JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

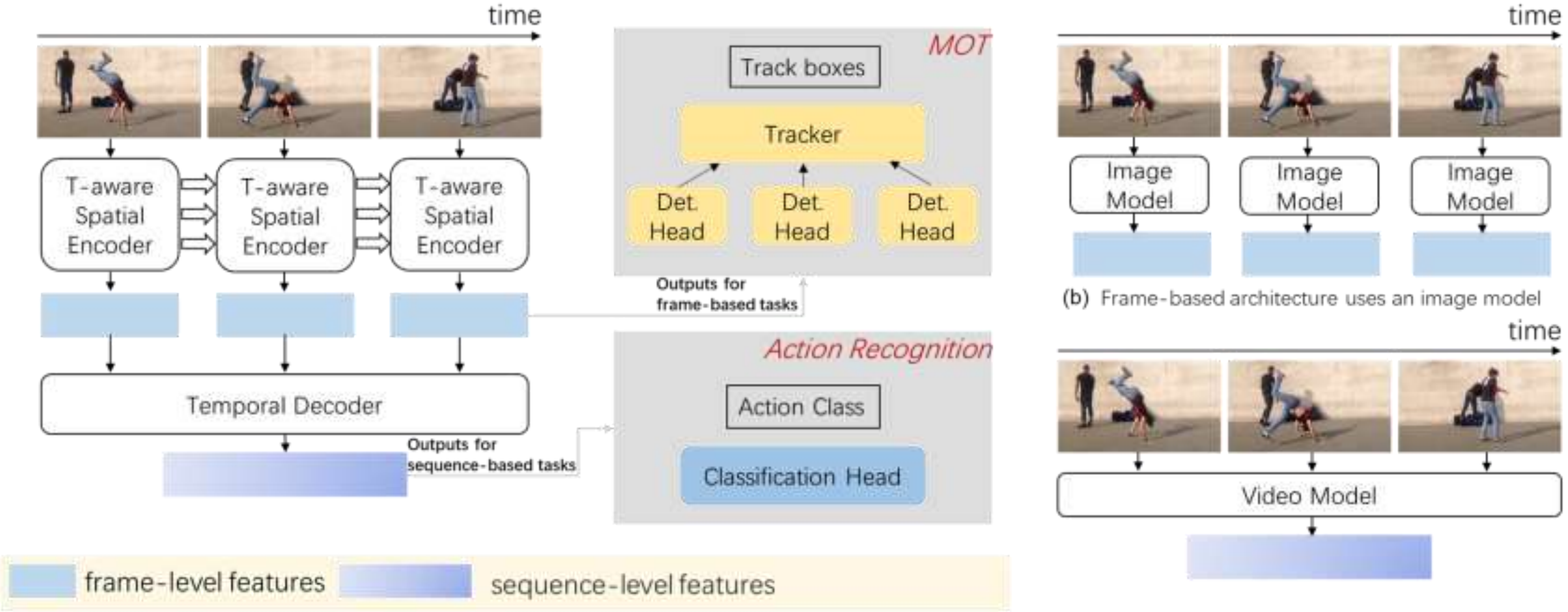
Streaming Video Model

Yucheng Zhao Chong Luo Chuanxin Tang Dongdong Chen Noel Codella Zheng-Jun Zha

WED-PM-215

Streaming Video Model

- Traditionally, two separate kinds of video models are used to solve the sequence-based tasks (e.g. action classification) and frame-based tasks (e.g. MOT).
- We propose a unified architecture, named streaming video model, for handling both types of tasks.

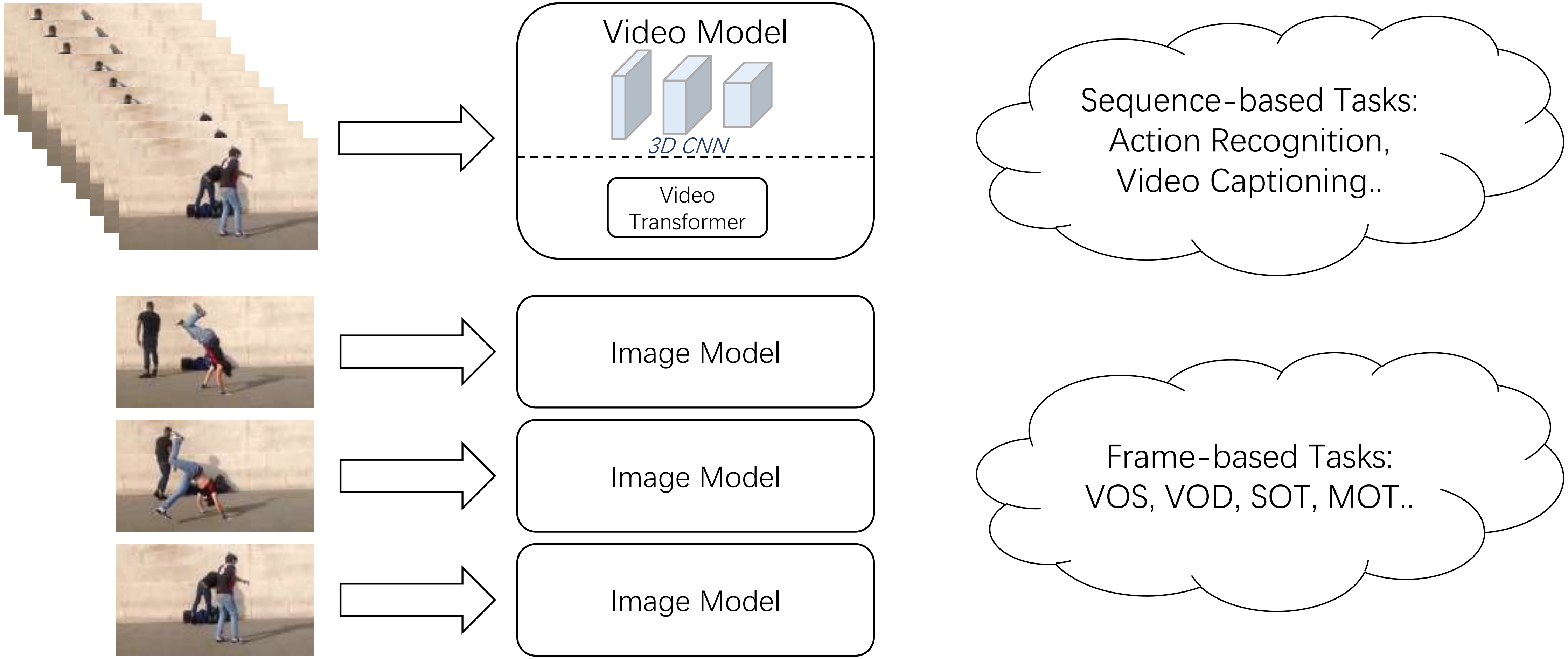


(a) Proposed streaming video model for both frame-based tasks and sequence-based tasks

(b) Frame-based architecture uses an image model

(c) Clip-based architecture uses a video model

Traditional Video Models



Video Model

3D CNN

Video Transformer

Sequence-based Tasks:
Action Recognition,
Video Captioning..

Image Model

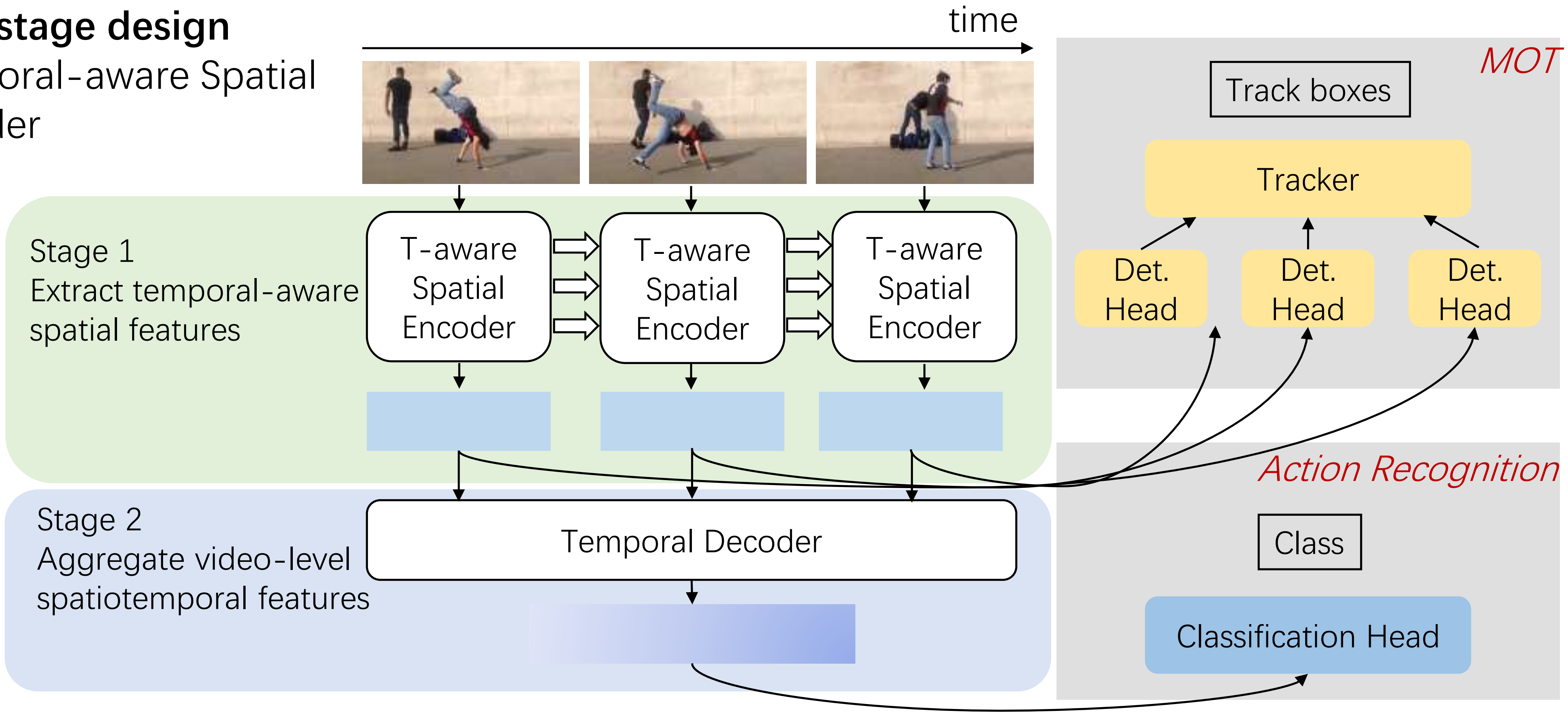
Image Model

Image Model

Frame-based Tasks:
VOS, VOD, SOT, MOT..

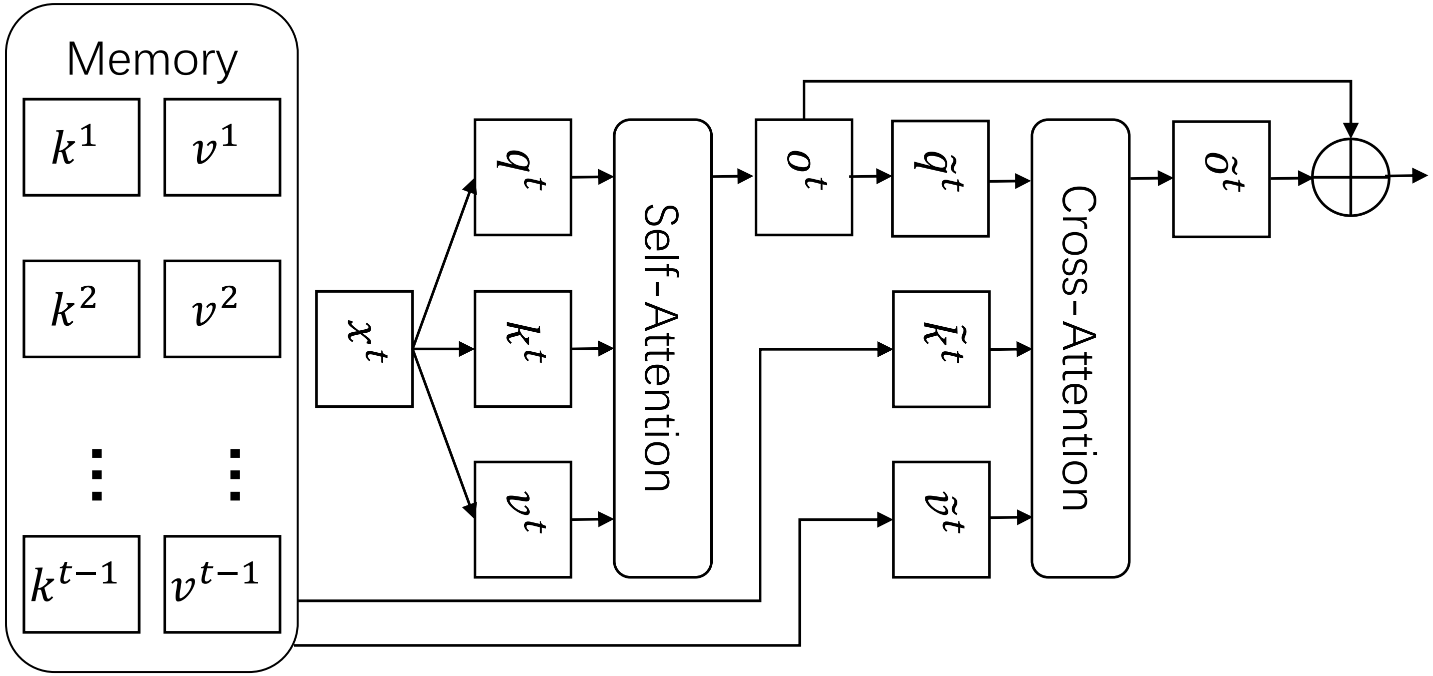
Method: Streaming Video Model

- Key Innovation
  - **Two-stage design**
  - Temporal-aware Spatial Encoder



Method: Streaming Video Model

- Key Innovation
  - Two-stage design
  - **Temporal-aware Spatial Encoder**



Streaming T2D Attention:

1. Intra-frame self-attention

$$q_t = x_t W_q; k_t = x_t W_k; v_t = x_t W_v, \tag{4}$$

$$o_t = \text{Attention}(q_t, k_t, v_t), \tag{5}$$

2. Building memory pool

$$\tilde{k}^t = [\text{sg}(k^1), \text{sg}(k^2), ..., \text{sg}(k^{t-1}), \text{sg}(k^t)], \tag{6}$$

$$\tilde{v}^t = [\text{sg}(v^1), \text{sg}(v^2), ..., \text{sg}(v^{t-1}), \text{sg}(v^t)]. \tag{7}$$

3. Inter-frame cross-attention

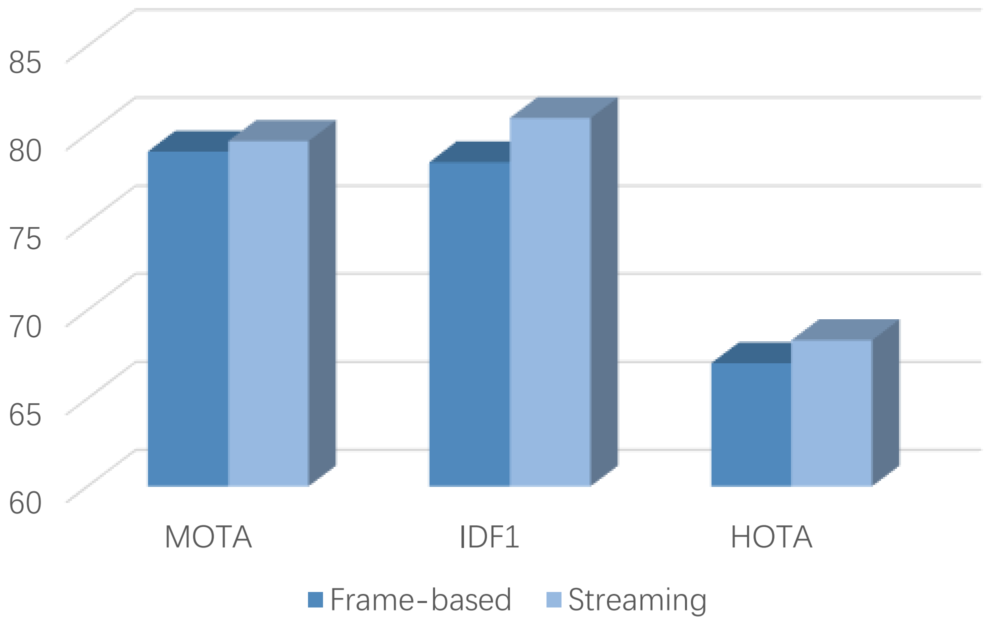
$$\tilde{q}^t = o_t \widetilde{W}_q \tag{8}$$

$$\tilde{o}^t = \text{Attention}(\tilde{q}_t, \tilde{k}_t, \tilde{v}_t). \tag{9}$$

T2d: Spatiotemporal feature learning based on triple 2d decomposition. Zhao, et al

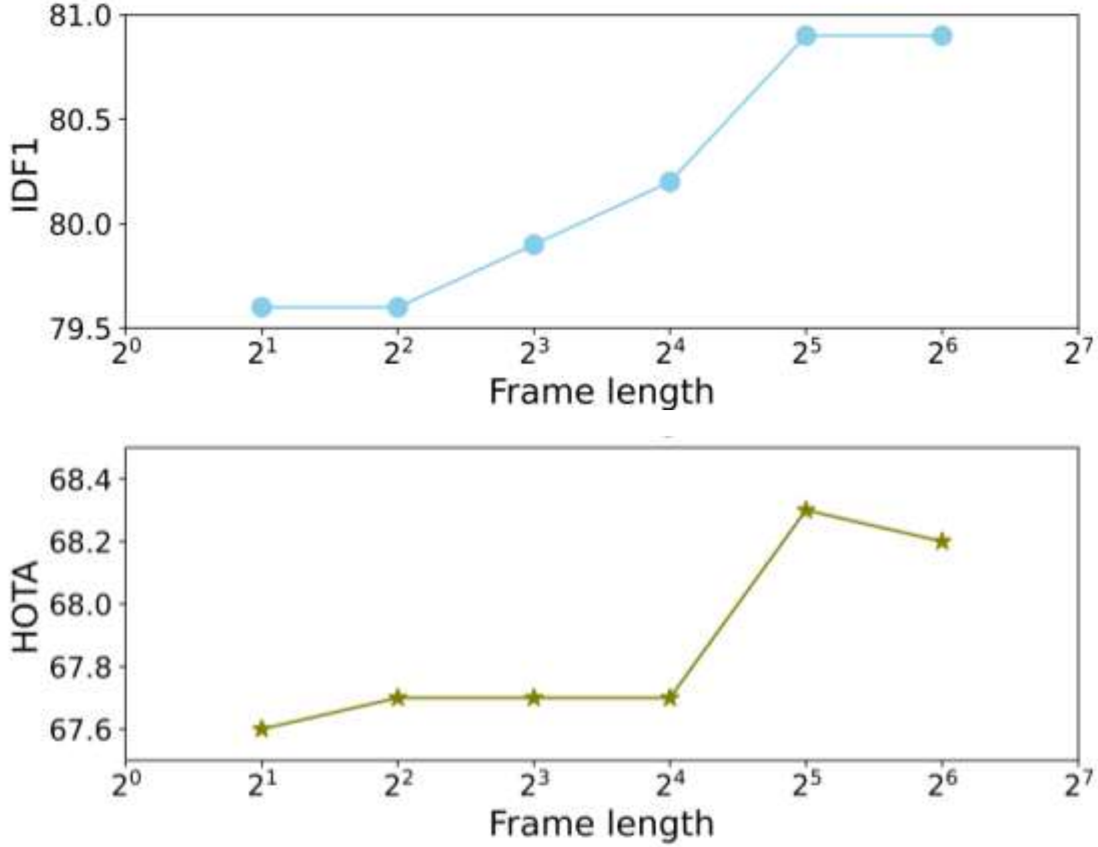
Experiments: Multiple Object Tracking

Exp 1: Comparison between streaming model and frame-based model on MOT17



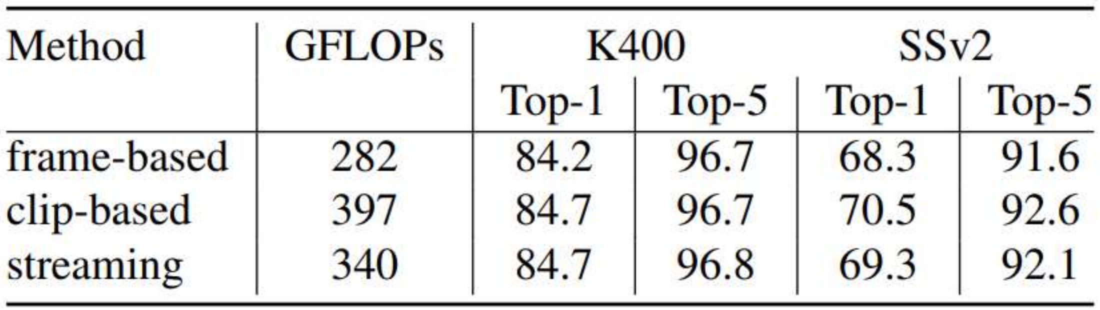
✓ **Temporal information MATTERs in MOT.**

Exp 2: Influence of test-time memory length on MOT17



Experiments: Video Action Recognition

Exp3: Comparison among streaming model, clip-based model, and frame-based model on action recognition



Method	GFLOPs	K400		SSv2	
		Top-1	Top-5	Top-1	Top-5
frame-based	282	84.2	96.7	68.3	91.6
clip-based	397	84.7	96.7	70.5	92.6
streaming	340	84.7	96.8	69.3	92.1

✓ **Streaming video model shows competitive performance on the sequence-based tasks.**

Conclusion

- we propose the idea of streaming video models that aim to unify the treatment of both frame-based and sequence-based video understanding tasks, which in the past were handled by separate models.
- We present an implementation named streaming video Transformer and conduct comprehensive experiments on multiple benchmarks.
- Experimental results demonstrate our proposed model achieves remarkable performance on both action recognition and multiple object tracking.
- To the best of our knowledge, our work is the first deep learning architecture that unifies video understanding tasks.

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

SCAN ME

Thank you for you listening. For more details, please visit:
https://arxiv.org/abs/2303.17228

Streaming Video Model

Microsoft

WED-PM-215