Generative Approach for Probabilistic Human Mesh Recovery using Diffusion Models

Hanbyel Cho
KAIST, South Korea
tlrl4658@kaist.ac.kr

Junmo Kim
KAIST, South Korea
junmo.kim@kaist.ac.kr

Abstract

This work focuses on the problem of reconstructing a 3D human body mesh from a given 2D image. Despite the inherent ambiguity of the task of human mesh recovery, most existing works have adopted a method of regressing a single output. In contrast, we propose a generative approach framework, called "Diffusion-based Human Mesh Recovery (Diff-HMR)" that takes advantage of the denoising diffusion process to account for multiple plausible outcomes. During the training phase, the SMPL parameters are diffused from ground-truth parameters to random distribution, and Diff-HMR learns the reverse process of this diffusion. In the inference phase, the model progressively refines the given random SMPL parameters into the corresponding parameters that align with the input image. Diff-HMR, being a generative approach, is capable of generating diverse results for the same input image as the input noise varies. We conduct validation experiments, and the results demonstrate that the proposed framework effectively models the inherent ambiguity of the task of human mesh recovery in a probabilistic manner. Code is available at https://github.com/hanbyel0105/Diff-HMR.

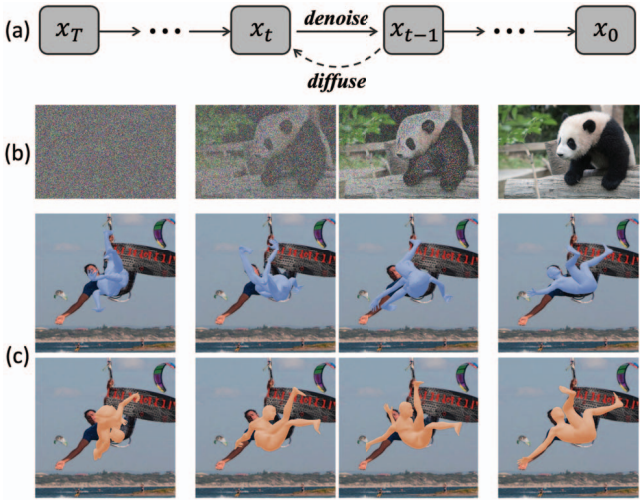

Figure 1: **Diffusion model for human mesh recovery.** (a) A diffusion model defined by a Markov chain, consisting of a forward process ("*diffuse*") and a reverse process ("*denoise*"). (b) Diffusion model for the task of image generation. (c) We propose Diff-HMR, which formulates human mesh recovery as a denoising diffusion process, enabling the generation of multiple plausible meshes based on variations in a random noise input of SMPL.

1. Introduction

Human Mesh Recovery (HMR) is a task of regressing three-dimensional human body model parameters, such as SMPL [24], from a given 2D image. Along with joint-based methods [29, 6, 23], HMR is a fundamental task in computer vision and holds great significance in applications like computer graphics and VR/AR. Despite significant advances in HMR, it remains a challenging problem due to the inherent ambiguity caused by the loss of depth information and occlusions in 2D images. Most existing approaches [14, 15, 4, 38, 5, 36] have relied on single-output regression, limiting their ability to explain uncertainty in the HMR process. Consequently, these methods often fail to reconstruct diverse and plausible 3D human body meshes that accurately represent the actual posture of the human.

To overcome these limitations, we propose a generative approach framework called "Diffusion-based Human Mesh Recovery (*Diff-HMR*)". Unlike traditional *perception*-based approaches [14, 15, 38, 20] that regress a single output from the given image, our proposed framework draws inspiration from the denoising diffusion process [10] and outputs SMPL parameters in a *generative* way from random noise input. By exploring multiple plausible outcomes during the reconstruction process based on the input noise varying, Diff-HMR can probabilistically model the inherent ambiguity in human mesh recovery.

Specifically, during the training phase, the SMPL parameters are diffused from ground-truth parameters to random distribution, and Diff-HMR learns the reverse of this diffusion process. In the inference phase, the model progressively refines the initial random SMPL parameters to the

parameters corresponding to the given input image. As a generative approach, Diff-HMR can produce diverse plausible human body meshes for the same input image as the input noise varies as shown in Figure 1 (c).

Additionally, to overcome the limitation of the diffusion model, which is sensitive to signal-to-noise ratio (SNR), we adopt 6D rotation representations [39] instead of Axis-angle representations as joint angle representations for SMPL. We conduct experiments on various datasets [37, 22] and verify the efficacy of Diff-HMR. Our contributions are as follows:

- We propose a novel generative HMR framework, Diff-HMR, which utilizes the denoising diffusion process to generate multiple 3D human meshes from a given image, effectively modeling the inherent ambiguity of HMR.
- By adopting 6D rotation representations for SMPL joint angles, Diff-HMR can maintain a suitable signal-to-noise ratio (SNR), enabling stable training of diffusion models.
- The experimental results show that Diff-HMR effectively models the inherent ambiguity of HMR, generating multiple plausible outcomes from a given image.

2. Related Work

Monocular 3D Human Mesh Recovery can be categorized into optimization-based, regression-based, and hybrid approaches. Optimization-based approaches [3, 17] fit SMPL parameters to minimize errors between reconstructed meshes and 3D/2D evidence like keypoints or silhouettes. Regression-based approaches [14, 27, 4, 28, 5, 38] leverage deep neural networks to improve inference speed and reconstruction quality by directly inferring SMPL from input images. In recent studies, hybrid approaches [15, 13] that combine both optimization and regression-based methods have emerged. This approach offers a more accurate pseudo-ground-truth SMPL for 2D images. Despite these advancements, such single-output regression methods still limit their ability to explain the uncertainty in the HMR. To overcome this, we propose Diff-HMR that produces multiple plausible outcomes from random noise of SMPL, which provides comprehensive modeling in probabilistic manner.

Probabilistic Inference for Human Body Mesh. Estimating a person's 3D pose from a single image remains challenging due to inherent ambiguities such as occlusions and depth ambiguities. To address this problem, Jahangiri *et al*. [12] adopted a compositional model and utilized anatomical evidence to infer multiple hypotheses of 3D poses. In addition, Li *et al*. [18] used a Mixture Density Network, considering the Gaussian kernel centroids as individual hypotheses. Sharma *et al*. [33] utilized CVAE *et al*. [35] to generate hypotheses and generated the final output by using the weighted average of hypotheses based on joint-ordinal relations. In the field of human mesh recovery, following this trend, recent works such as Biggs *et al*. [2] have utilized Normalizing flow to output N pre-defined predictions, and Kolotouros *et al*. [16] directly outputs likelihood values. However, despite the excellent capabilities of a denoising diffusion process for probability distribution modeling, it has not yet been applied to probabilistic HMR.

Denoising Diffusion Probabilistic Models. Denoising diffusion probabilistic models (DDPMs) [10], also known as diffusion models, are generative models composed of two stages: a forward process ("*diffuse*") and a reverse process ("*denoise*"). In the forward process, the input data is diffused into a random distribution through multiple steps by adding Gaussian noise. In the reverse process, the model learns to reverse the forward process, denoising the noised data, and consequently learns the distribution of data. DDPMs have shown excellent performance in density estimation and have recently been applied to tasks such as image/text generation [34, 25, 7, 30, 19, 21], despite the computational cost of the sampling process. In the context of joint-based 3D human pose estimation, DDPMs have been applied to tackle ambiguities [8, 32]. However, there is no method for human mesh recovery yet; therefore, in this work, we propose a probabilistic HMR method that considers multiple outputs by applying DDPMs for the first time.

3. Method

The overall framework of our Diff-HMR is depicted in Figure 2. In this section, we first recapitulate DDPMs [10] and then explain the model architecture of Diff-HMR.

3.1. Preliminary: DDPMs

Denoising diffusion probabilistic models (DDPMs) are composed of two stage, each defined as a Markov chain: the *forward process* and the *reverse process*.

Forward Process. Let $\boldsymbol{x}^{(0)}$ be the observed original data, and it follows an unknown distribution $q(\boldsymbol{x}^{(0)})$. The goal of the forward process, referred to as the diffusion process, is to transform $\boldsymbol{x}^{(0)}$ into a Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ by gradually adding Gaussian noise over pre-defined T steps.

At each step t, the data is progressively disturbed by adding noise according to the following equation:

$$q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(t-1)}) \sim \mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{x}^{(t-1)}, \beta_t\mathbf{I}) \tag{1}$$

where $t = 1, \cdots, T$ and β_t is pre-defined noise schedule; We can rewrite $\boldsymbol{x}^{(t)}$ as $\boldsymbol{x}^{(t)} = \sqrt{1-\beta_t}\boldsymbol{x}^{(t-1)} + \sqrt{\beta_t}\boldsymbol{\epsilon}_t$, where $\boldsymbol{\epsilon}_t$ follows a Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

For any given t, since $\boldsymbol{\epsilon}_t$ are *i.i.d.*, we can directly generate $\boldsymbol{x}^{(t)}$ from $\boldsymbol{x}^{(0)}$ in a closed form as follows:

$$q(\boldsymbol{x}^{(t)}|\boldsymbol{x}^{(0)}) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2}$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

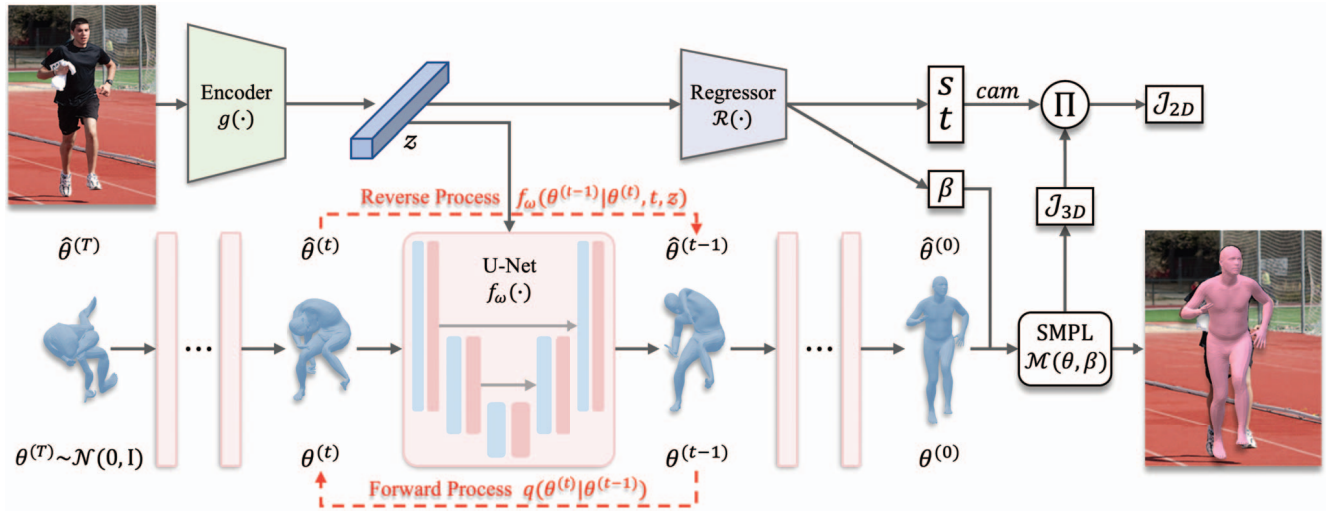

Figure 2: **Illustration of Diff-HMR architecture.** For a given image, Diff-HMR can predict multiple plausible human body meshes that a person in the image may take. To model the ambiguity of the posture taken by the person, Diff-HMR progressively generates the final human mesh $\hat{\boldsymbol{\theta}}^{(0)}$ from the noise SMPL pose parameters $\hat{\boldsymbol{\theta}}^{(T)}$ that follows $\mathcal{N}(\mathbf{0}, \mathbf{I})$. During the *reverse process*, the denoising network $f_{\boldsymbol{\omega}}(\cdot)$ denoises $\hat{\boldsymbol{\theta}}^{(T)}$ as the correct SMPL pose $\hat{\boldsymbol{\theta}}^{(0)}$, referring to the feature vector $\boldsymbol{z}$ extracted from the input image as $f_{\boldsymbol{\omega}}(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^{(t)}, t, \boldsymbol{z})$. In order to train $f_{\boldsymbol{\omega}}(\cdot)$, the *forward process* gradually perturbs the ground-truth SMPL pose $\boldsymbol{\theta}^{(0)}$ as $q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t-1)})$ to produce the noised SMPL pose $\boldsymbol{\theta}^{(t)}$ required for training.

By using Eq. 2, we can efficiently generate noised samples $\boldsymbol{x}^{(t)}$ during the training phase. In Diff-HMR, we employ ground-truth SMPL pose parameters $\boldsymbol{\theta}^{(0)}$ as $\boldsymbol{x}^{(0)}$ to get the noised SMPL pose $\boldsymbol{\theta}^{(t)}$ in the forward process.

Reverse Process. Let us define the joint distribution $p_{\boldsymbol{\omega}}(\boldsymbol{x}^{(0:T)})$, parameterized by the learnable parameters $\boldsymbol{\omega}$, which aims to mimic the distribution $q(\boldsymbol{x}^{(0)})$ as follows:

$$p_{\boldsymbol{\omega}}(\boldsymbol{x}^{(0:T)}) = p(\boldsymbol{x}^{(T)}) \prod_{t=1}^{T} p_{\boldsymbol{\omega}}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)}) \tag{3}$$

where $p_{\boldsymbol{\omega}}(\boldsymbol{x}^{(t-1)}|\boldsymbol{x}^{(t)}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\omega}}(\boldsymbol{x}^{(t)}, t), \boldsymbol{\Sigma}_{\boldsymbol{\omega}}(\boldsymbol{x}^{(t)}, t))$ and $p(\boldsymbol{x}^{(T)}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The goal of the reverse process is to find $\boldsymbol{\omega}$ that maximizes $p(\boldsymbol{x}^{(0)})$ when $\boldsymbol{x}^{(0)}$ follows $q(\boldsymbol{x}^{(0)})$.

To achieve this, we need to learn $\boldsymbol{\mu}_{\boldsymbol{\omega}}(\boldsymbol{x}^{(t)}, t)$ (*mean*) and $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}(\boldsymbol{x}^{(t)}, t)$ (*variance*). However, according to DDPMs, variance can be expressed as $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}(\boldsymbol{x}^{(t)}, t) = \beta_t \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\mathbf{I}$, which only depends on t; Thus we only need to learn mean.

Furthermore, since the mean can be reparameterized as $\boldsymbol{\mu}_{\boldsymbol{\omega}}(\boldsymbol{x}^{(t)}, t) = \frac{1}{\sqrt{\alpha_t}}(\boldsymbol{x}^{(t)} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\omega}}(\boldsymbol{x}^{(t)}, t))$, we can simply train a neural network $\boldsymbol{\epsilon}_{\boldsymbol{\omega}}(\boldsymbol{x}^{(t)}, t)$ parameterized by $\boldsymbol{\omega}$ to predict the noise $\boldsymbol{\epsilon}$ at each reverse step.

In Diff-HMR, the model extracts a feature vector $\boldsymbol{z}$ from the input image and predicts conditioned noise $\boldsymbol{\epsilon}$ on the $\boldsymbol{z}$ to generate the pose corresponding to the person of the input image. Therefore, the training objective for the reverse process in Diff-HMR is $\mathbb{E}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\omega}}(\boldsymbol{x}^{(t)}, t, \boldsymbol{z})\|^2]$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{x}^{(t)}$ is as the following equation:

$$\boldsymbol{x}^{(t)} = \sqrt{\bar{\alpha}_t}\boldsymbol{x}^{(0)} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}. \tag{4}$$

3.2. Diffusion-based Human Mesh Recovery

In this work, we propose a diffusion-based human mesh recovery (*Diff-HMR*) by incorporating denoising diffusion probabilistic models (DDPMs) [10] into the existing HMR framework to address the inherent ambiguity of the HMR.

Overall Framework. To tackle the inherent ambiguity of human poses in the given image, Diff-HMR gradually generates the correct human mesh $\hat{\boldsymbol{\theta}}^{(0)}$ corresponding to the image from noise SMPL pose parameters $\hat{\boldsymbol{\theta}}^{(T)}$ as in the reverse process in DDPMs. To this end, we first extract the image feature $\boldsymbol{z}$ from the given image $\mathbf{I}$ using the ResNet-50 [9]-based image encoder $g(\cdot)$ as $\boldsymbol{z} = g(\mathbf{I}) \in \mathbb{R}^{2048\times7\times7}$.

At each reverse step t, the denoising network $f_{\boldsymbol{\omega}}(\cdot)$ with 1D U-Net [31] architecture is conditioned on the time step t and image feature $\boldsymbol{z}$ to predict the noise $\hat{\boldsymbol{\epsilon}}_t$ necessary to estimate the previous pose $\hat{\boldsymbol{\theta}}^{(t-1)}$ from $\hat{\boldsymbol{\theta}}^{(t)}$. To train $f_{\boldsymbol{\omega}}(\cdot)$, the ground-truth noise $\boldsymbol{\epsilon}_t$ is generated from the ground-truth SMPL pose $\boldsymbol{\theta}^{(0)}$ through the forward process using Eq. 4. We define loss function for typical DDPMs training that predicts the noise $\hat{\boldsymbol{\epsilon}}_t$ as $\mathcal{L}_{\text{diff}} = \mathbb{E}[\|\boldsymbol{\epsilon}_t - f_{\boldsymbol{\omega}}(\boldsymbol{\theta}^{(t)}, t, \boldsymbol{z})\|^2]$ and use this for training of the denoising network $f_{\boldsymbol{\omega}}(\cdot)$.

Training Objective. In addition to using $\mathcal{L}_{\text{diff}}$, we also impose constraints on the final SMPL pose $\hat{\boldsymbol{\theta}}^{(0)}$ and the derived 2D ($\mathcal{J}_{2D}$) and 3D ($\mathcal{J}_{3D}$) joints as in conventional HMR methods [14, 5], which are represented as $\mathcal{L}_{\text{hmr}}$. However, since $f_{\boldsymbol{\omega}}(\cdot)$ predicts the noise at each time step t, we cannot directly access $\hat{\boldsymbol{\theta}}^{(0)}$. To overcome this, we deduce $\hat{\boldsymbol{\theta}}^{(0)}$ from the predicted noise $\hat{\boldsymbol{\epsilon}}_t$ by reparameterizing Eq. 4 as

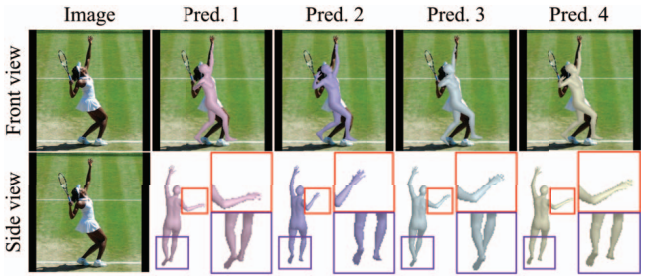

Figure 3: **Predicted meshes obtained by varying seeds.** Diff-HMR outputs a plausible human body mesh of different possibilities for each sampled seed $\hat{\boldsymbol{\theta}}^{(T)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓
SPIN [15] ICCV'19	96.9	59.2	116.4
Biggs *et al*. [2] NeurIPS'20	N/A	59.9	N/A
ProHMR [16] ICCV'21	N/A	59.8	N/A
Diff-HMR, $n = 1$	98.9	58.5	114.6
Diff-HMR, $n = 5$	96.3	57.0	111.8
Diff-HMR, $n = 10$	<u>95.5</u>	<u>56.5</u>	<u>110.9</u>
Diff-HMR, $n = 25$	**94.5**	**55.9**	**109.8**

Table 1: **Multiple hypotheses results on 3DPW.** Values are in mm. n denotes the number of inferences performed by varying the seed $\hat{\boldsymbol{\theta}}^{(T)}$. We report the error calculated based on the minimum error among n samples. Best in bold, second-best underlined.

$\hat{\boldsymbol{\theta}}^{(0)} = \frac{1}{\sqrt{\bar{\alpha}_t}}\hat{\boldsymbol{\theta}}^{(t)} - (\sqrt{\frac{1}{\bar{\alpha}_t} - 1}) f_{\boldsymbol{\omega}}(\hat{\boldsymbol{\theta}}^{(t)}, t, \boldsymbol{z})$. Thus, our overall loss function is $\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{hmr}}$.

Since the ambiguity of human mesh recovery is mainly related to pose, Diff-HMR predicts SMPL shape parameters $\boldsymbol{\beta}$ and weak-perspective camera parameters $\boldsymbol{\pi} \in [s(\textit{scale}), t(\textit{translation})]$ simply using MLP-based regressor $\mathcal{R}(\cdot)$ as $\{\boldsymbol{\beta}, \boldsymbol{\pi}\} = \mathcal{R}(\boldsymbol{z})$. From predicted results, we can get mesh vertices $M = \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{6890\times 3}$ where $\mathcal{M}$ is transform function and get 3D joints $\mathcal{J}_{\text{3D}} = WM$ and 2D joints $\mathcal{J}_{\text{2D}} = \boldsymbol{\Pi}(\mathcal{J}_{\text{3D}})$ where W and $\boldsymbol{\Pi}$ denote pre-trained linear regressor and projection function, respectively.

4. Experiments

4.1. Datasets and Evaluation Metrics

In order to train our model, we use H36M [11] and MPI-INF-3DHP [26] as 3D datasets and pseudo-ground-truth SMPL annotated COCO [22] and MPII [1] as 2D datasets. For quantitative evaluation, we use 3DPW [37] test split and use the Mean Per Joint Position Error (MPJPE), Procrustes-Aligned MPJPE (PA-MPJPE), and Per-Vertex Error (PVE) as our evaluation metrics. For qualitative evaluation, we use the sample image from the validation set of COCO [22].

4.2. Experimental Results

This section includes basic experiments to validate Diff-HMR. We show quantitative and qualitative results and explore the efficacy of using 6D rotation representations as joint angle representations for SMPL in diffusion models.

Method	3DPW [37]	MPI-INF-3DHP [26]
Axis-angle repr.	116.3	146.0
6D rotation repr. [39]	**64.5**	**92.5**

Table 2: **Performance comparison based on SMPL joint angle representations.** We report PA-MPJPE in mm. Diff-HMR shows better results when adopting 6D rotation representations as joint angle representations. Note that only the COCO [22] is used for training here. Best in bold.

Quantitative results. Table 1 shows quantitative results of Diff-HMR. When n is 1, Diff-HMR shows comparable performance to existing HMR methods [15, 2, 16]. Moreover, as n increases to 5, 10, and 25, the errors decrease significantly, indicating that Diff-HMR has strong representational power for modeling the inherent ambiguity of HMR.

Qualitative results. Figure 3 shows human mesh reconstruction results of Diff-HMR. As the input noise SMPL pose parameters $\hat{\boldsymbol{\theta}}^{(T)}$ vary along Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the model exhibits diverse inference outcomes for probabilistically ambiguous body parts.

Efficacy of using 6D rotation representations. For stable training of diffusion models [10], the ratio between the magnitude of the original data and the injected noise, known as the signal-to-noise ratio (SNR), is crucial. To incorporate diffusion models into HMR, we adopt 6D rotation representations [39] with the same scale as Gaussian distribution as joint angle representations of SMPL pose parameters. As shown in Table 2, we can confirm that it is more reasonable to adopt 6D rotation than Axis-angle representations.

5. Conclusion and Future Work

Conclusion. In this work, we focus on the task of human mesh recovery, which reconstructs 3D human body mesh from a given 2D observation. To probabilistically model the inherent ambiguity of the task, we propose a generative approach framework, called "*Diff-HMR*" that takes advantage of the denoising diffusion process to account for multiple plausible outcomes. To overcome the limitation of the diffusion model, which is sensitive to the signal-to-noise ratio (SNR), we adopt 6D rotation representations instead of Axis-angle representations as joint angle representations for SMPL. The evaluations on benchmark datasets demonstrate that the proposed framework successfully models the inherent ambiguity of the task of human mesh recovery.

Future Work. We will focus on improving the conditioning module of the 1D U-Net to better understand the spatial context of the input image. By enhancing the module, we can expect Diff-HMR to infer more plausible results, especially in occluded situations.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 4

[2] Benjamin Biggs, Sébastien Ehrhart, Hanbyul Joo, Benjamin Graham, Andrea Vedaldi, and David Novotny. 3D multibodies: Fitting sets of plausible 3D models to ambiguous image data. In *NeurIPS*, 2020. 2, 4

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 2

[4] Hanbyel Cho, Jaesung Ahn, Yooshin Cho, and Junmo Kim. Video inference for human mesh recovery with vision transformer. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6, 2023. 1, 2

[5] Hanbyel Cho, Yooshin Cho, Jaesung Ahn, and Junmo Kim. Implicit 3d human mesh recovery using consistency with pose and shape from unseen-view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21148–21158, June 2023. 1, 2, 3

[6] Hanbyel Cho, Yooshin Cho, Jaemyung Yu, and Junmo Kim. Camera distortion-aware 3d human pose estimation in video with optimization-based meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11169–11178, 2021. 1

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021. 2

[8] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13041–13051, June 2023. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 2, 3, 4

[11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 4

[12] Ehsan Jahangiri and Alan Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. pages 805–814, 10 2017. 2

[13] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *3DV*, 2020. 2

[14] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Regognition (CVPR)*, 2018. 1, 2, 3

[15] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1, 2, 4

[16] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 2, 4

[17] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6050–6059, 2017. 2

[18] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9887–9895, 2019. 2

[19] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomput.*, 479(C):47–59, mar 2022. 2

[20] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 1

[21] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-lm improves controllable text generation. *ArXiv*, abs/2205.14217, 2022. 2

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft coco: Common objects in context. In *ECCV*. European Conference on Computer Vision, September 2014. 2, 4

[23] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-ching Cheung, and Vijayan Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5064–5073, 2020. 1

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1

[25] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR), pages 11461–11471, June 2022. 2

[26] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 4

[27] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. Verona, Italy, 2018. 2

[28] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[29] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 1

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 2

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 3

[32] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. *arXiv preprint arXiv:2303.11579*, 2023. 2

[33] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[34] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. 2

[35] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 2

[36] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey, 2023. 1

[37] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2, 4

[38] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 1, 2

[39] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 4