Pattern Recognition 138 (2023) 109358

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Enhancing egocentric 3D pose estimation with third person views

Ameya Dhamanaskar, Mariella Dimiccoli*, Enric Corona, Albert Pumarola, Francesc Moreno-Noguer

Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Carrer Llorens i Artigas 4-6, Barcelona, 08028, Spain

ARTICLE INFO

Article history:
Received 14 June 2022
Revised 9 January 2023
Accepted 21 January 2023
Available online 25 January 2023

Keywords:
3D pose estimation
Self-supervised learning
Egocentric vision

ABSTRACT

We propose a novel approach to enhance the 3D body pose estimation of a person computed from videos captured from a single wearable camera. The main technical contribution consists of leveraging high-level features linking first- and third-views in a joint embedding space. To learn such embedding space we introduce *First2Third-Pose*, a new paired synchronized dataset of nearly 2000 videos depicting human activities captured from both first- and third-view perspectives. We explicitly consider spatial- and motion-domain features, combined using a semi-Siamese architecture trained in a self-supervised fashion. Experimental results demonstrate that the joint multi-view embedded space learned with our dataset is useful to extract discriminatory features from arbitrary single-view egocentric videos, with no need to perform any sort of domain adaptation or knowledge of camera parameters. An extensive evaluation demonstrates that we achieve significant improvement in egocentric 3D body pose estimation performance on two unconstrained datasets, over three supervised state-of-the-art approaches. The collected dataset and pre-trained model are available for research purposes.[1]

1. Introduction

Egocentric computer vision is gaining attention in recent years, since it may enable several important developments in the fields of healthcare, robotics and augmented reality [1–3]. For many of these applications, estimating the 3D full body pose of the person wearing the camera is of special interest because it conveys rich information about his/her activities, interactions and behaviour. However, inferring the first-person's 3D body pose from a given egocentric video sequence is a challenging problem in computer vision since wearable cameras are typically worn on the chest or the head, and have almost no view of the camera wearer's body. As a consequence, state-of-the-art approaches for third-person body pose estimation are not suited to the egocentric domain, and dedicated methods need to be developed.

Despite the relevance of the problem, egocentric body pose estimation has received little attention so far [4–9]. Prior work has shown the importance of leveraging egomotion and the coarse scene structure to predict the body pose behind the camera [4]. More recently, the body pose of a second-person (as opposed to the first-person wearing the camera), as observed in a first-person video stream, has been used to enhance egocentric pose estimation during dyadic interactions [7]. While these methods rely only on the egocentric video data itself to estimate the egopose, another line of work uses simulated data to learn a control policy that is ultimately transferred to egocentric videos.

Specifically, Yuan and Kitani [5] considers simulated data to learn a control policy accounting for the physics that underlies the kinematics of the motion and hence can generate physically plausible first-person body poses. In the same spirit, Yuan and Kitani [6] learns a control policy from unsegmented motion-captured data without the need for domain adaptation to transfer them to real egocentric data. Lately, fostered by potential applications in the field of augmented and virtual reality, the egocentric 3D pose has been estimated from wearable cameras with eye-fish lenses, which ensures a larger body part view [8,9] by incorporating motion priors learned from motion-captured data. All these methods, however, either rely solely on information available in first-person videos or leverage simulated data for egopose from existing motion-captured data or humanoid simulator.

In this paper, we explore whether visual cues on third-person videos can help to improve the robustness of 3D pose detectors on

* Corresponding author.
E-mail addresses: adhamanaskar@iri.upc.edu (A. Dhamanaskar), mdimiccoli@iri.upc.edu (M. Dimiccoli), ecorona@iri.upc.edu (E. Corona), apumarola@iri.upc.edu, apumarola@fb.com (A. Pumarola), fmoreno@iri.upc.edu (F. Moreno-Noguer).

[1] https://github.com/nudlesoup/First2Third-Pose

https://doi.org/10.1016/j.patcog.2023.109358

Fig. 1. First-person (left) and third-person (right) perspectives represent the two sides of the same coin. Our work considers how a joint embedding space between these two worlds can facilitate egocentric 3D body pose estimation. The skeleton in the centre has been estimated from an egocentric video by relying on our joint embedding space.

first-person views. For this purpose, we rely on a new, real dataset of synchronized paired first- and third-person videos to learn a joint embedded space that we leverage to enhance 3D egopose estimation (see Fig. 1). To learn the embedding space, we train in a self-supervised fashion, a two-stream semi-Siamese convolutional neural network for the task of discriminating if two views (first- and third-views) correspond to the same 3D skeleton. At inference time, the embedded space is used to extract image features allowing for the discrimination of different 3D human poses from new, unpaired egocentric videos. Our approach does not require domain adaptation to transfer to new unpaired egocentric datasets and can be applied to egocentric videos captured in the wild, for which the camera parameters are typically unknown. Experimental results on two real datasets show that the use of joint first- and third-embedded features have a significant benefit for 3D egopose estimation. Specifically, we reduce the error of 12.71% and 7.51% on average respectively on two datasets over three state-of-the-art approaches.

Our key contributions can be summarized as follows:

- We demonstrate for the first time the benefit of linking first- and third-person views for the task of 3D egocentric body pose estimation.
- We collect and make publicly available *First2Third-Pose*, a large dataset consisting of nearly 2000 synchronized first- and third-views videos, capturing 14 people performing 40 different activities in 4 different scenarios.
- We show that our dataset is useful to compute a joint embedded space between first- and third-views from which to extract discriminative features for estimating 3D egopose from arbitrary egocentric videos, without any knowledge of the camera parameters.
- We achieve consistent and significant performance improvement on two real datasets, *First2Third-Pose* and [4], over three existing baseline methods [4,7,10].

2. Related work

3D pose estimation from third-view. Learning to estimate 3D body pose from a single RGB image from a third-view (assuming the person is seen in the image) is a long-standing problem in computer vision. Most approaches in this area follow a fully-supervised pipeline and use images annotated with 3D poses to train a deep neural network that regresses the 3D pose directly from images [11–13] or via an intermediate step that estimates the 2D pose [14–16] The architectures explored so far for this task range from Euclidean CNN to the more recent non-Euclidean Graph-Convolutional Network (GCN) [17]. Some representative examples are as follows : Li et al. [11] proposed a deep neural network approach for maximum-margin structured learning that learns jointly the feature representations for image and pose as well as the score function. Tome et al. [14] proposes the first CNN architecture for jointly estimating 2D landmarks and 3D human pose. Cai et al. [16] models spatial-temporal dependencies between different joints of temporal consecutive skeletons through a GCN approach and consolidates features across scales via a hierarchical "local-to-global" architecture.

However, since all these approaches require a large amount of annotated data for training, they are typically trained on datasets acquired in controlled indoor environments, for which it is easy to use motion capture systems [18,19].

Along with another body of work devoted to 3D mesh reconstruction, the 3D skeleton is often explicitly taken into account. Sun et al. [20] proposed a module for disentangling the skeleton from the rest part of the human 3D mesh, hence building a bridge between 2D/3D pose estimation and 3D mesh recovery. Wang et al. [21] directly infer sequential 3D body models by extracting local features of a sequence of point clouds and regressing 3D coordinates of mesh vertices at different resolutions from the latent features of point clouds. In [22] expressive body motion capture including 3d hands, face, and body are estimated from a single image as a form of shape and pose parameters of the SMPL-X 3D model of the human body. Xu et al. [23] proposed a resolution-aware network for 3D human pose and shape estimation that can handle arbitrary-resolution input with one single model.

Interestingly, to reduce the need for expensive 3D annotations and hence enable training on 3D body pose datasets acquired in the wild, several approaches only use 2D weak annotations [24–26] or used weak [27,28] and self-supervised based methods [29]. Iqbal et al. [27] proposed using unlabeled multi-view data for training in an end-to-end manner by enforcing the 3D poses estimated from different views to be consistent. Cai et al. [28] proposed to use depth images captured by commodity RGB-D cameras at training time to alleviate the burden of costly 3D annotations in large-scale real datasets. Jenni and Favaro [29] proposed a self-supervised approach to learn feature representations suitable for 3D pose estimation, that uses as a pretext task the detection of synchronized views (which are always related by a rigid transformation).

Another self-supervised method for 3D pose estimation was proposed in Rodhin et al. [30], where the latent 3D representation is learned by reconstructing one view from another. However, differently from Jenni and Favaro [29] their approach strictly relies on the knowledge of the camera extrinsic parameters and background images and therefore are not suited to datasets captured in the wild.

In this paper, we use first- and third-person paired data not only to get weak annotations for training but also to learn a multi-view embedding space in a self-supervised fashion, which we further exploit to enhance 3D pose estimations. Differently from Jenni

and Favaro [29], where the pretext task translates into a classification of rigid versus non-rigid motion, in our case, there is no direct link between the image information of the two types of images (first and third view). In addition, in contrast to [30], our approach does not require knowledge of the camera parameters nor of background images.

3D pose estimation from first-view. Inferring human poses from egocentric images or videos is a problem that has been looked into only recently. Early works focused on estimating gestures and hand poses assuming that arms were partially visible [31–33]. In [34] several body-mounted cameras on a person's joints were used to infer body joint locations via a structure-from-motion approach. Jiang and Grauman [4] has been pioneering in showing that it is possible to estimate the invisible full body pose of the camera wearer directly from egocentric videos. This work considered dynamic motion signatures and static scene structure cues to build a motion graph from the training data and recovered the pose sequence by solving for the optimal pose path. More recently, Ng et al. [7] leveraged the visible body pose of a person interacting with the camera wearer to improve the wearer's pose estimation.

Other methods use a humanoid simulator in a control-based approach [5,6] to estimate the 3D body pose of a camera wearer. Yuan and Kitani [5] learns a control policy on simulated data in a two-stage imitation learning process that yields physically valid 3D pose sequences. This is evaluated quantitatively only on synthetic sequences. On the same line, Yuan and Kitani [6] proposed an approach that can learn a Proportional Derivative control-based policy and a reward function from unsegmented motion-captured data and estimate various complex human motions in real-time without the need to perform domain adaptation.

More recent approaches estimate egopose from video captured by a head-mounted and front-facing fisheye camera [8,9,35,36], which better simulates augmented and virtual devices. Mo2Cap2 [35] and xR-egopose [36] estimate the local 3D body pose in egocentric camera space, whereas [9] proposes a method to estimate it in the world coordinate system. This is achieved by leveraging the 2D and 3D keypoints from CNN detection as well as VAE-based motion priors learned from a large motion-captured dataset. Jiang and Ithapu [8] leverages both the dynamic motion information obtained from camera SLAM and the occasionally visible body parts to predict jointly head and body pose. Unlike any of the existing methods, our approach exploits the underlying connection between first- and third-views for 3D egopose estimation.

Linking first-person and third-person perspectives. Previous work has demonstrated the benefits of linking first-person and third-person perspectives for different tasks. Soran et al. [37] showed the potential of combining a single wearable camera and multiple static cameras to better understand action recognition. More recently, Sigurdsson et al. [38] introduced a large-scale dataset of paired first- and third-person videos and used it to learn a joint multi-view representation and transfer knowledge from the third-person to the first-person domain for the task of zero-shot action recognition. A combination of first-person views from two social partners has been explored for recognizing micro-actions and reactions during social interactions [39] as well as to improve activity recognition of two partners engaged in the same activity [40]. In [41], an embedding space shared by first- and third-person videos is learned to match camera wearers between third and first-person. Ego-exo [42] is a framework to create strong video representations for downstream egocentric understanding tasks by leveraging traditional third-view large-scale datasets.

In any event, and to the best of our knowledge, the potential of linking the first-person and third-person perspectives for the 3D egocentric body pose estimation we propose in this paper has never been explored so far.

Fig. 2. Capture setup used for our *First2Third-Pose* dataset. In addition to a head-mounted wearable camera, two static cameras are used to capture side and front views.

3. First2Third-Pose dataset

We next introduce *First2Third-Pose*, a large dataset of short videos covering a variety of human pose types and including multi-third-person-views in addition to a first-person view.

Dataset collection. We built a multi-view synchronized dataset wherein we capture 14 people (in turns) of varying heights, weights and genders while performing 40 different activities in both indoor and outdoor environments (lab, streets, parks, corridors, basketball courts, and parking). Every individual is asked to wear a head-mounted camera which captures his/her egocentric view. We use the Go Pro Hero 4 in a normal view setting which records in 1920×1080 resolution at 25 fps. All indoor and outdoor environments are equipped with two static cameras that capture the side and front views. We use the Go Pro Hero 3 in 1920×1080 resolution at 25 fps to record the side view and the Sony DSLR in 1920×1080 resolution at 25 fps to record the front view. An example of outdoor capture setup is shown in Fig. 2. The 'Lab' scene is equipped with an additional Amcrest camera that we use under the wide setting with a frame rate of 20 fps and 2304×1296 resolution. The top view captures a perspective from above the individual; the side view captures the 3rd person perspective from either left/right side recorded parallel to the individual; the front view captures the 3rd person view of the individual from the front, and the egocentric view captures the 1st person perspective of the individual. Each person performs about 40 activities in two indoor and two outdoor locations. We record a total of 1950 activity sequences lasting between 8 to 25 seconds, with a total duration of 2.5 hours. The activities include sports actions such as *e.g.* boxing, basketball, soccer and day-to-day tasks like reading, typing or sitting on a couch/chair/ground. Examples of synchronized viewpoints for different activities in our dataset are shown in Fig. 3.

Post-processing. To enable synchronization across multiple views, we recorded one video for each view in a location (ego, top, front, side). Before starting to enact each activity, the participants are asked to clap. This clap is visually captured in an egocentric camera and heard across multiple views. The sound of the clap is used to determine the starting point of each activity. We use the front views to check how long the activity has been performed. This time duration is noted and the video is manually scanned to find starting points of the activity. The activity is then clipped out from the main video and annotated using the name, activity class performed, location and the view that the video presents. Each video clip has one activity class. This is done for all the front, ego, top and side views.

Dataset annotation. The body pose is represented by $J = 17$ 3D joint positions. Similarly to previous work on egocentric pose estimation [6], our ground truth is estimated from front views. However, instead of using only 2D joints as in [6], we first estimated

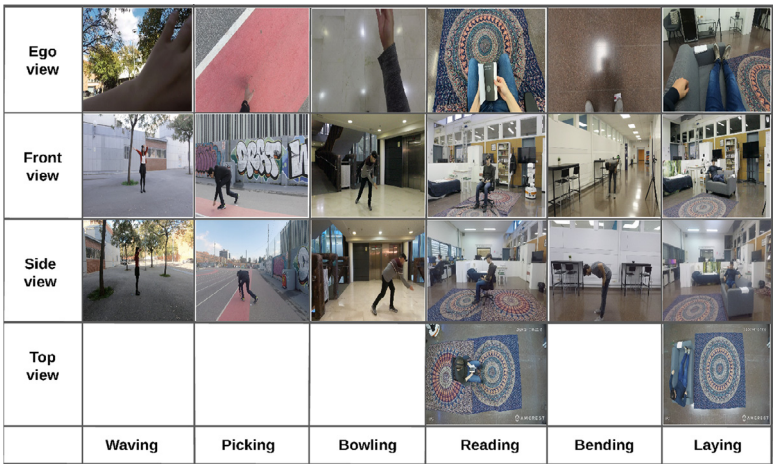

Fig. 3. Example of synchronized viewpoints in our dataset for different activities.

Table 1
Comparison of publicly available datasets for 3D egocentric pose estimation.

Dataset	#activities	#videos	#people	#view	#body joints	#scenes	#fps	#camera loc.	#duration	#GT
MotionGraph [4]	19	18	10	1	25	4 indoor	30	chest	1-3m	KineticsV2
You2me [7]	4	42	10	1	25	indoor	30	chest	2m	Kinect&Panoptic
[6]	8	24	5	2	22	2(indoor/outdoor)	30	head	20s	3rd-view(2D)
Ours	**40**	**1950**	**14**	**4**	**17**	**4(indoor/outdoor)**	**25**	**head**	**8-25s**	**3rd-view(3D)**

Table 2
Comparison of datasets combining first-person videos with other's viewpoint synchronized videos.

Dataset	#activities	#videos	#people	#view	#scenes	#fps	#add.camera location	#duration	full pose visible on add. camera	Task
[43]	12	20	11	3	3	60	hands	30s	-	gesture rec.
[37]	28	140	5	4	1(Lab)	30fps	side,back,top	1-2s	-	action rec.
[41]	uncons.	7	4	3	2	30	side	5-10m	-	pers. disam.
[41]	uncons.	7	4	3	2	30	side	5-10m	-	pers. disam.
[39]	7	1226	6	2	6	60	head	1.5s	-	act./react.
[40]	4	24	4	2	3	30	head	90s	-	act.rec.
Ours	**40**	**1950**	**14**	**4**	**4(in/out)**	**25**	**side,front,top**	**8-25s**	**yes**	**3d pose est.**

2D body poses by using Detectron2 [44], and then we obtained 3D body estimations via a pre-trained lifting model [45]. As the latter method reports an average error of 4–5 cm on large-scale datasets, we can assume that at most the same holds in our case. Visual inspection of 3D pose predictions on the test set corroborated the plausibility of such an assumption. It is worth remarking that as our dataset was captured in multiple scenarios, the camera extrinsic parameters are not available. Therefore, the computation of the ground truth could not benefit from triangulation methods by using multiple views. However, the ground truth annotations are not required to compute the joint embedded space we exploit in the proposed approach.

Dataset comparisons. In Table 1 we summarize the characteristics of our proposed dataset with respect to other available benchmarks for 3D egopose estimation. It can be observed that our dataset scales existing ones in terms of the number of videos and presents more variability in terms of background scenes, number of participants and activities. Furthermore, it provides multiple views of the same scene (egocentric, top, side, front). Therefore, it is currently the largest and most comprehensive dataset for 3D egopose estimation from videos.

In Table 2 we compare qualitatively existing datasets with synchronized videos including at least one first-person viewpoint, and we show that our dataset is the only one suited for the task of 3D egopose estimation. We stress that while other paired datasets with first- and third-person views currently exist, e.g. [38], we did not include them in Table 2 since they are not synchronized and therefore not suitable for our task. All our data will be made publicly available upon acceptance.

4. Approach

Given an egocentric video sequence as input, our goal is to estimate the 3D body joints of the camera wearer as output. More formally, for each egocentric video frame at time t, the output is a set of J joint 3D coordinates corresponding to the skeleton of the camera wearer at frame t, with shape $p_t \in \mathbb{R}^{3J}$.

Our key insight is to build image features allowing us to discriminate different 3D human poses by projecting and aligning data from the first and third-view onto a shared representation space. Formally, our objective is to learn functions $f_1 : \mathcal{R}^F \rightarrow \mathcal{R}^J$ and $f_2 : \mathcal{R}^T \rightarrow \mathcal{R}^J$ which map first and third views corresponding to the same 3D pose respectively onto nearby points in a joint embedded space. Positive first-third pairs are extracted from synchronized videos and fed into a two-stream Siamese architecture with a first-view subnetwork and a third-view subnetwork, each producing 64-D embeddings. A curriculum-based mining schedule is used to select appropriate negative pairs which are then trained using a contrastive loss, as detailed below. Our contrastive loss evaluates if pairs of first- and third-person views are synchronized. To solve such a synchronization task, the network must learn what these extremely different views have in common when they are

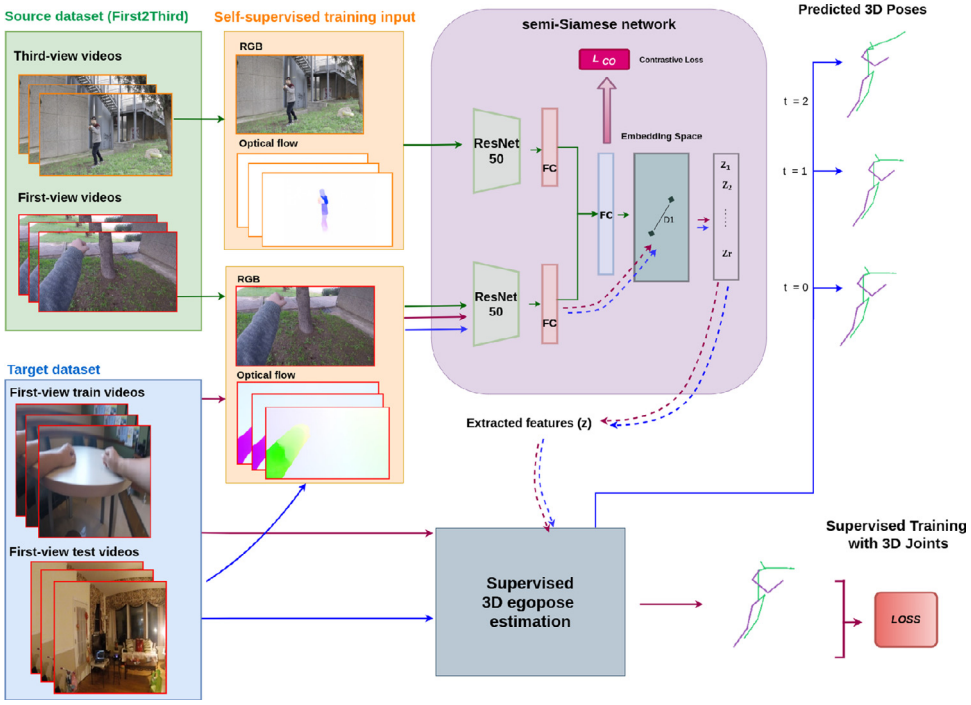

Fig. 4. Our model uses a semi-Siamese architecture to learn to detect if a pair of first- and third-view videos of the *First2Third-Pose* paired source dataset are synchronized or not, by minimizing a contrastive loss (green arrows). This pretext task leads to learning a joint embedding space, where the gap between the first-view and third-view worlds is minimized. The so-learned joint embedding space can in principle be leveraged by any supervised method for 3D egopose estimation on a target dataset, without a need for domain adaptation. At both train time (brown arrows) and test time (blue arrows), the semi-Siamese network is used for feature projection onto the learned joint embedded space. z is 64-dimensional vector, obtained once removed the softmax layer of the Siamese network was pre-trained with our dataset.

synchronized: the 3D human body pose of the person is visible in the third-view but invisible in the first-view. Consequently, the proposed synchronization pretext task translates well to the downstream task of 3D egopose estimation.

4.1. Learning a joint first/third view embedding

First and third-person perspectives are very different in appearance (see Fig. 3). However, previous work [38,41] has shown that it is possible to find spatial and motion domain feature correspondences between video types in a self-supervised fashion. While this has proved to be useful in the context of activity recognition and second-person disambiguation, its usefulness for 3D body pose estimation has never been investigated before. Our goal is therefore to attain a semantically meaningful space where paired first- and third-person videos are close to each other while the proximity of unpaired videos is avoided. The pretext task is to classify paired views into synchronized and unsynchronized, which in turn corresponds to determining if pairs of first-person and third-person views correspond to the same 3D human pose. Similarly to Fan et al. [41], this is achieved by training a two-stream semi-Siamese Convolutional Neural Network (CNN) with a ResNet50 backbone [46]. Differently from Fan et al. [41], where self-supervision is performed separately on the spatial and motion domains, we treat them jointly and we minimize a single contrastive loss. Since first- and third-person views are very different in appearance, parameter sharing is allowed only for the last fully connected layer. The semi-Siamese architecture is shown in the upper part of Fig. 4.

At training time, we feed to the network paired exemplars of first-view and third-view, consisting of stacked RGB and optical flow frames. To estimate optical flow on your dataset, we used the FlowNet2 [47] architecture pre-trained on the FlyingChairs [48] and subsequently fine-tuned on the Things3D [49] datasets. Specifically, for each given video frame, we computed the optical flow field as a forward pass for a window of length 11 centered on it. We corroborated the quality of the estimation by visually inspecting the results. We then minimize a contrastive loss by measuring the Euclidean distance for positive exemplars and a hinge loss for negative ones. Our loss function is as follows:

$$L = \sum_{j}^{M} y_j ||x_j^f - x_j^t||^2 + (1 - y_j) \max(0, m - ||x_j^f - x_j^t||)^2$$

where M is the number of frames in a batch, m is a predefined constant margin, x^f and x^t indicate first and third views (stacked RGB and optical flow frames) respectively, y_i is an indicator that takes value 1 if x_j^f and x_j^t are synchronized exemplars, and 0 otherwise. Following its effectiveness in self-supervised learning as an alternative to random exploration, we adopted a curriculum learning strategy for training [50]. Concretely, we defined it as *Easy negatives* pairs of first- and third video clips corresponding to the same person doing a different activity in the same environment.*Hard negatives* were defined as pairs of first- and third video clips corresponding to the same person doing the same action in the same environment at different intervals of time.

In the following subsection, we aim at leveraging this joint representation space learned with our source *First2Third-Pose* dataset to estimate the 3D body pose of the person behind the camera in first-view videos from a target dataset.

4.2. Transfer to 3D egopose estimation

As a byproduct of learning to classify paired first- and third-view videos into synchronized and unsynchronized, the representation gap between the two perspective views is minimized. Therefore, the representation space shared by the first and third-view enables learning a semantically rich space to represent 3D human

pose. From this low-dimension representation we can learn to predict 3D human pose regression via supervised learning.

We used the semi-Siamese network pre-trained on the pretext task with the source *First2Third-Pose* dataset to extract features, which are then useful for 3D pose estimation from egocentric videos. More specifically, stacked RGB and optical flow frames from the unpaired egocentric video are fed into the first-view stream of the network in a forward pass. The bottom part of Fig. 4 shows how to target train and test egocentric video datasets that can be used as input to our pre-trained semi-Siamese network to extract features that are subsequently used as additional features by a supervised 3D egopose estimation method. The network can in principle be used as a feature extractor by an arbitrary supervised model for 3D egopose estimation, at both training and test time. At training time, the target dataset is used as input for both the supervised method at hand and the pre-trained Siamese network. However, the latter is used only to perform a forward pass allowing to project of the input videos into the shared representation space and hence obtaining discriminative features used as additional input for the supervised approach. In our experiments, we will show that even if the embedding space has been learned relying on the *First2Third-Pose* source dataset, it transfers well on a different target egocentric video dataset, without a need for domain adaptation.

5. Experiments

Implementation details. We train a semi-Siamese network based on the ResNet50 backbone that takes an input RGB frame on each stream, stacked to the optical flow fields for a set of 10 consecutive frames. We used FlowNet2 [47] to estimate optical flow. The output of the ResNet in each stream is fed to a fully connected layer of dimension 100. The last common fully connected layer has a dimension of 64.

We generate the training data by splitting our Multiview dataset into 150k training frames and 40k testing frames. The train set includes activities performed by 10 people (8 actors, 2 actresses), while the test set includes activities performed by 4 unseen people (2 actors, 2 actresses). To train the Siamese Network we generate positive and negative image pairs, where the positive pairs are generated by taking synchronized first- and third-view (front) video frames. As we adopted a curriculum learning strategy for training, we generated negative pairs in two ways. Easy negative pairs correspond to first- and third-view videos of the same person doing different activities in the egocentric and front images but in the same environment. Hard negative pairs correspond to shifted time intervals in paired first-and-ego views. We follow curriculum learning to train the Siamese network for 2 epochs. We train the network using easy negative pairs for the first epoch and use the hard negative pairs for the second epoch. We use the contrastive loss with a margin of 0.9 to train the network using these pairs. Training time is 96 hours on a single GPU for 2 epochs. The learning rate is set to 0.0001 with the momentum 0.9 and the weight decay 5e-4. Each predicted 3D body pose has the hip joint positioned at the origin of the coordinates system. The first axis is parallel to the ground and points to the wearer's facing direction. The second axis is parallel to the ground and points to the left hip. The third axis is perpendicular to the ground and points in the direction of the spine. To account for the variability in human dimensions, we normalize each skeleton for scale based on the individual's shoulder width.

Datasets. In addition to our *First2Third-Pose* dataset described in Section 3, we use the dataset introduced in [4]. The first-view for the two datasets has been captured wearing the camera on the head for our dataset and on the chest for the other. More details about the dataset [4] can be found in Table 1. Difference in appearance with our source *First2Third-Pose* dataset can be appreciated in Fig. 4 (see source and target first-view videos). Figure 6 illustrates the difference in appearance between two datasets used in the paper for the same activity: our proposed *First2Third-Pose* captured by a head-mounted camera, and the Invisible pose dataset [4] captured by a chest-mounted camera. **Baselines.** We considered two state-of-the-art methods for 3D egocentric pose estimation [4,7], and we considered also a baseline method tailored for 3D pose estimation from third-view videos [10] that we adapted to our task. **MotionGraph** [4] is currently the state-of-the-art method for predicting 3D body pose from real egocentric videos without a second interacting person. We used the publicly available author's code[2] to extract static and motion features from our dataset, and modified the MotionGraph dynamic programming algorithm to account for the features extracted by relying on our joint embedded space. We retrained the model for both datasets, using the 300 quantized poses as used in [4]. **You2me** [7] has been recently proposed as a method able to account for the visible second person interacting with the camera wearer, as it leverages his/her 3D pose estimates to improve egopose predictions. Even if there are no second persons in our *First2Third-Pose* dataset, this approach is still a valid alternative to state-of-the-art MotionGraph, since the use of a recurrent long short-term memory (LSTM) network ensures smooth frame-to-frame 3D body pose transitions. We used the author's code[3], that takes as input motion and appearance-based features, and used additional features vector extracted leveraging our learned joint embedding as input to the LSTM for each frame. We trained this model for both datasets, using 700 quantized poses for the upper body and 100 for the lower body, as in [7]. We also adapted and trained from scratch the 3D human pose estimation baseline method proposed in [10][4], **DeconvNet**, that adds deconvolutional layers to ResNet. We altered the output space for 3D joints and minimized the mean-squared error on the training set. We found this off-the-shelf deep pose method extremely effective for ego-pose estimation, especially on our *First2Third-Pose* dataset that, being large scale, is well suited for end-to-end learning.

Evaluation metric. Each skeleton is rotated so the shoulder is parallel to the yz plane and the body centre is at the origin. The error is then computed as the Euclidean distance between the predicted 3D joints and the ground truth, averaged over the full sequence and scaled to centimetres based on a reference shoulder distance of 30 cm.

Results. In Tables 3 and 4 we show the average error per joint and for all joints (in cm) obtained on the test set of our dataset and on the dataset [4] respectively. We denote by MotionGraph-SS, you2me-SS and Deconvnet-SS our Self-supervised approach based on the methods MotionGraph [4], you2me [7] and DeconvNet [10], respectively. In addition, following competitive methods [4,7], we present results in terms of errors averaged separately for the upper body (Neck, Head, Thorax, Spine, Shoulders, Elbows, Wrists, Hands) and lower body joints (Hips, Knees, Ankles, Feet). Overall, these tables show that leveraging features extracted from the common representation space learnt with our proposed dataset consistently gives better results over existing supervised methods for 3D egopose estimation. The fact that our approach results effectively also on the dataset [4], demonstrates that the features extracted relying on the learned joint embedding space can be efficiently transferred to arbitrary egocentric videos, without the need for domain adaptation.

[2] http://www.hao-jiang.net/egopose/index.html
[3] https://github.com/facebookresearch/you2me
[4] https://github.com/una-dinosauria/3d-pose-baseline

Table 3
Average joint error in the *First2Third* Dataset, in cm.

	First2Third Dataset											
	Hip	Neck	Head	Shoulders	Elbows	Wrists	Thorax	Knees	Feet	UppBody	LowBody	Avg
MotionGraph [4]	3.40	8.03	12.40	7.74	20.93	39.13	10.73	32.50	53.81	16.98	25.63	20.54
MotionGraph-SS	**3.37**	**5.96**	**9.57**	**6.23**	**18.86**	**31.09**	**8.38**	**28.66**	**45.14**	**13.89**	**22.04**	**17.25**
you2me [7]	2.96	10.45	15.31	9.81	20.15	34.52	13.13	26.31	48.25	17.78	23.16	20.00
you2me-SS	2.99	**8.66**	**13.26**	**8.23**	**18.90**	**33.61**	**11.34**	27.08	**47.74**	**15.63**	**21.33**	**18.03**
Deconvnet [10]	3.15	5.72	8.96	5.77	17.13	31.26	7.81	27.84	45.48	13.35	21.85	16.85
Deconvnet-SS	**3.12**	**5.69**	9.30	**5.71**	**14.89**	**27.10**	8.03	**24.00**	**40.09**	**12.09**	**18.64**	**14.78**

Table 4
Average joint error in the Invisible Poses Dataset [4], in cm.

	Invisible Poses Dataset [4]												
	Hip	Neck	Head	Shoulders	Elbows	Wrists	Hands	Knees	Ankle	Feet	UppBody	LowBody	Avg
MotionGraph [4]	2.24	19.40	21.60	16.23	17.06	24.02	27.27	24.78	32.29	34.13	22.09	20.76	21.61
MotionGraph-SS [4]	**2.14**	**17.20**	**19.50**	**14.23**	**14.81**	**21.29**	**24.32**	**23.32**	**30.43**	**32.29**	**19.65**	**19.60**	**19.64**
you2me [7]	2.14	15.50	17.10	15.17	19.34	28.54	32.23	22.48	33.94	36.63	24.53	21.55	23.38
you2me-SS	**1.97**	**15.00**	**17.00**	**13.87**	**16.19**	**24.04**	**27.19**	23.27	**32.80**	**35.37**	**20.94**	**20.75**	**20.87**
Deconvnet [10]	**2.58**	17.70	21.30	12.65	13.81	20.87	23.46	22.00	27.30	29.04	18.46	17.98	18.29
Deconvnet-SS	2.64	**16.00**	**18.80**	**12.48**	**13.59**	**20.31**	**22.82**	**21.49**	**26.20**	**27.74**	**18.07**	**17.31**	**17.85**

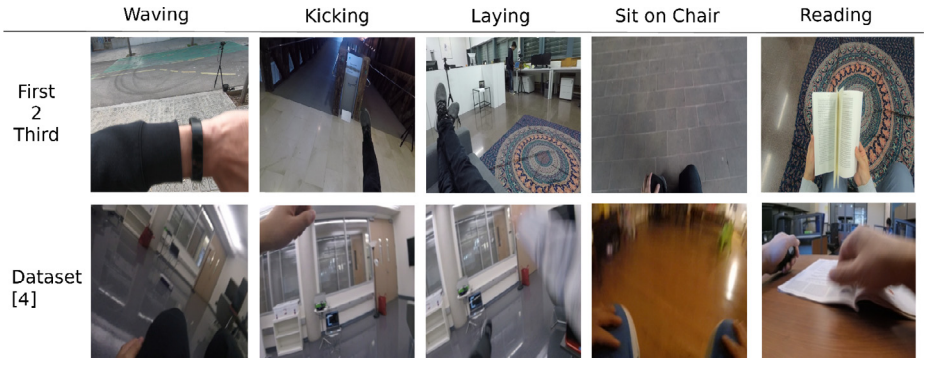

Fig. 5. Visual comparisons of predicted skeletons for three different activities. GT: ground truth. DeconvNet-SS: proposed method. MotionGraph: state-of-the-art [4]. DeconvNet: end-to-end baseline [10]. you2me: [7] baseline. Test videos are from the *First2Third-Pose* dataset test split.

Table 5
Comparison of time complexity (FLOPS) and actual run-times for training one epoch on the *First2Third-Pose* dataset and for inference on a single image.

Method	Tr. time (h/epoch)	Inf. time (sec/image)	FLOPS (GMac)
MotionGraph [4]	0.10	0.01	0.45
You2me [7]	2.50	0.17	12.74
DeconvNet [10]	1.24	0.28	9.95
MotionGraph-SS	0.25	0.01	0.65
You2me-SS	26.06	0.30	12.74
DeconvNet-SS	13.00	0.49	9.95
Siamese component	40.33	0.06	9.82 (Tr.)/4.91 (Inf.)

Figure 5 shows qualitatively that using features extracted by leveraging the learned joint embedding space indeed gives better results.

In Table 5, we also considered the time complexity of the proposed approach with respect to the methods we compare to, measured in terms of Floating Point Operations per Second (FLOPS) computed by using a dedicated PyTorch library[5] for the neural network based models. For the Motion graph, whose dynamic programming code released by the authors only includes vector-to-vector operations, we report the FLOPS corresponding to the sum of these operations during the execution of the program. As it is standard when computing FLOPS, we assume that all input features have been pre-computed and loaded. As the size of the self-supervised features $\mathbf{z}$ is relatively small compared to the rest of the input features, its effect on the FLOPS counting is negligible for the considered methods. However, given that precisely the same number of FLOPS may have radically different run times, we also report the run times for training and inference. We performed all the experiments on a single workstation equipped with an Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz, 128GB RAM 2133 MHz, NVIDIA Tesla K40c with 2880 CUDA core and operating system Ubuntu 14.04. It can be observed that, for the methods based on neural networks, while the run time at training time is significantly increased, at inference time only a very little increment is observed for getting the projection of the ego-view on the shared representation space. Therefore, at inference time, the proposed approach can be considered suitable for real-time applications.

[5] https://pypi.org/project/ptflops/

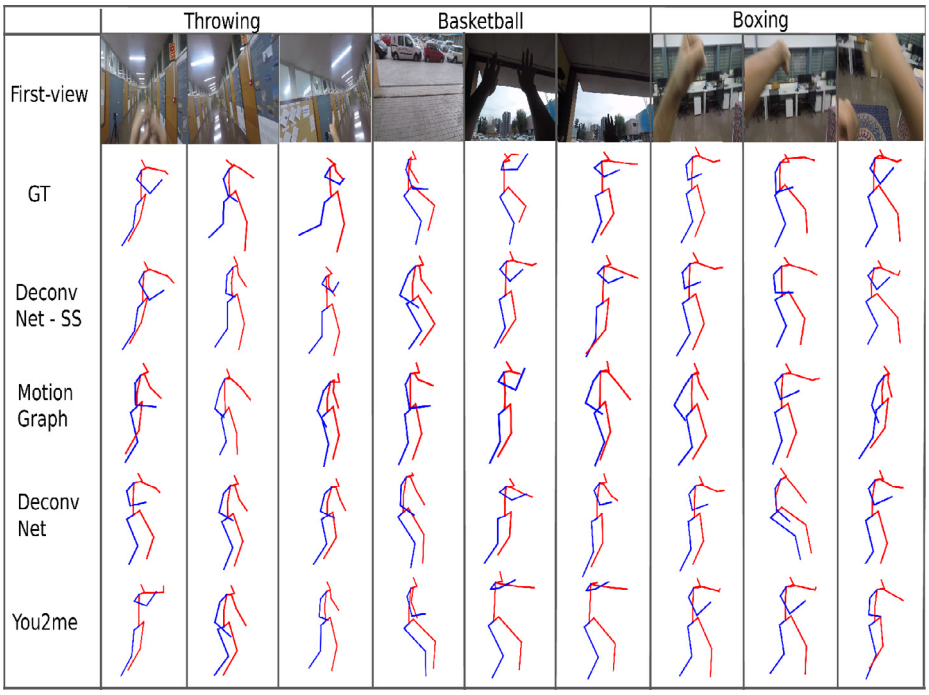

Fig. 6. Examples of activities captured in the two datasets: *First2Third-Pose* (top) and Invisible Pose Dataset [4] (bottom).

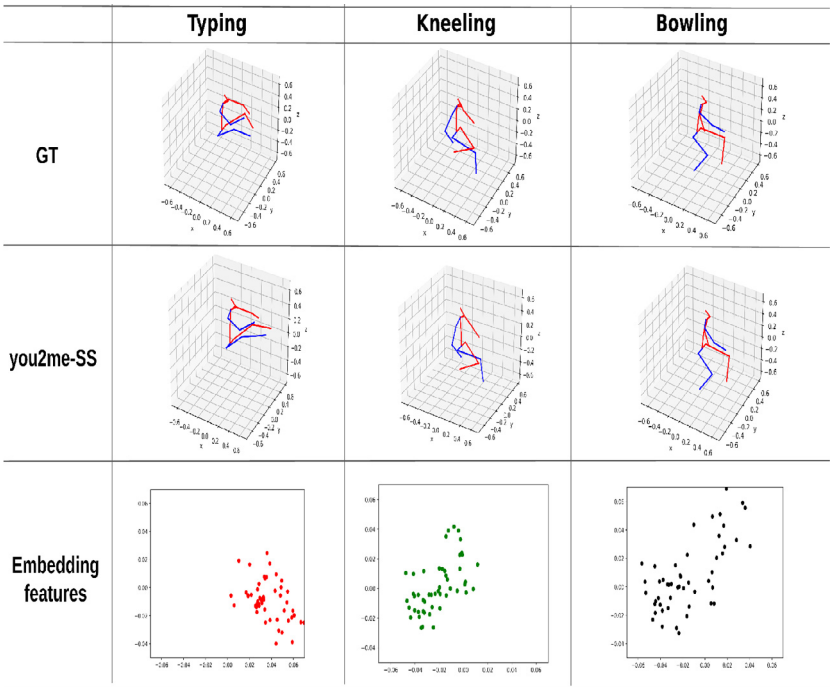

Fig. 7. Top: Ground truth 3D body pose for three different activities. Middle: 3D body pose estimated by using our self-supervised approach based on the method you2me, denoted as you2me-SS. Bottom: Scatter plots of embedding features obtained by using with you2me-SS, Similar body poses have similar features.

Insights on the joint embedded space. To verify that the features obtained using our embedded space, say embedded features, are discriminative for 3D egopose estimation, we first apply PCA to the feature matrix to reduce the feature dimension to two, and then we visualized the features of videos corresponding to different activities via a scatter plot (each dot is a frame). In Fig. 7 we show example poses for three different activities (ground truth and prediction) together with the scatter plot of corresponding embedding features for the surrounding 1-second video segment. The 3D skeleton corresponding to the activity *typing* differs from both *bowling* and *kneeling*, while those of *bowling* and *kneeling* are more similar. This is also reflected by the corresponding features.

6. Model interpretation

To shed light on the structure of the learned embedded space, we evaluated numerically the distance between corresponding first and third view embeddings of the same action. In addition, we visualized the embedding vectors corresponding to different activities captured from the same point of view (first-or third). In Fig. 8

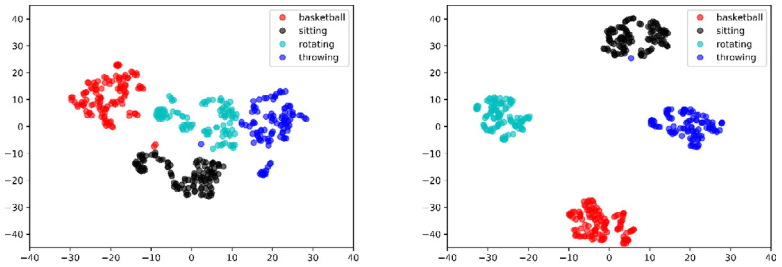

Fig. 8. t-SNE of four different activities captured from first- (a) and third-person (b) views in the joint space (each dot is a video).

Table 6
Average CCA coefficient among first and third view pairs of embeddings for four different classes of activities.

	Class based Activities			
Top *backslash* Front	Complex	Hands & Feet	Vertical Movement	Whole Body
Complex	**0.58**	0.15	-0.07	0.01
Hands & Feet	0.11	**0.62**	-0.01	0.00
Vertical Movement	-0.06	-0.08	**0.43**	-0.01
Whole Body	0.04	-0.06	0.07	**0.52**

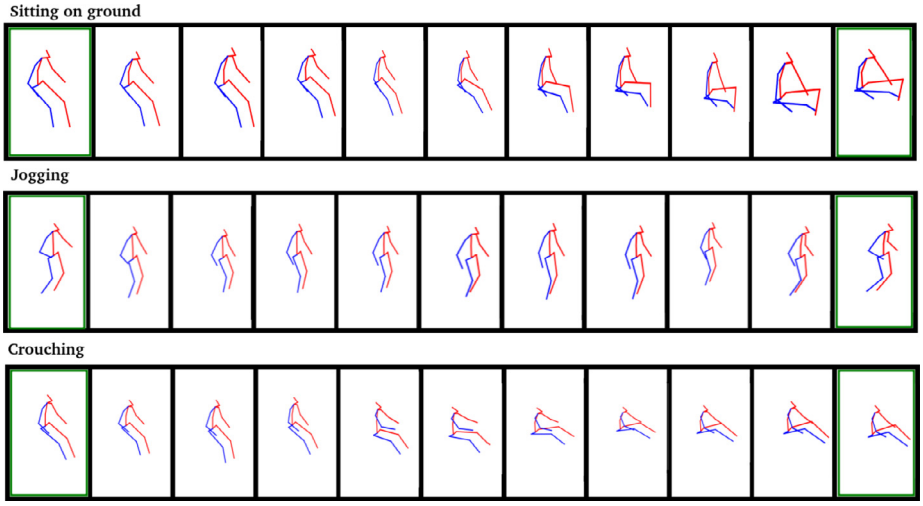

Fig. 9. Examples of transversal for the activities *sit to ground*, *jogging*, and *crouching*. The endpoints are highlighted in green.

we visualize via t-SNE the 1st- and 3rd-views embedding relative to 4 different actions in the joint space (each dot is a video). In both domains, the embedding of different activities are well separated, meaning they are discriminative for activity recognition. This also suggests that the embedded space can be useful for egocentric action recognition.

Furthermore, we examined the relationship between first-view and third-view feature projections in our embedding space by using Canonical Correlation Analysis (CCA) [51]. To make the results more easily interpretable, we grouped the set of activities captured by our dataset into four classes, depending on the type of body movement: 1) short actions involving the whole body (rotating, walking, jogging, etc.); 2) actions requiring a vertical movement of the body or body parts (crouching, sit on the ground, pick object right, etc.); 3) short actions requiring mainly the movement of hands or feet (throwing, bowling, etc.); 4) activities made of a sequence of several short actions (basketball, exercise, etc.). Table 6 reports the CCA coefficient among first-third view pairs of embeddings for each of these four classes. The correlation between first- and third-view embedding is strong, meaning that the features encode relevant geometrical information linking the two views.

Finally, to get insights into the smoothness of the shared representation space, we analyzed several straight Euclidean transversals. As endpoints, we considered two frames of the same video, x_i and x_j and computed their projection onto the joint space, say z_i and z_j. We then obtained the corresponding skeletons p_i and p_j by using a supervised 3D egopose method (DeconvNet) pretrained on the same dataset (*First2Third-Pose*). Afterwards, we obtained points on the same line in the joint embedded space by interpolating between the two endpoint's corresponding latent vectors as $z_t = z_i + (z_j - z_i)\beta$, where β is a real number corresponding to the slope of the line. The input features for the 3D egopose supervised model, say $\Phi(x_t)$, are also obtained as an interpolation from $\Phi(x_i)$ and $\Phi(x_j)$, and fed to the network together with the latent features z_t. In Fig. 9, we present some illustrations. The first and last skeletons, highlighted in green, represent the endpoints. The visualizations of the corresponding skeletons lying on such Eu-

clidean transversal, obtained by incrementing β from zero to one with step 0.1, clearly show that the learned latent space can be considered to a large extent smooth.

7. Conclusion

In this paper, we explored for the first time how to exploit the link between first- and third-view perspectives for the task of egocentric 3D pose estimation. We proposed a versatile framework to build image features that help to discriminate different 3D human poses from egocentric videos even in a target dataset different from the source dataset used to obtain the joint embedded space. Additionally, we built and made publicly available *First2Third-Pose*, a large and synchronized dataset of first- and third-view videos capturing 14 people performing overall 40 different activities. Currently, this is the only 3D pose dataset with synchronized first and third-views videos.

To bridge the heterogeneity gap between the two views, we proposed a self-supervised representation learning approach that learns to transform data samples from different views into a common embedding space, which is subsequently employed to extract features from unpaired and unseen egocentric videos. We provided insights into the structure of the joint learned feature space, through both data visualization and data analytical tools. These insights suggest that the learned feature space well separates different actions and may be therefore potentially useful also for skeleton-based action recognition. We tested our approach on three state-of-the-art methods and two real datasets. Experimental results demonstrated that the joint embedding space learned with *First2Third-Pose* can be used to enhance supervised state-of-the-art egopose estimation methods on different datasets, without the need for domain adaptation or knowledge of camera parameters. Further research will investigate how to further close the gap between first- and third-view, and how to benefit both first- and third-view domains.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data available on the project page: https://github.com/nudlesoup/First2Third-Pose.

Acknowledgments

This work has been partially supported by projects PID2020-120049RB-I00 and PID2019-110977GA-I00 funded by MCIN/ AEI /10.13039/501100011033 and by the “European Union NextGenerationEU/PRTR”, as well as by grant RYC-2017-22563 funded by MCIN/ AEI /10.13039/501100011033 and by “ESF Investing in your future”, and network RED2018-102511-T funded by MCIN/ AEI.

References

[1] M. Kutbi, X. Du, Y. Chang, B. Sun, N. Agadakos, H. Li, G. Hua, P. Mordohai, Usability studies of an egocentric vision-based robotic wheelchair, ACM Trans. Hum.-Robot Interact. (THRI) 10 (1) (2020) 1–23.
[2] M. Dimiccoli, Computer vision for egocentric (first-person) vision, in: Computer Vision for Assistive Healthcare, Elsevier, 2018, pp. 183–210.
[3] H. Liang, J. Yuan, D. Thalmann, N.M. Thalmann, AR in hand: egocentric palm pose tracking and gesture recognition for augmented reality applications, in: Proceedings of the ACM International Conference on Multimedia (ACMMM), 2015, pp. 743–744.
[4] H. Jiang, K. Grauman, Seeing invisible poses: estimating 3D body pose from egocentric video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3501–3509.
[5] Y. Yuan, K. Kitani, 3D ego-pose estimation via imitation learning, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 735–750.
[6] Y. Yuan, K. Kitani, Ego-pose estimation and forecasting as real-time PD control, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 10082–10092.
[7] E. Ng, D. Xiang, H. Joo, K. Grauman, You2Me: inferring body pose in egocentric video via first and second person interactions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9890–9900.
[8] H. Jiang, V.K. Ithapu, Egocentric pose estimation from human vision span, in: Proceedings of the International Conference on Computer Vision (ICCV), 2021, pp. 11006–11014.
[9] J. Wang, L. Liu, W. Xu, K. Sarkar, C. Theobalt, Estimating egocentric 3D human pose in global space, in: Proceedings of the International Conference on Computer Vision (ICCV), 2021, pp. 11500–11509.
[10] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 466–481.
[11] S. Li, W. Zhang, A.B. Chan, Maximum-margin structured learning with deep networks for 3D human pose estimation, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2848–2856.
[12] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, P. Fua, Structured prediction of 3D human pose with deep neural networks, in: Proceedings of the British Machine Vision Conference (BMVC), 2016, pp. 130–141.
[13] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, A. Sharma, A. Jain, Learning 3D human pose from structure and motion, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 668–683.
[14] D. Tome, C. Russell, L. Agapito, Lifting from the deep: Convolutional 3D pose estimation from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2500–2509.
[15] F. Moreno-Noguer, 3D human pose estimation from a single image via distance matrix regression, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2823–2832.
[16] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, N.M. Thalmann, Exploiting spatial-temporal relationships for 3D pose estimation via graph convolutional networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2272–2281.
[17] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, CoRR (2016) arXiv:1609.02907.
[18] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, C. Theobalt, Monocular 3D human pose estimation in the wild using improved CNN supervision, in: Proceedings of the International Conference on 3D Vision (3DV), 2017, pp. 506–516.
[19] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3. 6M: large scale datasets and predictive methods for 3D human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 36 (7) (2013) 1325–1339.
[20] Y. Sun, Y. Ye, W. Liu, W. Gao, Y. Fu, T. Mei, Human mesh recovery from monocular images via a skeleton-disentangled representation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5349–5358.
[21] K. Wang, J. Xie, G. Zhang, L. Liu, J. Yang, Sequential 3D human pose and shape estimation from point clouds, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7275–7284.
[22] Y. Rong, T. Shiratori, H. Joo, FrankMocap: fast monocular 3D hand and body motion capture by regression and integration, in: IEEE International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 1749–1759.
[23] X. Xu, H. Chen, F. Moreno-Noguer, L.A. Jeni, F. De la Torre, 3D human pose, shape and texture from low-resolution images and videos, IEEE Trans. Pattern Anal. Mach.Intell. (TPAMI) (2021) Inpress.
[24] X. Sun, J. Shang, S. Liang, Y. Wei, Compositional human pose regression, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2602–2611.
[25] G. Pavlakos, X. Zhou, K. Daniilidis, Ordinal depth supervision for 3D human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7307–7316.
[26] X. Zhou, X. Sun, W. Zhang, S. Liang, Y. Wei, Deep kinematic pose regression, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 186–201.
[27] U. Iqbal, P. Molchanov, J. Kautz, Weakly-supervised 3D human pose learning via multi-view images in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 5243–5252.
[28] Y. Cai, L. Ge, J. Cai, J. Yuan, Weakly-supervised 3D hand pose estimation from monocular RGB images, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 666–682.
[29] S. Jenni, P. Favaro, Self-supervised multi-view synchronization learning for 3D pose estimation, in: Proceedings of the Asian Conference on Computer Vision (ACCV), 2020, pp. 170–187.
[30] H. Rhodin, M. Salzmann, P. Fua, Unsupervised geometry-aware representation for 3D human pose estimation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 750–767.
[31] C. Li, K.M. Kitani, Model recommendation with virtual probes for egocentric hand detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2624–2631.
[32] C. Li, K.M. Kitani, Pixel-level hand detection in ego-centric videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3570–3577.

[33] G. Rogez, J.S. Supancic, D. Ramanan, First-person pose recognition using egocentric workspaces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4325–4333.
[34] T. Shiratori, H.S. Park, Y. Sheikh, J.K. Hodgins, et al., Motion capture from body mounted cameras, Google Patents, Patent 8,786,680, 2014.
[35] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, P. Fua, H.-P. Seidel, C. Theobalt, Mo2Cap2: real-time mobile 3D motion capture with a cap-mounted fisheye camera, IEEE Trans. Vis. Comput.Graph. (TVCG) 25 (5) (2019) 2093–2101.
[36] D. Tome, P. Peluse, L. Agapito, H. Badino, xR-EgoPose: egocentric 3D human pose from an HMD camera, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019, pp. 7728–7738.
[37] B. Soran, A. Farhadi, L. Shapiro, Action recognition in the presence of one egocentric and multiple static cameras, in: Proceedings of the Asian Conference on Computer Vision (ACCV), 2014, pp. 178–193.
[38] G.A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, K. Alahari, Actor and observer: Joint modeling of first and third-person videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 7396–7404.
[39] R. Yonetani, K.M. Kitani, Y. Sato, Recognizing micro-actions and reactions from paired egocentric videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2629–2638.
[40] S. Bambach, D.J. Crandall, C. Yu, Viewpoint integration for hand-based recognition of social interactions from a first-person view, in: Proceedings of the ACM International Conference on Multimodal Interaction (ICMI), 2015, pp. 351–354.
[41] C. Fan, J. Lee, M. Xu, K. Kumar Singh, Y. Jae Lee, D.J. Crandall, M.S. Ryoo, Identifying first-person camera wearers in third-person videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5125–5133.
[42] Y. Li, T. Nagarajan, B. Xiong, K. Grauman, Ego-Exo: transferring visual representations from third-person to first-person videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6943–6953.
[43] C.-S. Chan, S.-Z. Chen, P.-X. Xie, C.-C. Chang, M. Sun, Recognition from hand cameras: a revisit with deep learning, in: Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 505–521.
[44] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, 2019, (https://github.com/facebookresearch/detectron2).
[45] D. Pavllo, C. Feichtenhofer, D. Grangier, M. Auli, 3D human pose estimation in video with temporal convolutions and semi-supervised training, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7753–7762.
[46] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[47] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, FlowNet 2.0: Evolution of optical flow estimation with deep networks, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2462–2470.
[48] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P.v.d. Smagt, D. Cremers, T. Brox, FlowNet: learning optical flow with convolutional networks, in: IEEE International Conference on Computer Vision (ICCV), 2015.
[49] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, T. Brox, A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimationarXiv:1512.02134
[50] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the International Conference on Machine Learning (ICML), 2009, pp. 41–48.
[51] H. Hotelling, Relations between two sets of variates, in: Breakthroughs in Statistics, Springer, 1992, pp. 162–190.

Ameya Dhamanaskar is currently a master student at Arizona University. He graduated with a BEng in Electrical and Electronics from Birla Institute of Technology and Science, Pilani. During one year (2019–2020), he was a Research Assistant at Institute of Robotics and Industrial Informatics (UPCCSIC). Prior to that, he was a Software Developer at Tesco Technology for one year and a Research Assistant at the Cognitive Computing Group in Central Electronics Engineering Research Institute (CSIR-CEERI), where he worked on improving signature verification for low power devices.

Mariella Dimiccoli is a Ramón y Cajal fellow at the Institute of Robotics and Industrial Informatics (CSIC-UPC). She holds a MSc degree in Computer Engineering from the Polytechnic University of Bari, Italy, and a PhD from the Technical University of Catalonia, Spain. During her career, she worked in renowned research institutions within and outside Europe, including the ENSSaclay, France, the University Pompeu Fabra, the Computer Vision Center, Spain, and the University of Texas at Austin, USA. She was the recipient of several competitive research grants including a Marie-curie cofund fellowship and the prestigious Ramón y Cajal grant. Throughout these experiences, her research interests have revolved around image and video understanding. She received an honorable mention award at ICIP 2019 and a best paper award at IbPRIA 2017.

Enric Corona is currently a PhD student at the Institute of Robotics and Industrial Informatics (CSICUPC) under the supervision of Francesc Moreno- Noguer and Guillem AlenyÀ. Before that, he was a master student at the University of Toronto ans worked as data scientist at IntelliSense.io. In 2019, he was a visiting PhD student at Naver Labs Europe.

Albert Pumarola is a Computer Vision Researcher at Facebook RealityLabs. He received the PhD degree from the Institute of Robotics and Industrial Informatics (CSIC-UPC) in 2021. His research interests are in Computer Vision and Deep Learning, with focus in developing novel generative approaches for photorealistic virtual human avatars. He is a recipient of the Best Paper Award Honorable Mention ECCV'18.

Francesc Moreno-Noguer is a Research Scientist of the Spanish National Research Council at the Institut de Robòtica i Informàtica Industrial. His research interests include the fields of computer vision and machine learning. He received the Polytechnic University of Catalonia's Doctoral Dissertation Extraordinary Award, several best paper awards (e.g. ECCV 2018 Honorable mention, ICCV 2017 workshop in Fashion, Intl. Conf. on Machine Vision applications 2016), outstanding reviewer awards at ECCV 2012, CVPR 2014/2021 and ICCV 2021, and Google and Amazon Faculty Research Awards in 2017 and 2019, respectively. He has (co)-authored over 150 publications in refereed journals and conferences (including 10 IEEE Transactions on PAMI, 5 Intl. Journal of Computer Vision, 25 CVPR, 8 ICCV and 8 ECCV).