Mo^2Cap^2: Real-time *Mo*bile 3D *Mot*ion *Cap*ture with a *Cap*-mounted Fisheye Camera

Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Pascal Fua, *Fellow, IEEE*, Hans-Peter Seidel, and Christian Theobalt, *Member, IEEE*
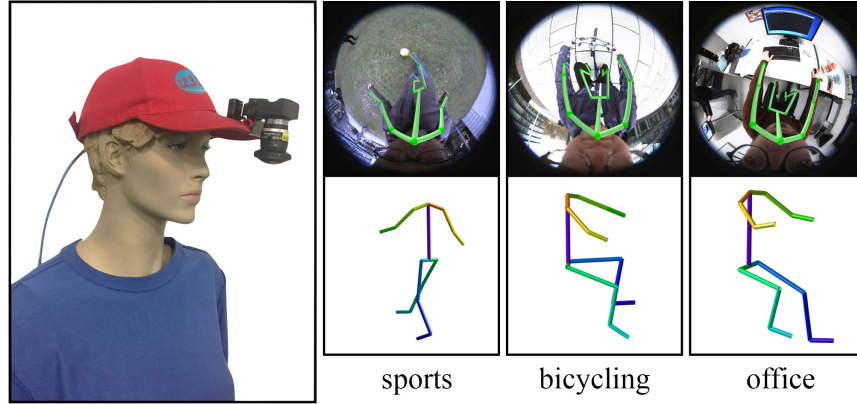
Fig. 1. Our real-time mobile 3D pose estimation system is based on a single monocular cap-mounted fisheye camera that is attached to a standard baseball cap. The setup is lightweight and enables 3D pose estimation in everyday situations.

Abstract— We propose the first real-time system for the egocentric estimation of 3D human body pose in a wide range of unconstrained everyday activities. This setting has a unique set of challenges, such as mobility of the hardware setup, and robustness to long capture sessions with fast recovery from tracking failures. We tackle these challenges based on a novel lightweight setup that converts a standard baseball cap to a device for high-quality pose estimation based on a single cap-mounted fisheye camera. From the captured egocentric live stream, our CNN based 3D pose estimation approach runs at 60 Hz on a consumer-level GPU. In addition to the lightweight hardware setup, our other main contributions are: 1) a large ground truth training corpus of top-down fisheye images and 2) a disentangled 3D pose estimation approach that takes the unique properties of the egocentric viewpoint into account. As shown by our evaluation, we achieve lower 3D joint error as well as better 2D overlay than the existing baselines.

Index Terms— Egocentric, Monocular, Mobile motion capture

---- ✦ ----

1 INTRODUCTION

The goal of this work is to solve the problem of *mobile* 3D human pose estimation in a wide range of activities performed in unconstrained real world scenes, such as walking, biking, cooking, doing sports and office work. The resulting 3D pose can be used for action recognition, motion control, and performance analysis in fields such as sports, animation and health-care. A real-time solution to this problem is also desirable for many virtual reality (VR) and augmented reality (AR) applications.

Such 3D human pose estimation in daily real world situations imposes a unique set of requirements on the employed capture setup and algorithm, such as: mobility, real-time performance, robustness to long capture sequences and fast recovery from tracking failures. In the past, many works for outside-in 3D human pose estimation have been pro-

---

- *Weipeng Xu, Avishek Chatterjee, Hans-Peter Seidel and Christian Theobalt are with the Max-Planck Institute for Informatics. Emails: wxu, achatter, hpseidel, theobalt@mpi-inf.mpg.de.*
- *Michael Zollhöfer is with Stanford University. E-mail: zollhoefer@cs.stanford.edu.*
- *Helge Rhodin and Pascal Fua are with the CVLAB of EPFL. E-mails: helge.rhodin, pascal.fua@epfl.ch.*

posed, which use a single or multiple cameras placed statically around the user [13, 32, 33, 37, 40, 50]. However, daily real world situations make outside-in capture setups impractical, since they are immobile, can not be placed everywhere, require a recording space without occluders in front of the subject, and have only a small recording volume.

Motion capture systems based on body-worn sensors, such as inertial measurement units (IMUs) [63] or multi-camera structure-from-motion (SFM) from multiple limb-mounted cameras [46], support mobile capturing. However, these setups are expensive, require tedious pre-calibration, and often require pose optimization over the entire sequence, which prevents real-time performance. Most closely related to our approach is the EgoCap [39] system that is based on two head mounted fisheye cameras. While it alleviates the problem of a limited capture volume, the setup is quite heavy and requires uncomfortable, obtrusive large extension-sticks. EgoCap also requires dedicated 3D actor model initialization based on keyframes, does not run at real-time rates for the full body, and has not been shown to be robust on very long sequences.

In contrast, we tackle the unique challenges of real-time ubiquitous mobile 3D pose estimation with a novel lightweight hardware setup (see Fig. 1) that converts a standard baseball cap to a device for accurate 3D human pose estimation using a single fisheye camera. Our approach fulfills all requirements mentioned at the outset: 1) Our hardware setup is compact, lightweight and power efficient, which makes it suited for daily mobile use. 2) Our approach requires no actor calibration and works for general and dynamic backgrounds, which enables free roaming during daily activities. 3) From the live stream of the cap-
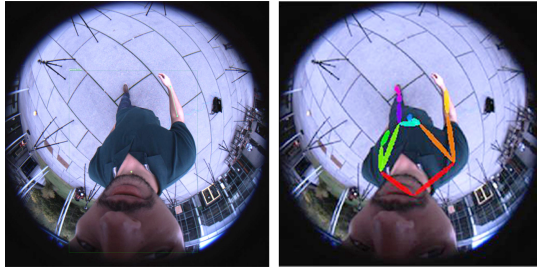
Fig. 2. The state-of-the-art 2D human pose estimator Mask R-CNN [17] trained on the COCO dataset [29] fails on images captured by our setup (left). Our 2D pose estimation results (right).

mounted camera, our approach estimates 3D human pose at 60 Hz. 4) Our online frame-by-frame pose estimation solution is suitable for capturing long sequences and automatically recovers from occasional failures.

As is true for most of the recent outside-in monocular 3D human pose estimation methods, our approach is also based on a deep neural network. However, existing methods do not apply well to our setting. First, their training data is captured with regular cameras and mostly from chest high viewpoints. Thus, they fail on our images, which are captured from a top-down view and exhibit a large radial distortion (see Fig. 2). Second, most of the existing methods directly estimate 3D human pose in the form of 3D joint locations relative to the pelvis and do not respect the 2D-3D consistency. This not only makes them yield bad 2D overlay of 3D pose results on the images, but also makes the 3D pose estimation less accurate, since even a small 2D displacement translates to a large 3D error due to the short focal length of the fisheye camera. Third, the close proximity of the camera to the head creates a strong perspective distortion, resulting in a large upper body and very small lower body in the images, which makes the estimation of the lower body less accurate. To solve these problems, we propose a new ground truth training corpus of top-down fisheye images and, more importantly, a novel 3D pose estimation algorithm based on a CNN that is specifically tailored to the uniqueness of our camera position and optics. Specifically, instead of directly regressing the 3D joint locations, we disentangle the 3D pose estimation problem to the following three subproblems: 1) 2D joint detection from images with large perspective and radial distortions, which is solved with a two-scale location invariant convolutional network, 2) absolute camera-to-joint distance estimation, which is solved with a location sensitive distance module that exploits the spatial dependencies induced by the radial distortion and fixed camera placement relative to the head and 3) recovering the actual joint position by back-projecting the 2D detections using the distance estimate and the optical properties of the fisheye lens. Our disentangled approach leads to not only accurate 3D pose estimation, but also good 2D overlay of results, since, by construction, the 3D joint locations will exactly re-project to the corresponding 2D detections.

To the best of our knowledge, our work is the first approach that performs real-time mobile 3D human pose estimation from a single egocentric fisheye camera. Our qualitative and quantitative evaluations demonstrate that the proposed approach outperforms the baseline methods on our test set. Our datasets and code are publicly available at http://gvv.mpi-inf.mpg.de/projects/wxu/Mo2Cap2/.

2   RELATED WORK

In the following, we categorize relevant motion capture approaches in terms of the employed setup.

Studio and Multi-view Motion Capture   Multi-view motion capture in a studio typically employs ten or more cameras. For marker-based systems the subject has to be instrumented, e.g., with a marker or LED suit. Marker-less motion-capture algorithms overcome this constraint [6, 15, 19, 25, 34, 48, 49, 53, 59, 61], with recent work [2, 7, 13, 37, 40, 42, 50] even succeeding in outdoor scenes and using fewer cameras. The static camera setup ensures high accuracy

but imposes a constrained recording volume, has high setup time and cost, and breaks when the subject is occluded in crowded scenes. On the other hand, mobile hand-held solutions require a team of operators [16, 65, 69]. In contrast, our system does not require any additional operators than the user, a multi-camera setup or complicated synchronization and calibration of multi-camera systems. This makes our system more practical in everyday situations.

Monocular Human Pose Estimation   Monocular human pose estimation is a requirement for many consumer-level applications. For instance, human-computer interaction in living-room environments was enabled by real-time pose reconstruction from a single RGB-D camera [4, 47, 66]. However, active IR-based cameras are unsuitable for outdoor capture in sunlight and their high energy consumption limits their mobile application. Purely RGB-based monocular approaches for capture in more general scenes have been enabled with the advent of convolutional neural networks (CNNs) and large training datasets [11, 20, 44, 62]. Methods either operate directly on images [28, 32, 56, 57, 73], lift 2D pose detections to 3D [5, 10, 21, 68, 74], or use motion compensation and optical flow in videos [1, 58]. The most recent improvements are due to hierarchical processing [38, 60] and combining 2D and 3D tasks [33, 36, 72]. Our approach is inspired by the separation of 2D pose and depth estimation by [72], which, however, assumes an orthographic projection model that does not apply to the strong distortion of our fisheye-lens and is different in that it predicts relative, hip-centered depth instead of absolute distance. While these approaches enable many new applications, the camera is either fixed, which imposes a restricted capture volume, or needs to be operated by a cinematographer that follows the action. We build upon these monocular approaches. We generalize them to a head-mounted fisheye setup and address its unique challenges, such as the special top-down view and the large distortion in the images. The robustness and accuracy is significantly improved compared to the state-of-the-art by a new training dataset and by exploiting the characteristics of the head-mounted camera setup with a disentangled 3D pose estimation approach.

Body-worn Motion Sensors   For some studies, the restricted capture volume of static camera systems is overcome by using inertial measurement units (IMUs) [55, 63] or exoskeleton suits (e.g., METAmotion Gypsy). These form an inside-in arrangement, where the sensors are body-worn and capture body motion independent of external devices. Unfortunately, the sensor instrumentation and calibration of the subject cause long setup times and makes capturing multitudes of people difficult. Furthermore, IMU measurements require temporal integration to obtain position estimates, which is usually addressed by offline batch-optimization to minimize drift globally [63]. We aim at lower setup times and real-time reconstruction with minimal latency, e.g., for interactive virtual reality experiences.

Mobile Motion Capture   Self-contained motion capture in everyday conditions demands for novel concepts. By attaching 16 cameras to the subject's limbs and torso in an inside-out configuration Shiratori et al. recover the human pose by structure from motion on the environment, enabling free roaming in static backgrounds [46]. For dynamic scenes, vision-based inside-in arrangements have been proposed. The camera placement is task specific. Facial expression and eye gaze have been captured with a helmet-mounted camera or rig [24, 54, 64], hand articulation and action from head-mounted [51, 52, 67] or even wrist- or chest-worn cameras [26, 43]. The user's gestures and activity can also be recognized from a first-person perspective [8, 14, 27, 31, 35].

However, capturing accurate full body motion in such a body-mounted inside-in camera arrangement is considerably more challenging, as it is difficult to observe the whole body from such close proximity. Yonemoto et al. propose indirect inference of arm and torso poses from arm-only RGB-D footage [70] and Jiang attempted to reconstruct full-body pose by analyzing the egomotion and observed scene [22], but indirect predictions have low confidence and accuracy. A first approach towards direct full-body motion capture from the egocentric perspective was proposed by Rhodin et al. [39]. A 3D kinematic skeleton model is optimized to explain 2D features in each of the views of

a stereo fisheye camera mounted on head-extensions similar in structure to a selfie stick. While enabling free roaming many application scenarios are hampered by the bulky stereo hardware. This approach achieves interactive framerates for upper body tracking only, while ours enables real-time 3D pose estimation for the full human body. To be less intrusive, we propose a lightweight hardware setup based on a single cap-mounted fisheye camera, which requires an entirely different reconstruction algorithm since the optimization used by Rhodin et al. intrinsically requires stereo vision with a large baseline. Furthermore, our algorithm is capable of estimating 3D pose from a single frame, which reduces the chance of long-term tracking failures and enables capture of arbitrarily long sequences without manual intervention.

3 THE MO²CAP² APPROACH

Mo²Cap² is a real-time approach for mobile 3D human body pose estimation based on a single cap-mounted fisheye camera. Our headgear augments a standard baseball cap with an attached fisheye camera. It is lightweight, comfortable and very easy to put on. However, the usage of only one camera view, the very slanted and proximate viewpoint and the fisheye distortion makes 3D pose estimation extremely challenging. We address these challenges by a novel disentangled 3D pose estimation algorithm based on a CNN that is specifically tailored to our setup. We also contribute a large scale training corpus of synthetic top-down view fisheye images with ground truth annotations. It covers a wide range of body motion and appearance. In the following, we provide more details on these aspects.

3.1 Lightweight Hardware Setup

Our work is the first approach that performs 3D real-time human body pose estimation from a single head-mounted camera. Previous work [39] has demonstrated successful motion capture with a helmet-mounted fisheye stereo pair. While their results are promising, their setup has a number of practical disadvantages. Since they mount each of the cameras approximately 25 cm away from the forehead, the weight of the two cameras translates into a large moment, making their helmet quite uncomfortable to wear. Furthermore, their large stereo baseline of 30-40 cm in combination with the large forehead-to-camera distance forces the actor to stay far away from walls and other objects, which limits usability of the approach in many everyday situations.

In contrast, our setup is based on a single fisheye camera mounted to the brim of a standard baseball cap (see Fig. 1), which leads to a lightweight, comfortable and easy-to-use head-gear. Installed only 8cm away from the head, the weight of our camera (only 175g) translates to a very small moment, which makes our setup practical for many scenarios. Note, there exist even smaller/lighter cameras we could use, without making any algorithmic changes to our method. One could even integrate the small camera inside the brim, which would make the setup even lighter. Such engineering improvements are possible, but beyond the scope of this paper. Our fisheye camera has a 182° field of view in both the horizontal and vertical direction. This allows capturing the full body under a wide range of motion, including fully extended arms. However, our hardware setup also makes 3D pose estimation more challenging since 1) explicit depth is not available in our monocular setup and 2) due to the shorter forehead to camera distance, the body is viewed quite obliquely. Solving 3D pose estimation under such challenging conditions is the key contribution of our paper.

3.2 Synthetic Training Corpus

We now present our egocentric fisheye training corpus that enables training of a deep neural network that is tailored to our unique hardware setup. Capturing a large amount of annotated 3D pose data is already a mammoth task for outside-in setups and it is even harder for egocentric data. Since manual labeling in 3D space is impractical, [39] proposes to use marker-less multi-view motion capture with externally mounted cameras to get 3D annotations.

However, even with the help of such professional motion capture systems, acquiring a large number of annotated real-life training examples for the egocentric viewpoint is still a time consuming and tedious recording task. It requires to capture the training data in a complex

Fig. 3. Example images of our synthetically rendered fisheye training corpus. Our synthetic training corpus features a large variety of poses, human body appearance and realistic backgrounds.

multi-view studio environment and precise 6 DOF tracking of the cap-mounted camera, such that the 3D body pose can be reprojected to the egocentric viewpoint of interest. Furthermore, scalability to general scenes requires foreground/background augmentation, which typically relies on extra effort of capturing with green screen and image segmentation with color keying. Given the difficulty of capturing a large amount of training data, the EgoCap [39] system does not scale to a large corpus of motions and real world diversity of human bodies in terms of shape and appearance, as well as diversity of real scene backgrounds. Furthermore, their dataset cannot be directly used for our method, due to the different camera position relative to the head.

In contrast, we alleviate these difficulties by rendering a synthetic human body model from the egocentric fisheye view. Note that the success of any learning based method largely depends on how well the training corpus resembles the real world in terms of motion, body appearance and environment realism. Therefore, care must be taken to ensure that 1) the variety of motion and appearance is maximized and 2) that the differences between synthetic and real images are minimized. On one hand, to achieve a large variety of training examples, we build our dataset on top of the large scale synthetic human SURREAL dataset [62]. We animate characters using the SMPL body model [30] with uniformly sampled motions from the CMU MoCap dataset [12]. Body textures are chosen randomly from the texture set provided by the SURREAL dataset [62]. In total, we render 530,000 images (see Fig. 3), which encompass around 3000 different actions and more than 700 different body textures. On the other hand, to generate realistic training images, we mimic the camera, lighting and background of the real world scenario. Specifically, images are rendered from a virtual fisheye camera attached to the forehead of the character at a distance similar to the size of the brim of the used real world baseball cap. To this end, we calibrate the real world fisheye camera using the omni-directional camera calibration toolbox *ocamcalib* [45] and apply the intrinsic calibration parameters to the virtual camera. Characters are rendered using a custom shader that models the radial distortion of the fisheye camera. Note that the camera position with respect to the head might change slightly, due to the camera movements and varying wearing angles and positions of the cap. To simulate this effect, we add a random perturbation to the virtual fisheye camera position. Random spherical harmonics illumination is used with a special parameterization to ensure a realistic top down illumination. All images are augmented with the backgrounds chosen randomly from a set of more than 5000 indoor and outdoor ground plane images captured by our fisheye camera. To gather such background images, we attach the fisheye camera to a long stick to obtain images that do not show the person holding the camera. Furthermore, we applied a random gamma correction to the rendered images, such that the network becomes insensitive to the specific photometric response characteristics of the used camera.

Our synthetic dataset contains the ground truth annotation of 2D and 3D joint positions, which can be easily generated using the body model
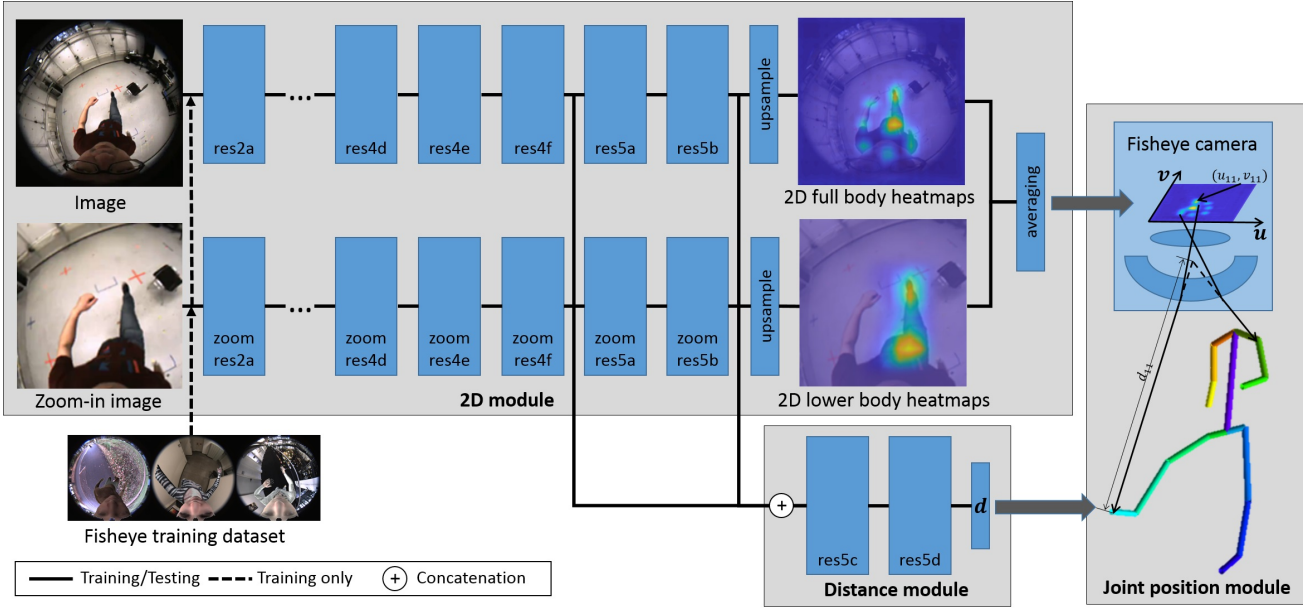
Fig. 4. Our disentangled 3D pose estimation method, which is specifically tailored to our cap-mounted fisheye camera setup, consists of three modules: The two branched **2D module** estimates the 2D joint location heatmaps of the full body from the original image and the lower body from the zoom-in image. The **distance module** estimates the distance between the camera and each joint. The **joint position module** recovers the actual joint position by back-projecting the 2D detection using the joint-to-camera distance estimate and the intrinsic calibration of the fisheye camera.

and the camera calibration. Specifically, we provide the joint positions of the following 15 body joints: *neck, shoulders, elbows, wrists, hips, knees, ankles and toes*. The 3D joint positions are with respect to the fisheye camera coordinate system for our egocentric setup. The joint-to-camera distances are computed based on the 3D joint position (see Sec. 3.3). The 2D joint position annotation is provided in the form of 2D heatmaps. To this end, we first project the ground truth 3D joint positions onto the image space using the camera calibration. The projections are then resized to a resolution of 64×64, and we put a Gaussian kernel of size 5×5 and standard deviation 0.8 at each of the 2D joint positions. Finally, the 2D heatmaps are further downsampled to 32×32.

3.3 Monocular Fisheye 3D Pose Estimation

Our disentangled 3D pose estimation method consists of three modules (see Fig. 4).

The **2D module** of our method estimates 2D heatmaps of the joint locations in image space, where we adopt a fully convolutional architecture that is suited for 2D detection problems. As mentioned before, the strong perspective distortion of our setup makes the lower body appear particularly small in the images and therefore leads to lower accuracy in the estimation of the lower body joints. To solve this problem, we propose a 2D pose estimation module consisting of two independently trained branches, which see different parts of the images. The original scale branch sees the complete images and predicts the 2D pose heatmaps of 15 joints in the full body. The zoom-in branch only sees the $2\times$ zoomed central part of the original images. This zoom-in branch predicts the 2D heatmaps of the 8 lower body joints (hips, knees, ankle and toes), since these joints project into this central region in most of the images captured by our cap-mounted camera. Our zoom-in branch yields more accurate results on the lower body than the original scale branch, since it sees the images at $2\times$ higher resolution. The lower body heatmaps from the two branches are then averaged.

The **distance module** performs a vectorized regression of per-joint absolute camera space depth, i.e., the distance between the camera and each joint, based on the higher and medium level features of the 2D module. In contrast to the fully convolutional architecture of our 2D module, here we use a fully connected layer that can exploit the spatial dependencies in our setup induced by the radial distortion and fixed camera placement relative to the head. Please note that absolute distance estimation is not practical for the classical outside-in

camera setup, where the subject is first cropped in 2D to a normalized pixel scale from which 3D pose is estimated, by which absolute scale information is lost.

At last, the **joint position module** recovers the actual joint position by back-projecting the 2D detections using the distance estimate and the intrinsic calibration (including the distortion-coefficients) of the fisheye camera. To this end, we first read out the u, v coordinates of each joint from the averaged heatmaps. Then, given the calibration of the fisheye camera [45], each 2D joint detection $[u, v]^T$ can be mapped to its corresponding 3D ray vector $[x, y, z]^T$ with respect to the fisheye camera coordinate system:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} u \\ v \\ f(\rho) \end{bmatrix}, \tag{1}$$

where $\rho = \sqrt{u^2 + v^2}$, $f(\rho) = \alpha_0 + \alpha_1 \rho + \alpha_2 \rho^2 + \alpha_3 \rho^3 + \alpha_4 \rho^4 + \dots$ is a polynomial function that is obtained from camera calibration. The 3D position of each joint P is obtained by multiplying the direction vector with the predicted absolute joint-to-camera distance d,

$$P = \frac{d}{\sqrt{x^2 + y^2 + z^2}} [x, y, z]^T. \tag{2}$$

Our disentangled 3D pose estimation method ensures that the 3D joint location will exactly reproject to its 2D detection, handles the scale difference between upper and lower body and leverages location dependent information of the egocentric setup as a valuable depth cue, and therefore results in more accurate 3D pose estimation than previous architectures trained on the same data.

Implementation of our network Each branch of our 2D module consists of 15 residual blocks [18] and performs a deconvolution and two convolutions to upsample the prediction to the heatmap size of 32×32 pixels given images of resolution 256×256 pixels as input. In addition to the euclidean ℓ_2 loss on the final heatmap predictions, we add two additional intermediate supervision losses (after 11 and 14 residual blocks) for faster convergence during training and to prevent vanishing gradients during back-propagation. The architecture of the distance module is based on 2 additional residual blocks, 2 convolution and 1 fully connected layer. We concatenate the output features of the
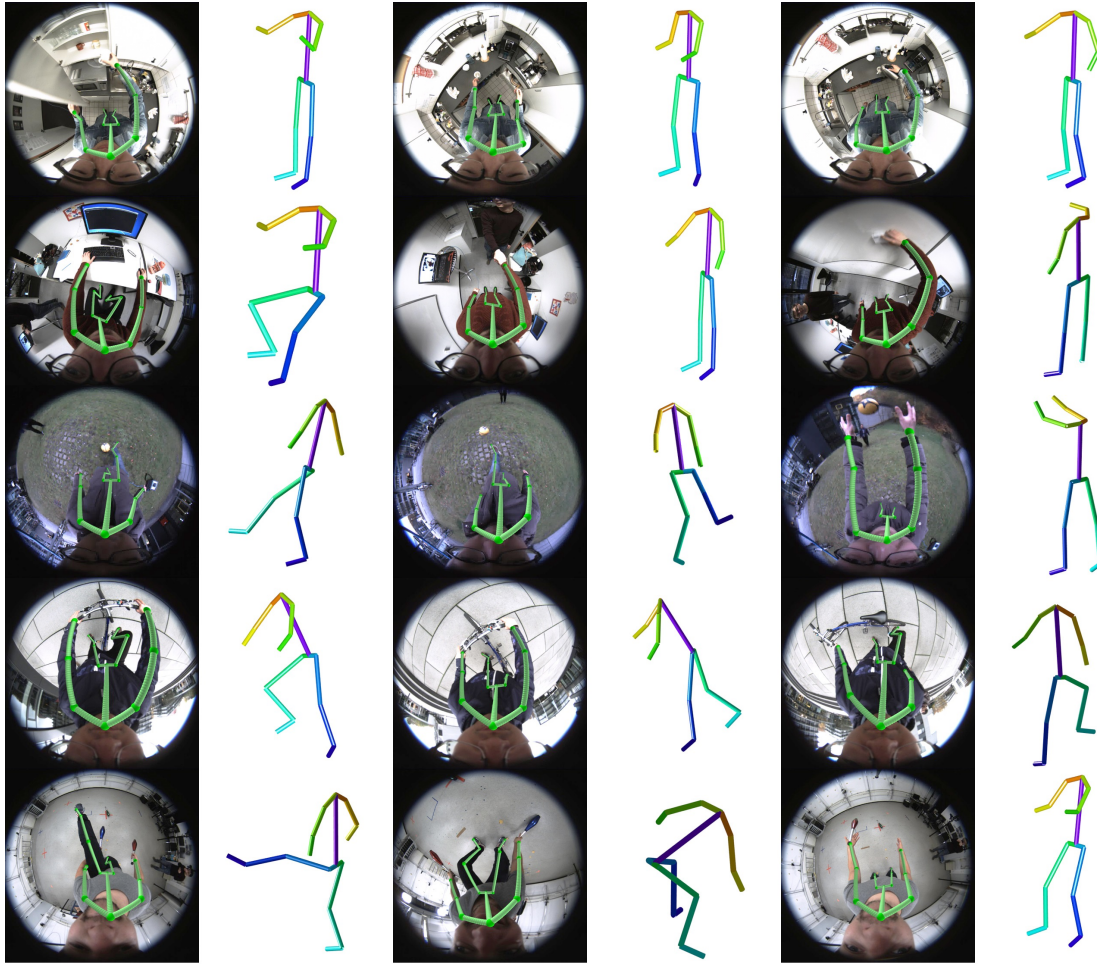
Fig. 5. Results in a variety of everyday situations. Left: our 3D pose results overlaid on the input images; Right: our 3D pose results from a side view.

13th and 15th residual blocks of the two 2D module branches, and pass it to the distance module.

Multi-stage Training Our training corpus is based on synthetically rendered images. To make our network better generalize to real world imagery, we train it in multiple stages using transfer learning. First, we pre-train the 2D module of our network on an outside-in pose estimation task based on the MPII Human Pose [3] and LSP [23] datasets. These real images with normal optics enable our network to learn good low-level features, which, at that feature level, are transferable to our egocentric fisheye setup. Afterwards, we fine tune the two branches of the 2D module separately on the images from our synthetically rendered fisheye training corpus and the $2\times$ zoomed version of them respectively. Note that in order to preserve the low level features learned from real images, we decrease the learning rate multiplier to 0.001 for the initial 13 residual blocks. Afterwards, we fix the weights of the 2D module and train our distance module. The Euclidean loss is used for the final loss and all intermediate losses. In all training stages, we use a batch size of 24, and we train the 2D module for 50k iterations, and the distance module for 70k iterations. For the fine tuning stages, we use a learning rate of 0.05. AdaDelta is used for optimization [71].

4 RESULTS

We study the effectiveness and accuracy of our approach in different scenarios. Our Mo^2Cap^2 system runs at 60 Hz on an Nvidia GTX 1080 Ti, which boils down to 16.7 ms for the forward pass. Thus, our approach can be applied in many applications in which real-time performance is critical, e.g., for motion control in virtual reality.

In the following, we first evaluate our approach qualitatively and quantitatively. Then, we demonstrate that our disentangled 3D human pose estimation approach leads to significant gains in reconstruction accuracy.

4.1 Qualitative Results

Our lightweight and non-intrusive hardware setup allows the users to capture general daily activities. To demonstrate this, we captured a test set of 5 activities, including both everyday and challenging motions, in unconstrained environments: 1) making tea in the kitchen, 2) working in the office, 3) playing football, 4) bicycling, and 5) juggling. Each sequence contains approximately 2000 frames. The sequences cover a large variety of motions and subject appearances (see Fig. 5 for examples). We can see that our method estimates accurate 3D poses for all sequences. Even interactions with other people or objects are captured, which is a challenge even for multi-view outside-in methods. Note that we capture the bicycling and juggling sequences to provide a comparison with the state-of-the-art egocentric 3D pose estimation approach of [39], since they also show results for these two actions. We can see that our monocular method yields comparable and sometimes more stable results than their binocular method. Also note, in contrast to [39], our method runs in real-time on the full body, and does not require 3D model calibration of the user or any optimization as post-process. The complete results on all sequences are shown in the supplementary video.

4.2 Quantitative Results

Existing, widely used data sets for monocular 3D pose estimation, e.g., Human3.6M [20], are designed for outside-in camera perspectives with normal optics, not our egocentric, body-worn fisheye setup. In turn, our absolute distance estimation without image cropping only applies to
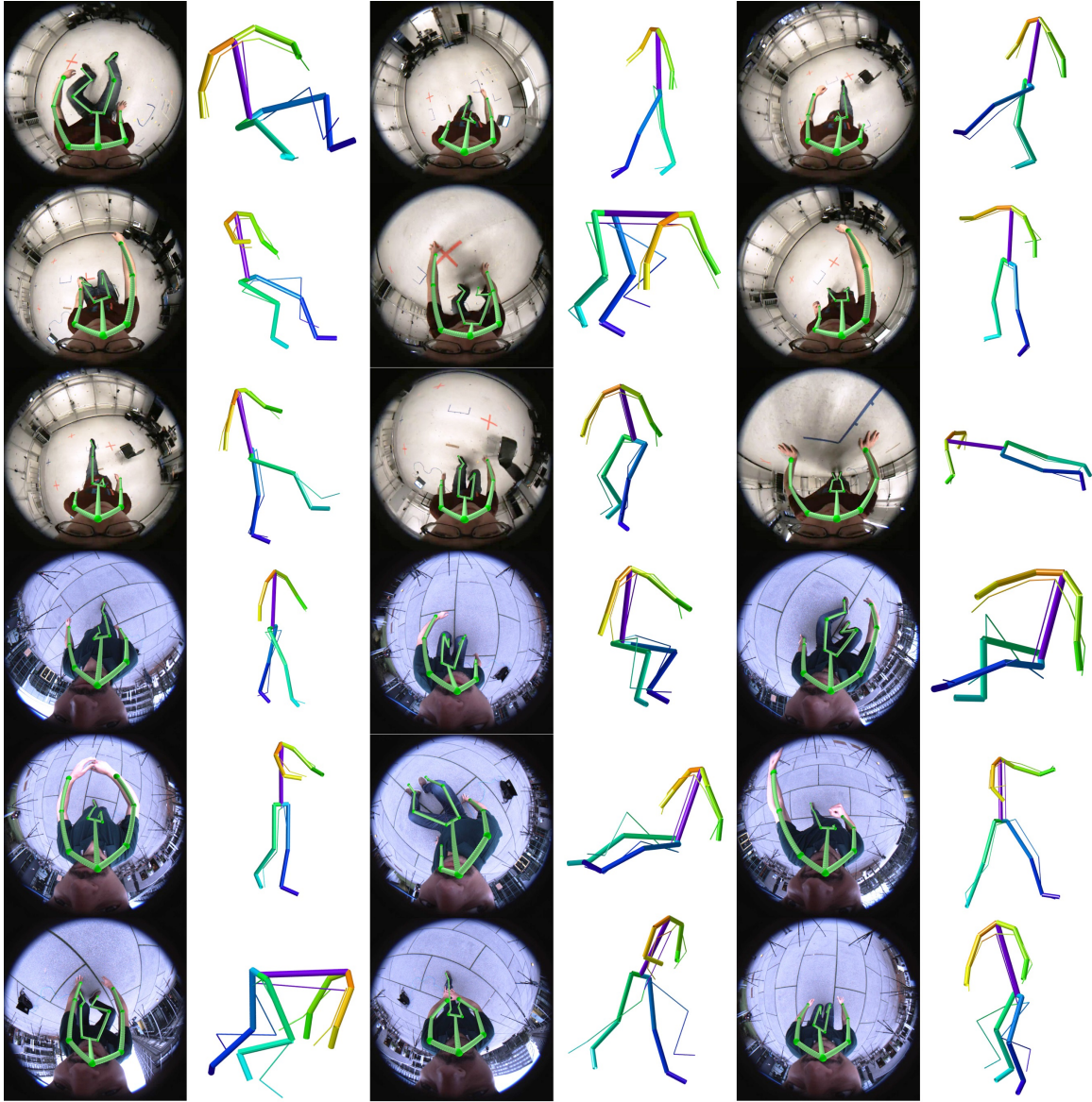
Fig. 6. Results on the indoor and outdoor sequences with ground truth. Left: 3D pose results overlaid on the input images; Right: 3D pose results from a side view, the thinner skeleton is the ground truth obtained using a commercial multi-view motion capture software.

body-mounted scenarios. In order to evaluate our method quantitatively, we therefore captured an extra test set with ground truth annotation containing 8 different actions across 5591 frames, recorded both indoors and outdoors with people in general clothing. The recorded actions include walking, sitting, crawling, crouching, boxing, dancing, stretching and waving. The 3D ground truth is recorded with a commercial external multi-view marker-less motion capture system [9]. Fig. 6 shows our 3D pose results overlaid on the input images (left) and from a side view (right), where the ground truth 3D pose is shown with the thinner skeleton. Since our method does not estimate the global translation and rotation of the body, in order to quantitatively compare our method to the ground truth, we apply Procrustes analysis to register our results to the ground truth. Following many other 3D pose estimation methods [32, 33], we rescale the bone length of our estimated pose to the "universal" skeleton for quantitative evaluation. The average per-joint 3D error (in millimeters) on different actions is shown in Tab. 1. Note that our accuracy is comparable with monocular outside-in 3D pose estimation approaches, even though our setting is much more challenging.

4.3 Influence of the Network Architecture

We also quantitatively compare our disentangled architecture to other state-of-the-art baseline approaches (see Tab. 1) on our egocentric fisheye data. The latter were originally developed for outside-in capture from undistorted camera views. Specifically, we compare to the vectorized 3D body pose prediction network of [32] (referred to as **3DV'17**) and the location map approach used in [33] (referred to as **VNect**). As all three methods are based on a ResNet, we modify their architectures to use the same number of ResNet blocks as ours for a fair comparison. We also apply the same intermediate supervision to all three method and use the same training strategy. We train all networks on our synthetic training corpus of egocentric fisheye images. One can see that our disentangled 3D pose estimation approach outperforms these two state-of-the-art network architectures by a large margin (indoors: 19.5%, outdoors: 14.7% over **3DV'17**) in terms of mean joint error (in mm), see Tab. 1. This demonstrates that our architecture is especially well suited for our monocular fisheye setup. In addition, our disentangled representation leads to good 2D overlay, since the 2D and 3D detections are consistent by construction. A comparison of the 2D overlay results of the three different methods is shown in Fig. 7. One

Table 1. Ground truth comparison on real world sequences. Our disentangled 3D pose estimation approach outperforms the vectorized 3D body pose prediction network of [32] and the location map approach used in [33], which are **trained on our dataset**, in terms of mean joint error (in mm) .

| Indoor	walking	sitting	crawling	crouching	boxing	dancing	stretching	waving	total
3DV'17 [32]	48.7571	101.2177	118.9554	94.9254	57.3380	60.9604	111.3591	64.4975	76.2813
VNect [33]	65.2818	129.5852	133.0847	120.3911	78.4339	82.4563	153.1731	83.9061	97.8454
Ours w/o zoom	47.0895	82.6745	98.9962	87.9168	58.7640	63.6811	109.2848	69.3515	70.1923
Ours w/o averaging	45.8356	77.6024	99.9472	83.8608	55.2959	60.5191	115.7854	66.972	68.1455
Ours	**38.4083**	**70.9365**	**94.3191**	**81.898**	**48.5518**	**55.1928**	**99.3448**	**60.9205**	**61.3977**
Outdoor	walking	sitting	crawling	crouching	boxing	dancing	stretching	waving	total
3DV'17 [32]	68.6660	114.8663	113.2263	118.5457	95.2946	72.9855	144.4816	72.4117	92.4635
VNect [33]	84.4322	167.8719	138.3871	154.5411	108.3584	85.0144	160.5673	96.2204	113.7492
Ours w/o zoom	69.3500	89.1967	99.7597	101.7018	105.7102	74.1185	134.5125	71.2431	87.3114
Ours w/o averaging	67.889	88.7139	99.2919	99.3326	106.3386	72.3075	136.4019	69.0395	86.3061
Ours	**63.1027**	**85.4761**	**96.6318**	**92.8823**	**96.0142**	**68.3541**	**123.5616**	**61.4151**	**80.6366**

| 3DV'17 | VNect | Ours |
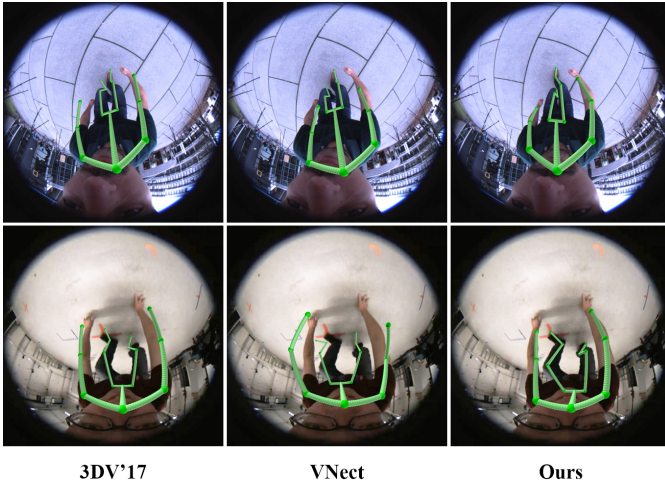
Fig. 7. Comparison of 3D pose results overlaid on the input images. Our results accurately overlay on the images, while the results of the baseline methods exhibit significant offsets.

can see that our 3D pose results accurately overlay on the images, while the results of the baseline methods exhibit significant offsets.

We further perform an ablation study to evaluate the importance of the zoom-in branch of our 2D module. We compare to two incomplete versions of our method: 1) with zoom-in branch completely removed (referred to as **Ours w/o zoom**) and 2) without averaging the heatmaps from the two branches, but only using those from the original scale branch (referred to as **Ours w/o averaging**). We can see from Fig. 8 that, benefiting from the zoom-in branch, our full method yields significantly better overlay of the lower body joints. Quantitatively, our disentangled strategy alone (**Ours w/o zoom**) obtains 6mm (8%) and 5mm (5.5%) improvement over **3DV'17** in the indoor and outdoor scenarios respectively. Using the features from the zoom-in branch in distance estimation (**Ours w/o averaging**) gains an additional improvement of 2mm and 1mm. Using the averaged heatmaps (our full method) yields 7mm (12.5%) and 6mm (7.6%) improvement. This evaluation shows that the 2D-3D consistency obtained by our disentangled strategy and the more accurate 2D prediction from the zoom-in branch are the key contributors to the overall improvement.

4.4 Discussion

We have demonstrated compelling real-time human 3D pose estimation results from a single cap-mounted fisheye camera. Nevertheless our approach still has a few limitations that can be addressed in follow-up work: 1) Similar to all other learning-based approaches, it does not generalize well to data far outside the span of the training corpus. This can be alleviated by extending the training corpus to cover larger variations in motion, body shape and appearance. Since we train on synthetically rendered data, this is easily possible. 2) The reconstruction of 3D body pose under strong occlusions is challenging, since such situations are highly ambiguous, such as when the arms are raised above the head and thus cannot be seen by the camera. In these cases, there are multiple distinct body poses that could give rise to the same observation, thus 3D pose estimation can fail. Fortunately, since our approach works on a per-frame basis, it can recover directly after the occluded parts become visible again. 3) Our per-frame predictions may exhibit some temporal instability, similar to previous single-frame methods. We believe that our approach could be easily extended by adding temporal stabilization as a post-process, or by using a recurrent architecture. Several typical failure cases are shown in Fig. 9. Despite these limitations, we believe, that we took an important step in the direction of real-time ubiquitous mobile 3D motion capture. Our current capture setup conveniently augments a widely used fashion item. In future work, we will explore the design space more broadly and also experiment with other unconventional body-mounted camera locations.

5 APPLICATIONS

Our egocentric 3D pose estimation system can be used for various applications such as action recognition, motion control, and performance analysis. Especially, our system provides a novel natural human-computer-interaction (HCI) solution for recent popular virtual reality (VR) and augmented reality (AR) systems. Specifically, body gesture-based HCI translates the natural movements of the user's body into tangible actions in a virtual world. This allows the users to immerse themselves in a virtual environment and interact with the virtual content more intuitively. Previous solutions typically rely on controllers or outside-in vision-based tracking systems. In contrast, our system provides a compact, inside-in and controller-free solution, which can be integrated in the VR headsets and therefore does not require external tracking devices. In VR games, our system allows the users to control a virtual character with their full body movements [41], instead of only hands for typical controller-based systems. The users will also have a better perception of their full body movements, which is important typically for many sports games. Similarly, the same technology can also be used for sports training or health-care, where the motion capture results can be used for motion analysis, performance monitoring or fall detection. In VR/AR-based telepresence applications, our system can be used to capture the body motion, which can then be used to animate an avatar, without complicated multi-camera motion capture systems. Importantly, benefiting from our mobile egocentric setup, users are not restricted to a fixed recording volume and therefore can roam freely while being captured.

6 CONCLUSION

We proposed the first real-time approach for 3D human pose estimation from a single fisheye camera that is attached to a standard baseball cap. Our novel monocular setup clearly improves over cumbersome existing technologies and is an important step towards practical daily full-body

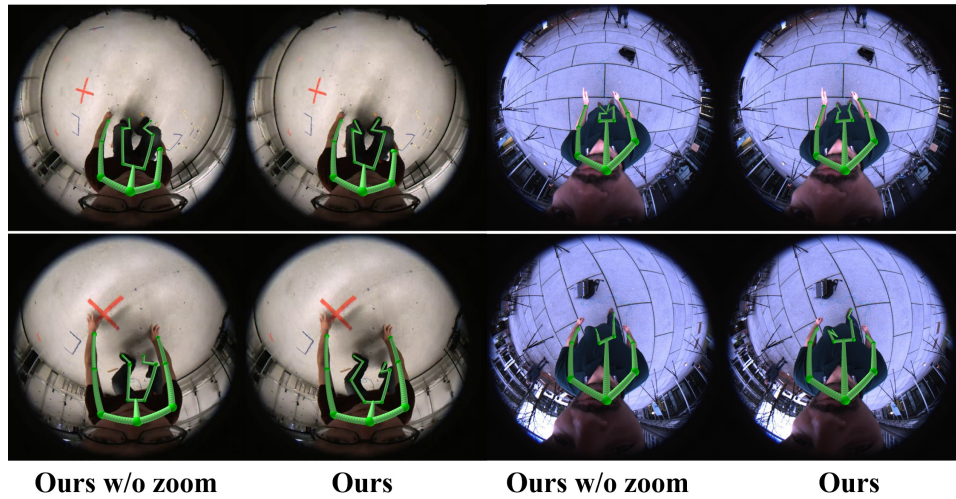| Ours w/o zoom | Ours | Ours w/o zoom | Ours |

Fig. 8. Benefiting from the zoom-in branch, our full method yields significantly better overlay of the lower body joints.
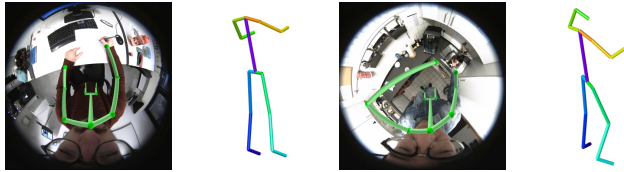

Fig. 9. Failure cases of our method. Left: Our method outputs a standing pose instead of a sitting pose, since the legs are completely occluded. Right: As the left arm is barely visible, our method aligns the arm to the edge of the cupboard.

motion capture. 3D pose estimation is based on a novel 3D pose regression network that is specifically tailored to our setup. We train our network an a new ground truth training corpus of synthetic top-down fisheye images, which we will make publicly available. Our evaluation shows that we achieve lower 3D joint error as well as better 2D overlay than exisiting baseline methods, when applied to the egocentric fisheye setting. We see our approach as the basis for many exciting new applications in several areas, such as action recognition, performance analysis, and motion control in fields such as sports, health-care, and virtual reality.

REFERENCES

[1] T. Alldieck, M. Kassubeck, B. Wandt, B. Rosenhahn, and M. Magnor. Optical flow-based 3d human motion estimation from monocular video. In *German Conference on Pattern Recognition*, pp. 347–360. Springer, 2017.

[2] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3D human pose estimation. In *BMVC*, 2009. doi: 10.5244/C .27.45

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, June 2014. doi: 10.1109/CVPR.2014.471

[4] A. Baak, M. Mller, G. Bharaj, H.-P. Seidel, and C. Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *ICCV*, 2011. doi: 10.1109/ICCV.2011.6126356

[5] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *ECCV*, 2016.

[6] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *CVPR*, Jun 1998. doi: 10.1109/CVPR.1998.698581

[7] M. Burenius, J. Sullivan, and S. Carlsson. 3D pictorial structures for multiple view articulated pose estimation. In *CVPR*, June 2013. doi: 10. 1109/CVPR.2013.464

[8] C. Cao, Y. Zhang, Y. Wu, H. Lu, and J. Cheng. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3763–3771, 2017.

[9] The Captury. http://www.thecaptury.com/.

[10] C.-H. Chen and D. Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *CVPR*, 2016.

[11] W. Chen, H. Wang, Y. Li, H. Su, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3D pose estimation. arXiv:1604.02703, April 2016.

[12] Carnegie Mellon University Motion Capture Database. http://mocap. cs.cmu.edu/.

[13] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *CVPR*, 2015.

[14] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, November 2011. doi: 10.1109/ICCV.2011.6126269

[15] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel. Optimization and filtering for human motion capture. *International Journal of Computer Vision*, 87(1–2):75–92, 2010.

[16] N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall, and H.-P. Seidel. Markerless motion capture with unsynchronized moving cameras. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 224–231. IEEE, 2009.

[17] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask r-cnn. In *ICCV*, pp. 2980–2988, Oct 2017. doi: 10.1109/ICCV.2017.322

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[19] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund. Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments. *IEEE Journal of Selected Topics in Signal Processing*, 6(5):538–552, 2012. doi: 10.1109/JSTSP.2012. 2196975

[20] C. Ionescu, I. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[21] E. Jahangiri and A. L. Yuille. Generating multiple hypotheses for human 3d pose consistent with 2d joint detections. In *ICCV*, 2017.

[22] H. Jiang and K. Grauman. Seeing invisible poses: Estimating 3D body pose from egocentric video. arXiv:1603.07763, 2016.

[23] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011. doi: 10.1109/CVPR.2011. 5995318

[24] A. Jones, G. Fyffe, X. Yu, W.-C. Ma, J. Busch, R. Ichikari, M. Bolas, and P. Debevec. Head-mounted photometric stereo for performance capture. In *CVMP*, 2011. doi: 10.1109/CVMP.2011.24

[25] 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 *ICCV*, December 2015. doi: 10. 1109/ICCV.2015.381

[26] D. Kim, O. Hilliges, S. Izadi, A. D. Butler, J. Chen, I. Oikonomidis, and P. Olivier. Digits: Freehand 3D interactions anywhere using a wrist-worn gloveless sensor. In *UIST*, 2012.

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

[28] S. Li and A. Chan. 3D Human Pose Estimation from Monocular Images with Deep Convolutional Neural Network. In *ACCV*, 2014.

[29] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollr. Microsoft coco: Common objects in context. arXiv:1405.0312, 2014.

[30] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[31] M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *CVPR*, 2016.

[32] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. doi: 10.1109/3dv.2017.00064

[33] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. In *ACM TOG Proc. SIGGRAPH 2017*, vol. 36, July 2017.

[34] T. B. Moeslund, A. Hilton, V. Krger, and L. Sigal, eds. *Visual Analysis of Humans: Looking at People*. Springer, 2011. doi: 10.1007/978-0-85729 -997-0

[35] K. Ohnishi, A. Kanehira, A. Kanezaki, and T. Harada. Recognizing activities of daily living with a wrist-mounted camera. In *CVPR*, 2016.

[36] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[37] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *CVPR*, 2017.

[38] A.-I. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *CVPR*, 2017.

[39] H. Rhodin, C. Richardt, D. Casas, E. Insafutdinov, M. Shafiei, H.-P. Seidel, B. Schiele, and C. Theobalt. Egocap: Egocentric marker-less motion capture with two fisheye cameras. *ACM Trans. Graph.*, 35(6):162:1–162:11, November 2016. doi: 10.1145/2980179.2980235

[40] H. Rhodin, N. Robertini, C. Richardt, H.-P. Seidel, and C. Theobalt. A versatile scene model with differentiable visibility applied to generative pose estimation. In *ICCV*, December 2015. doi: 10.1109/ICCV.2015.94

[41] H. Rhodin, J. Tompkin, K. I. Kim, E. de Aguiar, H. Pfister, H.-P. Seidel, and C. Theobalt. Generalizing wave gestures from sparse examples for real-time character control. *ACM Trans. Graph.*, 34(6):181:1–181:12, October 2015. doi: 10.1145/2816795.2818082

[42] N. Robertini, D. Casas, H. Rhodin, H.-P. Seidel, and C. Theobalt. Model-based outdoor performance capture. In *Proceedings of the 2016 International Conference on 3D Vision (3DV 2016)*, 2016.

[43] G. Rogez, M. Khademi, J. S. Supancic, III, J. M. M. Montiel, and D. Ramanan. 3D hand pose detection in egocentric RGB-D images. In *ECCV Workshops*, 2014.

[44] G. Rogez and C. Schmid. Mocap Guided Data Augmentation for 3D Pose Estimation in the Wild. In *NIPS*, 2016.

[45] D. Scaramuzza, A. Martinelli, and R. Siegwart. A toolbox for easily calibrating omnidirectional cameras. In *IROS*, 2006. doi: 10.1109/IROS. 2006.282372

[46] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins. Motion capture from body-mounted cameras. *ACM Transactions on Graphics*, 30(4):31:1–10, 2011. doi: 10.1145/2010324.1964926

[47] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, 2011.

[48] L. Sigal, A. O. Bălan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, pp. 4–27, 2010. doi: 10.1007/s11263-009-0273-6

[49] L. Sigal, M. Isard, H. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propa-

gation. *International Journal of Computer Vision*, 98(1):15–48, 2012. doi: 10.1007/s11263-011-0493-4

[50] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[51] S. Singh, C. Arora, and C. Jawahar. Trajectory aligned features for first person action recognition. *Pattern Recognition*, 62:45–55, 2017.

[52] S. Sridhar, F. Mueller, A. Oulasvirta, and C. Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, June 2015. doi: 10.1109/CVPR.2015.7298941

[53] C. Stoll, N. Hasler, J. Gall, H.-P. Seidel, and C. Theobalt. Fast articulated motion tracking using a sums of Gaussians body model. In *ICCV*, November 2011. doi: 10.1109/ICCV.2011.6126338

[54] Y. Sugano and A. Bulling. Self-calibrating head-mounted eye trackers using egocentric visual saliency. In *UIST*, 2015.

[55] J. Tautges, A. Zinke, B. Krüger, J. Baumann, A. Weber, T. Helten, M. Müller, H.-P. Seidel, and B. Eberhardt. Motion reconstruction using sparse accelerometer data. *ACM Transactions on Graphics*, 30(3):18:1–12, 2011. doi: 10.1145/1966394.1966397

[56] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua. Structured Prediction of 3D Human Pose with Deep Neural Networks. In *British Machine Vision Conference (BMVC)*, 2016.

[57] B. Tekin, P. Márquez-Neila, M. Salzmann, and P. Fua. Fusing 2D Uncertainty and 3D Cues for Monocular Body Pose Estimation. In *ICCV*, 2017.

[58] B. Tekin, A. Rozantsev, V. Lepetit, and P. Fua. Direct prediction of 3D body poses from motion compensated sequences. In *CVPR*, 2016.

[59] C. Theobalt, E. de Aguiar, C. Stoll, H.-P. Seidel, and S. Thrun. Performance capture from multi-view video. In R. Ronfard and G. Taubin, eds., *Image and Geometry Processing for 3-D Cinematography*, pp. 127–149. Springer, 2010. doi: 10.1007/978-3-642-12392-4_6

[60] D. Tome, C. Russell, and L. Agapito. Lifting From the Deep: Convolutional 3D Pose Estimation From a Single Image. In *CVPR*, 2017.

[61] R. Urtasun, D. J. Fleet, and P. Fua. Temporal motion models for monocular and multiview 3D human body tracking. *Computer Vision and Image Understanding*, 104(2):157–177, 2006.

[62] G. Varol, J. Romero, X. Martin, N. Mahmood, M. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[63] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll. Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. In *Computer Graphics Forum*, vol. 36, pp. 349–360. Wiley Online Library, 2017.

[64] J. Wang, Y. Cheng, and R. S. Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In *CVPR*, 2016.

[65] Y. Wang, Y. Liu, X. Tong, Q. Dai, and P. Tan. Outdoor markerless motion capture with sparse handheld video cameras. *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[66] X. Wei, P. Zhang, and J. Chai. Accurate realtime full-body motion capture using a single depth camera. *ACM Transactions on Graphics*, 31(6):188:1–12, 2012.

[67] W. Wu, C. Li, Z. Cheng, X. Zhang, and L. Jin. Yolse: Egocentric fingertip detection from single rgb images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 623–630, 2017.

[68] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall. A Dual-Source Approach for 3D Pose Estimation from a Single Image. In *CVPR*, 2016.

[69] G. Ye, Y. Liu, N. Hasler, X. Ji, Q. Dai, and C. Theobalt. Performance capture of interacting characters with handheld kinects. In *Computer Vision–ECCV 2012*, pp. 828–841. Springer, 2012.

[70] H. Yonemoto, K. Murasaki, T. Osawa, K. Sudo, J. Shimamura, and Y. Taniguchi. Egocentric articulated pose tracking for action recognition. In *International Conference on Machine Vision Applications (MVA)*, May 2015. doi: 10.1109/MVA.2015.7153142

[71] M. D. Zeiler. Adadelta: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[72] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Weakly-supervised transfer for 3d human pose estimation in the wild. *arXiv preprint arXiv:1704.02447*, 2016.

[73] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep Kinematic Pose Regression. In *ECCV Workshops*, 2016.

[74] X. Zhou, M. Zhu, S. Leonardos, K. Derpanis, and K. Daniilidis. Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In *CVPR*, 2016.