MHPro: Multi-Hypothesis Probabilistic Modeling for Human Mesh Recovery

Haibiao Xuan[1], Jinsong Zhang[1], and Kun Li[1,*]

College of Intelligence and Computing, Tianjin University, Tianjin
{hbxuan, jinszhang, lik}@tju.edu.cn

Abstract. Recovering 3D human meshes from monocular images is an inherently ambiguous and challenging task due to depth ambiguity, joint occlusion and truncation. However, most recent works avoid modeling uncertainty, typically obtaining a single reconstruction for a given input. In contrast, this paper presents the ambiguity of reception reconstruction and considers the problem as an inverse problem for which multiple feasible solutions exist. Our method, **MHPro**, first constructs a probability distribution and obtains a set of feasible recovery results (*i.e.* multi-hypotheses), from monocular images. Intra-hypothesis refinement is then performed to achieve independent feature enhancement. Finally, the multi-hypothesis features are aggregated by inter-hypothesis communication to recover the final 3D human mesh. The effectiveness of our method is validated on two benchmark datasets, Human3.6M and 3DPW, where experimental results show that our method achieves state-of-the-art performance and recovers more accurate human meshes. Our results validate the importance of intra-hypothesis refinement and inter-hypothesis communication in probabilistic modeling and show optimal performance across a variety of settings. Our source code will be available at *http://cic.tju.edu.cn/faculty/likun/projects/MHPro*.

Keywords: Human Mesh Recovery · Monocular Images · Multi-Hypothesis · Probabilistic Modeling.

1 Introduction

3D human mesh recovery from a single color image is a widely-studied problem in computer vision, as well as a vision task with a wide range of application scenarios, such as action recognition [1], human-computer interaction [2] and AR/VR [3]. However, human mesh recovery from a single image remains a challenging task and an inherently ill-posed problem due to depth ambiguity, joint occlusion and truncation.

Given a single image, recent literature for 3D human mesh recovery typically returns a single deterministic 3D mesh output [4, 13, 19, 21]. These efforts mainly consider that systems returning a single deterministic output, tend to be sufficiently convenient and make comparisons on benchmarks straightforward and

[*] Corresponding author

fairly. But this often leads to unsatisfactory results, especially for challenging input images. On the other hand, some scholars accept the ill-poseness from 2D to 3D and the uncertainty from ambiguity and occlusion, and successively propose to estimate probability distributions or generate multi-hypotheses [26–29]. Although these works have shown interesting potentials, they often rely on one-to-many mappings by adding multiple output heads to the existing architectures, which leads to potentially unscalable and poorly expressive multi-hypothesis output. Also, they suffer from an important shortcoming in failing to establish the relationship between the different hypothesis features, because it is essential to improve the expressiveness and performance of the model.

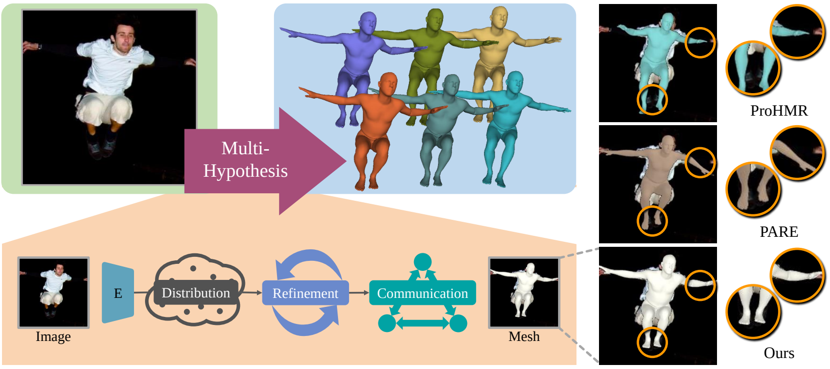

Fig. 1. We propose a multi-hypothesis method to recovering 3D human meshes from monocular images. Right: recovery results of the probabilistic method ProHMR [27], SOTA method PARE [21] and our method for a challenging image.

Our method aims to generate multi-hypotheses from the input monocular image and construct their relationships to enrich the diversity of features and obtain more accurate final results. To achieve this, we propose MHPro, which has many desirable properties missed in recent work. We first use a probabilistic model based on normalizing flow to regress a feasible pose distribution and generate multiple initial human mesh hypotheses, as depicted in Fig. 1. Then, we propose two transformer-based modules, the *Intra-hypothesis refinement* module and the *Inter-hypothesis communication* module, to construct hypothetical relationships and enhance feature representations. The former module focuses on refining the features of each single hypothesis, which models each hypothesis feature separately and enhances the information transfer within each hypothesis. In addition, for all the hypotheses to share their respective enhancements, a single fusion representation is converged from multi-hypotheses, and is then divided into several divergent hypotheses. But the relationship between different hypotheses is not sufficient. To address this, the latter module captures relationships and passes information among hypotheses. Finally, multi-hypotheses are aggregated to regress the final human mesh.

We conduct extensive experiments to demonstrate the validity of our proposed MHPro and the importance of refining and communicating the hypotheses. Experimental results demonstrate our ability to represent features and generate more accurate human mesh recovery results, especially for monocular image inputs including depth ambiguity, joint occlusion and truncation. Our contributions can be summarized as follows:

- We propose MHPro for human mesh recovery from monocular images. Our model can efficiently and adequately learn the feature representation of multi-hypotheses.
- We achieve a better representation of image features and establish strong relationships among hypotheses, using two transformer-based modules.
- Our MHPro achieves the best performance on the large-scale benchmark Human3.6M and the challenging 3DPW dataset, even for the cases of depth ambiguity, joint occlusion and truncation.

2 Related Work

In this section, we mainly discuss the human mesh recovery from monocular images. Due to limited space, here we only discuss the most relevant methods and suggest the interested readers refer to the recent surveys [5, 6]. Apart from this, the recent multi-hypothesis methods that have been introduced into human reconstruction, and transformer in computer vision, are presented here.

2.1 Human Mesh Recovery from Monocular Images

Recovering 3D human meshes from monocular images is quite challenging due to the inherent ambiguity in lifting 2D observations into 3D space, flexible body structures and insufficient annotated 3D data.

Previous methods proposed to use a parametric human model and estimate the pose and shape coefficients for human mesh recovery. SMPL [7] is one of the widely used parametric human models, which is also used in this work. Bogo *et al.* [8] proposed SMPLify to estimate 3D human mesh by fitting the SMPL model to the predicted 2D keypoints and minimizing the re-projection error. Lassner *et al.* [9] used silhouettes and 2D keypoints in the optimization procedure to capture the overall information from a simple 2D input. In turn, Song *et al.* [10] utilized the learning gradient descent method in the optimization process. These optimization-based methods are fragile and inefficient, require additional data, and struggle with time-consuming inference on image inputs. In contrast, regression-based methods [11–21] trained deep neural networks for regressing SMPL parameters directly from pixels and enhanced the robustness and plausibility of the results. For example, HMR [11], a regressor from 2D joints to SMPL parameters, used a discriminator of unpaired 3D data to encourage plausible poses. SPIN [13] revisited reconstruction methods that work with neural networks and extended SMPLify [8] to provide more supervision in the training loop. Unlike previous work, PARE [21] focused on predicting body-part-guided attention masks and achieved a degree of robustness to occlusion.

Although these methods have produced encouraging results and the issue of occlusion has been focused on, they are still not robust enough and produce only approximate single reconstructions. In this work, we generate multiple plausible hypotheses from monocular images with the help of probabilistic models, and further improve the model's recovery accuracy in cases of depth ambiguity, joint occlusion and truncation through refinement and communication.

2.2 Multi-Hypothesis Methods

Multiple hypothesis methods have been gradually introduced into 3D human pose estimation and human mesh reconstruction, to deal with the inherent ambiguities of the reconstructions described earlier, such as depth ambiguity, joint occlusion or truncation. Several recent works generated different hypotheses for this problem and demonstrated significant performance gains relative to a single solution [22–29]. For example, Li *et al.* [22] proposed multi-modal hybrid density networks to generate multiple feasible 3D pose hypotheses. Oikarinen *et al.* [27] followed conditional normalizing flows to model the conditional probability distribution, which makes for a more powerful and expressive model. Li *et al.* [29] proposed a multi-hypothesis transformer to learn the spatio-temporal representation of multiple plausible pose hypotheses and modeled multi-hypothesis features for accurate 3D human pose estimation from monocular videos. Unlike their work, we attempt not just to generating plausible human pose and shape, but to establish strong relationships between hypothesis features and achieve effective modeling of different features through intra-hypothesis refinement and inter-hypothesis communication.

2.3 Transformer in Computer Vision

Transformer [30], an encoder-decoder model, is first proposed in NLP field. Inspired by its achievements, the transformer, equipped with a powerful multi-head self-attention mechanism, has received increasing research attention in the computer vision community. Vision Transformer (ViT) [31] considered an image as a 16x16 patch sequence, and trained a standard transformer architecture for image classification. METRO [32] achieved progressive dimensionality reduction using a multi-level transformer for pose estimation. In addition, transformer has also achieved impressive results in many downstream tasks, including image generation [33], denoising [34], object detection [35], video inpainting [36], *etc.*

3 Method

Our aim is to achieve higher performance in human mesh recovery from monocular images. Fig. 2 shows the framework of our method. In our method, we extract image features from a given input image, establish a pose distribution, construct hypothetical relationships, enhance feature representations, and finally output accurate recovery results. Our method consists of three steps: 1) probabilistic modeling and initial hypothesis generation (Sec. 3.2); 2) Intra-hypothesis Refinement (Sec. 3.3); 3) Inter-hypothesis Communication (Sec. 3.4).

3.1 Preliminary

SMPL Model. SMPL [7] provides a differentiable function $\mathcal{M}(\theta, \beta)$ which takes body pose parameters $\theta \in R^{72}$ and shape parameters $\beta \in R^{10}$ as inputs and outputs the body mesh $M \in R^{6890\times 3}$. While θ represents the global body rotation and the relative rotation of 23 joints in axis-angle format, β represents the first 10 coefficients of a PCA shape space, controlling the shape of the body. Given the mesh M, 3D joint locations can be obtained using a linear regressor, $\mathbf{J}^{\mathrm{3D}} = JM$, where $J \in R^{L\times 6890}$ is a regression matrix for L joints.

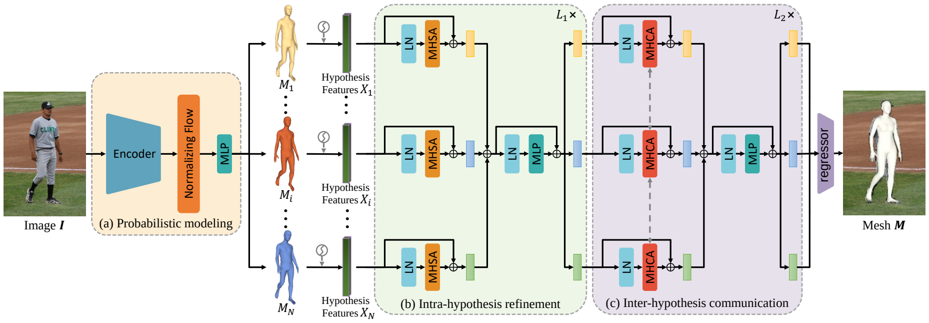
Image I
Encoder
Normalizing Flow
MLP
(a) Probabilistic modeling
M_1
M_i
M_N
Hypothesis Features X_1
Hypothesis Features X_i
Hypothesis Features X_N
$L_1\times$
LN
MHSA
MLP
(b) Intra-hypothesis refinement
$L_2\times$
MHCA
(c) Inter-hypothesis communication
regressor
Mesh M

Fig. 2. Overview of the proposed method. Given an input monocular image $\mathbf{I}$, we perform probabilistic modeling (a) with normalizing flows to extract image features, predict a pose distribution and generate multiple initial human mesh hypotheses (N indicates the number of hypotheses), input these multi-hypotheses into *Intra-hypothesis refinement* module (b) for independent refinement and feature enhancement, use *Inter-hypothesis communication* module (c) to implement their mutual communication and finally regress to obtain the recovered human mesh $\mathbf{M}$.

Transformer. Our refinement and communication of multi-hypotheses are based on the transformer architecture, as it performs well in feature representation and information stabilisation in propagation. Here we briefly describe Multi-Head Self-Attention (MHSA) and Multi-Layer Perceptron (MLP).

MHSA. In the MHSA, the inputs $X \in R^{n\times d}$ are linearly mapped to queries $Q \in R^{n\times d}$, keys $K \in R^{n\times d}$, and values $V \in R^{n\times d}$, where n is the sequence length and d is the dimension. Then, Q, K, and V, are split into h different subspaces, so that self-attention can be performed on them independently. Finally, the outputs from the different subspaces are concatenated to form the final result $Y \in R^{n\times d}$. The scaled dot-product attention can be expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{1}$$

MLP. The MLP consists of two linear layers, which are used for non-linearity and feature transformation:

$$\text{MLP}(X) = \sigma\left(XW_1 + b_1\right)W_2 + b_2, \tag{2}$$

where σ is activation function, $W_1 \in R^{d\times d_m}$ and $W_2 \in R^{d_m\times d}$ are the weights of the two linear layers respectively, and $b_1 \in R^{d_m}$ and $b_2 \in R^{d}$ are the bias terms.

3.2 Probabilistic Modeling

Given a monocular RGB image $\mathbf{I}$ as input, we attempt to learn a distribution of plausible poses conditional on $\mathbf{I}$ to obtain initial multiple plausible hypotheses. Inspired by ProHMR [27], we first encode the input image $\mathbf{I}$ using a CNN g to obtain image features $f_{\mathbf{I}}$. Subsequently, the probability distribution of the human pose $p_{\Theta|\mathbf{I}}(\theta \mid f_{\mathbf{I}} = g(\mathbf{I}))$ is modeled using Conditional Normalizing Flows. Unlike

ProHMR, we adopt probabilistic modeling only to obtain feasible initial multiple hypotheses, rather than focusing on one-to-many mappings.

Normalizing Flow models are used to transform arbitrary complex distributions into a simple base distribution $p_Z(z)$ by constructing a series of reversible transformations. Each building block f_i consists of 3 basic transformations:

$$f_i = f_{\mathrm{AC}} \circ f_{\mathrm{LT}} \circ f_{\mathrm{IN}}, \tag{3}$$

where $f_{\mathrm{IN}}(\mathbf{z}) = \mathbf{a} \odot \mathbf{z} + \mathbf{b}$ (Instance Normalization), $f_{\mathrm{LT}}(\mathbf{z}) = W\mathbf{z} + \mathbf{b}$ (Linear Transformation) and $f_{\mathrm{AC}} = [\mathbf{z}_{1:k}, \mathbf{z}_{k+1:d} + \mathbf{t}(\mathbf{z}_{1:d}, \mathbf{c})]$ (Additive Coupling). In addition, we combined four building blocks as above to obtain our flow model.

Meanwhile, the flow model allows not only for fast computation of probability distributions, but also for fast sampling from the distributions to obtain multiphypotheses. In order not to lose generality, we consider the case where no other additional information is available, so instead of taking a direct mode computation from the output probability distribution with $\theta_I^* = \operatorname{argmax}_\theta p_{\Theta \mid f_{\mathbf{I}}}(\theta \mid f_{\mathbf{I}})$, we sample the distribution to select the larger probability N hypotheses. Therefore, the samples $\theta_i, i \in [1, 2, ..., N]$ drawn from the output distribution are:

$$\theta_i \sim p_{\Theta \mid \mathbf{I}}(\theta \mid f_{\mathbf{I}}). \tag{4}$$

Then, we use MLP to estimate the SMPL shape β_i and camera parameters π_i using image features f_I and pose θ_i as input:

$$[\beta_i, \pi_i] = \mathrm{MLP}\left(f_I, \theta_i\right). \tag{5}$$

To summarize, we use probabilistic models to obtain a conditional probability distribution of poses, as well as sampling and estimation to obtain the initial human mesh hypotheses $M_i(\theta_i, \beta_i, \pi_i)$. However, these hypotheses are discrepant and insufficient for feature representation and need further enhancement.

3.3 Intra-hypothesis Refinement

After obtaining multiple human mesh recovery hypotheses $M_i(\theta_i, \beta_i, \pi_i)$, we first maintain its mesh information via a learnable positional embedding and encode its features $X_i, i \in [1, 2, ..., N]$ as subsequent inputs, for each hypothesis. To refine single-hypothesis features and enhance those coarse representations independently, the *Intra-hypothesis refinement* module feeds the encoded hypothesis features X_i into several parallel MHSA blocks, which can be represented as:

$$\widetilde{X}_i^l = X_i^{l-1} + \mathrm{MHSA}\left(\mathrm{LN}\left(X_i^{l-1}\right)\right), \tag{6}$$

where $l \in [1, 2, ..., L_1]$ is the index of *Intra-hypothesis refinement* module.

However it is not enough to process each hypothesis independently, the respective feature enhancements need to be shared. Thus, the hypothesis features are concatenated and fed into the MLP to mix themselves and forming refined hypothesis representations. The procedure can be represented as:

$$\begin{gathered} \widetilde{X}_{concat}^l = \mathrm{Concat}\left(\widetilde{X}_1^l, \widetilde{X}_2^l, \ldots, \widetilde{X}_N^l\right), \\ \mathrm{Concat}\left(\widetilde{X}_1^l, \widetilde{X}_2^l, \ldots, \widetilde{X}_N^l\right) = \widetilde{X}_{concat}^l + \mathrm{MLP}\left(\mathrm{LN}\left(\widetilde{X}_{concat}^l\right)\right), \end{gathered} \tag{7}$$

where $\mathrm{Concat}(\cdot)$ is the concatenation operation.

3.4 Inter-hypothesis Communication

To capture multi-hypothesis relationships mutually, we inherit the cross-attention mechanism from [38–40] and apply the Multi-Head Cross-Attention (MHCA) to model inter-hypothesis relationships.

Specifically, the multiple hypotheses feature X_i^l are alternately regarded as queries and keys, and fed into the MHCA:

$$X_i^l = X_i^{l-1} + \mathrm{MHCA}\left(\mathrm{LN}\left(X_1^{l-1}\right), \ldots, \mathrm{LN}\left(X_i^{l-1}\right), \ldots\right), \tag{8}$$

where $l \in [1, 2, ..., L_2]$ is the index of *Inter-hypothesis communication* module, $X_i^0 = \widetilde{X}_i^{L_1}$. As a result, MHCA passes information crosswise among hypotheses to significantly enhance feature representation and modelling capabilities.

Similarly, here we proceed to mix the hypothesis features obtained, as well as forming hypothesis representations after communication:

$$\begin{gathered} X_{concat}^l = \mathrm{Concat}\left(X_1^l, X_2^l, \ldots, X_N^l\right), \\ \mathrm{Concat}\left(X_1^l, X_2^l, \ldots, X_N^l\right) = X_{concat}^l + \mathrm{MLP}\left(\mathrm{LN}\left(X_{concat}^l\right)\right), \end{gathered} \tag{9}$$

where Concat$(\cdot)$ is the concatenation operation. Considering that the final single estimation result is obtained, the hypothesis features can be optionally not divided in the last MLP. Note that you can likewise choose to divide and thus obtain multiple reasonable results.

Finally, a regressor is applied to the output feature X^{L_2} to produce the 3D human mesh $M(\theta_i, \beta_i, \pi_i)$.

3.5 Loss Function

We introduce multiple losses as supervision for the probability distribution and mesh recovery hypotheses, respectively.

NLL loss. As with typical probabilistic models, we use NLL loss to minimize the negative log-likelihood: $\mathcal{L}_{nll} = -\ln p_{\Theta|\mathbf{I}}\left(\theta_{gt} \mid f_{\mathbf{I}}\right)$.

2D joint loss. A squared error reprojection loss is applied between the ground truth J_{2D} and estimated 2D joints $\hat{J}_{2D}$: $\mathcal{L}_{2D}(\theta, \beta, \pi) = \|J_{2D} - \hat{J}_{2D}\|_2$.

3D loss. When 3D annotations (3D joints and/or SMPL parameters) are available, 3D loss is applied to reduce the errors between the ground truth and estimated values: $\mathcal{L}_{3D}(\theta, \beta) = \|J_{3D} - \hat{J}_{3D}\|_2 + \|\beta - \hat{\beta}\|_2 + \|\theta - \hat{\theta}\|_2$.

Orth loss. The 6D representation proposed in [37] is used in our model to estimate the rotations. Since the absence of any constraint on the 6D representation leads to large differences between examples with partial 3D and or 2D annotations, we use L_{orth} to force the 6D representation of the samples drawn from the distribution to be close to the orthogonal 6D representation.

Overall: In total, the objective function of our model is:

$$\mathcal{L} = \lambda_{nll}\mathcal{L}_{nll} + \lambda_{2D}\mathcal{L}_{2D} + \lambda_{3D}\mathcal{L}_{3D} + \lambda_{orth}\mathcal{L}_{orth}, \tag{10}$$

where λ_{nll}, λ_{2D}, λ_{3D} and λ_{orth} represent the weights of the corresponding losses.

4 Experimental Results

4.1 Datasets

Following the settings of previous work [11, 13], our method is trained on a mixture of data from several datasets with 3D and 2D annotations, including Human3.6M [41], MPI-INF-3DHP [42], 3DPW [43], COCO [44], and MPII [45]. In addition, we report experimental results on the evaluation sets of Human3.6M [41] and 3DPW [43], and apply widely used evaluation metrics, including Mean Per Joint Position Error (MPJPE) and Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE).

4.2 Comparison

We compare our method with the previous state-of-the-art temporal and frame-based methods on Human3.6M and 3DPW datasets. As shown in Tab. 1, our method achieves state-of-the-art performance in terms of accuracy in both the indoor dataset Human3.6M and the challenging field dataset 3DPW. It is worth noting that, our method outperforms the state-of-the-art temporal method MAED [20], whereas our method is a frame-based approach.

Fig. 3 shows the qualitative results of our method on LSP dataset. We observe that our method can better extract and represent the image features, and achieve more accurate mesh recovery. Moreover, we show the recovery results of our model for challenging monocular image inputs including depth ambiguity, joint occlusion and truncation, in Fig. 4. It can be seen that our model is able to handle them well by refining and communicating multi-hypotheses. We refer to the project website for more qualitative results.

Table 1. Quantitative evaluation of state-of-the-art temporal and frame-based methods on Human3.6M and 3DPW datasets. The best results are highlighted in bold and "-" shows the results that are not available.

Method	Human3.6M		3DPW	
	MPJPE↓	PA-MPJPE↓	MPJPE↓	PA-MPJPE↓
Temporal				
VIBE [16]	65.9	41.5	93.5	56.5
Lee *et al.* [18]	58.4	38.4	92.8	52.2
MAED [20]	56.3	38.7	88.8	50.7
Frame-based				
SPIN [13]	62.5	41.1	96.9	59.2
I2L-MeshNet [15]	55.7	41.1	93.2	57.7
ProHMR [27]	-	41.2	-	59.8
PyMAF [19]	57.7	40.5	92.8	58.9
PARE [21]	-	-	84.3	51.2
Ours	**54.8**	**38.1**	**83.7**	**50.5**

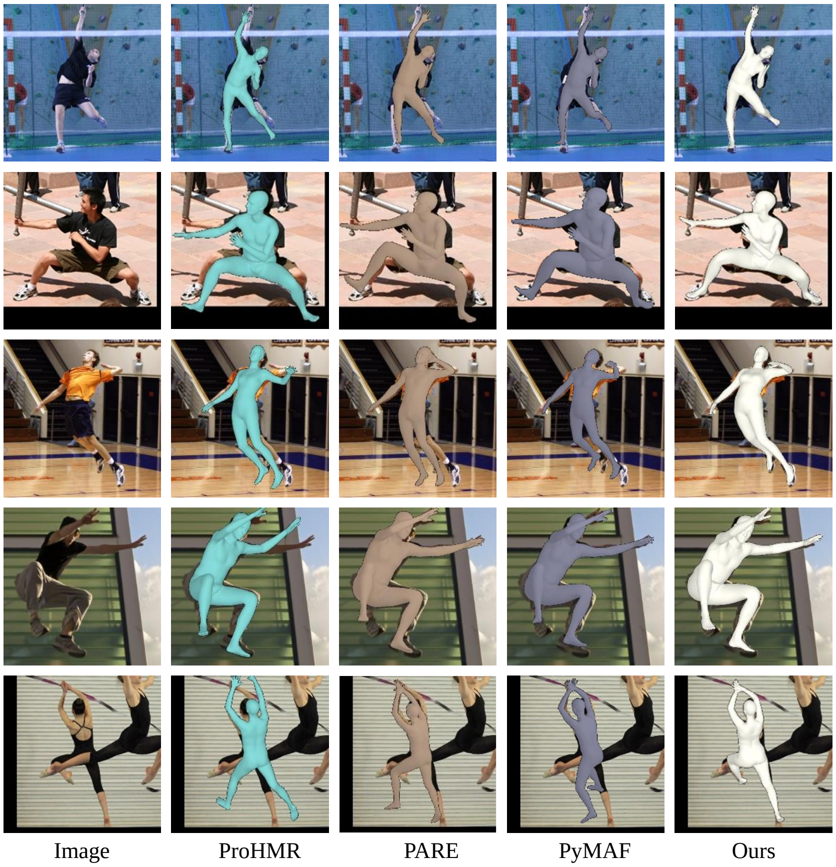

Fig. 3. Qualitative results on LSP dataset. From left to right: Input images, ProHMR [27] results, PyMAF [19] results, PARE [21] results, Our results.

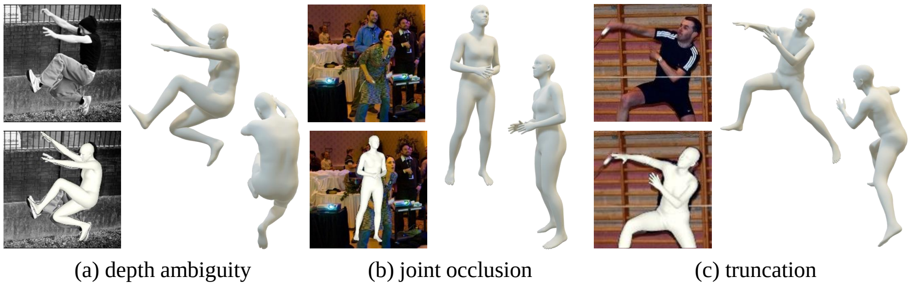

Fig. 4. Plausible human mesh recovery results generated by our method, especially for ambiguous parts with depth ambiguity, joint occlusion and truncation.

4.3 Ablation Study

We further conduct extensive ablation experiments on the effect of each key component and design in the proposed model. In the top part of Tab. 2, we report the results with different numbers of initial human mesh hypotheses. Experiments show that generating more hypotheses from the probabilistic model improves performance with a small increase in parameters, but becomes worse instead when $N > 8$. In the middle and bottom parts of Tab. 2, we report how the different parameters L_1 and L_2 impact the performance of our model, respectively. The validity and importance of our proposed modules for the experiment can be known from the results when $L_1 = 0$ or $L_2 = 0$, and the best performance of the model at $L_1 = 2$ and $L_2 = 2$.

Table 2. Ablation study on different parameters of our model, evaluated on Human3.6M. N is the hypothesis number, L_1 is the number of *Intra-hypothesis refinement* module and L_2 is the number of *Inter-hypothesis communication* module.

N	L_1	L_2	MPJPE↓	PA-MPJPE↓
6	2	2	60.1	44.3
8	2	2	**55.3**	**38.1**
12	2	2	58.7	40.2
20	2	2	61.6	42.1
8	2	0	70.2	50.8
8	2	1	65.9	46.4
8	2	2	**55.3**	**38.1**
8	2	3	59.5	43.7
8	0	2	65.3	47.1
8	1	2	58.3	40.2
8	2	2	**55.3**	**38.1**
8	3	2	60.7	42.5

5 Conclusion

In this paper, we present a multi-hypothesis and probabilistic model-based method, MHPro, for human mesh recovery from monocular images. Unlike most probabilistic modeling and multi-hypothesis methods, we propose to refine and communicate multi-hypothesis for a better image feature representation. Extensive experiments show that our method achieves state-of-the-art performance on two benchmark datasets and can better handle challenging images. Future work could consider continually extending our method to better exploit the ability of multi-hypotheses and promote recovery accuracy considering various ambiguities.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China (62171317 and 62122058).

References

1. Duan H, Zhao Y, Chen K, *et al.* Revisiting Skeleton-Based Action Recognition. In CVPR, 2022: 2969-2978.
2. Liu Y, Sivaparthipan C B, Shankar A. Human-computer Interaction Based Visual Feedback System for Augmentative and Alternative Communication. International Journal of Speech Technology, 2021: 1-10.
3. Weng C Y, Curless B, Kemelmacher-Shlizerman I. Photo Wake-Up: 3D Character Animation from a Single Photo. In CVPR, 2019: 5908-5917.
4. Khirodkar R, Tripathi S, Kitani K. Occluded Human Mesh Recovery. In CVPR. 2022: 1715-1725.
5. Zheng C, Wu W, Yang T, *et al.* Deep Learning-Based Human Pose Estimation: A Survey. ArXiv:2012.13392, 2020.
6. Tian Y, Zhang H, Liu Y, *et al.* Recovering 3D Human Mesh from Monocular Images: A Survey. ArXiv:2203.01923, 2022.
7. Loper M, Mahmood N, Romero J, *et al.* SMPL: A Skinned Multi-person Linear Model. In TOG, 2015, 34(6): 1-16.
8. Bogo F, Kanazawa A, Lassner C, *et al.* Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In ECCV, 2016: 561-578.
9. Lassner C, Romero J, Kiefel M, *et al.* Unite the People: Closing the Loop Between 3D and 2D Human Representations. In CVPR, 2017: 6050-6059.
10. Song J, Chen X, Hilliges O. Human Body Model Fitting by Learned Gradient Descent. In ECCV, 2020: 744-760.
11. Kanazawa A, Black M J, Jacobs D W, *et al.* End-to-end Recovery of Human Shape and Pose. In CVPR, 2018: 7122-7131.
12. Pavlakos G, Zhu L, Zhou X, *et al.* Learning to Estimate 3D Human Pose and Shape from a Single Color Image. In CVPR, 2018: 459-468.
13. Kolotouros N, Pavlakos G, Black M J, *et al.* Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. In ICCV, 2019: 2252-2261.
14. Kolotouros N, Pavlakos G, Daniilidis K. Convolutional Mesh Regression for Single-Image Human Shape Reconstruction. In CVPR, 2019: 4501-4510.
15. Moon G, Lee K M. I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image. In ECCV, 2020: 752-768.
16. Kocabas M, Athanasiou N, Black M J. VIBE: Video Inference for Human Body Pose and Shape Estimation. In CVPR, 2020: 5253-5263.
17. Jiang W, Kolotouros N, Pavlakos G, *et al.* Coherent Reconstruction of Multiple Humans from a Single Image. In CVPR, 2020: 5579-5588.
18. Lee G H, Lee S W. Uncertainty-Aware Human Mesh Recovery from Video by Learning Part-Based 3D Dynamics. In ICCV, 2021: 12375-12384.
19. Zhang H, Tian Y, Zhou X, *et al.* PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop. In ICCV, 2021: 11446-11456.
20. Wan Z, Li Z, Tian M, *et al.* Encoder-decoder with Multi-level Attention for 3D Human Shape and Pose Estimation. In ICCV, 2021: 13033-13042.
21. Kocabas M, Huang C H P, Hilliges O, *et al.* PARE: Part Attention Regressor for 3D Human Body Estimation. In ICCV, 2021: 11127-11137.
22. Li C, Lee G H. Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network. In CVPR, 2019: 9887-9895.
23. Li C, Lee G H. Weakly Supervised Generative Network for Multiple 3D Human Pose Hypotheses. ArXiv:2008.05770, 2020.

24. Biggs B, Novotny D, Ehrhardt S, *et al.* 3D Multi-bodies: Fitting Sets of Plausible 3D Human Models to Ambiguous Image Data. In NIPS, 2020, 33: 20496-20507.
25. Tuomas Oikarinen, Daniel Hannah, and Sohrob Kazerounian. GraphMDN: Leveraging Graph Structure and Deep Learning to Solve Inverse Problems. In IJCNN, 2021: 1-9.
26. Wehrbein T, Rudolph M, Rosenhahn B, *et al.* Probabilistic Monocular 3D Human Pose Estimation with Normalizing Flows. In ICCV, 2021: 11199-11208.
27. Kolotouros N, Pavlakos G, Jayaraman D, *et al.* Probabilistic Modeling for Human Mesh Recovery. In ICCV, 2021: 11605-11614.
28. Sengupta A, Budvytis I, Cipolla R. Hierarchical Kinematic Probability Distributions for 3D Human Shape and Pose Estimation from Images in the Wild. In ICCV, 2021: 11219-11229.
29. Li W, Liu H, Tang H, *et al.* MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation. ArXiv:2111.12707, 2021.
30. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is All You Need. In NIPS, 2017, 30.
31. Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv:2010.11929, 2020.
32. Lin K, Wang L, Liu Z. End-to-End Human Pose and Mesh Reconstruction with Transformers. In CVPR, 2021: 1954-1963.
33. Jiang Y, Chang S, Wang Z. TransGAN: Two Pure Transformers Can Make One Strong GAN, and That Can Scale Up. In NIPS, 2021, 34.
34. Chen H, Wang Y, Guo T, *et al.* Pre-trained Image Processing Transformer. In CVPR, 2021: 12299-12310.
35. Dai Z, Cai B, Lin Y, *et al.* UP-DETR: Unsupervised Pre-training for Object Detection with Transformers. In CVPR, 2021: 1601-1610.
36. Zeng Y, Fu J, Chao H. Learning Joint Spatial-Temporal Transformations for Video Inpainting. In ECCV, 2020: 528-543.
37. Zhou Y, Barnes C, Lu J, *et al.* On the Continuity of Rotation Representations in Neural Networks. In CVPR, 2019: 5745-5753.
38. Chen C F R, Fan Q, Panda R. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In ICCV, 2021: 357-366.
39. Wei X, Zhang T, Li Y, *et al.* Multi-Modality Cross Attention Network for Image and Sentence Matching. In CVPR, 2020: 10941-10950.
40. Hou R, Chang H, Ma B, *et al.* Cross Attention Network for Few-shot Classification. In NIPS, 2019, 32.
41. Ionescu C, Papava D, Olaru V, *et al.* Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. In TPAMI, 2013, 36(7): 1325-1339.
42. Mehta D, Rhodin H, Casas D, *et al.* Monocular 3D Human Pose Estimation in the Wild Using Improved CNN Supervision. In 3DV, 2017: 506-516.
43. Von Marcard T, Henschel R, Black M J, *et al.* Recovering Accurate 3D Human Pose in the Wild Using IMUs and a Moving Camera. In ECCV, 2018: 601-617.
44. Lin T Y, Maire M, Belongie S, *et al.* Microsoft COCO: Common Objects in Context. In ECCV, 2014: 740-755.
45. Andriluka M, Pishchulin L, Gehler P, *et al.* 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In CVPR, 2014: 3686-3693.