Exploiting Motion and Topology for Segmenting and tracking under Entanglement

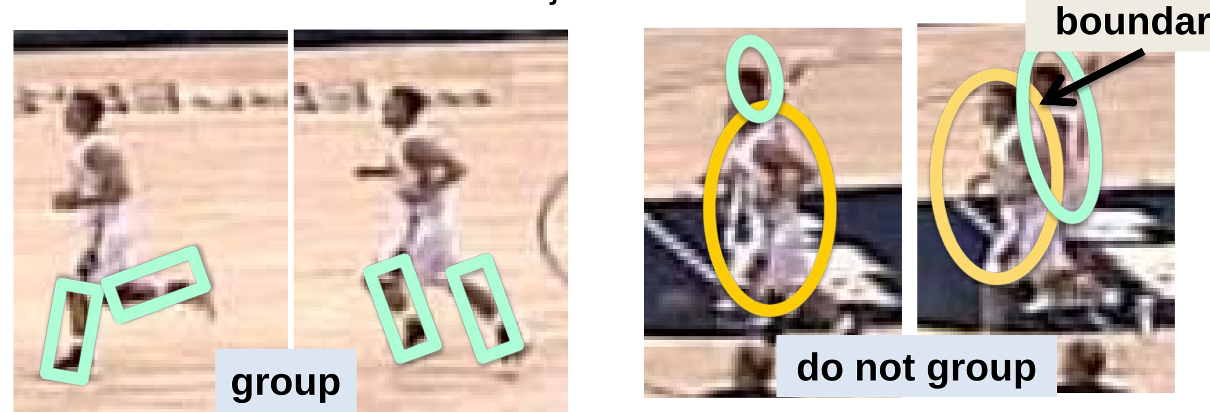
Katerina Fragkiadaki and Jianbo Shi

Penn University of Pennsylvania

CVPR 11 20 COLORADO 11

Problem

We want to segment interacting articulated bodies in videos.
• Motion is insufficient under object deformation / articulation.

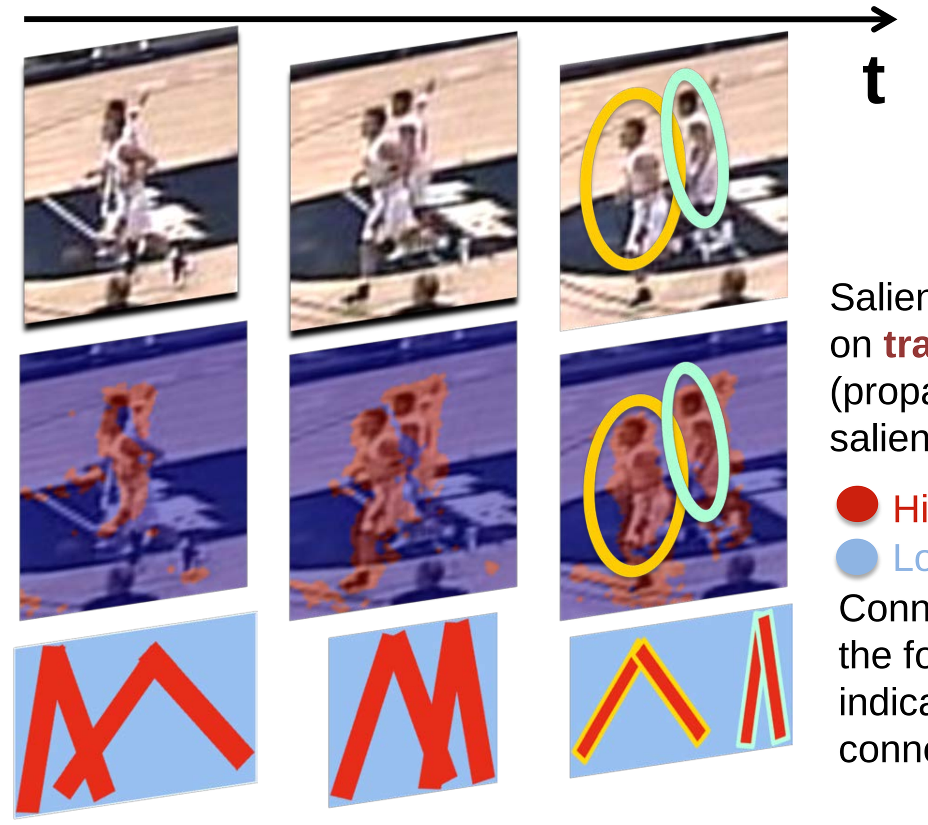


• Static image cues are often faint and unreliable.

Our contributions:

1. *Object connectedness* in large temporal context as complementary to motion for video segmentation. We attach connectedness constraints on pixel trajectories and gain large temporal support for their effective application.
2. *Trajectory saliency* for time consistent figure-ground segmentation that determines per frame object connectedness.

Object connectedness



Saliency maps based on **trajectory saliency** (propagating per frame saliency in time)

High saliency
Low saliency

Conn. components of the foreground maps indicate *per frame* connectedness.

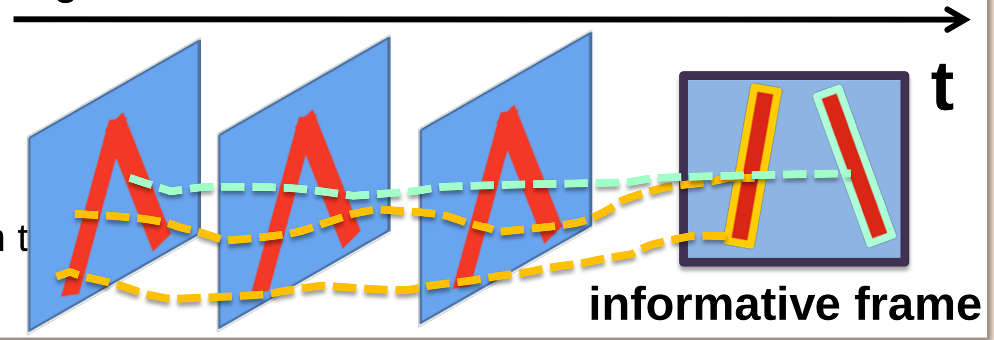
No disconnection: 1 body or 2 interacting agents?

Disconnection! 2 separate agents

Connectedness in time

Informative cues for correctly segmenting a video are **not uniformly distributed among video frames**.

Trajectories propagate connectedness across frames, from separation to entanglement.



Segmenting with Motion and Topology

We define trajecory $\mathrm{tr}^i = \{p_t^i\}, \quad i = 1 \cdots \mathcal{T}$, where p_t^i the pixel of tr^i at time t. We want to cluster trajectories into groups $C_\ell, \quad \ell = 1 \cdots K$. Our cues:

Attractions $\mathbf{A_{ij}}$ between similarly moving trajectories:

$\mathbf{A_{ij}} = \exp(-\frac{D_{ij}}{\sigma}), \quad D_{ij} = \bar{d} \cdot \max_{t \in t_1 \cdots t_2} ||\vec{u}_t^i - \vec{u}_t^j||^2$, $\bar{d}$: mean eucleidian distance.

Repulsions $\mathbf{R_{ij}}$ between trajectories of different connected components (CC) of any foreground frame map:

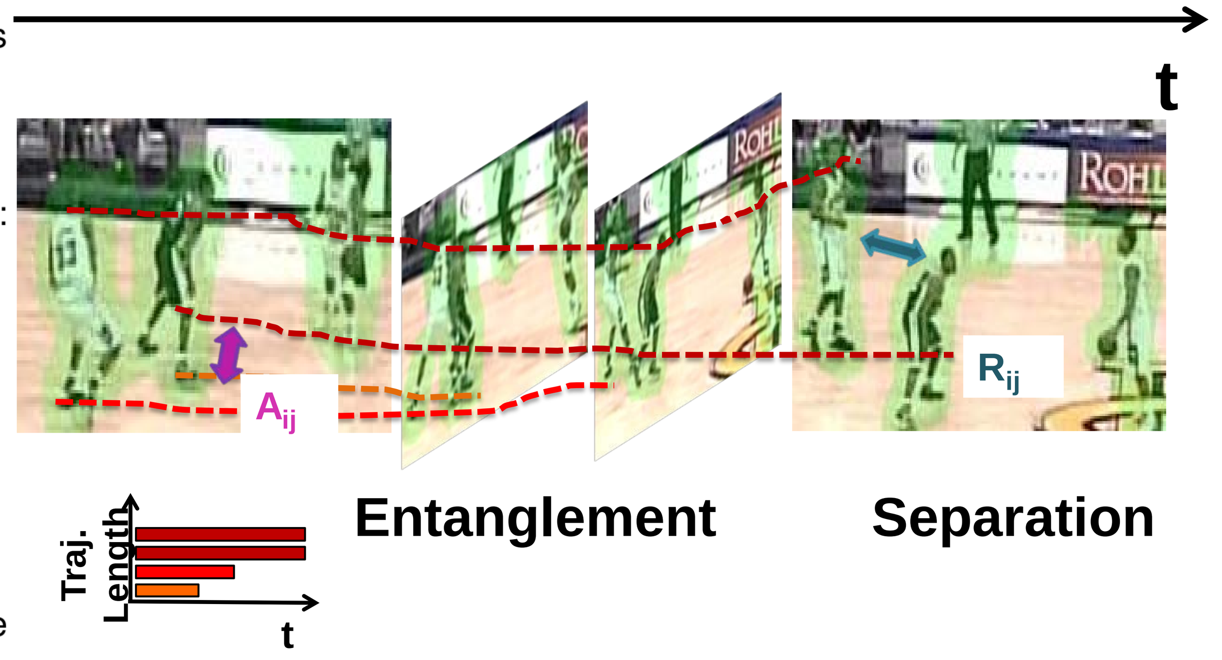
$$\mathbf{R_{ij}} = \begin{cases} 1 & \text{if } \exists\, t \text{ s.t. } CC(\mathrm{tr}^i) \neq CC(\mathrm{tr}^j) \\ 0 & \text{otherwise} \end{cases}$$

Our cost function maximizes within-group normalized attraction and between-group normalized repulsion:

$$\underset{X}{\text{maximize}} \quad \epsilon(X) = \frac{1}{K} \sum_{\ell=1}^{K} \frac{x_l^T \left(\mathbf{A} + \mathbf{D^R} - \mathbf{R}\right) x_l}{x_l^T \left(\mathbf{D^A} + \mathbf{D^R}\right) x_l}$$

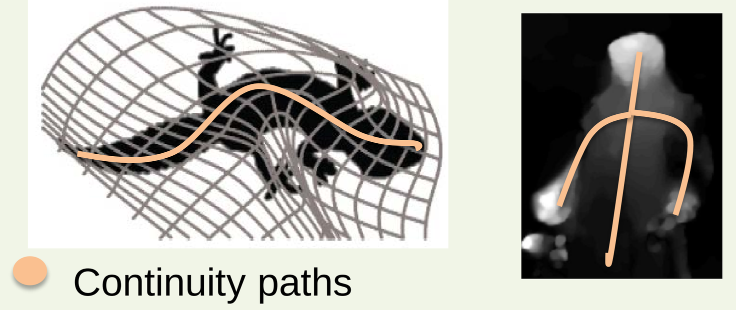
$$\text{subject to} \quad X \mathbf{1}_K = \mathbf{1}_{|\mathcal{T}|}$$

where x_ℓ is the binary indicator for group C_ℓ, $X = [x_1 \cdots x_K]$ the partition matrix and $\mathbf{D}^{\mathbf{A}}_{i,i} = \sum_j \mathbf{A}_{i,j}$ the degree matrix.



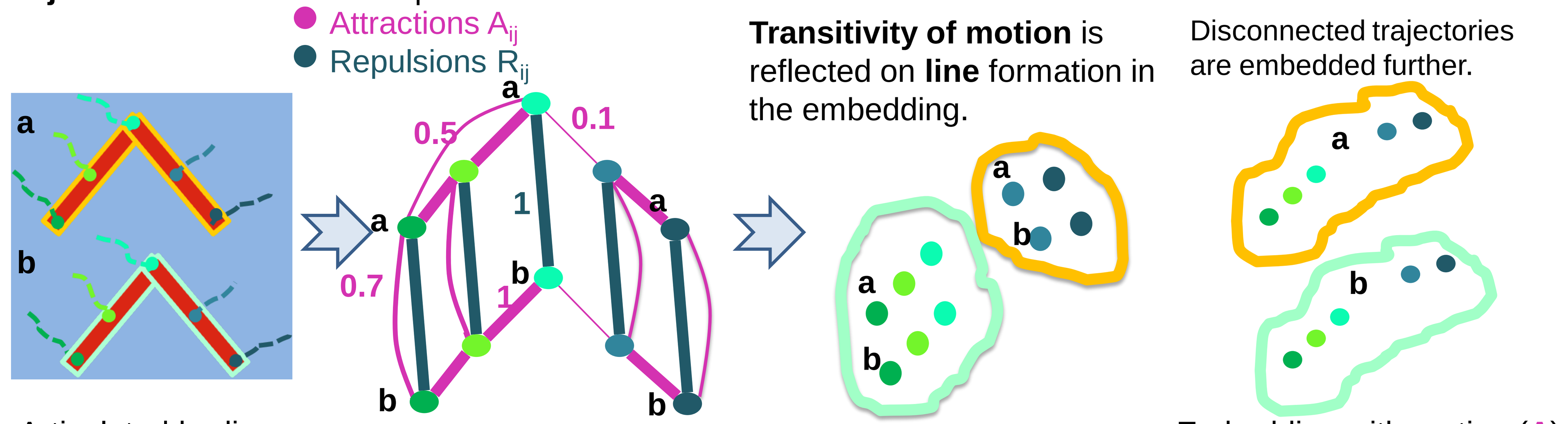
Object Deformation

Optical flow is not constant across object surfaces.



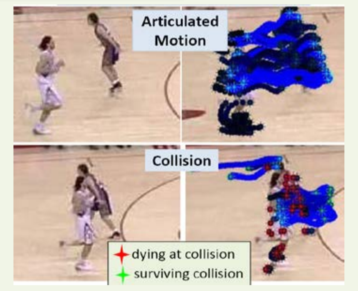
Smooth variation along deforming objects, motion **discontinuities** at joints. Model based clustering makes assumptions about data distributions (e.g. k-means assumes unimodal clusters)

Non parametric clustering based on normalized cut exploits transitivity from continuity of optical flow.
Object connectedness compensates for motion discontinuities.



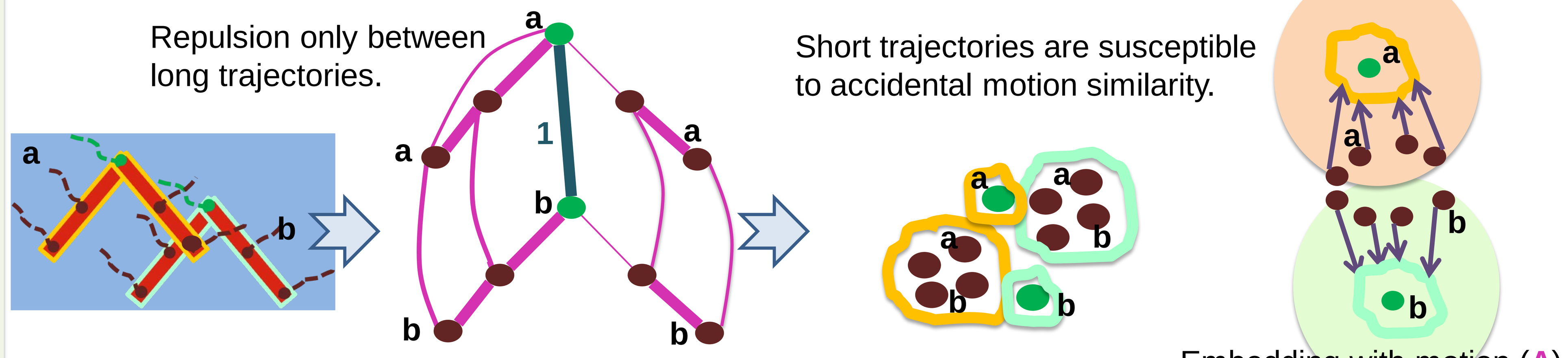
Trajectory asymmetry

Trajectories on articulated bodies vary in length due to self occlusions, extreme deformations and collisions.



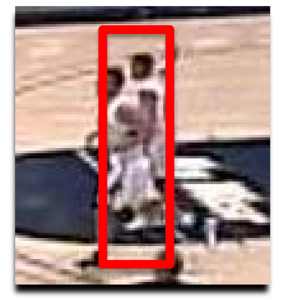
Longer trajectories live longer to see the informative splits between objects and propagate them in time.

Two step clustering: 1.Clustering of long trajectories provides the skeleton of the video scene.
2. Assignment of short trajectories to 'long' clusters based on embedding affinity.



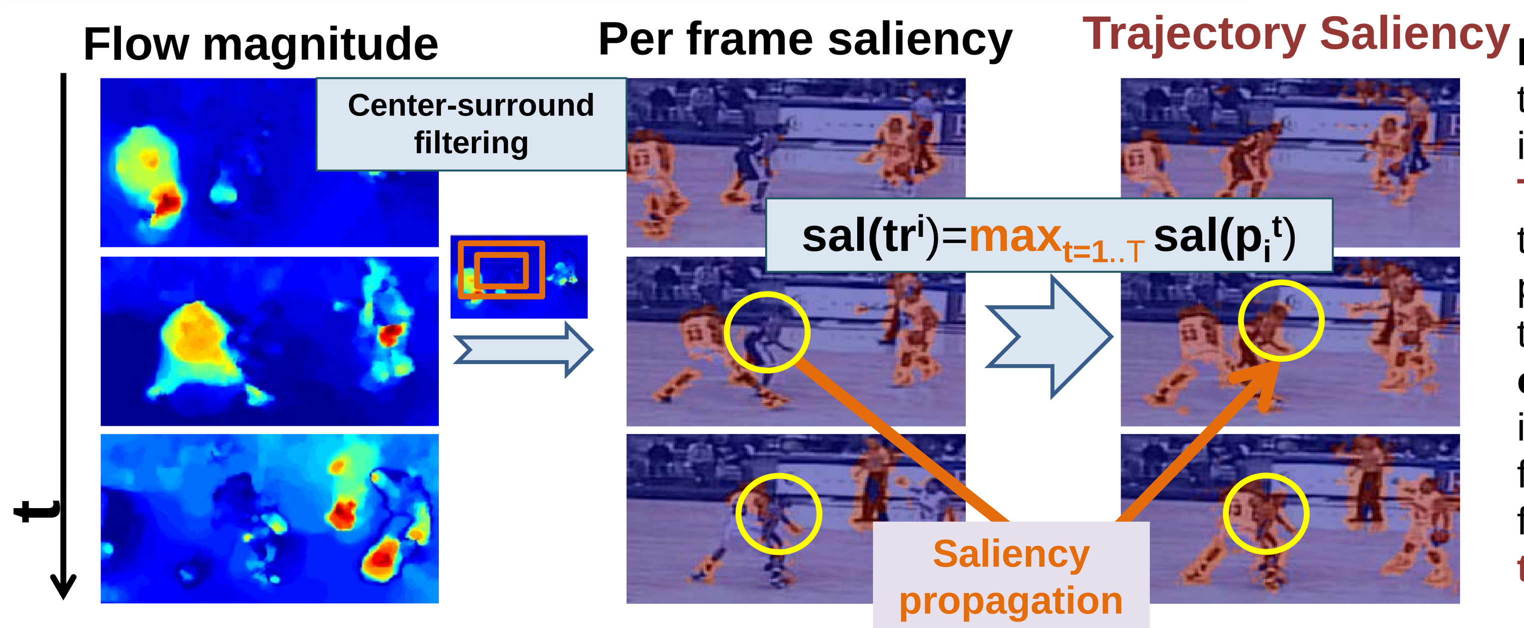
Occlusions

Partial occlusions cause problems to detectors that often fire in between the overlapping objects. Bounding boxes cover both bodies and the features extracted leaking across agents can easily cause drifting in detection based tracking.



Our resulting trajectory tracklets correctly segment overlapping, interacting objects during partial occlusions.

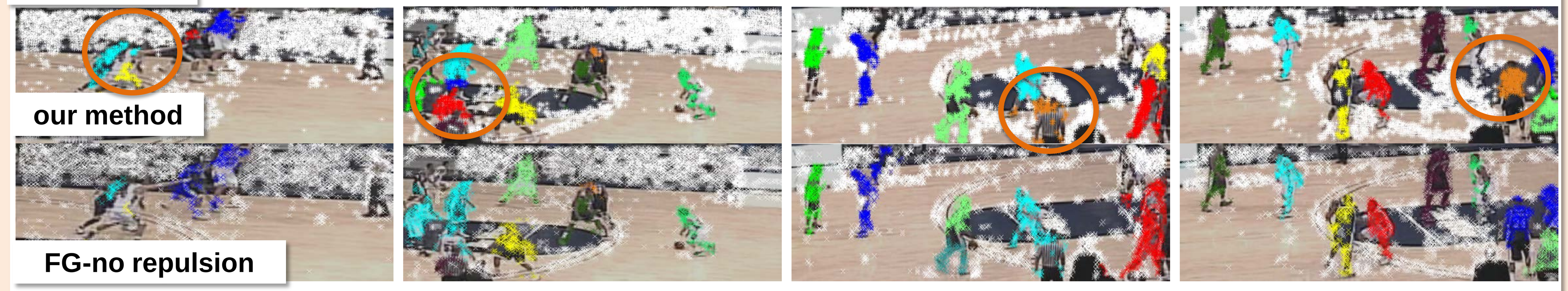
Figure-ground from Trajectory Saliency



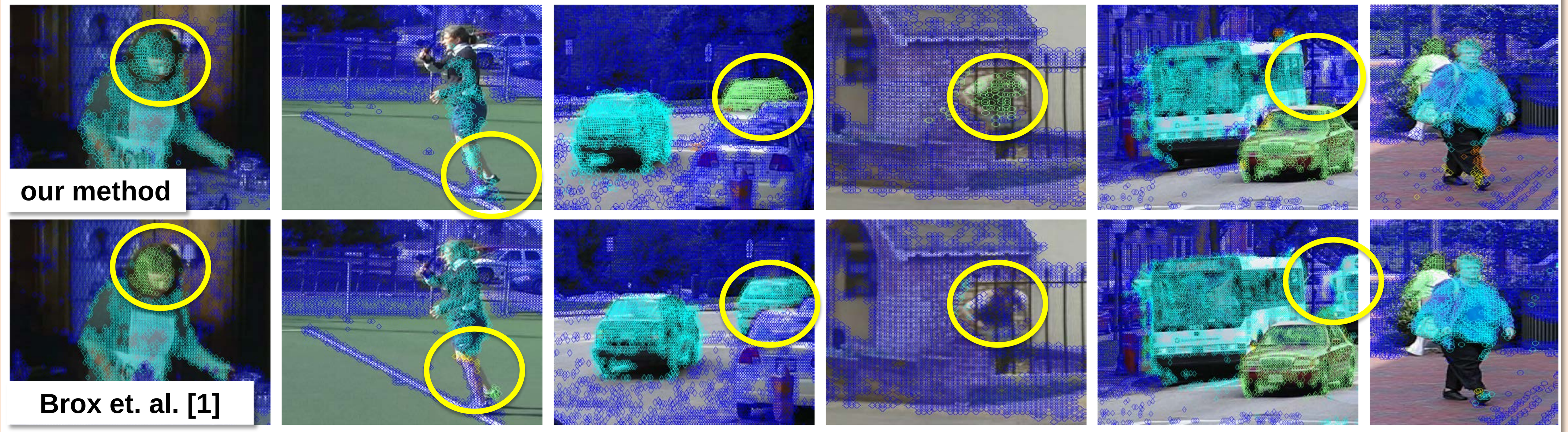
Pixel saliency: the local contrast of the motion of a pixel in comparison to its surroundings

Trajectory saliency: the maximum of the saliency values of the trajectory points. Trajectory saliency in contrast to pixel saliency offers **time consistent saliency** results: an agent is found salient even if he moves only for a subset of frames. We compute figure-ground by threshholding **trajectories saliencies.**

Results



	density	clustering error	per region clustering error	over-segmentation	recall	leakage	tracking time
our method	5.21%	**4.73%**	**20.32%**	1.57	**31.07%**	**16.52%**	**75.13%**
FG-$\tilde{r}$ - $\widetilde{asym}$	4.43%	11.13%	33.63%	1.29	20.41%	23.57%	50.77%
FG-r-$\widetilde{asym}$	3.28%	5.12 %	26.24%	2.07	18.89%	21.16%	46.63%
FG-$\tilde{r}$-asym	**5.57%**	12.91%	31.32%	1.36	26.95%	21.16%	65.79%
Brox et al. [ ]	0.57%	20.74%	86.43%	**0**	0.46 %	81.55%	1.03%



	density	clustering error	per region clustering error	over-segmentation	extracted objects
our method	3.22%	3.76%	**22.06%**	1.15	25
Brox et al. [1]	**3.32%**	**3.43%**	27.06%	**0.4**	**26**

References

1. Object segmentation by long term analysis of point trajectories. T. Brox, J. Malik ECCV 2010
2. Understanding popout through repulsion. Stella X. Yu and Jianbo Shi, CVPR 2001