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

Evonne Ng\(^1,2\)  Donglai Xiang\(^3\)  Hanbyul Joo\(^4\)  Kristen Grauman\(^2,4\)

evonne_ng@berkeley.edu  donglaix@cs.cmu.edu  hjoo@fb.com  grauman@cs.utexas.edu

\(^1\)UC Berkeley  \(^2\)UT Austin  \(^3\)Carnegie Mellon University  \(^4\)Facebook AI Research

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.

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 [63, 45, 58, 34, 24, 10], recent work suggests that an ego-video stream nonetheless offers implicit cues for first-person body pose [26, 69, 70]. 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 activities. As is well-known in cognitive science [47, 60, 8], 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.

To that end, we introduce “You2Me”: an approach for 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

\(^1\)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”. 
The observed second-person pose together with the camera motion and scene appearance to infer the latent ego-pose across an entire video sequence. See Figure 2.

To our knowledge, no prior work models interactions for ego-pose. Our key contribution is to leverage the action-reaction dynamics in dyadic interactions to estimate ego-pose from a monocular wearable video camera.

We validate our You2Me ego-pose approach on two forms of ground-truth capture—from Kinect sensors and a Panoptic Studio [28]—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 from 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.

2. Related work

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 [53, 50, 64, 13]. Recent approaches explore novel CNN-based methods, which have substantially improved the detection of visible body poses in images and video [10, 65, 71, 11, 34, 24, 18, 58, 36, 35, 30]. Our approach instead estimates the largely “invisible” first-person pose. Multi-person pose tracking investigates structure in human motion and inter-person interactions to limit potential pose trajectories [25, 10]. Beyond body pose, there is a growing interest in modeling human-human interactions [57, 23, 40] to predict pedestrian trajectories [1, 43, 2], analyze social behavior and group activities [43, 59, 6, 15, 23], and understand human-object interactions [61, 19, 12]. Our method also capitalizes on the structure in inter-person interactions. However, whereas these existing methods assume that all people are fully within view of the camera, our approach addresses interactions between an individual in-view and an individual out-of-view, i.e., the camera wearer.

Egocentric video Recent egocentric vision work focuses primarily on recognizing objects [14], activities [44, 16, 39, 56, 46, 37, 52, 51, 68], visible hand and arm pose [7, 31, 32, 49, 48], eye gaze [33], or anticipating future camera trajectories [41, 9]. In contrast, we explore 3D pose estimation for the camera wearer’s full body, and unlike any of the above, we show that the inferred body pose of another individual directly benefits the pose estimates.

First-person body pose from video Egocentric 3D full body pose estimation has received only limited attention [26, 55, 69]. The first attempt to the problem is the geometry-based “inside-out mocap” approach [55], 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.

Limited recent work extracts ego-pose from monocular first-person video [26, 69]. The method in [26] infers the poses of a camera wearer by optimizing an implicit motion graph with an array of hand-crafted cost functions, including a sit/stand classifier. In contrast, we propose an end-to-end learning method, which learns from the full visual frame. The method in [69] uses a humanoid simulator in a control-based approach to recover the sequence of actions affecting pose, and 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), our approach improves the prediction of upper-body joint locations during sequences when the camera has only subtle motions (like handshakes and other conversational gestures). Furthermore, unlike [69], 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 [26, 55, 69], our approach discovers the connection between the dynamics in inter-person interactions and egocentric body poses.

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, 17] and mutual gaze [66, 67] or shared gaze [42] from ego-video. More relevant
to our work, the activity recognition method of [68] 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) in two synchronized video clips to better classify them. However, whereas [68] 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. Furthermore, we infer body pose, whereas [68] classifies 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 model that uses first- and second-person features—both extracted from monocular egocentric video—to predict the 3D joints of the camera wearer. We first define the pose encoding (Sec 3.1) and 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 pose 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 articulated pose as opposed to recognizing an 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

Motion patterns observed from a 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 [26]. 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 [22]. Each element in the resulting \( 3 \times 3 \) 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 or sitting, they are more ambiguous for sequences with little motion in the egocentric video. To account for this, our second feature 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 [26] 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 elicit 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 the ball. Or more subtly, if we see a person turn 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. We use OpenPose [10] to infer interactee poses due to its efficiency and accuracy, though other third-person methods can also be employed. OpenPose provides real-time multi-person keypoint detection: given a stack of frames, it returns a corresponding stack of normalized 2D keypoint joint estimations (see Supp. file). 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”). We set missing or occluded joints in \( o_t \) to zeros, which can provide its own signal about the wearer’s proximity to the interactee (e.g., no legs visible when he/she is closer).

Note that our learning approach is flexible to the exact encodings of the ego- and second-person poses. Using normalized 2D keypoints for the second-person pose is compatible with using person-centric 3D coordinates for the ego-pose (cf. Sec. 3.1); whereas it would be problematic for a purely geometric method relying on spatial registration, for a learning-based approach mixing 2D and 3D in this way is consistent. Furthermore, while perfect 3D second-person poses would offer the most complete information, e.g., avoiding foreshortening, we find that state-of-the-art 3D methods [29, 62, 71] fail on our data due to extensive occlusions from the wearer’s hands. See Sec. 5 for experiments that concretely justify this design choice.

Figure 3 illustrates the complete set of features.

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 [20, 21] for our task. The LSTM learns the current state of the camera wearer, scene, and interactee, and uses this encoding to infer the camera wearer’s next 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 fine-grained classification task to train more robustly, as often reported in the literature (see last row of Table 1). Hence, we first quantize the space of training body poses to a fine-grained classification task to train more robustly. Our third and most important input consists of the interactee’s inferred pose even with slight occlusions.

Note that our learning approach is flexible to the exact encodings of the ego- and second-person poses. Using normalized 2D keypoints for the second-person pose is compatible with using person-centric 3D coordinates for the ego-pose (cf. Sec. 3.1); whereas it would be problematic for a purely geometric method relying on spatial registration, for a learning-based approach mixing 2D and 3D in this way is consistent. Furthermore, while perfect 3D second-person poses would offer the most complete information, e.g., avoiding foreshortening, we find that state-of-the-art 3D methods [29, 62, 71] fail on our data due to extensive occlusions from the wearer’s hands. See Sec. 5 for experiments that concretely justify this design choice.

Figure 3 illustrates the complete set of features.

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 [20, 21] for our task. The LSTM learns the current state of the camera wearer, scene, and interactee, and uses this encoding to infer the camera wearer’s next 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 fine-grained classification task to train more robustly, as often reported in the literature (see last row of Table 1). Hence, we first quantize the space of training body poses to a fine-grained classification task to train more robustly. Our third and most important input consists of the interactee’s inferred pose even with slight occlusions.

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. We use OpenPose [10] to infer interactee poses due to its efficiency and accuracy, though other third-person methods can also be employed. OpenPose provides real-time multi-person keypoint detection: given a stack of frames, it returns a corresponding stack of normalized 25
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 \( t-1 \) (initialized at \( t=0 \) as the average training pose). We learn a linear embedding for the pose indicator to map it to vector \( z_t \):

\[
z_t = \phi_z(p_{t-1}; W_z),
\]

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

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

\[
b_t = m_t \oplus o_t \oplus x_t \oplus z_t,
\]

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:

\[
h_t = LSTM(h_{t-1}, b_t; \theta_l),
\]

where \( \theta_l \) 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:

\[
\mathcal{L}(W_x, W_z, W_p, \theta_l) = - \sum_{t=1}^{N} \log(\sigma_p(W_p h_t)),
\]

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, such that each per-frame estimate is quite specific; on average the nearest quantized pose in the codebook is just 0.27 cm away per joint (see Supp. file). The inferred pose ID at time \( t \) (i.e., the arg max over the pose posteriors at that timestep) is taken as the input for \( z_{t+1} \) for the subsequent frame.

Rather than uniformly quantize all \( J \) joints, we perform a mixed granularity clustering to account for the more subtle pose changes concentrated on the upper body. Lower body poses across frames exhibit less variance, but upper body poses have more fine-grained important differences (e.g., from arm gestures). Hence, we use a sparser \( K \)-means clustering (\( K_{bot} = 100 \)) for the lower body joints and denser (\( K_{upp} = 700 \)) for the upper body. Any given pose is thus coded in terms of a 2D cluster ID.

At test time, we use the trained LSTM to predict the pose sequence. From time \( t=1 \) to \( t \), we use the predicted cluster \( p_{t-1} \) from the previous LSTM cell in Eq. 2. Figure 4Overview the model; see Supp. for architecture details.

4. You2Me Video Datasets

We present a first-person interaction dataset consisting of 42 two-minute sequences from one-on-one interactions between 10 individuals. We ask each individual (in turn) to wear a chest mounted GoPro camera and perform various interactive activities with another individual. We collect their egocentric video then synchronize it 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 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.
Panoptic Studio capture Our first capture mode uses a Panoptic Studio dome, following [28]. 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 interactee are then reconstructed at 30 fps, matching the ego-video frame rate. Each skeleton is parameterized by \( J = 19 \) 3D joint positions obtained using the method of [28]. 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.

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 wide field of view setting. Both people’s ground truth skeletons 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 limitations of the Kinect V2 sensor. While Kinect-sensed ground truth poses are noisier than those captured in the Panoptic Studio, prior work demonstrates that overall the Kinect poses are well aligned with human judgment of pose [26].

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, 27]) are captured in the third-person viewpoint. Existing egocentric datasets are either limited to visible hands and arms [32, 44], contain only single-person sequences [26, 5, 27], consist of synthetic test data [69], or lack body-pose joint labels [68]. All our data is publicly available.\(^2\) See Supp video for examples.

5. Experiments

We evaluate our approach on both the Panoptic Studio and Kinect captures. Our method is trained and tested in an 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.

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. This yields 3.2K training sequences and 2.3K test sequences. For the LSTM, we use an embedding dimension of \( E = 256 \), fixed hidden state dimension of \( D = 512 \), and batch size of 32. Learning rate is 0.001 for the first 10 epochs then decreased to 0.0001. In initial experiments, we found results relatively insensitive to values of \( K_{upp} \) from 500 to 900 and \( K_{bot} \) from 70 to 120, and fixed \( K_{upp} = 700 \) and \( K_{bot} = 100 \) for all results. Training time is 18 hours on a single GPU for 20 epochs; test time is 36 fps. See Supp.

Baselines We compare to the following methods:

- **Ego-pose motion graph (MotionGraph)** [26]: the current state-of-the-art method for predicting body pose from real egocentric video [26]. We use the authors’ code\(^3\) and retrain their model on our dataset. This method also outputs quantized poses; we find their method performs best on our data for \( K = 500 \).

- **Third-person pose deconv network (DeconvNet)** [63]: We adapt the human pose estimation baseline of [63] to our task.\(^4\) 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, altering the output space for 3D joints. 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.

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

\(^2\)http://vision.cs.utexas.edu/projects/you2me/
\(^3\)http://www.hao-jiang.net/code/egopose/ego_pose_code.tar.gz
\(^4\)https://github.com/leoxiaobin/pose.pytorch
teractee often moves out of view for long periods, explaining our 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). **Stand** outperforms other baselines but not our method. This is a well-known issue in motion forecasting: doing “nothing” can be better than doing something, since the average standing pose is in-between many test poses, representing a “safe” estimate [38]. Our method also outperforms **DeconvNet** [63], 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. Replacing our model’s classification objective with regression is significantly weaker, supporting our design choice (see last row Table 1).

Figure 5 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** baseline. Intuitively, gains happen for interactions with good body language links, such as mutually extending hands or smaller conversational gestures.

**Results** Table 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 the highest variance in everyday activity (head, elbow, wrists, hands) and the lower body joints (hips, knees, ankles, foot). See Supp. file for per-joint errors. Our approach outperforms **MotionGraph** [26] and **Ours w/o o2**. 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 our method’s impact is greatest on conversation sequences, and lowest on sports sequences (see Supp. file). 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. In sports sequences, on the other hand, the inter-
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 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 3 analyzes how the source of the second-person pose estimates affects our results. First, we substitute in for $o_t$ the 3D 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. Using 2D OpenPose for $o_t$ is better than using predicted 3D poses [71]; 3D pose from monocular data remains challenging. This justifies our use of 2D for $o_t$. 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.

Please see Supp. file for videos, per-joint and per-activity error breakdowns, depiction of the fine-grained quantization, impact of quantization on approach, and additional architecture details.

6. Conclusions

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 the promise of our idea, and we obtain state-of-the-art results for ego-pose.

Future work will include reasoning about the absence of second-person poses when interactions are not taking place, extending to sequences with multiple “second people”, and exploring how ego-pose estimates might reciprocate to boost second-person pose estimates.

Acknowledgements: We thank Hao Jiang for helpful discussions. UT Austin is supported in part by ONR PECASE and NSF IIS-1514118.
References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 2
[2] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 2
[3] S. Alletto, G. Serra, S. Calderara, and R. Cucchiara. Understanding social relationships in egocentric vision. Pattern Recognition, 2015. 2
[4] S. Alletto, G. Serra, S. Calderara, F. Solera, and R. 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. 2
[5] M. Andriluka, L. Pishchulin, P. Gehler, and B. 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
[6] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. 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. 2
[7] S. Bamshad, S. Lee, D. J. Crandall, and C. Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In International Conference on Computer Vision (ICCV), 2015. 2
[8] F. J. Bernieri, J. S. Reznick, and R. Rosenthal. Synchrony, pseudosynchrony, and dysynchrony: Measuring the entrainment process in mother-infant interactions. Journal of personality and social psychology, 1988. 1
[9] G. Bertasius, A. Chan, and J. Shi. Egocentric basketball motion planning from a single first-person image. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2
[10] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. 2018. 1, 2, 4, 5
[11] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 2
[12] Y.-W. Chao, Z. W. 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. 2
[13] G. Chéron, I. Laptev, and C. Schmid. P-cnn: Pose-based cnn features for action recognition. In IEEE International Conference on Computer Vision (ICCV), 2015. 2
[14] D. Damen, T. Leelasawasuk, O. Haines, A. Calway, and W. 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. 2
[15] Z. Deng, A. Vahdat, H. Hu, and G. 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. 2
[16] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In International Conference on Computer Vision (ICCV), 2011. 2
[17] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012. 2
[18] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran. Detect-and-track: Efficient pose estimation in videos. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2
[19] G. Gkioxari, R. Girshick, P. Dollár, and K. He. Detecting and recognizing human-object interactions. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2
[20] A. Graves. Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850, 2013. 4
[21] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In International Conference on Machine Learning (ICML), 2014. 4
[22] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. Cambridge university press, 2003. 3
[23] D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In European Conference on Computer Vision (ECCV), 2014. 2
[24] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepcut: A deeper, stronger, and faster multi-person pose estimation model. In European Conference on Computer Vision (ECCV), 2016. 1, 2
[25] U. Iqbal, A. Milan, and J. Gall. Posetrack: Joint multi-person pose estimation and tracking. In IEEE Conference on Computer Vision and Pattern Recognition (ICCV), 2017. 2
[26] 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. 1, 2, 3, 6, 7, 8
[27] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In British Machine Vision Conference (BMVC), 2010. 6
[28] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In IEEE International Conference on Computer Vision (ICCV), 2015. 2, 6
[29] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In European Conference on Computer Vision (ECCV), 2014. 2
[30] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In International Conference on Machine Learning (ICML), 2014. 4
[31] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. Cambridge university press, 2003. 3
[32] D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In European Conference on Computer Vision (ECCV), 2014. 2
[33] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepcut: A deeper, stronger, and faster multi-person pose estimation model. In European Conference on Computer Vision (ECCV), 2016. 1, 2
[34] U. Iqbal, A. Milan, and J. Gall. Posetrack: Joint multi-person pose estimation and tracking. In IEEE Conference on Computer Vision and Pattern Recognition (ICCV), 2017. 2
[35] 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. 1, 2, 3, 6, 7, 8
[36] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In British Machine Vision Conference (BMVC), 2010. 6
[37] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In IEEE International Conference on Computer Vision (ICCV), 2015. 2, 6
[38] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In European Conference on Computer Vision (ECCV), 2014. 2
[39] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In International Conference on Machine Learning (ICML), 2014. 4
[40] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. Cambridge university press, 2003. 3
[41] D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In European Conference on Computer Vision (ECCV), 2014. 2
[42] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepcut: A deeper, stronger, and faster multi-person pose estimation model. In European Conference on Computer Vision (ECCV), 2016. 1, 2
[43] U. Iqbal, A. Milan, and J. Gall. Posetrack: Joint multi-person pose estimation and tracking. In IEEE Conference on Computer Vision and Pattern Recognition (ICCV), 2017. 2
[44] 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. 1, 2, 3, 6, 7, 8
[45] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In British Machine Vision Conference (BMVC), 2010. 6
[46] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In IEEE International Conference on Computer Vision (ICCV), 2015. 2, 6
[47] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In European Conference on Computer Vision (ECCV), 2014. 2
[48] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In International Conference on Machine Learning (ICML), 2014. 4
[49] R. Hartley and A. Zisserman. Multiple view geometry in computer vision. Cambridge university press, 2003. 3
[50] D.-A. Huang and K. M. Kitani. Action-reaction: Forecasting the dynamics of human interaction. In European Conference on Computer Vision (ECCV), 2014. 2
[51] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepcut: A deeper, stronger, and faster multi-person pose estimation model. In European Conference on Computer Vision (ECCV), 2016. 1, 2
[65] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[66] Z. Ye, Y. Li, A. Fathi, Y. Han, A. Rozga, G. D. Abowd, and J. M. Rehg. Detecting eye contact using wearable eye-tracking glasses. In *ACM Conference on Ubiquitous Computing (UbiComp)*, 2012. 2

[67] Z. Ye, Y. Li, Y. Liu, C. Bridges, A. Rozga, and J. 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. 2

[68] R. Yonetani, K. M. Kitani, and Y. Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 6

[69] Y. Yuan and K. Kitani. 3d ego-pose estimation via imitation learning. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 6

[70] Y. Yuan and K. Kitani. Ego-pose estimation and forecasting as real-time pd control. In *ICCV*, 2019. 1

[71] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 4, 8