

基于视频内人体细粒度部位信息注意力分配的 三维人体网格重建

龚永义 秦哲 丁若尧

(广东外语外贸大学信息科学与技术学院 广州 510006)

摘要 当视频中人物的某些身体部位超出镜头(EFoV, Exceeding Field of View)时, 给基于视频的三维人体网格建模任务带来了新的挑战。本文中我们提出视频内细粒度人体部位信息注意力分配(BAA, Body part Attention Allocation)模型框架, 通过选用人体虚拟标记(Virtual Markers)作为三维人体的中间表示, 将基于 EFoV 视频的三维人体网格建模问题转化为人体虚拟标记点集合序列的恢复问题。具体的, 首先, 我们设计了 Rand Mask 策略, 通过对完整的人体虚拟标记点集合按人体部位进行随机遮蔽, 模拟视频中人体部位超出镜头时导致的信息缺失。其次, 设计置信度增强过滤模块(CEF, Confidence Enhanced and Filtering), 对前置方法估计的虚拟标记置信度进行过滤增强, 进而摆脱对前置方法的过度依赖。最后, 设计虚拟标记注意力分配模块(AIVM, Attention In Virtual Markers), 根据增强后的虚拟标记置信度情况, 为各虚拟标记点分配合适的注意力, 进而合理地利用序列中虚拟标记的互补关系恢复出完整的虚拟标记序列, 并以此估计最终的 3D 人体网格模型。我们将三维运动捕捉(MoCap)数据集中的 SMPL 人体网格进行采样, 得到对应的虚拟标记点集合数据, 人体 SMPL 模型与虚拟标记点集合共同构成了模型训练中的数据集。

关键词 三维人体网格重建; 虚拟标记; 注意力分配; 视频信息缺失

中图法分类号 TP391 DOI号 10.11897/SP.J.1016.2026.00622

Video-Based 3D Human Mesh Reconstruction with Fine-Grained Body Part Attention Allocation

GONG Yong-Yi QIN Zhe DING Ruo-Yao

(School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006)

Abstract As a key task in computer vision, video-based 3D human mesh modeling has wide-ranging applications in fields such as virtual reality, game development and movie Visual Effects production. In cases where human body information in video is fully available, existing research works have already achieved significant success. However, when certain body parts of a person in video Exceeding the Field of View (EFoV), new challenges arise for video-based 3D human mesh reconstruction, since certain human body information is missing. To address these challenges, in this paper, we propose a Body part Attention Allocation (BAA) framework for fine-grained human body part information in video. This framework adopts human body Virtual Markers as an intermediate representation of the 3D human body and performs part-level masking on these markers to simulate the body parts exceeding the field of view. Subsequently, by explicitly utilizing the masking information to allocate corresponding attention to the virtual markers in each frame, the framework fully leverages the complementary relationships between

body parts across video frames, restores a reasonable sequence of complete human body virtual marker sets, and thereby reformulates the 3D mesh reconstruction problem under the EFoV condition as a sequence recovery task of Virtual Markers sets. More specifically, we first design a Rand Mask strategy, which randomly masks subsets of a full body parts Virtual Markers set, to simulate the information loss caused by the body parts exceeding the field of view. Secondly, we develop a Confidence Enhanced and Filtering (CEF) module. This module receives the masking matrix output by the Rand Mask module, adds Gaussian noise, and adjusts the scale of the input data to ensure the consistency of different feature dimensions. After a series of enhancement and filtering processes, the module filters and enhances the virtual marker confidence scores estimated by the preceding method, thereby reducing excessive reliance on the preceding method. Finally, to fully apply the confidence information to fine-grained virtual markers, we propose an Attention In Virtual Markers (AIVM) module. Based on the enhanced virtual marker confidence, this module uses a multi-head cross-attention mechanism, which enables the model to parallelly enhance the relationships between virtual markers in each video frame through confidence information across multiple subspaces. Then, through a self-attention mechanism, the model further refines the virtual markers at the marker-level through self-enhancement. After the AIVM module allocates appropriate attention to each virtual marker, the complementary relationships of virtual markers in the sequence are used to restore the complete virtual marker sequence, based on which the final 3D human SMPL mesh model is estimated. Finally, we introduce a Skinned Multi-Person Linear (SMPL) regression module based on a multi-layer perceptron (MLP) to estimate the SMPL model parameter sequence from the complete human body virtual marker set sequence, and generate the final 3D human SMPL mesh model. We sample the SMPL human meshes from the 3D Motion Capture (MoCap) dataset to obtain the corresponding virtual marker set data. These Virtual Markers sets and their corresponding SMPL models are jointly used to supervise our model training.

Key words 3D human mesh reconstruction; virtual markers; attention allocation; missing video information

1 引言

基于视频的三维人体网格模型序列建模,在计算机图形学、计算机视觉等领域有重要应用。与基于图像进行三维人体网格建模^[1-12]不同,基于视频进行三维人体网格序列建模^[13-22]是一项更复杂的任务,人们不仅需要估计视频中每一帧的人体网格模型,还需要考虑序列中人物运动的连续性^[23]。

基于视频的人体网格模型估计工作已取得了一定的进展^[13-18],但是当视频中某些帧内人体信息不完整时,重建合理的完整人体模型依然存在一定挑战。视频内人体信息不完整的情形可分为两种:镜头内因遮挡导致的人体信息缺失(图 1(a))和人体超出镜头造成的信息缺失(图 1(b))。其中,镜头内遮挡造成的信息缺失是因外部物体遮挡,遮挡物的形状和位置可能提供线索,且被遮挡部分与帧内

可见部分的连贯性较大,可通过人体部位连续性等信息进行重建。而人体超出镜头时,由于镜头的物理限制导致超出视野范围的部分完全无法获取直接信息,需要依赖运动模型、历史帧信息及人体先验知识等来预测和重建超出部分。对于第一种情形,现有的数据集提供了包括镜头内完整人体信息的详尽标注,因此研究人员通过对这些数据集有监督的训练网络模型,通过分析帧内可见的人体信息来推断被遮挡部分的人体结构^[9],或者利用视频帧间人体运动的连续性来重建该帧内人体的完整信息^[13-14,17-18],最终达到重建人体网格的目的。对于第二种情形,即人体信息超出视角范围(Exceeding Field of View, EFoV)时,目前公开的数据集中极少包含镜头外的人体标注信息,限制了现有工作预测镜头外人体部位信息的能力,这给针对 EFoV 情形的研究造成了极大的困难。当视频中人体超出镜头时,现有的模型方法无法对此问题进行有效的处

理。例如 Ma 等人^[24]利用基于视频帧估计的虚拟标记直接进行人体网格重建，其主要关注镜头内可见人体的建模准确性，镜头外部位的处理方式主要是根据当前帧中可见部位进行简单推测，这会导致整体的建模结果出现偏差；Baradel 等人^[14]则主要是以初步推测出的人体网格作为基础，利用相近帧人体网格模型的连续性，填补或者修正错误的人体模型，以得到平滑和连贯的网格序列，这种方式过于依赖初步推测出的人体网格，当视频中较多帧处于 EFoV 状态时甚至无法重建出人体网格。因此，如何基于 EFoV 视频重建合理的完整三维人体网格模型仍然是一项急需解决的问题。

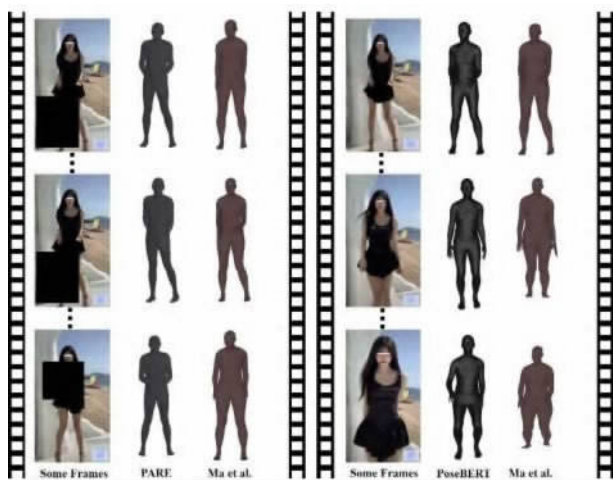


图 1 视频帧内人体信息不完整的两种情形

我们注意到 Ma 等人^[24]提出的人体虚拟标记具备两个特点：(1) 人体虚拟标记点集合包含了丰富且直接的人体姿势和形状信息，利用完整的人体虚拟标记点集合，能够可靠地估计出对应的 SMPL 模型参数；(2) 虚拟标记点集合与人体部位显式对应，可通过遮蔽虚拟标记点来模拟人体部位超出镜头时造成的信息缺失。由此，我们提出视频内细粒度人体部位信息注意力分配模型(BAA, Body part Attention Allocation)，选用虚拟标记点集合作为人体信息的中间表示，并对其进行遮蔽来模拟视频中超出镜头的部位，之后通过利用各帧虚拟标记点的互补关系，恢复出合理的完整人体虚拟标记点集合序列，从而将基于 EFoV 视频的人体网格建模问题的核心转化为虚拟标记点集合序列的恢复问题。

综上所述，我们的主要贡献如下：

(1) 提出 BAA 模型，用于在 EFoV 视频中进行合理的完整人体网格序列建模，并巧妙地将此建模

问题的关键转化为虚拟标记点集合序列的恢复问题。BAA 模型的核心包括置信度增强过滤模块 (Confidence Enhanced and Filtering, CEF) 和虚拟标记注意力分配模块 (Attention In Virtual Markers, AIVM)，其中 CEF 模块用于对前置方法估计的虚拟标记置信度进行过滤增强，AIVM 模块通过利用视频帧间人体虚拟标记点集合信息的互补关系恢复出合理的完整的人体虚拟标记点集合，最终由完整的虚拟标记信息估计对应的人体网格形状序列。

(2) 提出 Rand Mask 策略，通过将视频序列中人体虚拟标记点信息根据部位进行遮蔽来模拟 EFoV 视频中人体部位超出镜头的情形，并将遮蔽信息作为 CEF 模块的监督数据。

(3) 我们在 AMASS^[25]数据集的基础上，根据人体 SMPL^[26]参数化网格采样出其对应的人体的虚拟标记点集合^[24]，二者统一起来作为本文训练过程中使用的数据集，不需要任何繁琐的图像或视频姿态注释。

本文的其余部分组织如下：第 2 节是对相关研究工作的介绍；第 3 节介绍了我们整体的网络架构及遮蔽恢复的设计思路，并介绍了虚拟标记的基本知识；第 4 节对实验结果进行展示和分析；第 5 节对我们的工作进行了总结。

2 相关工作

在本节中，我们重点介绍与我们工作相关的一些研究进展，包括基于图像和视频的人体网格模型、姿态估计以及对应的序列生成任务等。

2.1 基于单张图像估计三维人体姿态或人体网格模型

基于单张图像的三维人体姿态或网格形状估计方法大多是采用端到端的方式，即从输入图像训练网络模型估计人体姿态、网格形状或者参数化模型的参数等。

估计参数化人体网格模型。Kanazawa 等人^[4]提出了一种端到端框架 HMR，用于从单个 RGB 图像重建人体的完整 3D 网格，在网络中引入一个经过训练的鉴别器，该鉴别器使用一个大型 3D 人体网格数据库来判断人体形状和姿态参数是否真实，缓解有限的三维标注问题。Pavlakos 等人^[5]提出了一种基于卷积网络的直接从单幅彩色图像中估计人体三维姿态和形状预测方法，通过 2D 关键点和轮廓

对人体网格参数进行预测, 在训练时逐顶点对生成的全身 3D 网格进行监督。Kolotouros 等人^[8]提出的方法 SPIN, 融合了回归方法和基于优化的方法来训练用于 3D 人体姿势和形状估计的深度网络。使用回归网络为优化例程提供初始估计, 然后将模型拟合到循环中, 并为网络的训练提供基于模型的监督, 优化模块和回归模块就形成了一个自我完善的循环, 二者紧密协作, 共同受益。Zhang 等人^[6]提出一种名为 DaNet 的分解聚合网络, 采用密集对应映射, 密集地在二维像素点和三维顶点之间架起一座桥梁(IUVmaps)作为中间表示, 便于二维到三维映射的学习。DaNet 的预测模块被分解为一个全局流和多个局部流, 分别实现对形状和姿态预测的全局和细粒度感知, 并提出了一种位置辅助旋转特征细化策略, 以利用身体关节之间的空间关系。Kocabas 等人^[9]提出了部位引导注意机制(PARE), 通过利用关于单个身体部位的可见性的信息, 同时利用来自邻近身体部位的信息来预测被遮挡的部位, 实现了在图像中存在遮挡的情况下对人体建模的需求。

估计非参数化人体网格。Kolotouros 等人^[3]提出了一种名为 Graph-CMR 的基于 Graph-CNN 的方法, 保留 SMPL 模板网格的拓扑结构, 将基于图像的特征附加到网格顶点上, 每个顶点的回归目标是其对应的人体模型的 3D 点位置。在恢复了网格的完整 3D 几何结构后, 如果仍然需要特定的模型参数化, 可以从顶点位置可靠地回归出 SMPL 模型参数。Lin 等人^[7]提出一种基于 Transform 的方法 METRO, 从单个输入图像重建人体姿势和网格。METRO 使用 Transform 编码器联合建模顶点-顶点和顶点-关节的相互作用, 并同时输出 3D 关节坐标和网格顶点。Ma 等人^[24]提出一种人体姿态中间表示方式——虚拟标记, 用于人体捕捉, 并通过虚拟标记表示中的系数矩阵编码了网格顶点之间的空间关系, 进而得到对应的人体网格。

估计三维人体骨架姿态。与二维人体姿态估计相比^[27], 三维人体姿态估计面临着诸多更为复杂的挑战。尽管如此, 仍有一些研究取得了显著进展。例如, Ke 等人^[28]提出了 DetPoseNet, 一个由粗到细的多尺度的端到端多人检测和姿态估计框架。Li 等人^[29]提出的 PolarPose, 在极坐标下进行二维回归, 通过将二维偏移回归任务简化为分类任务来解决长期回归的挑战, 从而实现更精确的关键点定位。Hassan 等人^[30]提出了一种高阶图网络, 利用规则矩

阵分裂结合权值和邻接调制进行三维人体姿态估计, 不仅捕捉身体关节之间的长期依赖关系, 还捕捉相邻关节和远处关节之间的不同关系。Cheng 等人^[31]介绍了一种基于双网络的单目视频多人姿态估计方法, 综合了自上而下和自下而上两种姿态估计方法的优势, 进一步提高姿态估计的准确性。

尽管上述方法对静态图像是有效的, 但将他们应用于视频时, 很难在视频序列中产生时间上相关和平滑的三维人体姿态或网格模型, 即可能发生抖动、不连续、不合理的三维人体运动。

2.2 基于视频估计三维人体姿态或网格形状序列

与基于图像的方法不同, 基于视频的方法不仅需要估计出人体姿态或者网格信息, 还需要考虑视频各帧之间的关系, 以防止出现抖动、不连贯等问题。

估计人体三维网格模型序列。Wei 等人^[13]通过分层注意集成的方式有效结合了相邻的过去和未来的特征表示, 利用从人类运动中观察到的视觉线索, 自适应地重新校准序列中需要注意的范围, 以捕捉运动的连续性依赖性, 并加强时间相关性和细化特征表示, 产生时间相干的人体模型估计。Baradel 等人^[14]提出一种后处理的模型 PoseBERT, 通过对 MoCap 数据加入噪声和掩码的方式进行掩码建模和去噪训练, 在视频帧中利用注意力来推测各帧中人体信息, 以至于视频中部分帧中人体缺失的情况下也可以重建出连贯的人体模型。Kanazawa 等人^[15]提出了一种基于卷积的时间编码器, 通过进一步估计相邻的过去和未来帧中的 SMPL 参数来学习人体运动运动学。Sun 等人^[16]提出了一个骨架分离框架, 将任务分为多层次的空间和时间子问题。他们进一步提出了一种无监督的对抗性训练策略, 即时间打乱和顺序恢复, 以鼓励时间特征学习。Kocabas 等人^[17]提出了一种由双向门控循环单元(GRU)组成的时间编码器, 将静态特征编码为一系列时间相关的潜在特征, 用于回归 SMPL 模型参数。他们进一步整合了对抗性训练策略, 利用 AMASS 数据集^[25]来区分真实的人体运动和由回归器估计的运动, 以鼓励生成合理的 3D 人体运动。Doersch 等人^[19]通过结合 CNN 和长短期记忆(LSTM)网络, 利用二维关键点热图和光流序列的信息对模型进行训练, 证明了考虑预处理的运动信息可以提高 SMPL 参数估计。Luo 等人^[32]提出了一个两阶段模型, 首先通过变分运动估计器来估计粗糙的三维人

体运动, 然后使用运动残差回归器来细化运动估计。Choi 等人^[18]提出了一种时间一致网格恢复 (TCMR) 系统, 该系统使用基于 GRU 的时间编码器和三种不同的编码策略, 以鼓励网络更好地学习时间特征。此外, 他们还提出了一种时间特征集成方案, 用于结合了三个时间编码器的输出, 以帮助 SMPL 参数回归器估计准确和平滑的三维人体姿态和形状。

估计三维人体骨架姿态序列。Li 等人^[33]提出了一个名为 MHFormer 的多假设转换器, 它学习了多个看似合理的假设的时空表示。生成多个初始假设表示后将多个假设合并为一个收敛表示, 然后再将其划分为几个发散的假设, 然后再学习跨假设交流, 聚合多假设特征, 合成最终的 3D 姿态。用于解决从单目视频中估计 3D 人体姿势时存在多个可行解(假设)的问题。Xue 等人^[34]提出了局部感知的时间注意模块来分别提取每个身体部分的时间依赖性, 充分考虑局部运动的一致性, 并利用长距离上的部分相关进一步提高三维姿态估计。Zhang 等人^[35]提出了 DG-Net, 引入了动态空间/时间图卷积 (DSG/DTG) 来动态地识别每个视频样本的空间/时间人体关节亲和力, 并通过自适应学习视频中的空间/时间关节关系来估计三维姿态, 减少了在将 2D 姿势提升到 3D 姿势时的深度模糊和/或运动不确定性。Wu 等人^[36]从三维人体姿态中自由度 (DOF) 较高的关节出发, 提出了一个肢体姿态感知框架, 包括运动学约束感知网络和轨迹感知时间模

块, 以提高肢体关节位置的三维预测精度。

尽管上述方法在处理视频时取得了一些成功, 但在实际中, 当视频不能始终呈现完整人体, 即某些镜头中有人体部位超出镜头范围的情况下, 现有的方法仍难以估计合理的完整三维人体姿态或者网格序列。这可能是由于现有工作更多关注视频中可见人体部分的建模准确性^[9,14], 对不可见部分的处理方式并未充分利用连续帧间人体部位对应的互补信息。

3 方 法

本节介绍我们的方法, 视频内细粒度人体部位信息注意力分配 (BAA) 模型为方法的核心, BAA 借助虚拟标记作为人体信息的中间表示, 将基于 EFoV 视频的人体网格建模问题, 转化为虚拟标记点集合序列的恢复问题。

3.1 模型概述

如图 2 所示, 本文方法的核心是 BAA 模型。BAA 模型由 Embedding 模块、置信度增强过滤模块 (CEF) 以及注意力分配模块 (AIVM) 组成。其中, Embedding 模块负责将遮蔽后的虚拟标记数据嵌入到 D 维空间中, 并添加位置编码; CEF 模块负责对初步的置信度信息进行增强过滤; AIVM 模块则利用置信度信息辅助恢复出合理的完整虚拟标记点集合。

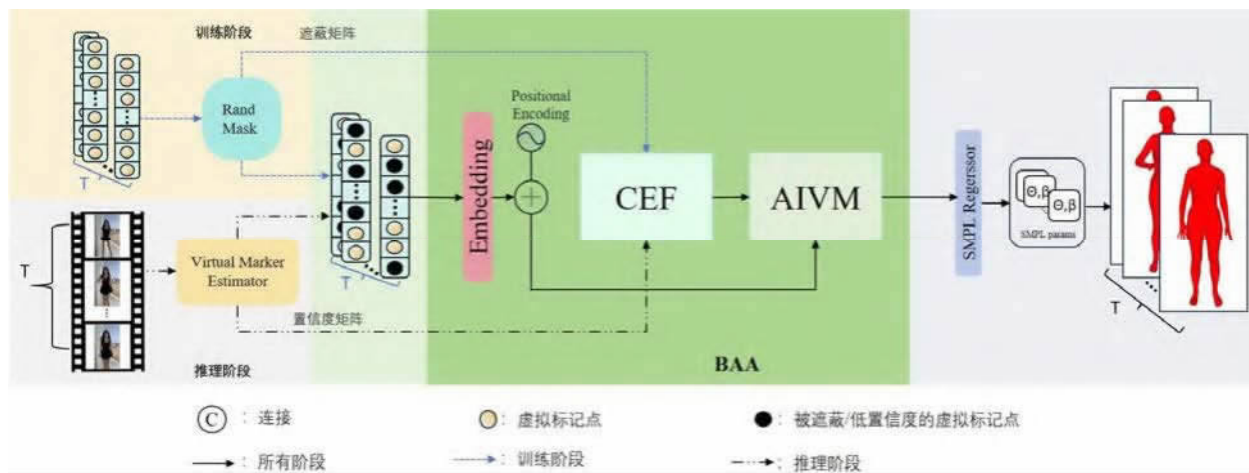


图2 本文方法的框架流程

本文方法的流程分为两种情形: 在训练阶段, 由 Rand Mask 模块负责对完整的虚拟标记信息根据人体部位进行随机遮蔽, 模拟视频中人体部位超出镜头的情况, 从而得到受损的虚拟标记数据, 以及

mask matrix; 在推理阶段, 使用 Ma 等人^[24]提出的虚拟标记估计器从视频中估计得到初步受损的虚拟标记数据以及 confidence matrix。然后, 由 BAA 负责恢复出完整的人体虚拟标记序列, 并最终通过

SMPL 参数回归模块估计最终的 SMPL 模型。我们给出了 BAA 模型的算法伪代码展示，如算法 1 所示。

算法 1. BAA 模型的算法伪代码

输入：虚拟标记点 V

输出：SMPL 模型参数 θ

1. procedure BAA(V)
2. $[V_i^{mask}, C_i] \leftarrow \text{Rand Mask}(V_i)$ //编码并加入位置信息
3. FOR $t \leftarrow 1, T$ DO
4. $x_t \leftarrow e(V_i^{mask}) + PE_t$ //编码并加入位置信息
5. $conf_t \leftarrow \text{CEF}(C_t)$ //对置信度信息增强和过滤
6. $V_L \leftarrow \text{AIVM}([conf, x])$ //恢复虚拟标记信息
7. $\theta \leftarrow \text{SMPLRegressor}([x_{init}, V_L])$ //得到 SMPL 参数
8. END FOR

3.2 人体的虚拟标记表示

如图 3(c)，虚拟标记由人体的主干及四肢的表面点和关节点组成，基于完整的人体虚拟标记点集合^[24]，可以估计出对应的人体 SMPL 网格模型。我

们用 $V = \{v_1, v_2, \dots, v_k\} \in R^{k*3}$ ，表示一个人体模型所对应的虚拟标记点集合，每个人体部位都有相应的虚拟标记点与之对应， k 表示虚拟标记点集合中虚拟标记点的个数，本文中 k 取 81。

针对 EFoV 视频的人体网格建模问题，选用虚拟标记作为人体信息的中间表示有以下优势：(1) 相较于图 3(a) 展示的 SMPL 模型参数，人体虚拟标记与人体部位信息显式对应，通过遮蔽虚拟标记点，可以直观地模拟镜头中人体部位的缺失现象；(2) 相较于图 3(b) 的人体网格点，人体虚拟标记有相似的表达能力，但在点的数量上有很大优势；(3) 相较于图 3(d) 展示的人体骨架点，人体虚拟标记包含了丰富且直接的人体姿势和形状信息，利用完整的虚拟标记，能够更可靠地估计出对应的 SMPL 模型参数。

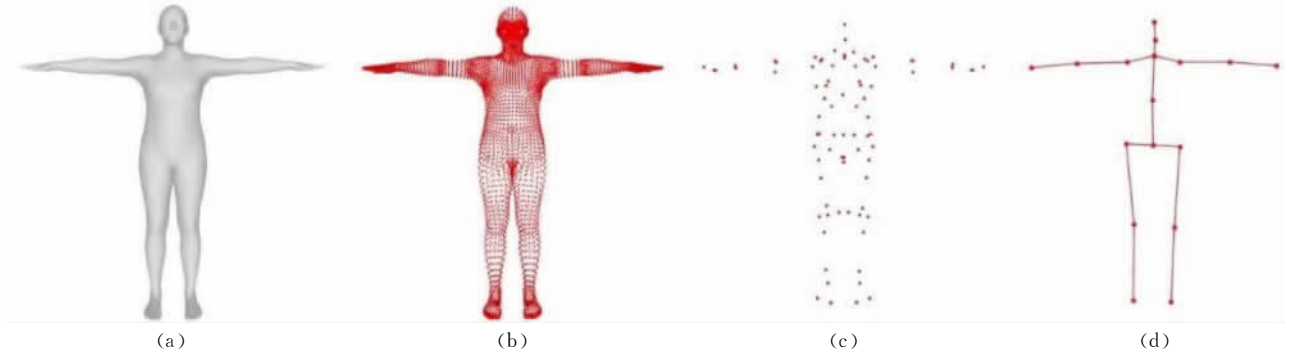


图 3 人体信息表示的各种方法的可视化展示((a) SMPL 参数化人体模型，由形状参数 $\beta \in R^{10}$ 和姿态参数 $\theta \in R^{24*3}$ 控制；(b) 人体网格点 M ，包含丰富的人体表面位置信息，连接相邻的点可构成人体网格， $M \in R^{6890*3}$ ；(c) 虚拟标记点集合 V ，包含足够的人体表面信息以及人体骨架信息， $V \in R^{81*3}$ ；(d) 人体骨架点 K ，仅涵盖人体骨架信息， $K \in R^{17*2}$)

3.3 Rand Mask 策略

Rand Mask 策略用于解决模型训练时的数据构建问题。在训练过程中，数据集中人体网格以及对应的虚拟标记点集合都是完整的合理数据，于是，为了模拟人体超出镜头的情形，我们设计了 Rand Mask 策略，用于将初始的完整虚拟标记点集合序列数据进行一系列的人体部位随机遮蔽。

与现有的某些工作如 PoseBERT^[14] 中对某帧中信息进行完全遮蔽不同，如图 4，我们可视化地对比了的 Rand Mask 策略与现有的传统 Bert 系列^[37] 工作的遮蔽方式，Rand Mask 策略只对序列中某些人体部位所对应的虚拟标记点进行遮蔽。设 $V_i \in R^{k*3}$ 表示第 i 帧的虚拟标记点集合， $V_i = \{v_i^1, v_i^2, \dots, v_i^k\}$ ， $i = 1, 2, \dots, T$ ，其中 v_i^j 表示第 i 帧中第 j 个虚拟标记点。我们对虚拟标记点集合根据人体部位进行了分类，包括手臂、腿、躯干等部位，每个部位包含

不同的虚拟标记点。在遮蔽时，Rand Mask 随机选择人体部位，并只对被选择的部位所包含的虚拟标记点进行遮蔽。Rand Mask 由式(1)表示：

$$[V_i^{mask}, C_i] = \text{RandMask}(V_i) \quad (1)$$

其中，

$$V_i^{mask} = \{v_i^1, v_i^2, \dots, \hat{v}_i^m, \dots, \hat{v}_i^n, \dots, v_i^k\} \quad (2)$$

$$C_i = \{c_i^1, c_i^2, \dots, c_i^k\} \quad (3)$$

$V_i^{mask} \in R^{k*3}$ ， \hat{v}_i^m 表示第 i 帧中第 m 个虚拟标记点被遮蔽后的数据， $i = 1, 2, \dots, T$ 。为了更真实的模拟实际情况中人体部位超出镜头时虚拟标记估计器^[24] 造成的误差，我们的遮蔽方式是在虚拟标记点的坐标中添加随机的扰动。 $C_i \in R^k$ 则是第 i 帧中对应的虚拟标记点的遮蔽向量，若 c_i^j 为 0 则表示第 i 帧中第 j 个虚拟标记点被遮蔽， c_i^j 为 1 则表示未被遮蔽。需要注意的是同一个部位在所有帧中不会始终被遮蔽。

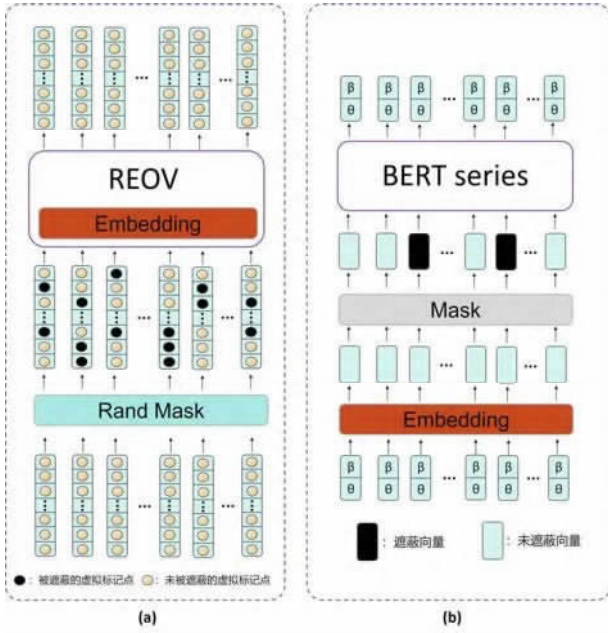


图4 Rand Mask策略应用于模型训练时的具体情况及与现有Mask策略的对比

必须特别指出, Rand Mask仅应用在训练阶段, 在推理阶段中, 我们采用虚拟标记估计器^[24]得到的每个虚拟标记点的置信度来替换其对应的 C_i 。

通过Rand Mask策略得到的遮蔽后的虚拟标记点信息用于强化AIVM模块的恢复能力, 得到的置信度信息则用于监督CEF模块的置信度修正能力。

3.4 Embedding模块

如图4所示, 与基于传统Transformer^[37]系列的模型的做法不同, Rand Mask施加在虚拟标记点集合后再对其进行Embedding。若先对虚拟标记点集合进行Embedding, 遮蔽时就无法精准控制到人体部位。于是, 我们首先将完整的虚拟标记点集合逐帧进行遮蔽得到 V_{mask} , 然后使用线性投影将每帧中的 V_{mask} 嵌入到D维空间中。此外我们将可学习的一维位置编码PE添加到Embedding中^[14]。在训练中对序列中每个帧都添加了特定的位置编码, 嵌入序列 $X = \{x_1, \dots, x_T\}$ 由下式给出:

$$x_i = e(V_i^{mask}) + PE_i \quad (4)$$

其中, $e(V_i^{mask})$ 表示对输入的每个虚拟标记点集合进行线性投影编码, PE_i 则是对应的位置编码, $i = 1, 2, \dots, T$ 。

3.5 CEF模块

置信度增强与过滤模块(CEF模块)用于将前置步骤得到的置信度信息进行增强和过滤, 使模型降低对前置方法估计虚拟标记点置信度的依赖, 其输入为Rand Mask策略生成的初始遮蔽矩阵(或推理阶段由估计器输出的置信度矩阵), 输出为经过修

正后的置信度分布。具体而言, CEF模块针对的是虚拟标记点级别的置信度特征(即每个标记点的可见性概率), 而非图像或3D坐标本身。训练阶段, 首先接收来自Rand Mask模块的遮蔽矩阵, 并对其进行高斯加噪处理。这一步骤至关重要, 因为Rand Mask模块输出的零一数据过于理想化, 不利于模型的鲁棒性训练, 且在实际应用中, 模型无法得到完美的零一表示的置信度。因此, 通过在训练阶段引入高斯噪声, 不仅增强了模型对不确定性的鲁棒性, 还模拟了实际应用中置信度数据的不完美性。需要指出的是, 在实际推理阶段, 由虚拟标记估计器产生的置信度数据不需要额外添加高斯噪声。

对于CEF模块的设计, 如图5所示, 在接收Rand Mask模块输出的遮蔽矩阵并添加高斯噪声之后, 数据进入层归一化(L-Norm)层, 该层通过调整输入数据的尺度, 确保了不同特征维度的一致性, 为后续处理提供了稳定的基础。接着, 数据通过过滤机制进行初步筛选, 这有助于筛选出关键信息并抑制噪声, 进一步净化数据。之后, Tanh激活函数被应用于数据, 引入非线性特性, 使得模型能够学习和模拟更复杂的函数映射。紧接着, 线性层对特征进行线性变换, 为最终的输出做准备。最后, 再经过Tanh函数将线性层的输出尽可能转换为增强后的置信度数据, 这些数据不仅更加鲁棒, 而且更能反映真实世界中的不确定性。

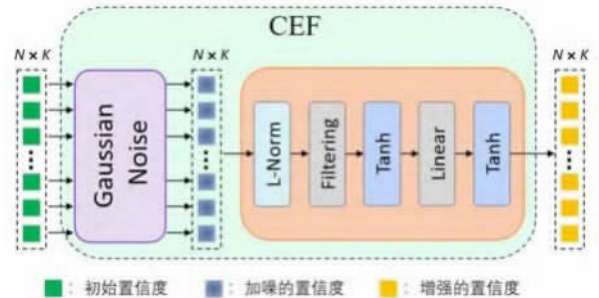


图5 CEF模块设计, 用于对置信度数据进行增强和过滤

3.6 AIVM模块

为了充分地将置信度信息应用到细粒度的虚拟标记点上, 我们引入了一种新的注意力计算方法, 将置信度信息多次融入注意力矩阵的构建过程中, 如图6所示。具体而言, AIVM模块首先接收编码后的人体虚拟标记点信息, 并将其转化为查询向量。同时, 利用CEF模块输出的置信度信息构建键和值向量。这样, 模型就得到了用于计算注意力的查询、键、值三元组。通过这种设计, AIVM模

块能够根据每帧中虚拟标记点的置信度，通过交叉注意力自适应地调整注意力的分配，从而更准确地利用帧内信息。紧接着，将交叉注意力输出的信息再次通过自注意力进行进一步的增强。

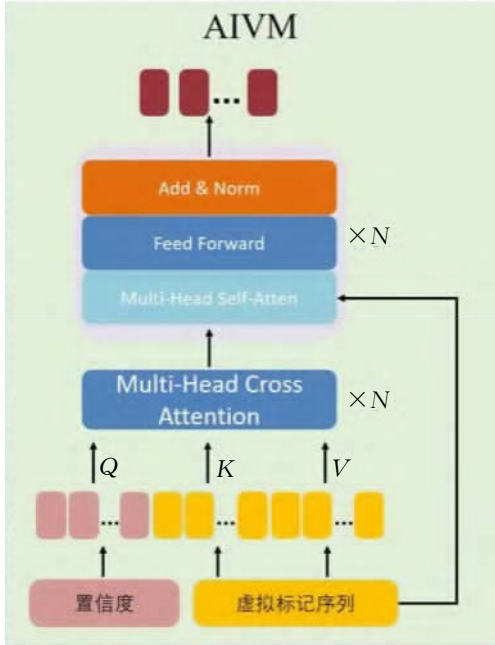


图6 AIVM模块设计，用于将增强后的置信度信息应用到细粒度的人体虚拟标记点上

在AIVM模块的架构中，如图6所示，首先使用多头交叉注意力机制，它首先使得模型在多个子空间中并行地通过置信度信息对各帧中虚拟标记点之间的关系进行增强。随后，通过自注意力层，模型能够更好地在虚拟标记点层面进行进一步自我增强细化。接着，通过前馈网络层对自注意力层的输出进行非线性变换，增强了模型的表达能力。最后，通过残差和归一化层，模型的稳定性和泛化能力得到了进一步提升。此外，AIVM模块循环迭代多次，这意味着模块还会多次重复执行，以逐步提高重建的准确性。每次迭代都基于前一次迭代的输出，不断优化虚拟标记点信息，进而输入SMPL回归层进行人体网格重建。

在AIVM的交叉注意力机制中，通过每帧中虚拟标记点信息和对应的增强后置置信度信息进行注意力分配，注意力 A_i 由下式给出：

$$A_i(Q, K, V) = \text{soft max} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (5)$$

其中，

$$Q = \text{Reshape}(\text{Linear}(\text{Conf})) \quad (6)$$

$$K = \text{Reshape}(VM) \quad (7)$$

$$V = \text{Reshape}(VM) \quad (8)$$

其中， $A_i = \{A_1, \dots, A_T\}$ ， Q 是由特征编码模块对增强后的置信度信息(Conf)进行编码后得到， K 和 V 则是对虚拟标记点集合数据(VM)进行编码得到。交叉注意力之后的自注意力层中， Q ， K ， V 均由虚拟标记点集合数据(VM)编码得到。

通过这种创新的注意力机制和迭代过程，AIVM模块显著提升了模型在处理EFoV视频时的性能，特别是在视频中某帧内突然缺少某些部位信息的情况下。它不仅能够更准确地重建人体网格，还能够在模型训练和推理过程中，有效地利用高置信度的虚拟标记点信息，从而实现更鲁棒的人体网格重建。

3.7 SMPL模型参数估计

对于SMPL回归层，我们使用与HMR^[4]中相似架构的多层感知器(MLP)，不同的是， Θ 参数序列的回归依赖于人体虚拟标记点集合序列。SMPL回归模块接收BAA的结果，用于从完整的人体虚拟标记点集合序列估计SMPL模型参数序列，SMPL回归模块的映射关系如式(9)：

$$\Theta = \text{SMPLRegressor}([\Theta_{init}, V_L]) \quad (9)$$

其中， $[\cdot, \cdot]$ 表示连接， Θ 表示回归得到的SMPL网格模型参数， Θ_{init} 表示初始SMPL模型参数， V_L 表示第 L 层的人体集合序列 V 估计。

3.8 损失函数

我们的模型训练采用了三种数据监督形式：虚拟标记点集合的准确性、SMPL参数及模型网格的精确性、增强虚拟标记点置信度的准确性。这些约束共同构成了最终的损失函数，如式(10)所示：

$$L = L_{VM} + L_{SMPL} + L_{Conf} \quad (10)$$

式中的 L_{VM} 负责度量模型恢复出的虚拟标记点集合序列与真实数据之间的差异。BAA模型采用L1范数来计算这一差异，因为它对异常值不敏感，能够提供一种稳健的性能评估方式：

$$L_{VM} = \|V - V^*\|_1 \quad (11)$$

其中， V 为模型预测的人体虚拟标记点集合， V^* 为GT的虚拟标记点集合。

L_{SMPL} 则从SMPL网格模型的层面进行监督，主要比较SMPL参数、SMPL网格点以及基于SMPL网格采样得到的虚拟标记点信息，整体的SMPL方面的监督如下：

$$L_{SMPL} = \epsilon \|\Theta - \Theta^*\|_2 + \sigma \|V_{M2VM} - V^*\|_1 + \alpha \|M_{\Theta2M} - M_{\Theta2M}^*\|_1 \quad (12)$$

其中, SMPL 模型参数 $\Theta = \{\beta, \theta\}$, 包括形状参数 β , 姿势参数 θ , 与 GT 形状参数 β^* , GT 姿势参数 θ^* 之间采用 MSE 损失。 V_{M2VM} 表示从 SMPL 模型网格映射回的虚拟标记点集合, $M_{\Theta2M}$ 则是模型所估计的 SMPL 参数对应的网格, 用于与 GT 参数对应的网格之间计算差异。 ϵ, σ, α 为每个损失所占的权重, 用于平衡不同损失项对总损失的贡献。

最后, BAA 模型添加了 L_{Conf} 在虚拟标记点置信度层面进行监督, 用于约束 CEF 模块的增强能力:

$$L_{Conf} = \|C_1 - C_1^*\|_1 \quad (13)$$

其中, C_1 为 CEF 模块输出的增强后的 Virtual Markers 点置信度, C_1^* 为 Rand Mask 策略对虚拟标记点集合的遮蔽信息。二者之间使用 L1 范数来度量。

通过这三个层面约束的共同监督, BAA 模型不仅确保了中间数据虚拟标记点集合的准确性, 而且在 SMPL 模型参数和网格层面上保证了精确性。此外, 模型对虚拟标记点置信度数据的识别能力也有进一步的加强。这些改进使得 BAA 模型在处理 EFoV 视频时, 能够提供更加准确和可靠的人体网格重建结果。

4 实验与分析

我们采用 SMPL 参数模型作为最终的人体网格表示, 虚拟标记作为人体中间表示来训练 BAA。在第 4.1 节中, 我们介绍了实验中用于模型训练以及实践时的数据来源, 并介绍了实验中所使用的评价指标。在第 4.2 节中, 我们分析并展示了 BAA 模型的可视化效果; 在第 4.3 节中, 我们通过定量分析, 从量化数据层面对比了一些现有工作; 在第 4.4 节中, 我们通过一系列的消融实验, 来确认模型各组成部分的有效性; 在第 4.5 节中, 我们详细介绍了 BAA 模型训练时的参数设置。

4.1 数据集及评价指标

据我们所了解, 现有的公开人体运动数据集中, 没有超出镜头的人体部位注释数据, 因此, BAA 模型训练中所用的数据是从 AMASS^[25] 中部分动捕数据集(其中包括 CMU、KIT、HDM05、Eyes_Japan_Dataset、SFU、HumanEva、SSM、SOMA 等 16 个数据集)处理得到的纯粹人体虚拟标记序列及对应的网格参数序列信息, 没有额外的图像或者视频注释。在 AMASS 数据集中提供的参数化人体 SMPL 网格模型及其参数的基础上, 我们利用 Ma 等人^[24]

提供的方式, 采样出人体网格模型对应的人体虚拟标记, 这些信息统一起来作为我们训练时所使用的数据集, 其中包含 10 042 个长度不同人物运动序列信息。这种非端到端的训练方式在一定程度上降低了对设备的要求。在实际应用中, 给定视频后, 根据 Ma 等人^[24] 提供的虚拟标记估计器可以从视频中得到初步的人体虚拟标记序列表示, 作为 BAA 需要的数据进行后续的处理。

我们对模型得到的完整虚拟标记点集合以及最终的 SMPL 模型应用 MPJPE 指标对模型的性能进行评估, MPJPE 指标衡量了预测位置与真实位置之间的平均距离误差。MPJPE 的值越低, 表示模型的姿势预测越准确, 具体的计算方式如下:

$$MPJPE_{VM} = 1000 * \text{mean} \left(\sum_{i=1}^T \sqrt{(VM_i - VM_i^*)^2} \right) * n \quad (14)$$

$$MPJPE_{SMPL} = 1000 * \text{mean} \left(\sum_{i=1}^T \sqrt{(SMPL_i - SMPL_i^*)^2} \right) * n \quad (15)$$

其中, VM_i 是由模型生成的第 i 帧中的 Virtual Markers 点集合, VM_i^* 是数据集中的真实 Virtual Markers 点集合数据。 $SMPL_i$ 是从生成的第 i 帧中的 SMPL 模型中采样的 Virtual Markers 点集合, $SMPL_i^*$ 是数据集中相应的真实网格点对应的 Virtual Markers 点集合。 n 表示在单人视频中识别错误导致的多人个数。

此外, 我们提出了重建率指标来辅助衡量各方法的重建能力。重建率指标定义如下:

$$RR = \frac{\sum_{i=1}^T F_i^m}{\sum_{i=1}^T F_i^o} \quad (16)$$

其中, F_i 表示视频第 i 帧是否成功重建出对应的人体网格。 $F_i=1$ 表示该帧重建出了对应的人体模型, $F_i=0$ 则表示重建失败。 F^m 对应被 mask 处理后的视频, F^o 对应原始视频。

4.2 定性分析

为了验证我们方法的有效性, 我们设计了相关的定性实验。在图 7 中, 我们对比了 PoseBERT^[14]、Ma 等人^[24] 以及 BAA 模型在 EFoV 视频中的人体网格重建效果, 选择这两个模型是因为他们的工作或多或少与我们有一定的关联性。需要指出的是, 图中展示的伪 GT 数据为该视频帧中人体完整时对应的建模结果, None 表示该方法在测试使用的 EFoV 情形视频中报错。

结果表明，在视频某帧中人体部位超出镜头时，BAA模型展现出了更优秀的可视化效果。此外，在视频中人体信息完整的情况下，BAA模型的重建能力依然表现出色。这表明 BAA 模型不仅在处理

EFoV 视频时表现出色，而且在常规的人体网格重建任务中也能提供高质量的结果。这种在不同情况下都能保持高性能的能力，充分证明了 BAA 模型的鲁棒性和适应性。



图7 不同方法在 EFoV 视频中重建效果对比

4.3 定量分析

为了对比各方法在 EFoV 状态视频中的重建能力，我们采用了一种折中的做法：首先对原始的包含所有人体信息的视频进行建模，得到各个方法在

人体完整时重建的人体网格序列作为该方法的伪 GT 数据；然后，将每个视频处理成 EFoV 状态的视频，并重建出该情形下对应的人体网格序列与伪 GT 数据进行比较。为了尽可能地让现有方法能够

重建出人体模型，我们仅将每个视频中 10% 的帧进行 mask 处理，mask 的范围也仅限于视频上半部分或者下半部分。

我们通过一系列定量实验对比 MPS^[13]、PoseBERT^[14]、VIBE^[17]、TCMR^[18]、Ma 等人^[24]，以及我们方法在 EFoV 情形下的网格重建结果。表 1 中报告了每个方法在视频处理前后进行建模的结果误差指标 MPJPE 以及重建率。定量实验的结果表明了不同方法在处理 EFoV 视频时的重建性能差异，并且清晰地凸显了 BAA 模型的优势。需要指出的是，根据我们的调研，尚未找到与本研究一样针对 EFoV 问题的方法，因此当前对比的方法仅为一些基于视频重建中较为代表性的工作，其中 PoseBERT^[14]利用了 Transformer 架构的优势，将其应用于人体模型重建工作，VIBE^[17]和 TCMR^[18]则是视频连贯性方面比较突出的工作，Ma 等人^[24]的方法作为我们方法的前置工作，本文中进行了对比。我们方法中对传统 Transformer 架构进行了修改，将遮蔽数据作为一种显式的信息参与模型推测，很大程度上增加了模型鲁棒性。

表 1 不同方法在完整人体运动视频和 EFoV 情形时的建模误差

Method	MPJPE(↓)	RR(↑)
PoseBERT ^[13]	None	None
Ma et al. ^[20]	68.928	1.0000
MPS ^[12]	115.630	0.9481
TCMR ^[17]	139.229	0.9481
VIBE ^[16]	69.290	0.9429
Ours	44.041	1.0000

注：体现为 MPJPE 和重建率。MPJPE 指标越小越好，重建率越大越好，None 表示在处理过的视频中该方法失效。

4.4 消融实验

在本节中，我们研究了不同的策略选择、结构设计以及超参数对模型的影响，并报告模型在各种情形下恢复出虚拟标记点集合序列及人体网格序列所对应的 MPJPE 指标，从而验证我们方法的合理性。需要指出的是，我们测试中使用的输入数据是对原始干净数据根据默认遮蔽策略得到的，且测试数据不曾出现在我们的训练集中。

4.4.1 掩蔽方式对模型的影响

在表 2 中，我们详细列出了在不同遮蔽策略下，BAA 模型恢复虚拟标记点集合序列及人体 SMPL 网格序列时的 MPJPE 指标表现。这些策略涵盖了对整个序列实施随机部位遮蔽(默认策略)、对 50%

的序列应用随机部位遮蔽、对 25% 的序列应用随机部位遮蔽，以及对所有序列执行随机点遮蔽。实验结果表明，根据部位实施遮蔽策略对模型恢复虚拟标记点信息的重要性。此外，在所有视频帧中一致应用部位随机遮蔽策略会显著增强了模型在处理超出视野情况时的适应性和鲁棒性。最后，在图 8(a)中，以折线图的形式展示了各策略在模型性能方面的表现。

表 2 对 BAA 模型遮蔽策略的消融实验

	MPJPE _{VM}	MPJPE _{SMPL}
Mask rand points	16.58	17.90
Mask 25% sequence	12.54	13.94
Mask 50% sequence	12.43	14.29
Mask 100% sequence (default)	10.73	12.14

4.4.2 结构设计对模型的影响

如图 8(b)和图 9 所示，不同的架构设计对模型也会有不同的影响。首先，去除 AIVM 模块会使得 BAA 模型退化到 BERT 系列模型的注意力机制，这在一定程度上会降低模型的性能。从图 9 中可知当模型加上 AIVM 模块时，对有效帧的注意力数值显著提高，并且根据不同的遮蔽对应帧的权重也会相对应调整以保证得到正确的 VM 点；而去掉模块的模型中注意力数值都很低且得到的所有视频帧 VM 点分布接近，这样的结果显然是错误的。AIVM 的注意力机制是根据 VM 点的缺失情况驱动的，注意力模式不是固定的，而是根据每一帧的具体情况(遮挡部分)动态变化的，实现跨越时间的精准修复。例如当腿部遮蔽时模型会主动在整个 64 帧序列中搜索其他有用的参考帧进行修复，以此得到最终正确的 VM 点。

其次，图 8(b)的实验表明，省略对虚拟标记点集合序列的编码虽然简化了模型结构，降低了模型的复杂性，但这是以牺牲模型的恢复能力为代价的，这进一步证实了对序列数据进行编码是提升模型性能不可或缺的步骤。此外，消融研究还验证了位置编码在模型中的重要性。实验表明，去除位置编码会削弱网络对时间信息的学习能力，这对于理解和预测人体姿态的变化至关重要。位置编码为模型提供了关于序列中各帧相对位置的关键信息，从而增强了模型对时序动态的理解。综上所述，我们在表 3 中报告了不同架构设计对应的 MPJPE 指标，这些指标为上述论述提供了量化的证据，展示了各种架构设计选择对模型性能的具体影响。

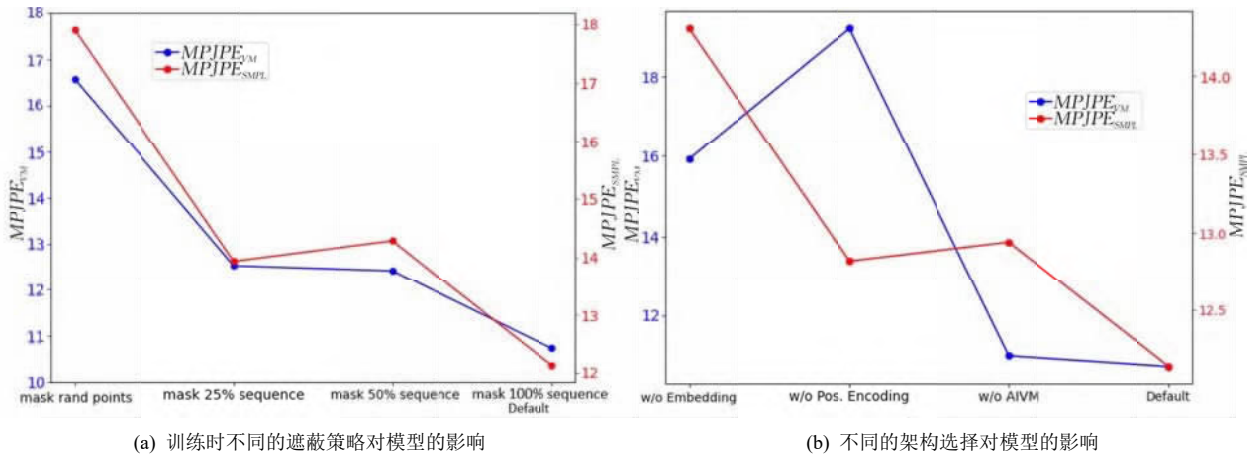


图8 模型消融分析

表3 对BAA模型结构设计的消融实验

	$MPJPE_{VM}$	$MPJPE_{SMPL}$
w/o AIVM	11.00	12.94
w/o Embedding	15.93	14.30
w/o Pos. Encoding	19.19	12.82
default	10.73	12.14

4.4.3 超参数的影响

在模型大小方面，我们研究了不同的迭代模块数、嵌入层维度以及序列长度对模型的影响。我们默认的BAA选择迭代模块次数为 $L=4$ 和嵌入维度为 $D=512$ ，选择这些超参数是因为增加模型的复杂性并不能带来成比例的改进，且相对于更少的迭代次数和更小的嵌入维度，默认策略会带来更大的收益。对于模型训练时的序列长度，我们设置为 $sl=64$ 为默认策略，因为较长的序列虽然会带来微弱的改进，但在模型训练过程中会增加不成比例的消耗，而较短的序列则会带来更大的误差。表4以及图10中展示了这三种参数在不同的选择下对模型重建能力的影响。

表4 对BAA模型超参数的消融实验

	$MPJPE_{VM}$	$MPJPE_{SMPL}$
$L=1$	24.26	17.05
$L=2$	16.60	14.84
$L=8$	11.17	16.74
$D=128$	14.95	15.10
$D=256$	11.92	12.82
$D=1024$	38.09	28.10
$sl=16$	16.37	16.93
$sl=32$	10.97	13.59
$sl=128$	11.15	13.40
Default ($L=4, D=512, sl=64$)	10.73	12.14

4.5 实现细节和设置

我们的BAA模型使用PyTorch框架，在单张NVIDIA GeForce RTX 4090显卡上进行大约12h的训练。我们在表5中报告了用于训练BAA的详细超参数设置。

表5 模型训练时的超参数及设置

Hyper-parameters	fine-tuning	Hyper-parameters	fine-tuning
Optimizer	AdamW	Decay step	[15, 85, 110, 125]
Base learning rate	5e-4	Batch size	32
Learning rate ^β	0.4	Epochs	130
Weight decay	1e-4	Warmup epochs	1

5 结论

在本文中我们提出了一种面向EFoV视频估计合理三维人体网格模型序列的方法。这项工作的主要贡献在于设计了RandMask策略以及BAA模型。RandMask对输入的虚拟标记点集合进行随机的部位遮蔽以模拟实际中的超出镜头的人体部位，BAA模型根据视频帧中人体部位的可见程度来分配注意力，以达到合理利用各帧信息的目的。此外，训练过程只需要人体虚拟标记点集合和网格参数信息来拟合模型，所以我们不需要有额外的图像或者视频注释，这在减少了数据标注的工作量的同时，也使得我们的模型方法可以在低算力场景下进行。

然而在研究过程中，我们发现BAA模型仍有以下几个方向可以开展后续研究工作：

(1) 手部和面部等细节部位的精确重建一直是该领域的难点，本文中使用的SMPL模型仅表达人体的整体姿势，没有手部脸部以及衣物细节。未来

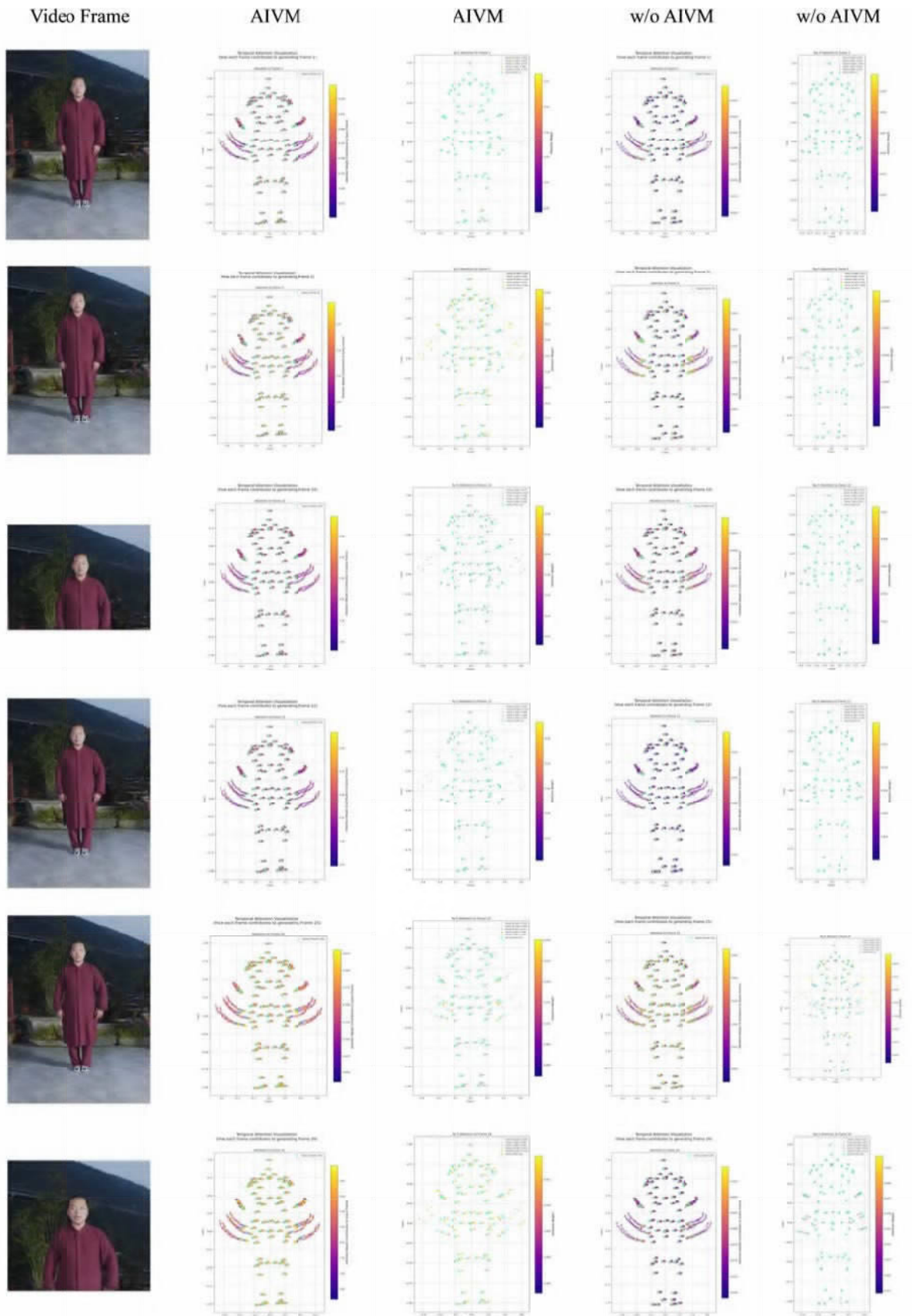


图9 跨帧注意力热力图(图中随机选取了部分视频帧进行可视化, 第一列为视频帧, 第二列展示了当Frame t (查询帧)有遮挡时, 注意力机制如何从其他清晰的帧($t-k, t+j$)中获取信息来修复姿态。图中的坐标点代表某一帧中的人体虚拟标记点。图中的颜色代表该帧的注意力权重, 数值越大颜色越亮表示该帧对于生成当前帧的贡献越大、越重要。第三列热图中只显示了注意力权重最高的5个帧)

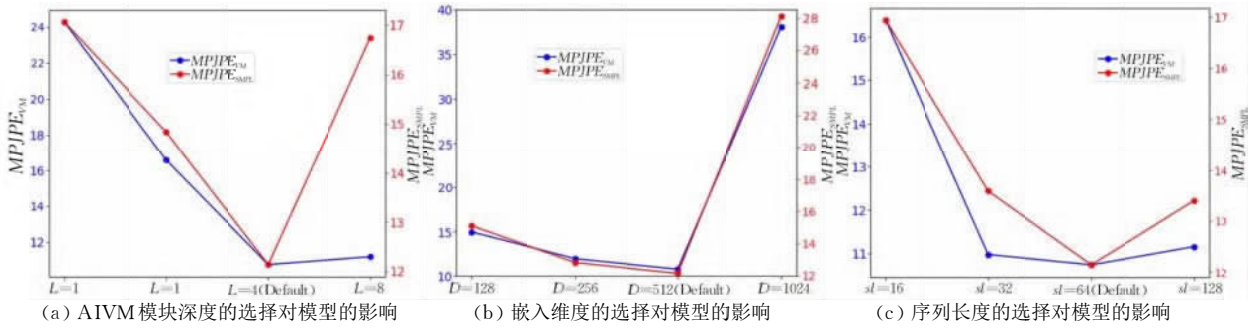


图10 对模型超参数的消融分析

的工作可以尝试添加人体细节的估计, 以实现更加真实和细致的人体网格模型。

(2)当前的运动人体网格模型重建仅基于视频中的信息, 未来可尝试结合多模态数据(如语音、文字等)进行人体网格的驱动。通过提供更丰富的信息, 提高人体网格模型的准确性和可操控性。未来的研究会继续探索如何有效融合不同模态。

参考文献

- [1] Bogo F, et al. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 561-578
- [2] Zanfir A, Maroiu E, Sminchisescu C. Monocular 3D pose and shape estimation of multiple people in natural scenes: The importance of multiple scene constraints//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2148-2157
- [3] Kolotouros N, Pavlakos G, Daniilidis K. Convolutional mesh regression for single-image human shape reconstruction//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4501-4510
- [4] Kanazawa A, et al. End-to-end recovery of human shape and pose//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7122-7131
- [5] Pavlakos G, Zhu L, Zhou X, et al. Learning to estimate 3D human pose and shape from a single color image//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 459-468
- [6] Zhang H, Cao J, Lu G, et al. Learning 3D human shape and pose from dense body parts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(5): 2610-2627
- [7] Lin K, Wang L, Liu Z. End-to-end human pose and mesh reconstruction with transformers//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 1954-1963
- [8] Kolotouros N, Pavlakos G, Black M J, et al. Learning to Reconstruct 3D human pose and shape via model-fitting in the loop//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 2252-2261
- [9] Kocabas M, Huang C-H P, Hilliges O, et al. PARE: Part attention regressor for 3D human body estimation//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 11107-11117
- [10] Tian, Y, et al. Recovering 3D human mesh from monocular images: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15406-15425
- [11] Li Z, Liu J, Zhang Z, et al. Cliff: Carrying location information in full frames into human pose and shape estimation//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 590-606
- [12] Wang L, Liu X, Ma X, et al. A progressive quadric graph convolutional network for 3D human mesh recovery. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(1): 104-117
- [13] Wei W-L, Lin J-C, Liu T-L, et al. Capturing humans in motion: Temporal-attentive 3D human pose and shape estimation from monocular video//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 13201-13210
- [14] Baradel F, Brégier R, Groueix T, et al. PoseBERT: A generic transformer module for temporal 3D human modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 12798-12815
- [15] Kanazawa A, Zhang J Y, Felsen P, et al. Learning 3D human dynamics from video//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5614-5623
- [16] Sun Y, Ye Y, Liu W, et al. Human mesh recovery from monocular images via a skeleton-disentangled representation//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 5348-5357
- [17] Kocabas M, Athanasiou N, Black M J. VIBE: Video inference for human body pose and shape estimation//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, Seattle, USA, 2020: 5252-5262
- [18] Choi H, Moon G, Chang J Y, et al. Beyond static features for

- temporally consistent 3d human pose and shape from a video//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 1964-1973
- [19] Doersch C, Zisserman A. Sim2real transfer learning for 3D human pose estimation: Motion to the rescue//Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019: 12929-12941
- [20] Dang Y, Yin J, Zhang S. Relation-based associative joint location for human pose estimation in videos. *IEEE Transactions on Image Processing*, 2022, 31(1): 3973-3986
- [21] Feng R, Gao Y, Tse T H E, et al. DiffPose: SpatioTemporal Diffusion Model for Video-Based Human Pose Estimation//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 14815-14826
- [22] Zhu W, Ma X, Liu Z, et al. MotionBERT: A unified perspective on learning human motion representations//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 15039-15053
- [23] Toshpulatov M, et al. Human pose, hand and mesh estimation using deep learning: A survey. *The Journal of Supercomputing*, 2022, 78(6): 7616-7654
- [24] Ma X, Su J, Wang C, et al. 3D human mesh estimation from virtual markers//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 534-543
- [25] Mahmood N, Ghorbani N, Troje N F, et al. AMASS: Archive of motion capture as surface shapes//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 5441-5450
- [26] Loper M, Mahmood N, Romero J, et al. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 2015, 34(6): 248:1-248:16
- [27] LI Jia-Ning, WANG Dong-Kai, ZHANG Shi-Liang. Deep-learning-based 2D human pose estimation: Present and future. *Chinese Journal of Computers*, 2024, 47(1): 231-250(in Chinese)
(李佳宁, 王东凯, 张史梁. 基于深度学习的二维人体姿态估计: 现状及展望. *计算机学报*, 2024, 47(1): 231-250)
- [28] Ke L, Chang M-C, Qi H, et al. DetPoseNet: Improving multi-person pose estimation via coarse-pose filtering. *IEEE Transactions on Image Processing*, 2022, 31(3): 2782-2795
- [29] Li J, Wang Y, Zhang S. PolarPose: Single-stage multi-person pose estimation in polar coordinates. *IEEE Transactions on Image Processing*, 2023, 32(1): 1108-1119
- [30] Hassan M T, Ben Hamza A. Regular splitting graph network for 3D human pose estimation. *IEEE Transactions on Image Processing*, 2023, 32(1): 4212-4222
- [31] Cheng Y, Wang B, Tan R T. Dual networks based 3D multi-person pose estimation from monocular video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(2): 1636-1651
- [32] Luo Z, Golestaneh S A, Kitani K M. 3D human motion estimation via motion compression and refinement//Proceedings of the Asian Conference on Computer Vision. Kyoto, Japan, 2020: 324-340
- [33] Li W, Liu H, Tang H, et al. Mhformer: Multi-hypothesis transformer for 3d human pose estimation//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New Orleans, USA, 2022: 13137-13146
- [34] Xue Y, Chen J, Gu X, et al. Boosting monocular 3D human pose estimation with part aware attention. *IEEE Transactions on Image Processing*, 2022, 31(1): 4278-4291
- [35] Zhang J, Wang Y, Zhou Z, et al. Learning dynamical human-joint affinity for 3D pose estimation in videos. *IEEE Transactions on Image Processing*, 2021, 30(9): 7914-7925
- [36] Wu L, Yu Z, Liu Y, et al. Limb pose aware networks for monocular 3D pose estimation. *IEEE Transactions on Image Processing*, 2022, 31(12): 906-917
- [37] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008



GONG Yong-Yi, Ph.D., professor. His current research interests include computer vision and machine learning.

QIN Zhe, M.S. His main research interests include computer vision and machine learning.

DING Ruo-Yao, Ph.D., associate professor. His research interests include text mining, machine learning, pattern recognition.

Background

Video-based 3D human mesh reconstruction is a fundamental and important task in computer vision. It supports

multiple downstream tasks and enables a wide range of real-world applications. With the rapid advancement of deep

learning techniques, significant progresses have been made in video-based 3D human mesh reconstruction. However, based on our comprehensive investigation, when certain parts of human body Exceeding the Field of View (EFoV), reconstructing accurate and complete human meshes remains a difficult problem. We posit that the primary challenge lies in the inherent inability to capture direct visual information for body parts that extend beyond the camera's field of view, due to the physical constraints of camera field. Consequently, reconstructing these occluded regions necessitates the integration of motion priors, temporal cues from adjacent frames and other information. Furthermore, the scarcity of publicly available datasets with explicit annotations for out-of-view body parts significantly increased the difficulty to accurately infer and reconstruct these missing regions.

In this paper, we propose a method to reconstruct 3D human

mesh under the EFoV condition. Firstly, we define the notion of EFoV condition and clearly distinguish it from the occlusion problem. Secondly, we review related works in image-based and video-based human mesh reconstruction and pose estimation. Thirdly, we present our proposed method specifically designed to handle the EFoV scenarios in videos and compare our method with several related approaches. Experimental results demonstrate that our proposed method can reconstruct more complete and reasonable human meshes under the EFoV conditions. Finally, we summarize our contributions and discuss potential research directions for future work.

This work was supported by the Basic and Applied Basic Research Foundation of Guangdong Province (2019-A1515011078).