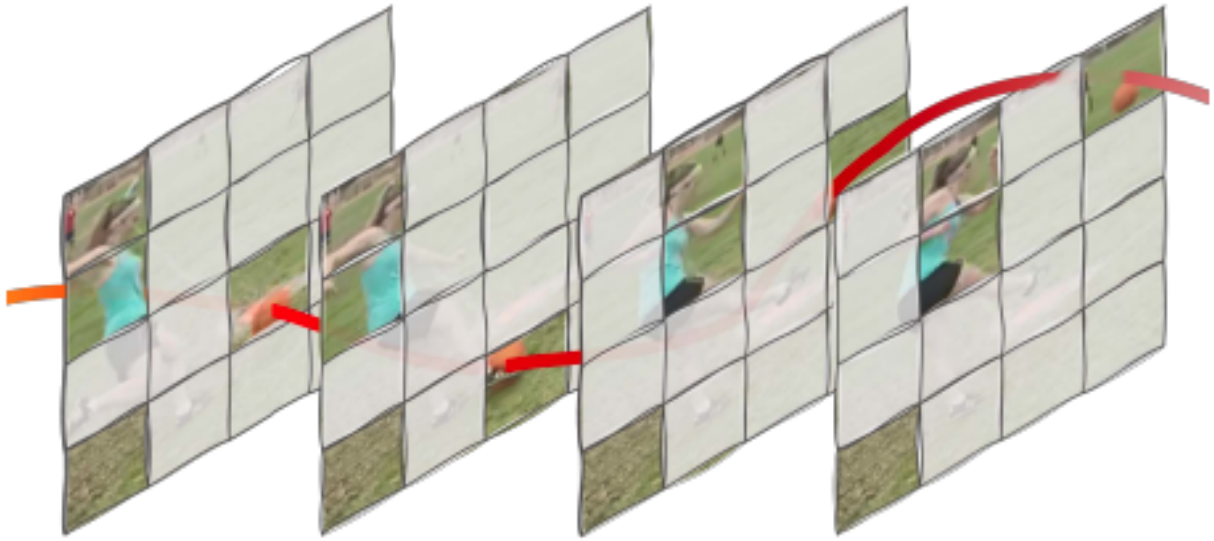
Background: VideoMAE v2



Dual masking: improve the efficiency

VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking, Limin Wang, et al., CVPR 2023

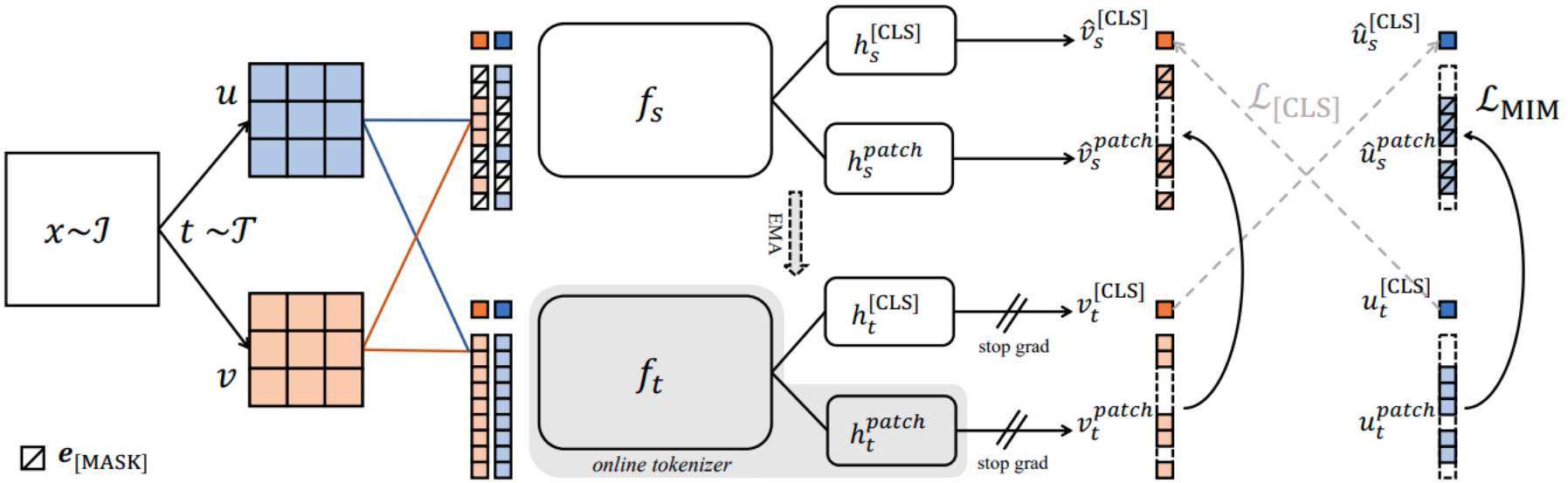
Background: MGMAE



- Motion-guided masking
- Optical flow

MGMAE: Motion Guided Masking for Video Masked Autoencoding, Bingkun Huang, et al., ICCV 2023

Background: iBOT

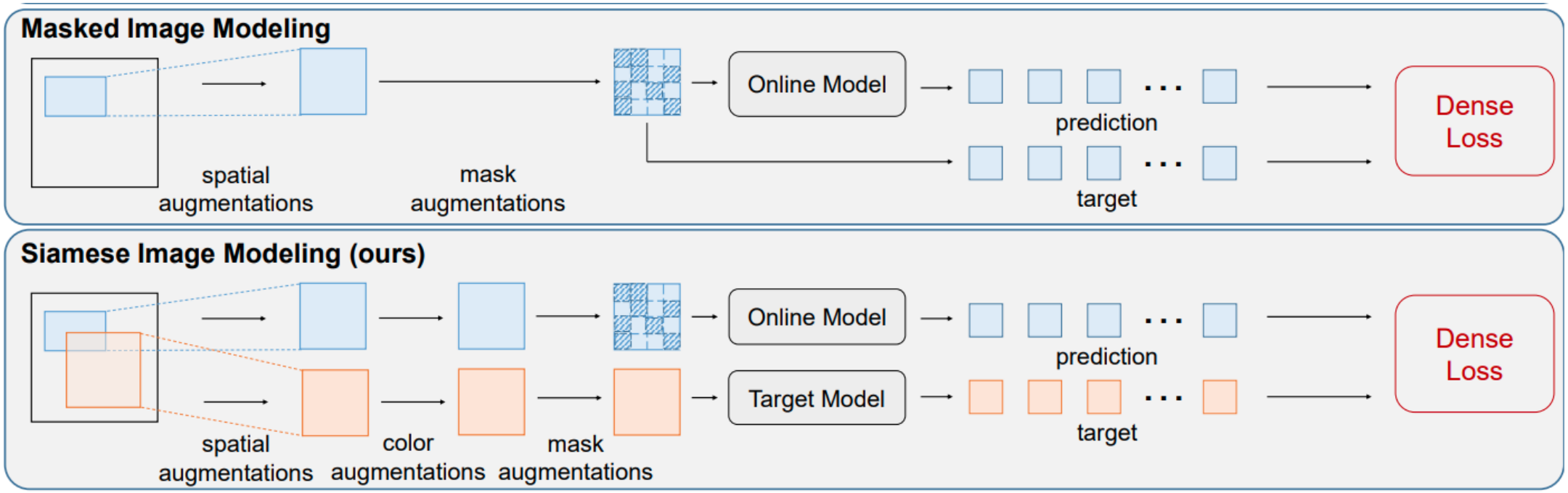


$$\mathcal{L}_{[\mathrm{CLS}]} = -P_{\boldsymbol{\theta}'}^{[\mathrm{CLS}]}(\boldsymbol{v})^{\mathrm{T}} \log P_{\boldsymbol{\theta}}^{[\mathrm{CLS}]}(\boldsymbol{u}).$$

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

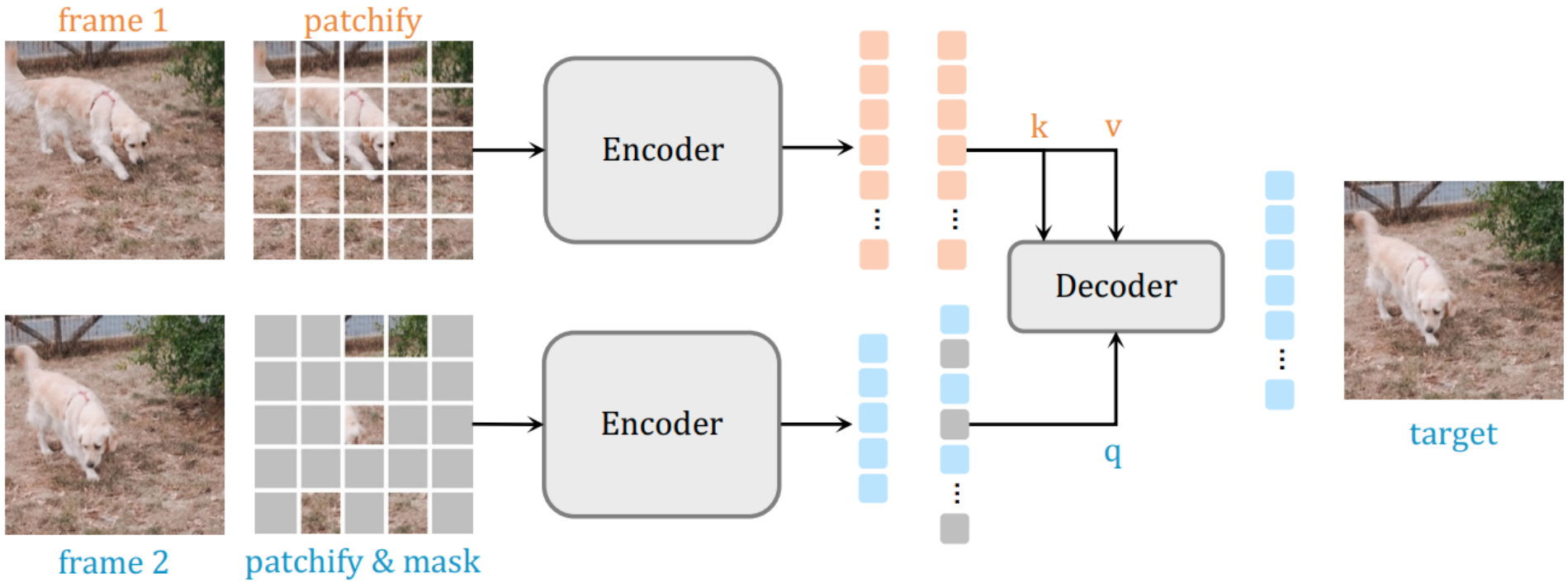
iBOT: Image BERT Pre-Training with Online Tokenizer, Jinghao Zhou, et al., ICLR 2022

Background: Siamese Image Modeling



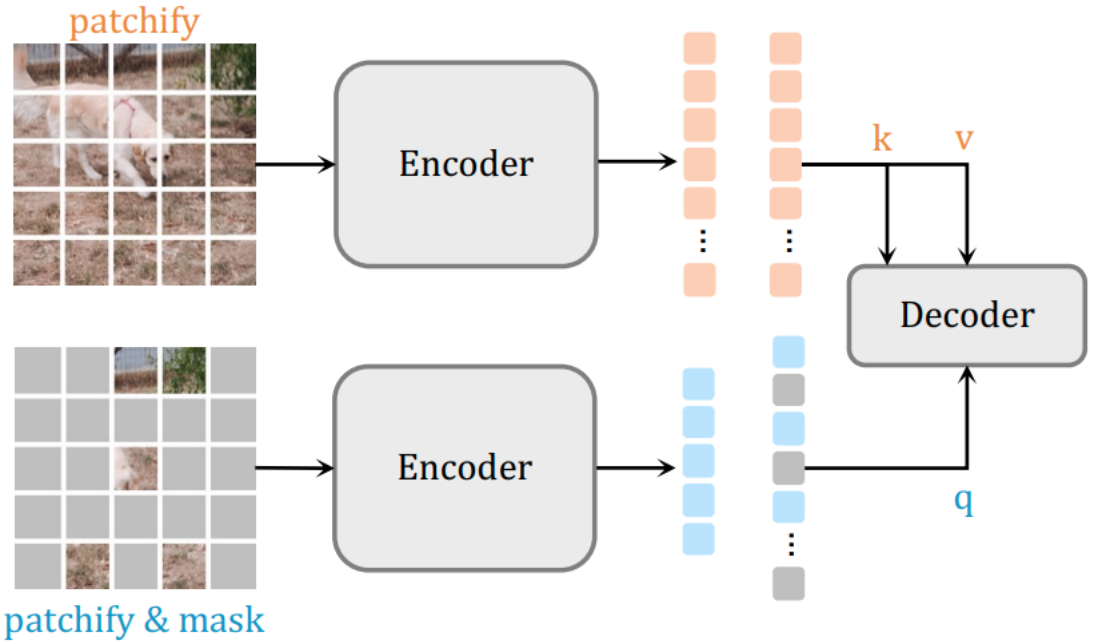
Siamese Image Modeling for Self-Supervised Vision Representation Learning, Chenxin Tao, et al., CVPR 2023

SiamMAE: Overview



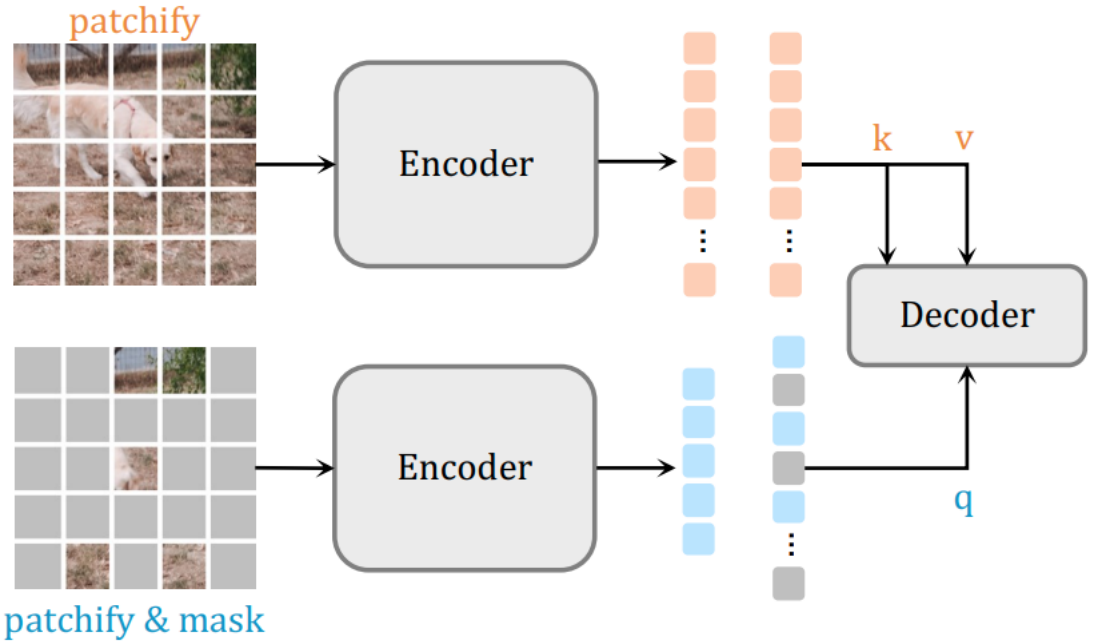
Different views: different frames

SiamMAE: Method



- Randomly selecting two frames f_1 and f_2
- Patchify: following ViT
- Masking: asymmetric masking
  - f_1 0% and f_2 95%

SiamMAE: Method



- Encoder: siamese encoder
- Decoder: cross-self decoder
  - similar to the original Transformer
- Loss: l_2 loss

SiamMAE: Experiments

北京大学 PEKING UNIVERSITY

- Backbone: ViT-16/S, for a fair comparison with ResNet-50 (21M vs. 23M)

- Pretraining:
  - Dataset: Kinetics-400
  - Frame gap: 4 to 48 frames (30 fps)
  - Augmentations: random resized cropping, horizontal flipping
  - Batch size: 2048
  - Epochs: 2000 (ablations 400)
  - GPU: 4 Nvidia Titan RTX GPUs

SiamMAE: Experiments

北京大学 PEKING UNIVERSITY

- Evaluation methodology: video label propagation

  - Video object segmentation
    - Dataset: Densely Annotated VIdeo Segmentation (DAVIS)
    - Metrics: $\mathcal{J}_m$, $\mathcal{F}_m$, $\mathcal{J}\&\mathcal{F}_m$
  - Human pose propagation
    - Dataset: Joint-annotated Human Motion Data Base (JHMDB)
    - Metrics: PCK
  - Semantic part propagation:
    - Dataset: Video Instance-level Parsing (VIP)
    - Metrics: mIoU

SiamMAE: Experiments

北京大学 PEKING UNIVERSITY

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: Experiments

北京大学 PEKING UNIVERSITY

- Ablation studies:
  - Encoder
    - siamese encoder
    - joint encoder: concatenate two frames
  - Decoder
    - joint decoder
    - cross decoder: only cross-attention layers
    - cross-self decoder
  - Masking
    - random, grid
    - symmetric, asymmetric

SiamMAE: Experiments

encoder	decoder	mask ratio	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
joint	joint	0.50 (s)	51.8	50.7	52.9
joint	joint	0.75 (s)	55.4	54.3	56.6
joint	joint	0.90 (s)	51.9	50.8	52.9
siam	cross-self	0.95 (a)	**58.1**	**56.6**	**59.6**

(a) **FrameMAE.** Simple extension of MAEs to frames does not work.

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**

(b) **Encoder-decoder design.** The combination of a siamese encoder and a cross-self decoder works the best.

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**

(c) **Symmetric masking.** Symmetric random masking degrades performance.

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**

(d) **Asymmetric masking.** Extremely high asymmetric masking is essential.

SiamMAE: Experiments

北京大学 PEKING UNIVERSITY

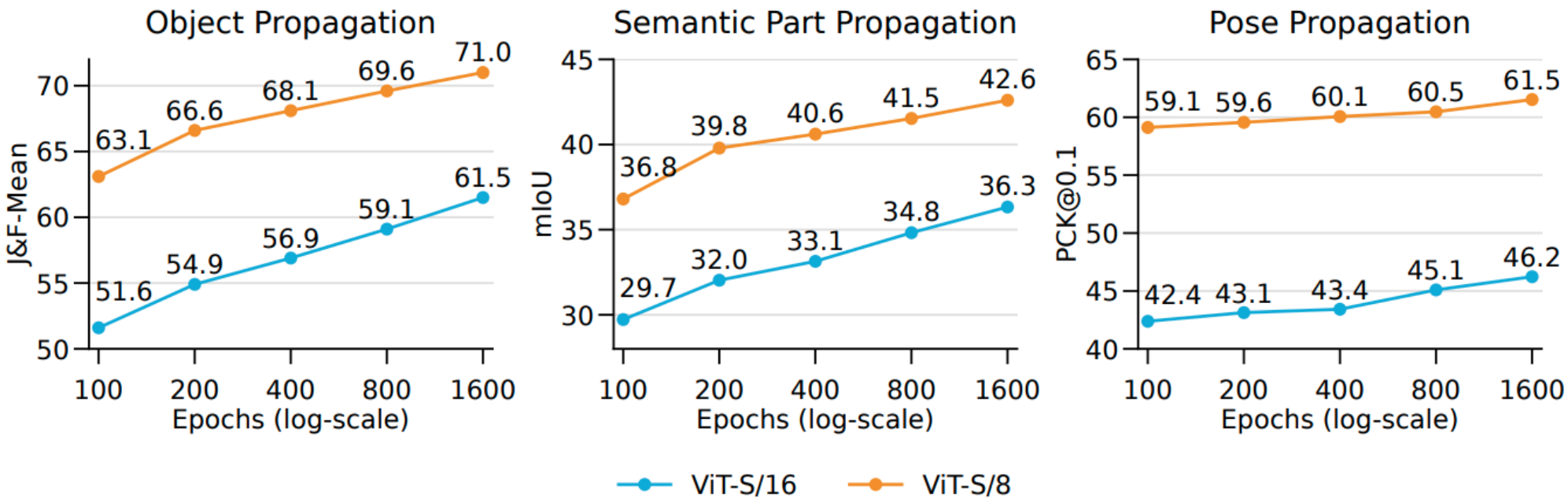
spatial	color	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
		56.8	55.5	58.1
✓		**58.1**	**56.6**	**59.6**
	✓	55.8	54.6	57.0
✓	✓	56.7	55.4	57.9

(a) **Data augmentation.** SiamMAE requires minimal data augmentation.

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**

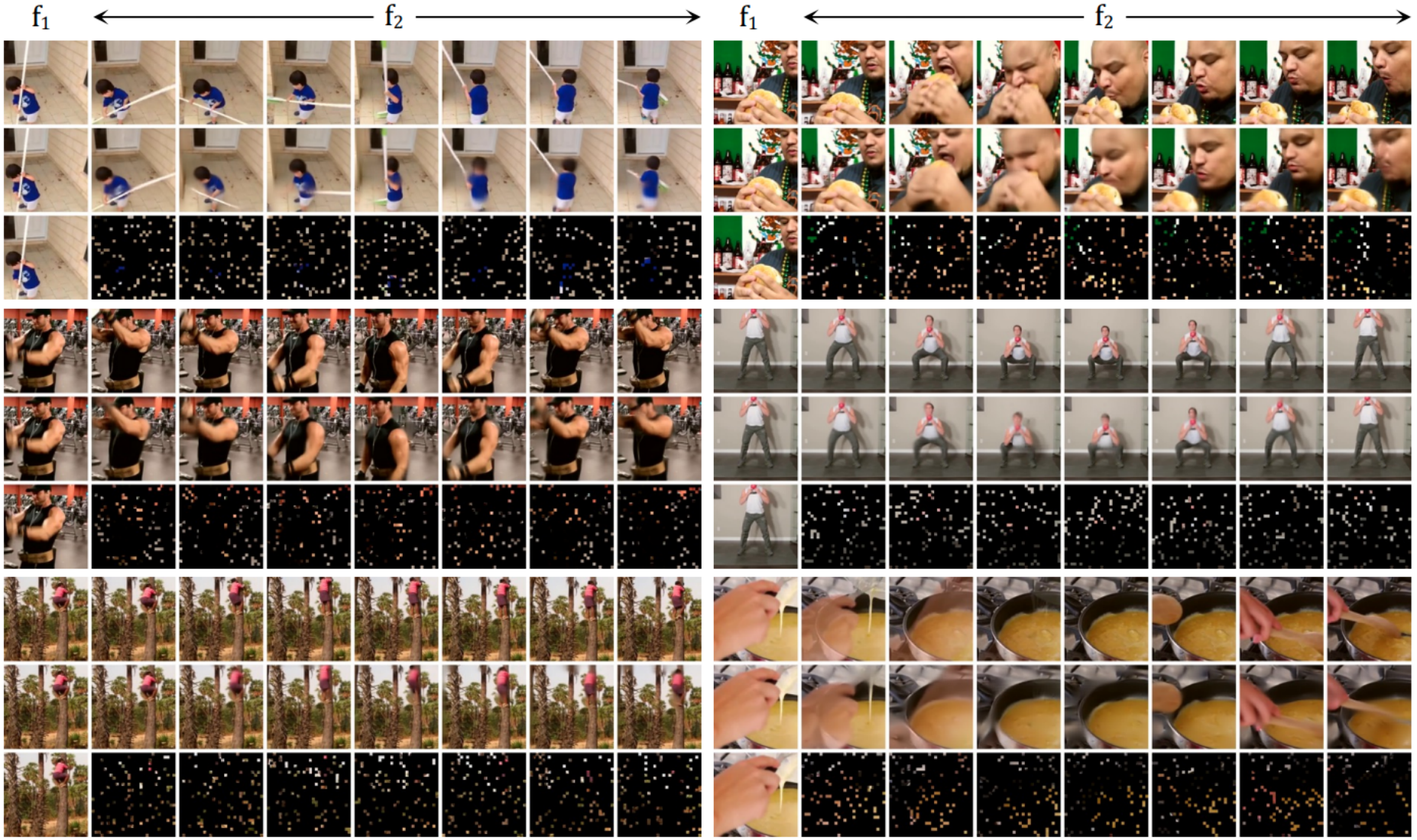
(b) **Frame sampling.** Random frame gap works the best.

SiamMAE: Experiments



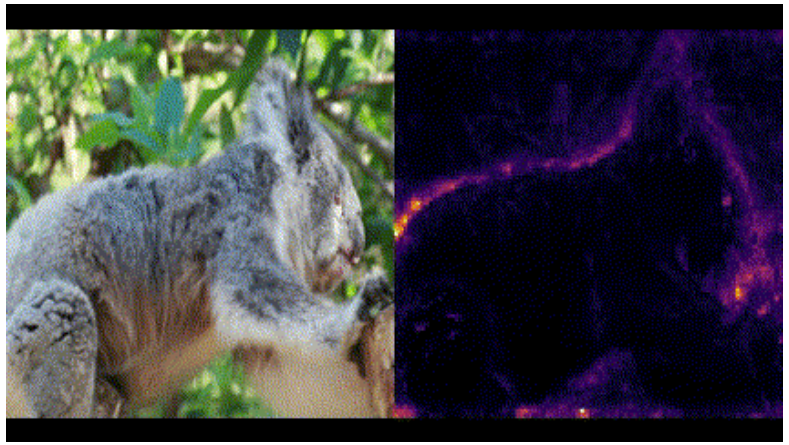
Longer training and smaller patch sizes lead to improved performance.

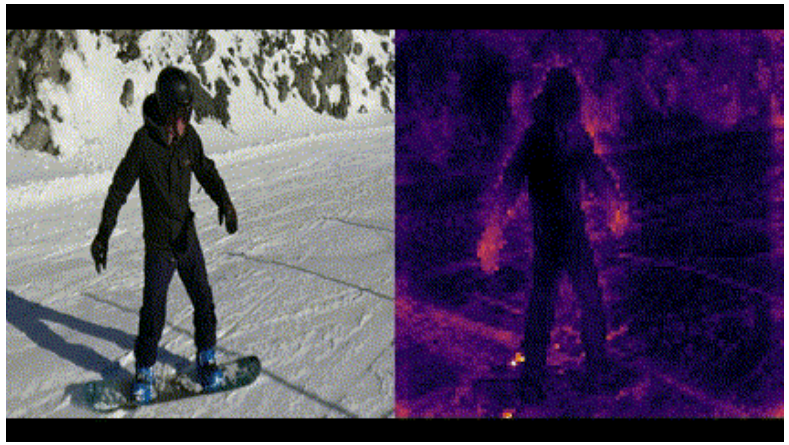
SiamMAE: Visualization Results



Kinetics-400

SiamMAE: Visualization Results



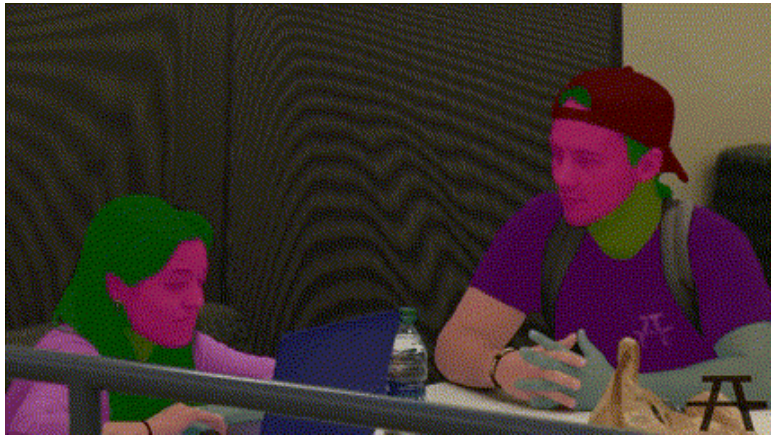


Attention Map

SiamMAE: Visualization Results

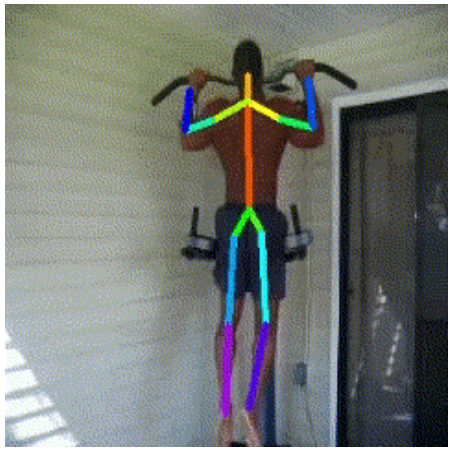
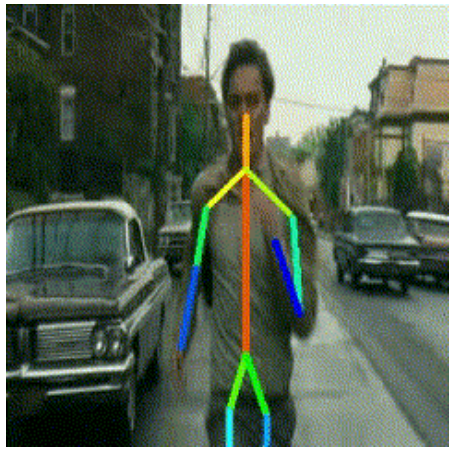
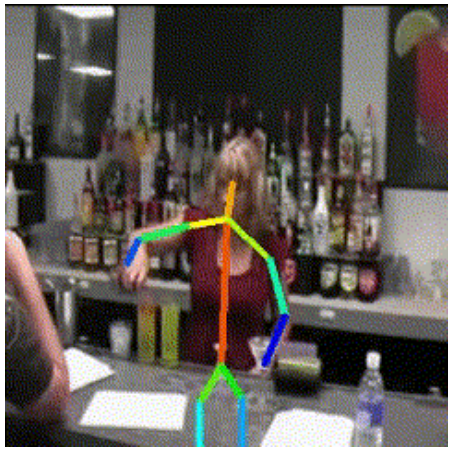
DAVIS: Video Object Segmentation

SiamMAE: Visualization Results



VIP: Semantic Part Propagation

SiamMAE: Visualization Results


JHMDB: Pose Keypoint Propagation

Conclusion

北京大学 PEKING UNIVERSITY

- A simple method for representation learning from videos
- Core idea: learning correspondence among frames, but treating them differently
- Why does it work?
  - Lack of theoretical analysis
- Further investigation: predicting multiple future frames based on past frames

Thanks for listening!