You2Me: Inferring Body Pose in Egocentric Video via First and Second Person Interactions

by

Evonne Ng

evonne_ng@utexas.edu

Supervised by:
Dr. Kristen Grauman

Department of Computer Science

Abstract

The body pose of a person wearing a camera is of great interest for applications in augmented reality, healthcare, and robotics, yet much of the person's body is out of view for a typical wearable camera. We propose a learning-based approach to estimate the camera wearer's 3D body pose from egocentric video sequences. Our key insight is to leverage interactions with another person—whose body pose we *can* directly observe—as a signal inherently linked to the body pose of the first-person subject.

We show that since interactions between individuals often induce a well-ordered series of back-and-forth responses, it is possible to learn a temporal model of the interlinked poses even though one party is largely out of view. We demonstrate our idea on a variety of domains with dyadic interaction and show the substantial impact on egocentric body pose estimation, which improves the state of the art.

Acknowledgments

I am extremely grateful to have had Dr. Kristen Grauman as my advisor. I would like to thank her for her mentorship and the generous support and dedication she provided that made this work possible.

I would like to thank Hanbyul Joo and Donglai Xiang for helping me in the collection of the datasets, and Hao Jiang for helpful discussions.

Contents

List of Figures

1 Introduction

Wearable cameras are becoming an increasingly viable platform for entertainment and productivity. In augmented reality (AR), wearable headsets will let users blend useful information from the virtual world together with their real first-person visual experience to access information in a timely manner or interact with games. In healthcare, wearables can open up new forms of remote therapy for rehabilitating patients trying to improve their body's physical function in their own home. In robotics, wearables could simplify video-based learning from demonstration.

In all such cases and many more, the camera receives a first-person or "egocentric" perspective of the surrounding visual world. A vision system analyzing the egocentric video stream should not only extract high-level information about the visible surroundings (object, scenes, events), but also the current state of the person wearing the camera. In particular, the *body pose* of the camera wearer is of great interest, since it reveals his/her physical activity, postures, and gestures. Unfortunately, the camera wearer's body is often largely out of the camera's field of view. While this makes state-of-the-art third-person pose methods poorly suited [10, 23, 31, 39, 50, 54], recent work suggests that an ego-video stream nonetheless offers implicit cues for first-person body pose [25, 59]. However, prior work restricts the task to static environments devoid of inter-person interactions, forcing the algorithms to rely on low-level cues like apparent camera motion or coarse scene layout.

Our idea is to facilitate the recovery of 3D body pose for the camera wearer (or "ego-pose" for short) by paying attention to the *interactions* between the first and second person as observed in a first-person video stream.[1] Inter-person interactions are extremely common and occupy a large part of any individual's day-to-day activ-

[1]Throughout, we use "second person" to refer to the person the camera wearer is currently interacting with; if the wearer is "I", the interactee or partner in the interaction is "you".

Figure 1: **Daily Inter-person Interactions** – Inter-person interactions are common in daily activity and offer rich signals for perception. Our work considers how interactions viewed from a first-person wearable camera can facilitate egocentric 3D body pose estimation.

ities. As is well-known in cognitive science [8, 41, 52], human body pose is largely influenced by an inherent synchronization between interacting individuals. For instance, a person who sees someone reaching out their hand for a handshake will most likely respond by also reaching out their hand; a person animatedly gesturing while telling a story may see their interacting partner nod in response; children playing may interact closely with their body motions. See Figure 1.

This motivates us to build a model which accounts for both the action and reaction dynamics inherent within a dyadic interaction sequence while predicting a camera wearer's pose. To that end, we introduce "You2Me": an approach to ego-pose estimation that explicitly captures the interplay between the first and second person body poses. Our model uses a recurrent neural network to incorporate cues from the *observed* second-person pose together with the camera motion and scene appearance to infer the *latent* ego-pose across the entire video sequence. See Figure 2.

We validate our You2Me ego-pose approach on two forms of ground-truth capture—from Kinect sensors and a Panoptic Studio [27]—on video data spanning 10 subjects and several interaction domains (conversation, sports, hand games, and ball tossing). Our results demonstrate that even though the first-person's body is largely out of view, the inferred second-person pose provides a useful prior on likely interactions, significantly boosting the estimates possible with the ego-camera motion and scene context alone. Furthermore, our You2Me approach outperforms the state-of-the-art approach for ego-pose as well as a current standard deep third-person pose method when adapted to our setting.

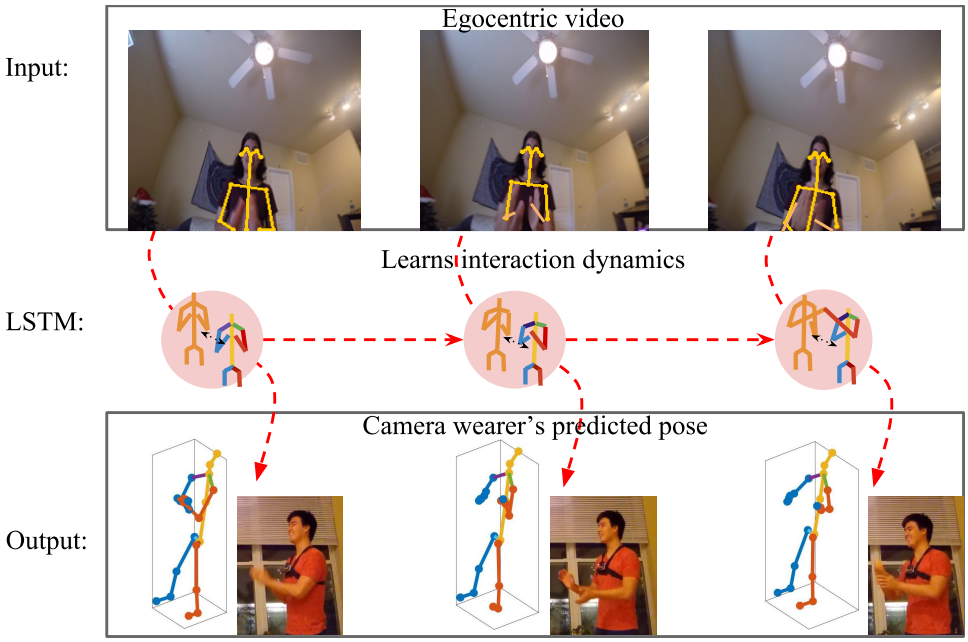

Figure 2: **Concept Overview** – Our goal is to infer the full 3D body pose sequence of a person from their egocentric video, captured by a single chest-mounted camera. We propose an LSTM that focuses on human-human interaction dynamics to predict the wearer's pose by taking into account the interactee's pose, which is visible from the ego-view. The figure shows the input video with the interactee's (second-person) pose highlighted, and the output 3D joint predictions of the wearer's pose with corresponding pictures of the camera wearer. Note that our approach sees only the egocentric video (top); it does not see the bottom row of images showing the "first person" behind the camera.

2 Related Work

We focus on the relatively unexplored problem of predicting the invisible egocentric full-body human pose from a single egocentric video stream. To contextualize our idea, we review related works involving first- and third-person body pose estimation, egocentric video analysis, and social signals in first-person video.

2.1 Third-person body pose and interactions

There is extensive literature on human body pose estimation from the traditional third-person viewpoint, where the person is entirely visible [45]. Recent approaches explore novel CNN-based methods, which have substantially improved the detection of *visible* body poses in images and video [10, 11, 17, 23, 31, 50, 55, 60]. Our approach instead estimates the largely "invisible" first-person pose. Multi-person pose tracking work investigates structure in human motion and inter-person interactions in order to limit the potential pose trajectories [10, 24]. Beyond body pose, there is a growing interest in modeling human-human interactions [22, 34, 49] to predict pedestrian trajectories [1, 2, 37], analyze social behavior and group activities [6, 14, 22, 37, 51], and understand human-object interactions [12, 18, 53]. Our method also capitalizes on the structure in inter-person interactions. However, whereas these existing methods assume that all people are fully within the view of the camera, our approach addresses interactions between an individual in-view and an individual *out-of-view*, i.e., the camera wearer.

2.2 Egocentric video analysis

Recent egocentric vision work focuses primarily on recognizing objects [13], activities [16, 32, 33, 38, 40, 43, 44, 48, 58], visible hand and arm poses [7, 28, 29,

42], eye gaze [30], or anticipating future camera trajectories [9, 35]. In contrast, we explore 3D pose estimation for the camera wearer's full body, and unlike any of the above methods, we show that the inferred body pose of another individual during an interaction directly benefits the pose estimates.

2.3 First-person body pose from video

Egocentric 3D full body pose estimation has received only limited attention [25, 47, 59]. The first attempt to the problem is the geometry-based "inside-out mocap" approach [47], which uses structure from motion (SfM) to reconstruct the 3D location of 16 body mounted cameras placed on a person's joints. In contrast, we propose a learning-based solution, and it requires only a single chest-mounted camera, which makes it more suitable and comfortable for daily activity.

More recently, two methods based on monocular first-person video have been proposed [25, 59]. The method in [25] infers the poses of a camera wearer by using both homographies and static visual cues to optimize an implicit motion graph. The method in [59] uses a humanoid simulator in a control-based approach to recover the sequence of actions affecting pose, and it is evaluated quantitatively only on synthetic sequences. Whereas both prior learning-based methods focus on sweeping motions that induce notable camera movements (like bending, sitting, walking, running), our approach improves the prediction of upper-body joint locations during sequences when the camera remains relatively still (like handshakes and other conversational gestures). Furthermore, unlike [59], our method does not require a simulator and does all its learning directly from video accompanied by ground truth ego-poses. Most importantly, unlike any of the existing methods [25, 47, 59], our approach discovers the connection between the dynamics in inter-person interactions and egocentric body poses.

2.4 Social signals in first-person video

Being person-centric by definition, first-person video is naturally a rich source of social information. Prior work exploring social signals focuses on detecting social groups [3, 4, 15] and mutual gaze [56, 57] or shared gaze [36] from ego-video. More relevant to our work, the activity recognition method of [58] uses paired egocentric

videos to learn gestures and micro-actions in dyadic interactions. That approach captures the correlations among inter-person actions (*e.g.*, pointing, passing item, receiving item, etc.) in two synchronized novel video clips to better classify them. However, whereas [58] requires two egocentric videos at test time, our approach relies only on a single ego-video. While eliminating the second camera introduces new technical challenges (since we cannot view both the action and response), it offers greater flexibility as we do not have to synchronize footage or require another individual to wear a camera. Furthermore, our approach infers body pose, whereas [58] classifies clips into a fixed vocabulary of seven actions.

3 Our Approach

The goal is to take a single first-person video as input, and estimate the camera wearer's 3D body pose sequence as output. Our main insight is to leverage not only the appearance and motion evident in the first-person video, but also an estimate of the second-person's body poses.

In this section, we present a recurrent neural network model that utilizes first- and second-person features—both extracted from monocular egocentric video—to predict the 3D joints of the camera wearer. After defining the pose encoding (Sec 3.1), we define the three inputs to our network (Sec 3.2 to 3.4), followed by the recurrent long short-term memory (LSTM) network that uses them to make sequential predictions for a video (Sec 3.5).

3.1 Problem formulation

Given N video frames from a chest-mounted camera, we estimate a corresponding sequence of N 3D human poses. Each output pose $p_t \in \mathbb{R}^{3J}$ is a stick figure skeleton of 3D points consisting of J joint positions for the predicted body pose of the camera wearer at frame t. Note that our goal is to infer pose as opposed to classifying the action.

Each predicted 3D body joint is positioned in a person-centric coordinate system with its origin at the camera on the wearer's chest. The first axis is parallel to the ground and points towards the direction in which the wearer is facing. The second axis is parallel to the ground and lies along the same plane as the shoulder line. The third axis is perpendicular to the ground plane. To account for people of varying sizes, we normalize each skeleton for scale based on the shoulder width of the individual.

3.2 Dynamic first-person motion features

As shown in [25], motion patterns observed from the first-person camera offer a strong scene-independent cue about the camera wearer's body articulations, despite the limbs themselves largely being out of the field of view. For example, a sudden drop in elevation can indicate movement towards a sitting posture, or a counterclockwise rotation can indicate shoulders tilting to the left.

To capture these patterns, we construct scene-invariant dynamic features by extracting a sequence of homographies between each successive video frame, following [25]. While a homography is only strictly scene invariant when the camera is purely rotating, the egocentric camera translates very little between successive frames when the frame rate is high. These homographies facilitate generalization to novel environments, since the motion signals are independent of the exact appearance of the scene.

We estimate the homography from flow correspondences by solving a homogeneous linear equation via SVD [21]. Each element in the resulting 3×3 homography matrix is then normalized by the top-left corner element. The stack of normalized homographies over a given duration is used to represent the global camera movement within the interval. For frame f_t at timestep t in a given video, the motion representation is constructed by calculating the homographies between successive frames within the interval $[f_{t-15}, f_t]$. We then vectorize the homographies and combine them into a $m_t \in \mathbb{R}^{135}$ vector, which represents a half-second interval of camera movements preceding frame f_t (for 30 fps video).

3.3 Static first-person scene features

While the dynamic features reveal important cues for sweeping actions that induce notable camera movements, such as running, walking, or sitting and standing, they are more ambiguous for sequences with little motion in the egocentric video. To account for this, our second feature input attends to the appearance of the surrounding scene. In everyday life, many static scene structures are heavily associated with certain poses. For example, if the camera wearer leans forward to touch his/her toes, the egocentric camera may see the floor; if the camera wearer stands while looking at a computer monitor, the egocentric camera will see a different image than if the camera

wearer sits while looking at the same monitor. As with the dynamic features above, the surrounding scene provides cues about ego-pose without the camera wearer's body being visible.

To obtain static first-person scene features, we use a ResNet-152 model pre-trained on ImageNet. Dropping the last fully connected layer on the pre-trained model, we treat the rest of the ResNet-152 as a fixed feature extractor for video frames. Given frame f_t, we run the image through the modified ResNet-152, which outputs $s_t \in \mathbb{R}^{2048}$. Whereas the ego-pose method of [25] relies on a standing vs. sitting image classifier to capture static context, we find our full visual encoding of the scene contributes to more accurate pose learning. Note that this feature by default also captures elements of the second-person pose; however, without extracting the pose explicitly it would be much more data inefficient to learn it simply from ResNet features, as we will see in results.

3.4 Second-person body pose interaction features

Our third and most important input consists of the "second-person" pose of the person with whom the camera wearer is interacting. Whereas both the dynamic and static features help capture poses that come from larger common actions, we propose to incorporate second-person pose to explicitly account for the *interaction dynamics* that influence gestures and micro-actions performed in sequence between two people engaged in an interaction.

In human-human interactions, there is a great deal of symbiosis between both actors. Specific actions solicit certain reactions, which in turn influence the body pose of the individual. For example, if we see an individual windup to throw a ball, our natural response is to raise our arms to catch or block the ball. Or, more subtly, if we see a person turn slightly to look at a passerby, we may turn to follow their gaze. By understanding this dynamic, we can gather important ego-pose information for the camera wearer by simply observing the visible pose of the person with whom he/she interacts.

Thus, our third feature records the interactee's inferred pose. Still using the egocentric video, we estimate the pose of the interactee in each frame. Here we can leverage recent successes for pose estimation from a third-person perspective: unlike the camera wearer, the second person *is* visible, i.e., the ego-camera footage

gives a third-person view of the interactee. In particular, we use OpenPose [10] to infer interactee poses due to its efficiency and accuracy, though other third-person methods could also be employed. OpenPose provides real-time multi-person keypoint detection: given a stack of frames, it returns a corresponding stack of 25 2D keypoint joint estimations. For each frame f_t, we flatten the output 25 keypoint estimates into a vector $o_t \in \mathbb{R}^{50}$ (denoted o_t for "other"). Note that our learning approach is flexible to the exact encodings of the ego- and second-person poses, i.e., it is fine for the second-person pose estimate to be 2D keypoints while the ego-pose is expressed in 3D. As we will see in experiments, the second-person pose is crucial in improving ego-pose prediction.

To handle scenarios when the interactee is occluded or moves out of view, we simply set o_t to be the zero vector to represent a missing second-person skeleton. While the pose of the interactee provides crucial information to the LSTM (defined next), we find that the priors learned from the LSTM are strong enough to continue to predict accurate poses for a short period ($\lessapprox 15$ frames) with no view of the interactee (cf. Sec. 5). Figure 3 illustrates the complete set of features for some ego-frames.

3.5 Recurrent neural network for pose inference

All three video-based cues defined above serve as input to a recurrent neural network to perform pose estimation for the full sequence. In particular, we define a Long Short-Term Memory (LSTM) network [19, 20] for our task. The LSTM learns the current state of the camera wearer, scene, and interactee, and uses this encoding to predict the camera wearer's future poses. The LSTM's hidden state captures the sequential patterns of linked body poses that result from inter-person back-and-forth responses.

While the LSTM can be trained to perform regression on the real-valued coordinates of the body pose, we found a classification task to train more robustly (as often reported in the literature). Hence, we first quantize the space of training body poses into a large number ($K = 500$) of fine-grained poses using K-means. Now the task is to map to the closest possible quantized pose at each time step. We visualize the granularity of differences between each cluster center in Figure 4. With 500 pose clusters, we get a good diversity of poses, enough to reasonably capture all possible poses in the training set. Additionally, the poses are fine-grained enough to accurately

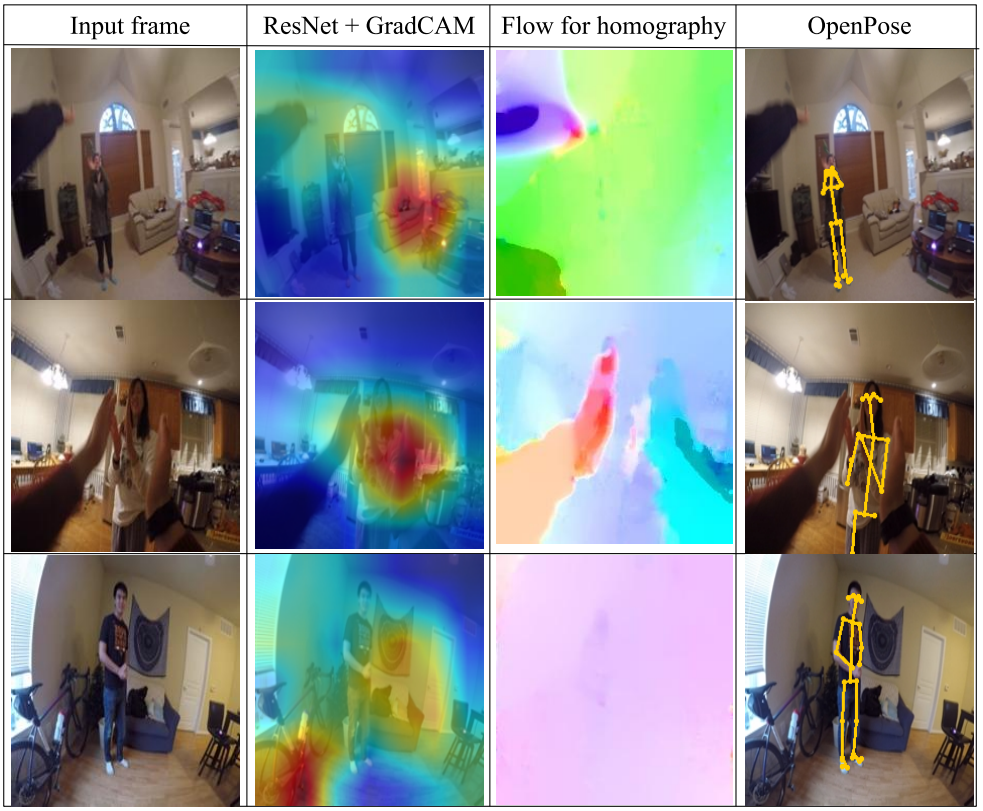

Figure 3: **Features Extracted for You2Me** – Visualization of features extracted from ego-video frames. The ResNet Grad-CAM [46] heatmaps suggest that when a person is further away, the focus is on static objects in the room (couch, bike, wall rug) which help capture coarse posture, but when the interactee is closer, the focus is more on the person, which influences finer details. While the flow/homography does especially well capturing motion from the camera wearer's hands, many sequences lack global motion and produce flows similar to the bottom row example. OpenPose generates a 2D representation of the interactee's pose even with slight occlusions.

capture smaller movements of the arms and the legs (gesturing, micro-actions), or intermediate poses in larger movements (swinging, walking, sitting).

Given a hidden state dimension of D, the hidden state vector $h_t \in \mathbb{R}^D$ of the LSTM at time t captures the cumulative latent representation of the camera wearer's pose at that instant in the video. For each frame f_t, we extract the homography matrix m_t, the ResNet-152 scene feature vector s_t, and the second-person joint position vector o_t. To provide a more compact representation of the scene to the LSTM (useful to

Figure 4: 120 randomly chosen pose clusters out of 500 possible clusters.

conserve GPU memory), we project s_t to a lower-dimensional embedding $x_t \in \mathbb{R}^E$:

$$x_t = \phi_x(s_t; W_x), \tag{3.1}$$

where W_x is of size $E \times 2048$ and consists of the embedding weights for $\phi_x(.)$. The embedding is then passed through a batch normalization layer.

The LSTM uses the wearer's pose in the previous frame p_{t-1} as input for the current frame. Let p_{t-1} be a K-dimensional one-hot vector indicating the pose for the camera wearer at the previous frame $t-1$. We learn a linear embedding for the pose indicator to map it to vector z_t:

$$z_t = \phi_z(p_{t-1}; W_z), \tag{3.2}$$

where W_z is of size $E \times K$ and consists of the learned embedding weights for $\phi_z(.)$.

All the features are concatenated (indicated by operation $\oplus$) into a single vector $b_t \in \mathbb{R}^{135+50+2E}$:

$$b_t = m_t \oplus o_t \oplus x_t \oplus z_t, \tag{3.3}$$

which is then used as input to the LSTM cell for the corresponding prediction at time t. This introduces the following recurrence for the hidden state vector:

$$h_t = \text{LSTM}(h_{t-1}, b_t; \theta_l), \tag{3.4}$$

where θ_l denotes the LSTM parameters.

We define the loss for the network as the cross entropy loss across an entire sequence for predicting the correct (quantized) pose in each frame. Specifically, the loss $\mathcal{L}$ for a video of length N is:

$$\mathcal{L}(W_x, W_z, W_p, \theta_l) = -\sum_{t}^{N} \log(\sigma_P(W_p h_t)), \tag{3.5}$$

where $\sigma_P(\cdot)$ is the softmax probability of the correct pose "class", and W_p is the linear classifier layer of dimension $K \times D$. Recall that the quantization is fine-grained ($K = 500$ pose clusters) such that this estimate is quite specific; on average the nearest quantized pose in the codebook is just 0.35 cm away per joint (see Figure 4). The inferred pose ID at time t (i.e., the arg max over the pose posteriors at that

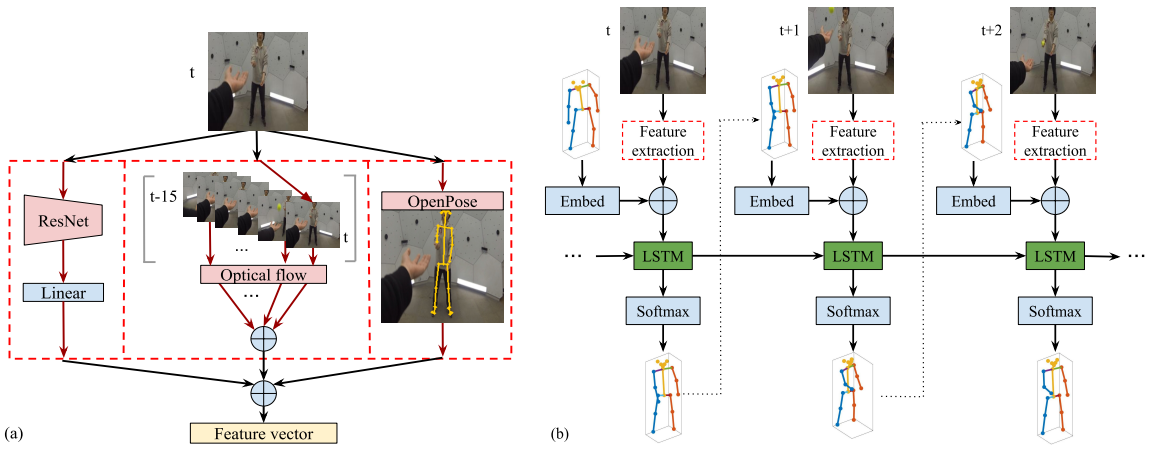

Figure 5: **You2Me Network Architecture** – Network architecture for our You2Me approach. (a) For each video frame, we extract three features. ResNet provides static visual cues about the scene. Stacked homographies for the past 15 frames provide motion cues for the ego-camera. Finally, we extract the inferred 2D pose of the visible interactee with OpenPose [10]. All three features are concatenated ($\oplus$) and fed into the LSTM. (b) illustrates our LSTM, which takes as input the feature vector from (a) and an embedding of the camera wearer's pose estimated from the previous frame. Outputs from the LSTM produce ego-pose predictions, assigning one of the 500 possible quantized body poses to each frame.

timestep) is taken as the input for z_{t+1} for the subsequent frame.

At test time, we use the trained LSTM model to predict the sequence of poses. From time $t-1$ to t, we use the predicted cluster $\hat{p}_{t-1}$ from the previous LSTM cell in Eq. 3.2. Figure 5 overviews the LSTM.

4 You2Me Video Datasets

We present a first-person interaction dataset consisting of 42 two-minute sequences from one-on-one interactions between 10 different individuals. We asked each individual (in turn) to wear a chest mounted GoPro camera and perform various interactive activities with another individual. We collect egocentric video captured by the camera, which is then synchronized with the body-pose ground truth for both the camera wearer and the individual standing in front of the camera. The dataset captures four classes of activities: *hand games*, *tossing and catching*, *sports*, and *conversation*. The classes are broad enough such that intra-class variation exists. For example, the sports category contains instances of (reenacted) basketball, tennis, boxing, etc.; the conversation category contains instances of individuals playing charades, selling a product, negotiating, etc. In about 50% of the frames, no first-person body parts are visible. To ensure that our approach is generalizable, we employ two methods of capture, as detailed next.

4.1 Panoptic Studio capture

Our first capture mode uses a Panoptic Studio dome, following [27]. The studio capture consists of 14 sequences recorded in 1920×1080 resolution at 30 fps using the GoPro Hero3 chest mounted camera on the medium field of view setting. The ground truth skeletons of the camera wearer and the individual in front of the camera are then reconstructed at 30 fps, matching the frame rate at which we extract the video. Each skeleton is parameterized by $J = 19$ 3D joint positions obtained using the method of [27]. Capturing video in the dome offers extremely accurate ground truth, at the expense of a more constrained background environment. A total of six participants of different height, body shape, and gender enacted sequences from each of the four activity classes.

4.2 Kinect capture

Our second capture mode uses Kinect sensors for ground truth poses. The Kinect capture consists of 28 sequences also recorded in 1920×1080 resolution at 30 fps. We use the GoPro Hero4 chest mounted camera on the wide field of view setting, and both people's ground truth skeleton poses are captured at 30 fps using the Kinect V2 sensor. The pose is represented by $J = 25$ 3D joint positions defined in the MS Kinect SDK. Given the greater mobility of the Kinect in contrast to the Panoptic Studio, we ask four participants to enact sequences from each of the activity classes in various places such as offices, labs, and apartment rooms. The videos from this dataset are taken in unconstrained environments but are all indoors due to the limitations of the Kinect V2 sensor. While Kinect-sensed ground truth poses are more noisy than those captured in the Panoptic Studio, prior work demonstrates that overall the Kinect poses are very well aligned with human judgments of pose [25].

We stress that our method uses *only the egocentric camera video as input* at test time for both datasets. Further, we emphasize that no existing dataset is suitable for our task. Existing pose detection and tracking datasets (*e.g.*, [5, 26]) are captured in the third-person viewpoint. Existing egocentric datasets are either limited to visible hands and arms [29, 38], contain only single-person sequences [5, 25, 26], consist of synthetic test data [59], or lack body-pose joint labels [58]. All our data will be made publicly available.

5 Experiments

We evaluate our approach on both the Panoptic Studio and Kinect captures. For both sets, each video clip contains a single execution of an activity. Our method is trained and tested in a activity-agnostic setting: the training and test sets are split such that each set contains roughly an equal number of sequences from each activity domain (conversation, sports, etc.). For the Panoptic Studio, we train on 7 sequences and test on 7. For the Kinect set, we train on 18 sequences and test on 10 that are recorded at locations not seen in the training set. For both, we ensure that the people appearing in test clips do *not* appear in the training set.

5.1 Implementation Details

We generate training data by creating sliding windows of size 512 frames with an overlap of 32 frames for each sequence in the training set. For the LSTM, we use an embedding dimension of $E = 256$ and a fixed hidden state dimension of $D = 512$. Batch size is 32 and learning rate is 0.001 for the first 10 epochs then decreased to 0.0001. The model was trained on a single GPU with PyTorch. In initial experiments, we found results relatively insensitive to values of K between 300 and 600, and fixed $K = 500$ for all results. Run-time for our method averages 36 fps.

Furthermore, rather than feeding in raw video to the LSTM, we first perform some preprocessing on the images. Each raw video is extracted at a frame rate of 30 fps. The frames are then resized to 224 x 224 x 3 images and normalized with a mean of $[0.485, 0.456, 0.406]$ and standard deviation of $[0.229, 0.24, 0.225]$ across the three channels. A stack of these preprocessed images serves as input to the LSTM.

5.2 Baselines

We compare to the following methods:

- **Ego-pose motion graph (*MotionGraph*) [25]**: the current state-of-the art method for predicting body pose from real egocentric video [25]. We use the authors' code[1] and retrain their model on our dataset. This method also outputs quantized poses; we use the identical 500 pose clusters as for our method.
- **Third-person pose deconv network (*DeconvNet*) [54]**: We adapt the human pose estimation baseline of [54] to our task.[2] Their approach adds deconvolutional layers to ResNet, and achieves the state-of-the-art on the 2017 COCO keypoint challenge. We use the same network structure presented in the baseline, but retrain it on our egocentric dataset. While this network is intended for detecting visible poses in third-person images, it is useful to gauge how well an extremely effective off-the-shelf deep pose method can learn from ego-video.
- **Ours without pose information (*Ours w/o* o_t)**: This is a simplified version of our model in which we do not feed the second-person 2D joints to the LSTM. The remaining network is unchanged and takes as input the extracted image features and homographies. This ablation isolates the impact of modeling interactee poses versus all remaining design choices in our method.
- **Always standing (*Stand*)** and **Always sitting (*Sit*)**: a simple guessing method (stronger than a truly random guess) that exploits the prior that most poses are somewhere near a standing or a sitting pose. The standing and sitting poses are averaged over the training sequences.

5.3 Evaluation Metric

We rotate each skeleton so the shoulder is parallel to the yz plane and the body center is at the origin, then calculate error as the Euclidean distance between the predicted 3D joints and the ground truth, averaged over the sequence and scaled to

[1] `http://www.hao-jiang.net/code/egopose/ego_pose_code.tar.gz`
[2] `https://github.com/leoxiaobin/pose.pytorch`

	Kinect			Panoptic		
	Upp	Bot	All	Upp	Bot	All
Ours	**17.0 (1.3)**	**14.9 (1.2)**	**15.5 (1.2)**	**10.2 (3.1)**	**14.7 (4.5)**	**11.9 (3.5)**
Ours w/o o_t	25.7 (2.0)	18.9 (2.3)	22.0 (1.9)	16.8 (2.4)	20.5 (3.3)	18.2 (1.7)
MotionGraph [25]	24.4 (2.4)	15.7 (1.3)	21.2 (1.9)	11.9 (2.8)	20.7 (3.0)	15.2 (2.8)
DeconvNet [54]	26.0 (1.2)	20.3 (0.8)	23.3 (1.2)	18.3 (1.2)	21.2 (3.5)	19.4 (1.8)
Stand	27.8 (3.5)	23.1 (1.6)	25.4 (2.1)	10.6 (4.4)	18.5 (8.2)	13.5 (5.5)
Sit	21.8 (1.0)	43.3 (1.8)	28.5 (1.2)	17.3 (2.1)	28.9 (1.4)	21.6 (1.4)

Table 5.1: Average joint error (cm) for all methods on the two dataset captures. Our approach is stronger than the existing methods, and the second-person pose is crucial to its performance. Standard errors for all methods on the two dataset captures are displayed in parentheses. The standard error of our approach is comparable with the other methods

centimeters (cm) based on a reference shoulder distance of 30 cm. Note that the predicted joints are always some cluster center, whereas the ground truth is the exact pose (non-quantized); so, even if we predict the best discrete pose nearest to the ground truth, there will be non-zero error.

5.4 Results

Table 5.1 shows that the proposed method consistently gives better results than all of the competing methods. We show errors averaged over all J joints, and separately for the upper body joints which have highest variance in everyday activity (head, elbow, wrists, hands) and the lower body joints (hips, knees, ankles, foot). Our approach outperforms **MotionGraph** [25] and **Ours w/o o_t.** This result supports our key technical novelty of modeling mutual pose interactions between the first and second person. Our method's improvement is even more significant in the upper body joints, which agrees with the fact that the most highly correlated inter-person poses occur with gestural motions of the head and arms. The results show that the information provided by the pose of the interactee is essential for deriving accurate body pose estimates for the camera wearer.

We find that our method's impact is greatest on the conversation sequences, and lowest on the sports sequences. This suggests that during conversation sequences which involve less global motion, second-person pose provides essential information for more accurate upper body ego-pose predictions. Sports sequences, on the other hand, often have the interactee moving out of view for long periods, explaining our

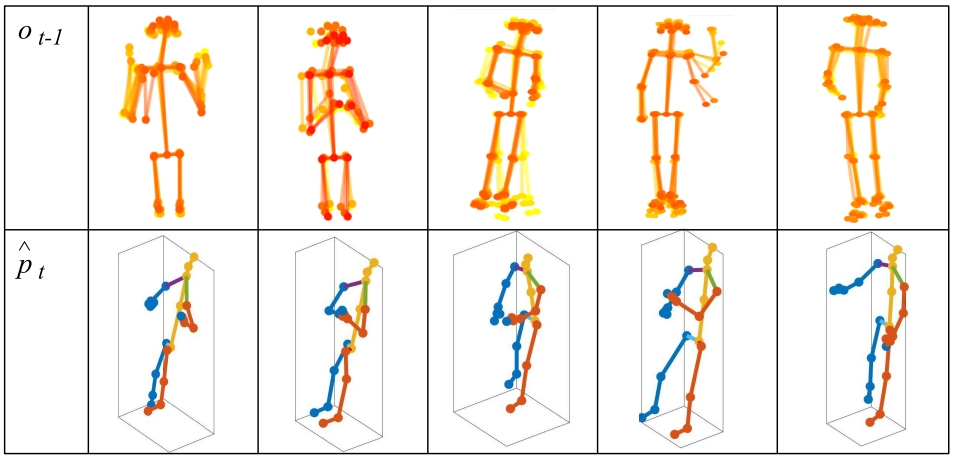

Figure 6: **Common Second Person Priors for Sample Pose Clusters** – Most common second-person 2D poses (top) seen immediately preceding a given predicted 3D pose cluster (bottom) for test sequences. You2Me captures useful interaction links like mutual reaches or tied conversation gestures.

method's lower degree of impact for sports.

While **Sit** and **Stand** offer a reasonable prior for most test frames, our method still makes significant gains on them, showing the ability to make more informed estimates on the limbs (e.g., 10 cm better on average for the upper body keypoints). **Sit** has a much larger lower body and overall error than any other method, which is in line with the distribution of the test data. Our method also outperforms **DeconvNet** [54], which suggests that approaches for detecting poses from a third-person point of view do not easily adapt to handle the first-person pose task.

Figure 6 shows examples of the linked poses our method benefits from. We display the second-person pose estimates immediately preceding various ego-pose estimates for cases where our method improves over the **Ours w/o** o_t baseline. Intuitively, gains happen for interactions with good body language links, such as mutually extending hands or smaller conversational gestures.

Figures 7 and 8 show example success and failure cases for our approach, respectively. In Figure 7, our method outperforms **MotionGraph** [25] in predicting upper body movements of the camera wearer, e.g., better capturing the swing of an arm before catching a ball or reaching out to grab an object during a conversation. The failures in Figure 8 show the importance of the second-person pose to our approach. Analyzing the frames with the highest errors, we find failure cases occur primarily when the camera wearer is crouched over, the camera is pointed towards the floor,

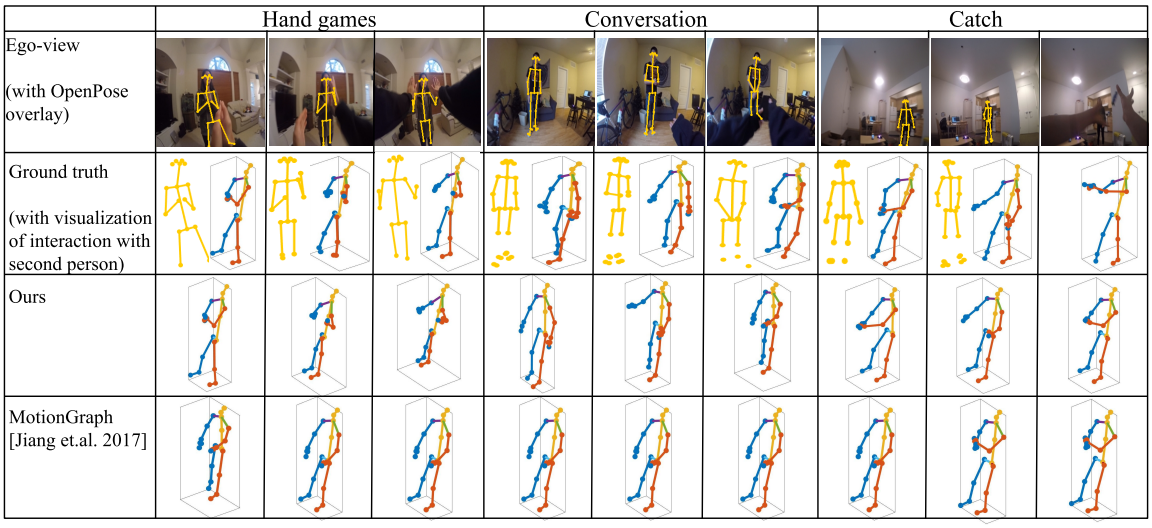

Figure 7: **Success Cases for You2Me Results** – Example inferred poses for three different activity domains trained in a domain-agnostic setting. Row 1: ego-video view with OpenPose overlay (input to our method is only the raw frame). Row 2: 3D ground truth poses in multicolor, displayed as interacting with the 2D OpenPose skeletons in yellow. Note: for ease of viewing, we show them side by side. Row 3: results from our approach. Row 4: MotionGraph [25] results. In the last column, the interactee is fully occluded in the ego-view, but our predicted pose is still accurate.

or the view of the interactee is obstructed. While our LSTM has enough priors to continue to accurately predict poses for a few frames without the interactee pose, absent second person poses over extended periods are detrimental. We also provide a supplemental video[3] demonstrating our approach on video sequences of various test subjects and capture locations.

We show examples of success cases across the four different action domains: *conversation, sports, hand games, and ball tossing*. In both the Kinect and Panoptic Studio captures, our method is able to perform well. Most notably, our approach is able to determine when the camera wearer is going to squat or sit, when they are raising their hand to receive or catch an item, and when they are gesturing as part of a conversation.

Consistent with the quantitative results provided by Figure 7, compared against the baselines, we notice a significant difference between our approach and the **MotionGraph** [25] . While our approach is able to detect when the camera wearer is

[3] `http://vision.cs.utexas.edu/projects/you2me/demo.mp4`

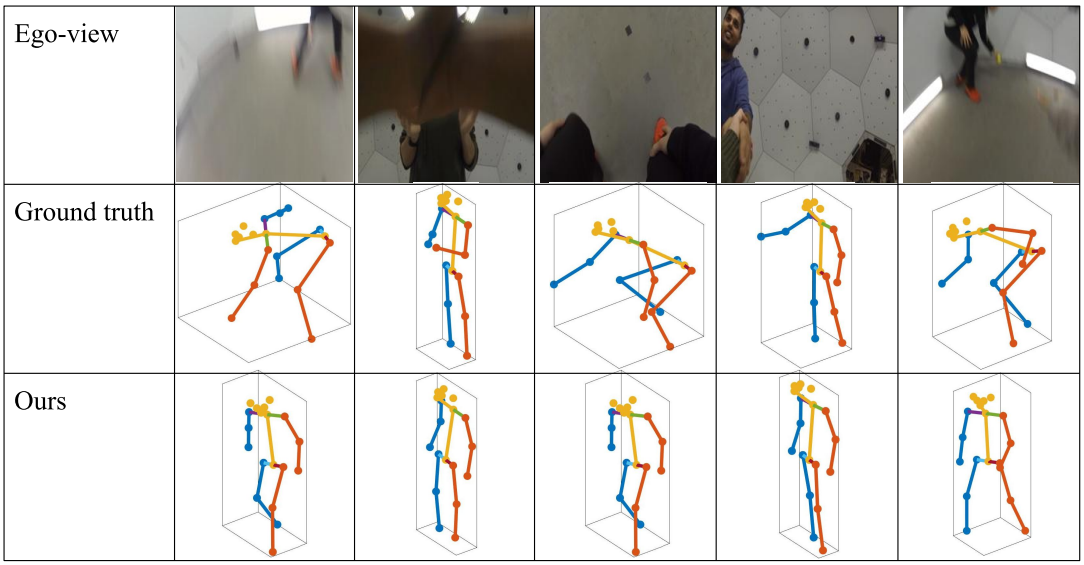

Figure 8: **Failure Cases for You2Me Results** – Example failure cases. Typical failure cases are when the ego-view points at the ground or at feet, lacking the interactee's pose for a long duration.

	Kinect			Panoptic		
	Upp	Bot	All	Upp	Bot	All
Ours	17.0	**14.9**	**15.5**	**10.2**	**14.7**	**11.9**
w/o x_t	**16.7**	16.3	16.1	10.7	15.3	12.4
w/o o_t	25.7	18.9	22.0	16.8	20.5	18.2
w/o both	20.9	17.7	19.4	17.4	19.4	18.1

Table 5.2: Ablation study to gauge the importance of the second-person pose features o_t and scene features x_t. Error in cm.

clapping as part of a hand game, the **MotionGraph** [25] fails to do so. Similarly, when we remove the OpenPose features, **Ours w/o** o_t is also unable to capture when a person's hand is raised. However, our approach is even able to detect when the camera wearer is returning a single handed clap or a double handed clap in the hand-game.

Table 5.2 shows an ablation study, where we add or remove features from our LSTM to quantify the impact of the second-person pose. Recall that o_t is the second-person pose and x_t is the ResNet scene feature. The results indicate that **Ours** and the **w/o** x_t model, which both use the second-person pose (OpenPose estimates), consistently outperform the **w/o** o_t and **w/o both** models that lack the second-person pose estimate. Moreover, the results show that the addition of o_t most significantly

	Kinect			Panoptic		
	Upp	Bot	All	Upp	Bot	All
o_t	17.0	15.0	15.5	10.2	14.7	11.9
GT	**16.2**	**14.9**	**15.1**	**8.3**	**13.5**	**10.2**
Still	22.6	15.0	18.9	25.6	24.6	25.2
Zero	23.7	27.5	20.0	18.8	21.8	19.9
Random	19.5	17.7	18.0	22.3	17.6	18.9

Table 5.3: Effects of second-person pose source. Error in cm.

improves upper body predictions. The features of the interactee captured by the ResNet (**w/o** o_t) do not sufficiently capture the information encoded in the explicit pose estimate.

Table 5.3 analyzes to what extent the imperfect second-person pose estimates affect our results. First, we substitute in for o_t the ground truth (GT) skeleton of the interactee, i.e., the true pose for the second person as given by the Panoptic Studio or Kinect. We see that more accurate second-person poses can further improve results, though the margins are smaller than those separating our method from the baselines. Next, to confirm our network properly learns a correlative function between the interactee pose and the ego-pose, we feed *incorrect* values for o_t: either the average standing pose (Still), empty poses (Zero), or random poses from another sequence of another class (Random). In all cases, the network produces poorer results, showing that our method is indeed leveraging the true structure in interactions.

6 Conclusions

With the growing usage of wearable cameras across entertainment, healthcare, and gaming industries, there has been developing interest in accurately predicting the body pose of a camera wearer from a single egocentric video stream. Accurately predicting the ego-pose can reveal the individual's physical activities, postures, and gestures, making it possible for the system to directly interact with or assist the wearer. To this end, we presented the You2Me approach to predict a camera wearer's pose given video from a single chest-mounted camera. Our key insight is to capture the ties in interaction between the first (unobserved) and second (observed) person poses. Our results on two capture scenarios from several different activity domains demonstrate that promise of our idea, and we obtain state-of-the-art results for ego-pose.

Despite the demonstrated successes of our approach, we acknowledge noticeable weaknesses in our approach. As mentioned earlier, our method performs poorly when the second-person pose is not visible. We also notice that our method is prone to producing unsmooth output, which could be caused by a loss in granularity from clustering the set of all possible poses. Finally, our approach occasionally produces lags in the pose estimation sequences. For example, the wearer may reach out to initiate a hand-shake and our method may not predict the hand-shake pose sequence until the interactee's hand is already outstretched. In this case, the network instead assumes that the wearer is returning a handshake. We hypothesize that this is because the network is best at predicting the wearer's pose sequence only after the second-person's entire interacting pose sequence has already unfolded. To address this issue, our future work could explore a bidirectional LSTM, which involve using a reversed copy of the input sequence as an additional layer to reason about the full context of the whole video sequence.

Future work will include better reasoning about the absence of second-person

poses when interactions are not taking place. For instance, when the interactee leaves from a conversation or is obstructed from the view of the camera in a sports scene, our method should be able to reason about the poses from only motion and static cues. Furthermore, we would be interested in extending our method to handle sequences with multiple "second people". While dyadic interactions are common in daily activity, more commonly, people interact with multiple individuals all at once. Larger social settings such as conference meetings or team sports offer a richer set of social signals which may be jointly exploited to improve the ego-pose estimation of the camera wearer. Finally, we are also interested in exploring how ego-pose estimates might reciprocate to boost second-person pose estimates. Our method already suggests that the symmetry of poses between the interactee and the camera wearer can be used to improve the pose estimation of the wearer. Similar to the joint learning framework used in [58], we suppose that these ego-pose improvements can in turn be used to further improve existing second-person pose estimation techniques.

7 Bibliography

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. "Social lstm: Human trajectory prediction in crowded spaces". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[2] Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. "Socially-aware large-scale crowd forecasting". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[3] Stefano Alletto, Giuseppe Serra, Simone Calderara, and Rita Cucchiara. "Understanding social relationships in egocentric vision". In: *Pattern Recognition* (2015).

[4] Stefano Alletto, Giuseppe Serra, Simone Calderara, Francesco Solera, and Rita Cucchiara. "From ego to nos-vision: Detecting social relationships in first-person views". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2014.

[5] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. "2d human pose estimation: New benchmark and state of the art analysis". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[6] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. "Social scene understanding: End-to-end multi-person action localization and collective activity recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[7] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions". In: *International Conference on Computer Vision (ICCV)*. 2015.

[8] Frank J Bernieri, J Steven Reznick, and Robert Rosenthal. "Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions." In: *Journal of personality and social psychology* (1988).

[9] Gedas Bertasius, Aaron Chan, and Jianbo Shi. "Egocentric basketball motion planning from a single first-person image". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[10] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields". In: 2018.

[11] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. "Human pose estimation with iterative error feedback". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[12] Y-W. Chao, Z. Wang Y. He, J. Wang, and J. Deng. "HICO: A Benchmark for Recognizing Human-Object Interactions in Images". In: *IEEE International Conference on Computer Vision (ICCV)*. 2015.

[13] Dima Damen, Teesid Leelasawassuk, Osian Haines, Andrew Calway, and Walterio W Mayol-Cuevas. "You-Do, I-Learn: Discovering Task Relevant Objects and their Modes of Interaction from Multi-User Egocentric Video." In: *British Machine Vision Conference (BMVC)*. 2014.

[14] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[15] Alircza Fathi, Jessica K Hodgins, and James M Rehg. "Social interactions: A first-person perspective". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[16] Alireza Fathi, Ali Farhadi, and James M Rehg. "Understanding egocentric activities". In: *International Conference on Computer Vision (ICCV)*. 2011.

[17] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. "Detect-and-track: Efficient pose estimation in videos". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[18] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. "Detecting and recognizing human-object interactions". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[19] Alex Graves. "Generating sequences with recurrent neural networks". In: *arXiv preprint arXiv:1308.0850* (2013).

[20] Alex Graves and Navdeep Jaitly. "Towards end-to-end speech recognition with recurrent neural networks". In: *International Conference on Machine Learning (ICML)*. 2014.

[21] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[22] De-An Huang and Kris M Kitani. "Action-reaction: Forecasting the dynamics of human interaction". In: *European Conference on Computer Vision (ECCV)*. 2014.

[23] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. "Deepercut: A deeper, stronger, and faster multi-person pose estimation model". In: *European Conference on Computer Vision (ECCV)*. 2016.

[24] Umar Iqbal, Anton Milan, and Juergen Gall. "Posetrack: Joint multi-person pose estimation and tracking". In: *IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*. 2017.

[25] H. Jiang and K. Grauman. "Seeing Invisible Poses: Estimating 3D Body Pose from Egocentric Video". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[26] Sam Johnson and Mark Everingham. "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation." In: *British Machine Vision Conference (BMVC)*. 2010.

[27] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. "Panoptic Studio: A Massively Multiview System for Social Motion Capture". In: *IEEE International Conference on Computer Vision (ICCV)*. 2015.

[28] Cheng Li and Kris M Kitani. "Model recommendation with virtual probes for egocentric hand detection". In: *IEEE International Conference on Computer Vision (ICCV)*. 2013.

[29] Cheng Li and Kris M Kitani. "Pixel-level hand detection in ego-centric videos". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013.

[30] Yin Li, Alireza Fathi, and James M Rehg. "Learning to predict gaze in egocentric video". In: *IEEE International Conference on Computer Vision (ICCV)*. 2013.

[31] Ita Lifshitz, Ethan Fetaya, and Shimon Ullman. "Human pose estimation using deep consensus voting". In: *European Conference on Computer Vision (ECCV)*. 2016.

[32] Minghuang Ma, Haoqi Fan, and Kris M Kitani. "Going deeper into first-person activity recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[33] T. McCandless and K. Grauman. "Object-Centric Spatio-Temporal Pyramids for Egocentric Activity Recognition". In: *British Machine Vision Conference (BMVC)*. 2013.

[34] Louis-Philippe Morency. "Modeling human communication dynamics [social sciences]". In: *IEEE Signal Processing Magazine* (2010).

[35] H. S. Park, J-J. Hwang, Y. Niu, and J. Shi. "Egocentric future localization". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[36] H. S. Park and Jl Shi. "Social Saliency Prediction". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[37] Stefano Pellegrini, Andreas Ess, and Luc Van Gool. "Improving data association by joint modeling of pedestrian trajectories and groupings". In: *European Conference on Computer Vision (ECCV)*. 2010.

[38] Hamed Pirsiavash and Deva Ramanan. "Detecting activities of daily living in first-person camera views". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[39] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. "Deepcut: Joint subset partition and labeling for multi person pose estimation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[40] Yair Poleg, Ariel Ephrat, Shmuel Peleg, and Chetan Arora. "Compact cnn for indexing egocentric videos". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2016.

[41] Geovany A Ramirez, Tadas Baltrušaitis, and Louis-Philippe Morency. "Modeling latent discriminative dynamic of multi-dimensional affective signals". In: *International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2011.

[42] Grégory Rogez, James S Supancic, and Deva Ramanan. "First-person pose recognition using egocentric workspaces". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[43] Michael S Ryoo and Larry Matthies. "First-person activity recognition: What are they doing to me?" In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013.

[44] Michael S Ryoo, Brandon Rothrock, and Larry Matthies. "Pooled motion features for first-person videos". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[45] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A Kakadiaris. "3d human pose estimation: A review of the literature and analysis of covariates". In: *Computer Vision and Image Understanding* (2016).

[46] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *IEEE International Conference on Computer Vision (ICCV)*. 2017.

[47] Takaaki Shiratori, Hyun Soo Park, Leonid Sigal, Yaser Sheikh, and Jessica K Hodgins. "Motion capture from body-mounted cameras". In: *ACM Transactions on Graphics (TOG)*. 2011.

[48] Ekaterina H Spriggs, Fernando De La Torre, and Martial Hebert. "Temporal segmentation and activity classification from first-person sensing". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2009.

[49] Alexandros Stergiou and Ronald Poppe. "Understanding human-human interactions: a survey". In: *arXiv preprint arXiv:1808.00022* (2018).

[50] Alexander Toshev and Christian Szegedy. "DeepPose: Human pose estimation via deep neural networks". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014.

[51] Adrien Treuille, Seth Cooper, and Zoran Popović. "Continuum crowds". In: (2006).

[52] Alessandro Vinciarelli, Hugues Salamin, Anna Polychroniou, Gelareh Mohammadi, and Antonio Origlia. "From nonverbal cues to perception: personality and social attractiveness". In: *Cognitive Behavioural Systems*. 2012.

[53] Yifan Wang, Jie Song, Limin Wang, Luc Van Gool, and Otmar Hilliges. "Two-Stream SR-CNNs for Action Recognition in Videos". In: *British Machine Vision Conference (BMVC)*. 2016.

[54] Bin Xiao, Haiping Wu, and Yichen Wei. "Simple baselines for human pose estimation and tracking". In: *European Conference on Computer Vision (ECCV)*. 2018.

[55] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. "Learning feature pyramids for human pose estimation". In: *IEEE International Conference on Computer Vision (ICCV)*. 2017.

[56] Zhefan Ye, Yin Li, Alireza Fathi, Yi Han, Agata Rozga, Gregory D Abowd, and James M Rehg. "Detecting eye contact using wearable eye-tracking glasses". In: *ACM Conference on Ubiquitous Computing (UbiComp)*. 2012.

[57] Zhefan Ye, Yin Li, Yun Liu, Chanel Bridges, Agata Rozga, and James M Rehg. "Detecting bids for eye contact using a wearable camera". In: *IEEE International Conference and Workshops on Automatic Face & Gesture Recognition (FG 2015)*. 2015.

[58] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. "Recognizing micro-actions and reactions from paired egocentric videos". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[59] Ye Yuan and Kris Kitani. "3D Ego-Pose Estimation via Imitation Learning". In: *European Conference on Computer Vision (ECCV)*. 2018.

[60] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. "Towards 3d human pose estimation in the wild: a weakly-supervised approach". In: *IEEE International Conference on Computer Vision (ICCV)*. 2017.