

Received March 21, 2022, accepted May 18, 2022, date of publication May 23, 2022, date of current version May 27, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3177623

Silhouette-Based 3D Human Pose Estimation Using a Single Wrist-Mounted 360° Camera

RYOSUKE HORI¹, (Student Member, IEEE), **RYO HACHIUMA¹**, (Member, IEEE),
MARIKO ISOGAWA², (Member, IEEE), **DAN MIKAMI³**, (Member, IEEE),
AND HIDEO SAITO¹, (Senior Member, IEEE)

¹Graduate School of Science and Technology, Keio University, Yokohama, Kanagawa 223-8522, Japan

²Computer and Data Science Laboratories, Nippon Telegraph and Telephone Corporation, Yokosuka, Kanagawa 239-0847, Japan

³Faculty of Informatics, Kogakuin University, Tokyo 163-8677, Japan

Corresponding author: Ryosuke Hori (hori-rysk@keio.jp)

This work was supported in part by the Japan Science and Technology Agency (JST)-Mirai Program, Japan, under Grant JPMJMI19B2.

ABSTRACT In this paper, we propose a framework for 3D human pose estimation using a single 360° camera mounted on the user's wrist. Perceiving a 3D human pose with such a simple setup has remarkable potential for various applications (*e.g.*, daily-living activity monitoring, motion analysis for sports training). However, no existing method has tackled this task due to the difficulty of estimating a human pose from a single camera image in which only a part of the human body is captured, and because of a lack of training data. We propose a method for translating wrist-mounted 360° camera images into 3D human poses. Since we are the first to try this task, we cannot use existing datasets. To address this issue, we use synthetic data to build our own dataset. This solution, however, creates a different problem, that of a domain gap between synthetic data for training and real image data for inference. To resolve this problem, we propose silhouette-based synthetic data generation created for this task. Extensive experiments comparing our method with several baseline methods demonstrated the effectiveness of our silhouette-based pose estimation approach.

INDEX TERMS 3D human pose estimation, domain adaptation, 360° camera, data synthesis, silhouette.


I. INTRODUCTION

Human Pose Estimation (HPE) has long been studied in the computer vision community [1], [2]. In particular, 3D pose estimation using a monocular camera, which is one of the most widely used visual sensors in the world, has been actively studied [3] because of its usefulness in various fields such as animation, virtual reality, healthcare, and sports. In the past, many third-person perspective 3D human pose estimation methods have been proposed that use an *outside-in* arrangement of single or multiple cameras statically placed around the user [4]–[7]. However, since the third-person perspective method can only estimate the movements within the range of vision of the externally placed camera, it is often not practical in real-world situations where people move around in a large space.

In contrast, there are first-person perspective (also known as “egocentric”) 3D human pose estimation methods [8]–[16], in which cameras or inertial measurement

units (IMUs) are attached to the body. These can obtain 3D pose data in various activities in the real world due to their mobility and flexibility, enabling new applications such as motion recognition and performance analysis in the fields of sports and healthcare. In egocentric human pose estimation, it is important to minimize the number of devices in terms of usability. For this reason, mainstream methods in recent years have been *inside-out* methods [12], [17]–[20] in which an outward-facing camera is attached to the body, and *inside-in* methods [8], [13]–[15], [21]–[23] in which a camera with a fisheye lens is attached to the body to capture the camera wearer's body with its wide field of view. However, thus far, there is no method with a more practical camera setting, *i.e.*, a single wrist-mounted camera that could be introduced in smartwatches in the future (see Fig. 1-(a)).

Therefore, we propose a framework for estimating 3D human poses from images taken with a single wrist-mounted camera. This task is quite challenging as some human body parts are hidden from the camera's line of sight. We make use of a single 360° camera and a neural network-based framework that leverages time-series features to estimate a

The associate editor coordinating the review of this manuscript and approving it for publication was John See .

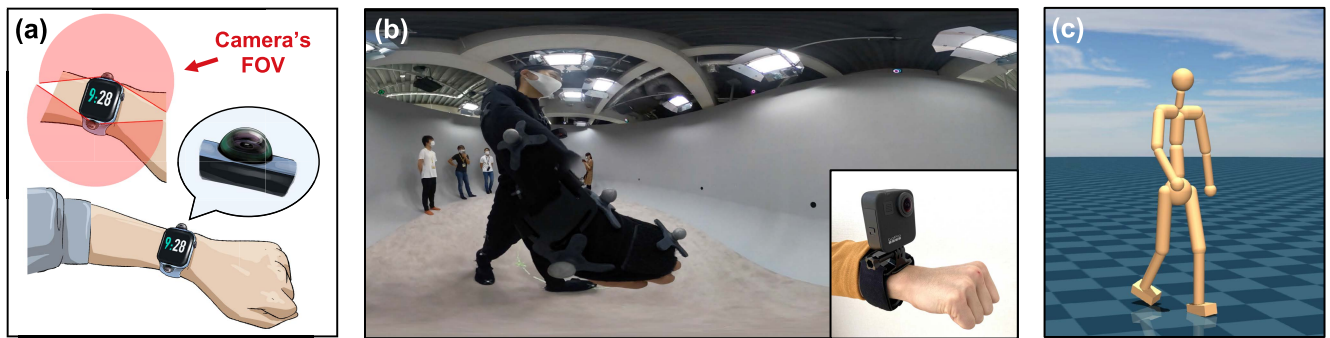


FIGURE 1. Concept of our 3D human pose estimation method using a single wrist-mounted 360° camera. (a) Future vision depicting a smartwatch with a 360-degree camera. The area shown in red represents the camera's field of view (FOV); (b) camera view with a 360° camera (GoPro MAX) mounted on the wrist, which is the camera configuration of our method; (c) 3D pose of the camera wearer estimated from (b) image by our 3D human pose estimation framework.

3D human pose with only limited visual information (see Fig. 1-(b), 1-(c), and Fig. 2). The difficulty is how to prepare the training data; there is no existing dataset for 3D human pose estimation with wrist-mounted cameras. Existing works with body-mounted cameras that faced this issue tackled it by using synthetic data [13]–[15], [20]–[23]. However, training on synthetic datasets, which lack diversity in subject appearance, lighting, or background variation, can inherently cause domain gaps and restrict generalizability [24], [25].

To overcome these issues, we also describe a simplified method for generating training data. In contrast to the conventional approach of creating realistic and diverse large-scale datasets, we propose an approach that bridges the domain gap by using binary silhouette images for both training and inference data. We train the network using silhouette equirectangular image sequences generated using only existing motion capture (MoCap) data. During inference, the 3D poses are estimated from silhouette equirectangular image sequences obtained by our proposed silhouetting process from real images captured by the 360° camera. The proposed method for generating synthetic data is quite simple and thus can be easily applied to any other approach that uses varied camera positions or a different number of cameras.

We validate our approach for egocentric pose estimation on a large motion capture dataset and an in-the-wild dataset consisting of various human motions (walking, crouching, jumping, raising hands, and motion transitions). Experiments comparing the accuracy of pose estimation with several different types of data show the effectiveness of our silhouette-based approach. Experiments comparing the accuracy of pose estimation and the acceleration of the output pose with several different pose estimation models show that our model, which utilizes time-series features, generates smooth and accurate poses.

To summarize, our contributions are as follows:

- We are the first to propose a 3D human pose estimation framework using a single wrist-mounted camera, which contributes to many important applications.

- To bridge the domain gap between synthetic data and real-world data, we describe a method of silhouette-based human pose estimation. Our simplified method of generating silhouette-based training data has the potential to be used for other camera configurations.
- We provide extensive experiments and show the effectiveness of our silhouette-based pose estimation approach and the pose estimation model that leverages time-series features.
- We build a training dataset, composed of 54K frames of silhouette images, for our new problem setting. We make it publicly available to promote progress in the area of egocentric motion capture. The dataset is available at <http://hvrl.ics.keio.ac.jp/hori/iee-access/sil-hpe/index.html>

A conference version of this paper exists [26]. This journal version extends it by adding a new method for generating pseudo-silhouettes based on a conditional adversarial network that improves our human pose estimation model. In addition, this paper provides results from more extensive experiments conducted by adding the number of subjects in the dataset, baseline methods, and evaluation metrics.

II. RELATED WORK

A. 3D HUMAN POSE ESTIMATION

Vision-based 3D human pose estimation tasks have been developed for decades [1]–[3] because of their potential applications in many fields such as animation, virtual reality, healthcare, and sports. The well-known commercial products for 3D HPE are Vicon [27] with optical sensors and Capture [28] with multiple cameras. However, they only work in a limited space captured with multiple camera sensors or require special markers to be attached to the human body. Therefore, 3D HPE methods using monocular cameras, which are the most widely used visual sensors, have been actively studied to obtain human pose data in more diverse real-world scenarios. They can be classified into two types based on the positional relationship between the cameras

and the human body: third-person perspective methods and egocentric methods.

Third-person perspective 3D HPE is a method of estimating human poses from images or videos using the outside-in arrangement, *i.e.*, one or more cameras are placed around the person. In recent years, many pose estimation methods using deep learning networks have been proposed [4], [6], [29]–[33]. Although these methods have made it possible to apply them in various real-world situations, they still have a limitation in that the user can only move within the field of view of the externally placed cameras. Therefore, they are not practical in environments where people move around in a large space or are occluded by objects.

On the other hand, egocentric 3D HPE is a method for estimating poses using body-mounted cameras, without the need for external cameras. Due to its mobility and flexibility, the egocentric method can obtain 3D pose data in various real-world activities, enabling new applications in fields such as sports, animation, and healthcare. We tackled this task in the hardware configuration of a wrist-mounted 360° camera. The related egocentric methods are presented in the following subsection (Sec. II-B).

B. EGOCENTRIC HUMAN POSE ESTIMATION

In early research, the egocentric HPE methods were proposed for human activity recognition, with most of them [34]–[37] detecting only the upper body. Estimating the whole body in an egocentric setting is a more challenging task, and many researchers have dealt with this challenge through various approaches as follows:

1) POSE ESTIMATION USING MULTIPLE SENSOR DEVICES

In early work, Shiratori *et al.* [38] reconstructed human poses by attaching 16 cameras to the limbs and torso of a subject and performing Structure from Motion (SfM) of the environment. Rhodin *et al.* [8] pioneered the use of top-down view cameras for full-body pose estimation and achieved egocentric markerless motion capture with two helmet-mounted fisheye cameras. Cha *et al.* [10] proposed a method to capture the user's body pose, facial expression, and surrounding environment from multiple cameras attached to the user's glasses, assuming that they will be integrated into AR glasses in the future. Later, Cha *et al.* [11] also proposed a standalone real-time system that dynamically captures a person in 3D using only multiple head-mounted cameras and IMUs worn on the wrist and ankle. Recently, Guzov *et al.* [16] combined accurate pose estimation using IMUs with camera localization using a head-mounted first-person view camera to estimate the user's global position and posture in a large-scale 3D space. However, the setup of these methods is technically expensive, as it requires tedious pre-calibration or pose optimization throughout the sequence. We adopt a method that uses a single camera, which is superior in this respect.

2) POSE ESTIMATION USING A SINGLE CAMERA

There are many studies that have taken on the more difficult task of whole-body pose estimation using only one camera, such as an inside-out configuration where the body is almost invisible to the camera [12], [17]–[20], or a configuration where a wide-angle camera is attached to capture the whole body [13]–[15], [21]–[23].

Jiang and Grauman [17] estimate full-body pose by leveraging both learned dynamic and scene classifiers and pose coupling over a long time. Yuan and Kitani [12], [18] proposed DeepRL-based methods for estimating and forecasting both accurate and physically plausible 3D egopose sequences without observing the camera wearer's body. Ng *et al.* [19] estimate the camera wearer's 3D body pose from egocentric video sequences by leveraging the pose of the interacting person. Jiang and Ithapu [20] tackled the egopose estimation from a natural human vision span by taking advantage of both the dynamic features from visual simultaneous localization and mapping (SLAM) and body shape imagery.

Xu *et al.* [13] proposed the first real-time 3D pose estimation system using a single egocentric fisheye lens camera mounted on a cap. In the same camera position, Hu *et al.* [15] synthesize free-viewpoint avatars, through a rendering process that includes texture synthesis, pose construction, and neural image translation. Tome *et al.* [14], [21] presented a method for full-body pose estimation from monocular images captured from a downward-looking fish-eye camera installed on a Head Mount Display (HMD). Hwang *et al.* [23] proposed a multimodal human motion capture system that estimates 3D body posture, head posture, and camera posture in real-time using a single RGB camera with an ultra-wide-angle fisheye lens mounted on the chest. Wang *et al.* [22] estimate temporally stable global 3D human poses from a single head-mounted fisheye camera by leveraging the 2D and 3D keypoints from CNN detection as well as VAE-based motion priors.

Thus far, these existing methods mount the camera on the user's head or chest. This paper explores the potential for another camera setting for user-mounted cameras, *i.e.*, a single wrist-mounted camera that is considered more practical because it could be introduced in smartwatches in the future.

C. DATA SYNTHESIS FOR EGOCENTRIC HUMAN POSE ESTIMATION

In learning-based methods, it is important to collect a large amount of pose data for training in order to estimate various poses. Since existing datasets consisting of third-person perspective images cannot be used for egocentric HPE using a fisheye camera, it is necessary to create a unique dataset. However, capturing a large amount of annotated 3D pose data is a huge task, and manual labeling in 3D space is impractical. To address this difficulty, Rhodin *et al.* [8] proposed markerless multi-view motion capture with an

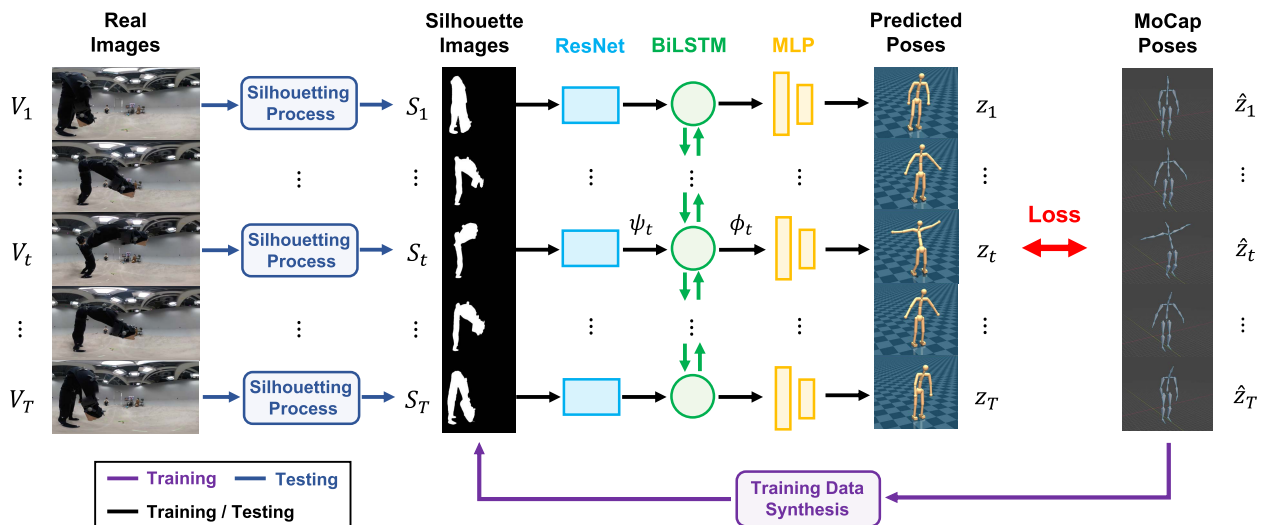


FIGURE 2. Overview of our 3D pose estimation method using a single wrist-mounted 360° camera. The network extracts time-series features from a silhouette image sequence and predicts a 3D human pose sequence. The input data for the network during training is a synthetic silhouette image sequence generated by animating a humanoid using MoCap data as described in Sec. III-B. During inference, the input is a silhouette image sequence obtained by our silhouetting process from equirectangular images captured by a 360° camera in a real environment, as described in Sec. III-C.

external camera to acquire 3D annotations. In recent years, to further reduce the effort required for data annotation, some existing methods have generated synthetic data to build the training dataset [13]–[15], [20]–[23]. However, training on synthetic datasets, which lack diversity and photorealism in subject appearance, lighting, and background variation, can inherently cause domain gaps and restrict generalizability [24], [25]. Therefore, it is necessary to build a large dataset with maximum diversity in motion and appearance and minimal differences between synthetic images and real images.

Xu *et al.* [13] created the Mo²Cap² dataset, which consists of a total of 530K images rendering the SMPL body model [39] characters animated using 3K different motions from the CMU MoCap dataset [40]. The images were created using more than 700 body textures from the SURREAL dataset [41] and more than 5,000 background images taken indoors and outdoors with a fisheye camera mounted on a long pole. Hu *et al.* [15] and Wang *et al.* [22] also used this dataset. The xR-EgoPose Synthetic dataset created by Tome *et al.* [14], [21] consists of 383K frames of synthetic images generated using 23 male and 23 female characters of various skin tones, clothing, and motions. To maximize the photorealism of the synthetic dataset, they animated the characters in Maya [42] using actual mocap data [43], and used a standardized physically based rendering setup with V-Ray. Hwang *et al.* [23] created a composite dataset of 680K frames. They used the SMPL models to create humanoid models of two males and two females with various body shapes and appearances, and animated the models using the CMU MoCap dataset. The clothing textures were randomly selected and rendered from the SURREAL dataset, while the background textures were

sampled and applied to the omnidirectional images from the SUN360 dataset [44]. Jiang *et al.* [20] generated images using Blender [45] and the CMU MoCap dataset for a total of approximately 10 hours. Randomly selected human meshes from 190 different mesh models were applied to each MoCap sequence, and the background was randomly used from the ADE20K dataset [46].

In contrast to these approaches of building large datasets of realistic and diverse synthetic images, we took on an “opposite” approach. We propose to use synthetic binary silhouette images, inspired by the method of Xu *et al.* [47] who used low-dimensional synthetic data to fill the domain gap in pedestrian trajectory estimation. This silhouette image can be easily generated from existing MoCap data (see Sec. III-B) and is easily adaptable to changes in camera position. As the data are fully silhouette-based, it reduces the problem of domain gaps between synthetic data and real-world data.

III. METHOD

As shown in Fig. 2, our goal is to estimate 3D human poses $z_{1:T}$ from a real equirectangular image sequence $V_{1:T}$ by our network described in Sec. III-A. The network is trained only on synthetic silhouette data generated at a lower cost than conventional methods by the method in Sec. III-B. During inference, the silhouetting process presented in Sec. III-C is applied to real equirectangular images to bridge the domain gap between synthetic data and real-world data. As an optional feature, we also propose a method for adding noise to the synthetic training data to improve the robustness of the model against noise caused by the silhouetting process and the smoothness of the output pose during inference. Details of this method are in Sec. III-D.

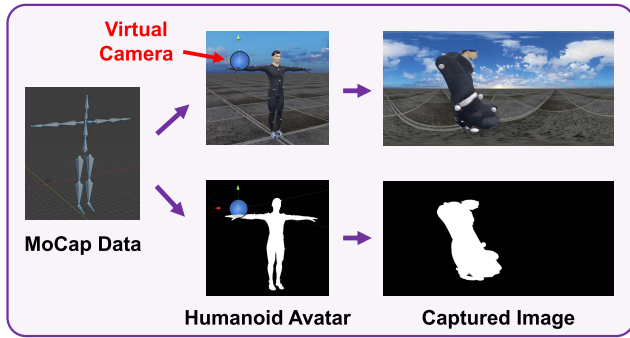


FIGURE 3. Overview of training data synthesis. We animated a humanoid avatar using MoCap data and captured equirectangular images with a virtual 360° camera fixed at the avatar’s wrist position. To capture binary silhouette images, we set the model’s body to white and the background to black.

A. HUMAN POSE ESTIMATION NETWORK

Our network \mathcal{F} takes the input of the silhouette equirectangular image sequence $S_{1:T}$ and predicts the humanoid state $z_{1:T}$ at each frame, as shown in Fig. 2. The humanoid state z_t consists of the pose p_t (position and orientation of the root, and joint angles) and velocity v_t (linear and angular velocities of the root, and joint velocities). Since human motion is temporally continuous and smooth, the change in poses in nearby past and future frames can be utilized to estimate human poses from images. Therefore, the network extracts image-by-image features from given sequential images and estimates 3D human poses based on their temporal context. This network is based on Yuan and Kitani’s work [12]. The use of time-series features enables smoother and more natural motion estimation than image-by-image pose estimation.

The model encodes the silhouette image by ResNet-18 [48] to extract the feature vector $\psi_{1:T} \in \mathbb{R}^{128}$ and feeds it to Bidirectional Long-Short Term Memory (BiLSTM) to generate the visual context $\phi_{1:T} \in \mathbb{R}^{128}$ for each frame. We then feed it to the Multilayer Perceptrons (MLPs) and predict the humanoid state $z_{1:T}$. The Mean Squared Error (MSE) is used as the loss function:

$$L(\zeta) = \frac{1}{T} \sum_{t=1}^T \|\mathcal{F}(S_{1:T})_t - \hat{z}_t\|^2, \quad (1)$$

where ζ is the parameter of this network \mathcal{F} , and \hat{z}_t is the ground-truth humanoid state. The optimal \mathcal{F}^* can be obtained by an SGD-based method.

B. TRAINING DATA SYNTHESIS

Fig. 3 shows an overview of our process of generating an equirectangular image sequence corresponding to the MoCap data for training the network. In a virtual environment such as one provided by Unity [49], we animate a humanoid avatar by applying the MoCap data. A virtual camera with a 360° field of view is fixed at the wrist of the avatar. Since this virtual camera can be attached to any part of the avatar’s body, our method can be easily adapted to any camera position, not only the wrist. By capturing images while

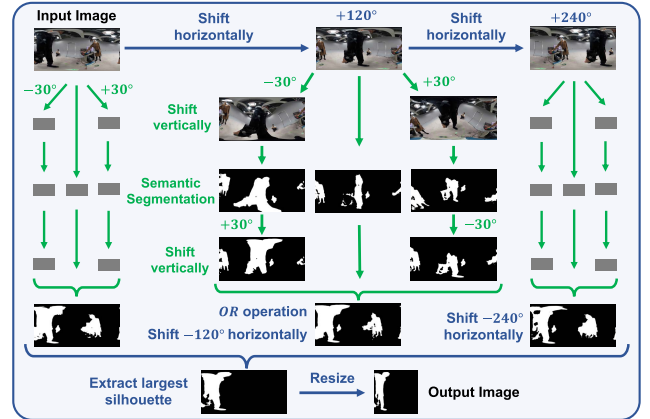


FIGURE 4. Our silhouetting process, which converts an equirectangular image captured by a wrist-mounted 360° camera into a binary silhouette image. We extract the silhouette of the camera wearer from the region labeled as human, which is estimated by an existing semantic segmentation model.

keeping them horizontal, we generate RGB equirectangular images as shown in the upper right of Fig. 3.

To improve the performance of generalization on real images during inference, conventional methods generated a large amount of synthetic RGB image data using photorealistic and diverse backgrounds and humanoid textures. In contrast, our approach is to bridge the domain gap by using binary silhouette images for both training and inference data. To this end, we generate silhouetted equirectangular images as shown in the lower right of Fig. 3, by setting the background to black and the avatar’s body texture to white, and capturing images with light shining from all directions. By binarizing those output images and resizing them to the input size of the network, we obtain a set of 3D poses and equirectangular image sequences for training.

C. SILHOUETTING PROCESS FOR INFERENCE

We propose a silhouetting process for silhouette-based pose estimation, which converts an equirectangular image captured by a wrist-mounted 360° camera into a binary image of a human silhouette. We extract the silhouette of the camera wearer from the region labeled as human, which is estimated by an existing learned semantic segmentation model. In this study, we used the Swin Transformer [50] model trained on the ADE20K [46] dataset, which is publicly available [51], for semantic segmentation. Since this process only uses the trained model of the semantic segmentation method, it can be replaced by any existing method. The issue here is that images taken with a 360° camera are highly distorted; the top and bottom of the images are stretched out, which is very different from normal-perspective view-images such as ADE20K. Therefore, it is difficult to extract human silhouettes by simply applying semantic segmentation. To solve this issue, we apply semantic segmentation not only to the original equirectangular images but also to the equirectangular images shifted vertically. Then we merge the results of the semantic segmentation to obtain the network input.

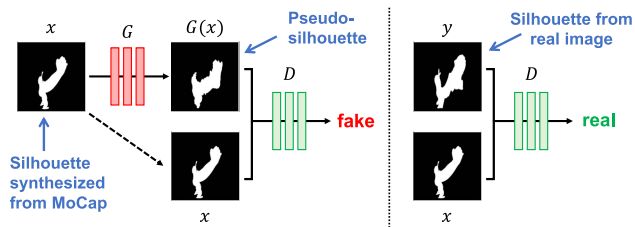


FIGURE 5. Training the pix2pix model to map our synthetic silhouette images with smooth contour to noisy silhouette images obtained through our silhouetting process. The discriminator, D , learns to classify between fake (synthesized by the generator) and real (smooth silhouette image, noisy silhouette image) tuples. The generator, G , learns to fool the discriminator.

Specifically, we first generate three equirectangular images by shifting them horizontally (yaw axis) at 120° intervals. Second, we generate two images by shifting each one by $\pm 30^\circ$ vertically (pitch axis). Third, we apply semantic segmentation to each image and shift them vertically back to the equirectangular images of the original vertical angle. Fourth, in each of the three horizontally shifted images, we combine the three generated silhouette images via the *OR* operation and shift them horizontally back to the same position as the input image. Finally, we extract the largest silhouette from the three images and resize it for the input of the network.

D. PSEUDO-SILHOUETTE GENERATION FOR NOISE ROBUSTNESS

As an optional feature, we also propose a method of adding generated images of noisy silhouettes (hereafter, “pseudo-silhouettes”) to the training data to improve the robustness of the model against silhouette noise during inference. Although our silhouetting process aims to bridge the domain gap between the synthetic training data and the real data for inference by using silhouettes, it is still difficult to completely bridge the gap. The synthetic training data in our method are silhouette images with smooth contours, while the silhouette of the input image during inference has a rough contour (hereafter, “silhouette noise”). The silhouette noise is caused by applying semantic segmentation trained on general perspective view images to equirectangular images and is an inaccurate contour with some missing parts as in the bottom image of Fig. 4, or some protruding parts as in the top right image of Fig. 5. These differences between the training data and test data affect the accuracy of pose estimation as a domain gap.

Therefore, we improve the robustness of the model by adding pseudo-silhouette images to the training data. The pseudo-silhouette image is generated by converting the smooth silhouette synthetic image to a noisy silhouette image using the pix2pix [52] model, an image-to-image translation method with conditional adversarial networks. The silhouette noise can be reproduced by the pix2pix model trained to transfer the synthetic silhouette image generated using MoCap data to the silhouette image obtained from the real

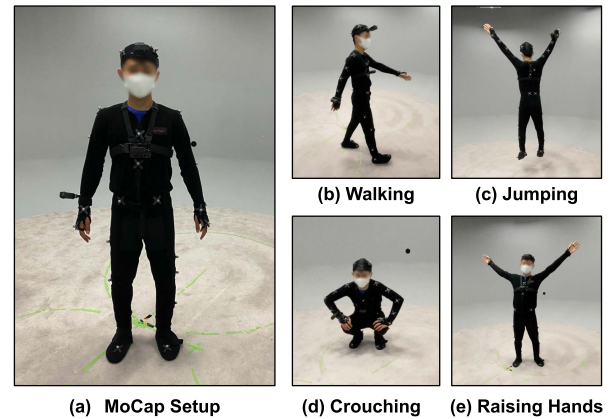


FIGURE 6. Motion capture for constructing our MoCap Training and Test Data. Subjects wore motion capture markers and the 360° camera as shown in Fig. (a) and were asked to perform actions as shown in Figs. (b)–(e).

image (see Fig. 5). By adding this pseudo-silhouette image to the training data, even if the image sequence contains silhouette noise during inference, it is less likely to be affected by it, resulting in the estimation of more correct poses and smoother motions.

IV. EXPERIMENT

A. DATASET

- **MoCap Training and Test Data:** We used OptiTrack to capture the motion data to construct the dataset (see Fig. 6). Four subjects each wore a 360° camera on their wrist and were asked to perform a variety of actions including walking, jumping, crouching, raising hands, and transitioning between all of these motions. Each take lasted about 5 min and a total of 9 takes were captured, of which 7 takes were used as training data and 2 takes as test data. After removing the parts of inaccurate motions caused by motion capture failures, the training data consists of 54K frames and the test data consists of 14K frames.
- **In-the-wild Data:** We also collected in-the-wild data to verify the effectiveness of our method in a real-world environment. As in the previous MoCap training and test data collection, the subject wore the 360° camera on the wrist and was asked to perform a variety of actions. This dataset consisted of 11 videos each lasting about 5 sec. As it is hard to obtain ground-truth 3D poses in a real-world environment, following Yuan and Kitani [12], we captured side-view poses of the subject, which were used for quantitative evaluation based on 2D keypoints.

B. IMPLEMENTATION DETAILS

The weights of ResNet-18 [48] used in our 3D human pose regressor were pre-trained with ImageNet [53]. The input equirectangular image was resized to 224×224 . The Adam [54] optimizer was employed at the learning rate of $1e - 4$. When training this network, for each time step we sampled data fragments in turn for 120 frames (4 sec) and

padding 10 frames of visual features ψ_t on both sides to reduce the estimation error on the boundary frames when computing ϕ_t . We used Unity for the virtual environment to generate the training data and MuJoCo [55] to visualize estimated human poses that consisted of 52 Degrees of Freedom (DoFs) and 19 rigid bodies.

The pix2pix model for generating the pseudo-silhouettes was trained using our MoCap Training Data. The model was trained to learn a translation from the silhouette image generated by animating the avatar using the MoCap to the silhouette image obtained from the real image by our silhouette process. To align the position of the silhouettes in the real and virtual images, we shifted these images so that the *Sum of Squared Differences (SSD)* was minimized. When training our pose estimation network, we added a pseudo-silhouette image with a probability of one-fourth to the image sequence selected with a probability of one-half at each iteration.

C. EVALUATION METRIC

To evaluate the accuracy of the pose estimation and the smoothness of the motion of our approach, we use the following evaluation metrics. For the estimated and ground-truth keypoints, we set the hip keypoint as the origin and scaled the coordinate to make the height between the shoulder and hip equal to 0.5 [m].

- **Mean Per-Joint Position Error (E_{MPJPE}):** A pose-based metric used by Isogawa *et al.* [56] that measures the Euclidean distance between the estimated 3D pose and the ground-truth 3D pose in mm. We use it for evaluating the accuracy of pose estimation on MoCap test data. This metric is defined as

$$E_{MPJPE} = \frac{1}{TJ} \sum_{t=1}^T \sum_{j=1}^J \| (x_t^j - x_t^{root}) - (\hat{x}_t^j - \hat{x}_t^{root}) \|_2, \quad (2)$$

where x_t^j is the j -th joint position of the estimated pose, and \hat{x}_t^j is the ground truth. x_t^{root} and \hat{x}_t^{root} represent the root joint position of the estimated and ground-truth poses, respectively.

- **2D Keypoint Error (E_{key}):** A pose-based metric proposed by Yuan and Kitani [12] that measures the Euclidean distance between the estimated 2D pose and the ground-truth 2D pose in mm. We use it for evaluating the accuracy of posture estimation on the in-the-wild data. The metric is defined as

$$E_{key} = \frac{1}{TJ} \sum_{t=1}^T \sum_{j=1}^J \| y_t^j - \hat{y}_t^j \|_2, \quad (3)$$

where y_t^j is the j -th 2D keypoint of the estimated pose obtained by projecting the 3D joints to an image plane with a side-view camera, and \hat{y}_t^j is the ground truth extracted with OpenPose [57].

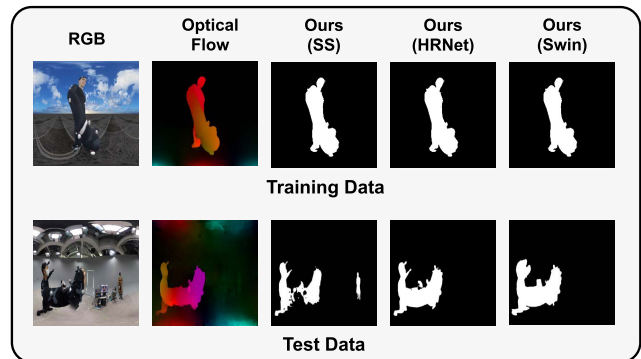


FIGURE 7. Examples of training and test data for each method in experiment 1. We used the same silhouette images for training our silhouette-based methods (SS, HRNet, and Swin).

- **Acceleration Error (E_{acc}):** A metric proposed by Kanazawa *et al.* [58] that measures the average difference between ground truth and predicted acceleration of each joint in mm/s^2 . For the MoCap test data, we used the original metric E_{acc3d} for 3D poses, and for the in-the-wild data, we used the metric E_{acc2d} adapted to 2D data. The metric is defined as

$$E_{acc} = \frac{1}{TJ} \sum_{t=1}^T \sum_{j=1}^J \| a_t^j - \hat{a}_t^j \|_2, \quad (4)$$

where a_t^j and \hat{a}_t^j are the acceleration of the j -th keypoint of the estimated pose and ground truth pose, respectively.

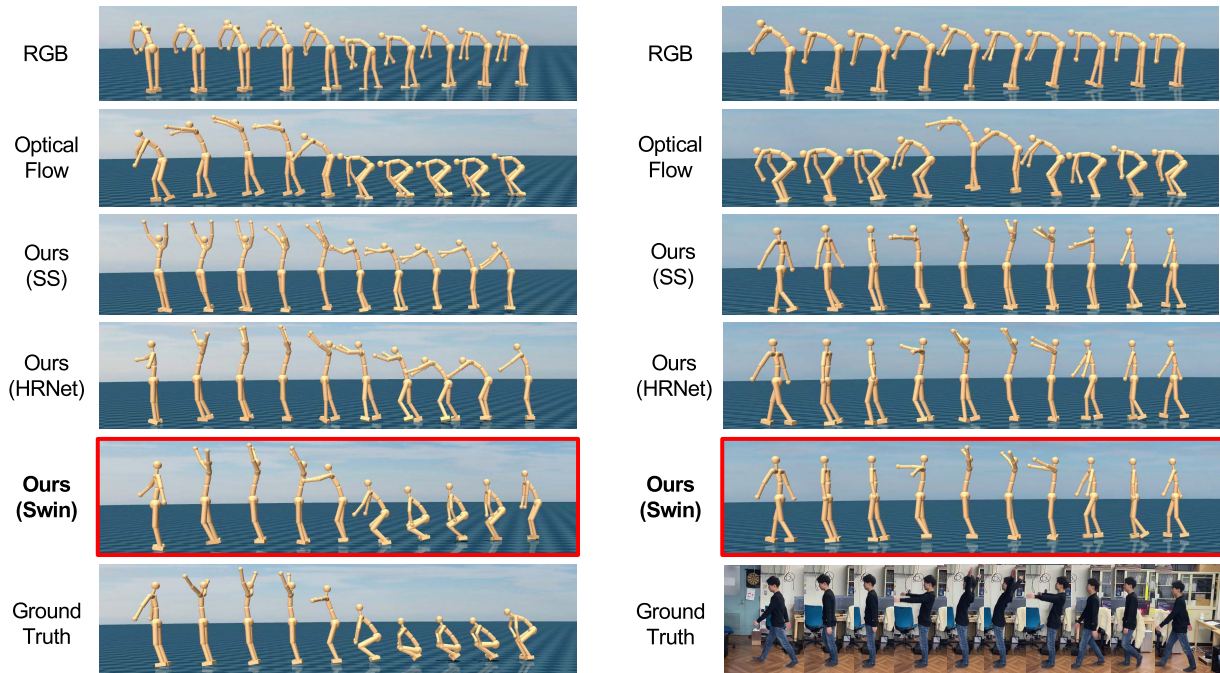
D. EXPERIMENT 1: INVESTIGATING THE EFFECTIVENESS OF OUR SILHOUETTE-BASED POSE ESTIMATION

To investigate the effectiveness of our silhouette-based approach, we evaluate the accuracy of the pose estimation when using our network on several different types of data. We compare our method (denoted as **Ours (Swin)** in Table. 1) against the following four baseline methods. Fig. 7 shows examples of input data used for training and testing, respectively, for each baseline method.

- **RGB:** A method that trains our network on synthetic RGB data and tests it on real RGB data taken with a 360° camera.
- **Optical Flow:** A method that uses optical flow obtained by PWC-Net [59] from image sequences to train our network. This is the same method as PoseReg proposed by Yuan and Kitani [12]. The optical flows for training and testing were obtained from the image sequences of the RGB method above, respectively.
- **Ours (SS):** A method that trains the network on our synthetic silhouette data and tests it on silhouette images that are human-labeled regions extracted by applying semantic segmentation to real equirectangular images.
- **Ours (HRNet):** A method proposed in [26] that trains the network on our synthetic silhouette data and tests it on silhouette images obtained by our silhouetting

TABLE 1. Quantitative results of evaluating the accuracy of pose estimation on the MoCap test data and the in-the-wild data in Experiment 1. Bold values show the best scores.

Method	MoCap Test Data (E_{MPJPE})					In-the-wild Data (E_{key})				
	Walk	Crouch	Jump	Raise Hands	All Frames	Walk	Crouch	Jump	Raise Hands	All Frames
RGB	209.7	209.7	215.5	258.6	212.8 ± 69.5	330.1	263.1	245.1	175.8	225.6 ± 82.3
Optical Flow	200.7	166.4	222.0	197.4	197.1 ± 77.1	359.3	306.0	320.1	361.4	337.4 ± 67.3
Ours (SS)	141.6	142.5	181.4	128.1	144.0 ± 67.2	119.0	134.5	149.2	109.8	130.4 ± 61.1
Ours (HRNet)	99.0	160.2	163.9	110.4	113.3 ± 61.5	88.2	125.6	121.5	103.1	110.4 ± 42.5
Ours (Swin)	83.1	104.7	116.1	95.5	89.4 ± 34.9	88.3	119.3	117.0	95.0	106.1 ± 51.3

**FIGURE 8.** Qualitative results for each method on the MoCap test data (left) and the in-the-wild data (right) in Experiment 1.

process using a semantic segmentation model based on HRNet.

E. EXPERIMENT 2: INVESTIGATING THE EFFECTIVENESS OF OUR POSE ESTIMATION NETWORK

To verify the effectiveness of our pose estimation model that uses time-series features, we conduct an experiment to evaluate the accuracy of the pose estimation and the smoothness of the output pose. However, there is currently no 3D human pose estimation method using a wrist-mounted 360° camera, and there is no publicly available code for other egocentric 3D human pose estimation methods applicable to this experiment. Therefore, as baseline methods, we used third-person-view-based existing 3D human pose estimation networks. Please note that we re-trained those networks with the same synthetic silhouette image dataset as our method for a fair comparison. We evaluate the following four baseline methods, including our model trained with pseudo-silhouette images in the training data.

- **Mobile Human Pose:** A model proposed by Choi *et al.* [60] for real-time 3D human pose estimation

from a single image, which is the most precise and compact model that can be implemented in mobile devices.

- **Integral Human Pose:** A model proposed by Sun *et al.* [32] for 3D human pose estimation from a single image, which is a simple and effective integral regression model that unifies the heatmap representation and the joint regression approach, sharing the merits of both.
- **Ours w/o Pseudo-Silhouette (PS):** Our proposed model, which was trained using only smooth synthetic silhouette images without pseudo-silhouette images.
- **Ours w/ Pseudo-Silhouette (PS):** Our proposed model aims to be robust against silhouette noise by mixing pseudo-silhouette images in the training data.

V. RESULTS

A. RESULTS OF EXPERIMENT 1

Quantitative results of the pose estimation on the MoCap test data and the in-the-wild data are shown in Table 1, and qualitative results are shown in Fig. 8. The results show that our method outperforms the baseline methods.

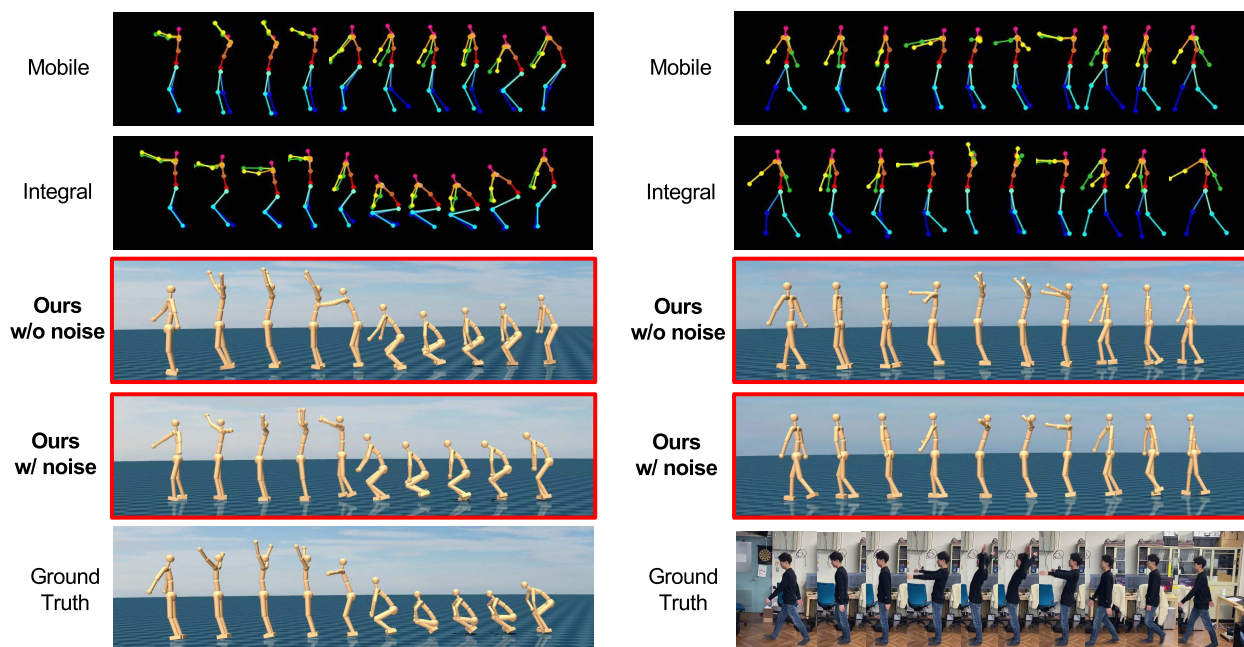


FIGURE 9. Qualitative results for each method on the MoCap test data (left) and the in-the-wild data (right) in Experiment 2.

The RGB method had large errors due to the continuous output of incorrect poses such as bending forward throughout the sequence. This occurred because the synthetic training data was not sufficiently photorealistic or diverse with respect to the humanoid avatar’s appearance, background, and lighting, *i.e.*, there was a large domain gap between the test data collected in the real environment.

Although the optical flow method occasionally estimated correct walking and crouching motions, the overall output was quite unstable and inaccurate. This is because the real-world data contains complex optical flows such as objects and lighting around the subject, which bring a domain gap between the training data. In order to bridge this gap, it is necessary to diversify the background objects and lighting in the training data.

The SS method also had lower estimation accuracy than the method using our silhouetting process. Due to the distortion of the equirectangular images, the segmentation of human regions often failed, resulting in very noisy data with undesired regions other than the camera wearer’s body.

In contrast, our proposed methods using silhouetting processes (HRNet and Swin) were able to generate 3D human poses close to the ground truth. In particular, using the Swin Transformer model, *i.e.*, a state-of-the-art semantic segmentation method, we were able to correctly extract the silhouette and estimate the pose with high accuracy.

These results show that our silhouette-based pose estimation method using synthetic training data and a silhouetting process for inference works well in bridging the domain gap between synthetic data and real data.

TABLE 2. Quantitative evaluation results of pose estimation accuracy and joint acceleration on the MoCap test data and the in-the-wild data in Experiment 2.

Method	MoCap Test Data		In-the-wild Data	
	E_{MPJPE}	E_{acc3d}	E_{key}	E_{acc2d}
Mobile	99.7 ± 48.3	55.8	120.8 ± 58.2	85.0
Integral	83.2 ± 44.8	46.5	107.5 ± 49.2	78.5
Ours w/o PS	89.4 ± 34.9	24.3	106.1 ± 38.5	51.3
Ours w/ PS	86.2 ± 35.4	17.5	106.1 ± 47.6	48.1

B. RESULTS OF EXPERIMENT 2

For Experiment 2, quantitative results are shown in Table 2 and qualitative results are shown in Fig. 9. From the results in Table 2, the highest accuracy of pose estimation was achieved by the Integral method for the MoCap Test Data and our method (Ours w/ PS) for the In-the-wild Data. In contrast to the Integral method, which estimates joint positions, our method, which estimates joint angles, can cause accumulation of joint angle estimation errors in end joints such as hands and feet. While this may be the reason for the inferior accuracy of our method, the differences between them were not very large.

On the other hand, in the comparison of acceleration, our proposed method is much smoother than the Mobile and Integral methods. This is because the Mobile and Integral methods estimate the pose frame-by-frame, and the accuracy of the pose estimation is significantly decreased for images with severe noise or occlusion. In contrast, our network utilizes time-series features for inference, which results in smooth and stable motion estimation even when

such images are included in the sequence. In addition, our method (Ours w/ PS) has the lowest error among the E_{acc1} , indicating that the smoothness of the output poses is improved by mixing pseudo-silhouette images during training. These results demonstrate the effectiveness of our proposed method, *i.e.*, a pose estimation network that takes advantage of time series features and pseudo-silhouette images to improve the robustness of the model against silhouette noise.

VI. DISCUSSION AND LIMITATIONS

We proposed a new 3D pose estimation method using only a single 360° camera attached to the wrist for various applications such as motion recognition and motion analysis in the fields of sports, medicine, animation, and others. However, in order to use it for applications in real-world scenarios, it will be necessary to resolve the following issues.

A. PRACTICAL SIZE AND WEIGHT OF SENSOR DEVICES

Although we proposed an approach that utilizes a minimal number of devices, the commercial 360° camera we used in our experiment was not small and light enough to be usable on a daily basis on a wrist. However, since there are already smartwatches equipped with cameras, we believe that we may soon see a wristwatch-type device equipped with a 360° camera that can capture the wearer's body with a wide field of view, as shown in Fig. 1-(a).

B. VERSATILE AND HIGHLY ACCURATE POSE ESTIMATION

Although it is important to estimate various human poses for practical use, our model is currently not capable of estimating all human motions. However, since our method uses only MoCap data to generate synthetic training data, it is quite easy to prepare a more diverse data set using publicly available MoCap data. We will verify how well our method is able to estimate complex or unusual motions in a future study.

As for accuracy, there is a limit to the accuracy of silhouette extraction because we are using a semantic segmentation model trained on existing third-person images for egocentric equirectangular images. However, by taking advantage of the fact that foreground extraction using deep learning is not affected by domain shift [61], we have a possibility to easily create a deep learning model that can extract silhouettes with high accuracy using synthetic data.

It is also important to be able to estimate the user's global position. Several methods using camera localization algorithms have already been proposed [11], [16], [22], and there is a possibility of utilizing them in our method. However, we need to address the issue that the camera moves in a more complex manner since it is attached to the wrist in our method as opposed to the conventional methods where the camera is attached to the head or chest.

C. REAL-TIME POSE ESTIMATION

Real-time pose estimation technology enables a variety of new applications. Although our method can estimate the poses from silhouette images in real-time (about 100 FPS on

an Intel Core i9 and a GeForce RTX 3090), the silhouetting process runs at about 0.1 FPS. We believe that this can be improved by designing the silhouette extraction network described in Sec. VI-B to be able to process in real time.

VII. CONCLUSION

We presented a framework for estimating 3D human poses using a single wrist-mounted 360° camera. Our pose estimation network is trained only on synthetic silhouette image data generated in the virtual environment. For inference, our method uses binary silhouette images generated via a silhouetting process that takes real equirectangular images as input. Our silhouette-based method could reduce data generation costs and bridges the domain gap between synthetic and real data, which has been an issue in previous research. As an optional feature, we also proposed pseudo-silhouette image generation for noise robustness using conditional adversarial networks. Although we are currently assuming that there are real images synchronized with the MoCap data, we plan to train a generalized model by generating a larger dataset in the future. The experimental results have demonstrated the effectiveness of our silhouette-based approach and our pose estimation network, which takes advantage of the time-series features of the image sequences.

Our proposed framework can be easily extended to other camera positions. We hope that our method will become a stepping stone not only to further research on 3D pose estimation using wearable cameras but also to the discovery of new applications in the real world.

REFERENCES

- [1] Z. Liu, J. Zhu, J. Bu, and C. Chen, "A survey of human pose estimation: The body parts parsing based methods," *J. Vis. Commun. Image Represent.*, vol. 32, pp. 10–19, Oct. 2015.
- [2] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Comput. Vis. Image Understand.*, vol. 192, Mar. 2020, Art. no. 102897.
- [3] N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3D human pose estimation: A review of the literature and analysis of covariates," *Comput. Vis. Image Understand.*, vol. 152, pp. 1–20, Nov. 2016.
- [4] 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 *Proc. Int. Conf. 3D Vis. (DV)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2017, pp. 506–516. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/3DV.2017.00064>
- [5] 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," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Jul. 2017. [Online]. Available: <http://gvp.mpi-inf.mpg.de/projects/VNect/>
- [6] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Harvesting multiple views for marker-less 3D human pose annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1253–1262.
- [7] H. Rhodin, F. Meyer, J. Sporri, E. Müller, V. Constantin, P. Fua, I. Katircioglu, and M. Salzmann, "Learning monocular 3D human pose estimation from multi-view images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8437–8446.
- [8] 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.*, vol. 35, no. 6, pp. 1–11, Nov. 2016, doi: [10.1145/2980179.2980235](https://doi.org/10.1145/2980179.2980235).
- [9] Y. Huang, M. Kaufmann, E. Aksan, M. J. Black, O. Hilliges, and G. Pons-Moll, "Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time," *ACM Trans. Graph.*, vol. 37, pp. 185:1–185:15, Nov. 2018.

- [10] Y.-W. Cha, T. Price, Z. Wei, X. Lu, N. Rewkowski, R. Chabra, Z. Qin, H. Kim, Z. Su, Y. Liu, A. Ilie, A. State, Z. Xu, J.-M. Frahm, and H. Fuchs, "Towards fully mobile 3D face, body, and environment capture using only head-worn cameras," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 11, pp. 2993–3004, Nov. 2018.
- [11] Y.-W. Cha, H. Shaik, Q. Zhang, F. Feng, A. State, A. Ilie, and H. Fuchs, "Mobile. Egocentric human body motion reconstruction using only eyeglasses-mounted cameras and a few body-worn inertial sensors," in *Proc. IEEE Virtual Reality 3D User Interfaces (VR)*, Mar. 2021, pp. 616–625.
- [12] Y. Yuan and K. Kitani, "Ego-pose estimation and forecasting as real-time PD control," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10081–10091.
- [13] W. Xu, A. Chatterjee, M. Zöllhöfer, H. Rhodin, P. Fua, H.-P. Seidel, and C. Theobalt, "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.
- [14] D. Tome, T. Alldieck, P. Peluse, G. Pons-Moll, L. Agapito, H. Badino, and F. De la Torre, "SelfPose: 3D egocentric pose estimation from a headset mounted camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 2, 2020, doi: [10.1109/TPAMI.2020.3029700](https://doi.org/10.1109/TPAMI.2020.3029700).
- [15] T. Hu, K. Sarkar, L. Liu, M. Zwicker, and C. Theobalt, "EgoRenderer: Rendering human avatars from egocentric camera images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14528–14538.
- [16] V. Guzov, A. Mir, T. Sattler, and G. Pons-Moll, "Human POSEitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4316–4327.
- [17] H. Jiang and K. Grauman, "Seeing invisible poses: Estimating 3D body pose from egocentric video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3501–3509.
- [18] Y. Yuan and K. Kitani, "3D ego-pose estimation via imitation learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 763–778.
- [19] E. Ng, D. Xiang, H. Joo, and K. Grauman, "You2Me: Inferring body pose in egocentric video via first and second person interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9887–9897.
- [20] H. Jiang and V. K. Ithapu, "Egocentric pose estimation from human vision span," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11006–11014.
- [21] D. Tome, P. Peluse, L. Agapito, and H. Badino, "XR-EgoPose: Egocentric 3D human pose from an HMD camera," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7727–7737.
- [22] J. Wang, L. Liu, W. Xu, K. Sarkar, and C. Theobalt, "Estimating egocentric 3D human pose in global space," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11500–11509.
- [23] D.-H. Hwang, K. Aso, Y. Yuan, K. Kitani, and H. Koike, "MonoEye: Multimodal human motion capture system using a single ultra-wide fisheye camera," in *Proc. 33rd Annu. ACM Symp. User Interface Softw. Technol.*, Oct. 2020, pp. 98–111.
- [24] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 97–105.
- [25] J. N. Kundu, N. Lakkakula, and V. B. Radhakrishnan, "UM-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1436–1445.
- [26] R. Hori, R. Hachiuma, H. Saito, M. Isogawa, and D. Mikami, "Silhouette-based synthetic data generation for 3D human pose estimation with a single wrist-mounted 360° camera," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1304–1308.
- [27] *Vicon*. Accessed: May 21, 2022. [Online]. Available: <https://www.vicon.com/>
- [28] *Captury*. Accessed: May 21, 2022. [Online]. Available: <https://captury.com/>
- [29] 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.*, Jun. 2018, pp. 7122–7131.
- [30] 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. (CVPR)*, Jul. 2017, pp. 5689–5698.
- [31] 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. (CVPR)*, Jul. 2017, pp. 1263–1272.
- [32] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 529–545.
- [33] 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. (ICCV)*, Jun. 2017, pp. 2659–2668.
- [34] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 407–414.
- [35] G. Rogez, J. S. Supancic, and D. Ramanan, "First-person pose recognition using egocentric workspaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4325–4333.
- [36] H. Yonemoto, K. Murasaki, T. Osawa, K. Sudo, J. Shimamura, and Y. Taniguchi, "Egocentric articulated pose tracking for action recognition," in *Proc. 14th Int. Conf. Mach. Vis. Appl. (MVA)*, May 2015, pp. 98–101.
- [37] B. L. Bhatnagar, S. Singh, C. Arora, and C. V. Jawahar, "Unsupervised learning of deep feature representation for clustering egocentric actions," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1447–1453, doi: [10.24963/ijcai.2017/200](https://doi.org/10.24963/ijcai.2017/200).
- [38] 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, pp. 1–10, Jul. 2011, doi: [10.1145/2010324.1964926](https://doi.org/10.1145/2010324.1964926).
- [39] 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, pp. 248:1–248:16, Oct. 2015.
- [40] *Carnegie Mellon University Motion Capture Database*. Accessed: May 21, 2022. [Online]. Available: <http://mocap.cs.cmu.edu/>
- [41] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4627–4635.
- [42] *Maya*. Accessed: May 21, 2022. [Online]. Available: <https://www.autodesk.com/products/maya/>
- [43] *Animated 3D Characters*. Accessed: May 21, 2022. [Online]. Available: <https://www.mixamo.com/>
- [44] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba, "Recognizing scene viewpoint using panoramic place representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2695–2702.
- [45] *Blender*. Accessed: May 21, 2022. [Online]. Available: <https://www.blender.org>
- [46] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 302–321, Mar. 2019.
- [47] Y. Xu, V. Roy, and K. Kitani, "Estimating 3D camera pose from 2D pedestrian trajectories," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2568–2577.
- [48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [49] *Unity*. Accessed: May 21, 2022. [Online]. Available: <https://unity.com/>
- [50] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.
- [51] *MMSegmentation Contributors*. (2020). *MMSegmentation: OpenMMLab Semantic Segmentation ToolBox and Benchmark*. [Online]. Available: <https://github.com/open-mmlab/mms Segmentation>
- [52] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.
- [53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
- [55] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 5026–5033.
- [56] M. Isogawa, Y. Yuan, M. O'Toole, and K. Kitani, "Optical non-line-of-sight physics-based 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7011–7020.
- [57] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

- [58] 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. (CVPR)*, Jun. 2019, pp. 5607–5616.
- [59] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.
- [60] S. Choi, S. Choi, and C. Kim, "MobileHumanPose: Toward real-time 3D human pose estimation in mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2328–2338.
- [61] M. Rakesh, J. N. Kundu, V. Jampani, and V. B. Radhakrishnan, "Aligning silhouette topology for self-adaptive 3D human pose recovery," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 4582–4593. [Online]. Available: <https://openreview.net/forum?id=rsNBA9gtDf4>



RYOSUKE HORI (Student Member, IEEE) received the B.E. degree in information and computer science from Keio University, Japan, in 2021, where he is currently pursuing the M.Sc.Eng. degree in science and technology. His research interests include 3D human pose estimation and 3D hand reconstruction.



RYO HACHIUMA (Member, IEEE) received the B.E., M.Sc.Eng., and Ph.D. degrees in information and computer science from Keio University, Japan, in 2016, 2017, and 2021, respectively. His research interests include 3D object recognition, 3D object tracking, and simultaneous localization and mapping.



MARIKO ISOGAWA (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Osaka University, Japan, in 2011, 2013, and 2019, respectively. Since 2013, she has been with Nippon Telegraph and Telephone Corporation as a Researcher. She was a Visiting Scholar with Carnegie Mellon University, USA, from 2019 to 2020. Her research interests include computer vision and pattern recognition.



DAN MIKAMI (Member, IEEE) received the B.E. and M.E. degrees from Keio University, Japan, in 2000 and 2002, respectively, and the Ph.D. degree from the University of Tsukuba, Japan, in 2012. He joined Nippon Telegraph and Telephone Corporation (NTT), in 2002. He has been an Associate Professor with the Faculty of Informatics, Kogakuin University, since 2021. His research interests include computer vision, virtual reality, and computer-aided motor learning.



HIDEO SAITO (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Keio University, Japan, in 1992. Since 1992, he has been with the Faculty of Science and Technology, Keio University. From 1997 to 1999, he joined the Virtualized Reality Project at the Robotics Institute, Carnegie Mellon University, as a Visiting Researcher. Since 2006, he has been a Full Professor with the Department of Information and Computer Science, Keio University. His research interests include computer vision and pattern recognition, and their applications to augmented reality, virtual reality, and human–robotic interaction. His recent activities in academic conferences include being the Program Chair of ACCV 2014, the General Chair of ISMAR 2015, and the Program Chair of ISMAR 2016.

...