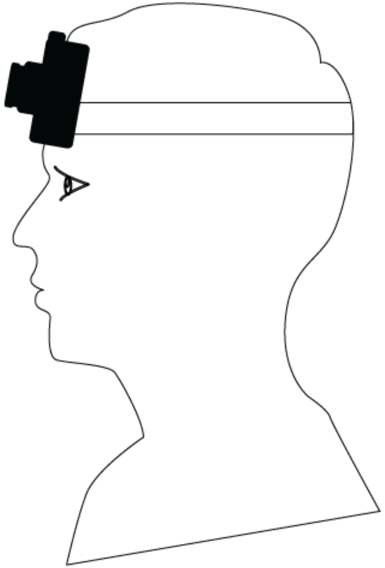
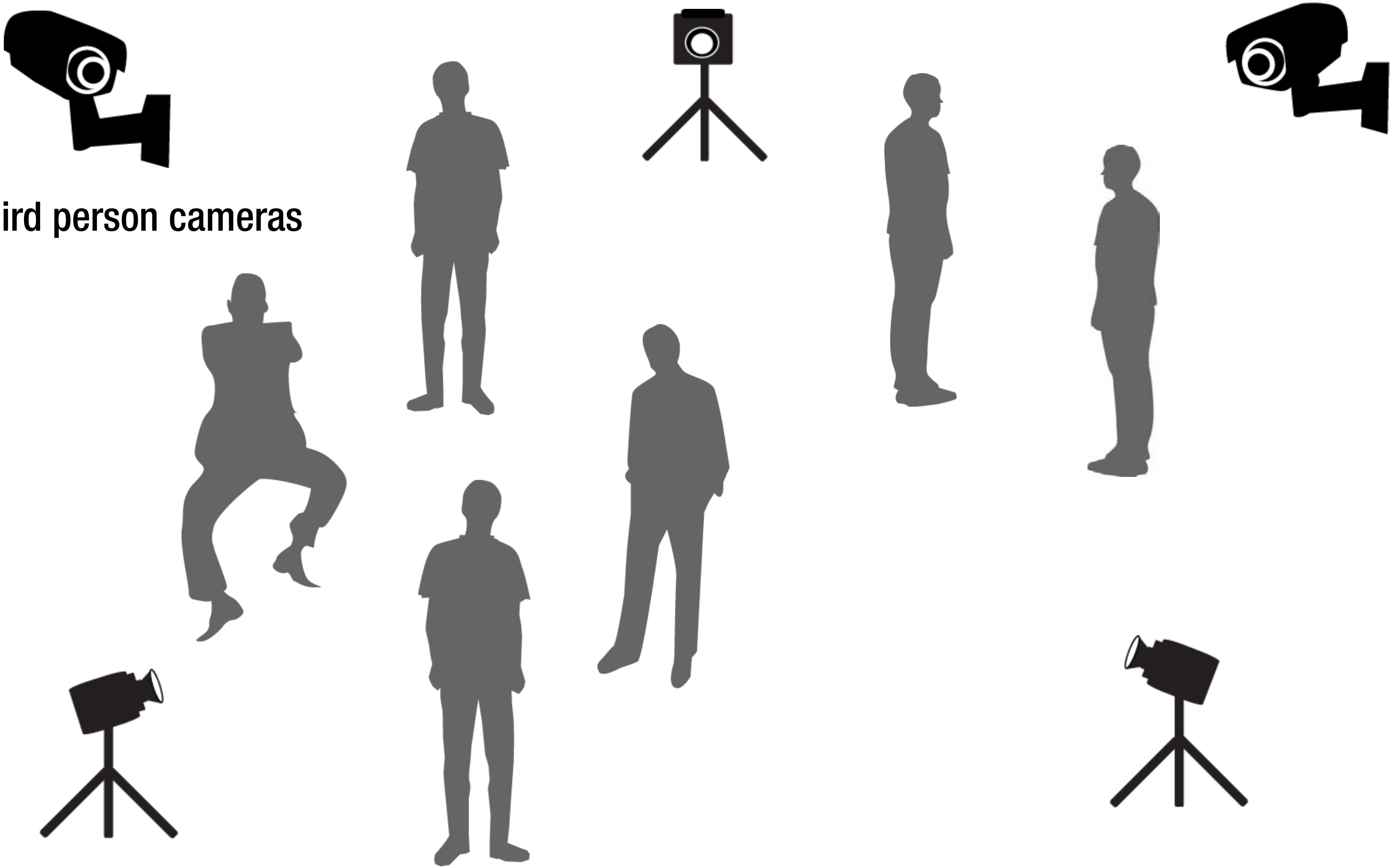
Social Attention

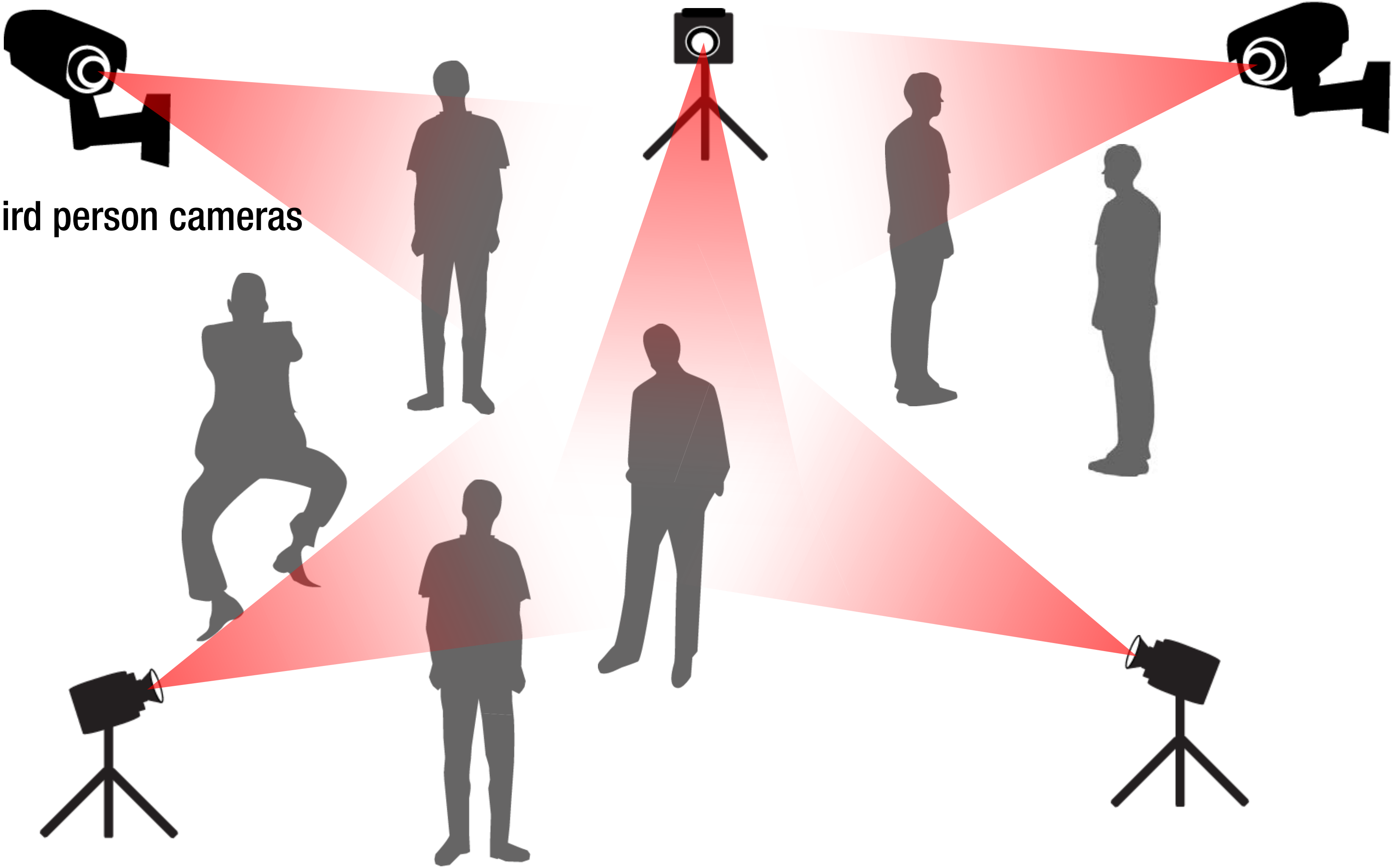
:What can first person cameras tell us about our social interactions?

First person cameras are ideal sensors to measure social behaviors.

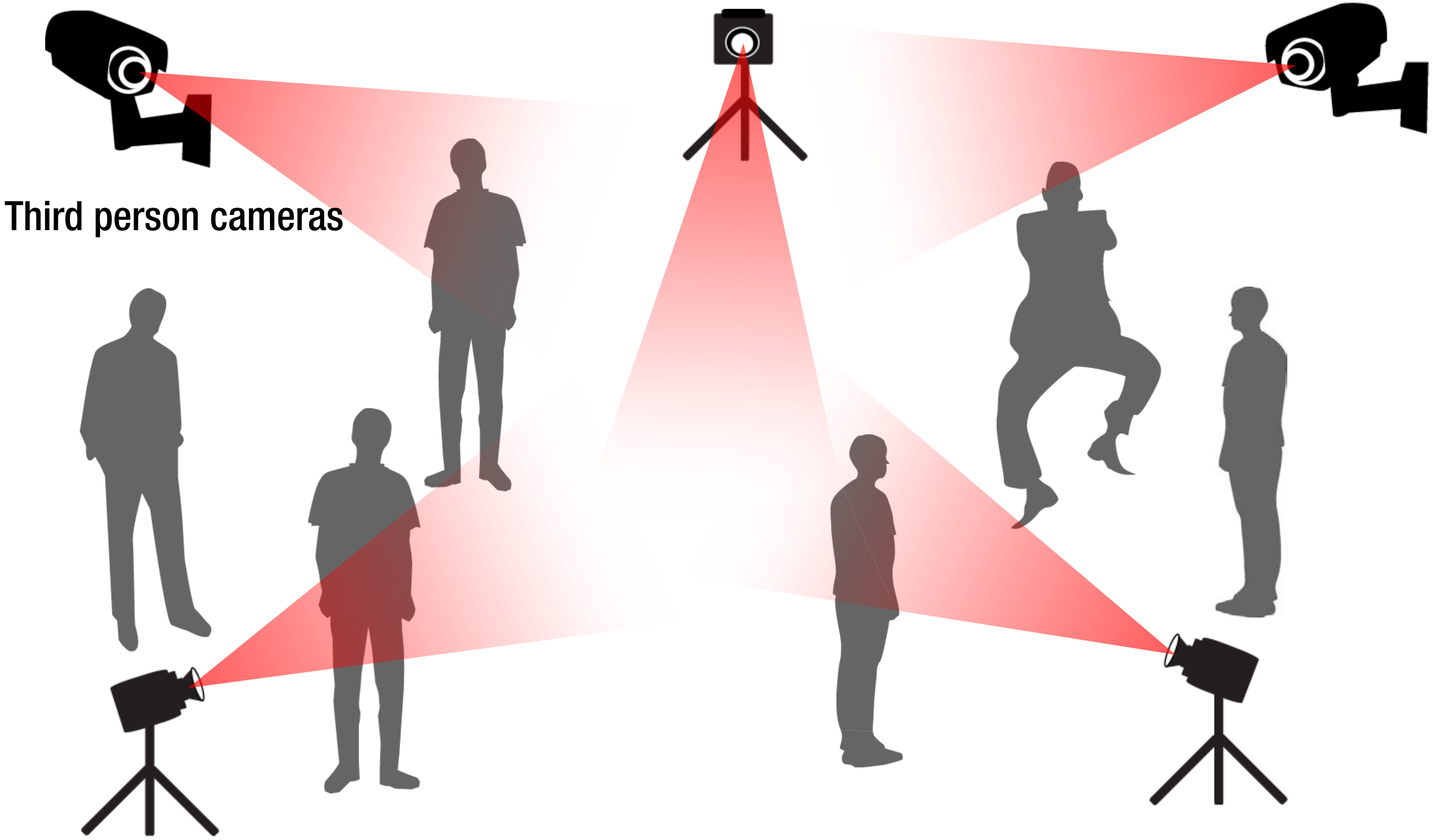




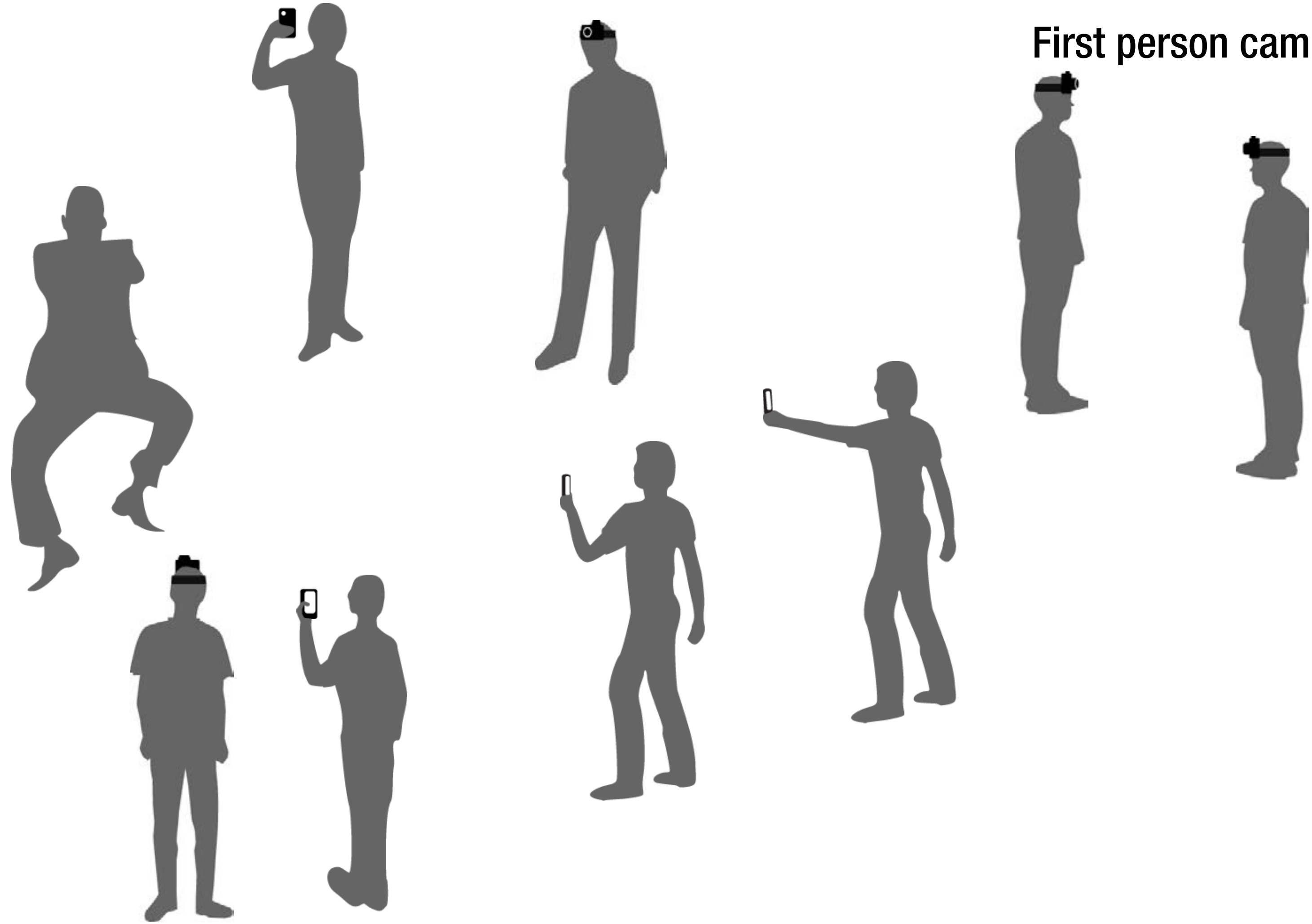
Third person cameras

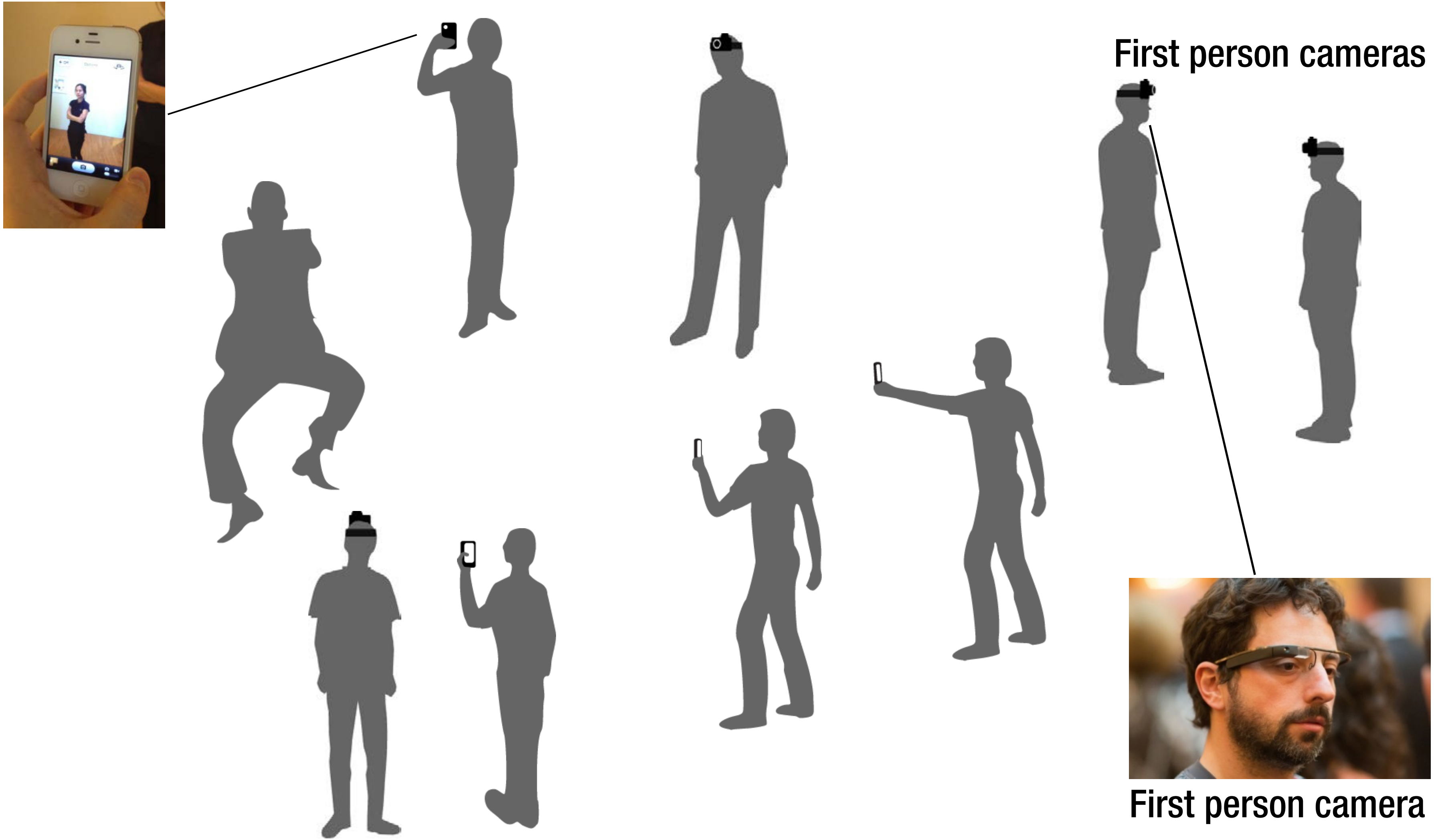


Third person cameras


Third person cameras

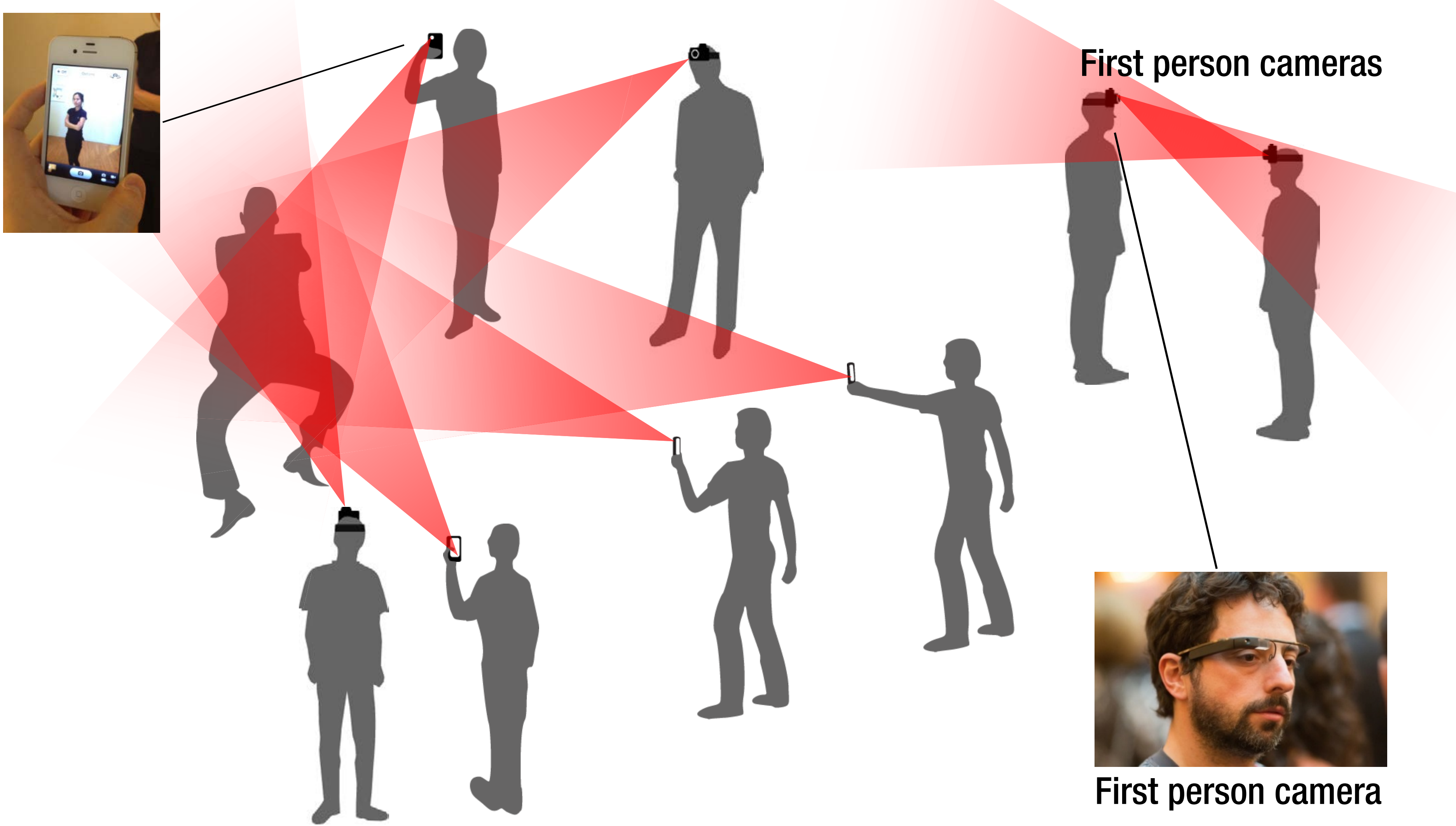
First person cameras





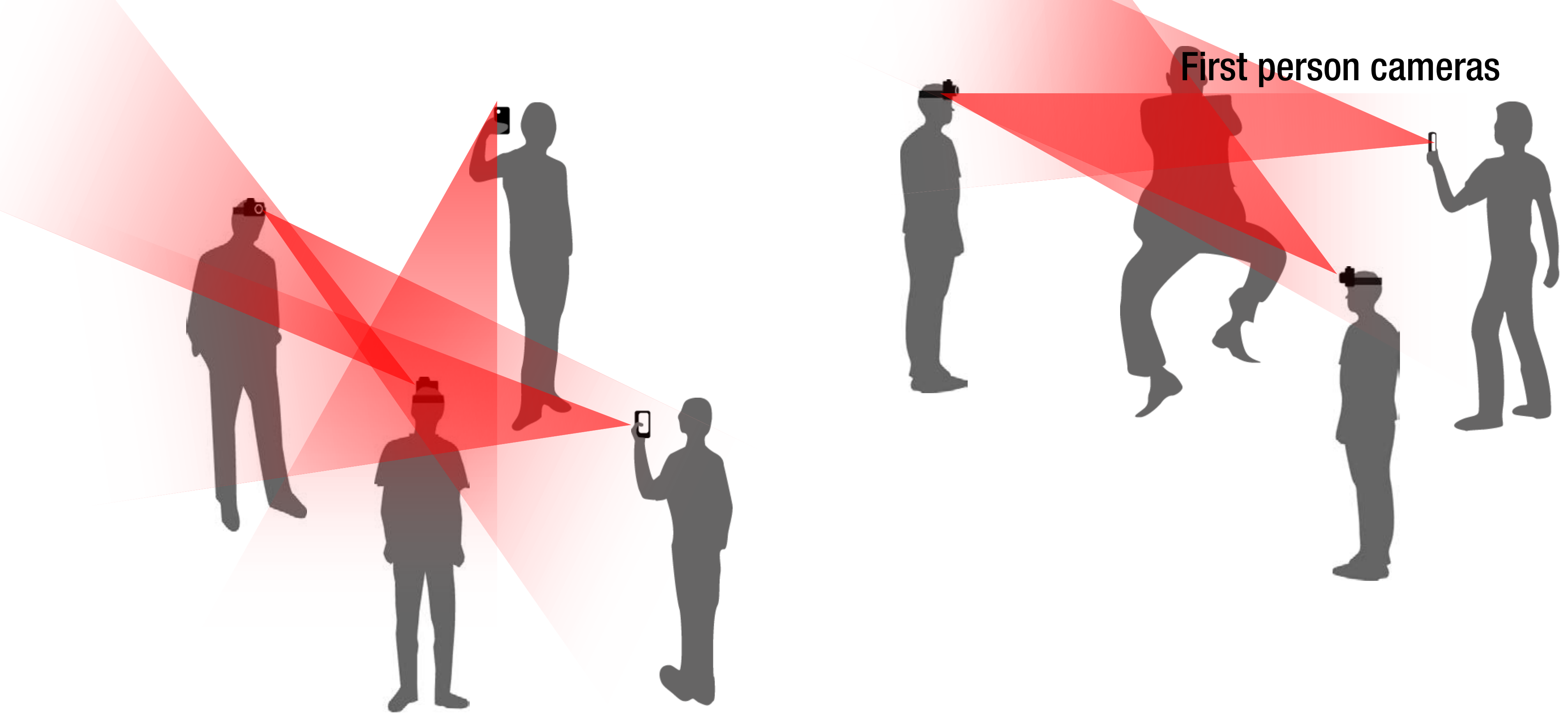
First person cameras

First person camera

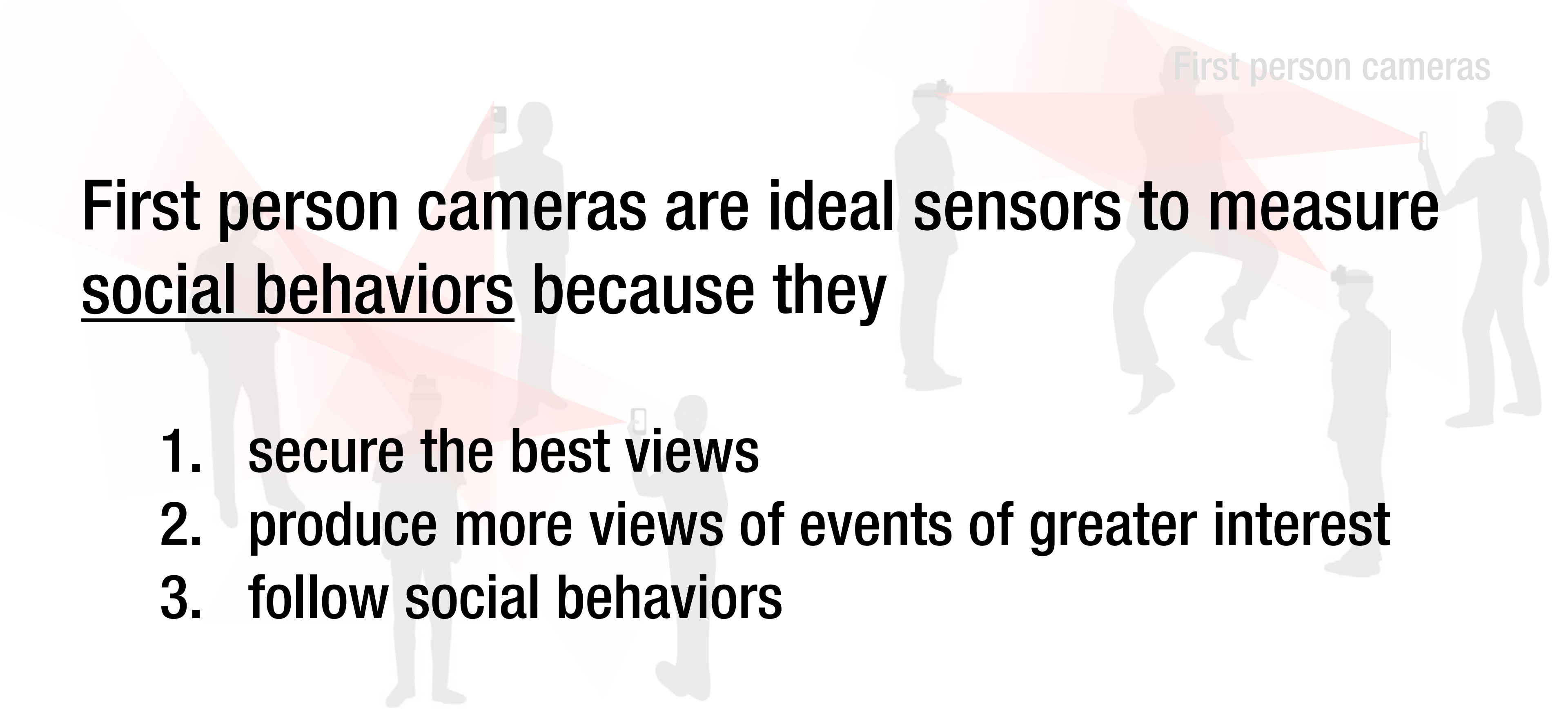


First person cameras

First person camera

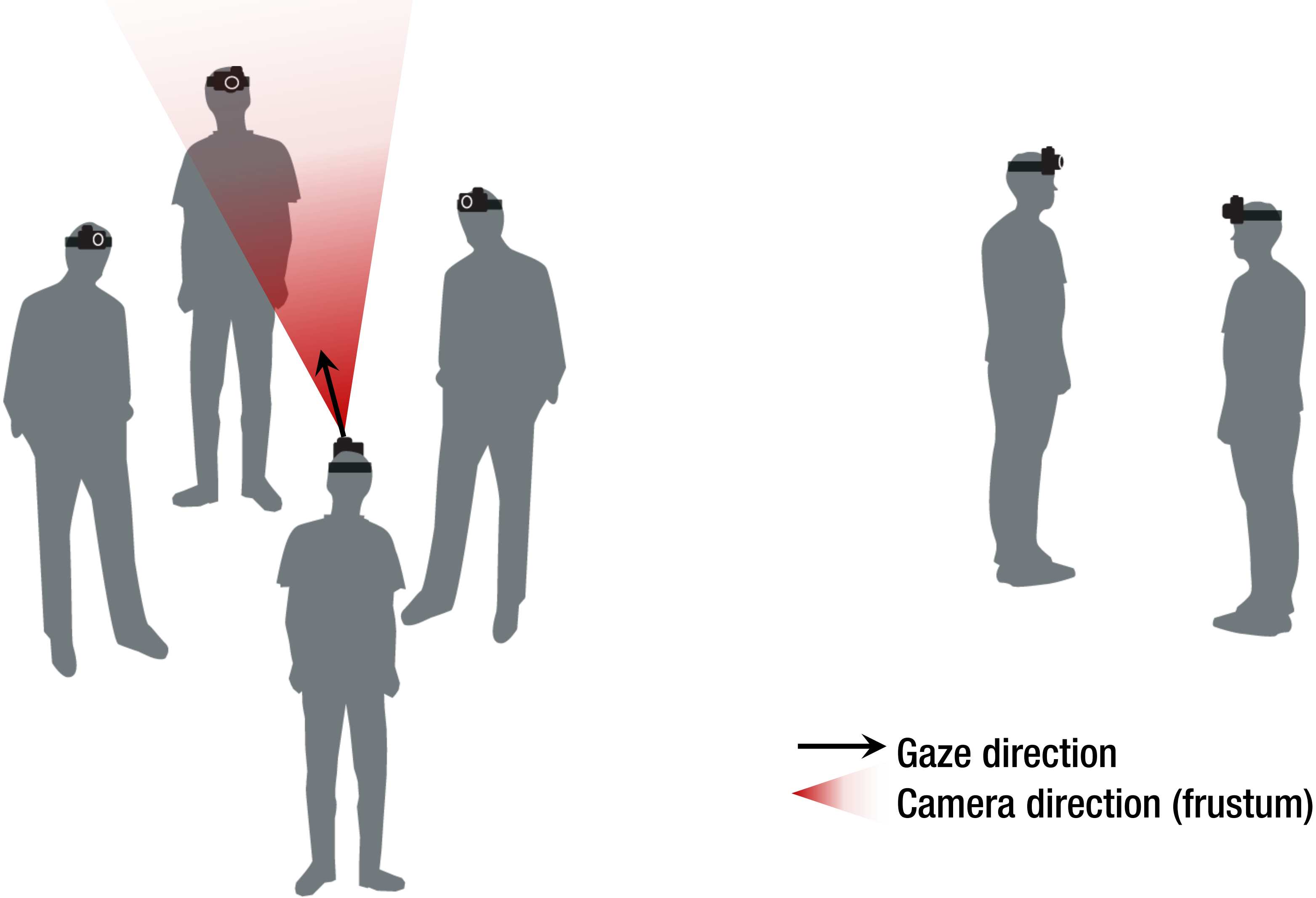


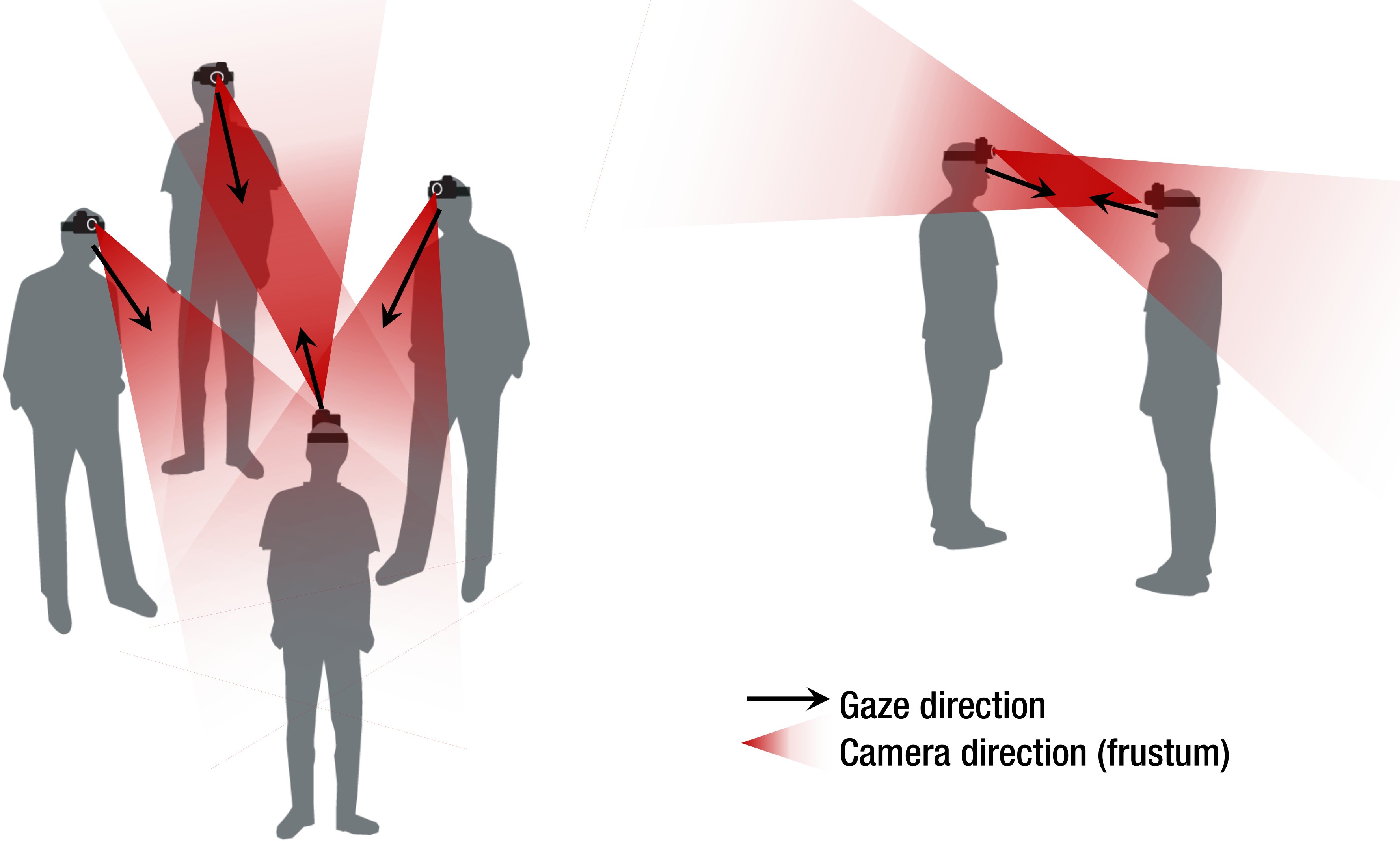
First person cameras

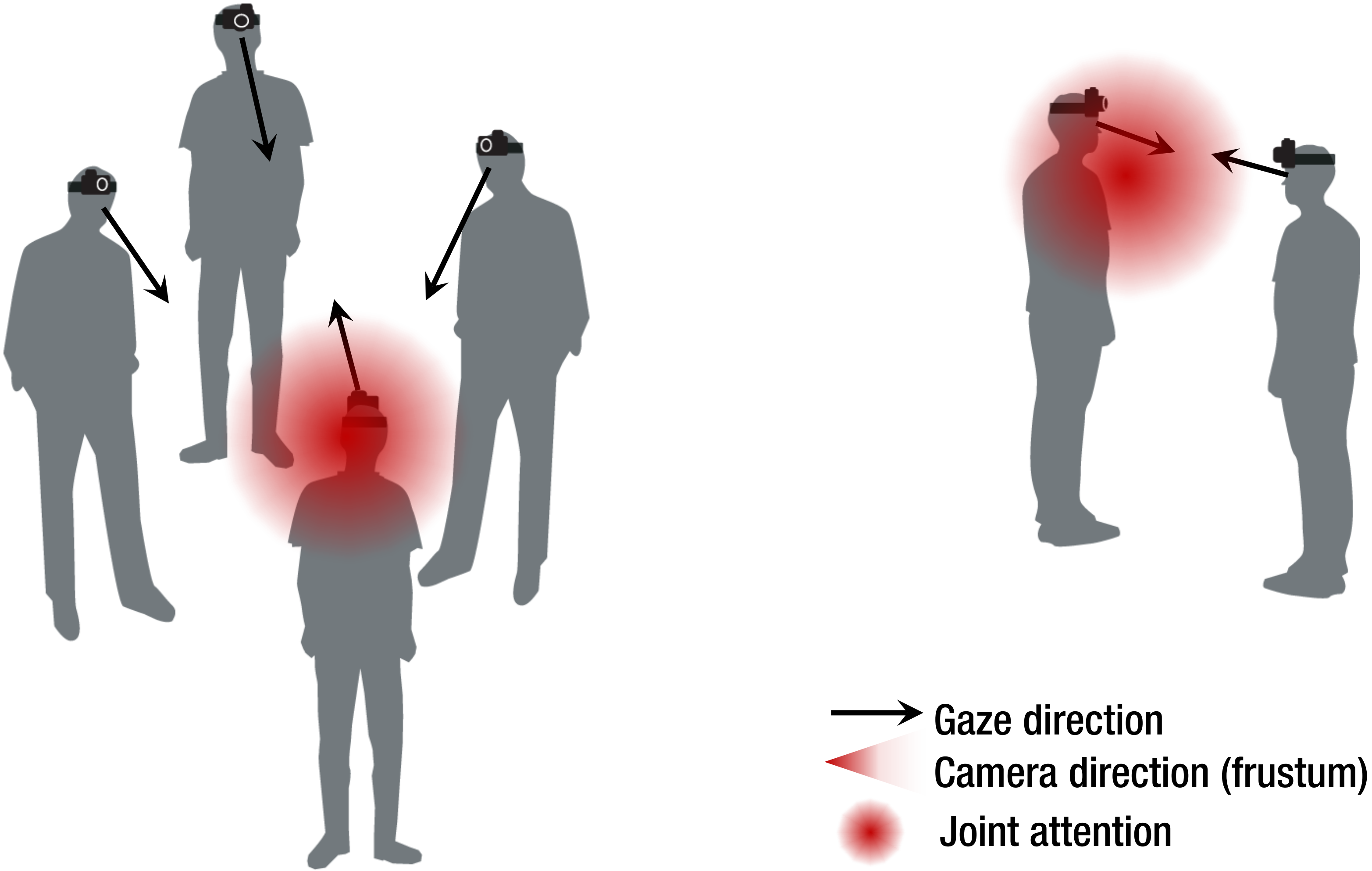


First person cameras are ideal sensors to measure social behaviors because they

1. secure the best views
2. produce more views of events of greater interest
3. follow social behaviors


Gaze direction
Camera direction (frustum)


Gaze direction
Camera direction (frustum)

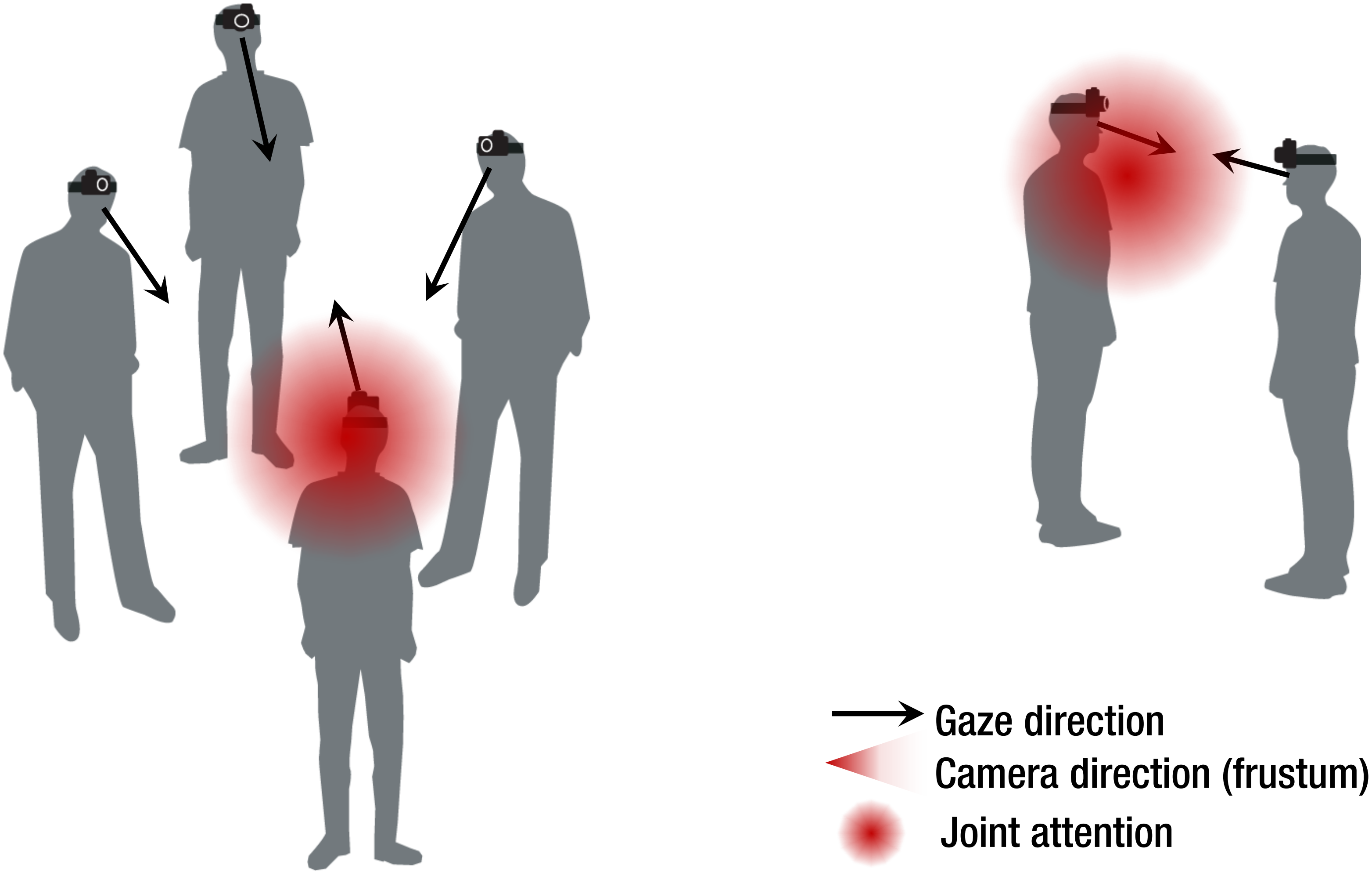


Gaze direction
Camera direction (frustum)
Joint attention

Joint Attention

Autism spectrum disorder (ASD)

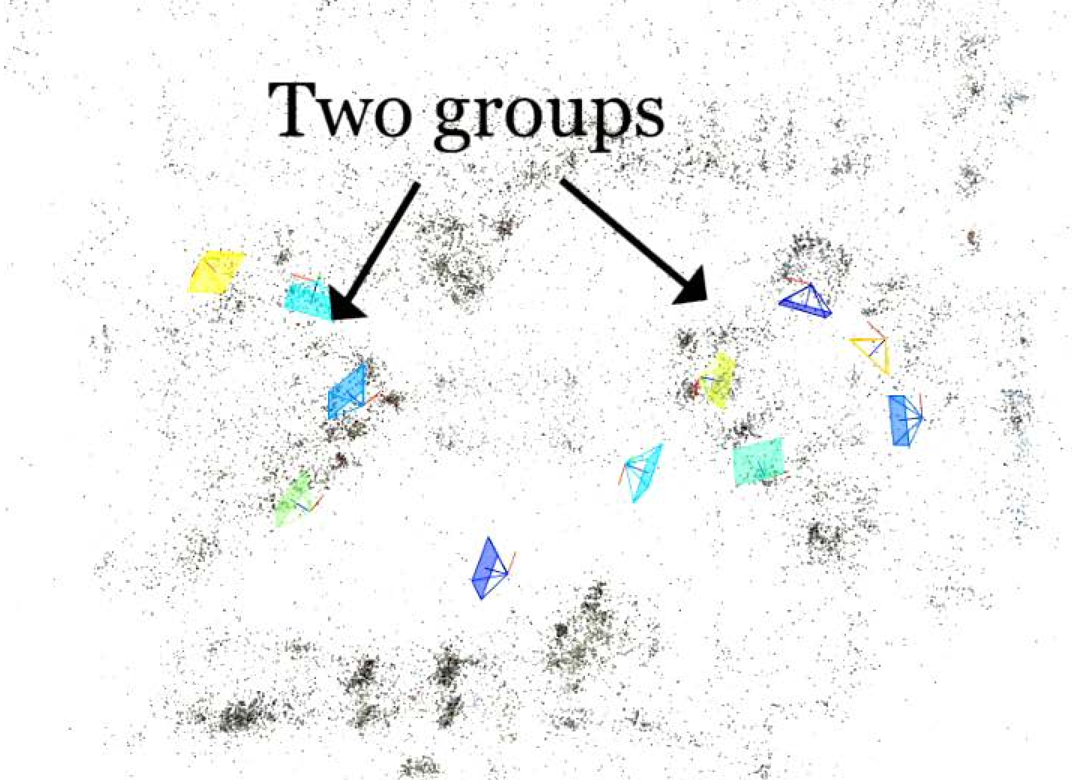
Attention deficit hyperactivity disorder (ADHD)


Gaze direction
Camera direction (frustum)
Joint attention

Input Video: Meeting Scene

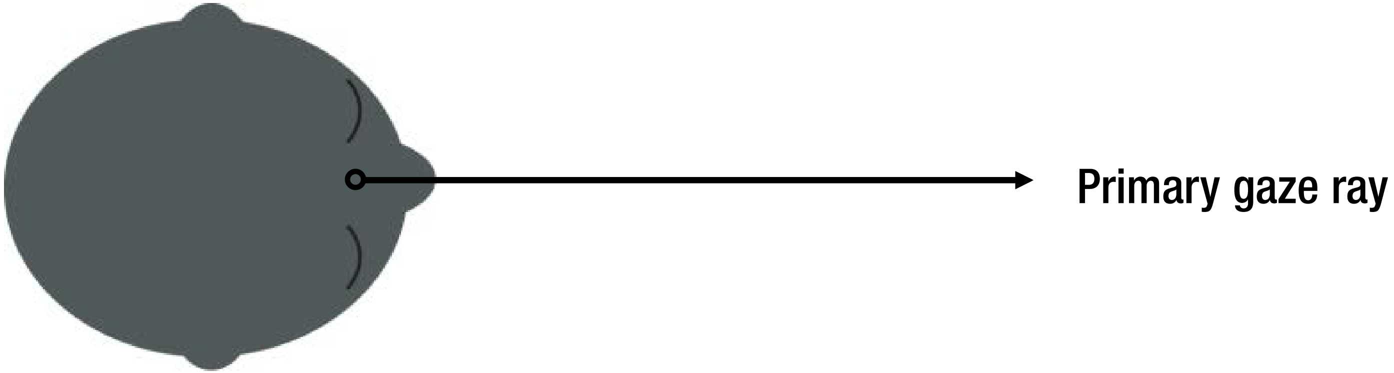
1x speed

3D Camera Pose Estimation
(Structure from motion)

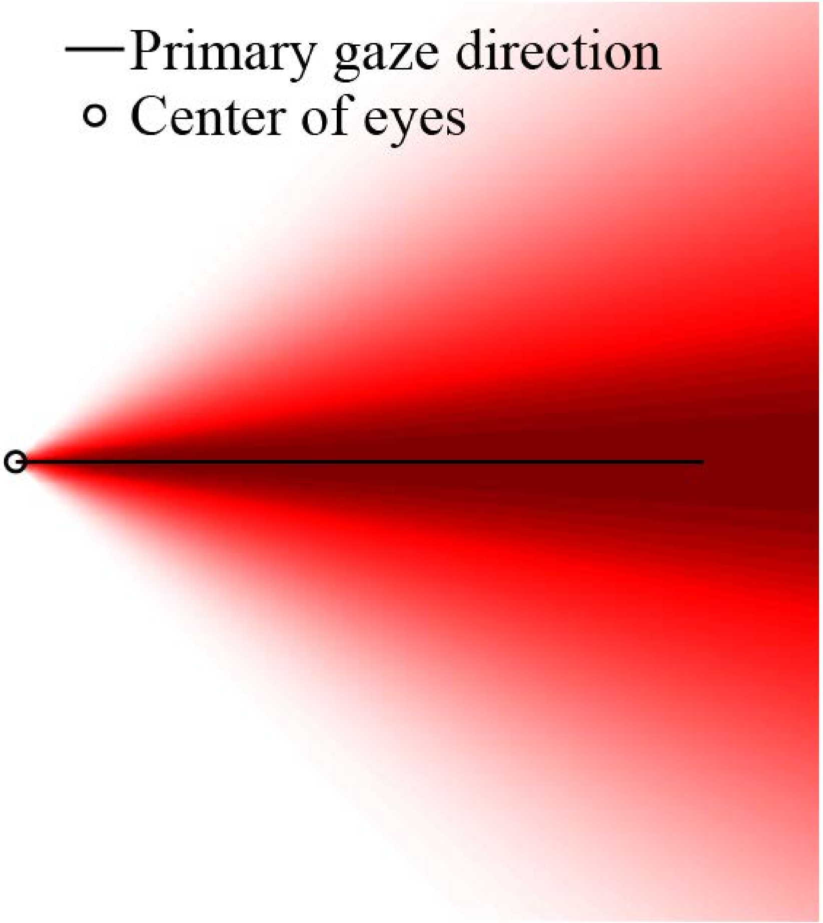


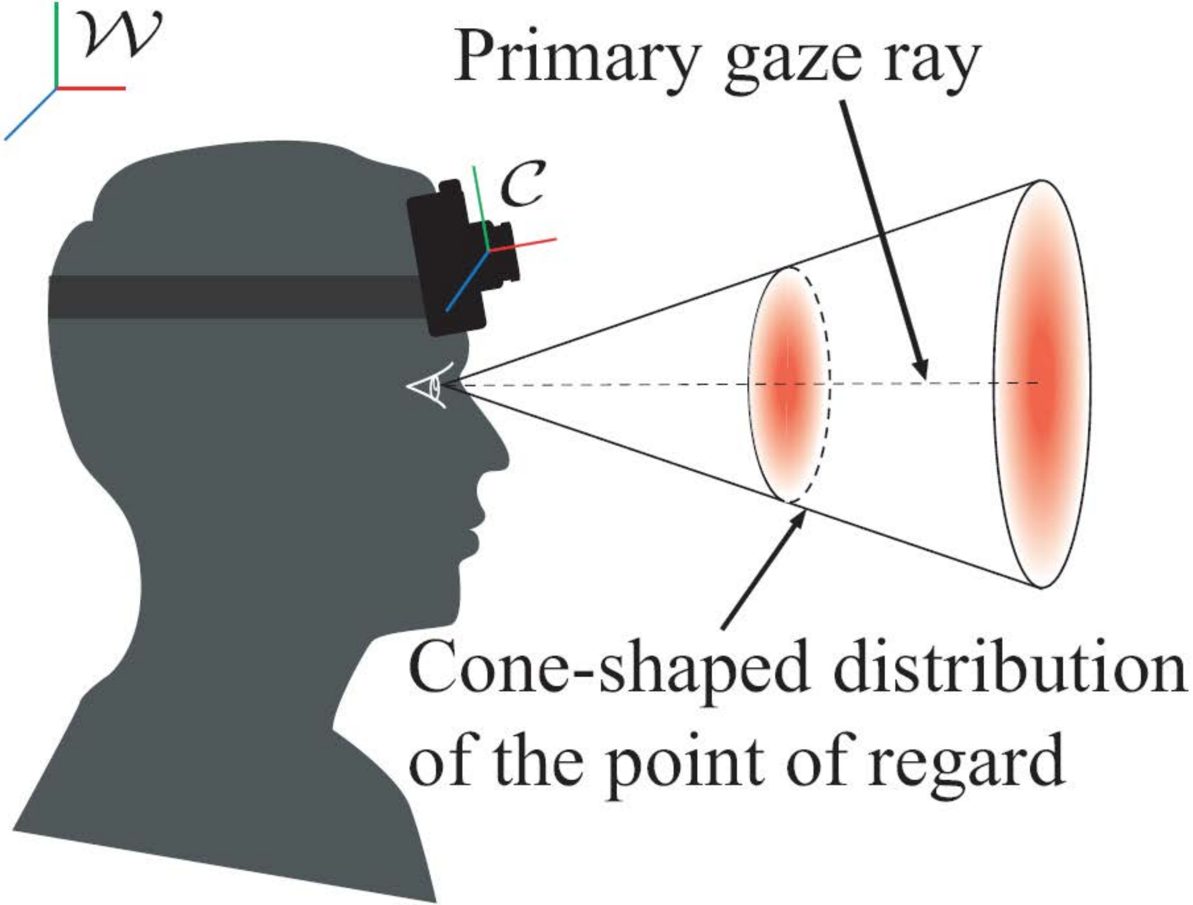
Two groups

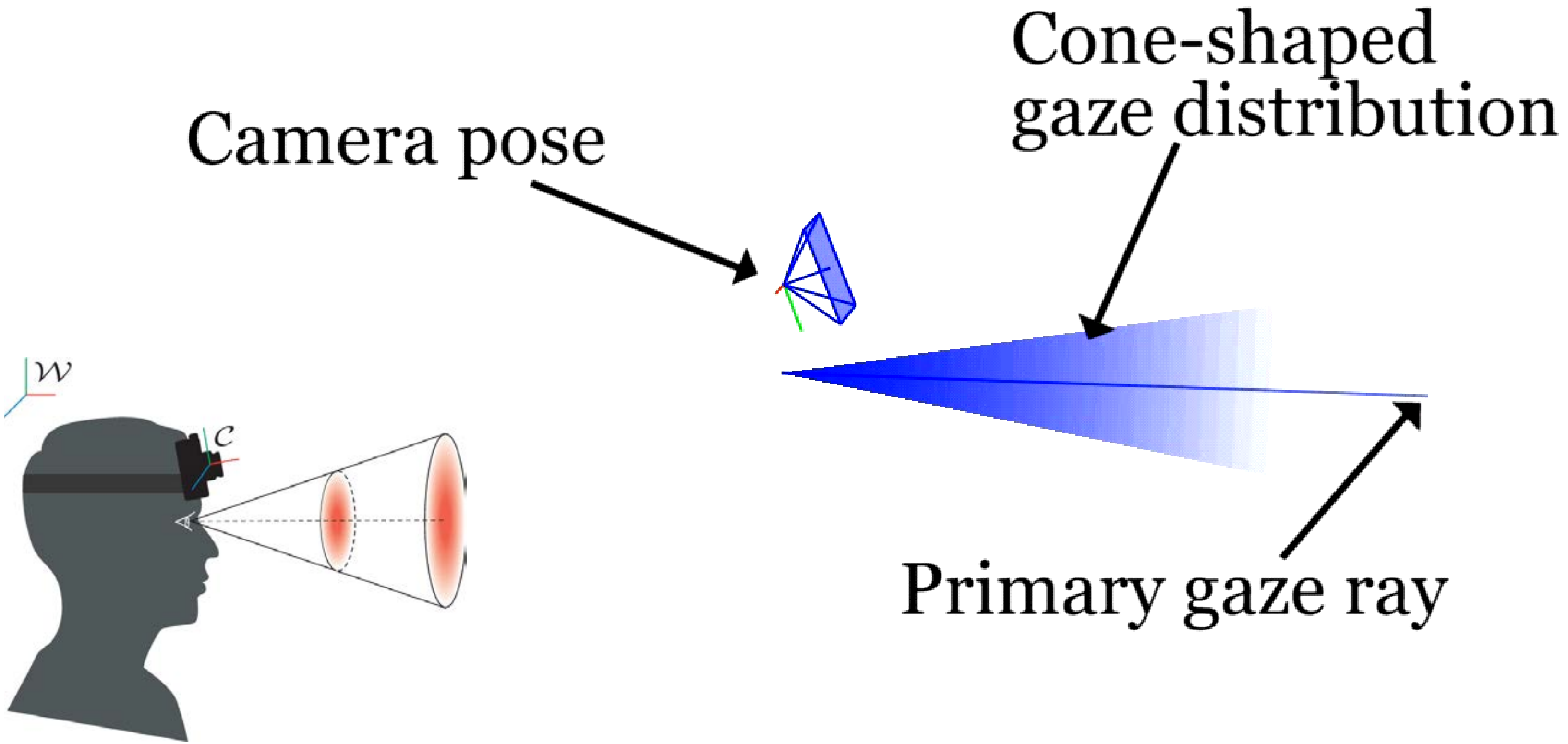
6x speed



Eye-in-head motion


— Primary gaze direction
○ Center of eyes


$\mathcal{W}$
Primary gaze ray
$\mathcal{C}$
Cone-shaped distribution
of the point of regard

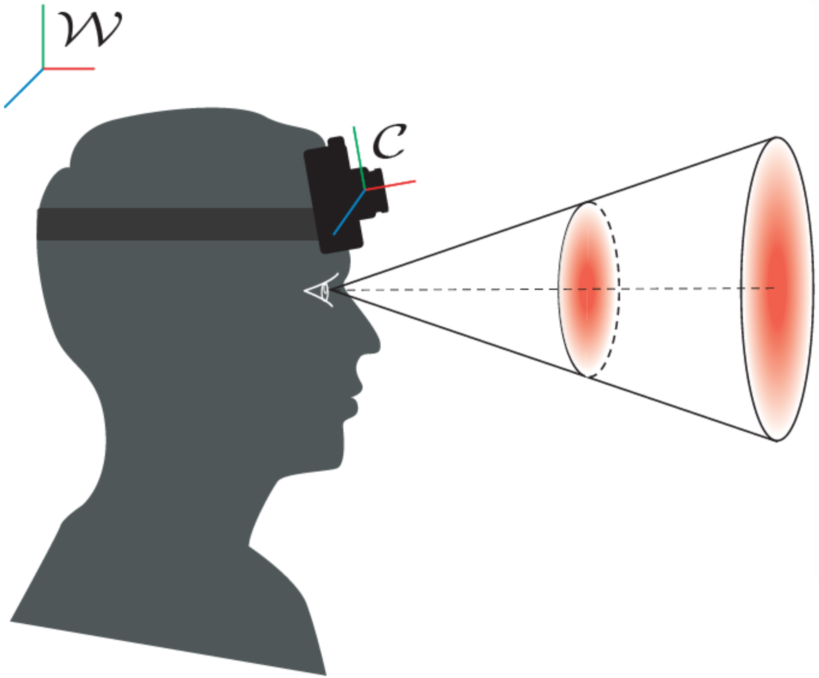


Camera pose

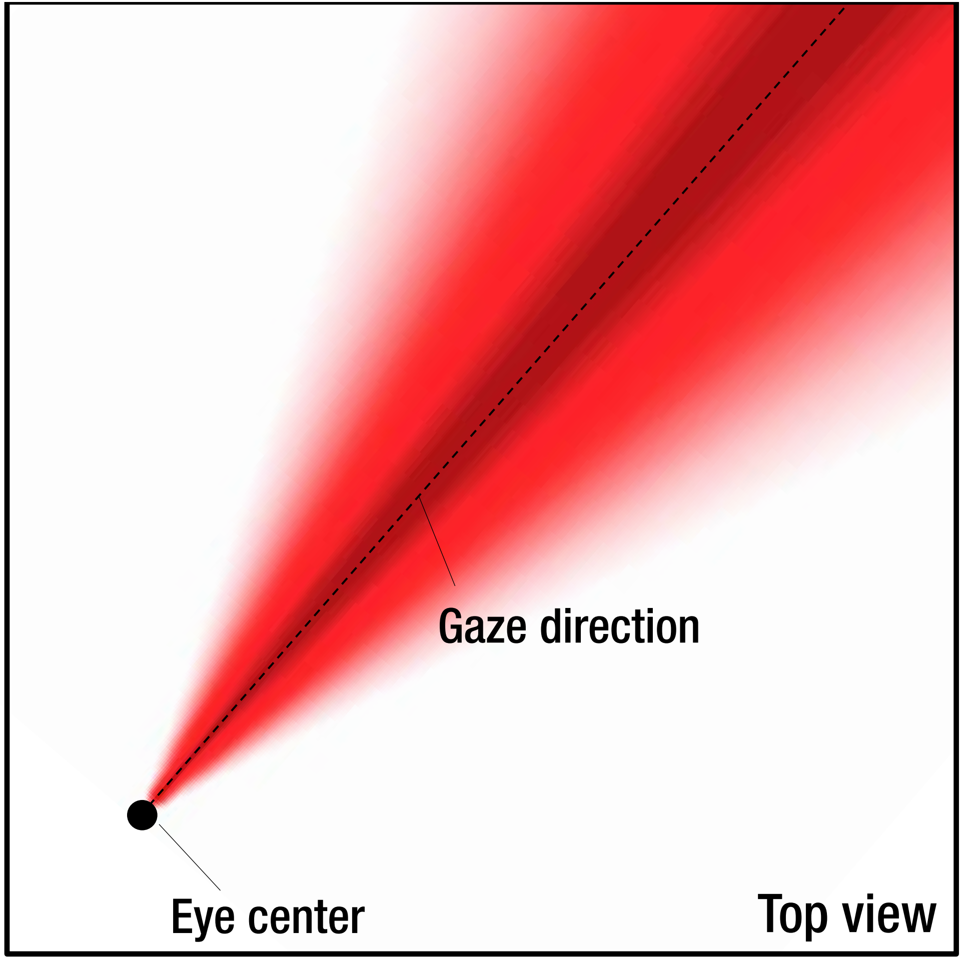
Cone-shaped gaze distribution

Primary gaze ray

3D gaze model registration



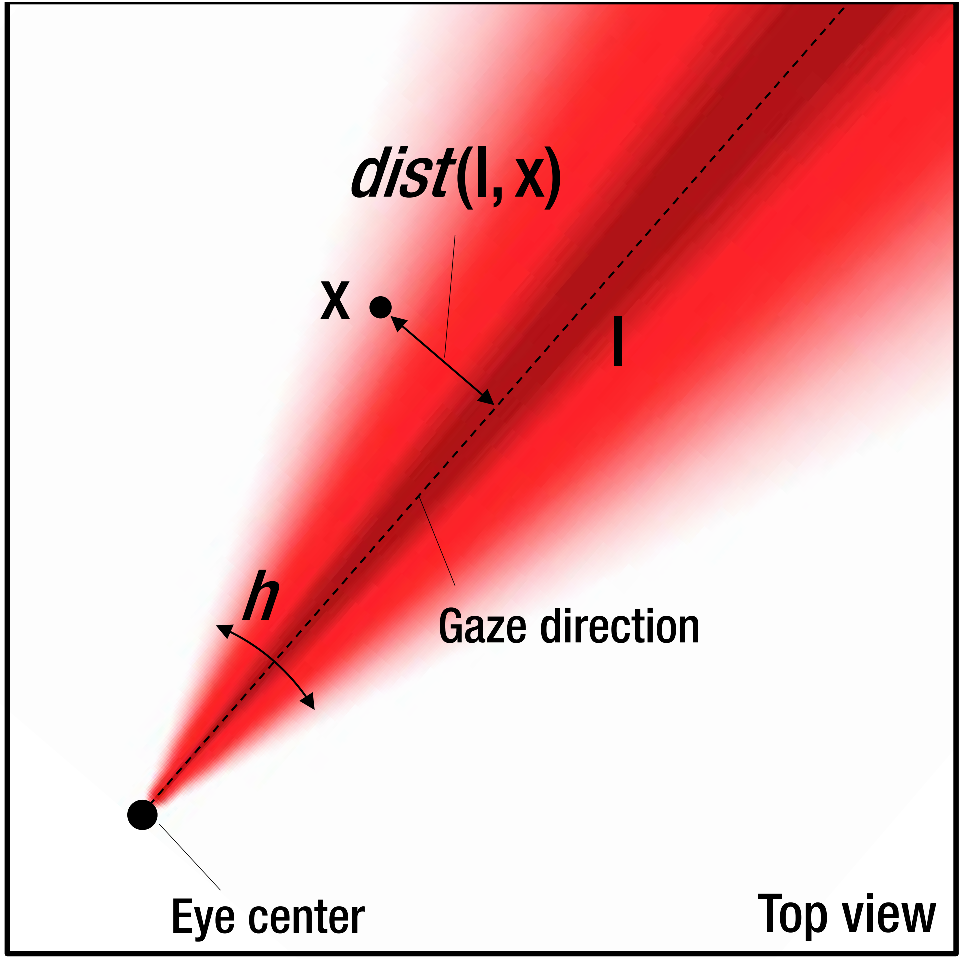
Cone-shaped gaze model



Attention density field:

$$f(\mathrm{x}) = k\left(\frac{\mathrm{dist}(\mathrm{l},\mathrm{x})}{h}\right)$$

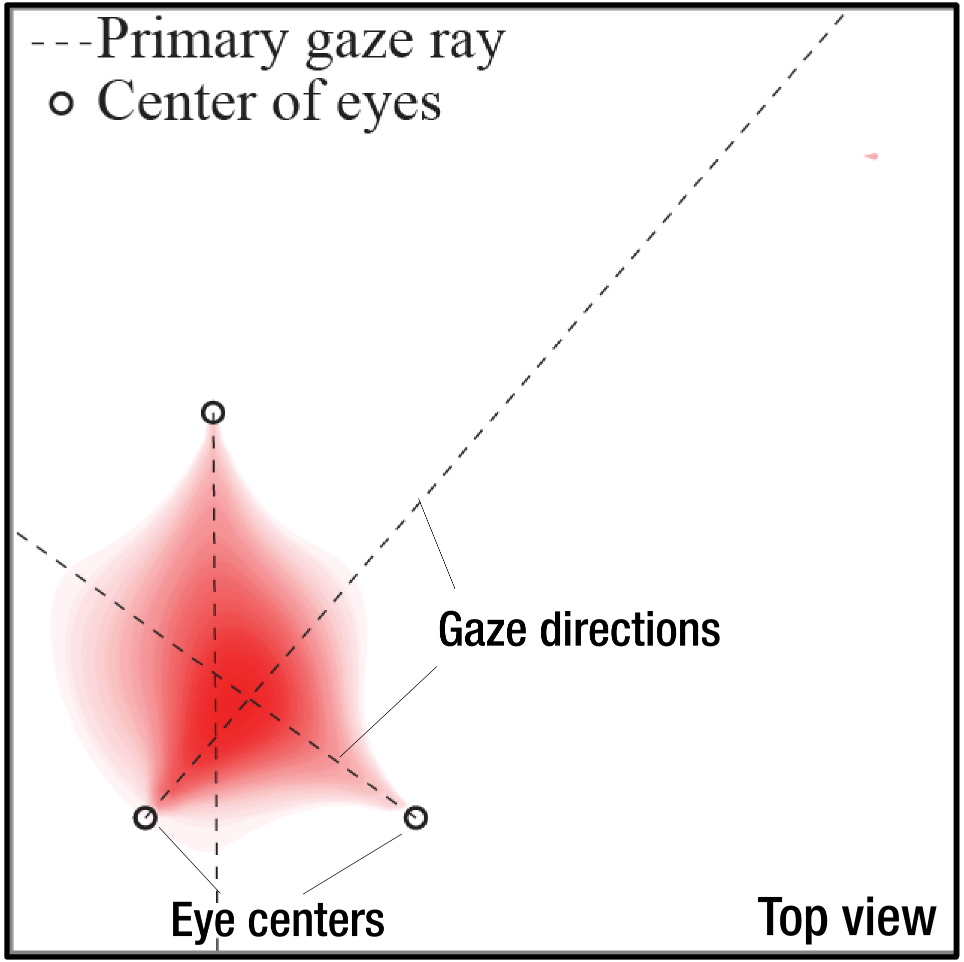
where k is a kernel density function (●).



Attention density field:

$$f(\mathrm{x}) = \sum_i k\left(\frac{\mathrm{dist}(\mathrm{l}_i, \mathrm{x})}{h}\right)$$

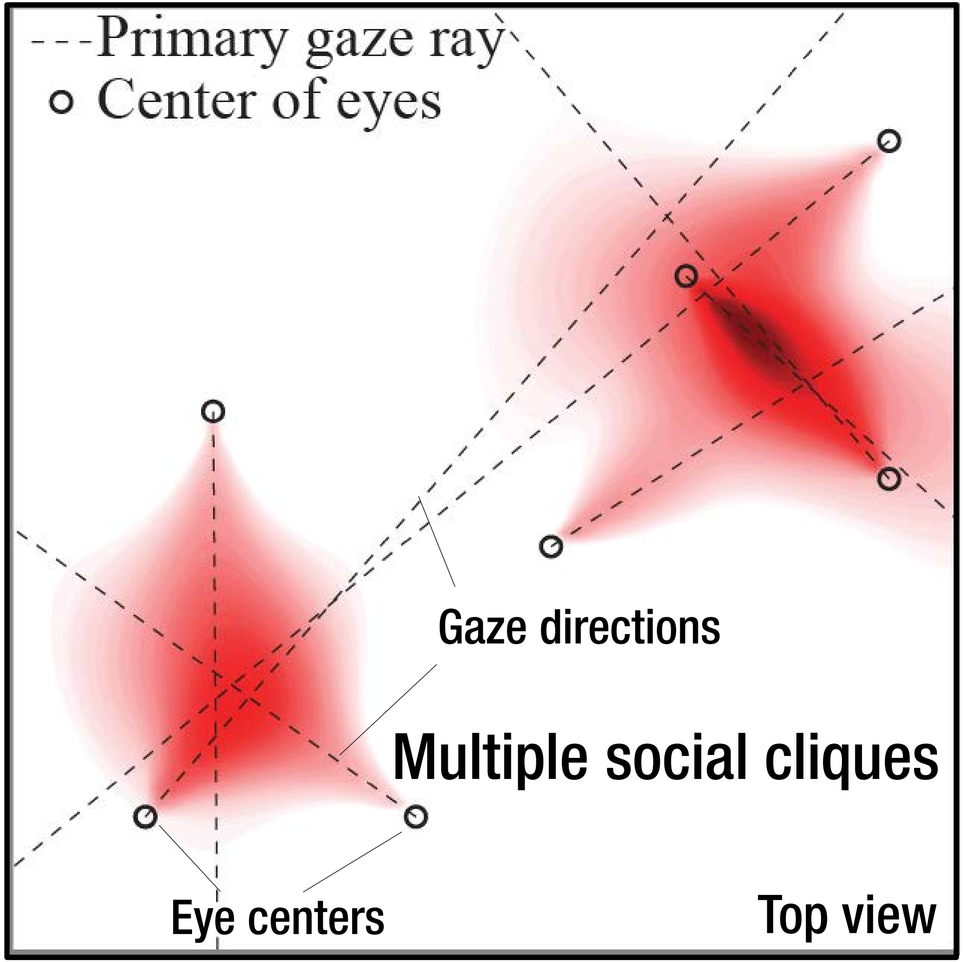
where k is a kernel density function (●).



Attention density field:

$$f(\mathrm{x}) = \sum_i k\left(\frac{\mathrm{dist}(\mathrm{l}_i, \mathrm{x})}{h}\right)$$

where k is a kernel density function (●).



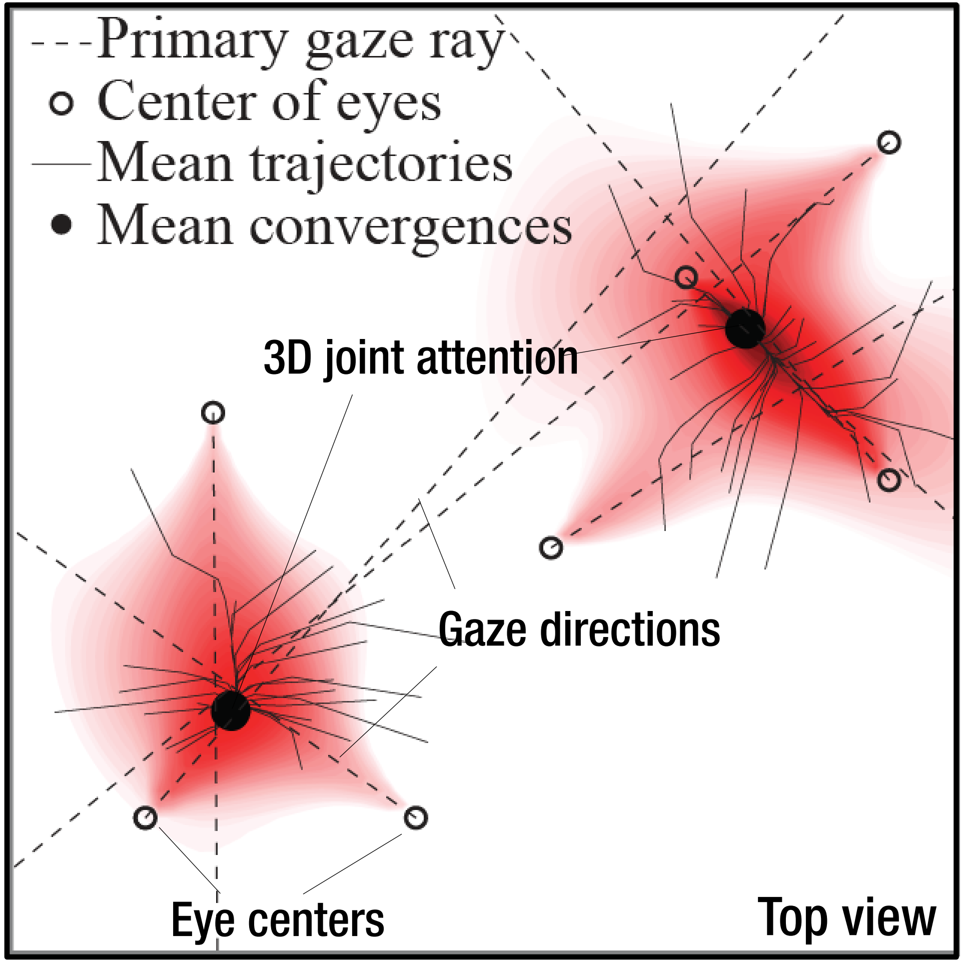
Attention density field:

$$f(\mathrm{x}) = \sum_i k\left(\frac{\mathrm{dist}(\mathrm{l}_i\,, \mathrm{x})}{h}\right)$$

where k is a kernel density function (●).

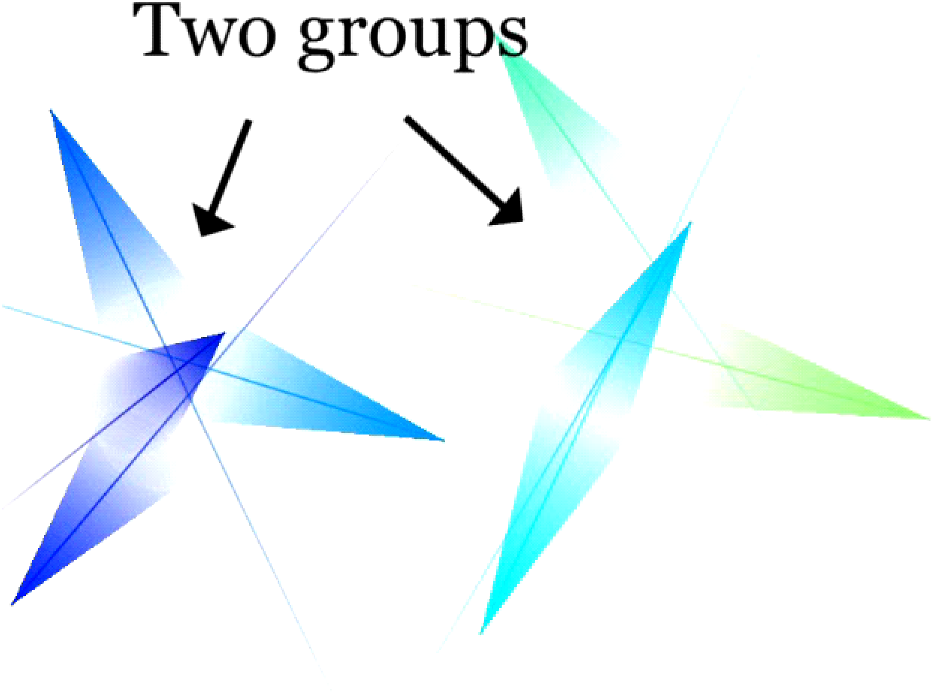
3D joint attention

≈ Modes of attention density field

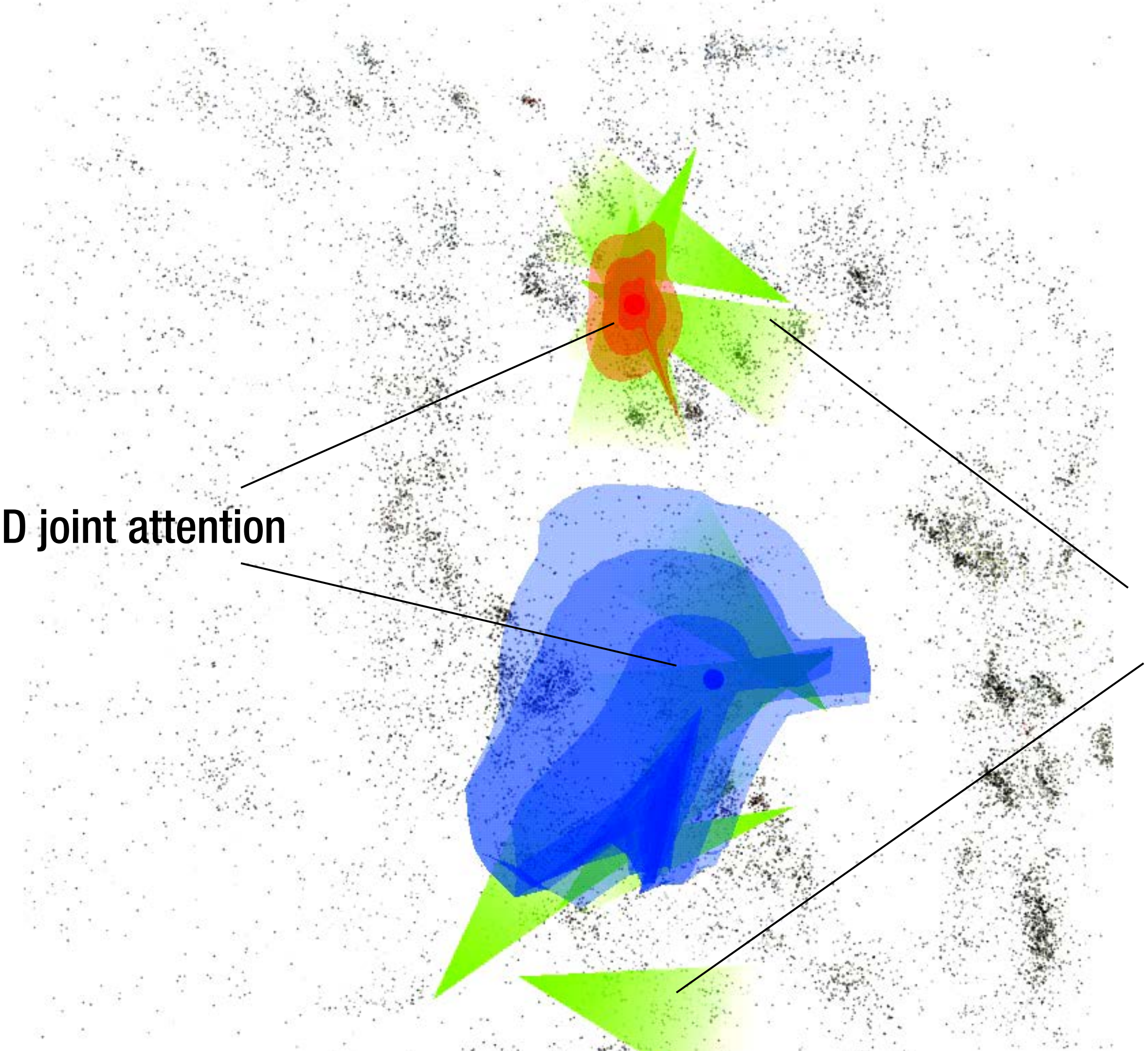


Mode-seeking: Gaze Concurrences

Two groups

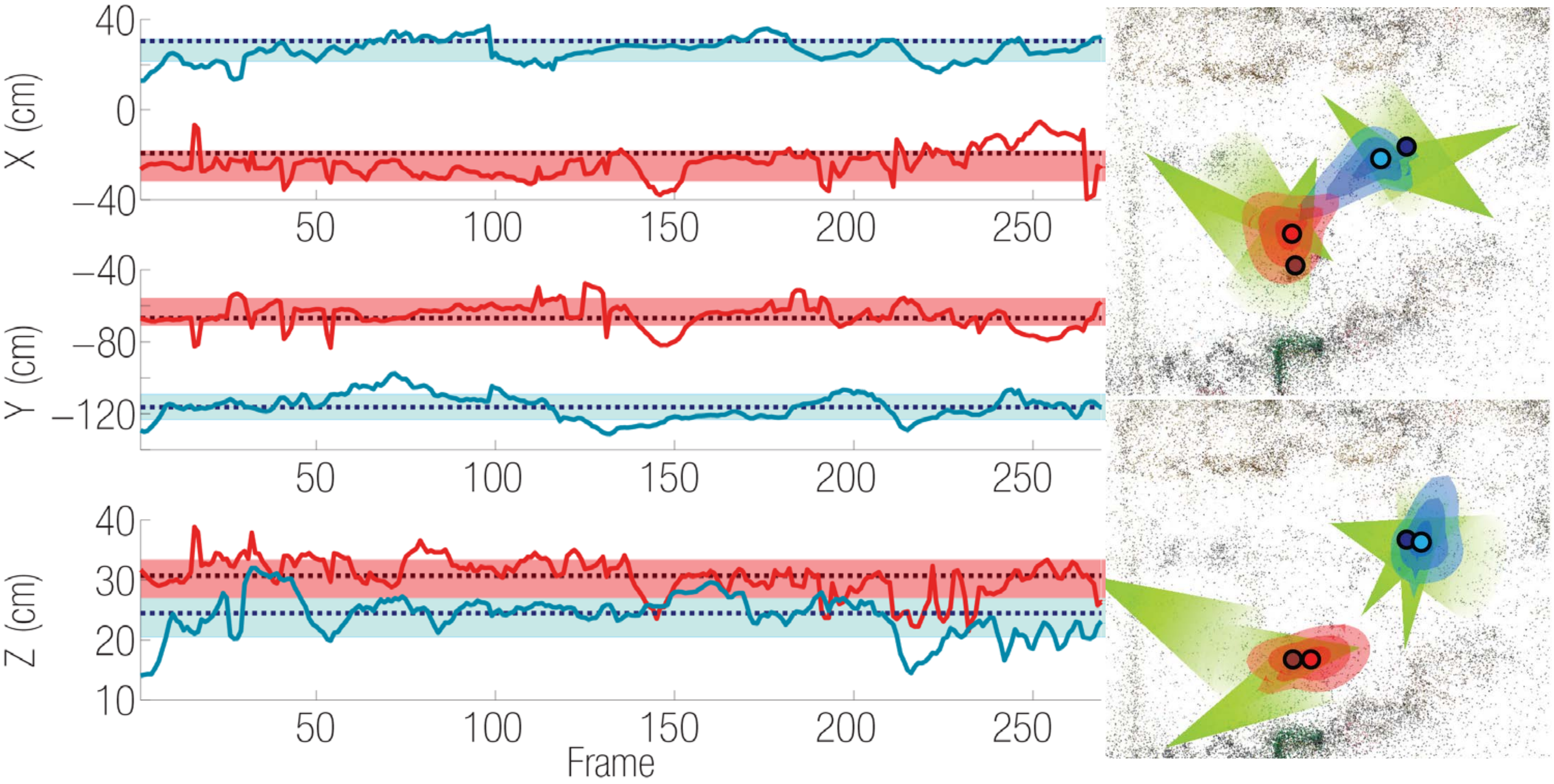


Multiple groups



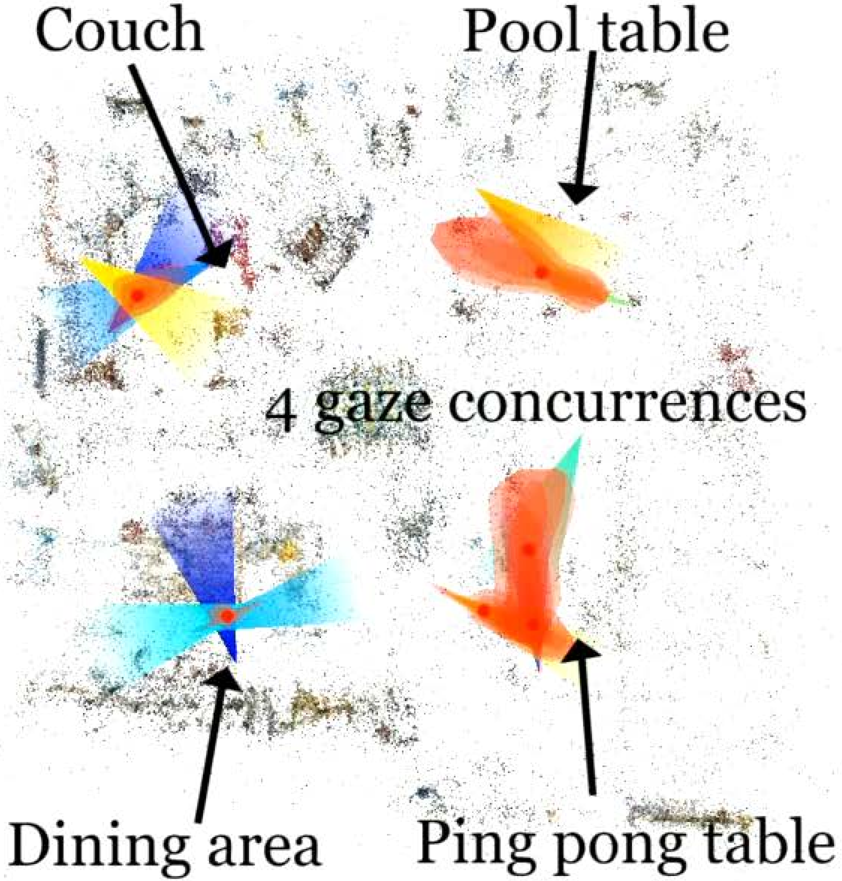
3D joint attention

Gaze directions

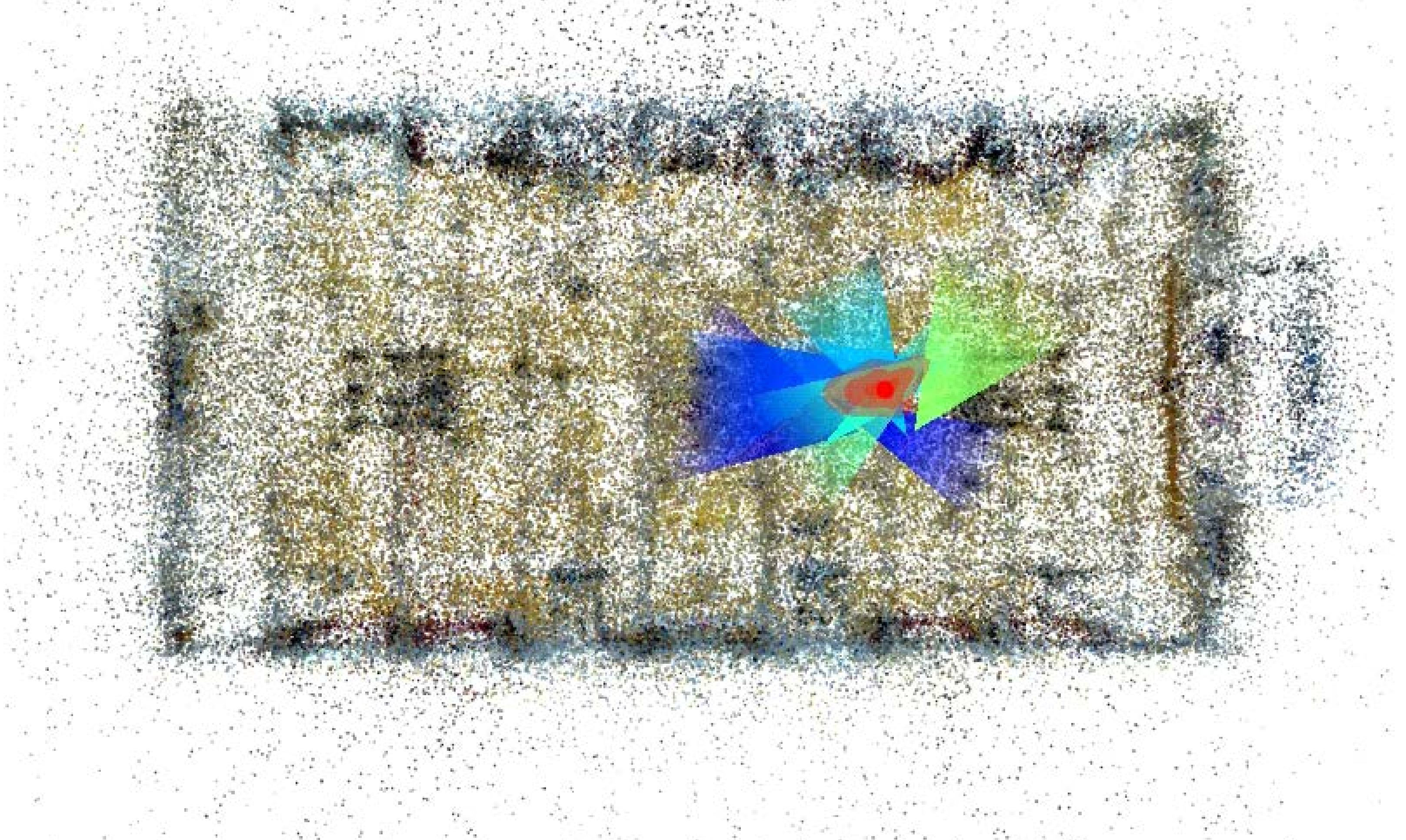

40
0
−40
X (cm)
50
100
150
200
250
−40
−80
−120
Y (cm)
50
100
150
200
250
40
30
20
10
Z (cm)
50
100
150
200
250
Frame

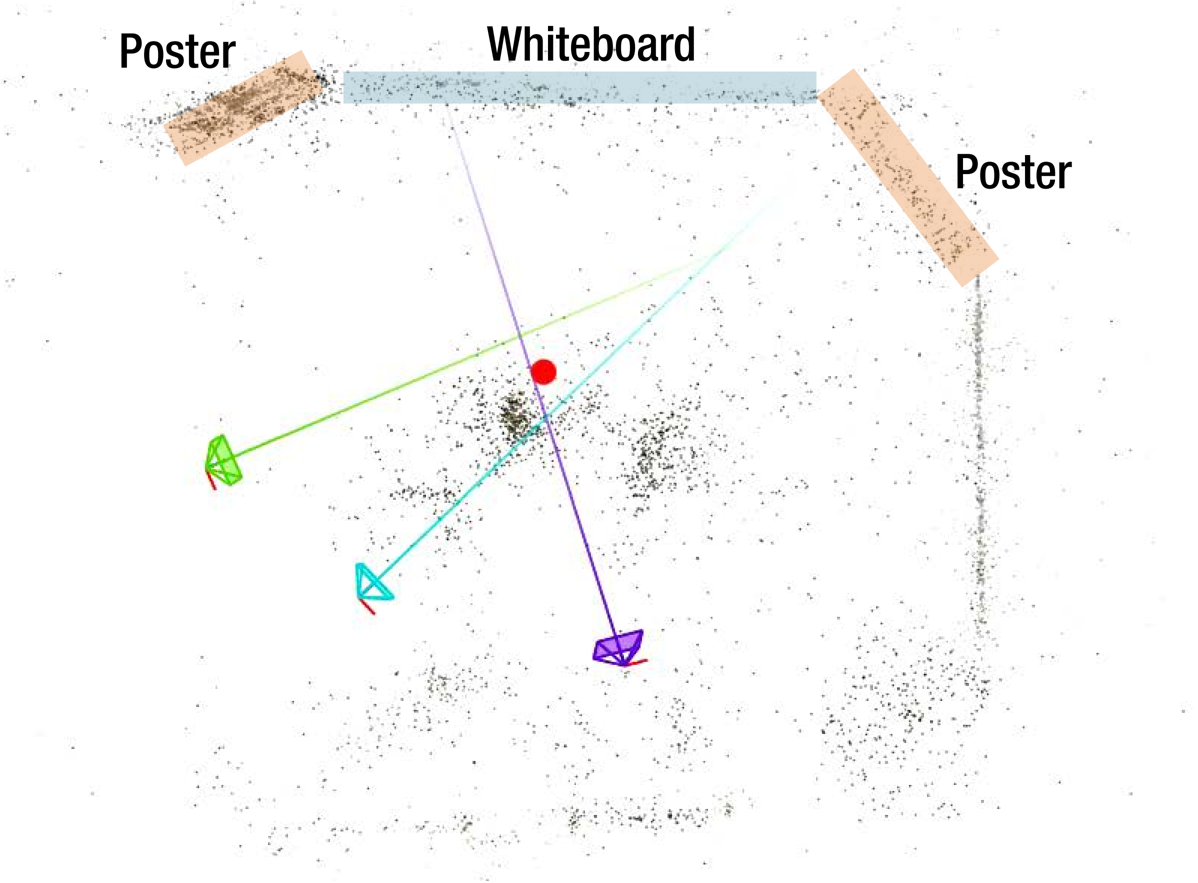
Party Scene: 4 groups in a room

1x speed

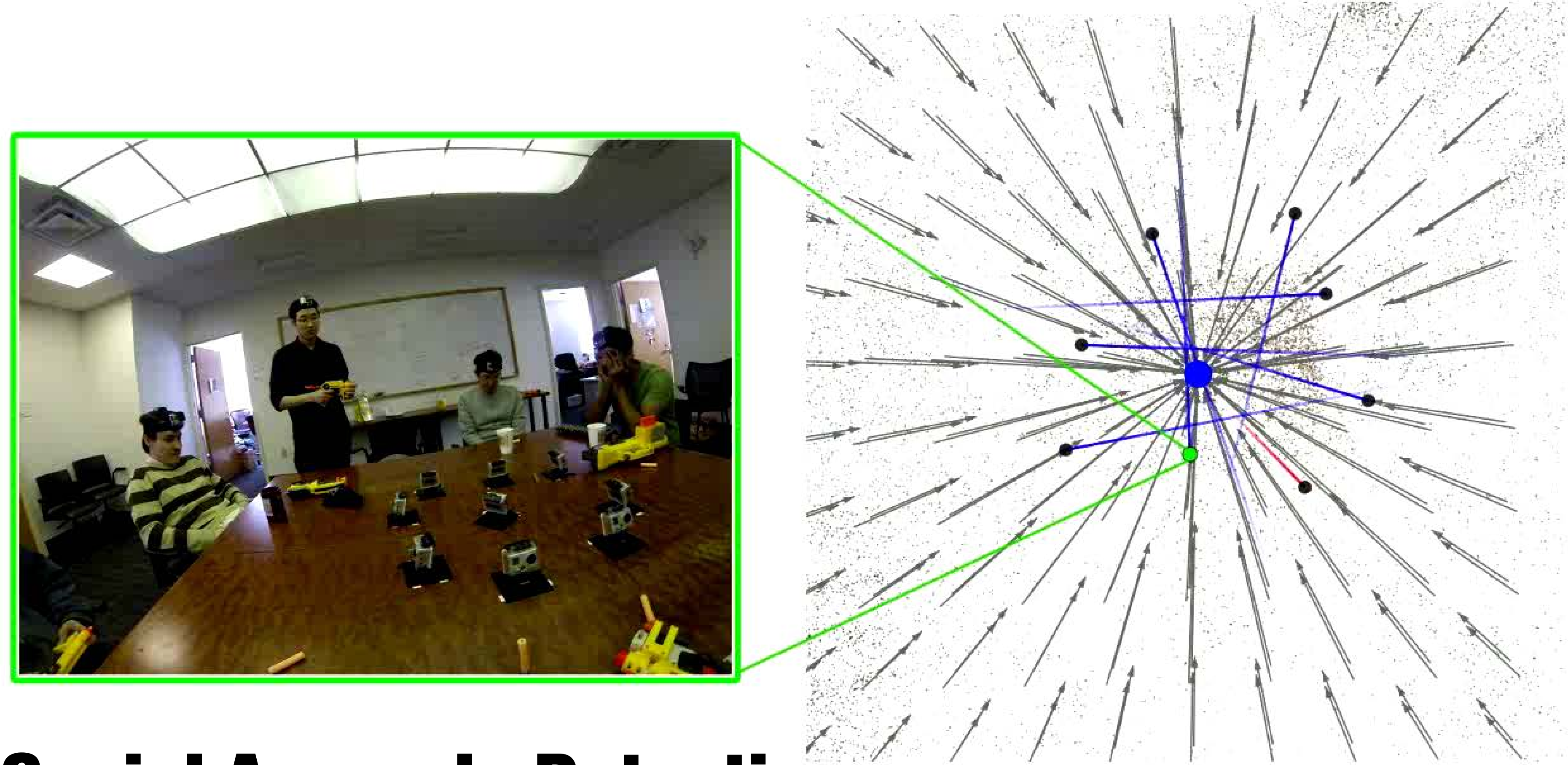


GO TARTANS



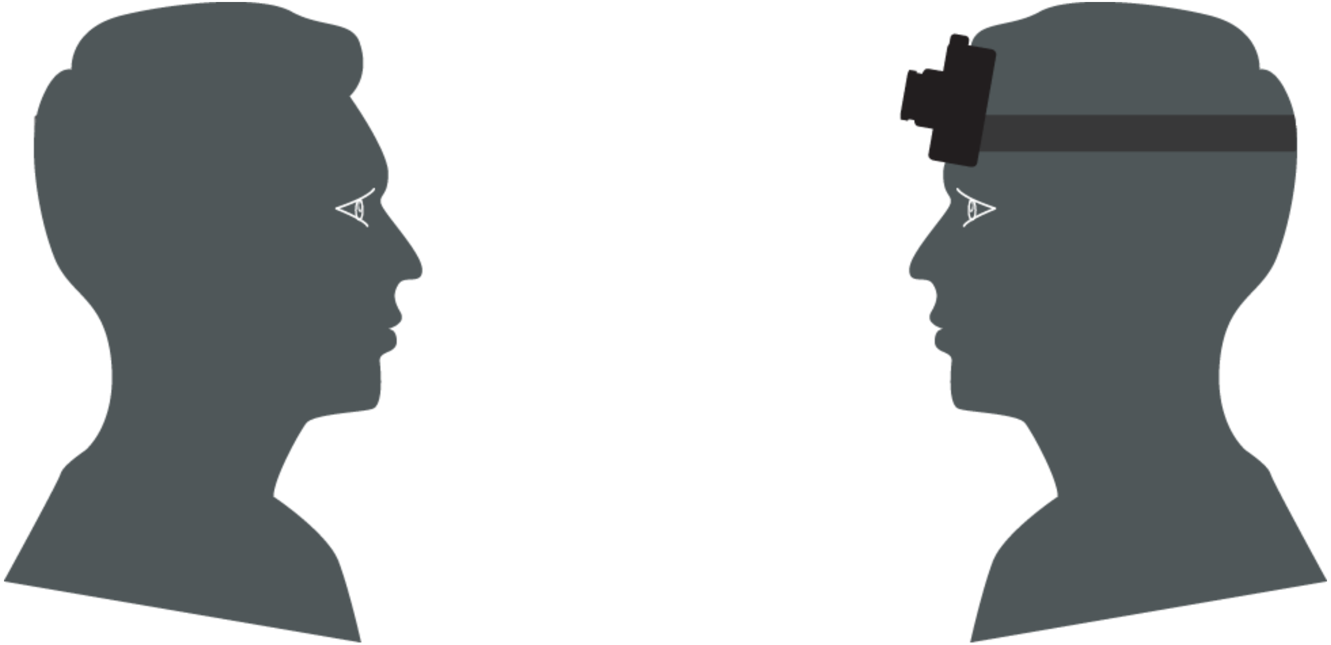


Realtime Joint Attention

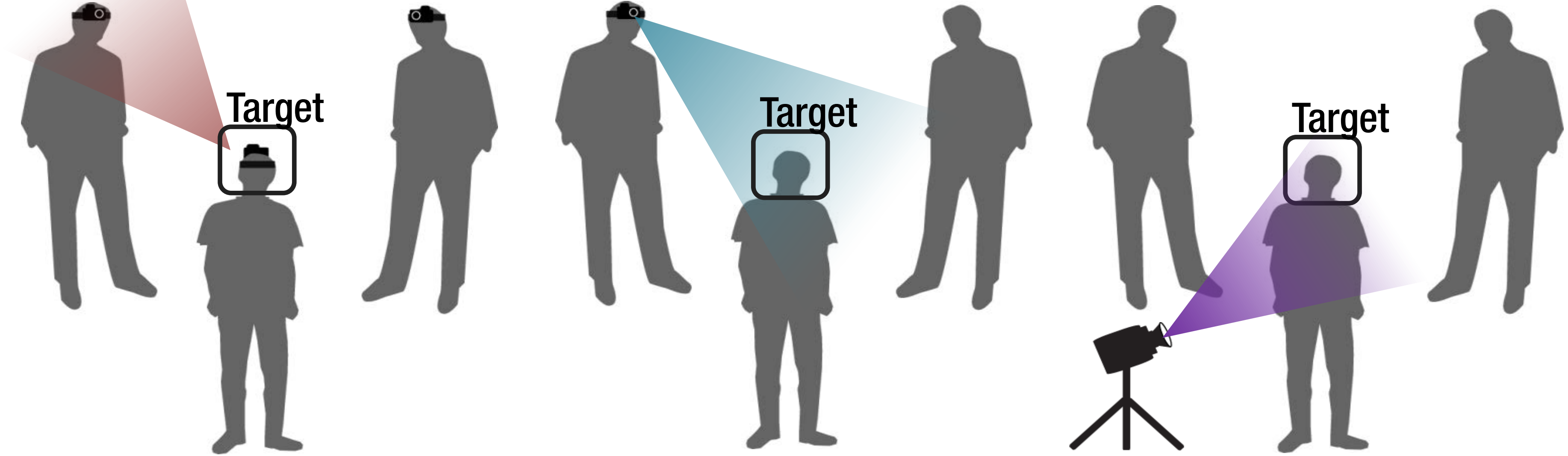


Social Anomaly Detection

→ Predicted gaze direction



Can we predict without first person cameras?



First person

Second person

Third person

3D estimation error < 5cm

Measurement accuracy

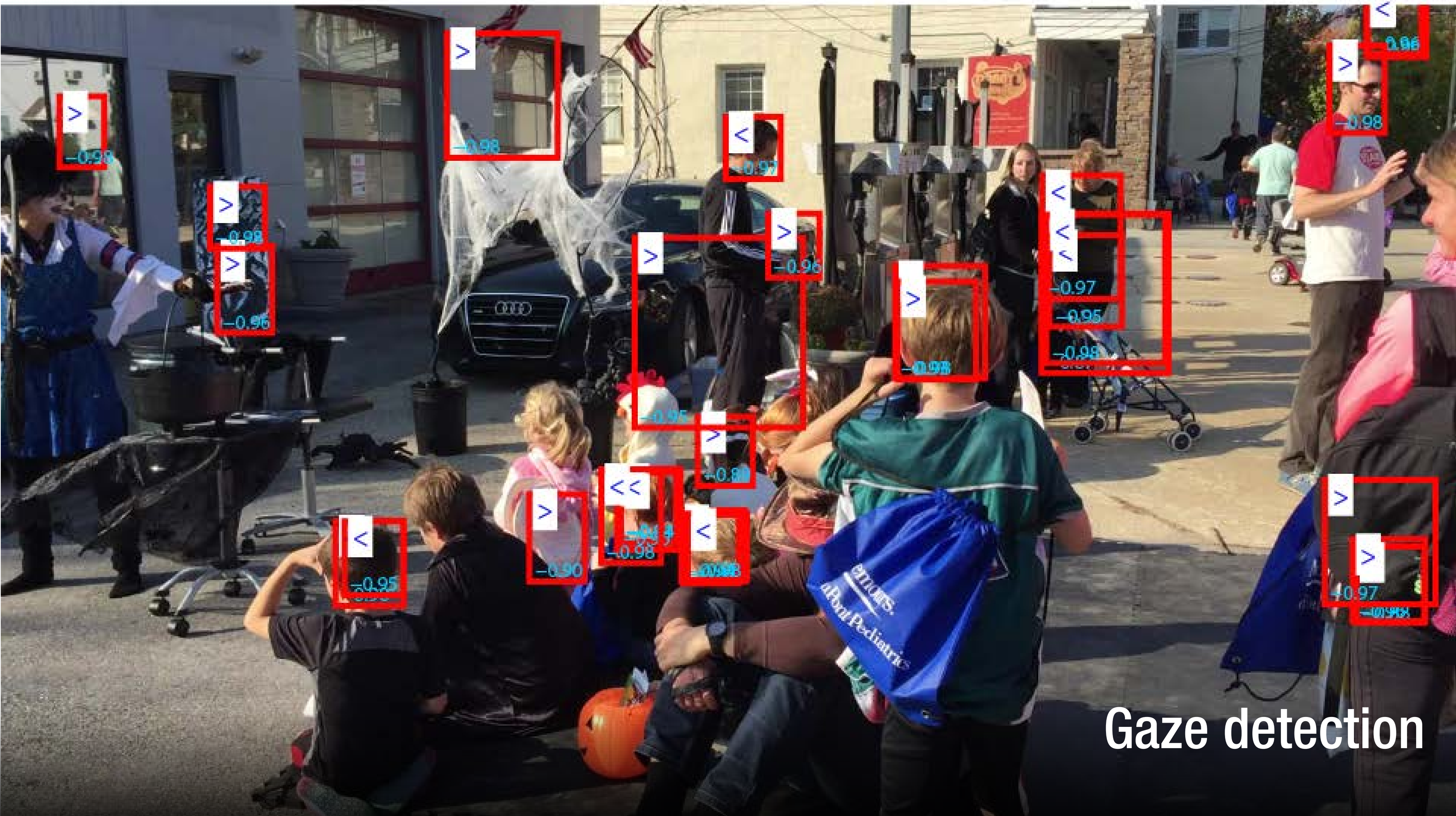
Learning

Prediction

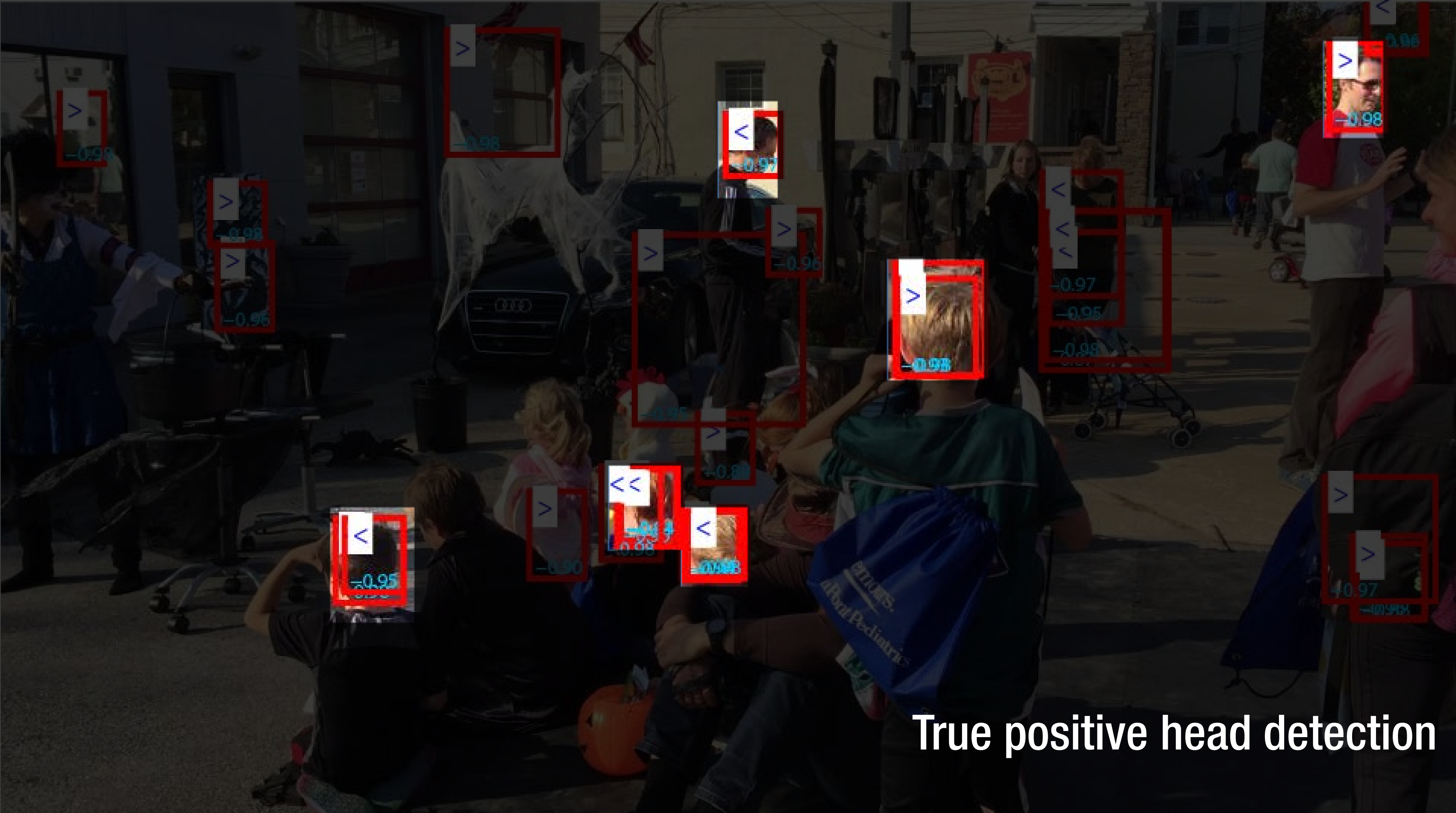
Noninvasiveness

Nemours.
duPont Pediatrics

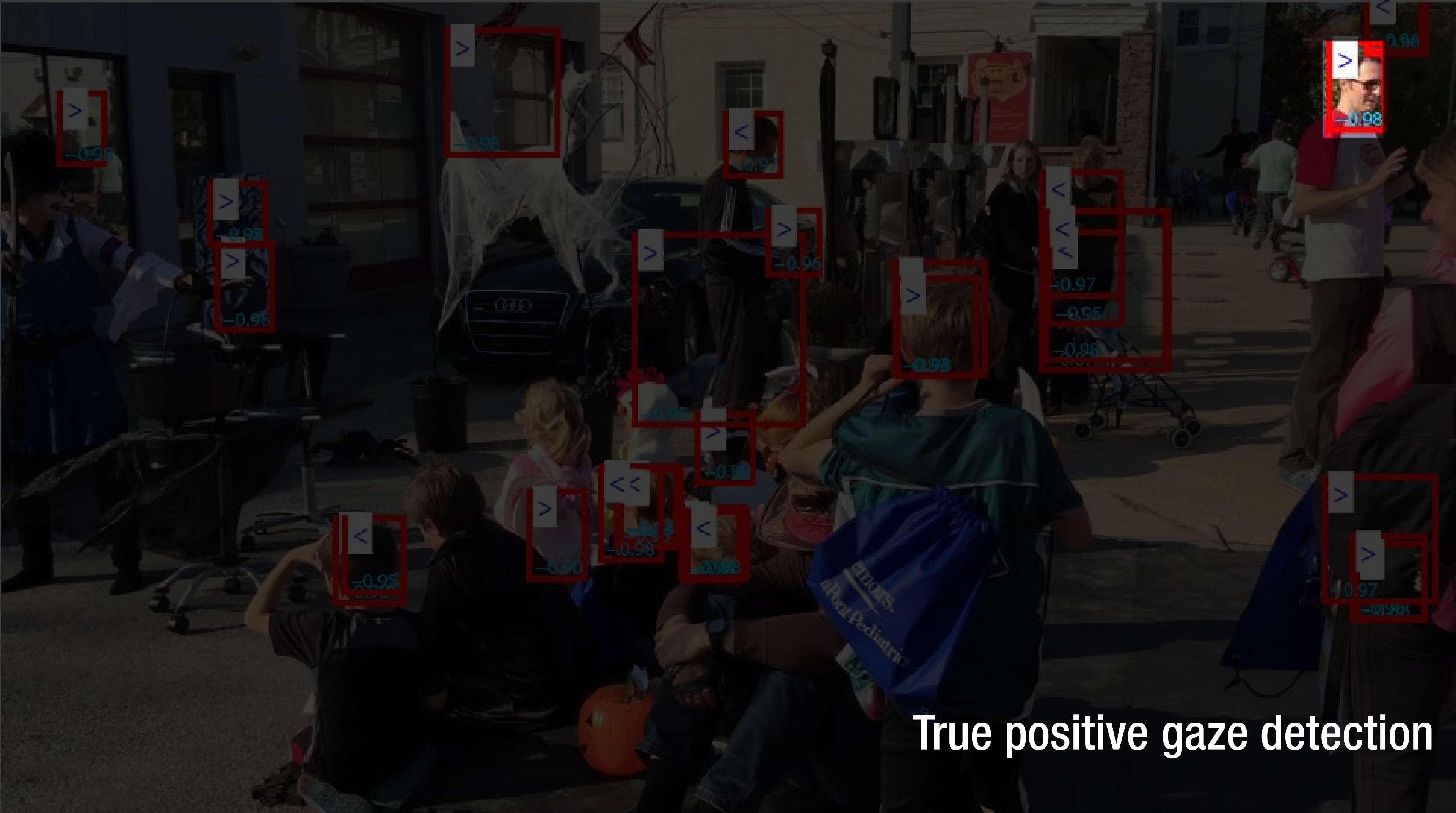
emours.
duPont Pediatrics



Gaze detection

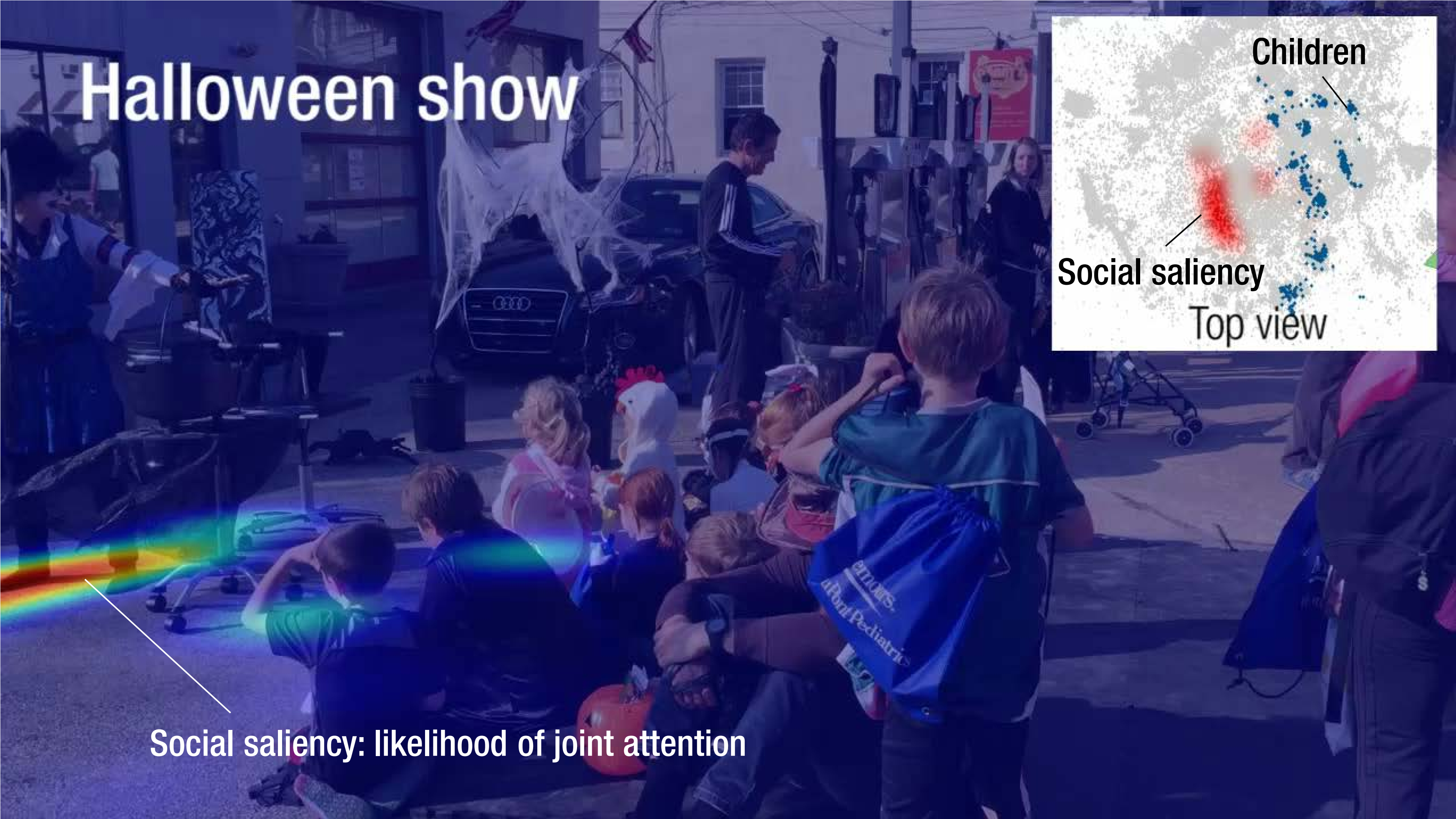


True positive head detection



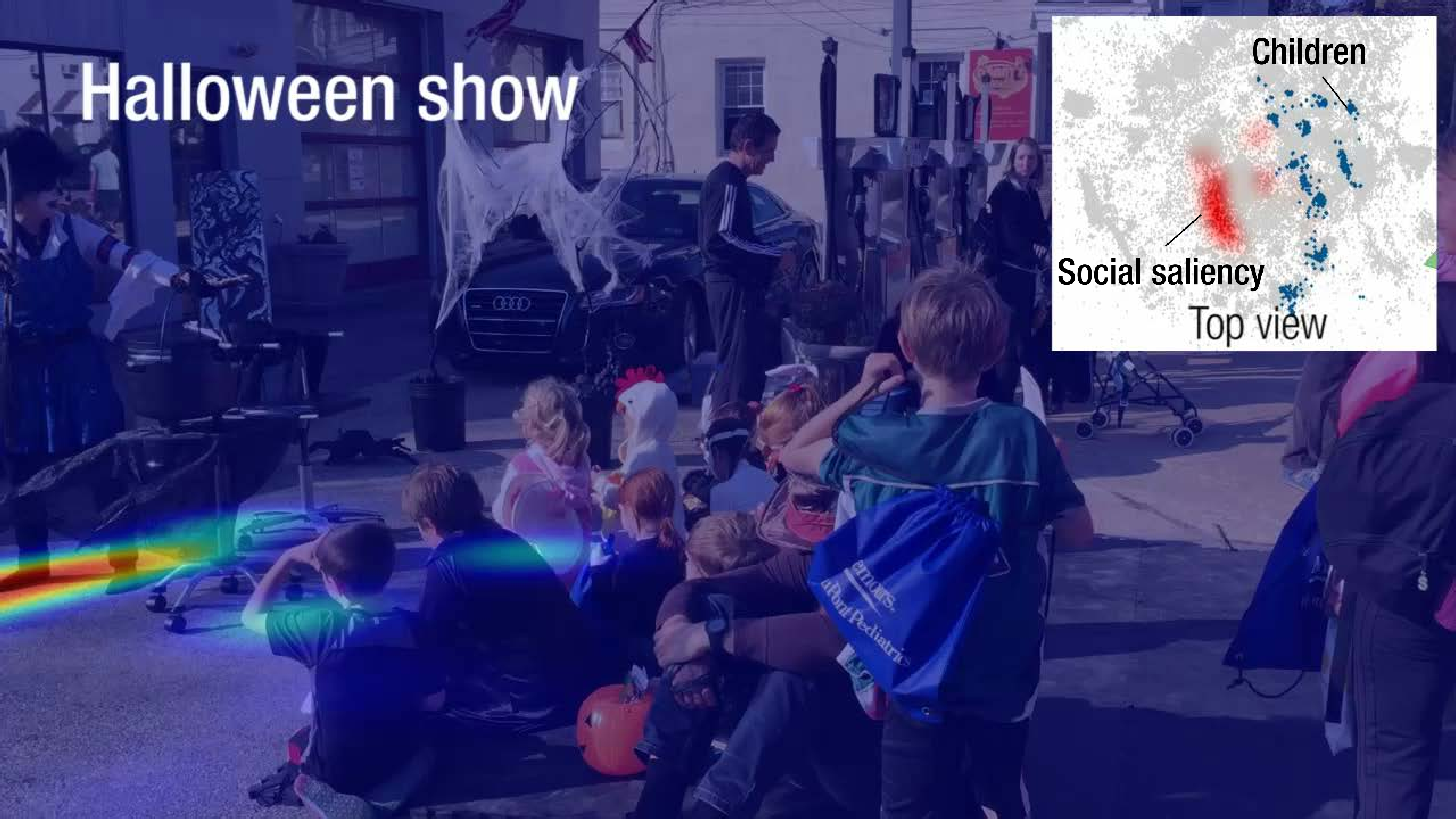
True positive gaze detection

Halloween show



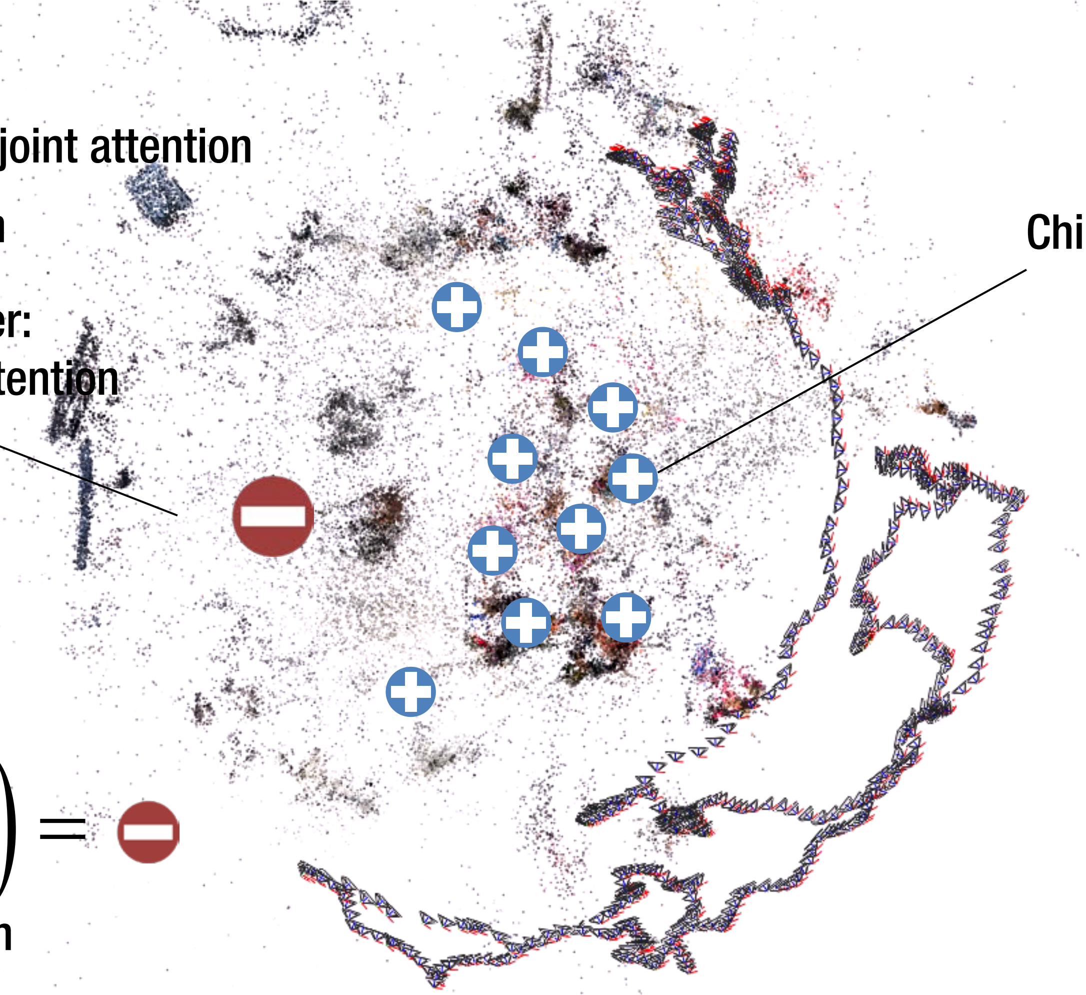
Social saliency: likelihood of joint attention

Halloween show



⊖ : Ground truth joint attention

⊕ : Head location



Location of performer:
Ground truth joint attention

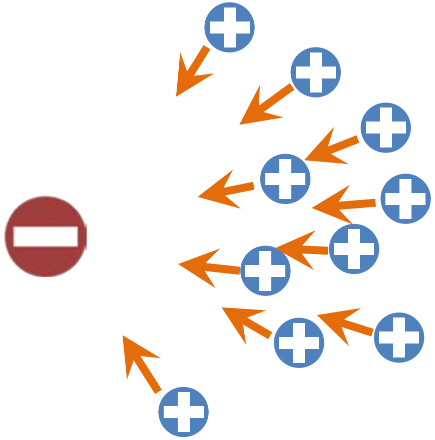
Children

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

where N is the number of social members.

⊖ : Ground truth joint attention

⊕ : Head location



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

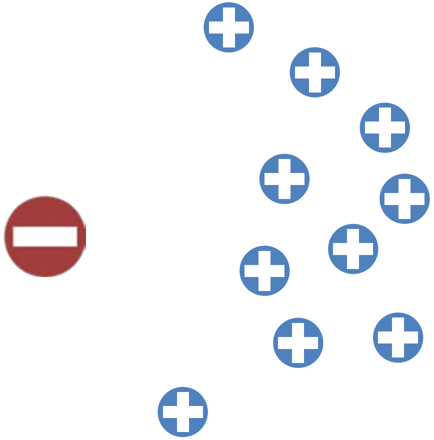
where N is the number of social members.

cf. $$g\left(\left\{\oplus\longrightarrow_i\right\}_{i=1}^{N}\right) = \ominus$$

Geometric localization: deterministic

⊖ : Ground truth joint attention

⊕ : Head location

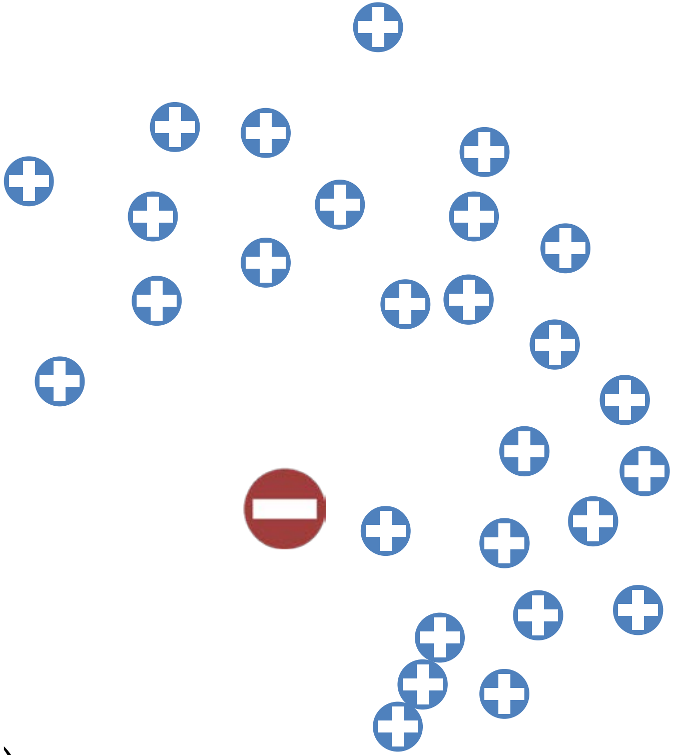


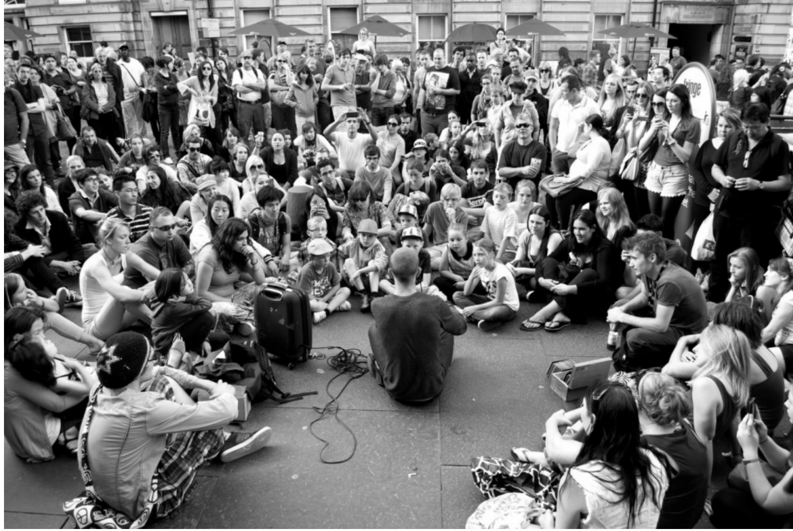
$$g\left(\underbrace{\left\{\oplus_i\right\}_{i=1}^{N}}_{\text{Social formation}}\right) = \ominus$$

where N is the number of social members.

cf. $$g\left(\left\{\oplus\!\rightarrow_i\right\}_{i=1}^{N}\right) = \ominus$$

Geometric localization: deterministic





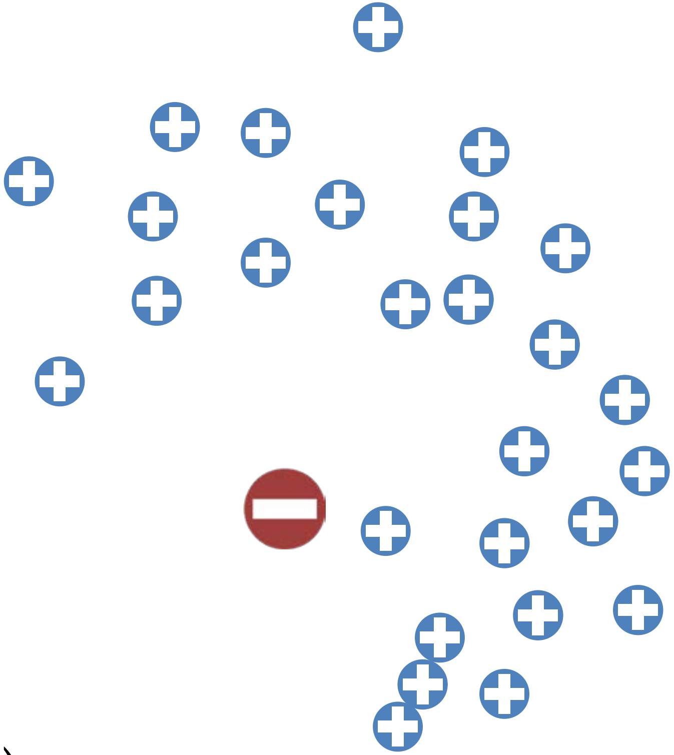
Scale variation

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

where N is the number of social members.

⊖ : Ground truth joint attention

⊕ : Head location



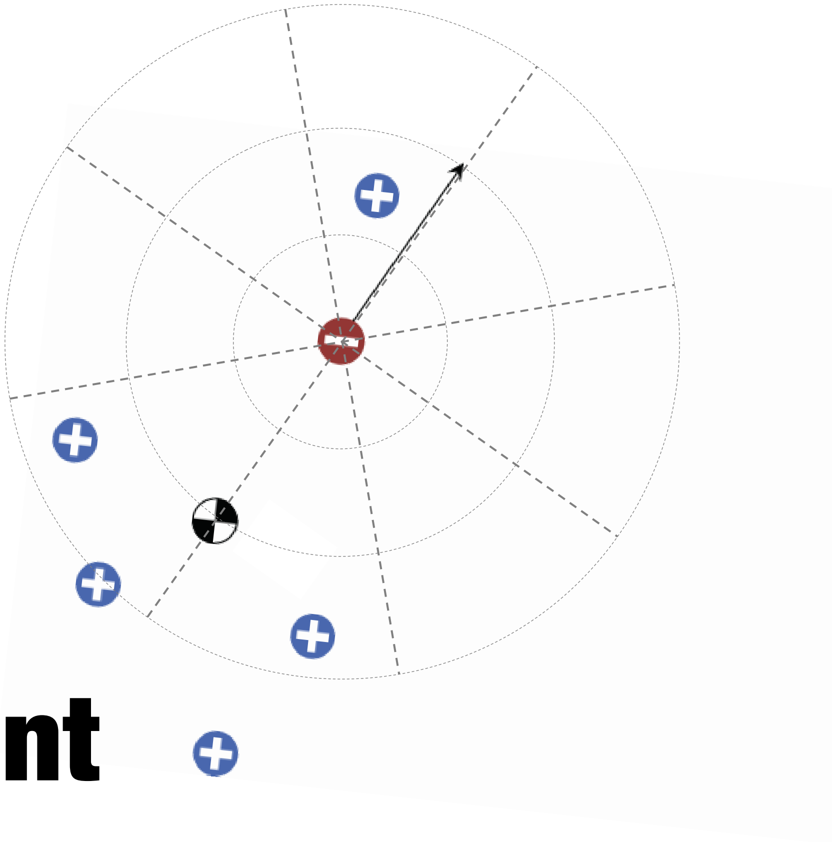
$$g\Big(\underset{\text{Social formation}}{\underline{\{\oplus_i\}_{i=1}^{N}}}\Big) = \ominus$$

where *N* is the number of social members.

Scale variation
Orientation variation

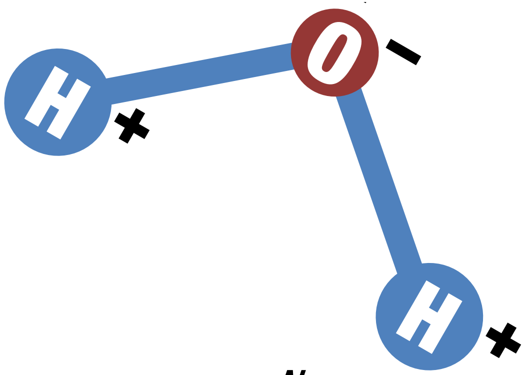
⊖ : Ground truth joint attention

⊕ : Head location



Representation: Social Dipole Moment

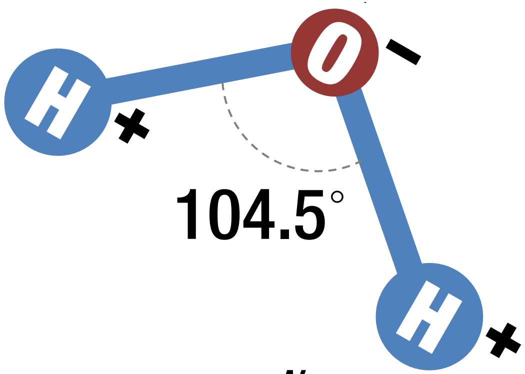
Water molecule, H_2O



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

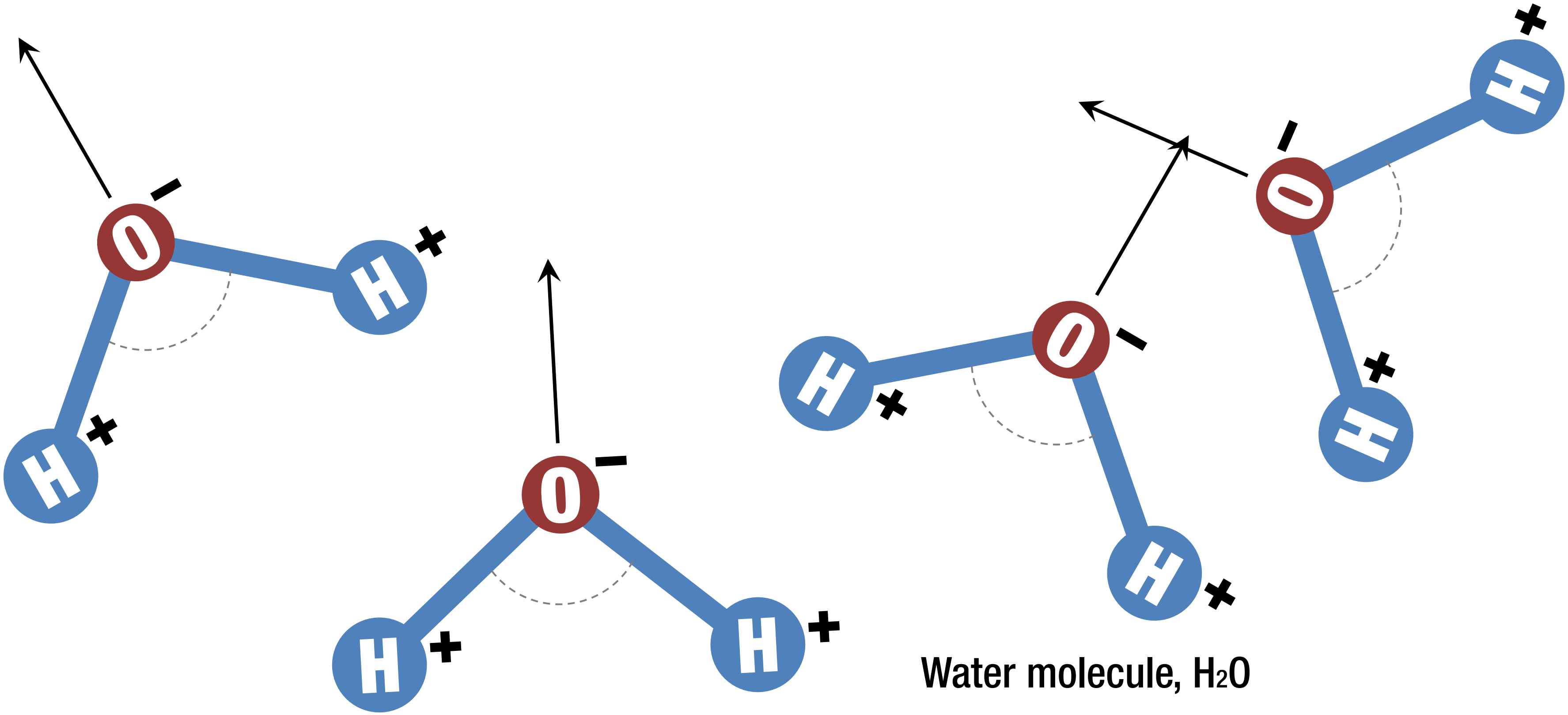
Electric dipole moment

Water molecule, H_2O

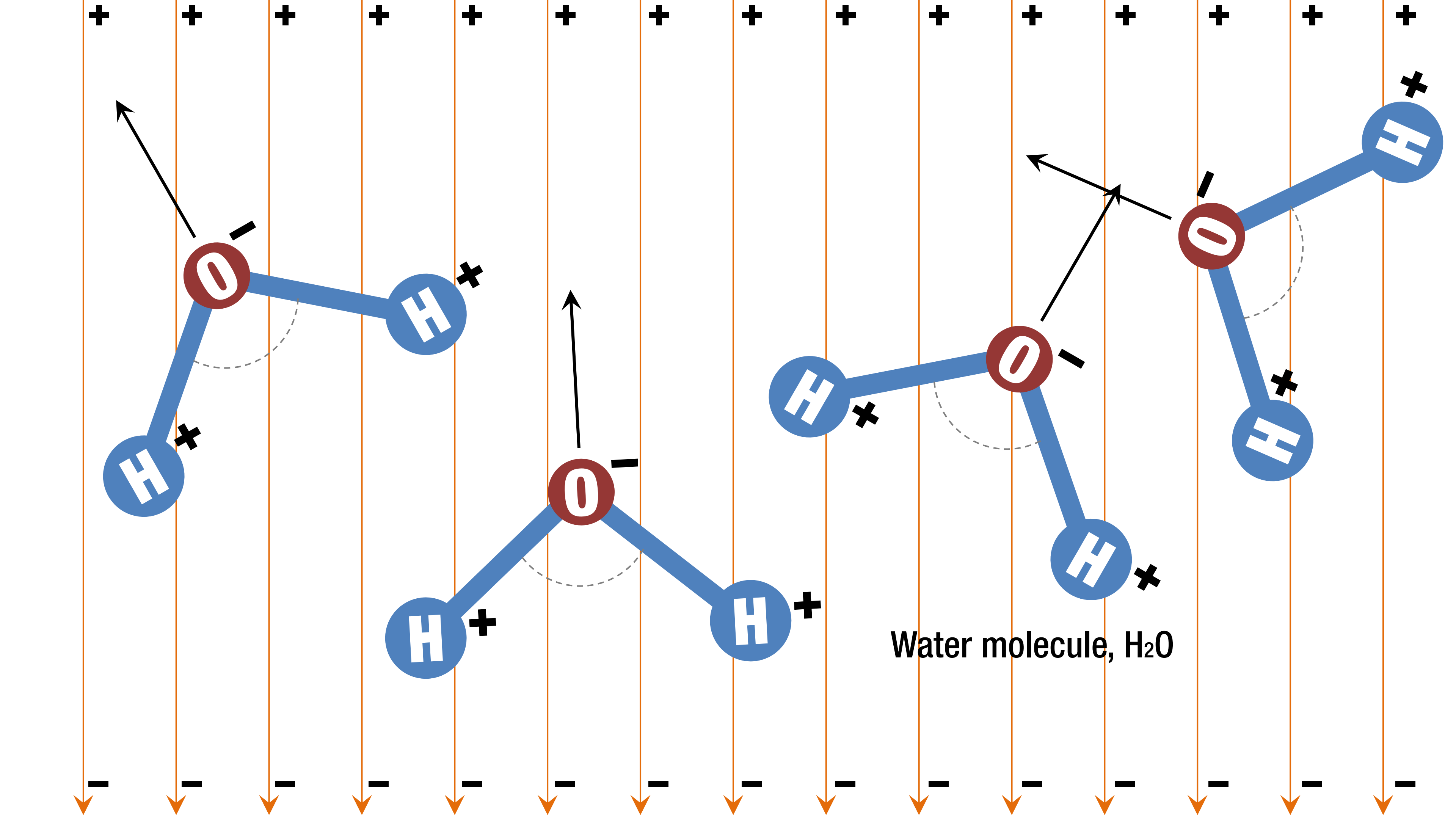


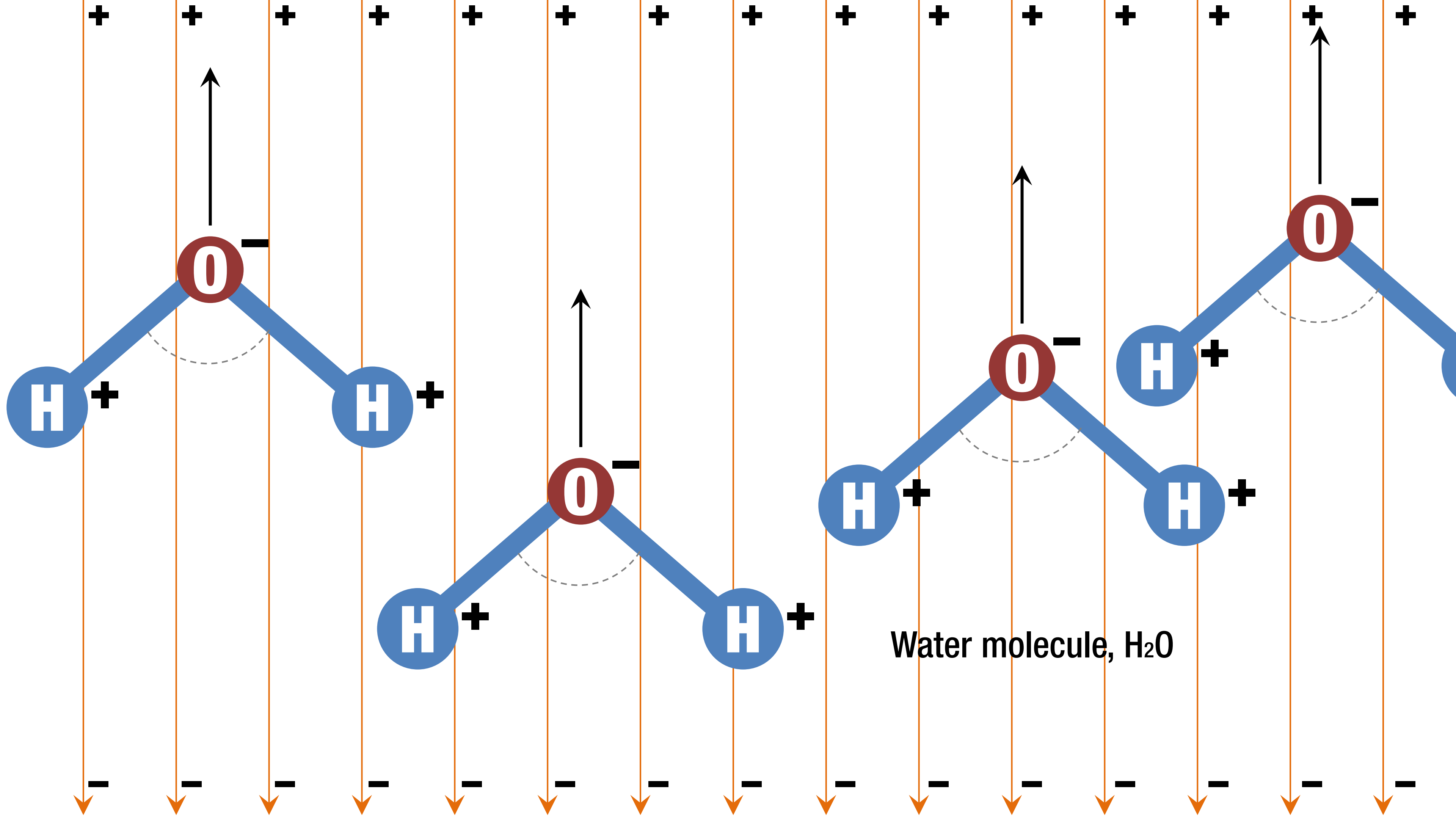
$$\mathbf{q}_e = \sum_{i}^{N} (\mathbf{e} - \mathbf{p}_i)$$

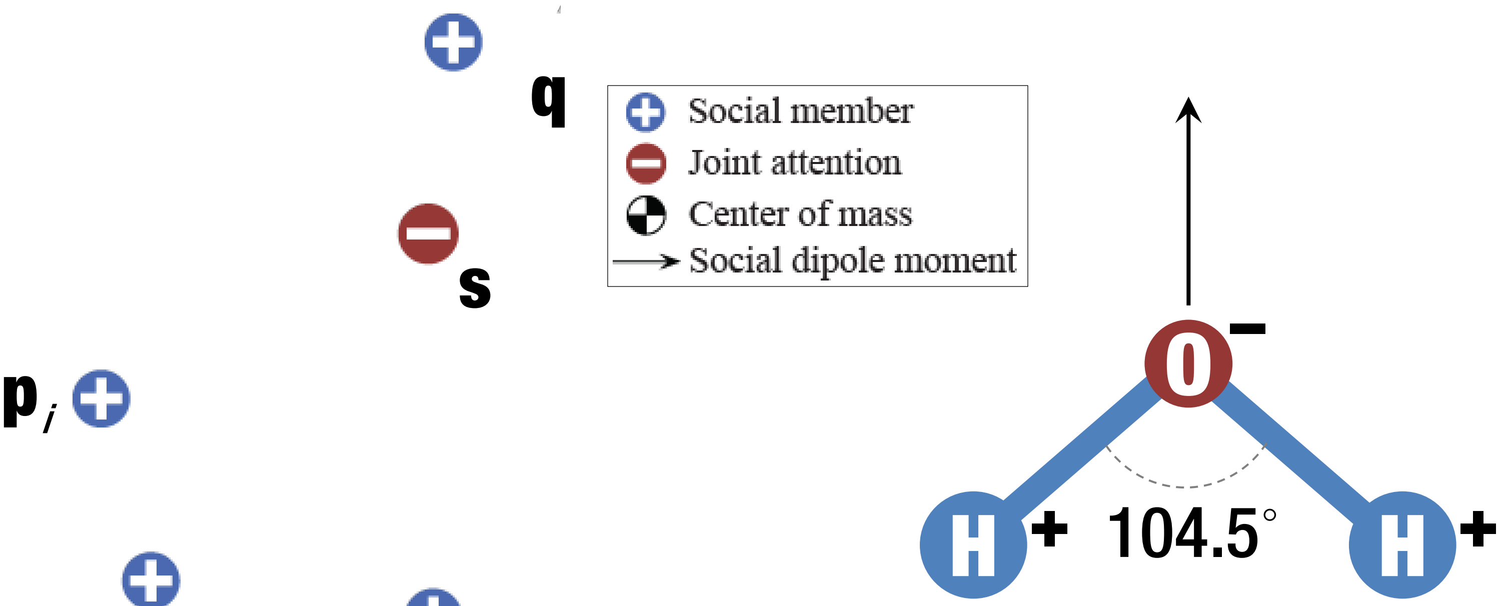
Electric dipole moment



Water molecule, H_2O


+
−
O
H
Water molecule, H₂O


+
−
O
H
Water molecule, H_2O

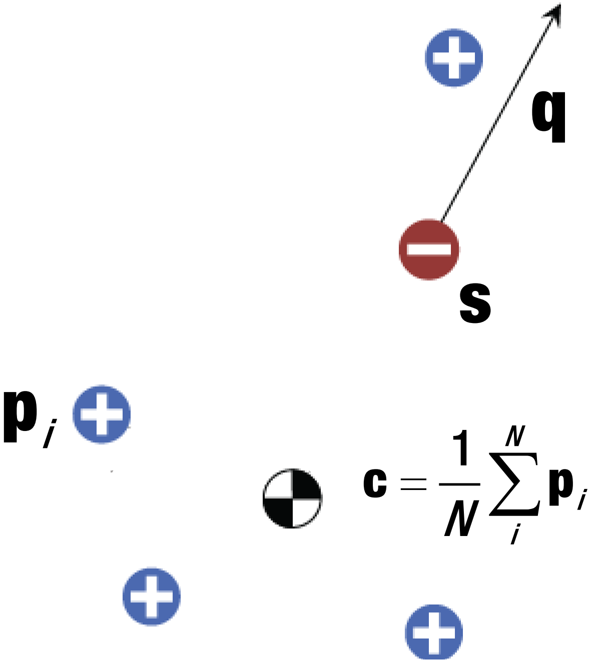


Social dipole moment

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

Electric dipole moment

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



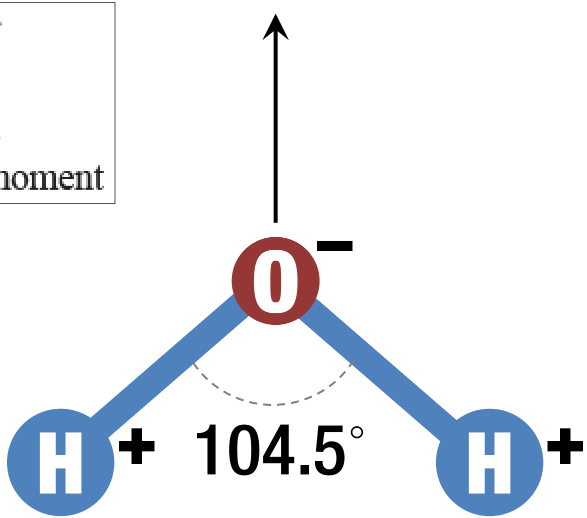
$$\mathbf{c} = \frac{1}{N}\sum_{i}^{N}\mathbf{p}_i$$

⊕ Social member

⊖ Joint attention

◐ Center of mass

→ Social dipole moment

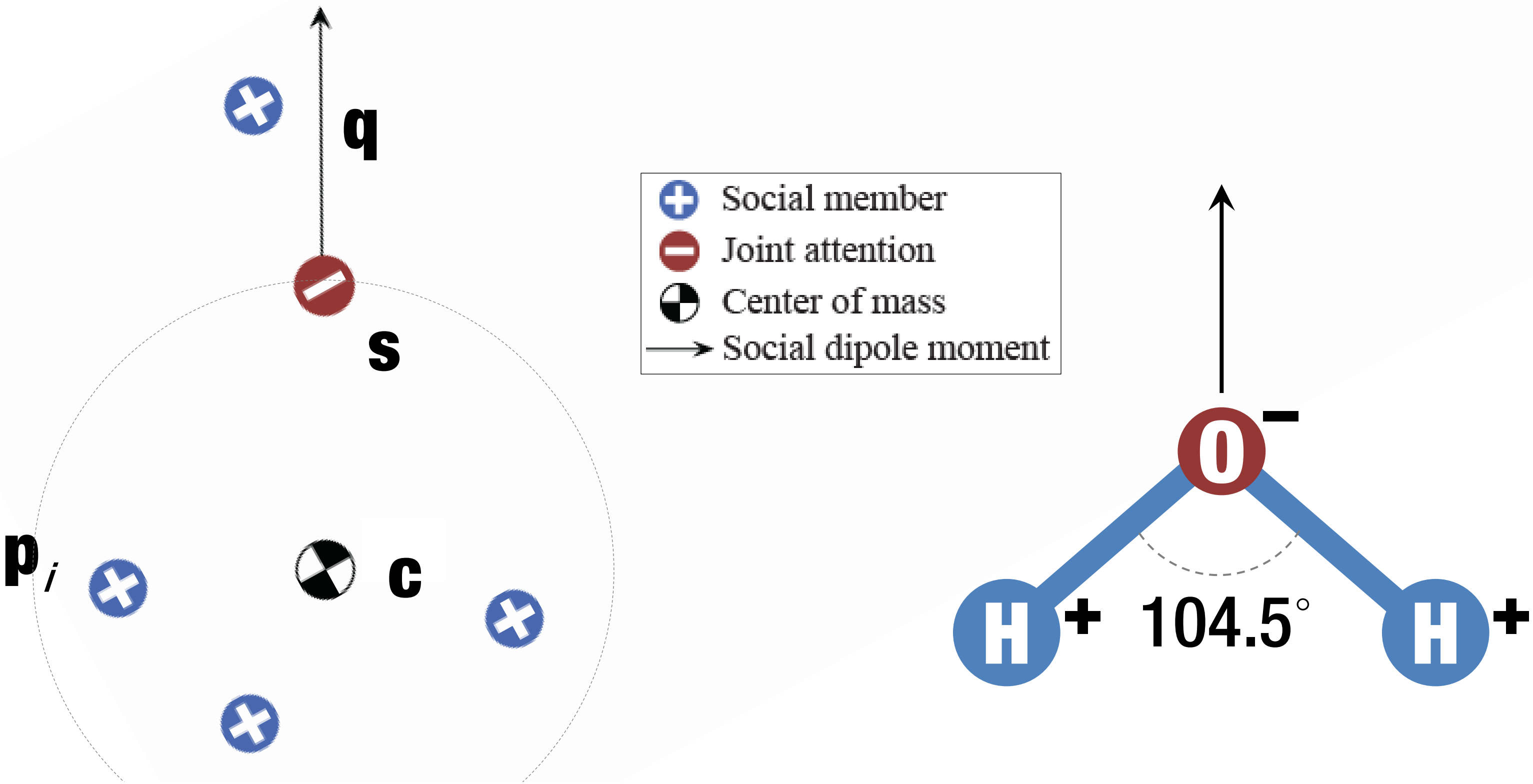


Social dipole moment

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

Electric dipole moment

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

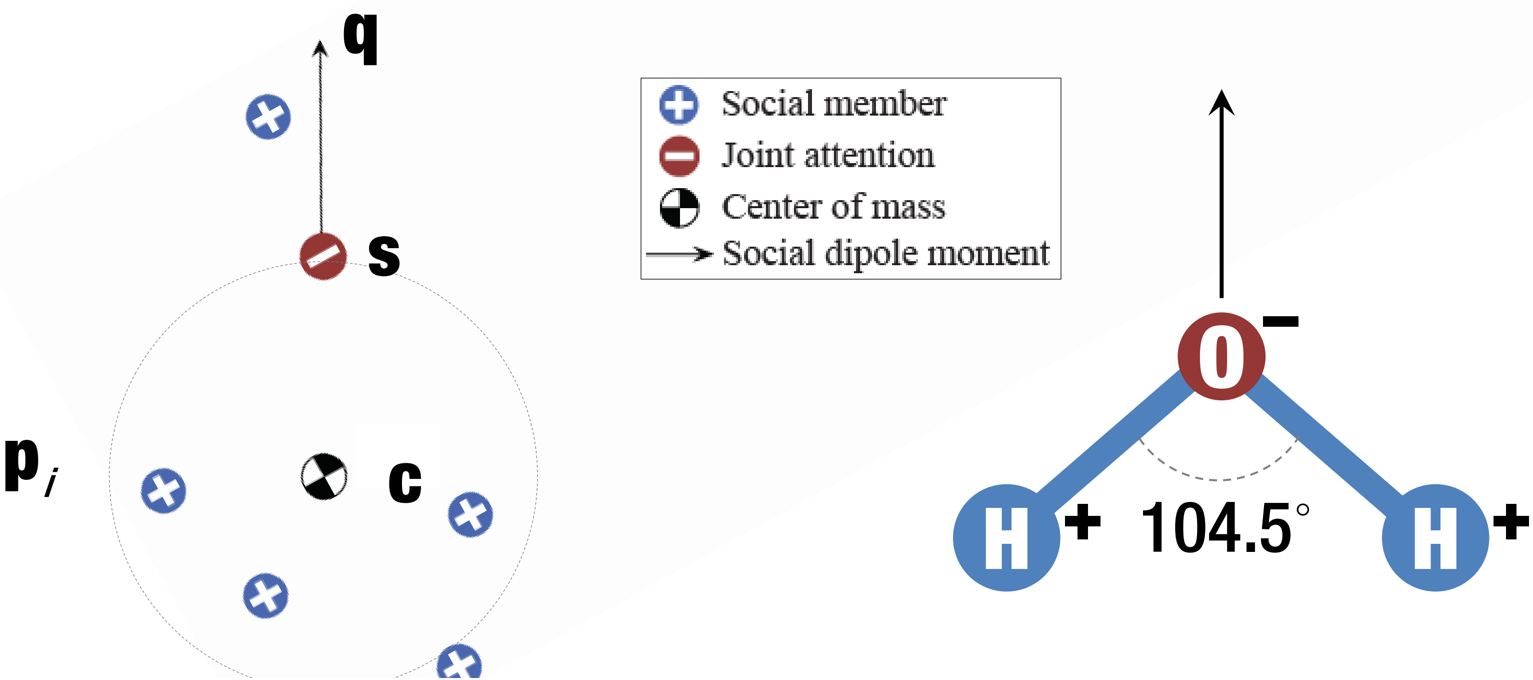


Social dipole moment

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

Electric dipole moment

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

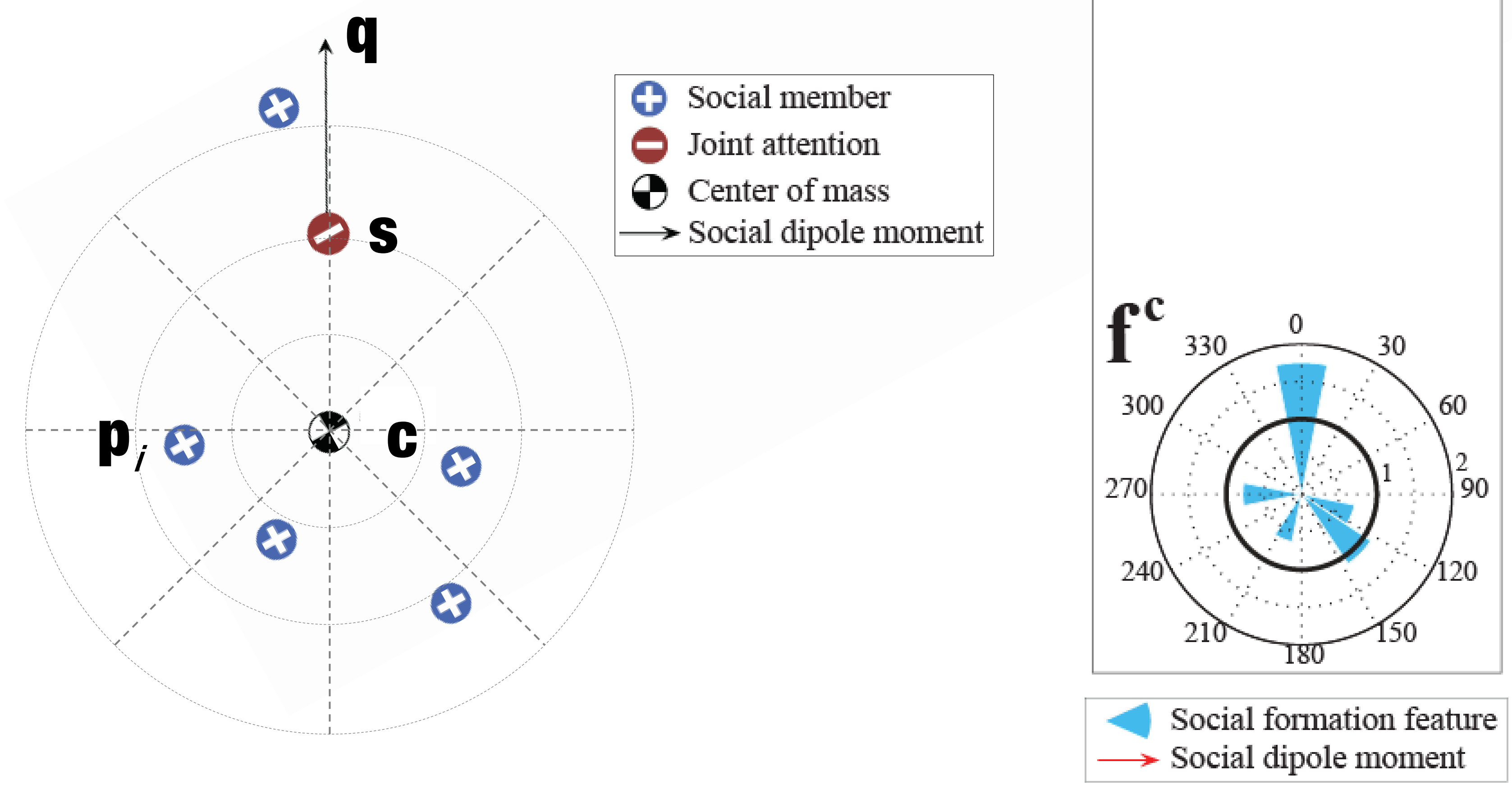


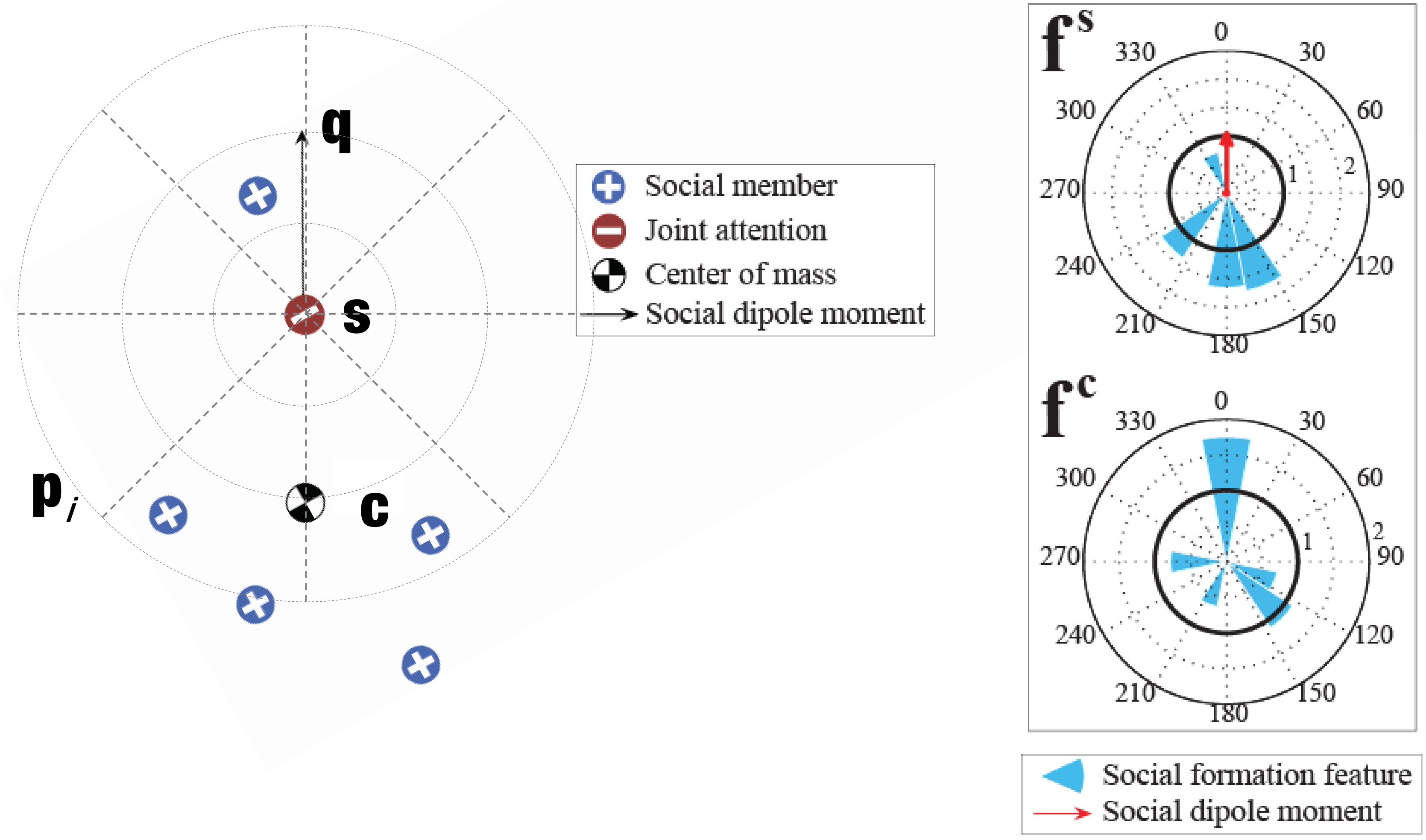
Social dipole moment

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

Electric dipole moment

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

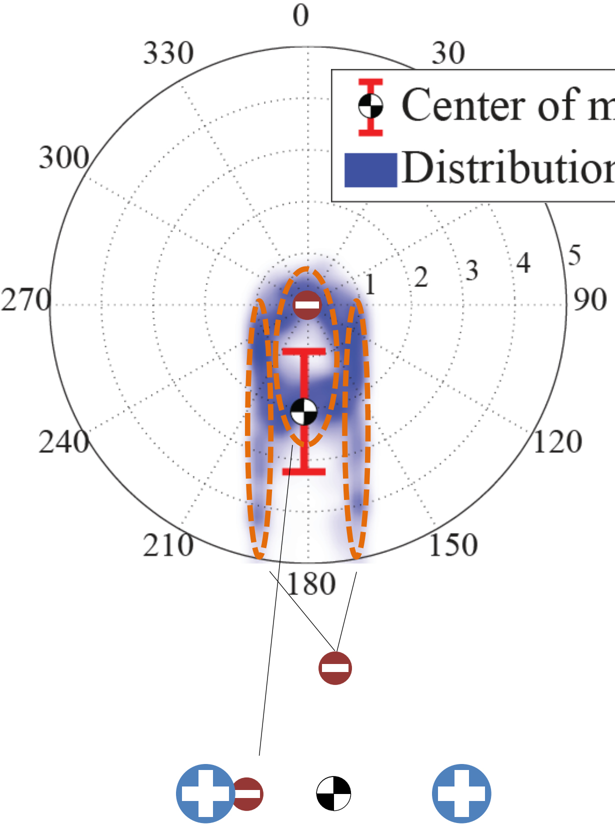

q
s
p_i
c
Social member
Joint attention
Center of mass
Social dipole moment
f^c
0
30
60
90
120
150
180
210
240
270
300
330
1
2
Social formation feature
Social dipole moment

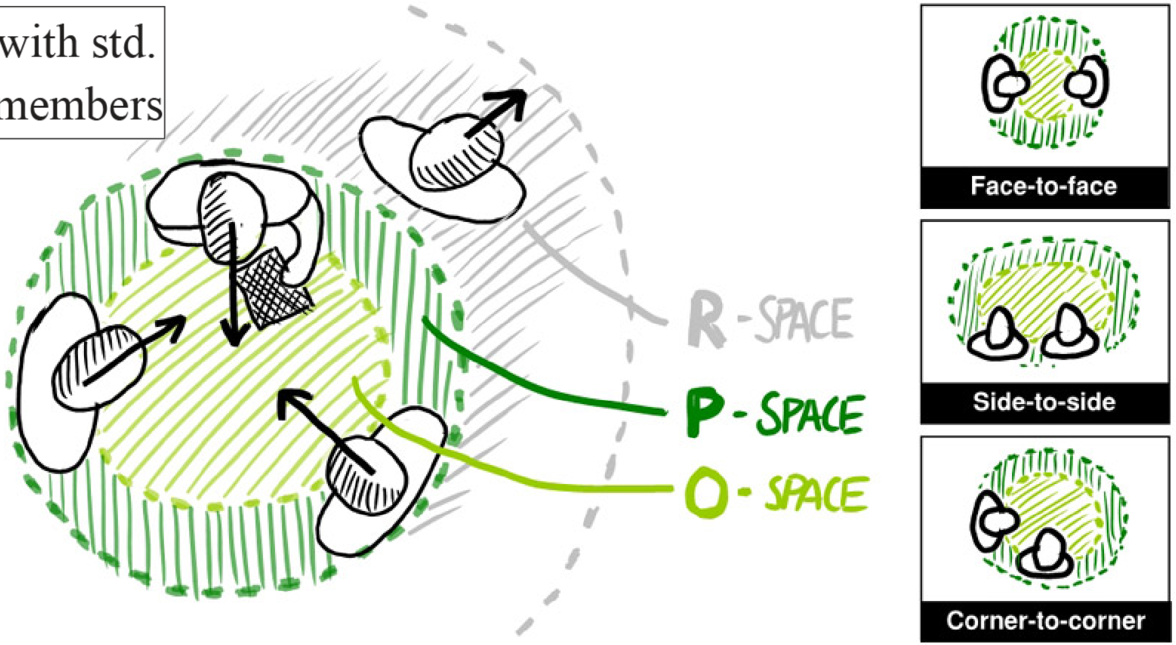

q
s
$\mathbf{p}_i$
c
Social member
Joint attention
Center of mass
Social dipole moment
$\mathbf{f}^s$
0
30
60
90
120
150
180
210
240
270
300
330
1
2
$\mathbf{f}^c$
0
30
60
90
120
150
180
210
240
270
300
330
1
2
Social formation feature
Social dipole moment

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

Dyadic interaction

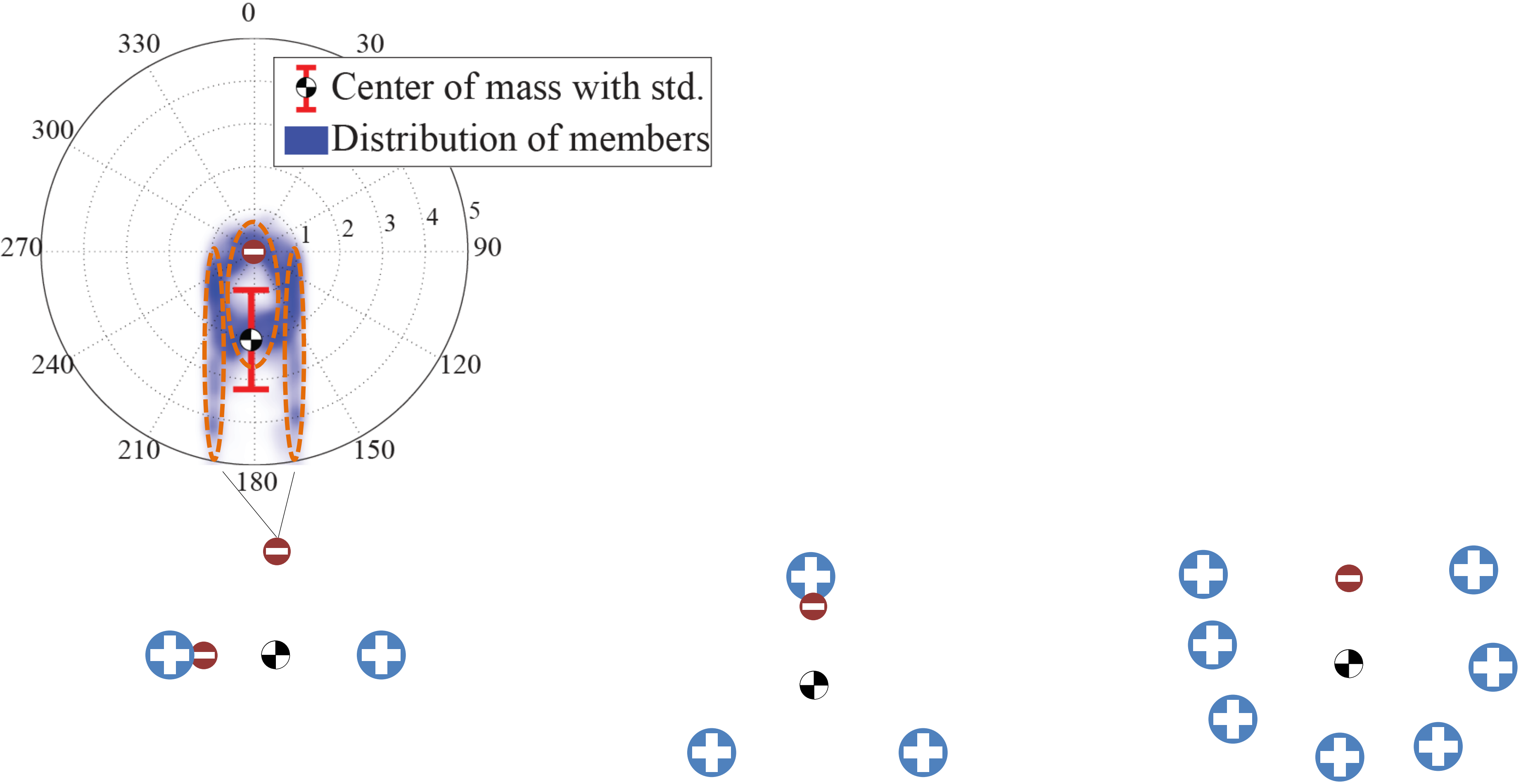




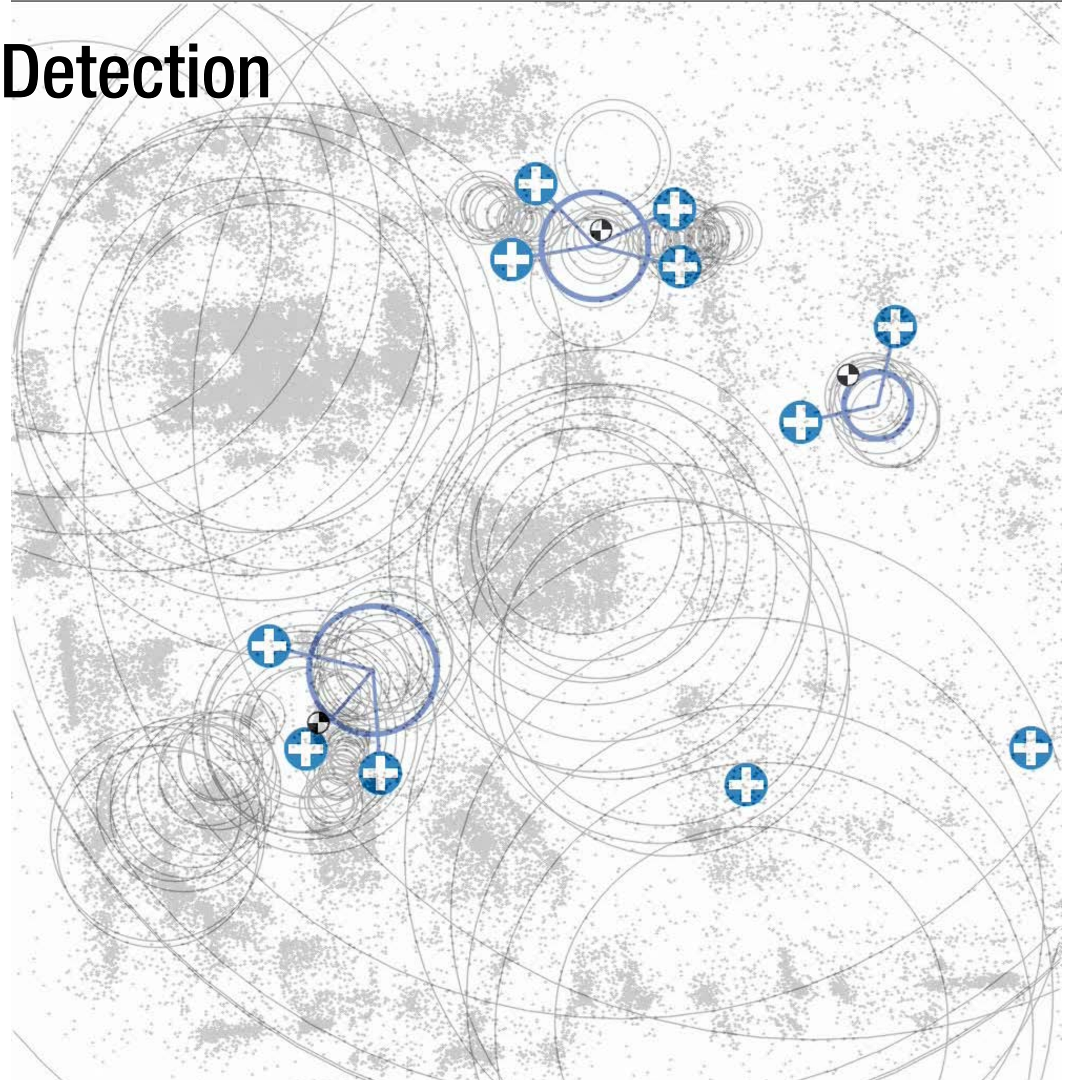
Social formation theory (F-formation)*

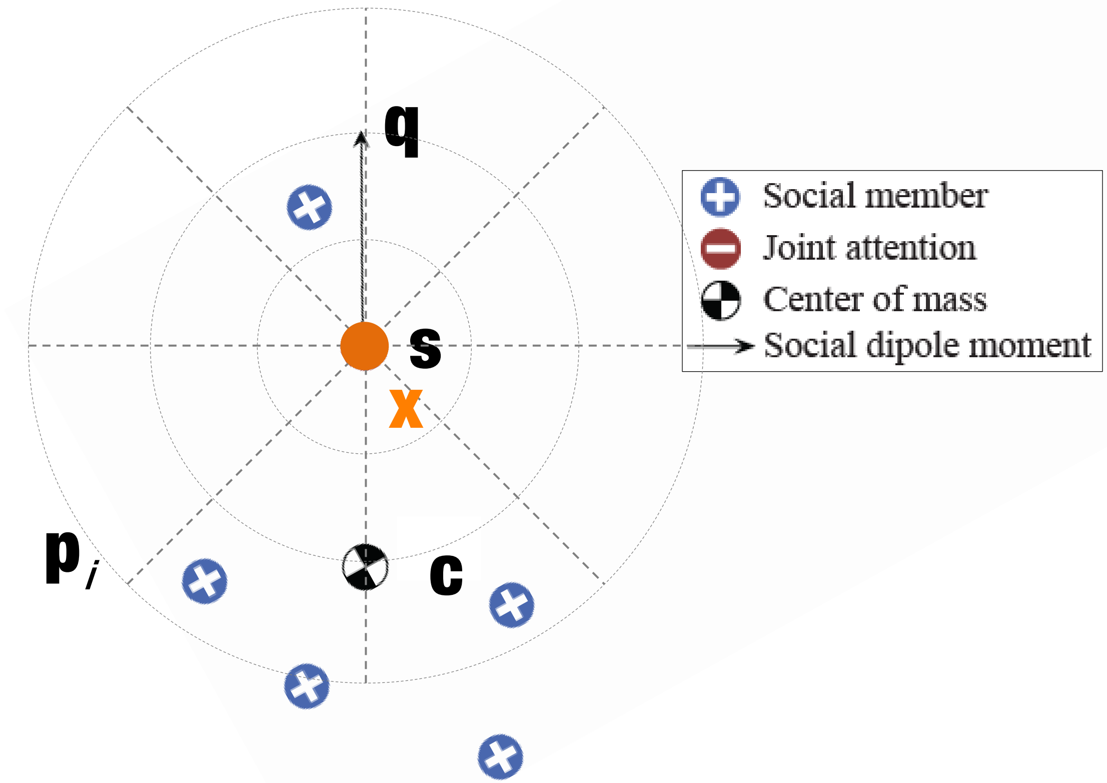
*A. Kendon, “Conducting Interaction: Patterns of Behavior in Focused Encounters”, Cambridge University Press, 1990.

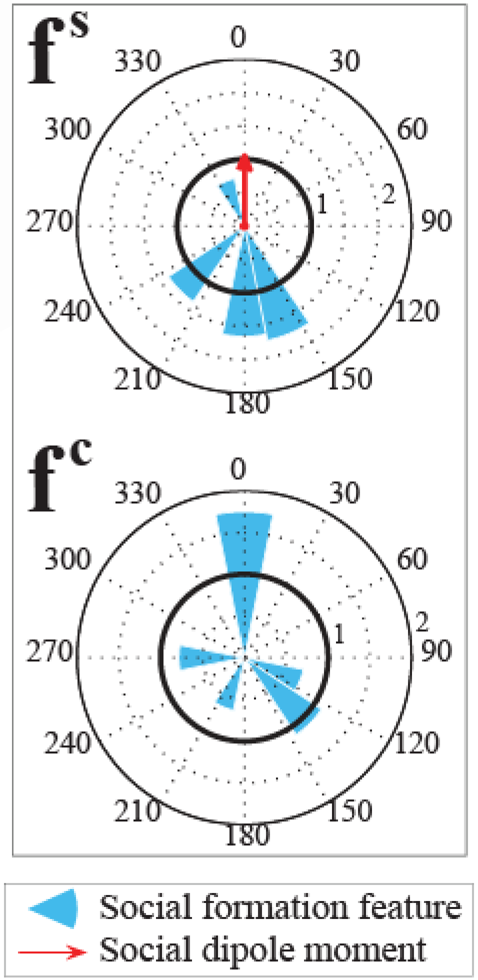
Dyadic interaction



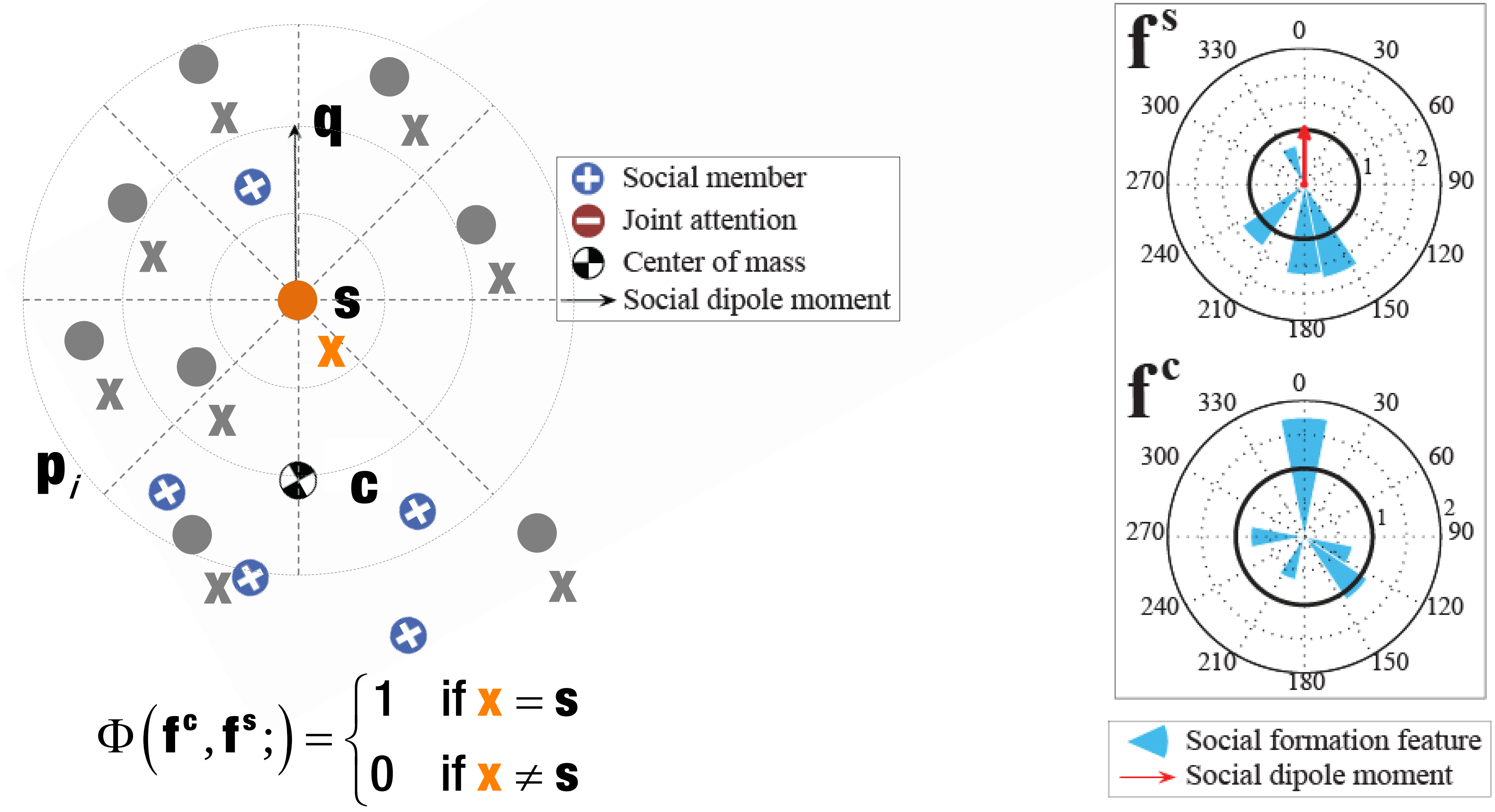
Social Group Detection



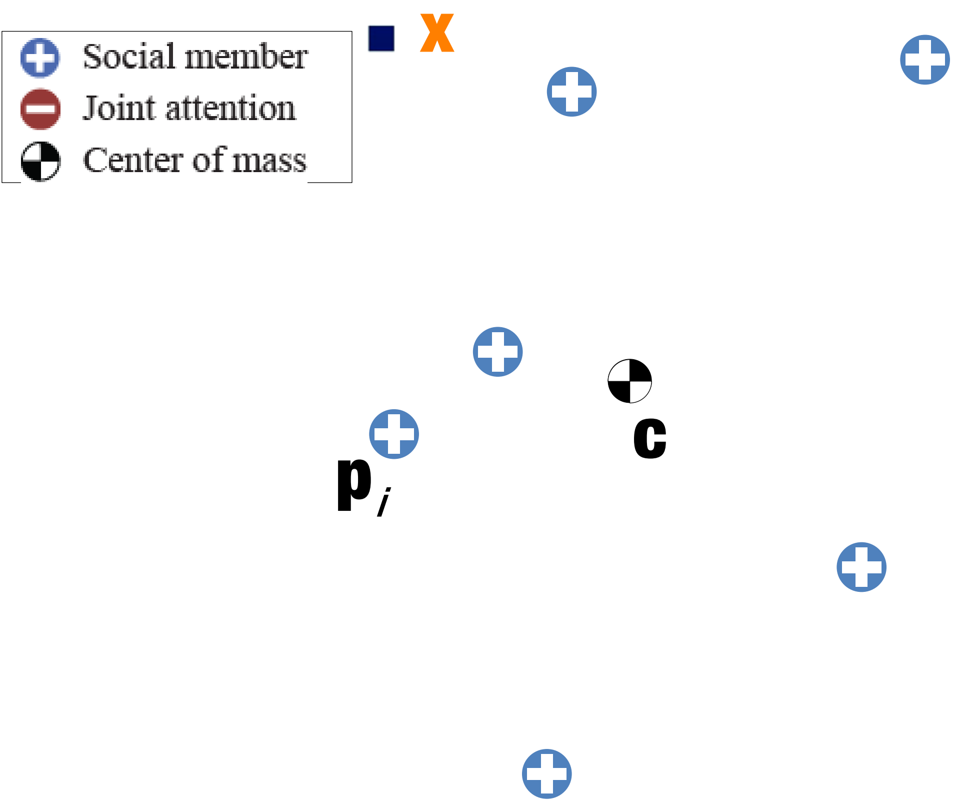




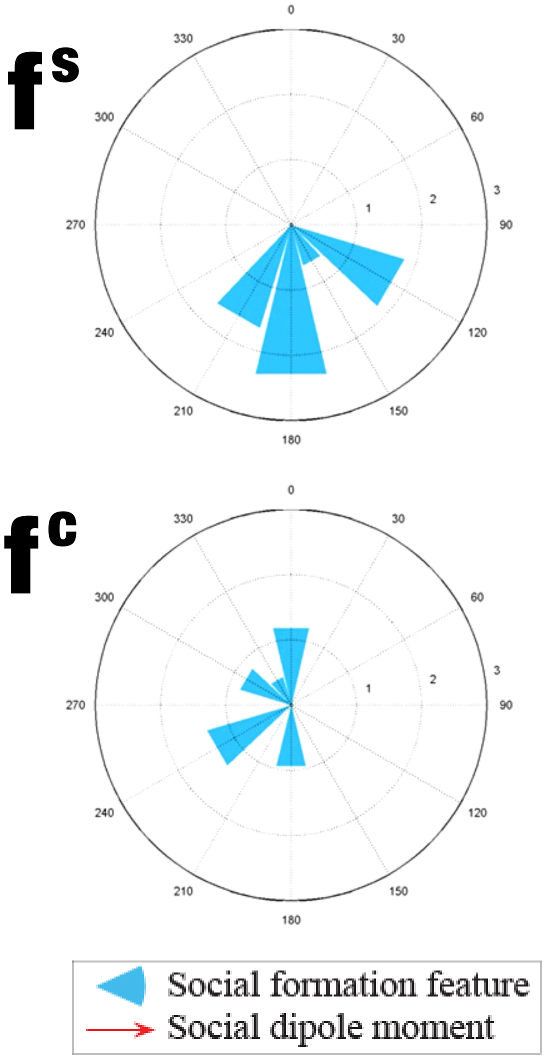
$$\Phi\left(\mathbf{f}^{c},\mathbf{f}^{s};\right)=\begin{cases}1 & \text{if } \mathbf{x}=\mathbf{s}\\ & \end{cases}$$

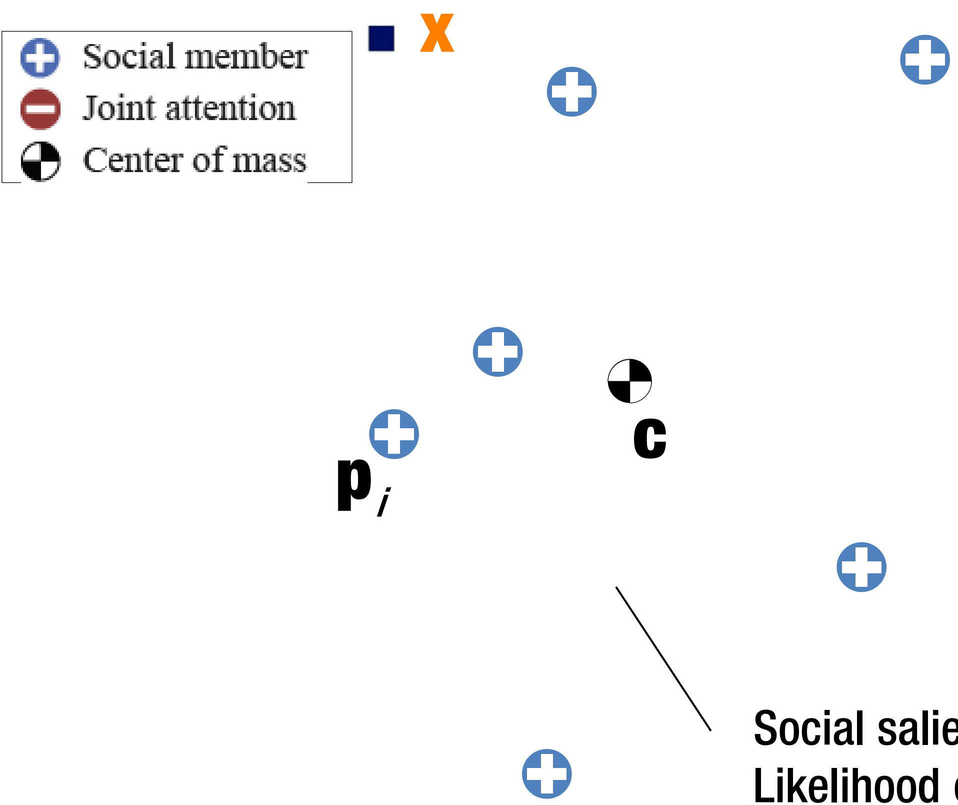


$$\Phi\left(\mathbf{f}^c, \mathbf{f}^s;\right) = \begin{cases} 1 & \text{if } \mathbf{x} = \mathbf{s} \\ 0 & \text{if } \mathbf{x} \neq \mathbf{s} \end{cases}$$



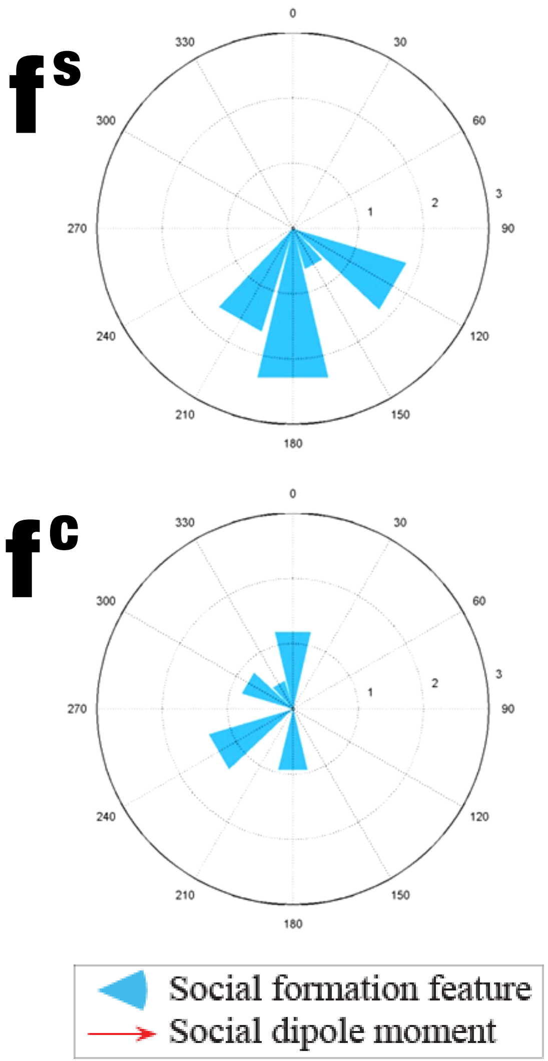
$$\Phi\left(\mathbf{f}^{c},\mathbf{f}^{s};\right)=\begin{cases}1 & \text{if } \mathbf{x}=\mathbf{s}\\ 0 & \text{if } \mathbf{x}\neq\mathbf{s}\end{cases}$$

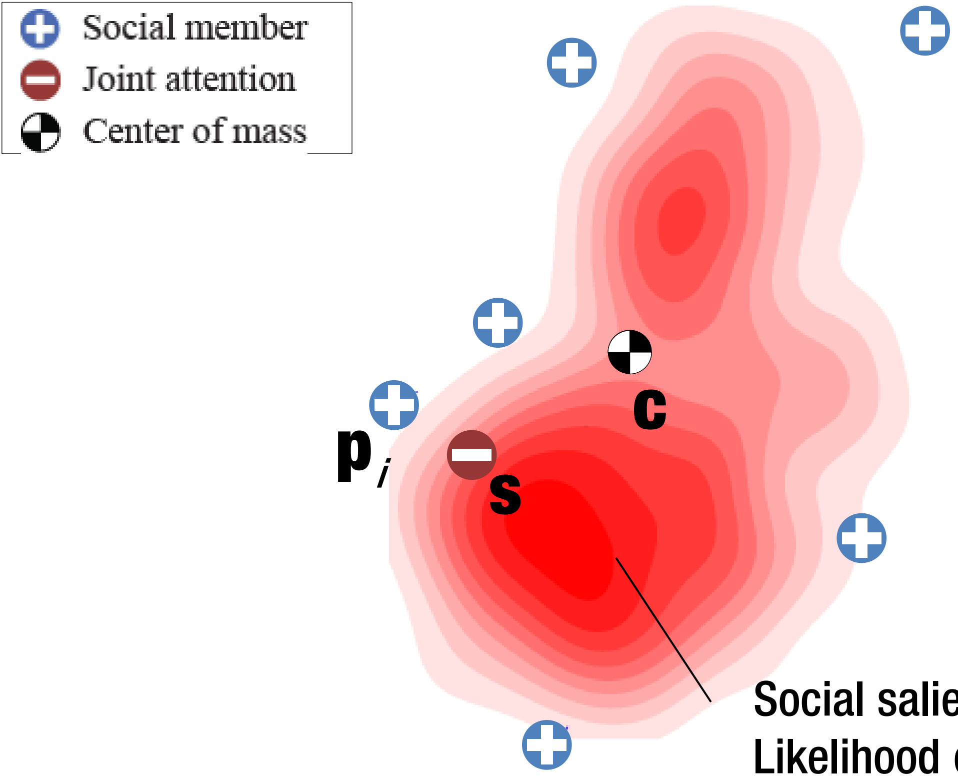




Social saliency:
Likelihood of joint attention

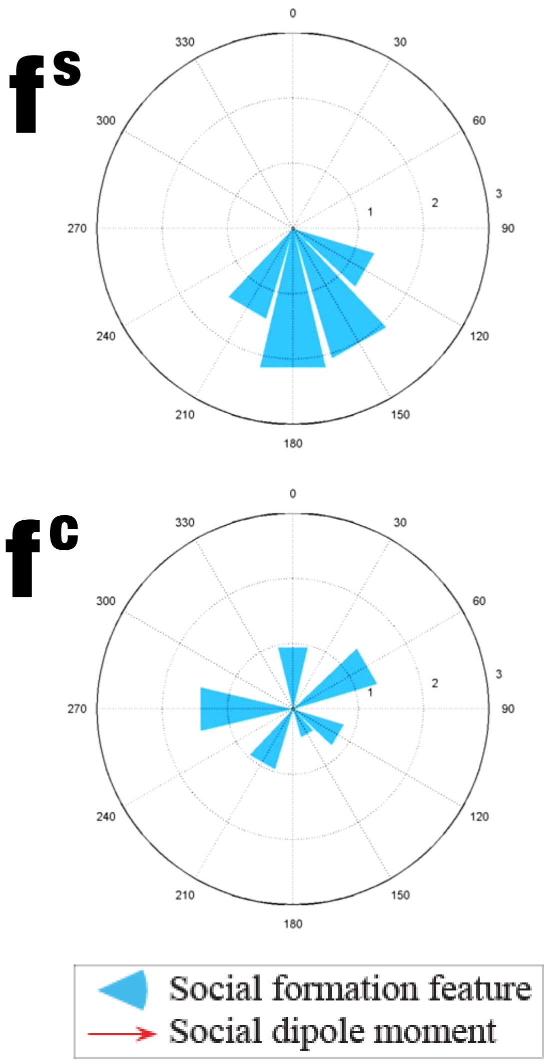
$$\Phi\left(\mathbf{f}^{c},\mathbf{f}^{s};\right)=\begin{cases}1 & \text{if } \mathbf{x}=\mathbf{s}\\ 0 & \text{if } \mathbf{x}\neq\mathbf{s}\end{cases}$$

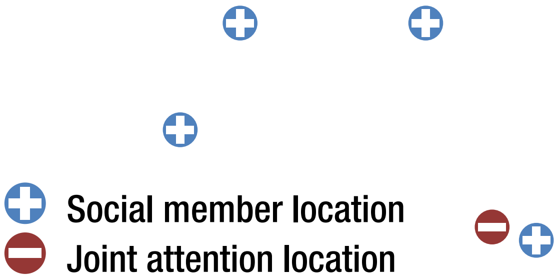




Social saliency:
Likelihood of joint attention

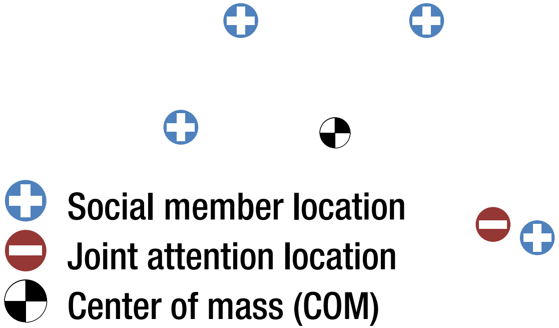
$$\Phi\left(\mathbf{f}^{c}, \mathbf{f}^{s};\right) = \begin{cases} 1 & \text{if } \mathbf{x} = \mathbf{s} \\ 0 & \text{if } \mathbf{x} \neq \mathbf{s} \end{cases}$$

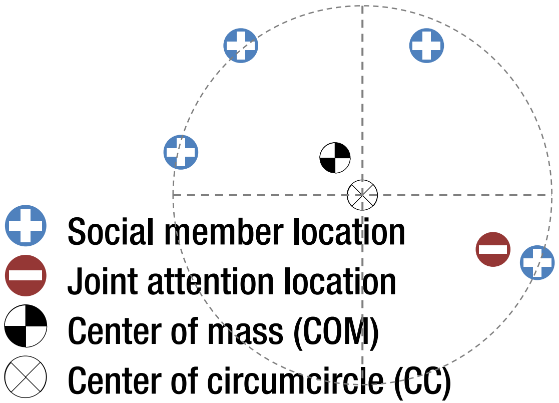


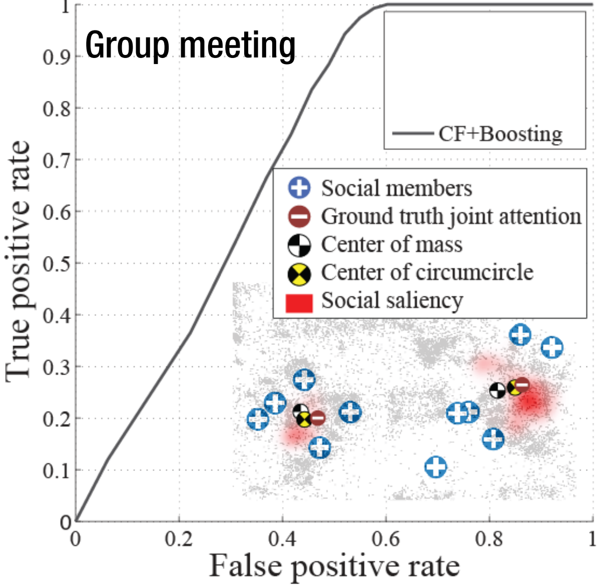


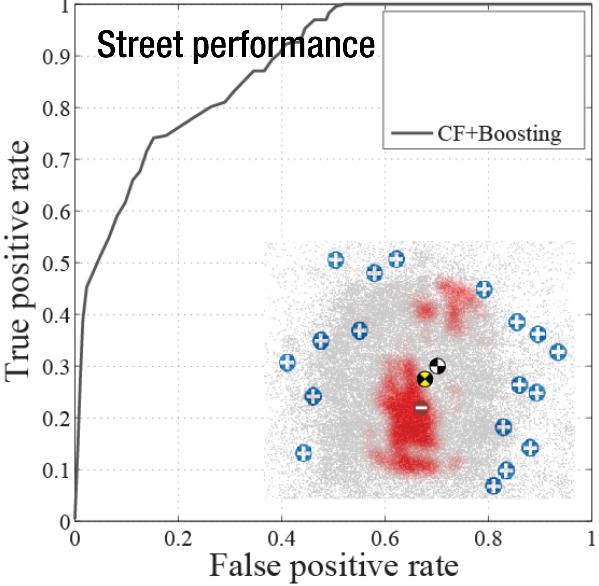
⊕ Social member location

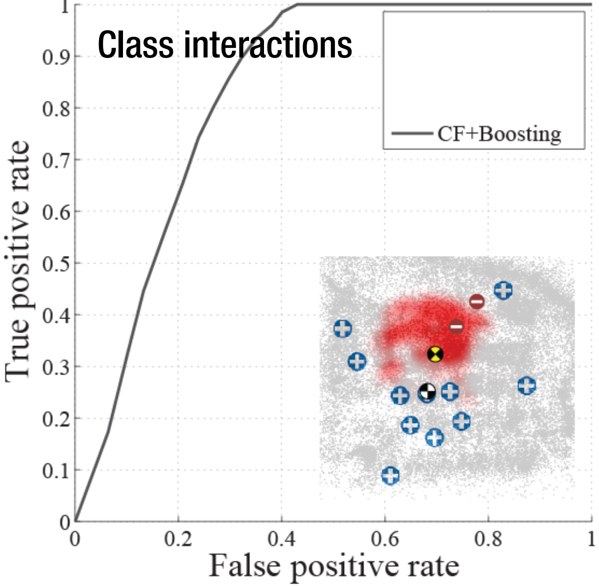
⊖ Joint attention location


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

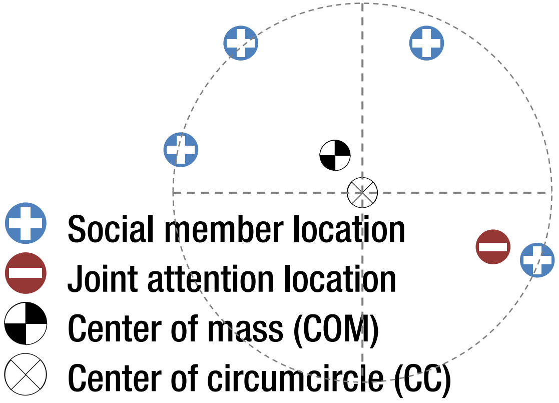

Social member location
Joint attention location
Center of mass (COM)
Center of circumcircle (CC)







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

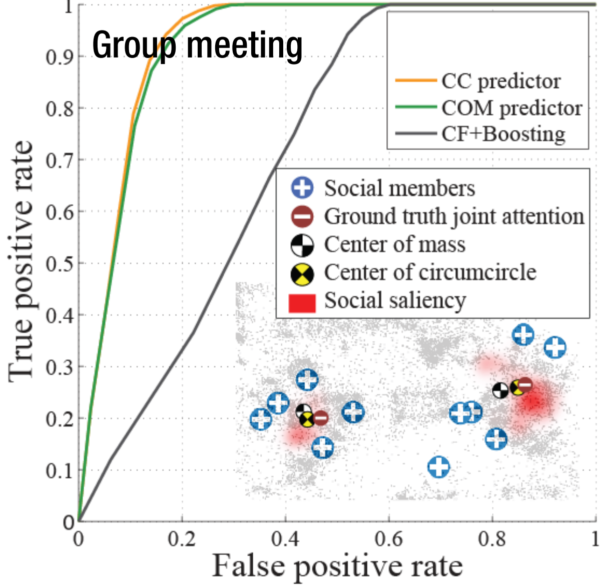


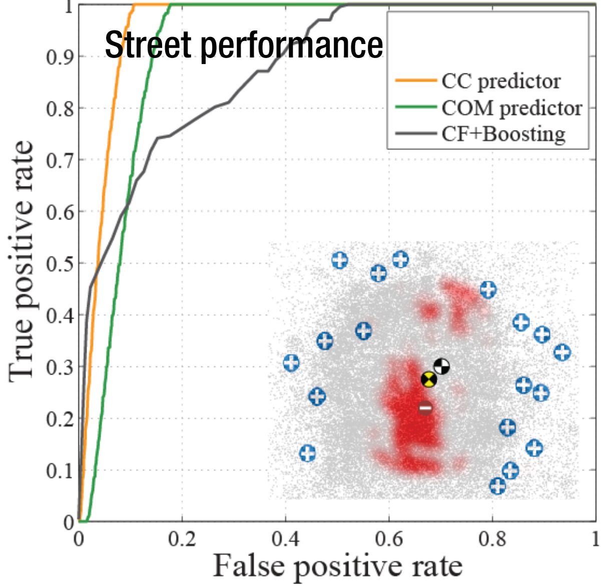
Social member location

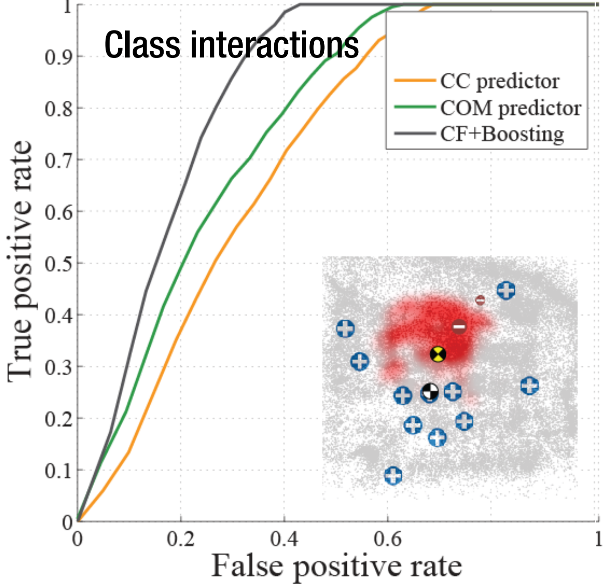
Joint attention location

Center of mass (COM)

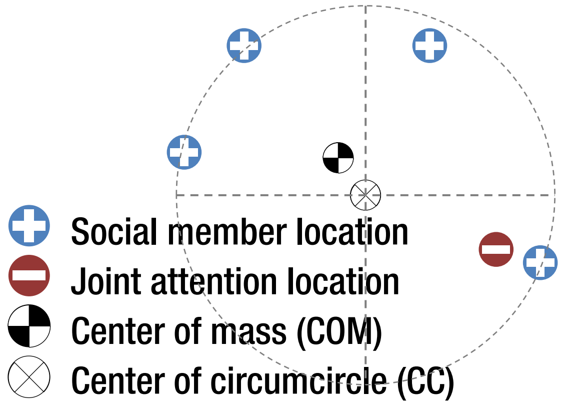
Center of circumcircle (CC)





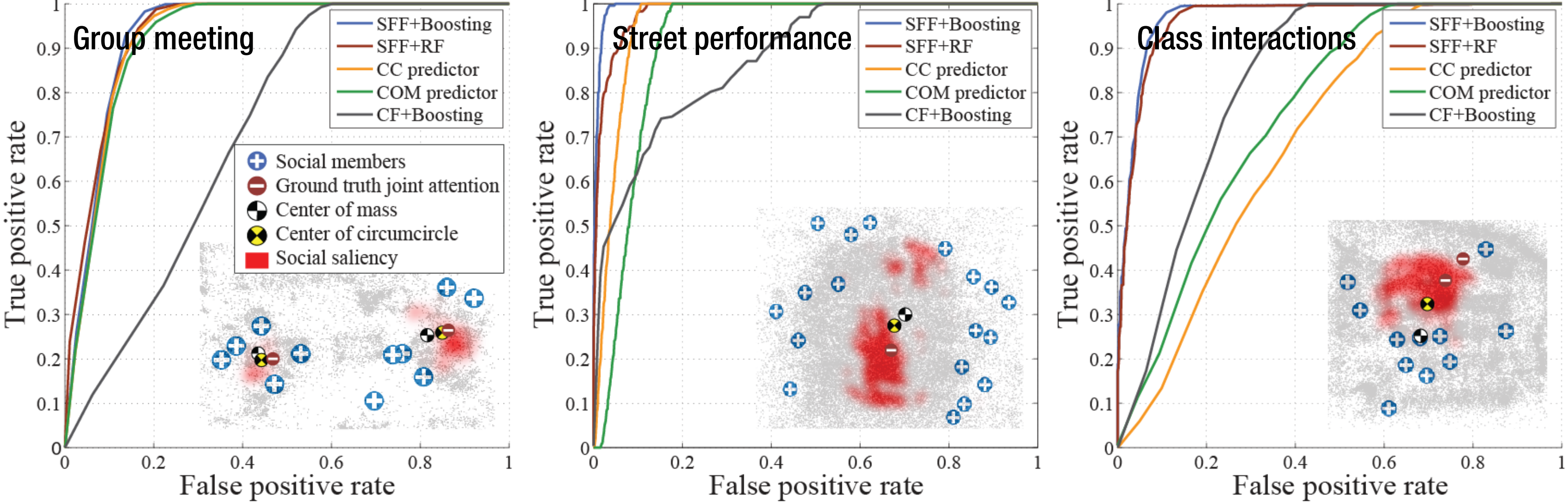


CF: Context feature (Lan et al., PAMI 2012)

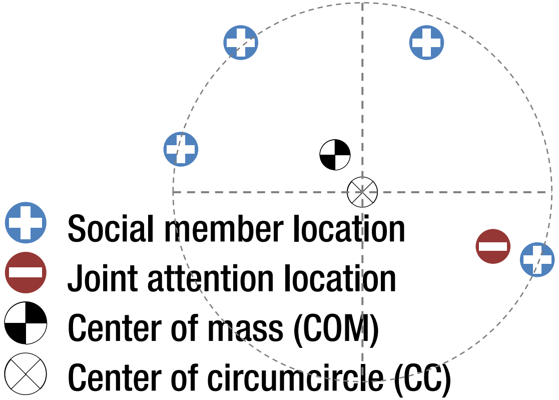


Mean average precision

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

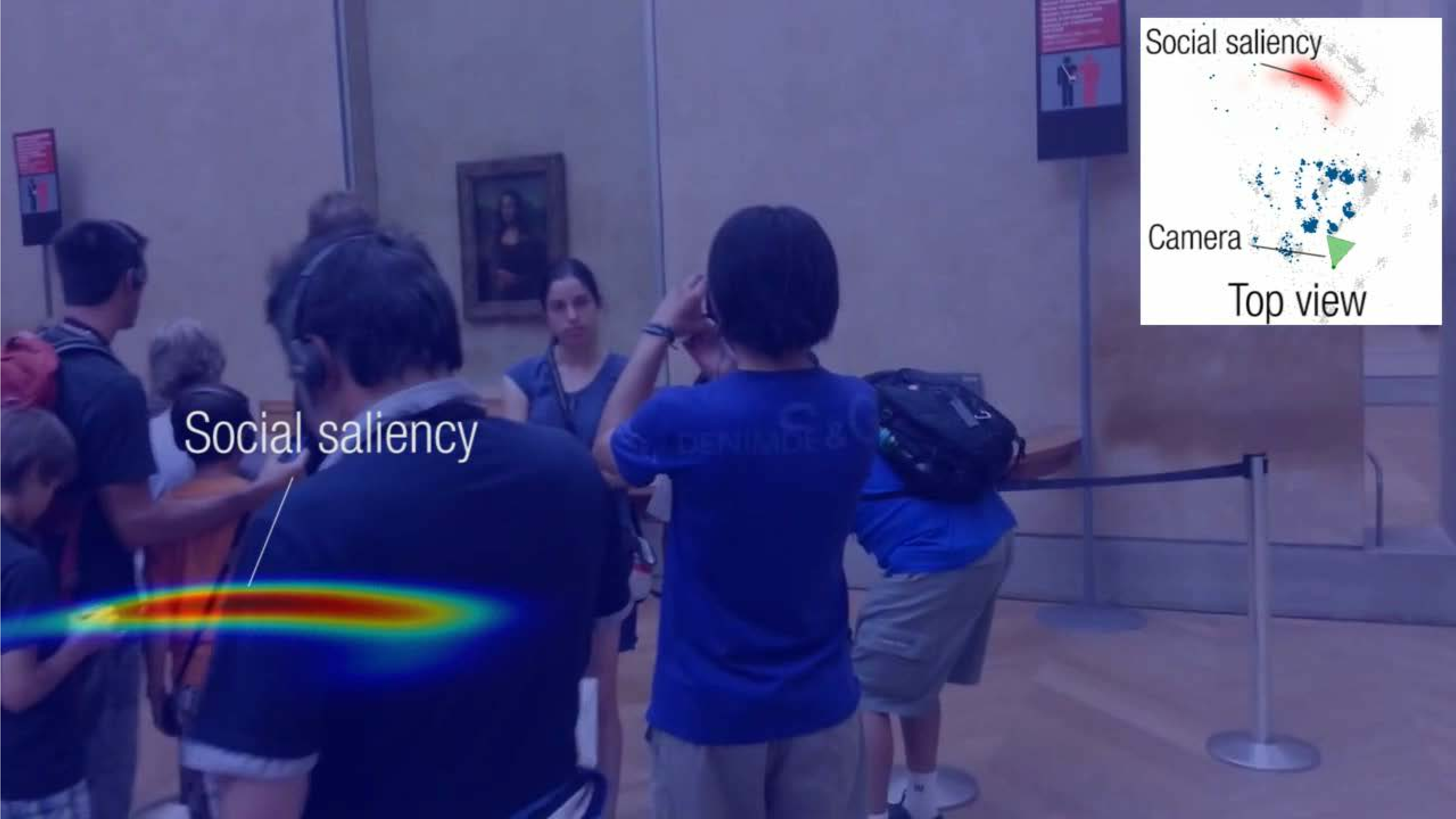


CF: Context feature (Lan et al., PAMI 2012)



Mean average precision

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

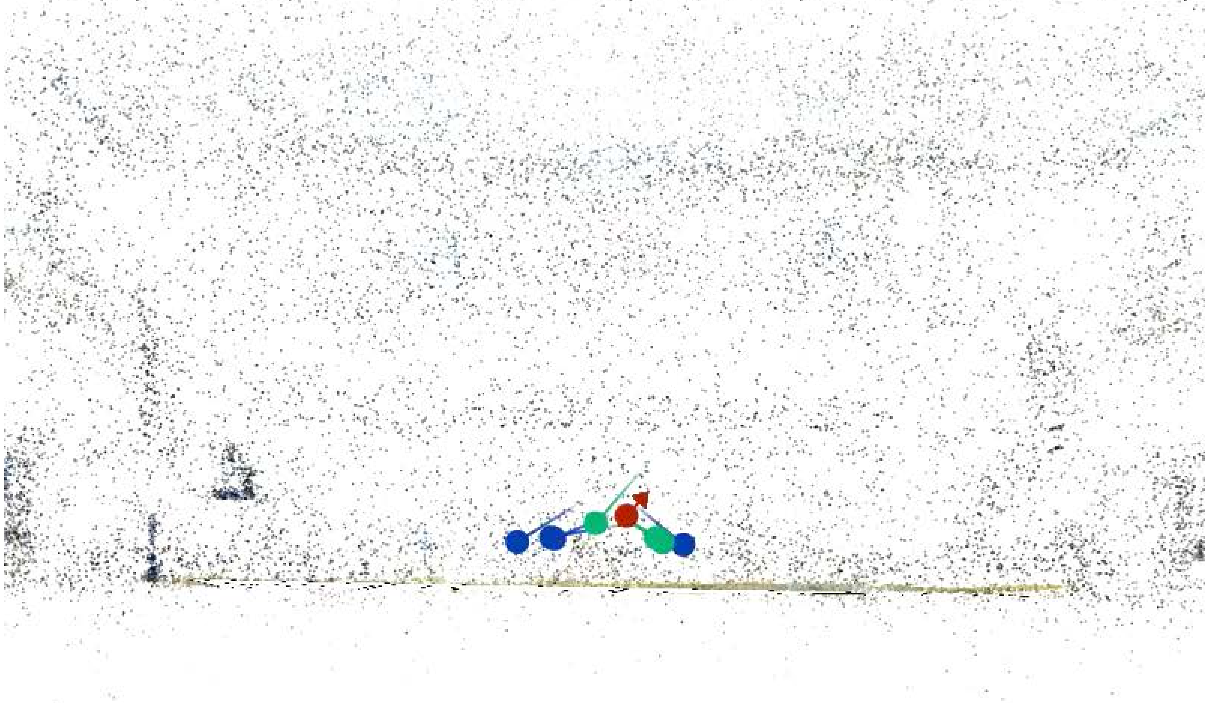

Social saliency
Social saliency
Camera
Top view

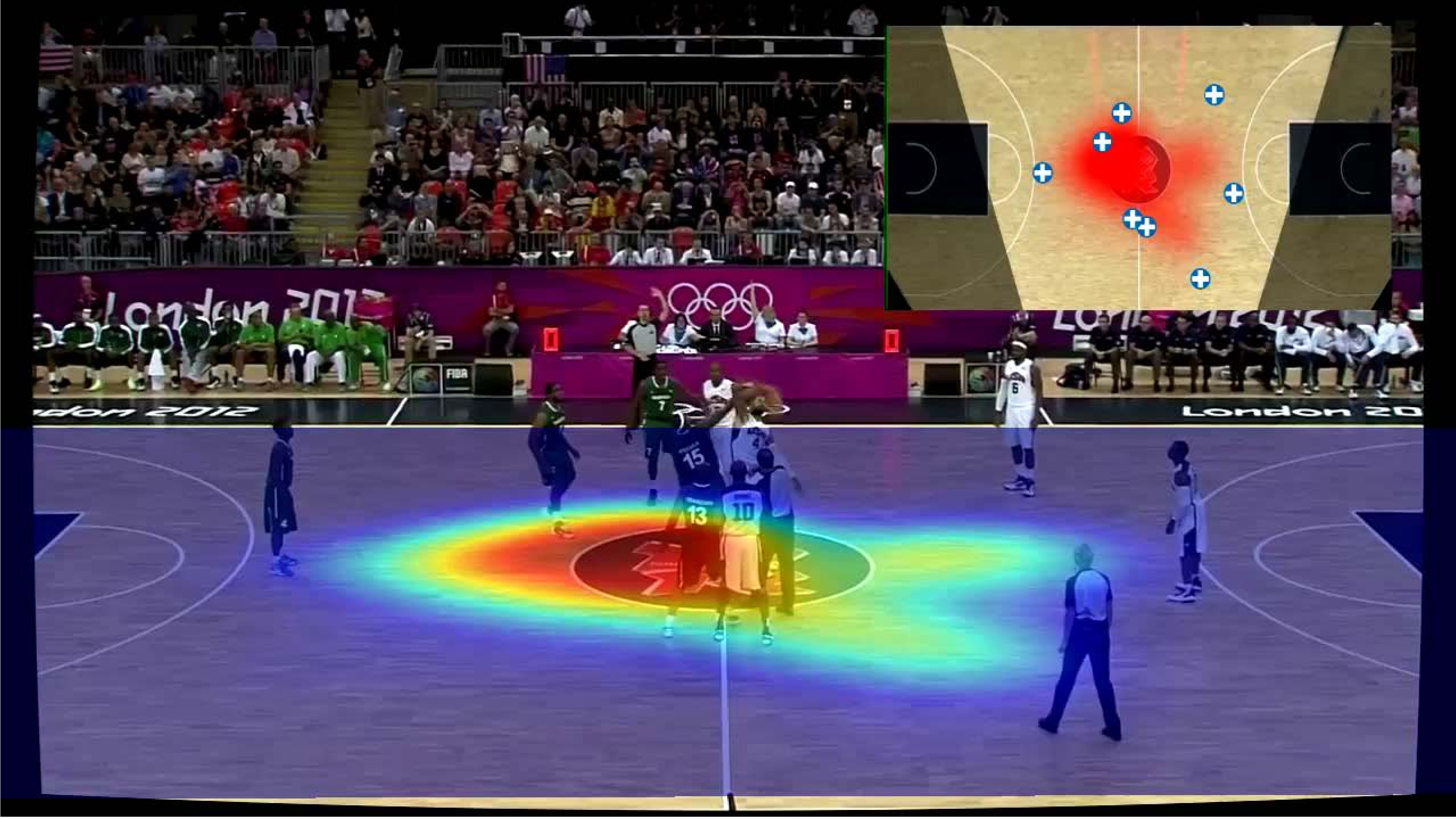
Time Square

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

USA 5
NGR 0
9:25 1ST
Input Video
London 2012
London 2012
London 2012
4

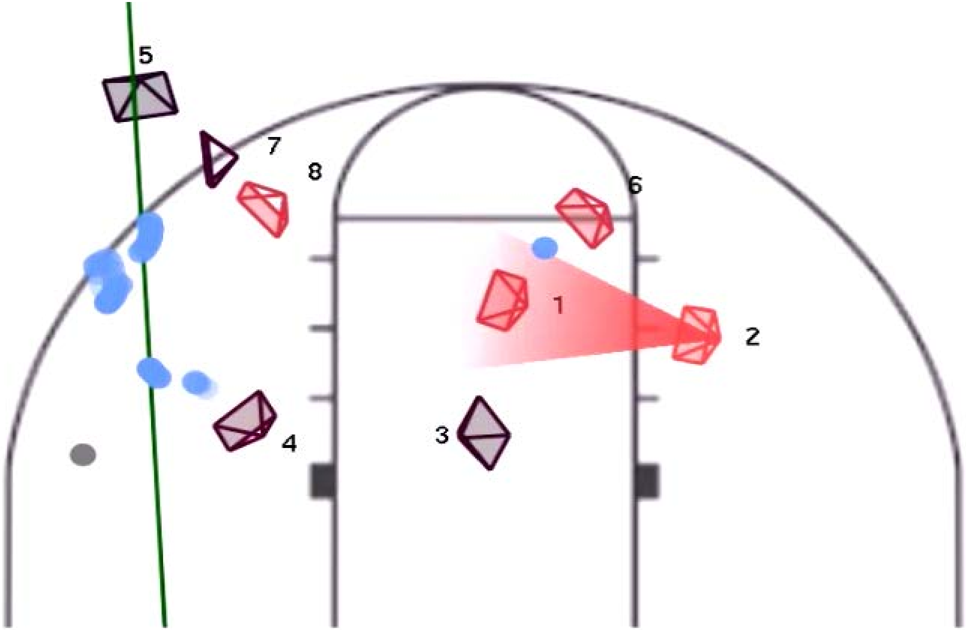
First Person Basketball Data

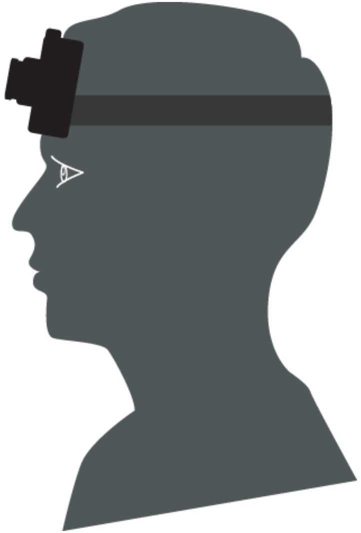


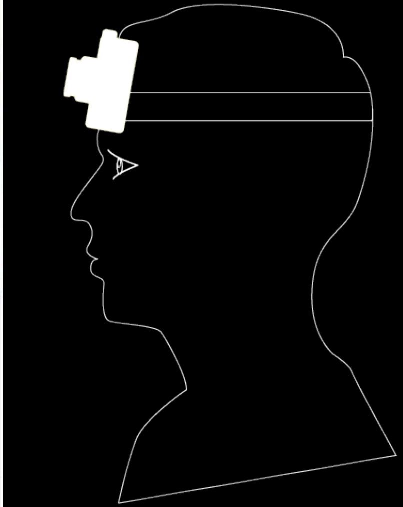


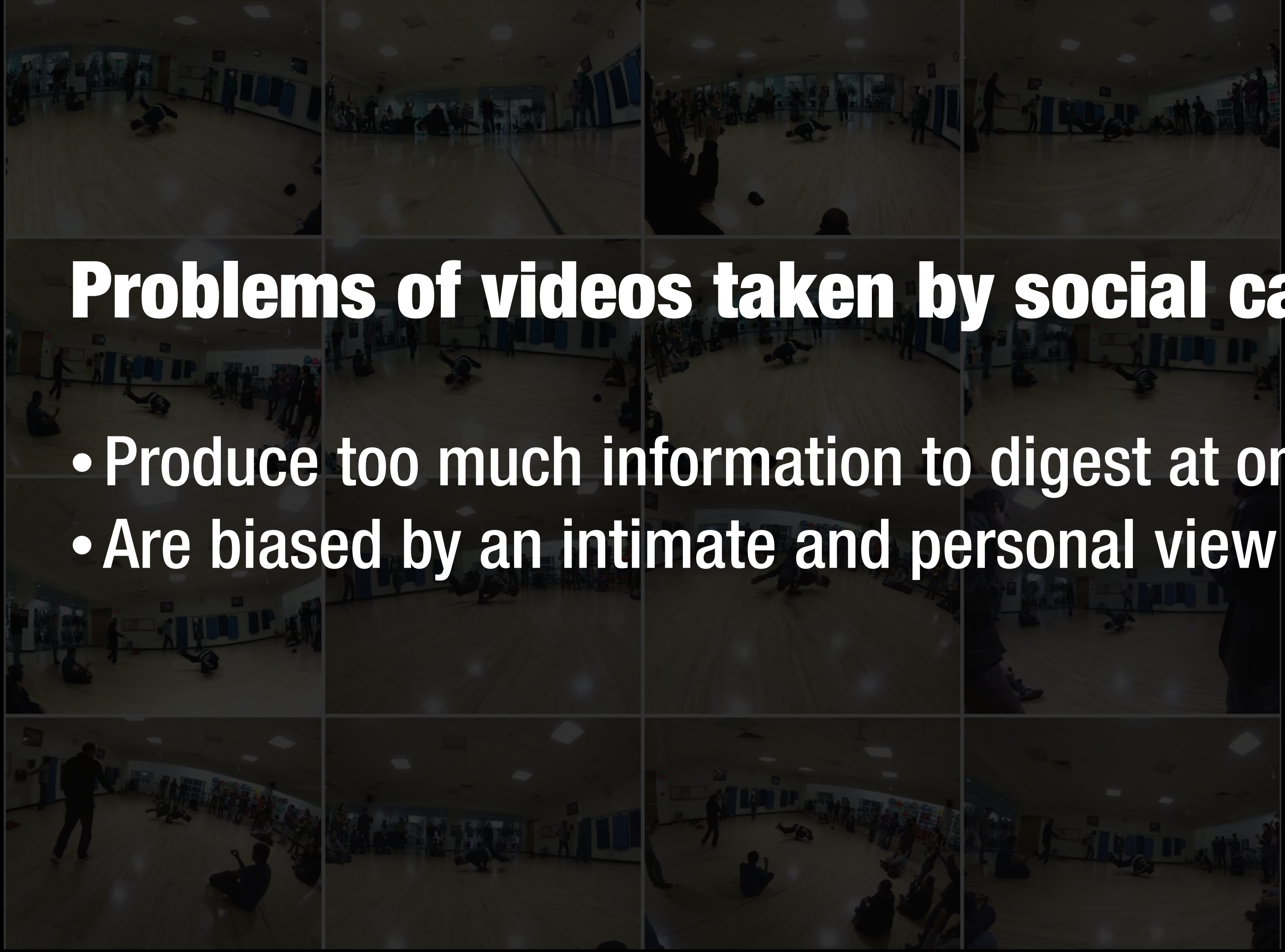
London 2012
FIBA
London 2012
London 2012
15
13
10
7
6

How would this joint attention be useful for social event?





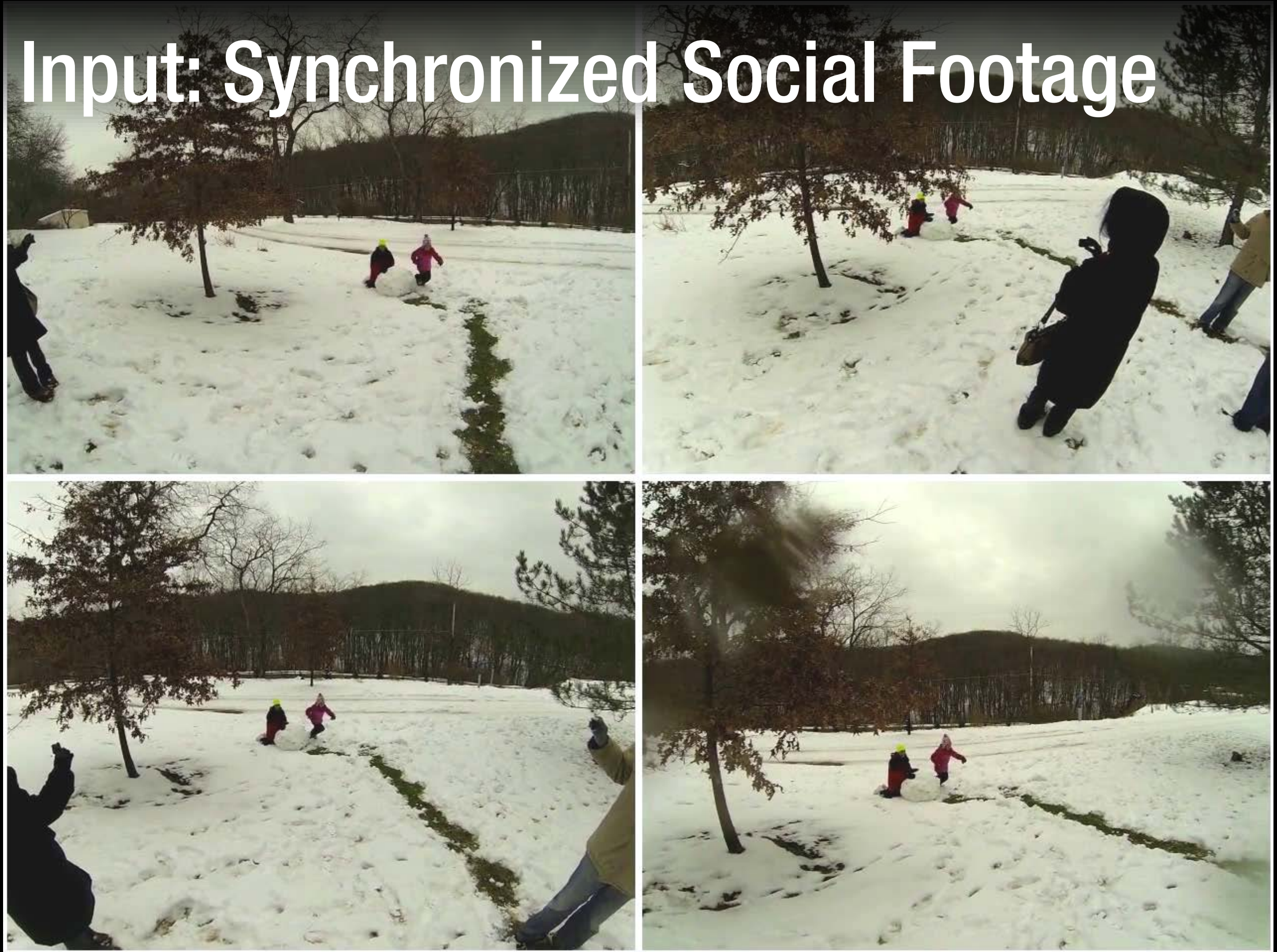




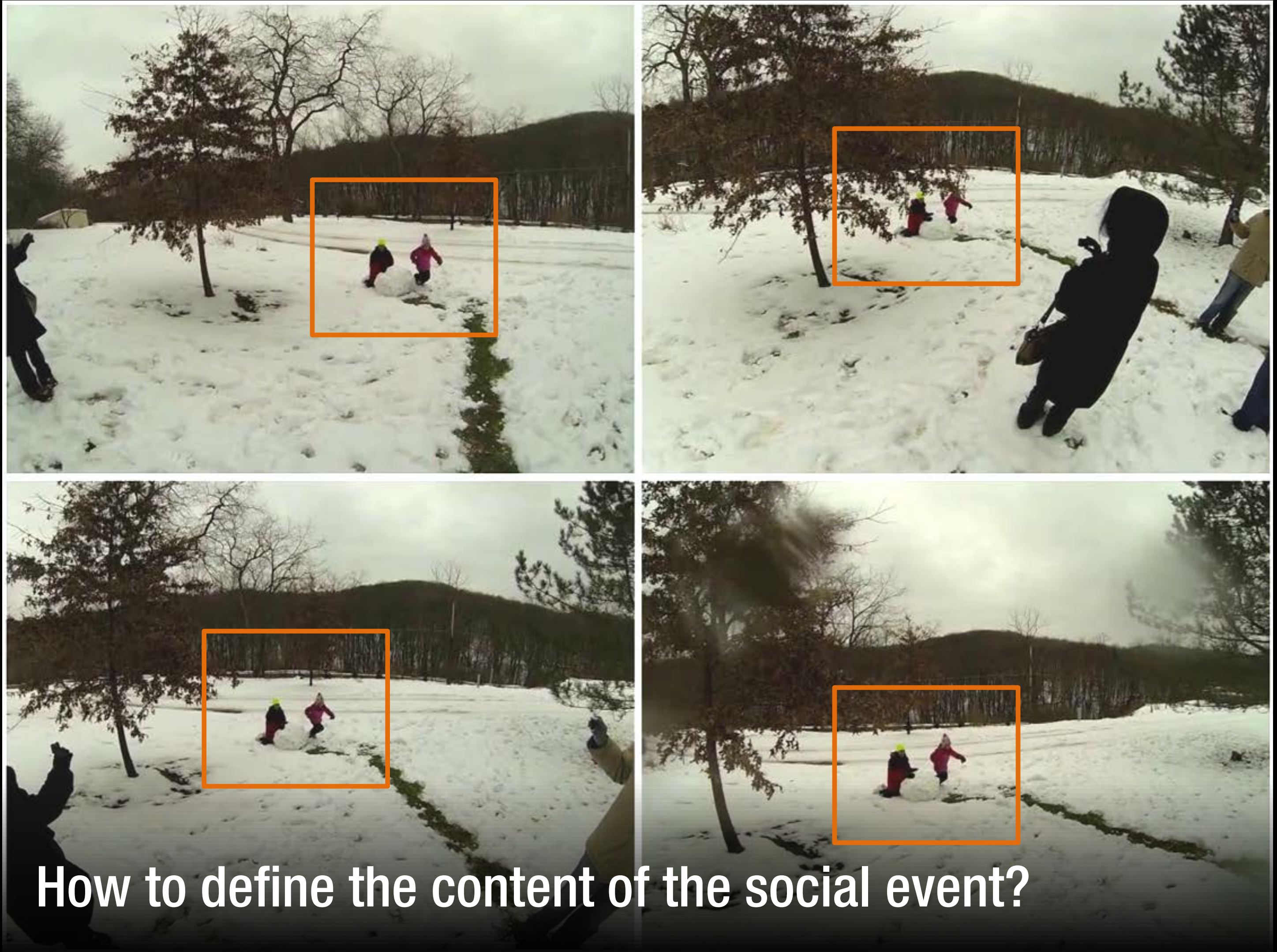
Problems of videos taken by social cameras:

- Produce too much information to digest at once
- Are biased by an intimate and personal view

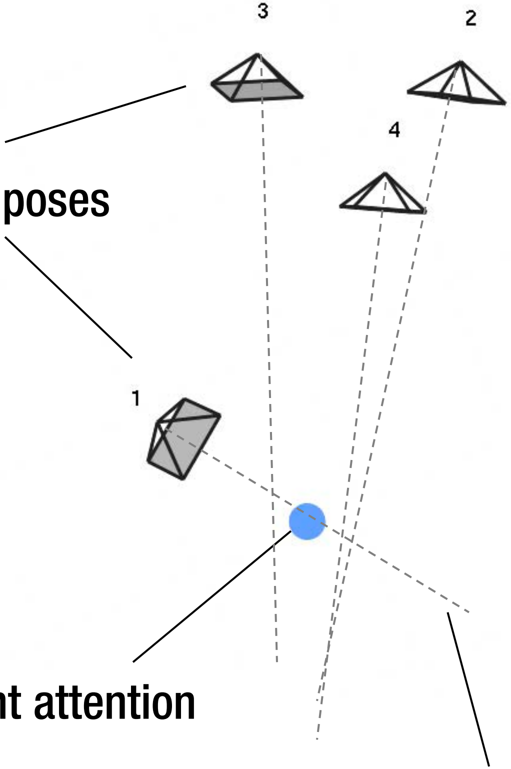
Input: Synchronized Social Footage



Output: Edited Video



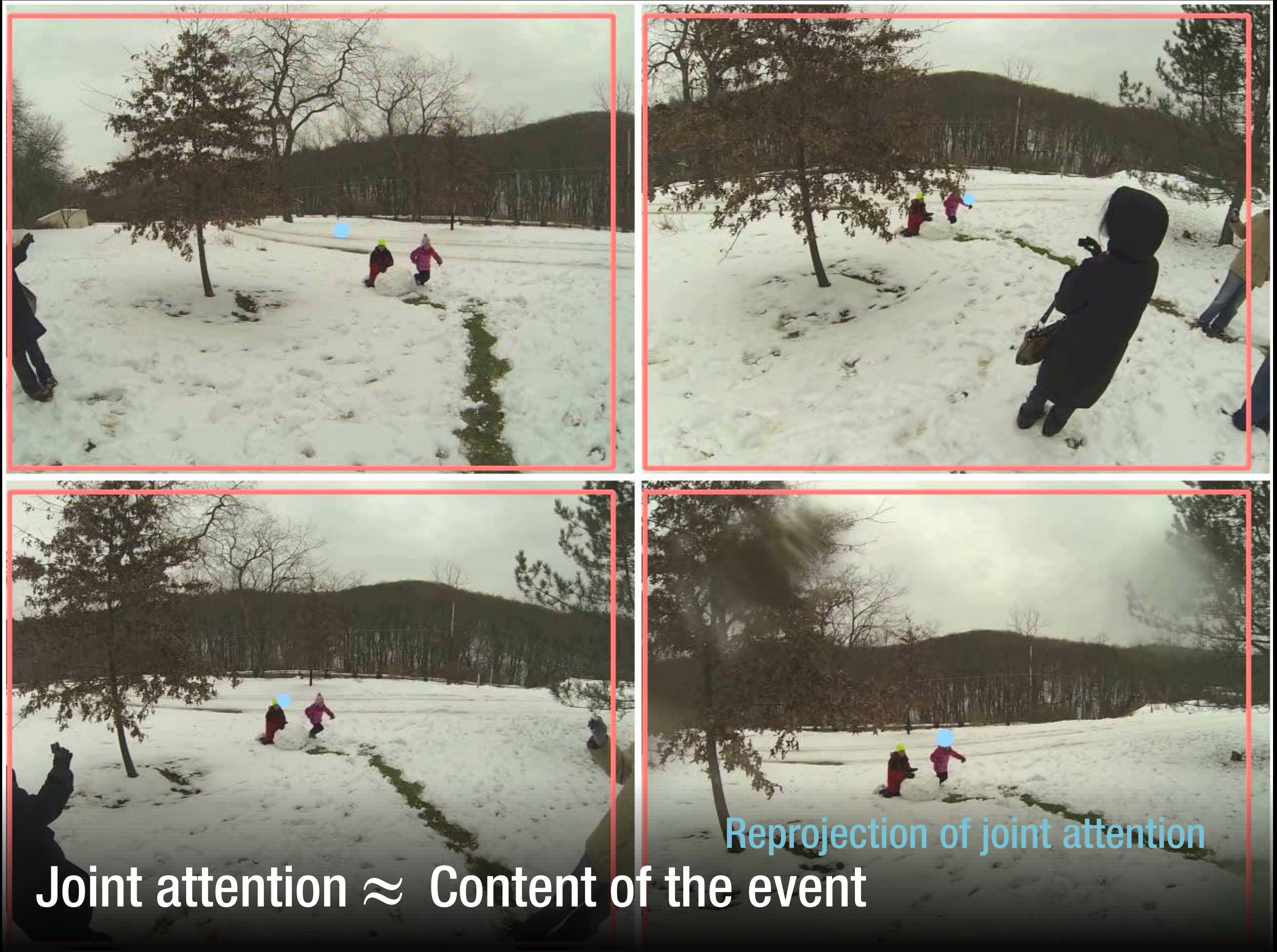
How to define the content of the social event?



3D camera poses

3D joint attention

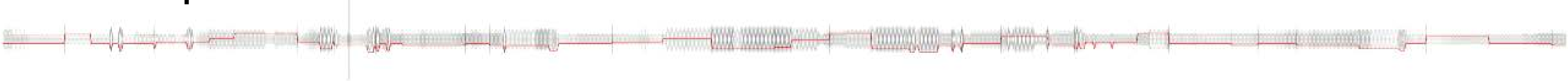
Camera optical axis



Reprojection of joint attention

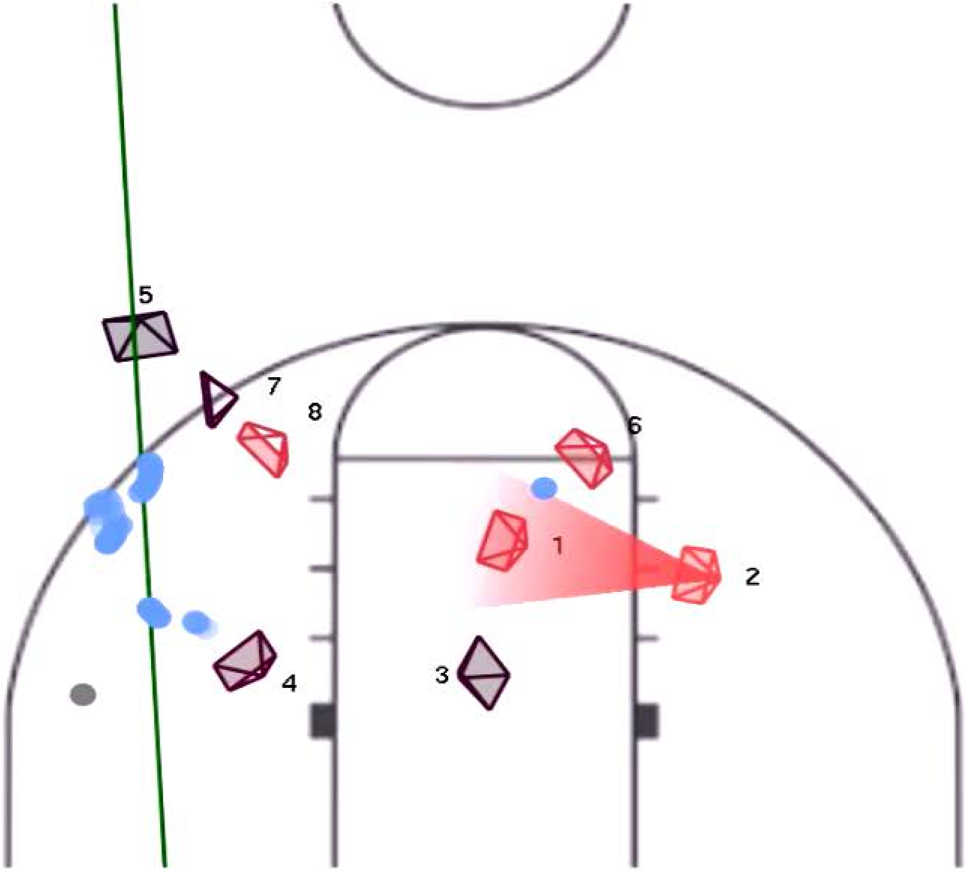
Joint attention ≈ Content of the event

Input video feeds



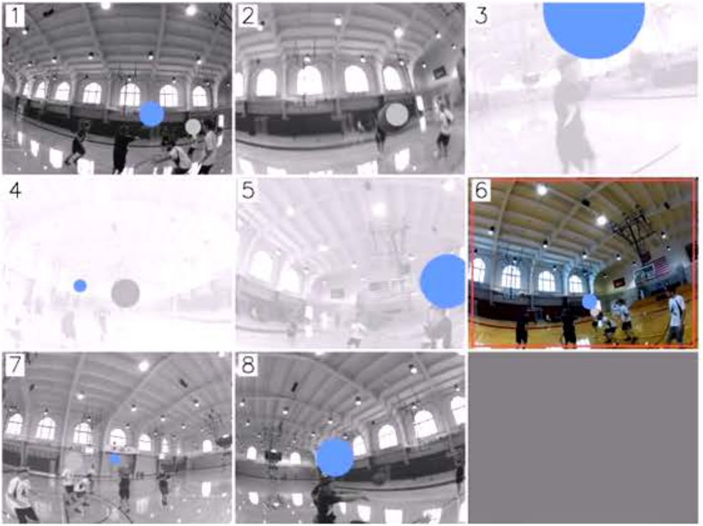
Timeline

Ours
GO
TARTANS

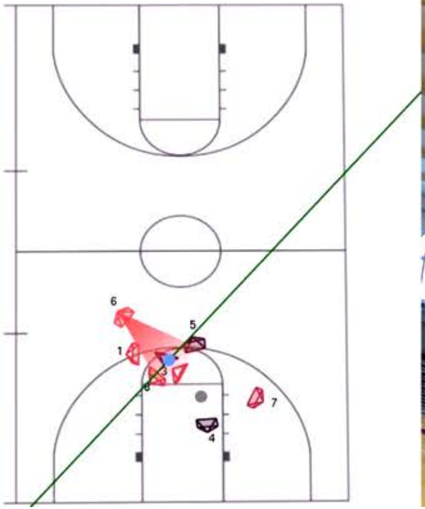


Selected camera: 2

Carnegie Mellon
Tartans
Carnegie Mellon
NCAA

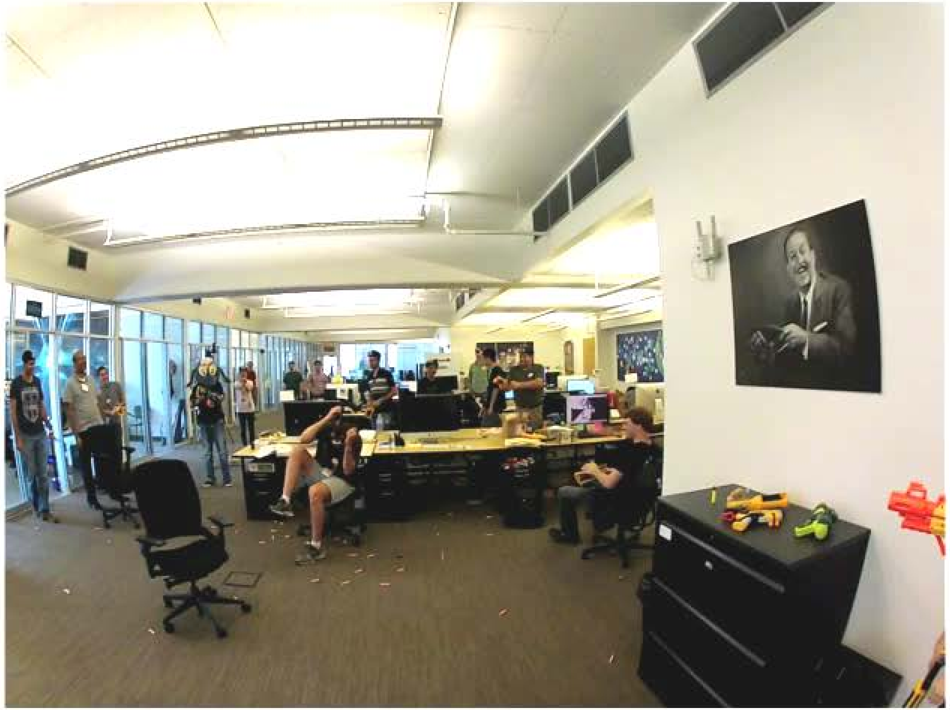


Video feeds

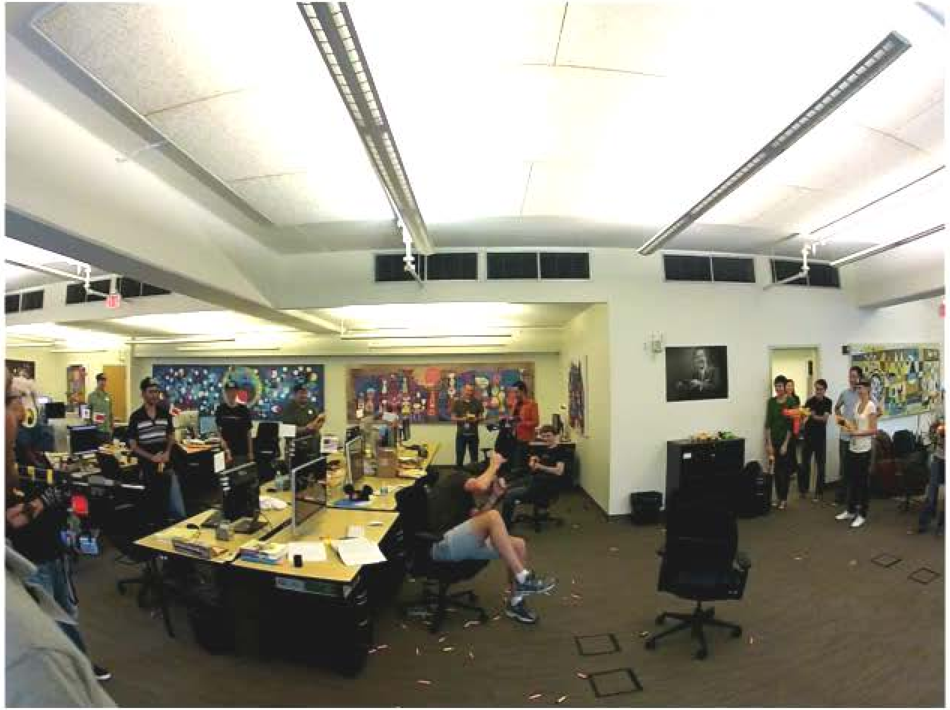


3D geometry

Output video

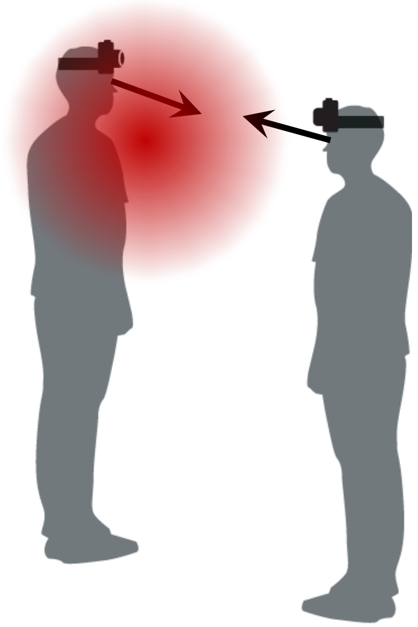


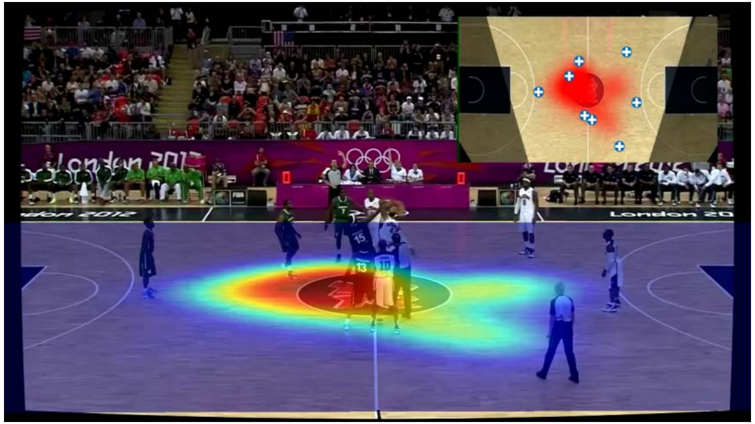
Our method

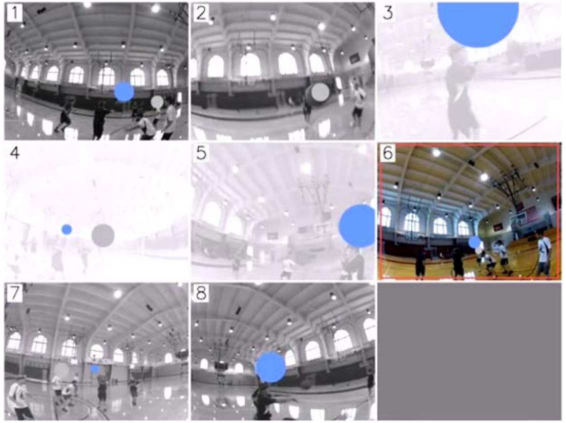


Professional Editor

First person cameras are ideal sensors to measure social behaviors.

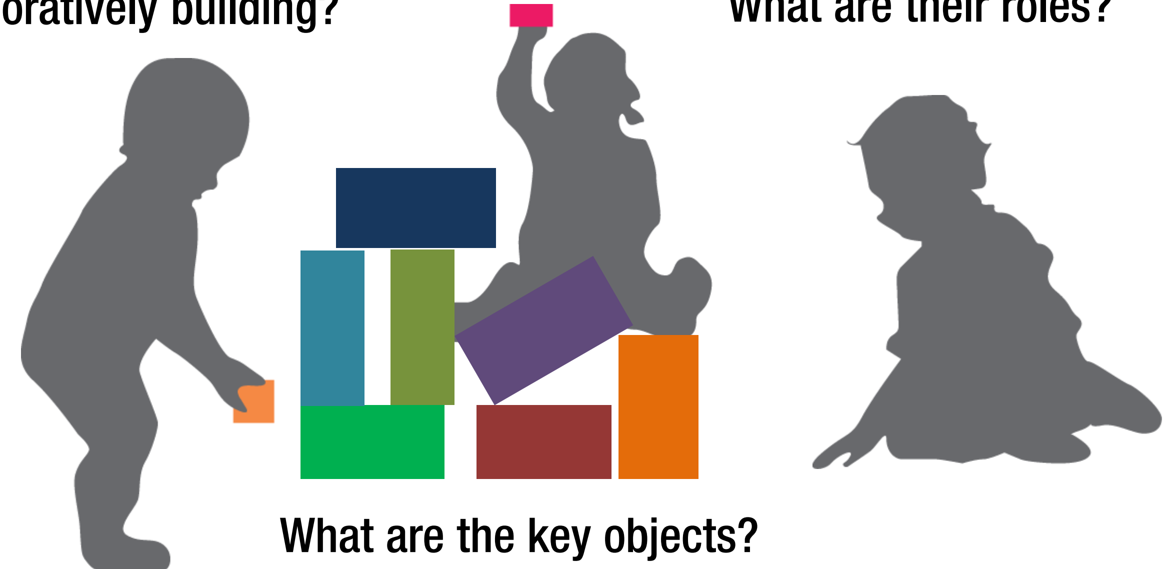






What are they collaboratively building?

What are their roles?



What are the key objects?

Social Attention

:What can first person cameras tell us about our social interactions?