arXiv:2304.01199v2 [cs.CV] 7 Aug 2023

三维姿态和跟踪对人体动作识别的益处

Jathushan Rajasegaran[1,2], Georgios Pavlakos[1], Angjoo Kanazawa[1],Christoph Feichtenhofer[2],
Jitendra Malik[1,21]加州大学伯克利分校， [2]Meta AI、FAIR

摘要

本文研究了使用跟踪和三维姿态进行动作识别的益处。为此，我们采用拉格朗日视角，沿着人体运动轨迹分析动作，而非在空间中的固定位置进行分析。采取这种方法，我们可以利用人物的轨迹片段来预测其动作。基于这一思路，我们首先展示使用三维姿态推断动作的益处，并研究人物之间的交互。随后，我们提出一种拉格朗日动作识别模型，通过融合轨迹片段上的三维姿态与上下文化外观来进行识别。为此，我们的方法在 *AVA v2.2* 数据集上，无论是在仅使用姿态的设置还是标准基准设置中，均达到了当前最先进水平。在仅使用姿态线索推理动作时，我们的姿态模型比相应的当前最先进方法高出 ***+10.0*** 个平均精度均值；而我们的融合模型则比最佳的当前最先进模型高出 ***+2.8*** 个平均精度均值。我们的最佳模型在 ***AVA 2.2*** 数据集上取得了 ***45.1*** 的平均精度均值。代码和结果见：*https://brjathu.github.io/LART*

1. 引言

在流体力学中，传统上会区分流场的拉格朗日描述与欧拉描述。引用 维基百科词条中的说法，“流场的拉格朗日描述是一种观察流体运动的方式：观察者跟随单个流体微团，观察它如何随时间在空间中移动。描绘单个流体微团的位置随时间的变化，就能得到该微团的迹线。这可以想象成坐在船上，顺流而下。流场的欧拉描述则是一种观察流体运动的方式，它关注流体流经的空间中特定位置随时间发生的变化。这可以想象成坐在河岸上，观看水流经过某个固定位置。”

这些概念与我们分析人类活动视频的方式密切相关。在欧拉视角下，我们会关注特定位置的特征向量，例如（x，y）或（x，y，z），并在固定于该位置的空间坐标中考察其随时间的变化。在拉格朗日视角下，我们会跟踪某个人物在时空中的运动并跟踪相应特征向量在时空中的变化。

尽管早期的活动识别文献例如，[13, 21, 61] 通常采用拉格朗日视角，但自基于三维时空卷积的神经网络问世以来，例如， [58], 欧拉视角已成为当前最先进方法（如 SlowFast 网络）中的标准做法 [18]。即使转向 Transformer 架构之后 [14, 60]，欧拉视角依然延续至今。这一点值得注意，因为 Transformer 的标记化步骤为我们提供了重新审视这一问题的机会：“视频分析中，什么可以作为词语的对应物？”。Dosovitskiy 等人 [12] 提出图像块是一个不错的选择，而将这一思路延伸至视频，则表明时空立方体同样适用于视频。

相反，在这项工作中，我们采用拉格朗日视角来分析人体动作。这意味着我们关注某个实体随时间变化的轨迹。这里，实体可以是低层级的，例如，一个像素或一个图像块；也可以是高层级的，例如一个人物。由于我们关注的是理解人体动作，因此选择以“人类实体”为处理对象。为此，我们开发了一种方法，处理视频中人物的轨迹，并利用这些轨迹识别他们的动作。我们利用最近提出的三维跟踪方法 PHALP [50] 和 HMR 2.0 [22]来恢复这些轨迹。如图1所示，PHALP 通过将人物提升到三维空间，从视频中恢复人物轨迹片段。这意味着我们既能在一系列帧中关联人物，也能获取他们的三维表示。给定这些人物的三维表示（即三维姿态和三维位置），我们将其用作每个标记的基本内容。这样，我们便能构建一个灵活的系统，其中模型（此处为 Transformer）以对应不同人物的标记作为输入，并能获取他们的身份、三维姿态和三维位置。拥有场景中人物的三维位置，使我们能够学习人物之间的交互。基于这种标记化方式的模型能够利用三维跟踪和姿态，并且优于只能获取姿态信息的先前基线方法 [9, 53]。

尽管人体姿态随时间的变化是一个强有力的信号，但有些动作需要更多关于外观和场景的上下文信息。因此，还需要将姿态与直接来自像素的人物及场景外观信息进行融合。

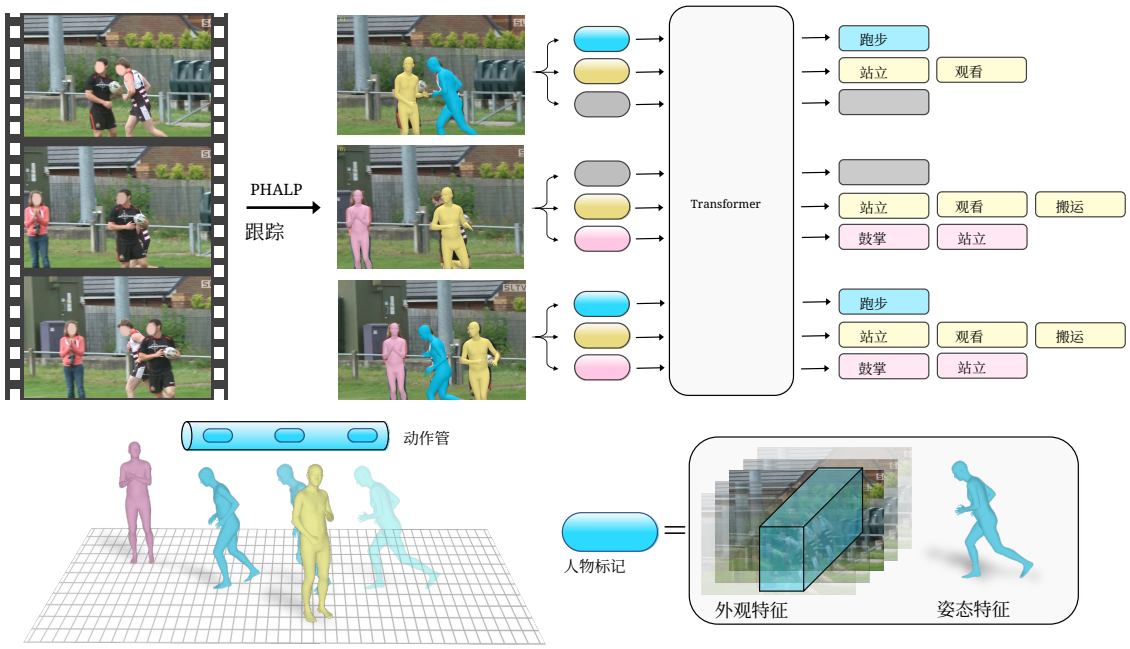


图 1. **我们方法的概述：** 给定一个视频，首先，我们使用跟踪算法跟踪每个人物（例如 PHALP [50]）。然后，将轨迹中的每个检测结果标记化，以表示以人体为中心的向量（例如姿态、外观）。为了表示三维姿态，我们使用 SMPL [40]参数和估计的人物三维位置；对于上下文化外观，我们使用 MViT [14]（在 MaskFeat 上预训练 [66]）得到的特征。随后，我们训练一个 Transformer 网络，利用这些轨迹预测动作。请注意，在第二帧中，我们没有检测到蓝色人物，因此在这些位置，我们会传入一个掩码标记，以填补漏检。

为此，我们还采用当前最先进的动作识别模型 [14, 39]，以拉格朗日框架从人物和场景的上下文化外观中提供互补信息。具体而言，我们在每个轨迹片段的轨迹上密集运行此类模型，并记录轨迹片段周围的上下文化外观特征。因此，我们的词元包含人物三维姿态的显式信息，以及由动作识别骨干网络 [14]处理的、从像素中密集采样的外观信息。与以往的当前最先进水平相比，我们的完整系统在具有挑战性的 AVA v2.2 数据集上大幅提升了 **2.8 平均精度均值**。

总的来说，我们的主要贡献是提出一种方法，突出跟踪和三维姿态对人体动作理解的影响。为此，我们提出了**拉**格朗日**动**作**识**别与**跟**踪（**LART**）方法，利用人物的轨迹片段来预测其动作。我们的基线版本利用视频中人物的轨迹片段轨迹和三维姿态表示，性能优于以往使用姿态信息的基线方法。此外，我们证明，动作识别的拉格朗日视角可以轻松与仅依赖视频外观和上下文的传统基线方法相结合，相较于主流范式取得显著提升，与主流范式相比取得了显著提升。

2. 相关工作

三维人体重建： 许多相关工作使用 SMPL人体模型 [40] 从图像中重建三维人体。最初，相关方法依赖于基于优化的方法，例如 SMPLify [6]，；但自 HMR [28], 问世以来，人们对直接根据人物对应图像作为输入来回归 SMPL参数 [40] 的方法产生了浓厚兴趣。许多后续工作在原始模型的基础上进行了改进，通过估计更准确的姿态 [36]或形状 [8], 来提高模型的鲁棒性 [47], 、纳入侧面信息 [35, 37], 、探索不同的架构选择 [34, 72], 等。

虽然这些工作提升了基本的单帧重建性能，但与此同时，也有研究致力于从视频输入中对人体进行时序重建。HMMR模型 [29]在HMR图像特征上使用卷积时序编码器 [28]，以重建人体随时间的变化。其他方法则研究了循环编码器 [33]或Transformer编码器 [47]。近期研究不再对图像特征进行时序池化，而是直接使用SMPL参数进行时序编码 [3, 51]。

上述类别中的时序方法有一个共同假设：它们能够获取视频中人物的轨迹片段。这意味着它们依赖跟踪方法，而这些方法大多在二维域上运行 [4,15,43,70]，并因此引入了许多错误。为克服这一局限，近期研究 [49,50] 利用三维人体恢复方面的进展，实现了更稳健的视频身份跟踪。具体而言，Rajasegaran提出的PHALP方法等人 [50] 能够在各种场景中实现稳健跟踪，包括野外视频和电影。本文使用PHALP系统，从大规模视频数据集中发现较长的跟踪轨迹片段。这样，我们便能训练所提出的方法，使其根据三维姿态输入识别动作。

动作识别： 早期的动作识别研究依赖 HOG3D [32],立方体 [11] 和密集轨迹 [61,62]等手工设计的特征。深度学习兴起后，三维卷积网络成为动作识别的主要骨干网络 [7,56,58]。然而，三维卷积模型以相似的方式处理空间和时间，因此，为了解决这一问题，研究者提出了双流架构 [54]。在双流网络中，一条通路专门用于提取运动特征，通常以光流作为输入。计算光流这一要求使得这些模型难以进行端到端学习。另一方面，SlowFast 网络 [18]仅使用视频流，但采用不同的帧率，使其能够从快速通路中学习运动特征，并通过横向连接融合空间和时序信息。近年来，随着 Transformer 架构的发展，许多研究开始使用 Transformer 主干网络进行动作识别 [1,5,14,44]。

尽管上述工作主要关注动作识别的模型架构，但另一类研究则探讨行为主体与物体之间更细粒度的关系 [55,63,64,73]。非局部网络 [63]利用自注意力对视频中的实体进行推理，并学习长程关系。ACAR [45]通过先在边界框区域内进行池化来提取主体-上下文特征，再学习行为主体之间的高层关系，从而对主体-上下文-主体关系进行建模。与 ACAR 相比，除轨迹身份外，我们的方法并未显式设计任何关于行为主体关系的先验。

沿着这一思路，一些研究利用人体姿态来理解动作 [9,53,59,67,71]。PoTion [9] 通过将时间依赖关系着色，采用基于关键点的姿态表示。最近，JMRN [53] 提出了一种关节运动重加权网络，用于分别学习关节轨迹，再融合这些信息以推断关节间运动。尽管这些研究依赖二维关键点，并采用特定架构对表示进行编码，我们则使用更明确的三维 SMPL 参数。

3. 方法

理解人体动作需要解读多种信息来源 [31]。这些信息包括头部朝向和视线方向、人体姿态及其动态、与物体或其他人或动物的互动、整体场景、活动上下文（例如自己或他人刚刚做出的动作）等。有些动作仅凭姿态及其动态便可识别，Johansson 等人 [27]的研究对此进行了证明：人们只需观察运动的点光源，就能出色地识别行走、跑步和爬行。然而，理解复杂动作需要综合推理多种信息来源。例如，要识别某人正在用刀切番茄，看到刀和番茄会有所帮助。

这里有许多可供选择的设计方案。是否应采用“解耦”的表征，以模块化方式显式表示姿态、交互物体等元素？还是应将视频像素输入一个大容量的神经网络模型，并依赖它自行判断哪些信息具有判别价值？本文研究了两种方案：a) 由 HMR 模型 [22,28] 重建的人体姿态；以及 b) 由 MViT 模型 [14] 计算的人体姿态及其上下文外观。

给定一个包含 T 帧的视频，我们首先使用 PHALP [50], 跟踪每个人物，从而获得每个人物随时间变化的唯一身份。设时刻 $t \in [1,2,3,...T]$ 的人物 $i \in [1,2,3,...n]$ 由人物向量 $\mathbf{H}_t^i$ 表示。这里， n 是每帧中的人物数量。构造这个人物向量时，使其包含以人为中心的表征 $\mathbf{P}_t^i$ 以及一些经过上下文处理的外观信息$\mathbf{Q}_t$。

$$\mathbf{H}_t^i = \{\mathbf{P}_t^i, \mathbf{Q}_t^i\}. \tag{1}$$

由于我们通过跟踪获知每个人物的身份，因此可以为每个人物创建动作管道 [21] 表征。设 $\mathbf{\Phi}_i$ 为人物 i 的动作管道，那么该动作管道包含此人物随时间变化的所有人物向量。

$$\mathbf{\Phi}_i = \{\mathbf{H}_1^i, \mathbf{H}_2^i, \mathbf{H}_3^i, ..., \mathbf{H}_T^i\}. \tag{2}$$

给定这种表征，我们训练模型 **LART**，使其根据动作管道（跟踪轨迹）预测动作。在这项工作中，我们使用标准的 Transformer [60] 对网络 $\mathcal{F}$建模，这使我们能够在跟踪轨迹因遮挡、检测失败等原因而不连续时屏蔽注意力。有关网络架构的更多详情，请参阅附录。

$$\mathcal{F}\big(\mathbf{\Phi}_1, \mathbf{\Phi}_2, ..., \mathbf{\Phi}_i, ..., \mathbf{\Phi}_n; \Theta\big) = \widehat{Y}_i. \tag{3}$$

其中， Θ 是模型参数， $\widehat{Y}_i = \{y_1^i, y_2^i, y_3^i, ..., y_T^i\}$是某条跟踪轨迹的预测结果， y_t^i 是该跟踪轨迹 i 在时间 t的预测动作。在为目标人物 i预测动作时，模型可以利用其他人的动作进行推理。最后，我们使用二元交叉熵损失来训练模型，并使用平均精度均值（mAP）进行评估。

3.1. 基于三维姿态的动作识别

在本节中，我们研究以人为中心的姿态表示对动作识别的影响。为此，我们考虑一种仅包含姿态表示的人物向量，$\mathbf{H}_t^i = \{\mathbf{P}_t^i\}$。而$\mathbf{P}_t^i$ 通常可以包含关于人物的任何信息，在本文中，我们训练了一个仅使用姿态的模型 **LART-pose** ，该模型基于 SMPL [40] 模型使用人物的三维人体姿态。这包括不同身体部位的关节角度，$\theta_t^i \in \mathcal{R}^{23\times3\times3}$ 并被视为一种模态无关表示，这意味着我们会对所有身体部位进行预测，即使其中一些部位在图像中可能被遮挡或截断。由于全局身体朝向 $\psi_t^i \in \mathcal{R}^{3\times3}$ 与身体姿态分开表示，因此我们的身体表示不受视频特定视角的影响。除三维姿态外，我们还使用人物在摄像机视图中的三维位置（该位置也由 PHALP 模型预测 L_t^i）。这样便可以考虑不同人物在三维空间中的相对位置。更具体地说，每个人表示为：

$$\mathbf{H}_t^i = \mathbf{P}_t^i = \{\theta_t^i, \psi_t^i, L_t^i\}. \tag{4}$$

假设给定视频中有n 个轨迹片段 $\{\mathbf{\Phi}_1, \mathbf{\Phi}_2, \mathbf{\Phi}_3, ..., \mathbf{\Phi}_n\}$。为了研究轨迹片段 i中的动作，我们将人物 i 视为关注人物；获取其他轨迹片段有助于解读人物 i的人-人交互。因此，为了预测所有 n 个轨迹片段的动作，我们需要进行 n 次前向传播。如果人物 i是关注人物，那么我们会随机采样 $N-1$个其他轨迹片段，并将其与 $\mathbf{\Phi}_i$一起输入模型 $\mathcal{F}(;\ \Theta)$。

$$\mathcal{F}(\mathbf{\Phi}_i, \{\mathbf{\Phi}_j | j \in [N]\}; \Theta) = \widehat{Y}_i \tag{5}$$

因此，模型会看到 N 个轨迹片段，并预测主轨迹（关注人物）的动作。为此，我们首先将所有人物向量通过线性层进行标记化，并将其投影到 $f_{proj}(\mathcal{H}_t^i) \in \mathcal{R}^d$ 个 d 维空间中。随后，我们添加用于 a) 时间和 b) 轨迹片段 ID 的位置嵌入。对于时间和轨迹片段 ID，我们使用二维正弦和余弦函数进行位置编码 [65]，，将人物 i 指定为零号$^{\text{th}}$ 轨迹，其余轨迹片段则使用轨迹片段 ID $\{1, 2, 3, ..., N-1\}$。

$$\begin{aligned}
PE(t, i, 2r) &= \sin(t/10000^{4r/d}) \\
PE(t, i, 2r+1) &= \cos(t/10000^{4r/d}) \\
PE(t, i, 2s + D/2) &= \sin(i/10000^{4s/d}) \\
PE(t, i, 2s + D/2 + 1) &= \cos(i/10000^{4s/d})
\end{aligned}$$

其中，t 表示时间索引，i 表示轨迹编号，r，$s \in [0, d/2)$指定维度，D 表示标记的维度。

添加时间和身份的位置编码后，每个人物标记都会传递给 Transformer 网络。$(t + i \times N)^{th}$ 标记表示为：

$$token_{(t+i\times N)} = f_{proj}(\mathcal{H}_t^i) + PE(t, i, :) \tag{6}$$

我们的目标人物表述使我们能够利用场景中的其他演员，更准确地预测主要行为者的动作。当场景中有多个演员时，了解一个人的动作有助于预测另一个人的动作。场景中演员的某些动作彼此相关（例如，跳舞、打斗）；有时，人们会做出互惠动作（例如，说话和倾听）。在这些情况下，了解一个人的动作有助于更有把握地预测另一个人的动作。

3.2. 基于外观和3D姿态的动作

虽然人体姿态在理解动作方面起着关键作用，但更复杂的动作需要结合场景及其上下文进行推理。因此，本节将探讨结合姿态与上下文外观特征对动作识别的益处，并训练模型**LART** ，使其能够利用整个跟踪轨迹上的三维姿态和外观。对于每条跟踪轨迹，我们以频率 f_s 运行2D动作识别模型（即MaskFeat [66]预训练MViT [14]），并存储分类层之前的特征向量。例如，考虑一条包含 $\{D_1^i, D_2^i, D_3^i, ..., D_T^i\}$个检测结果的跟踪轨迹 $\mathbf{\Phi}_i$。我们获取2D动作识别模型对位于 $\{t, t + f_{FPS}/f_s, t + 2f_{FPS}/f_s, ...\}$的检测结果所作的预测。这里，$f_{FPS}$表示帧在屏幕上的显示速率。由于这些动作识别模型在一定程度上能够捕捉时序信息，因此$\mathbf{Q}_{t-f_{FPS}/2f_s}^i$至$\mathbf{Q}_{t+f_{FPS}/2f_s}^i$共享相同的外观特征。假设我们有一个预训练的动作识别模型 $\mathcal{A}$，它接收帧序列以及位于中间帧的检测边界框，那么$\mathbf{Q}_t^i$ 的特征向量表示为：

$$\mathcal{A}\big(D_t^i, \{I\}_{t-M}^{t+M}\big) = \mathbf{U}_t^i$$

其中，$\{I\}_{t-M}^{t+M}$ 是图像帧序列，$2M$ 是动作识别模型所观察到的帧数，$\mathbf{U}_t^i$ 是上下文外观向量。请注意，由于动作识别模型会查看完整的图像帧，因此该表示隐含地包含了关于场景、物体和运动的信息。然而，我们认为，以人为中心的姿态表示与从卷积网络或 Transformer 网络中提取的特征向量包含正交信息。例如，三维姿态是一种几何表示，而$\mathbf{U}_t^i$ 更多地体现光度信息；SMPL参数包含更多关于人类动作/姿态的先验，并且是非模态的，而外观表示则是从原始像素中学习得到的。现在我们同时拥有

人物向量$\mathbf{H}_t^i$中的以姿态为中心的表示和以外观为中心的表示：

$$\mathbf{H}_t^i = \{\underbrace{\theta_t^i, \psi_t^i, L_t^i}_{\mathbf{P}_t^i}, \underbrace{\mathbf{U}_t^i}_{\mathbf{Q}_t^i}\} \tag{7}$$

因此，每个人都由其三维姿态、三维位置、外观和场景内容来表示。我们采用上一节所述的相同流程，添加位置编码，并使用姿态+外观标记训练 Transformer 网络 $\mathcal{F}(\Theta)$。

4. 实验

我们在 AVA [23] 上的多种设置中评估了我们的方法。AVA [23] 提出了一个动作检测问题：需要在时空体中定位人物，并为其标注动作标签。该数据集以 1 赫兹的频率提供标注，每个行为主体有 1 个姿态动作标签、最多 3 个可选的人-物交互标签，以及最多 3 个可选的人-人交互标签。在评估中，我们使用 AVA v2.2 标注，并遵循 [23]中的标准协议。我们在 60 个类别上测量平均精度均值（mAP），帧级交并比阈值为 0.5。此外，我们还在 AVA-Kinetics [38]数据集上评估了我们的方法。该数据集为 Kinetics 视频提供时空定位标注。

我们使用 PHALP [50] 跟踪 AVA 数据集中的人物。PHALP 属于检测式跟踪范式，并使用 Mask R-CNN [25] 检测场景中的人物。在训练阶段，由于边界框标注仅以 1 赫兹的频率提供，我们使用 Mask R-CNN 检测结果处理中间帧，并每隔 30 帧使用一次真实值边界框。在验证阶段，我们使用 [45] 所用的边界框，并采用相同的策略完成跟踪。我们在 Kinetics-400 [30]和 AVA [23]上运行了 PHALP。这两个数据集都包含超过 100 万条跟踪轨迹，平均长度为 3.4 秒，检测结果超过 1 亿个。总计，我们使用了约 900 小时的跟踪轨迹，比以往工作多约 40 倍 [29]。更多详情请参见表 1。

跟踪使我们能够对动作进行密集训练。由于每个帧中的每个行为主体都有对应的词元，我们可以假设人体动作在一个*1* 秒窗口 [23]内保持不变，从而监督每个词元。首先，我们在 Kinetics-400 数据集 [30]和 AVA [23] 数据集上预训练模型。我们以 1 赫兹的频率在 Kinetics-400 的每条跟踪轨迹上运行 MViT [14] （在 MaskFeat 上预训练 [66]），以生成伪真实标注。每 30 帧共享同一组标注，我们使用二元交叉熵损失对模型进行端到端训练。随后，我们使用自己生成的轨迹，在 AVA 的真实动作标签上微调预训练模型。推理时，我们取一条跟踪轨迹，并从同一视频中随机采样 $N-1$ 其他跟踪轨迹，将其输入模型。我们对

数据集	# 片段	# 轨迹	# 边界框
AVA [23]	184k	320k	32.9m
Kinetics [30]	217k	686k	7140万
总计	400k	1m	1.043亿

表1. **AVA 上的跟踪统计 [23] 和 Kinetics-400 [30]：**我们报告了 PHALP [50] 为每个数据集返回的轨迹数量（ m：百万）。这总计超过900小时的跟踪轨迹，平均长度为3.4秒（存在重叠）。

模型	Pose	OM	PI	PM	mAP
PoTion [9]	2D	-	-	-	13.1
JMRN [53]	2D	7.1	17.2	27.6	14.1
LART-pose	3D (n=1)	12.0	22.0	46.6	22.9
LART-pose	3D (n=5)	13.3	25.9	48.7	24.1

表 2. **基于三维姿态的 AVA 动作识别：** 我们在 AVA 数据集上评估以人为中心的表征 [23]。其中，*OM*：物体操作，*PI*：人物交互，*PM*：人物移动。LART-pose 在人物移动任务上的表现可达到 MViT 模型的约 80%，且无需查看场景信息。

长度为 12 帧的序列上的预测头进行平均池化，并在中心帧上进行评估。有关模型架构、超参数和训练流程/训练时间的更多详情，请参阅附录 A1。

4.1. 基于三维姿态的动作识别

本节将讨论使用三维姿态线索时，我们的方法在 AVA 动作识别上的表现，对应第 3.1 节。我们在 Kinetics-400 和 AVA 数据集上训练三维姿态模型 **LART-pose**。对于 Kinetics-400 的轨迹，我们使用 MaskFeat [66] 伪真实标签；对于 AVA 轨迹，则使用真实标签进行训练。我们训练单人模型和多人模型，以研究人物随时间变化的交互以及人与人之间的交互。我们的方法在多人（N=5）设置下达到 24.1 平均精度均值（见表2）。尽管这一表现远低于当前最先进水平，但这是三维模型首次在 AVA 数据集上达到超过 15.6 平均精度均值的成绩。请注意，AVA 上首次报告的成绩为 15.6 平均精度均值 [23],，而我们的三维姿态模型已经超过了这一基线。

我们在三个 AVA子类别（物体操作（*OM*）、人物交互（*PI*）和人物移动（*PM*））上评估了我们方法的性能。对于人物移动任务（包括跑步、站立、坐着等动作），三维姿态模型达到 48.7 的平均精度均值。相比之下，MaskFeat 在这一子类别中的表现为 58.6 的平均精度均值。这表明，三维姿态模型的表现约为一个强大的当前最先进水平模型的 80%

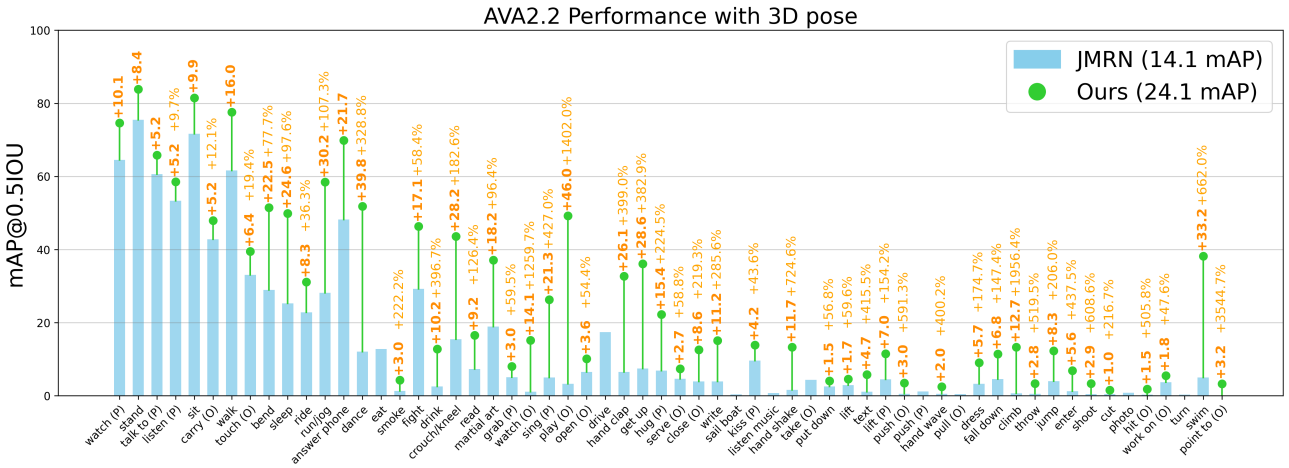


图2. **AVA 各类别的表现：**我们展示了 JMRN [53] 和 LART-pose 在 AVA 的 60 个类别上的表现（平均精度和相对增益）。对于站立、坐着和行走等基于姿态的类别，我们的三维姿态模型仅通过观察随时间变化的三维姿态，平均精度均值就能达到 60 以上。将多条轨迹作为输入进行建模后，我们的模型能够理解人物之间的交互。例如，跳舞(+30.1%）、武术(+19.8%）和拥抱(+62.1%）等活动，相较于当前最先进的纯姿态模型，具有较大的相对增益。我们仅绘制平均精度均值增益大于或小于 1 的结果。

模型。在人与人交互类别上，我们的多人模型相比单人模型提升了 +2.1 的平均精度均值，表明多人模型能够捕捉人与人之间的交互。如图2所示，对于跳舞、打斗、抬起一个人和握手等人与人之间的交互类别，多人模型的表现远优于当前最先进水平的纯姿态模型。例如，在跳舞类别中，多人模型提升了 +39.8 的平均精度均值；在拥抱类别中，相对增益超过 +200%。此外，在人物交互类别中，多人模型相较于单人模型的增益最大。

另一方面，在这三项任务中，物体操作的得分最低。由于我们没有对物体进行显式建模，模型无法获知正在操作的是哪个物体，也不知道该物体与人物之间存在何种关联。不过，有些任务在与物体交互时会呈现独特的姿态，例如接电话或搬运物体；了解姿态有助于识别动作，因此该项的平均精度均值达到 13.3。

4.2. 基于外观和3D姿态的动作

虽然三维姿态模型的性能约为当前最先进方法的 50 %，但它不会推理场景上下文。为对此进行建模，我们将以人为中心的3D表示与来自 MaskFeat 的特征向量拼接起来 [66] ，如第 3.2 节所述。Mask- Feat 采用 MViT 2 [39]作为骨干网络，并学习场景及其上下文化外观的强大表示。首先，我们使用伪真实标签在 Kinetics-400 [30] 和 AVA [23] 数据集上预训练该模型，然后，我们使用真实动作标注，在 AVA 轨迹上微调该模型。

在表3 中，我们将我们的方法与其他最先进的方法进行了比较。总体而言，与 Video MAE **+2相比，我们的方法的平均精度均值提高了 .8** 。 [17, 57]此外，如果使用来自 AVA-Kinetics 的额外标注进行训练，我们的方法可达到**42.3 平均精度均值**。图 3 展示了我们的方法与 MaskFeat [66]在各类别上的性能对比。总体而言，我们的方法在 60 个类别中的 56 个类别上提升了性能。对于某些类别（例如，打斗、拥抱、攀爬），我们的方法将性能提高了超过 +5平均精度均值。在表4中，我们在 AVA-Kinetics [38]数据集上评估了我们的方法。与此前最先进的方法相比，我们的方法的平均精度均值提高了 +1.5。

在图4中，我们展示了 MViT [14]和我们方法的定性结果。如图所示，显式获取场景中每个人的轨迹，使我们能够对拥抱和打斗等动作做出更有把握的预测，因为近距离互动很容易解读。此外，骑马和攀爬等动作也能从随时间变化的显式三维姿态中受益。最后，三维网格的模态完整性也使我们能够在发生遮挡时做出更准确的预测。

4.3. 消融实验

跟踪的影响：目前所有关于动作识别的研究都没有显式地将人物随时间关联起来。它们只使用中间帧边界框来预测动作。例如，当一个人物跑步穿过场景，从左向右移动时，根据中间帧边界框裁剪出的特征体不太可能包含关于该人物的全部信息。不过，如果我们能够跟踪这个人物

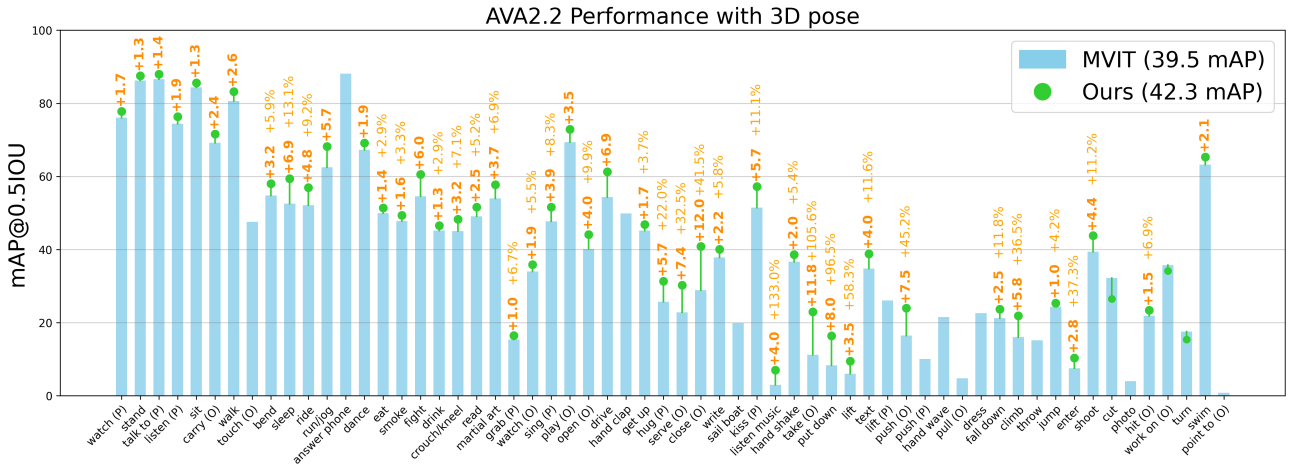


图3。 **与最先进方法的比较：** 我们展示了 MViT 的类别级性能（平均精度和相对增益） [14]（在 MaskFeat 上预训练 [66]）以及我们的方法。与 MViT 相比，我们的方法在 60 个类别中的 50 多个类别上取得了更好的性能。尤其是在跑步、打斗、拥抱和睡觉等动作上，我们的方法取得了超过 **+5 平均精度均值**。这表明，在动作识别中，能够使用显式轨迹和三维姿态具有明显优势。我们仅绘制增益高于或低于 1 个平均精度均值的类别。

模型	预训练	mAP
SlowFast R101, 8×8 [18]	K400	23.8
MViTv1-B, 64×3 [14]		27.3
SlowFast 16×8 +NL [18]	K600	27.5
X3D-XL [16]		27.4
MViTv1-B-24, 32×3 [14]		28.7
对象Transformer [68]		31.0
ACAR R101, 8×8 +NL [45]		31.4
ACAR R101, 8×8 +NL [45]	K700	33.3
MViT-L↑312, 40×3 [39],	IN-21K+K400	31.6
MaskFeat [66]	K400	37.5
MaskFeat [66]	K600	38.8
Video MAE [17,57]	K600	39.3
Video MAE [17,57]	K400	39.5
Hiera [52]	K700	42.3
LART- MViT	K400	42.6 (+2.8)
LART- Hiera	K700	45.1 (+2.5)

表3. **在 AVA 2.2 上与最先进方法的比较：**。我们的模型使用来自 MaskFeat 的特征 [66]，采用全裁剪推理。与 Video MAE 相比 [17,57]，我们的方法提升了 **+2.8 平均精度均值**，使用 Heira [52] 骨干网络时，我们的模型达到 45.1 平均精度均值，提升了 **+2.5 平均精度均值**。

我们就能轻松确定其随时间变化的精确位置，从而为模型提供更具局部性的信息来预测动作。

为此，我们首先使用评估中采用的相同检测边界框来评估 MaskFeat [66]，结果为 40.2 平均精度均值。以此作为我们系统的基线，我们训练了一个仅使用 MaskFeat

模型	mAP
SlowFast [18]	32.98
ACAR [45]	36.36
RM [19]	37.34
LART+MViT	**38.91**

表4. **AVA-Kinetics 数据集上的性能**。我们使用单个模型（不使用集成）在 AVA-Kinetics [38] 上评估了模型的性能，并将其与此前单模型的当前最先进水平进行了比较。

特征作为输入、但随时间变化的模型。通过这种方式，我们可以衡量跟踪对动作识别的影响。不出所料，如表 5 所示，在训练 MaskFeat 时加入跟踪后，模型的表现 +1 比基线高出 2 个平均精度均值。这清楚地表明，使用跟踪有助于动作识别。具体而言，利用轨迹可以定位人物随时间的位置变化，进而提供关节角度随时间变化的二阶信号。此外，了解每个人的身份也能提供区分不同人物的信号，有助于学习人物之间的交互。

姿态的影响： 我们工作的第二项贡献是将三维姿态信息用于动作识别。如第 4.1 节所述，仅使用三维姿态即可在 AVA 数据集上达到 24.1 的平均精度均值。虽然难以精确衡量三维姿态和 2D 特征各自的贡献，但我们将本方法与仅使用 MaskFeat 和跟踪轨迹训练的模型进行比较，两者唯一的区别在于是否使用三维姿态。如表 5 所示，加入三维姿态后，平均精度均值提升了 +0.8。与.....相比，这一提升相对较小

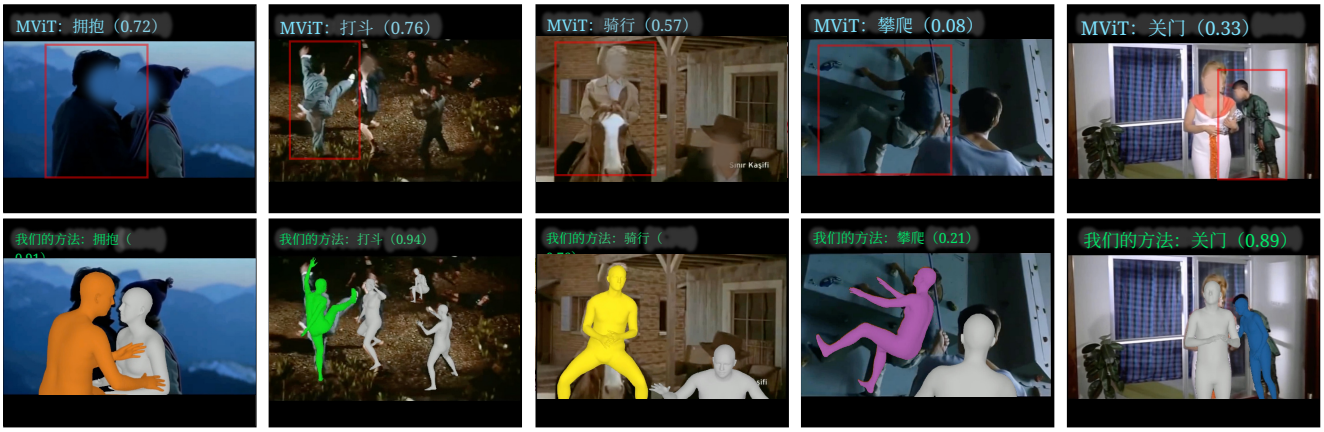


图 4. **定性结果：** 我们展示了 MViT 的预测结果 [14] 以及我们模型在 AVA v2.2 验证样本上的预测结果。彩色网格所标示的人物是我们识别其动作的关注人物，而带有灰色网格的人物则是辅助人物。前两列展示了获取其他人物的动作轨迹管对的益处。第一列中，橙色人物与另一位呈拥抱姿势的人物非常接近，因此更容易以较高概率预测为拥抱。同样，在第二列中，多人之间明确的交互，以及得知其他人也在打斗，都提高了打斗动作的置信度，使其高于二维识别模型对绿色人物的预测。第三列和第四列展示了在动作识别中，显式建模随时间变化的三维姿态（使用跟踪轨迹）的益处。其中，黄色人物处于骑行姿势，而紫色人物向上看，双腿位于垂直平面上。最后一列展示了用模态无关表示来表征人物的益处。此处，蓝色人物的手被遮挡，因此二维识别模型无法完整识别该动作。然而，SMPL 网格不受模态限制，因此手部仍然存在，从而提高了对关门这一动作标签的预测概率。

关于跟踪的使用，我们认为，借助更稳健、更准确的三维姿态系统，这一问题可以得到改善。

模型	OM	PI	PM	mAP
MViT	32.2	41.1	58.6	40.2
MViT +跟踪	33.4	43.0	59.3	41.4 (+1.2)
MViT +跟踪 + 姿态	34.4	43.9	59.9	42.3 (+0.9)

表5。**主要组件的消融实验：**我们使用相同的检测结果，对跟踪和三维姿态的作用进行消融实验。首先，我们仅使用轨迹上的 MViT 特征，以评估跟踪的贡献。然后，我们加入三维姿态特征，以研究三维姿态对动作识别的贡献。

4.4. 实现细节

在姿态模型和姿态+外观模型中，我们采用相同的标准 Transformer 架构 [60]，包含 16 层和 16 个注意力头。两个模型的嵌入维度均为 512。训练时使用 0.4 的掩码比例，测试时则使用相同的掩码标记来填补缺失的检测结果。Transformer 的输出标记被传入线性层，以预测 AVA 动作标签。我们使用 MViT [14] 的预测结果作为伪监督，在 Kinetics 上对模型进行 30 个训练轮次的预训练，随后使用 AVA 真实标签在 AVA 上进行少量轮次的微调。我们使用 AdamW [41] 训练模型，基础学习率为 0.001 以及 betas $= (0.9, 0.95)$。我们采用带线性预热的余弦退火调度。更多细节请参见附录。

4.5. Hiera 骨干网络结果

在本节中，我们展示了使用 Hiera [52] 提取上下文化外观特征时，各类别的表现。我们的模型比**Hiera基线高出 2.5 个平均精度均值**。总体而言，我们的方法取得了 45.1 的平均精度均值，并在多个动作类别上取得了显著提升。表 6展示了 Hiera 模型和 LART 的各任务表现。在所有三个类别中，LART 的表现都远优于基线模型，并且在人物交互任务上取得了显著提升。

模型	OM	PI	PM	mAP
MViT [66]	32.2	41.1	58.6	40.5
LART+MViT	34.5	44.0	59.9	42.3
Hiera [52]	34.5	43.0	62.3	42.6
LART+Hiera	37.3	45.7	63.8	45.1

表6. **骨干网络效果：** AVA 数据集 [23] 包含*OM*：物体操作、*PI*：人物交互，以及*PM*：人物移动任务。我们比较了骨干网络（MViT [66]和 Hiera [52]）的动作检测性能，以及与 LART 结合时的性能。

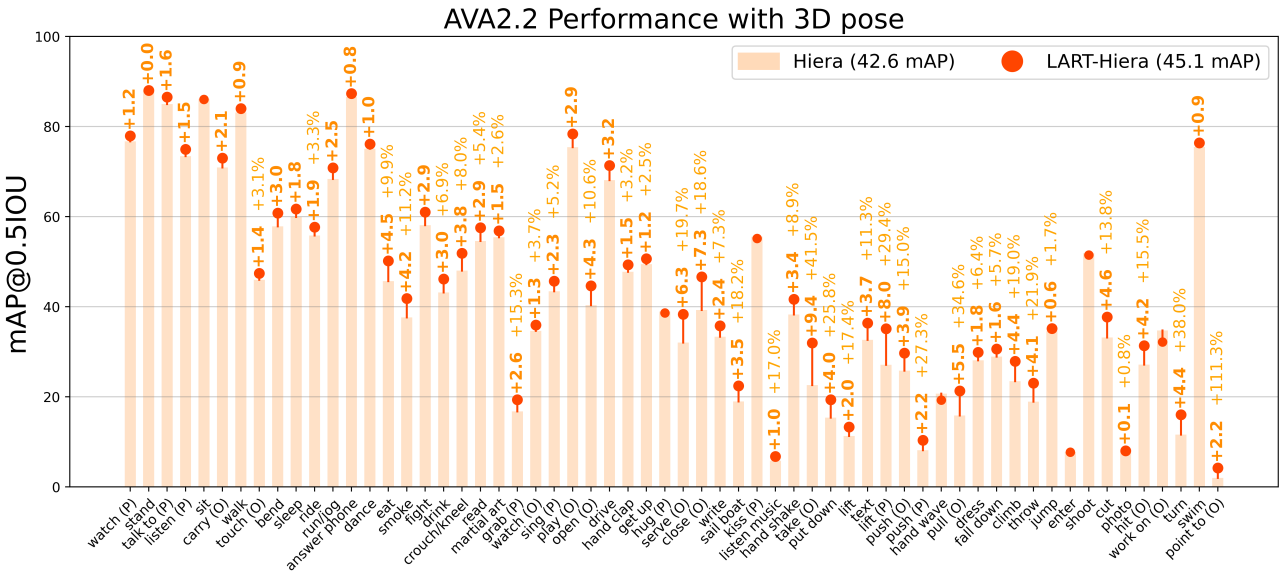


图5。 **与 Hiera 的比较 [52]:** 我们展示了 Hiera [52]（使用 MAE 进行预训练 [24]）与我们方法的类别级性能（平均精度和相对增益）。在 60 个类别中，我们的方法有 53 个类别的性能优于 Hiera。尤其是在涉及物体交互和姿态变化的动作中，例如跑步、打斗，我们的方法在平均精度均值上取得了超过 10% 的相对增益。这表明，获取显式轨迹和三维姿态有益于动作识别。我们仅绘制高于或低于 1 个平均精度均值点的增益。

5. 结论

本文探讨了三维跟踪和姿态对人体动作识别任务的益处。借助当前最先进的人物跟踪方法 PHALP [50], , 我们训练了一个 Transformer模型，其输入词元表示人物在每个时刻的状态。我们研究了词元内容的两种设计选择。首先，使用人物三维姿态信息时，在 AVA v2.2 数据集上，我们的表现比以姿态信息进行动作识别的先前基线方法高出 8.2 个平均精度均值。其次，我们还提出将姿态信息与来自典型动作识别骨干网络的上下文化外观信息相融合 [14]，并将其应用于短轨迹序列。采用该模型后，我们在 AVA v2.2 上比先前的当前最先进水平提升了 2.8 个平均精度均值。使用 Heira [52] 等更强的骨干网络时，LART 在 AVA v2.2 动作检测任务上达到 45.1 的平均精度均值。动作识别仍有许多未来研究方向和改进空间。例如，可以通过更具表达力的人体三维重建（例如，使用 SMPL-X 模型 [46] 以同时捕捉手部），并显式建模场景中的物体（可能将“管状结构”的思路扩展到物体），来提升更细粒度任务的表现。

致谢： 本研究得到 FAIR-BAIR项目和 ONR MURI（ N00014-21- 1-2801）的支持。我们感谢 Shubham Goel 提供了有益的讨论。

参考文献

[1] Anurag Arnab、Mostafa Dehghani、Georg Heigold、Chen Sun、Mario Lu˘ci´c 和 Cordelia Schmid。《ViViT: 视频视觉Transformer》。发表于*ICCV*，2021年。[2] Ji mmy Lei Ba、Jamie Ryan Kiros 和 Geoffrey E Hin- ton。《层归一化》。*arXiv*预印本 *arXiv:1607 06450*，2016年。. [3] Fabien Baradel、Thibault Groueix、Philippe Weinzaepfel、Romain Br´egier、Yannis Kalantidis 和 Gr´egory Rogez。《利用MoCap数据进行人体网格恢复》。发表于*3DV*，2021年。[4] Philipp Bergmann、Tim Meinhardt 和 Laura Leal-Taixe。《无须繁琐操作的跟踪》。发表于*ICCV*，2019年。[5] Gedas Bertasius、Heng Wang 和 Lorenzo Torresani。《视频理解只需要时空注意力吗？》。发表于*ICML*，2021年。[6] Federica Bogo、Angjoo Kanazawa、Christoph Lassner、Peter Gehler、Javier Romero 和 Michael J Black。《坚持使用SMPL：从单张图像自动估计三维人体姿态与形状》。发表于*ECCV*，2016年。[7] Joao Carreira 和 Andrew Zisserman。《动作识别何去何从？一种新模型与Kinetics数据集》。发表于*CVPR*，2017年。[8] Vasileios Choutas、Lea M¨uller、Chun-Hao P Huang、Siyu Tang、Dimitrios Tzionas 和 Michael J Black。《利用度量与语义属性实现精准的三维人体形状回归》。发表于*CVPR*，2022年。

[9] Vasileios Choutas、Philippe Weinzaepfel、J´erˆome Revaud 和 Cordelia Schmid。《PoTion：用于动作识别的姿态运动表示》。发表于*CVPR*，2018年。[10] Kevin Clark、Minh-Thang Luong、Quoc V Le 和 Christo- pher D Manning。《Electra：将文本编码器预训练为判别器而非生成器》。*arXiv*预印本 *arXiv:2003.10555*，2020年。[11] Piotr Doll´ar、Vincent Rabaud、Garrison Cottrell 和 Serge Belongie。《利用稀疏时空特征进行行为识别》。发表于*2005年IEEE*视觉监控与跟踪及监控性能评估国际研讨会，2005年。[12] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Syl- vain Gelly、Jakob Uszkoreit 和 Neil Houlsby。《一张图像胜过16×16个词：面向大规模图像识别的Transformer》。2021年。[13] Alexei A Efros、Alexander C Berg、Greg Mori 和 Jitendra Malik。《远距离动作识别》。发表于*ICCV*，2003年。[14] Haoqi Fan、Bo Xiong、Karttikeya Mangalam、Yanghao Li、Zhicheng Yan、Jitendra Malik 和 Christoph Feichtenhofer。《多尺度视觉Transformer》。发表于*ICCV*，2021年。[15] Hao-Shu Fang、Shuqin Xie、Yu-Wing Tai 和 Cewu Lu。《RMPE：区域多人姿态估计》。发表于*ICCV*，2017年。[16] Christoph Feichtenhofer。《X3D：扩展架构以实现高效视频识别》。发表于*CVPR*，2020年。[17] Christoph Feichtenhofer、Haoqi Fan、Yanghao Li 和 Kaim- ing He。《作为时空学习器的掩码自编码器》。发表于*NeurIPS*，2022年。[18] Christoph Feichtenhofer、Haoqi Fan、Jitendra Malik 和 Kaiming He。《用于视频识别的SlowFast网络》。发表于*ICCV*，2019年。[19] Yutong Feng、Jianwen Jiang、Ziyuan Huang、Zhiwu Qing、Xiang Wang、Shiwei Zhang、Mingqian Tang 和 Yue Gao。《时空动作定位中的关系建模》。发表于*CVPRW*，2021年。[20] Golnaz Ghiasi、Yin Cui、Aravind Srinivas、Rui Qian、Tsung- Yi Lin、Ekin D Cubuk、Quoc V Le 和 Barret Zoph。《简单的复制粘贴是一种强大的实例分割数据增强方法》。发表于*CVPR*，2021年。[21] Georgia Gkioxari 和 Jitendra Malik。《查找动作管道》。发表于*CVPR*，2015。[22] Shubham Goel、Georgios Pavlakos、Jathushan Rajasegaran、Angjoo Kanazawa 和 Jitendra Malik。《4D人体：利用Transformer重建和跟踪人体》。发表于*ICCV*，2023年。[23]Chunhui Gu、Chen Sun、David A Ross、Carl Von- drick、Caroline Pantofaru、Yeqing Li、Sudheendra Vijaya- narasimhan、George Toderici、Susanna Ricco、Rahul Suk- thankar、Cordelia Schmid 和 Jitendra Malik。《AVA：一个包含时空定位的原子视觉动作的视频数据集》。发表于*CVPR*，2018年。[24] Kaiming He、Xinlei Chen、Saining Xie、Yanghao Li、Piotr Doll´ar 和 Ross Girshick。《掩码自编码器是可扩展的视觉学习器》。发表于*CVPR*，2022年。[25] Kaiming He、Georgia Gkioxari、Piotr Doll´ar 和 Ross Gir- shick。《Mask R-CNN》。发表于*ICCV*，2017年。[26] Gao Huang、Yu Sun、Zhuang Liu、Daniel Sedra 和 Kilian Q Weinberger。《随机深度深度网络》。发表于*ECCV*，2016年。[27] Gunnar Johansson。生物运动的视觉感知及其分析模型。发表于《知觉与心理物理学》，14(2):201–211，1973年。[28] Angjoo Kanazawa、Michael J Black、David W Jacobs 和 Jitendra Malik。《人体形状与姿态的端到端恢复》。发表于*CVPR*，2018年。[29] Angjoo Kanazawa、Jason Y Zhang、Panna Felsen 和 Jiten- dra Malik。《从视频中学习三维人体动态》。发表于*CVPR*，2019年。[30] Will Kay、Joao Carreira、Karen Simonyan、Brian Zhang、Chloe Hillier、Sudheendra Vijayanarasimhan、Fabio Viola、Tim Green、Trevor Back、Paul Natsev 等。《Kinetics 人体动作视频数据集》。发表于*arXiv* 预印本 *arXiv:1705 06950*，2017。。[31] Machiel Keestra。《理解人体动作：整合含义、机制、原因与情境》。2015年。[32] Alexander Klaser、Marcin Marszałek 和 Cordelia Schmid。《基于三维梯度的时空描述符》。发表于*BMVC*，2008年。[33] Muhammed Kocabas、Nikos Athanasiou 和 Michael J Black。《VIBE：用于人体姿态与形状估计的视频推断》。发表于*CVPR*，2020年。[34] Muhammed Kocabas、Chun-Hao P Huang、Otmar Hilliges 和 Michael J Black。《PARE：用于三维人体估计的部件注意力回归器》。发表于*ICCV*，2021年。[35] Muhammed Kocabas、Chun-Hao P Huang、Joachim Tesch、Lea Muller、Otmar Hilliges 和 Michael J Black。《SPEC：借助估计相机在野外场景中观察人物》。发表于*ICCV*，2021年。[36] Nikos Kolotouros、Georgios Pavlakos、Michael J Black 和 Kostas Daniilidis。《通过循环中的模型拟合学习重建三维人体姿态与形状》。发表于*ICCV*，2019年。[37] Nikos Kolotouros、Georgios Pavlakos、Dinesh Jayaraman 和 Kostas Daniilidis。《用于人体网格恢复的概率建模》。发表于*ICCV*，2021年。[38] Ang Li、Meghana Thotakuri、David A Ross、Jo˜ao Car- reira、Alexander Vostrikov 和 Andrew Zisserman。《AVA-Kinetics 局部化人体动作视频数据集》。发表于*arXiv* 预印本 *arXiv:2005.00214*，2020年。[39] Yanghao Li、Chao-Yuan Wu、Haoqi Fan、Karttikeya Man- galam、Bo Xiong、Jitendra Malik 和 Christoph Feichten-hofer。《MViTv2：用于分类与检测的改进型多尺度视觉Transformer》。发表于*CVPR*，2022年。[40] Matthew Loper、Naureen Mahmood、Javier Romero、Gerard Pons-Moll 和 Michael J Black。《SMPL：蒙皮多人线性模型》。发表于《*ACM* 图形学汇刊》（*TOG*），34(6):1–16，2015年。[41] Ilya Loshchilov 和 Frank Hutter。《解耦权重衰减正则化》。发表于*arXiv* 预印本 *arXiv:1711.05101*，2017年。[42] Ilya Loshchilov 和 Frank Hutter。《SGDR：带热重启的随机梯度下降》。发表于*ICLR*，2017年。

[43] Tim Meinhardt、Alexander Kirillov、Laura Leal-Taixe 和 Christoph Feichtenhofer。《TrackFormer：使用Transformer进行多目标跟踪》。发表于*CVPR*，2022年。[44] Daniel Neimark、Omri Bar、Maya Zohar 和 Dotan Asselmann。《视频变换器网络》。发表于*ICCV*，2021年。[45] Junting Pan、Siyu Chen、Mike Zheng Shou、Yu Liu、Jing Shao 和 Hongsheng Li。《行为主体-情境-行为主体关系网络：用于时空动作定位》。发表于*CVPR*，2021年。[46]Georgios Pavlakos、Vasileios Choutas、Nima Ghorbani、Timo Bolkart、Ahmed AA Osman、Dimitrios Tzionas 和 Michael J Black。《表现力人体捕捉：从单张图像恢复三维手部、面部与人体》。发表于*CVPR*，2019年。[47] Georgios Pavlakos、Jitendra Malik 和 Angjoo Kanazawa。《从多张图像中恢复人体网格》。发表于*CVPR*，2022年。[48] Ilija Radosavovic、Raj Prateek Kosaraju、Ross Girshick、Kaiming He 和 Piotr Doll´ar。《设计网络设计空间》。发表于*CVPR*，2020年。[49] Jathushan Rajasegaran、Georgios Pavlakos、Angjoo Kanazawa 和 Jitendra Malik。《利用三维表示跟踪人物》。发表于*NeurIPS*，2021年。[50] Jathushan Rajasegaran、Georgios Pavlakos、Angjoo Kanazawa 和 Jitendra Malik。《通过预测三维外观、位置与姿态来跟踪人物》。发表于*CVPR*，2022年。[51] Davis Rempe、Tolga Birdal、Aaron Hertzmann、Jimei Yang、Srinath Sridhar 和 Leonidas J Guibas。《HuMoR：用于鲁棒姿态估计的三维人体运动模型》。发表于*ICCV*，2021年。[52] Chaitanya Ryali、Yuan-Ting Hu、Daniel Bolya、Chen Wei、Haoqi Fan、Po-Yao Huang、Vaibhav Aggarwal、Arkabandhu Chowdhury、Omid Poursaeed、Judy Hoffman、Jitendra Ma- lik、Yanghao Li 和 Christoph Feichtenhofer。《Hiera：无需繁复设计的分层视觉Transformer》。发表于*ICML*，2023年。[53] Anshul Shah、Shlok Mishra、Ankan Bansal、Jun-Cheng Chen、Rama Chellappa 和 Abhinav Shrivastava。《姿态与关节感知动作识别》。发表于*WACV*，2022年。[54] Karen Simonyan 和 Andrew Zisserman。《用于视频动作识别的双流卷积网络》。发表于*NIPS*，2014年。[55] Chen Sun、Abhinav Shrivastava、Carl Vondrick、Rahul Sukthankar、Kevin Murphy 和 Cordelia Schmid。《关系动作预测》。发表于*CVPR*，2019年。[56] Graham W Taylor、Rob Fergus、Yann LeCun 和 Christoph Bregler。《时空特征的卷积学习》。发表于*ECCV*，2010年。[57] Zhan Tong、Yibing Song、Jue Wang 和 Limin Wang。《VideoMAE：掩码自编码器是数据高效的自监督视频预训练学习器》。发表于*NeurIPS*，2022年。[58] Du Tran、Lubomir Bourdev、Rob Fergus、Lorenzo Torresani 和 Manohar Paluri。使用三维卷积网络学习时空特征。发表于 ICCV，2015年。[59] G¨ul Varol、Ivan Laptev、Cordelia Schmid 和 Andrew Zis- serman。利用合成的人体，从未见过的视角识别动作。IJCV，2021年。[60] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention is All You Need。发表于 NIPS，2017年。[61] Heng Wang、A. Klaser、C. Schmid 和 Cheng-Lin Liu。通过密集轨迹进行动作识别。发表于 CVPR，2011年。[62] Heng Wang 和 Cordelia Schmid。使用改进轨迹进行动作识别。发表于 ICCV，2013年。[63] Xiaolong Wang、Ross Girshick、Abhinav Gupta 和 Kaim- ing He。非局部神经网络。发表于 CVPR，2018年。[64] Xiaolong Wang 和 Abhinav Gupta。将视频表示为时空区域图。发表于 ECCV，2018年。[65] Zelun Wang 和 Jyh-Charn Liu。使用深度神经网络和序列级训练将数学公式图像转换为 LaTeX序列。国际文档分析与识别期刊（IJDAR），24(1)：63–75页，2021年。[66] Chen Wei、Haoqi Fan、Saining Xie、Chao-Yuan Wu、Alan Yuille 和 Christoph Feichtenhofer。用于自监督视觉预训练的掩码特征预测。发表于 CVPR，2022年。[67] Philippe Weinzaepfel 和 Gr´egory Rogez。Mimetics：迈向对上下文之外人体动作的理解。IJCV，2021年。[68] Chao-Yuan Wu 和 Philipp Krahenbuhl。迈向长视频理解。发表于 CVPR，2021年。[69] Yuxin Wu、Alexander Kirillov、Francisco Massa、Wan-Yen Lo 和 Ross Girshick。Detectron2。https://github. com/facebookresearch/detectron2，2019年。[70] Yuliang Xiu、Jiefeng Li、Haoyu Wang、Yinghong Fang 和 Cewu Lu。Pose Flow：高效的在线姿态跟踪。发表于 BMVC，2018 年。[71] An Yan 、 Yali Wang 、 Zhifeng Li 和 Yu Qiao。PA3D：用于视频识别的姿态-动作三维机器。发表于 CVPR，2019年。[72] Hongwen Zhang、Yating Tian、Xinchi Zhou、Wanli Ouyang、Yebin Liu、Limin Wang 和 Zhenan Sun。PyMAF：通过金字塔网格对齐反馈循环进行三维人体姿态和形状回归。发表于 ICCV，2021年。[73] Yubo Zhang、Pavel Tokmakov、Martial Hebert 和 Cordelia Schmid。用于动作检测的结构化模型。发表于 CVPR，2019年。

A. 补充材料

由于篇幅限制，本文未能在正文中介绍我们方法的更多细节，因此本文档将对此作补充。我们补充介绍了方法的实验和实现细节（第 2 和第 3 节），进一步说明了论文中的实验（第 4 节），并列出了 Kinetics-400 和 AVA 的训练配置，以便复现实验。

B. 附加结果

B.1. 对场景中的多个人物进行建模

正如第 3 节所述，只要内存足够，我们的模型可以接收任意数量的人物（轨迹）。尽管我们通过简单的随机采样来寻找场景中的辅助人物，但了解其他人物所在的位置以及他们正在做什么，可能是预测目标人物正在做什么的重要信号。我们通过改变人物的最大上下文范围来训练多个模型 ($n \in [1, 2, 3, 4, 5]$)。当 $n = 1$时，LART 只观察关注人物，而有关场景和其他人物的信息则通过上下文化外观向量输入。然而，当 n较大时，其他人物的姿态、位置和外观会被明确提供给模型。如图 6 所示，随着我们增加 LART-pose 模型的人物上下文范围，其在 AVA 数据集上的性能单调提升，并开始趋于饱和于 $n > 4$。

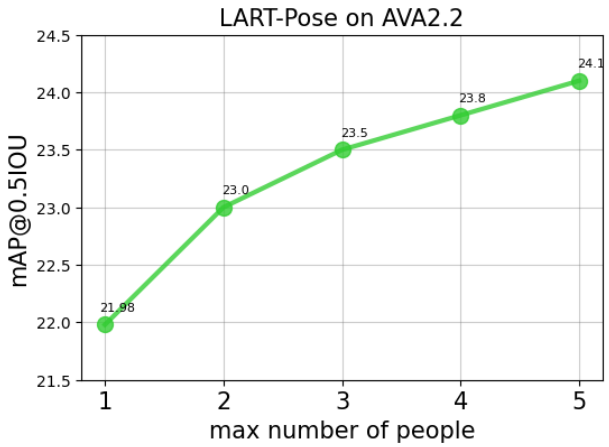


图 6. **LART-pose 的性能与场景中的人物数量之间的关系：**我们展示了 LART-pose 在不同 n（模型在每帧中看到的最多人物数量）下的性能。该图表明，当我们增加模型在推理目标人物时能够看到的人物数量，性能会持续提升。虽然我们的多人模型非常简单；我们只是为场景中的其他人物输入额外的词元，但模型能够从大规模训练数据中理解这些交互。

这一观察突显了对场景中的多个人物进行建模并联合学习其交互的益处。然而，只有在我们能够跟踪场景中的每个人物时，才能实现这些益处。我们观察到，采用简单的随机采样时，性能在 $n = 5$ 处趋于饱和，但这一观察可能受到 AVA 数据集的严重偏差影响。大多数电影中的角色并不多，而电影拍摄方式所造成的偏差（例如，亲吻和拥抱场景通常采用特写镜头）也会影响这些结果。需要注意的是，所有这些模型的参数数量相同，并且训练时长也相同（在同一数据集上训练 30 个训练轮次）。

总之，我们的研究结果表明，对多人交互进行建模可以显著提升动作识别模型的性能，尤其是对于涉及近距离人-人交互或群体活动的动作。不过，饱和点可能因数据集以及场景拍摄方式所带来的偏差而异。

B.2. 单人对比多人结果

在上一节中，我们讨论了对多人交互进行建模的益处。本节将研究从单人模型转为多人模型能带来多大的性能提升。我们在图=1中比较 LART-pose (n=5) 和 LART-pose (n7) 的表现。与单人模型相比，多人模型的平均精度均值总体提升了**2.0**。

仔细考察各类别的表现后可以发现，涉及近距离人-人交互的类别从多人训练中获益显著。例如，拥抱、亲吻、握手、抬起一个人和听一个人说话等动作，相较于单人模型提升了**10% 以上**。这些交互发生在彼此距离很近的情况下，因此，明确了解另一人物的 3D 位置、姿态和动作，有助于识别主要行为者的动作。

除了近距离交互动作外，群体动作也能从利用其他人物动作的上下文中获益，从而推断目标人物的动作。例如，跳舞 和游泳通常属于群体活动，了解其他人在做什么有助于推断动作。

与单人模型相比，该多人模型在 50 多个类别上取得了提升，其中 30 多个类别的提升超过**1 平均精度均值**。然而，正如上一节关于数据集偏差所述，这些结果可能会因数据集类型不同而略有差异，例如体育数据集。举例来说，在体育场景中，可能需要观察超过 5 个人，才能识别运动员的动作；而且体育场景的拍摄方式与电影截然不同。

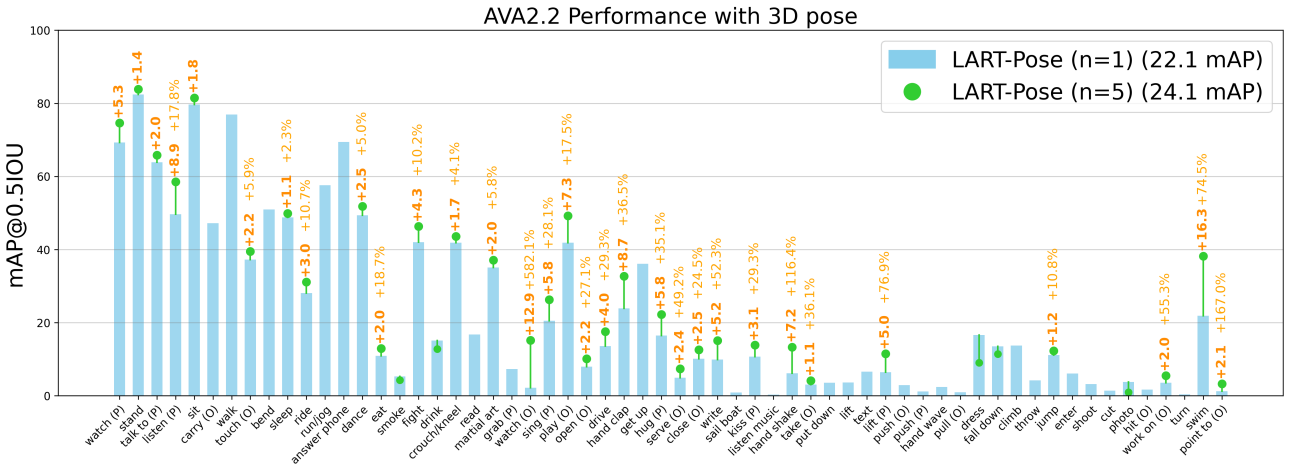


图7. **AVA 上按类别划分的性能：**我们比较了单人设置（$n = 1$）和多人设置（$n = 5$）下的 LART-pose。在超过 50 个类别上，我们的多人模型优于单人模型；在某些人与人互动类别上，多人模型的相对增益约为 **10%**。

总的来说，我们的研究结果表明，对多人交互进行建模可以显著提升动作识别模型的性能，尤其是对于涉及人物之间紧密交互或群体活动的动作。

C. 实现细节

我们的完整跟踪动作识别系统整合了多个子系统，以结合二维检测、识别、三维重建和跟踪方面的最新进展。整体流程可分为两部分：**a)** 帧到实体和**b)** 实体到动作。

第一部分是从帧中提取实体（此处我们将实体=视为人物）。为此，我们采用当前最先进的跟踪算法 PHALP [50]。PHALP 的第一步是使用 Mask R-CNN 检测每一帧中的人物 [25]。我们使用了 Detectron2 的 [69]新基线模型，这些模型采用简单复制粘贴数据增强 [20] 进行训练，并使用 RegNet- 4gf [48]骨干网络来执行检测任务。检测出每一帧中的人物后，PHALP 使用 HMR [22, 28, 49] 重建每个人物的三维形态。随后，系统会预测每个人物未来的位置、姿态和外观，以解决关联问题。PHALP 使用匈牙利匹配来求解三维检测结果与三维预测结果之间的关联。最后，跟踪过程会返回一组轨迹，使我们能够随时间访问这些实体（人物）。

在论文的第二部分，我们收集轨迹，并用它们训练一个用于动作识别的 Transformer 模型。网络架构以及训练和推理流程的更多细节将在后续章节中讨论。这部分是我们利用 Transformer 模型从三维轨迹中识别动作这一方法的关键组成部分。

C.1. PHALP轨迹片段

在这项工作中，我们对 Kinetics-400 [30] 和 AVA [23]中的每个人物进行跟踪。为此，我们使用了近期提出的三维跟踪算法 PHALP [50]。PHALP 能够稳健地跟踪自然场景中的人物，并生成其三维表示。不过，AVA 的动作真实标注以每秒 1 帧的频率提供，形式为边界框。顺带一提，我们不使用 AVA 数据集中的跟踪真实标注，该标注同样仅以每秒 1 帧的频率提供。首先，当视频帧没有真实标注时，我们使用 PHALP 检测模型（例如，Mask R-CNN）检测人物。如果该帧确实有标注，我们就直接采用真实值边界框，不使用 Mask R-CNN 的检测结果。由于 AVA 仅提供边界框标注，而 PHALP [50]需要边界框和掩码，因此我们使用 Detectron2 [69] 通过 Mask R-CNN 从边界框中提取掩码。对于验证集，我们使用了 ACAR [45], 提供的检测结果，这些结果同样仅每 30 帧提供一次。因此，我们采用类似的策略，从每秒 1 帧可用的边界框中获取跟踪轨迹。对于 Kinetics，我们对整个序列运行 PHALP 跟踪，序列通常由时长为 10 秒的片段组成。然而，由于 AVA 的视频远长于 Kinetics（15 分钟），我们以评估帧为中心，在 4 秒的时间窗口内运行跟踪器。

C.2. 架构细节

在所有实验中，我们使用一个标准的 Transformer [60] 架构，包含 16 层，宽度为 512。每层有 16 个自注意力头，后接层归一化 [2], ，以及一个后接层归一化的 2 层多层感知机。我们训练所有模型时，最大序列长度设为 128

帧/人物。换句话说，每个轨迹片段都会裁剪至 128 帧长的序列。我们使用的唯一数据增强方式是随机选择序列的起始位置进行裁剪。Transformer 块后接一个用于预测 AVA 动作类别的线性层。我们使用二元交叉熵损失训练所有模型。

在训练阶段，我们使用两种注意力掩码。首先，由于轨迹片段可能会因遮挡和漏检而不连续，我们会完全屏蔽这些词元对应的自注意力。这些词元不参与损失计算，因此轨迹片段的这一部分不会影响训练。第二种掩码用于模拟测试阶段出现的这类漏检。我们随机选择少量词元（根据掩码比例），并用可学习的掩码词元替换人物向量。在自注意力层中，我们对注意力进行掩码，使这些被掩码的词元可以关注其他词元，而其他词元不能关注这些被掩码的词元。与第一种掩码不同，我们会对这些被掩码词元的预测计算损失，因为只要有可用的检测结果，就会有伪真实标注或真实标签可用于训练。

在推理阶段，我们不使用任何注意力掩码。不过，某些轨迹片段的检测结果会不连续。在这些位置，我们使用学习到的掩码词元来补全轨迹片段的预测。由于我们要对每一帧密集地预测动作标签，因此会对以人工标注的检测结果为中心的 12 个词元进行平均池化，以尽量缩小人工标注与模型预测之间的差距。

C.2.1 基于 3D 姿态的动作识别

本小节讨论仅利用随时间变化的 3D 姿态信息进行动作识别时所采用的网络架构。3D 姿态包含 226 个参数：其中 207 个（$23\times3\times3$）参数用于表示关节角度，9 个用于表示人物的全局朝向，10 个用于表示身体形状。此外，人物在相机坐标系中的 3D 平移由 3 个参数表示。总的来说，在该系统中，人物向量的维度为 229。该向量由一个具有两个隐藏层的多层感知机编码，并投影为 256 维向量。随后，将投影后的人物向量输入 Transformer。我们还采用了正文（第3节 3.1）中讨论的三种位置编码，分别用于时间、跟踪轨迹和空间。

C.2.2 基于 3D 姿态与外观的动作识别

为编码强大的上下文化外观特征，我们使用了经 MaskFeat [14]预训练的 MViT [66]。用于 AVA 的 MViT 模型以帧序列和一帧中间帧的边界框作为输入，预测目标人物的动作标签（这是采用欧拉式方法预测动作的典型示例）。本文使用一个 MViT-L 40×3模型，以 4 秒视频片段中按 3 帧时间步长采样的 40 帧，以及中间帧中某个人物的边界框作为输入。在 MViT 分类器的线性层之前，这会生成一个 1152 维特征向量。我们将这个 1152 维特征向量用作上下文化外观特征，并通过一个包含两个隐藏层的多层感知机将其编码为 256 维向量。现在，我们有一个姿态向量（256 维，来自上一节）和一个外观向量（256 维）。我们将这两个向量拼接起来，构建用于结合外观的三维姿态人物向量，并将最终的 512 维向量输入 Transformer。

C.3. 训练方案

如正文第3节所述，我们首先使用从 PHALP 获得的跟踪轨迹，在 Kinetics-400 数据集上对我们的方法进行预训练 [50]。除非人物被遮挡，或因检测系统失效而未被检测到，否则每个轨迹片段在每一帧中都有一次检测。我们将这些检测边界框作为 MViT [14] 模型的输入，并为轨迹片段生成伪真实动作标签。生成标签后，我们以轨迹片段为输入、动作标签为输出，对模型进行端到端训练。我们采用表7中的训练配置，在 Kinetics-400 轨迹片段上预训练模型。模型在 Kinetics 轨迹片段上完成预训练后，我们使用真实动作标签，在 AVA 轨迹片段（由 PHALP 生成）上对模型进行微调。最后，在微调阶段，我们采用逐层衰减 [10] 和路径丢弃 [26]。

配置	Kinetics-400	AVA
优化器	AdamW [41]	
优化器动量	$\beta_1, \beta_2 = 0.9, 0.95$	
权重衰减	0.05	
学习率调度	余弦衰减 [42]	
预热轮次	5	
随机失活	0.1	
基础学习率	1e-3	1e-3
逐层衰减 [10]	-	0.9
批量大小	64	64
训练轮数	30	30
路径丢弃 [26]	-	0.1
掩码比例	0.4	0.0

表7. **训练配置：** 我们报告了用于在 Kinetics-400 和 AVA 数据集上训练模型的训练配置。