ScoreHypo: Probabilistic Human Mesh Estimation with Hypothesis Scoring

Yuan Xu[1] Xiaoxuan Ma[1] Jiajun Su[5] Wentao Zhu[1] Yu Qiao[6*] Yizhou Wang[1, 2, 3, 4*]

[1] Center on Frontiers of Computing Studies, School of Computer Science, Peking University
[2] Inst. for Artificial Intelligence, Peking University [3] Nat'l Eng. Research Center of Visual Technology
[4] Nat'l Key Lab of General Artificial Intelligence [5] International Digital Economy Academy (IDEA)
[6] School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University

xuyuan@stu.pku.edu.cn, {maxiaoxuan, wtzhu, yizhou.wang}@pku.edu.cn,
sujiajun@idea.edu.cn, qiaoyu@sjtu.edu.cn

Abstract

Monocular 3D human mesh estimation is an ill-posed problem, characterized by inherent ambiguity and occlusion. While recent probabilistic methods propose generating multiple solutions, little attention is paid to obtaining high-quality estimates from them. To address this limitation, we introduce ***ScoreHypo****, a versatile framework by first leveraging our novel* ***HypoNet*** *to generate multiple hypotheses, followed by employing a meticulously designed scorer,* ***ScoreNet****, to evaluate and select high-quality estimates. ScoreHypo formulates the estimation process as a reverse denoising process, where HypoNet produces a diverse set of plausible estimates that effectively align with the image cues. Subsequently, ScoreNet is employed to rigorously evaluate and rank these estimates based on their quality and finally identify superior ones. Experimental results demonstrate that HypoNet outperforms existing state-of-the-art probabilistic methods as a multi-hypothesis mesh estimator. Moreover, the estimates selected by ScoreNet significantly outperform random generation or simple averaging. Notably, the trained ScoreNet exhibits generalizability, as it can effectively score existing methods and significantly reduce their errors by more than* 15%*. Code and models are available at* *https://xy02-05.github.io/ScoreHypo*.

1. Introduction

Recovering 3D human mesh from a single 2D image presents a fundamental and challenging problem in various human-centered applications, such as motion analysis [1, 14] and avatar animation [68, 70, 79, 81]. Recent advancements in this field have primarily focused on enhancing the accuracy of producing a single deterministic estimate [8, 9, 22, 39, 46, 48, 76, 80]. Nonetheless, the process of mapping 2D to 3D inherently suffers from several issues including depth ambiguity and occlusion, which are prevalent in real-world environments. As a result, exploring multiple feasible solutions becomes a more appropriate and effective strategy for addressing the monocular mesh estimation challenge.

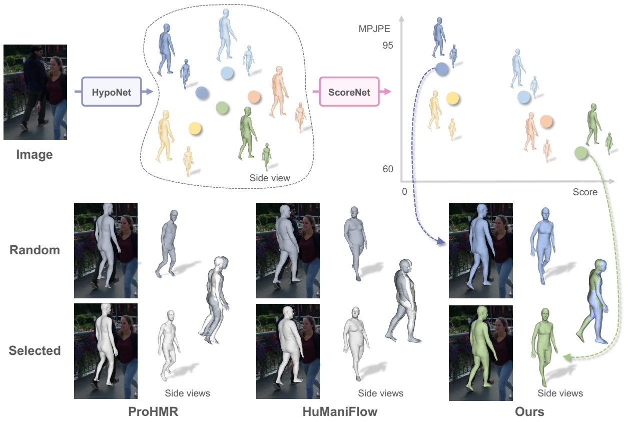

Figure 1. **Top:** Our proposed versatile framework ScoreHypo. HypoNet first generates multiple hypotheses that align with the image cues. Then ScoreNet evaluates and ranks them to identify superior estimates. **Bottom:** Qualitative comparison of (Random) randomly selected estimates to (Selected) the selected estimates by our ScoreNet. We visualize the projection view and two side views. The view setting is kept the same across different methods. Our selected result (green) exhibits the most reasonable poses compared to the randomly generated ones and previous works [31, 56]. Please zoom in to see the subtle differences.

Recently, there has been an increasing interest in proba-

[*] Corresponding authors.

bilistically modeling this task by leveraging generative techniques to produce multiple solutions. For instance, ProHMR [31] and HuManiFlow [56] both propose to map the image to a distribution of 3D human meshes using Normalizing Flows [27, 54]. Some approaches further employ diffusion models [20] to enhance the generation process [16]. However, a significant limitation of these probabilistic approaches is their limited emphasis on obtaining high-quality estimates from the generated hypotheses, which may encompass infeasible solutions. Currently, the selection mechanism typically involves either choosing the estimate with minimal error compared to the ground-truth (GT) [31] or simply averaging all the estimates [16]. Despite providing multiple possible estimates, the practical applicability of these methods in real-world scenarios is hindered by the absence of a robust mechanism for selecting more reliable estimates.

To overcome these limitations, we propose an innovative and versatile framework, named **ScoreHypo**, that not only enables the generation of multiple viable hypotheses but also incorporates a robust and generalized selection module, as illustrated in Figure 1 (top). To achieve this, we first employ *HypoNet* to generate multiple hypotheses by formulating the estimation problem as a progressive denoising process [20]. To effectively guide the denoising process, HypoNet conditions on the multi-scale image features and employs cross-attention mechanisms [65]. The multi-scale features provide both global and local pixel-aligned features which enables HypoNet to generate feasible estimates that visually align with 2D image cues. Nevertheless, these feasible solutions vary in probability when taking into account fine-grained visual cues and common sense. For example, consider the blue estimated result, where the human body exhibits a subtle forward inclination when observed from a side viewpoint in Figure 1. This subtle discrepancy contradicts the visual portrayal depicted in the image and poses challenges to the selection process.

In light of this, we develop a novel perspective to design a *ScoreNet* within the same framework. The ScoreNet functions as a critical scorer that ranks all the estimates and finally yields a single solution. To achieve this, we train the ScoreNet with a bundle set of diverse hypotheses generated by HypoNet. By incorporating the differentiable pairwise probabilistic ranking cost [7], we establish the probability distribution of each hypothesis pair's relative quality. Once trained, ScoreNet can score and rank over multiple estimates as shown in Figure 1. It can be seen that the selected estimate (green body) exhibits more reasonable orientations and inclinations. More importantly, our ScoreNet demonstrates strong generalizability, effectively scoring existing probabilistic methods [31, 56] to assist in selecting more reasonable results and significantly improve their performance (please refer to Table 3 in Section 4.4).

To conclude, our contributions are three-fold:

- We introduce ScoreHypo, a novel and versatile framework that combines probabilistic mesh estimation with a robust and generalized hypothesis selection module.
- We propose HypoNet, utilizing multi-scale image features and cross-attention mechanisms to generate feasible 3D estimates aligned with 2D image cues. It outperforms existing state-of-the-art methods on benchmark datasets.
- We present ScoreNet, a robust and generalizable module that effectively selects high-quality results. It is noteworthy that ScoreNet significantly improves the performance of existing probabilistic methods by more than 15%.

2. Related work

2.1. 3D Human Mesh Estimation

Deterministic estimation Most 3D human mesh estimation methods [22, 29, 43, 45, 64] output a single solution given a monocular image. Pioneer methods [6, 32, 53] optimize the 3D human parametric models such as SMPL [42] to align with 2D observations. For example, SMPLify [6] optimizes the SMPL parameters by minimizing the distance between the fitted 2D keypoints with the detected 2D keypoints. However, the optimization process is prone to get trapped in local optima due to the influence of initialization. Recent works [9, 22, 30, 33, 34, 37, 38, 73, 75] shift to using deep networks to estimate the human mesh and show promising results. Classical work HMR [22] proposes to learn the mapping from image space to the parameter space. However, the mapping is highly non-linear [33, 48, 75] which causes performance degradation. HybrIK [33] proposes to first estimate 3D human pose and then obtain the mesh using Inverse Kinematics, eliminating the difficulty of regressing the SMPL pose parameters.

Probabilistic estimation While multi-hypothesis 3D pose estimation has been extensively studied [13], there is relatively little research on probabilistic 3D human mesh estimation. Biggs *et al.* [5] extend HMR [22] to predict a discrete set of multiple hypotheses. Sengupta *et al.* [55] and ProPose [15] use the matrix Fisher distribution to model the distribution of pose rotations. ProHMR [31] and HuManiFlow [56] employ Normalizing Flows [27, 54] to model the plausible 3D human model parameter space. HMDiff [16] employs diffusion models [20] to estimate the plausible human meshes. In contrast, our method not only generates a set of plausible estimates but also provides a scoring module for selecting the more suitable estimates.

2.2. Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs), which are first introduced by Sohl-Dickstein *et al.* [57], are a type of generative model for image generation and undergo significant improvements with the development of acceleration

[20, 58], and enhancement [3, 50]. DDPMs learn the target distribution and generation by progressively introducing noise and denoising in both forward and reverse processes. This iterative denoising generation imparts knowledge about the distribution, ultimately yielding the generation of high-quality samples. These advancements have contributed to the growing popularity of DDPMs, sparking increased exploration in various fields and tasks such as image inpainting [44], semantic segmentation [4, 71], video generation [19, 72], and motion generation [24, 63]. In the realm of 3D human mesh estimation, where inherent depth ambiguity poses a challenge, probabilistic generation methods are well-suited. Motivated by the promising performances of Diffusion Models, we leverage them to formulate the task of 3D human mesh estimation.

3. Method

To address the limitations of existing methods in 3D mesh estimation from single RGB images, we propose a versatile framework **ScoreHypo** as depicted in Figure 2. The framework comprises two key components: *HypoNet* and *ScoreNet*. The HypoNet and ScoreNet share the same architecture design within the versatile framework elegantly, comprising a LatentNet and a Transformer-based network named HypoFormer and ScoreFormer, respectively. This design allows for HypoNet to generate multiple hypotheses based on the RGB image input, and allows for ScoreNet to score and rank these estimates to select a more suitable hypothesis. In the following, we present the model design and workflow of HypoNet and ScoreNet in Sections 3.1 and 3.2, respectively.

3.1. HypoNet

Problem formulation Monocular 3D human mesh estimation poses significant challenges due to depth ambiguity and self-occlusions, often resulting in multiple feasible solutions. To address this challenge, we draw inspiration from recent advancements in diffusion models [20, 59], which have demonstrated exceptional capabilities in generating diverse and high-quality solutions. The diffusion models [20, 59] achieve this by decomposing the generation process into multiple intermediate denoising steps.

Motivated by this, we formulate the estimation process as a reverse diffusion process, where we progressively denoise a Gaussian noise conditioned on the input image to recover the 3D human mesh. The whole framework of *HypoNet* consists of two key processes: a *forward diffusion process* and a *reverse sampling process*. (1) The forward diffusion process perturbs the 3D human mesh from the data distribution towards a Gaussian prior distribution by gradually adding noise to the GT meshes. (2) On the other hand, the reverse sampling process involves sampling Gaussian noise and progressively denoising it step-by-step. This process allows us to obtain a feasible 3D mesh estimate from the data distribution. By formulating the mesh estimation task as a reverse sampling process conditioned on a single image, we leverage the denoising capabilities of HypoNet, which are learned from the forward diffusion process. In the following, we provide a formal introduction to both the forward diffusion and reverse sampling processes. Additionally, we present the architecture design and training details of HypoNet.

Forward diffusion Following [20], starting from $\mathbf{x}_0$, a sample drawn from the data distribution, we establish a time-dependent diffusion process by noisy samples $\{\mathbf{x}_t\}_{t=0}^{T}$, where T denotes the total number of timesteps. Over the course of this process, we introduce standard Gaussian noise to the GT data $\mathbf{x}_0$, gradually transforming it into a Gaussian distribution $\mathbf{x}_T \sim p_T$:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) := \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{1}$$

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2}$$

where $\{\beta_t\}_{t=1}^{T}$ denotes the variance schedule. Thanks to the additivity of independent Gaussian distributions and reparameterization [26], the perturbation of $\mathbf{x}_t$ can be formulated as:

$$q(\mathbf{x}_t|\mathbf{x}_0) := \sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, \tag{3}$$

where $\alpha_t := 1 - \beta_t$, $\overline{\alpha_t} := \prod_{s=1}^{t} \alpha_s$, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Generation through reverse sampling By reversing the diffusion perturbing process [20], we can get a data sample $\mathbf{x}_0$ by denoising from a Gaussian distribution $\mathbf{x}_T \sim p_T$. To expedite the reverse process, we leverage the Denoising Diffusion Implicit Models (DDIM) [58] framework, which allows for denoising with fewer steps. This is achieved by defining a subset $\tau \subset \{0, ..., T\}$ that maintains denoising quality while reducing computational overhead. The reverse process can be defined as follows:

$$\begin{aligned}\mathbf{x}_{\tau_{i-1}} = &\sqrt{\overline{\alpha}_{\tau_{i-1}}}(\frac{\mathbf{x}_{\tau_i} - \sqrt{1-\overline{\alpha}_{\tau_i}}\hat{\boldsymbol{\epsilon}}_{\tau_i}}{\sqrt{\overline{\alpha}_{\tau_i}}}) \\ &+ \sqrt{1-\overline{\alpha}_{\tau_{i-1}} - \sigma_{\tau_i}^2}\hat{\boldsymbol{\epsilon}}_{\tau_i} + \sigma_{\tau_i}\boldsymbol{\epsilon}_{\tau_i},\end{aligned} \tag{4}$$

where τ_i, τ_{i-1} are the adjacent timesteps in the subset τ, $\sigma_{\tau_i}(\eta) = \eta\sqrt{(1-\overline{\alpha}_{\tau_{i-1}})/(1-\overline{\alpha}_{\tau_i})}\sqrt{1-\overline{\alpha}_{\tau_i}/\overline{\alpha}_{\tau_{i-1}}}$, and $\boldsymbol{\epsilon}_{\tau_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. During the reverse process, we need to know $\hat{\boldsymbol{\epsilon}}_{\tau_i}$ for each timestep. Therefore, we train a neural network $h_\theta(\mathbf{x}_t, t|\mathbf{c})$ to estimate it, where θ denotes the parameters, $\mathbf{c}$ denotes the image condition. We detail $h_\theta(\mathbf{x}_t, t|\mathbf{c})$ in the following.

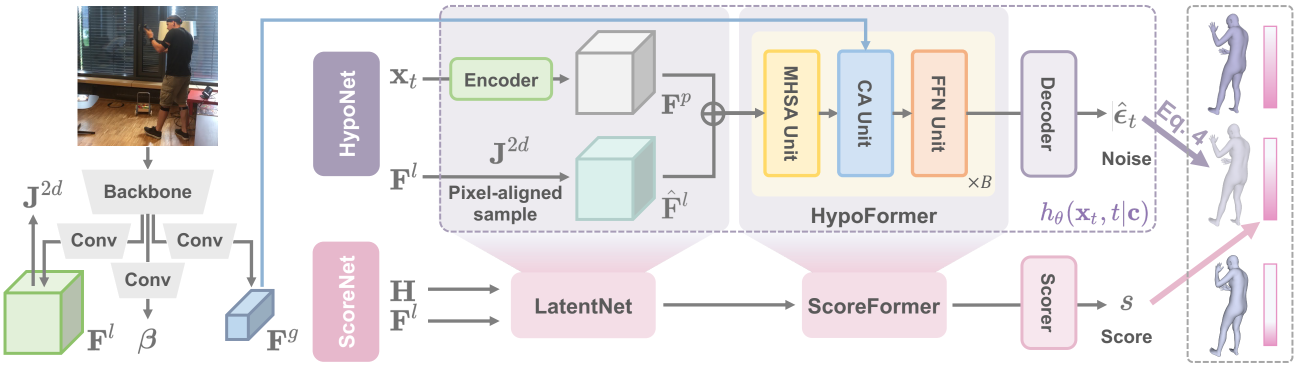

Figure 2. An overview of ScoreHypo, which consists of two core components, HypoNet and ScoreNet. HypoNet generates multiple estimations that align with image conditions through a diffusion process, while ScoreNet scores and selects a more suitable estimation based on the image cues.

Architecture design Figure 2 shows an overview of the whole framework. The noise estimator $h_\theta(\mathbf{x}_t, t|\mathbf{c})$ is the core network of HypoNet, which estimates the noise from the noisy data input $\mathbf{x}_t$. The noise estimator $h_\theta(\mathbf{x}_t, t|\mathbf{c})$ is composed of LatentNet, HypoFormer, and a decoder. We first introduce how we construct the data samples $\mathbf{x}_0$ and thus define the noisy sample $\mathbf{x}_t$, then present how to process the image as a condition $\mathbf{c}$, and finally introduce the network designs.

We use the SMPL model [42] to represent 3D human mesh which is parameterized by the pose $\boldsymbol{\theta} \in \mathbb{R}^{72}$ and shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$ parameters. Following [33], we decompose the SMPL pose parameters $\boldsymbol{\theta}$ into swing and twist representations. The swing representation could be derived from the 3D body joint positions $\mathbf{J} \in \mathbb{R}^{J\times 3}$ [33] in a closed-form, where J is the number of SMPL joints. The twist representation $\Phi \in \mathbb{R}^{\varphi\times 2} = \{(cos(\phi_i), sin(\phi_i))\}_{i=1}^{\varphi}$ denotes the twist rotation, where φ and ϕ_i denote the number of body-parts and 1-DoF twist rotation around i^{th} body-part, respectively. We construct the data sample $\mathbf{x}_0 = \{\tilde{\mathbf{J}}, \tilde{\Phi}\}$ as the combination of GT 3D joint positions and twist, where ˜ denotes GT. The forward diffusion process $\{\mathbf{x}_t\}_{t=0}^{T}$ is then defined according to Eq. 3 for all timesteps $t \in \{0, ..., T\}$. We use an encoder which is a multilayer perceptron (MLP) to map $\mathbf{x}_t$ to a high-dimensional feature vector $\mathbf{F}^p \in \mathbb{R}^{C^l \times (J+\varphi)}$.

To guide the diffusion process, we propose to use multi-scale image features as the condition $\mathbf{c} := \{\mathbf{F}^g, \mathbf{F}^l\}$. Two convolutional heads are deployed to obtain the low-resolution *global* feature $\mathbf{F}^g \in \mathbb{R}^{C^g\times H^g\times W^g}$ and the high-resolution *local* feature $\mathbf{F}^l \in \mathbb{R}^{C^l\times H^l\times W^l}$ after a CNN backbone, where C^* and $H^* \times W^*$ denotes corresponding feature channel and feature resolution, respectively. We enforce the local feature $\mathbf{F}^l$ to regress the 2D body joints $\mathbf{J}^{2d} \in \mathbb{R}^{J\times 2}$. Concretely, we obtain heatmaps by applying a convolution layer to $\mathbf{F}^l$, from which $\mathbf{J}^{2d}$ is regressed through the spatial integral technique in a differentiable manner [61]. We omit this process for clarity in Figure 2. We sample the local feature $\mathbf{F}^l$ according to the predicted 2D joint positions $\mathbf{J}^{2d}$ and obtain pixel-aligned features for each joint. Additionally, we use the midpoint position of a 2D joint pair to sample pixel-aligned features for φ body limbs to get the twist features. The combined features for J joints and φ limbs are denoted as $\hat{\mathbf{F}}^l \in \mathbb{R}^{C^l\times(J+\varphi)}$.

The LatentNet outputs the concatenated features of $\mathbf{F}^p$ and $\hat{\mathbf{F}}^l$, which is then fed into HypoFormer. The HypoFormer is a Transformer-encoder [65] based network, containing B basic blocks. Each basic block is built upon three units: a Multi-Head Self Attention (MHSA) unit, a Cross-Attention (CA) unit, and a Feed-Forward-Network [65] unit. In the CA unit, HypoFormer treats the global image feature $\mathbf{F}^g$ as the key and value features, while the output of the previous MHSA unit is the query feature. By using the cross-attention mechanism, we effectively guide the diffusion process to align with the image cues.

Finally, a decoder network [12] is deployed to estimate the noise $\hat{\boldsymbol{\epsilon}}_t$. The decoder network is an MLP.

Training losses To train the HypoNet, we randomly sample a timestep $t \in \{1, ..., T\}$ to get the perturbed noisy sample $\mathbf{x}_t$ according to Eq. 3. The overall loss function of HypoNet is defined as:

$$\mathcal{L}^H = \lambda_{noise}\mathcal{L}_{noise} + \lambda_{\boldsymbol{\beta}}\mathcal{L}_{\boldsymbol{\beta}} + \lambda_{2d}\mathcal{L}_{2d}, \tag{5}$$

where $\mathcal{L}_{noise}$ is the Mean Squared Error (MSE) loss between the predicted noise $\hat{\boldsymbol{\epsilon}}_t$ and the sampled noise $\boldsymbol{\epsilon}$:

$$\mathcal{L}_{noise} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}[||\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_t||^2]. \tag{6}$$

We estimate the SMPL shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ directly from the image and define $\mathcal{L}_{\boldsymbol{\beta}}$ as:

$$\mathcal{L}_{\boldsymbol{\beta}} = ||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2, \tag{7}$$

where ˜ denotes the GT. Additionally, we enforce the MSE loss between the predicted and GT 2D joint coordinates:

$$\mathcal{L}_{2d} = ||\tilde{\mathbf{J}}^{2d} - \mathbf{J}^{2d}||^2. \tag{8}$$

λ_{noise}, λ_{β} and λ_{2d} are constant coefficients.

3.2. ScoreNet

Once the HypoNet is trained, according to Eq. 4, HypoNet can produce a diverse set of plausible estimates that align with the input image given a random noise. However, the inherent ambiguity in the 2D to 3D lifting and the demands of real-world applications make it crucial to assess the quality of the generated estimates and propose a more reasonable and reliable one. However, it is challenging because the generated samples are already reasonably aligned with the 2D observations and previous probabilistic works [5, 31, 56] pay little attention to this. The subtle differences necessitate the model to have a keen perception of both the 2D observations and the 3D mesh priors.

In response to this challenge, we propose a robust module within the same framework that could effectively select high-quality results from multiple feasible hypotheses. We elegantly share the architecture design of HypoNet but simply change the input to the hypothesis $\mathbf{H}$ and the task-specific head to a Scorer network. We follow the same mesh decomposition and denote each $\mathbf{H} = \{\mathbf{J}, \Phi\}$ as the combination of the denoised joint $\mathbf{J}$ and twist Φ. ScoreNet assesses each hypothesis conditioned on the image features $\mathbf{c}$ in the same way as HypoNet does, and finally uses a Scorer network to assign a score s, unveiling its quality level. The Scorer is implemented as an MLP.

Training Given a set of hypotheses $\{\mathbf{H}_m\}_{m=0}^{M}$ generated from HypoNet, where M represents the number of hypotheses, our goal is to train ScoreNet to assign a corresponding score $\{s_m \in \mathbb{R}\}_{i=0}^{M}$ to each hypothesis conditioned on the corresponding image. We expect to assign higher scores to hypotheses with higher quality. We measure the hypothesis quality by utilizing common mesh evaluation metrics including Mean Per Vertex Error (MPVE) Q^v [33, 52] and Mean Per Joint Position Error (MPJPE) Q^j [22, 33, 38], which are defined as:

$$Q^j = \frac{1}{J}||\tilde{\mathbf{J}} - \mathbf{J}||_2^2, \quad Q^v = \frac{1}{V}||\tilde{\mathbf{V}} - \mathbf{V}||_2^2, \tag{9}$$

where $\mathbf{V} \in \mathbb{R}^{V\times 3}$ denotes the mesh obtained by the SMPL models [42], and V denotes the number of vertices.

To ensure the score accurately reflects the subtle quality differences, we model the training process of ScoreNet as the learning of the probability distribution of the relative quality differences among different hypotheses. We adopt a differentiable pairwise probabilistic ranking cost function [7] to define the relative quality difference probability P_{mn} between hypotheses $\mathbf{H}_m$ and $\mathbf{H}_n$ based on their corresponding scores s_m and s_n as follows:

$$P_{mn} := P(\mathbf{H}_m \succ \mathbf{H}_n) = \frac{1}{1 + e^{-\sigma\cdot(s_m - s_n)}}, \tag{10}$$

where $\mathbf{H}_m \succ \mathbf{H}_n$ indicates that the quality of $\mathbf{H}_m$ is higher than that of $\mathbf{H}_n$, and σ is a hyperparameter. The target probability $\tilde{P}_{mn}$ is designed as:

$$\tilde{P}_{mn} = \begin{cases} 1, & \mathbf{H}_m \succ \mathbf{H}_n \\ \frac{1}{2}, & \mathbf{H}_m = \mathbf{H}_n . \\ 0, & \mathbf{H}_m \prec \mathbf{H}_n \end{cases} \tag{11}$$

During training, we use the cross-entropy cost function [7] to fit the learned probability to the GT probability:

$$\begin{aligned} C_{mn}(P_{mn}, \tilde{P}_{mn}) := &- \tilde{P}_{mn}\log P_{mn} \\ &- (1 - \tilde{P}_{mn})\log(1 - P_{mn}). \end{aligned} \tag{12}$$

We define two target probabilities $\tilde{P}_{mn}^{j}$ and $\tilde{P}_{mn}^{v}$ based on MPVE Q^v and MPJPE Q^j quality measures, respectively. For $\tilde{P}_{mn}^{j}$, we define $\mathbf{H}_m \succ \mathbf{H}_n$ when $Q_m^j < Q_n^j$, *i.e.*, hypothesis $\mathbf{H}_m$ has a lower MPJPE error than hypothesis $\mathbf{H}_n$. Similarly, for $\tilde{P}_{mn}^{v}$, we define $\mathbf{H}_m \succ \mathbf{H}_n$ when $Q_m^v < Q_n^v$. The training loss is defined as follows:

$$\begin{aligned} \mathcal{L}_{rank} &= \sum_{m=0}^{M} \sum_{n=0, n\neq m}^{M} (\lambda_j \mathcal{L}_j + \lambda_v \mathcal{L}_v), \\ \mathcal{L}_j &= C_{mn}(P_{mn}, \tilde{P}_{mn}^{j}), \\ \mathcal{L}_v &= C_{mn}(P_{mn}, \tilde{P}_{mn}^{v}), \end{aligned} \tag{13}$$

where λ_j and λ_v are constant coefficients of the two cross-entropy costs.

The overall training loss of ScoreNet is defined as:

$$\mathcal{L}^S = \lambda_{rank}\mathcal{L}_{rank} + \lambda_{2d}\mathcal{L}_{2d}, \tag{14}$$

where λ_{rank} and λ_{2d} denote constant coefficients of the respective loss.

Inference In the inference phase, given a set of hypotheses $\{\mathbf{H}_m\}_{m=0}^{M}$ and the corresponding image $\mathbf{I}$, ScoreNet outputs the corresponding scores $\{s_m \in \mathbb{R}\}_{m=0}^{M}$. We sort them based on the scores and select the top K hypotheses to aggregate the final output by taking the average.

4. Experiments

4.1. Datasets and Metrics

H3.6M [21] dataset is a large-scale indoor 3D human dataset, where SMPL parameters are obtained from MoSh [41]. We follow the standard train-test split, using subjects (S1, S5, S6, S7, S8) for training and (S9, S11) for testing. Consistent with previous works [10, 22, 38, 39], we report the Mean Per Joint Position Error (MPJPE) and PA-MPJPE for SMPL poses derived from the meshes. We also provide the Mean Per Vertex Error (MPVE) for the entire mesh.

Method	H3.6M			3DPW		
	MPVPE↓	MPJPE↓	PA-MPJPE↓	MPVPE↓	MPJPE↓	PA-MPJPE↓
SMPLify [6] ECCV'16	-	-	82.3	-	-	-
HMR [22] CVPR'18	96.1	88.0	56.8	152.7	130.0	81.3
GraphCMR [30] CVPR'19	-	-	50.1	-	-	70.2
SPIN [29] ICCV'19	-	-	41.1	116.4	96.9	59.2
Pose2Mesh [10] ECCV'20	85.3	64.9	46.3	106.3	88.9	58.3
I2L-MeshNet [48] ECCV'20	65.1	55.7	41.1	110.1	93.2	57.7
HybrIK [33] CVPR'21	65.7	54.4	34.5	86.5	74.1	45.0
METRO [39] CVPR'21	-	54.0	36.7	88.2	77.1	47.9
PARE [28] ICCV'21	-	-	-	88.6	74.5	46.5
PyMaf [76] ICCV'21	-	57.7	40.5	110.1	92.8	58.9
CLIFF [37] ECCV'22	-	47.1	32.7	81.2	69.0	43.0
FastMETRO [9] ECCV'22	-	52.2	33.7	84.1	73.5	44.6
DeFormer [73] CVPR'23	-	44.8	31.6	82.6	72.9	44.3
POTTER [78] CVPR'23	-	56.5	35.1	87.4	75.0	44.8
ImpHMR [8] CVPR'23	-	-	-	87.1	74.3	45.4
NIKI [34] CVPR'23	-	-	-	86.6	71.3	40.6
Zolly [67] ICCV'23	-	49.4	32.3	76.3	65.0	39.8
Biggs *et al.* [5] NeurIPS'20 ($M = 10$)	-	59.2	42.2	-	79.4	56.6
Biggs *et al.* [5] NeurIPS'20 ($M = 25$)	-	58.2	42.2	-	75.8	55.6
Sengupta *et al.* [55] CVPR'21 ($M = 25$)	-	-	-	-	75.1	47.0
ProHMR [31] ICCV'21 ($M = 10$)	-	-	38.3	-	-	54.6
HuManiFlow [56] CVPR'23 ($M = 100$)	-	-	-	-	65.1	39.9
HMDiff [16] ICCV'23 ($M = 25$)	-	49.3	32.4	82.4	72.7	44.5
Ours ($M = 10$)	52.5	42.4	29.0	79.8	68.5	41.0
Ours ($M = 100$)	47.5	38.4	26.0	73.4	63.0	37.6
Ours ($M = 200$)	**46.4**	**37.4**	**25.3**	**71.9**	**61.8**	**36.1**

Table 1. Comparison to the state-of-the-arts on H3.6M [21] and 3DPW [66] datasets. The top and bottom blocks show deterministic and probabilistic methods, respectively.

3DPW [66] is an outdoor 3D human dataset that provides SMPL annotations. Following the previous works [28, 38, 39, 74], we use the training set of 3DPW for model training and evaluate its performance on the test set. We apply the same evaluation metrics as used for H3.6M [21].

4.2. Implementation Details

Following previous work [17, 28, 39, 55], our approach is trained on a mixture of data with 3D and 2D annotations, including H3.6M [21], 3DPW [66], MPI-INF-3DHP [47], MPII [2], COCO [40] and UP-3D [32] datasets. Only the training sets are used, following the standard split protocols.

We employ HRNet [60] as the CNN backbone and use the GT box to crop the human region, resizing the image to 256 × 256. The sizes of the two feature maps are $C^l = 256$, $H^l = W^l = 64$, $C^g = 512$, and $H^g = W^g = 8$. The number of body joints and twists are $J = 29$ and $\varphi = 23$, respectively, and their definitions follow HybrIK [33]. Both HypoFormer and ScoreFormer have $B = 6$ basic blocks.

We train HypoNet for 50 epochs using the Adam optimizer [25]. The initial learning rates for the backbone and the HypoNet are set to 0.0002 and 0.0005, respectively. We decay them by 0.5 at epochs 20, 30 and 40. In the inference process, we employ the accelerated sampling strategy from DDIM [58], generating hypotheses in 4 steps, and set $\eta = 0$. We train ScoreNet for 10 epochs with the same initial learning rates, and each sample has $M = 15$ hypotheses. We decay the learning rates by 0.5 at epoch 5. For inference, we set the aggregate number $K = 5$.

4.3. Comparison to State-of-the-art

We compare our method to the state-of-the-art methods on H3.6M [21] dataset and 3DPW [66] dataset, as presented in Table 1. Following the conventions of standard multi-hypothesis approaches [5, 31, 36], we generate multiple estimates aligned with the image by HypoNet and report the minMPJPE, minMPVE of the M hypotheses. Our method consistently outperforms all probabilistic state-of-the-art methods, such as ProHMR [31] and HuManiFlow [56], by a substantial margin when sampling $M = 10$ and $M = 100$ hypotheses. As we increase the number of hypotheses, our method exhibits significant improvements, showcasing the scalability and superiority of our multi-hypothesis mesh estimator HypoNet. We offer additional comparisons, including using different training datasets and different methods for generating M hypotheses, in the supplementary material.

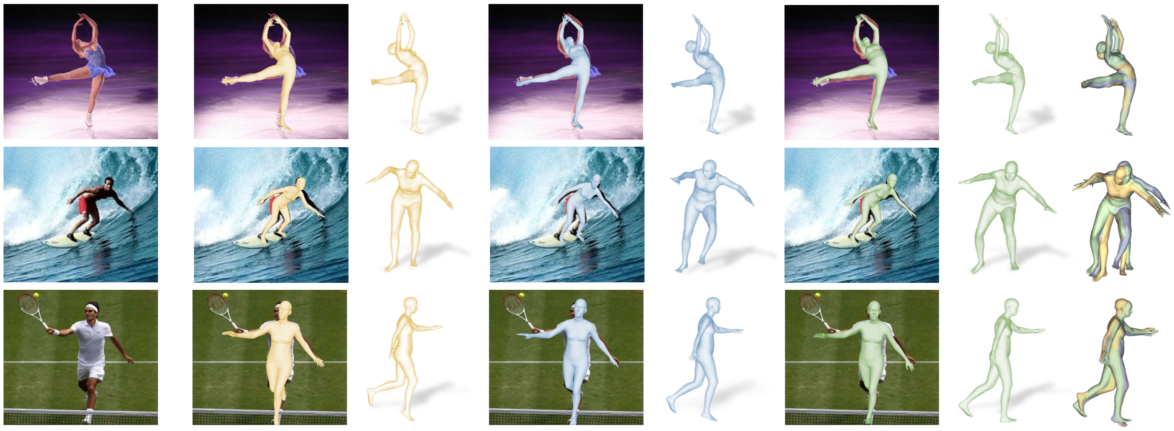

Figure 3. Qualitative results on challenging in-the-wild images. The yellow and blue-colored meshes are the generated results of HypoNet, while the green ones are the final results selected by ScoreNet. The last column overlaps the multiple estimates to unveil their differences.

Method	MPVPE↓	MPJPE↓	PA-MPJPE↓
(a) *w/o* $\mathbf{F}^g$	87.8	75.9	46.4
(b) *w/o* $\mathbf{F}^l$	93.1	80.8	47.8
(c) HypoNet *(full)*	**86.6**	**74.1**	**44.8**

Table 2. Evaluation of HypoNet components on 3DPW [66] testset.

4.4. Ablation Study

Effectiveness of HypoNet To validate the effectiveness of the 2D image condition guidance in the diffusion process of HypoNet, we compare our approach to two baselines in Table 2 on the 3DPW test set. In baseline (a), we remove the introduction of the global feature $\mathbf{F}^g$ as well as the cross-attention unit to assess the role of global features. In baseline (b), we eliminate the local features $\mathbf{F}^l$. To mitigate the impact of randomness, we report metrics based on a single sample generated without any noise. Table 2 demonstrates the effectiveness of both global and local features to the generation capability of HypoNet, highlighting the effectiveness of incorporating multi-scale image features with cross-attention mechanisms [65]. In Figure 3, we showcase the multi-hypotheses generated by HypoNet on real-world images, all of which align well with the 2D observations. We also provide a side view to demonstrate the 3D feasibility. Notably, HypoNet exhibits robust generalization even in highly challenging and complex scenes, as exemplified by the first row depicting figure skating.

Effectiveness of ScoreNet To evaluate the effectiveness of our ScoreNet, we design four selection strategies on the 3DPW [66] dataset in Table 3. In strategies (a) and (b), the output results are samples generated from the Gaussian noise and zero noise ϵ_0 using the HypoNet, respectively. In strategy (c), we simply average the multi-hypotheses generated by HypoNet as the final output. In strategy (d), we use our trained ScoreNet to select from M hypotheses generated by HypoNet. The rightmost column block shows the results of our framework, which shows that our ScoreNet can effectively select higher-quality hypotheses, outperforming all the other selection strategies. With an increase in the number of hypotheses, ScoreNet consistently enhances the performance.

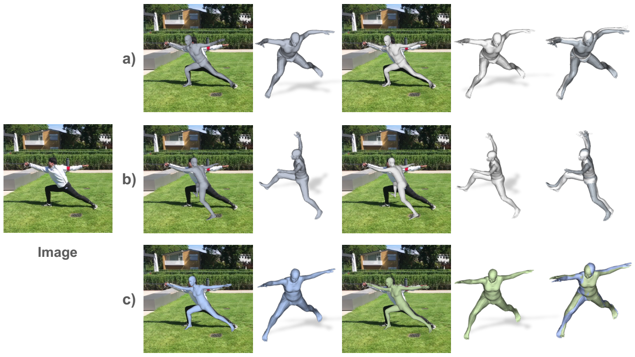

Figure 4. Qualitative comparison of (a) ProHMR [31], (b) HuManiFlow [56] and (c) our method on 3DPW [66] test set. The left and right columns denote the randomly selected estimate and the selected estimate by our ScoreNet, respectively. Side views (*w.* shadow) and the overlapped meshes are shown for a better view of subtle differences.

Furthermore, ScoreNet demonstrates robustness and generalizability which could effectively improve the SOTA probabilistic methods, including ProHMR [31] and HuManiFlow [56], in a seamless plug-and-play manner. Without the need for fine-tuning, our ScoreNet significantly boosts their performance, reducing more than 13.7% and 17.2% on MPVE for these two methods compared to (a) random selection. Besides, the notable improvement also indicates that ProHMR [31] and HuManiFlow [56] exhibit inferior quality in generating multiple hypotheses compared to HypoNet, resulting

Method	M	ProHMR [31]			HuManiFlow [56]			Ours		
		MPVE↓	MPJPE↓	PA-MPJPE↓	MPVE↓	MPJPE↓	PA-MPJPE↓	MPVE↓	MPJPE↓	PA-MPJPE↓
(a) Random	1	120.7	105.1	66.4	112.9	93.8	60.9	93.8	78.3	49.2
(b) Zero noise ϵ_0	1	111.0	98.1	59.0	100.8	86.3	53.5	86.6	74.1	44.8
(c) Average	10	111.9	98.7	59.6	100.9	86.5	53.6	87.2	74.6	45.3
	100	110.9	98.0	58.8	99.9	85.8	52.8	86.3	73.9	44.8
(d) ScoreNet	10	109.2	96.6	58.8	98.2	84.6	52.7	86.1	73.6	44.6
	100	104.2	92.6	56.9	93.5	81.0	50.6	**84.6**	**72.4**	**44.5**

Table 3. Ablation study on the effectiveness of ScoreNet on 3DPW [66] dataset.

Method	MPVE↓	MPJPE↓	PA-MPJPE↓
I2L-MeshNet[48]	129.5	92.0	61.4
SPIN[29]	121.4	95.5	60.7
PyMAF[76]	113.7	89.6	59.1
ROMP[62]	-	91.0	62.0
OCHMR[23]	145.9	112.2	75.2
PARE[28]	101.5	83.5	57.0
3DCrowdNet[11]	101.5	83.5	57.1
JOTR[35]	92.6	75.7	52.2
Ours	**89.8**	**73.9**	**48.7**

Table 4. Comparison to the state-of-the-art methods on the challenging 3DPW-OC occlusion datatset [66, 77].

in more unreliable estimates.

In addition to Figure 1, Figure 4 displays another qualitative comparison of the 3DPW test set between (a) ProHMR [31], (b) HuManiFlow [56] and (c) our method. The left and right columns denote the randomly selected estimate and the selected estimate by our ScoreNet, respectively. It can be seen that the hypothesis generated by HypoNet exhibits higher quality compared to the other two methods, showing the powerful generation capabilities of HypoNet. The selected results by ScoreNet are more aligned on the 2D image and more reasonable in the 3D space. Please zoom in to observe our improvement over ProHMR [31], where the results selected by ScoreNet exhibit more accurate forward inclination angles of the body, demonstrating the strong generalization and robustness of ScoreNet. Since HuManiFlow [56] fails to generate any plausible estimates, our ScoreNet is unable to correct the estimates. Due to the page limit, please refer to the supplementary for more cases.

Robustness to occlusion To assess our robustness to occlusion, we conduct experiments on the object occlusion subset of 3DPW (3DPW-OC) [66, 77]. To ensure fairness, HypoNet and ScoreNet are not trained on the 3DPW training set. As shown in Table 4, our method achieves state-of-the-art performance on 3DPW-OC when using ScoreNet, highlighting its strong effectiveness in handling occluded scenarios. Figure 5 provides visualizations of qualitative results on the 3DPW-OC subset. Notably, HypoNet generates diverse and reasonable results (yellow and blue) even in highly self-occluded cases. ScoreNet further selects a more plausible estimate (green), such as a more reasonable head orientation in the third row.

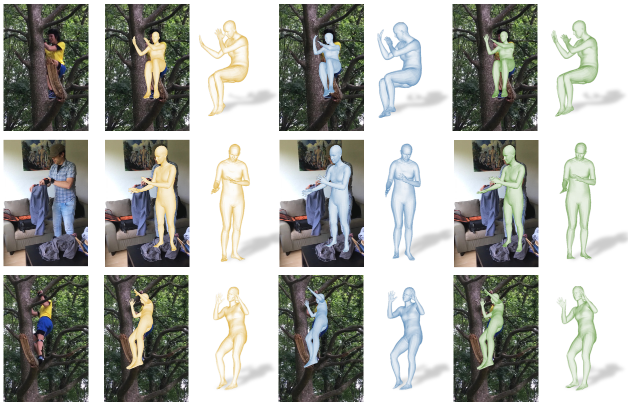

Figure 5. Qualitative results of our method on 3DPW-OC subset [66, 77]. Yellow and blue-colored meshes denote the hypotheses generated by HypoNet, and green ones denote the selected estimate by ScoreNet.

5. Conclusion

We present ScoreHypo, a versatile framework that combines probabilistic mesh estimation with a robust and generalized hypothesis selection module. We propose HypoNet, which leverages multi-scale image features to generate multiple accurate 3D estimates that align well with 2D image cues. HypoNet outperforms existing state-of-the-art methods on benchmark datasets, demonstrating its superior performance. In addition, we propose ScoreNet, a robust and generalizable module that effectively selects high-quality results. Notably, ScoreNet significantly improves the performance of existing probabilistic methods, showcasing its strong generalization ability and versatility. Moreover, our approach provides accurate and reliable solutions even in challenging real-world scenarios.

6. Acknowledgment

This work was supported by National Science and Technology Major Project (2022ZD0114904). We would like to thank Hai Ci and Xuehai Pan for their insightful discussions.

References

[1] Jake K Aggarwal and Quin Cai. Human motion analysis: A review. pages 428–440. Elsevier, 1999. 1

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014. 6, 9

[3] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. 34:17981–17993, 2021. 3

[4] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021. 3

[5] Benjamin Biggs, David Novotny, Sebastien Ehrhardt, Hanbyul Joo, Ben Graham, and Andrea Vedaldi. 3d multi-bodies: Fitting sets of plausible 3d human models to ambiguous image data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 20496–20507, 2020. 2, 5, 6, 10

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, pages 561–578, 2016. 2, 6, 9

[7] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005. 2, 5

[8] Hanbyel Cho, Yooshin Cho, Jaesung Ahn, and Junmo Kim. Implicit 3d human mesh recovery using consistency with pose and shape from unseen-view. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21148–21158, 2023. 1, 6

[9] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision (ECCV)*, pages 342–359, 2022. 1, 2, 6

[10] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision (ECCV)*, pages 769–787, 2020. 5, 6

[11] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3d human mesh from in-the-wild crowded scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1475–1484, 2022. 8

[12] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *International Conference on Computer Vision (ICCV)*, pages 2262–2271, 2019. 4

[13] Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, and Yizhou Wang. Gfpose: Learning 3d human pose prior with gradient fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4800–4810, 2023. 2

[14] Steffi L Colyer, Murray Evans, Darren P Cosker, and Aki IT Salo. A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. *Sports medicine-open*, 4(1):1–15, 2018. 1

[15] Qi Fang, Kang Chen, Yinghui Fan, Qing Shuai, Jiefeng Li, and Weidong Zhang. Learning analytical posterior probability for human mesh recovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8781–8791, 2023. 2

[16] Lin Geng Foo, Jia Gong, Hossein Rahmani, and Jun Liu. Distribution-aligned diffusion for human mesh recovery. In *International Conference on Computer Vision (ICCV)*, pages 9221–9232, 2023. 2, 6

[17] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *European Conference on Computer Vision (ECCV)*, pages 768–784. Springer, 2020. 6

[18] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 9

[19] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. 35:27953–27965, 2022. 3

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 2, 3

[21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 5, 6, 9

[22] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 1, 2, 5, 6, 9

[23] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1715–1725, 2022. 8

[24] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 8255–8263, 2023. 3

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[27] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. 29, 2016. 2

[28] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. Pare: Part attention regressor for 3d human body estimation. In *International Conference on Computer Vision (ICCV)*, pages 11127–11137, 2021. 6, 8, 10

[29] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 2, 6, 8, 9

[30] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4501–4510, 2019. 2, 6, 9

[31] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *International Conference on Computer Vision (ICCV)*, pages 11605–11614, 2021. 1, 2, 5, 6, 7, 8, 10

[32] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6050–6059, 2017. 2, 6, 9

[33] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3383–3393, 2021. 2, 4, 5, 6, 9

[34] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12933–12942, 2023. 2, 6

[35] Jiahao Li, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Jotr: 3d joint contrastive learning with transformers for occluded human mesh recovery. In *International Conference on Computer Vision (ICCV)*, pages 9110–9121, 2023. 8

[36] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13147–13156, 2022. 6

[37] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision (ECCV)*, pages 590–606, 2022. 2, 6, 9

[38] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *International Conference on Computer Vision (ICCV)*, pages 12939–12948, 2021. 2, 5, 6, 9

[39] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1954–1963, 2021. 1, 5, 6, 9

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 6, 9

[41] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):1–13, 2014. 5, 9

[42] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 2, 4, 5, 10

[43] Tianyu Luan, Yali Wang, Junhao Zhang, Zhe Wang, Zhipeng Zhou, and Yu Qiao. Pc-hmr: Pose calibration for 3d human mesh recovery from 2d images/videos. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 2269–2276, 2021. 2

[44] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, 2022. 3

[45] Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, and Yizhou Wang. 3d human mesh estimation from virtual markers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 534–543, 2023. 2

[46] Abed Malti. Robust monocular 3d human motion with lasso-based differential kinematics. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6607–6617, 2023. 1

[47] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision (3DV)*, pages 506–516, 2017. 6, 9

[48] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, pages 752–768, 2020. 1, 2, 6, 8

[49] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2299–2307, 2022. 9

[50] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pages 8162–8171. PMLR, 2021. 3

[51] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 9

[52] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 459–468, 2018. 5

[53] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[54] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning (ICML)*, pages 1530–1538. PMLR, 2015. 2

[55] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Hierarchical kinematic probability distributions for 3d human

shape and pose estimation from images in the wild. In *International Conference on Computer Vision (ICCV)*, pages 11219–11229, 2021. 2, 6

[56] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Humaniflow: Ancestor-conditioned normalising flows on so(3) manifolds for human pose and shape distribution estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4779–4789, 2023. 1, 2, 5, 6, 7, 8, 10

[57] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265. PMLR, 2015. 2

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 3, 6, 10

[59] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[60] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5693–5703, 2019. 6, 9

[61] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 4

[62] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *International Conference on Computer Vision (ICCV)*, pages 11179–11188, 2021. 8

[63] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[64] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 4, 7, 9

[66] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 6, 7, 8, 9, 10

[67] Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 3925–3935, 2023. 6

[68] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5908–5917, 2019. 1

[69] Yuxin Wu and Kaiming He. Group normalization. In *European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 9

[70] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[71] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2955–2966, 2023. 3

[72] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *Entropy*, 25 (10):1469, 2023. 3

[73] Yusuke Yoshiyasu. Deformable mesh transformer for 3d human mesh recovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17006–17015, 2023. 2, 6

[74] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Thundr: Transformer-based 3d human reconstruction with markers. In *International Conference on Computer Vision (ICCV)*, pages 12971–12980, 2021. 6

[75] Wang Zeng, Wanli Ouyang, Ping Luo, Wentao Liu, and Xiaogang Wang. 3d human mesh regression with dense correspondence. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7054–7063, 2020. 2

[76] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *International Conference on Computer Vision (ICCV)*, pages 11446–11456, 2021. 1, 6, 8

[77] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7376–7385, 2020. 8

[78] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1611–1620, 2023. 6

[79] Wentao Zhu, Zhuoqian Yang, Ziang Di, Wayne Wu, Yizhou Wang, and Chen Change Loy. Mocanet: Motion retargeting in-the-wild via canonicalization networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 3617–3625, 2022. 1

[80] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *International Conference on Computer Vision (ICCV)*, pages 15085–15099, 2023. 1

[81] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1