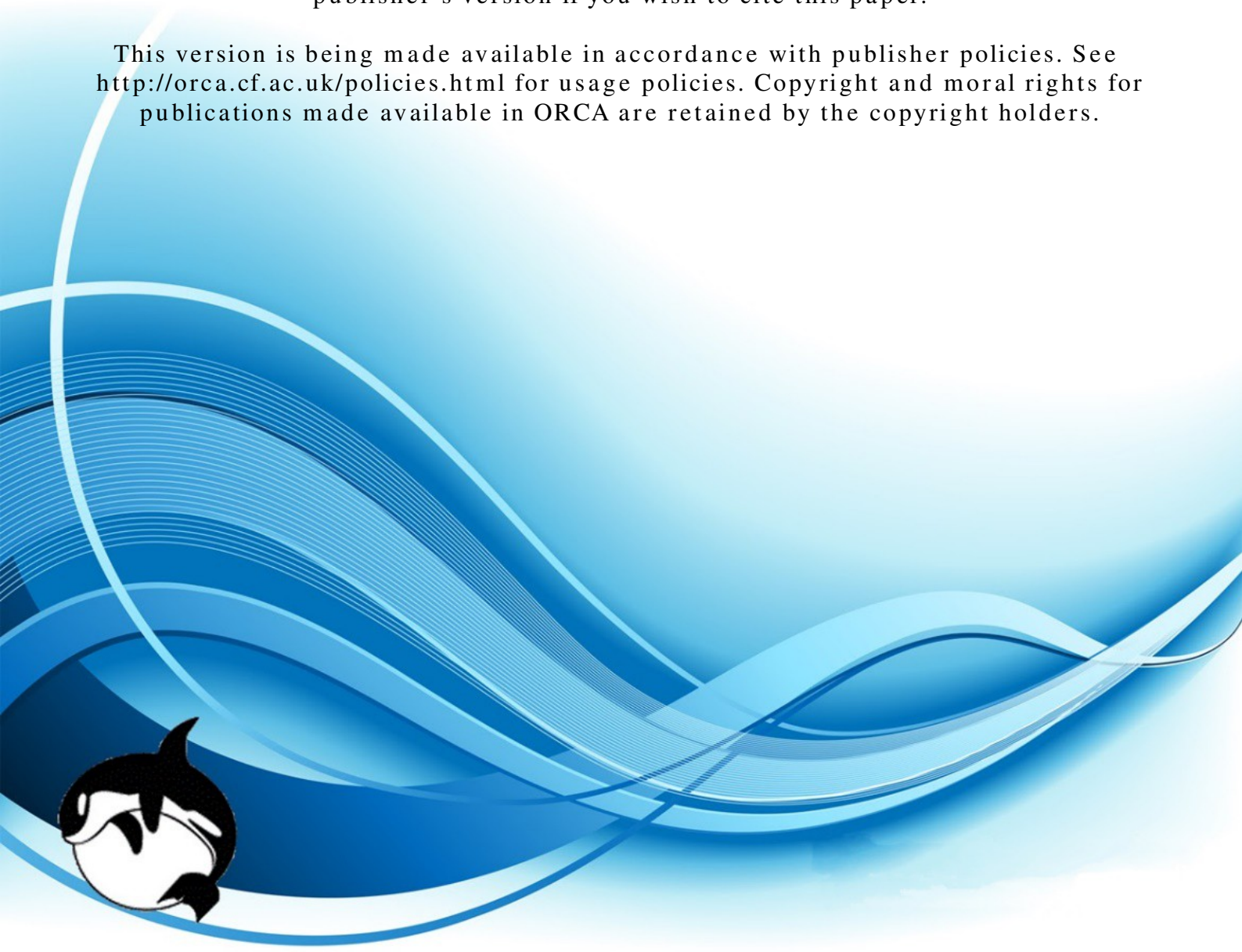


Received <day> <Month>, <year>; | Revised <day> <Month>, <year>; | Accepted <day> <Month>, <year>

DOI: xxx/xxxx

ORIGINAL RESEARCH

MH-HMR: Human Mesh Recovery from Monocular Images via Multi-Hypothesis Learning

Haibiao Xuan[1] | Jinsong Zhang[1] | Yu-Kun Lai[2] | Kun Li[1]

[1]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[2]School of Computer Science and Informatics, Cardiff University, Cardiff CF24 4AG, United Kingdom

Correspondence
Kun Li, College of Intelligence and Computing, Tianjin University, No.135 Yaguan Road, Jinnan District, Tianjin, China.
Email: lik@tju.edu.cn
Yu-Kun Lai, School of Computer Science and Informatics, Cardiff University, Senghenydd Road, Cardiff, CF24 4AG, UK.
Email: Yukun.Lai@cs.cardiff.ac.uk

Funding Information
National Key R&D Program of China, Grant Number: 2023YFC3082100;
National Natural Science Foundation of China, Grant Number: 62171317, 62122058;
Science Fund for Distinguished Young Scholars of Tianjin, Grant Number: 22JCJQJC00040.

Abstract

Recovering 3D human meshes from monocular images is an inherently ill-posed and challenging task due to depth ambiguity, joint occlusion, and truncation. However, most existing approaches do not model such uncertainties, typically yielding a single reconstruction for one input. In contrast, this paper embraces the ambiguity of the reconstruction and considers the problem as an inverse problem for which multiple feasible solutions exist. To address these issues, we propose a multi-hypothesis approach, MH-HMR, to efficiently model the multi-hypothesis representation and build strong relationships among the hypothetical features. Specifically, the task is decomposed into three stages: (1) generating a reasonable set of initial recovery results (*i.e.*, multiple hypotheses) given a single color image; (2) modeling intra-hypothesis refinement to enhance every single-hypothesis feature; and (3) establishing inter-hypothesis communication and regressing the final human meshes. Meanwhile, we further take advantage of multiple hypotheses and our recovery process to achieve human mesh recovery from multiple uncalibrated views. Compared with state-of-the-art methods, our approach MH-HMR achieves superior performance and recovers more accurate human meshes on challenging benchmark datasets like Human3.6M and 3DPW, while demonstrating our effectiveness across a variety of settings. The code will be publicly available for research purposes.

KEYWORDS:
Human Mesh Recovery, Monocular Images, Multi-Hypothesis

1 | INTRODUCTION

3D human mesh recovery from monocular images is a widely-studied problem and a popular research topic in computer vision, which can be the cornerstone for numerous applications including action recognition [1, 2, 3], human-computer interaction [4], augmented/virtual reality [5], *etc*. However, it remains a challenging task and an inherently ill-posed problem due to issues such as depth ambiguity in lifting 2D observation to 3D space, joint occlusion caused by flexible body structures, and truncation regarding insufficient input.

Given an input image, existing works for 3D human mesh recovery [6, 7, 8, 9] typically return a single 3D mesh output in a deterministic manner, largely due to its convenience in network designs, benchmark comparisons and downstream applications. But this often produces unsatisfactory results, especially for challenging input images. On the other hand, few methods recognize the ill-posedness and uncertainty of this problem, and successively propose to estimate probability distributions or explicitly generate multi-hypotheses [10, 11, 12, 13]. Despite their impressive performance, they tend to share feature extractors and add multiple output heads to existing architectures for one-to-many mappings, which leads to potentially non-scalable and inadequately expressive multi-hypothesis output. Apart from this, they fail to establish relationships among features of different hypotheses, which is a major problem that can significantly affect the performance and expressiveness of the model.

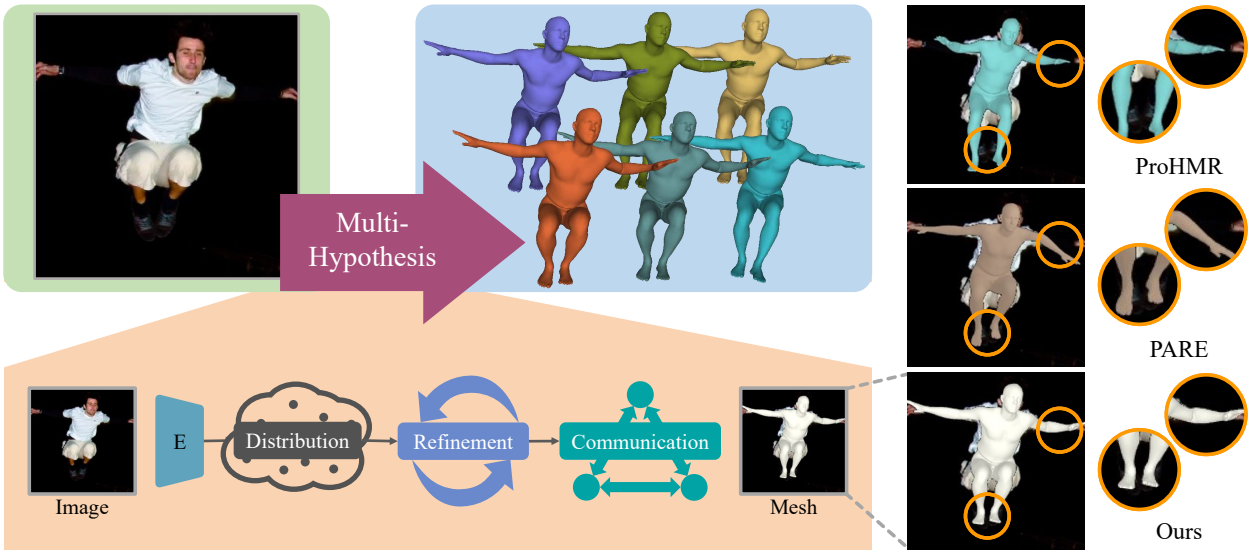


FIGURE 1 We propose, MH-HMR, to accurately recover a 3D human mesh given an input image. Right: results of the probabilistic method ProHMR [11], the state-of-the-art (SOTA) method PARE [9] and our approach for a challenging image.

Motivated by the above observation, we propose a novel multi-hypothesis approach, **MH-HMR**, to exploit image features and enhance feature learning for more accurate human mesh recovery. The central idea of our approach is to generate multiple feasible hypotheses from a single input image, progressively construct their relationships and integrate their respective feature expressiveness. In MH-HMR, 3D human evidence is initially extracted from the monocular image by a probabilistic model based on normalizing flow, and then fed into a feasible pose distribution regressor to obtain multiple initial hypotheses, as shown in Fig. 1 . In order to model multi-hypothesis consistencies and enhance those coarse representations, two transformer-based modules, namely the *Intra-hypothesis refinement* module and the *Inter-hypothesis communication* module, are proposed to construct hypothetical relationships and enhance feature learning. The former module focuses on refining every single-hypothesis feature, which models each hypothesis feature separately, enabling message passing within each hypothesis for feature enhancement. To exchange information across hypotheses, those multiple hypotheses are merged into a single fusion representation, and then partitioned into several divergent hypotheses. Meanwhile, the latter module is introduced to capture relationships and pass information among hypotheses so that our model can be aware of more accurate and plausible mesh features. Finally, we regress multiple feasible results or one definite result from the final multi-hypothesis features.

A preliminary version of our work has been presented in a conference paper [14]. In this paper, our work is extended from the following aspects: 1) Considering the important role of multi-hypothesis fusion and communication effects on our model performance, we propose the Hypothesis-Mixing Multi-Layer Perceptron to explore the relationship between channels with different hypotheses, and a new configuration of the Multi-Head Cross-Attention to achieve more thorough information exchanges among multi-hypotheses; 2) We demonstrate that our module designs and multi-hypothesis nature can effectively facilitate the multi-view fusion task by leveraging information from different views better; 3) We provide more details, more comprehensive experiments, and more thorough discussions to validate our performance.

Experimental results demonstrate our model has more learning ability for feature representation and can generate more accurate recovery results, especially for challenging monocular image inputs including cases with depth ambiguity, joint occlusion, and truncation, which demonstrates the robustness of our model. Fig. 1 gives an example. The code will be publicly available for research purposes.

Our contributions can be summarized as follows:

- We propose a novel multi-hypothesis approach, MH-HMR, for human mesh recovery, which can efficiently and adequately learn the feature representation of multiple hypotheses.
- We propose two transformer-based modules, the intra-hypothesis refinement module and the inter-hypothesis communication module, to achieve a better representation of image features and model the relationship among multi-hypothses.
- Our MH-HMR achieves superior performance on challenging benchmark datasets like Human3.6M and 3DPW, even for the cases with depth ambiguity, joint occlusion, and truncation.
- We demonstrate that our model can elegantly and efficiently leverage additional image information and handle the multi-view fusion task.

2 | RELATED WORK

In this section, we first mainly discuss the most relevant methods about human mesh recovery from monocular images and refer interested readers to the recent surveys [15, 16]. Then, we present the recent multi-hypothesis methods that have been introduced into human human pose estimation and mesh reconstruction, and conclude with a brief introduction to transformers in computer vision.

2.1 | Human Mesh Recovery from Monocular Images

Recovering 3D human meshes from monocular images is challenging because of the ambiguity in lifting 2D information into 3D space and the uncertainty caused by complex body variations and insufficient 3D annotations.

Recent works have made significant progress by using the pre-trained parametric human model such as SMPL [17] and estimating its hyper-parameters to represent the human body mesh. The optimization-based methods estimate the parameters of the body model iteratively, such that it is consistent with a set of features, like 2D keypoints, silhouettes and part segmentation. For example, Bogo *et al.* [18] propose SMPLify, a multi-stage optimization method that iteratively fits the SMPL model with 2D keypoints and minimizes the reprojection error to estimate a 3D human mesh. Lassner *et al.* [19] employ silhouettes together with 2D keypoints in the optimization procedure. Despite the well-aligned results can be obtained, these methods are sensitive to initialization, require additional data, and suffer from time-consuming fitting and inefficient inference. In contrast, taking advantage of the powerful nonlinear mapping capability of neural networks, regression-based methods [20, 21, 7, 22, 23, 24, 8, 25, 9, 26] train deep neural networks for regressing hyper-parameters directly from pixels. A canonical example is HMR [20], an end-to-end trainable human mesh recovery framework that utilizes the unpaired 3D annotations and penalizes implausible 3D human meshes with adversarial training. SPIN [7] combines HMR and SMPLify [18] in the training loop, resulting in better supervision for the network. PyMAF [8] proposes a mesh alignment feedback that leverages mesh-aligned evidence sampled from spatial feature maps to correct parameters in each loop. Unlike them, PARE [9] focuses on the partial occlusion problem, proposes a novel attention mechanism to predict body-part-guided attention masks, and uses information from neighbouring body parts to improve predictions for occluded parts.

Despite the promising results achieved by these methods, assuming only a single solution might be sub-optimal and becomes the bottleneck in this task. In our solution, multiple plausible hypotheses are generated from image features using probabilistic models and are enhanced to achieve a high-level and comprehensive perception.

2.2 | Multi-Hypothesis Methods

To cope with the inherent ambiguities of the reconstructions described earlier, multiple hypothesis methods have been gradually introduced into 3D human pose estimation and mesh reconstruction and achieve substantial performance gains.

Recently, a few approaches [27, 28, 10, 11, 12, 13, 29, 30] are proposed that generate different hypotheses using generative networks to cover the ambiguous nature. For instance, Li *et al.* [27] propose a mixture density network and learn the multi-modal posterior distribution to generate multiple feasible 3D pose parameters that are plausible estimates consistent with the ambiguous inputs, while Sengupta *et al.* [31] tackle this problem using simple multivariate Gaussian distributions. By contrast, Oikarinen *et al.* [11] model the conditional probability distribution using conditional normalizing flows, which makes the network even more powerful and expressive. Li *et al.* [13] design a multi-hypothesis transformer to exploit the spatio-temporal representation of multiple plausible pose hypotheses from monocular videos. Zheng *et al.* [29] take human silhouettes as input under the constraints of 2D joints and relative depth, and propose a two-stage weakly-supervised method to solve the multi-hypothesis problem of human pose and mesh. Holmquist *et al.* [30] introduce diffusion models into the multi-hypothesis method and combine an embedding transformer to represent the uncertainty in the 2D joint heatmaps.

Differently from these methods, the goal of MH-HMR is not only to generate plausible hypotheses (*i.e.*, one-to-many mappings), but also to establish strong relationships between hypothesis features and improve the representation ability (*i.e.*, many-to-one mappings). Therefore, MH-HMR can handle more ambiguous and complex images, and obtain stronger hypothesis features compared to existing methods, allowing for many downstream applications.

2.3 | Transformer in Computer Vision

Transformer [32], an encoder-decoder model, is first proposed in the natural language processing (NLP) field. Motivated by the achievements, various works start to apply transformer equipped with a powerful multi-head self-attention mechanism to the computer vision tasks. Vision Transformer (ViT) [33] treats an image as a 16×16 patch sequence, and apply a standard transformer architecture directly for image classification task. METRO [34] leverages a multi-level transformer to achieve progressive dimensionality reduction for pose estimation task. GLAMR [35] proposes a transformer-based motion-filling method to aid in global mesh recovery from monocular videos. In addition, the transformer has also achieved impressive results in many downstream tasks, including image generation [36], denoising [37], object detection [38], video inpainting [39], *etc*.

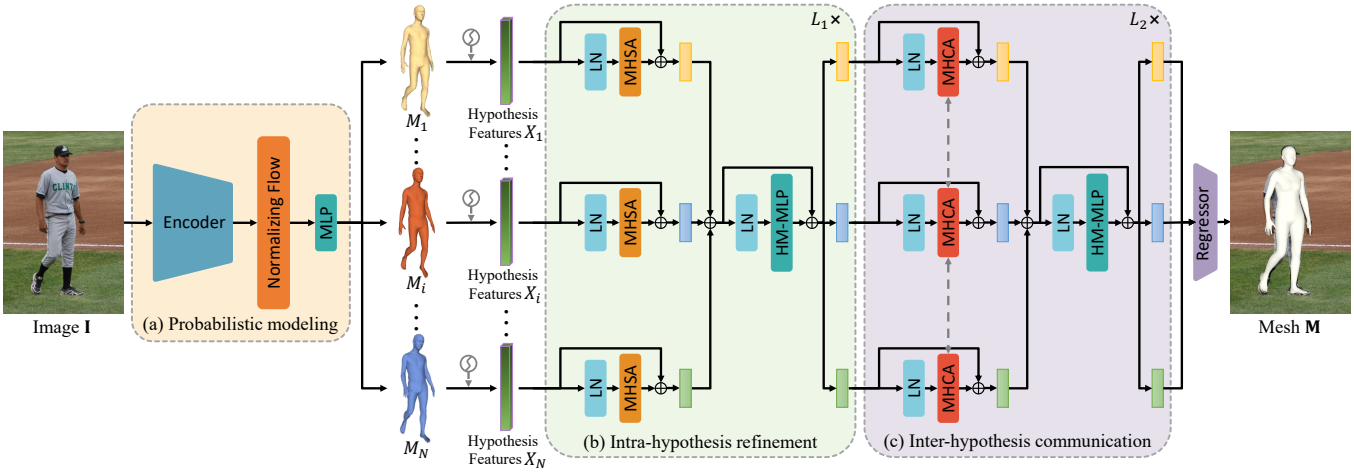


FIGURE 2 Overview of the proposed approach. Given an input monocular image **I**, we perform *probabilistic modeling* (a) with normalizing flows to extract image features, predict a pose distribution and generate multiple initial human mesh hypotheses (where N indicates the number of hypotheses), input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

3 | METHOD

Our goal is to leverage multi-hypothesis properties and relationships and recover a more accurate human mesh consistent with 2D image evidence. The overall framework of our approach, MH-HMR, is depicted in Fig. 2 . Our approach, MH-HMR, consists of three steps: 1) probabilistic modeling and initial hypothesis generation (Sec. 3.2); 2) intra-hypothesis refinement (Sec. 3.3); and 3) inter-hypothesis communication (Sec. 3.4). We discuss each component in more detail below.

3.1 | Preliminary

3.1.1 | SMPL Model

SMPL [17] is a classical parametric human body model. It defines a differentiable function $\mathcal{M}(\theta, \beta)$ that takes the pose parameters $\theta \in \mathbb{R}^{72}$ and the shape parameters $\beta \in \mathbb{R}^{10}$ as inputs and returns the body mesh $M \in \mathbb{R}^{6890\times3}$. θ represents the global body rotation and the relative rotation of 23 joints in axis-angle format, and β represents the first 10 coefficients of a PCA shape space, controlling the shape of the body. Given the mesh M, the SMPL 3D joint locations can be obtained using a pre-trained linear regressor, $J^{3D} = \mathcal{J}M$, where $\mathcal{J} \in \mathbb{R}^{K\times6890}$ is a regression matrix for K joints.

3.1.2 | Transformer

The transformer architecture is used for multi-hypothesis refinement and communication modules because it works well in feature representation and information stabilization in propagation. Here we briefly describe Multi-Head Self-Attention (MHSA) and Multi-Layer Perceptron (MLP).

MHSA. Given the inputs $X \in \mathbb{R}^{n\times d}$ where d is the hidden size, MHSA first linearly projects X to queries $Q \in \mathbb{R}^{n\times d}$, keys $K \in \mathbb{R}^{n\times d}$, and values $V \in \mathbb{R}^{n\times d}$, where n is the sequence length and d is the dimension. The scaled dot-product attention can be expressed as:

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

Then, MHSA splits the queries Q, keys K, and values V into h different subspaces as well as performs the attention in parallel. Finally, the outputs from the h different subspaces are concatenated to form the final result $Y \in \mathbb{R}^{n\times d}$.

MLP. The MLP used in our work consists of two linear layers (along with a nonlinear activation function between them), which are used for non-linearity and feature transformation:

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

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

3.2 | Probabilistic Modeling

Given a monocular RGB image **I** as input, our approach learns a distribution of plausible poses conditioned on **I** to obtain initial multiple plausible hypotheses. Inspired by ProHMR [11], we first encode the input image **I** using a Convolutional Neural Network (CNN) g and obtain image features $f_{\mathbf{I}}$. Then, the Conditional Normalizing Flow is applied to model the probability distribution of the human pose $p_{\Theta|\mathbf{I}}(\theta \mid f_{\mathbf{I}} = g(\mathbf{I}))$, due to their expressiveness and modeling capabilities. In contrast to ProHMR, we employ probabilistic modelling to extract image features and obtain multiple initial hypotheses that are both feasible to a certain extent and reflect different detailed features, rather than focusing on one-to-many mappings.

The normalizing flow is a series of reversible transformations that transforms arbitrary complex distributions into a simple base distribution $p_Z(z)$ (typically a standard multivariate Gaussian distribution). We combine four building blocks to obtain our flow model. 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 $f_{\mathrm{AC}} = \left[\mathbf{z}_{1:k}, \mathbf{z}_{k+1:d} + \mathbf{t}\left(\mathbf{z}_{1:d}, \mathbf{c}\right)\right]$ (Additive Coupling).

Moreover, the flow model provides fast computing of probability distributions as well as fast sampling from the distributions to produce multi-hypotheses. To ensure generality and robustness, we consider the case where no additional information is available. Thus, instead of taking a direct mode computation from the output probability distribution with $\theta_I^* = \mathrm{argmax}_\theta p_{\Theta|f_{\mathbf{I}}}(\theta \mid f_{\mathbf{I}})$, we sample the distribution to select N hypotheses with larger probabilities. The samples $\{\theta_i\}_1^N$ drawn from the output distribution are:

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

Then, we use an MLP to regress the SMPL shape $\{\beta_i\}_1^N$ and the camera parameters $\{\pi_i \in \mathbb{R}^3\}_1^N$ taking image features f_I and poses $\{\theta_i\}_1^N$ as input:

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

In summary, the probabilistic model based on the normalizing flow is used to construct conditional probability distributions of poses consistent with the input image, and then the initial N human mesh hypotheses $\{M_i(\theta_i, \beta_i, \pi_i)\}$ are produced by sampling and regression. However, these hypotheses, which include diverse and different image information, are not sufficient to represent the image features completely and accurately and therefore 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 adopt a learnable positional embedding inspired by [40] to maintain each mesh information, instead of using spatial information-dependent positional embedding. Then, we encode its features $\{X_i \in \mathbb{R}^C\}_1^N$ as subsequent inputs, where C is the embedding dimension.

The enhancement and information transfer of hypothesis features play an important role in achieving expressiveness and accuracy of the model. To refine the single-hypothesis feature and enhance those coarse representations independently, the *Intra-hypothesis refinement* module feeds the encoded hypothesis features $\{X_i\}_1^N$ into several parallel MHSA blocks (the structure of the MHSA block is shown in Fig. 3), which can be represented as:

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

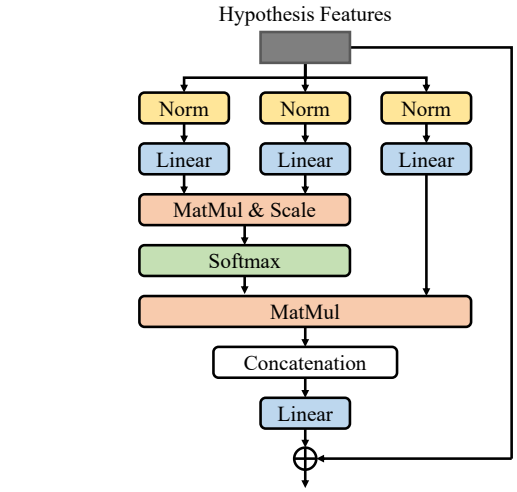


FIGURE 3 Multi-head self-attention (MHSA).

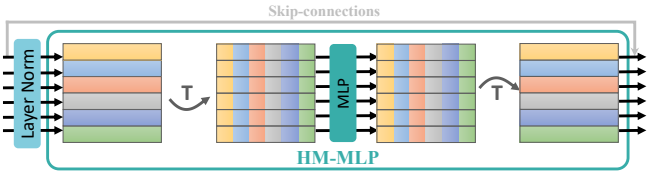


FIGURE 4 Hypothesis-Mixing MLP (HM-MLP).

where LN(·) is the LayerNorm layer, and $l \in [1, 2, ..., L_1]$ is the index of L_1 *Intra-hypothesis refinement* modules.

However, it is not enough to process each hypothesis independently, and the respective feature enhancements need to be shared. Thus, the hypothesis features are concatenated and fed into the Hypothesis-Mixing MLP (HM-MLP) to mix themselves and form the refined hypothesis representations.

The procedure can be represented as:

$$\begin{aligned} X_{concat}^{\prime l} &= \mathrm{Concat}\left(X_1^{\prime l}, X_2^{\prime l}, \ldots, X_N^{\prime l}\right) \\ X_{concat}^{l} &= X_{concat}^{\prime l} + \text{HM-MLP}\left(\mathrm{LN}\left(X_{concat}^{\prime l}\right)\right), \\ \left(Y_1^l, Y_2^l, \ldots, Y_N^l\right) &= \mathrm{Diverge}\left(X_{concat}^{l}\right), \end{aligned} \tag{7}$$

where $X_{concat}^{\prime l} \in \mathbb{R}^{C \times N}$, and Concat(·) and Diverge(·) are concatenation and division operations, respectively. HM-MLP(·) is the function of hypothesis-mixing MLP modified for the hypothetical features (as shown in Fig. 4), which explores the relationship between channels with different hypotheses.

3.4 | Inter-hypothesis Communication

To more explicitly incorporate differentiated feature representations and capture multi-hypothesis relationships mutually, we inherit the cross-attention mechanism from [41, 42, 43] and apply multiple Multi-Head Cross-Attention (MHCA) components in parallel. Note that although HM-MLP also plays a role in exchange, its more primary purpose is to fuse and repartition features. Thus, this communication module using cross-attention is still needed to achieve more effective message passing and stronger relationships.

The MHCA used in our conference version (denoted as MHCA-Conf) follows the common configuration of using the

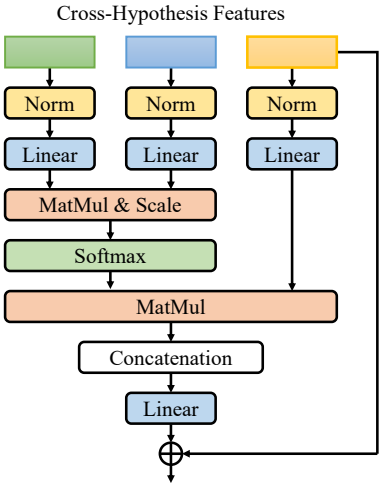


FIGURE 5 Multi-head cross-attention (MHCA).

same input between keys and values. However, this configuration tends to lead to inadequate communication between hypotheses and information transfer being trapped in localized areas. In addition to this, when the number of hypotheses is high, the need for more blocks takes up a larger number of parameters, affecting the efficiency of the model. Considering the above problems, we modify the conference version and adopt a more efficient strategy using different inputs (as shown in Fig. 5), to reduce the number of parameters and enhance the communication and transfer.

The multi-hypothesis features $\{Y_i \in \mathbb{R}^C\}_1^N$ are alternately regarded as queries and keys, and fed into the MHCA:

$$Y_i'^l = Y_i^{l-1} + \text{MHCA}\left(\text{LN}\left(Y_{i_1}^{l-1}\right), \text{LN}\left(Y_{i_2}^{l-1}\right), \text{LN}\left(Y_i^{l-1}\right)\right), \tag{8}$$

where Y_{i_1} and Y_{i_2} are the other two corresponding hypotheses, $l \in [1, 2, ..., L_2]$ is the index of L_2 *Inter-hypothesis communication* modules, and $Y_i^0 = X_i^{L_1}$. Finally, MHCA passes information among hypotheses in a crossing way to significantly enhance feature representation and modeling capabilities.

Similarly, we proceed to mix the obtained hypothesis features, and form the hypothesis representations after communication:

$$\begin{aligned} Y_{concat}'^l &= \text{Concat}\left(Y_1'^l, Y_2'^l, \ldots, Y_N'^l\right), \\ Y_{concat}^l &= Y_{concat}'^l + \text{HM-MLP}\left(\text{LN}\left(Y_{concat}'^l\right)\right), \\ \left(Z_1^l, Z_2^l, \ldots, Z_N^l\right) &= \text{Diverge}\left(Y_{concat}^l\right), \end{aligned} \tag{9}$$

where $Y_{concat}'^l \in \mathbb{R}^{C\times N}$, and Concat$(\cdot)$ and Diverge$(\cdot)$ are concatenation and division operations, respectively. We can choose whether to divide the hypothetical features in the last MLP to obtain multiple plausible results or a single final estimate.

Finally, a regressor is applied to the output feature $Z^{L_2} \in \mathbb{R}^{C\times N}$ to produce the 3D human mesh $M(\theta, \beta, \pi)$.

3.5 | Loss Function

To train our model, we apply multiple losses as supervision.

NLL loss. As with typical probabilistic models, our normalizing flow models are trained to minimize the negative log-likelihood of the ground truth θ_{gt}, *i.e.* the loss function is:

$$\mathcal{L}_{nll} = -\ln p_{\Theta|\mathbf{I}}\left(\theta_{gt} \mid f_{\mathbf{I}}\right). \tag{10}$$

2D loss. To penalize misalignment between the 2D projection and image evidences, we apply a squared reprojection error loss between the ground truth $J_{2D} \in \mathbb{R}^{K\times 2}$ and the estimated 2D keypoints $\hat{J}_{2D} \in \mathbb{R}^{K\times 2}$, where K is the number of joints of a person:

$$\mathcal{L}_{2D}(\theta, \beta, \pi) = \|J_{2D} - \hat{J}_{2D}\|_2. \tag{11}$$

3D loss. Additional 3D supervisions are added when 3D annotations (3D joints $J_{3D} \in \mathbb{R}^{K\times 3}$ and/or SMPL parameters θ, β) are available:

$$\mathcal{L}_{3D}(\theta, \beta) = \|J_{3D} - \hat{J}_{3D}\|_2 + \|\theta - \hat{\theta}\|_2 + \|\beta - \hat{\beta}\|_2. \tag{12}$$

Orthonormal loss. The 6D representation [44] is used to model rotations in our approach. Without any constraint restriction on the 6D representation, it would lead to a large difference between examples with full 3D SMPL parameter supervision and those with only 2D keypoint annotations. Thus, we use L_{orth} to force the 6D representation of the recovered samples to be close to the orthogonal 6D representation.

Our overall objective function is formulated as:

$$\mathcal{L} = \lambda_{nll}\mathcal{L}_{nll} + \lambda_{2D}\mathcal{L}_{2D} + \lambda_{3D}\mathcal{L}_{3D} + \lambda_{orth}\mathcal{L}_{orth}, \tag{13}$$

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

4 | APPLICATION: MULTI-VIEW FUSION

Multi-view fusion is a key technology for human mesh recovery from multi-view images. The ultimate goal is to recover a 3D body mesh in a world coordinate system from multiple cameras placed in natural environments. Although our model has been trained for single-image reconstruction, we can utilize existing module designs and multi-hypothesis features to obtain the refined pose and shape estimations of a person under multiple views. We address this problem with multi-hypothesis modeling, refinement and communication, which make the model pay attention to the consistency of body poses and shapes corresponding to different views.

Given uncalibrated multi-view images $\{\mathbf{I}_i\}_1^N$ of the same subject, we input them separately into *probabilistic modeling* (in Sec. 3.2) to obtain the initial SMPL body parameters and then partition those vectors of each frame as $\Theta_n = \{\theta_n^g, \theta_n^b, \beta_n\}$, where θ_n^g corresponds to the global rotation of the model, θ_n^b is the body pose and β_n is the body shape. Subsequently, the corresponding hypotheses for each frame are fed

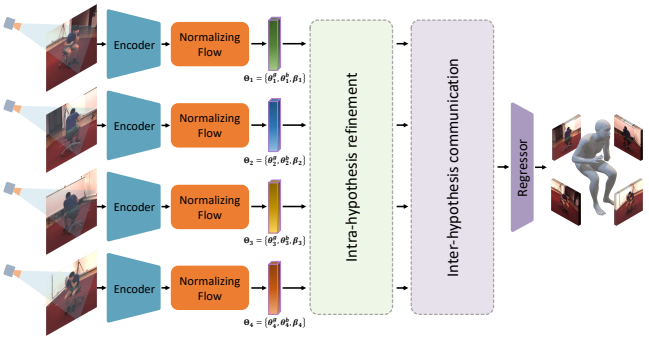


FIGURE 6 Our pipeline for the multiple view fusion task.

into the *Intra-hypothesis refinement* module (in Sec. 3.3) and the *Inter-hypothesis communication* module (in Sec. 3.4) in parallel, allowing the exchanges and fusion of image features in different views. Fig. 6 shows the overview of our proposed approach for the multi-view fusion task.

We refine and fuse multiple view information by minimizing the following loss:

$$\mathcal{L}_{mvf} = -\sum_{n=1}^{N} \ln p\left(\theta_n \mid f_{\mathbf{I}_n}\right) + \lambda_\theta \sum_{n=1}^{N} \left\|\theta_n^b - \tilde{\theta}^b\right\|_2^2 + \lambda_\beta \sum_{n=1}^{N} \left\|\beta_n - \tilde{\beta}\right\|_2^2, \quad (14)$$

where $\tilde{\theta}^b = \frac{1}{N}\sum_{n=1}^{N}\theta_n^b$ and $\tilde{\beta}^b = \frac{1}{N}\sum_{n=1}^{N}\beta_n$. The last two terms of the loss represent the squared distances between all the pose pairs and shape pairs, respectively.

5 | EXPERIMENTS

5.1 | Datasets and Metrics

Traing. Following previous works [20, 7], our approach uses mixed datasets with 3D and 2D annotations for training, including Human3.6M [45], MPI-INF-3DHP [46], 3DPW [47], LSP [48], MPII [49] and COCO [50].

Evaluation. We report the experiments results on the Human3.6M [45] and 3DPW [47] evaluation sets. We adopt the widely-used evaluation metrics for quantitative comparisons with previous methods including Mean Per Joint Position Error (MPJPE), Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE) , and Per Vertex Error (PVE).

5.2 | Implementation Details

The proposed MH-HMR model is implemented in PyTorch framework on a single NVIDIA RTX2080Ti GPU and validated on the ResNet-50 [51] backbone pre-trained on ImageNet [52]. We train our model with a batch size of 64 using the Adam optimizer [53] with the learning rate 0.0001 and the weight decay 0.0001. MH-HMR generates 8 initial hypotheses and contains 2 refinement modules and 2 communication modules. The loss weights are: λ_{nll}=0.001, λ_{2D}=0.01, λ_{3D}=0.05, and λ_{orth}=0.1. For the multi-view fusion task, we set λ_θ to 0.00 1 and λ_β to 0.0005. Our proposed method, MH-HMR, takes about 1.724 s to process one sample on the machine with an NVIDIA RTX 2080Ti GPU. For multi-view fusion task, MH-HMR takes about 2.131 s to process one sample.

FIGURE 7 Qualitative results on LSP [48] dataset. From left to right shows the input images, and the results of ProHMR [11], PyMAF [8], PARE [9] and Ours.

FIGURE 8 Qualitative results on LSP [48] dataset and Human3.6M [45] dataset. From left to right shows the input images, and the results of Ours-conf and Ours.

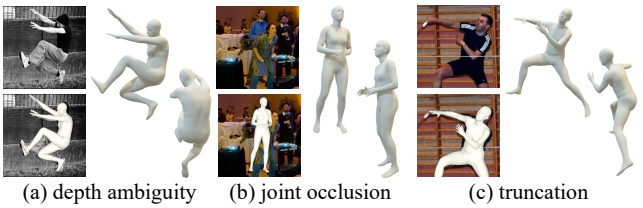


FIGURE 9 Plausible human mesh recovery results generated by our approach, especially for ambiguous parts with depth ambiguity, joint occlusion, and truncation.

TABLE 1 Quantitative comparison with the state-of-the-art temporal and frame-based methods on Human3.6M [45] and 3DPW [47] datasets. The best results are highlighted in bold and "-" represents that the results are not available.

Method	Human3.6M		3DPW		
	MPJPE ↓	PA-MPJPE ↓	MPJPE ↓	PA-MPJPE ↓	PVE ↓
Temporal					
VIBE [23]	65.9	41.5	93.5	56.5	113.4
TCMR [54]	62.3	41.1	95.0	55.8	111.3
Lee *et al.* [24]	58.4	38.4	92.8	52.2	106.1
MAED [25]	56.3	38.7	88.8	50.7	104.5
Frame-based					
SPIN [7]	62.5	41.1	96.9	59.2	135.1
I2L-MeshNet [22]	55.7	41.1	93.2	57.7	-
ProHMR [11]	-	41.2	-	59.8	-
ROMP [55]	-	-	89.3	53.5	103.1
THUNDR [56]	55.0	39.8	-	-	-
PyMAF [8]	57.7	40.5	92.8	58.9	110.1
PARE [9]	-	-	84.3	51.2	101.2
Baseline	56.2	40.6	86.9	53.1	100.2
Ours-Conf	54.8	38.1	83.7	50.5	94.4
Ours	**53.6**	**37.4**	**82.2**	**49.6**	**93.3**

5.3 | Comparison

We qualitatively and quantitatively compare our approach with the state-of-the-art temporal and frame-based methods, including MAED [25], SPIN [7], ProHMR [11], PyMAF [8], and PARE [9].

We present quantitative comparison results on Human3.6M and 3DPW datasets in Tab. 1 . Our MH-HMR achieves competitive or superior results compared with previous approaches. The methods reported in Tab. 1 are not strictly comparable because they may use different training data, learning rate schedules, or training epochs, *etc.*, which could affect their performance. For a fair comparison, we report the results of our baseline in Tab. 1 , which is trained under the same setting as MH-HMR and has the same network architecture as ProHMR [11]. In comparison with the baseline, MH-HMR reduces the MPJPE by 2.6 mm and 4.7 mm on Human3.6M and 3DPW datasets, respectively. From Tab. 1 , we can see that MH-HMR has more notable improvements on the metrics MPJPE and PVE. It is worth noting that, our MH-HMR outperforms the state-of-the-art temporal method MAED [25], despite the fact that our approach is frame-based.

Recovery results on the LSP [48] dataset are depicted in Fig. 7 for qualitative comparison, where MH-HMR convincingly performs better than the probabilistic method ProHMR [11], and the SOTA methods PyMAF [8] and PARE [9] by producing better aligned and more natural results.

As shown in Tab. 1 , compared to the conference version, we reduce the MPJPE by 1.2 mm and 1.5 mm on Human3.6M and 3DPW datasets, respectively. In addition to this, qualitative results are shown in Fig. 8 . They both demonstrate the validity and importance of the proposed extension HM-MLP and the new configuration of the MHCA.

Moreover, we show more recovery results of our model for challenging monocular image inputs including depth ambiguity, joint occlusion, and truncation, in Fig. 9 . It can be seen that our model is able to handle these cases well by refining and communicating multi-hypotheses.

More qualitative results can be found in the demo video [1].

5.4 | Ablation Study

We conduct several ablation studies to evaluate our approach in different settings and validate our contributions. All ablation approaches are trained and tested on Human3.6M [45], as it includes ground-truth 3D labels and is the most widely-used benchmark for 3D human mesh recovery.

Number of initial hypotheses. In MH-HMR, a larger number of initial hypotheses can provide more information on image features and more room for improvement subsequently, which is essential for better mesh recovery. However, an excessive number of initial hypotheses also tend to affect network efficiency and prevent adequate communication. To verify this, we report the performance of different variants with different numbers of hypotheses in probabilistic modeling in Tab. 2 -A. Experiments show that generating more hypotheses improves performance with a small increase in parameters, but becomes

[1]Our demo video at http://cic.tju.edu.cn/faculty/likun/projects/MH-HMR/imgs/demo.mp4

TABLE 2 Ablation study on different parameters of our model. N is the number of hypotheses, L_1 is the number of *Intra-hypothesis refinement* modules and L_2 is the number of *Inter-hypothesis communication* modules.

	N	L_1	L_2	MPJPE↓	PA-MPJPE↓
A	6	2	2	59.9	43.0
	8	2	2	**53.6**	**37.4**
	12	2	2	55.7	39.5
	20	2	2	60.3	40.6
B	8	2	0	67.8	46.4
	8	2	1	63.4	40.2
	8	2	2	**53.6**	**37.4**
	8	2	3	55.3	41.2
C	8	0	2	63.8	44.7
	8	1	2	56.1	39.6
	8	2	2	**53.6**	**37.4**
	8	3	2	55.9	42.0

worse instead for more than 8 hypotheses. Therefore, in our main experiments, we choose to use 8 initial assumptions as a good tradeoff between performance and complexity. Note that the performance of our approach can remain stable and advantageous with a small number of hypotheses.

Number of layers of two modules. Tab. 2 -B and Tab. 2 -C report how the different numbers of layers of refinement and communication modules impact the performance of our model. The results show that expanding the number of layers to 2 improves the performance, but stacking more modules does not lead to further improvements. Therefore, the optimal parameters for our model are $L_1 = 2$ and $L_2 = 2$.

Impact of HM-MLP. Reasonable hypothesis fusion settings help to fully utilize the capability of multiple hypotheses and improve the reliability of the mesh extracted from the hypothetical features. For deeper analysis and better quality of mesh recovery, we improve the MLP as HM-MLP for concatenation and division in modules, which is better adapted to the hypothetical features. As shown in Tab. 3 , when applying HM-MLP, the errors are reduced by 0.5 mm and 0.3 mm in MPJPE and PA-MPJPE, respectively. Meanwhile, HM-MLP is helpful in outputting reconstruction results that match the images in Fig. 10 , especially on the joints of the hands and feet.

Impact of configurations in MHCA. As described in Sec. 3.4, the common configuration tends to lead to inadequate communication between hypotheses and information transfer being trapped in localized areas. We adopt a more efficient configuration by using different inputs among queries, keys, and values. We can see from Tab. 4 that using the same input between keys and values in MHCA (*i.e.*, with MHCA-Conf) requires more parameters but cannot bring further performance gains. It illustrates the effectiveness of our efficient strategy in MHCA.

TABLE 3 Ablation study on HM-MLP.

	MPJPE↓	PA-MPJPE↓
Ours (w/ MLP)	54.1	37.7
Ours (w/ HM-MLP)	**53.6**	**37.4**

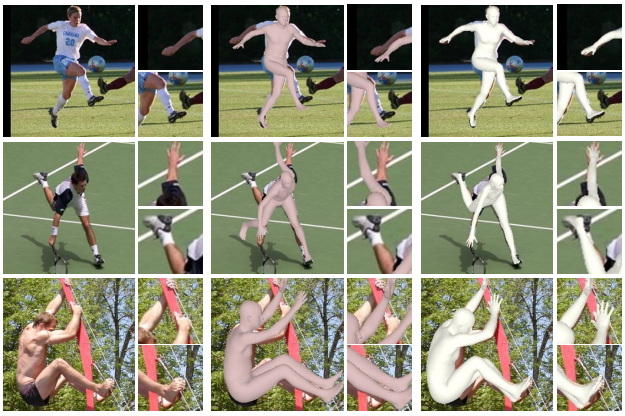


Image w/o HM-MLP Ours

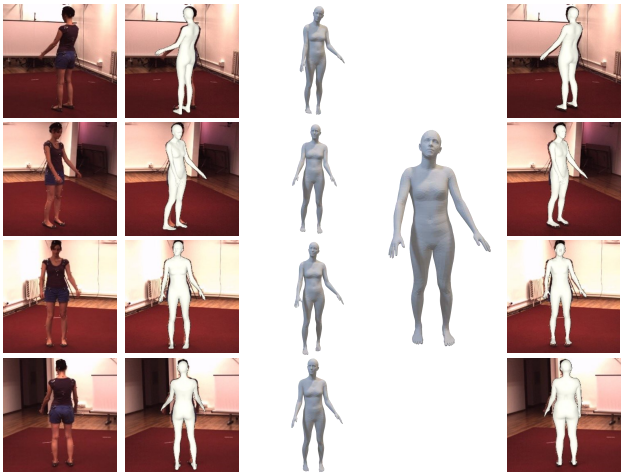
FIGURE 10 Ablation study on HM-MLP.

TABLE 4 Ablation study on different configurations in MHCA.

	Params(M)↓	MPJPE↓
Ours (w/ MHCA-Conf)	25.32	54.3
Ours (w/ MHCA)	**20.58**	**53.6**

TABLE 5 Quantitative comparison with the state-of-the-art methods on Human3.6M [45] for the multi-view fusion task.

Method	MPJPE↓	PA-MPJPE↓
Liang *et al.* [57]	79.8	45.1
Li *et al.* [58]	64.8	43.8
ProHMR [11]	62.2	34.5
Ours	**53.8**	**32.7**



Input Images Reconstructions (Camera view) Reconstructions (Canonical view) Multi-view Fusion After Fusion (Camera view)

FIGURE 11 Recovery results for the multi-view fusion task.

5.5 | Application: Multi-view Fusion

We also perform quantitative and qualitative evaluations to validate the effectiveness of MH-HMR for the multi-view fusion task. We present quantitative comparison results on Human3.6M [45] dataset in Tab. 5 . Compared to Li *et al.* [58] and ProHMR [11], our approach outperforms them in both MPJPE and PA-MPJPE. In Fig. 11 , we show that the refinement and communication modules based on our multi-hypotheses can be used to obtain more accurate mesh recovery by fusing information from multiple views. As shown in Fig. 11 , problems such as the upper limbs in the first view being largely occluded and the body in the other views having depth ambiguity, result in a less accurate recovered mesh. However, with the fusion of multiple views, the recovered mesh captures the real and natural pose and shape more faithfully.

6 | CONCLUSION

This paper presents MH-HMR, a novel multi-hypothesis approach that addresses the inverse problem of human mesh recovery from a monocular image by leveraging differential feature representations learned from image information and a series of feature enhancements to hypotheses, resulting in better accuracy and enhanced robustness. Unlike existing multi-hypothesis methods, we first employ a probabilistic model to generate multiple initial hypotheses, and further propose two transformer-based refinement and communication modules to establish information transfer and strong relationships among the hypotheses. Meanwhile, benefiting from the multi-hypothesis properties and our module designs, we demonstrate the effectiveness of our model in the multi-view fusion downstream task. We conduct extensive comparative experiments to demonstrate that MH-HMR achieves superior performance and can better handle challenging images, together with detailed ablation studies showing that each design contributes to our performance on the benchmark datasets.

Future work could consider continually extending and incorporating MH-HMR with recent progress to better exploit multi-hypothesis relationships and promote recovery accuracy while considering various ambiguities.

Acknowledgements

This work was supported in part by National Key R&D Program of China (2023YFC3082100), National Natural Science Foundation of China (62122058 and 62171317), and Science Fund for Distinguished Young Scholars of Tianjin (22JCJQJC00040).

References

1. Xing Y, Zhu J. Deep learning-based action recognition with 3D skeleton: a survey. *CAAI Transactions on Intelligence Technology.* 2021; 6: 80–92.

2. Zhang J, Ye G, Tu Z, et al. A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. *CAAI Transactions on Intelligence Technology.* 2022; 7(1): 46–55.

3. Duan H, Zhao Y, Chen K, Lin D, Dai B. Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2969–2978.

4. Liu Y, Sivaparthipan C, Shankar A. Human–computer interaction based visual feedback system for augmentative and alternative communication. *International Journal of Speech Technology.* 2022: 1–10.

5. Weng CY, Curless B, Kemelmacher-Shlizerman I. Photo wake-up: 3D character animation from a single photo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5908–5917.

6. Khirodkar R, Tripathi S, Kitani K. Occluded human mesh recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 1715–1725.

7. Kolotouros N, Pavlakos G, Black MJ, Daniilidis K. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 2252–2261.

8. Zhang H, Tian Y, Zhou X, et al. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11446–11456.

9. Kocabas M, Huang CHP, Hilliges O, Black MJ. PARE: Part attention regressor for 3D human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11127–11137.

10. Wehrbein T, Rudolph M, Rosenhahn B, Wandt B. Probabilistic monocular 3D human pose estimation with normalizing flows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11199–11208.

11. Kolotouros N, Pavlakos G, Jayaraman D, Daniilidis K. Probabilistic modeling for human mesh recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11605–11614.

12. Sengupta A, Budvytis I, Cipolla R. Hierarchical kinematic probability distributions for 3D human shape and pose estimation from images in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11219–11229.

13. Li W, Liu H, Tang H, Wang P, Van Gool L. MHFormer: Multi-hypothesis transformer for 3D human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13147–13156.

14. Xuan H, Zhang J, Li K. MHPro: Multi-hypothesis Probabilistic Modeling for Human Mesh Recovery. In: Artificial Intelligence: Second CAAI International Conference, CICAI 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part I. Springer. 2022: 216–228.

15. Zheng C, Wu W, Chen C, et al. Deep learning-based human pose estimation: A survey. *arXiv preprint arXiv:2012.13392.* 2020.

16. Tian Y, Zhang H, Liu Y, Wang L. Recovering 3D human mesh from monocular images: A survey. *arXiv preprint arXiv:2203.01923.* 2022.

17. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics.* 2015; 34(6): 1–16.

18. Bogo F, Kanazawa A, Lassner C, Gehler P, Romero J, Black MJ. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Proceedings of the European Conference on Computer Vision. Springer. 2016: 561–578.

19. Lassner C, Romero J, Kiefel M, Bogo F, Black MJ, Gehler PV. Unite the people: Closing the loop between 3D and 2D human representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6050–6059.

20. Kanazawa A, Black MJ, Jacobs DW, Malik J. End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7122–7131.

21. Pavlakos G, Zhu L, Zhou X, Daniilidis K. 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. 2018: 459–468.

22. Moon G, Lee KM. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single rgb image. In: Proceedings of the European Conference on Computer Vision. Springer. 2020: 752–768.

23. Kocabas M, Athanasiou N, Black MJ. VIBE: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5253–5263.

24. Lee GH, Lee SW. Uncertainty-aware human mesh recovery from video by learning part-based 3D dynamics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 12375–12384.

25. Wan Z, Li Z, Tian M, Liu J, Yi S, Li H. Encoder-decoder with multi-level attention for 3D human shape and pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 13033–13042.

26. Tu Z, Huang Z, Chen Y, et al. Consistent 3D hand reconstruction in video via self-supervised learning. In: Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023.

27. Li C, Lee GH. 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. 2019: 9887–9895.

28. Biggs B, Novotny D, Ehrhardt S, Joo H, Graham B, Vedaldi A. 3D multi-bodies: Fitting sets of plausible 3D human models to ambiguous image data. *Advances in Neural Information Processing Systems.* 2020; 33: 20496–20507.

29. Zheng X, Zheng Y, Yang S. Generating multiple hypotheses for 3D human mesh and pose using conditional generative adversarial nets. In: Proceedings of the Asian Conference on Computer Vision. 2022: 1709–1725.

30. Holmquist K, Wandt B. DiffPose: Multi-hypothesis human pose estimation using diffusion models. *arXiv preprint arXiv:2211.16487.* 2022.

31. Sengupta A, Budvytis I, Cipolla R. Synthetic training for accurate 3D human pose and shape estimation in the wild. In: 2020.

32. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in Neural Information Processing Systems.* 2017; 30.

33. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929.* 2020.

34. Lin K, Wang L, Liu Z. End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1954–1963.

35. Yuan Y, Iqbal U, Molchanov P, Kitani K, Kautz J. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11038–11049.

36. Jiang Y, Chang S, Wang Z. TransGAN: Two pure transformers can make one strong GAN, and that can scale up. *Advances in Neural Information Processing Systems.* 2021; 34: 14745–14758.

37. Li M, Fu Y, Zhang Y. Spatial-spectral transformer for hyperspectral image denoising. In: Proceedings of AAAI Conference on Artificial Intelligence. 2023.

38. Dai Z, Cai B, Lin Y, Chen J. UP-DETR: Unsupervised pre-training for object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1601–1610.

39. Zeng Y, Fu J, Chao H. Learning joint spatial-temporal transformations for video inpainting. In: Proceedings of the European Conference on Computer Vision. Springer. 2020: 528–543.

40. Li Y, Si S, Li G, Hsieh CJ, Bengio S. Learnable fourier features for multi-dimensional spatial positional encoding. *Advances in Neural Information Processing Systems.* 2021; 34: 15816–15829.

41. Chen CFR, Fan Q, Panda R. CrossVIT: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 357–366.

42. Wei X, Zhang T, Li Y, Zhang Y, Wu F. Multi-modality cross attention network for image and sentence matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 10941–10950.

43. Hou R, Chang H, Ma B, Shan S, Chen X. Cross attention network for few-shot classification. *Advances in Neural Information Processing Systems.* 2019; 32.

44. Zhou Y, Barnes C, Lu J, Yang J, Li H. On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5745–5753.

45. Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2013; 36(7): 1325–1339.

46. Mehta D, Rhodin H, Casas D, et al. Monocular 3D human pose estimation in the wild using improved cnn supervision. In: 2017 International Conference on 3D Vision (3DV). IEEE. 2017: 506–516.

47. Von Marcard T, Henschel R, Black MJ, Rosenhahn B, Pons-Moll G. Recovering accurate 3D human pose in the wild using imus and a moving camera. In: Proceedings of the European Conference on Computer Vision. 2018: 601–617.

48. Johnson S, Everingham M. Clustered pose and nonlinear appearance models for human pose estimation.. In: British Machine Vision Conference. 2010: 5.

49. Andriluka M, Pishchulin L, Gehler P, Schiele B. 2D human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 3686–3693.

50. Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context. In: Proceedings of the European Conference on Computer Vision. Springer. 2014: 740–755.

51. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770–778.

52. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE. 2009: 248–255.

53. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.* 2014.

54. Choi H, Moon G, Chang JY, Lee KM. Beyond static features for temporally consistent 3D human pose and shape from a video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1964–1973.

55. Sun Y, Bao Q, Liu W, Fu Y, Black MJ, Mei T. Monocular, one-stage, regression of multiple 3D people. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 11179–11188.

56. Zanfir M, Zanfir A, Bazavan EG, Freeman WT, Sukthankar R, Sminchisescu C. THUNDR: Transformer-based 3D human reconstruction with markers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 12971–12980.

57. Liang J, Lin MC. Shape-aware human pose and shape reconstruction using multi-view images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4352–4362.

58. Li Z, Oskarsson M, Heyden A. 3D human pose and shape estimation through collaborative learning and multi-view model-fitting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 1888–1897.