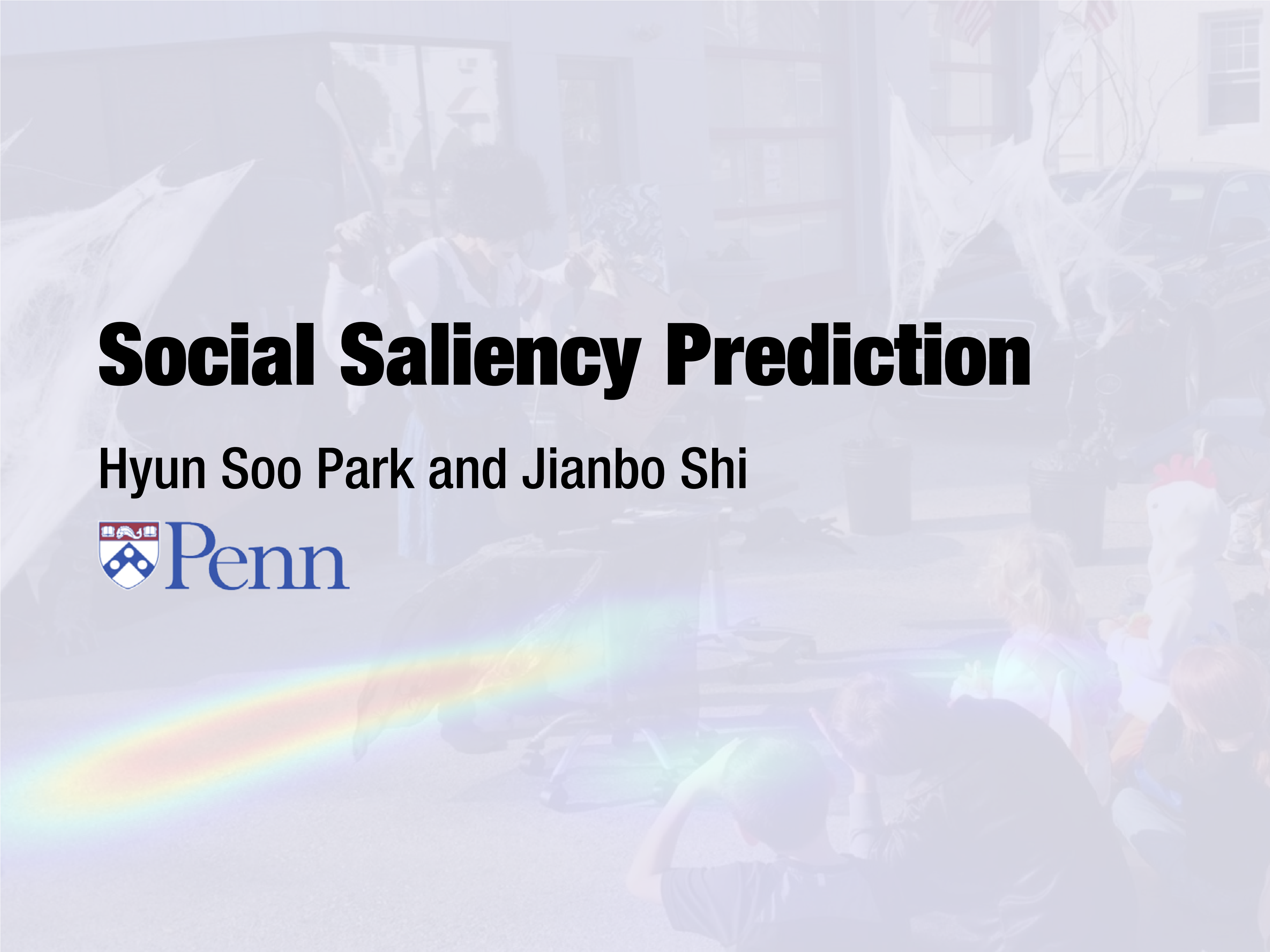
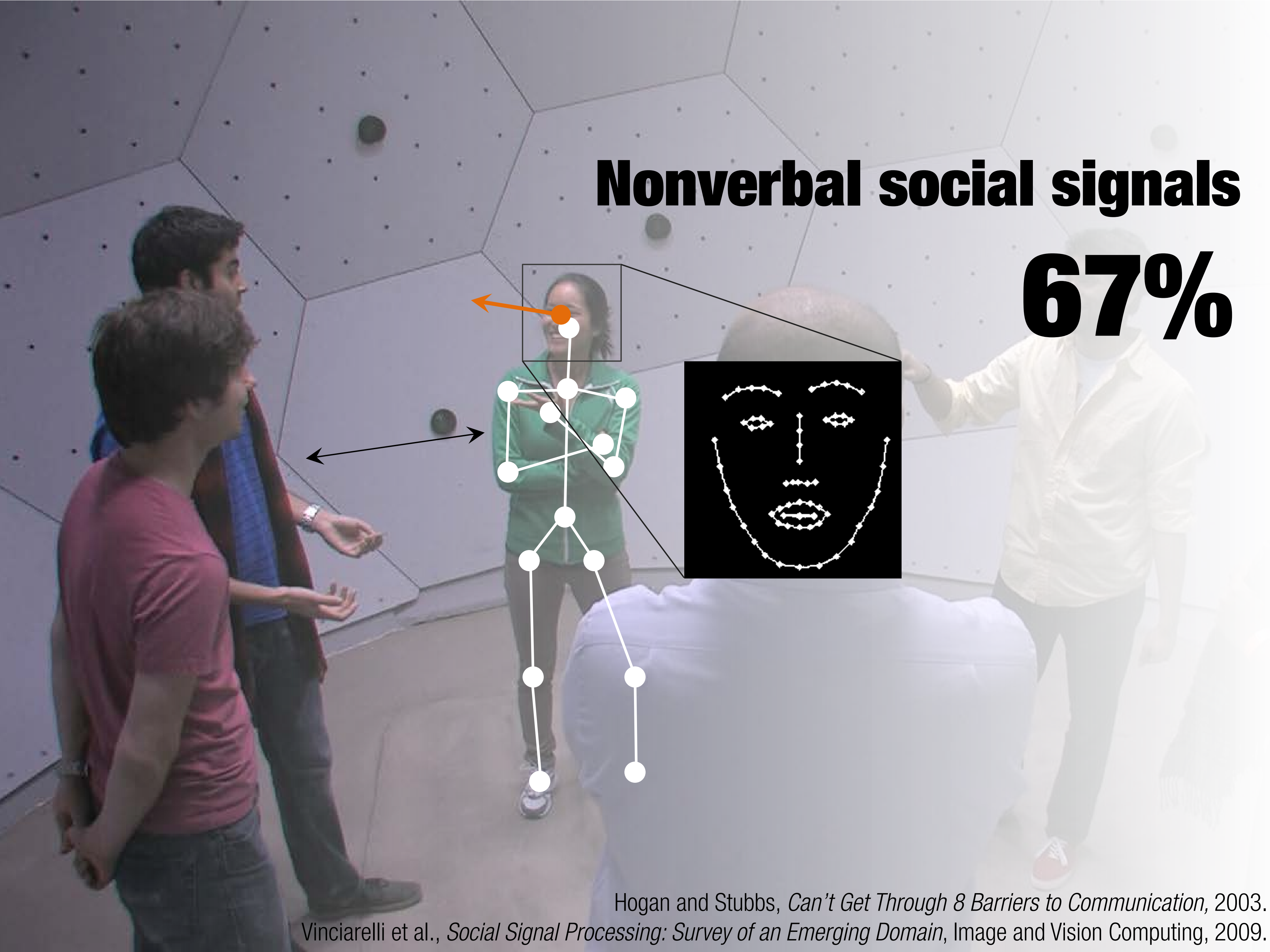


Social Saliency Prediction

Hyun Soo Park and Jianbo Shi

Penn



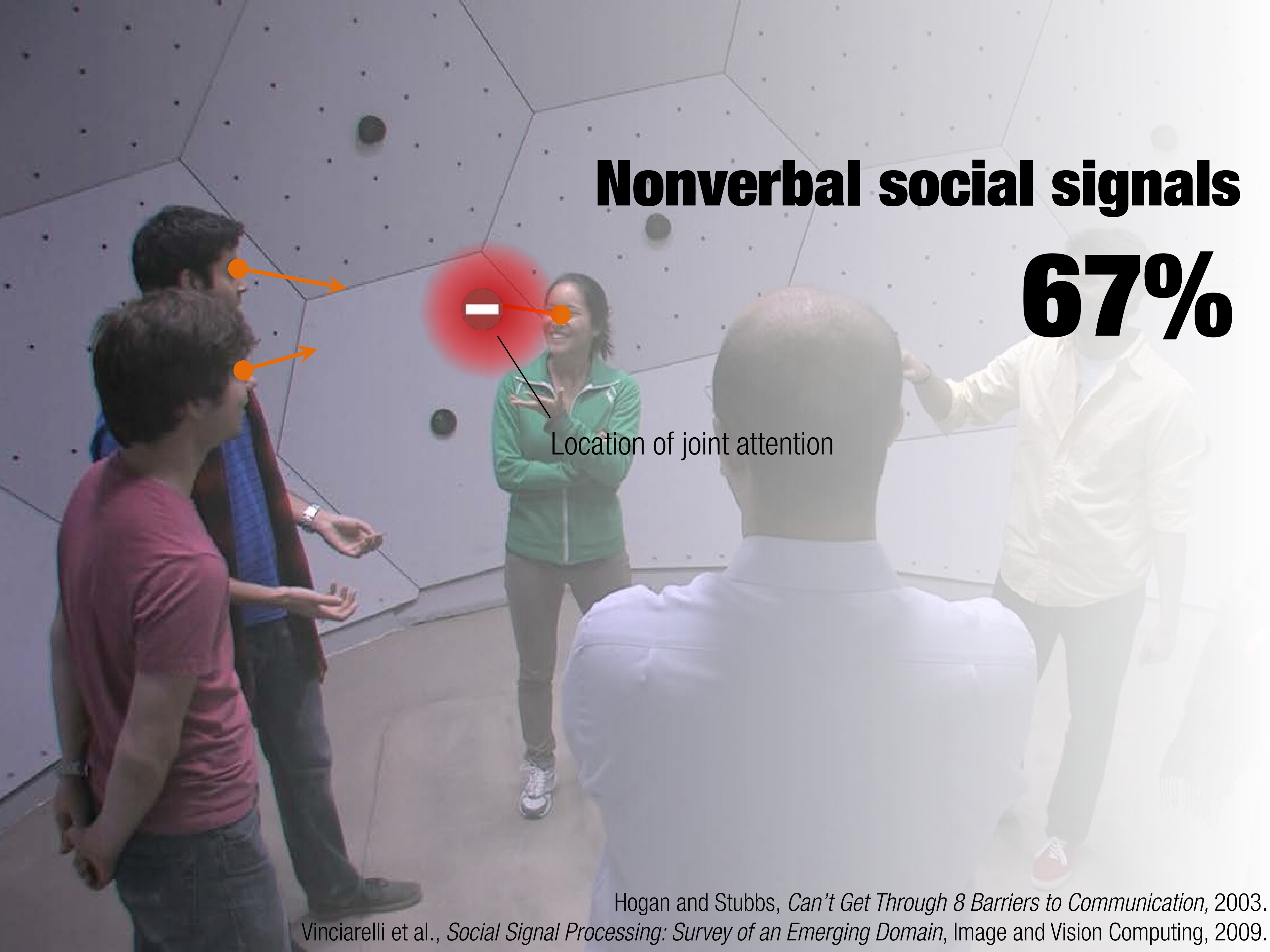
Nonverbal social signals

67%

Hogan and Stubbs, *Can't Get Through 8 Barriers to Communication,* 2003.
Vinciarelli et al., *Social Signal Processing: Survey of an Emerging Domain*, Image and Vision Computing, 2009.

Nonverbal social signals

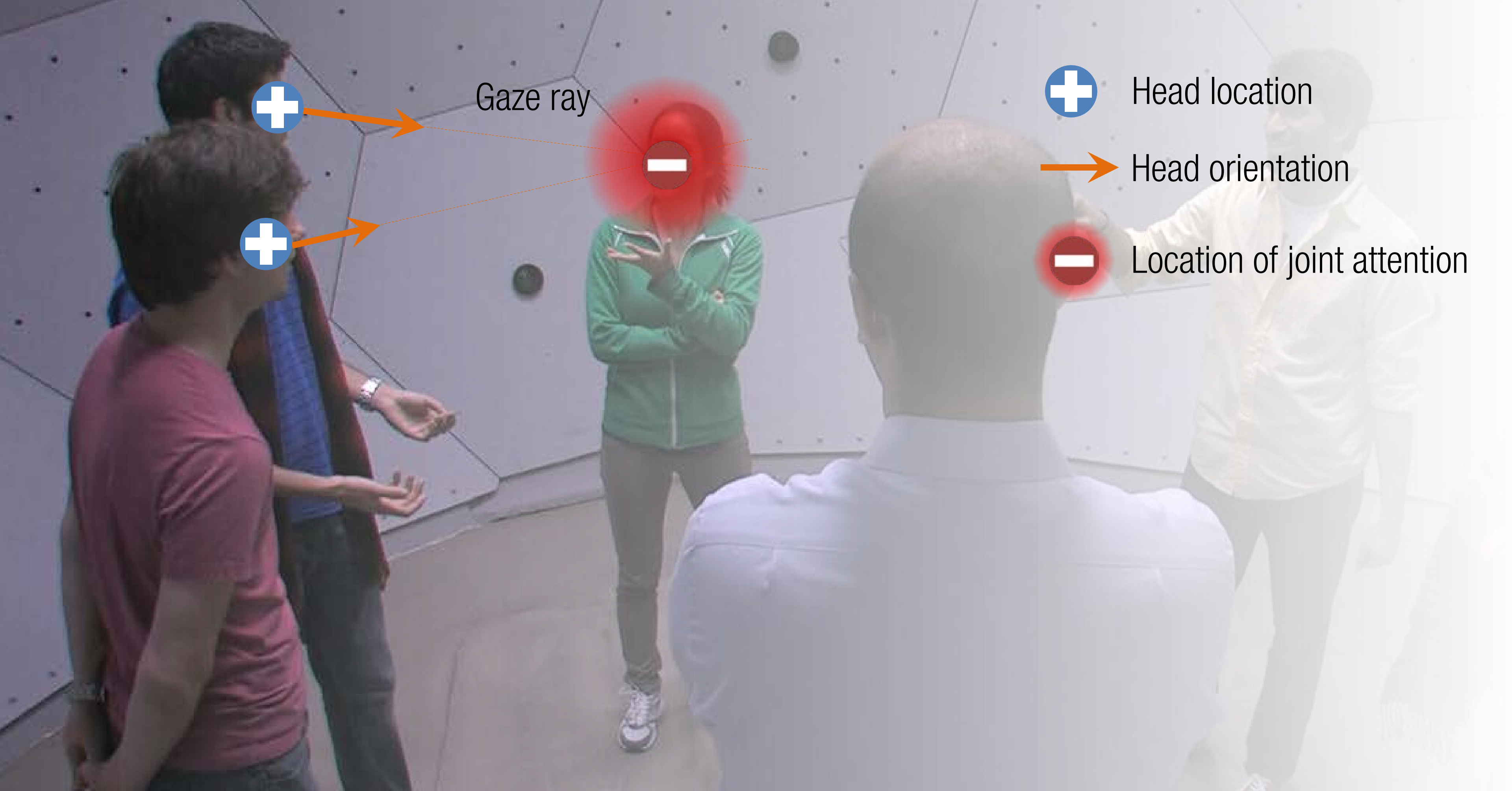
67%



Location of joint attention

Hogan and Stubbs, *Can't Get Through 8 Barriers to Communication,* 2003.
Vinciarelli et al., *Social Signal Processing: Survey of an Emerging Domain*, Image and Vision Computing, 2009.

Geometric Localization of Joint Attention



✚ Head location

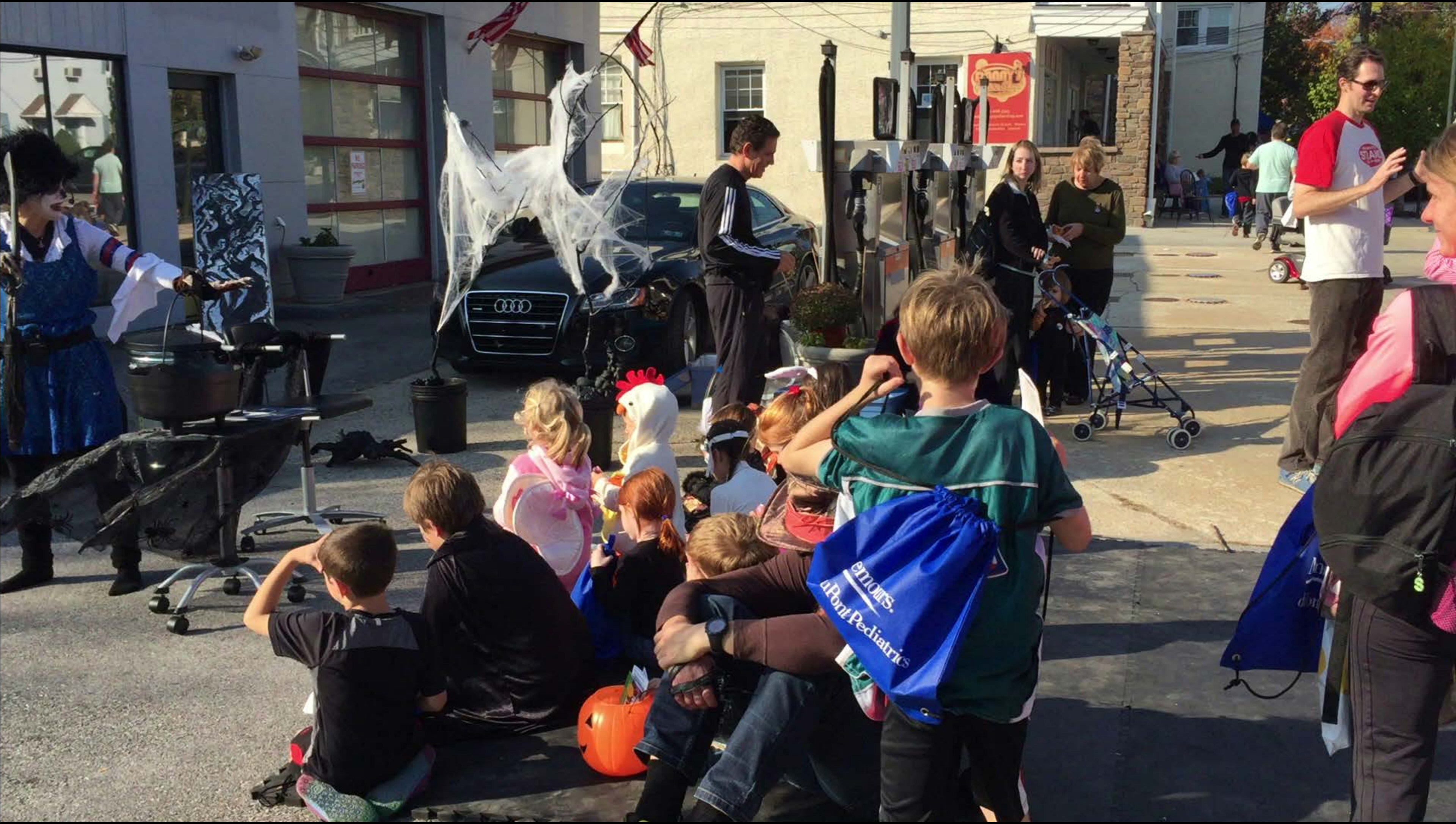
→ Head orientation

⊖ Location of joint attention

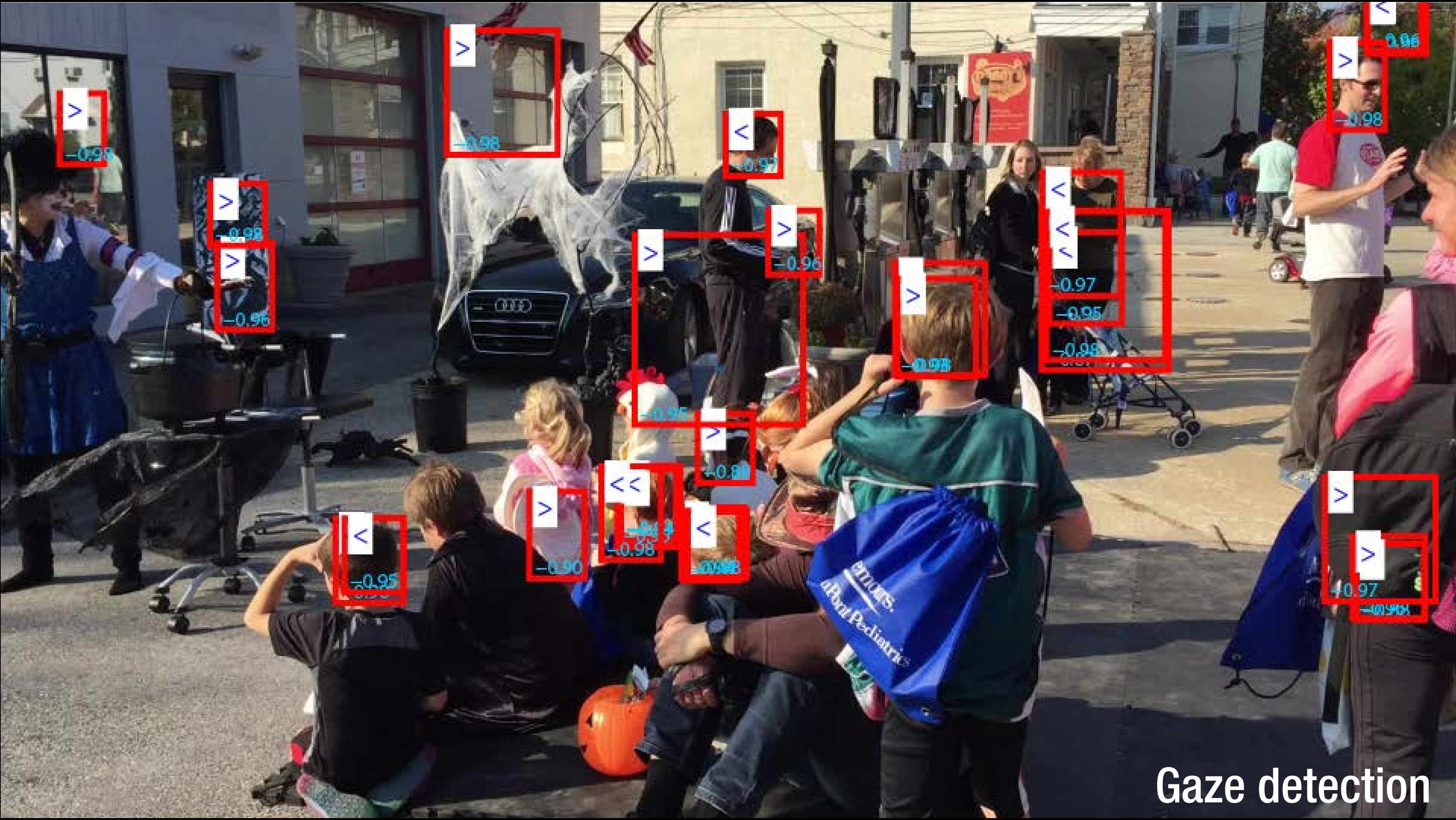
Patron-Perez et al., Structured Learning of Human Interactions in TV Shows, PAMI, 2012

Fathi et al., *Social Interactions: A First Person Perspective*, CVPR, 2012

Park et al., 3D Social Saliency from Head-mounted Cameras, NIPS, 2012



Challenges in Social Scenes



Gaze detection

Challenges in Social Scenes

Marin-Jimenez et al., *Detecting People Looking at Each Other in Videos*, IJCV 2014

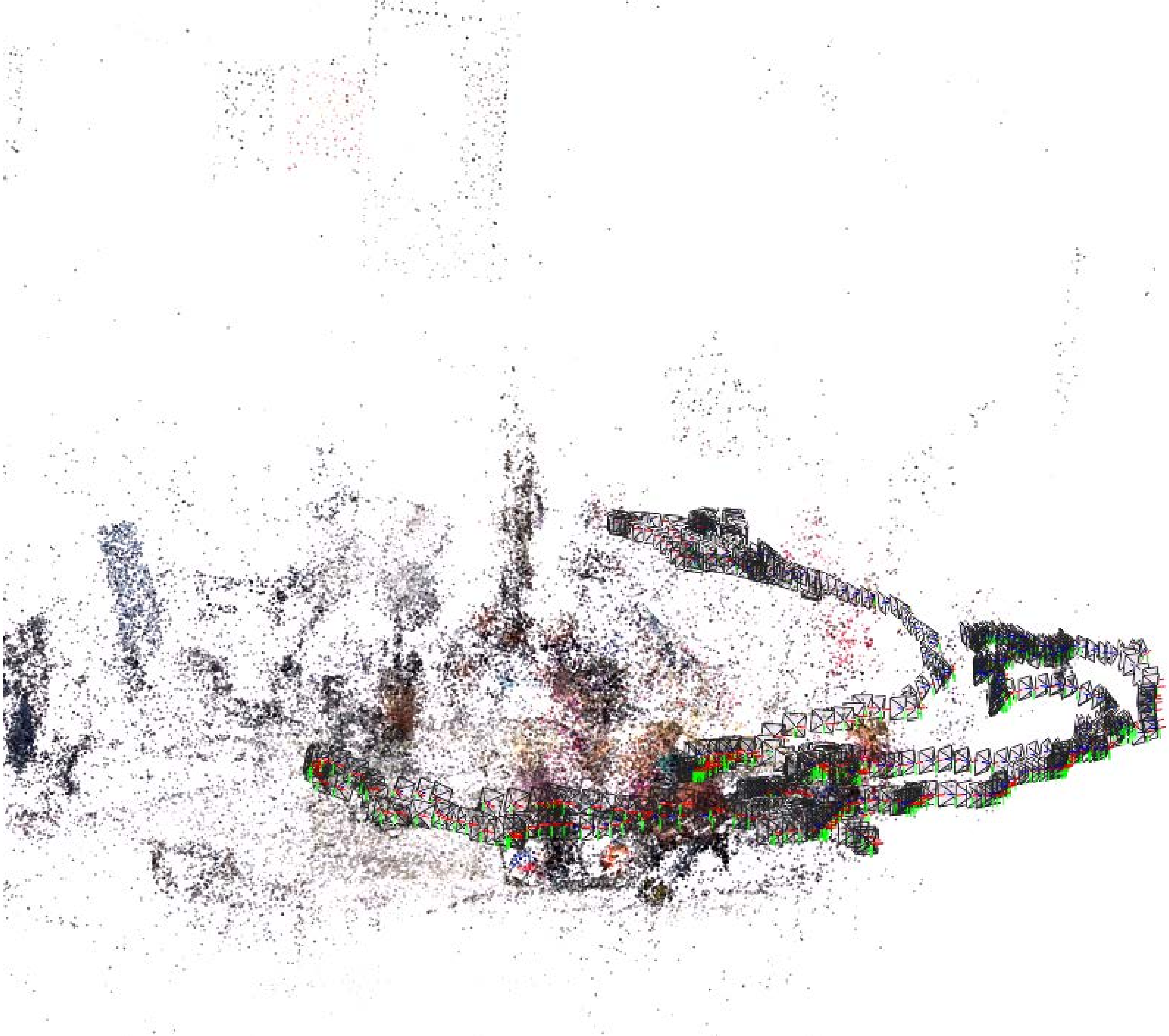
Can we localize joint attention without measuring gaze directions?

True positive head detection

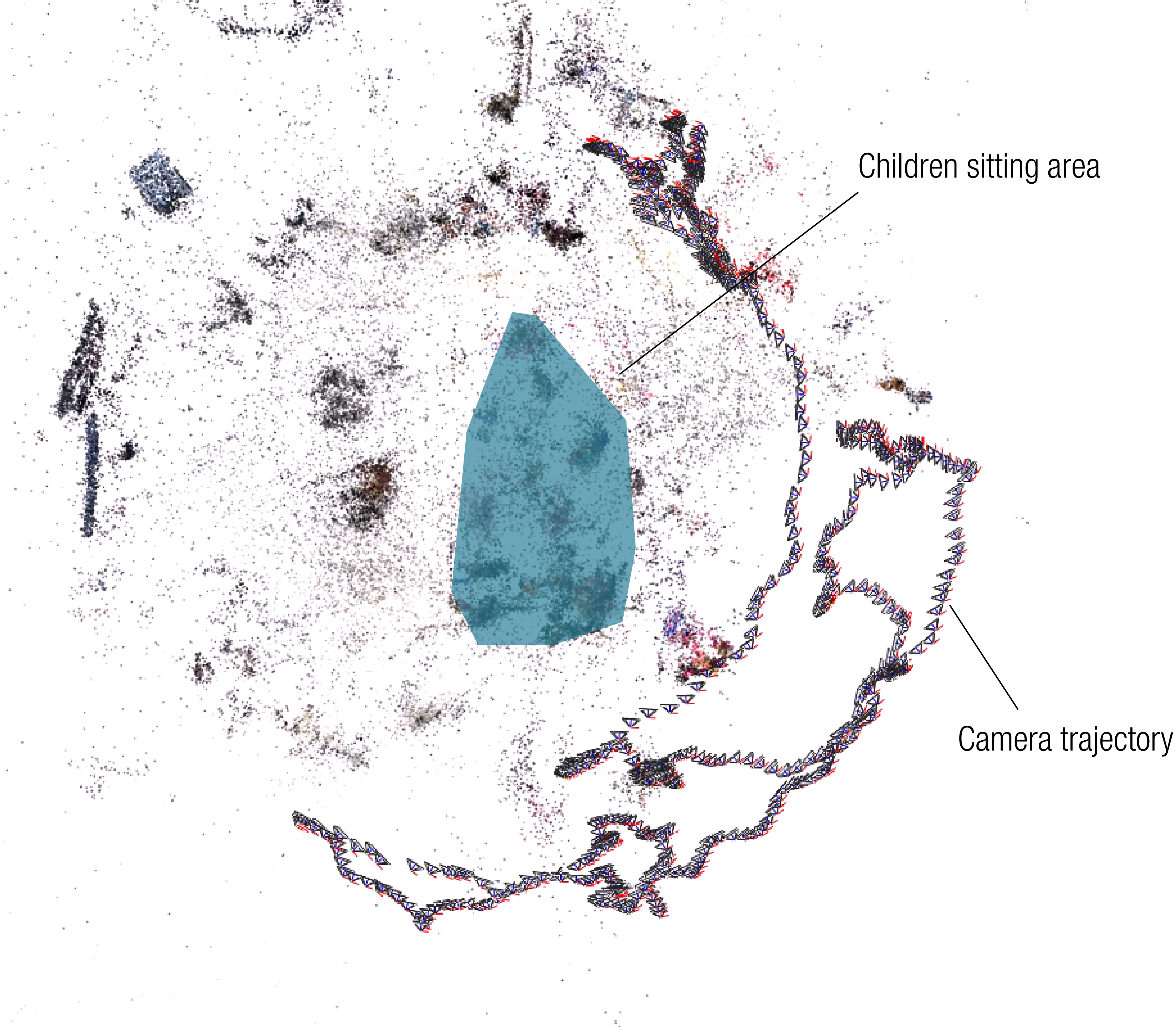
Challenges in Social Scenes

Marin-Jimenez et al., *Detecting People Looking at Each Other in Videos*, IJCV 2014

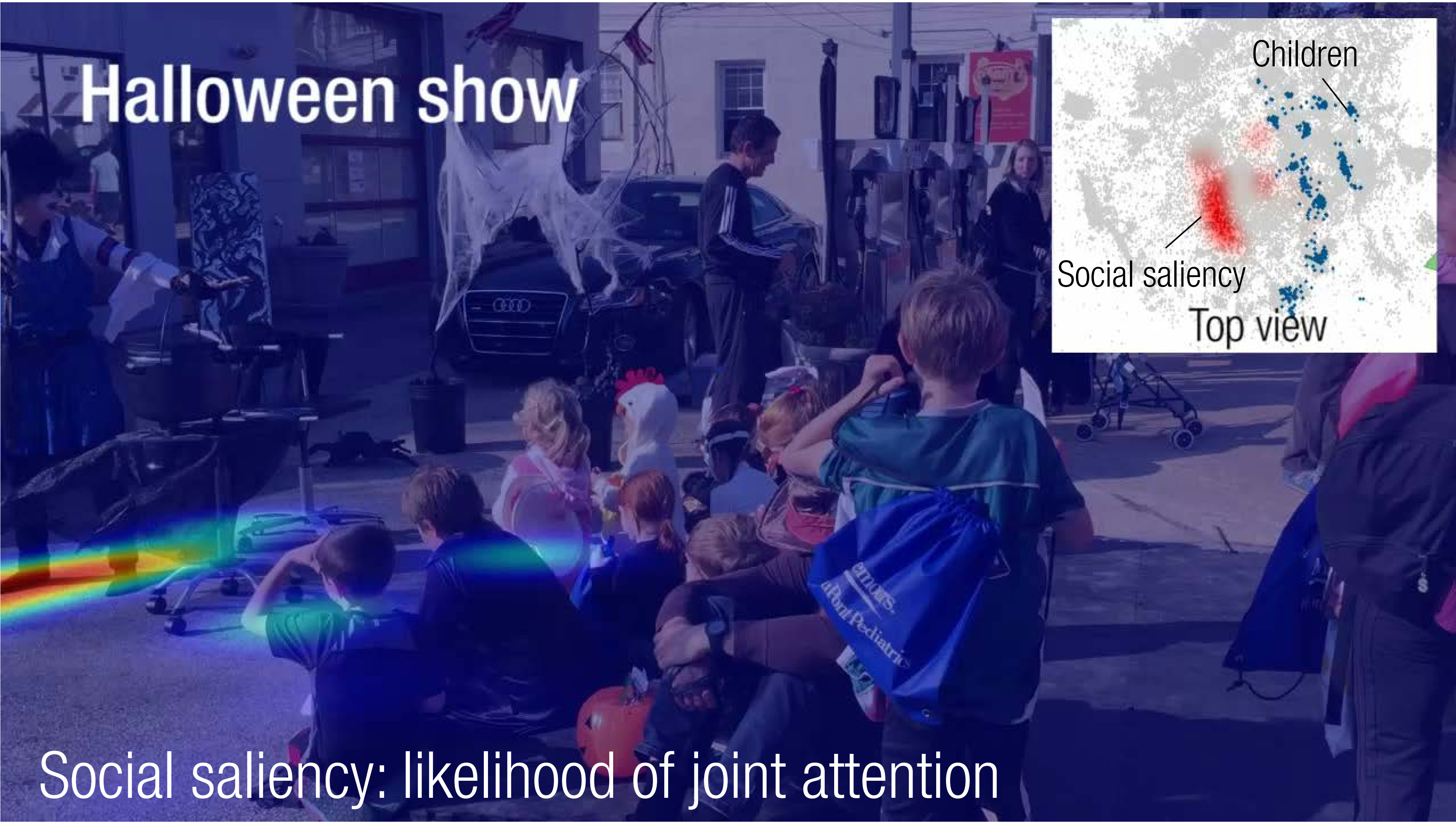
Input Video



Structure from Motion



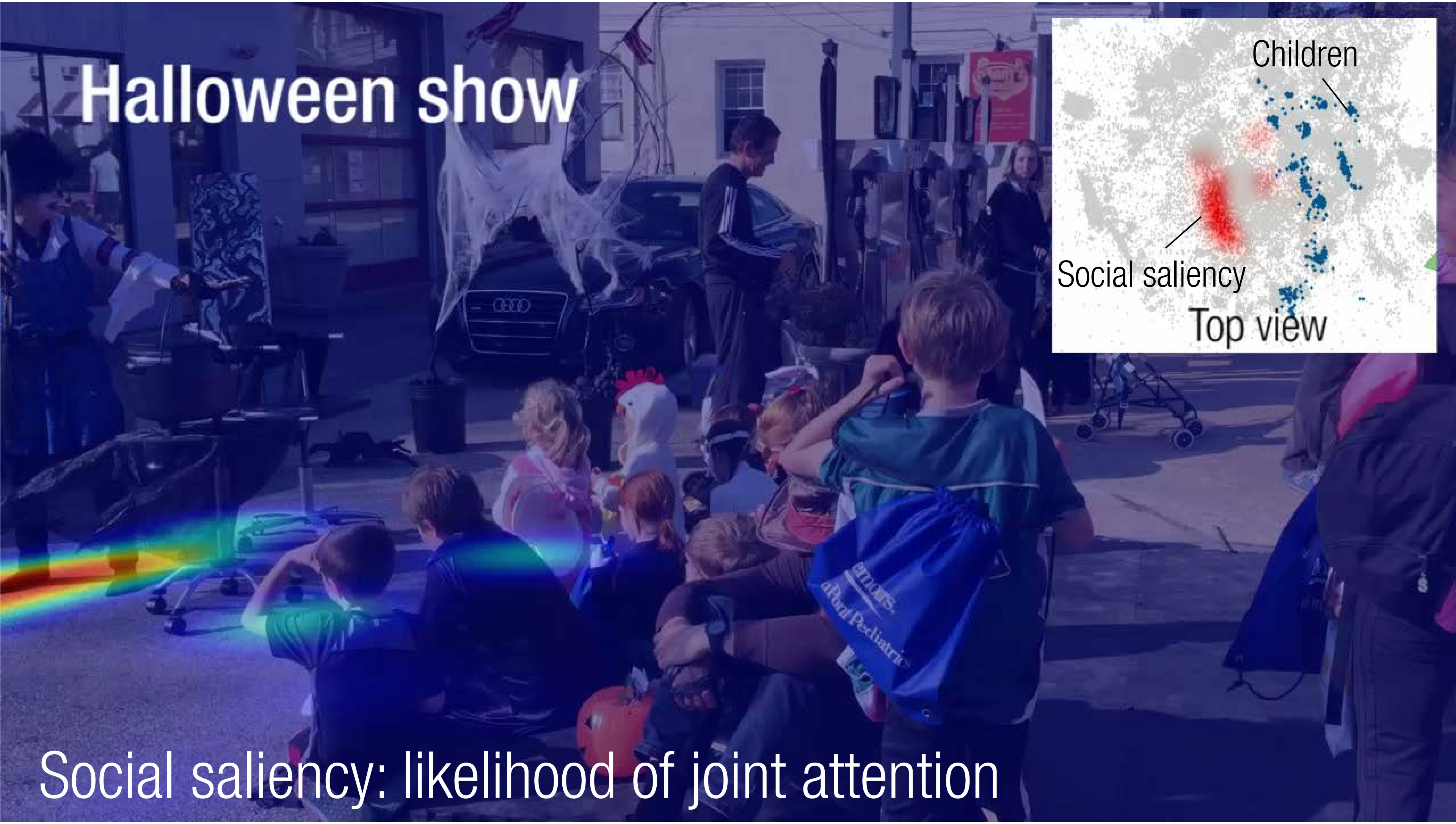
Children sitting area
Camera trajectory



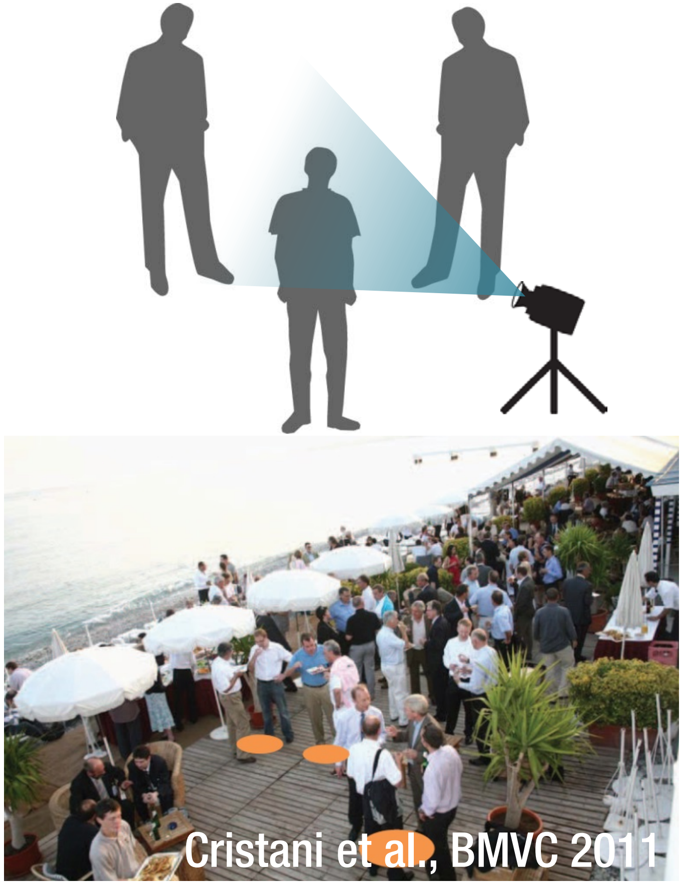
Halloween show

Social saliency: likelihood of joint attention

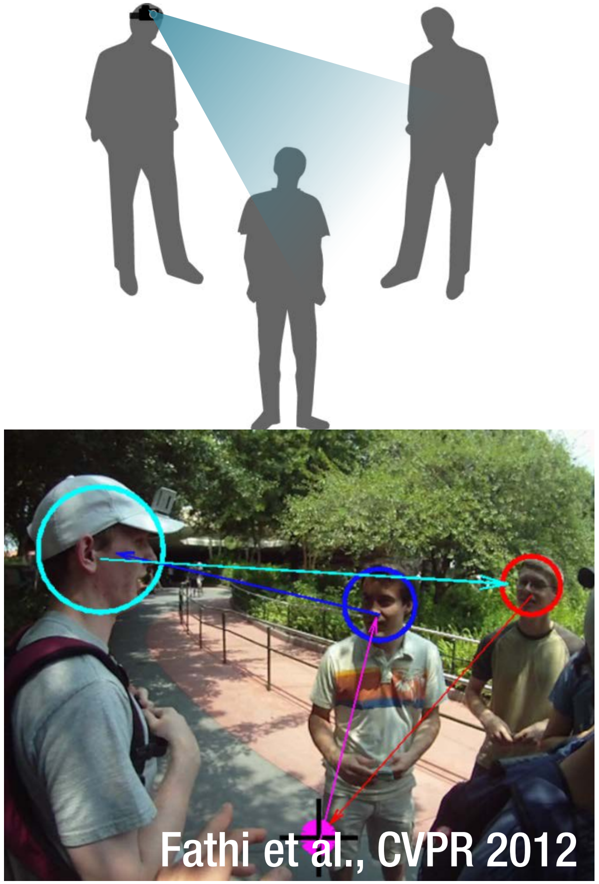
Output


Halloween show
Children
Social saliency
Top view
Social saliency: likelihood of joint attention

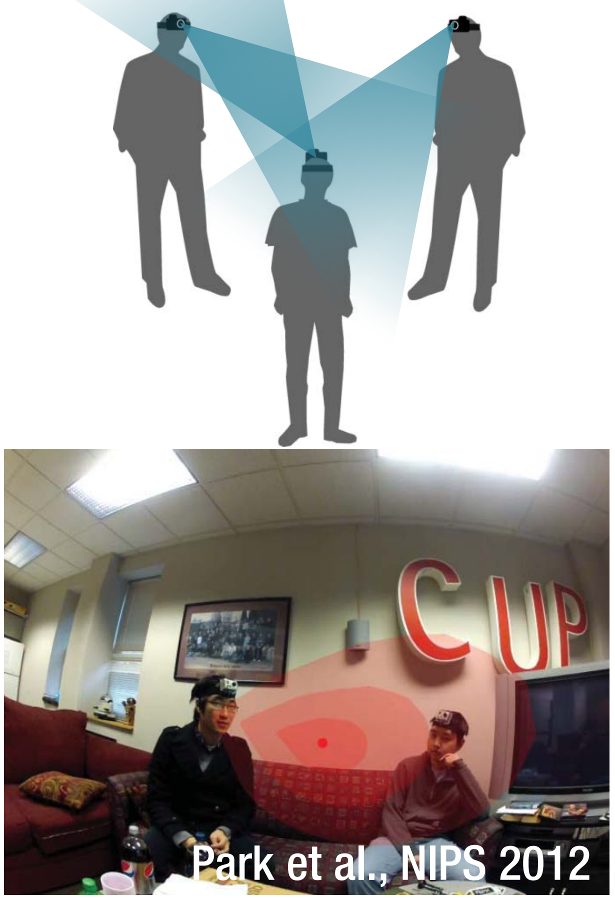
Output



Rodriguez et al., ICCV 2011
Lan et al., PAMI 2012
Chakraborty et al., CVPR 2013
Yang et al., CVPR 2011
Alahi et al., CVPR 2014
Choi et al., ECCV 2014



Li et al., ICCV 2013
Ryoo et al., CVPR 2013
Pusiol et al., CogSci 2014



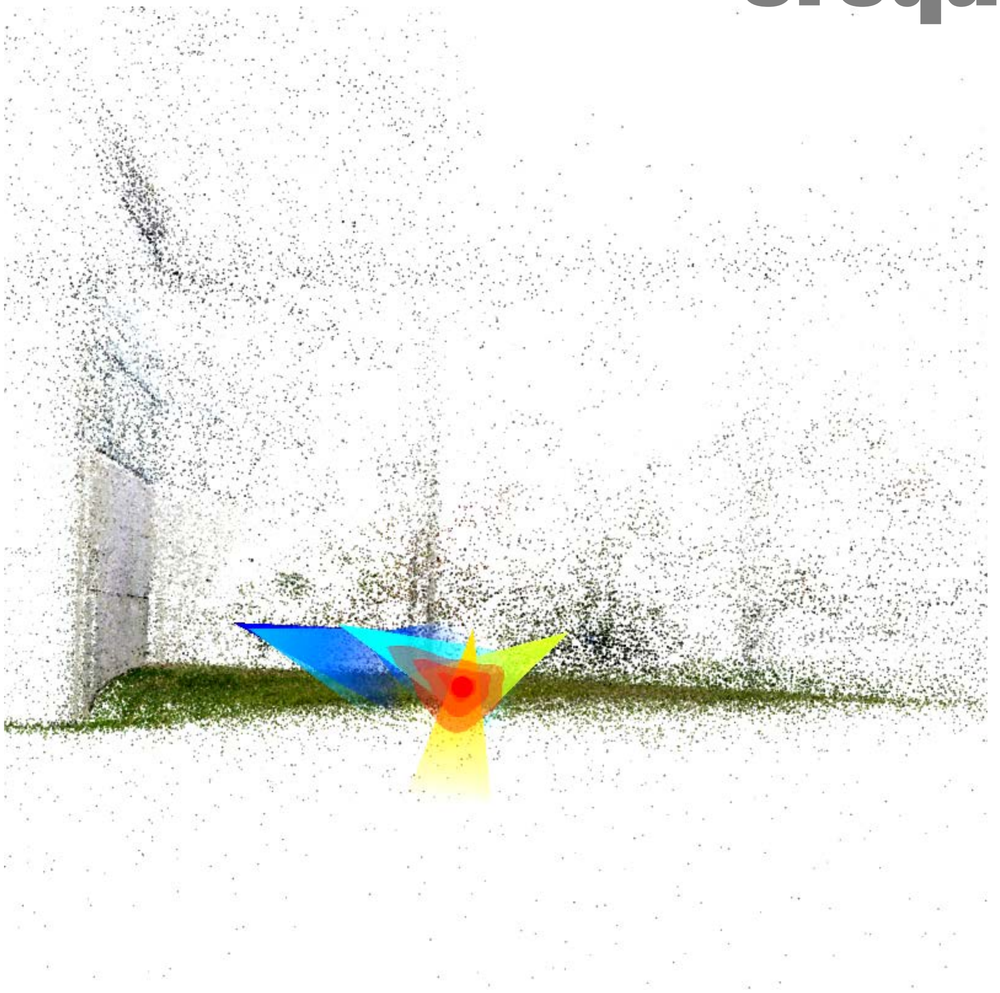
Arev et al., SIGGRAPH 2014

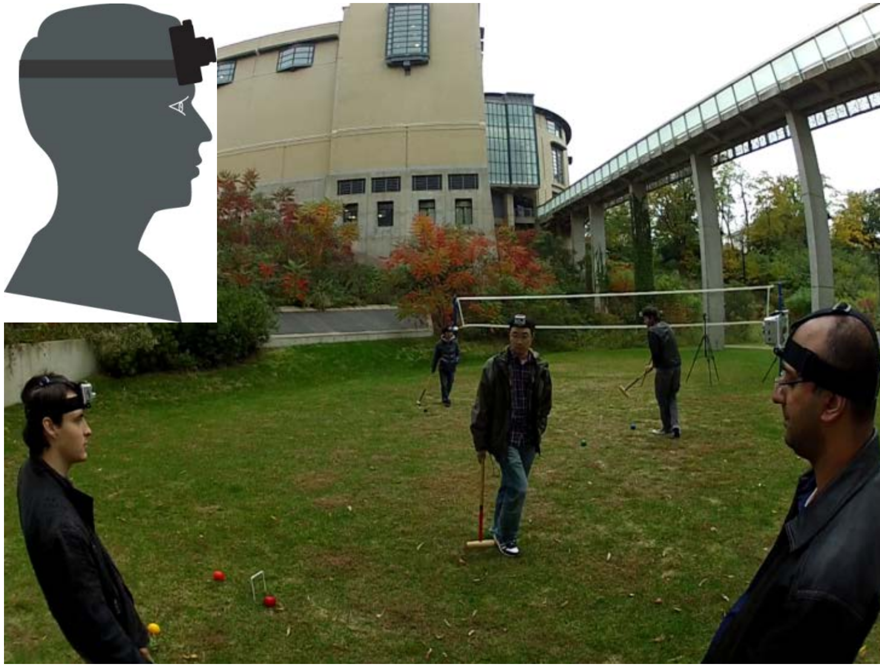
Third person view ⟷ First person view

Distance between face and camera

Joint Attention from First Person Cameras

Croquet

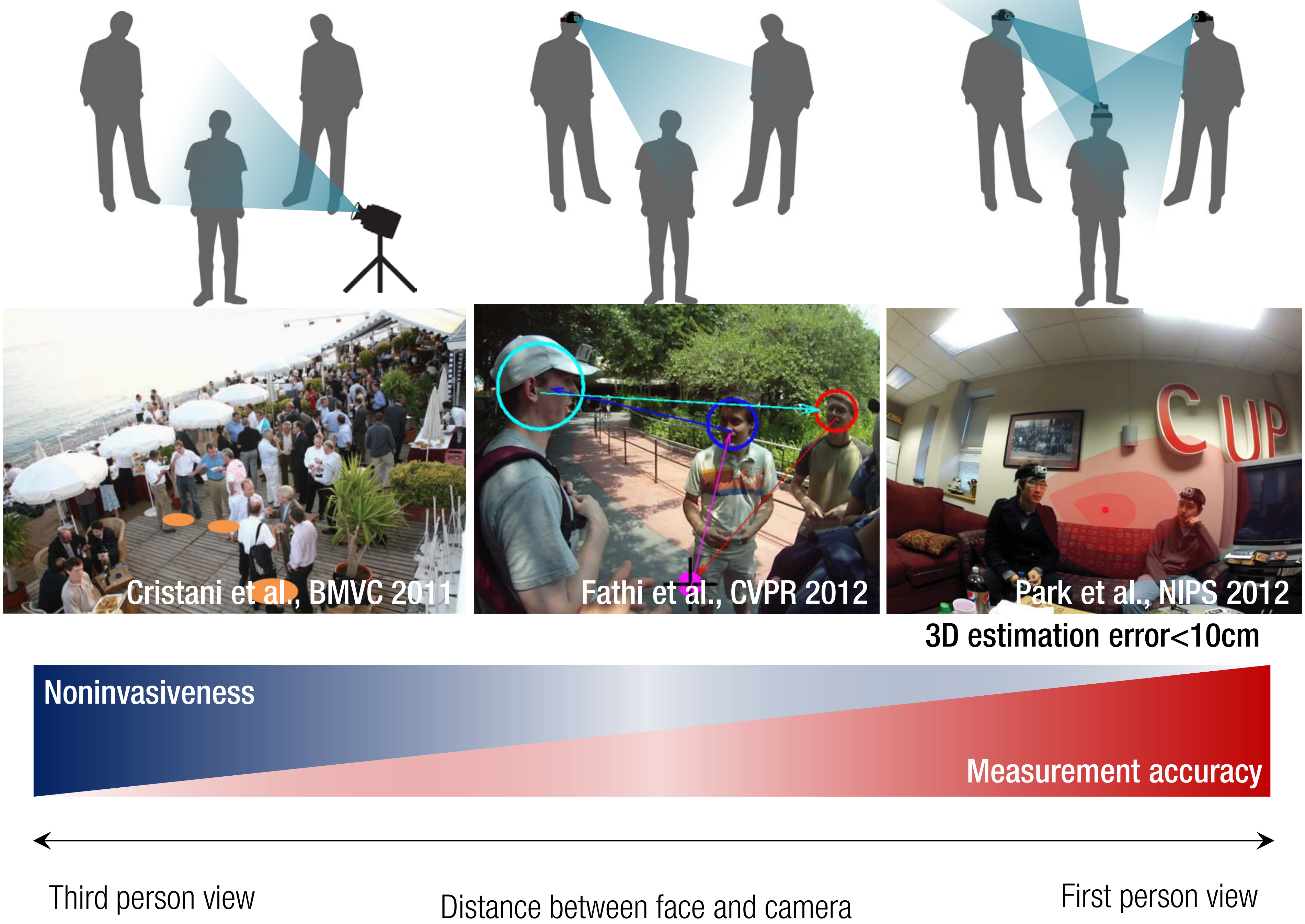




● : Joint attention

◀ : Head direction

Park et al., *3D Social Saliency from Head-mounted Cameras*, NIPS, 2012.



Cristani et al., BMVC 2011

Fathi et al., CVPR 2012

Park et al., NIPS 2012

3D estimation error<10cm

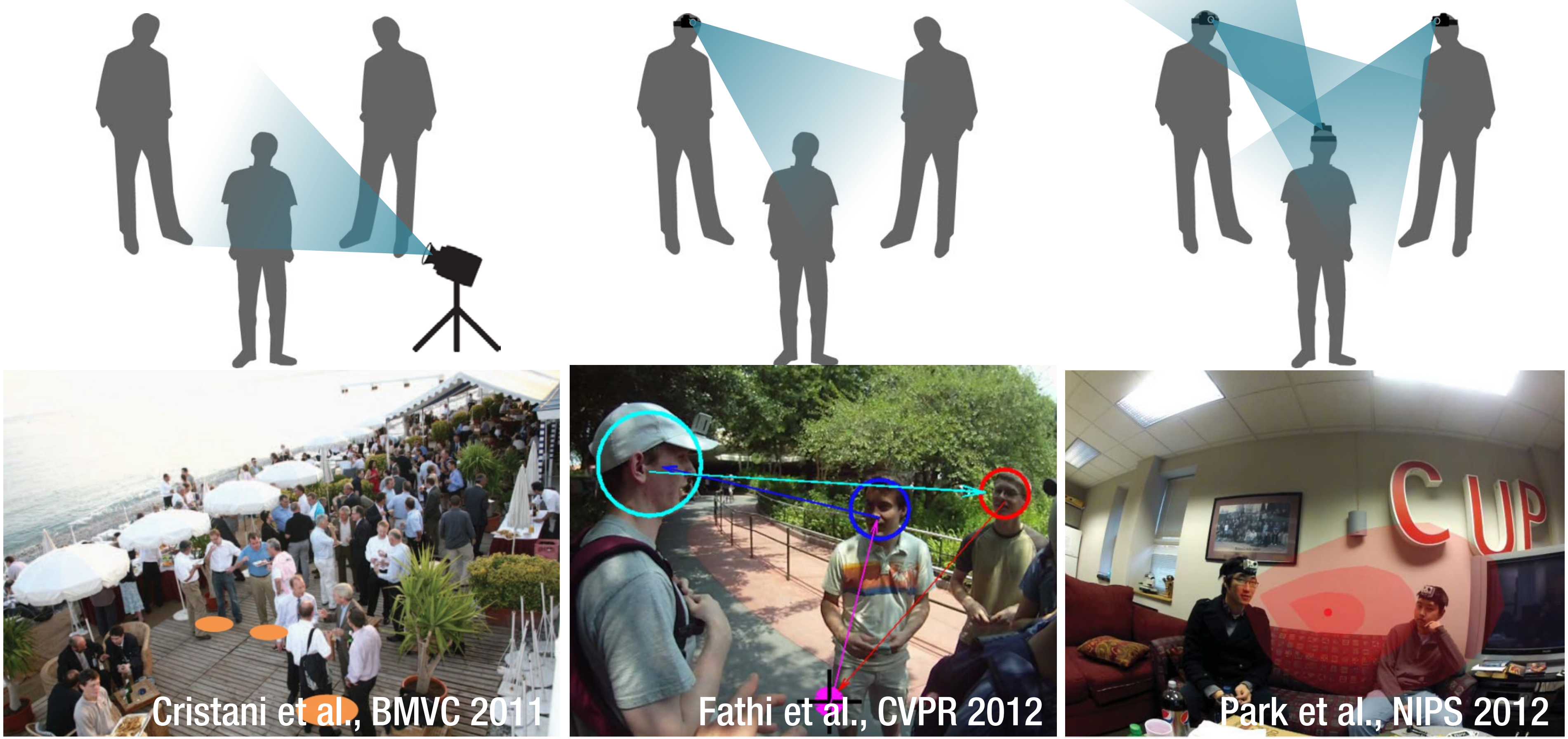
Noninvasiveness

Measurement accuracy

Third person view

Distance between face and camera

First person view



Cristani et al., BMVC 2011

Fathi et al., CVPR 2012

Park et al., NIPS 2012

3D estimation error<10cm

Noninvasiveness

Prediction

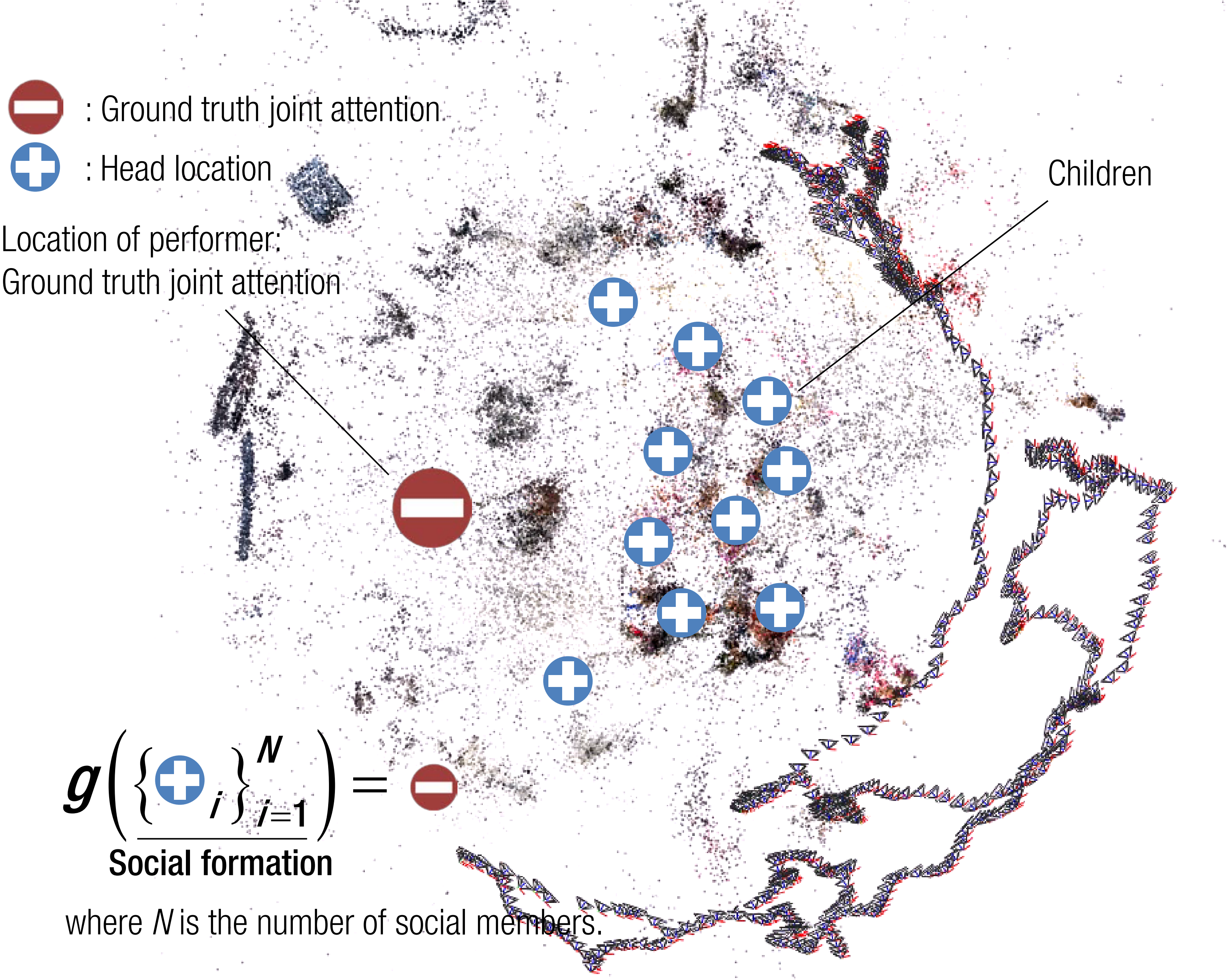
Learning

Measurement accuracy

Third person view

Distance between face and camera

First person view



⊖ : Ground truth joint attention

⊕ : Head location

Location of performer:
Ground truth joint attention

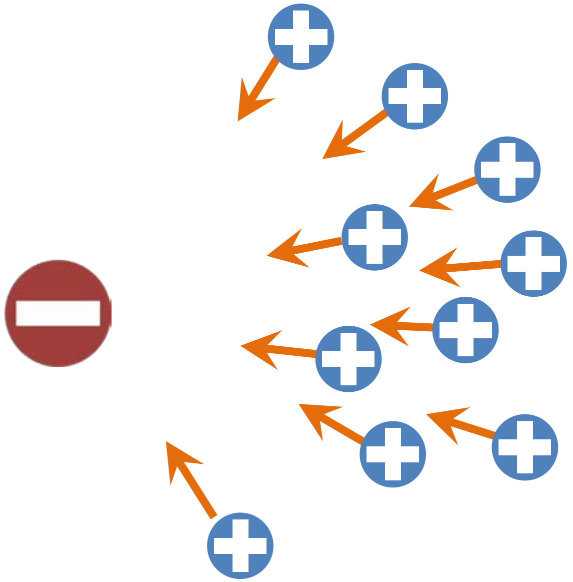
Children

$$g\Bigg(\underbrace{\{\oplus_i\}_{i=1}^{N}}_{\textbf{Social formation}}\Bigg) = \ominus$$

where N is the number of social members.

⊖ : Ground truth joint attention

⊕ : Head location



$$g\Big(\underbrace{\{\oplus_i\}_{i=1}^{N}}_{\textbf{Social formation}}\Big) = \ominus$$

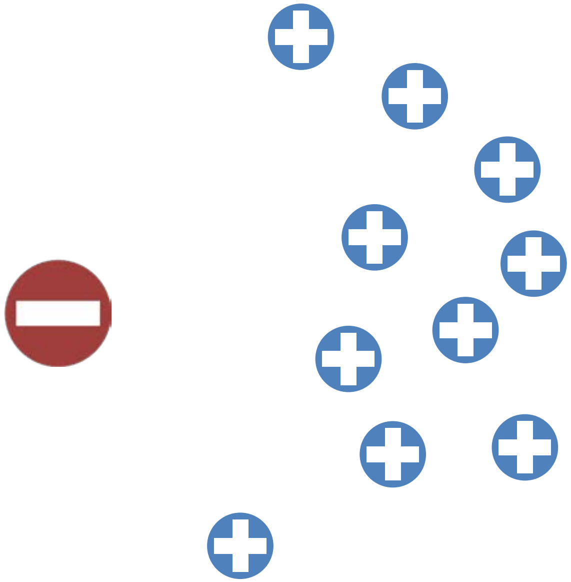
where N is the number of social members.

$$\text{cf.}\quad g\Big(\{\oplus\!\longrightarrow_i\}_{i=1}^{N}\Big) = \ominus$$

Geometric localization: triangulation

⊖ : Ground truth joint attention

⊕ : Head location



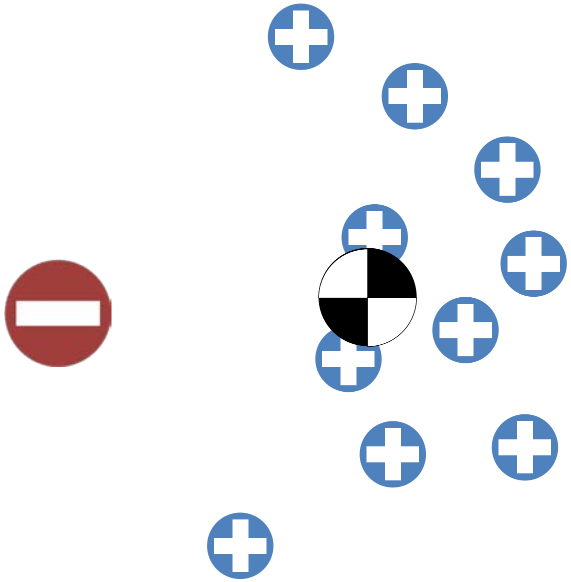
$$g\left(\underbrace{\left\{\oplus_i\right\}_{i=1}^{N}}_{\text{Social formation}}\right) = \ominus$$

where *N* is the number of social members.

⊖ : Ground truth joint attention

⊕ : Head location

◐ : Center of mass



$$g\left(\underset{\textbf{Social formation}}{\underline{\left\{\oplus_{i}\right\}_{i=1}^{N}}}\right)=\ominus$$

where N is the number of social members.

$$\text{ex. } g\left(\left\{\oplus_{i}\right\}_{i=1}^{N}\right)=\frac{1}{N}\sum_{i=1}^{N}\oplus=\ominus$$

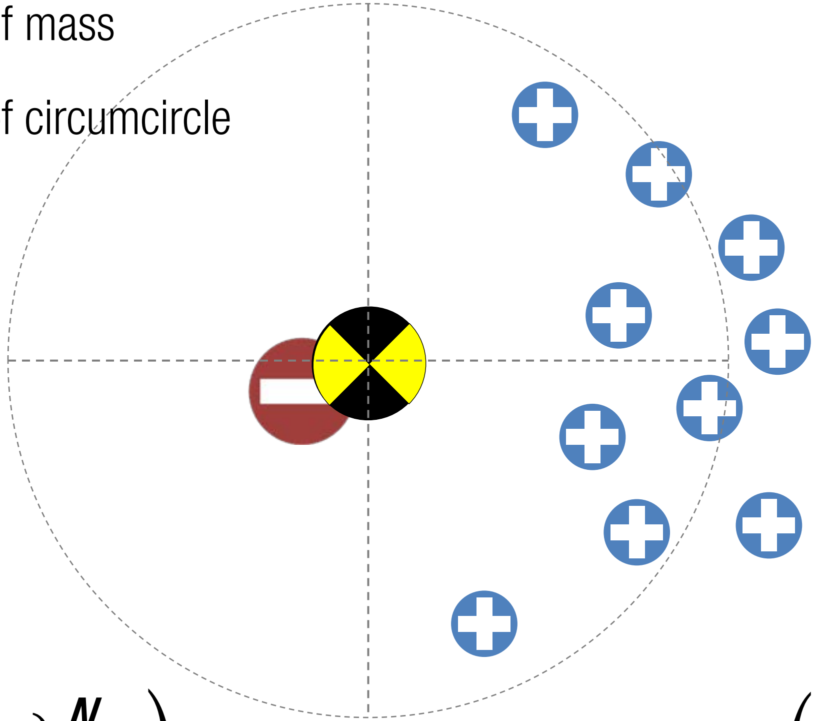
Geometric localization: center of mass ◐

⊖ : Ground truth joint attention

⊕ : Head location

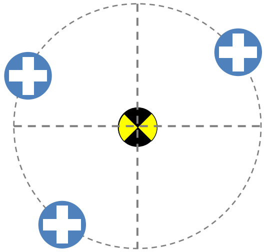
◐ : Center of mass

⊗ : Center of circumcircle



$$g\Big(\underbrace{\{\oplus_i\}_{i=1}^{N}}_{\textbf{Social formation}}\Big) = \ominus$$

where N is the number of social members.

cf. $g\Big(\{\oplus_i\}_{i=1}^{N}\Big) =$  $= \ominus$

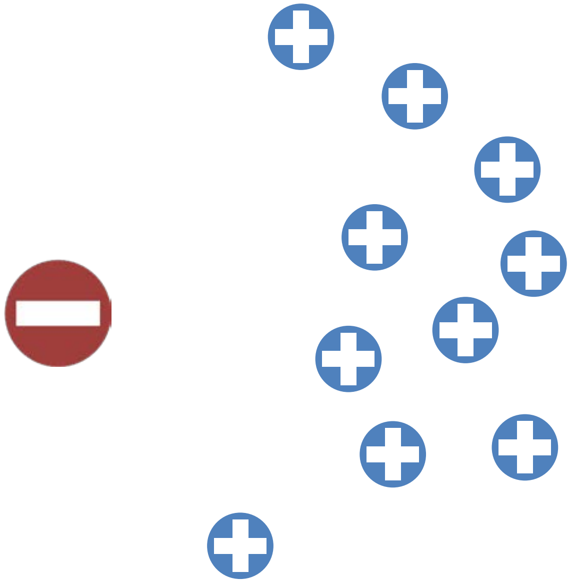
Geometric localization: center of circumcircle ⊗

⊖ : Ground truth joint attention

⊕ : Head location

◐ : Center of mass

⊗ : Center of circumcircle



$$g\Big(\underbrace{\{\oplus_i\}_{i=1}^{N}}_{\text{Social formation}}\Big) = \ominus$$

where N is the number of social members.

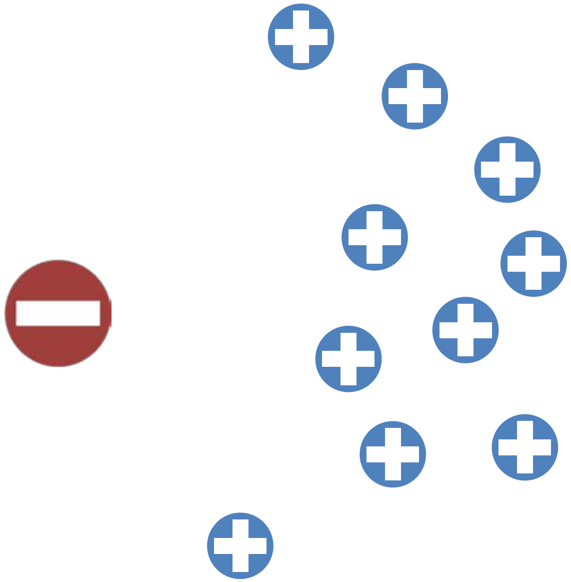
First Person Social Interaction Data

Scene	N	T(sec)	F
B-boy I	18	105	317
B-boy II	18	450	1351
B-boy III	18	160	528
B-boy IV	18	50	180
Surprise party	11	120	2227
Class	11	360	3590
Croquet	6	300	6000
Busker I	6	120	3566
Busker II	6	180	5394
Card game	3	180	768
Hide and seek	3	180	214
Block building	3	700	2702
Social game	8	450	2086
Meeting I	11	120	832
Meeting II	5	440	1120
Picnic	6	60	965
Musical	7	180	2184
Dance	6	180	5301
4 way party	11	180	1909
Snowman	4	753	8256

Total 49,490 social formations

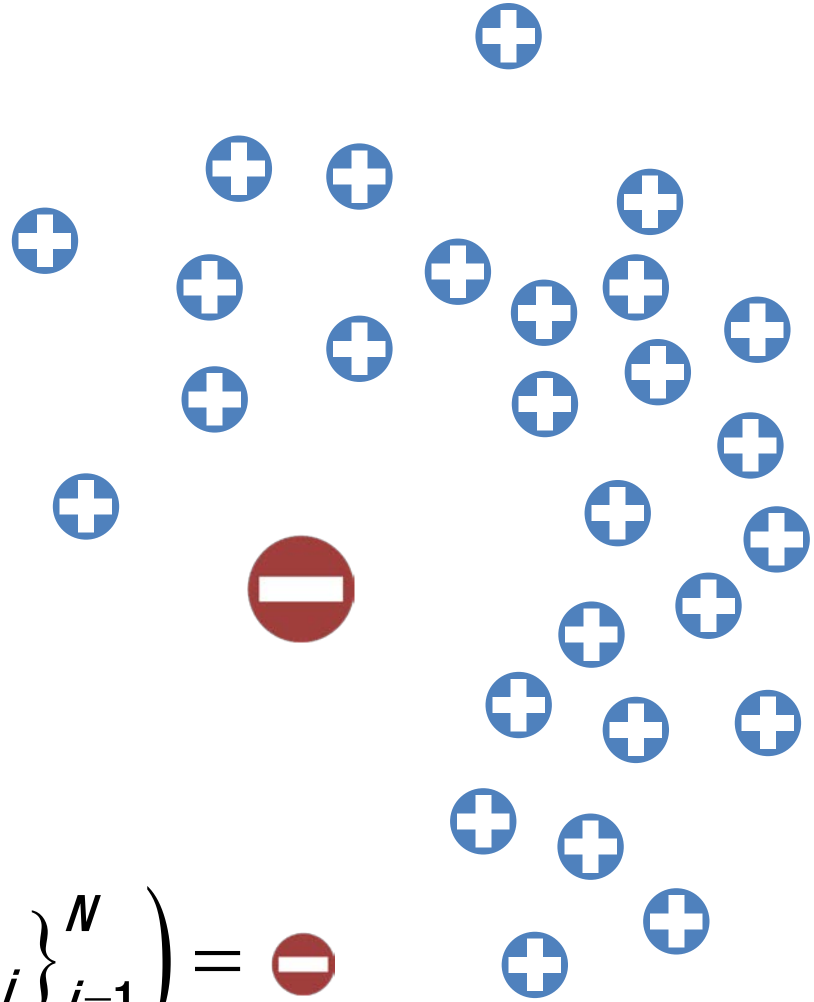
⊖ : Ground truth joint attention

⊕ : Head location



$$g\left(\underbrace{\left\{\oplus_i\right\}_{i=1}^{N}}_{\text{Social formation}}\right) = \ominus$$

where N is the number of social members.



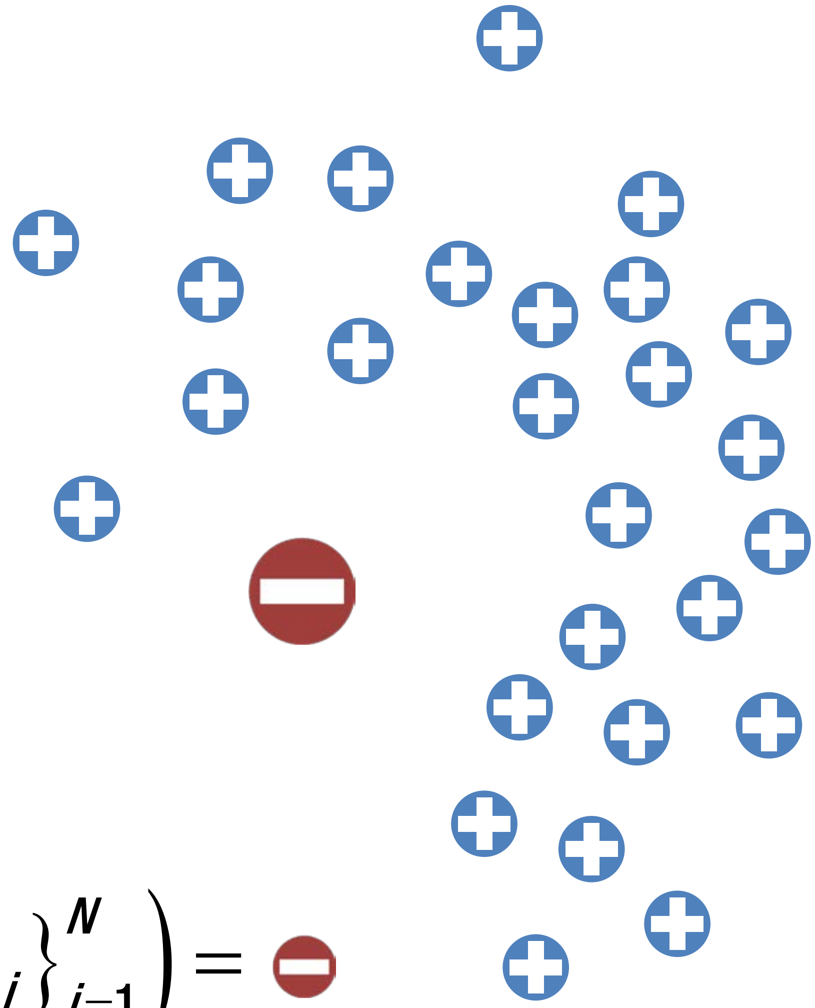
$$g\left(\underbrace{\left\{\oplus_i\right\}_{i=1}^{N}}_{\textbf{Social formation}}\right) = \ominus$$

where N is the number of social members.

Scale variation

⊖ : Ground truth joint attention

⊕ : Head location



$$g\left(\underbrace{\left\{\oplus_i\right\}_{i=1}^{N}}_{\textbf{Social formation}}\right) = \ominus$$

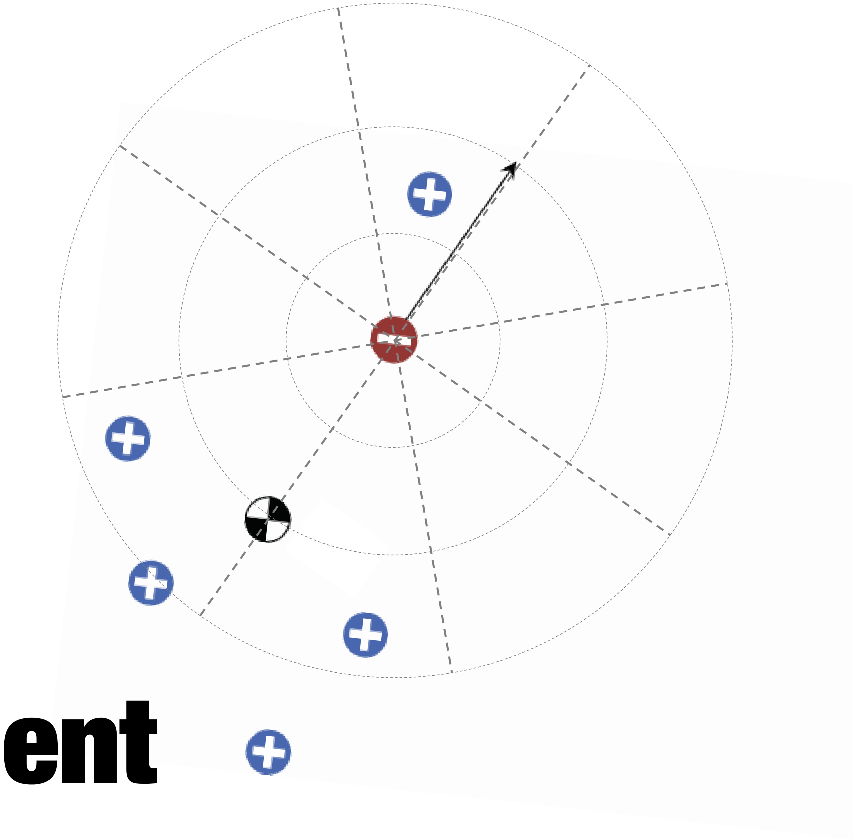
where *N* is the number of social members.

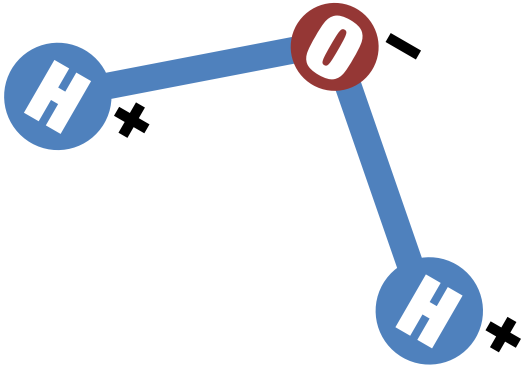
Scale variation
Orientation variation

⊖ : Joint attention

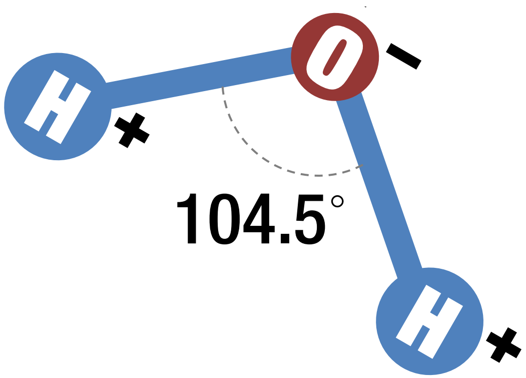
⊕ : Head location

Representation:
Social Dipole Moment





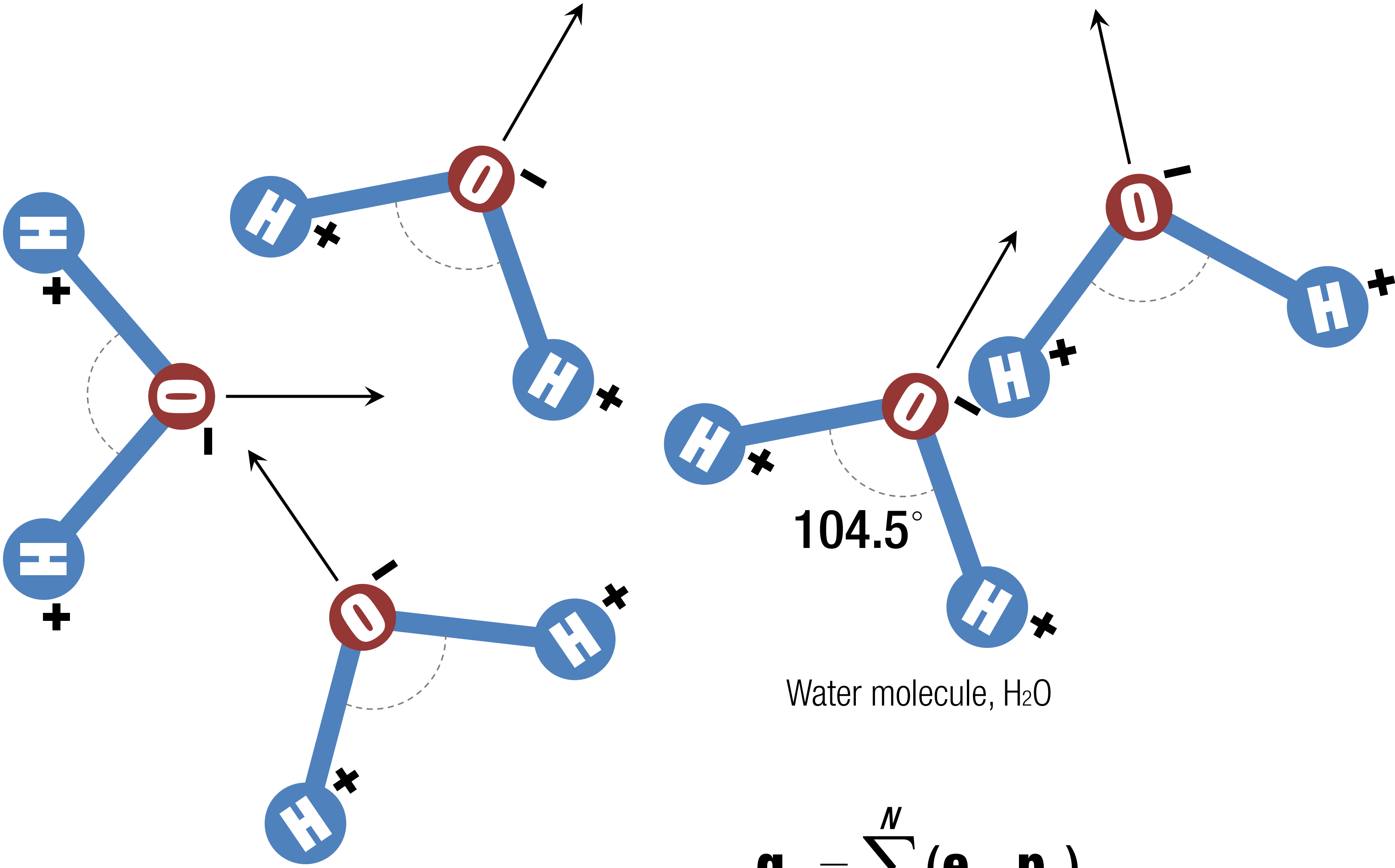
Water molecule, H_2O



Water molecule, H_2O

$$\mathbf{q}_e = \sum_{i}^{N} (\mathbf{e} - \mathbf{p}_i)$$

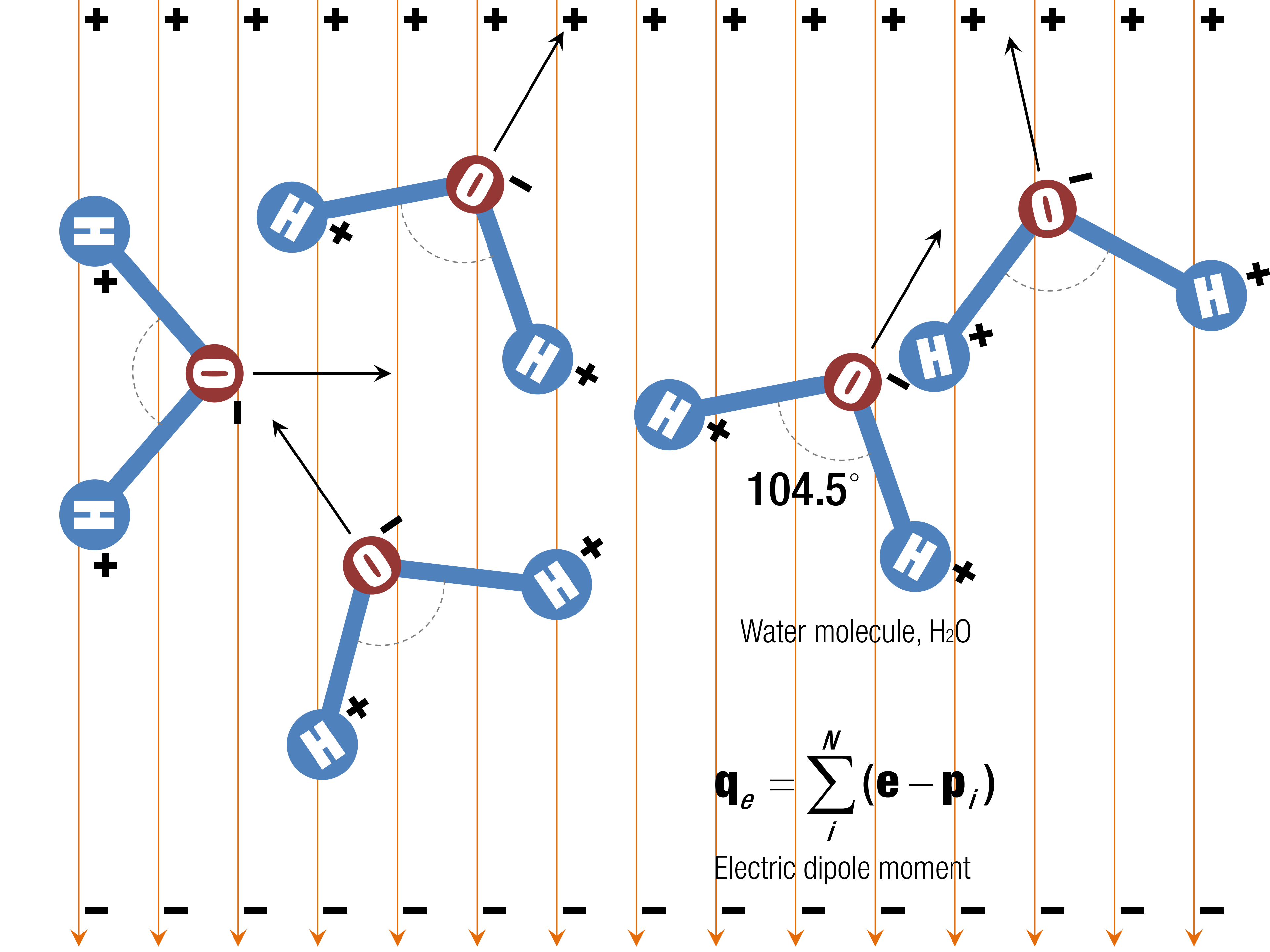
Electric dipole moment



Water molecule, H_2O

$$\mathbf{q}_e = \sum_{i}^{N} (\mathbf{e} - \mathbf{p}_i)$$

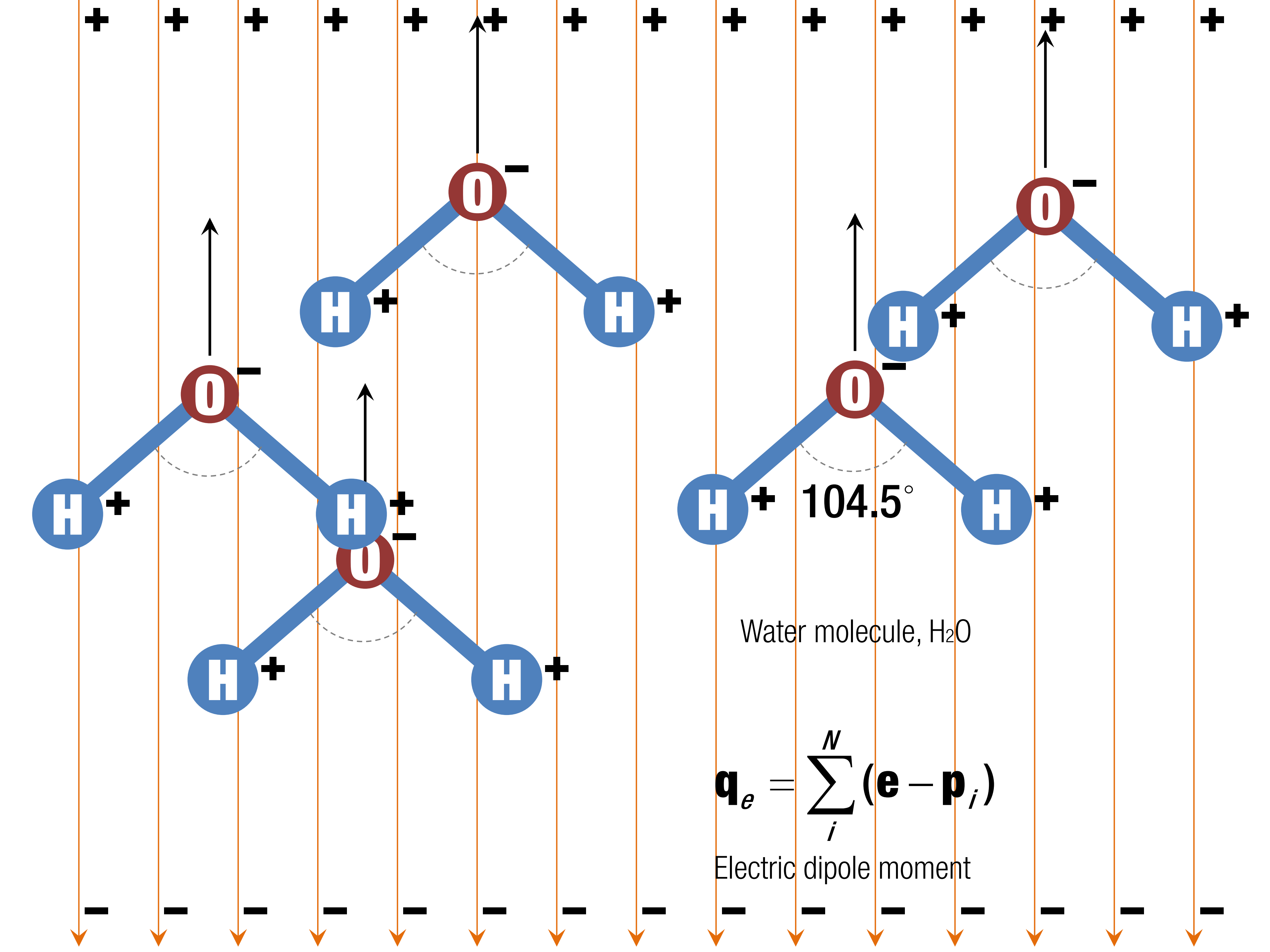
Electric dipole moment



Water molecule, H_2O

$$\mathbf{q}_e = \sum_{i}^{N} (\mathbf{e} - \mathbf{p}_i)$$

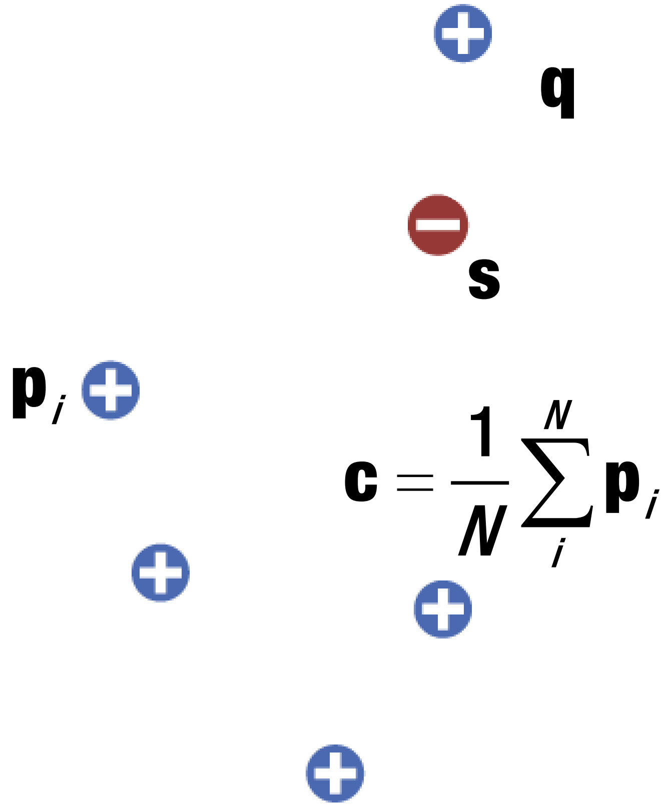
Electric dipole moment



Water molecule, H_2O

$$\mathbf{q}_e = \sum_{i}^{N} (\mathbf{e} - \mathbf{p}_i)$$

Electric dipole moment



$$\mathbf{c} = \frac{1}{N}\sum_{i}^{N}\mathbf{p}_i$$

⊖ : Ground truth joint attention

⊕ : Head location

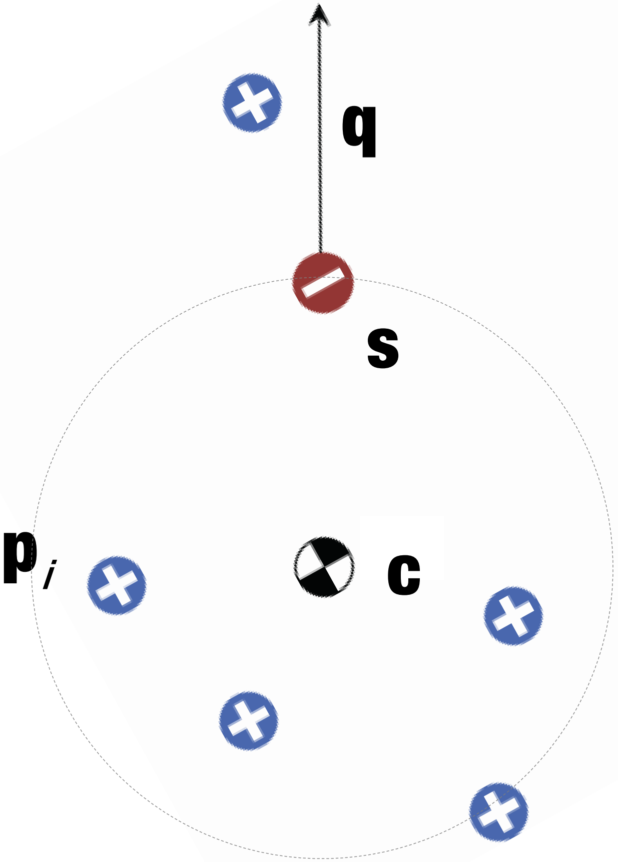
◐ : Center of mass

→ : Social dipole moment

$$\mathbf{q} = \mathbf{s} - \frac{1}{N}\sum_{i}^{N}\mathbf{p}_i = \mathbf{s} - \mathbf{c}$$

Social dipole moment

Orientation Normalization



⊖ : Ground truth joint attention

⊕ : Head location

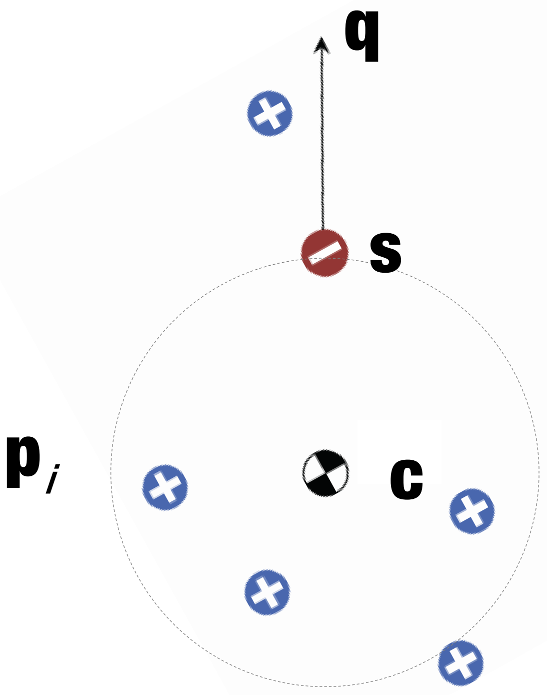
◑ : Center of mass

→ : Social dipole moment

$$\mathbf{q} = \mathbf{s} - \frac{1}{N}\sum_{i}^{N}\mathbf{p}_i = \mathbf{s} - \mathbf{c}$$

Social dipole moment

Scale Normalization



⊖ : Ground truth joint attention

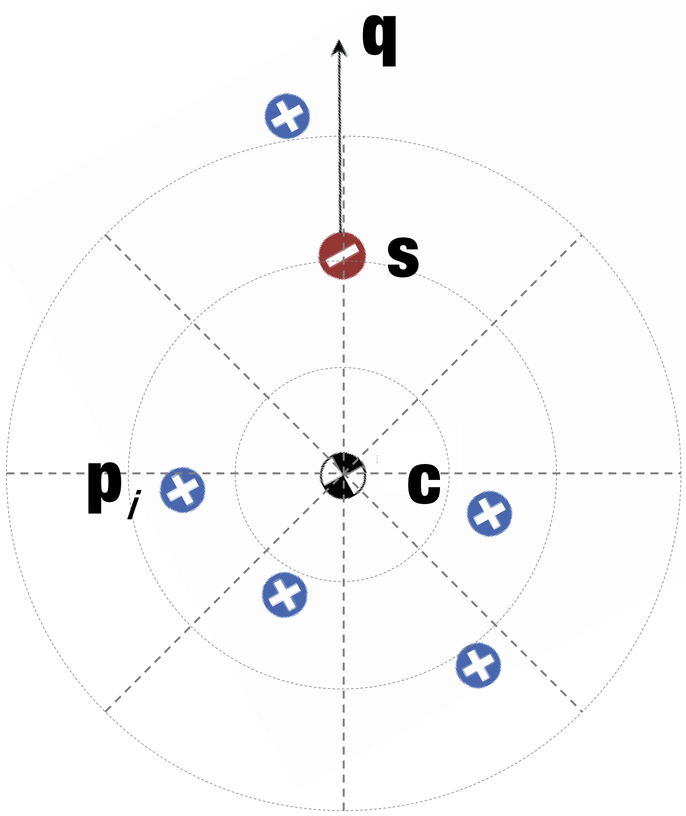
⊕ : Head location

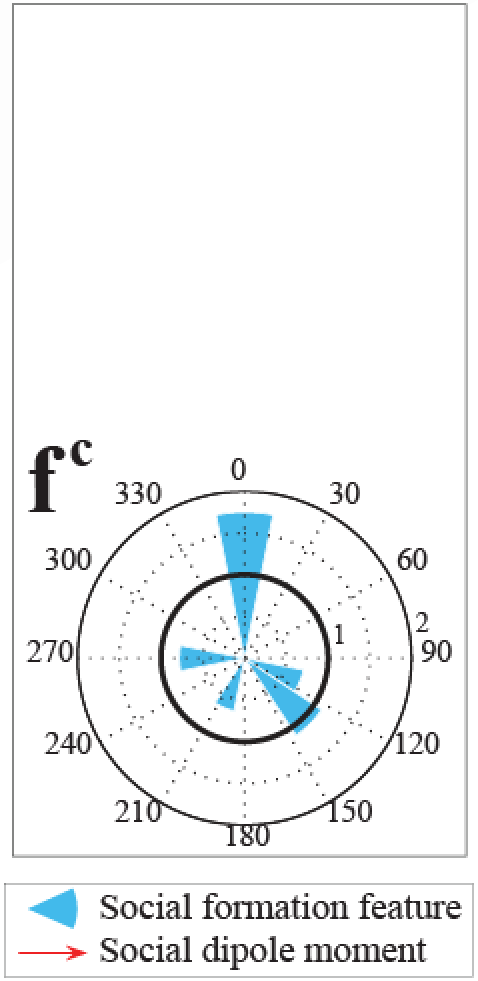
◕ : Center of mass

→ : Social dipole moment

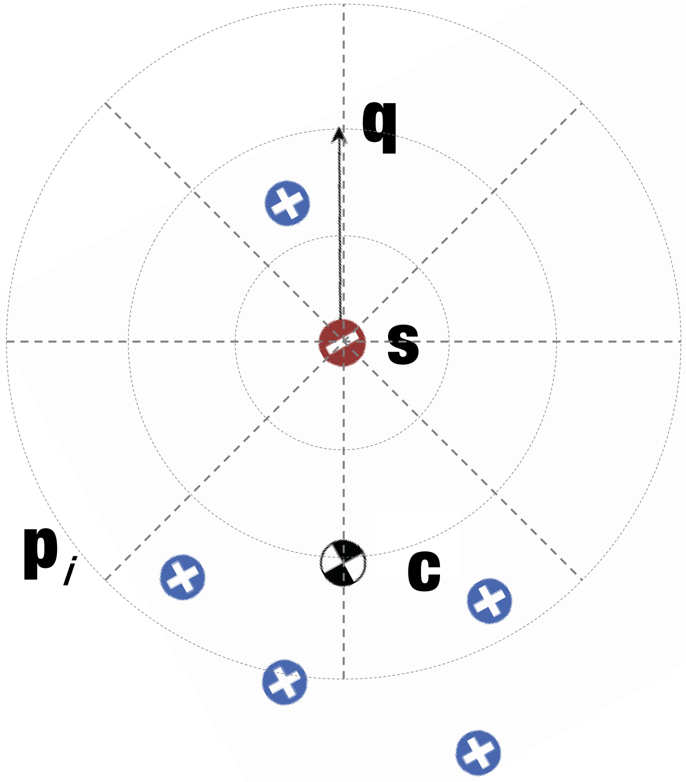
$$\frac{1}{N}\sum_{i}^{N}\left\|\mathbf{p}_i - \mathbf{c}\right\| = 1$$

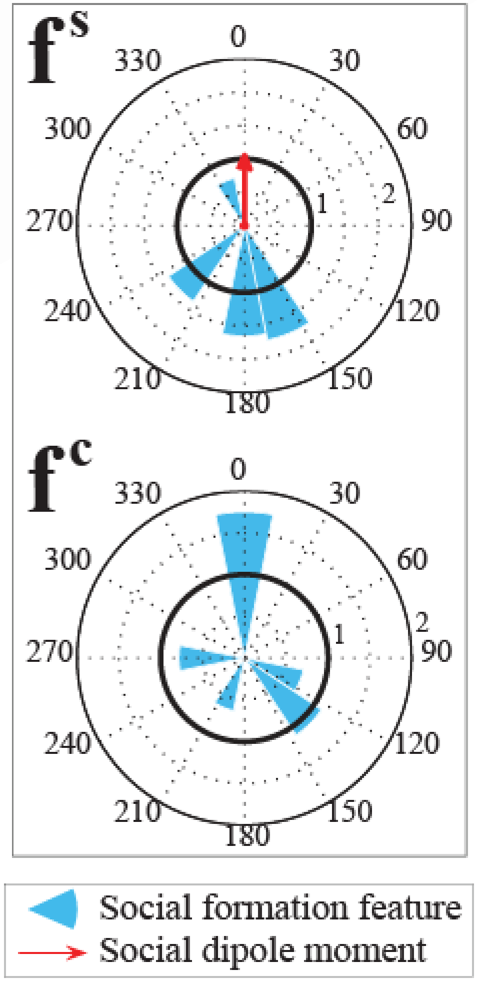
Social Formation Feature



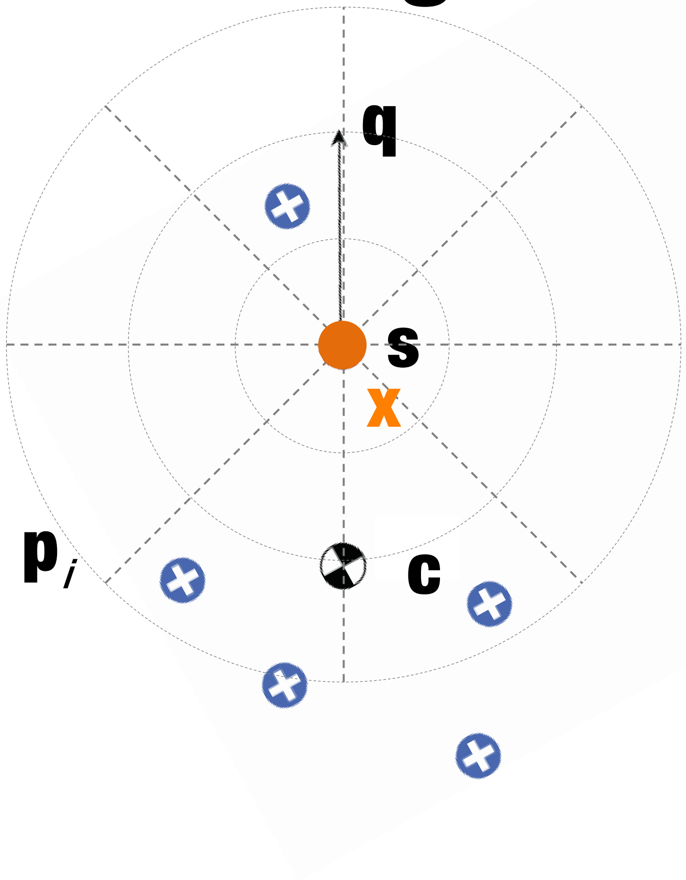


Social Formation Feature

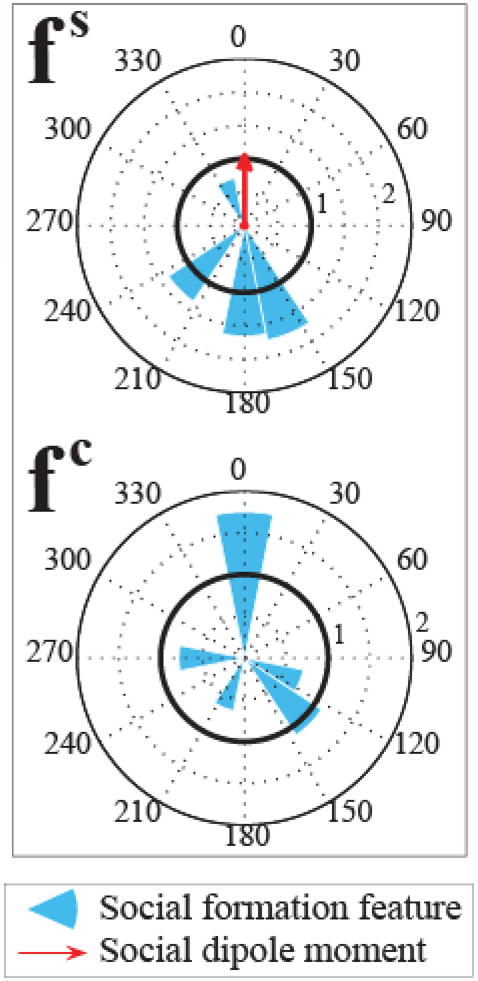




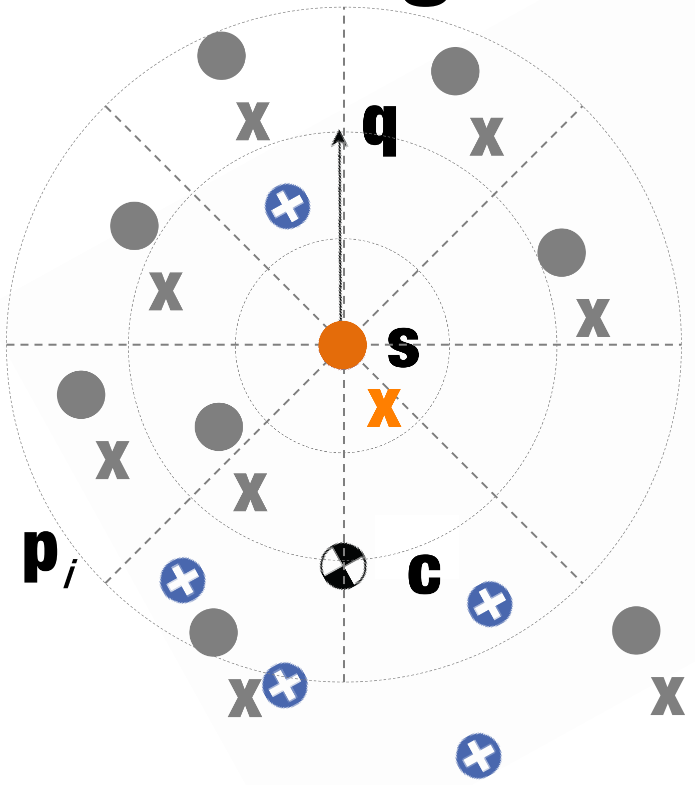
Learning Likelihood of Joint Attention

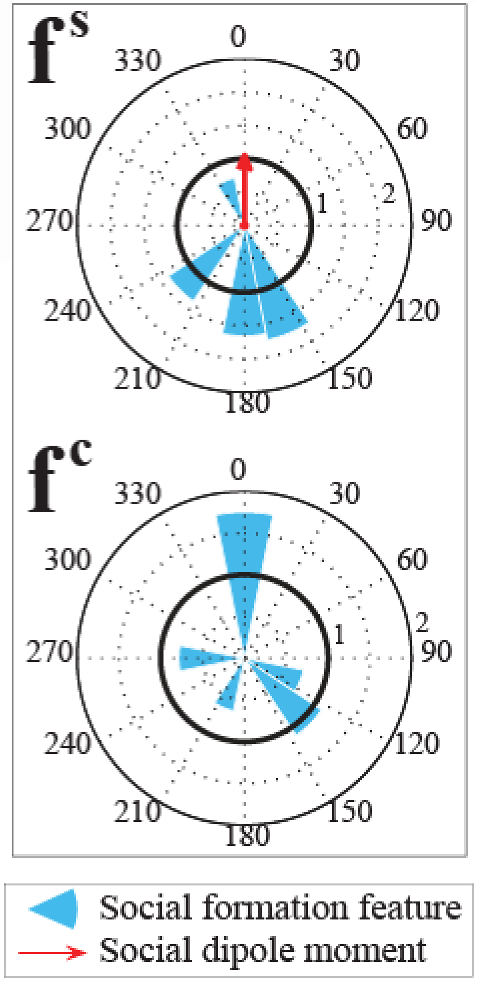


$$\Phi\left(\mathbf{f}^{c}, \mathbf{f}^{s}; \mathbf{x} = \mathbf{s}\right) = 1$$



Learning Likelihood of Joint Attention



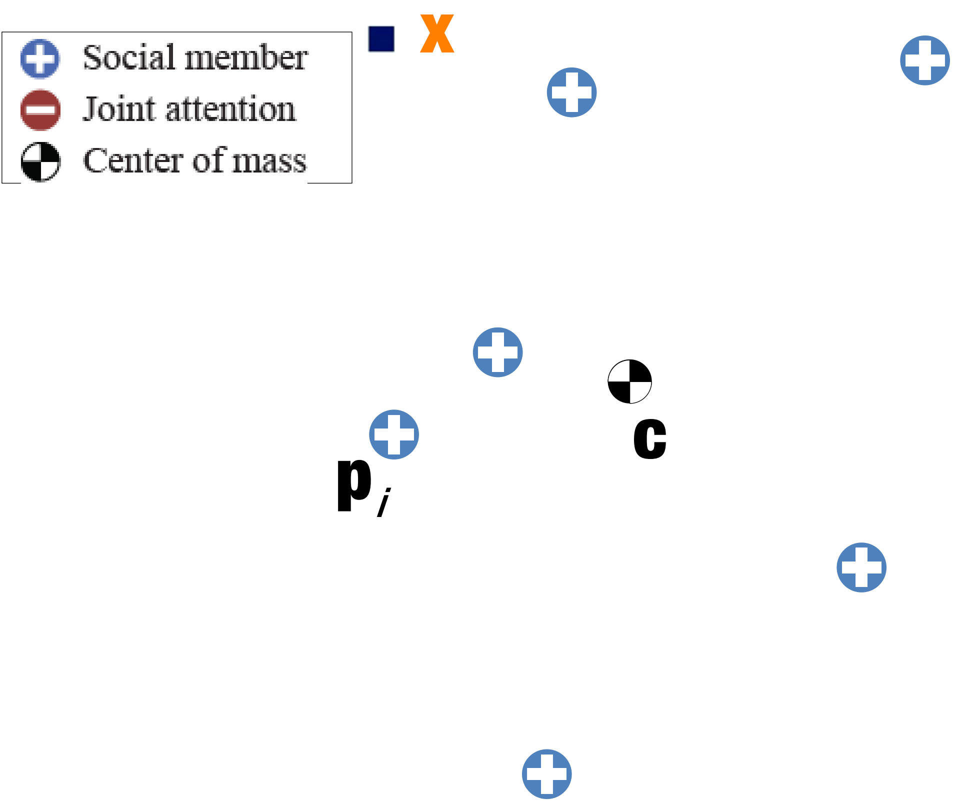


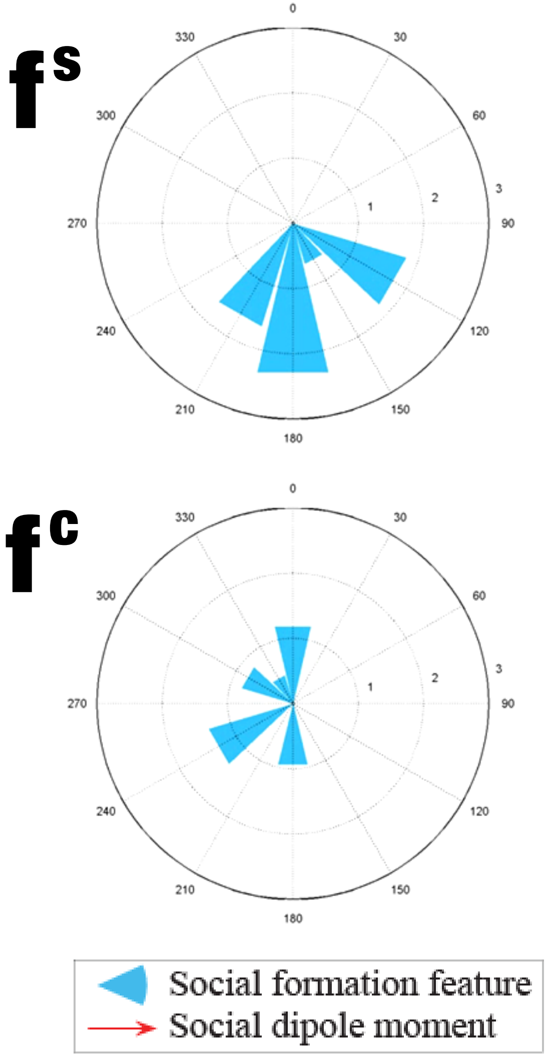
$$\Phi\left(\mathbf{f}^c, \mathbf{f}^s; \mathbf{x} = \mathbf{s}\right) = 1$$

$$\Phi\left(\mathbf{f}^c, \mathbf{f}^s; \mathbf{x} \neq \mathbf{s}\right) = 0$$

AdaBoost binary classifier

Joint Attention Prediction



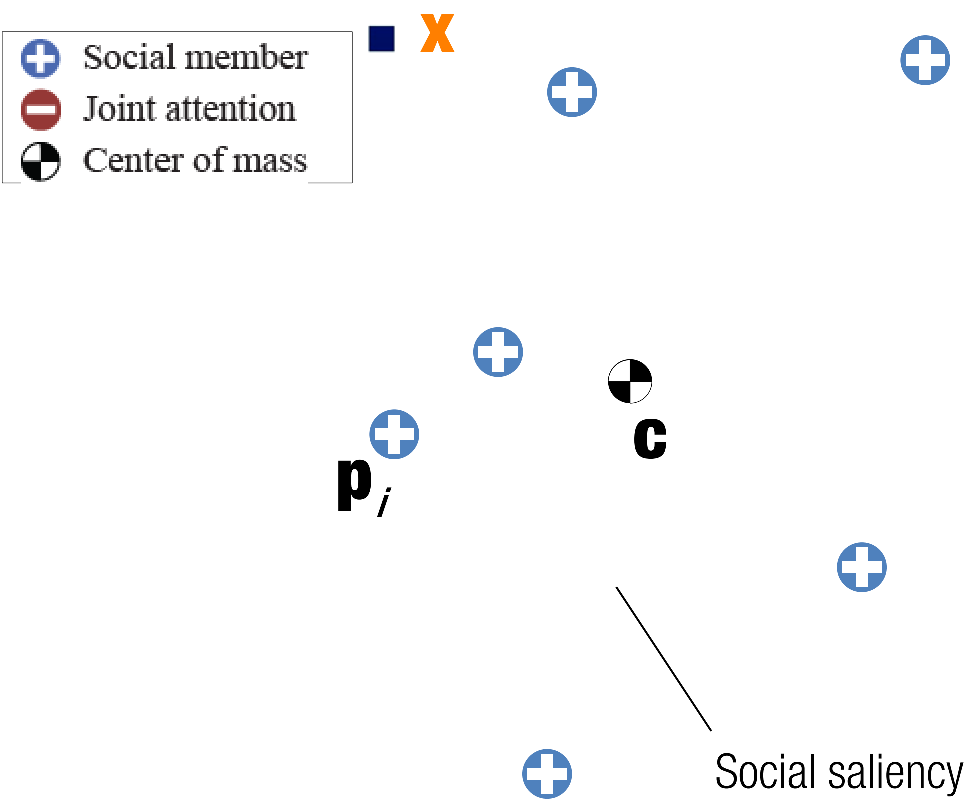


$$\Phi\left(\mathbf{f}^c, \mathbf{f}^s; \mathbf{x} = \mathbf{s}\right) = 1$$

$$\Phi\left(\mathbf{f}^c, \mathbf{f}^s; \mathbf{x} \neq \mathbf{s}\right) = 0$$

AdaBoost binary classifier

Joint Attention Prediction

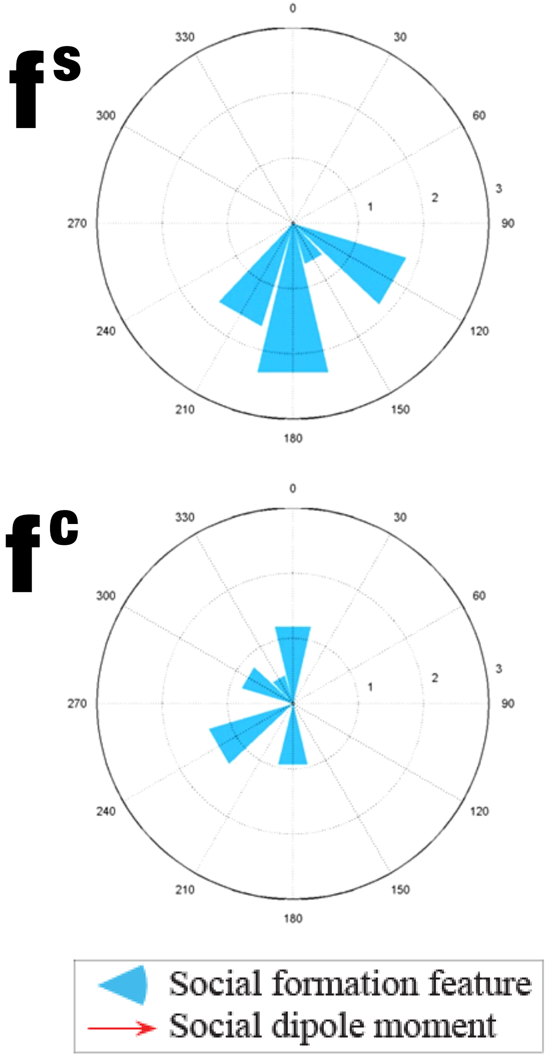


Social saliency:
Likelihood of joint attention

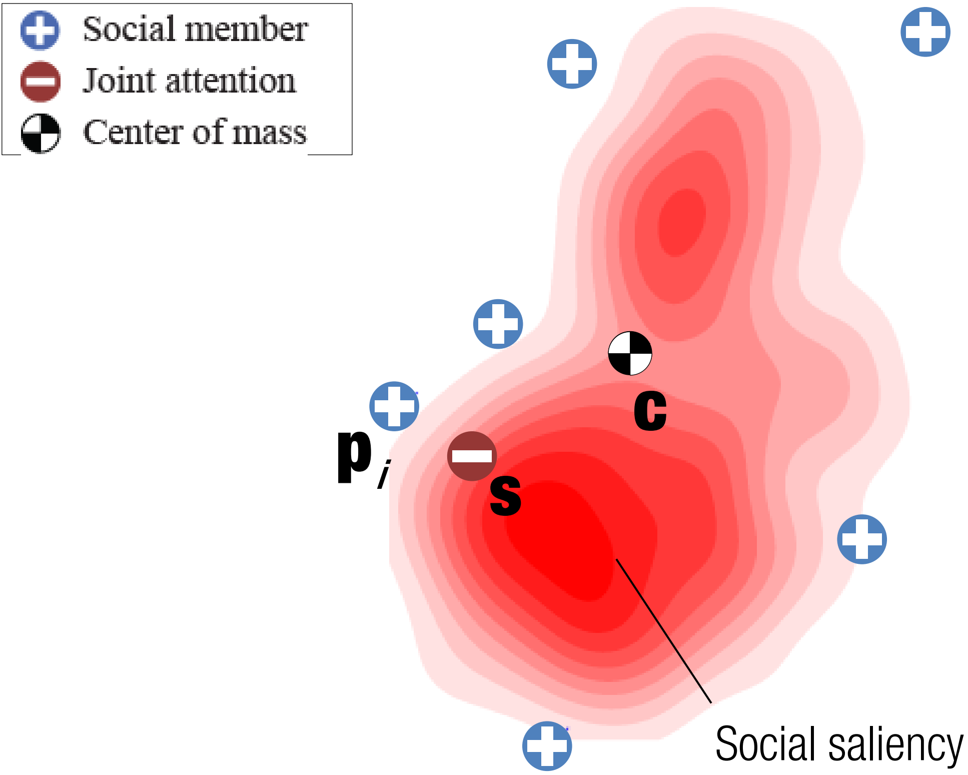
$$\Phi\left(\mathbf{f}^c, \mathbf{f}^s; \mathbf{x} = \mathbf{s}\right) = 1$$

$$\Phi\left(\mathbf{f}^c, \mathbf{f}^s; \mathbf{x} \neq \mathbf{s}\right) = 0$$

AdaBoost binary classifier



Joint Attention Prediction

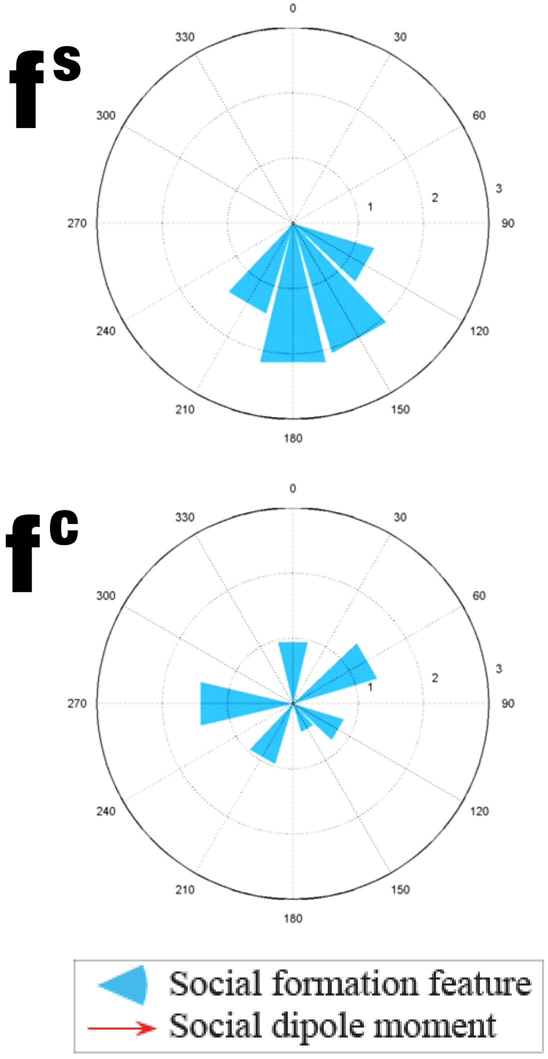


Social saliency:
Likelihood of joint attention

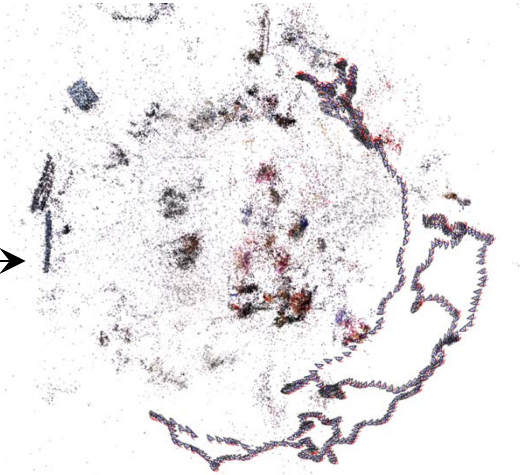
$$\Phi\left(\mathbf{f}^{\mathbf{c}}, \mathbf{f}^{\mathbf{s}}; \mathbf{x} = \mathbf{s}\right) = 1$$

$$\Phi\left(\mathbf{f}^{\mathbf{c}}, \mathbf{f}^{\mathbf{s}}; \mathbf{x} \neq \mathbf{s}\right) = 0$$

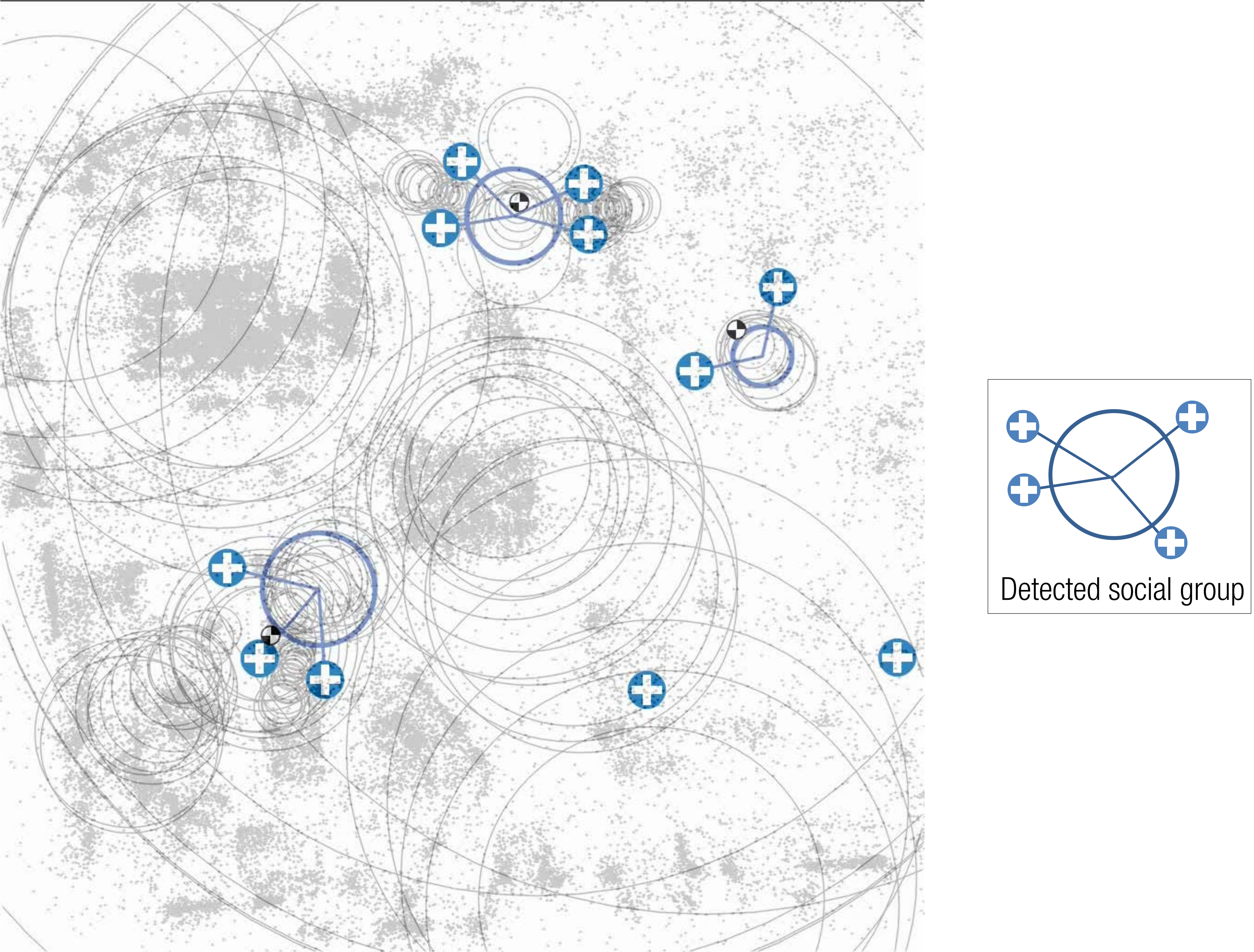
AdaBoost binary classifier

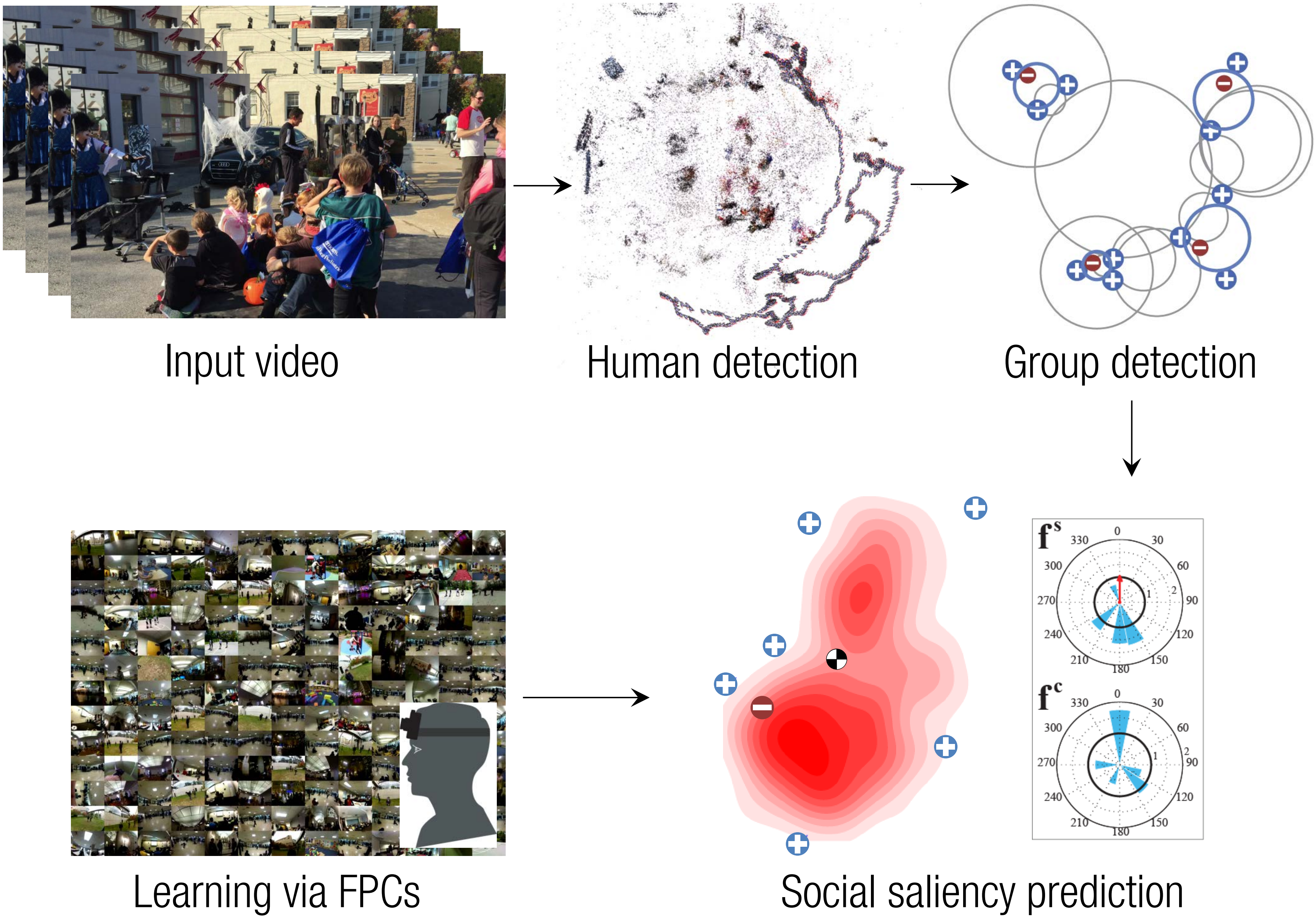


Input video



Human detection


Detected social group



Input video

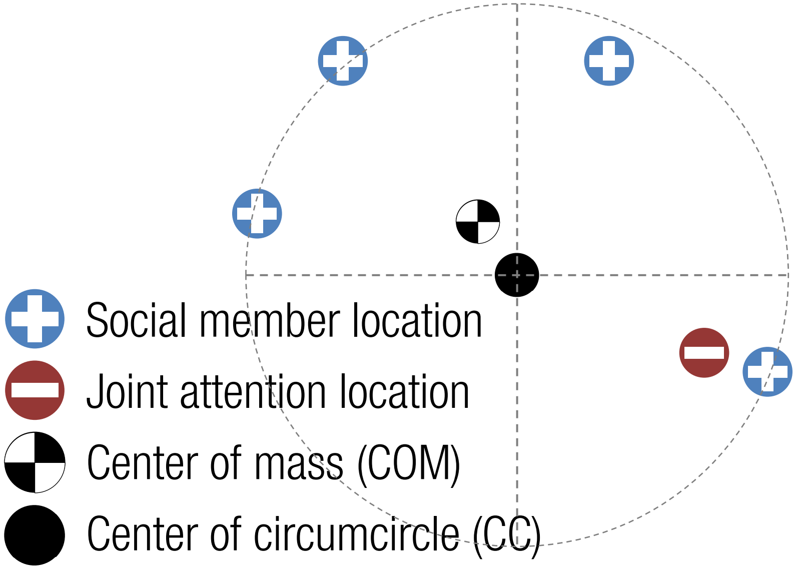
Human detection

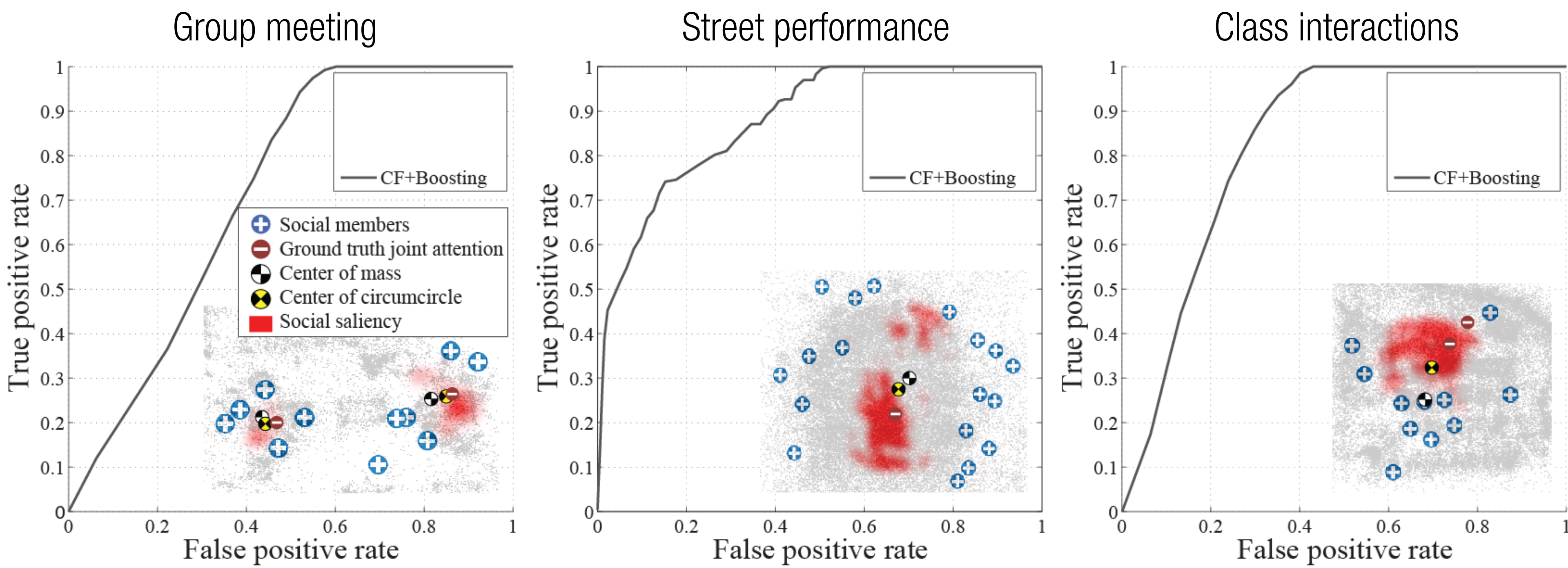
Group detection

Learning via FPCs

Social saliency prediction

Result

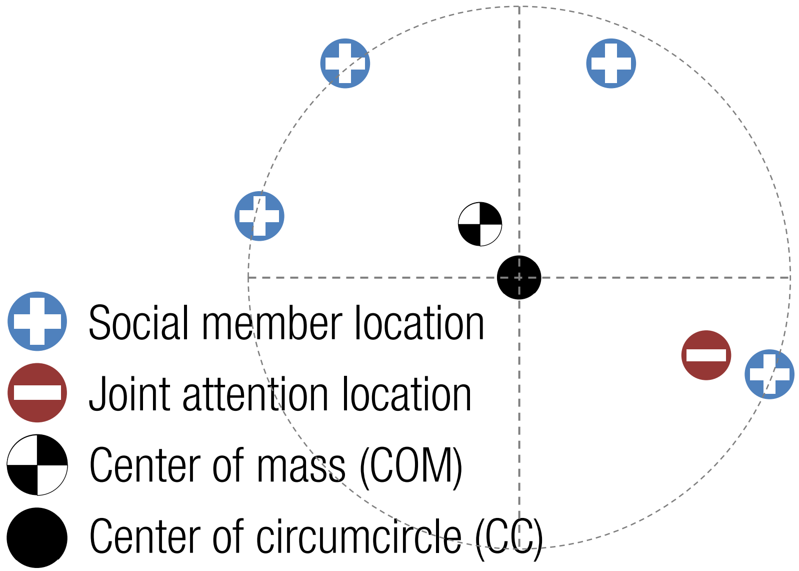

Social member location
Joint attention location
Center of mass (COM)
Center of circumcircle (CC)


Group meeting

Street performance

Class interactions

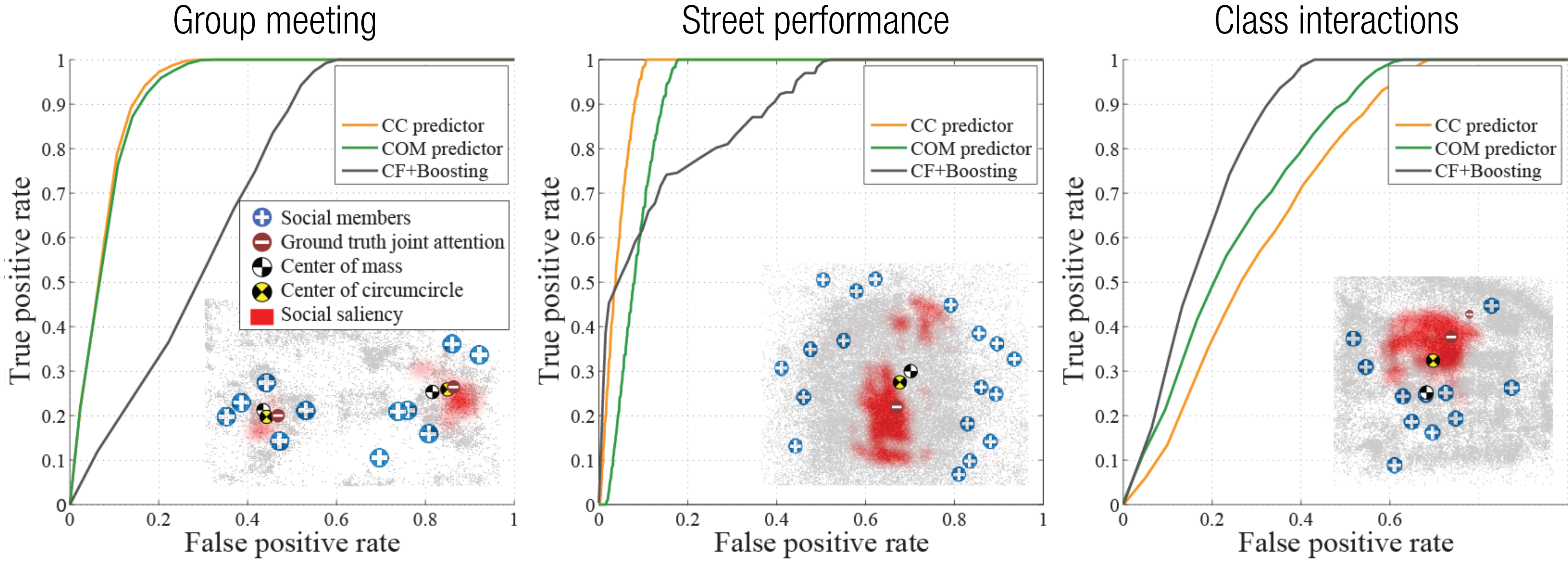
CF: Context feature (Lan et al., PAMI 2012)


Social member location

Joint attention location

Center of mass (COM)

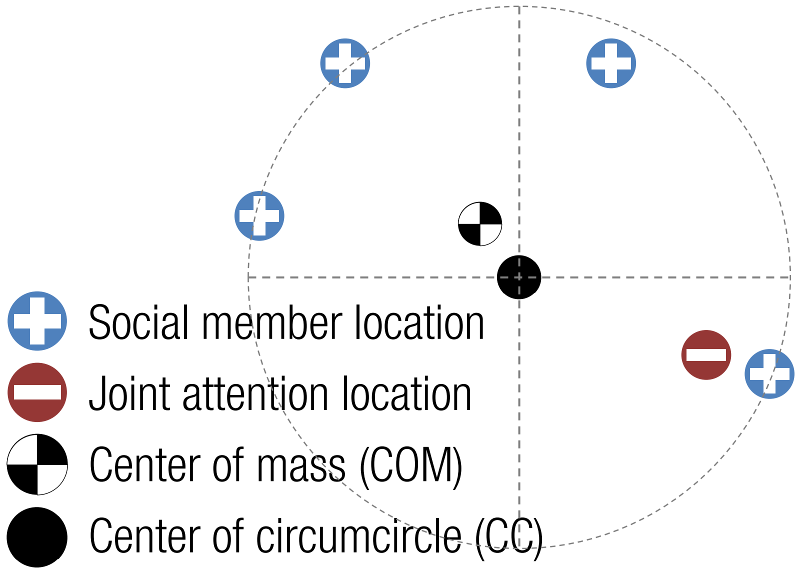
Center of circumcircle (CC)

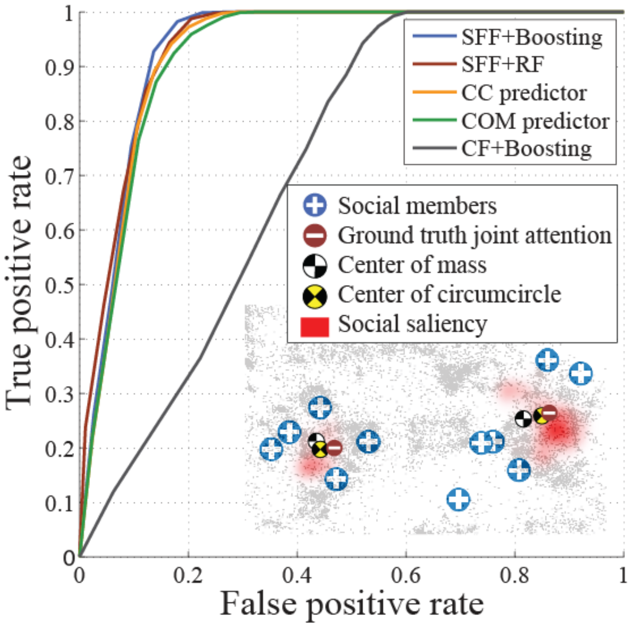

Group meeting

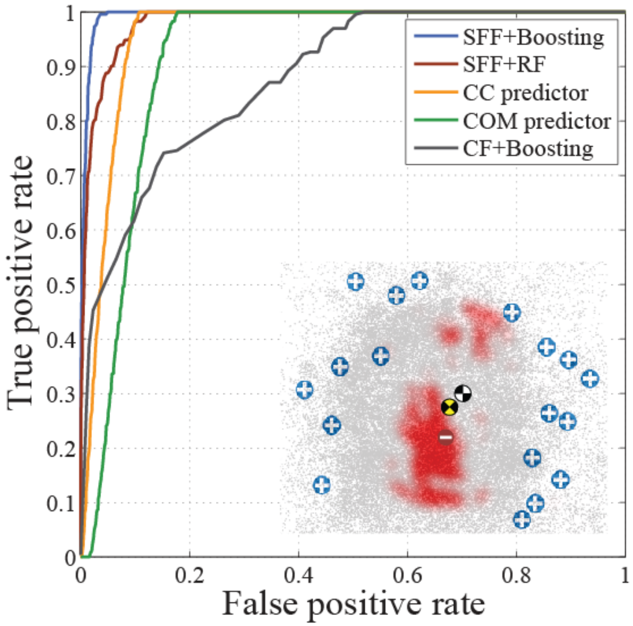
Street performance

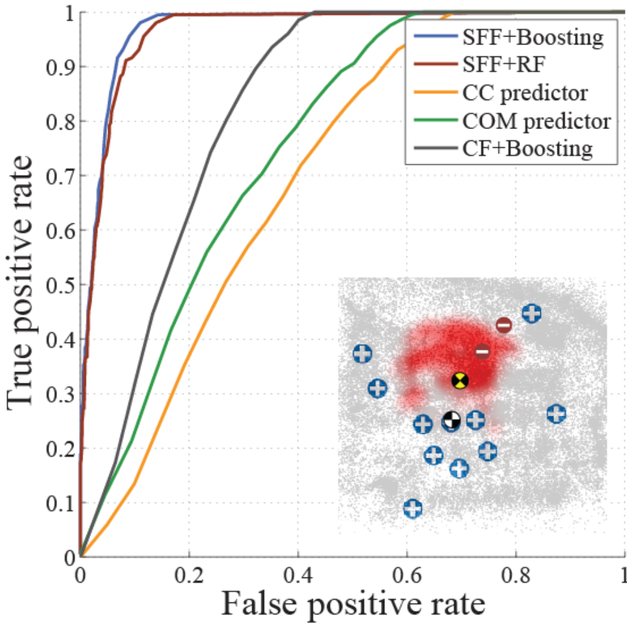
Class interactions

CF: Context feature (Lan et al., PAMI 2012)




Group meeting

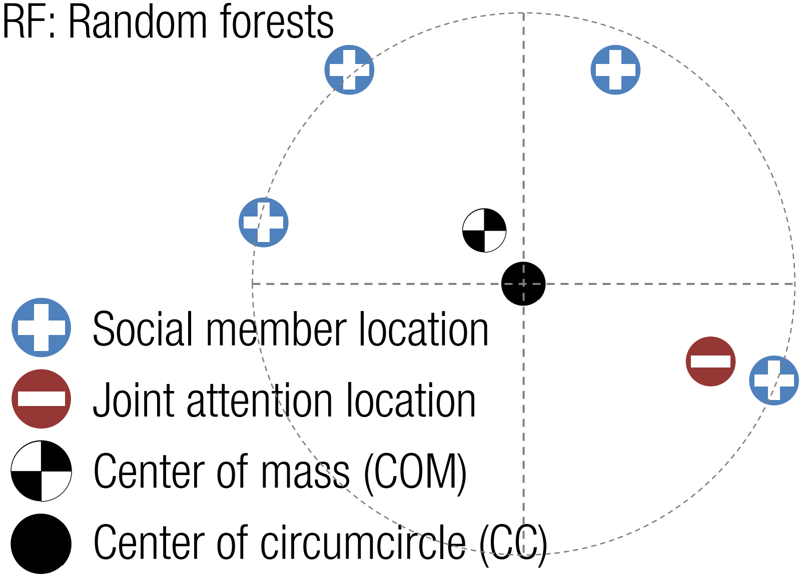

Street performance


Class interactions

CF: Context feature (Lan et al., PAMI 2012)

SSF: Social formation feature

RF: Random forests


Social member location

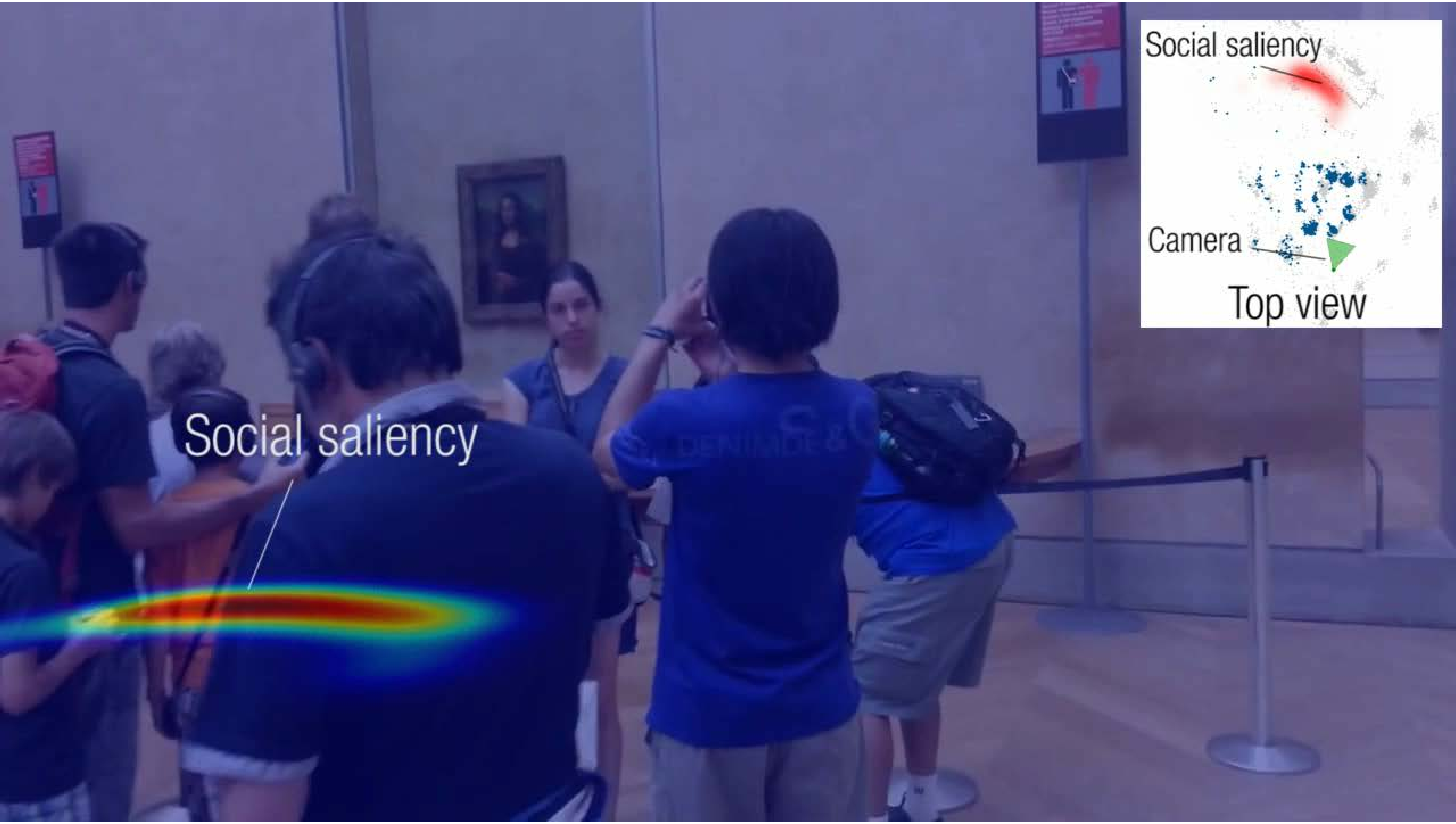
Joint attention location

Center of mass (COM)

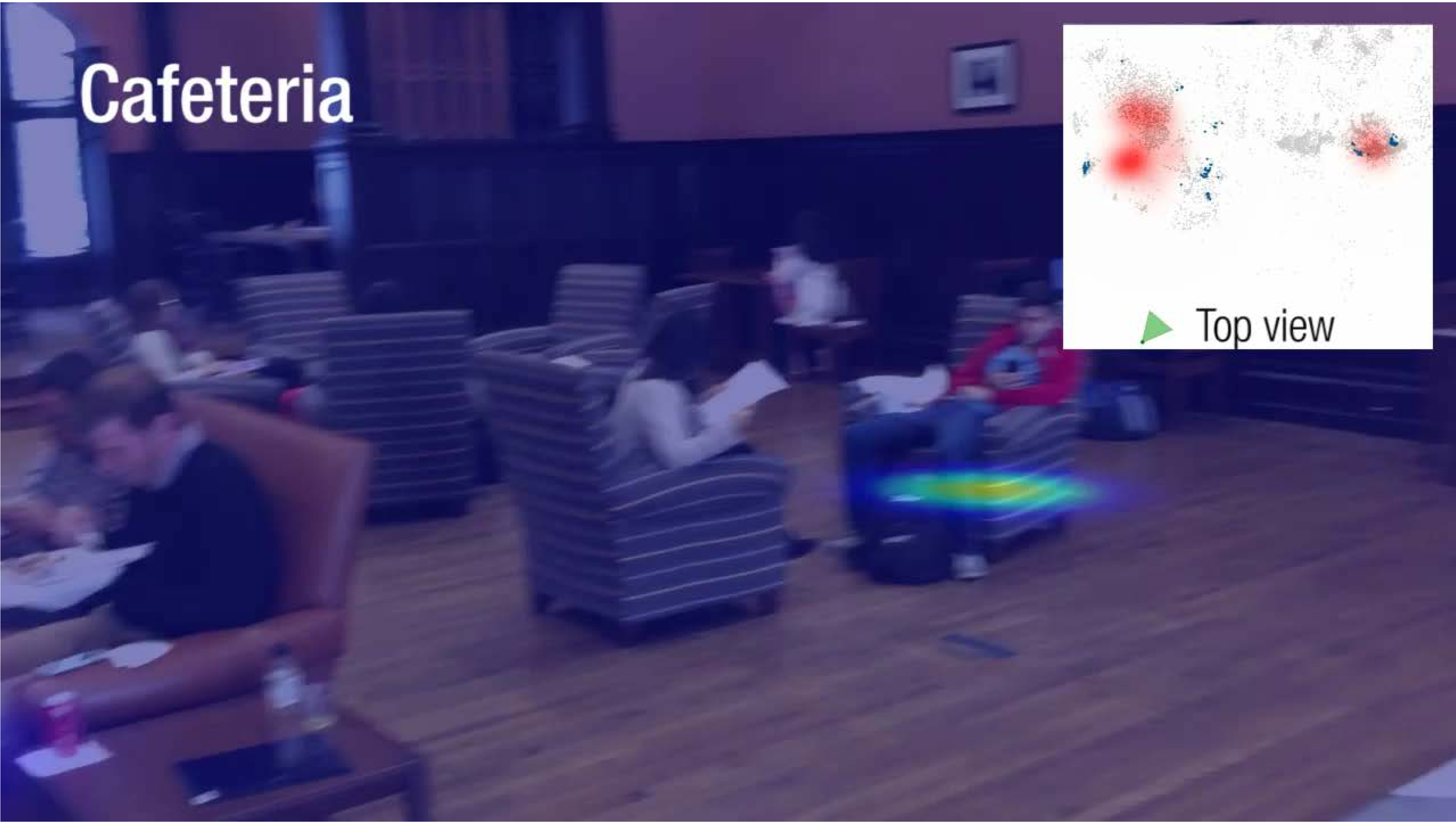
Center of circumcircle (CC)

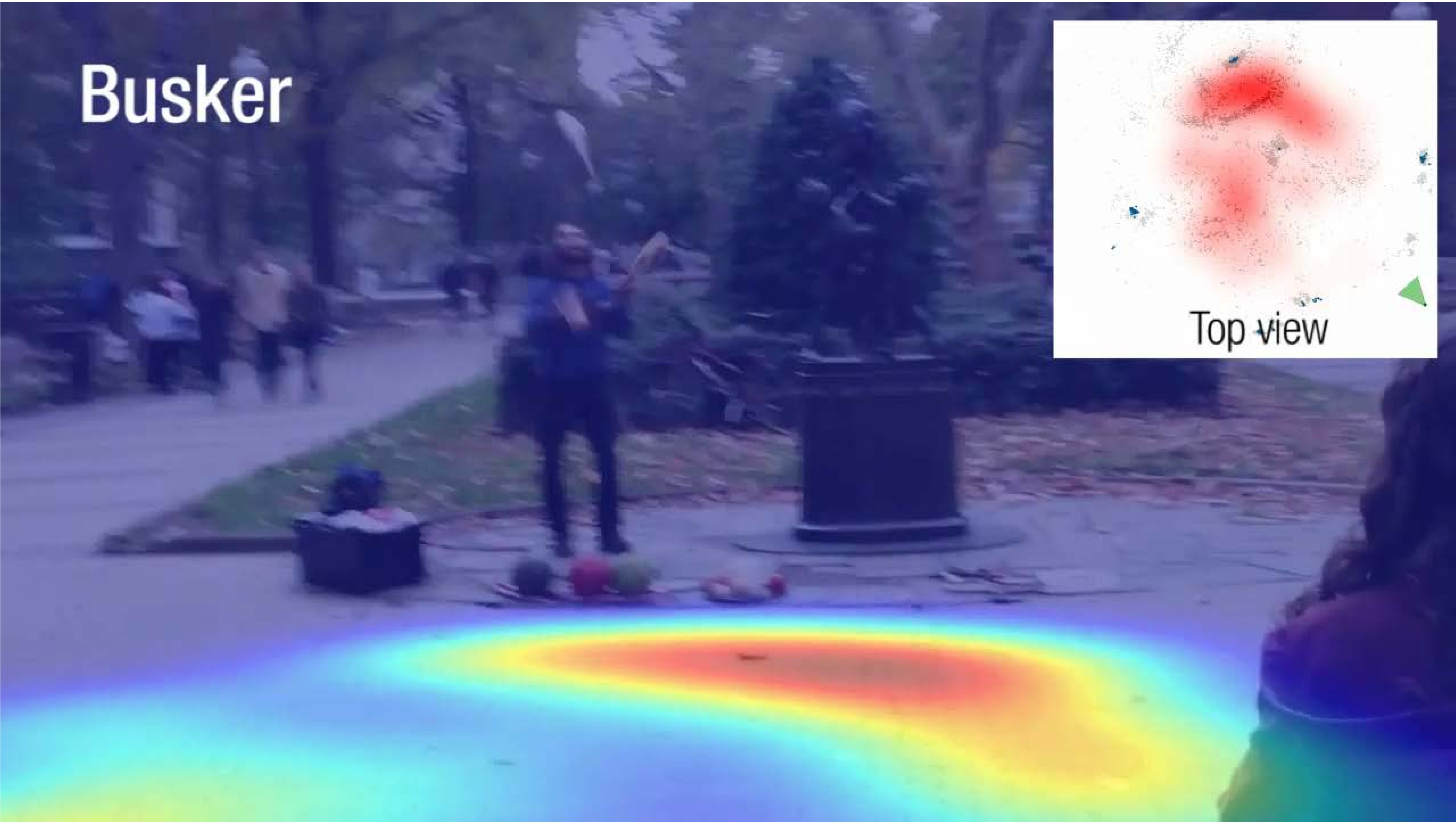
Scenes	SFF+Boosting	SFF+RF	CC	COM	CF
Dance	0.2769	0.1381	**0.3299**	0.0419	0.0106
Meeting I	0.2941	**0.3599**	0.2418	0.2350	0.0649
B-boy I	**0.7178**	0.6907	0.2078	0.1232	0.1225
Class	**0.7678**	0.7386	0.1445	0.2757	0.1873
Busker	0.2919	0.2059	**0.3432**	0.1929	0.0103
Picnic	**0.1364**	0.1349	0.1115	0.1808	0.0244
Social game	**0.5425**	0.4419	0.3461	0.2463	0.0020

Mean average precision


Social saliency
Camera
Top view
Social saliency

Cafeteria



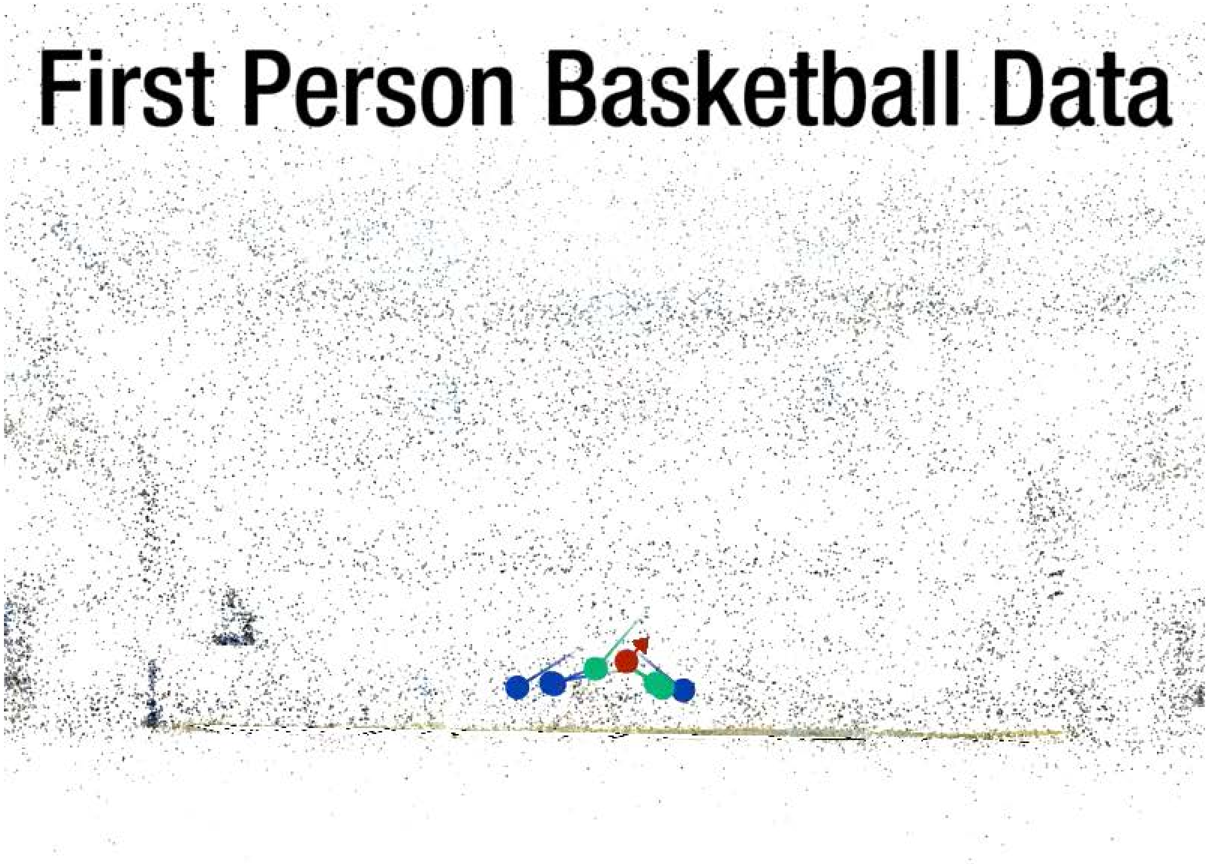

Busker
Top view

Time Square

Source: https://www.youtube.com/watch?v=ezyrSKgcyJw

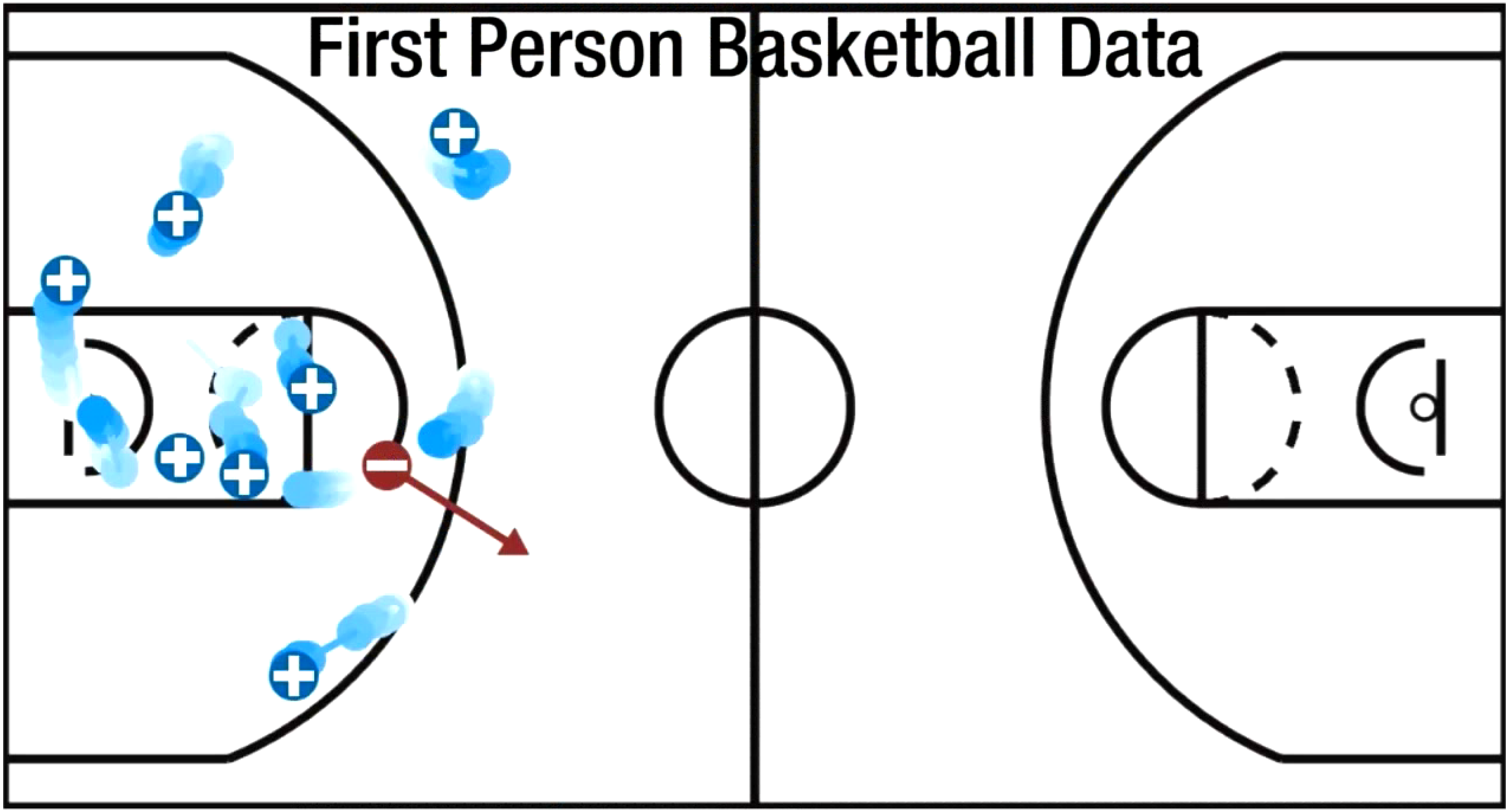
Basketball Scene Result

USA 5
NGR 0
9:25 1ST
Input Video
London 2012

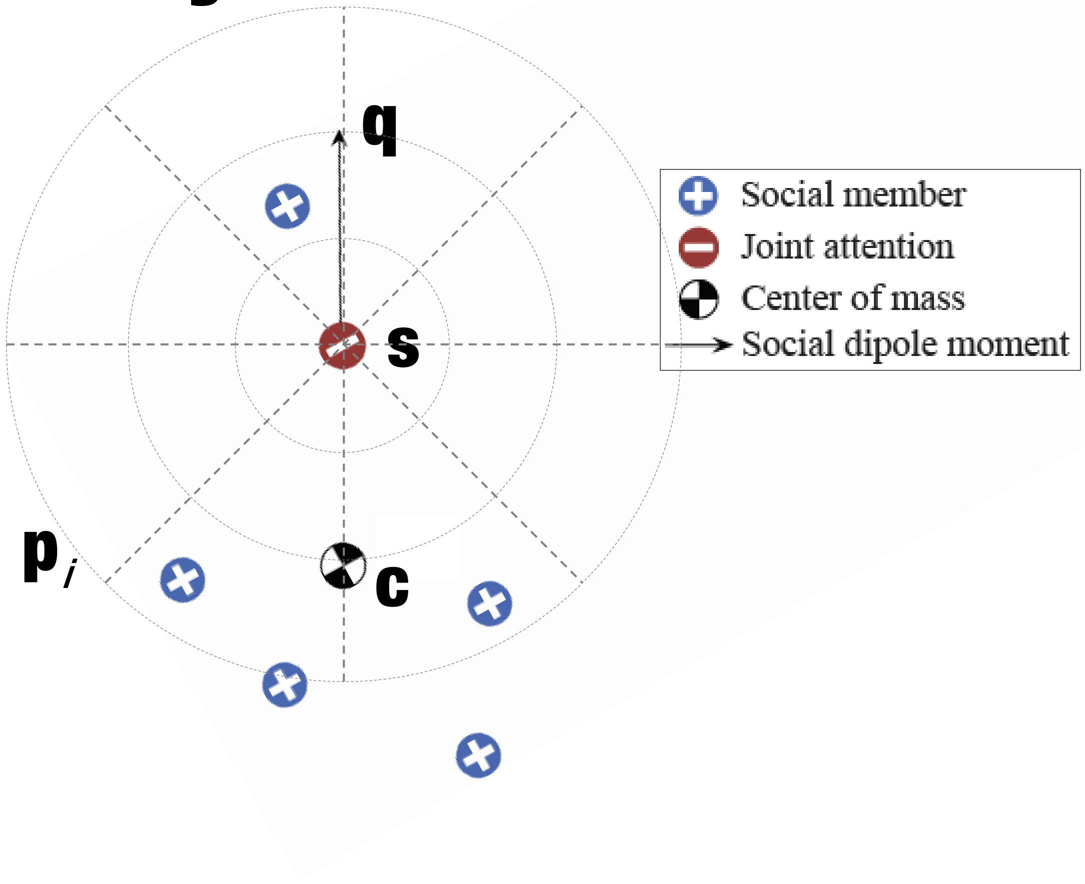


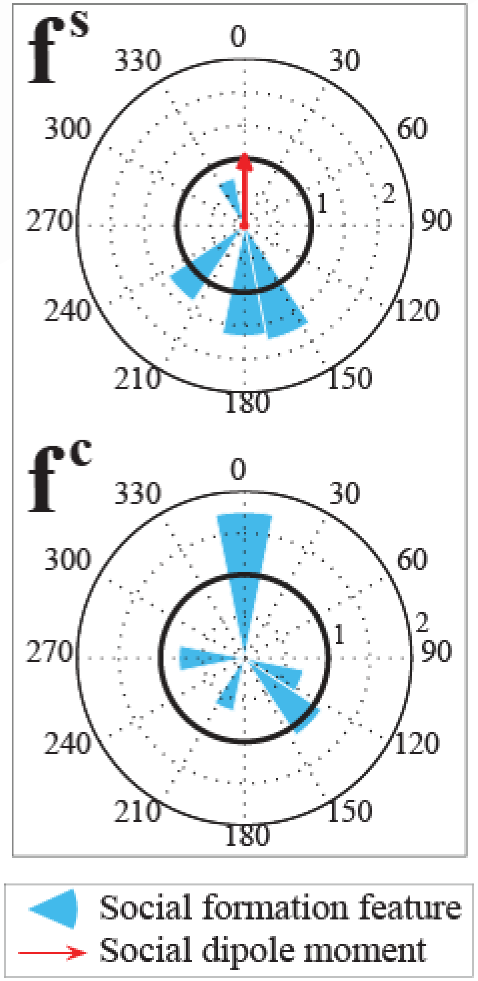
University team of Northwestern Polytechnical University (China)

First Person Basketball Data



Dynamic Joint Attention Prediction

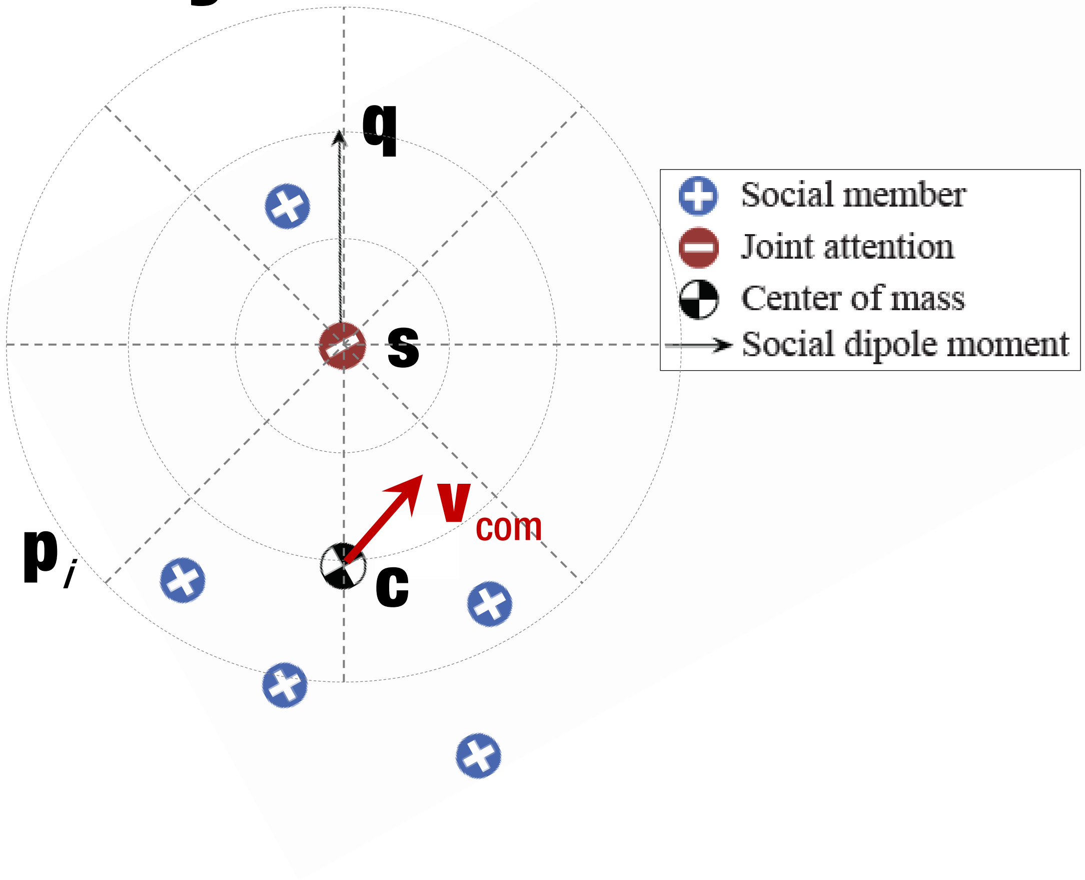


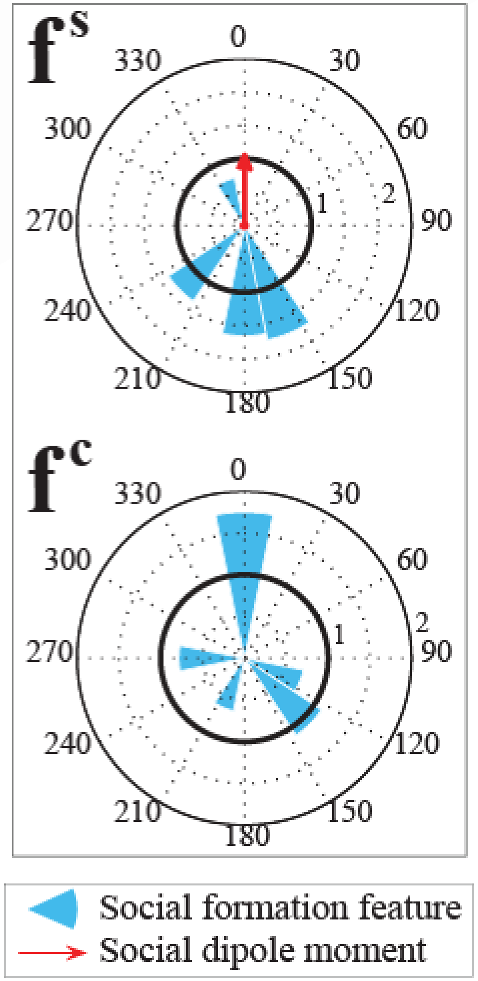


$$\Phi\left(\mathbf{f}^{c}, \mathbf{f}^{s}; \mathbf{x} = \mathbf{s}\right) = 1$$

$$\Phi\left(\mathbf{f}^{c}, \mathbf{f}^{s}; \mathbf{x} \neq \mathbf{s}\right) = 0$$

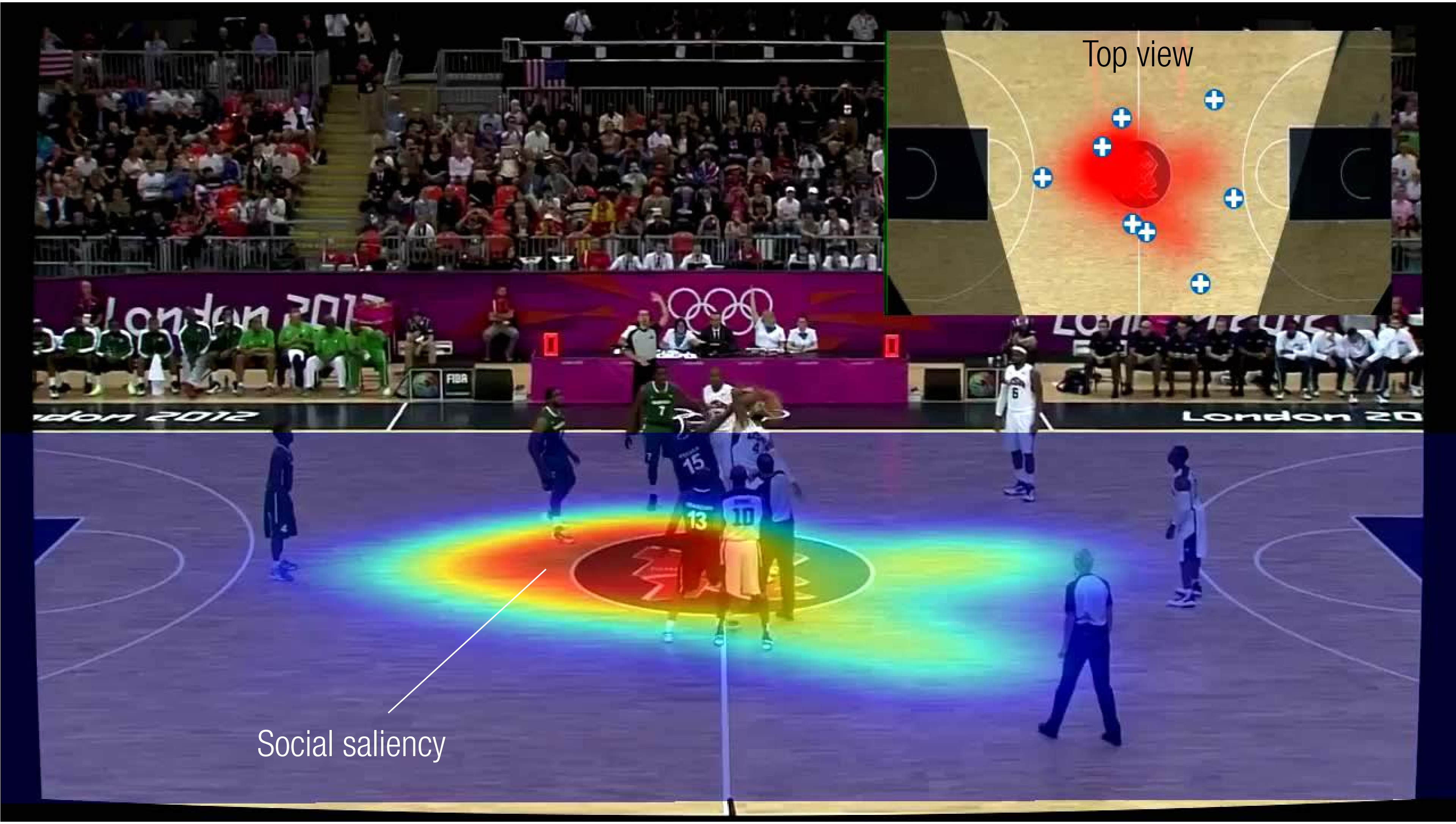
Dynamic Joint Attention Prediction





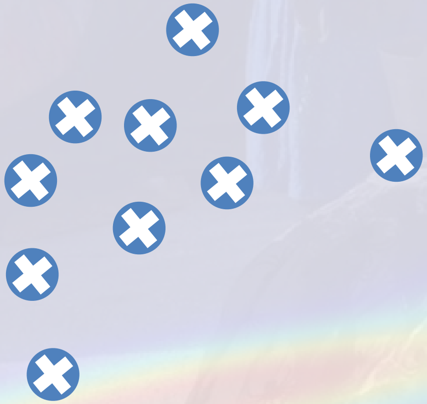
$$\Phi\left(\mathbf{f}^{c}, \mathbf{f}^{s}, \mathbf{v}_{com}; \mathbf{x} = \mathbf{s}\right) = 1$$

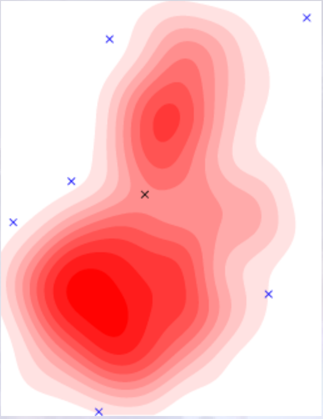
$$\Phi\left(\mathbf{f}^{c}, \mathbf{f}^{s}, \mathbf{v}_{com}; \mathbf{x} = \mathbf{s}\right) = 0$$



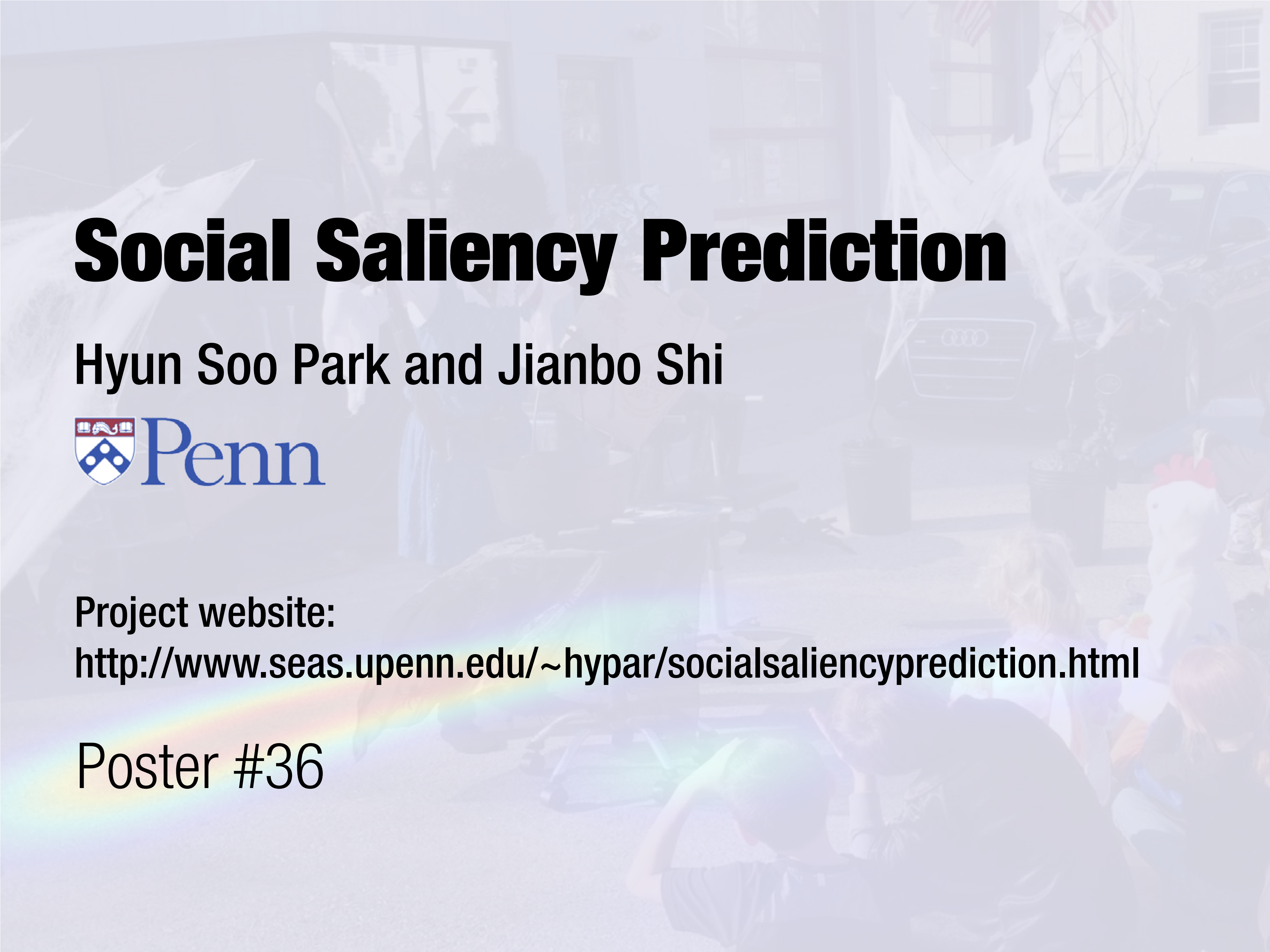
Person detector:
Yang and Ramanan, *Articulated Human Detection with Flexible Mixtures of Parts*, PAMI 2003.

Can we predict social saliency without measuring gaze directions?





Social formation ⟷ Social saliency



Social Saliency Prediction

Hyun Soo Park and Jianbo Shi

Penn

Project website:
http://www.seas.upenn.edu/~hypar/socialsaliencyprediction.html

Poster #36