



ORIGINAL RESEARCH

MH-HMR: Human mesh recovery from monocular images via multi-hypothesis learning

Haibiao Xuan¹  | Jinsong Zhang¹ | Yu-Kun Lai² | Kun Li¹ 

¹College of Intelligence and Computing, Tianjin University, Tianjin, China

²School of Computer Science and Informatics, Cardiff University, Cardiff, UK

Correspondence

Kun Li, College of Intelligence and Computing, Tianjin University, No. 135 Yaguan Road, Tianjin, China.

Email: lik@tju.edu.cn

Yu-Kun Lai, School of Computer Science and Informatics, Cardiff University, Senghennydd Road, Cardiff CF24 4AG, UK.

Email: Yukun.Lai@cs.cardiff.ac.uk

Funding information

National Key Research and Development Program of China, Grant/Award Number: 2023YFC3082100; National Natural Science Foundation of China, Grant/Award Numbers: 62122058, 62171317; Science Fund for Distinguished Young Scholars of Tianjin Municipality, Grant/Award Number: 22JCJQC00040

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, the ambiguity of the reconstruction is embraced and the problem is considered as an inverse problem for which multiple feasible solutions exist. To address these issues, the authors propose a multi-hypothesis approach, multi-hypothesis human mesh recovery (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 colour image; (2) modelling intra-hypothesis refinement to enhance every single-hypothesis feature; and (3) establishing inter-hypothesis communication and regressing the final human meshes. Meanwhile, the authors take further advantage of multiple hypotheses and the recovery process to achieve human mesh recovery from multiple uncalibrated views. Compared with state-of-the-art methods, the MH-HMR approach achieves superior performance and recovers more accurate human meshes on challenging benchmark datasets, such as Human3.6M and 3DPW, while demonstrating the effectiveness across a variety of settings. The code will be publicly available at <https://cic.tju.edu.cn/faculty/likun/projects/MH-HMR>.

KEYWORDS

3-D, computer vision, human reconstruction

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 [1–3], 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 [6–9] 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 recognise the ill-posedness and uncertainty of this problem, and successively propose to estimate probability distributions or explicitly generate multi-hypotheses [10–13]. 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

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

significantly affect the performance and expressiveness of the model.

Motivated by the above observation, we propose a novel multi-hypothesis approach, multi-hypothesis human mesh recovery (**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, a 3D human evidence is initially extracted from the monocular image by a probabilistic model based on normalising flow, and then fed into a feasible pose distribution regressor to obtain multiple initial hypotheses as shown in Figure 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 (MLP) to explore the relationship between channels with different hypotheses, and a new configuration of the Multi-Head Cross-Attention (MHCA) 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. Figure 1 gives an example. The code will be publicly available for research purposes.

Our contributions can be summarised 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-hypotheses.
- Our MH-HMR achieves superior performance on challenging benchmark datasets, such as Human3.6M and 3DPW, even for the cases with depth ambiguity, joint occlusion, and truncation.

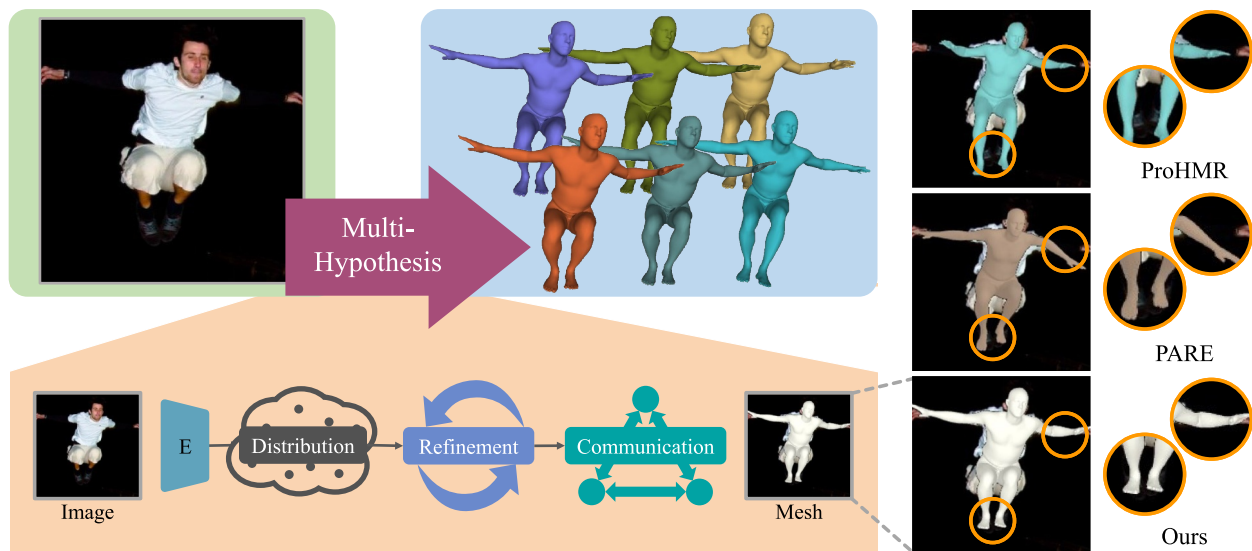


FIGURE 1 We propose, MH-HMR, to accurately recover a 3D human mesh given an input image. Right: results of the probabilistic method ProHMR [11], the state-of-the-art method PARE [9] and our approach for a challenging image. MH-HMR, multi-hypothesis human mesh recovery.

- 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 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 optimisation-based methods estimate the parameters of the body model iteratively, such that it is consistent with a set of features, such as 2D keypoints, silhouettes and part segmentation. For example, Bogo et al. [18] propose SMPLify, a multi-stage optimisation method that iteratively fits the SMPL model with 2D keypoints and minimises the reprojection error to estimate a 3D human mesh. Lassner et al. [19] employ silhouettes together with 2D keypoints in the optimisation procedure. Despite the well-aligned results can be obtained, these methods are sensitive to initialisation, require additional data, and suffer from time-consuming fitting and inefficient inference. In contrast, taking advantage of the powerful non-linear mapping capability of neural networks, regression-based methods [7–9, 20–26] 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 utilises the unpaired 3D annotations and penalises 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 [10–13, 27–30] 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, Kolotouros et al. [11] model the conditional probability distribution using conditional normalising 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.

Different 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 field. Motivated by the achievements, various works start to apply transformer equipped with a powerful Multi-Head Self-Attention (MHSA) mechanism to the computer vision tasks. Vision Transformer [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.

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 Figure 2. Our approach, MH-HMR, consists of three steps: (1) probabilistic modelling and initial hypothesis generation (Section 3.2); (2) intra-hypothesis refinement (Section 3.3); and (3) inter-hypothesis communication (Section 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 $M \in \mathbb{R}^{6890 \times 3}$. 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 $\mathcal{J} \in \mathbb{R}^{K \times 6890}$ 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 stabilisation in propagation. Here, we briefly describe MHSA and 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 follows:

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

Then, MHSA splits the queries Q , keys K , and values V into b different subspaces as well as performs the attention in parallel. Finally, the outputs from the b 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 non-linear activation function between them), which are used for non-linearity and feature transformation:

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

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 modelling

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 g and obtain image features f_I . Then, the Conditional Normalising Flow is applied to model the probability distribution of the human pose $p_{\Theta|I}(\theta|f_I = g(I))$, due to their expressiveness and modelling capabilities. In contrast to

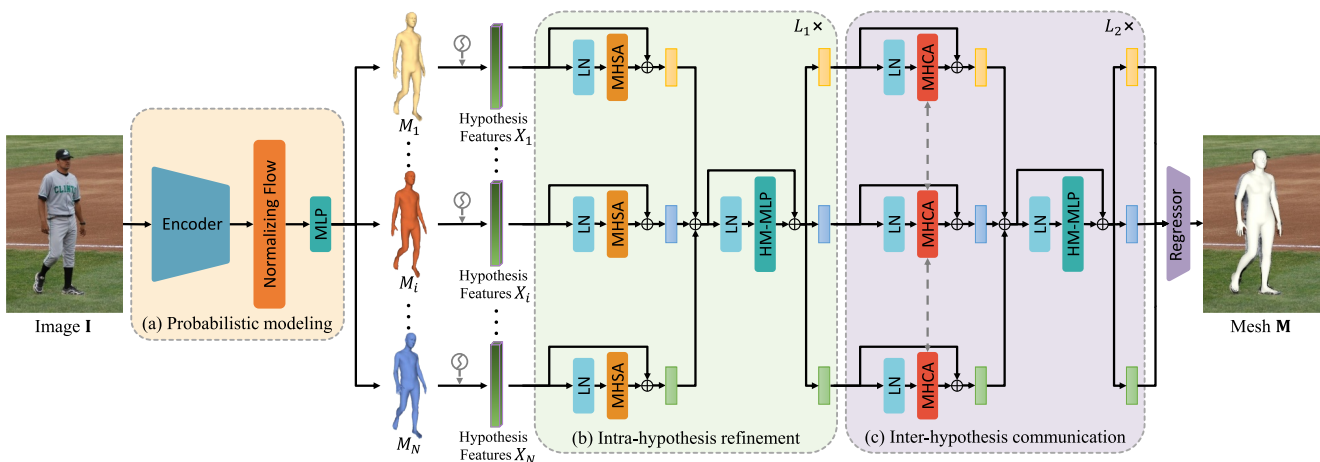


FIGURE 2 Overview of the proposed approach. Given an input monocular image I , we perform probabilistic modelling (a) with normalising flows to extract image features, predict a pose distribution and generate multiple initial human mesh hypotheses (where N indicates the number of hypotheses), input these 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 .

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 focussing on one-to-many mappings.

The normalising 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 three basic transformations:

$$f_i = f_{AC} \circ f_{LT} \circ f_{IN}, \quad (3)$$

where $f_{IN}(z) = \mathbf{a} \odot z + \mathbf{b}$ (Instance Normalisation), $f_{LT}(z) = Wz + \mathbf{b}$ (Linear Transformation) and $f_{AC} = [z_{1:k}, z_{k+1:d} + \mathbf{t}(z_{1:d}, \mathbf{c})]$ (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 $\theta_i^* = \operatorname{argmax}_{\theta} p_{\Theta|f_i}(\theta | f_i)$, we sample the distribution to select N hypotheses with larger probabilities. The samples $\{\theta_i\}_1^N$ drawn from the output distribution are as follows:

$$\theta_i \sim p_{\Theta|I}(\theta | f_i). \quad (4)$$

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:

$$[\beta_i, \pi_i] = \operatorname{MLP}(f_i, \theta_i). \quad (5)$$

In summary, the probabilistic model based on the normalising 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 ref. [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 Figure 3), which can be represented as follows:

$$X_i^l = X_i^{l-1} + \operatorname{MHSA}\left(\operatorname{LN}\left(X_i^{l-1}\right)\right), \quad (6)$$

where $\operatorname{LN}(\cdot)$ is the LayerNorm layer, and $l \in [1, 2, \dots, L_1]$ 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 follows:

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

where $X_{concat}^l \in \mathbb{R}^{C \times N}$, and $\operatorname{Concat}(\cdot)$ and $\operatorname{Diverge}(\cdot)$ are concatenation and division operations, respectively. $\operatorname{HM-MLP}(\cdot)$ is the function of HM-MLP modified for the hypothetical features (as shown in Figure 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 refs. [41–43] and apply multiple MHCA components in parallel. Note that

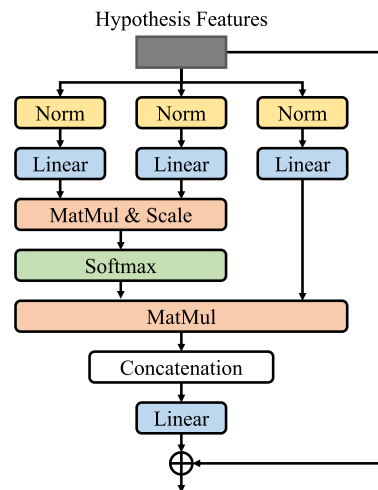


FIGURE 3 Multi-head self-attention.

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 same input between keys and values. However, this configuration tends to lead to inadequate communication between hypotheses and information transfer being trapped in localised 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 Figure 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:

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

where Y_{i_1} and Y_{i_2} are the other two corresponding hypotheses, $l \in [1, 2, \dots, L_2]$ is the index of L_2 *Inter-hypothesis communication* modules, and $Y_i^0 = X_i^{L_1}$. Finally, MHCA passes in-

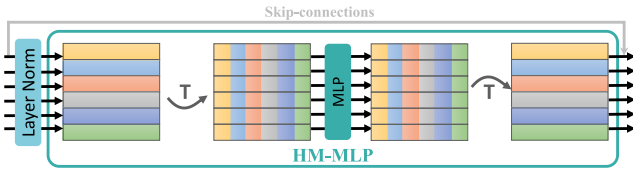


FIGURE 4 Hypothesis-mixing MLP.

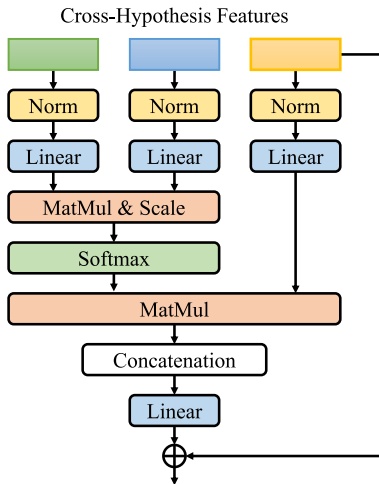


FIGURE 5 Multi-Head Cross-Attention.

formation among hypotheses in a crossing way to significantly enhance feature representation and modelling capabilities.

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

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

where $Y_{concat}^l \in \mathbb{R}^{C \times N}$, and $\text{Concat}(\cdot)$ and $\text{Diverge}(\cdot)$ 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.

3.5.1 | NLL loss

As with typical probabilistic models, our normalising flow models are trained to minimise the negative log-likelihood of the ground truth θ_{gt} , that is, the loss function is as follows:

$$\mathcal{L}_{nll} = -\ln p_{\Theta|I}(\theta_{gt} | f_I). \quad (10)$$

3.5.2 | 2D loss

To penalise 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:

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

3.5.3 | 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:

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

3.5.4 | 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 follows:

$$\mathcal{L} = \lambda_{\text{null}}\mathcal{L}_{\text{null}} + \lambda_{2D}\mathcal{L}_{2D} + \lambda_{3D}\mathcal{L}_{3D} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}}, \quad (13)$$

where λ_{null} , 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 utilise 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 modelling, 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 modelling* (in Section 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 into the *Intra-hypothesis refinement* module (in Section 3.3) and the *Inter-hypothesis communication* module (in Section 3.4) in parallel, allowing the exchanges and fusion of image features in different views. Figure 6 shows the overview of our proposed approach for the multi-view fusion task.

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

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

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

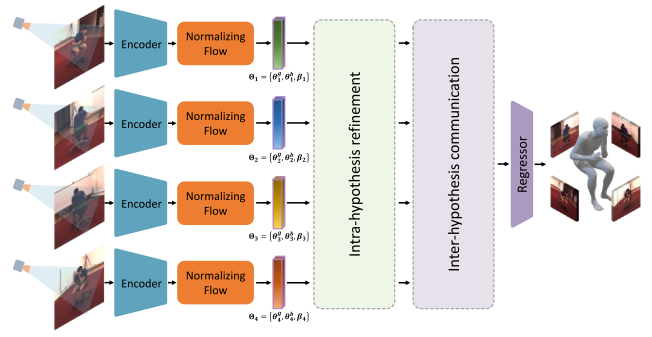


FIGURE 6 Our pipeline for the multiple view fusion task.

5 | EXPERIMENTS

5.1 | Datasets and metrics

5.1.1 | Training

Following previous works [7, 20], 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].

5.1.2 | Evaluation

We report the experiment 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 (PAM-JPE), 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 optimiser [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: $\lambda_{\text{null}} = 0.001$, $\lambda_{2D} = 0.01$, $\lambda_{3D} = 0.05$, and $\lambda_{\text{orth}} = 0.1$. For the multi-view fusion task, we set λ_{θ} to 0.001 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.

5.3 | Comparison

We qualitatively and quantitatively compare our approach with the state-of-the-art (SOTA) temporal and frame-based methods,

TABLE 1 Quantitative comparison with the state-of-the-art temporal and frame-based methods on Human3.6M [45] and 3DPW [47] datasets.

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

Note: The best results are highlighted in bold and ‘-’ represents that the results are not available.

Abbreviations: MPJPE, Mean Per Joint Position Error; PA-MPJPE, Procrustes-Aligned Mean Per Joint Position Error; PVE, Per Vertex Error.

including MAED [25], SPIN [7], ProHMR [11], PyMAF [8], and PARE [9].

We present quantitative comparison results on Human3.6M and 3DPW datasets in Table 1. Our MH-HMR achieves competitive or superior results compared with previous approaches. The methods reported in Table 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 Table 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 and 4.7 mm on Human3.6M and 3DPW datasets, respectively. From Table 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 SOTA temporal method MAED [25], despite the fact that our approach is frame-based.

Recovery results on the LSP [48] dataset are depicted in Figure 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 Table 1, compared to the conference version, we reduce the MPJPE by 1.2 and 1.5 mm on Human3.6M and 3DPW datasets, respectively. In addition to this, qualitative results are shown in Figure 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 Figure 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.

5.4.1 | 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

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



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.

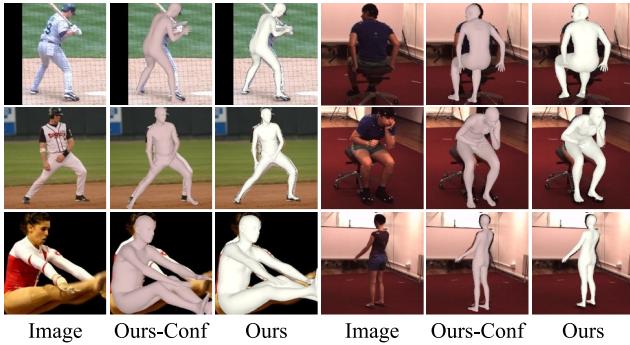


FIGURE 8 Qualitative results on LSP [48] dataset and Human3.6M [45] dataset. From left to right shows the input images, and the results of Ours-conf and Ours.

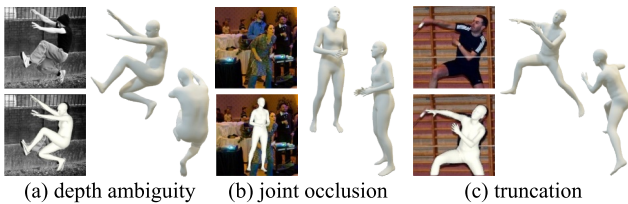


FIGURE 9 Plausible human mesh recovery results generated by our approach, especially for ambiguous parts with depth ambiguity, joint occlusion, and truncation.

probabilistic modelling in Table 2a. Experiments show that generating more hypotheses improves performance with a small increase in parameters, but becomes worse instead for

TABLE 2 Ablation study on different parameters of our model.

	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

Note: 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. And the best results are highlighted in bold.

Abbreviations: MPJPE, Mean Per Joint Position Error; PA-MPJPE, Procrustes-Aligned Mean Per Joint Position Error.

more than eight hypotheses. Therefore, in our main experiments, we choose to use eight 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.

5.4.2 | Number of layers of two modules

Table 2B,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 two 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 $L_2 = 2$.

5.4.3 | Impact of HM-MLP

Reasonable hypothesis fusion settings help to fully utilise 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 Table 3, when applying HM-MLP, the errors are reduced by 0.5 and 0.3 mm in MPJPE and PA-MPJPE, respectively. Meanwhile, HM-MLP is helpful in outputting reconstruction results that match the images in Figure 10, especially on the joints of the hands and feet.

5.4.4 | Impact of configurations in MHCA

As described in Section 3.4, the common configuration tends to lead to inadequate communication between hypotheses and information transfer being trapped in localised areas. We adopt a more efficient configuration by using different inputs among queries, keys, and values. We can see from Table 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.

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 Table 5. Compared to Li et al. [58] and ProHMR [11], our approach outperforms them in both

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

Note: The best results are highlighted in bold.

Abbreviations: HM-MLP, Hypothesis-Mixing Multi-Layer Perceptron; MLP, Multi-Layer Perceptron; MPJPE, Mean Per Joint Position Error; PA-MPJPE, Procrustes-Aligned Mean Per Joint Position Error.

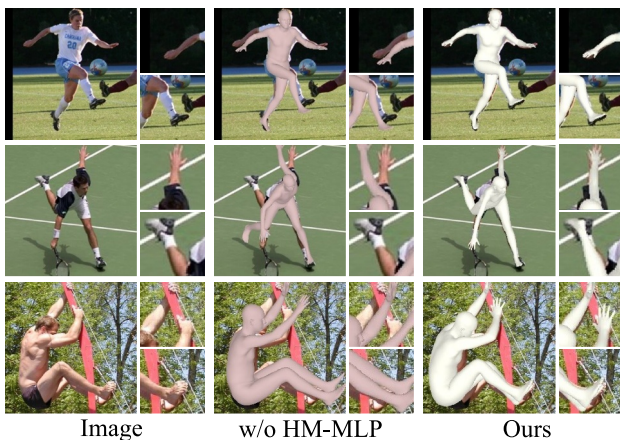


FIGURE 10 Ablation study on Hypothesis-Mixing MLP (HM-MLP).

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

Note: The best results are highlighted in bold.

Abbreviations: MHCA, Multi-Head Cross-Attention; MPJPE, Mean Per Joint Position Error.

MPJPE and PA-MPJPE. In Figure 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 Figure 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 learnt 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

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

Note: The best results are highlighted in bold.

Abbreviations: MPJPE, Mean Per Joint Position Error; PA-MPJPE, Procrustes-Aligned Mean Per Joint Position Error.

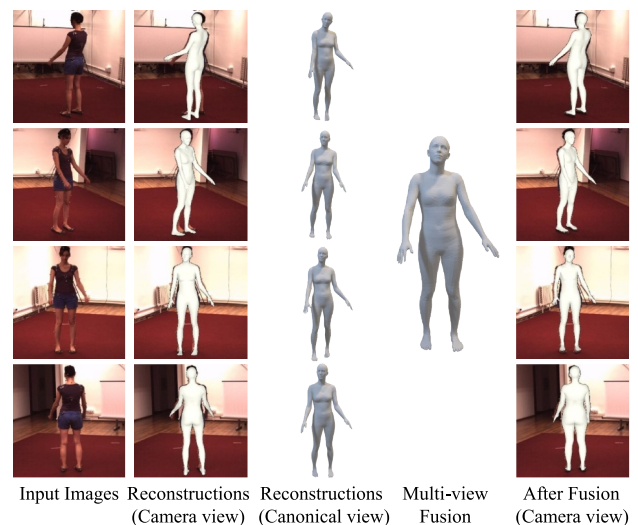


FIGURE 11 Recovery results for the multi-view fusion task.

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.

ACKNOWLEDGEMENT

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

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data subject to third party restrictions. The data that support the findings of this study are available from University of Pennsylvania, Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V., University of Leeds, and Microsoft. Restrictions apply to the availability of these data, which were used under licence for this study. Data are available at <http://vision.imar.ro/human3.6m/>, <https://vcai.mpi-inf.mpg.de/3dhp-dataset/>, <https://virtualhumans.mpi-inf.mpg.de/3DPW/>, <http://sam.johnson.io/research/lsp.html>, <http://human-pose.mpi-inf.mpg.de/>, and <https://cocodataset.org/>, with the permission of University of Pennsylvania, Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V., University of Leeds, and Microsoft.

ORCID

Haibiao Xuan  <https://orcid.org/0000-0002-3367-2773>

Kun Li  <https://orcid.org/0000-0003-2326-0166>

REFERENCES

- Xing, Y., Zhu, J.: Deep learning-based action recognition with 3D skeleton: a survey. *CAAI Trans. Intell. Technol.* 6(1), 80–92 (2021). <https://doi.org/10.1049/cit2.12014>
- Zhang, J., et al.: A spatial attentive and temporal dilated (SATD) GCN for skeleton-based action recognition. *CAAI Trans. Intell. Technol.* 7(1), 46–55 (2022). <https://doi.org/10.1049/cit2.12012>
- Duan, H., et al.: Revisiting skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2969–2978. IEEE, New Orleans (2022)
- Liu, Y., Sivaparthipan, C., Shankar, A.: Human–computer interaction based visual feedback system for augmentative and alternative communication. *Int. J. Speech Technol.* 25, 305–314 (2022)
- Weng, C.Y., 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, pp. 5908–5917. IEEE, Long Beach (2019)
- Khirodkar, R., Tripathi, S., Kitani, K.: Occluded human mesh recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1715–1725. IEEE, New Orleans (2022)
- Kolotouros, N., et al.: 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, pp. 2252–2261. IEEE, Seoul (2019)
- Zhang, H., 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, pp. 11446–11456. IEEE, Montreal (2021)
- Kocabas, M., et al.: PARE: Part attention regressor for 3D human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11127–11137. IEEE, Montreal (2021)
- Wehrbein, T., et al.: Probabilistic monocular 3D human pose estimation with normalizing flows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11199–11208. IEEE, Montreal (2021)
- Kolotouros, N., et al.: Probabilistic modeling for human mesh recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11605–11614. IEEE, Montreal (2021)
- 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, pp. 11219–11229. IEEE, Montreal (2021)
- Li, W., et al.: MHFormer: multi-hypothesis transformer for 3D human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13147–13156. IEEE, New Orleans (2022)
- Xuan, H., Zhang, J., Li, K.: MHPro: multi-hypothesis probabilistic modeling for human mesh recovery. In: Artificial Intelligence: Second CAAI International Conference, CICA 2022, Beijing, China, August 27–28, 2022, Revised Selected Papers, Part I, pp. 216–228. Springer, Beijing (2022)
- Zheng, C., et al.: Deep learning-based human pose estimation: a survey. arXiv preprint arXiv:2012.13392 (2020)
- Tian, Y., et al.: Recovering 3D human mesh from monocular images: a survey. arXiv preprint arXiv:2203.01923 (2022)
- Loper, M., et al.: SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34(6), 1–16 (2015). <https://doi.org/10.1145/2816795.2818013>
- Bogo, F., et al.: Keep it SMPL: automatic estimation of 3D human pose and shape from a single image. In: Proceedings of the European Conference on Computer Vision, pp. 561–578. Springer, Amsterdam (2016)
- Lassner, C., et al.: Unite the people: closing the loop between 3D and 2D human representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6050–6059. IEEE, Honolulu (2017)
- Kanazawa, A., et al.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7122–7131. IEEE, Salt Lake City (2018)
- Pavlakos, G., et al.: 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, pp. 459–468. IEEE, Salt Lake City (2018)
- 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: Proceedings of the European Conference on Computer Vision, pp. 752–768. Springer, Glasgow (2020)
- Kocabas, M., Athanasiou, N., Black, M.J.: VIBE: video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5253–5263. IEEE, Seattle (2020)
- Lee, G.H., Lee, S.W.: Uncertainty-aware human mesh recovery from video by learning part-based 3D dynamics. In: Proceedings of the IEEE/

- CVF International Conference on Computer Vision, pp. 12375–12384. IEEE, Montreal (2021)
25. Wan, Z., et al.: Encoder-decoder with multi-level attention for 3D human shape and pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13033–13042. IEEE, Montreal (2021)
 26. Tu, Z., et al.: Consistent 3D hand reconstruction in video via self-supervised learning. *Proc. IEEE Trans. Pattern Anal. Mach. Intell.* 45, 9469–9485 (2023)
 27. Li, C., Lee, G.H.: 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, pp. 9887–9895. IEEE, Long Beach (2019)
 28. Biggs, B., et al.: 3D multi-bodies: fitting sets of plausible 3D human models to ambiguous image data. *Adv. Neural Inf. Process. Syst.* 33, 20496–20507 (2020)
 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, pp. 1709–1725. Springer, Macao (2022)
 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. *arXiv preprint arXiv:2009.10013*. (2020)
 32. Vaswani, A., et al.: Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30, 6000–6010 (2017)
 33. Dosovitskiy, A., et al.: An image is worth 16 × 16 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, pp. 1954–1963. IEEE, Nashville (2021)
 35. Yuan, Y., et al.: GLAMR: global occlusion-aware human mesh recovery with dynamic cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11038–11049. IEEE, New Orleans (2022)
 36. Jiang, Y., Chang, S., Wang, Z.: TransGAN: two pure transformers can make one strong GAN, and that can scale up. *Adv. Neural Inf. Process. Syst.* 34, 14745–14758 (2021)
 37. Li, M., Fu, Y., Zhang, Y.: Spatial-spectral transformer for hyperspectral image denoising. In: Proceedings of AAAI Conference on Artificial Intelligence. AAAI, Washington (2023)
 38. Dai, Z., et al.: UP-DETR: unsupervised pre-training for object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1601–1610. IEEE, Nashville (2021)
 39. Zeng, Y., Fu, J., Chao, H.: Learning joint spatial-temporal transformations for video inpainting. In: Proceedings of the European Conference on Computer Vision, pp. 528–543. Springer, Glasgow (2020)
 40. Li, Y., et al.: Learnable Fourier features for multi-dimensional spatial positional encoding. *Adv. Neural Inf. Process. Syst.* 34, 15816–15829 (2021)
 41. Chen, C.F.R., 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, pp. 357–366. IEEE, Montreal (2021)
 42. Wei, X., et al.: Multi-modality cross attention network for image and sentence matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10941–10950. IEEE, Seattle (2020)
 43. Hou, R., et al.: Cross attention network for few-shot classification. *Adv. Neural Inf. Process. Syst.* 32, 4003–4014 (2019)
 44. Zhou, Y., et al.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5745–5753. IEEE, Long Beach (2019)
 45. Ionescu, C., et al.: Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* 36(7), 1325–1339 (2013). <https://doi.org/10.1109/tpami.2013.248>
 46. Mehta, D., et al.: Monocular 3D human pose estimation in the wild using improved CNN supervision. In: 2017 International Conference on 3D Vision (3DV), pp. 506–516. IEEE, Qingdao (2017)
 47. Von Marcard, T., et al.: Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In: Proceedings of the European Conference on Computer Vision, pp. 601–617. Springer, Munich (2018)
 48. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: British Machine Vision Conference, pp. 5. British Machine Vision Association, Aberystwyth (2010)
 49. Andriluka, M., et al.: 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3686–3693. IEEE, Columbus (2014)
 50. Lin, T.Y., et al.: Microsoft COCO: common objects in context. In: Proceedings of the European Conference on Computer Vision, pp. 740–755. Springer, Zurich (2014)
 51. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE, Las Vegas (2016)
 52. Deng, J., et al.: ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE, Miami (2009)
 53. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
 54. Choi, H., et al.: 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, pp. 1964–1973. IEEE, Nashville (2021)
 55. Sun, Y., et al.: Monocular, one-stage, regression of multiple 3D people. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11179–11188. IEEE, Montreal (2021)
 56. Zanfir, M., et al.: THUNDR: transformer-based 3D human reconstruction with markers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12971–12980. IEEE, Montreal (2021)
 57. Liang, J., Lin, M.C.: Shape-aware human pose and shape reconstruction using multi-view images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4352–4362. IEEE, Seoul (2019)
 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, pp. 1888–1897. IEEE, Waikoloa (2021)

How to cite this article: Xuan, H., et al.: MH-HMR: human mesh recovery from monocular images via multi-hypothesis learning. *CAAI Trans. Intell. Technol.* 1–12 (2024). <https://doi.org/10.1049/cit2.12337>