Egocentric Future Localization

Hyun Soo Park Jyh-Jing Hwang Yedong Niu Jianbo Shi
University of Pennsylvania
{hypar,jyh,yedniu,jshi}@seas.upenn.edu

Abstract

We presents a method for future localization: to predict plausible future trajectories of ego-motion in egocentric stereo images. Our paths avoid obstacles, move between objects, even turn around a corner into space behind objects. As a byproduct of the predicted trajectories, we discover the empty space occluded by foreground objects.

One key innovation is the creation of an EgoRetinal map, akin to an illustrated tourist map, that 'rearranges' pixels taking into accounts depth information, the ground plane, and body motion direction, so that it allows motion planning and perception of objects on one image space. We learn to plan trajectories directly *on this EgoRetinal map using first person experience of walking around in a variety of scenes. In a testing phase, given an novel scene, we find multiple hypotheses of future trajectories from the learned experience. We refine them by minimizing a cost function that describes compatibility between the obstacles in the EgoRetinal map and trajectories. We quantitatively evaluate our method to show predictive validity and apply to various real world daily activities including walking, shopping, and social interactions.*

1. Introduction

Consider a dynamic scene such as Figure 1 where you, as the camera wearer, plan to pass through the corridor in the shopping mall while others walk in different directions. You need to plan your trajectory to avoid collisions with others and objects such as walls and fence. Looking ahead, you would plan a trajectory that enters into the shop by turning left at the corner although such space cannot be seen directly from your perspective.

The fundamental problem we are interested in is *future localization*: where am I supposed to be after 5, 10, and 15 seconds? This challenging task requires understanding of the scene in terms of a long term temporal human behaviors, with missing data due to occlusions. We solve this future path prediction problem by learning from our experiences of walking around (in different scenes) with egocentric stereo cameras[1] as shown in Figure 2(a).

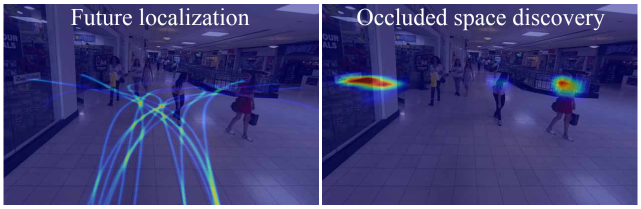


Figure 1. Where am I supposed to be after 5, 10, and 15 seconds? We present a method to predict a set of plausible future trajectories given a pair of egocentric stereo images. As a byproduct of the predicted trajectories, the occluded space by foreground objects such as the space inside of the shop or behind the ladies are discovered.

How past experiences lead to future path prediction into a novel scene?[2] There are two forms of learning signals: 1) what we see in egocentric visual images (on 2D retinal space), and 2) where we go through physical movements (on a 3D ground space). Learning from retinal visual images tell us about 'walking affordance': what are the walkable surfaces, which objects we can should avoid, which gap between objects we can pass through. Learning from physical movement tell us about 'spatial preferences': what is our sense of personal spaces with people, how we navigate around people and objects, and how we 'parse' a clutter scene and walk through it.

While 'walking affordance' and 'spatial preferences' can be learned in isolation, can we learn them jointly and use that information *directly* to effect path planning in a novel scene? Our goal of direct learning/reasoning is contrast to first extracting object semantic information then project them onto the 3D ground plane for motion planning (as done in robotics), which is sensitive to over-simplified semantic abstraction (not all icy roads are hard to walk on). This is also in contrast to directly learning paths in the image space then remap it to the 3D ground plane, which is sensitive to 3D geometrical alignment of the ground plane and objects in the novel scene.

We unify the spatial and perceptual learning signals by

[1] Any RGBD sensor such as Kinect is complimentary to our depth measurement.

[2] Note that a camera resectioning method such as perspective-n-point algorithms [22] does not apply as we predict a trajectory for a novel scene.

'rearranging' pixels into a synthetic EgoRetinal image, taking into accounts depth information, so that it is both easy for planning and for perception of objects. Inspired by proxemics [12] and Gibson's 'ground theory' of spatial perception [10], we represent the space around a camera wearer using an EgoRetinal map which reassembles an illustrated tourist map: an overhead map with objects seen from first person video projected onto it. 3D information is used to generate this map: 1) ground plane inferred from the RGBD data, and 2) the heading direction from instantaneous egomotion of walking. The pixels in input (retinal) image are rearranged according to its projection onto the ground plane, parametrized by a log-polar coordinate centered at the person's location and aligned with the direction of heading. Each pixel in EgoRetinal map retains its RGB value in the retinal image, and its object height of the ground computed from 3D depth. This 2.5D representation efficiently models a trajectory configuration space as it respects 3D distance and 2D image measurements.

A predictive future localization model is learned by exploiting in-situ first person stereo videos from various life logging activities. Given a testing EgoRetinal map, we find multiple hypotheses of future trajectories from the learned experience. We adapt these trajectories to the testing EgoRetinal map by minimizing a cost function that describes compatibility between the obstacles and trajectories.

Why EgoRetinal map? Two cues are strongly related to predict a trajectory of ego-motion, e.g., where is he or she going? (1) ego-cue: a vanishing point is often aligned with the direction of ego-motion, and 2D visual layout of the obstacles in the first person view implicitly encodes the semantics of the scene. (2) exo-cue: objects in a 3D scene such as road, buildings, and tables constrain the space where the wearer can navigate. Our EgoRetinal map representation exploits these two cues where we create an illustrated tourist map representation capturing both 2D visual arrangement of the obstacles (egocentric coordinate) and their 3D layout (exocentric coordinate). This representation allows us to analyze and understand different scene types and ego-motion in a unified coordinate system.

Contributions To our best knowledge, this is the first paper that tackles egocentric future localization via in-situ first person measurements. Core technical contributions of our paper are (a) an EgoRetinal map that encodes a spatial-visual distribution of objects with respect to an egocentric view, allowing us to apply perception and trajectory planning in a common coordinate system; (b) trajectory learning by inferring 'walking affordance' and 'spatial preferences' from past ego-walking experience; (c) occluded space discovery through trajectory prediction; and (d) the EgoMotion dataset with a depth and its long term camera trajectory, which includes diverse daily activities across camera wearers. Our EgoRetinal map representation significantly outperforms image based representation up to ×8 accuracy at 0-15 seconds.

2. Related Work

Our framework lies an intersection between behavior prediction and egocentric vision.

2.1. Human Behavior Prediction

Predicting where-to-go is a long standing task in behavioral science. This task requires to understand the interactions of agents with objects in a scene that afford a space to move. Pentland and Lin [32] modeled human behaviors using a hidden Markov dynamic model to recognize driving patterns. Such Markovian model is an attractive choice to encode human behaviors because it reflects the way humans make a decision [20, 23, 40]. These models, especially partially observable Markov decision process (POMDP), have influenced motion planning in robotics [19, 33, 35].

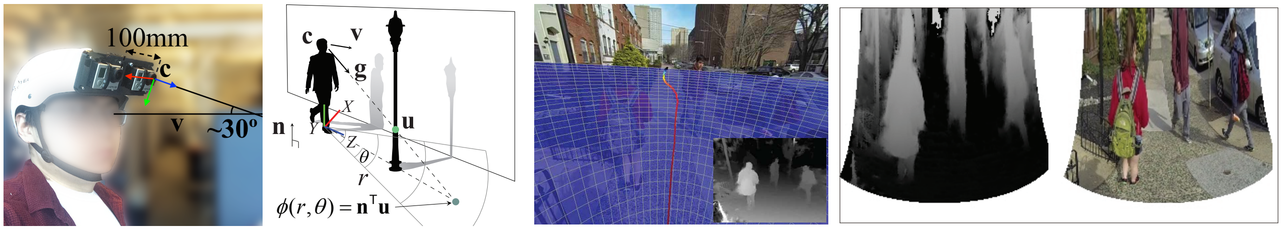
In computer vision, Ali and Shah [3] developed a flow field model that predicts spatial crowd behaviors for tracking extremely cluttered crowd scenes. Inspired by the social force model [13], Mehran et al. [27] predicted pedestrian behaviors in a crowd scene to detect abnormal behaviors, and Pellegrini et al. [31] used a modified model to track multiple agents. Vu et al. [38] predicted plausible activities from a static scene by associating the scene statistics and labeled actions. Our work is also closely related with path planning frameworks by Gong et al. [11] (path topology), Kitani et al. [16] (visual semantics), and Alahi et al. [2] (social affinity).

2.2. Egocentric Vision

A first person camera is an ideal camera placement to observe human activities because it reflects the attention of the camera wearer. This characteristics provides a powerful cue to understand human behaviors [6, 8, 15, 34, 36].

Traditional vision frameworks such as object detection, recognition, and segmentation frameworks have been integrated in first person data [6, 21, 24, 25, 34]. Notably, the relationship between visual semantics and egomotion has been recently studied [1, 14, 37]. In a social setting, first person cameras are used to capture person's visual attention, which allows localizing joint attention in social interactions. [7, 29, 30]. Such characteristics of first person cameras were used to generate interesting applications in vision [21, 39] and graphics [4, 17].

Unlike previous methods, our EgoRetinal map representation combines ego- and exo-coordinates, which allows us to perceive a scene and predict trajectories in a unified coordinate system. From this representation, we can understand scene dynamic affordance with respect to ego-coordinate, e.g., how does a person approaching to me affect my egomotion? Our method does not rely on prior processes such as semantic segmentation, object detection, or saliency prediction, which are often fragile to real world scenes or need manual annotations. Our trajectories are automatically annotated by structure from motion enabling a large scale prediction. We leverage the trajectory prediction to discover an



(a) Ego-stereo cameras (b) Geometry (c) Egocentric RGBD image (d) EgoRetinal map

Figure 2. (a) We use egocentric stereo cameras to capture our dataset. (b) We represent the space around a person using a trajectory configuration space called EgoRetinal map computed from (c) an egocentric RGBD image. The future trajectory is marked as a colored line. (d) The EgoRetinal map that is normalized to the ground plane and the direction of instantaneous ego-motion captures a likelihood of occlusion and its semantics. This EgoRetinal map is invariant to sudden gaze movement, camera placement offset, and scene orientation.

empty space that is not observable because of visual occlusion.

3. Representation of EgoRetinal Map

An EgoRetinal map is a trajectory configuration space [5] for space experienced from first-person view but visualized in an overhead bird-eye map, akin to an illustrated tourist map. There are three key ingredients in the EgoRetinal map.

First, inspired by Gibson's ground theory [10]—the ground plane plays a crucial role in our perception on a 3D spatial arrangement around us, we define the ground plane represented by (X^g, Z^g) coordinate from an input ego-centric depth map, $\mathcal{D}(x, y)$. The normal direction, $\mathbf{n}$, of the ground plane is aligned with the Y-axis (the gravity direction). The egocentric camera is located at $\mathbf{c} = \begin{bmatrix} 0 & h & 0 \end{bmatrix}^\mathsf{T}$ with h height above the ground plane as shown in Figure 2(b).

Second, the instantaneous ego-motion direction is used to define the Z^g direction of the EgoRetinal map. We identify the instantaneous ego-motion in 3D by projecting it onto the ground plane to define the Z^g direction, i.e., $Z^g = (\mathbf{v} - (\mathbf{n}^\mathsf{T}\mathbf{v})\mathbf{v})/\|\mathbf{v} - (\mathbf{n}^\mathsf{T}\mathbf{v})\mathbf{v}\|$ where $\mathbf{v} \in \mathbb{R}^3$ is the 3D instantaneous velocity. This representation is less sensitive to unintentional head vibration and intentional gaze direction movement, which both are much rapid movements comparing to the body motion.

Third, we project each pixel (x, y) in the egocentric image (retina) onto the ground plane coordinate system, (X^G, Z^G), and shown in Figure 2(b). We use a log-polar (r, θ) parametrization in the (X^g, Z^g) plane to obtain the EgoRetinal coordinate for each pixel, i.e., an injective map exists from the EgoRetinal map to the egocentric image: $f(r, \theta) = (x, y)$. In practice, we discretize the polar coordinate system by uniformly sampling in angle between $\pi/6$ and $5\pi/6$ and uniform sampling in the logarithm of radius as shown in Figure 2(c).

We construct an EgoRetinal map, i.e., $\mathcal{M}(r, \theta) = \begin{bmatrix} \phi(r,\theta) & \mathcal{I}(f(r,\theta))^\mathsf{T} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^4$ where $\mathcal{I}(x, y) \in \mathbb{R}^3$ is the RGB value of the egocentric image at (x, y). $\phi(r, \theta) \in \mathbb{R}$ measures the height of the point off ground $\mathbf{u}$, from the ground plane that intersects the ray, $\mathbf{g}$, from the center of eyes, $\mathbf{c}$, to (r, θ) with an occluding object, $\mathcal{O}$, i.e.,

$$\phi(r, \theta) = \mathbf{u}^\mathsf{T}\mathbf{n}, \tag{1}$$

where $\mathbf{u} = \min_{\lambda \in \mathcal{L}} \lambda\mathbf{g} + \mathbf{c}$ such that $\mathcal{L} = \{\lambda | \lambda\mathbf{g} + \mathbf{c} \subset \cup_{i=1}^{I} \mathcal{O}_i, \lambda > 0\}$. $\{\mathcal{O}_i\}_{i=1}^{O}$ is a set of objects in the scene.

We characterize a 3D spatial arrangement of objects with respect to the EgoRetinal coordinate:

1) Objects on the ground: EgoRetinal map produces an uniform sample of the space around us in terms of distance and direction as shown in the first column of Figure 3(a). In contrast, the Cartesian representation on ground plane (second column) drastically collapses image pixels of short range area to construct the configuration space. Also it is hard to represent destinations on the horizon, as they have infinite spatial extend. In the other hand, the Cartesian in image plane (third column) does not encode a 3D spatial layout and rapidly diminishes image pixels of long range area, which does not reflect 3D distance.

2) Objects off the ground: EgoRetinal map provides a cylindrical unwrapping of the image. It reduces foreshortening of surfaces that are facing us, thus creating a frontal view of the facade. This process normalizes the object appearances with respect to the camera location producing a less orientation-variant shape representation that makes the learning efficient.

3) Objects at height of the viewer: EgoRetinal map emphasizes higher objects up the the height of the camera as they are spatially enlarged relative to objects close to the ground. This characteristics allows us to prioritize obstacles when planning a trajectory, e.g., a curb is easily passed over.

The EgoRetinal representation supports learning *future localization* from first person videos by combining cues from 3D scene geometry and ego-motion direction. Its benefits include: 1) a coordinate system normalized by the direction of ego-motion provides a common 3D reference frame to learn; 2) overhead view representation removes the variations in first person 3D experience due to sudden gaze movements and camera placement offset, 3) the log-polar encoding and sampling gives more importance to nearby space, and 4) the depth masking encodes implicitly both roll and pitch angle of head, making it more situation aware.

3.1. Trajectory Representation

Let $\mathbf{X} = \begin{bmatrix} r_1 & \theta_1 & \cdots & r_F & \theta_F \end{bmatrix}^\mathsf{T}$ be a 2D trajectory on the ground plane where F is the time steps to predict and r_i and θ_i are distance (radial) and direction (angle) with respect to the person's feet location at the i^{th} time instance as shown in Figure 2(b). In practice, this trajectory can be obtained by projecting 3D camera poses between the $f+1$ and $f+F$ time instances at the f^{th} time instant onto the ground plane. This allows us to represent all trajectories in the same coordinate system (EgoRetinal space), which are normalized by the direction of instantaneous ego-motion.

Topological properties of trajectory Two trajectories may share the same topology with respect to a scene while their Euclidean distance remains large. For instance, two trajectories that move in parallel along a wide road are topologically same while trajectories that bifurcate at an Y-junction are topologically different. To encode such topological variance, we augment the orientation toward near objects at each point in trajectory measured by an obstacle image: $\mathbf{Y} = \begin{bmatrix} \mathbf{X}^\mathsf{T} & \boldsymbol{\Omega}^\mathsf{T} \end{bmatrix}^\mathsf{T}$ where $\boldsymbol{\Omega} = \begin{bmatrix} \omega_1 & \cdots & \omega_F \end{bmatrix}^\mathsf{T}$, and $\omega_i = \text{atan2}(\boldsymbol{v}_i \times \boldsymbol{z}_i, \boldsymbol{v}_i \cdot \boldsymbol{z}_i)$ where $\boldsymbol{v}_i \in \mathbb{R}^2$ is the unit tangential direction of trajectory at the i^{th} time instant[3]. $\boldsymbol{z} = \boldsymbol{y}/\|\boldsymbol{y}\|$ is a unit vector towards near obstacles, i.e., $\boldsymbol{y} = \left(\sum_{\boldsymbol{w}\in\mathcal{N}_\epsilon(\mathbf{X}_i)} \phi(\boldsymbol{w})\boldsymbol{w}\right) / \left(\sum_{\boldsymbol{w}\in\mathcal{N}_\epsilon(\mathbf{X}_i)} \phi(\boldsymbol{w})\right) - \mathbf{X}_i$ where $\mathbf{X}_i$ is the point on the trajectory at the i^{th} time instant and $\mathcal{N}_\epsilon$ is neighboring pixels of $\mathbf{X}_i$ with radius $\epsilon \propto 1/r$ in the EgoRetinal map. $\boldsymbol{\Omega}$ encodes an angular distribution of near obstacles with respect to the tangential direction of the trajectory as shown in the second column of Figure 3(b). This representation is invariant to a homotopy class between trajectories, i.e., line integral along a trajectory encodes the winding number of each homotopy class [11].

4. Prediction

A trajectory of ego-motion is associated with an EgoRetinal map, i.e., given a depth image, we know how we explored the space in the training data (Section 5). By leveraging a trajectory configuration space, or EgoRetinal map described in Section 3, in this section, we present a method to predict a set of plausible trajectories and to discover the occluded space that the predicted trajectories pass.

4.1. Ego-motion Prediction

Trajectory planning consists of coarse and fine levels. At a coarse level, we are interested in finding path satisfying the 'spatial preferences' induced by the scene type and spatial layout: what are possible destinations to move towards, whether we choose to move through a clutter environment of people, parked cars and furniture. We retrieve the coarse level plan by finding top m most similar EgoRetinal maps in the training set, and copy their trajectories directly as $\{\mathbf{X}_{D_i}\}_{i=1}^m$. Given the EgoRetinal map, we compute a feature $\mathbf{h} = \begin{bmatrix} h(\mathcal{M}_D)^\mathsf{T} & h(\mathcal{M}_{RGB})^\mathsf{T} \end{bmatrix}^\mathsf{T}$ where $h(\mathcal{M})$ is a feature representation of the image $\mathcal{M}$ computed by pretrained network [18]. $\mathcal{M}_D$ and $\mathcal{M}_{RGB}$ are the EgoRetinal map for depth and RGB images, respectively, where $\mathcal{M}_D$ is treated as an independent three-channel image. Note that other compact feature representations can be used as a complimentary. We further group trajectories sharing similar topological features into k where $k \ll m$ based on the trajectory representation described in Section 3.1 using a k-mean clustering algorithm.

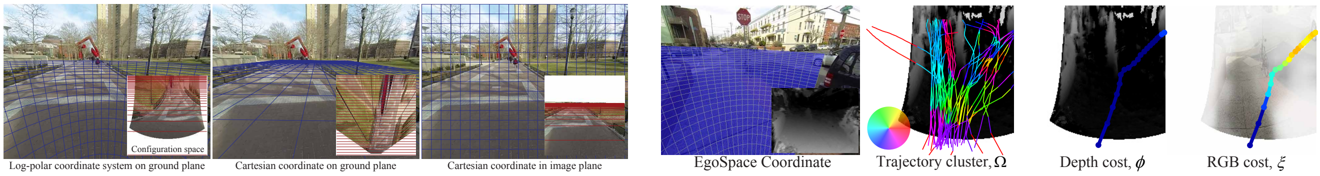
At a fine level, we ensure that the trajectory is physically feasible. The main requirement is learning a 'walking avoidance' probability function $\xi(r_i, \theta_i))$ on the EgoRetinal map using RGB values, i.e., $\mathcal{I}(f(r,\theta))$. We fine-tune the fully convolutional network [26] using our training data. Given an RGB EgoRetinal map, $\mathcal{M}_{RGB}$, we convolve a Gaussian along the ground truth trajectory and invert its intensity, producing a pixel-wise 'walking avoidance' map. We use 500×500 EgoRetinal image with 227×227 receptive field and modify the FC8 layer to predict a binary output. We automatically label pixels that trajectories have passed from the training data, which enables the network to predict a pixel-wise probability of walkability as shown in the forth column of Figure 3(b). In conjunction with ξ, we incorporate $\phi(r,\theta)$ that also indicates obstacles via the depth.

Estimating $\mathbf{X}$ that to find a path that stays in the ground plane while conforming both the obstacle map and the 'walking avoidance'. The trajectory minimizes the following cost function:

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \sum_i^F \left(\phi(r_i,\theta_i) + \xi(r_i,\theta_i)\right) + \lambda\|\mathbf{X} - \mathbf{X}_D^*\|^2 \\ \text{subject to} \quad & \mathbf{X}_D^* = \underset{\{\mathbf{X}_{D_j}\}_{j=1}^m}{\text{argmin}} \ \|\mathbf{X} - \mathbf{X}_{D_j}\|^2 \end{aligned} \tag{2}$$

where λ controls how much deviation allows from the retrieved trajectories. Equation (2) is used in robotics communities for various path planning tasks. However, this does not take into account the trajectory that is partially occluded by objects because the occluded part of the trajectory always produces higher cost. Instead, we introduce a novel cost function that minimizes a trajectory cost difference between the given image and the retrieved image from

[3] In practice, we use a trigonometric reprsentation instead of ω, i.e., $\begin{bmatrix} \cos\omega & \sin\omega \end{bmatrix}$, to avoid the singularity at 2π.


Configuration space
Log-polar coordinate system on ground plane
Cartesian coordinate on ground plane
Cartesian coordinate in image plane
EgoSpace Coordinate
Trajectory cluster, Ω
Depth cost, ϕ
RGB cost, ξ

(a) Trajectory configuration space comparison

(b) Trajectory representation

Figure 3. (a) We represent an EgoRetinal map on ground plane (first column). The blue grid shows sampling density to construct its representation and the red lines indicates unit 3D distance. The Cartesian on ground plane (second column) drastically collapses image pixels of short range area and more importantly, models one destination. The Cartesian in image plane (third column) does not model 3D spatial layout of scenes and rapidly diminishes image pixels of long range area. (b) A trajectory is represented by its polar coordinates $\mathbf{X}$ and angular distribution $\mathbf{\Omega}$ of near objects (second column) that allows us to cluster into a few multiple hypotheses. We predict a trajectory that minimizes depth and semantic incompatibility in the EgoRetinal space. Colormap represents incompatibility with respect to D (third column) and RGB (forth column) channel along the trajectory. For the RGB cost, we overlay with a probability of pixel-wise incompatibility learned by a fully convolutional neural network [26].

the database:

$$\underset{\mathbf{X}}{\text{minimize}} \quad \sum_{i}^{F} \left(h\left(\phi\left(r_i,\theta_i\right) - \phi_D\left(r_i,\theta_i\right)\right)\right. \\ \left. + h\left(\xi\left(r_i,\theta_i\right) - \xi_D\left(r_i,\theta_i\right)\right)\right) + \lambda\|\mathbf{X}-\mathbf{X}_D^*\|^2$$

$$\text{subject to} \quad \mathbf{X}_D^* = \underset{\{\mathbf{X}_{D_j}\}_{j=1}^{m}}{\text{argmin}} \quad \|\mathbf{X}-\mathbf{X}_{D_j}\|^2, \qquad (3)$$

where $h(\cdot)$ is the hinge loss function, and ϕ_D and ξ_D are the cost of the retrieved trajectory from its EgoRetinal map. This minimization finds a partially occluded trajectory as long as there exists a trajectory in the database that has similar occlusion cost.

4.2. Occluded Space Discovery

The predicted trajectories allow us to discover the hidden space occluded by foreground objects because the trajectories can be still predicted in the hidden space. We build a likelihood map of the occluded space as follows:

$$\psi(\mathbf{x}) = \frac{\sum_{j=1}^{J}\sum_{i=1}^{F} \exp\left(-\|\mathbf{x}-\mathbf{X}_{ij}\|^2/2\sigma^2\right)\phi\left(\mathbf{X}_{ij}\right)}{\sum_{j=1}^{J}\sum_{i=1}^{F} \exp\left(-\|\mathbf{x}-\mathbf{X}_{ij}\|^2/2\sigma^2\right)}, \qquad (4)$$

where $\psi(\mathbf{x}=(r,\theta))$ is the likelihood of the occluded space that a trajectory can pass through at the evaluating point $\mathbf{x}$ in the EgoRetinal map. $\mathbf{X}_{ij} = (r_{ij},\theta_{ij})$ is the i^{th} point of the j^{th} predicted trajectory, J is the number of predicted trajectories, and σ is the bandwidth for the Guassian kernel. Equation (4) takes into account the likelihood of the predicted trajectories weighted by the likelihood of the occlusion. $\psi(\mathbf{x})$ is high when many trajectories are predicted at $\mathbf{x}$ while $\phi(\mathbf{x})$ is high.

5. EgoMotion Dataset

We present a new dataset, EgoMotion dataset, captured by first person stereo cameras (GoPro Hero 3 cameras with 100mm baseline) as shown in Figure 2(a). This dataset includes various indoor and outdoor scenes such as Park, Malls, and Campus with various activities such as walking, shopping, and social interactions. The stereo cameras are calibrated prior to the data collection and synchronized manually with a synchronization. Detailed data analysis can be found in the supplementary material.

Depth Computation We compute disparity between the stereo pair after stereo rectification. A cost space of stereo matching is generated for each scan line and match each pixel by exploiting dynamic programming in a coarse-to-fine manner.

3D Reconstruction of Ego-motion We reconstruct a camera trajectory using a standard structure from motion pipeline with a few modifications to handle a large number of images[4]. We independently reconstruct partitioned dataset and merge them using overlapping images. Then, we project the reconstructed camera trajectory onto the ground plane estimated by fitting a plane using RANSAC [9].

Scenes We collect both indoor and outdoor data, which consists of 26 scenes with 65.5k frames of 9.1 hours long in total, including walking on campus, in parks and downtown streets, shopping in the mall, cafe and grocery, as well as taking public transportation. The data consists of various activities (walking, talking, and shopping), scenes (campus, park, malls, and downtown streets), cities, and time. The dataset is summarized in Table 1.

6. Result

We apply our method to predict ego-motion and hidden space in real world scenes by leveraging the EgoMotion dataset. Testing data are completely isolated from the training data in terms of geographical locations, i.e., a camera resectioning method such as perspective-*n*-point algorithms [22] does not apply.

[4] A 30 minute walking sequence at a 30 fps reconstruction rate produces HD 108,000 images.

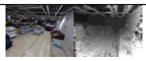
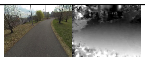
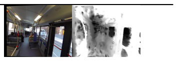
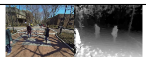
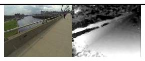
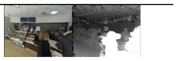
Image Disparity								
Scene	IKEA	Costco	Mall	Park	School1/2	Downtown1/2	Grocery1/2/3	Bus1/2
Frames	966	577	2683	3088	3754/3736	2856/3405	2858/2892/2834	2292/1850
Duration	08:03	04:49	22:22	25:44	31:17/31:08	23:48/28:23	23:49/24:06/23:37	19:06/15:25
Image Disparity								
Scene	Campus1/2/3	CVS1/2	Train 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

Table 1. EgoMotion dataset

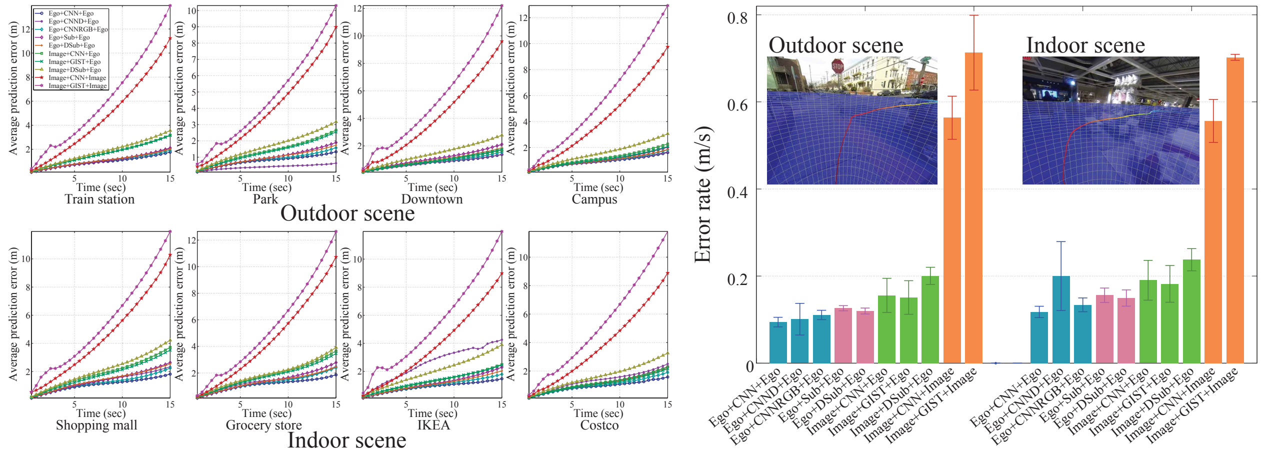


Figure 4. We compare our method (A1: Ego+CNN+Ego) with 9 baseline representations (please find baseline descriptions in Section 6.1; A2: Ego+CNND+Ego; A3: Ego+CNNRGB+Ego; A4: Ego+Sub+Ego; A5: Ego+DSub+Ego; A6: Image+CNN+Ego; A7: Image+GIST+Ego; A8: Image+DSub+Ego; A9: Image+CNN+Image; A10: Image+GIST+Image). Our EgoRetinal map representation significantly outperforms other representations. Comparing to image based prediction (A9 and A10), it produces ×8 accurate prediction. In the right column, we show error rate, i.e., how fast error increases. All errors are measured in the meteric scale.

6.1. Quantitative Evaluation

We quantitatively evaluate our trajectory prediction by comparing with ground truth trajectories. Multiple trajectories are often equally plausible, e.g., Y-junction, while one ground truth trajectory is available per image. To account multiple hypotheses, we find top $K = 10$ trajectories and use the trajectory that produces minimum error, i.e., $e = \min_j \|\mathbf{X} - \mathbf{Z}_j\|^2$ where $\mathbf{Z}_j$ is the j^{th} retrieved trajectory. Note that unlike previous approaches measured a spatial distance between trajectories [16][5], our evaluation measures a spatiotemporal distance between trajectories because the time scale also needs to be considered.

Total 10 representations (A1-A10 in Table 3)[6] are compared. The first column in Table 3 represents a trajectory configuration space which can be either EgoRetinal map or image space as shown in Figure 3(a). Also the feature representation can be either EgoRetinal map or image space. For each representation, we use various data type, e.g., RGB vs. RGBD vs. D. Coarse rep. stands for a feature vector constructed by uniform sampling of RGB or D data. GIST is a scene descriptor [28]. Note that A6, A7, and A8 use image features which are not normalized by ego-motion and ground plane. A9 and A10 are predictions on image domain without 3D information about the scene. For these predictions, we project the predicted trajectories in the image onto the ground plane to measure 3D distance between trajectories.

	Representation for perception							
	EgoRetinal map					Image space		
	CNN			Coarse rep.		CNN	GIST	Coarse rep.
Planning	RGBD	D	RGB	RGB	D	RGB	RGB	D
EgoRetinal	A1	A2	A3	A4	A5	A6	A7	A8
Image space						A9	A10	

Table 3. Baseline algorithms

We compare our method (A1) in 4 outdoor (Train staion, Park, Downtown, and Campus) and 4 indoor scenes (Shopping mall, Grocery store, IKEA, and Costco). Figure 6 shows average error between a retrieved trajectory and ground truth over time. A2 and A3 performs similar to ours and A4 and A5 are slightly worse. A6, A7, and A8 are 1.5-2 times worse than our method and image based prediction (A9 and A10) is 7-8 times worse than ours. A similar trend is also observed in error rate, i.e., how fast the error increases. Table 2 summarizes the error across scenes and

[5] A dynamic time warping was used to handle a time scale.

[6] These baseline algorithms are designed by ours because no previous algorithm exists to predict the trajectories of ego-motion

	0∼5 secs								5∼10 secs								10∼15 secs							
	O1	O2	O3	O4	I1	I2	I3	I4	O1	O2	O3	O4	I1	I2	I3	I4	O1	O2	O3	O4	I1	I2	I3	I4
Image+GIST+Image	2.00	1.66	1.77	1.69	2.03	1.66	2.04	1.68	5.49	4.22	4.83	5.25	4.96	4.86	4.95	4.91	10.78	8.04	9.45	10.22	9.38	9.70	9.34	9.36
Image+CNN+Image	1.28	1.20	1.12	1.11	1.44	1.18	1.11	1.05	4.28	3.60	3.69	3.77	4.12	4.08	3.55	3.49	8.63	6.97	7.45	7.57	7.95	8.28	7.01	6.91
Image+DSub+Ego	0.72	0.71	0.60	0.67	0.87	0.86	0.85	0.71	1.77	1.64	1.42	1.52	2.06	1.89	1.96	1.61	2.89	2.59	2.28	2.47	3.38	3.05	3.17	2.61
Image+GIST+Ego	0.64	0.56	0.39	0.49	0.72	0.74	0.62	0.50	1.57	1.29	0.88	1.04	1.73	1.72	1.34	1.07	2.56	2.06	1.43	1.67	2.82	2.74	2.02	1.70
Image+CNN+Ego	0.66	0.60	0.38	0.52	0.79	0.77	0.59	0.54	1.58	1.36	0.87	1.13	1.85	1.80	1.29	1.16	2.59	2.15	1.40	1.82	2.98	2.89	2.04	1.81
Ego+DSub+Ego	0.46	0.43	0.42	0.47	0.61	0.60	0.48	0.52	1.05	0.94	0.95	0.96	1.41	1.38	1.06	1.09	1.64	1.42	1.51	1.49	2.10	1.99	1.67	1.74
Ego+Sub+Ego	0.44	0.42	0.45	0.48	0.57	0.58	0.49	0.52	1.00	0.94	1.05	0.99	1.36	1.37	1.11	1.12	1.67	1.52	1.70	1.61	2.15	2.20	1.82	1.83
Ego+CNNRGB+Ego	0.44	0.41	0.39	0.47	0.55	0.57	0.48	**0.48**	1.00	0.87	0.85	0.96	1.21	1.28	1.02	1.00	1.56	1.26	1.27	1.46	1.78	1.90	1.50	1.52
Ego+CNND+Ego	0.47	**0.24**	**0.34**	0.46	0.62	0.68	1.11	0.61	1.05	**0.41**	0.79	0.91	1.37	1.41	2.78	1.38	1.66	**0.53**	1.15	1.37	1.97	1.95	3.79	2.01
Ego+CNN+Ego	**0.44**	0.40	0.38	**0.45**	**0.54**	**0.57**	**0.45**	**0.48**	**0.94**	0.81	**0.77**	**0.89**	**1.12**	**1.21**	**0.95**	**0.94**	**1.43**	1.11	**1.13**	**1.32**	**1.52**	**1.56**	**1.26**	**1.32**

Table 2. Mean error (m) (O1-O4: outdoor scenes, I1-I4: indoor scenes)

time.

Comparison with moving straight We compare with a linear trajectory, "moving straight" in terms of predictive precision—how often one of the predicted trajectories aligns with the ground truth trajectory, i.e., $prec. = \sum_{i=1}^{N} \mathcal{D}_i / N$, where N is the number of testing images. $\mathcal{D}_i = 1$ if $\min_k \max_t \|\widehat{\mathbf{X}}_t - \mathbf{X}_t^k\| < \epsilon$, and $\mathcal{D}_i = 0$ otherwise where $\mathbf{X}_t^k$ is the location at the t^{th} time instant of the k^{th} predicted trajectory and $\widehat{\mathbf{X}}$ is the ground truth trajectory. We set $\epsilon = 1.5m$. The predictive precision is summarized in Table 4 and our prediction (A1) clearly outperforms "Moving straight" prediction.

	Indoor			Outdoor		
	0∼5	5∼10	10∼15	0∼5	5∼10	10∼15
Moving straight	0.571	0.221	0.124	0.443	0.259	0.103
A10	0.507	0.379	0.229	0.535	0.391	0.267
A6+Depth	0.710	0.561	0.384	0.554	0.407	0.293
A1 (Eq. 2)	0.690	0.570	0.401	0.567	0.432	0.289
A1 (Eq. 3)	**0.825**	**0.693**	**0.482**	**0.683**	**0.538**	**0.373**

Table 4. Predictive precision

Occluded Space Discovery We quantitatively evaluate our occluded space discovery by measuring detection rate, D/N where D is the number of true positive detection and N the total number of detection produced by the space discovery. We threshold the likelihood of the occluded space, ψ, from Equation (4) and manually evaluate whether the detection is correct. Note that no ground truth label is available unless the camera wearer already had passed through the space. The detection rate in Table 5 indicates that our method predicts the outdoor scenes better than the indoor scenes. This is because the indoor scenes such as Grocery and IKEA, the camera wearer had a number of close interactions with objects such as shelves or products where the view of the scenes are substantially limited.

	Mall	Grocery	IKEA	Park	Train sta.	Campus
Detection rate	0.59	0.24	0.39	0.62	0.66	0.64

Table 5. Detection rate

6.2. Qualitative Evaluation

We apply our method on real world examples to predict a set of plausible trajectories of ego-motion and the occluded space by foreground objects. Our training dataset is completely separated from testing data, e.g., Grocery scene was trained to predict IKEA scene. Given a depth image, we estimate the ground plane by a RANSAC based plane fitting with gravity and height prior. This ground plane is used to define the EgoRetinal map.

Figure 1 and 5 illustrate our results from the EgoMotion dataset. In Figure 5, we show (1) image and ground truth ego motion; (2) input depth image; (3) EgoRetinal map overlaid with the predicted trajectories (purple) and ground truth trajectory (red); (4) projection of the trajectories (future localization); (5) projection of occluded space (Occlusion discovery). For all scenes, our method predicts the plausible trajectories that pass through unexplored space.

Obstacle Avoidance Our cost function in Equation (3) minimizes cost difference between trajectories from training data and testing data. This precludes a trajectory passing through an object unless the retrieved trajectory was partially occluded. EgoRetinal map captures the obstacle avoidance as shown in Downtown I, Campus II, Walmart, Costco, and so on. For Downtown II and Bench, we predict plausible trajectories while the ground truth trajectory cannot be correctly estimated.

Multiple Plausible Trajectories Our prediction produces a number of plausible trajectories that conform to the testing scene. Trifurcated trajectories in Street I, Department store I and II; bifurcated trajectories in Mall I and Parking lot; and multiple directions of trajectories in Costco and Mall III.

Occluded Space Discovery The space occluded by foreground objects is discovered by the predicted trajectories. The space inside of the shop and behind the person in Figure 1; the space occluded by moving persons in Campus I and II; the space behind the cars in Bus stop; the space inside of the shop in Mall I; the space inside the cloth in Department store II and III.

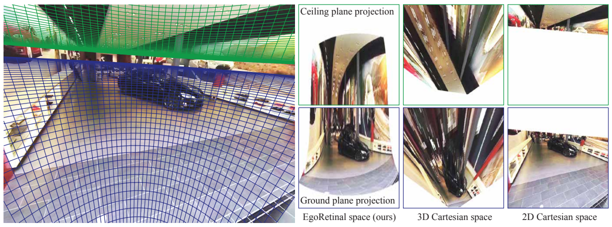
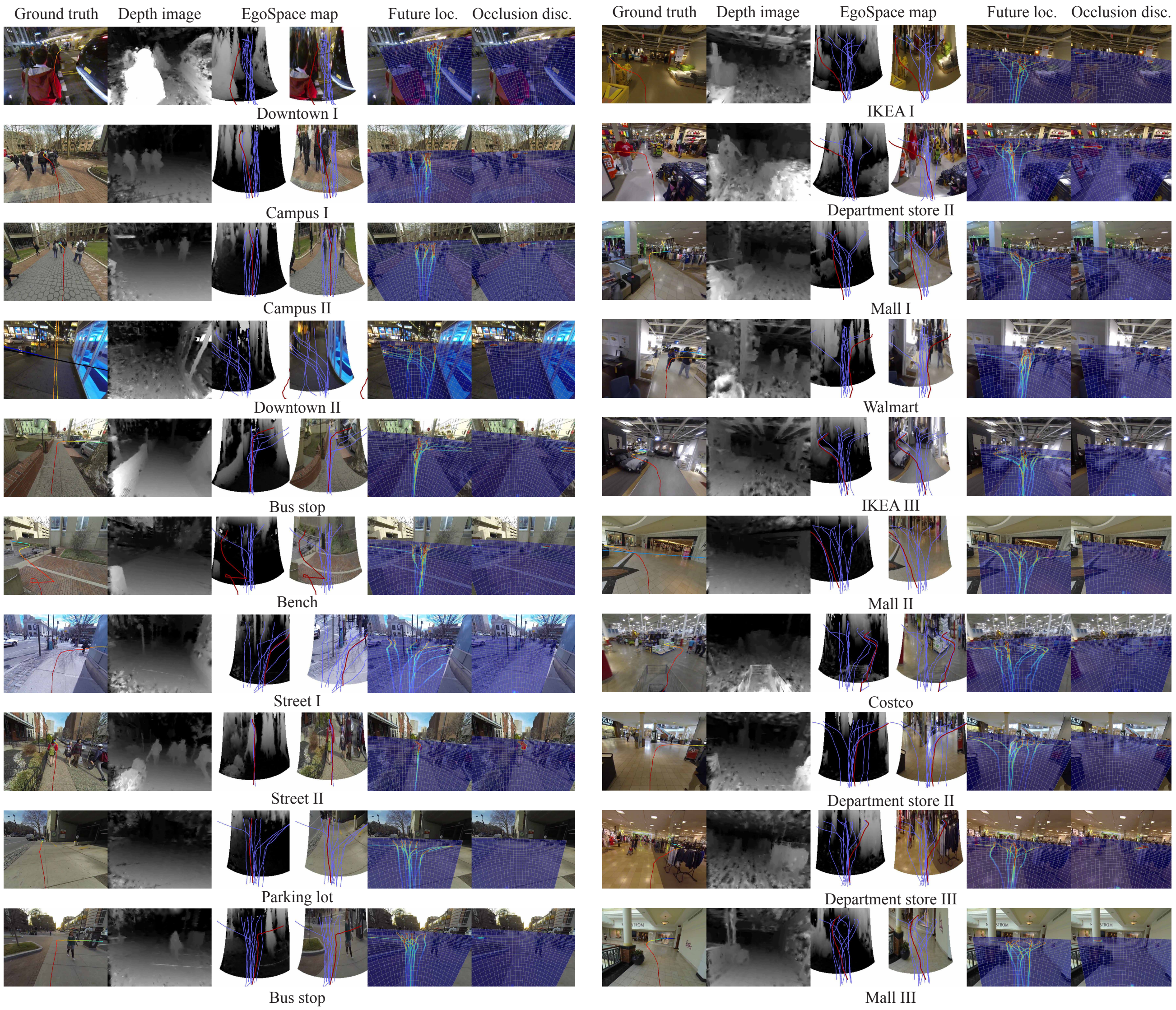


Figure 6. Our EgoRetinal map is beneficial to encode a configuration space above the vanishing line. Comparing to other representations, the EgoRetinal map continually links two different overhead views (ground plane and ceiling projections).

7. Discussion

In this paper, we present a method to predict ego-motion and occluded space by foreground objects from egocentric stereo images. An EgoRetinal map that encodes a likelihood of occlusion and its semantics is used to represent a



(a) Outdoor scene

(b) Indoor scene

Figure 5. Given an input RGBD image (the first and second column), we predict a set of plausible trajectories of ego-motion (the forth column) and discover the occluded space (the fifth column) using the EgoRetinal map (the third column: predicted purple trajectories and ground truth red trajectory). The first column shows an image with ground truth trajectory of ego-motion measured by 3D reconstruction of a first person camera (time is color-coded). For more scene description, see Section 6.2.

scene around a camera wearer. We associate a trajectory with the EgoRetinal map to predict a set of plausible trajectories. The trajectories that retrieved via a convolutional neural network are refined to conform with a testing scene. The occluded space is detected by measuring how often the predicted trajectories invade the occluded space.

References

[1] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *ICCV*, 2015. 2

[2] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In *CVPR*, 2014. 2

[3] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *ECCV*, 2008. 2

[4] I. Arev, H. S. Park, Y. Sheikh, J. K. Hodgins, and A. Shamir. Automatic editing of footage from multiple social cameras. *SIGGRAPH*, 2014. 2

[5] H. Choset, K. M. Lynch, S. Hutchinson, G. Kantor, W. Burgard, L. E. Kavraki, and S. Thrun. *Principles of Robot Mo-*

tion: Theory, Algorithms, and Implementations. MIT Press, 2005. 3

[6] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011. 2

[7] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interaction: A first-person perspective. In *CVPR*, 2012. 2

[8] A. Fathi and J. M. Rehg. Modeling actions through state changes. In *CVPR*, 2013. 2

[9] M. A. Fischler and R. C. Bolles. Modeling and prediction of human behavior. *Communications of the ACM*, 1981. 5

[10] J. J. Gibson. *The Perception of the Visual World*. Boston: Houghton Mifflin, 1950. 2, 3

[11] H. Gong, J. Sim, M. Likhachev, and J. Shi. Multi-hypothesis motion planning for visual object tracking. In *ICCV*, 2011. 2, 4

[12] E. T. Hall. A system for the notation of proxemic behaviour. *American Anthropologist*, 1963. 2

[13] D. Helbing and P. Molnár. Social force model for pedestrian dynamics. *Physics Review*, 1995. 2

[14] D. Jayaraman and K. Grauman. Learning image representations tied to ego-motion. In *ICCV*, 2015. 2

[15] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011. 2

[16] K. M. Kitani, B. Ziebart, J. D. Bagnell, and M. Hebert. Activity forecasting. In *ECCV*, 2012. 2, 6

[17] J. Kopf, M. Cohen, and R. Szeliski. First person hyperlapse videos. *SIGGRAPH*, 2014. 2

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 4

[19] H. Kurniawati, Y. Du, D. Hsu, and W. S. Lee. Motion planning under uncertainty for robotic tasks with long time horizons. In *Robotics Research*, 2009. 2

[20] R. Lee, D. H. Wolpert, S. Backhaus, R. Bent, J. Bono, and B. Tracey. Modeling humans as reinforcement learners: How to predict human behavior in multi-stage games. In *NIPS*, 2011. 2

[21] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *CVPR*, 2012. 2

[22] V. Lepetit, F. Moreno-Noguer, and P. Fua. Epnp: An accurate o(n) solution to the pnp problem. *IJCV*, 2009. 1, 5

[23] S. Levine and V. Koltun. Continuous inverse optimal control with locally optimal examples. In *ICML*, 2012. 2

[24] C. Li and K. M. Kitani. Pixel-level hand detection for egocentric videos. In *CVPR*, 2013. 2

[25] Y. Li, A. Fathi, and J. M. Rehg. Learning to predict gaze in egocentric video. In *ICCV*, 2013. 2

[26] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 4, 5

[27] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009. 2

[28] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 2001. 6

[29] H. S. Park, E. Jain, and Y. Shiekh. 3D social saliency from head-mounted cameras. In *NIPS*, 2012. 2

[30] H. S. Park and J. Shi. Social saliency prediction. In *CVPR*, 2015. 2

[31] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. Youll never walk alone: Modeling social behavior for multi-target tracking. In *ICCV*, 2009. 2

[32] A. Pentland and A. Lin. Modeling and prediction of human behavior. *Neural Computation*, 1995. 2

[33] J. Pineau and G. J. Gordon. Pomdp planning for robust robot control. In *Robotics Research*, 2007. 2

[34] H. Pirsiavash and D. Ramanan. Recognizing activities of daily living in first-person camera views. In *CVPR*, 2012. 2

[35] S. Ragi and E. K. P. Chong. Uav path planning in a dynamic environment via partially observable markov decision process. In *IEEE Transactions on Aerospace and Electronics Systems*, 2013. 2

[36] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me. In *CVPR*, 2013. 2

[37] K. K. Singh, K. Fatahalian, and A. A. Efros. Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *WACV*, 2016. 2

[38] T. Vu, C. Olsson, I. Laptev, A. Oliva, and J. Sivic. Predicting actions from static scenes. In *ECCV*, 2014. 2

[39] B. Xiong and K. Grauman. Detecting snap points in egocentric video with a web photo prior. In *ECCV*, 2014. 2

[40] B. Ziebart, A. Maas, J. Bagnell, and A. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008. 2