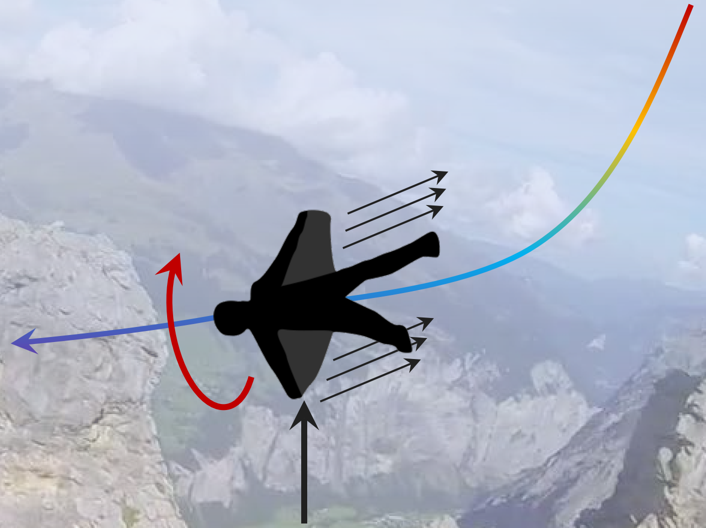
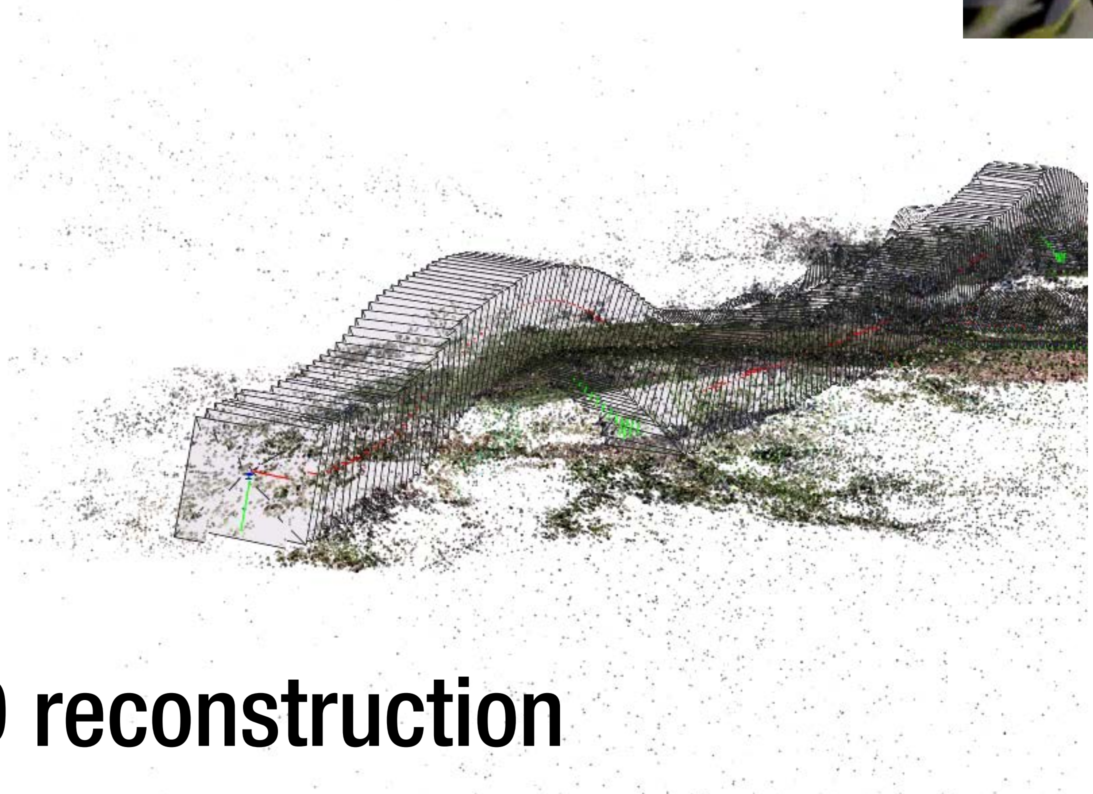
Visual Sensorimotor Behaviors II:

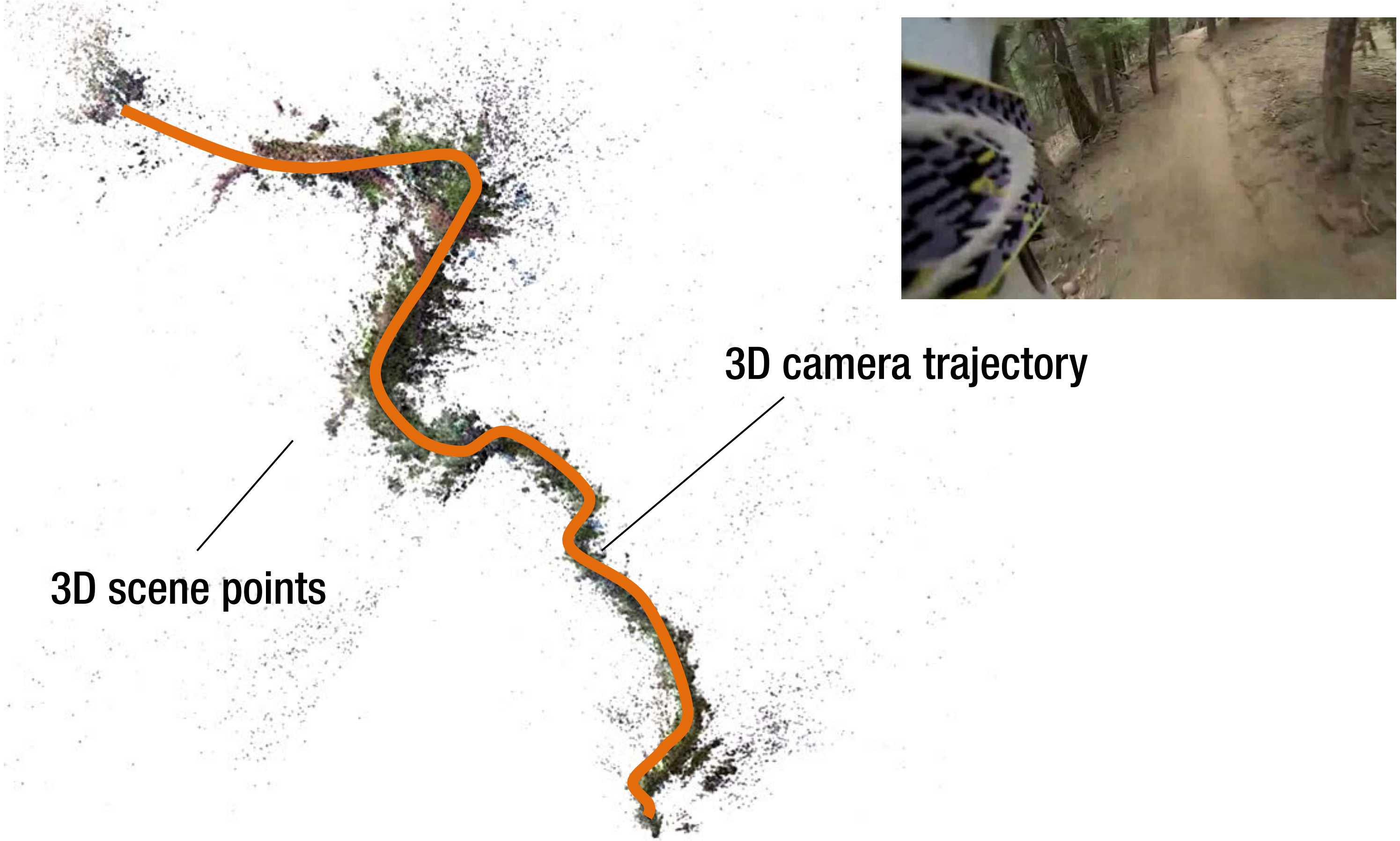
What can a first person video tell about how we control?




Input: first person video
https://www.youtube.com/watch?v=bKU0Z4bj7Mw



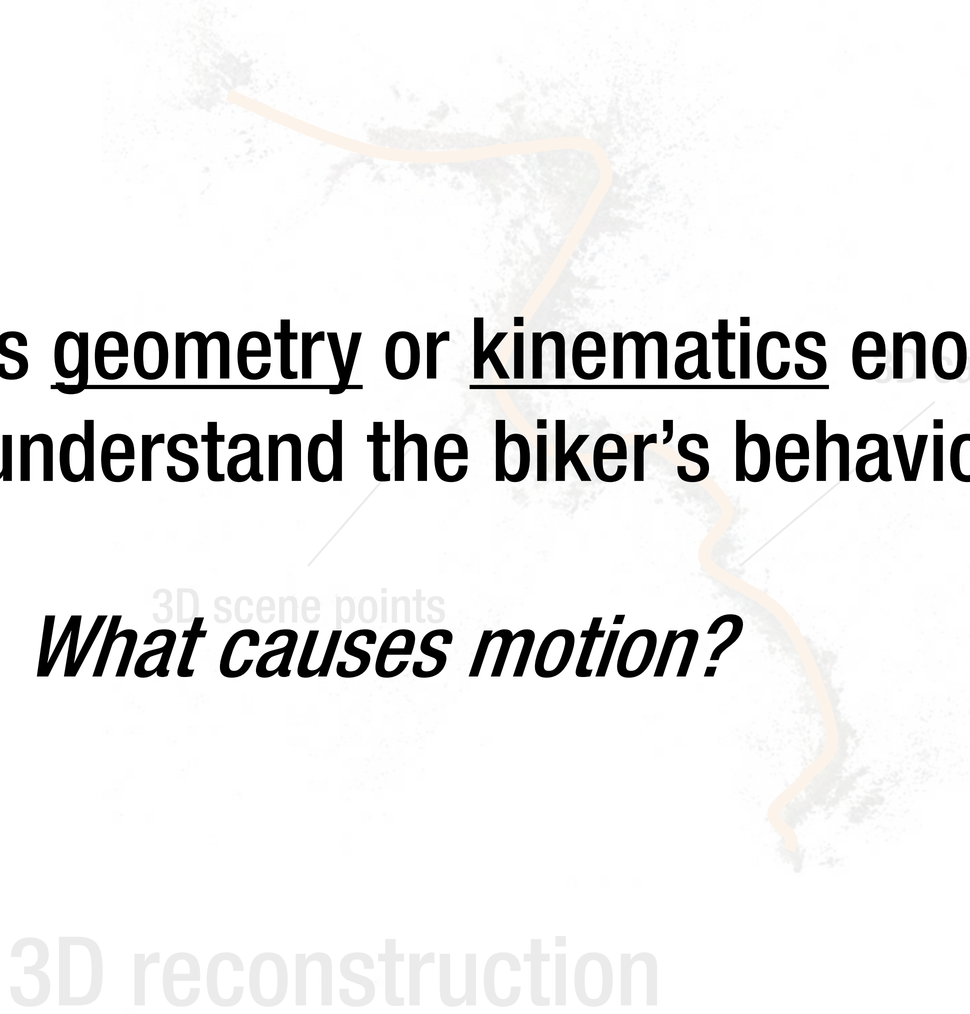
3D reconstruction



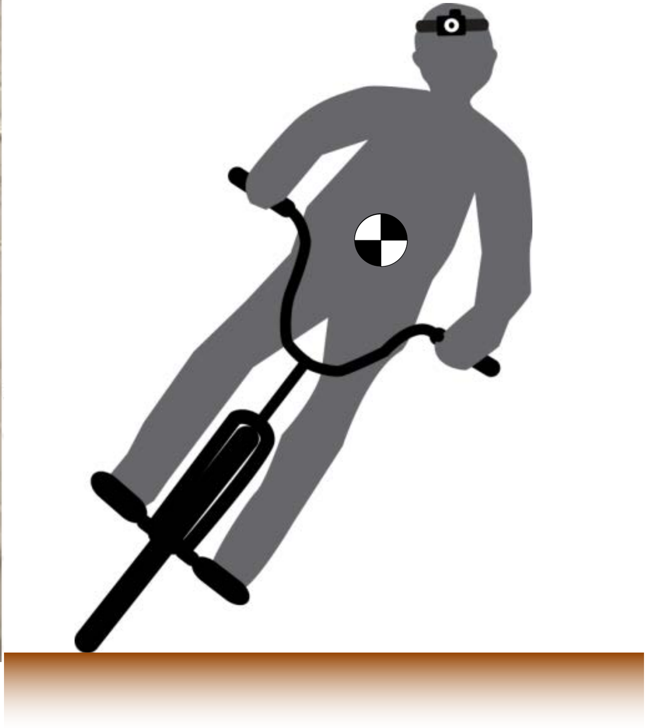
3D reconstruction

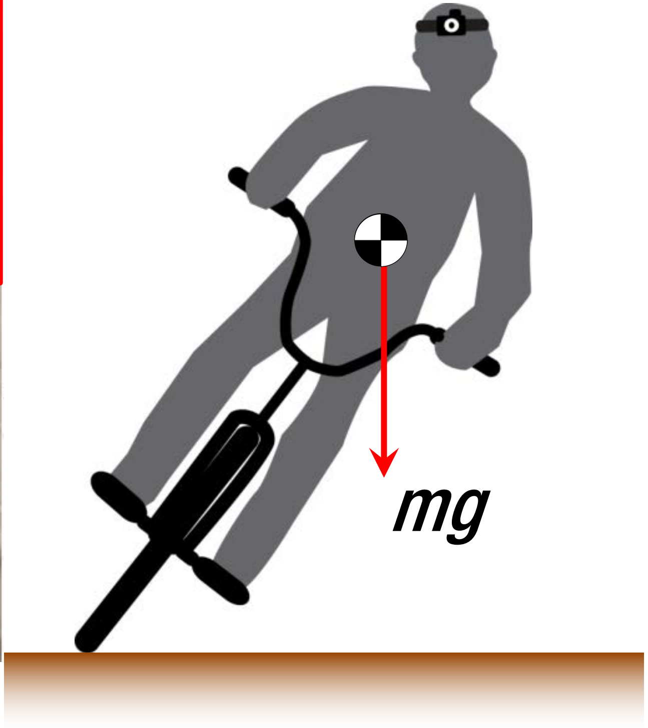
Is <u>geometry</u> or <u>kinematics</u> enough to understand the biker's behaviors?

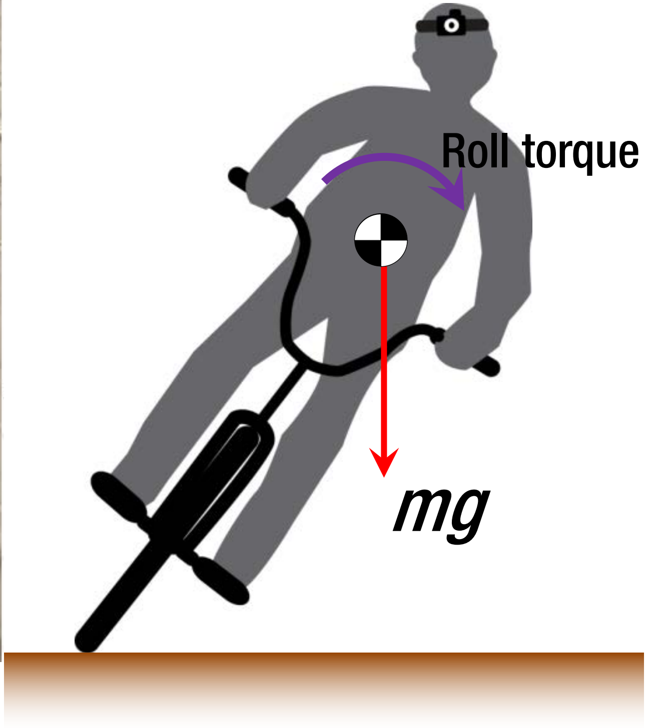
What causes motion?



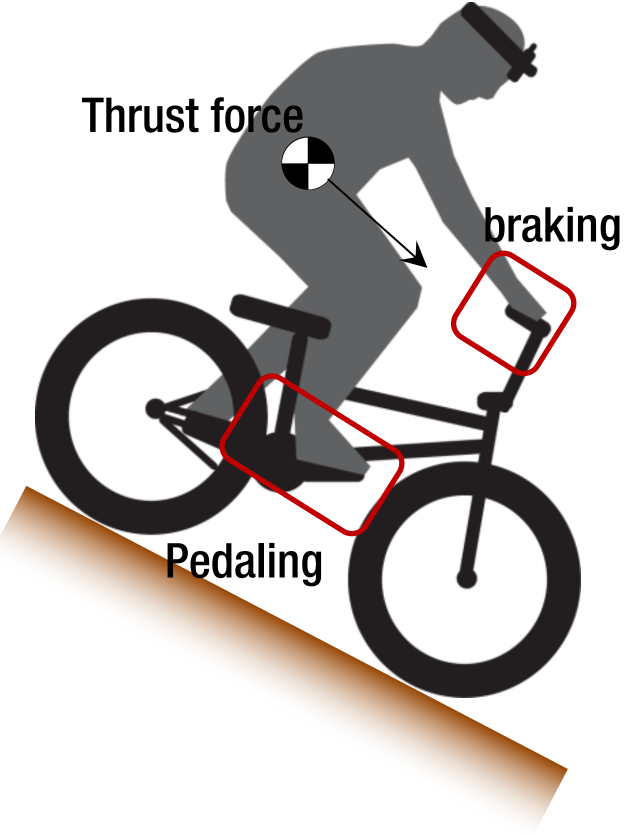

Time: 0.17sec
Speed: 2.0m/s
Air Drag: 1.77N
Gravity
Th: -177N
Ro: 91Nm
Pi: -93Nm
Lt: 4N
No: -842N
Yw: 13Nm




Time: 0.17sec
Speed: 2.0m/s
Air Drag: 1.77N
Gravity
Th: -177N
Ro: 91Nm
Pi: -93Nm
Lt: 4N
No: -842N
Yw: 13Nm

mg

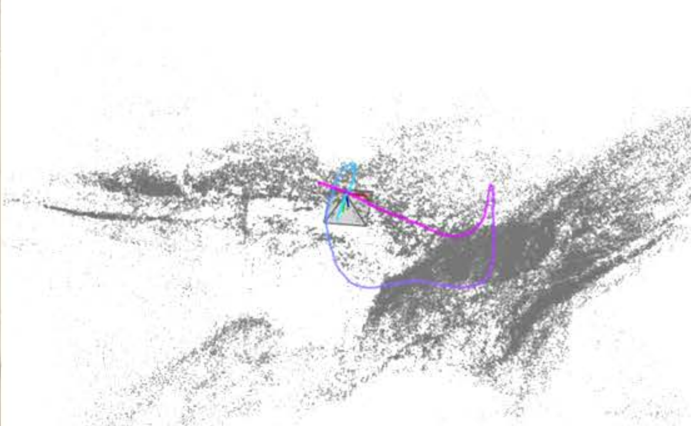

Time: 0.17sec
Speed: 2.0m/s
Air Drag: 1.77N
Gravity
Th: -177N
Ro: 91Nm
Pi: -93Nm
Lt: 4N
No: -842N
Yw: 13Nm

Roll torque
mg


Time: 0.17sec
Speed: 2.0m/s
Air Drag: 1.77N
Gravity
Th: -177N
Ro: 91Nm
Pi: -93Nm
Lt: 4N
No: -842N
Yw: 13Nm

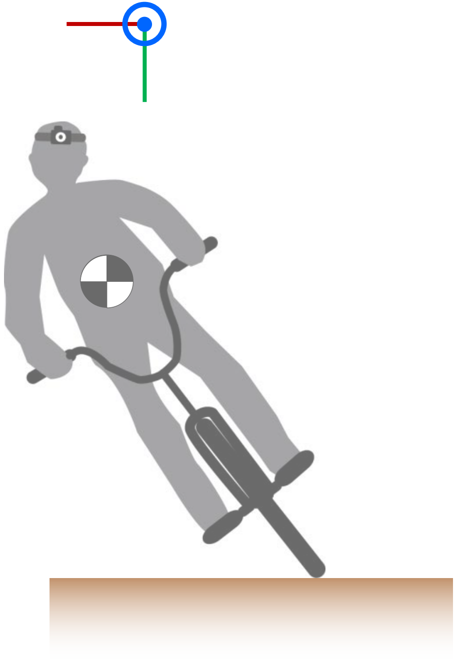

Thrust force
braking
Pedaling

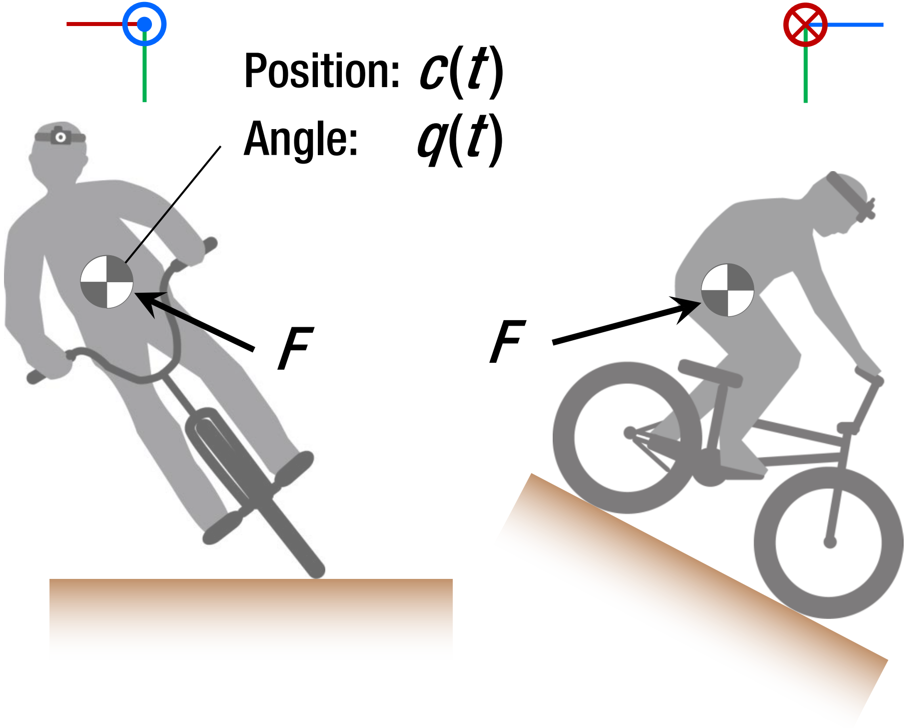

Time: 0.17sec
Speed: 2.0m/s
Air Drag: 1.77N
Gravity
Th: -177N
Ro: 91Nm
Lt: 4N
No: -842N
Pi: -93Nm
Yw: 13Nm


Air drag
Thrust force
braking
Pedaling



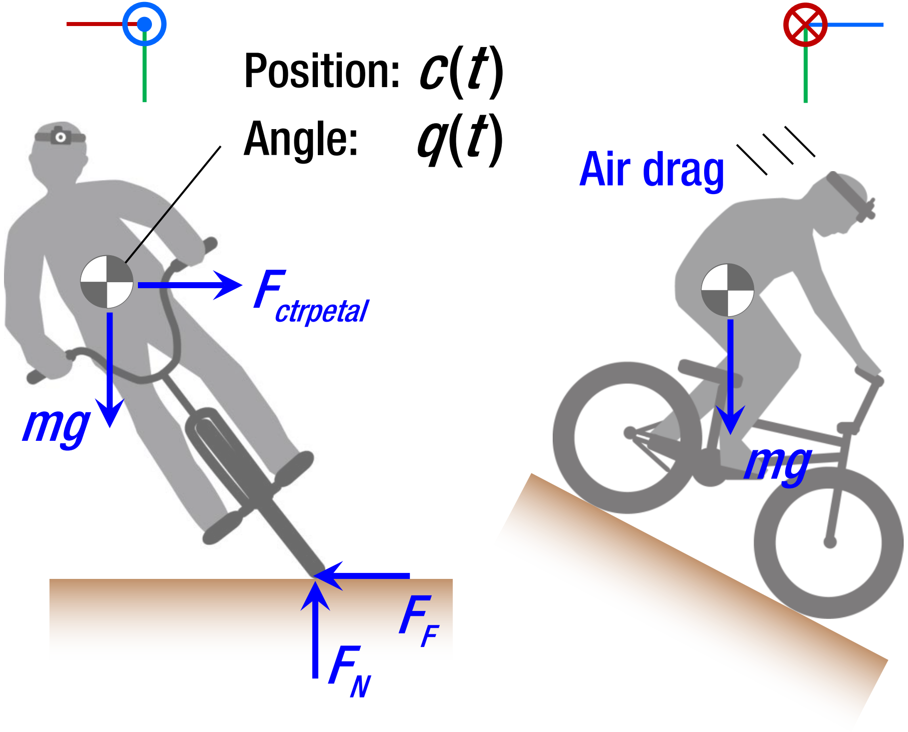
3D reconstruction





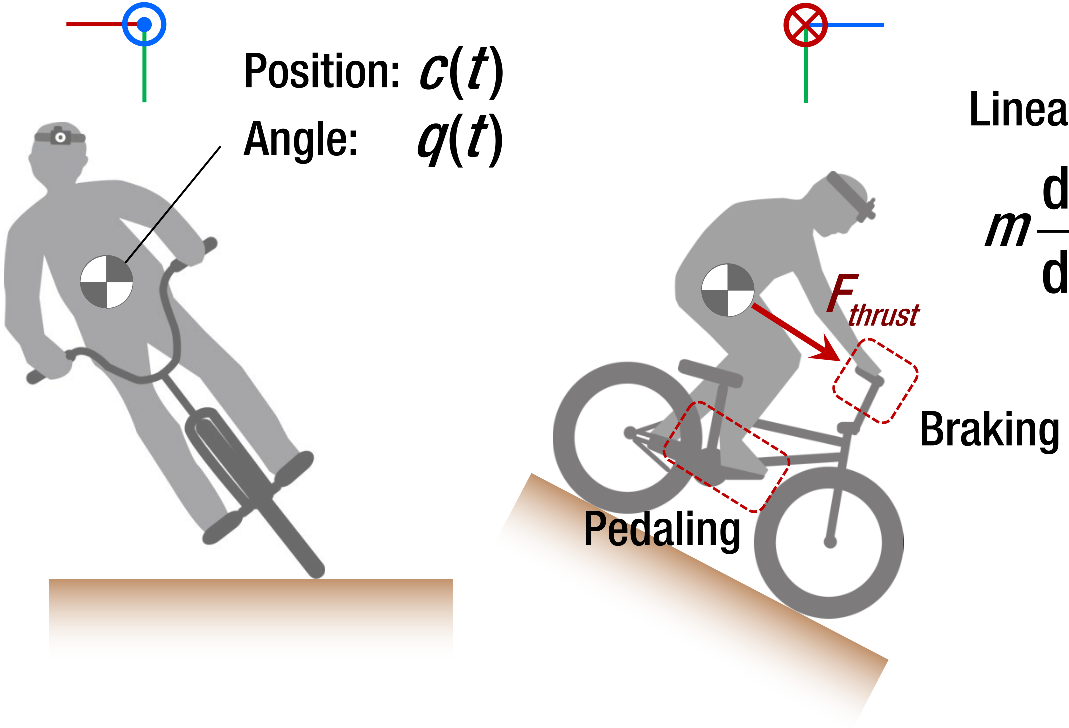
Linear force:

$$m\frac{\mathrm{d}^2 c}{\mathrm{d}t^2} = \underline{F}$$



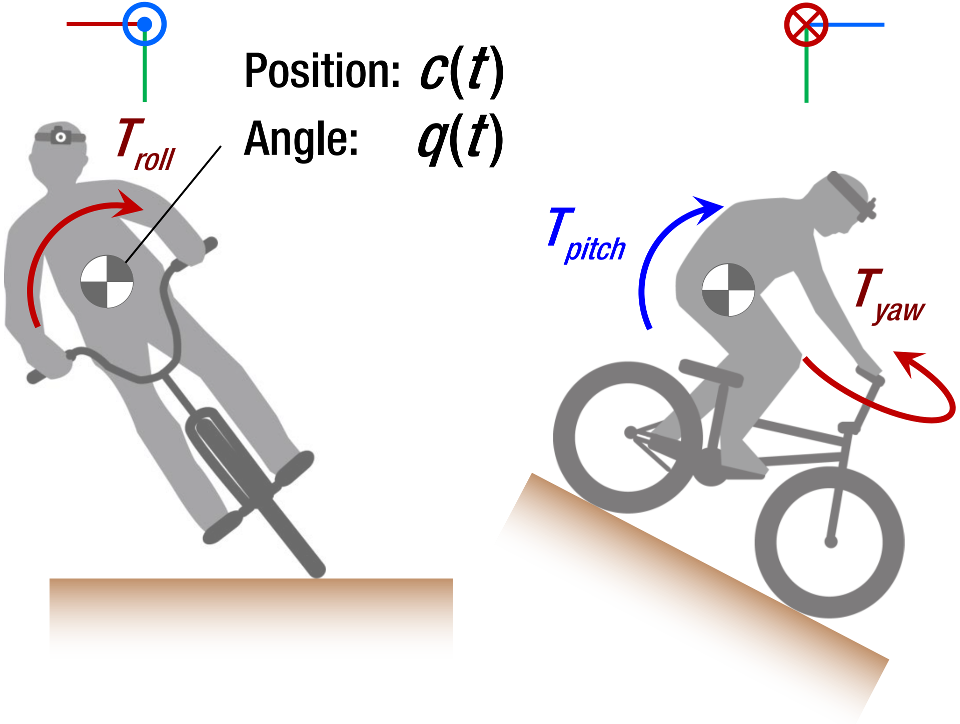
Linear force:

$$m\frac{d^2c}{dt^2} = F = \underline{\sum F_{\text{passive}}} + \sum F_{\text{active}}$$



Linear force:

$$m\frac{\mathrm{d}^2 c}{\mathrm{d}t^2} = F = \sum F_{\text{passive}} + \underline{\sum F_{\text{active}}}$$



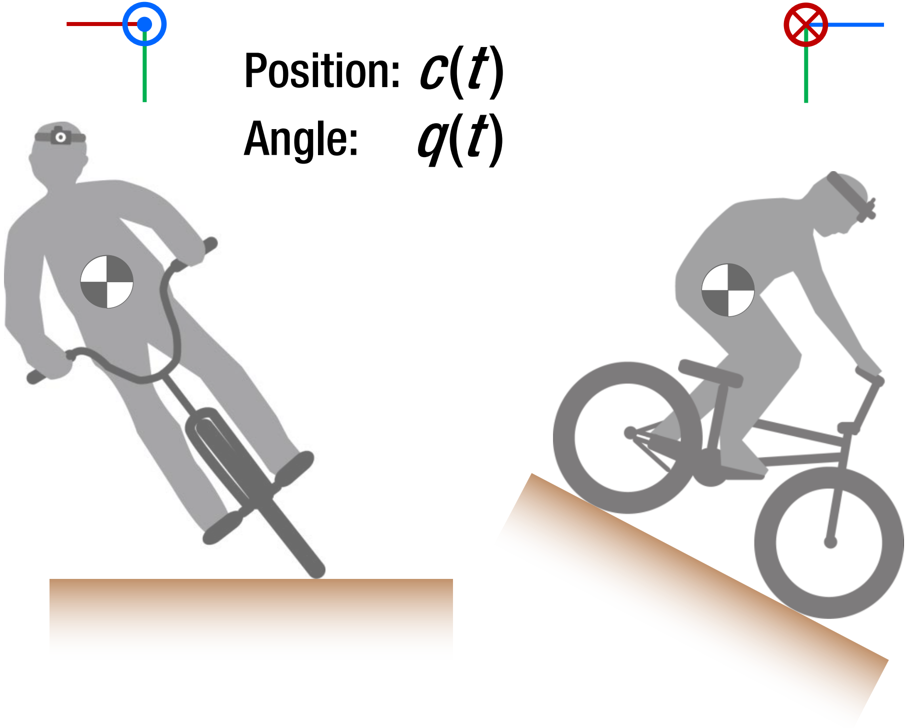
Linear force:

$$m\frac{\mathrm{d}^2 c}{\mathrm{d}t^2} = F = \sum F_{\mathrm{passive}} + \sum F_{\mathrm{active}}$$

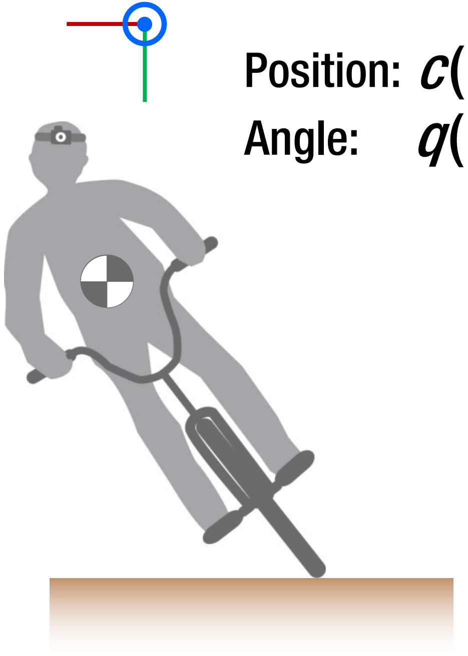
Angular force (torque):

$$J\frac{\mathrm{d}^2 q}{\mathrm{d}t^2} + \frac{\mathrm{d}q}{\mathrm{d}t} \times J\frac{\mathrm{d}q}{\mathrm{d}t} = \sum T_{\mathrm{passive}} + \sum T_{\mathrm{active}}$$

J : Moment of inertia



$$(c, q) = \mathrm{ODE}(F_{\text{passive}}, F_{\text{active}}, T_{\text{passive}}, T_{\text{active}})$$



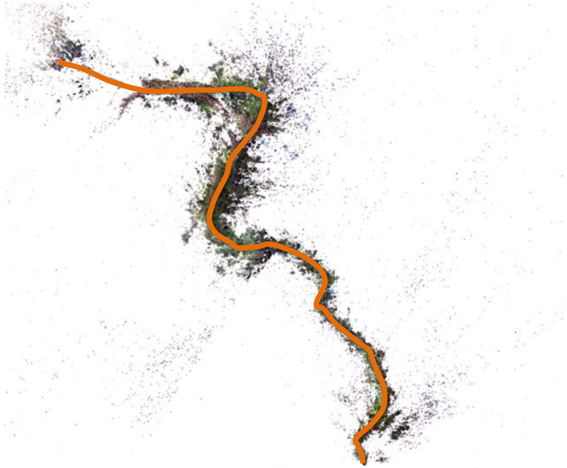
Position: $c(t)$

Angle: $q(t)$

$$(c,q) = \mathrm{ODE}(F_{\mathrm{passive}}, F_{\mathrm{active}}, T_{\mathrm{passive}}, T_{\mathrm{active}})$$

$$\approx (\mathrm{t},\mathrm{R})$$

where $\mathrm{P} = \mathrm{K}\begin{bmatrix}\mathrm{R} & \mathrm{t}\end{bmatrix}$

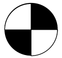


Inverse control:

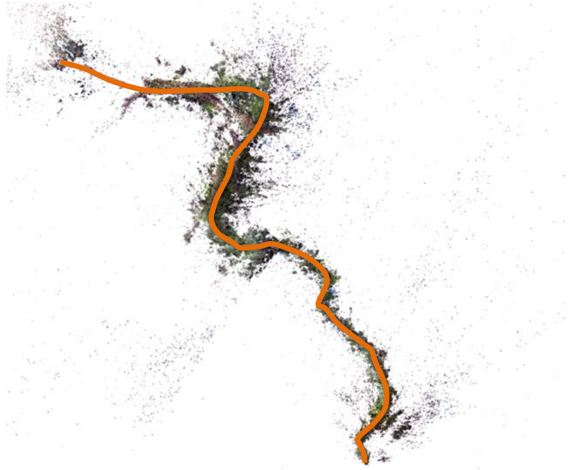
$$\underset{F,T,X}{\text{minimize}}\ \underbrace{E_{\text{SfM}}}_{\text{Reprojection error}}$$

subject to $(\mathrm{t},\mathrm{R}) = \mathrm{ODE}(F,T)$

$$(c,q) = \mathrm{ODE}(F_{\text{passive}}, F_{\text{active}}, T_{\text{passive}}, T_{\text{active}})$$

 $\approx (\mathrm{t},\mathrm{R})$ 

where $\mathrm{P} = \mathrm{K}\begin{bmatrix}\mathrm{R} & \mathrm{t}\end{bmatrix}$

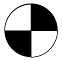


Inverse control:

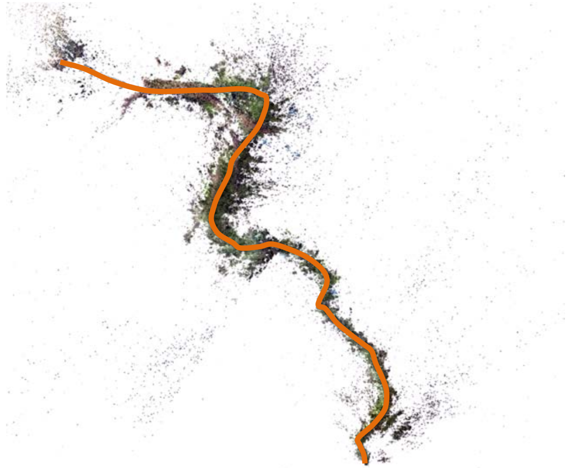
$$\underset{F,T,X}{\text{minimize}}\; E_{\text{SfM}} + \lambda \underbrace{E_{\text{reg}}(F,T)}_{\text{Temporal regularization}}$$

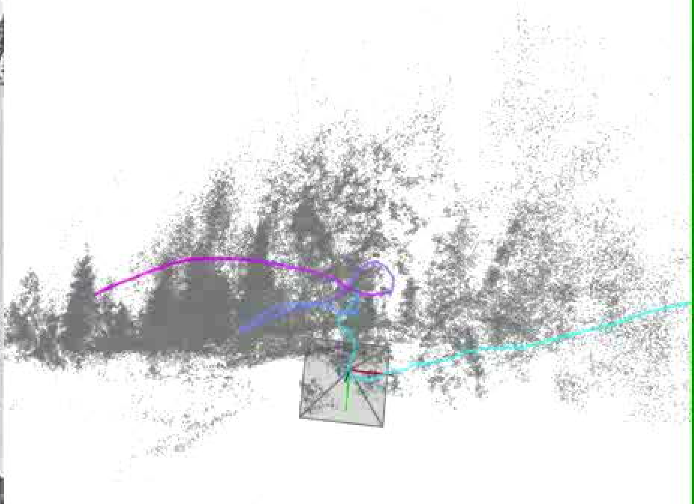
subject to $(\mathrm{t},\mathrm{R}) = \mathrm{ODE}(F,T)$

$$(c,q) = \mathrm{ODE}(F_{\text{passive}}, F_{\text{active}}, T_{\text{passive}}, T_{\text{active}})$$

 $\approx (\mathrm{t},\mathrm{R})$ 

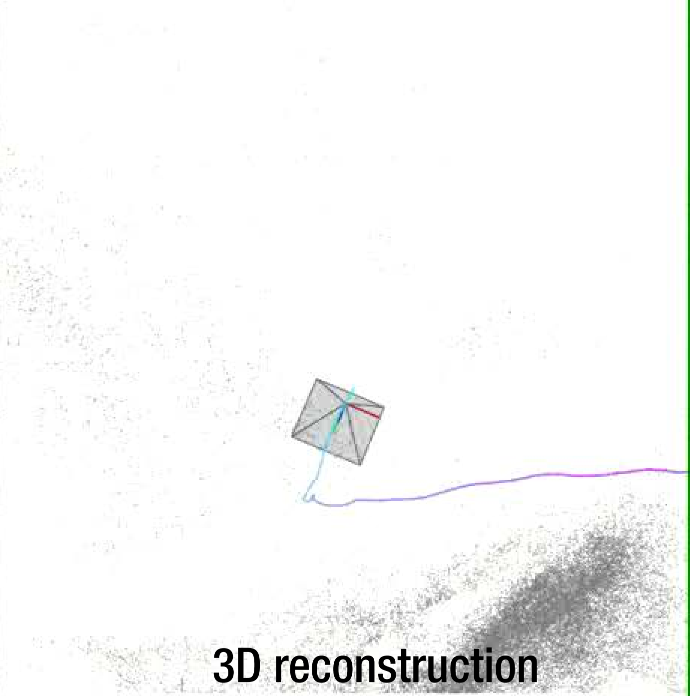
where $\mathrm{P} = \mathrm{K}\begin{bmatrix}\mathrm{R} & \mathrm{t}\end{bmatrix}$





3D reconstruction

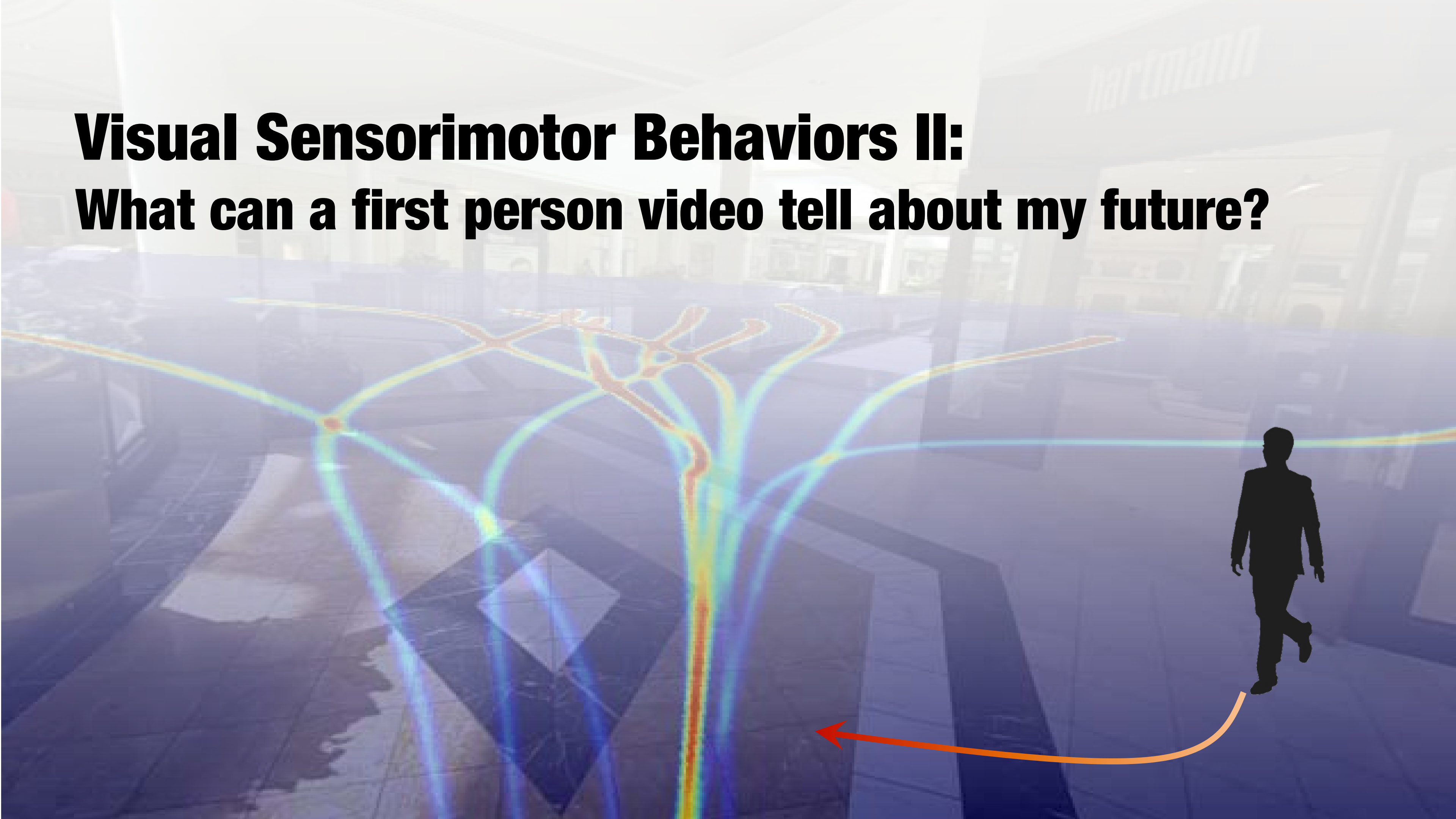
https://www.youtube.com/watch?v=pCcuKClUpLs

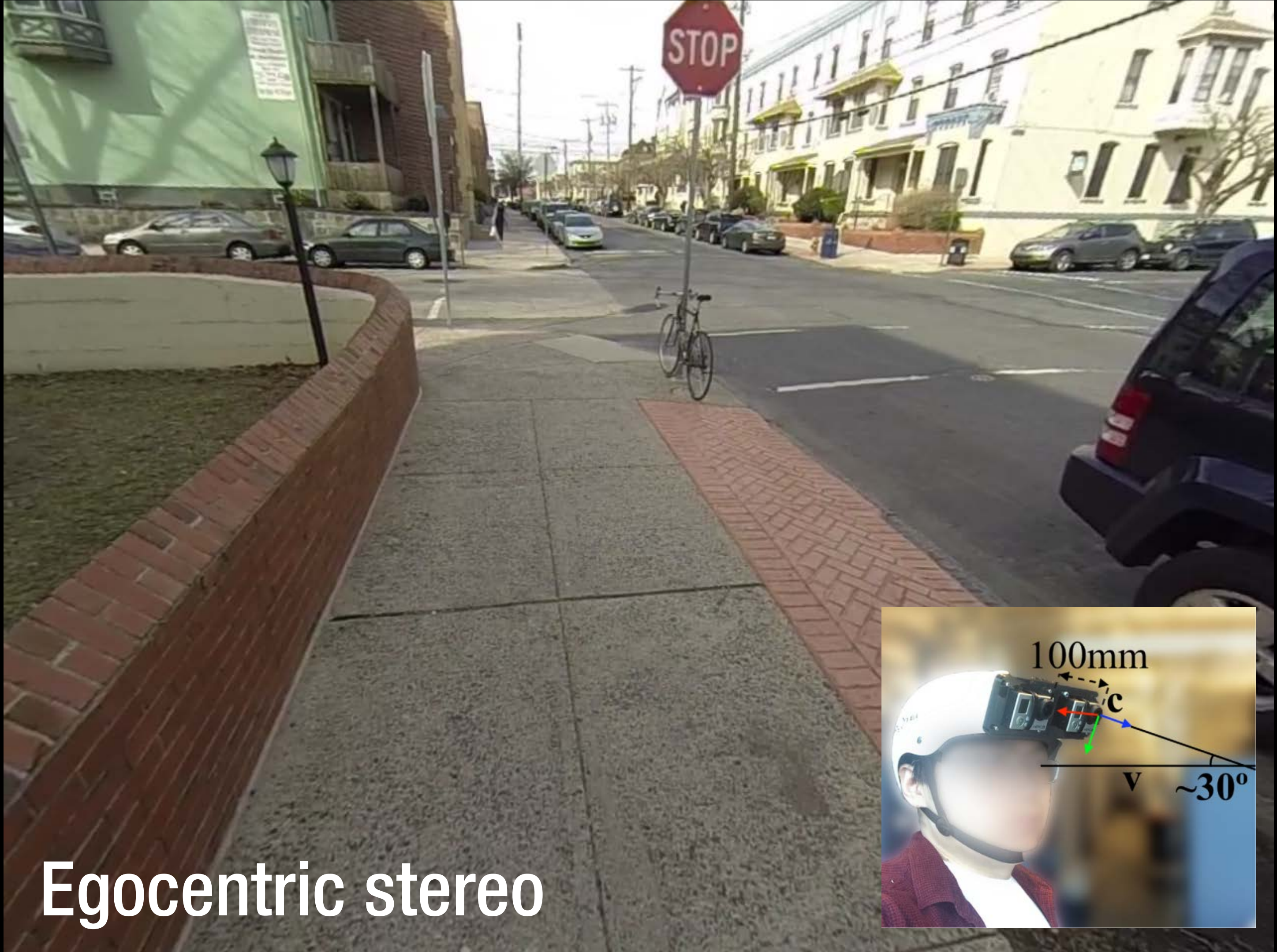


https://www.youtube.com/watch?v=rnvvsjstveM

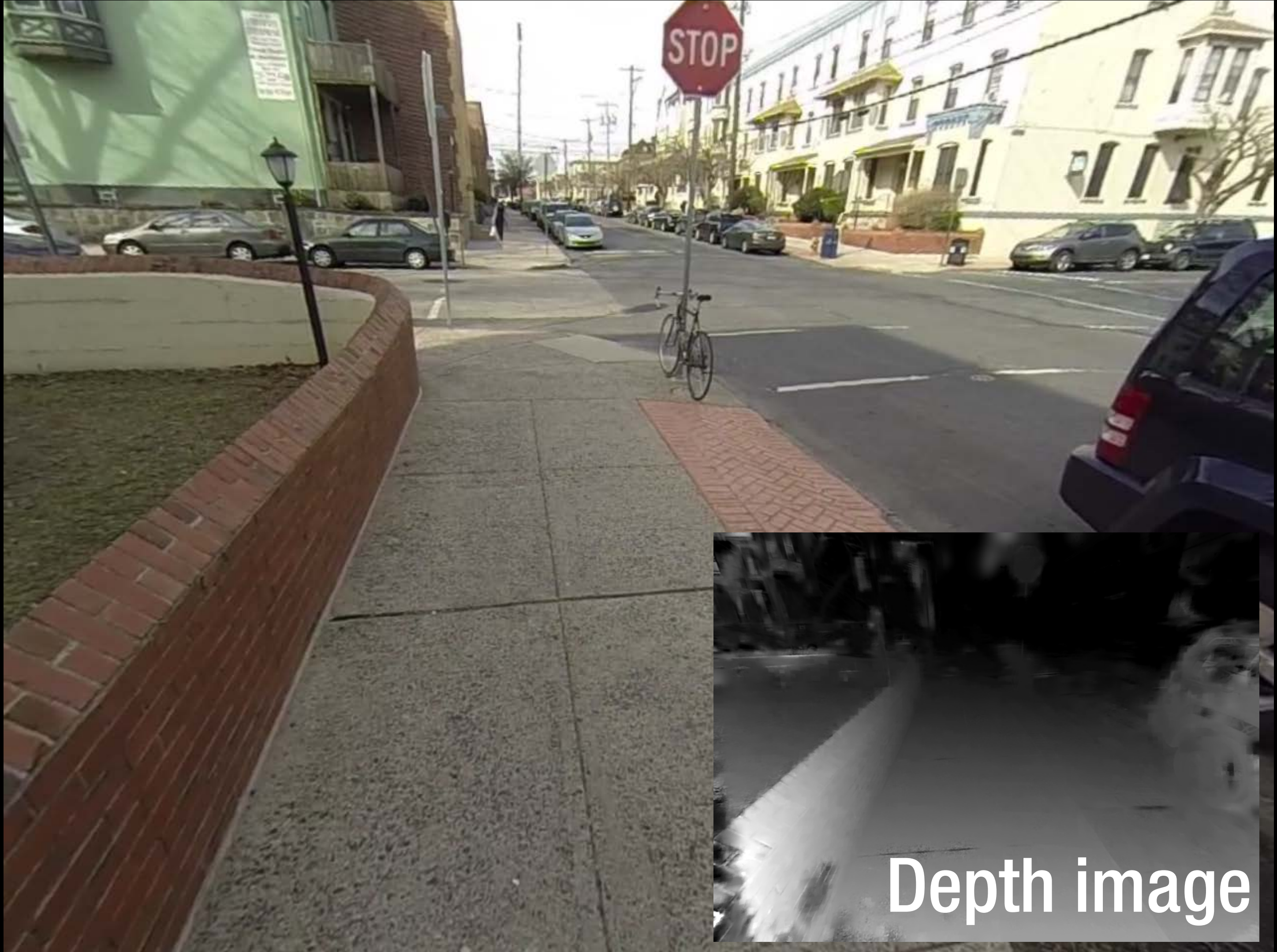
Visual Sensorimotor Behaviors II:

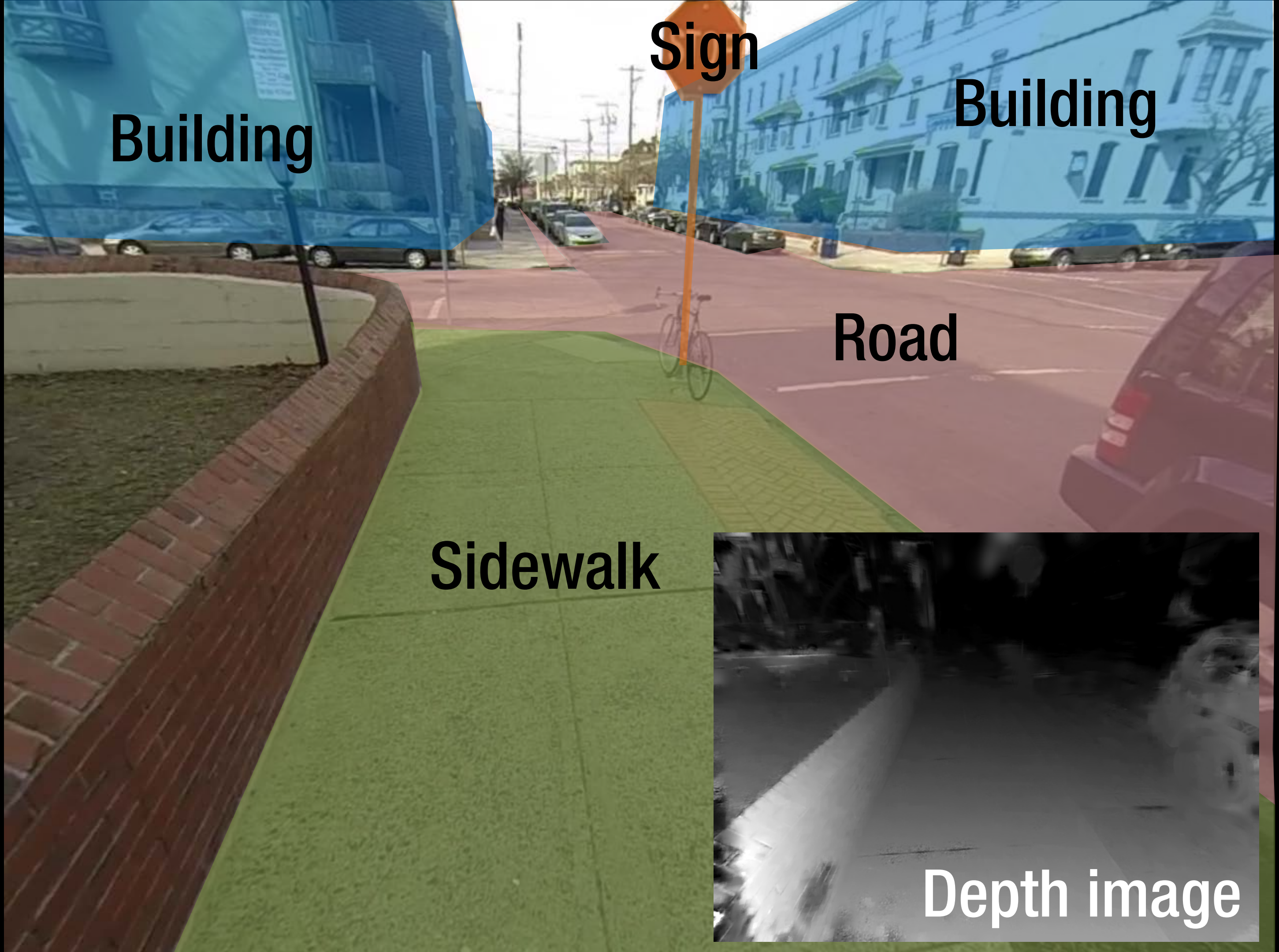
What can a first person video tell about my future?

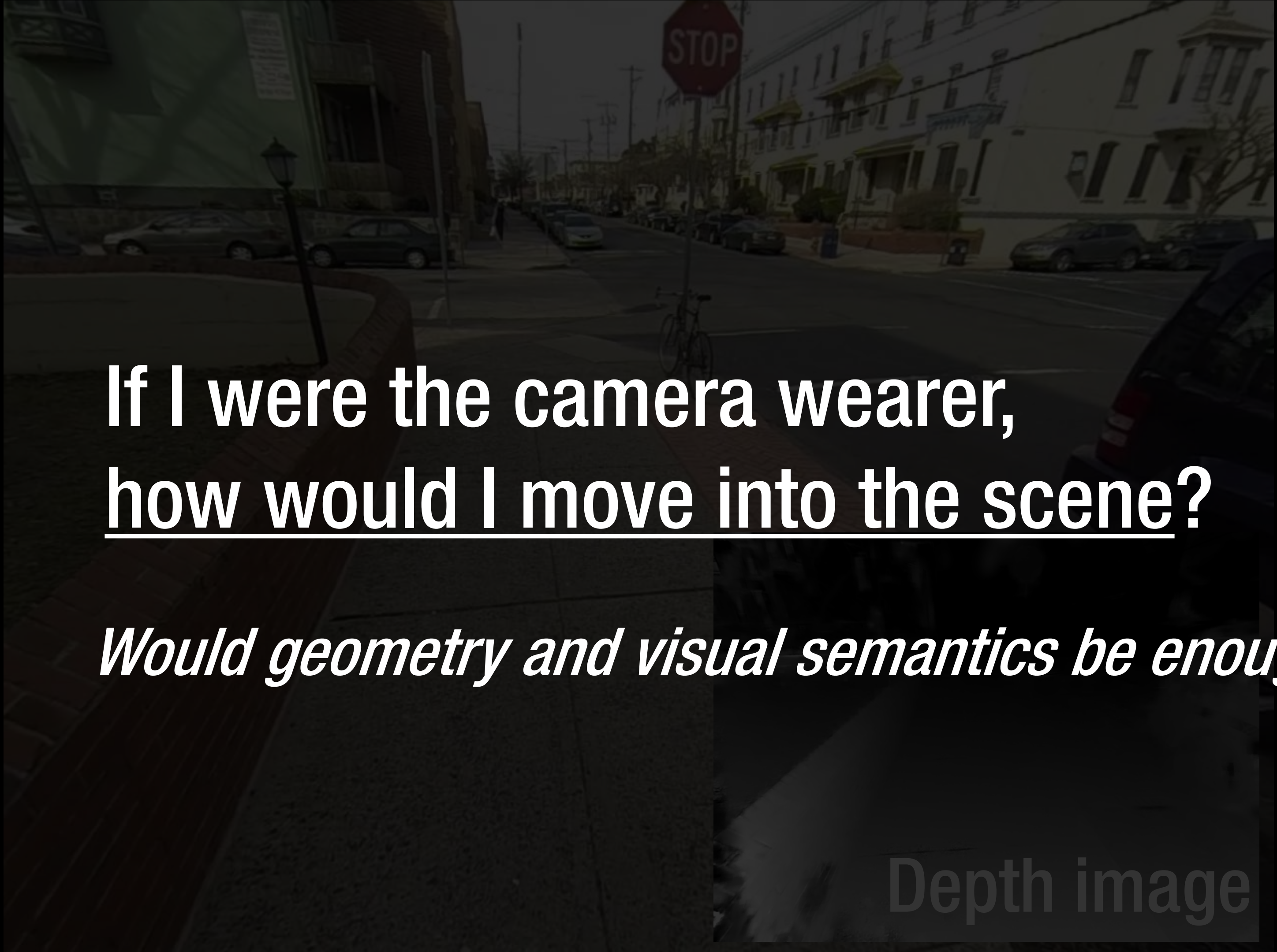




Egocentric stereo

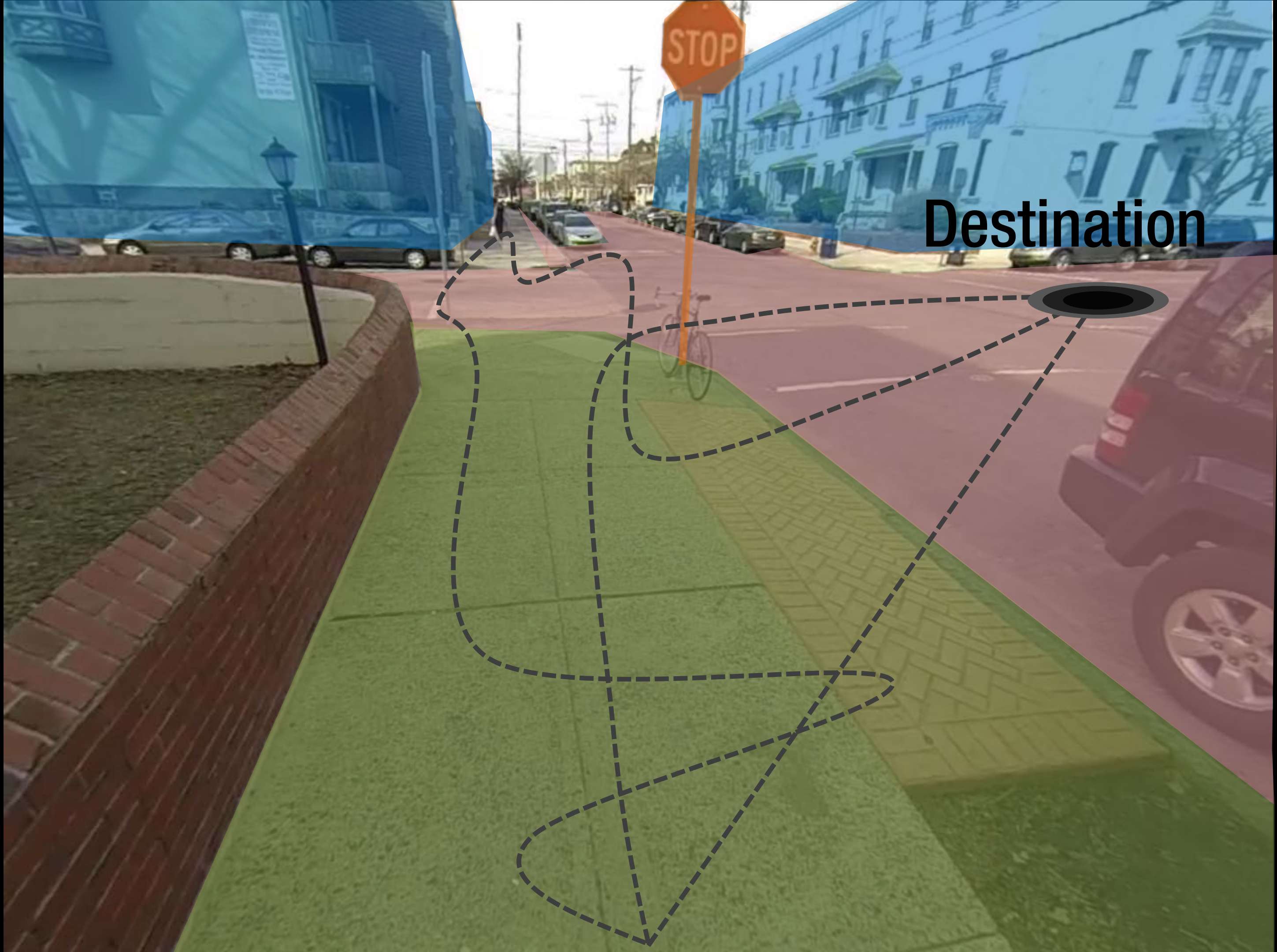

STOP
Depth image

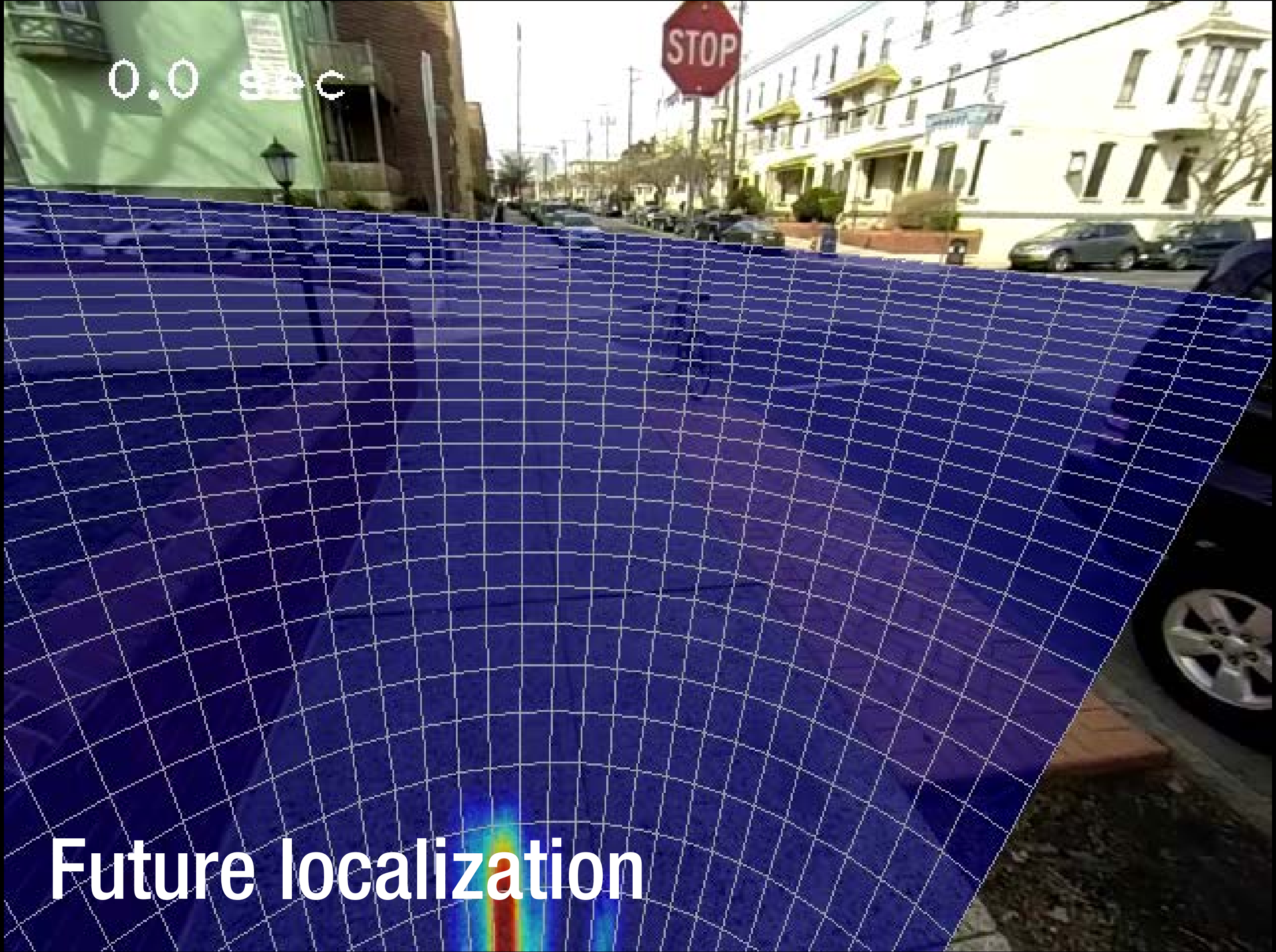

Sign
Building
Building
Road
Sidewalk
Depth image

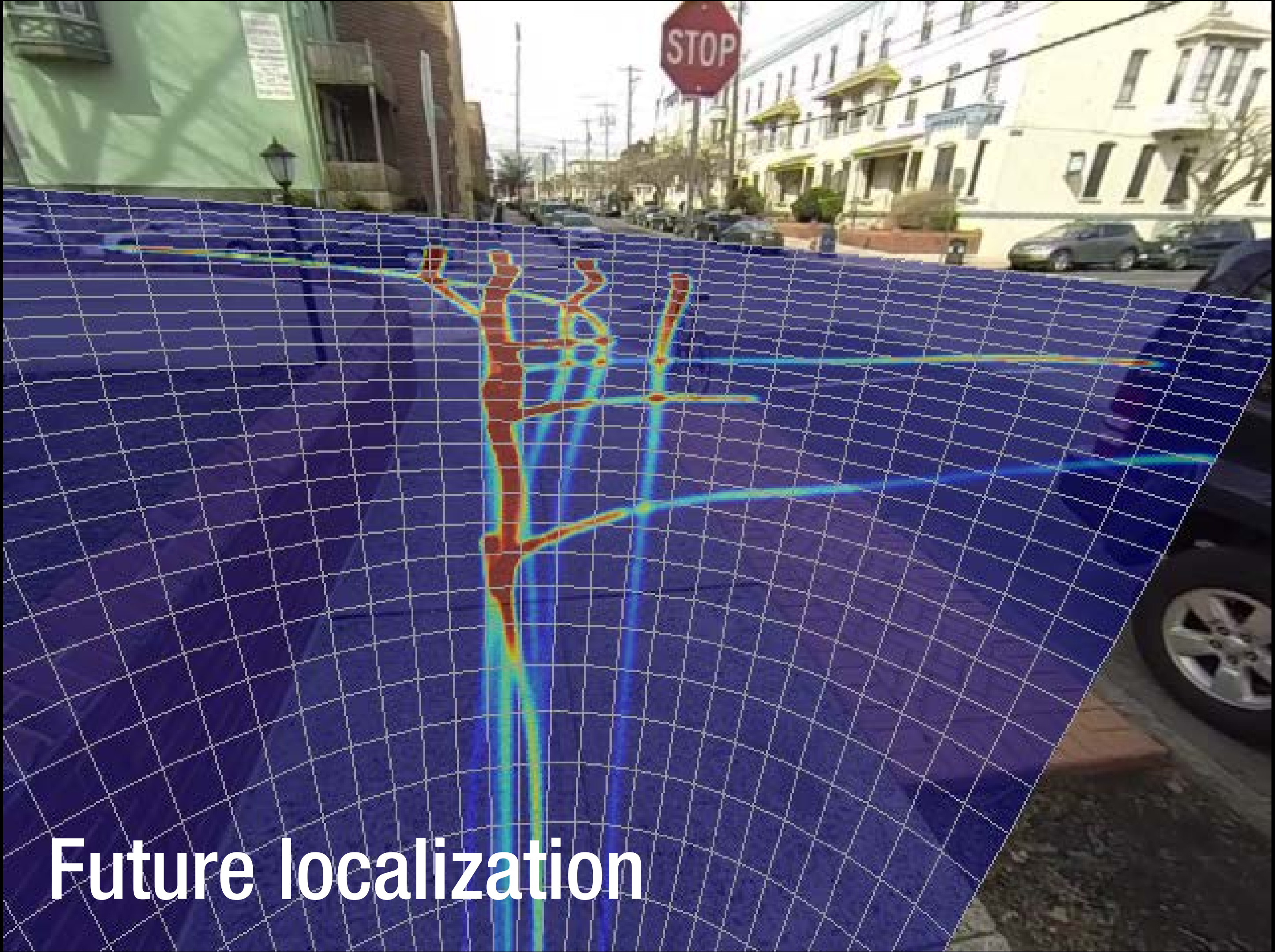


If I were the camera wearer,
how would I move into the scene?

Would geometry and visual semantics be enough?


STOP
Destination


0.0 sec
STOP
Future localization


STOP
Future localization

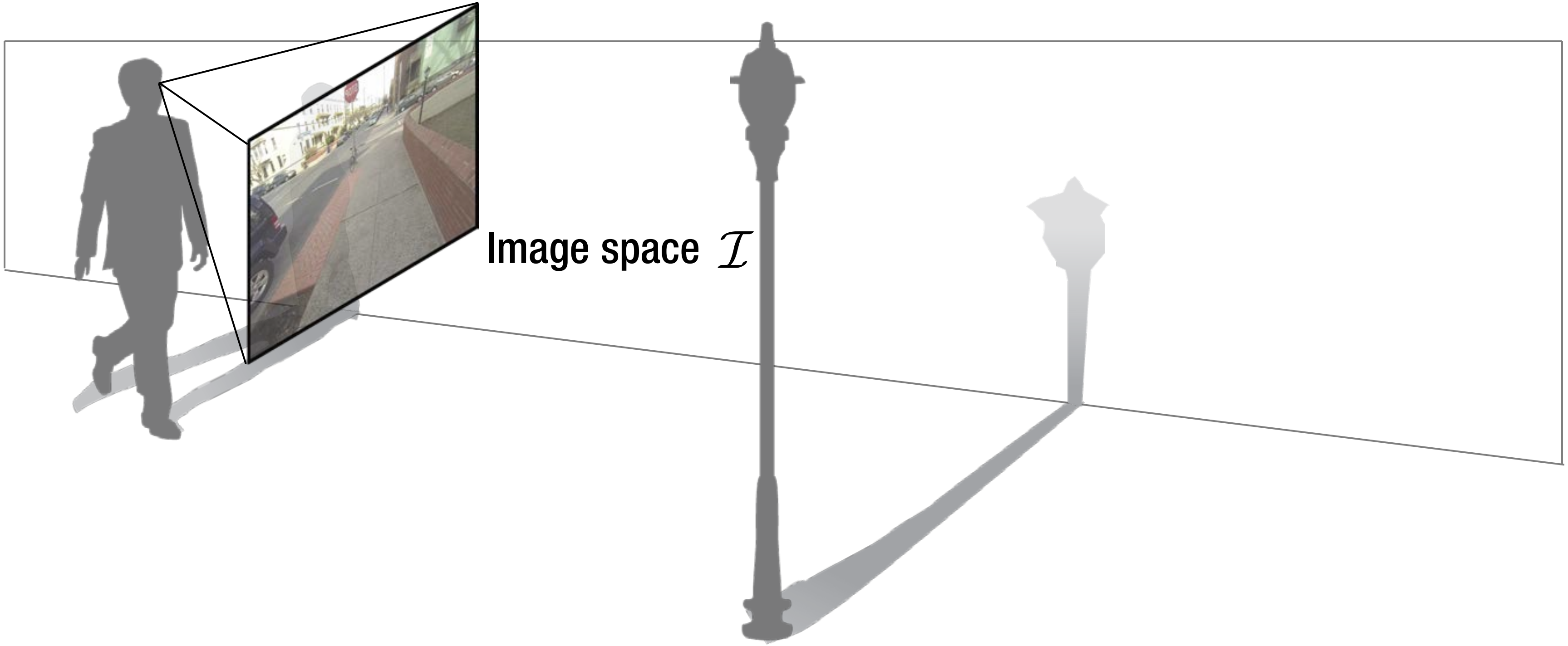


Image space $\mathcal{I}$

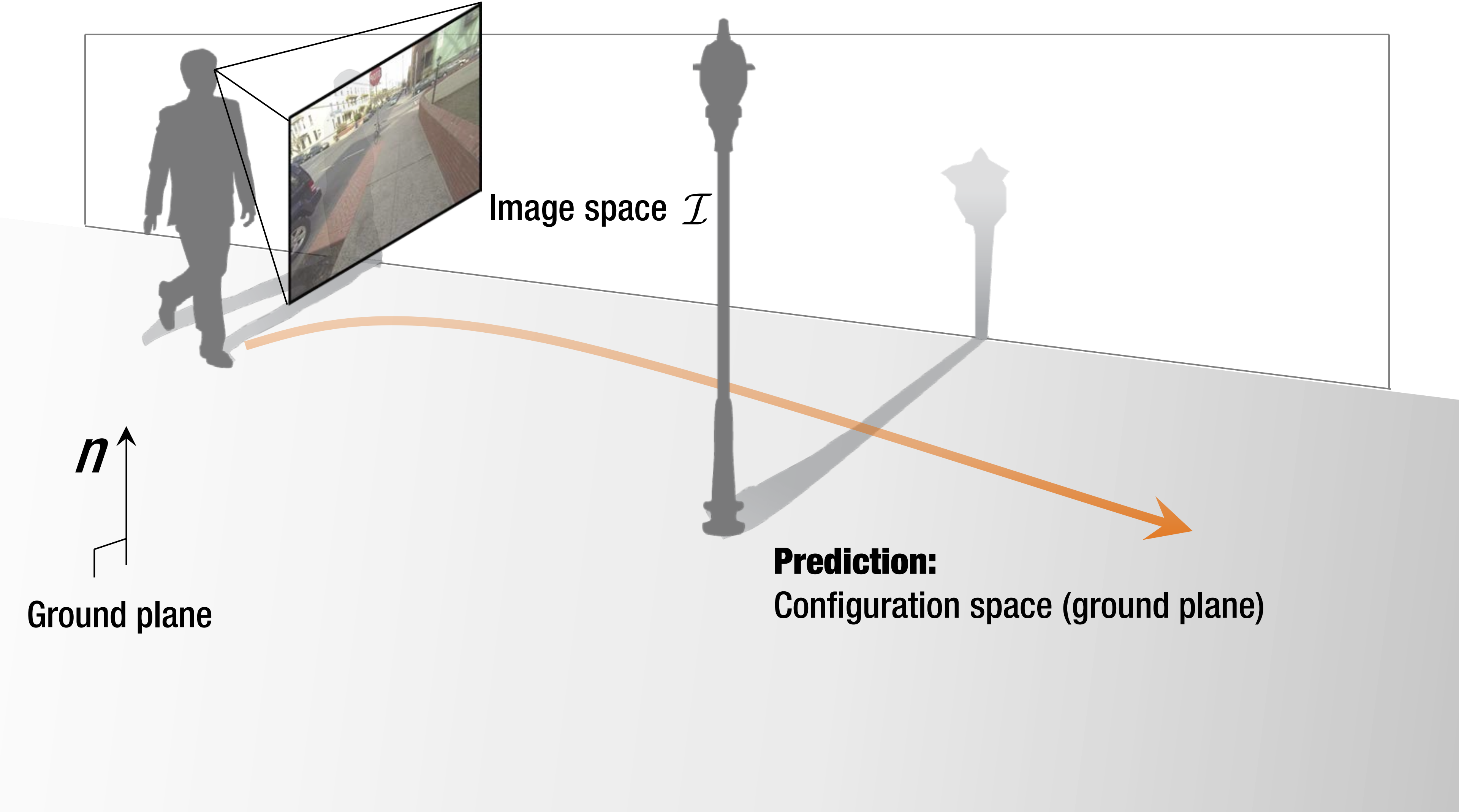
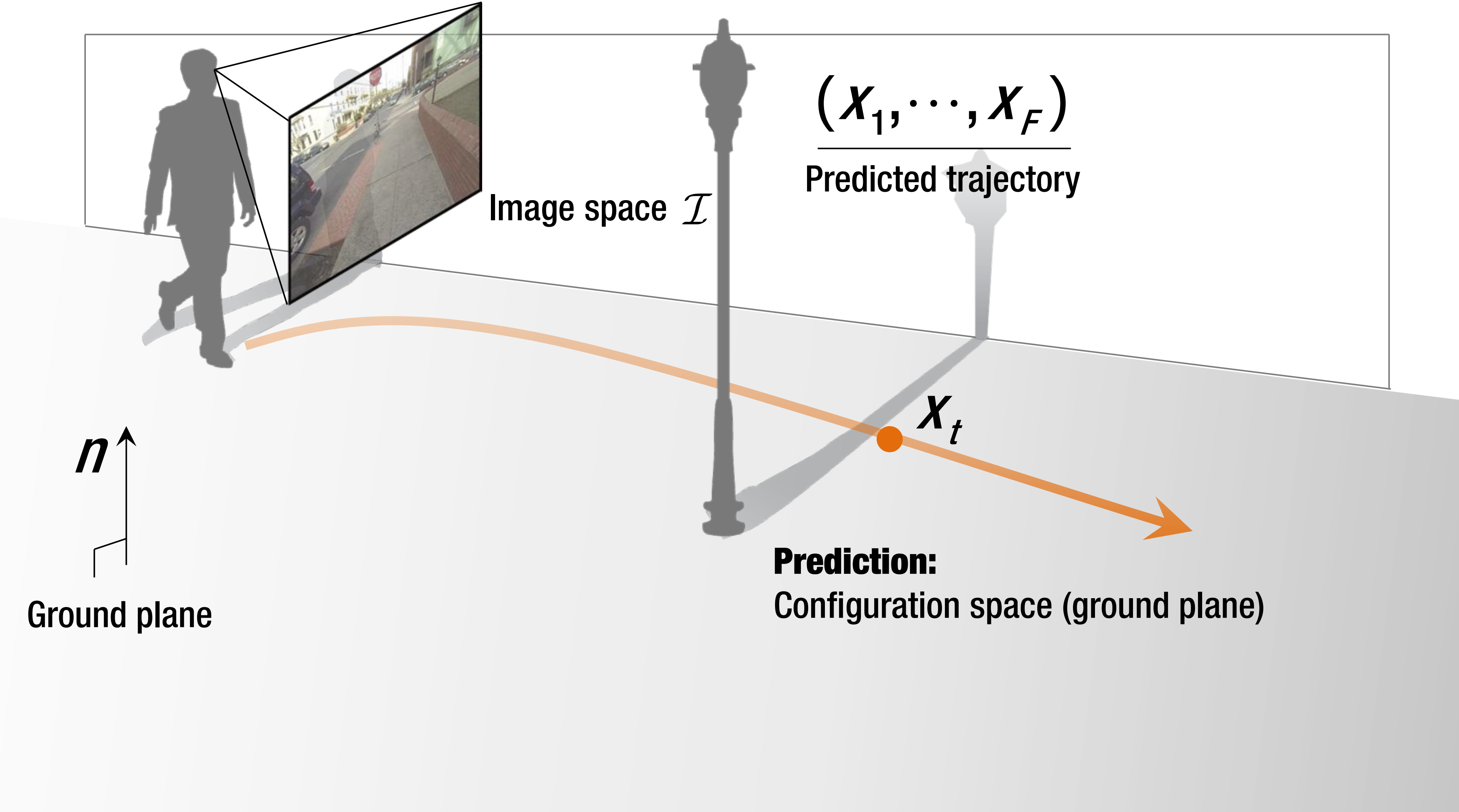
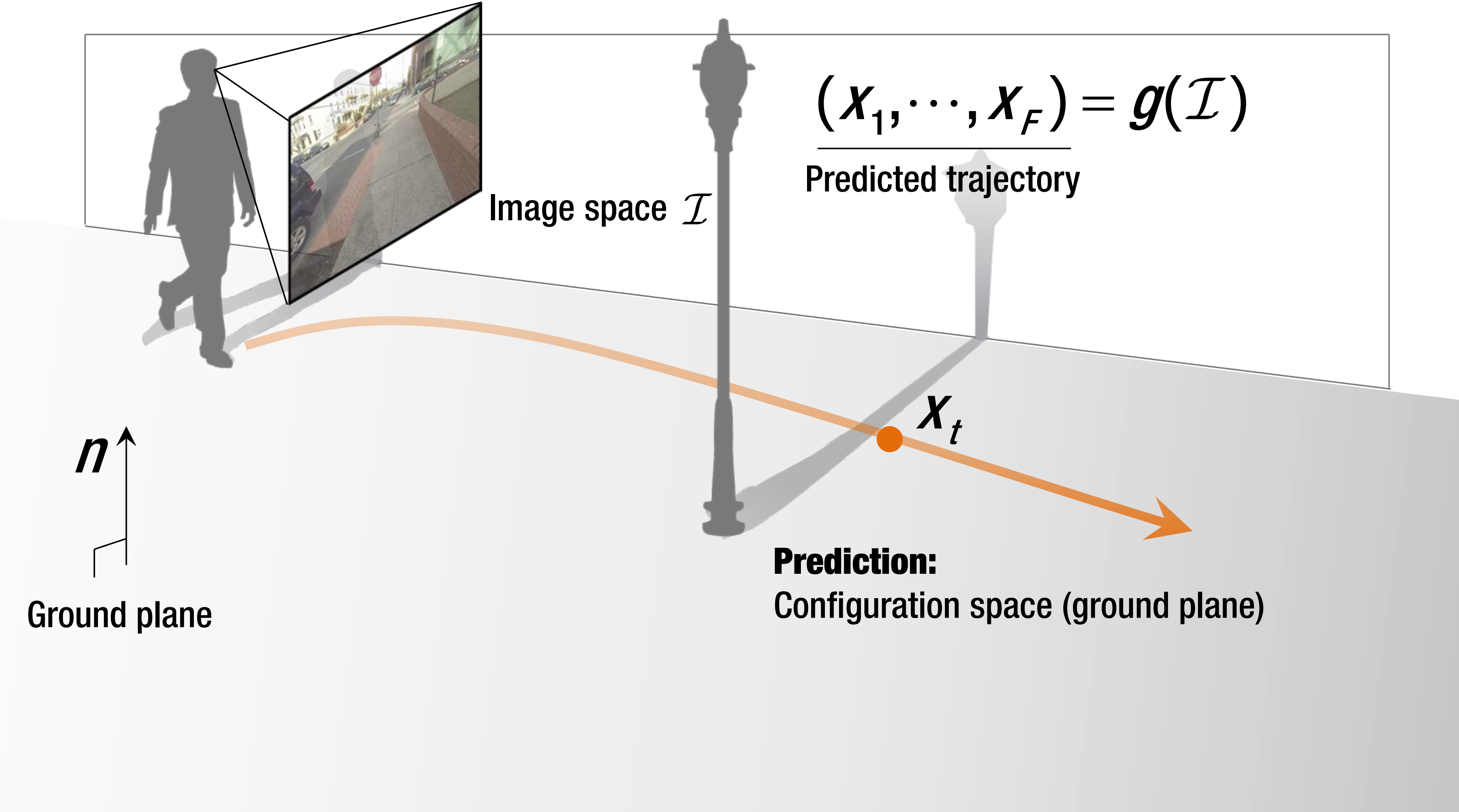


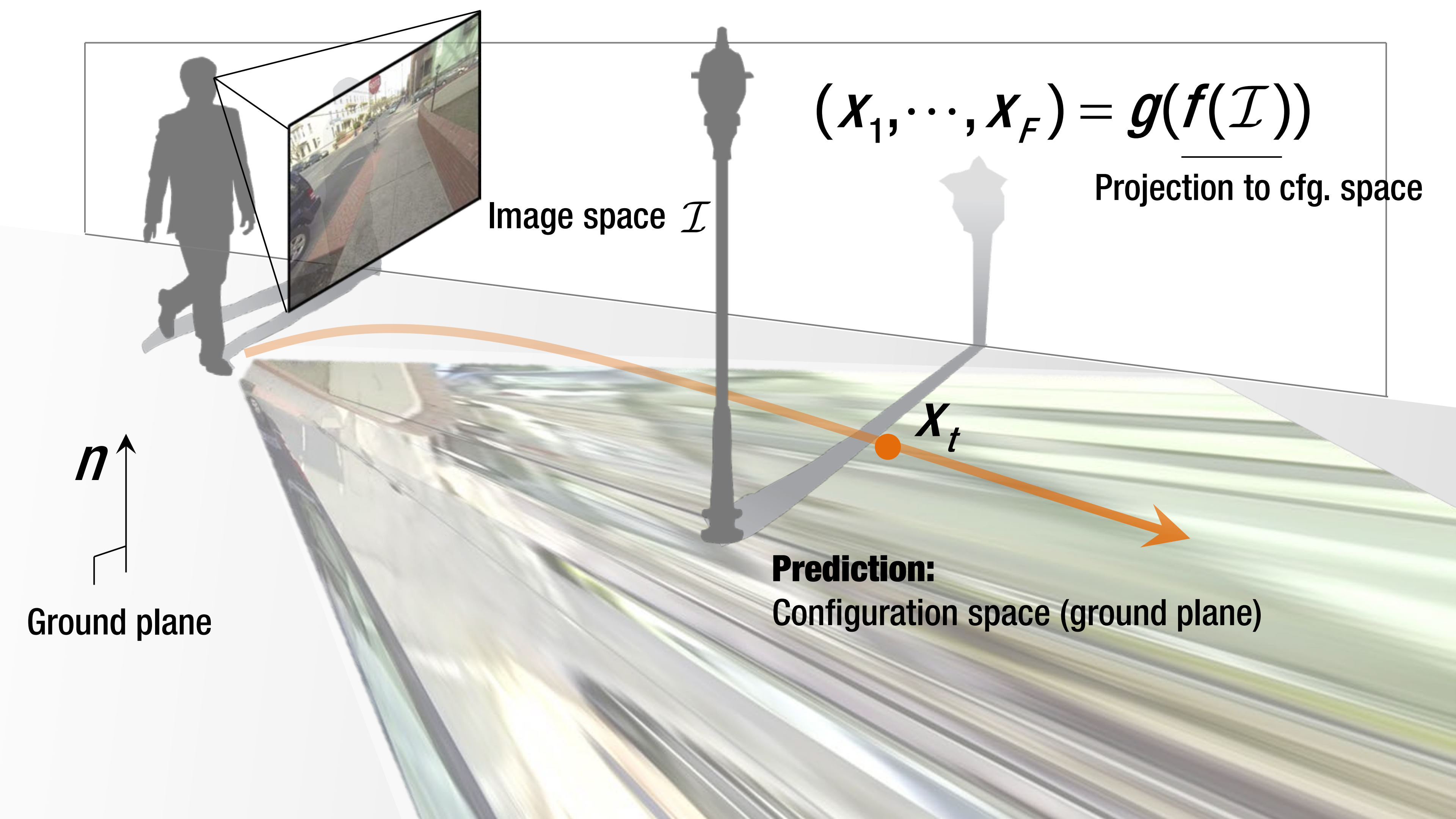
Image space $\mathcal{I}$
n
Ground plane
Prediction:
Configuration space (ground plane)



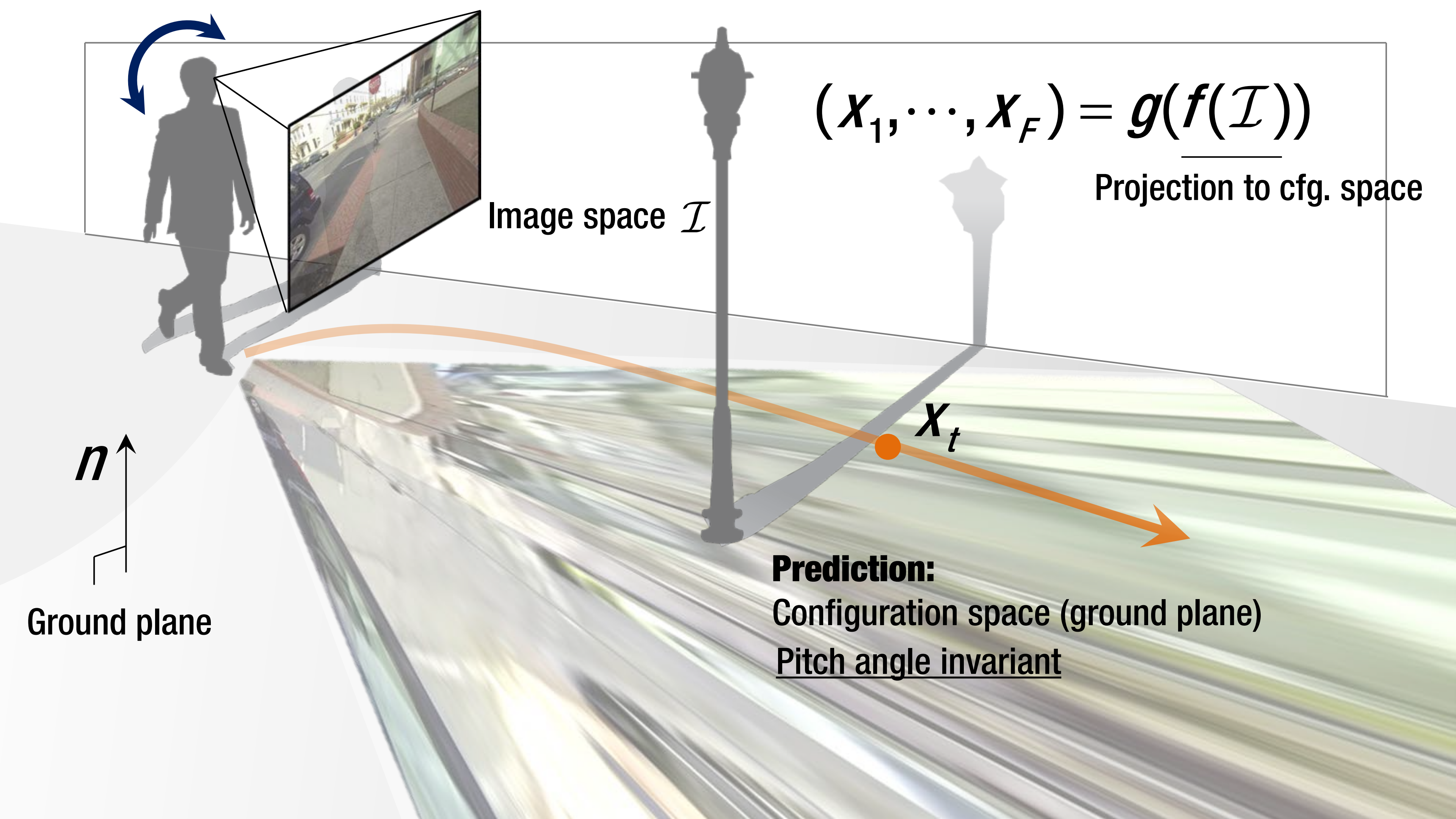
$(X_1, \cdots, X_F)$
Predicted trajectory
Image space $\mathcal{I}$
X_t
n
Ground plane
Prediction:
Configuration space (ground plane)



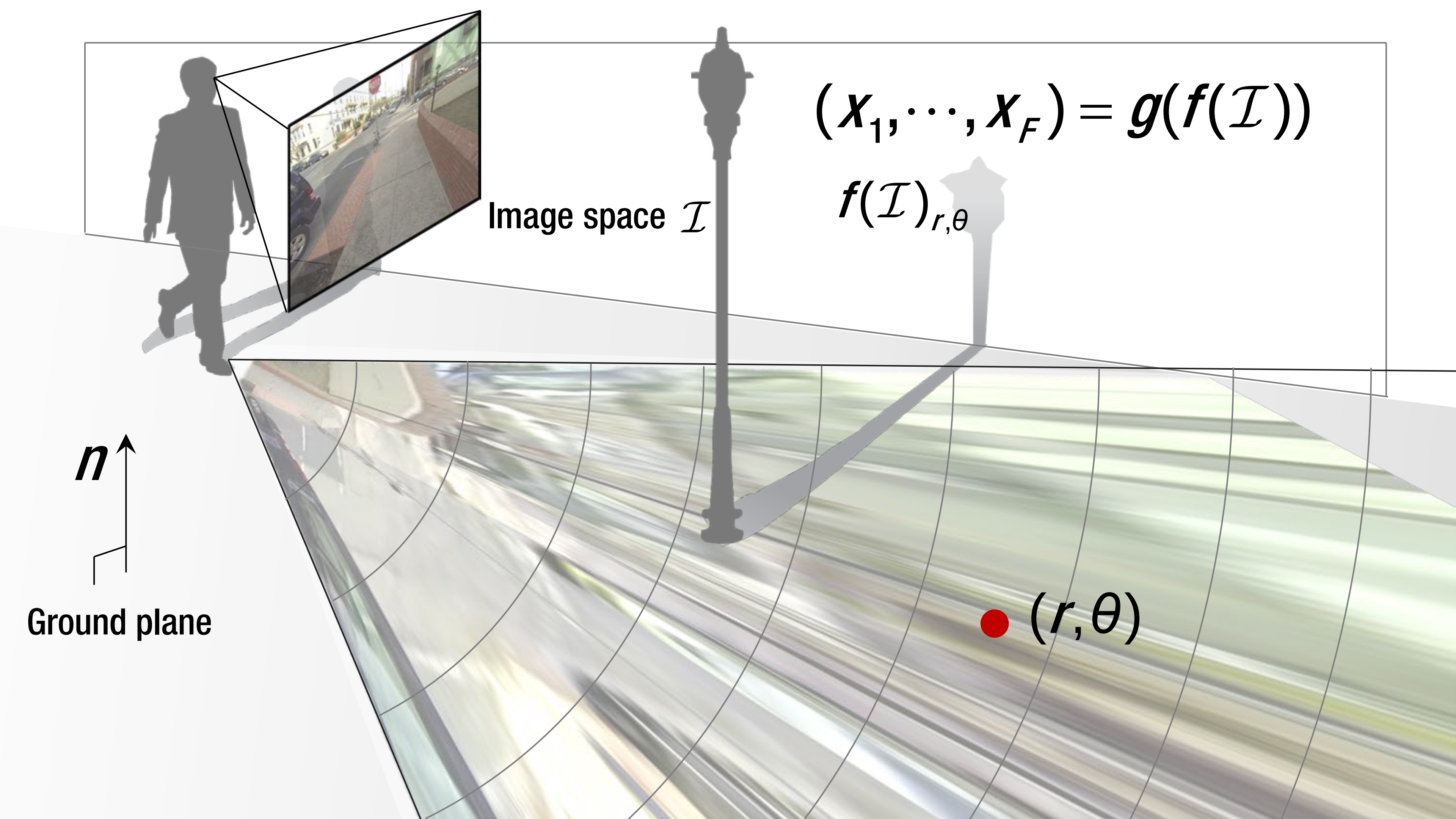
$(x_1, \cdots, x_F) = g(\mathcal{I})$
Predicted trajectory
Image space $\mathcal{I}$
x_t
n
Ground plane
Prediction:
Configuration space (ground plane)



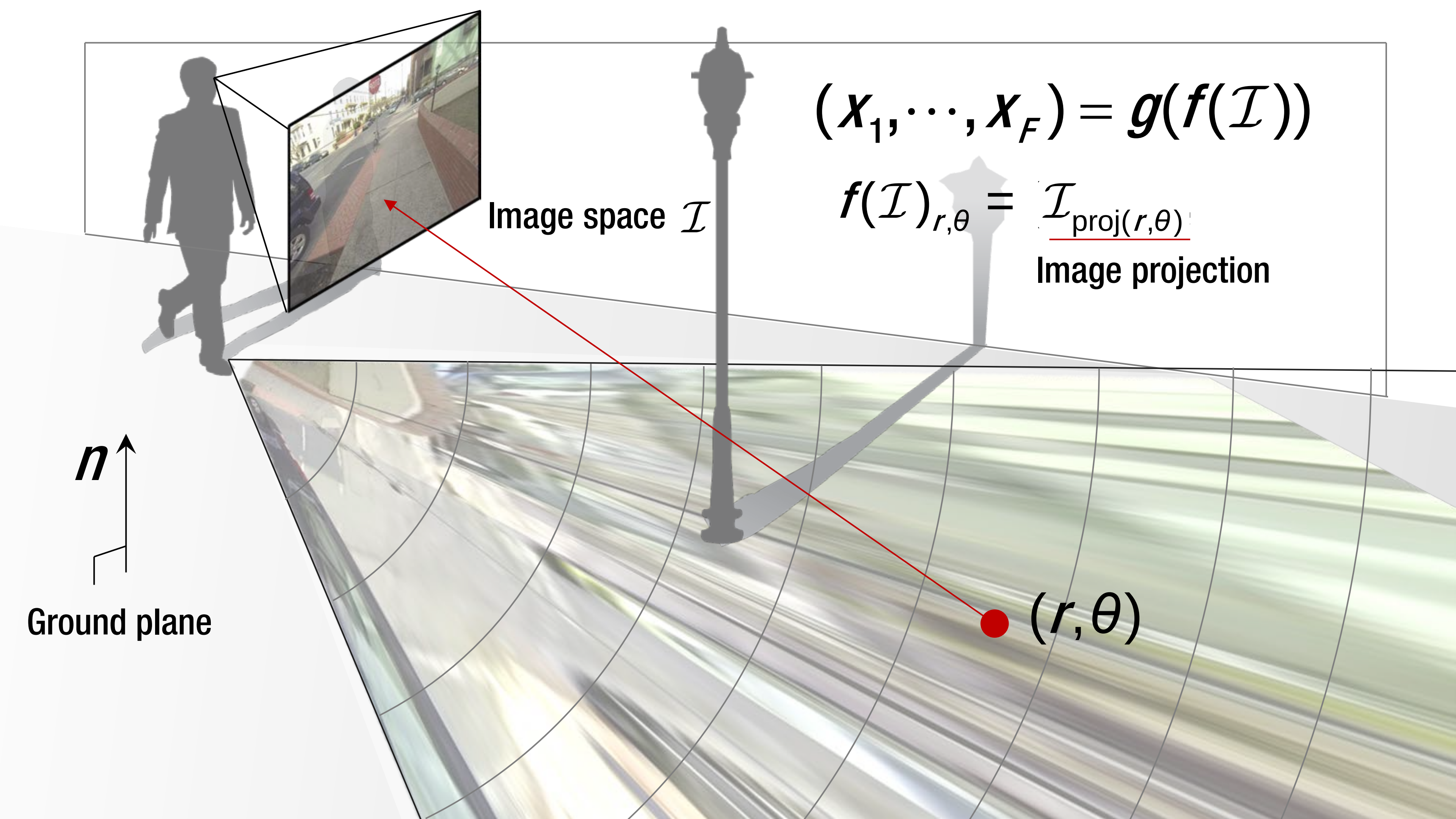
$(x_1, \cdots, x_F) = g(f(\mathcal{I}))$
Projection to cfg. space
Image space $\mathcal{I}$
x_t
n
Ground plane
Prediction:
Configuration space (ground plane)



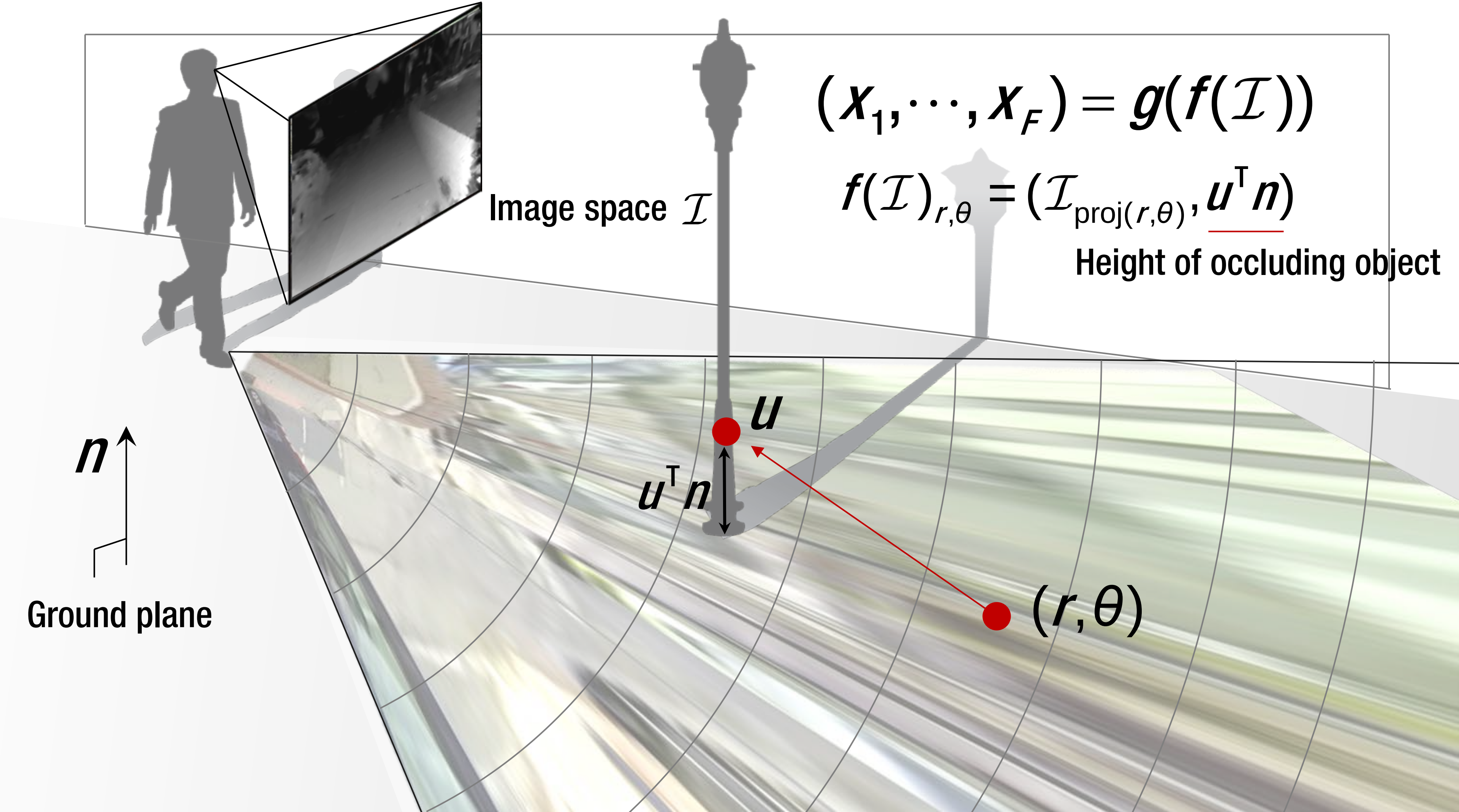
$(x_1, \cdots, x_F) = g(\underline{f(\mathcal{I})})$
Projection to cfg. space
Image space $\mathcal{I}$
x_t
n
Ground plane
Prediction:
Configuration space (ground plane)
Pitch angle invariant



$(x_1, \cdots, x_F) = g(f(\mathcal{I}))$
Image space $\mathcal{I}$
$f(\mathcal{I})_{r,\theta}$
n
Ground plane
(r, θ)



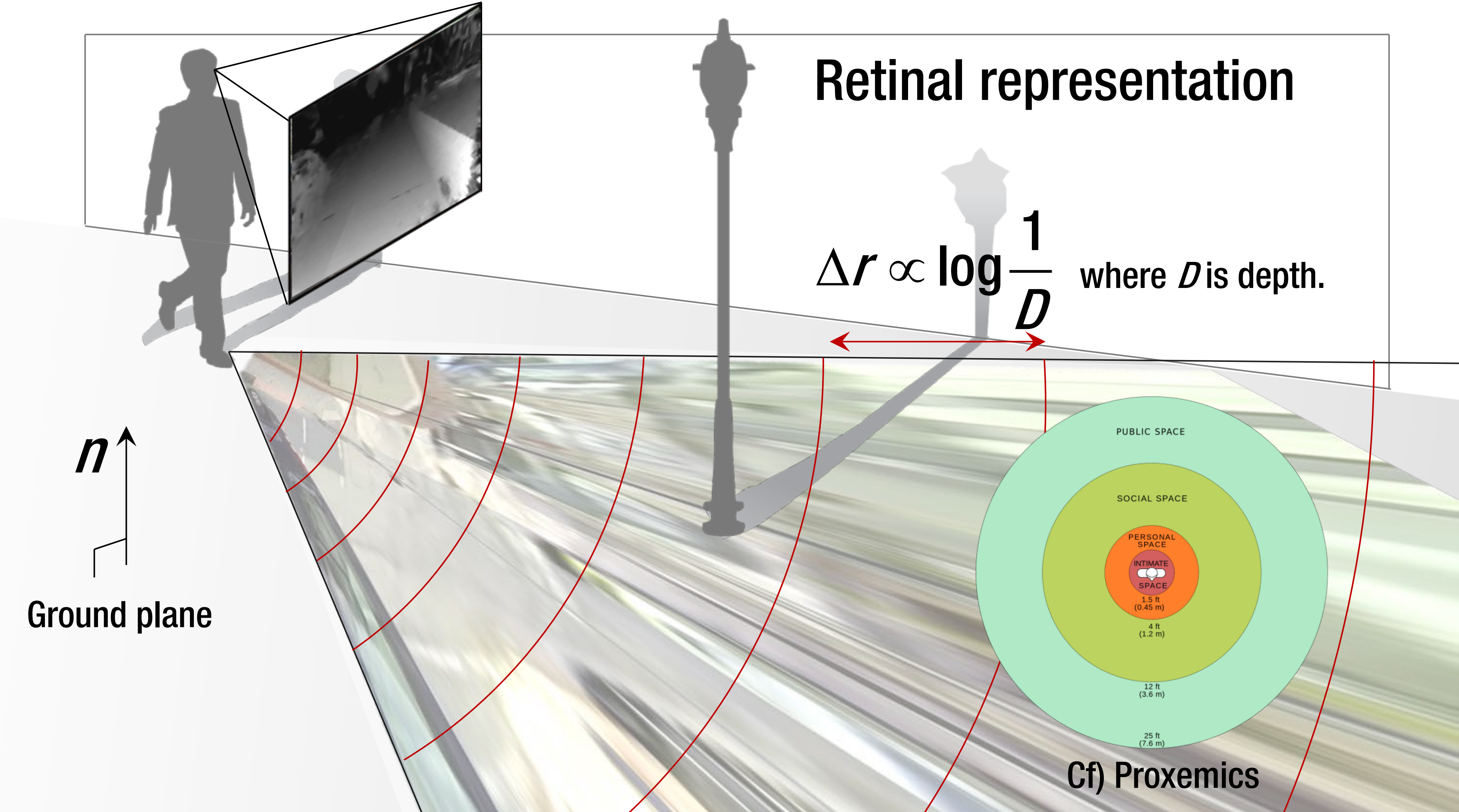
$(x_1, \cdots, x_F) = g(f(\mathcal{I}))$
$f(\mathcal{I})_{r,\theta} = \mathcal{I}_{\mathrm{proj}(r,\theta)}$
Image projection
Image space $\mathcal{I}$
n
Ground plane
(r,θ)



$(x_1, \cdots, x_F) = g(f(\mathcal{I}))$
$f(\mathcal{I})_{r,\theta} = (\mathcal{I}_{\text{proj}(r,\theta)}, u^{\mathsf{T}} n)$
Height of occluding object
Image space $\mathcal{I}$
u
$u^{\mathsf{T}} n$
n
(r,θ)
Ground plane

Retinal representation

$$\Delta r \propto \log \frac{1}{D}$$ where D is depth.

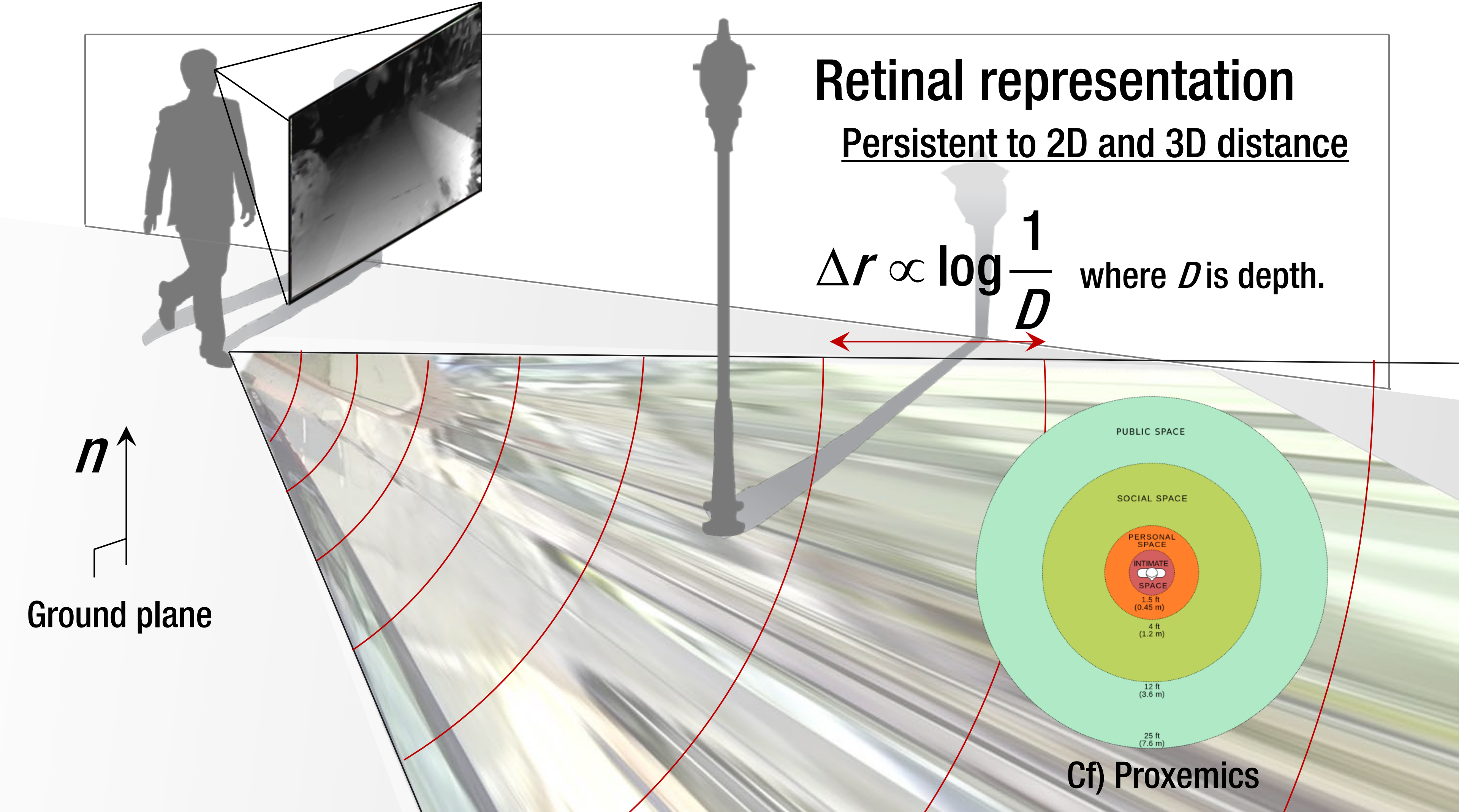


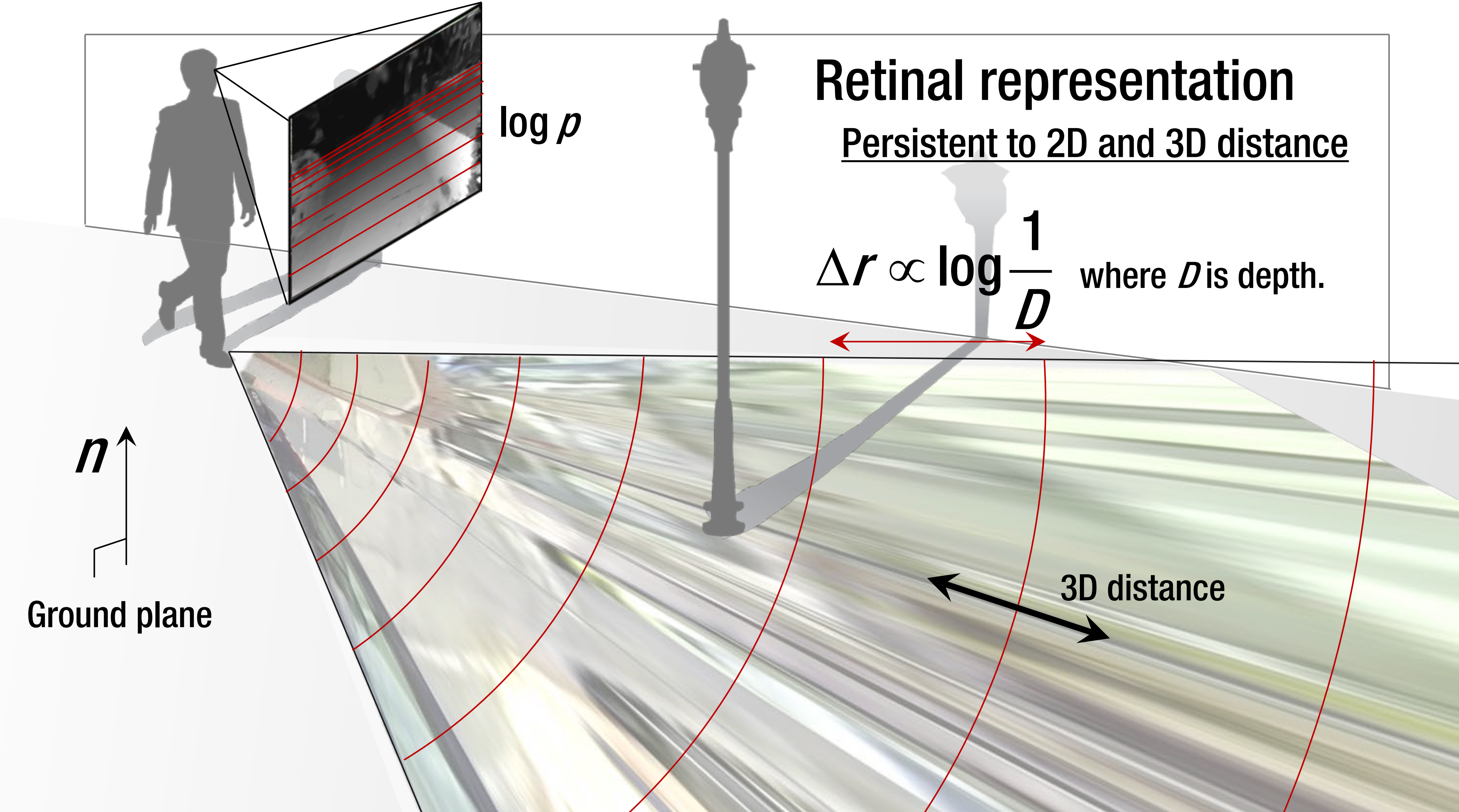
Cf) Proxemics

Retinal representation

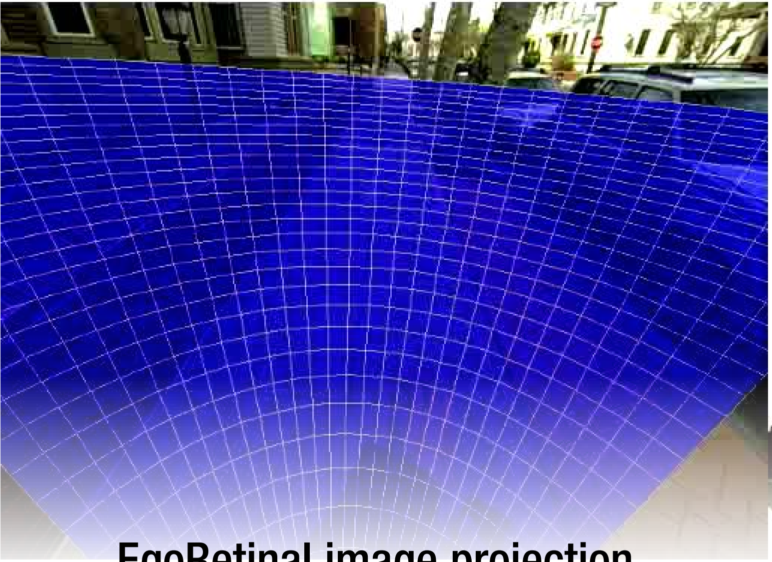
Persistent to 2D and 3D distance

$$\Delta r \propto \log \frac{1}{D}$$ where D is depth.



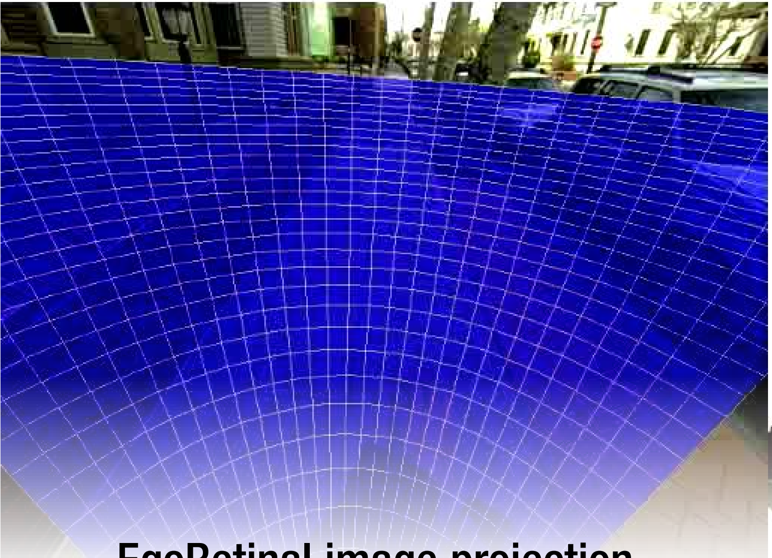


Retinal representation
Persistent to 2D and 3D distance
log p
$\Delta r \propto \log \frac{1}{D}$ where D is depth.
n
Ground plane
3D distance



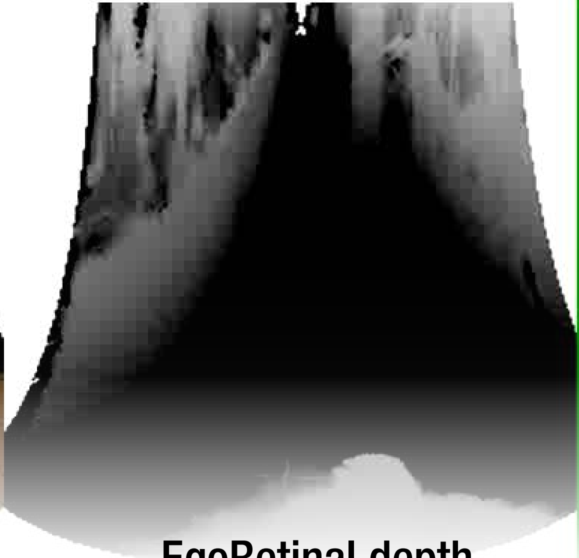
EgoRetinal image projection

EgoRetinal RGB

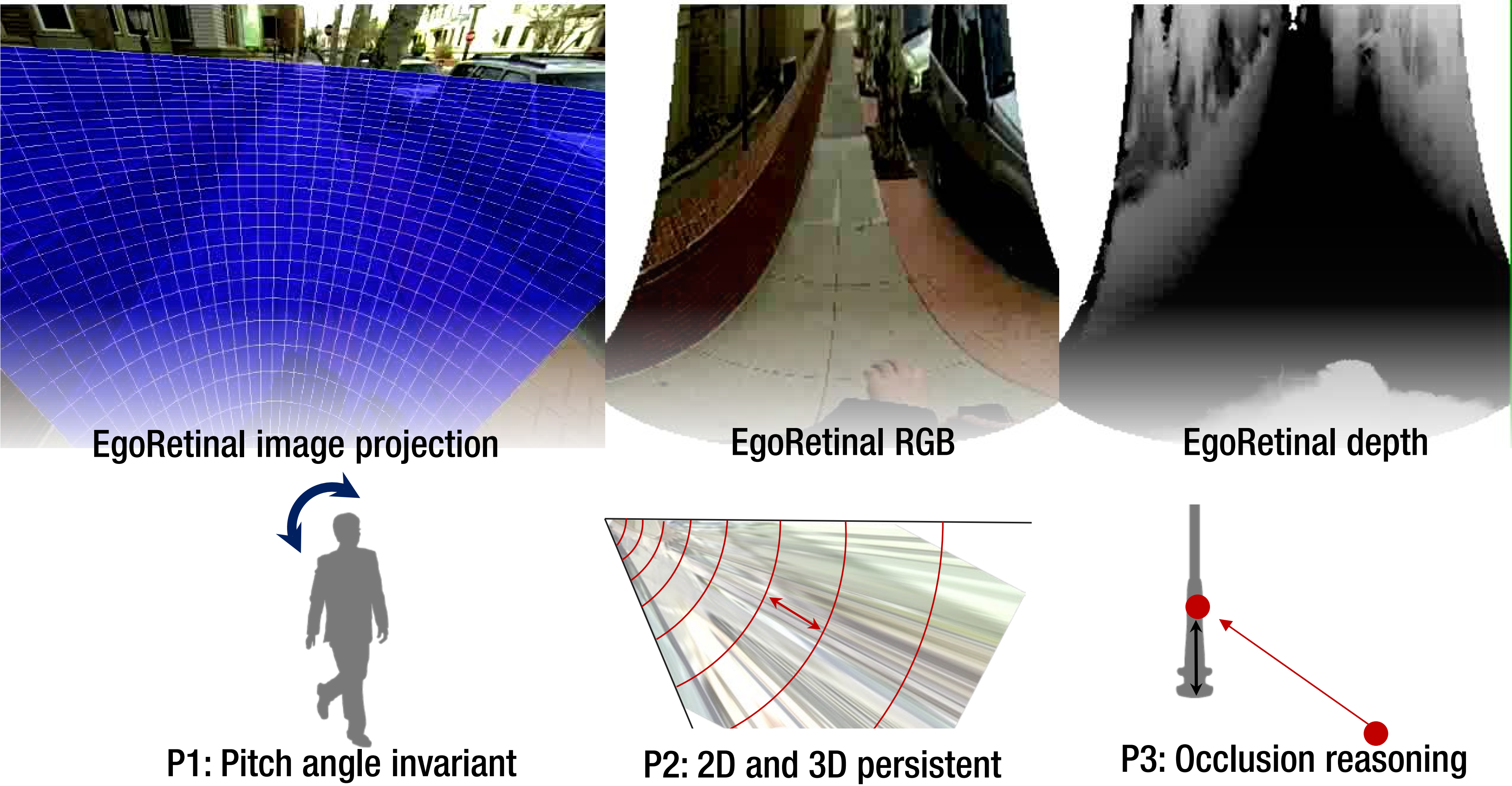


EgoRetinal image projection

EgoRetinal RGB



EgoRetinal depth


EgoRetinal image projection
EgoRetinal RGB
EgoRetinal depth
P1: Pitch angle invariant
P2: 2D and 3D persistent
P3: Occlusion reasoning

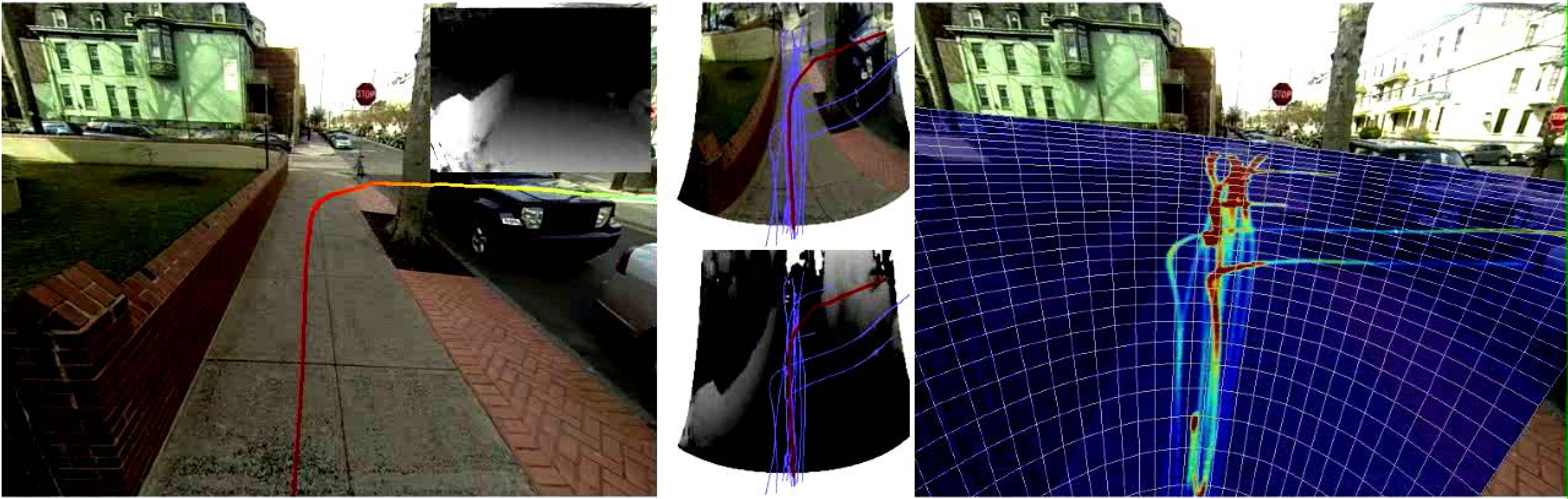
EgoMotion Dataset (outdoor)

Image Disparity								
Scene	IKEA	Costco	Mall	Park	School1/2	Downtown1/2	Grocery1/2/3	Bus1/2
Frames	966	577	2683	3088	3754/3736	2856/3405	2858/2892/2834	2292/1850
Duration	08:03	04:49	22:22	25:44	31:17/31:08	23:48/28:23	23:49/24:06/23:37	19:06/15:25

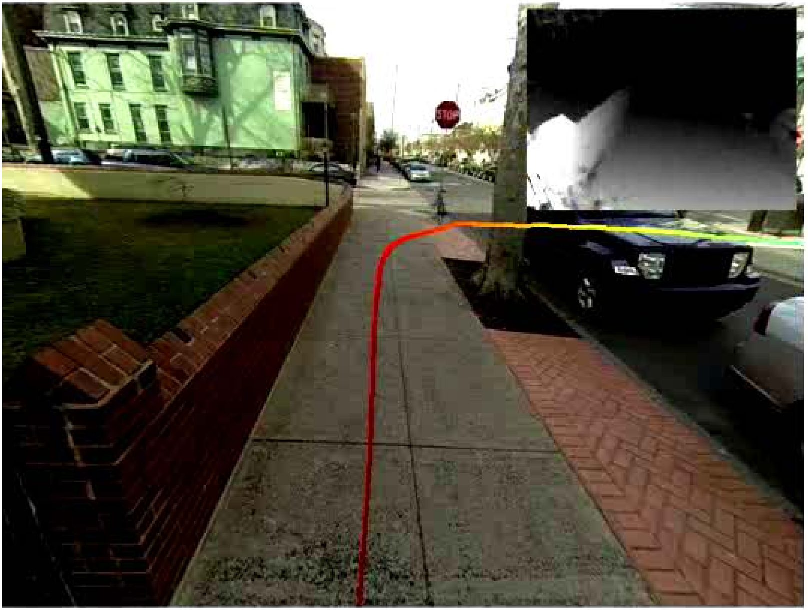
Image Disparity								
Scene	Campus1/2/3	CVS1/2	Train Sta.1/2	River1/2	Dep. store	Library	Apartment	Caffe
Frames	2607/1884/1975	2359/3337	4034/2568	3378/2250	2250	1255	2050	1550
Duration	21:44/15:42/16:28	19:40/27:49	33:37/21:24	28:09/18:45	13:20	10:30	17:05	13:00

Dataset summary

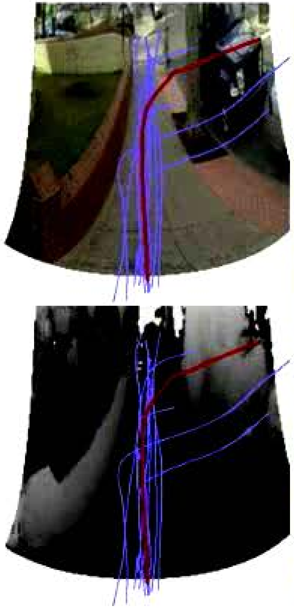
- 1280x960 stereo (100mm baseline, ~15m depth resolution)
- 26 scenes (13 indoor, 13 outdoor)
- 65.5k frames (9.1 hours)



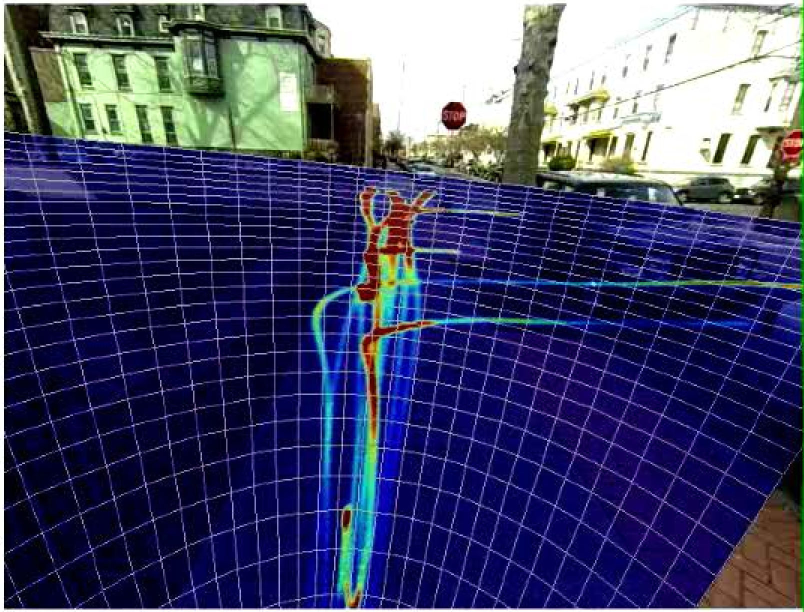
Ground truth EgoRetinal map Predicted trajectories



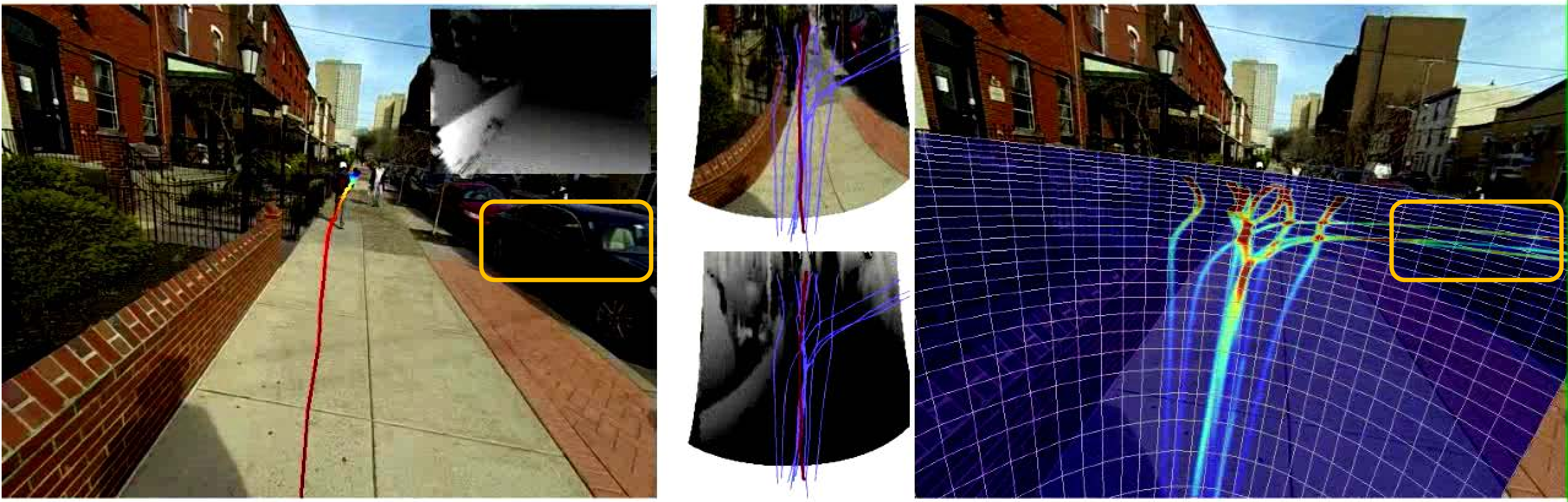
Ground truth



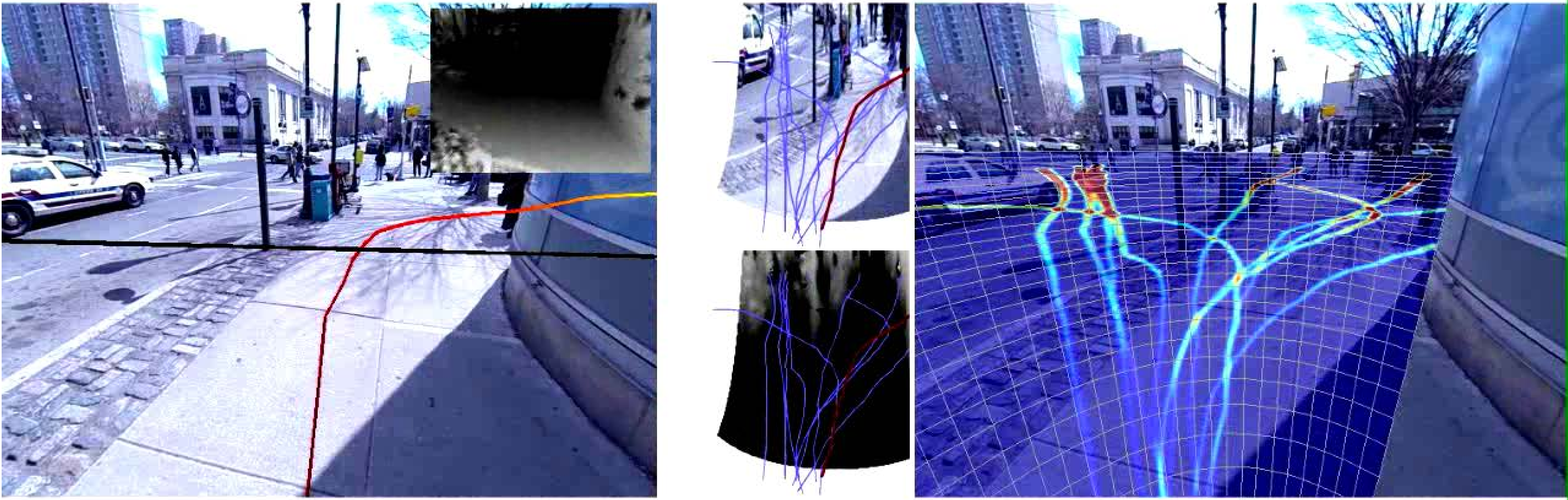
EgoRetinal map



Predicted trajectories



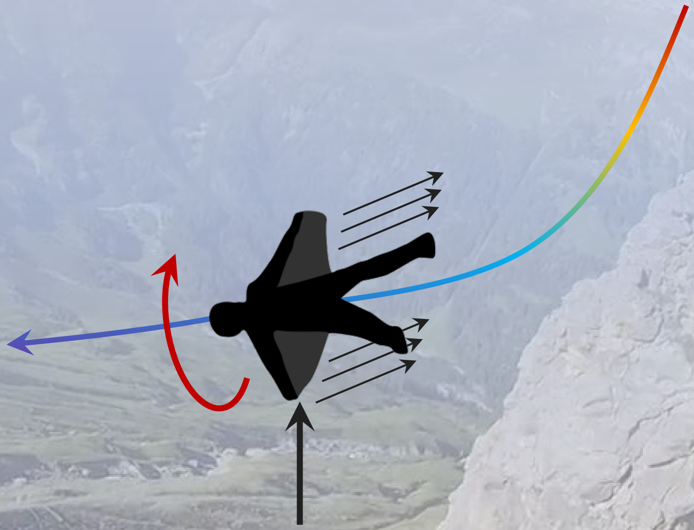
Ground truth EgoRetinal map Predicted trajectories



Ground truth EgoRetinal map Predicted trajectories

Visual Sensorimotor Behaviors II:
Control and Planning

Force from Motion
(ORAL) Afternoon, Wed, June 29



Egocentric Future Localization
(ORAL) Morning, Thr, June 30