

SelfPose: 3D Egocentric Pose Estimation From a Headset Mounted Camera

Denis Tome¹, Thiemo Alldieck², Patrick Peluse, Gerard Pons-Moll, Lourdes Agapito, Hernan Badino, and Fernando de la Torre

Abstract—We present a new solution to egocentric 3D body pose estimation from monocular images captured from a downward looking fish-eye camera installed on the rim of a head mounted virtual reality device. This unusual viewpoint leads to images with unique visual appearance, characterized by severe self-occlusions and strong perspective distortions that result in a drastic difference in resolution between lower and upper body. We propose a new encoder-decoder architecture with a novel multi-branch decoder designed specifically to account for the varying uncertainty in 2D joint locations. Our quantitative evaluation, both on synthetic and real-world datasets, shows that our strategy leads to substantial improvements in accuracy over state of the art egocentric pose estimation approaches. To tackle the severe lack of labelled training data for egocentric 3D pose estimation we also introduced a large-scale photo-realistic synthetic dataset. *xR-EgoPose* offers 383K frames of high quality renderings of people with diverse skin tones, body shapes and clothing, in a variety of backgrounds and lighting conditions, performing a range of actions. Our experiments show that the high variability in our new synthetic training corpus leads to good generalization to real world footage and to state of the art results on real world datasets with ground truth. Moreover, an evaluation on the Human3.6M benchmark shows that the performance of our method is on par with top performing approaches on the more classic problem of 3D human pose from a third person viewpoint.

Index Terms—3D human pose estimation, egocentric, VR/AR, character animation

1 INTRODUCTION

THE advent of *xR* technologies (such as AR, VR, and MR) has led to a wide variety of applications in areas such as entertainment, communication, medicine, CAD design, art, and workspace productivity. These technologies mainly focus on immersing the user in a virtual space using a head mounted display (HMD) which renders the environment from the specific viewpoint of the user. However, current solutions have so far focused on the video and audio aspects of the user's perceptual system, leaving a gap in the touch and proprioception senses. Partial solutions to proprioception have been limited to the use of controller devices to track and render hand positions in real time. The 3D pose of the rest of the body is then inferred from inverse kinematics of the head and hand poses [1], but this often results in inaccurate estimates of the body configuration with a large loss

of signal that impedes compelling social interaction [2] and even leads to motion sickness [3].

Fig. 1 illustrates the problem that this paper addresses: the goal is to infer 2D and 3D pose information, such as joint positions and rotations, from an egocentric camera perspective, necessary to transfer the motion from the original user to a *generic avatar* or to gather user pose information.

The monocular camera used in our configuration is mounted on the rim of a HMD (as shown in Fig. 1a), approximately 2cm away from an average size nose, looking down. Fig. 2 provides a more clear visualization of the unique visual appearance of the images that the camera sees for different body configurations — the top row shows which body parts would become self-occluded from an egocentric viewpoint. The continuous gradation from bright red to dark green encodes the increasing pixel resolution for the corresponding colored area.

There are several challenges that contribute to the difficulty of this problem: (1) Strong perspective distortions occur, due to the fish-eye lenses and the proximity of the camera to the face. This results in images with strong radial distortion and drastic difference in image resolution between the upper and lower body (as visible in Fig. 2 — bottom row). Due to this, state-of-the-art approaches for 2D body pose estimation [4] from a frontal or 360 degree yaw view, fail on this type of images; (2) There are many instances where body self-occlusion occurs, especially in the lower-body (see right images of Fig. 3), which demands strong spatial awareness of joint locations; (3) Egocentric 3D body pose estimation is a relatively unexplored problem in computer vision, hence the scarce availability of publicly accessible labeled datasets; (4) As shared by traditional 3D

- Denis Tome and Lourdes Agapito are with Department of Computer Science, University College London, WC1E 6BT London, UK. E-mail: {D.tome, l.agapito}@cs.ucl.ac.uk.
- Thiemo Alldieck is with Computer Graphics Lab, Technical University of Braunschweig, 38106 Braunschweig, Germany. E-mail: alldieck@cg.cs.tu-bs.de.
- Patrick Peluse and Hernan Badino are with Facebook Reality Lab, Facebook Inc, Menlo Park, CA 94025. E-mail: {ppeluse, hernan.badino}@fb.com.
- Gerard Pons-Moll is with Perceiving Systems, Max Planck for Intelligent Systems, 70569 Stuttgart, Germany. E-mail: gpons@mpi-inf.mpg.de.
- Fernando de la Torre is with the Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: fforre@cs.cmu.edu.

Manuscript received 21 Feb. 2020; revised 8 Aug. 2020; accepted 5 Oct. 2020. Date of publication 0 . 0000; date of current version 0 . 0000.

(Corresponding author: Denis Tome.)

Recommended for acceptance by M. Farinella, D. Crandall, D. Damen, A. Furnari, and K. Grauman.

Digital Object Identifier no. 10.1109/TPAMI.2020.3029700

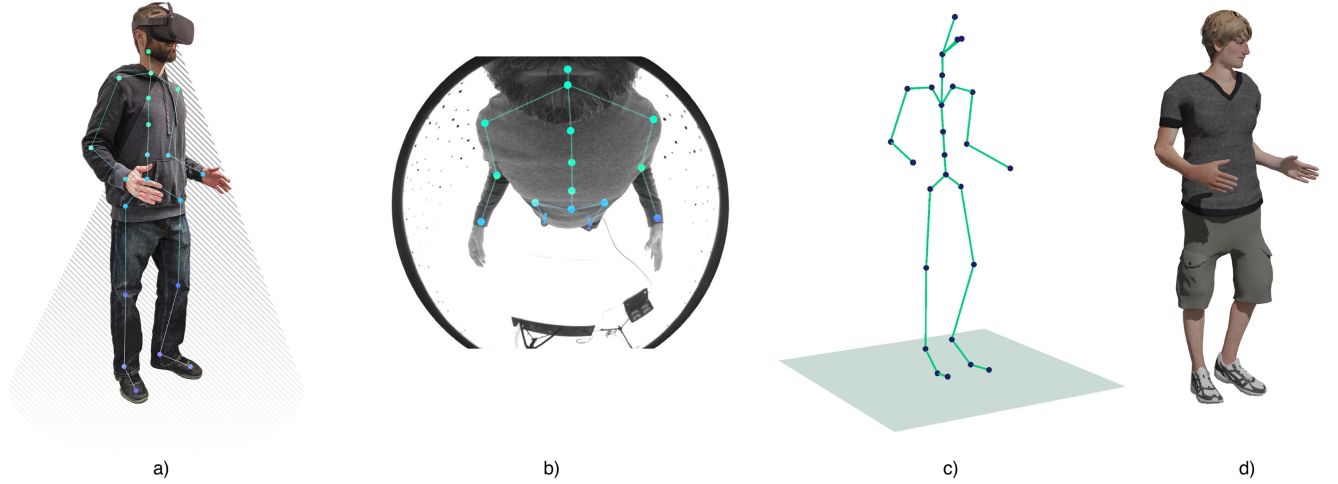


Fig. 1. Egocentric human pose estimation: driving an avatar from an egocentric camera perspective. b) Egocentric perspective of the pose visualized in a) from an external point of view; c) 3D joint locations predicted from the input RGB only-information shown in b); d) synthetic character driven from the local joint rotations estimated alongside the 3D locations.

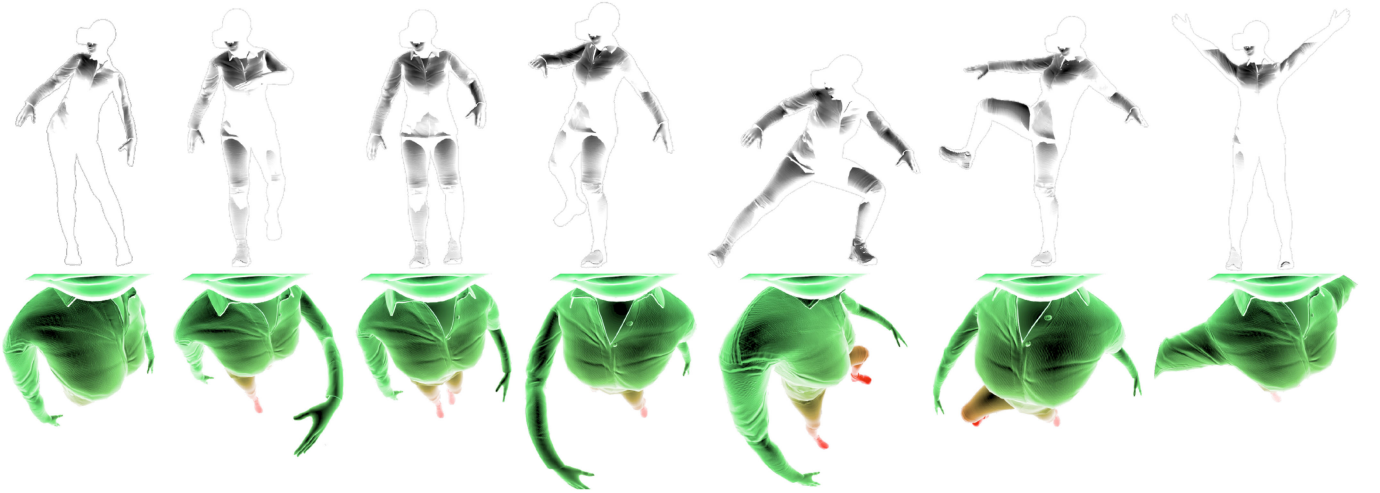


Fig. 2. Visualization of different poses with the same character. *Top*: poses rendered from an external camera viewpoint. White represents occlusion, which is body parts that would not be visible from the egocentric perspective. *Bottom*: poses rendered from the egocentric camera viewpoint. Color gradient indicates the density of image pixels for each area of the body: *green* is higher pixel density, whereas *red* is lower density. This figure illustrates the challenges faced in egocentric human pose estimation: severe self-occlusions, extreme perspective effects and lower pixel density for the lower body.

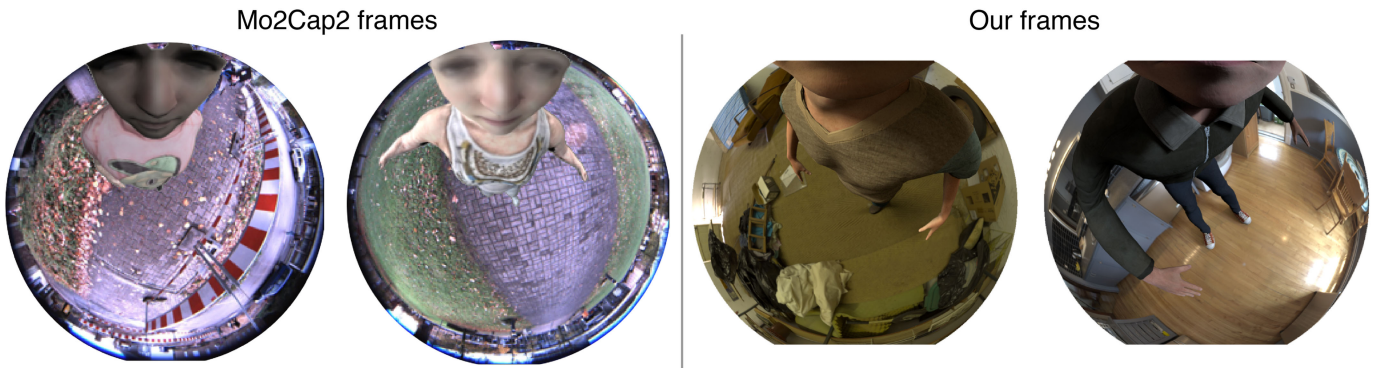


Fig. 3. Example images from our xR -EgoPose Dataset compared with the competitor Mo2Cap2 dataset [5]. The quality of our frames is far superior than the randomly sampled frames from mo2cap2, where the characters suffer color matching with respect to the background light conditions.

body pose estimation, natural ambiguity is present when lifting 2D joint positions in 3D.

The unusual visual egocentric appearance calls for a new approach and a new training corpus. This paper tackles both.

Our novel neural network architecture encodes the difference in uncertainty between upper and lower body joints caused by the varying resolution, extreme perspective effects and self-occlusions.

We conducted quantitative and qualitative evaluations on both synthetic and real-world benchmarks with ground truth 3D annotations, showing that our approach outperforms previous egocentric state-of-the-art Mo²Cap² [5] by *more than 25 percent*. In addition, we achieve state-of-the-art performance on the more standard front-facing cameras 3D human pose reconstruction scenario, without any architecture modifications, performing second best after [6] on the Human3.6M benchmark [7].

Our ablation studies show that the introduction of our novel *multi-branch* decoder to reconstruct the 2D input heatmaps and rotations, is responsible for the drastic improvements in 3D pose estimation. Furthermore, the contribution of each of the branches is analyzed, providing tools to control the level of uncertainty embedded in the latent space.

2 RELATED WORK

We describe related work on monocular (single-camera) marker-less 3D human pose estimation focusing on two distinct capture setups: *outside-in* approaches where an external camera viewpoint is used to capture one or more subjects from a distance – the most commonly used setup; and *first person* or egocentric systems where a head-mounted camera observes the own body of the user. While our paper focuses on the second scenario, we build on recent advances in CNN-based methods for human 3D pose estimation. We also describe approaches that incorporate wearable sensors for first person human pose estimation.

Monocular 3D Pose Estimation From an External Camera Viewpoint. The advent of convolutional neural networks and the availability of large 2D and 3D training datasets [7], [8] has recently allowed fast progress in monocular 3D pose estimation from RGB images captured from external cameras. Two main trends have emerged: (i) fully supervised regression of 3D joint locations directly from images [9], [10], [11], [12], [13], [14] and (ii) pipeline approaches that decouple the problem into the tasks of 2D joint detection followed by 3D lifting [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. Progress in fully supervised approaches and their ability to generalize has been severely affected by the limited availability of 3D pose annotations for in-the-wild images. This has led to significant efforts in creating photo-realistic synthetic datasets [25], [26] aided by the recent availability of parametric dense 3D models of the human body learned from body scans [27]. On the other hand, the appeal of two-step decoupled approaches comes from two main advantages: the availability of high-quality off-the-shelf 2D joint detectors [28], [29], [30], [31] that only require easy-to-harvest 2D annotations, and the possibility of training the 3D lifting step using 3D mocap datasets and their ground truth projections without the need for 3D annotations for images. Even simple architectures have been shown to solve this task with a low error rate [15]. Recent advances are due to combining the 2D and 3D tasks into a joint estimation [4], [32], [33], [34] and using weakly [35], [36], [37], [38], [39] or self-supervised losses [40], [41], [42], [43], [44] or mixing 2D and 3D data for training [6], [42], [45], [46].

First Person 3D Human Pose Estimation. While capturing users from an egocentric camera perspective for activity recognition has received significant attention in recent

years [47], [48], [49], most methods detect, at most, only upper body motion (hands, arms or torso). Capturing full 3D body motion from head-mounted cameras is considerably more challenging. Some head-mounted capture systems are based on RGB-D input and reconstruct mostly hand, arm and torso motions [50], [51]. Jiang and Grauman [52] reconstruct full body pose from footage taken from a camera worn on the chest by estimating egomotion from the observed scene, but their estimates lack accuracy and have high uncertainty. Yuan *et al.* [53], [54] instead explores a different solution by moving away from kinematics-based representations and using a control-based representation of humanoid motion, commonly used in robotics. A step towards dealing with large parts of the body not being observable was proposed in [55] but for external camera viewpoints. Rhodin *et al.* [56] pioneered the first approach towards full-body capture from a helmet-mounted stereo fish-eye camera pair. The cameras were placed around 25 cm away from the user’s head, using telescopic sticks, which resulted in a fairly cumbersome setup for the user but with the benefit of capturing large field of view images where most of the body was in view. A monocular head-mounted systems for full-body pose estimation has more recently been demonstrated by Xu *et al.* [5], who propose a real-time compact setup mounted on a baseball cap, although in this case the egocentric camera is placed a few centimeters further from the user’s forehead than in our proposed approach. Our approach substantially outperforms Xu *et al.*’s method [5] by at least 20 percent on both indoor and outdoor sequences from their real world evaluation dataset. In this journal paper, we go beyond our previous conference paper [57]. First, we perform an extensive analysis on deep architectures for the task of egocentric pose estimation, and show that UNet architectures significantly outperform the originally proposed ResNet architecture [57], specifically for transfer learning from synthetic to real data. Second, we propose a new model which additionally predicts per part rotations. In contrast to [57], this allows us to animate virtual characters, which is necessary for many applications.

3D Pose Estimation From Wearable Devices. Inertial Measurement Units (IMUs) worn by the subject provide a camera-free alternative solution to first person human pose estimation. However, such systems are intrusive and complex to calibrate. While reducing the number of sensors leads to a less invasive configuration [58], [59] recovering accurate human pose from sparse sensor readings becomes a more challenging task. Video data can be fused with IMU [60], [61], [62], [63] to improve accuracy, but these approaches require line of sight with an external camera. An alternative approach, introduced by Shiratori *et al.* [64] consists of a multi-camera structure-from-motion (SFM) approach using 16 limb-mounted cameras. Still very intrusive, this approach suffers from motion blur, automatic white balancing, rolling shutter effects and motion in the scene, making it impractical in realistic scenarios.

3 xR-EGOPOSE SYNTHETIC DATASET

Ego-3D posed estimation from HMC is a relatively new research problem in computer vision, and to the best of our knowledge there is only one dataset available to analyze the algorithms, see Fig. 3. Existing databases are not rich enough

TABLE 1
Total Number of Frames Per Action and Their
Distribution Between Train and Test Sets

Action	N. Frames	Size Train	Size Test
Gaming	24,019	11,153	4,684
Gesticulating	21,411	9,866	4,206
Greeting	8,966	4,188	1,739
Lower Stretching	82,541	66,165	43,491
Patting	9,615	4,404	1,898
Reacting	26,629	12,599	5,104
Talking	13,685	6,215	2,723
Upper Stretching	162,193	114,446	46,468
Walking	34,989	24,603	9,971

Everything else not mentioned is validation data.

to provide statistical significant analysis due to the scarcity of data. This section describes a photo-realistic synthetic egocentric dataset with ground-truth data, that overcomes some of the limitations of existing approaches.

The design of this dataset focuses on scalability, with augmentation of characters, environments, and lighting conditions. A rendered scene is generated from a random selection of characters, environments, lighting rigs, and animation actions. The animations are obtained from mocap data. A small random displacement is added to the positioning of the camera on the headset to simulate the typical variation of the pose of the headset with respect to the head when worn by the user.

Characters. To improve the diversity of body types, from a single character, we generate additional *skinny short*, *skinny tall*, *full short*, and *full tall* versions. The height distribution of ranges from 155 cm to 189 cm.

Skin. Color tones include *white* (Caucasian, freckles or Albino), *light-skinned European*, *dark-skinned European* (darker Caucasian, European mix), *Mediterranean or olive* (Mediterranean, Asian, Hispanic, Native American), *dark brown* (Afro-American, Middle Eastern), and *black* (Afro-American, African, Middle Eastern). Additionally, we built random skin tone parameters into the shaders of each character used with the scene generator.

Clothing. Clothing types include athletic pants, jeans, shorts, dress pants, skirts, jackets, T-Shirts, long sleeves, and tank tops. Shoes include sandals, boots, dress shoes, athletic shoes, crocs. Each type is rendered with different texture and colors.

Actions. The type of actions are listed in Table 1.

Images. The images have a resolution of 1024×1024 pixels and 16-bit color depth. For training and testing, we downsample the color depth to 8 bit. The frame rate is 30 fps. *RGB*, *depth*, *normals*, *body segmentation*, and *pixel world position* images are generated for each frame, with the option for exposure control for augmentation of lighting. Metadata is provided for each frame including 3D joint positions, height of the character, environment, camera pose, body segmentation, and animation rig.

Render Quality. Maximizing the photo-realism of the synthetic dataset was our top priority. Therefore, we animated the characters in Maya using actual mocap data [65], and used a standardized physically based rendering setup with V-Ray. The characters were created with global custom shader settings applied across clothing, skin, and lighting of environments for all rendered scenes.

3.1 Training, Test, and Validation Sets

The dataset has a total size of 383K frames, with 23 male and 23 female characters, divided into three sets: *Train-set*: 252K frames; *Test-set*: 115K frames; and *Validation-set*: 16K frames. The gender distribution is: *Train-set*: 13M/11F, *Test-set*: 7M/5F and *Validation-set*: 3M/3F. Table 1 provides a detailed description of the partitioning of the dataset according to the different actions.

4 ARCHITECTURE

This section describes the deep learning architecture for 3D pose estimation. The proposed architecture (Fig. 4), is a two step approach consisting of two main modules: i) the first module detects 2D heatmaps of the locations of the body joints in image space. We experiment with different standard architectures, please refer to Section 5 for details;

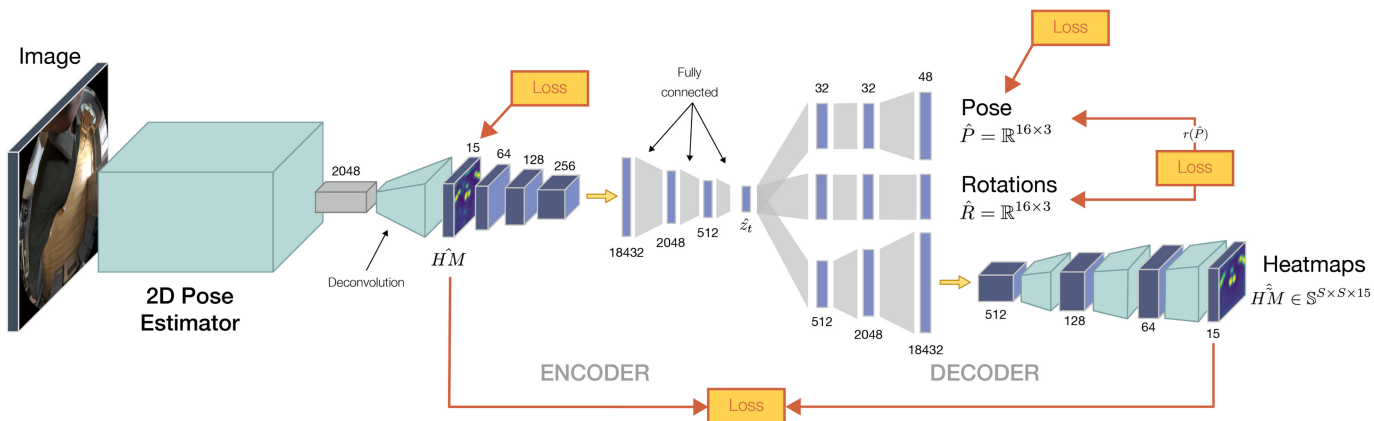


Fig. 4. Proposed architecture for egocentric 3D human pose estimation consisting of two modules: a) interchangeable 2D pose detector that predicts heatmaps from the input RGB image; b) multi-branch auto-encoder that finds a representation of poses which includes also a level of uncertainty of predictions per joint. Alongside the main branch, for 3D joint location prediction, two auxiliary branches as used at training-time to improve latent space distribution. Branch ii) estimates local joint rotations, forcing them to be consistent with those rotations extracted by the predicted pose from i); branch iii) forces the latent space to include a level of uncertainty of the 2D joint locations by reconstructing the given predicted heatmaps from the pose embedding. These additional branches have demonstrated considerable improvements with respect to a standard AE architecture, as shown in Section 5.

ii) the second one takes as input the 2D heatmap predictions generated from the preceding module and regresses the 3D coordinates of the body joints, local joint rotations according to the skeleton hierarchy and reconstructed heatmap predictions, using a novel *multi-branch auto-encoder* architecture.

One of the most important advantages of this pipeline approach is that 2D and 3D modules can be trained independently according to the available training data. For instance, if a sufficiently large corpus of images with 3D annotations is not available, the 3D lifting module can be trained independently using 3D mocap data and its projected heatmaps. Once the two modules are pretrained the entire architecture can be fine-tuned end-to-end since it is fully differentiable. The *multi-branch* auto-encoder module gives also the ability of having multiple representations of the pose: e.g., joint positions, local rotations, etc. A further advantage of this architecture is that the second and third branches are only needed at training time (see Section 4.2) and can be removed at test time, guaranteeing the better performance and a faster execution.

4.1 2D Pose Detection

Given an RGB image $\mathbf{I} \in \mathbb{R}^{368 \times 368 \times 3}$ as input, the 2D pose detector infers 2D poses, represented as a set of heatmaps $\mathbf{HM} \in \mathbb{R}^{47 \times 47 \times 15}$, one for each of the body joints. For this task we have experimented with different standard architectures including *ResNet 50* [66] and *U-Net* [67]. For a detailed analysis, please refer to Section 5.

The models were trained using normalized input images, obtained by subtracting the mean value and dividing by the standard deviation, and using the mean square error of the difference between the ground truth heatmaps and the predicted ones as the loss

$$L_{2D} = \text{mse}(\mathbf{HM}, \widehat{\mathbf{HM}}). \quad (1)$$

4.2 2D-to-3D Mapping

The 3D pose module takes as input the 15 heatmaps computed by the first module and outputs the final 3D pose $\mathbf{P} \in \mathbb{R}^{16 \times 3}$ as a set of joint locations. Note that the number of output 3D joints is 16 since we include the head which despite being out of the field of view it can be regressed in 3D.

In most pipeline approaches the *3D lifting module* usually is given as input the 2D joint pixel positions in the image of the detected the 3D position. Instead, similarly to Pavlakos *et al.* [39], our approach predicts the 3D pose from input heatmaps, not just 2D locations. The main advantage is that these heatmaps carry important information relative to the *uncertainty of the 2D* pose estimations. Furthermore, due to the unique architecture, it is possible to change the different levels or representation of a pose, afterwards.

The main novelty of the proposed architecture (see Fig. 4), is that we ensure that the uncertainty information expressed in the heatmap representations does not get lost but it is preserved in the pose embedding. While the encoder takes as input a set of heatmaps and encodes them into the embedding $\hat{\mathbf{z}}$, the decoder has multiple branches – *1st* regresses the 3D pose from $\hat{\mathbf{z}}$; *2nd* estimates the local joint rotations (with respect to the parent node); and *3rd* reconstructs the input heatmaps. The purpose of this branch is to

force the latent vector to encode the probability density function of the estimated 2D heatmaps.

The overall loss function for the auto-encoder is expressed as

$$L_{AE} = \lambda_p (\|\mathbf{P} - \hat{\mathbf{P}}\|^2 + W(\mathbf{P}, \hat{\mathbf{P}})) + \lambda_r \|\hat{\mathbf{R}} - r(\hat{\mathbf{P}})\|^2 + \lambda_{hm} \|\widehat{\mathbf{HM}} - \widetilde{\mathbf{HM}}\|^2, \quad (2)$$

with \mathbf{P} the ground truth; $\hat{\mathbf{R}}$ the predicted local joint rotations and $r(\hat{\mathbf{P}})$ the function that estimates local joint rotations from a given pose; $\widehat{\mathbf{HM}}$ is the set of heatmaps regressed by the decoder from the latent space and $\widetilde{\mathbf{HM}}$ are the heatmaps regressed by 2D pose estimator module (see Section 4.1). Different local joint rotation representations were tested and ultimately a Quaternion representation was chosen due to the stability of the rotations during training, leading to more robust models. The rotation branch also helps generating better results as shown in Section 5 with smoother transitions on consecutive frames on poses estimated frame-by-frame.

Finally W is the regularizer over the 3D poses

$$W(\mathbf{P}, \hat{\mathbf{P}}) = \lambda_\theta \theta(\mathbf{P}, \hat{\mathbf{P}}) + \lambda_L L(\mathbf{P}, \hat{\mathbf{P}}),$$

with

$$\theta(\mathbf{P}, \hat{\mathbf{P}}) = \sum_l \frac{\mathbf{P}_l \cdot \hat{\mathbf{P}}_l}{\|\mathbf{P}\| * \|\hat{\mathbf{P}}_l\|} \quad L(\mathbf{P}, \hat{\mathbf{P}}) = \sum_l \|\mathbf{P}_l - \hat{\mathbf{P}}_l\|,$$

corresponding to the cosine-similarity error and the limb-length error, with $\mathbf{P}_l \in \mathbb{R}^3$ the l th limb of the pose. An important advantage of this loss is that the model can be trained on a mix of 3D and 2D datasets simultaneously: if an image sample only has 2D annotations then $\lambda_p = 0$ and $\lambda_r = 0$, such that only the heatmaps are contributing to the loss. In Section 5.7 we show how having a larger corpus of 2D annotations can be leveraged to improve final 3D body pose estimates.

4.3 Training Details

The model has been trained on the entire training set for 3 epochs, with a learning rate of $1e-3$ using batch normalization on a mini-batch of size 16. The deconvolutional layer used to identify the heatmaps from the features computed by *ResNet* has kernel size = 3 and stride = 2. The convolutional and deconvolutional layers of the encoder have kernel size = 4 and stride = 2. Finally, all the layers of the encoder use leaky ReLU as activation function with 0.2 leakiness. The λ weights used in the loss function were identified through grid search and set to $\lambda_{hm} = 10^{-3}$, $\lambda_p = 10^{-1}$, $\lambda_r = 10^{-1}$, $\lambda_\theta = -10^{-2}$ and $\lambda_L = 0.5$. The model has been trained from scratch with Xavier weight initializer.

5 EXPERIMENTAL EVALUATION

In the following, we thoroughly evaluate our proposed approach on our novel *xR-EgoPose* dataset, we perform parameter and architecture ablations, and we evaluate on the real-world Mo²Cap² test-set [5] which includes 2.7K frames of real images with ground truth 3D poses of two

TABLE 2
Quantitative Evaluation With Martinez *et al.* [15], a State-of-the-Art Approach Developed for Front-Facing Cameras

Approach	Gaming	Gesticulating	Greeting	Lower Stretching	Patting	Reacting	Talking	Upper Stretching	Walking	All (mm)
Martinez [15]	109.6	105.4	119.3	125.8	93.0	119.7	111.1	124.5	130.5	122.1
Ours — p3d	138.3	108.5	100.3	133.3	117.8	175.6	93.5	129.0	131.9	130.4
Ours — p3d+rot	110.7	90.9	91.9	119.1	98.6	106.8	86.9	88.0	88.2	91.2
Ours — p3d+hm	56.0	50.2	44.6	51.1	59.4	60.8	43.9	53.9	57.7	58.2
Ours — p3d+hm+rot	60.4	54.6	44.7	56.5	57.7	52.7	56.4	53.6	55.4	54.7

Both upper and lower body reconstructions are shown as well. A comparison with our own architecture where different configurations are analyzed. Specifically, the impact of the additional branches is evaluated. Note how the competing approach fails consistently across different actions in lower body reconstructions. This experiment emphasizes how, even a state-of-the-art 3D lifting method developed for external cameras fails on this challenging task. It also emphasizes the contribution of encoding uncertainty for achieving low-reconstruction errors.

people captured in indoor and outdoor scenes. In addition, we show qualitative results on our controlled small-scale real-world dataset and demonstrate how our approach can be used to animate virtual characters for xR telepresence. Finally, we evaluate quantitatively on the Human3.6M dataset to show that our architecture generalizes well without any modifications to the case of an external camera viewpoint.

Evaluation Protocol. Unless otherwise mentioned, we report the Mean Per Joint Position Error - MPJPE

$$E(\mathbf{P}, \hat{\mathbf{P}}) = \frac{1}{N_f} \frac{1}{N_j} \sum_{f=1}^{N_f} \sum_{j=1}^{N_j} \|\mathbf{P}_j^{(f)} - \hat{\mathbf{P}}_j^{(f)}\|_2, \quad (3)$$

where $\mathbf{P}_j^{(f)}$ and $\hat{\mathbf{P}}_j^{(f)}$ are the 3D points of the ground truth and predicted pose at frame f for joint j , out of N_f number of frames and N_j number of joints.

To ensure high reproducibility of our results on our novel synthetic xR -EgoPose dataset, we first evaluate our method on a randomly initialized *ResNet 50*. We intentionally do not perform any pre-training strategies given that, as we show in Section 5.3, this affects the final results. Our goal is to establish our xR -EgoPose dataset as a benchmark and therefore report reproducible numbers that have been computed using a standard network architecture, trained with a simple protocol, cf. Section 4.3.

5.1 Evaluation on Our Egocentric Synthetic Datasets

Evaluation on xR -EgoPose Test-Set. First, we evaluate our approach on the test-set of our synthetic xR -EgoPose dataset. We show qualitative results in Fig. 9. Unfortunately, it was not possible to establish a comparison on our dataset with state of the art monocular egocentric human pose estimation methods such as Mo^2Cap^2 [5] given that their code has not been made publicly available. Instead we compare with Martinez *et al.* [15], a recent state of the art method for a traditional external camera viewpoint. For a fair comparison, the training-set of our xR -EgoPose dataset has been used to re-train the model of Martinez *et al.* This way we can directly compare the performance of the 2D to 3D modules.

Table 2 reports the MPJPE (Eq. (3)) for both methods showing that our approach (Ours-dual-branch) outperforms Martinez *et al.*'s by 36.4 percent in the upper body reconstruction, 60 percent in the lower body reconstruction, and 52.3 percent overall, showing a considerable improvement.

Reconstruction Errors Per Joint Type. Table 3 reports a decomposition of the reconstruction error into different

individual joint types. The highest errors are in the hands and feet. This observation is in accordance with the fact that hands and feet are often not or only barely visible. Hands can go out of the camera field of view e.g., by lifting or stretching the arms or may be occluded by the body. Feet are only visible when the subject looks slightly down and always cover only a very small portion of the image, due to the strong distortion. Nevertheless, our method always predicts plausible poses, even for high occlusions as displayed in Figs. 6, 9 and 10.

Effect of the Decoder Branches. Table 2 reports an ablation study to compare the performance of three versions of our approach. We report results using: i) only 3D pose supervision only (Ours — p3d); ii) additional supervision on regressed rotations (Ours — p3d+rot); iii) and on regressed heatmaps (Ours — p3d+hm); finally for our novel *multi-branch* auto-encoder supervised on all three signals (Ours — p3d+hm+rot).

The overall average error of the single branch encoder is 130.4 mm, far from the 54.7 mm error achieved by our novel *multi-branch* architecture. The dual branch encoders produce an error of 91.2 mm and 58.2 mm, respectively. Ours results clearly demonstrate that all branches contribute to our final result. Both, forcing the network to encode uncertainty of the 2D joint estimates by regressing heatmaps, as well as preserving the limb orientation information by regressing rotations, helps to estimate better 3D poses.

Encoding Uncertainty in the Latent Space. Fig. 5 demonstrates the ability of our approach to encode the uncertainty of the input 2D heatmaps in the latent vector. Examples of input 2D heatmaps and those reconstructed by the second branch of the decoder are shown for comparison.

TABLE 3
Average Reconstruction Error Per Joint Using Eq. (3), Evaluated on the Entire Test-Set (see Section 3) With Model Trained Using Only Synthetic Data

Joint	Error (mm)	Joint	Error (mm)
Left Leg	34.33	Right Leg	33.85
Left Knee	62.57	Right Knee	61.36
Left Foot	70.08	Right Foot	68.17
Left Toe	76.43	Right Toe	71.94
Neck	6.57	Head	23.20
Left Arm	31.36	Right Arm	31.45
Left Elbow	60.89	Right Elbow	50.13
Left Hand	90.43	Right Hand	78.28

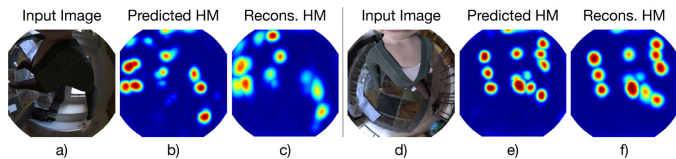


Fig. 5. Reconstructed heatmaps generated by the decoder branch which can reproduce the correct uncertainty of the 2D input predictions from the pose embedding.

5.2 Character Animation Using Estimated Rotations

The pose embedding estimation generated by the *multi-branch* auto-encoder architecture contains the relevant essential information of a pose, which grants us the ability to change / add a representation based on a specific application. Specifically, the introduction of the rotation branch improves the overall reconstruction error, as demonstrated in Table 2, and it is a pose definition usable for character animation.

The joint rotations estimated by the rotation-branch are expressed as local-rotations of each joint with respect to the parent node according to the skeleton hierarchy. Several rotation representations have been tested, including Euler angles, Rotation Matrices, Quaternions and the approach proposed by Zhou *et al.* [68]. We have not noticed any relevant improvements between Quaternions and [68], however the latter demands a larger number of components per joint to express rotations.

Example frames showing the *driven character* compared against the original animation are shown in Fig. 6. Notice

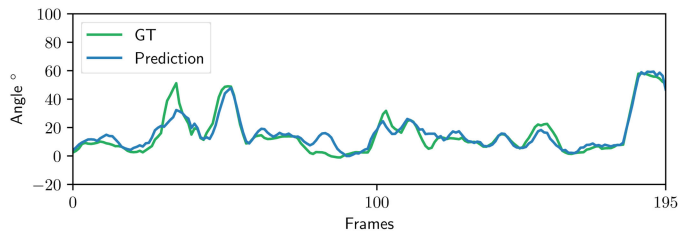


Fig. 7. Analysis of the angle predictions through time for the Right Foot in sequence of the test-set.

how the model is able to reliably estimate the correct rotations even for poses where the avatar’s limbs fall outside of the camera’s field-of-view. Furthermore, there is temporal consistency between poses in consecutive frames despite estimations being computed frame-by-frame.

Fig. 7 shows joint angle predictions, estimated from input images, through time. Specifically, joint angles are consistent with the ground truth. The rotations are smooth and limited “jittering” artefacts are introduced by the network in the predictions.

5.3 Heatmap Estimation: Architecture Ablation

So far, we have used the established *ResNet 50* [66] architecture in all our experiments. In order to study the effect of the heatmap estimation network, we experiment with different architectures and initialization strategies. Specifically, we experiment with *ResNet 50* [66] and *U-Net* [67]. We use *ResNet 50* in two variants: randomly initialized using Xavier

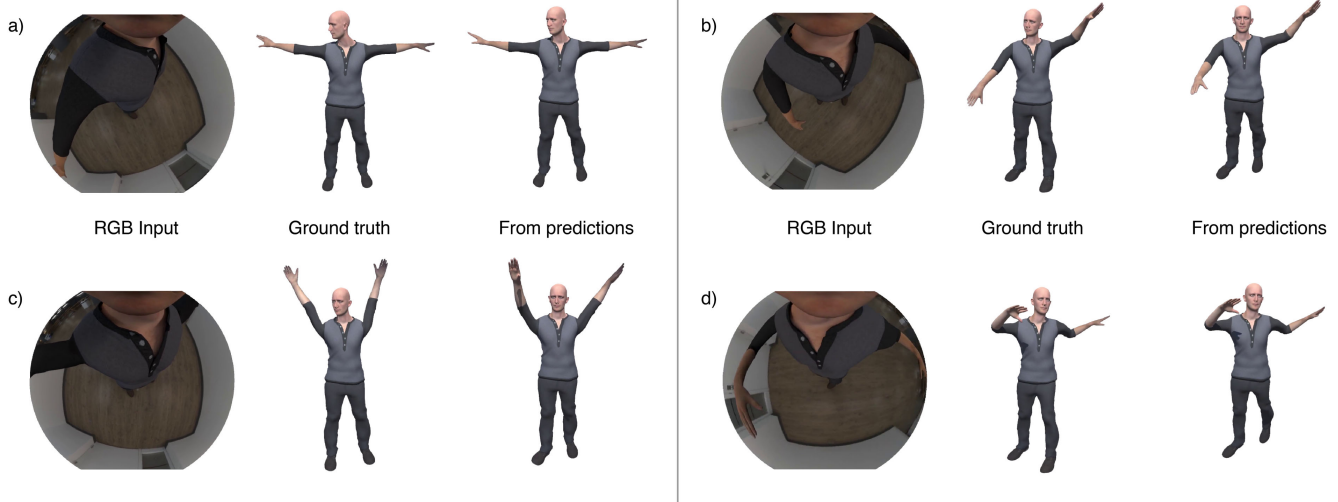


Fig. 6. Character animation from the joint local rotation predictions computed from the input image. Notice how the model is able to retrieve most of the desired information even when limbs fall outside the camera field of view.

TABLE 4
Performance Analysis: Different Combinations of 2D Pose Detectors Combined With the *Multi-Branch* Lifting Network

Configuration	Gaming	Gesticulating	Greeting	Lower Stretching	Patting	Reacting	Talking	Upper Stretching	Walking	All (mm)
ResNet 50	60.4	54.6	44.7	56.5	57.7	52.7	56.4	53.6	55.4	54.7
ResNet 50 (p)	51.6	44.6	64.6	52.4	50.8	44.0	46.5	51.4	52.8	51.1
U-Net (p)	52.5	49.2	72.0	37.3	53.0	44.4	46.1	39.3	37.2	41.0

All variants have been trained and tested on the synthetic dataset. Variants with (p) have been pre-trained on ImageNet.

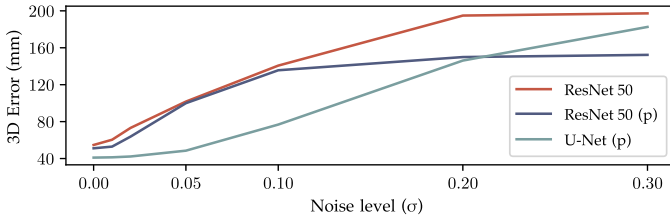


Fig. 8. Performance of our proposed pipeline using different 2D pose estimation networks under the influence of white Gaussian noise in the image domain. Networks with (p) have been pretrained on ImageNet.

initialization [69] and pre-trained on ImageNet [70]. The *U-Net* is composed from a *ResNet 18* backbone encoder, pre-trained on ImageNet, and a randomly initialized decoder. The *ResNet 50* consists of 24.2 million trainable parameters. The *U-Net* contains 18.3 million parameters. All variants produce the same heatmap resolution for better comparison. The lifting networks share the same architecture and number of parameters, but have been trained specifically for each 2D pose estimation network, to accommodate its unique heatmap properties. We additionally experimented with *ResNet 101* [66], *Convolutional Pose Machines* [28], and *Stacked Hourglass Network* [29]. These experiments resulted in comparable performance at a higher computational cost compared to *ResNet 50*, and are therefore not further discussed.

Our experiments suggest that pre-training helps. The full pipeline using a pre-trained *ResNet 50* improves the MPJPE error to 51.1 mm, compared to 54.7 for random initialization, see Table 4. While a recent work [71] suggests that pre-training usually is not necessary, the authors describe two aspects where pre-training does help. First, pre-training helps faster convergence. Second, for small datasets, pre-training helps to improve accuracy. While our synthetic dataset is large, it features less variability in scenes and subjects, compared to large real-world datasets like e.g., MPII [8].

In a next step, we experiment using a *U-Net* for 2D pose estimation. Using a *U-Net* architecture boosts the performance of our pipeline and significantly improves the MPJPE error to 41.0 mm. Empirically, we found that the *U-Net*-based 2D pose estimator also generalizes, to a certain extent, to real data, predicting plausible heatmaps for unseen data, while only having been trained on our synthetic dataset. The *Resnet 50*-based estimator fails without prior refinement. We hypothesize, that the improved performance, and the observed behavior on real images, demonstrate better generalization properties of the *U-Net*. To support our hypothesis, we

TABLE 5
Average Reconstruction Error Per Joint Using Eq. (3), Evaluated on the Entire Test-Set When the Model Architecture Differs Based on the Size of the Embedding \hat{z}

\hat{z} size	Error (mm)		
	ResNet50	ResNet50 (p)	UNet (p)
10	70.6	61.0	45.8
20	67.3	52.5	45.3
50	54.7	51.1	41.0
70	55.7	54.5	41.6
100	58.9	54.2	41.3
500	61.0	56.0	41.2

perform an additional experiment. We add white Gaussian noise to the test images of our synthetic dataset and measure the performance of our pipeline using the different 2D pose estimation networks. In Fig. 8 we plot the MPJPE error under various levels of noise. Notably, the error of the *U-Net*-based pipeline increases slowly, while *Resnet 50*-based pipelines produce large errors already under small noise levels. This behavior supports our hypothesis that the *U-Net* architecture features better generalization properties.

5.4 Lifting Network: Parameter Ablation

In order to validate the architecture design choices of our *multi-branch* 3D pose lifting network, we perform an ablation study of two main parameters.

First, we find the optimal size of the embedding \hat{z} , that encodes the 3D pose, the joint rotations, and the 2D pose uncertainty. Table 6 lists the MPJPE error using different sizes for \hat{z} for all three different heatmap estimation networks. Regardless of the choice of the heatmap estimation network, we find that $\hat{z} \in \mathbb{R}^{50}$ produces the best results. Smaller embeddings produce significantly higher errors, while larger embeddings only slightly impair the results.

Further, we study how the dimensions of the regressed heatmaps \mathbf{HM} influence the results, see Table 5. Unsurprisingly, we find that regressing the full heatmap produces the best results. This is in accordance with the experiments in Section 5.5, where we show that encoding uncertainty via regressing heatmaps helps over using them only as input.

To contribute towards fostering fairness in Computer Vision and Machine Learning we analyze the performance of the proposed models on our diverse dataset based on different skin tones. A comparison is shown in Table 7.

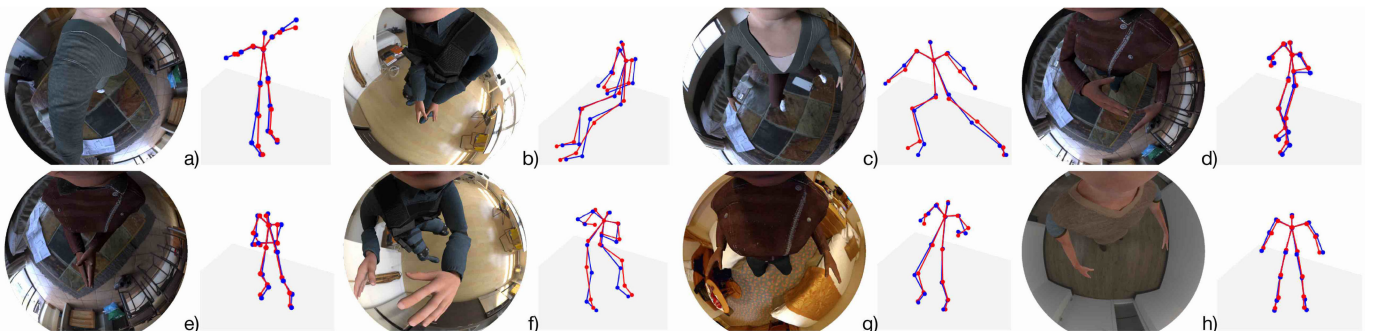


Fig. 9. Qualitative results on synthetic images from our synthetic test-set. Notice that the poses are expressed with respect to the camera reference system. Blue poses represent ground truth, whereas poses in red correspond to predictions.

TABLE 6
Average Reconstruction Error Per Joint Using Eq. (3), Evaluated on the Entire Test-Set for Different Heatmap (HM) Reconstruction Sizes

HM size	Error (mm)		
	ResNet50	ResNet50 (p)	UNet (p)
48	54.7	51.1	41.0
36	57.8	59.6	44.2
24	59.9	57.7	43.8
16	61.2	56.8	41.4
8	61.4	56.7	41.7

Notice how little uncertainty information still has dramatic impact on the reconstruction accuracy.

5.5 Evaluation on Egocentric Real Datasets

Comparison With Mo²Cap² [5]. We compare the results of our approach with those given by our direct competitor, Mo²Cap², on their real world test set including both indoor and outdoor sequences. For a fair comparison, we train our model solely on their provided synthetic training data (cf. Fig. 3). Table 8 reports the MPJPE errors for both methods. Our dual-branch approach substantially outperforms Mo²Cap² [5] in both indoor and outdoor scenarios. Here again, our approach using the *U-Net* model pre-trained on ImageNet produces the best results. However, indoors in a more controlled setting, both our architecture variants are almost on par. Note that comparison with the stereo egocentric system EgoCap [56] on their dataset is not meaningful, due to the hugely different camera position relative to the head (their stereo cameras are 25 cm from the head).

Evaluation on xR-EgoPose^R. The $\sim 10K$ frames of our small scale real-world data set were captured from a fish-eye camera mounted on a VR HMD worn by three different actors wearing different clothes, and performing 6 different actions. The ground truth 3D poses were acquired using a custom mocap system. The network was trained on our synthetic corpus (xR-EgoPose) and fine-tuned using the data from two of the actors. The test set contained data from the unseen third actor. xR-EgoPose^R is too small for meaningful numerical evaluation. However, we show qualitative examples of the input views and the reconstructed poses in Fig. 10. These results show good

TABLE 7
Model Evaluation Based on Skin Tones

Skin tone	Error (mm)		
	ResNet50	ResNet50 (p)	UNet (p)
White	42.7	46.5	46.3
Light European	61.9	58.2	43.5
Dark European	63.6	52.0	35.6
Dark brown	22.5	28.7	27.5
Black	89.0	68.8	42.7

generalization of the model (trained mostly on synthetic data) to real images.

5.6 Evaluation on Front-Facing Cameras

Comparison on Human3.6M Dataset. We show that our proposed approach is not specific for the egocentric case, but also provides excellent results in the more standard case of front-facing cameras. For this evaluation, we chose the Human3.6M dataset [7], [73]. We used two evaluation protocols. *Protocol 1* has five subjects (S1, S5, S6, S7, S8) used in training, with subjects (S9, S11) used for evaluation. The MPJPE error is computed on every 64th frame. *Protocol 2* contains six subjects (S1, S5, S6, S7, S8, S9) used for training, and the evaluation is performed on every 64th frame of Subject 11 (Procrustes aligned MPJPE is used for evaluation). The results are shown in Table 9 from where it can be seen that our approach is on par with state-of-the-art methods, scoring second overall within the non-temporal methods.

5.7 Mixing 2D and 3D Ground Truth Datasets

An important advantage of our architecture is that the model can be trained on a mix of 3D and 2D datasets simultaneously: if an image sample only has 2D annotations but no 3D ground truth labels, the sample can still be used, only the heatmaps will contribute to the loss. We evaluated the effect of adding additional images with 2D but no 3D labels on both scenarios: egocentric and front-facing cameras. In the egocentric case we created two subsets of the xR-EgoPose test-set. The first subset contained 50 percent of all the available image samples with both 3D and 2D labels. The

TABLE 8
Quantitative Evaluation on Mo²Cap² Dataset [5], Both Indoor and Outdoor Test-Sets

INDOOR	walking	sitting	crawling	crouching	boxing	dancing	stretching	waving	total (mm)
3DV'17 [14]	48.76	101.22	118.96	94.93	57.34	60.96	111.36	64.50	76.28
VCNet [72]	65.28	129.59	133.08	120.39	78.43	82.46	153.17	83.91	97.85
Xu [5]	38.41	70.94	94.31	81.90	48.55	55.19	99.34	60.92	61.40
Ours - ResNet 50	38.39	61.59	69.53	51.14	37.67	42.10	58.32	44.77	48.16
Ours - U-Net (p)	45.83	47.24	47.35	45.15	48.72	47.00	46.15	46.45	46.61
OUTDOOR	walking	sitting	crawling	crouching	boxing	dancing	stretching	waving	total (mm)
3DV'17 [14]	68.67	114.87	113.23	118.55	95.29	72.99	114.48	72.41	94.46
VCNet [72]	84.43	167.87	138.39	154.54	108.36	85.01	160.57	96.22	113.75
Xu [5]	63.10	85.48	96.63	92.88	96.01	68.35	123.56	61.42	80.64
Ours - ResNet 50	43.60	85.91	83.06	69.23	69.32	45.40	76.68	51.38	60.19
Ours - U-Net (p)	53.96	52.24	55.50	55.65	54.38	54.48	54.46	56.12	54.61

Our approach outperforms all competitors by more than 21.6 percent (13.24 mm) on indoor data and more than 25.4 percent (20.45 mm) on outdoor data when using only the provided synthetic data for training the model. Similarly to other experiments we provide in Section 5, when using a pre-trained U-Net model with the configuration defined as in Section 5.3, results improve even further: 24.9 percent (14.79 mm) and 32.28 percent (26.03 mm) respectively.

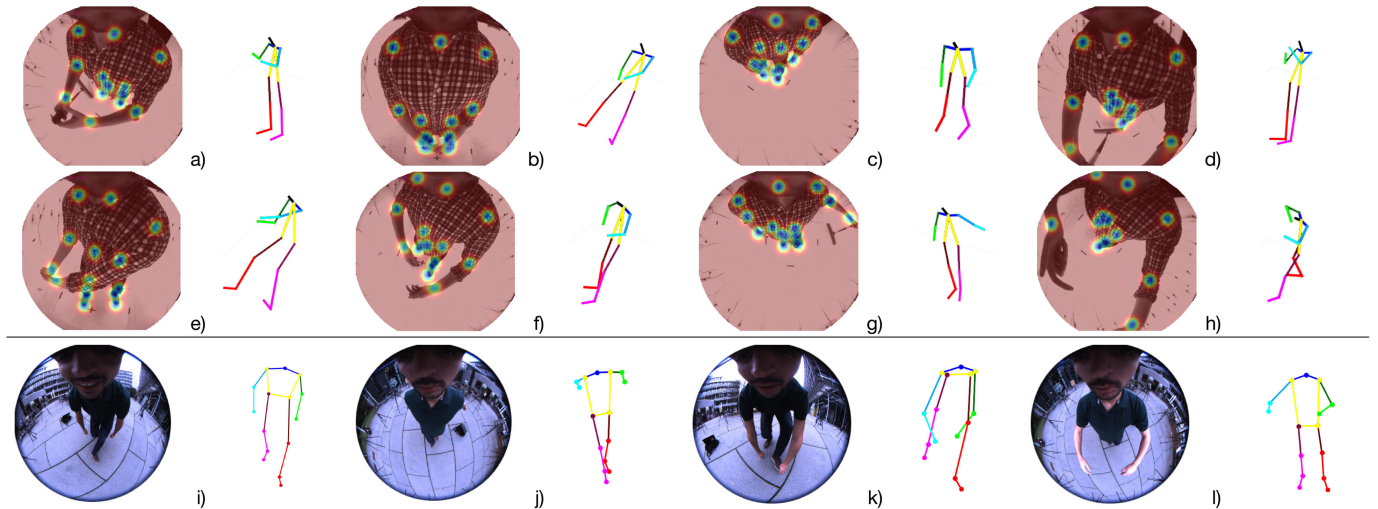


Fig. 10. Qualitative results on real images captured in a studio (top) and reconstructions of images in the wild from Mo²Cap² [5] (bottom). Notice that the poses are expressed with respect to the camera reference system.

TABLE 9
Comparison With Other State-of-the-Art Approaches on the Human3.6M Dataset (Front-Facing Cameras)

Protocol #1	Chen [74]	Hossain [75]*	Dabral [76]*	Tome [36]	Moreno [16]	Kanazawa [77]	Zhou [78]	Jahangiri [79]	Mehta [14]	Martinez [15]	Fang [80]	Sun [81]	Sun [6]	Ours
Errors (mm)	114.2	51.9	52.1	88.4	87.3	88.0	79.9	77.6	72.9	62.9	60.4	59.1	49.6	51.3
Protocol #2	Yasin [82]	Hossain [75]*	Dabral [76]*	Rogez [25]	Chen [74]	Moreno [16]	Tome [36]	Zhou [78]	Martinez [15]	Kanazawa [77]	Sun [81]	Fang [80]	Sun [6]	Ours
Errors (mm)	108.3	42.0	36.3	88.1	82.7	76.5	70.7	55.3	47.7	58.8	48.3	45.7	40.6	42.3

Approaches with * make use of temporal information. No specific modifications have been applied to our architecture: UNet 2D pose detector pre-trained on ImageNet has been used to estimate joint-heatmaps fed through our dual-branch auto-encoder architecture, since rotation information is not available for these data.

second contained 100 percent of the image samples with 2D labels, but only 50 percent of the 3D labels. Effectively the second subset contained twice the number of images with 2D annotations only. Table 10 a compares the results between the subsets, where it can be seen that the final 3D pose estimate benefits from additional 2D annotations. Equivalent behavior is seen on the Human3.6M dataset. Table 10 b shows the improvements in reconstruction error when additional 2D annotations from COCO [83] and MPII [8] are used.

6 CONCLUSION

We have presented a solution to the problem of 3D body pose estimation from a monocular camera installed on a HMD. Given a single image, our fully differentiable network estimates heatmaps and uses them as an intermediate representation to regress 3D poses using a novel *multi-*

branch auto-encoder. This new architecture design was fundamental for accurate reconstructions in our challenging dataset, with over 24 percent accuracy improvement on competitor datasets and that proves to generalize to the more generic 3D human pose estimation from front-facing cameras task with state-of-the-art performance. We have shown how the proposed architecture can be used to drive a virtual avatar directly from the estimations of the network, a fundamental step towards telepresence in virtual or augmented reality.

Finally, we have also introduced the *xR-EgoPose* dataset, a new large scale photo-realistic synthetic dataset that was essential for training and will be made publicly available to promote research in this exciting area. While our results are state-of-the-art, there are a few failures cases due to extreme occlusion and the inability of the system to measure hands when they are out of the field of view. Adding additional cameras to cover more field of view and enable multi-view sensing is the focus of our future work.

REFERENCES

- [1] <https://medium.com/@DeepMotionInc/how-to-make-3-point-tracked-full-body-avatars-in-vr-34b3f6709782>. How to make 3 point tracked full-body avatars in VR, <https://medium.com/@deepmotioninc/how-to-make-3-point-tracked-full-body-avatars-in-vr-34b3f6709782>. Accessed: Mar. 19, 2019. [Online]. Available: <https://medium.com/@DeepMotionInc/how-to-make-3-point-tracked-full-body-avatars-in-vr-34b3f6709782>

TABLE 10

Having a Larger Corpus of 2D Annotations can be Leveraged to Improve Final 3D Pose Estimation

3D	2D	Error (mm)	Training dataset	Error (mm)
50%	50%	68.04	H36M	67.9
50%	100%	63.98	H36M + COCO + MPII	53.4

(a) *xR-EgoPose*

(b) Human3.6M

- [2] U. Hess, K. Kafetsios, H. Mauersberger, C. Blaison, and C.-L. Kessler, "Signal and noise in the perception of facial emotion expressions: From labs to life," *Pers. Soc. Psychol. Bull.*, vol. 42, no. 8, pp. 1092–1110, 2016.
- [3] J. T. Reason and J. J. Brand, *Motion Sickness*. Cambridge, MA, USA: Academic Press, 1975.
- [4] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net++: Multi-person 2D and 3D pose detection in natural images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1146–1161, 2020.
- [5] W. Xu et al., "Mo²Cap²: Real-time mobile 3D motion capture with a cap-mounted fisheye camera," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 5, pp. 2093–2101, May 2019.
- [6] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 529–545.
- [7] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.
- [8] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3686–3693.
- [9] S. Li and A. B. Chan, "3D human pose estimation from monocular images with deep convolutional neural network," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 332–347.
- [10] S. Park, J. Hwang, and N. Kwak, "3D human pose estimation using convolutional neural networks with 2D pose information," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016, pp. 156–169.
- [11] B. Tekin, I. Katircioglu, M. Salzmann, V. Lepetit, and P. Fua, "Structured prediction of 3D human pose with deep neural networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 130.1–130.11. [Online]. Available: <https://dx.doi.org/10.5244/C.30.130>
- [12] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei, "Deep kinematic pose regression," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 186–201.
- [13] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1263–1272.
- [14] D. Mehta et al., "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *Proc. Int. Conf. 3D Vis.*, 2017, pp. 506–516.
- [15] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2659–2668.
- [16] F. Moreno-Noguer, "3D human pose estimation from a single image via distance matrix regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1561–1570.
- [17] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3D human pose from 2D image landmarks," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 573–586.
- [18] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1446–1455.
- [19] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis, "Sparse representation for 3D shape estimation: A convex relaxation approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1648–1661, Aug. 2017.
- [20] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "Sparseness meets deepness: 3D human pose estimation from monocular video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4966–4975.
- [21] 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 *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 561–578.
- [22] M. Sanzari, V. Ntouskos, and F. Pirri, "Bayesian image based 3D pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 566–582.
- [23] G. Pavlakos et al., "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 975–10 985.
- [24] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3D people models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8387–8397.
- [25] G. Rogez and C. Schmid, "MoCap-guided data augmentation for 3D pose estimation in the wild," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3108–3116.
- [26] G. Varol et al., "Learning from synthetic humans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4627–4635.
- [27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, 2015, Art. no. 248.
- [28] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4724–4732.
- [29] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 483–499.
- [30] L. Pishchulin et al., "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4929–4937.
- [31] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1302–1310.
- [32] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net: Localization-classification-regression for human pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1216–1224.
- [33] D. Mehta et al., "XNect: Real-time multi-person 3D human pose estimation with a single RGB camera," *ACM Trans. Graphics*, vol. 39, no. 4, p. 17, Jul. 2020. [Online]. Available: <http://gvv.mpi-inf.mpg.de/projects/XNect/>. doi: 10.1145/3386569.3392410.
- [34] D. Mehta et al., "Single-shot multi-person 3D pose estimation from monocular RGB," in *Proc. Int. Conf. 3D Vis.*, 2018, pp. 120–130.
- [35] J. Wu et al., "Single image 3D interpreter network," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 365–382.
- [36] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3D pose estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2500–2509.
- [37] H.-Y. F. Tung, A. W. Harley, W. Seto, and K. Fragkiadaki, "Adversarial inverse graphics networks: Learning 2D-to-3D lifting and image-to-image translation from unpaired supervision," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4354–4362.
- [38] D. Drover, C.-H. Chen, A. Agrawal, A. Tyagi, and C. P. Huynh, "Can 3D pose be learned from 2D projections alone?" in *Proc. Eur. Conf. Comput. Vis.*, 2018.
- [39] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3D human pose and shape from a single color image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 459–468.
- [40] H.-Y. Tung, H.-W. Tung, E. Yumer, and K. Fragkiadaki, "Self-supervised learning of motion capture," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5242–5252.
- [41] H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation for 3D human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 750–767.
- [42] M. Omran, C. Lassner, G. Pons-Moll, P. Gehler, and B. Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *Proc. Int. Conf. 3D Vis.*, 2018, pp. 484–494.
- [43] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3D human dynamics from video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5607–5616.
- [44] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7122–7131.
- [45] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll, "Learning to reconstruct people in clothing from a single RGB camera," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1175–1186.
- [46] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, "Multi-garment net: Learning to dress 3D people from images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5419–5429.
- [47] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 407–414.
- [48] M. Ma, H. Fan, and K. M. Kitani, "Going deeper into first-person activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1894–1903.
- [49] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3783–3791.
- [50] H. Yonemoto, K. Murasaki, T. Osawa, K. Sudo, J. Shimamura, and Y. Taniguchi, "Egocentric articulated pose tracking for action recognition," in *Proc. 14th IAPR Int. Conf. Mach. Vis. Appl.*, 2015, pp. 98–101.

- [51] G. Rogez, J. S. Supancic, and D. Ramanan, "First-person pose recognition using egocentric workspaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4325–4333.
- [52] H. Jiang and K. Grauman, "Seeing invisible poses: Estimating 3D body pose from egocentric video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3501–3509.
- [53] Y. Yuan and K. Kitani, "3D ego-pose estimation via imitation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 735–750.
- [54] Y. Yuan and K. Kitani, "Ego-pose estimation and forecasting as real-time PD control," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 10081–10091.
- [55] M. Amer, S. V. Amer, and A. Maria, "Deep 3D human pose estimation under partial body presence," in *Proc. 25th IEEE Int. Conf. Image Process.*, 2018, pp. 569–573.
- [56] H. Rhodin *et al.*, "EgoCap: Egocentric marker-less motion capture with two fisheye cameras," *ACM Trans. Graph.*, vol. 35, no. 6, 2016, Art. no. 162.
- [57] D. Tome, P. Peluse, L. Agapito, and H. Badino, "xR-EgoPose: Egocentric 3D human pose from an HMD camera," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7728–7738.
- [58] T. von Marcard, B. Rosenhahn, M. J. Black, and G. Pons-Moll, "Sparse inertial poser: Automatic 3D human pose estimation from sparse IMUs," *Comput. Graph. Forum*, vol. 36, no. 2, pp. 349–360, 2017.
- [59] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, "Deep inertial poser learning to reconstruct human pose from sparseinertial measurements in real time," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 185:1–185:15, Nov. 2018.
- [60] C. Malleson, J. Collomosse, and A. Hilton, "Real-time multi-person motion capture from multi-view video and IMUs," *Int. J. Comput. Vis.*, vol. 128, pp. 1594–1611, 2020.
- [61] T. von Marcard, G. Pons-Moll, and B. Rosenhahn, "Human pose estimation from video and IMUs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1533–1547, Aug. 2016.
- [62] G. Pons-Moll *et al.*, "Outdoor human motion capture using inverse kinematics and von mises-fisher sampling," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1243–1250.
- [63] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 614–631.
- [64] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. K. Hodgins, "Motion capture from body-mounted cameras," *ACM Trans. Graph.*, vol. 30, no. 4, 2011, Art. no. 31.
- [65] Animated 3D characters. Accessed: Mar. 19, 2019. [Online]. Available: <https://www.mixamo.com/>
- [66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [67] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
- [68] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5745–5753.
- [69] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
- [70] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
- [71] K. He, R. Girshick, and P. Dollár, "Rethinking ImageNet pre-training," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4918–4927.
- [72] D. Mehta *et al.*, "VNect: Real-time 3D human pose estimation with a single RGB camera," *ACM Trans. Graph.*, vol. 36, no. 4, 2017, Art. no. 44.
- [73] C. Sminchisescu, C. Ionescu, and F. Li, "Latent structured models for human pose estimation," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2220–2227.
- [74] C.-H. Chen and D. Ramanan, "3D human pose estimation= 2D pose estimation+ matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7035–7043.
- [75] M. R. I. Hossain and J. J. Little, "Exploiting temporal information for 3D human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 69–86.
- [76] R. Dabral, A. Mundhada, U. Kusupati, S. Afaque, and A. Jain, "Structure-aware and temporally coherent 3D human pose estimation," vol. 3, no. 4, p. 6, 2017, *arXiv: 1711.09250*.
- [77] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7122–7131.
- [78] X. Zhou, M. Zhu, G. Pavlakos, S. Leonardos, K. G. Derpanis, and K. Daniilidis, "MonoCap: Monocular human motion capture using a CNN coupled with a geometric prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 901–914, Apr. 2019.
- [79] E. Jahangiri and A. L. Yuille, "Generating multiple diverse hypotheses for human 3D pose consistent with 2D joint detections," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 805–814.
- [80] H. Fang, Y. Xu, W. Wang, X. Liu, and S. Zhu, "Learning pose grammar to encode human body configuration for 3D pose estimation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.
- [81] X. Sun, J. Shang, S. Liang, and Y. Wei, "Compositional human pose regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2602–2611.
- [82] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall, "A dual-source approach for 3D pose estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4948–4956.
- [83] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.



different applications ranging from robotics to VR/AR for headset mounted camera systems.



Denis Tome is currently working toward the PhD degree and is also a member of the Vision and Imaging Science Group, University College London (UCL), U.K., since 2016, under the supervision of Prof. Lourdes Agapito and Dr. Gabriel Brostow. He is currently a research scientist at Epic Games, working at the intersection between computer vision and graphics. His main research interests include 3D human pose reconstruction, both sparse and dense, from simplistic configuration (monocular) to more complex ones (multiview), for different applications ranging from robotics to VR/AR for headset mounted camera systems.

Thiemo Alldieck is currently working toward the PhD degree from the Computer Graphics Lab, TU Braunschweig, Germany, since 2016. He is currently a research intern at Facebook Reality Labs. Since 2018, he is also affiliated with the "Real Virtual Humans" group at Max Planck for Informatics (MPII) in Saarbrücken, Germany. His research interests include the intersection between computer vision, graphics, and machine learning and focuses on human pose, shape, and clothing reconstruction from monocular images and video.



Patrick Peluse received the BFA degree in visual effects/animation from the Academy of Art University, San Francisco, California, in 2008. He has worked in computer graphics for more than 15 years with current interests in simulation to solve real world problems. Currently, he is a technical artist at Facebook Reality Labs, where his research interests include both real-time and offline data generation for quality assurance and hardware prototyping. He has worked as video game programmer, UI/UX conceptual designer for AR/VR, and data generator / content creator with companies Autodesk M&E, Digital Domain, Circuits for Fun, Meta Vision, Magic Leap, and Oculus.



Gerard Pons-Moll is currently the head of the Emmy Noether independent research group Real Virtual Humans, senior researcher at the Max Planck for Informatics (MPII) in Saarbrücken, Germany, and junior faculty at Saarland Informatics Campus. His research interests include the intersection of computer vision, computer graphics and machine learning – with special focus on analyzing people in videos, and creating virtual human models by looking at real ones. His work has received several awards including the prestigious Emmy Noether Grant (2018), a Google Faculty Research Award (2019), a Facebook Reality Labs Faculty Award (2018), and recently the German Pattern Recognition Award (2019), which is given annually by the German Pattern Recognition Society to one outstanding researcher in the fields of Computer Vision and Machine Learning. His work got best papers awards at BMVC'13, Eurographics'17 and 3DV'18 and he served as area chair for ECCV'18, 3DV'19, SCA'18'19, FG'20 and will serve as an area chair for CVPR'21, ECCV'20, and 3DV'20.



Lourdes Agapito received the BSc, MSc, and PhD degrees from the Universidad Complutense, Madrid, Spain, in 1991, 1992, and 1996, respectively and was a postdoctoral fellow at the University of Oxford's Robotics Research Group (1997-2001). She is currently a professor of 3D Vision and the head of the Vision and the Imaging Science Group in the Department of Computer Science, University College London (UCL), U.K. She is also co-founder of London-based startup Synthesia Technologies and was an ERC Grant

holder (2008-14). Her research interests lie at the intersection of computer vision, graphics and machine learning; more specifically 3D reconstruction from video, 3D shape modelling, weakly supervised learning for 3D vision, human pose estimation and video synthesis. She has served as program chair for the top computer vision conferences (CVPR'16, ICCV'21), workshops chair for ECCV'14 and area chair for CVPR (3x), ECCV (2x), ICCV (1x). She is associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and the *International Journal of Computer Vision*. She was keynote speaker at ICRA'17.



Hernan Badino received the PhD degree in computer sciences from Goethe University Frankfurt, Germany, in 2008. He is currently a research scientist at Facebook Reality Labs. During his PhD, he worked with the Image Based Environment Perception Group at Daimler AG and joined the Carnegie Mellon University, Pittsburgh, Pennsylvania, in 2009 as a postdoctoral fellow. In 2012, he was appointed a faculty position at the Robotics Institute at the Carnegie Mellon University, Pittsburgh, Pennsylvania where he worked on visual based localization, sensor-fusion, ego-pose estimation, object detection, and tracking, and real-time embedded solutions for visual-based pose estimation. He joined Facebook Reality Labs, in 2015 and since then, he has been working at the intersection of computer vision, motion capture, and telepresence systems towards the goal of achieving social presence in artificial reality.



Fernando de la Torre received the BSc degree in telecommunications, and the MSc and PhD degrees in electronic engineering from the La Salle School of Engineering, Ramon Llull University, Barcelona, Spain, in 1994, 1996, and 2002, respectively. He is currently an associate research professor at the Robotics Institute at Carnegie Mellon University, Pittsburgh, Pennsylvania. His research interests include the fields of computer vision and machine learning. Currently, he is directing the Component Analysis Laboratory (<http://ca.cs.cmu.edu>) and the Human Sensing Laboratory (<http://humansensing.cs.cmu.edu>) at Carnegie Mellon University, Pittsburgh, Pennsylvania. He has more than 100 publications in refereed journals and conferences. He has organized and co-organized several workshops and has given tutorials at international conferences on the use and extensions of component analysis.

► **For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.**