Tracking Everything Everywhere All at Once

Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li
Bharath Hariharan, Aleksander Holynski, Noah Snavely

ICCV 2023 (Oral)

Presented by: Soumitri Chattopadhyay

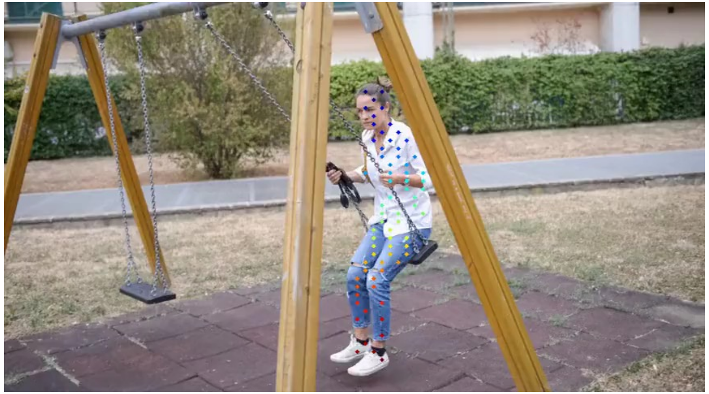
Some trivia to start with…

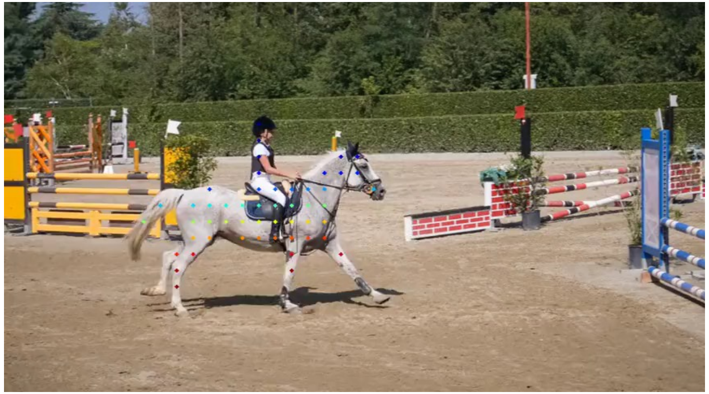
- Probable inspiration of the title of the paper, a 2022 film that won big in the last Academy Awards!

- But, we are not going to discuss movies…

In a nutshell...

- Track *every pixel* in each frame
    - Not simply object/semantic concepts; goes even finer to pixel-space
- *Globally consistent* motion representation
    - Pixel level motion trajectory across full video
- Can *handle occlusion* to a great extent
    - A limitation for several motion estimation methods
- Test-time optimization
    - Needs to be optimized per-video basis





Motivation

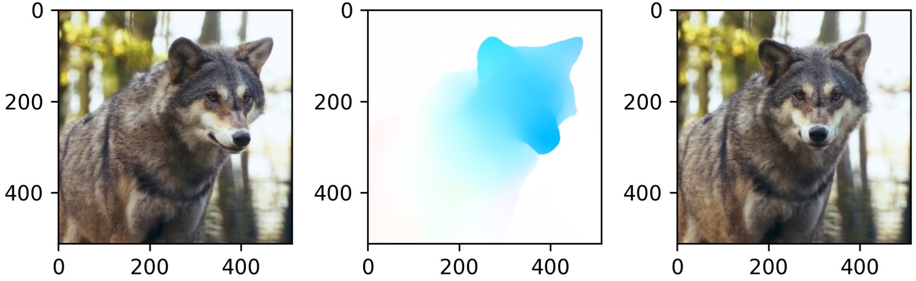
- Prior motion estimation methods fail at modeling motion of a video holistically
  - Sparse feature tracking models only certain key points/pixels which are easy to track (edges, corners, etc.)
  - Optical flow being a pairwise motion field is most effective for nearby/consecutive frames and hence works well only for smaller temporal windows
  - Chaining approaches that can compute multi-frame flow fields are prone to drift errors which get accumulated with each frame.

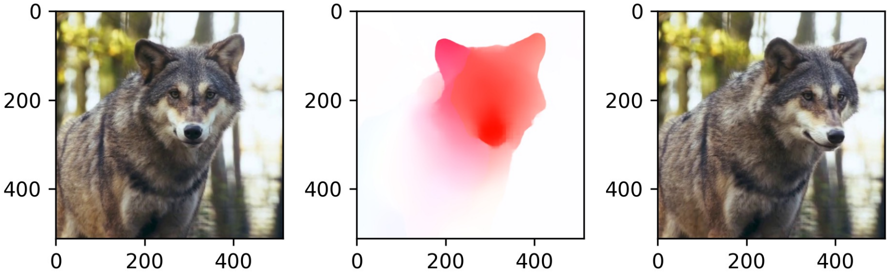
Challenges

- To maintain accurate trajectories across long video sequences
- Tracking points through episodes of occlusion
- Ensuring both spatial and temporal coherence (i.e. trajectories to be continuous and consistent in space)

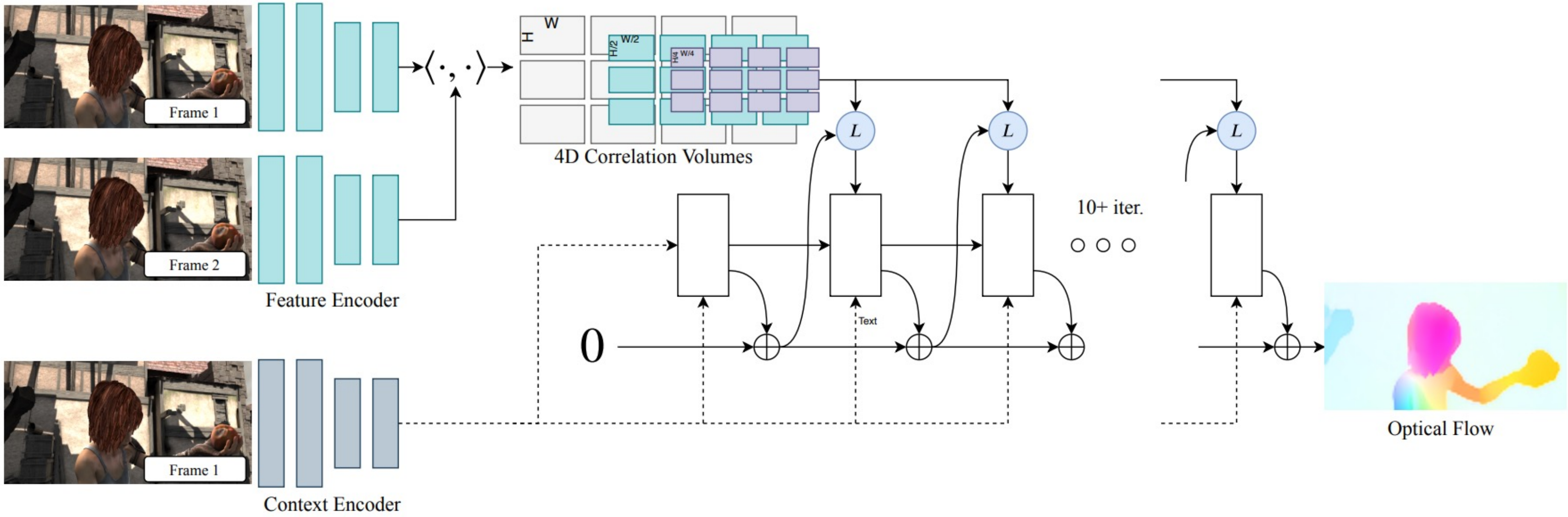
Background – Optical Flow

- It is a motion field comprising vectors at each spatial location that denote the direction of motion and also its magnitude, computed between a pair of images.
- Colour denotes direction of motion, while its intensity denotes the magnitude.





RAFT: A SoTA Optical Flow estimator



- Constructs a 4D correlation volume using inner products of all pairs of feature vectors.
- This is pooled at multiple scales to generate multiscale volumes.
- The optical flow is *recurrently* refined based on looking up values from the set of correlation volumes.

Image Credits: Zachary Teed and Jia Deng, RAFT: Recurrent All-pairs Field Transforms for Optical Flow, ECCV 2020

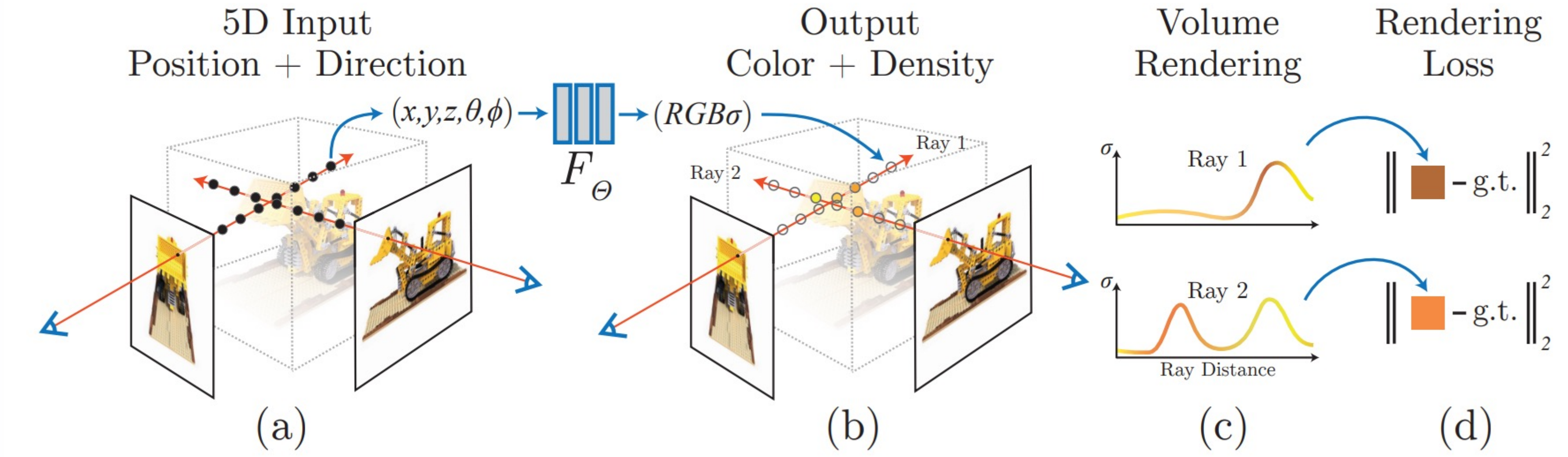
Methodology

- OmniMotion proposes a 3D mapping approach for motion estimation in videos
- The principle is based on bijective mapping of the pixels of a given frame to a global structure that preserves the motion behaviour across all frames

("Bijective" means a function in which each data point in its domain can be mapped to *exactly* one point in its co-domain, and for every point 'y' in its co-domain there exists atleast one 'x' such that f(x)=y)

- Specifically, each of the (local) frames are represented as a "quasi 3D" coordinate frame that are mapped to the global canonical 3D volume
- The bijections are parameterized by an invertible NN comprising affine transformations
- The mapping of local to global volumes and its subsequent optimization shares similarities with how NeRFs work.

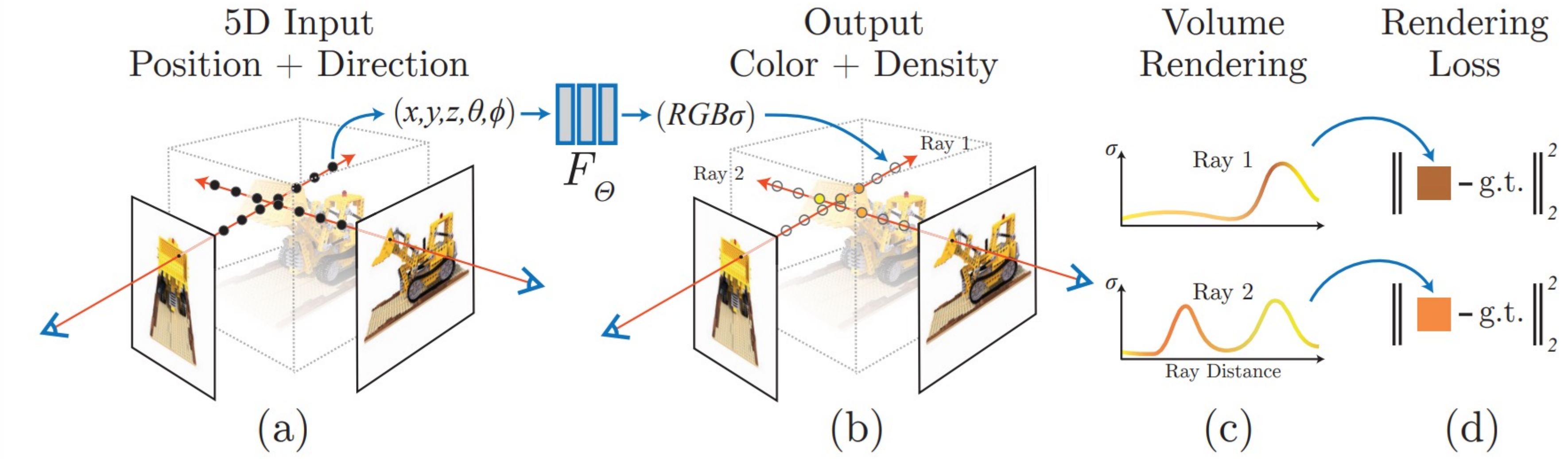
Background – Neural Radiance Fields (NeRFs)



- NeRFs are used for synthesis of novel views, given multiple views of a scene.
- A ray of light is passed through the given view, and points are sampled from its trajectory. Each point is denoted by spatial coordinates (x,y,z) and its viewing direction (θ,Φ), which is fed into an MLP to yield colour and volume density at that point.

Image Credits: Ben Mildenhall, and others, NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020

Background – Neural Radiance Fields (NeRFs)

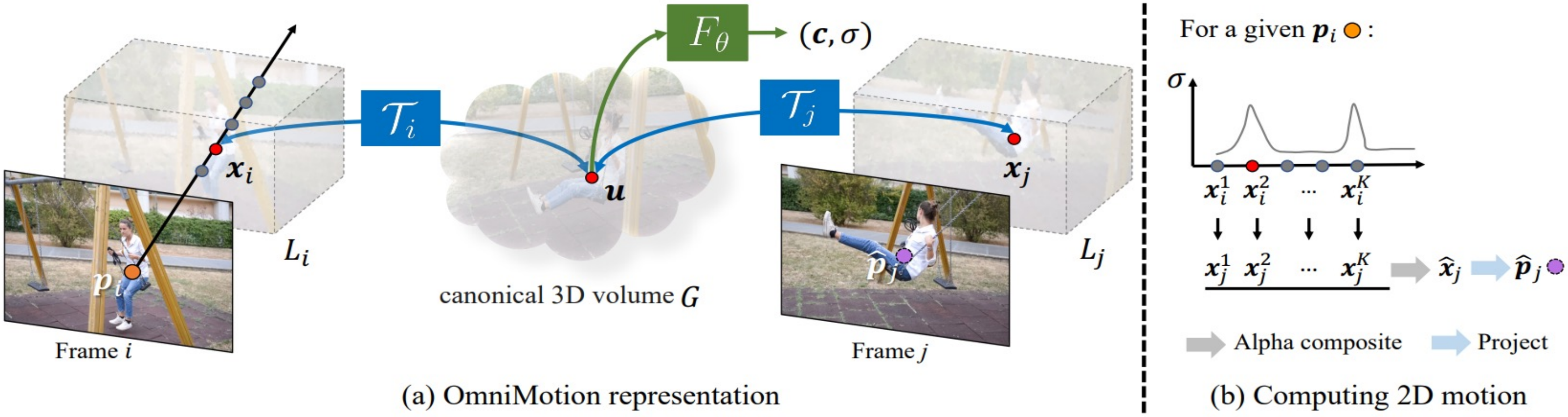


- Based on the density, volume rendering is performed in 3D space and the MLP F_θ is optimized using an L2 loss between GT and rendered volume.

- A similar principle is employed in OmniMotion too!

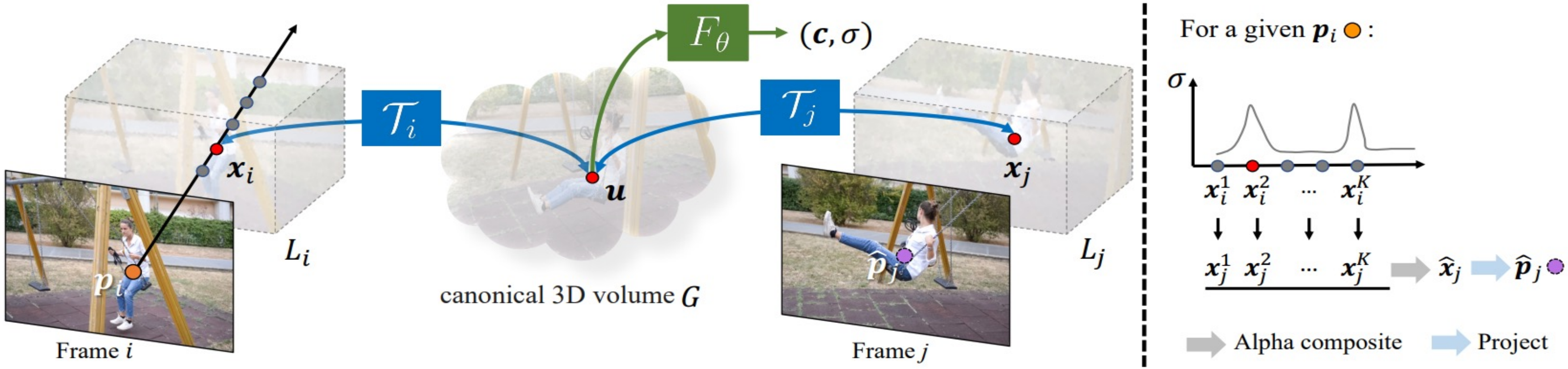
Image Credits: Ben Mildenhall, and others, NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020

Methodology (contd.)



- A ray of light is shot normally [0,0,1] through the source pixel p_i and points $\{x_i\}$ are sampled from it, which are mapped to u in the canonical volume G using the INN.
- Next, similar to NeRFs, the colour and density of u is estimated using F_θ, which is a 3-layer GaborNet MLP.

Methodology (contd.)



(a) OmniMotion representation

(b) Computing 2D motion

- Now taking inverse of the INN we can obtain $\{x_j\}$ from $\{x_i\}$.
- Thus, we essentially obtain a volume density in G corresponding to each point in $\{x_i\}$ which correspond one-one to $\{x_j\}$.
- The set of points $\{x_j\}$ are alpha composited and projected onto target pixel p_j (corresponding to queried location p_i).

Training Objectives

- Flow loss
  - Primary loss function
  - Minimizes MAE between predicted flow and pseudo-GT flow (from RAFT)

$$\mathcal{L}_{\text{flo}} = \sum_{\boldsymbol{f}_{i\to j}\in\Omega_f} ||\hat{\boldsymbol{f}}_{i\to j} - \boldsymbol{f}_{i\to j}||_1$$

- Photometric loss
  - Enforces color consistency between target (predicted) and source pixels using MSE loss

$$\mathcal{L}_{\text{pho}} = \sum_{(i,\boldsymbol{p})\in\Omega_p} ||\hat{\boldsymbol{C}}_i(\boldsymbol{p}) - \boldsymbol{C}_i(\boldsymbol{p})||_2^2$$

- Regularization loss
  - Penalizes large accelerations between adjacent frame pixels to ensure temporal smoothness

$$\mathcal{L}_{\text{reg}} = \sum_{(i,\boldsymbol{x})\in\Omega_x} ||\boldsymbol{x}_{i+1} + \boldsymbol{x}_{i-1} - 2\boldsymbol{x}_i||_1$$

- Auxiliary photometric loss (NOT in paper!)
  - Pairwise photometric loss computed using randomly sampled pixels (not adjacent pixels)

$$\mathcal{L}_{\text{pgrad}} = \sum_{\Omega_p} ||(\hat{\boldsymbol{C}}_i(\boldsymbol{p}_1) - \hat{\boldsymbol{C}}_i(\boldsymbol{p}_2)) - (\boldsymbol{C}_i(\boldsymbol{p}_1) - \boldsymbol{C}_i(\boldsymbol{p}_2))||_1$$

Implementation

- Hard sample mining
    - ○ Pairwise flows can be notoriously noisy, especially when they are temporally far apart (large displacement)
    - ○ A naïve solution would be to filter out the erroneous flows
    - ○ This, however, would create an imbalance in the sample distribution, since only simple motions whose flows are highly reliable would remain
    - ○ Instead, hard sample mining is adopted where flows are cached periodically and those with high errors are sampled more frequently.
    - ○ Error maps are computed using Euclidean distance, and they are employed to adjacent frames where flow fields are highly reliable

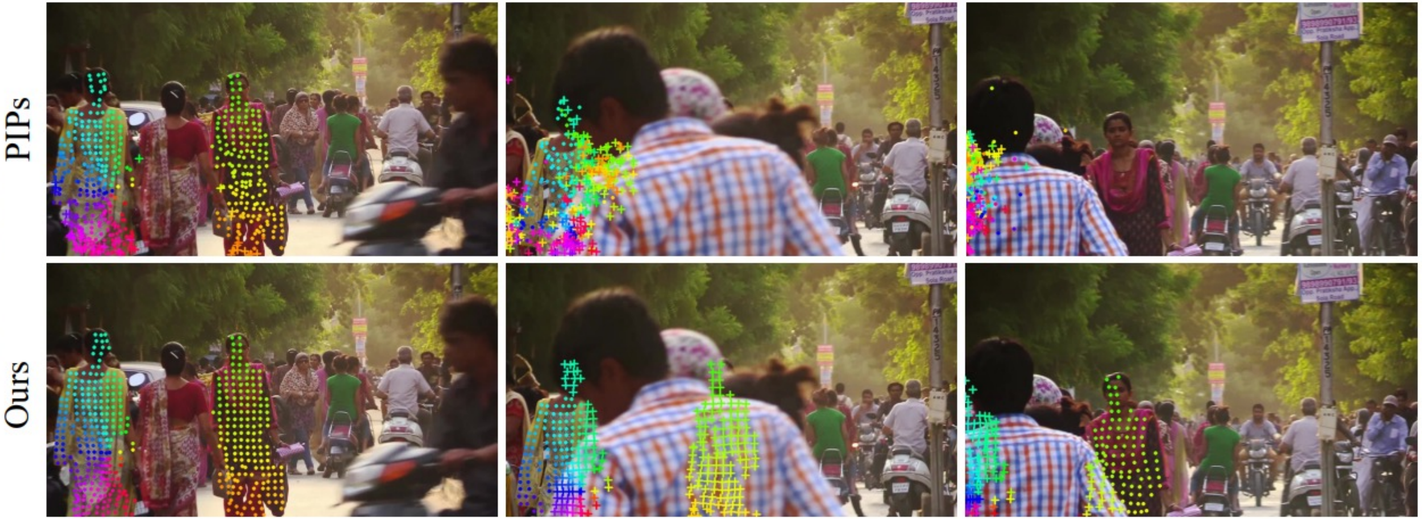
- Training configuration
    - ○ Adam, 200k iterations
    - ○ K=32 points sampled from each ray
    - ○ 256 pairs of correspondences across 8 pairs of frames

- Datasets
    - ○ DAVIS, Kinetics, RGB-Stacking

Qualitative Results

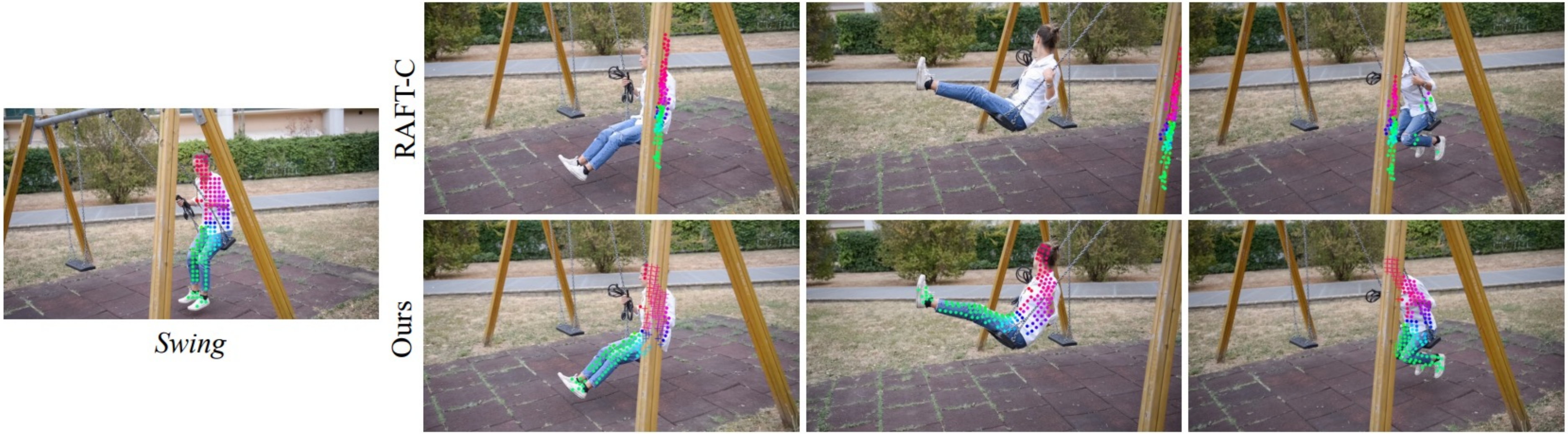
- Effectively tracks occluded pixels in scenes

India

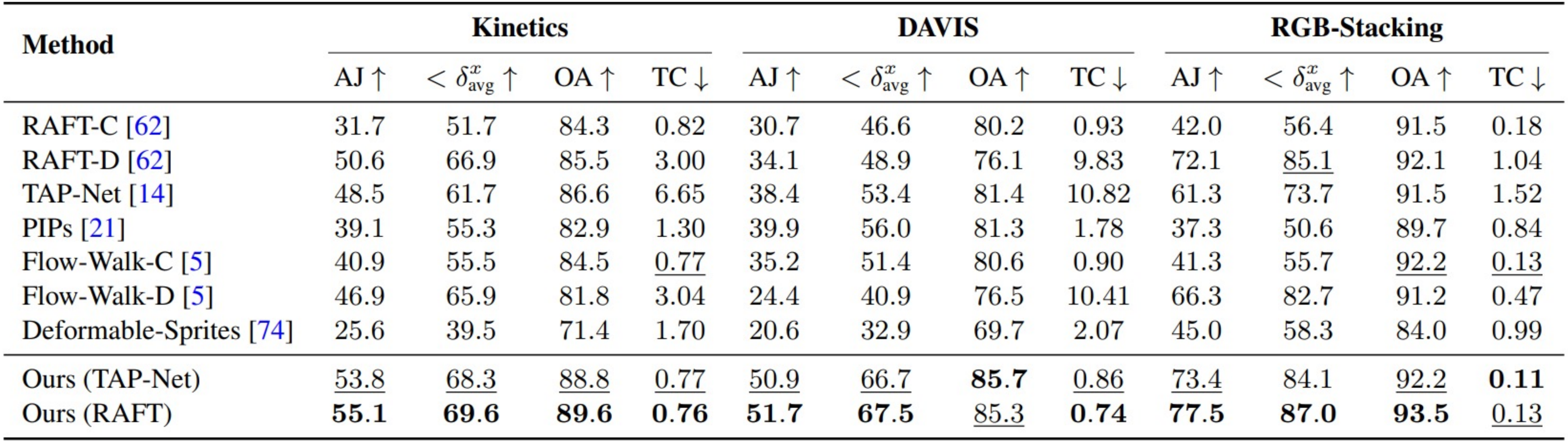


Qualitative Results

- Effectively tracks occluded pixels in scenes



Quantitative Results – SoTA comparison

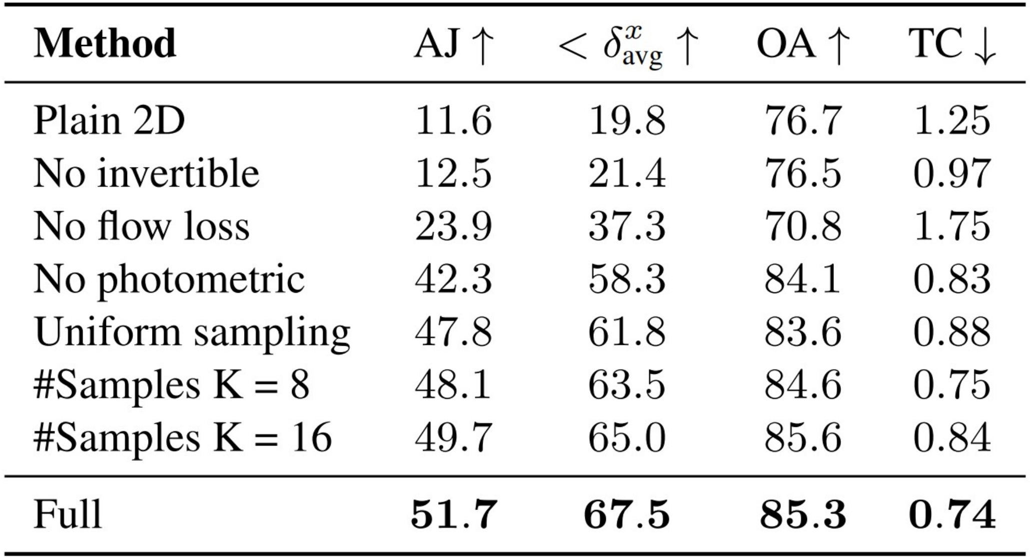


Method	Kinetics				DAVIS				RGB-Stacking			
	AJ ↑	$< \delta^x_{\text{avg}}$ ↑	OA ↑	TC ↓	AJ ↑	$< \delta^x_{\text{avg}}$ ↑	OA ↑	TC ↓	AJ ↑	$< \delta^x_{\text{avg}}$ ↑	OA ↑	TC ↓
RAFT-C [62]	31.7	51.7	84.3	0.82	30.7	46.6	80.2	0.93	42.0	56.4	91.5	0.18
RAFT-D [62]	50.6	66.9	85.5	3.00	34.1	48.9	76.1	9.83	72.1	85.1	92.1	1.04
TAP-Net [14]	48.5	61.7	86.6	6.65	38.4	53.4	81.4	10.82	61.3	73.7	91.5	1.52
PIPs [21]	39.1	55.3	82.9	1.30	39.9	56.0	81.3	1.78	37.3	50.6	89.7	0.84
Flow-Walk-C [5]	40.9	55.5	84.5	0.77	35.2	51.4	80.6	0.90	41.3	55.7	92.2	0.13
Flow-Walk-D [5]	46.9	65.9	81.8	3.04	24.4	40.9	76.5	10.41	66.3	82.7	91.2	0.47
Deformable-Sprites [74]	25.6	39.5	71.4	1.70	20.6	32.9	69.7	2.07	45.0	58.3	84.0	0.99
Ours (TAP-Net)	53.8	68.3	88.8	0.77	50.9	66.7	**85.7**	0.86	73.4	84.1	92.2	**0.11**
Ours (RAFT)	**55.1**	**69.6**	**89.6**	**0.76**	**51.7**	**67.5**	85.3	**0.74**	**77.5**	**87.0**	**93.5**	0.13

- OmniMotion essentially builds on SoTA optical flow baselines (RAFT/TapNet) and considerably improves upon their performances on Kinetics and DAVIS
- TC improvement is reported; but it is almost directly related to the explicit regularization loss term
- Gains on some metrics are small, especially in RGB-Stacking

Quantitative Results – Ablation study

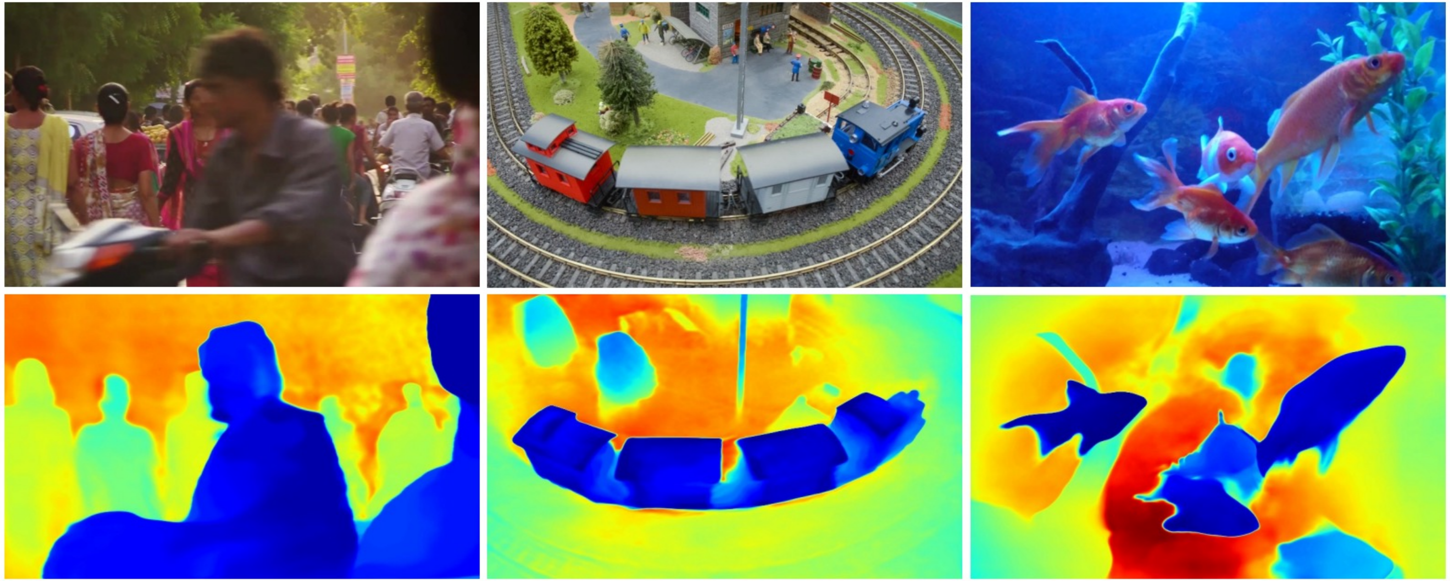
- Ablation study done on DAVIS dataset
- As expected, absence of flow loss greatly hurts temporal coherence/occlusion.
- Hard mining is essential to ensure a proper distribution of motion samples; uniform sampling reduces temporal coherence.



Method	AJ ↑	$< \delta^x_{\text{avg}} \uparrow$	OA ↑	TC ↓
Plain 2D	11.6	19.8	76.7	1.25
No invertible	12.5	21.4	76.5	0.97
No flow loss	23.9	37.3	70.8	1.75
No photometric	42.3	58.3	84.1	0.83
Uniform sampling	47.8	61.8	83.6	0.88
#Samples K = 8	48.1	63.5	84.6	0.75
#Samples K = 16	49.7	65.0	85.6	0.84
Full	**51.7**	**67.5**	**85.3**	**0.74**

Further Analysis

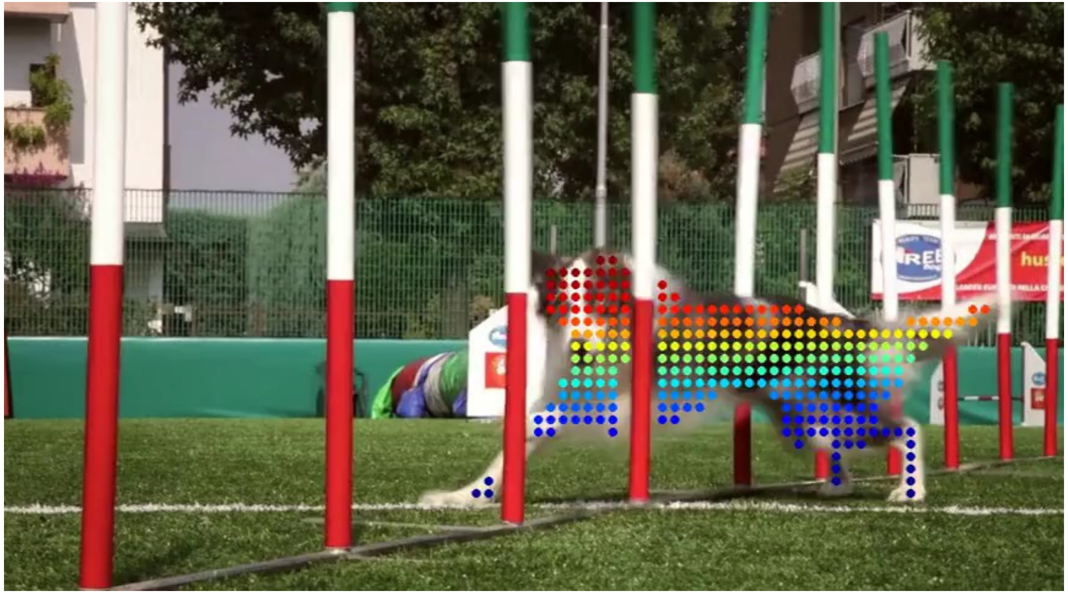
- Pseudo-depth maps generated by the model shows high quality separation of foreground and background
- This relative ordering of surfaces is crucial for tackling occlusion



Failure Cases

- Fails to track pixel trajectories in very fast motion videos

  This may be due to the regularization loss in the training objective, since it explicitly penalizes high magnitude displacements in adjacent frames, which is likely to be present in fast motion videos



Final thoughts…

- The bijective mapping from a local quasi-3D space to a global canonical 3D volume is very interesting and may be extended to motion editing/video generation tasks

- Qualitative results depicted show a lot of promise, while the quantitative gains over SoTA are non-trivial

- A major limitation however is – it needs to be run separately on each video
    - the training time is roughly 9 hours on a single A6000 GPU
    - This hinders the scalability of OmniMotion

- It is yet to be tested on longer videos, which would aptly justify its proposal

- Ethical concern regarding the fourth loss term in the training objective (?!)

Thank you!