2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

EgoHMR: Egocentric Human Mesh Recovery via Hierarchical Latent Diffusion Model

Yuxuan Liu, Jianxin Yang, Xiao Gu, Yao Guo, *Member*, *IEEE*, and Guang-Zhong Yang, *Fellow*, *IEEE*

***Abstract*— Egocentric vision has gained increasing popularity in social robotics, demonstrating great potentials for personal assistance and human-centric behavior analysis. Holistic perception of human body itself is a prerequisite for downstream applications, including action recognition and anticipation. Extensive research has been performed for human mesh recovery from the exocentric images captured from a third-person view, but limited studies are conducted for heavily distorted yet occluded egocentric images. In this paper, we propose Egocentric Human Mesh Recovery (EgoHMR), a novel hierarchical network based on latent diffusion models. Our method takes a single egocentric frame as the input and it can be trained in an end-to-end manner without supervision of 2D pose. The network is built upon the latent diffusion model by incorporating both global and local features in a hierarchical structure. To train the proposed network, we generate weak labels from synchronized exocentric images. The proposed method can perform human mesh recovery directly from egocentric images and detailed quantitative and qualitative experiments have been conducted to demonstrate the effectiveness of the proposed EgoHMR method.**

I. Introduction

Supported by recent advances of AI and sensing technologies, social robots, either wearable or contact-free, are being actively explored for performing long-term monitoring and supporting daily activities for personalized intervention and assistance [1], [2]. Among different sensing modalities, vision is one of the mainstream techniques for research in social robotics, which gives sufficient visual cues of the human target and surrounding environments [3]–[6].

Egocentric vision is an emerging topic in social robotics, which represents the processing and analysis of images captured by either a head-mounted or a chest-mounted camera [7]. Compared to the conventional third-person-view vision with camera [5], [6], [8], egocentric vision is more flexible and convenient since the egocentric camera can move along with the human target in a free-living environment [7], [9]. According to the viewpoint of the body-mounted camera, the egocentric vision can be divided into the camera looking outwards [10], [11] and the camera looking downwards [12]–[14]. When the egocentric camera looks outwards, it can imitate the visual perception of the human by focusing more on the surrounding environments but less on the wearer him/herself. When the egocentric camera looks downwards, especially with a fisheye lens to enlarge the field of view, more information about the human target can be captured. Consequently, egocentric vision is advantageous for long-term human-centric perception, bringing new opportunities for social robotics in terms of human behavior analysis and social activities understanding [9], [14], [15].

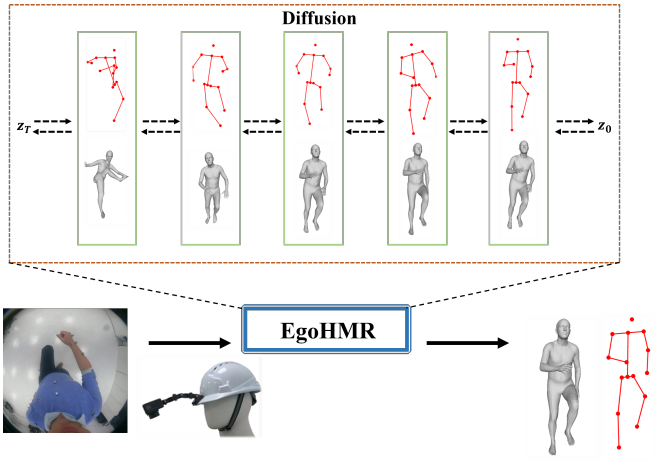

Fig. 1. Illustration of our proposed method for human mesh recovery based on a single egocentric image via latent diffusion. The upper box shows the process of the latent diffusion. The model step by step decodes the feature from the latent space and finally outputs the estimated human mesh and 3D pose. It can be found that from the early stage (the left part) the reconstruction is bad and unreasonable while from the final stage (the right part) the reconstruction is relatively accurate.

Egocentric human body reconstruction is one of the most significant prerequisites in human behavior analysis [16], [17], which refers to modeling the 3D human body of the person wearing the camera from the egocentric images. 3D human reconstruction is greatly influenced by the development of powerful human body models like SMPL [18]. The 3D human body reconstruction with these parametric models can be simplified into the task of predicting the shape and pose parameters and fitting in the defined body model to recover human meshes. Although there has been a tremendous process in 3D human modeling, the existing methods relying on images captured from the third-person-view camera easily fail to predict the accurate human body when directly applied to egocentric images. Recent research efforts have been largely devoted to human 3D pose estimation from egocentric images. Many related works focus on training the network either from synthetic images via fully supervised methods [12], [13], [19] or from real-world

This work was supported by the National Natural and Science Foundation of China under grant 62203296, Shanghai Pujiang Program under grant 22PJ1405500, the Science and Technology Commission of Shanghai Municipality under grant 20DZ2220400, and the Shanghai Pilot Program for Basic Research - Shanghai Jiao Tong University under grant 21TQ1400203. (*Corresponding authors: Yao Guo and Guang-Zhong Yang*)

Y. Liu, J. Yang, Y. Guo, G.-Z. Yang are with Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China, 200240. ({20000905lyx, jianxinyang, yao.guo, gzyang}@sjtu.edu.cn). X. Gu is with the Hamlyn Centre for Robotic Surgery, Imperial College London, London, UK. (xiao.gu17@imperial.ac.uk).

images via weakly supervision and domain adaptation [20], [21]. Compared to human pose estimation, the reconstruction of human mesh can provide more sufficient characterization of the human body but there is a lack of relevant research.

There are several inherent challenges for egocentric human body reconstruction. First, although increasing real-world egocentric datasets have been released, the annotations of egocentric images for body reconstruction are hard to acquire. Second, the 2D/3D pose serves as the strong information prior for most human mesh recovery methods from the third-person-view images while 2D/3D human pose can not be easily estimated from the egocentric images. Third, there exist severe human body self-occlusions from the egocentric view looking downwards, especially for the lower limbs, leading to significant uncertainties for egocentric pose estimation or body reconstruction.

To address aforementioned challenges, we propose a diffusion based method for human mesh recovery from the egocentric images, namely EgoHMR. To overcome the difficulties in acquiring ground truth for training, we leverage the extra third-person-view camera to provide weak supervision. Thus, we choose the ECHP [22] and EgoPW datasets [20] to train our model since these two datasets contain the synchronized images captured from both egocentric and third-person views. To avoid the needs for 2D/3D pose priors, our proposed method directly recovers the human mesh from the egocentric images and can be trained in an end-to-end manner. Furthermore, we design the network built upon a probabilistic diffusion model to characterize the uncertainty of the egocentric images and estimate the shape and pose parameters based on parametric human model SMPL [18]. To the best of our knowledge, this is the first work to perform the egocentric human mesh recovery from a fisheye camera looking downwards. It should be pointed out that our method can also be applied for egocentric human pose estimation by obtaining the joint positions directly from the SMPL model.

In summary, the main contributions of this paper are:

- We propose an end-to-end hierarchical diffusion probabilistic model for human mesh recovery from an egocentric image captured by a camera looking downwards.
- Our method skips the inaccurate and tedious process for egocentric 2D pose estimation, and achieves the comparable results with state-of-the-art methods which require 2D cues as supervision.
- Quantitative and qualitative results on different datasets demonstrate the effectiveness of our proposed hierarchical network for both human mesh recovery and human pose estimation.

II. Related Works

A. Egocentric 3D Human Body Modeling

With the progressing popularity of egocentric vision, 3D human body modeling has attracted great attention recently, which is the prerequisite for performing egocentric-related tasks, e.g., the human-centric action recognition, behavior analysis and social interaction.

Increasing research interests have been gained on 3D human pose estimation from egocentric images. On the one hand, some approaches focus on egocentric stereo vision via learning the feature correspondence to improve the performance. Rhodin et al. [23] was the first to perform this task from a head-mounted stereo fisheye camera system. Zhao et al. [24] proposed to utilize body part information to deal with the problem of body self-occlusions and limited body coverage. Akada et al. [19] proposed a new large-scale naturalistic dataset built on synthetic environments for stereo egocentric pose estimation. On the other hand, since the stereo egocentric vision system is inconvenient, more and more methods have appeared for egocentric pose estimation from monocular images. Tome et al. [12] and Xu et al. [13] proposed two large synthetic datasets for training the model to estimate the 3D human pose based on the 2D heatmaps. However, there exist serious domain shifts between the synthetic and real-world images. Wang et al. [20] combined the exocentric and egocentric real-world images together to generate the plausible weak labels to train the model on the real-world dataset.

In addition, some researchers perform the task of human mesh recovery from egocentric images. Liu et al. [25] proposed a simple yet effective optimization-based method to reconstruct 3D human body meshes from monocular egocentric videos by leveraging human-scene interaction constraints. Zhang et al. [26] proposed a new large-scale dataset for human pose and shape estimation from egocentric views. However, existing methods mostly reconstruct the body meshes of the interacting person from the second-person view in the egocentric videos, instead of the person wearing the egocentric camera from the first-person view. Consequently, our proposed method is designed for recovering human meshes of the person wearing the camera.

B. Diffusion Probabilistic Models

The diffusion model aims to learn the probabilistic distribution over gradually adding noises on inputs [27]. During the forward process, the original input is projected to the Gaussian noise by adding noise step by step, while during the inverse process the model denoises the Gaussian noise to generate the expected output. Since the processes of noise adding and removing enable the generated images highly robust to domain shifts, diffusion models have shown excellent performance in the field of image generation [28]–[30] and super-resolution [31]. Meanwhile, diffusion models have also been explored in 3D domains for point clouds generation and completion [32], [33], which show impressive ability to model point-voxel representations. Inspired by [28], diffusion models can be applied to the latent space for parameters regression as well. In this paper, we leverage the diffusion model to characterize the latent feature space of the egocentric images, and then predict the shape and pose parameters of the human target.

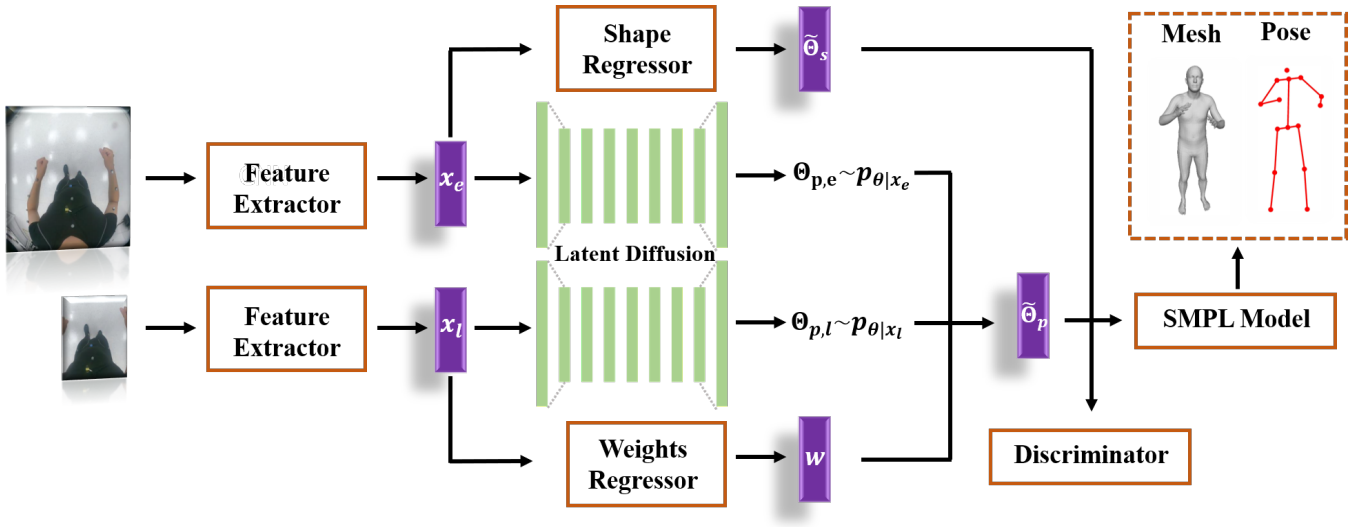

Fig. 2. Overview of the proposed EgoHMR network. The black arrows indicate the direction of the information flows. The egocentric image is fed into the feature extractor to obtain the global feature and local feature from the original image and the zoomed image, respectively. Then the hierarchical latent diffusion model is implemented to estimate the body pose parameters. A weights regressor is designed to estimate the weights to balance the lower limbs pose from two different branches. Meanwhile, the shape parameters are extracted by a regressor and a discriminator is introduced to make the estimated shape and pose more plausible. Finally, the output of our network is the human mesh based on SMPL model and 3D human pose can also be extracted from it.

III. Methodology

A. Problem Statement

In this paper, we aim to perform Human Mesh Recovery (HMR) from an egocentric RGB image captured by a head-mounted fisheye camera. We denote the captured image at each frame as I_e, and the zoomed image which contains the lower limb as I_l. The zoomed image facilitates the model to concentrate more on the details of the lower limbs, as they are always occluded by the upper body in egocentric vision. As shown in Fig. 2, our proposed model consists of five parts, i.e., a pair of feature extractors, a pair of latent diffusion models, a shape regressor, a weights regressor and a discriminator. Given the egocentric image pair $\{I_e, I_l\}$, the network estimates the SMPL parameters $\{\Theta_s, \Theta_p\}$ to recover the human mesh, where Θ_s and Θ_p refer to the shape parameters and pose parameters respectively. The positions of 3D body joints can be directly derived from the SMPL model as for egocentric human pose estimation.

B. Conditional Latent Diffusion Model

Before introducing the proposed network architecture, we give an overview of the conditional latent diffusion model. Consider that in latent space there are a series of latent variables $\{z_T, z_{T-1}, ..., z_1, z_0\}$ with decreasing levels of noise under the condition denoted as x, where z_0 is output of the diffusion model and z_T is sampled from standard normal distribution $z_T \sim N(0, \mathbf{I})$. The ground truth forward transition distribution is denoted as $q(z_t|z_{t-1}, x)$, which represents gradually adding noise to the former variable, and the diffusion model aims to learn the inverse conditional transition distributions $p_\theta(z_{t-1}|z_t, x)$ to recover the original variable from the noise sampled from normal distribution, where θ refers to the parameters of the diffusion models. According to Markov transition probabilities, the transition probabilities can be written as:

$$q(z_{0:T}, x) = q(z_0)\Pi_{t=1}^{T} q(z_t|z_{t-1}, x) \tag{1}$$

$$p_\theta(z_{0:T}, x) = p(z_T)\Pi_{t=1}^{T} p_\theta(z_{t-1}|z_t, x) \tag{2}$$

Let we denote $\{\beta_1, \beta_2, ..., \beta_T\}$ as a sequence of coefficients to control the process of adding noise, following to the formulations in [27], we simplify the inverse transition distribution:

$$q(z_t|z_{t-1}, x) \sim N(\sqrt{1-\beta_t} z_{t-1}, \beta_t \mathbf{I}) \tag{3}$$

$$p_\theta(z_{t-1}|z_t, x) \sim N(\mu_\theta(z_t, t, x), \beta_t^2 \mathbf{I}) \tag{4}$$

$$\mu_\theta(z_t, t, x) = \frac{1}{\sqrt{1-\beta_t}}(z_t - \frac{\beta_t}{\sqrt{1-\Pi_s^t(1-\beta_s)}}\epsilon_\theta(z_t, x, t)) \tag{5}$$

where $\epsilon_\theta(z_t, x, t)$ can be calculated as the output from the reparameterized model and $\mu_\theta(z_t, t, x_t)$ represents the mean of the estimated distribution. The training objective is to learn the marginal likelihood of $p_\theta(z, x)$ and it equals to constrain the $\mathcal{L}_2$ loss between output $\epsilon_\theta(z_t, x, t)$ and noise ϵ sampled from standard normal distribution:

$$\mathcal{L}^{diff} = ||\epsilon_\theta(z_t, x, t) - \epsilon||_2 \tag{6}$$

During the inverse process, the diffusion model generate the final output z_0 from the sampled noise $\epsilon \sim N(0, \mathbf{I})$ according to the learned distribution $p_\theta(z_{t-1}|z_t, x)$ step by step in Eq.(7).

$$z_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(z_t - \frac{\beta_t}{\sqrt{1-\Pi_s^t(1-\beta_s)}}\epsilon_\theta(z_t, x, t)) + \sqrt{\beta_t}\epsilon \tag{7}$$

C. Model Design

The overview of our designed model is shown in Fig.2. The input of our model is the pair of an egocentric image and its zoomed patch which contains lower limbs of the human target. Given the input pair $\{I_e, I_l\}$, we encode it with two separate feature extractors denoted as $\{E_e(\cdot), E_l(\cdot)\}$ and obtain the global feature $x_e = E_e(I_e)$ and the local feature $x_l = E_l(I_l)$, respectively. The input size of the original image is 384×384 and the zoomed image is 56×56.

After obtaining global feature and local feature, we implement two latent diffusion models $\{D_e(\cdot), D_l(\cdot)\}$ to learn the conditional distribution $p_{\theta|x_e}(z, x_e)$ and $p_{\theta|x_l}(z, x_l)$. The final output of the two diffuion models are SMPL pose parameters $\{\Theta_{p,e}, \Theta_{p,l}\}$. The pose parameters in $\Theta_{p,e}$ contain $N_e = 24$ body joints while in $\Theta_{p,l}$ only contain $N_l = 9$ lower body joints. To balance the pose parameters from different diffusion models, we design a reweight module to regress the average weights. The reweight module $R(\cdot)$ is built based on MLP and takes the local feature as input and regresses the weight $w = R(x_l)$ for nine lower body joints. To make the weight plausible, the last layer of the MLP is sigmoid function so that the range of the weights is limited to $[0, 1.0]$. The averaged pose parameters are:

$$\tilde{\Theta}_p = (1 - w)\Theta_{p,e} + w\Theta_{p,l} \tag{8}$$

The shape parameters $\tilde{\Theta}_s$ are inferred by passing the global feature x_e via a MLP based regressor $S(\cdot)$, i.e., $\tilde{\Theta}_s = S(x_e)$. A discriminator is also implemented to make the estimated pose and shape parameters more plausible. The estimated SMPL parameters $\{\tilde{\Theta}_p, \tilde{\Theta}_s\}$ are fed into SMPL models to generate the human mesh as well as the 3D human pose.

Finally we introduce the loss functions in our proposed model. Since there is no available ground truth under the egocentric view, we supervise our proposed model with the generated weak labels. The weak labels can be derived from the synchronized third-person-view images from the existing exocentric human mesh recovery method [34]. Assume that we have collected the weak labels of SMPL parameters $\{\Theta_p^w, \Theta_s^w\}$ of the input egocentric images. To train the diffusion models, the diffusion loss $\{\mathcal{L}_l^{diff}, \mathcal{L}_e^{diff}\}$ is implemented for $\{D_e(\cdot), D_l(\cdot)\}$:

$$\mathcal{L}_l^{diff} = ||\epsilon_\theta(z_t, x_l, t) - \epsilon||_2, \mathcal{L}_e^{diff} = ||\epsilon_\theta(z_t, x_e, t) - \epsilon||_2 \tag{9}$$

We also explicitly supervise the estimated parameters with weak labels and denote this loss function as $\mathcal{L}^{pm}$:

$$\mathcal{L}^{pm} = ||\Theta_p^w - \tilde{\Theta}_p||_2 + ||\Theta_s^w - \tilde{\Theta}_s||_2 \tag{10}$$

Similar to [35], we add extra constraints to the predicted pose parameters with loss function $\mathcal{L}^{orth}$ which forces the estimated pose parameters to be close to the orthonormal 6D representations and adversarial learning loss $\mathcal{L}^{adv}$ to generate more reliable parameters. Eventually, the training loss function $\mathcal{L}$ is:

$$\mathcal{L} = \mathcal{L}_l^{diff} + \mathcal{L}_e^{diff} + \mathcal{L}^{pm} + \mathcal{L}^{orth} + \mathcal{L}^{adv} \tag{11}$$

During the inference stage, our proposed hierarchical model can generate the results when the egocentric image and the zoomed lower body part are fed into the network. Meanwhile, our model can also work when only the egocentric image is taken as input without the zoomed part, and we set the weights w of the lower body part to zero.

IV. Experiments

A. Datasets

Since traditional methods trained on exocentric images fail to recover the egocentric human mesh, it is difficult to directly generate weak labels from the egocentric images. Consequently, we choose the **ECHP** [22] and **EgoPW** [20] datasets which contain the synchronized images from the third-person view and the egocentric view. We apply PARE [34] on the exocentric images to generate the weak labels $\{\Theta_p^w, \Theta_s^w\}$. For the ECHP dataset, the data was collected from two third-person-view cameras and one head-mounted egocentric camera, and there are 9 different subjects performing 10 daily actions. For the EgoPW dataset, the data was collected from one third-person-view camera and one head-mounted egocentric camera, which contains 97 sequences of 10 actors performing 20 different daily actions. Following the experiments in [20], we test the model on the real-world dataset from [36] about 12k frames.

B. Implementation Details

The feature extractor $\{E_e(\cdot), E_l(\cdot)\}$ has the same structure as ResNet50 which is built upon several basic CNN blocks with convolution layers, batchnorm layers and relu activation function. The output size of feature extractor is 2048. For other MLP based models, we implement the basic MLP layer proposed in [37]. The diffusion model $\{D_e(\cdot), D_l(\cdot)\}$ is built upon three basic MLP layers with input size 2048 and iteration times $T = 10$. The regressors for shape parameters and lower limb weight are both built upon one basic MLP layers with input size 2048 and output 10 shape parameters and 9 lower limb weights. The discriminator has the same structure in [38] with 2D convolution layers and MLP layers. The models mentioned above are implemented by PyTorch and trained 30 epochs. We apply Adam for optimization with a learning rate of 0.0001.

C. Evaluation Metrics

Since it is hard to acquire the ground truth of human mesh for egocentric images, we give the quantitative results of 3D human pose estimation obtained from the egocentric human mesh recovery for comparisons. The evaluation protocol we use is **PA-MPJPE**, which calculates the Mean Per Joint Position Error after applying Procrustes Analysis between the ground truth and the estimated results. Since the results of our model depend on the sampling noise values, we sample 10 times and report the mean value to represent performance of the model.

TABLE I

RESULTS IN MILLIMETERS (mm) ON EGOPW DATASET

Approaches	All
Tome [12]	112.0
Xu [13]	102.3
Kolotouros [35]	88.6
Wang† [20]	**84.2**
w/o $D_l(\cdot)$	88.0
w/o $R(\cdot)$	90.2
w/o $\mathcal{L}_l^{diff}$	89.2
w/o $\mathcal{L}^{pm}$	120.4
Ours	<u>85.8</u>

† More training data (training on both synthetic and real-world images).

TABLE II

RESULTS IN MILLIMETERS (mm) ON ECHP DATASET

Approaches	All	Squatting	Walking	Dancing	Stretching	Waving	Boxing	Kicking	Touching	Clamping	Knocking
Tome [12]	73.9	78.4	72.6	79.8	78.2	63.9	71.9	73.1	67.7	81.1	72.0
Xu [13]	71.2	70.7	76.8	70.1	69.5	65.5	68.1	72.0	65.8	85.9	66.0
Kolotouros [35]	<u>66.8</u>	77.6	75.8	79.4	67.8	52.9	63.3	69.5	57.3	53.8	58.8
w/o $D_l(\cdot)$	69.7	69.3	84.9	79.8	72.3	54.6	68.3	70.7	61.2	55.3	64.3
w/o $R(\cdot)$	67.5	71.6	81.6	76.3	68.5	49.9	68.0	69.9	54.9	56.0	61.2
w/o $\mathcal{L}_l^{diff}$	68.9	75.6	79.8	77.2	68.3	51.6	68.2	71.8	56.3	64.0	62.2
w/o $\mathcal{L}^{pm}$	149.5	138.2	148.0	147.6	156.0	147.2	155.4	141.2	149.9	157.7	155.3
Ours	**66.3**	71.2	77.7	73.5	70.0	50.3	65.0	68.1	56.7	56.5	60.4

D. Comparison Methods and Ablation Studies

On one hand, we compare our proposed method with several egocentric 3D human pose estimation methods. Tome [12] proposed a three-branch encoder-decoder network for pose estimation and Xu [13] proposed a two-branch network with both original and zoomed images as input, which is similar to the hierarchical network we propose. Wang [20] proposed to learn the different representations between synthetic and real-world images, exocentric and egocentric images, which achieved the state-of-the-art performance on egocentric human pose estimation. The three methods all require egocentric 2D pose as supervision cues while our proposed method directly estimates 3D pose without 2D supervision. Since the source code of Wang has not been published yet, we use the results reported in their original paper. Wang trained their method with the combination of synthetic images and real-world images while our proposed method only depends on the real-world images, and we report the result without learning domain adaptation to make the fair comparison. On the other hand, we compare our model with the human mesh recovery method and do ablation studies. We compare our method with [35]. Kolotouros [35] proposed to implement normalizing flows to characterize the probability of human pose estimation and the learned distribution can be optimized for different downstream tasks. Here we choose the mode value of the normal distribution as the sample value to represent the learned model. We also conduct several ablation studies to demonstrate the effectiveness of different modules. The comparison methods and ablation studies are noted as follows:

- Tome [12], a three-branch encoder-decoder network for egocentric pose estimation trained with 2D pose and 3D pose as supervision.
- Xu [13], a two-branch network for egocentric pose estimation with original and zoomed images as input which requires 2D pose and 3D pose as supervision.
- Wang [20], the state-of-the-art method for egocentric pose estimation with three different kinds of images as input which requires 2D pose and 3D as weak supervision.
- Kolotouros [35], an approach learns the distribution of plausible 3D poses based on normalizing flows and can be optimized to different downstream applications.
- w/o $D_l(\cdot)$, an ablated model by removing the branch of zoomed images and the estimation only depends on the global feature.
- w/o $R(\cdot)$, an ablated model by removing the reweight process and manually select the w as 0.5.
- w/o $\mathcal{L}_l^{diff}$, an ablated model by removing the loss function of the second diffusion model.
- w/o $\mathcal{L}^{pm}$, an ablated model by removing the loss function of explicitly supervising pose parameters.

V. RESULTS AND ANALYSIS

A. Quantitative Results

We report the quantitative results of human pose estimation since the two datasets provide ground truth 3D pose for evaluation. Without further clarification, the **bold** and the <u>underline</u> values in the table refer to the best and the second best performance in each column, respectively. We report all the results in millimeters (mm).

1) EgoPW dataset: We evaluate our proposed method on the public EgoPW dataset and report the PA-MPJPE for the average performance in the upper part of Table I. It can be found that our method achieves the comparable performance (PA-MPJPE=85.8mm) with the state-of-the-art model. Note that our method has two potential advantages over theirs. The method proposed by Wang et.al. is supervised under both 2D pose and 3D pose while ours skips the inaccurate process of 2D pose estimation, and their method trains the model with synthetic dataset Mo^2Cap^2 [13] and real-world dataset EgoPW [20] while ours only trains on EgoPW [20] with less training data. It can also be found that the performance of our method is much better than other egocentric pose estimation methods [12], [13] with the improvement PA-MPJPE about 25mm. Our method also performs better than the existing human mesh recovery method Kolotouros et.al. [35] with an improvement of about 3mm.

2) ECHP dataset: We evaluate our proposed method on the ECHP dataset and report the PA-MPJPE for the average performance and for each action in the upper part of Table II. It can be found that our proposed method achieves the

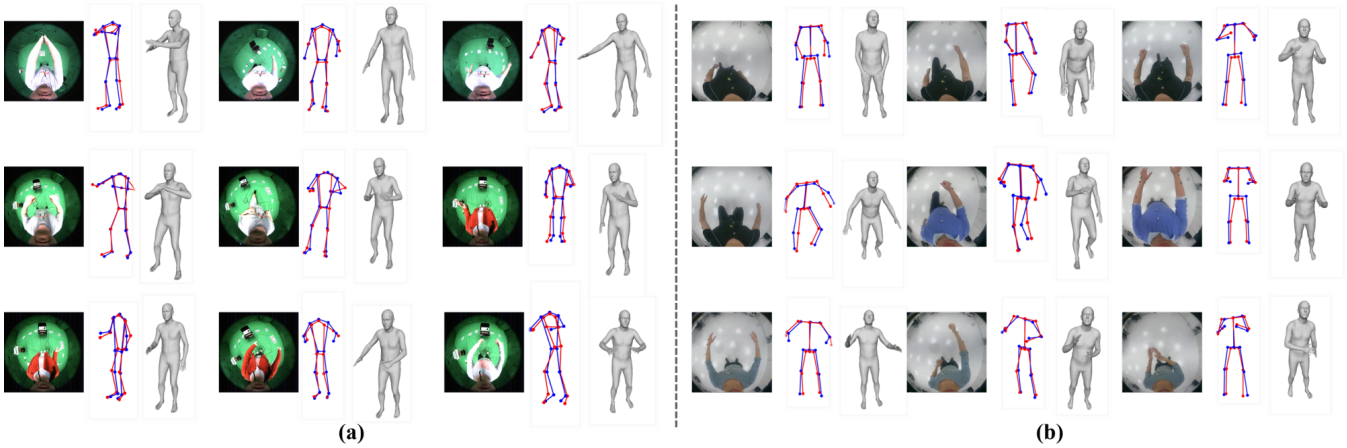

Fig. 3. Visualization results of our proposed method on test data. (a) On EgoPW test dataset. (b) On ECHP test dataset. The egocentric images for test do not have the corresponding exocentric images for generating weak labels. For human pose estimation, the red color is the predicted 3D pose by our method and the blue color is the ground truth.

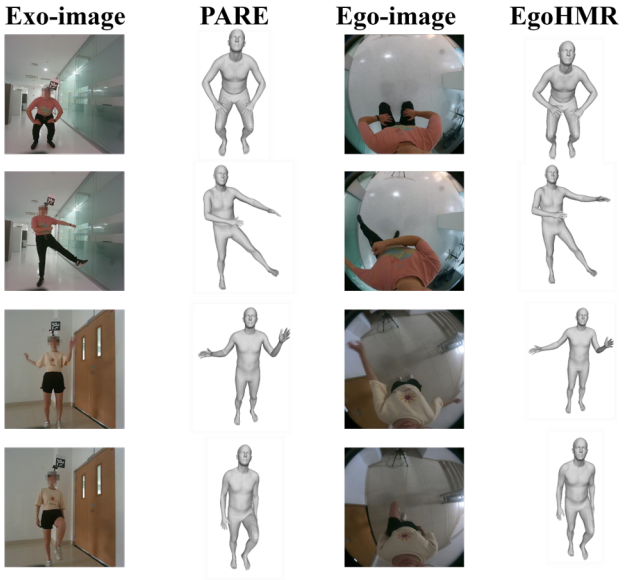

Fig. 4. Visualization results of our proposed method compared with state-of-the-art exocentric human mesh recovery method [34]. **Note that these images are not used for training**. Our method can achieve the comparable performance while only taking the egocentric images as input.

best performance (PA-MPJPE=66.3mm) compared to three different methods [12], [13], [35]. Among ten different actions, our method achieves the best performance or the second performance on eight actions, demonstrating the effectiveness of our proposed method.

3) Ablation studies: We conduct four ablation studies to evaluate the performance of different parts in our network architecture and report the PA-MPJPE in the lower part of Table I and Table II. It can be found that different parts of our proposed method all play crucial roles to perform the task. The hierarchical structure $D_l(\cdot)$ performs better than the single branch because the zoomed image provides the local feature of the lower body part which promotes the model to estimate the lower joints. The reweight process $R(\cdot)$ makes the model learn the weights distribution instead of selecting appropriate weights manually. The loss function $\mathcal{L}_l^{diff}$ helps the latent diffusion model conditioned on local feature converge faster. It can also be found that the loss function $\mathcal{L}^{pm}$ which explicitly supervises the pose and shape parameters with weak labels improves the performance best according to four ablation studies. It implies that weak labels play important roles in the egocentric human mesh recovery and the explicit supervision enhances the performance.

B. Qualitative Results

Fig. 3 presents the visualization results of our proposed method for egocentric human mesh recovery as well as human pose estimation on two datasets. The left part refers to the performance of the model on EgoPW dataset and the right part on ECHP dataset. For human pose estimation, we visualize both the estimated 3D pose and the ground truth in red color and blue color, respectively. Fig. 4 shows the results of our proposed method with existing exocentric human mesh recovery method [34]. Note that images in Fig. 4 are not used for training the model. Given a single egocentric image, it can be seen that EgoHMR can estimate reasonable human mesh and 3D pose for different actions, even for the images with severe body occlusions, which demonstrates the good performance of our proposed method.

VI. Conclusions

This paper proposes a hierarchical latent diffusion model called EgoHMR to perform the task of egocentric human mesh recovery from a single RGB image. We take the first attempt to perform the human mesh recovery from a single egocentric image without 2D pose supervision and the results on different datasets demonstrate the effectiveness of our proposed method. Specifically, we generate weak labels from the synchronized exocentric images to get rid of the ground truth acquisition, and we also design a hierarchical latent diffusion network to incorporate the global feature from the original egocentric image and the local feature from the zoomed image focusing on the lower body part. In future work, we are going to add temporal constraints to the 3D human modeling and incorporate the egocentric human mesh recovery with more tasks in social robotics.

REFERENCES

[1] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science robotics*, vol. 3, no. 21, p. eaat5954, 2018.

[2] Y. Guo, W. Chen, J. Zhao, and G.-Z. Yang, "Medical robotics: opportunities in china," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 361–383, 2022.

[3] Y. Guo, Y. Li, and Z. Shao, "Rrv: A spatiotemporal descriptor for rigid body motion recognition," *IEEE transactions on cybernetics*, vol. 48, no. 5, pp. 1513–1525, 2017.

[4] S. Liu, G. Tian, Y. Zhang, and P. Duan, "Scene recognition mechanism for service robot adapting various families: A cnn-based approach using multi-type cameras," *IEEE Transactions on Multimedia*, vol. 24, pp. 2392–2406, 2021.

[5] Y. Guo, F. Deligianni, X. Gu, and G.-Z. Yang, "3-d canonical pose estimation and abnormal gait recognition with a single rgb-d camera," *IEEE Robotics and Automation letters*, vol. 4, no. 4, pp. 3617–3624, 2019.

[6] J. Yang, Y. Liu, X. Gu, G.-Z. Yang, and Y. Guo, "Posesdf: Simultaneous 3d human shape reconstruction and gait pose estimation using signed distance functions," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 1297–1303.

[7] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 995–19 012.

[8] X. Gu, Y. Guo, G.-Z. Yang, and B. Lo, "Cross-domain self-supervised complete geometric representation learning for real-scanned point cloud based pathological gait analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1034–1044, 2021.

[9] S. Alletto, G. Serra, S. Calderara, and R. Cucchiara, "Understanding social relationships in egocentric vision," *Pattern Recognition*, vol. 48, no. 12, pp. 4082–4096, 2015.

[10] H. Jiang and K. Grauman, "Seeing invisible poses: Estimating 3d body pose from egocentric video," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3501–3509.

[11] Y. Yuan and K. Kitani, "Ego-pose estimation and forecasting as real-time pd control," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 082–10 092.

[12] 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 Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[13] W. Xu, A. Chatterjee, M. Zollhoefer, H. Rhodin, P. Fua, H.-P. Seidel, and C. Theobalt, "Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 5, pp. 2093–2101, 2019.

[14] Y. Liu, J. Yang, X. Gu, Y. Guo, and G.-Z. Yang, "Ego+x: An egocentric vision system for global 3d human pose estimation and social interaction characterization," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 5271–5277.

[15] D. Kim, B. B. Kang, K. B. Kim, H. Choi, J. Ha, K.-J. Cho, and S. Jo, "Eyes are faster than hands: A soft wearable robot learns user intention from the egocentric view," *Science Robotics*, vol. 4, no. 26, p. eaav2949, 2019.

[16] X. Wang, L. Zhu, H. Wang, and Y. Yang, "Interactive prototype learning for egocentric action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8168–8177.

[17] Y. Wu, L. Zhu, X. Wang, Y. Yang, and F. Wu, "Learning to anticipate egocentric actions by imagination," *IEEE Transactions on Image Processing*, vol. 30, pp. 1143–1152, 2020.

[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics (TOG)*, vol. 34, no. 6, pp. 1–16, 2015.

[19] H. Akada, J. Wang, S. Shimada, M. Takahashi, C. Theobalt, and V. Golyanik, "Unrealego: A new dataset for robust egocentric 3d human motion capture," in *European Conference on Computer Vision (ECCV)*, 2022.

[20] J. Wang, L. Liu, W. Xu, K. Sarkar, D. Luvizon, and C. Theobalt, "Estimating egocentric 3d human pose in the wild with external weak supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 157–13 166.

[21] A. Dhamanaskar, M. Dimiccoli, E. Corona, A. Pumarola, and F. Moreno-Noguer, "Enhancing egocentric 3d pose estimation with third person views," *arXiv preprint arXiv:2201.02017*, 2022.

[22] Y. Liu, J. Yang, X. Gu, Y. Chen, Y. Guo, and G.-Z. Yang, "Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning," *IEEE Transactions on Multimedia*, pp. 1–12, 2023.

[23] 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 Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–11, 2016.

[24] D. Zhao, Z. Wei, J. Mahmud, and J.-M. Frahm, "Egoglass: Egocentric-view human pose estimation from an eyeglass frame," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 32–41.

[25] M. Liu, D. Yang, Y. Zhang, Z. Cui, J. M. Rehg, and S. Tang, "4d human body capture from egocentric video via 3d scene grounding," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 930–939.

[26] S. Zhang, Q. Ma, Y. Zhang, Z. Qian, T. Kwon, M. Pollefeys, F. Bogo, and S. Tang, "Egobody: Human body shape and motion of interacting people from head-mounted devices," in *European conference on computer vision (ECCV)*, Oct. 2022.

[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[29] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[30] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 696–10 706.

[31] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *arXiv preprint arXiv:2104.07636*, 2021.

[32] S. Luo and W. Hu, "Diffusion probabilistic models for 3d point cloud generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2837–2845.

[33] L. Zhou, Y. Du, and J. Wu, "3d shape generation and completion through point-voxel diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5826–5835.

[34] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "Pare: Part attention regressor for 3d human body estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 127–11 137.

[35] N. Kolotouros, G. Pavlakos, D. Jayaraman, and K. Daniilidis, "Probabilistic modeling for human mesh recovery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 605–11 614.

[36] J. Wang, L. Liu, W. Xu, K. Sarkar, and C. Theobalt, "Estimating egocentric 3d human pose in global space," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 500–11 509.

[37] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2640–2649.

[38] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7122–7131.