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'lEng. 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 2Dimage 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 advance- ments in this field have primarily focused on enhancing

*Corresponding authors.

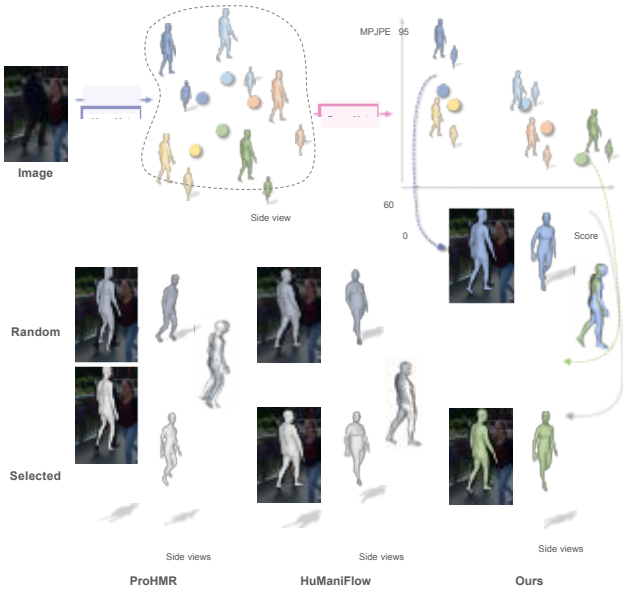

Figure 1. **Top:** Our proposed versatile framework ScoreHypo. HypoNet first generates multiple hypotheses that align with the image cues. Then ScoreNet evaluates and ranks them to identify superior estimates. **Bottom:** Qualitative comparison of (Random) randomly selected estimates to (Selected) the 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.

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 themonocular mesh estimation challenge.

Recently, there has been an increasing interest in proba-

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 notewor- thy 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 promis- ing 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