DropMAE: Masked Autoencoders with Spatial-Attention Dropout for Tracking Tasks

Qiangqiang Wu[1] Tianyu Yang[2*] Ziquan Liu[1] Baoyuan Wu[4] Ying Shan[3] Antoni B. Chan[1]

[1]Department of Computer Science, City University of Hong Kong

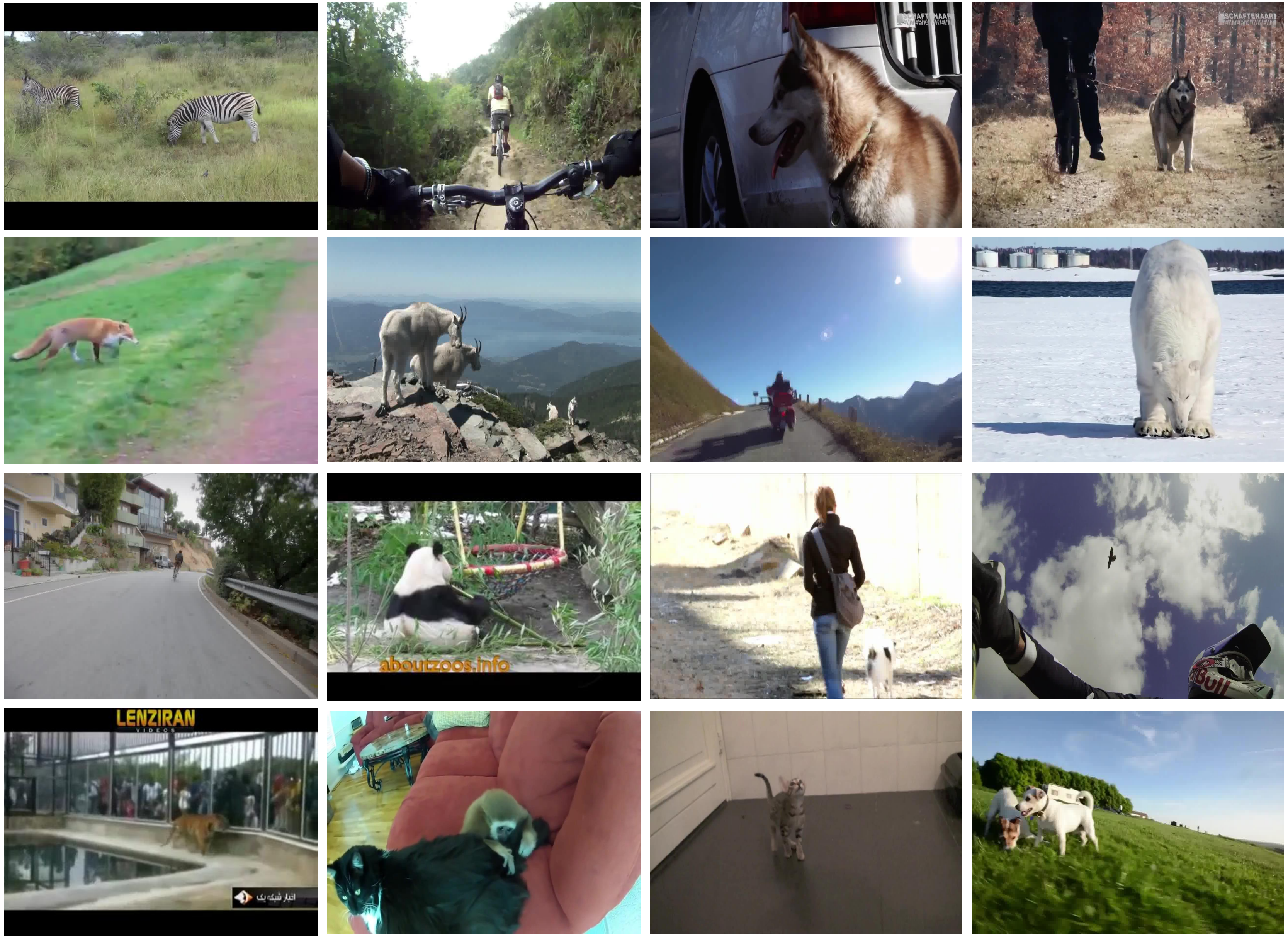
[2]International Digital Economy Academy [3]Tencent AI Lab

[4]School of Data Science, The Chinese University of Hong Kong, Shenzhen

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

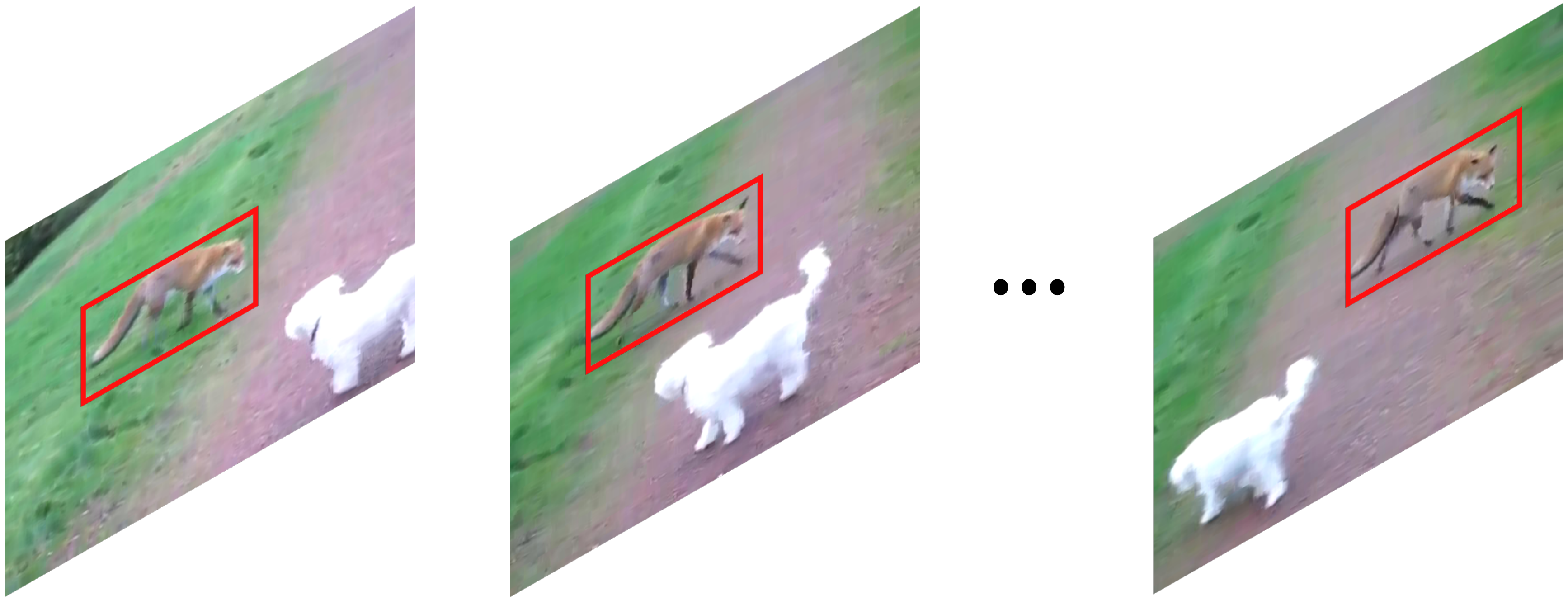
CityU
香港城市大學
City University of Hong Kong

Motivation: Large-Scale Unlabeled Videos



Motivation: The Excellent Power of MAE Pre-training

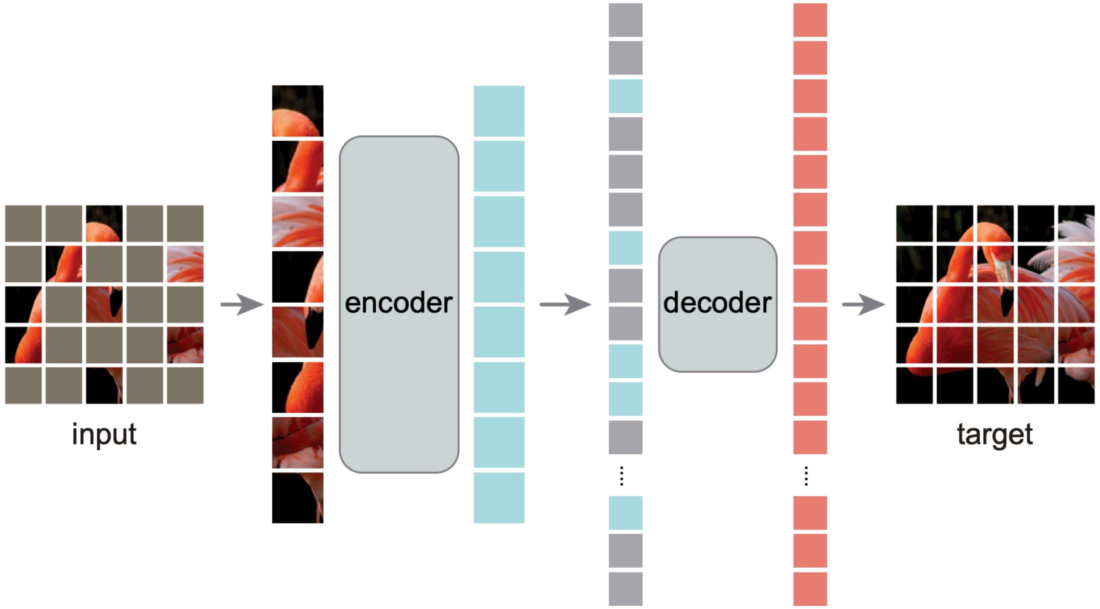
- Lack of Applications in Matching-based Downstream Tasks:
  - Video Object Tracking (VOT)
  - Video Object Segmentation (VOS)



Masked autoencoders are scalable vision learners. CVPR 2022, K. He et al.

Baseline Method

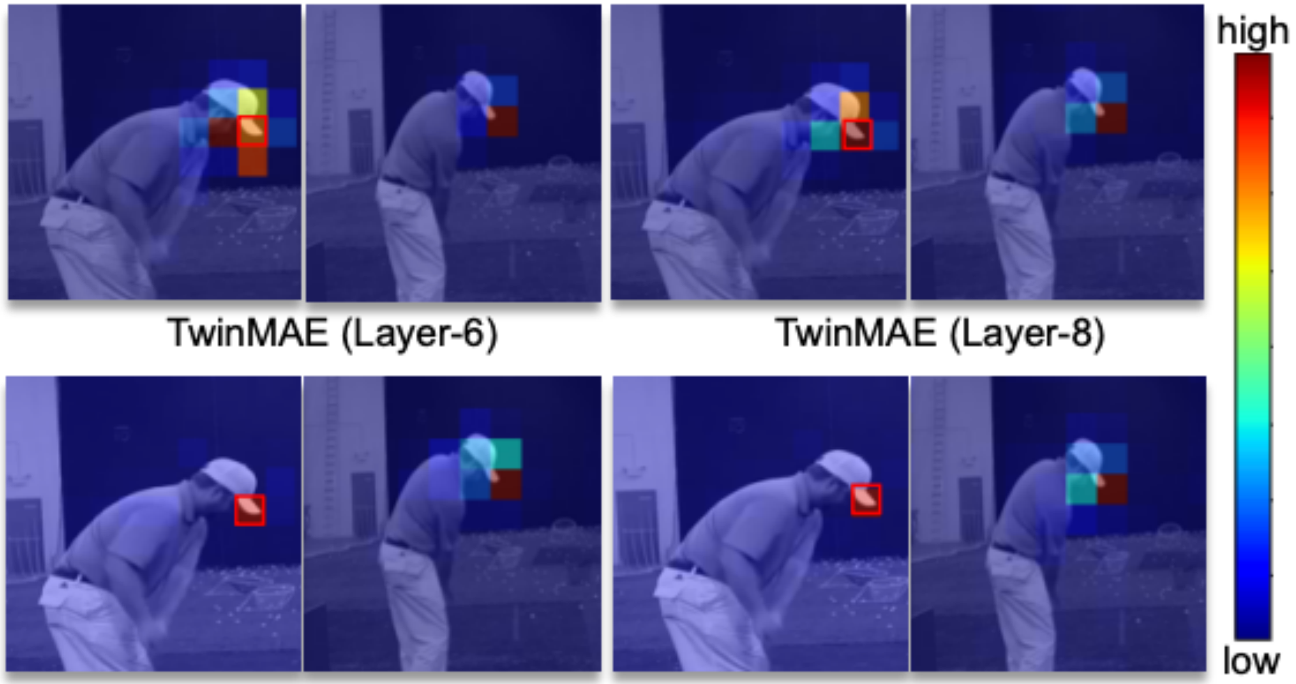
- MAE Pipeline:



- TwinMAE Baseline
  - Randonly sample 2-frames in a video.
  - Perform random mask on the sampled dual frames.
  - Input the masked frames to TwinMAE for reconstruction.
  - Trained on Kinetic datasets.

Visualization

- TwinMAE
  - Reconstruction heavily relies on within-frame patches or spatial cues, which may lead to sub-optimal temporal representations for matching-based video tasks.
  - Suboptimal temporal correspondence learning.

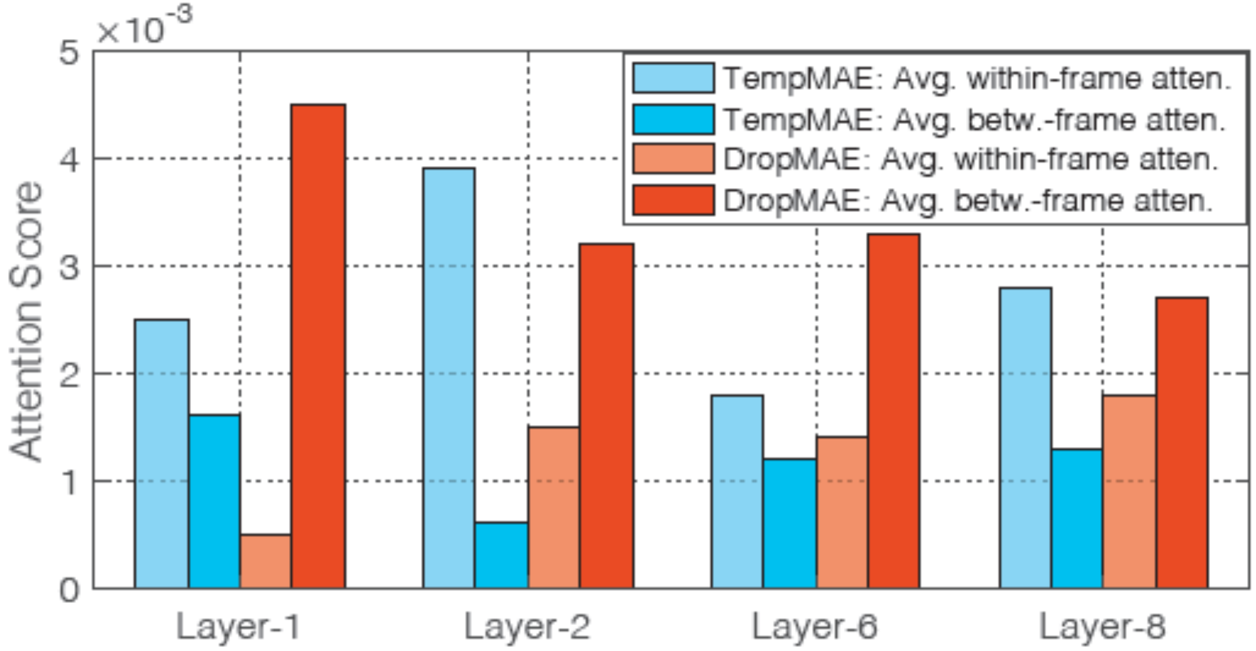


TwinMAE (Layer-6) TwinMAE (Layer-8)

DropMAE (Layer-6) **DropMAE** (Layer-8)

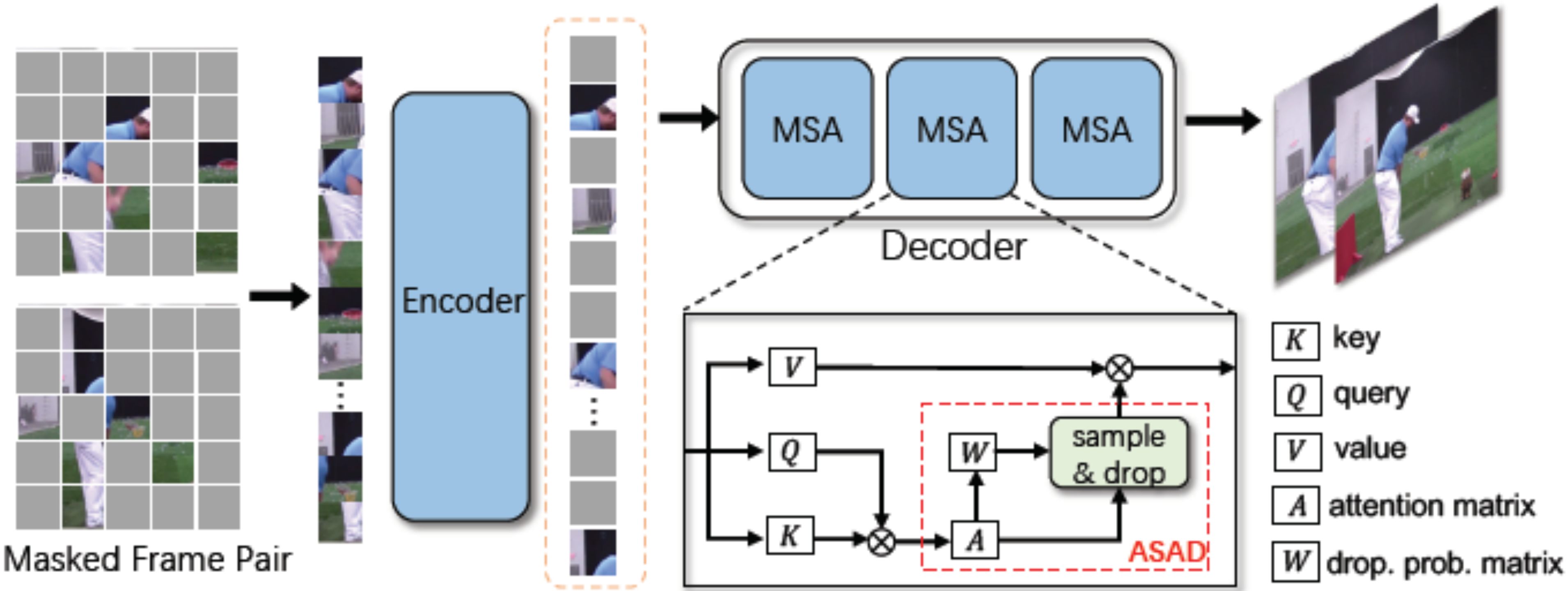
Visualization

- The average within-frame and between-frame attention scores obtained by TwinMAE and DropMAE in different decoder layers are shown in below.
- The attention score is calculated on 20 randomly sampled K400 validation videos, and is averaged on all heads and locations.



Overall Pipeline

- DropMAE
  - Transformer Encoder.
  - Transformer Decoder.
  - Adaptive Spatial Attention Dropout (ASAD) Module.



Adaptive Spatial Attention Dropout

- Focus more on temporal cues for reconstruction
  - **Goal:** facilitate the **temporal correspondence learning** in masked video pre-training.

- Temporal matching probability
  - Intuitively, a query token that has a strong match in the other frame should be a good candidate for ASAD, since in the absence of within frame cues, it can still be reconstructed well using the **temporal cues** in the other frame.
  - Here, we define a **temporal matching function** $f_{tem}(\cdot)$ to measure the temporal matching probability of the i-th query token:

$$f_{tem}(i) = \max_{j \in \Omega_t(i)} (\hat{A}_{i,j}), \quad \hat{A} = \mathrm{softmax}_{\mathrm{row}}(A),$$

Where A is the attention matrix of one head in a decoder layer, $\Omega_t(i)$ denotes the temporal index set of the i-th query token.

Adaptive Spatial Attention Dropout

- Overall Dropout Probability Measurement
  - The overall **spatial-attention dropout probability** at location (i, j) is measured by using both the temporal matching probability and the normalized spatial importance:

$$W_{i,j} = f_{tem}(i) \frac{\hat{A}_{i,j}}{\sum_{j \in \Omega_s(i)} \hat{A}_{i,j}},$$

where $\Omega_s(i)$ is the spatial index set that contains all the other token indices (i.e., excluding the query index itself) in the same frame as the i-th query.

- Sampling for Dropout
  - We draw N_d elements from a multinomial distribution based on the dropout probability matrix W.

Visualization of $f_{tem}(\cdot)$

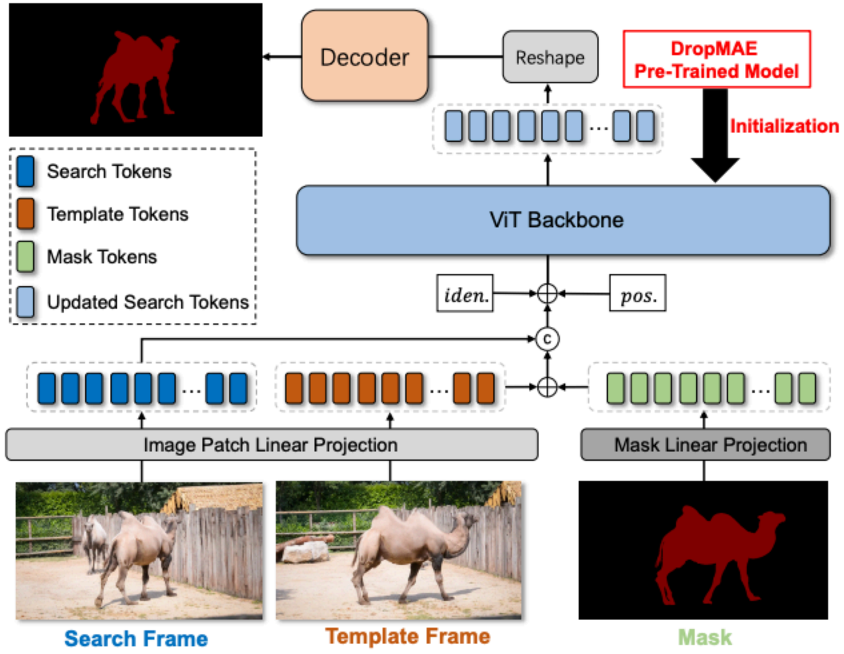
- Visualization of the **temporal matching function** on an example frame pair. A large value indicates that the i-th pixel matches well to a pixel in the other frame.

(a) layer-4 (b) layer-6

Downstream Tasks

- Video Object Tracking (VOT)
  - Use the state-of-the-art tracker OSTrack as our baseline.
  - Replace its pre-trained ViT model as our DropMAE ViT model.
  - Fine-tuning on VOT task following the convention.

- Video Object Segmentation (VOS)
  - Build a ViT-based VOS baseline.
  - Fine-tuning on VOS task.



VOS Framework

Joint feature learning and relation modeling for tracking: A one-stream framework. ECCV 2022, B. Ye et al.

Experimental Results

- Comparison with the other pre-training approaches on VOT/VOS.

Methods	Pre-training Data	Epochs	Pre-train. Time (h)	GOT-10k (VOT)			DAVIS-17 (VOS)		
				AO	$SR_{0.5}$	$SR_{0.75}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
No Pre-training	-	-	-	62.7	72.8	53.7	69.5	66.9	72.2
Supervised IN1k [75]	IN1K	300	-	69.7	79.0	65.6	78.0	74.8	81.1
Supervised IN21k [68]	IN21K	80	-	70.2	80.7	65.4	78.5	75.4	81.7
CLIP [67]	IN1K	32	-	67.4	76.8	60.0	73.6	70.5	76.7
MOCO-v3 [13]	IN1K	300	-	70.1	80.1	65.3	78.4	75.4	81.5
BeiT [2]	IN1K	800	103.1	67.4	76.8	60.0	76.1	72.7	79.4
MAE [37]	IN1K	1600	84	73.7	83.2	**70.8**	81.7	78.5	84.9
TwinMAE	K400	400	20.7	72.2	83.2	65.9	79.3	76.4	82.3
TwinMAE	K400	800	41.3	72.9	83.6	68.5	80.7	77.9	83.6
TwinMAE	K400	1600	82.7	74.2	84.9	69.4	81.2	78.1	84.2
DropMAE	K400	400	21.1	73.2	83.9	67.5	81.3	78.5	84.0
DropMAE	K400	800	42.2	74.8	85.4	70.5	82.7	**79.7**	85.6
DropMAE	K400	1600	84.4	**75.8**	**86.4**	**72.0**	**83.1**	**80.2**	**86.0**
DropMAE	K700	800	92.4	**75.9**	**86.8**	**72.0**	**83.0**	**80.2**	**85.7**

Experimental Results

- Comparison with state-of-the-art VOT approaches on four large-scale challenging datasets.

Method	Source	GOT-10k [40]			TNL2K [82]		$\text{LaSOT}_{\text{ext}}$ [28]			LaSOT [29]		
		AO	$SR_{0.5}$	$SR_{0.75}$	AUC	P	AUC	P_{Norm}	P	AUC	P_{Norm}	P
SiamFC [3]	ECCVW16	34.8	35.3	9.8	29.5	28.6	23.0	31.1	26.9	33.6	42.0	33.9
MDNet [60]	CVPR16	29.9	30.3	9.9	-	-	27.9	34.9	31.8	39.7	46.0	37.3
ECO [20]	ICCV17	31.6	30.9	11.1	32.6	31.7	22.0	25.2	24.0	32.4	33.8	30.1
SiamPRN++ [43]	CVPR19	51.7	61.6	32.5	41.3	41.2	34.0	41.6	39.6	49.6	56.9	49.1
DiMP [4]	ICCV19	61.1	71.7	49.2	44.7	43.4	39.2	47.6	45.1	56.9	65.0	56.7
SiamR-CNN [77]	CVPR20	64.9	72.8	59.7	52.3	52.8	-	-	-	64.8	72.2	-
LTMU [19]	CVPR20	-	-	-	48.5	47.3	41.4	49.9	47.3	57.2	-	57.2
Ocean [107]	ECCV20	61.1	72.1	47.3	38.4	37.7	-	-	-	56.0	65.1	56.6
TrDiMP [79]	CVPR21	67.1	77.7	58.3	-	-	-	-	-	63.9	-	61.4
TransT [14]	CVPR21	67.1	76.8	60.9	50.7	51.7	-	-	-	64.9	73.8	69.0
AutoMatch [105]	ICCV21	65.2	76.6	54.3	47.2	43.5	37.6	-	43.0	58.3	-	59.9
STARK [95]	ICCV21	68.8	78.1	64.1	-	-	-	-	-	67.1	77.0	-
KeepTrack [57]	ICCV21	-	-	-	-	-	48.2	-	-	67.1	77.2	70.2
MixFormer-L [18]	CVPR22	70.7	80.0	67.8	-	-	-	-	-	70.1	79.9	76.3
SBT [90]	CVPR22	70.4	80.8	64.7	-	-	-	-	-	66.7	-	71.1
UAST [101]	ICML22	63.5	74.1	51.4	-	-	-	-	-	57.1	-	58.7
SwinTrack-384 [50]	NeurIPS22	72.4	80.5	67.8	**55.9**	**57.1**	49.1	-	55.6	**71.3**	-	76.5
AiATrack [33]	ECCV22	69.6	80.0	63.2	-	-	47.7	55.6	55.4	69.0	79.4	73.8
CIA50 [65]	ECCV22	67.9	79.0	60.3	50.9	57.6	-	-	-	66.2	-	69.6
SimTrack-L [10]	ECCV22	69.8	78.8	66.0	55.6	55.7	-	-	-	70.5	79.7	-
OSTrack-384 [100]	ECCV22	**73.7**	**83.2**	**70.8**	**55.9**	56.7	**50.5**	**61.3**	**57.6**	71.1	**81.1**	**77.6**
DropTrack	**Ours**	**75.9**	**86.8**	**72.0**	**56.9**	**57.9**	**52.7**	**63.9**	**60.2**	**71.8**	**81.8**	**78.1**

Experimental Results

- Comparison with state-of-the-art VOS approaches.

Method	Source	OL	M	S	DAVIS-2016 [64] $\mathcal{J}\&\mathcal{F}$	DAVIS-2016 [64] $\mathcal{J}$	DAVIS-2016 [64] $\mathcal{F}$	DAVIS-2017 [66] $\mathcal{J}\&\mathcal{F}$	DAVIS-2017 [66] $\mathcal{J}$	DAVIS-2017 [66] $\mathcal{F}$
RANet [83]	ICCV19			✓	85.5	85.5	85.4	65.7	63.2	68.2
STM [62]	ICCV19		✓	✓	89.3	**88.7**	89.9	81.8	79.2	84.3
FRTM [69]	CVPR20	✓	✓		83.5	83.6	83.4	76.7	73.9	79.6
TVOS [104]	CVPR20		✓		-	-	-	72.3	69.9	74.7
LWL [5]	ECCV20	✓	✓		-	-	-	81.6	79.1	84.1
CFBI [98]	ECCV20		✓		89.4	88.3	**90.5**	81.9	79.1	84.6
UniTrack [84]	NeurIPS21		✓		-	-	-	-	58.4	-
STCN⁻ [16]	NeurIPS21		✓		-	-	-	**82.5**	79.3	**85.7**
SSTVOS [27]	CVPR21		✓		-	-	-	**82.5**	**79.9**	85.1
SWEM⁻ [52]	CVPR22		✓		**89.5**	-	-	81.9	-	-
RTS [63]	ECCV22	✓	✓		-	-	-	80.2	77.9	82.6
OSMN [56]	TPAMI18				73.5	74.0	72.9	54.8	52.5	57.1
FAVOS [17]	CVPR18				81.0	82.4	79.5	58.2	54.6	61.8
VideoMatch [39]	ECCV18				-	81.0	-	56.5	-	-
SiamMask [80]	CVPR19				69.8	71.7	67.8	56.4	54.3	58.5
D3S [54]	CVPR20				74.0	75.4	72.6	60.8	57.8	63.8
Siam R-CNN [54]	CVPR20				-	-	-	70.6	66.1	75.0
Unicorn [94]	ECCV22				87.4	86.5	88.2	69.2	65.2	73.2
DropSeg	**Ours**				**92.1**	**90.9**	**93.3**	**83.0**	**80.2**	**85.7**

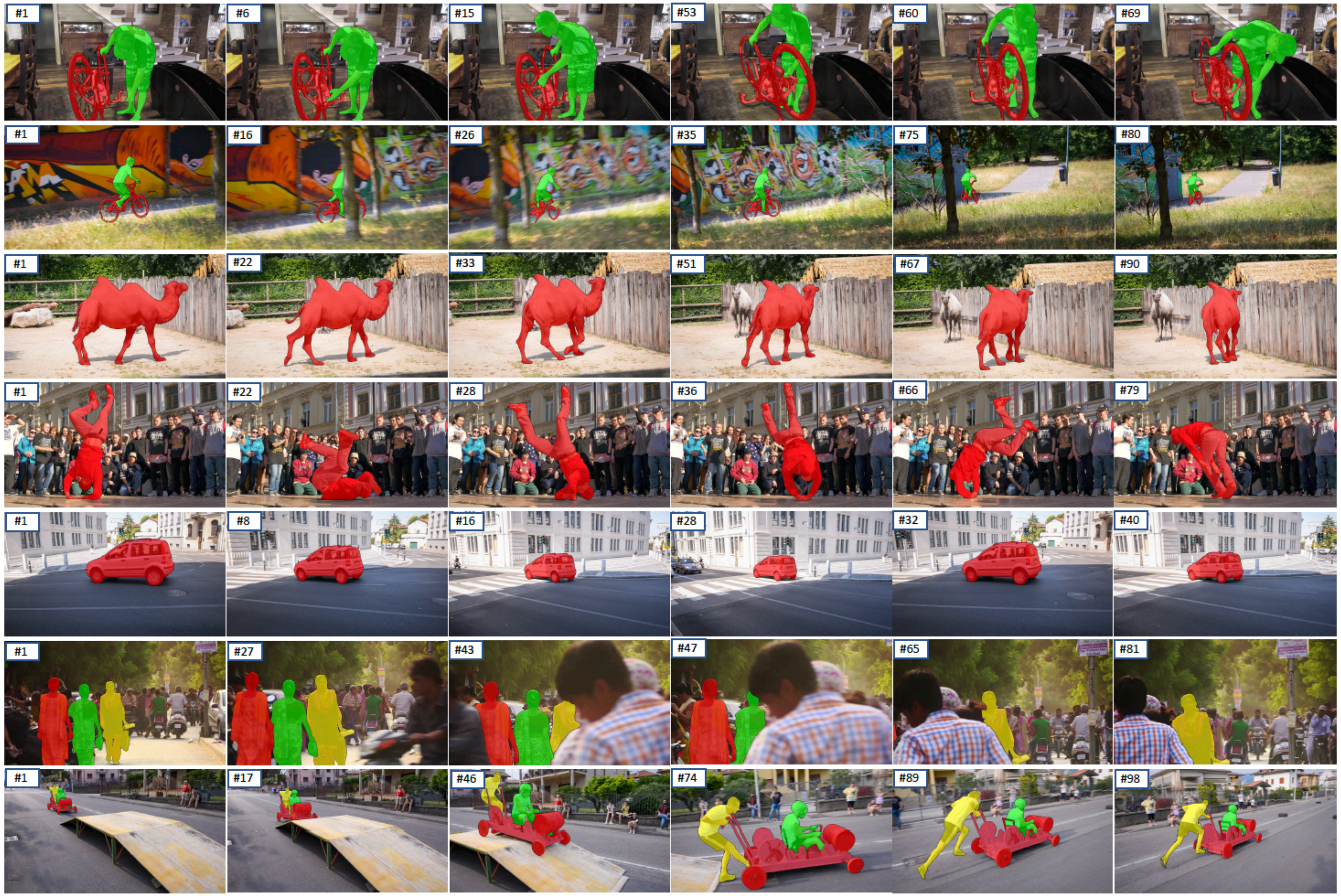
Data Sources

- **Motion diversity** in pre-training videos is more important than scene diversity for improving the performance on VOT and VOS.

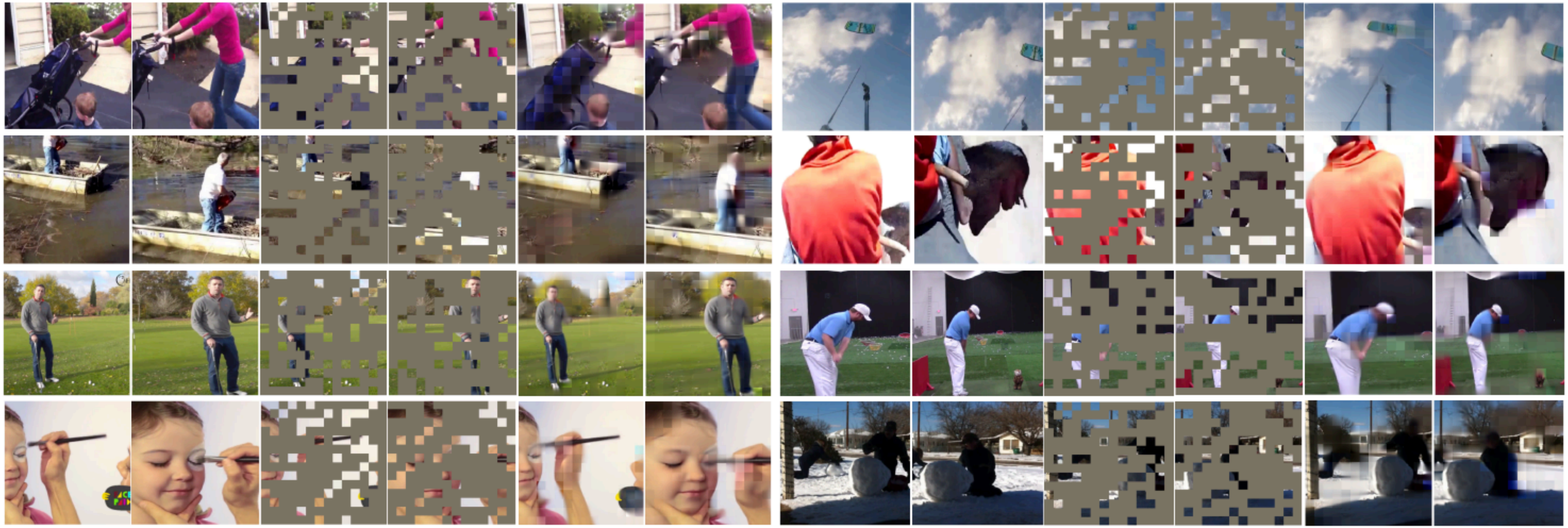
Datasets	No. Videos	No. Actions	VOT			VOS
			AO	$SR_{0.5}$	$SR_{0.75}$	$\mathcal{J}\&\mathcal{F}$
K400 [40]	240,000	400	73.2	83.9	67.5	82.7
K600 [8]	390,000	600	74.5	85.5	69.5	82.8
K700 [9]	526,768	700	**75.6**	**86.2**	**71.4**	**83.0**
MiT [53]	802,244	339	75.1	85.5	70.6	82.8
WebVid [1]	240,000	-	72.8	83.4	67.3	81.5
WebVid [1]	960,000	-	73.4	85.0	69.5	82.9

Qualitative Results: VOT

Qualitative Results: VOS



Qualitative Results: Frame Reconstruction



THANKS

https://github.com/jimmy-dq/DropMAE.git