摘要

在计算机视觉领域，如何在非受限场景下准确感知和理解人类的姿态与形态仍然面临诸多挑战。非受限场景通常指真实世界中自然采集的图像或视频，其中包含复杂背景、遮挡、视角变化以及形态各异的个体，使得感知任务更加复杂。这与受控实验室环境形成鲜明对比，后者通常依赖标记点、深度传感器或多视角系统进行精确建模。然而，在实际应用中，如体育动作分析、智能监控等，人类的感知往往发生在无标记的自然环境中，难以依赖额外的传感器信息。因此，研究适用于非受限场景的三维人体姿态与形态建模方法，不仅在虚拟现实、智能家居、人机交互等领域具有重要的理论意义和实践价值，也为计算机视觉在更广泛的真实场景中开辟了新的应用方向。

三维人体的姿态与形态重建通常通过估计完整的人体网格来实现，该网格同时编码了人体的姿态和形态信息。然而，由于单目图像本质上缺乏深度信息，这一任务极具挑战。传统的骨架表示虽然能够简洁地表达人体姿态，但难以全面刻画人体的完整形态，从而限制了三维重建的精度。针对该任务中存在的人体中间表示能力弱、姿态歧义严重及效率瓶颈等关键问题，本文围绕“非受限场景下的人体姿态与形态重建”展开系统性研究，提出一系列创新方法，以分别提升模型的准确性、鲁棒性以及效率，核心贡献包括以下四个方面：

（1）提出虚拟标记点表示以提升三维人体网格重建的精度。近年来，体素化三维姿态估计算法取得了显著进展，一些研究尝试将三维骨架作为中间表示，以此回归三维人体网格。然而，骨架表示难以完整表达人体形态，导致网格重建精度受限。相比之下，先进的运动捕捉系统通过在人体表面布置密集标记点，可精确获取非刚性运动并重建逼真的人体网格，但此类方法依赖物理标记点，难以直接应用于非受限场景的图像。针对这一问题，本文提出了一种新的中间表示——虚拟标记点，从大规模运动捕捉数据中学习得到一组稀疏地分布于人体表面的关键点，模拟物理标记点的效果。这些标记点可直接从单目图像中检测，并通过插值重建完整且真实的人体网格。基于该虚拟标记点表示，本文提出了一个能够从非受限图像中检测虚拟标记点并用其重建三维人体网格的算法框架。实验结果表明，该方法在多个数据集上均优于现有最先进方法，特别是在包含多样化人体形态的数据集上取得了显著优势。

（2）通过概率建模提高三维人体网格估计的遮挡鲁棒性。单目三维人体网格重建是一个高度病态的问题，图像中缺乏深度信息，往往存在遮挡等问题，导致重建结果具有固有的不确定性与多解性。在真实的非受限场景中，人体被遮挡的现象十分常见，这对三维网格重建提出了额外挑战。本文发现，在上述框架下，直接进行确定性回归

来估计三维虚拟标记点的方式，在存在遮挡时准确性可能会显著下降。为此，本文进一步提出了一种基于概率建模的框架，将三维虚拟标记点的估计建模为一个基于图像输入的生成过程。具体而言，引入条件去噪模型来建模三维虚拟标记点的估计，使得在存在遮挡的情况下，仍能生成多个合理的三维人体网格，并确保其与输入图像保持一致。实验表明，该概率建模框架在多个基准数据集上均超越现有方法，并在包含多样化人体形态的数据集中取得了显著的性能提升。此外，该方法在遮挡场景下表现尤为突出，能够更准确地建模数据分布，从而提升网格重建的鲁棒性，确保即使在部分关键点缺失的情况下，依然能够生成合理且与输入图像一致的人体网格。

（3）提出多假设生成与评分机制以进一步提升网格估计的稳定性与准确性。针对单目三维人体重建中的不确定性与多解性问题，本文从另一个角度出发，提出一种双阶段的多假设生成与评分机制，以全面提升估计结果的稳定性与准确性。首先，设计多假设生成模块，从输入图像出发，通过反向去噪过程生成一组多样化且与图像线索高度一致的三维人体网格候选解，显著提升了解的覆盖范围与合理性。随后，引入评分模块，对所有候选解进行系统性评估与排序，从中挑选出高质量的最终估计结果。实验表明，该方法作为一种多假设估计框架，在多个基准数据集上均优于现有的概率估计方法，生成结果在准确性、多样性方面均表现优异。进一步地，评分模块具有良好的泛化能力，不仅能筛选自身生成的结果，也能有效评估并优化已有方法的人体网格估计质量，帮助多个现有方法显著降低重建误差。该机制为单目三维人体网格估计提供了一种稳健、高效的解法，特别适用于遮挡严重与姿态复杂的非受限场景。

（4）提出基于尺度自适应策略的高效多人三维人体网格估计方法，以实现非受限场景下的实时重建。在前述方法有效提升单人重建的准确性与鲁棒性的基础上，如何在保证精度的同时实现高效建模，尤其在多人、复杂场景下获得实时性能，仍是实际应用中的关键问题。当前一阶段方法在高分辨率输入下可获得较高精度，但同时带来显著的计算开销。本文观察到图像中个体的尺度对估计性能影响显著，尤其是远距离或身高较小的个体。为此，本文提出一种基于个体尺寸动态调整图像分辨率的机制，通过定义尺度自适应标识，使不同尺度的个体分别以不同分辨率处理，背景区域则通过特征蒸馏简化冗余信息。该策略能够更有效地分配计算资源，提升小尺度个体的建模能力，同时降低整体计算成本。实验表明，该方法在保持高精度的同时实现了实时性能，在多个数据集上均达到与当前最优方法相当的水平，显示出良好的实用前景。

综上所述，本文围绕非受限场景下的三维人体姿态与形态重建问题，系统性地提出了一系列具有层次递进关系的创新方法，依次解决表示局限性、深度歧义性、解选择困难及建模效率低下等关键挑战。首先，提出虚拟标记点表示，显著增强了三维人体网格的表达能力与重建精度，为后续建模奠定基础；随后，引入基于条件生成的概率

建模框架，在虚拟标记点基础上，有效建模遮挡环境下的估计歧义，提升鲁棒性；在此基础上，进一步设计多假设评分机制，从多个候选解中系统性筛选最优估计结果，提高重建稳定性与可靠性；最后，针对多人复杂场景，提出尺度自适应策略，在保障精度的同时显著提升建模效率。上述研究形成一套从表达建模到不确定性处理、再到高效推理的完整方法链，为真实世界中的高质量人体感知提供了坚实的研究基础与实践经验。

关键词：计算机视觉，三维人体建模，姿态与形态估计

Reconstruction of 3D Human Pose and Shape In-the-Wild

Xiaoxuan Ma (Computer Application Technology)
Directed by: Prof. Yizhou Wang

ABSTRACT

In computer vision, accurately capturing human pose and shape under unconstrained scenarios remains a significant challenge. Unconstrained scenarios typically refer to images or videos captured in natural, real-world environments, often involving complex backgrounds, occlusions, varying viewpoints, and diverse human appearances. This stands in stark contrast to controlled laboratory settings, which rely on physical markers, depth sensors, or multi-view systems for precise capturing. However, in real-world applications such as sports analysis and intelligent surveillance, human perception must often be performed in markerless environments, where additional sensor data is unavailable. Therefore, developing robust 3D human pose and shape estimation methods tailored for unconstrained scenarios is of great theoretical and practical importance across domains such as virtual reality, smart environments, and human–computer interaction. It also opens new directions for deploying computer vision systems in broader real-world contexts.

Reconstructing 3D human pose and shape is typically achieved by estimating a full-body mesh that simultaneously encodes both pose and shape information. However, this task is highly challenging due to the inherent lack of depth information in monocular images. While traditional skeletal representations can succinctly describe human poses, they fall short in capturing detailed body shapes, thus limiting the accuracy of 3D reconstruction. To address key challenges in this task—including weak intermediate representations, severe pose ambiguity, and computational bottlenecks—this work conducts a systematic study on 3D human pose and shape reconstruction in in-the-wild scenarios and proposes a series of innovative methods aimed at improving the accuracy, robustness, and efficiency of the model. The core contributions can be summarized in the following four aspects:

(1) A novel virtual marker representation is proposed to improve the accuracy of 3D human mesh reconstruction. While voxel-based 3D pose estimation methods have made notable progress in recent years, several works attempt to use 3D skeletons as intermediate representations to regress full human meshes. However, skeletons fail to capture detailed body shape,

limiting reconstruction precision. In contrast, advanced motion capture systems deploy dense physical markers on the human body to capture non-rigid motion and produce highly accurate meshes, but these systems depend on physical instrumentation and are not suited for in-the-wild scenarios. To address this, we introduce a novel intermediate representation, virtual markers, which are sparse keypoints learned from large-scale motion capture data and distributed across the human surface to emulate the effects of physical markers. These markers can be directly detected from monocular images and used to reconstruct realistic meshes via interpolation. Based on this representation, we develop a simple yet effective framework that detects virtual markers from unconstrained images and reconstructs the 3D human mesh. Experimental results show that our method outperforms state-of-the-art approaches across multiple benchmarks, particularly in datasets with diverse body shapes.

(2) A probabilistic modeling framework is proposed to improve robustness under occlusion in monocular 3D mesh estimation. Monocular 3D human mesh reconstruction is inherently ill-posed, plagued by depth ambiguity, and frequent occlusion, especially in unconstrained scenarios. We observe that directly regressing virtual markers in a deterministic manner can suffer from severe accuracy degradation in the presence of occlusion. To overcome this, we propose a probabilistic framework that formulates the estimation of virtual markers as a conditional generation process from input images. Specifically, we introduce a conditional denoising model to model the distribution of virtual markers, allowing for the generation of multiple plausible and image-consistent meshes even under partial observations. Experiments demonstrate that this probabilistic approach achieves superior performance over existing methods on several benchmarks, particularly in datasets featuring diverse body types and heavy occlusions. The proposed method shows strong robustness by more accurately modeling the underlying data distribution, enabling plausible reconstructions even when keypoints are partially missing.

(3) A multi-hypothesis generation and scoring mechanism is proposed to further enhance the stability and accuracy of mesh estimation. To address the inherent uncertainty and multi-modality in monocular 3D human reconstruction, we propose a dual-stage framework consisting of multi-hypothesis generation and a dedicated scoring mechanism. First, a reverse denoising process is used to generate a diverse set of mesh candidates that are consistent with image cues, significantly expanding the solution space. Then, a scoring module is introduced to systematically evaluate and rank all candidates, selecting high-confidence predictions as final outputs. Experimental results confirm that this multi-hypothesis framework outperforms

existing probabilistic methods on multiple benchmarks, achieving higher accuracy and diversity. Importantly, the scoring module exhibits strong generalization ability: it can not only rank its own generated hypotheses but also effectively evaluate outputs from other methods, leading to significant error reduction. This mechanism offers a robust, efficient solution for monocular 3D human mesh estimation, especially in complex scenes with severe occlusions or challenging poses.

(4) A scale-adaptive strategy is proposed for efficient multi-person mesh estimation in real-time under unconstrained conditions. Building on the accuracy and robustness of the aforementioned single-person methods, we further address the challenge of efficiency in multi-person settings. While high-resolution inputs can enhance reconstruction quality, they also incur high computational costs. We observe that the scale of individuals in the image significantly affects estimation performance, especially for small or distant subjects. To this end, we introduce a scale-adaptive resolution adjustment strategy that dynamically adjusts input resolution per person based on estimated scale. Additionally, background regions are simplified via feature distillation to reduce redundancy. This approach enables more efficient resource allocation, enhances modeling quality for small-scale individuals, and reduces overall computation cost. Experiments show that our method achieves real-time performance without sacrificing accuracy, matching state-of-the-art results on multiple datasets and demonstrating strong practical potential.

In summary, this thesis systematically proposes a series of innovative methods with a progressive structure to address key challenges in 3D human pose and shape reconstruction under in-the-wild conditions, tackling limitations in representation, depth ambiguity, solution selection, and modeling efficiency. First, we introduce a virtual marker representation that significantly enhances the expressiveness and reconstruction accuracy of 3D human meshes, laying a solid foundation for subsequent modeling. Building upon this, we develop a conditional generative probabilistic framework to effectively capture the inherent ambiguity under occlusions, thereby improving robustness. On top of that, a multi-hypothesis scoring mechanism is designed to systematically select the optimal estimation from multiple candidates, enhancing the stability and reliability of the reconstruction. Finally, we propose a scale-adaptive strategy tailored for complex multi-person scenarios, which substantially improves modeling efficiency while maintaining high accuracy. Together, these contributions form a complete methodological pipeline—from representation modeling to uncertainty handling and efficient inference—providing a solid research foundation and practical insights for high-quality human

perception in real-world applications.

KEY WORDS: Computer Vision, 3D Human Modeling, Pose and Shape Estimation