Tracking by Animation: Unsupervised Learning of Multi-Object Attentive Trackers

Zhen He[1,2,3] Jian Li[2] Daxue Liu[2] Hangen He[2] David Barber[3,4]

[1]Beijing Institute of Basic Medical Sciences
[2]National University of Defense Technology
[3]University College London
[4]The Alan Turing Institute

CVPR
LONG BEACH
CALIFORNIA
June 16-20, 2019

1. Introduction

Problems

- Existing multi-object tracking (MOT) approaches are usually based on supervised learning, while collecting training labels for videos is expensive.
- Most of the approaches consider the task as detection and tracking and solve them independently, usually leading to sub-optimal solutions.

Motivation

- Removing the need of using training labels.
- Building an end-to-end model to guarantee optimality.

Key Idea

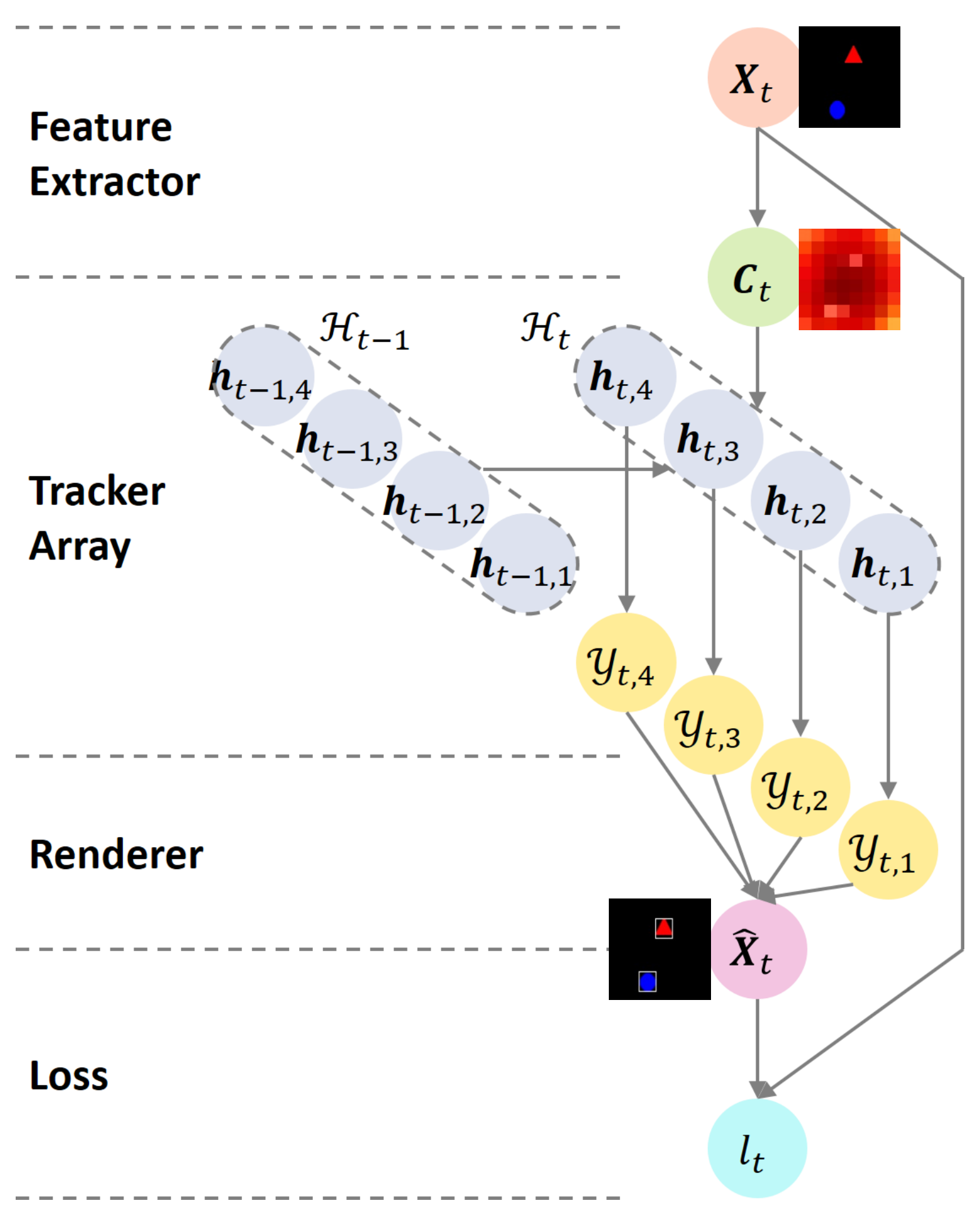
- Firstly, we use a recurrent neural model (a feature extractor and a tracker array) to extract (track) multiple objects $\mathcal{Y}_{t,1}, \mathcal{Y}_{t,2}, \ldots, \mathcal{Y}_{t,I}$ from each raw video frame $\boldsymbol{X}_t$.
- Each object $\mathcal{Y}_{t,i}$ is represented as a quintuple of confidence $y^c_{t,i}$, layer $\boldsymbol{y}^l_{t,i}$, pose $\boldsymbol{y}^p_{t,i}$, shape $\boldsymbol{Y}^s_{t,i}$, and appearance $\boldsymbol{Y}^a_{t,i}$.
- To train the model without object labels, we use a neural renderer to convert (animate) these extracted objects into a reconstructed frame $\widehat{\boldsymbol{X}}_t$, and then calculate the reconstruction error (loss) l_t between $\boldsymbol{X}_t$ and $\widehat{\boldsymbol{X}}_t$.
- By minimizing l_t, the recurrent neural model can learn to extract correct object representations which correspond to a correct reconstruction (i.e. $\widehat{\boldsymbol{X}}_t = \boldsymbol{X}_t$).

Contributions

- We proposed a Tracking-by-Animation (TBA) framework to achieve unsupervised end-to-end learning of MOT.
- We proposed a Reprioritized Attentive Tracking (RAT) method to achieve robust data association.
- We applied our model to real-world video surveillance tasks and showed its efficacy and practicality.

2. Method

Tracking by Animation (TBA)



Overview of the TBA framework, where the tracker number $I = 4$.

Reprioritized Attentive Tracking (RAT)

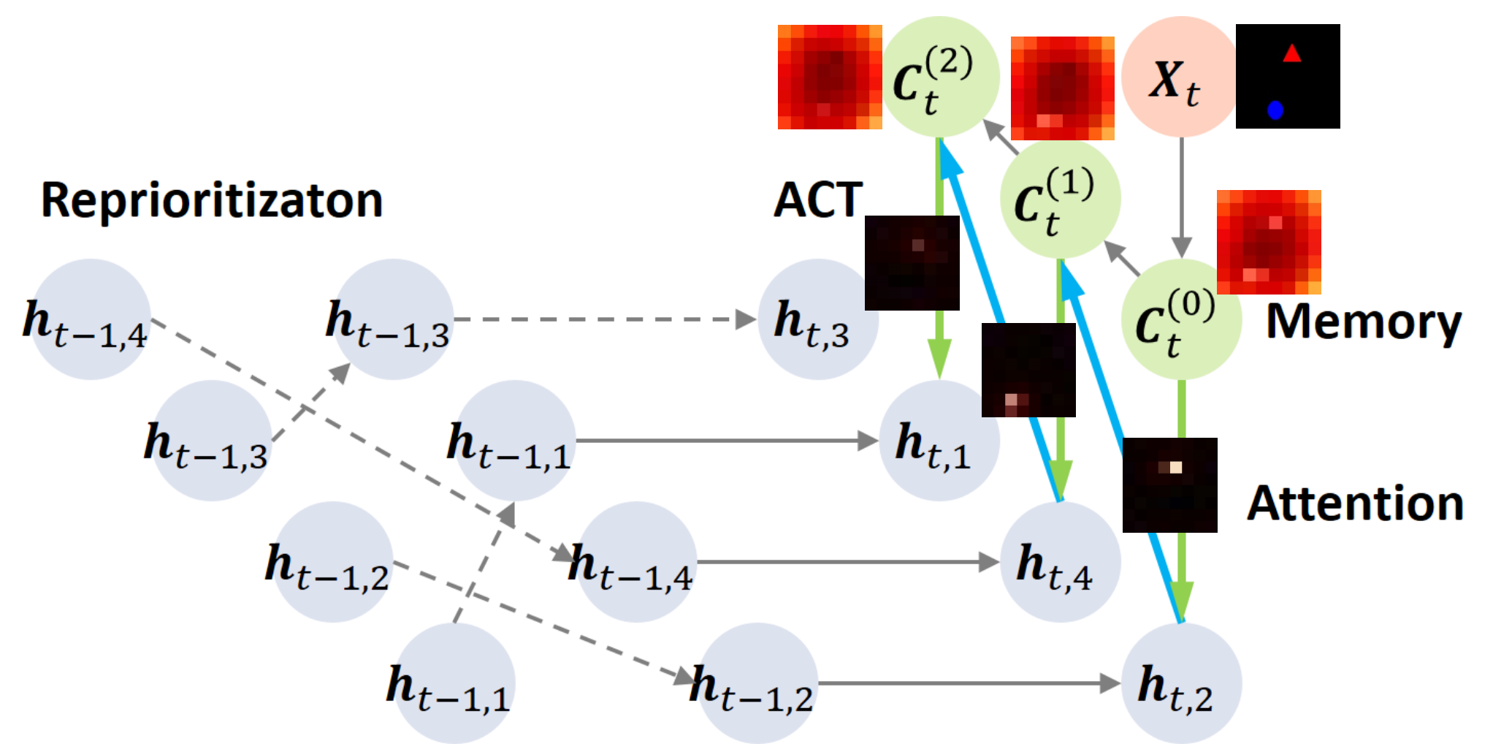


Illustration of the RAT with the tracker number $I = 4$. Green/Blue bold lines denote attentive read/write operations on memory. Dashed arrows denote copy operations. At time t, the iteration is performed by 3 times and terminated at Tracker 1.

Renderer

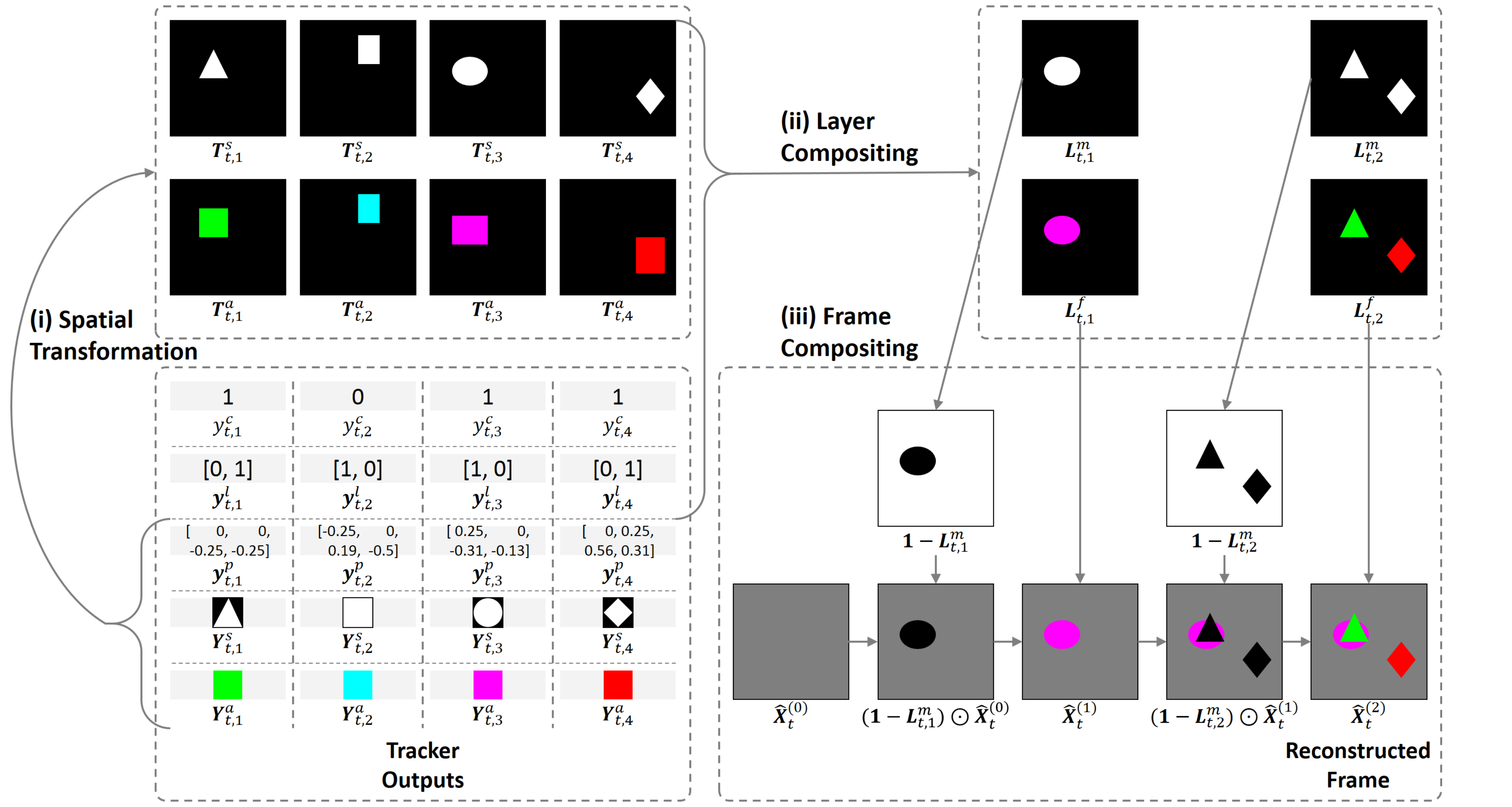
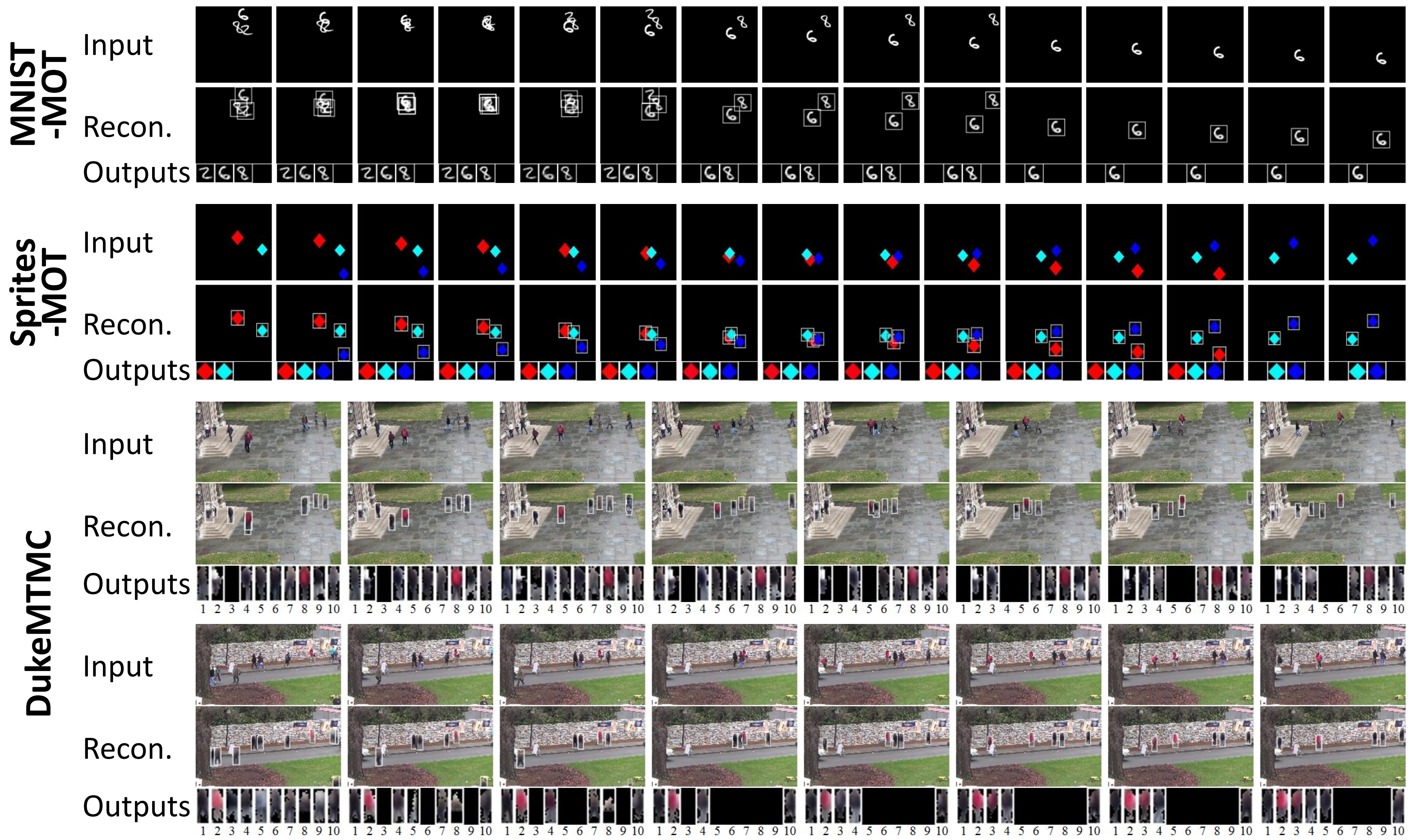


Illustration of the rendering process converting the tracker outputs into a reconstructed frame at time t, where the tracker number $I = 4$ and the layer number $K = 2$.

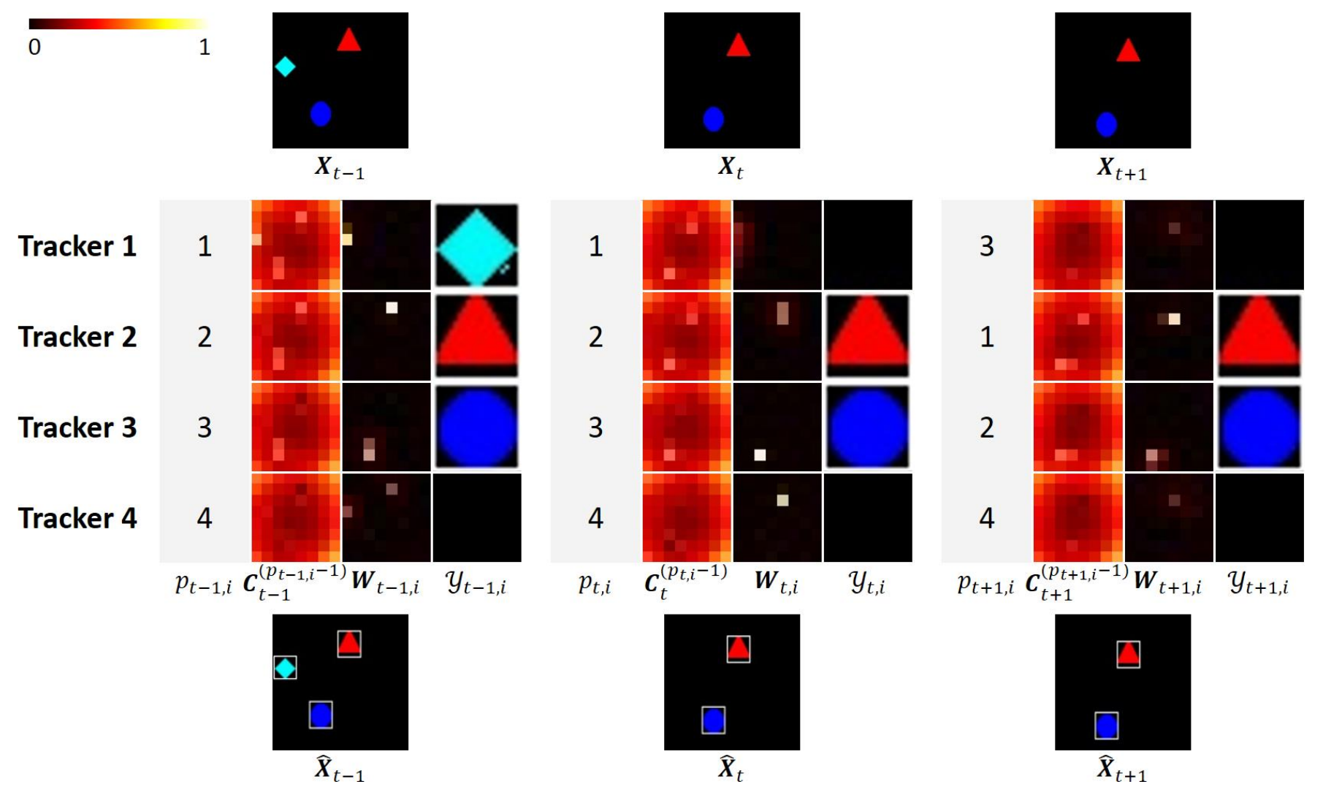
3. Experiments

Qualitative Results



Qualitative results of TBA. For each sequence, we show the input frames (top), reconstructed frames (middle), and the tracker outputs (bottom). For each frame, tracker outputs from left to right correspond to trackers 1 to I, respectively. Each tracker output $\mathcal{Y}_{t,i}$ is visualized as $(y^c_{t,i}\, \boldsymbol{Y}^s_{t,i} \odot \boldsymbol{Y}^a_{t,i}) \in [0,1]^{U \times V \times D}$.

Visualizing the RAT



Visualization of the RAT on Sprites-MOT. Both the memory $\boldsymbol{C}_t$ and the attention weight $\boldsymbol{W}_{t,i}$ are visualized as $M \times N$ (8× 8) matrices, where for $\boldsymbol{C}_t$ the matrix denotes its channel mean $\frac{1}{S}\sum_{s=1}^{S} \boldsymbol{C}_{t,1:M,1:N,s}$ normalized in $[0, 1]$.

Please scan the QR code or check our project page (https://github.com/zhen-he/tracking-by-animation) for video demos and the source code.