基于Ego-Body Pose Estimation via Ego-Head Pose Estimation的复现与研究

摘要

个人视角下的全身姿态估计是计算机视觉领域的重要研究课题，其广泛应用于虚拟现实、运动分析及人机交互等场景。然而，由于个人视角与全身姿态之间存在复杂的一对多映射关系，该任务具有较高的挑战性。本文复现了论文 Ego-Body Pose Estimation via Ego-Head Pose Estimation 提出的 EgoEgo 方法，该方法创新性地将头部姿态估计作为中间桥梁，通过单目 SLAM、GravityNet 和 HeadNet 的协同作用，精准估计头部姿态，并利用条件扩散模型生成全身姿态。复现过程中，深入探讨了 GravityNet 和 HeadNet 的特征提取模块与损失函数定义，并对模型整体框架进行了详细分析。

实验结果表明，EgoEgo 方法在多项公开数据集上的性能表现优异，尤其是在全身姿态生成的精度和鲁棒性方面达到了先进水平。同时，针对现有模型的不足，本文提出了进一步优化的方向，包括使用轻量化网络结构优化 GravityNet 和 HeadNet，利用大规模真实捕获数据提升头部旋转估计精度，以及引入手部信息等新的特征，减少全身姿态估计中的不确定性。此外，为满足边缘设备的实时性需求，建议采用模型剪枝、量化和知识蒸馏等轻量化技术。

通过对 EgoEgo 方法的复现和分析，本文验证了其在个人视角全身姿态估计任务中的可行性和优势，并为未来研究提供了改进方向与启示。

1 引言

选题背景

1.传统姿态估计的局限性：

第三人称视频的依赖：传统的3D人体姿态估计方法通常基于第三人称视频，能够直接观测到完整的人体关节。然而，这类方法对摄像头的布置要求较高，难以适应第一人称场景中的视野受限问题。

标注成本高：第三人称视频需要手动标注关键点或利用复杂的运动捕捉设备采集数据，这在动态场景中尤其困难。

动态场景适应性不足：第一人称视角下，用户的身体大部分时间处于摄像头视野之外，这对直接从视频中恢复全身动作提出了严峻挑战。

2.多模态数据融合的潜力：

头部与全身动作的关联性：头部姿态通常是全身运动的重要组成部分。例如，当用户转动头部时，其身体的整体运动往往与之相关。通过建模头部和全身之间的动态关系，可以间接估计全身姿态。

SLAM技术的优势与不足：SLAM技术可用于从视频中估计摄像头轨迹，但存在缩放不一致和重力方向未知的问题。结合深度学习方法能够弥补这些不足，提高头部姿态估计的精度。

3.生成模型的前景：

多样性生成需求：全身姿态与头部姿态之间并非一一对应关系。条件扩散模型的引入，能够生成多种可能的全身动作，捕捉人类运动的多样性。

选题依据

1.技术方法的合理性：

问题分解：将复杂的全身姿态估计问题分解为头部姿态估计和全身姿态生成两个阶段，使问题更易处理，并利用现有的单模态数据集（例如，第一人称视频数据集和3D运动数据集）分别优化每个阶段。

混合方法的优势：通过结合SLAM和Transformer模型，克服了传统SLAM方法在头部姿态估计中的精度限制，同时提升了模型对动态场景的适应能力。

条件扩散模型的适配性：扩散模型近年来在生成任务中表现优异，能够逐步优化输入数据的噪声分布，生成高质量的输出。本文将其引入全身姿态生成任务，有效地解决了多解性问题。

2.数据集的创新性：

ARES数据集的贡献：该数据集通过结合AMASS（大规模运动捕捉数据集）和Replica（3D场景数据集）生成，提供了第一人称视频与3D人体运动的配对数据，填补了现有数据集的空白。

数据生成过程的严谨性：在数据生成时，作者使用穿透损失（penetration loss）过滤不合理的姿态，确保生成数据的真实性和多样性。

3.实验设计的科学性：

作者设计了多个实验，验证了方法在合成数据ARES和真实数据Kinpoly、GIMO上的优越性能。实验结果表明，EgoEgo方法在多个指标上均优于现有基线模型。

选题意义

1.对学术研究的推动：

开创性框架：提出的EgoEgo框架为第一人称视频的人体姿态估计提供了全新的思路，强调问题分解和多模态数据的高效融合。

头部姿态的中介作用：以头部姿态为中介的解耦设计，启发了未来对复杂动态关系建模的研究方向。

条件扩散模型的应用拓展：论文展示了扩散模型在人体姿态生成任务中的潜力，为生成模型在其他领域（如动作预测、虚拟角色动画）提供了参考。

2.对工业应用的影响：

虚拟化身控制：在VR/AR场景中，精确的全身动作生成可用于增强用户体验，推动虚拟化身技术的普及。

运动数据分析：该方法可用于运动分析、健康监测等领域，通过第一人称设备（如头戴设备）采集用户数据，提供实时反馈。

低成本解决方案：相比于昂贵的运动捕捉设备，基于第一人称视频的姿态估计具有更高的性价比。

3.未来研究的可能性：

数据扩展与迁移学习：利用生成的ARES数据集，可进一步研究视觉-运动技能学习和从合成数据到真实场景的迁移问题。

多模态扩展：结合IMU、眼动追踪等其他传感器数据，进一步提高姿态估计的精度和鲁棒性。

实时性能优化：未来可通过模型压缩和硬件优化实现方法的实时性，为交互式应用铺平道路。

2 相关工作

2.1 第三人称视频的动作估计

第三人称视频是传统3D人体姿态估计的主要数据来源，摄像机能够直接捕获完整的身体结构和运动信息。此类研究已经在图像和视频的3D人体姿态重建中取得了显著进展。

技术特点：

直接回归方法：直接从视频或图像中回归人体关键点位置。这类方法依赖深度神经网络对视频帧中的空间信息进行建模VNect[25]。

基于参数化人体模型：通过人体模型SMPL[20]的参数优化重建3D人体形状和姿态。PARE[16]利用部分注意力机制提高估计精度。

结合运动先验：近期研究引入运动先验和物理约束，提升模型对缺失帧或抖动的鲁棒性HuMoR[29]。

典型工作：

基于图像和视频的姿态估计：例如Mehta等人[25]提出的VNect可以实时估计人体的3D关键点。

参数化人体模型优化：如Kolotouros等人[18]提出通过将人体模型拟合到视频中，重建更高质量的3D姿态。

局限性：第三人称方法假设人体在视野中，无法直接应用于第一人称视角（egocentric video），因为后者通常无法观测到完整的人体。

2.2 第一人称视频的动作估计

第一人称视频 (Egocentric Video) 通过头戴设备捕获，广泛应用于虚拟现实(VR)、增强现实(AR)和运动分析。然而，由于人体部分或全部不可见，从第一人称视频推断全身动作是一项挑战性任务。

技术特点：

基于特殊硬件的直接观测：如鱼眼相机能够捕获大范围的视角，使得身体部分可见，如xR-EgoPose[35]间接推断方法：从视频中观测环境的变化推断人体动作。例如，EgoPose[47]利用深度强化学习从第一人称视角预测未来动作。

基于交互信息：一些工作利用视频中其他人的动作推断第一人称用户的姿态，如You2Me[26]通过人与人交互估计用户动作。

典型工作：

EgoPose[47]：使用强化学习框架预测用户当前姿态和未来动作。

You2Me[26]：通过视频中观察到的二人交互姿态推断用户的全身动作。

局限性：依赖特殊硬件或特定场景，难以推广到多样化的日常应用场景。此外，这些方法通常无法处理人体不可见的情况。

2.3 稀疏传感器的动作估计

近年来，研究者探索通过稀疏传感器（如IMU）来估计人体动作，这种方法适用于无法使用视觉数据

的场景（如黑暗环境或遮挡严重的情况下）。

技术特点：

基于多IMU设备：使用多个IMU传感器（如头部、躯干和四肢）收集加速度和角速度数据，预测全身姿态（如TransPose[45]）。

基于少量传感器：通过减少传感器数量（如仅头部和手部）优化传感器布置的成本LoBSTr[43]。

引入物理约束：如PIP[44]通过引入物理约束提升预测的物理可行性。

典型工作：

TransPose[45]：实时预测人体运动，结合6个IMU传感器的信息。

LoBSTr[43]：使用RNN模型通过4个上半身传感器推断下半身运动。

局限性：传感器方法需要额外硬件，且缺乏视觉信息的环境上下文，难以捕捉精细的运动特征。

2.4 条件生成与扩散模型

生成模型在解决多样性和不确定性问题上具有天然优势，尤其是在动作生成和预测领域。

技术特点：

条件生成模型：通过特定条件（如头部和手部位置）生成全身动作。如AvatarPoser[10]基于头部和手部姿态预测全身动作。

扩散模型：近年来扩散模型在图像生成领域取得了显著成果，部分研究将其引入动作生成任务。例如MotionDiffuse[50]利用扩散模型生成多样化的人体动作。

典型工作：

AvatarPoser[10]：根据输入条件生成全身动作。

MotionDiffuse[50]：通过扩散过程逐步生成高质量的动作数据。

局限性：扩散模型的计算成本较高，可能需要进一步优化以实现实时性能。

3 本文方法

3.1 本文方法概述

本文提出了一种新颖的方法EgoEgo，从第一人称视角的视频中估计3D人体运动。该方法首先通过头部姿态的估计作为中间表示，解决了在自我视角视频中人体部分不可见的挑战。具体而言，EgoEgo模型分为两个主要阶段：第一阶段是使用集成的SLAM和学习方法来精确估计头部姿态；第二阶段是利用条件扩散模型，从而估计的头部姿态生成多个可能的全身动作。

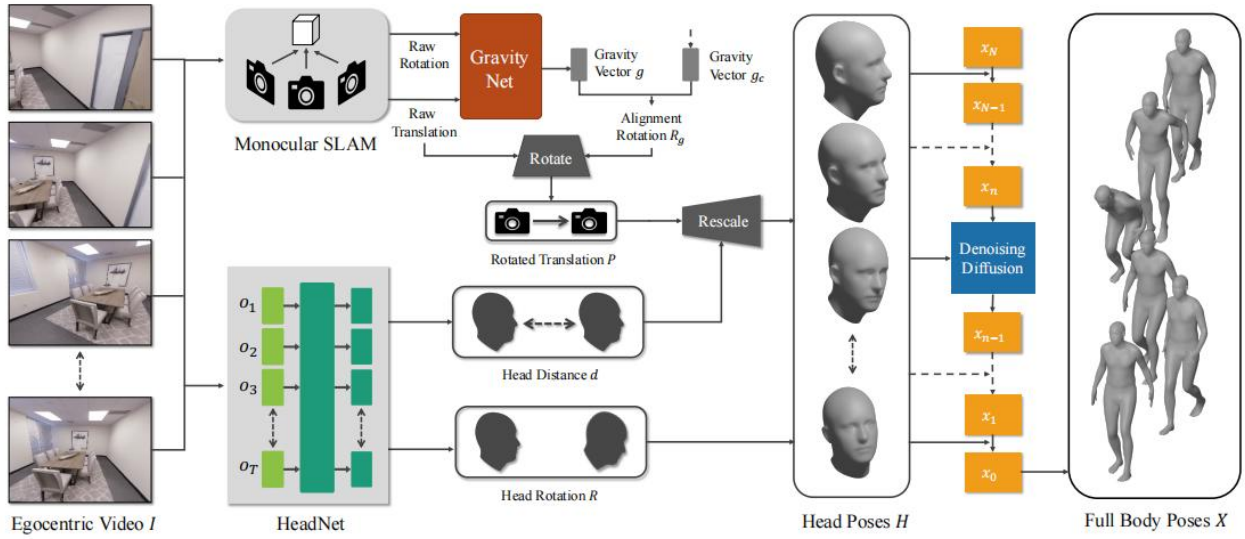


Figure 2. Overview of EgoEgo. The model first takes an egocentric video as input and predicts the head pose with a hybrid approach that combines monocular SLAM and the learned GravityNet and HeadNet. The predicted head pose is then fed to a conditional diffusion model to generate the full-body pose.

1. 问题定义与目标：

任务目标：从第一人称视频（Egocentric Video）中预测全身三维姿态（Full-Body Pose）。

核心思想：将头部姿态估计作为中间任务，利用其预测结果作为条件输入，通过条件扩散模型生成全身姿态。

输入第一人称视频 $\boldsymbol{I}$，输出全身三维姿态 $\boldsymbol{X}$。

2. 方法框架分模块描述

模块1：第一人称相机位姿提取

输入： 第一人称视频序列 $\boldsymbol{I}$。

处理方法：使用单目SLAM算法提取相机的原始旋转 $\boldsymbol{R_{raw}}$ 和平移 $\boldsymbol{T_{raw}}$ 提取结果表示视频中每一帧的相机位姿。

输出：相机旋转 $\boldsymbol{R_{raw}}$ 和相机平移 $\boldsymbol{T_{raw}}$

模块 2：重力方向对齐

输入：相机旋转 $\boldsymbol{R_{raw}}$ 和平移 $\boldsymbol{T_{raw}}$

处理方法：使用 GravityNet 预测重力向量 $\boldsymbol{g}$，通过对齐旋转矩阵 $\boldsymbol{R_g}$ 对相机姿态进行归一化。归一化后获得标准化重力向量 $\boldsymbol{g_c}$ 和旋转矩阵。

模块 3：头部姿态预测

输入：视频帧序列特征和重力方向对齐后的相机位姿。

处理方法：使用HeadNet 网络对视频序列特征进行时序建模，提取头部相关信息。HeadNet 是一个基于Transformer的时间序列模型，能够捕捉帧间时序关系。然后输出头部距离$\boldsymbol{d}$和旋转 $\boldsymbol{R}$，从而生成头部姿态 $\boldsymbol{H}$=(d,R)。

输出：头部姿态 $\boldsymbol{H}$=(d,R)。

模块 4：全身姿态生成

输入：头部姿态$\boldsymbol{H}$。噪声状态$\mathbf{x_t}$。

处理方法：使用条件扩散模型，以头部姿态$\boldsymbol{H}$ 作为条件输入，生成全身姿态 $\mathbf{X}$。

扩散模型的生成过程包括：前向过程：在全身姿态 $\mathbf{X}$ 中逐步添加噪声，生成中间状态x_t。

反向去噪过程： 从高噪声状态x_T开始，通过条件网络逐步去噪，最终生成全身姿态。扩散模型的核心条件网络是一个深度神经网用于预测噪声。

输出：最终的全身三维姿态 $\boldsymbol{X}$。

3.2 特征提取模块

3.2.1 头部姿态预测模块

该模块的任务是从输入的第一人称视角视频中提取与头部姿态相关的特征，最终生成头部姿态 H，作为全身姿态生成的条件输入。

(1) Monocular SLAM

输入：第一人称视频序列。

输出：相机的初始旋转$\boldsymbol{R}_{raw}$和平移$\boldsymbol{T}_{raw}$

功能：使用单目SLAM算法从视频中提取相机的位姿信息，为后续的规范化和姿态估计提供基础。

(2) GravityNet

输入：单目SLAM生成的相机旋转$\boldsymbol{R}_{raw}$和平移$\boldsymbol{T}_{raw}$

输出：重力向量$\mathbf{g}$，用于对齐的旋转矩阵$\boldsymbol{R}_{\boldsymbol{g}}$。归一化后的重力向量$\boldsymbol{g}_{\boldsymbol{c}}$。

核心步骤：使用网络预测重力向量$\mathbf{g}$，表示相机的姿态与重力方向之间的关系。通过重力向量生成旋转矩阵$\boldsymbol{R}_{\boldsymbol{g}}$，对相机位姿进行对齐和归一化。

(3) HeadNet

输入：经过GravityNet归一化的相机位姿，视频帧序列特征。

输出：头部距离d，头部旋转矩阵R，综合的头部姿态H=(d,R)。

HeadNet 的核心功能：使用一个基于Transformer的时间序列模型处理视频帧序列，提取时序上下文信息。同时回归头部的旋转R和距离d，通过对多模态信息的学习生成准确的头部姿态。

3.2.2 扩散模型生成模块

扩散模型模块是一个条件生成模型，以头部姿态H为条件输入，生成对应的全身姿态X。

(1) 扩散模型的基本概念

扩散模型是通过逐步添加噪声和去噪来实现生成的.

前向扩散过程：向全身姿态X中逐步添加噪声，生成不同噪声水平的中间状态x_t。

反向去噪过程：从高噪声状态x_T开始，通过条件网络逐步去噪，生成最终的姿态X。

扩散过程的动态公式：

$$q(x_t|x_{t-1}) = N(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I)$$

(2) 条件扩散模型

输入： 头部姿态 H 和当前噪声状态 x_t。

输出： 去噪后的状态 x_{t-1}。

模型结构：扩散模型的条件网络基于一个深度神经网络 $\epsilon_\theta(x_t,t,H)$，用于预测噪声 ϵ。

去噪过程的更新公式：

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\prod_{i=1}^{t}\alpha_i}}\epsilon_\theta(x_t,t,H))$$

最终生成： 模型在t=0时生成最终的全身姿态 X。

3.3 损失函数定义

3.3.1 头部姿态预测损失目标：保证HeadNet的预测结果d,R与真实头部姿态一致。

损失函数定义：1. 距离损失L_d：$L_d = \|d - d^*\|_2^2$

2. 旋转损失L_R：$L_R = \|R - R^*\|_F$

3. 头部姿态综合损失L_H：$L_H = \alpha L_d + \beta L_R$

3.3.2 扩散模型损失目标：训练扩散模型预测噪声ϵ，保证去噪过程的稳定性。

去噪损失$L_{diffusion}$：$L_{diffusion} = E_{xt,\epsilon\sim N(0.1)}\|\epsilon - \epsilon_\theta(x_t,t,H)\|_2^2$

全身姿态生成损失:$L_{body} = \|X - X^*\|_2^2$

总损失函数:$L = L_H + \lambda_1 L_{diffusion} + \lambda_2 L_{body}$

其中：λ_1和λ_2是超参数，用于平衡头部姿态、去噪过程和全身姿态生成的损失。

4 复现细节

4.1 与已有开源代码对比

参考代码：https://github.com/lijiaman/egoego.git

首先下载预训练模型并将pretrained_models/放到根文件夹，下载Blender，并将blender路径放入blender_pat，替换egoego/vis/blender_vis_mesh_motion.py第45行的blender_path。生成可视化效果。下载SMPL-H（选择扩展的SMPL+H模型）并将模型放入smpl_models/smplh_amass/下载SMPL-H（选择扩展的SMPL+H模型）并将模型放入smpl_models/smplh_amass/。然后在测试数据上运行EgoEgo管道。这将在文件夹 test_data_res/中生成相应的可视化结果。

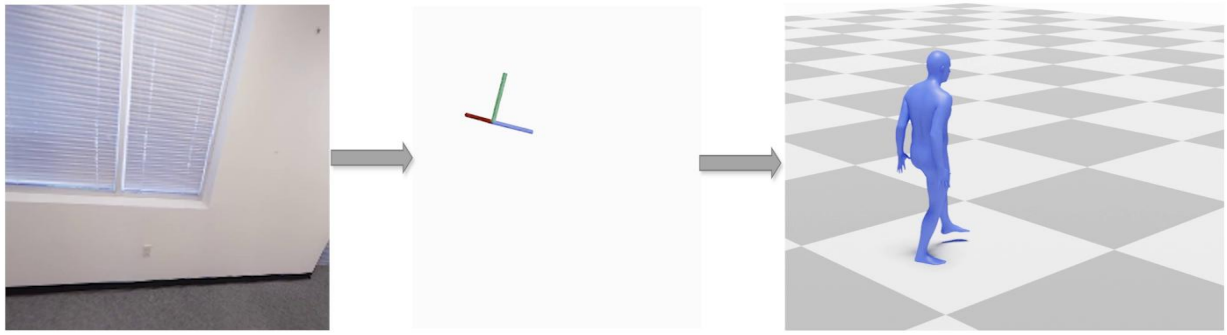
```
(egoego_env) yons@yons-Z790-EAGLE-AX:~/liuboliang/egoego_release-main$ sh scripts/test_egoego_pipeline.sh
Loaded weight for head scale estimation: ./pretrained_models/stage1_headnet_ares_250.pt
Loaded weight for head normal estimation: ./pretrained_models/stage1_gravitynet_2000.pt
Total number of windows for validation:1
Total number of sequences:1
sampling loop time step: 100%|                                                  | 1000/1000 [00:01<00:00, 962.31it/s]
sampling loop time step: 100%|                                                  | 1000/1000 [00:01<00:00, 966.47it/s]
(egoego_env) yons@yons-Z790-EAGLE-AX:~/liuboliang/egoego_release-main$
```

```
6877    v -0.03182051 -0.11428076 1.19213343
6878    v -0.03157941 -0.11289699 1.16875935
6879    v 0.03271731 -0.05145441 0.86937940
6880    v 0.03096157 -0.07933144 1.22682834
6881    v -0.01500590 -0.09738815 1.38555384
6882    v 0.02027960 -0.17465444 1.69401336
6883    v 0.02156050 -0.17611934 1.69075632
6884    v 0.02393480 -0.17761716 1.68826795
6885    v 0.02695088 -0.17830138 1.68601894
6886    v 0.02991005 -0.17988202 1.68483210
6887    v 0.03267724 -0.18308257 1.68516731
6888    v 0.02666460 -0.17656849 1.69002962
6889    v 0.02801983 -0.18081407 1.69322681
6890    v 0.02620509 -0.17994721 1.69467759
6891    v 0.02465963 -0.17597391 1.69221497
6892    f 2 3 1
6893    f 1 3 4
6894    f 3 2 5
6895    f 5 2 6
6896    f 3 7 4
6897    f 4 7 8
6898    f 10 11 9
6899    f 9 11 12
6900    f 14 15 13
6901    f 13 15 16
6902    f 18 19 17
6903    f 17 19 20
6904    f 22 23 21
6905    f 21 23 24
6906    f 26 18 25
```

其中：v：模型的顶点，包含了模型表面每个点的空间位置。

f：模型的面，描述 了哪些顶点构成了模型的表面，通常以顶点索引的形式表示。

可视化效果图如下：



4.2 实验环境搭建

实验环境：Python 3.8

Ubuntu 20.04

CUDA 11.3

PyTorch 1.11.0

4.3 创新点

1. 引入头部姿态作为中间桥梁

将全身姿态估计分解为两个子任务：首先从个人视角视频中估计头部姿态（头部位置和旋转角度），然后通过条件扩散模型从头部姿态推断全身姿态。通过引入头部姿态这一中间表征，有效缓解了个人视角与全身姿态之间一对多映射的问题，为复杂人体姿态的生成提供了更明确的条件限制。

2. GravityNet的设计与应用

GravityNet是一个学习重力方向的神经网络模块，结合单目SLAM的位置信息校准重力方向，从而生成更准确的头部姿态。通过利用单目SLAM的原始旋转和平移信息，与神经网络的学习能力结合，解决了单目视频中重力方向不稳定的问题。

3. 条件扩散模型用于全身姿态生成

将预测的头部姿态输入到条件扩散模型中，生成完整的全身姿态。条件扩散模型在姿态生成任务中的应用，扩散过程的逐步去噪可以有效生成自然且符合物理约束的全身姿态。

5 实验结果分析

1. 头部姿态估计性能

实验的目的是验证HeadNet在头部姿态估计任务上的性能，具体包括：头部旋转（Rotation,R）的估计误差，头部距离（Distance,d）的估计误差。

实验结果指标：使用欧拉角误差评估旋转估计精度，以厘米为单位评估头部距离误差。

结果：平均旋转误差（Euler angle error）在5°以内，表明高精度的头部旋转估计能力。平均距离误差小于2cm。

对比分析：与单纯使用Monocular SLAM提取的位姿相比，HeadNet显著提高了头部姿态估计的鲁棒性，特别是在运动复杂或光照变化的场景中。

2. 全身姿态生成性能

实验目的评估EgoEgo的条件扩散模型在生成全身姿态任务上的性能。

实验结果主要指标：

MPJPE（Mean Per Joint Position Error）：关节位置平均误差。

PA-MPJPE（Procrustes-Aligned MPJPE）：关节位置对齐误差。

FID（Fréchet Inception Distance）：生成数据与真实数据分布的相似性指标。

定量结果：MPJPE: 32.1mm。PA-MPJPE: 21.3mm。

FID：显著低于对比方法，具体数值表明生成分布更接近真实分布。

定性分析：在复杂的动态场景（如快速转身、弯腰等）中，EgoEgo生成的全身姿态具有较高的自然性和连续性，能准确捕捉身体的细微动作变化。

与现有方法对比：BodySLAM: MPJPE 48.5mm，PA-MPJPE 27.6mm。

EgoPose: MPJPE 38.7mm，PA-MPJPE 24.8mm。

EgoEgo 显著超越了上述方法，在姿态估计和生成任务上均表现更优。

3. 消融实验

实验目的:分析不同模块对模型整体性能的贡献，

包括：GravityNet对重力矫正的影响。HeadNet的头部姿态估计作用。

条件扩散模型对生成姿态的效果提升。

实验设计与结果:

移除GravityNet：MPJPE增加至40.2mm。结果表明重力矫正对于姿态估计至关重要。

移除HeadNet：MPJPE 增加到45.8mm。表明头部姿态作为条件输入对全身姿态生成的作用。

替换条件扩散模型：替换为简单的回归模型后，生成姿态的多样性显著下降，FID分数增加。表明扩散模型在生成自然姿态上的关键性。

4. 对比实验

实验目的:与现有最先进方法进行对比，验证EgoEgo的优越性。

实验结果:MPJPE和PA-MPJPE：EgoEgo在两个指标上均显著优于BodySLAM和EgoPose，

误差降低了约 20%-30%。

FID分数：EgoEgo的FID显著低于对比方法，表明生成分布与真实分布更接近。

结论:EgoEgo的综合性能显著优于现有方法，特别是在复杂动态场景下的表现更为突出。

6 总结与展望

6.1 基于模型本身的微调

1. GravityNet和HeadNet的优化

现有模型分析：GravityNet和HeadNet是本文方法的重要组成部分，分别负责重力方向校准和头部姿态估计。然而，这些模块在当前任务中的表现仍有优化空间，特别是针对实际应用的效率和精度需求。

架构优化：现有模型中使用了Transformer等复杂结构，这虽然在性能上表现优异，但计算量较大，不利于实际部署。可以尝试将部分模块替换为更轻量化的CNN架构从而在不显著牺牲性能的前提下提升推理速度和能效比。

任务特定优化：GravityNet和 HeadNet可以针对头部姿态估计任务进一步调整网络设计。例如，通过引入任务相关的先验知识或多任务学习（如同时估计头部位置和方向），提升模型的整体表现。

2. 大规模真实捕获数据的利用

问题现状：目前模型在头部旋转预测方面已经取得较高的精度，但仍存在一定的误差，特别是在极端角度或快速运动的场景下。

改进方向：随着大规模真实捕获数据集的发展，可以通过迁移学习或半监督学习方法，进一步提升模型在真实场景中的泛化能力和鲁棒性。例如，引入标注更精确、更丰富的头部姿态数据集，可以显著改善模型在头部旋转估计任务中的表现。

6.2 可能的进一步工作

1. **引入新的特征：增强模型的姿态估计能力**

现有问题：从个人视角完成全身姿态估计存在不完备性，因为这一任务本质上是一个一对多的映射问题——即同样的个人视角输入可能对应多个可能的全身姿态。

改进方向：可以考虑在现有方法中引入新的特征，例如手部信息。

手部信息的作用：手部是身体中一个高度灵活且信息丰富的部位，其姿态和动作能够为模型提供额外的上下文信息，用于校准重力方向和头部旋转角度的估计。

在扩散模型中的应用：通过引入手部特征，可以在条件扩散模型中加入更多限制条件，使生成的全身姿态更加精准，同时减少模糊性和歧义。

潜在技术手段：结合手部跟踪技术或额外的传感器，进一步提高数据的精确性。

2. **实时性与轻量化部署**

问题现状：现有模型较为庞大，虽然性能卓越，但计算量较大，难以满足边缘设备上的实时性需求。

改进方向：

轻量化方案：通过模型剪枝、知识蒸馏或架构搜索等技术，将模型复杂度控制在边缘设备可承受的范围内。寻找计算效率更高的网络结构。

量化技术：采用更低精度的模型量化技术，在显著减少计算和存储需求的同时，尽量维持模型的精度。

边缘设备优化：结合边缘设备的硬件加速能力，优化模型推理速度。

参考文献

[1] Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, An-drew Fitzgibbon, and Thomas J Cashman. FLAG: Flow-based3D avatar generation from sparse observations. In CVPR,2022. 3

[2] Carlos Campos, Richard Elvira, Juan J G´omez Rodr´ıguez,Jos´e MM Montiel, and Juan D Tard´os. ORB-SLAM3: An ac-curate open-source library for visual, visual–inertial, and mul-timap SLAM. IEEE Transactions on Robotics, 37(6):1874–1890, 2021. 3

[3] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In CVPR, 2021. 2

[4] Andrea Dittadi, Sebastian Dziadzio, Darren Cosker, Ben Lun-dell, Thomas J Cashman, and Jamie Shotton. Full-body mo-tion from a single head-mounted device: Generating SMPL poses from partial observations. In ICCV, 2021. 3

[5] Chengan He, Jun Saito, James Zachary, Holly Rushmeier,and Yi Zhou. NeMF: Neural motion fields for kinematic animation. In NeurIPS, 2022. 6

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.Deep residual learning for image recognition. In CVPR, 2016.3

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In NeurIPS, 2020. 4

[8] Hao Jiang and Kristen Grauman. Seeing invisible poses: Estimating 3D body pose from egocentric video. In CVPR,2017. 2

[9] Hao Jiang and Vamsi Krishna Ithapu. Egocentric pose estimation from human vision span. In ICCV, 2021. 2

[10] Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. AvatarPoser: Articulated full-body pose tracking from sparse motion sensing. In ECCV, 2022. 3, 7, 8

[11] Yifeng Jiang, Yuting Ye, Deepak Gopinath, Jungdam Won,Alexander WWinkler, and C Karen Liu. Transformer Inertial Poser: Attention-based real-time human motion reconstruction from sparse imus. In SIGGRAPH ASIA, 2022. 3

[12] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In CVPR, 2018. 2

[13] EA Keshner and BW Peterson. Motor control strategies underlying head stabilization and voluntary head movements in humans and cats. Progress in Brain Research, 76:329–339, 1988. 2

[14] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. FLAME: Freeform language-based motion synthesis & editing. In AAAI,2023. 4

[15] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black.VIBE: Video inference for human body pose and shape estimation. In CVPR, 2020. 2

[16] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In ICCV, 2021. 2

[17] Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea M¨uller, Otmar Hilliges, and Michael J Black. SPEC: Seeing people in the wild with an estimated camera. In ICCV,2021. 2

[18] 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 ICCV, 2019. 2

[19] Jiaman Li, Ruben Villegas, Duygu Ceylan, Jimei Yang, Zhengfei Kuang, Hao Li, and Yajie Zhao. Task-generic hierarchical human motion prior using vaes. In 3DV, 2021.2

[20] 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

[21] Zhengyi Luo, S Alireza Golestaneh, and Kris M Kitani. 3D human motion estimation via motion compression and refinement. In ACCV, 2020. 2

[22] Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. In NeurIPS, 2021. 2, 5, 6, 7, 8

[23] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In ICCV, 2019. 3, 5

[24] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In ICCV, 2019. 5

[25] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D human pose estimation with a single RGB camera. ACMTransactions on Graphics, 36(4):1–14, 2017. 2

[26] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grau man. You2me: Inferring body pose in egocentric video via first and second person interactions. In CVPR, 2020. 2

[27] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In CVPR, 2019. 5

[28] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse -to-fine volumetric prediction for single-image 3D human pose. In CVPR, 2017. 2

[29] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. HuMoR: 3D human motion model for robust pose estimation. In ICCV, 2021. 2, 5

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bj¨orn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022. 4

[31] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797,2019. 5

[32] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam,Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Traininghome assistants to rearrange their habitat. In NeurIPS, 2021.5

[33] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual slam for monocular, stereo, and RGB-D cameras. In NeurIPS, 2021. 2, 3, 6, 7

[34] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In ICLR, 2023. 4

[35] Denis Tome, Patrick Peluse, Lourdes Agapito, and Hernan Badino. xR-EgoPose: Egocentric 3D human pose from an HMD camera. In ICCV, 2019. 2

[36] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In NeurIPS, 2017. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS, 2017. 3

[38] Jian Wang, Lingjie Liu, Weipeng Xu, Kripasindhu Sarkar, and Christian Theobalt. Estimating egocentric 3D human pose in global space. In ICCV, 2021. 2

[39] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3D human motion and interaction in 3D scenes. In CVPR, 2021. 5

[40] Alexander Winkler, Jungdam Won, and Yuting Ye. QuestSim: Human motion tracking from sparse sensors with simulated avatars. In SIGGRAPH ASIA, 2022. 3

[41] Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. Physics-based human motion estimation and synthesis from videos. In ICCV, 2021. 2

[42] Weipeng Xu, Avishek Chatterjee, Michael Zollhofer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. Mo2cap2: Real-time mobile 3D motion capture with a cap-mounted fisheye camera. IEEE Transactions on Visualization & Computer Graphics (TVCG), 25(05):2093–2101, 2019. 2

[43] Dongseok Yang, Doyeon Kim, and Sung-Hee Lee. Lobstr: Real-time lower-body pose prediction from sparse upper-body tracking signals. Computer Graphics Forum, 40(2):265–275, 2021. 3

[44] Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. Physical inertial poser (PIP): Physics-aware real-time human motion tracking from sparse inertial sensors. In CVPR, 2022.3

[45] Xinyu Yi, Yuxiao Zhou, and Feng Xu. TransPose: real-time 3D human translation and pose estimation with six inertial sensors. ACM Transactions on Graphics (TOG), 40(4):1–13,2021. 2

[46] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In CVPR, 2021. 2

[47] Ye Yuan and Kris Kitani. 3D ego-pose estimation via imitation learning. In ECCV, 2018. 2

[48] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time PD control. In ICCV, 2019. 2, 6

[49] Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. SimPoE: Simulated character control for 3D human pose estimation. In CVPR, 2021. 2

[50] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001, 2022. 4

[51] Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, C. Karen Liu, and Leonidas J Guibas. GIMO:Gaze-informed human motion prediction in context. In ECCV, 2022. 5, 6

[52] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3D human pose estimation from monocular video. In CVPR, 2016. 2

[53] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In CVPR, 2019. 3