Probabilistic Human Mesh Recovery in 3D Scenes from Egocentric Views

Siwei Zhang\textsuperscript{1} Qianli Ma\textsuperscript{1,3} Yan Zhang\textsuperscript{1} Sadegh Aliakbarian\textsuperscript{2} Darren Cosker\textsuperscript{2} Siyu Tang\textsuperscript{1}
\textsuperscript{1}ETH Zürich \textsuperscript{2}Microsoft \textsuperscript{3}Max Planck Institute for Intelligent Systems
\{siwei.zhang, qianli.ma, yan.zhang, siyu.tang\}@inf.ethz.ch
\{coskerdarren, saliakbarian\}@microsoft.com

Figure 1: We propose EgoHMR, a novel scene-conditioned probabilistic method to recover the human mesh from an egocentric view image (typically with the body truncated) given the 3D environment. EgoHMR efficiently leverages the scene geometry and achieves significantly more plausible human-scene interactions compared to baseline methods, ProHMR [32] and the extended ProHMR-scene (Sec. 5.3), even with severe body truncations.

Abstract

Automatic perception of human behaviors during social interactions is crucial for AR/VR applications, and an essential component is estimation of plausible 3D human pose and shape of our social partners from the egocentric view. One of the biggest challenges of this task is severe body truncation due to close social distances in egocentric scenarios, which brings large pose ambiguities for unseen body parts. To tackle this challenge, we propose a novel scene-conditioned diffusion method to model the body pose distribution. Conditioned on the 3D scene geometry, the diffusion model generates bodies in plausible human-scene interactions, with the sampling guided by a physics-based collision score to further resolve human-scene interpenetrations. The classifier-free training enables flexible sampling with different conditions and enhanced diversity. A visibility-aware graph convolution model guided by per-joint visibility serves as the diffusion denoiser to incorporate inter-joint dependencies and per-body-part control. Extensive evaluations show that our method generates bodies in plausible interactions with 3D scenes, achieving both superior accuracy for visible joints and diversity for invisible body parts. The code is available at https://sanweiliti.github.io/egohmr/egohmr.html.

1. Introduction

With the rapid development of Augmented and Virtual Reality (AR/VR) devices, understanding human actions, behaviors and interactions of our social partners (“interactee”) from the egocentric view is a vital component for head-mounted devices (HMDs) to truly become a virtual companion for humans. The first step towards the automatic perception of human behaviors and interactions for HMDs is to estimate 3D human pose and shape (i.e. human mesh recovery) of the interactee from egocentric view images.

The research community has extensively studied human mesh recovery (HMR) from a single RGB image (usually captured with third-person view cameras) by predicting SMPL [39, 50] parameters from global image features [8, 24, 27, 30, 45]. However, their performance degrades significantly in egocentric images [78]. A major challenge presented in egocentric scenarios is frequent body truncation [38, 78], when individuals interact with each other within close proximity, while the HMD camera has a limited field-of-view. Since human poses are highly reliant on the surrounding environment, 3D scene structures can potentially provide strong cues to infer invisible body parts, which is crucial for precise understanding of human behaviors: sitting on a chair and standing on the ground may indicate distinct intentions and future behaviors. Following
PROX [15], we assume a rough 3D scene structure is available, as such information nowadays can be easily obtained with commodity sensors. Furthermore, visual localization and mapping, one of the extensively studied topics in computer vision, is progressing rapidly for head-mounted devices [52]. Therefore, we make the assumption that a coarse 3D model of the scene and the localization of the egocentric camera are readily available (e.g. in HoloLens2 [1]). In this paper, we focus on the challenging problem of estimating 3D human bodies that are heavily truncated from egocentric views due to the proximity between people and the motion of embodied cameras.

Given an egocentric image and the corresponding 3D environment, what should an ideal human mesh recovery method achieve? In contrast to previous studies that mostly pursue full-body pose accuracy, we argue the following properties are desired in the egocentric scenario: (1) natural and plausible human-scene interactions; (2) accurate body pose estimations consistent with image observations for visible joints; (3) a comprehensive conditional distribution to generate diverse and plausible poses for unobserved body parts. Several recent studies have attempted to tackle pose ambiguities caused by occlusions and body truncations. However, some methods can only produce a discrete number of body poses [3, 34, 48], ignoring the continuous nature of the pose manifold. Other approaches model the continuous body pose distribution via conditional variational autoencoder [53] or normalizing flows [32, 68], but with limited expressiveness. Furthermore, the 3D environment is often ignored although it provides strong cues for inferring missing body parts in the image. Existing works for human pose estimation [54, 55, 79] or generation [14, 80–82] in 3D scenes typically cannot deal with truncated bodies in images.

To address these issues, we introduce a novel scene-conditioned probabilistic approach, the first method to recover human mesh in 3D scenes from the egocentric view image. Inspired by the recent diffusion models that can generate high fidelity images [9, 17, 18, 59] and human motions [6, 25, 41, 63, 77] with flexible conditioning, our model is trained with a conditional diffusion framework, leveraging both the classifier-free guidance [18] and the classifier-guided diffusion sampling [9, 57, 60] for efficient scene conditioning. By training the model conditioning on a human-centric scene encoding, the diffusion denoiser can generate body poses with plausible human-scene interactions even with highly truncated bodies (see Fig. 1). The classifier-free training enables flexible sampling from models with or without image conditions, achieving accurate estimations of visible body parts while generating diverse plausible results for unseen body parts. On top of that, the diffusion sampling process is guided by the gradients of a physics-based collision score guidance in the classifier-guided sampling manner, which further resolves human-scene inter-penetrations without requiring additional time-consuming postprocessing.

To facilitate the learning of the multimodal distributions of the invisible body parts, we introduce a visibility-aware graph convolution model as the diffusion denoiser network, to explicitly guide the network to learn more expressive distribution for invisible body parts. Unlike existing methods [24, 32, 36, 68] that simply condition the full body pose on a global image feature, we condition the pose diffusion of each body joint on the joint visibility mask and the 3D scene feature, in addition to the image feature. With such explicit visibility guidance, the model learns to estimate accurate body pose for visible joints while encouraging diversity for truncated body parts. We adopt a graph convolution network (GCN) [26, 75, 84] to better incorporate local dependencies between highly-relevant body parts (e.g. knee-foot) according to the human kinematic tree.

In summary, our contributions are: 1) a novel scene-conditioned diffusion model for probabilistic human mesh recovery in the 3D environment from egocentric images; 2) a physics-based collision score that guides the diffusion sampling process to further resolve human-scene interpenetrations; 3) a visibility-aware GCN architecture that incorporates inter-joint dependencies for the pose diffusion, and enables per-body-part control via the per-joint visibility conditioning. With extensive evaluations, the proposed method demonstrates superior accuracy and diversity of generated human bodies from egocentric images, in natural and plausible interactions with the 3D environment.

2. Related Work

Human mesh recovery from a single image. Given a single RGB image, the task of deterministic 3D human mesh recovery has been widely studied in the literature, with regression-based methods [7, 8, 11, 24, 27, 28, 30, 31, 35–37, 45, 49, 62, 71, 76, 83], optimization-based methods [4, 10, 33, 50, 69] or hybrid methods [23, 29, 58], mostly adopting parametric 3D body models [4, 70] to represent the 3D human mesh. Most of recent regression-based methods train neural networks to regress SMPL [4] parameters from images [11, 24, 28, 36, 51]. Model-free methods [7, 8, 37, 45] regress human mesh vertices without relying on parametric body models. Optimization-based methods iteratively optimize the body model parameters to fit 2D keypoints, human silhouettes, etc. Some resort to self-contact [46], scene [15, 69] or object [61] constraints to model interactions. Hybrid methods seek to combine the strengths of both regression and optimization, for instance, SPIN [29], which integrates an optimization loop into the deep learning framework. However, these methods often lack robustness when dealing with truncated bodies, a typical challenge in egocentric scenarios, and cannot
model the diverse nature of plausible missing body parts. 3D scene information is mostly disregarded, resulting in unrealistic human-scene interactions. In this work, we address these challenges by proposing a scene-conditioned generative model for human mesh recovery in 3D scenes.

**Multi-hypothesis for 3D human pose estimation.** Due to the limited image observation and depth ambiguity, estimating 3D human pose from a single image can have multiple potential solutions, especially when body truncation is presented. Recent works seek to model this task as a generative process or predict multiple hypotheses of possible poses. A discrete number of hypotheses are generated in [3, 21, 34, 48]. To learn the continuous distribution of possible body poses, conditional variational autoencoders [53] or conditional normalizing flows [32, 68] are adopted to generate unlimited number of hypotheses, but relying on additional scoring functions for pose selection [33], or discriminators to stabilize the training [32, 68]. Nevertheless, the 3D scene constraint is usually not taken into account in this line of works. We leverage the diffusion process to model the inherent pose ambiguity, with flexible scene-conditioning strategies.

**Egocentric human pose estimation.** Most of existing works in egocentric human pose estimation targets the camera wearer’s 3D pose [13, 22, 40, 56, 64, 65, 73]. 3D pose estimation for a social partner from the egocentric view has been receiving growing attention [38, 47, 72]. However, the 3D scene geometry is neglected in [47, 72], while [38] employs the time-consuming optimization, without modelling the probabilistic distribution of unseen body parts.

**Diffusion models for human motion/pose generation.** Denoising Diffusion Models [17, 57, 59, 60] are becoming a popular choice for human motion generative models. Typically trained with the DDPM formulation [17] and based on Transformer-backbones, different methods vary on their choice of motion representation (e.g. skeletons [41, 77], joint rotations [2, 25, 66] or latent variables [6]) and downstream tasks. To generate plausible human-scene interactions, SceneDiffuser [19] condition the motion/pose diffusion model on 3D scenes and further guide the model with physics-based objectives at inference time. In a similar fashion, PhysDiff [74] incorporates a physics-based motion projection module in the inference loop to resolve typical flaws in motion synthesis such as foot-ground penetrations. While these methods all target human action synthesis, we take a different perspective and leverage Diffusion Models for a *perception* task — as a multi-proposal formulation to address the ambiguity in the egocentric pose estimation.

### 3. Preliminaries

**SMPL body model** [39]. We model the human mesh with the parametric SMPL body model. SMPL parametrizes a body as a function \( M_b(\gamma, \beta, \theta) \) of the global translation \( \gamma \in \mathbb{R}^3 \), body shape \( \beta \in \mathbb{R}^{10} \), and full body pose \( \theta \in \mathbb{R}^{24 \times 3} \) of 23 body joints, plus the global orientation, returning a mesh \( M_b \) with 6890 vertices.

**Conditional Diffusion Model.** We use DDPM [17] as our formulation of pose diffusion model. DDPM learns a distribution of body poses \( \theta \) through a forward diffusion process and an inverse denoising (sampling) process. The forward diffusion process is a Markov chain of added Gaussian noise over \( t \in \{1, \ldots, T\} \) steps:

\[
q(\theta_t|\theta_{t-1}) = N(\theta_t; \sqrt{\alpha_t} \theta_{t-1}, (1 - \alpha_t)I),\]

where the variance \( 1 - \alpha_t \in (0, 1) \) increases with \( t \) according to a pre-defined schedule, and \( T \) is the total number of noise corruption steps.

At the core of the inverse process is a denoiser neural network \( D(\cdot) \), which is trained to remove the added Gaussian noise based on condition signal \( c \) at each step \( t \). Following the design of recent motion diffusion models [63], the denoiser predicts the clean signal itself: \( \hat{\theta}_t = D(\theta_t, t, c) \).

The Gaussian forward diffusion process has a special property that allows directly sampling \( \theta_t \) from \( \theta_0 \):

\[
\theta_t = \sqrt{\alpha_t} \theta_0 + \sqrt{1 - \alpha_t} \epsilon, \quad \text{where } \epsilon \sim N(0, I). \]  (2)

With this, training the denoiser amounts to sampling a random \( t \in \{0, \ldots, T - 1\} \), adding noise according to Eq. (2), and optimizing the *simple objective* [17]:

\[
\mathcal{L}_{\text{simple}} = E_{\theta_0 \sim q(\theta_0|c), t \sim [1, T]} \left[ ||\theta_0 - D(\theta_0, t, c)||^2 \right]. \]  (3)

In this paper we use the subscript \( t \) to denote the diffusion timestep and superscript \( j \) for the body joint index.

### 4. Method

We introduce a novel scene-conditioned diffusion method to recover the human mesh in 3D environments from the egocentric image. Fig. 2 shows the overview of our model. Given an egocentric image \( I \) with a truncated body, the camera intrinsics \( K = (f, c_x, c_y) \), and the 3D scene point cloud \( S \in \mathbb{R}^{N \times 3} \) with \( N \) points, our goal is to learn the conditional distribution of body poses \( p(\theta|I, S, K) \), with the objective of generating humans that naturally interact with the scene while being aligned with the image observation. The body translation \( \gamma \) and shape \( \beta \) are learnt in a deterministic way following [32]. Here \( f \) and \( (c_x, c_y) \) denote the focal length and the camera center, respectively.

The diffusion denoising process (Sec. 3) is modeled by a visibility-aware graph convolution network (GCN), which achieves per-body-part control with the per-joint conditioning on the input image, scene geometry and joint visibility (Sec. 4.1). The diffusion model is trained in a classifier-free manner [18], enabling flexible sampling from both models.
that include or exclude image conditions to enhance diversity and expressiveness for truncated body parts. Guided by the gradients of a physics-based human-scene collision score, we further resolve human-scene inter-penetrations during the diffusion sampling (Sec. 4.2). The training objectives are explained in Sec. 4.3.

4.1. Scene-conditioned Pose Diffusion Model

3D scene conditioning. To model fine-grained human-scene interactions, only a small region of the scene surrounding the human is relevant. An accurate global body translation provides better reasoning about local scene geometries, thus crucial for accurate and scene-plausible local body pose modelling. To this end, we propose a body translation estimator (see Fig. 2) to estimate the translation $\hat{\gamma}$ by extracting features from $\mathcal{I}$ and the full 3D scene point cloud $\mathcal{S}$. The global scene features alleviate the scale ambiguity and localize the body in the 3D environment accurately (please refer to Supp. Mat. for more details). Given the estimated body translation $\hat{\gamma}$, scene vertices in a $2 \times 2m$ region around the human are selected and considered as the scene conditioning for the diffusion denoising process. Furthermore, the selected scene vertices are translated by $\hat{\gamma}$ such that the scene vertex origin is around the body pelvis. The translated scene point cloud is encoded by a scene encoder $E_S$ into a localized human-centric scene feature $E_S(S, \hat{\gamma})$ for modelling detailed interactions between the body and the 3D scene.

Per-joint conditioning. A ResNet50 backbone [16] $E_T$ is used to encode the cropped input image $I$ into a latent feature $E_T(I) \in \mathbb{R}^{2048}$ as the image conditioning. However, re-encoding the full body pose directly from the global image feature as in previous works [24, 29, 32] loses the explicit joint visibility information. With the body truncated in the image, different properties are desired for different body parts: the model should achieve pose accuracy for observed body joints, and pose expressiveness for truncated body parts. To achieve the per-body-part control, the 2D joint visibility mask $\mathcal{V} = (v^0, ..., v^j, ..., v^{J-1}) \in \mathbb{R}^J$ is extracted from the image via OpenPose [5].

Apart from above conditions, the bounding box information $\mathcal{B}$ is additionally fed to the network inspired by [36], serving as a global-location-aware feature to better model global information:

$$\mathcal{B} = (b_x, b_y, b) / f, \quad (4)$$

where $b_x$, $b_y$, $b$ denote the bounding box center coordinates and bounding box size in the full image, respectively.

The full condition $c^j$ for $j$-th joint is formulated as the concatenation of: the image feature (masked out by $v^j$ if the joint is not visible), 3D scene feature, estimated body translation, bounding box feature, camera intrinsics and the diffusion timestep $t$:

$$c^j = (E_T(I) \cdot v^j, E_S(S, \hat{\gamma}), \hat{\gamma}, \mathcal{B}, K, t). \quad (5)$$

By conditioning on a per-joint basis, the model achieves flexible control over each body part according to the joint visibility. This allows for both accurate pose estimation of visible body parts, and flexibility for invisible body parts.

Diffusion denoiser architecture. At diffusion timestep $t$, for the noisy body pose parameters $\theta_t$, the diffusion de-
noiser \( D \) predicts the clean pose parameter \( \hat{\theta}_0 \) conditioned on \( c \). Previous works either use a single MLP for the full body pose [24, 32, 36, 68] with redundant dependencies between irrelevant joints, or employ a separate MLP for each joint [27] which fully ignores joint dependencies. Instead, we adopt a graph convolution network (GCN) [26, 84] as the denoising function \( D \). Given the predefined human kinematic tree, GCNs models inter-joint dependencies by treating the human skeleton as a graph, where each node corresponds to a body joint. Specifically, for each joint \( j \), the noisy pose parameters \( \theta'_j \) is embedded and concatenated with the per-joint conditioning \( c'_j \) as the feature for node \( j \) in the GCN (see Fig. 2). We adopt the modulated GCN [84] for the implementation. Different from vanilla GCN [26], the modulated GCN learns node-specific modulation vectors to disentangle joint feature transformations, such that different relations can be learnt between joints. It also modulates the affinity matrix with learnable parameters to model additional edges beyond the human skeleton [84]. Please refer to Supp. Mat. for implementation details.

**Shape estimator.** As the body shape \( \beta \) is a global feature and more robust to body truncations, an MLP branch (the body shape estimator in Fig. 2) predicts \( \beta \) from the conditions \((E_T(\mathcal{I}), E_S(S, \gamma), \beta, K)\) in a deterministic way.

**4.2. Scene-guided Sampling**

We use the DDPM [17] sampler at inference time. At each sampling step \( t \), the denoiser predicts \( \theta_0 = D(\theta_t, t, c) \), which is noised back to \( \theta_{t-1} \) by sampling from the Gaussian distribution:

\[
\theta_{t-1} \sim \mathcal{N}(\mu_t(\theta_t, \theta_0), \Sigma_t),
\]

where \( \mu_t(\theta_t, \theta_0) \) is a linear combination of \( \theta_t \) and \( \theta_0 \), and \( \Sigma_t \) is a scheduled Gaussian distribution as in [17]. The sampling is iteratively repeated from \( t = T - 1 \) until \( t = 0 \).

**Classifier-free guidance.** The diffusion model \( D \) is trained with classifier-free guidance [18] by randomly masking out the image feature \( E_T(\mathcal{I}) \) for all joints with a probability of 5% during training, such that the model also learns the pose distribution independent from the image, enabling flexible sampling. By combining the poses of visible joints sampled from the full-conditioned model and the poses of invisible joints sampled from the model excluding image conditions:

\[
\hat{\theta}_0 = D(\theta_t, t, c) \odot V + D(\theta_t, t, c_{\phi}) \odot (1 - V),
\]

we achieve better per-body-part control, with enhanced diversity for invisible body parts while maintaining the accuracy of visible body parts (see Sec. 5.5 for the ablation study). Here \( \odot \) denotes multiplying the joint visibility mask with the pose of each joint, and \( c_{\phi} \) is the condition excluding image features.

**Collision score guided sampling.** The proposed scene encoding in Sec. 4.1 implicitly incorporates the 3D environment constraint into the diffusion conditioning. However, the network cannot fully resolve the human-scene collisions in a fine-grained way. Another way to incorporate conditions into the diffusion model is the score-function based guidance. As shown in previous works [9, 57, 60], a pretrained classifier can be adopted to provide class information by adding the gradients of the classifier at each sampling step to the diffusion model, to guide the sampling process towards the class label. Inspired by this, we propose a collision score guided sampling strategy to further resolve implausible human-scene inter-penetrations, by leveraging implicit body representations [43, 44]. We use COAP [43] to model neural articulated occupancy of human bodies as the zero-level set \( f_\theta(q|\mathcal{G}) = 0 \), where \( q \) is the input query point and \( \mathcal{G}(\theta, \beta) \) is the input bone transformations. By querying each scene vertex near the body to see if it is inside of the human volume, the collision score is defined as:

\[
J(\theta) = \mathbb{1}_{\{ \mathcal{S} \}} \sum_{q \in \mathcal{S}} \sigma(f_\theta(q|\mathcal{G})) I_{f_\theta(q|\mathcal{G}) > 0},
\]

where \( \sigma(\cdot) \) stands for the sigmoid function. The gradient \( \nabla J(\theta) \) efficiently guides the diffusion sampling process to further alleviate the human-scene collisions (see Sec. 5.5 for the ablation study) by modifying Eq. (6) to:

\[
\theta_{t-1} \sim \mathcal{N}(\mu_t(\theta_t, \theta_0) + a\Sigma_t \nabla J(\theta_t), \Sigma_t),
\]

where the guidance is modulated by \( \Sigma_t \) and a scale factor \( a \).

**4.3. Training Objectives**

The diffusion model is trained with \( L_{\text{simple}} \) in Eq. (3) for body pose \( \hat{\theta}_0 \), together with the the 3D joint loss \( L_{\text{3D}} \), the 2D keypoint re-projection loss \( L_{\text{2D}} \), calculated in the full image frame, and the shape loss \( L_\beta \):

\[
\begin{align*}
L_{\text{3D}} & = \|FK(\hat{\theta}_0, \beta) - FK(\theta_0, \beta)\|^2, \\
L_{\text{2D}} & = \|\Pi_K(J_{3D} + \gamma) - J_{\text{est}}\|^2, \\
L_\beta & = \|\beta - \hat{\beta}\|^2.
\end{align*}
\]

Here \( FK(\cdot) \) denotes the SMPL joint regressor to obtain 3D joint coordinates. \( J_{\text{est}} \) is the 2D body keypoint detections, and \( \Pi_K \) denotes the 3D to 2D projection in the full image with camera intrinsics \( K \). The overall loss is defined as:

\[
L = L_{\text{simple}} + \lambda_{3D}L_{\text{3D}} + \lambda_{2D}L_{\text{2D}} + \lambda_\beta L_\beta + \lambda_{\text{coll}} L_{\text{coll}} + \lambda_{\text{orth}} L_{\text{orth}},
\]

where \( L_{\text{orth}} \) is a regularizer to force the generated 6D pose rotation representation to be orthonormal, \( L_{\text{coll}} \) is the scene collision loss the same as defined in Eq. (8), and \( \lambda_s \) are the corresponding weight factors for each loss term.
5. Experiments

5.1. Dataset

Most of existing datasets for 3D human pose evaluation are either captured from the third-person view with the full body visible [15, 20, 42, 67], or do not contain the 3D environment [47]. The most relevant dataset is EgoBody [78], a large-scale egocentric dataset capturing 3D human motions during social interactions in 3D scenes. Each sequence involves two people interacting with each other in 3D environments, and one subject wears a head-mounted camera to capture the egocentric images containing the other interactee. Due to the close social interaction distance, the captured egocentric images typically exhibit strong body truncations for the interactee. We use the official train / test splits for training and evaluation, which include 90,120 / 62,140 frames, respectively.

5.2. Evaluation Metrics

Strong body truncations pose inherent ambiguity for invisible body parts, while the visible body parts are relatively deterministic given the image observations, thus evaluating the accuracy of full body joints together as in [3, 32] is not suitable for such tasks. We propose to evaluate the proposed method from the following aspects.

Accuracy. We employ Mean Per-Joint Position Error (MPJPE) and Vertex-to-Vertex (V2V) errors in mm to evaluate the accuracy of estimated body pose and shape. For MPJPE, we report different variations: G-MPJPE as the joint error in the global coordinate; MPJPE as the pelvis-aligned joint errors, and PA-MPJPE as the joint errors after Procrustes Alignment [12]. We report mean MPJPE and V2V metrics over n samples for visible joints. Nevertheless, such evaluation protocol assumes a single correct ground truth prediction, which does not hold for invisible joints. Instead of the mean error, we report min-of-n MPJPE for invisible body parts following [3]. To be specific, among the generated n samples for each input image, we select the one with the minimum MPJPE for all evaluated methods.

Physical plausibility. We evaluate the collision and contact scores between the body mesh and the 3D scene. As the clean scene signed distance field (SDF) is not available in EgoBody, we define the collision score as the ratio of the number of scene vertices inside of the body mesh to the number of all scene vertices. Note that the absolute scale of the collision score may vary with different size of the 3D scene, but within the same scene surface, the relative comparison between different methods indicates different collision levels. We use the cropped scene as described in Sec. 4.1 for evaluation (with a fixed size of 20,000 vertices). Following [19], the contact score is defined as 1 if the body contacts with the scene in a distance within a threshold (2 cm), otherwise 0. We report the mean collision score and contact score over all test samples. Note that a model with better physical plausibility should achieve both a lower collision score and a higher contact score.

Diversity. Diversity is evaluated for invisible body parts to demonstrate the expressiveness of the methods, by calculating the standard deviation (std) and Average Pairwise Distance (APD) of invisible body joints [19]. A higher diversity is desired, yet there is usually a trade-off between diversity and accuracy/plausibility. An ideal model should predict diverse poses for invisible body parts, with accurate visible body poses and plausible human-scene interactions.

5.3. Baselines and Our Method

As there is no existing work for our task, we compare our method with the following baselines: (1) ProHMR [32], the state-of-the-art method for probabilistic human mesh recovery from a single image based on conditional normalizing flows; (2) ProHMR-scene, in which we enhance the ProHMR framework by incorporating the scene geometry S with a global scene feature encoder. The ProHMR framework is trained with 2D and 3D supervisions for the sample drawn from z = 0 (referred as the ‘mode’), such that the ‘mode’ can be used as a predictive model (see [32] for details). Different training weights for the samples drawn from z ̸= 0 yields a trade-off between the accuracy and diversity for the generated multiple hypotheses. Here z denotes the latent space of the normalizing flow. To demonstrate this trade-off, we train the ProHMR-scene baseline with different training configurations. For samples drawn from z ̸= 0, we supervise with: 1) only 2D re-projection loss (the same as in [32]), denoted as -orig; 2) 3D supervision with small weights, denoted as -weak-3D; 3) 3D supervision with large weights, denoted as -strong-3D. Here the 3D supervision refers to the loss on 3D body joints and SMPL parameters. Please refer to Supp. Mat. for details of the baseline models.

5.4. Results

Tab. 1 shows the quantitative results. Compared to the baselines, our method achieves the best G-MPJPE, MPJPE and V2V metrics for visible joints, demonstrating accurate estimations of the global translation, local body pose for visible joints and body shape. For truncated body joints, our method obtains the best min-of-n MPJPE, indicating that our generated results better covers the ground truth distribution for truncated body parts than baselines, thanks to the efficient scene conditioning. Our method also outperforms the baselines with the higher contact ratios and significantly better collision scores. In particular, our collision score decreased 45% compared with ProHMR, and 37% compared with the best ProHMR-scene model. This indicates that the proposed method can effectively leverage the scene conditions and resolve body-scene inter-penetrations.
Table 1: **Evaluation for accuracy, physical plausibility and diversity on EgoBody.** Here -vis and -invis denote visible and invisible joints, respectively. ‘coll’ and ‘contact’ refer to the collision and contact scores, respectively. The percentage for the collision score denotes how much improvement the corresponding method has achieved compared with the ProHMR baseline. The best results are in boldface.

<table>
<thead>
<tr>
<th>n</th>
<th>Method</th>
<th>G-MPJPE ↓</th>
<th>MPIPE ↓</th>
<th>PA-MPJPE ↓</th>
<th>V2V ↓</th>
<th>min-of-n MPJPE ↓</th>
<th>coll ↓</th>
<th>contact ↑</th>
<th>std ↑</th>
<th>APD ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>ProHMR [32]</td>
<td>181.33</td>
<td>72.92</td>
<td>46.57</td>
<td>90.88</td>
<td>128.56</td>
<td>0.00346 (-)</td>
<td>0.94</td>
<td>26.53</td>
<td>32.95</td>
</tr>
<tr>
<td></td>
<td>ProHMR-scene-orig</td>
<td>133.72</td>
<td>73.13</td>
<td>48.87</td>
<td>91.59</td>
<td>117.88</td>
<td>0.00305 (12%)</td>
<td>0.96</td>
<td>35.42</td>
<td>44.02</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>128.62</td>
<td>65.70</td>
<td>47.17</td>
<td>82.93</td>
<td>107.77</td>
<td>0.00191 (45%)</td>
<td>0.99</td>
<td>29.59</td>
<td>35.15</td>
</tr>
<tr>
<td>10</td>
<td>ProHMR [32]</td>
<td>181.70</td>
<td>73.65</td>
<td>47.06</td>
<td>91.84</td>
<td>121.67</td>
<td>0.00346 (-)</td>
<td>0.94</td>
<td>29.26</td>
<td>35.06</td>
</tr>
<tr>
<td></td>
<td>ProHMR-scene-orig</td>
<td>134.57</td>
<td>74.42</td>
<td>49.63</td>
<td>93.21</td>
<td>110.64</td>
<td>0.00305 (12%)</td>
<td>0.96</td>
<td>39.07</td>
<td>46.86</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>128.60</td>
<td>65.69</td>
<td>47.20</td>
<td>82.94</td>
<td>100.75</td>
<td>0.00191 (45%)</td>
<td>0.99</td>
<td>29.59</td>
<td>35.15</td>
</tr>
<tr>
<td>20</td>
<td>ProHMR [32]</td>
<td>181.87</td>
<td>73.99</td>
<td>47.29</td>
<td>92.28</td>
<td>115.48</td>
<td>0.00346 (-)</td>
<td>0.93</td>
<td>30.55</td>
<td>36.14</td>
</tr>
<tr>
<td></td>
<td>ProHMR-scene-orig</td>
<td>134.97</td>
<td>75.03</td>
<td>49.98</td>
<td>93.98</td>
<td>103.81</td>
<td>0.00306 (12%)</td>
<td>0.96</td>
<td>40.82</td>
<td>48.32</td>
</tr>
<tr>
<td></td>
<td>ProHMR-scene-weak-3D</td>
<td>131.51</td>
<td>67.81</td>
<td>46.94</td>
<td>85.05</td>
<td>111.68</td>
<td>0.00313 (10%)</td>
<td>0.96</td>
<td>21.79</td>
<td>25.67</td>
</tr>
<tr>
<td></td>
<td>ProHMR-scene-strong-3D</td>
<td>134.40</td>
<td>66.42</td>
<td>45.96</td>
<td>82.99</td>
<td>123.34</td>
<td>0.00316 (9%)</td>
<td>0.97</td>
<td>13.01</td>
<td>15.23</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>128.60</td>
<td>65.69</td>
<td>47.19</td>
<td>82.94</td>
<td>94.96</td>
<td>0.00191 (45%)</td>
<td>0.99</td>
<td>30.23</td>
<td>35.14</td>
</tr>
</tbody>
</table>

Figure 3: **Qualitative results on EgoBody dataset.** The left part shows the comparison between the baseline methods and our proposed method. The right part demonstrates the effectiveness of our proposed scene collision score guidance $J_\theta$ to further resolve human-scene inter-penetrations. Red circles point to human-scene collisions.
Table 2: Ablation study. \( \mathcal{J}_\theta \) and CF denote the collision score guidance and classifier-free guidance during the diffusion sampling (Sec 4.2). \( L_{\text{coll}} \) denotes the collision loss for training (Eq. (11)). All results are reported for \( n = 20 \).

<table>
<thead>
<tr>
<th>Method</th>
<th>MPJPE ↓</th>
<th>min-of-n MPJPE ↓</th>
<th>coll ↓</th>
<th>contact ↑</th>
<th>std ↑</th>
<th>APD ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>65.69</td>
<td>94.96</td>
<td>0.00191</td>
<td>0.99</td>
<td>30.23</td>
<td>35.14</td>
</tr>
<tr>
<td>Ours w/o ( \mathcal{J}_\theta )</td>
<td>65.09</td>
<td>97.20</td>
<td>0.00225</td>
<td>0.99</td>
<td>21.53</td>
<td>25.32</td>
</tr>
<tr>
<td>Ours w/o ( \mathcal{J}_\theta ) w/o CF</td>
<td>64.58</td>
<td>98.82</td>
<td>0.00229</td>
<td>0.99</td>
<td>15.75</td>
<td>19.56</td>
</tr>
<tr>
<td>Ours w/o ( \mathcal{J}<em>\theta ) w/o CF w/o ( L</em>{\text{coll}} )</td>
<td>64.91</td>
<td>96.26</td>
<td>0.00302</td>
<td>0.99</td>
<td>22.94</td>
<td>21.79</td>
</tr>
</tbody>
</table>

Figure 4: Multiple samples. Given the same input, our method can generate diverse poses for invisible body parts.

while achieving better contact relations between the body and the 3D environment. Fig. 3 (left) shows the qualitative results and the comparison with the baseline method.

From the results among different ProHMR-scene baselines (\( n = 20 \)) in Tab. 1, we can clearly see a trade-off between the accuracy and diversity: stronger 3D supervision (ProHMR-scene-\textit{strong-3D}) has better accuracy for visible body parts (MPJPE-\textit{vis}), yet lower diversity (std-\textit{invis} and APD-\textit{invis}) for invisible joints. The reason behind this is that the ProHMR framework, like many other existing works, condition the generative process of full body pose only on the global image feature. In contrast, our method leverages the joint visibility and the GCN architecture, achieving the best MPJPE for visible joints, diverse poses for invisible joints, and natural human-scene interactions (see Fig. 4), while the baseline with a similar diversity level (ProHMR) and the baseline with the highest diversity (ProHMR-scene-\textit{orig}) exhibit much higher MPJPE errors and more severe human-scene collisions. There is a very slight increase for PA-MPJPE of our model compared to the ProHMR-scene-\textit{strong/weak-3D} baselines, while we obtain a much higher diversity for invisible body joints. This reveals that our method achieves flexible per-body-part control to balance between the accuracy and diversity for different body parts.

To validate the reconstruction accuracy, the full-body accuracy of the proposed method is also evaluated and compared against existing deterministic baselines on EgoBody dataset. The results show that our method even achieves higher full-body accuracy than deterministic methods, which demonstrates that the proposed model achieves diversity without sacrificing the reconstruction accuracy. In addition, the proposed method is also evaluated on PROX [15] dataset, a monocular RGB-based third-person view dataset with human-scene interactions, and our method outperforms the baseline methods in terms of pose accuracy, human-scene interaction plausibility and pose diversity. Please refer to Supp. Mat. for detailed results of evaluations on PROX and comparison with deterministic methods on EgoBody.

5.5. Ablation Study

To investigate the influences of different training and sampling schemes for the diffusion model, we perform an ablation study and the results are shown in Tab. 2. The scene collision score efficiently resolves the human-scene interpenetrations from two different perspectives: via the collision loss \( L_{\text{coll}} \) during training and via the explicit collision score guidance \( \mathcal{J}_\theta \) during the diffusion sampling (Eq. (9)). Fig. 3 (right) shows that the collision score guidance \( \mathcal{J}_\theta \) improves the human-scene interaction relationships. The classifier-free guidance during the diffusion sampling in Eq. (7) significantly improves the sample diversity, by leveraging the invisible joints sampled from the model trained excluding the image conditions. It is worth mentioning that the collision score guidance \( \mathcal{J}_\theta \) also contributes to the sample diversity. This is not surprising as the gradients of \( \mathcal{J}_\theta \) can potentially optimize the body pose to different directions given different initial body poses.

Compared with the trade-off in the ProHMR-based methods where the accuracy sacrifices a lot for diversity (and vice versa), the classifier-free guidance and the collision score guidance in our model only bring marginal changes for the MPJPE of visible body parts, in the trade-off for significantly enhanced physical plausibility, sample diversity and better \textit{min-of-n} MPJPE for unobserved joints.
6. Conclusion

This paper introduces a novel scene-conditioned probabilistic approach for human mesh recovery from the egocentric image given the 3D scene geometry. The proposed diffusion model efficiently leverages the scene conditions and joint visibility to achieve per-body-part control and model human-scene interactions. The classifier-free guidance and the collision score guidance enables flexible sampling process with further enhanced diversity and physical plausibility. We demonstrate that our model generates human bodies in realistic interactions with the 3D environment, with diverse poses for invisible body parts while conforming to the image observation. Nevertheless, the current method also has limitations. For example, currently only a single frame is considered as the input. A promising future direction is to further exploit temporal information to reconstruct temporally consistent human motions from the egocentric view.

Acknowledgements. This work was supported by Microsoft Mixed Reality & AI Zurich Lab PhD scholarship. Qianli Ma is partially funded by the Max Planck ETH Center for Learning Systems. We sincerely thank Korrawe Karunratanakul, Marko Mihajlovic and Shaofei Wang for the fruitful discussions.

References


Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations, 2021. 2, 3, 5


Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3D human body shape and pose prediction. 2017. 2


