智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

视觉感知人景互影响的人体动作预测方法

李沁, 陈飞扬, 彭晗, 王勇, 刘利枚, 张伟

引用本文:
李沁, 陈飞扬, 彭晗, 等. 视觉感知人景互影响的人体动作预测方法[J]. 智能系统学报, 2025, 20(4): 1010-1023.
LI Qin, CHEN Feiyang, PENG Han, et al. Human motion prediction method with visual perception of human-scene mutual influence[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(4): 1010-1023.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202411016

DOI: 10.11992/tis.202411016
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250226.1252.005

视觉感知人景互影响的人体动作预测方法

李沁[1,2], 陈飞扬[1], 彭晗[1], 王勇[3], 刘利枚[1], 张伟[4]
(1. 湖南工商大学 人工智能与先进计算学院, 湖南 长沙 410205; 2. 湘江实验室, 湖南 长沙 410205; 3. 中南大学 自动化学院, 湖南 长沙 410083; 4. 字节跳动, 广东 深圳 518063)

摘 要: 场景信息驱动人类调整动作轨迹, 对人体动作预测影响较大。当前研究仅捕获场景信息更新动作特征, 忽略了场景与动作的互影响关系。为此, 提出一种视觉感知人景互影响的人体动作预测方法。提取动作特征和场景特征, 然后循环执行场景信息捕获单元和场景适应度增强单元。前者捕获影响动作的场景信息, 后者利用该信息更新动作特征以增强场景适应性。完成循环后, 得到场景适应型动作特征。基于该特征执行噪声逆扩散完成动作预测。在 3 个数据集上进行实验, 结果表明本文方法的预测误差低于当前主流方法, 验证了其有效性。本文方法将为真实场景中的人体动作预测提供更加可靠的解决方案。

关键词: 人体动作预测; 场景信息; 视觉感知; 动作特征; 场景特征; 人景互影响; 场景适应性; 噪声逆扩散
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2025)04−1010−14

中文引用格式: 李沁, 陈飞扬, 彭晗, 等. 视觉感知人景互影响的人体动作预测方法 [J]. 智能系统学报, 2025, 20(4): 1010–1023.
英文引用格式: LI Qin, CHEN Feiyang, PENG Han, et al. Human motion prediction method with visual perception of human-scene mutual influence[J]. CAAI transactions on intelligent systems, 2025, 20(4): 1010–1023.

Human motion prediction method with visual perception of human-scene mutual influence

LI Qin[1,2], CHEN Feiyang[1], PENG Han[1], WANG Yong[3], LIU Limei[1], ZHANG Wei[4]
(1. School of Artificial Intelligence and Advanced Computing, Hunan Technology and Business University, Changsha 410205, China; 2. Xiangjiang Laboratory, Changsha 410205, China; 3. School of Automation, Central South University, Changsha 410083, China; 4. ByteDance, Shenzhen 518063, China)

Abstract: Scene information drives humans to adjust motion trajectories and greatly influences human motion prediction. Current research only updates motion features with scene information and ignores their mutual influences. Hence, a human motion prediction method with visual perception of human-scene mutual influence is proposed in this paper. Motion and scene features are extracted, and scene information capture and adaptability enhancement are iteratively executed. The former captures scene information affecting human motions, whereas the latter updates motion features with the information to enhance their scene adaptability. After the iteration, the scene-adaptive action features are obtained. Noise inverse diffusion is performed based on the features to complete motion prediction. Experiments conducted on three datasets demonstrate that the proposed method has lower prediction error than the current methods, which verifies its effectiveness. The proposed method provides a more reliable solution for human motion prediction in real scenes.

Keywords: human motion prediction; scene information; visual perception; motion features; scene features; human-scene mutual influence; scene adaptability; noise inverse diffusion

人体动作预测旨在利用已观测人体动作预测未来的动作走向。该研究在机器人应用中扮演着至关重要的角色, 显著提升了人机交互的效率与安全性[1-3]。例如, 在工业生产中, 机器人通过预测工人动作提前调整自身位置避免碰撞; 在康复治疗中, 机器人通过预测患者的动作反应实现实时的动作调整和自适应响应, 从而提供安全、精

收稿日期: 2024−11−15. **网络出版日期**: 2025−02−26.
基金项目: 国家自然科学基金项目 (62202161); 湖南省教育厅科学研究项目 (20A125, 22A0460, 23B0597, 24B0584); 湘江实验室重大项目 (23XJ01007, 23XJ01009); 湖南省自然科学基金项目 (2025-JJ60384).
通信作者: 刘利枚. E-mail: seagullm@163.com.

准的辅助治疗。

当前的人体动作预测研究普遍采用编码−解码双阶段流程: 1) 对已观测人体动作进行编码, 提取动作特征; 2) 基于动作特征进行动作解码, 预测未来的动作走向[4]。在早期研究中, 研究者们主要利用传统的概率模型, 如高斯模型[5] 和马尔可夫模型[6], 来解析已观测人体动作中的时序动态特性, 并以此作为动作特征实现后续预测。然而, 这些模型因受限于其假设条件和统计结构, 仅适用于预测变化简单且具备明显周期性规律的人体动作[4]。近年来, 随着深度学习技术的迅猛发展, 多种深度神经网络模型, 如循环神经网络 (recurrent neural networks, RNNs)[7-13]、卷积神经网络 (convolutional neural networks, CNNs)[14-16]、图卷积网络 (graph convolutional networks, GCNs)[4, 17-23] 等, 在人体动作预测方面展现出超越传统概率模型的显著优势。上述模型凭借其深层次非线性网络结构, 学习人体动作中复杂的语义关联和时−空动态特性, 从而有效提升人体动作预测的准确性。

尽管基于深度学习的方法已取得显著成效, 但其在复杂真实场景中的表现仍有待提升。原因在于, 当前方法忽略了场景信息对人体动作预测的影响。具体而言, 人体动作并非孤立行为, 而是由大脑深层的动机驱动, 旨在满足个体的心理需求和目标。真实场景中, 复杂的场景信息, 包括空间布局、周围物体, 以及其他人类的活动状态, 都会影响个体的行为决策, 促使大脑调整行为动机, 改变原本的动作轨迹, 以更好地适应复杂场景的影响[24]。最近, 一些研究人员尝试捕获与人体存在接触关联的场景信息来引导预测未来人体动作[25-30]。然而, 这一改进在预测精度上仍显不足, 因为它忽视了人类大脑感知世界的基本方式: 视觉反馈。不同于接触关联, 人类大脑主要依靠视觉信息来洞察环境, 预判可能的动作趋势及潜在后果。例如, 面对前方障碍物, 人们通常在视觉指引下即时调整行进路线, 避免碰撞, 而非等到接触后才做出反应; 准备拾取物品时, 个体依据视觉感知到的物品尺寸与形态, 预先调整手部与臂部姿势, 确保精准抓握。由此可见, 行为动机受视觉反馈的影响, 能够引导更合理且及时的动作变化。最近, 文献 [31] 做了初步尝试, 但其在捕获存在视觉反馈的场景信息时没有考虑人体动作与场景信息的互影响关系, 导致预测结果仍无法适应真实场景。

基于此, 本文提出了一种视觉感知人景互影响的人体动作预测方法 (human motion prediction method with visual perception of human-scene mutual influence, VPHSI)。给定已观测的人体动作和场景 RGB 图像, 首先利用预训练编码器分别提取这两类数据的特征, 即动作特征和场景特征。然后, 执行基于人景互影响的动作特征更新过程, 以增强动作特征的场景适应度。这一过程循环执行两个计算单元: 场景信息捕获单元和场景适应度增强单元。前者捕获场景特征中对人体行为具有显著影响的场景信息; 而后者利用捕获的场景信息增强动作特征的场景适应度, 实现对动作特征时−空动态性的有效更新。基于上述过程可以得到场景适应型动作特征。接着, 将该特征作为驱动条件执行噪声逆扩散过程以生成未来人体动作。

本文贡献总结如下:

1) 提出了基于人景互影响的动作特征更新方法。该方法捕捉人体动作的全局−局部动态特性, 用于感知场景图像中存在人体行为影响关联的场景信息, 再基于该信息探索具备场景适应性的人体关节点空间表征, 用于更新动作特征的时−空动态特性。重复上述过程, 得到能够有效适应真实场景的动作特征, 保障后续动作预测的准确性和真实性;

2) 设计了动作特征驱动的噪声逆扩散方法预测未来动作。在场景适应型动作特征的驱动下, 噪声逆扩散方法能够捕捉未来动作与场景信息之间的关联关系, 从而学习到适应真实场景的未来动作时−空变化趋势, 实现真实、准确的人体动作预测;

3) 大量实验表明, VPHSI 在面对不同场景时均展现出优于当前算法的人体动作预测性能。这一结果证明了 VPHSI 在应对复杂真实场景时具备较高的预测精度, 为人体动作预测的实际应用提供了强有力的支持。

1　相关工作

人体动作具备显著的时序动态特性。基于此, 当前研究普遍采用 RNNs 来提取人体动作的时序动态特征用于未来动作预测[7-13]。Martinez 等[7] 使用长短期记忆模块来学习人体动作时序依赖表示, 用于预测不同时长的未来动作。Wolter 等[8] 提出了一种融合复数值−范数保持状态转换策略的新型门控循环单元。该单元有效增强了模型在处理长时序人体动作依赖性时的鲁棒性。Wang 等[11] 提出了一种位置−速度循环编码−解码

神经网络 (position-velocity recurrent encoder-decoder, PVRED), 利用姿态数据、位置信息和位置嵌入向量预测姿态变化速度进而得到未来动作变化趋势。并且, 作者还将单位四元数空间中的均方误差损失函数用于模型训练, 从而确保模型能够学习复杂人体动作的运动模式。

尽管当前基于 RNNs 的方法在预测未来人体动作方面已展现出一定成效, 但其在捕获关节点空间变化规律方面仍存在不足, 这限制了人体动作预测性能的进一步提升。为此, 一些研究人员尝试用其他深度神经网络来捕获人体动作的时−空动态变化规律。例如, Liu 等[14] 基于 CNNs 分别提取人体动作的时−空耦合特征、动态局部−全局特征和全局时序共现特征, 来对人体动作变化形态进行建模, 用于后续的预测任务。Tang 等[15] 提出了一种新颖的时序融合 CNNs, 通过融合人体动作双流关节点信息来预测未来人体动作。时序融合模块包括时序关联模块和时−空强化模块。时序关联模块用于保障双流动作初步预测的时序连续性, 而时−空强化模块则用于提升时−空动作特征的耦合性。上述基于 CNNs 的方法虽然能够对关节点空间布局进行编码, 但忽略了各关节点之间的关联关系。为此, 越来越多的研究逐渐深入探讨关节点关联关系及其对动作时−空变化规律的表征。一类常用的方法是将人体骨架空间分布看作是一类图, 其中关节点作为图的节点而关节点之间的物理连接作为图的边。然后, 利用 GCNs 捕获图中各节点的关联关系用于未来动作预测[4,17-23]。例如, 贺朵[17] 提出一种时空可分离图卷积网络 (space-time-separable graph convolutional network, STSGCN)。该网络将时空图连接矩阵分解为时间亲和矩阵和空间亲和矩阵, 分别捕获相同关节点的时序关联关系和不同关节点的空间关联关系, 用于动作预测。Mao 等[18] 将人体姿态视为由不同对关节点连接而成的图结构, 以此表征各关节点之间的空间关系; 同时, 设计了一类自适应 GCNs 来学习这类图结构的空间分布特性及其语义关联关系。Li 等[19] 提出了一类基于语义关联注意机制的多阶多尺度特征融合网络 (semantic correlation attention-based multiorder multiscale feature fusion network, SCAFF)。该网络提取人体动作中关节点和骨头的时−空依赖特征, 并基于语义相关注意机制捕获不同身体部位与动作时间的语义相关性, 进而对关节点和骨头的时−空依赖特征进行语义增强, 提升其姿态描述能力。

近期研究引入了生成对抗网络 (generative adversarial networks, GANs)、Transformer 和扩散模型等方法, 用于进一步提高人体动作预测性能[32-35]。Liu 等[32] 提出了一种简单有效的复合 GANs 结构, 由对应不同身体部位的局部 GANs 组成, 并通过全局 GANs 进行聚合。局部 GANs 基于局部低维特征生成局部动作, 专注于细化特定身体部位的结构和细节; 全局 GANs 在高维特征空间中对局部动作进行整合和优化, 从而确保动作生成结果的多样性和全面性。Zhao 等[33] 提出了一种基于双向 Transformer 的新型生成对抗网络。该网络包含基于 Transformer 构建的动作生成器和双鉴别器。动作生成器采用了双向处理机制, 其中前向传播过程利用历史人体动作预测未来动作走向, 而后向传播过程则以未来动作预测结果重构历史人体动作。这一机制旨在强化模型的时间一致性。双鉴别器由序列级鉴别器和帧级鉴别器组成, 分别从整体动作一致性和单帧精细度两个层面对人体动作预测结果进行评估。Barquero 等[34] 提出了一类新扩散模型 BeLFusion。该模型在潜在动作表示空间中执行动作生成过程。该空间中的动作特征能够解耦人体动作坐标表示, 使得行为模式能够在不受表现形式限制的情况下自由变化, 从而丰富了动作预测的维度和形式。

尽管上述研究在人体动作预测方面取得了显著进展, 但未考虑场景信息对动作预测的影响, 从而限制了其在实际场景中的应用价值。为此, 一些研究尝试探究场景信息与人体动作的接触关联, 用于更新动作变化规律, 以提高人体动作预测的场景适应能力。Corona 等[25] 构建语义图模型并通过图注意迭代机制学习人体与已知场景信息的关系。语义图模型中的节点表示场景中的人或物体, 边表示它们之间的交互关系。Hassan 等[26] 提出了基于数据驱动的随机动作合成方法。该方法以 3D 空间作为输入, 利用条件变分自编码机对人体动作进行建模, 再基于给定的交互目标进行多点采样以生成多个合理的接触点和方向; 接着, 以交互目标为导向驱动人体动作预测。为了进一步提升预测方法的场景适应能力, 一些研究人员不再依赖已知的场景信息, 而是通过算法设计自动捕捉与人体存在接触关联的场景信息[27-30]。例如, Cao 等[27] 提出了一种新颖的三阶段框架。给定场景图像和已观测人体动作, 该框架对空间中可能的接触点进行采样, 规划朝向每个接触点的人体动作轨迹, 并预测每条路径对应的动作变化过程。Mao 等[28] 提出了基于距离的接触图, 可

自动捕获每个人体部位和 3D 场景空间点在任意时刻的接触关联；接着，作者开发了两阶段网络。该网络首先根据历史人体部位接触图和 3D 点云预测未来动作对应的人体部位接触图，然后基于该接触图生成未来动作轨迹。最近，Gao 等[31]将人体动作预测任务看作是一个以人体动作和视觉反馈场景信息为条件的联合推理问题，并通过潜在空间中的条件驱动扩散过程生成未来人体动作。该方法首次提出捕捉具有视觉反馈的场景信息。相较于接触关联，视觉反馈更符合人类基于视觉感知场景信息更新行为动机的过程。然而，该方法尚未考虑人体动作与场景信息的互影响关系，导致预测结果仍无法适应真实场景。

综上所述，当前的人体动作预测方法主要聚焦于学习人体骨架的关节点关联关系及其时序变化规律。这些方法虽然能够有效学习人体动作的变化规律，但未考虑场景信息对动作变化的影响。针对这一问题，一些研究尝试探究场景信息与人体的接触关联或视觉反馈，但仍未考虑人体动作与场景信息的互影响关系，导致场景适应能力有限。

2　基于视觉感知人景互影响的人体动作预测

人体动作预测任务描述如下。待输入数据为已观测人体动作和场景 RGB 图像。$\boldsymbol{X}_{1:N}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$表示已观测人体动作。其中$N$为时间维度；$\boldsymbol{x}_n\in\mathbb{R}^{J\times 3}$是第 n 个已观测人体状态，由 J 个人体关节点的 3D 位置组成。$\boldsymbol{I}\in\mathbb{R}^{W\times H\times 3}$表示场景图像。任务目标是根据$\boldsymbol{X}_{1:N}$和$\boldsymbol{I}$预测未来人体动作。

VPHSI 结构如图 1 所示，包括 3 个阶段：双模态特征提取、基于人景互影响的动作特征更新以及动作特征驱动噪声逆扩散。

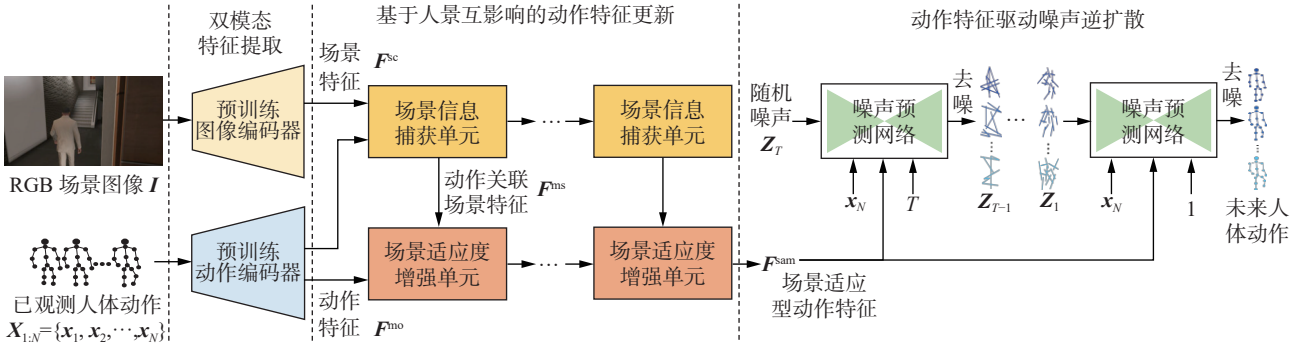

图 1　VPHSI 框架

Fig. 1　Framework of VPHSI

2.1　双模态特征提取

给定$\boldsymbol{X}_{1:N}$和$\boldsymbol{I}$，VPHSI 提取这两类模态的特征。具体地，使用预训练图像编码器提取$\boldsymbol{I}$的场景特征，记为$\boldsymbol{F}^{\mathrm{sc}}\in\mathbb{R}^{S\times D^{\mathrm{sc}}}$。$S$和$D^{\mathrm{sc}}$分别为空间尺寸和特征维度。同时，使用预训练动作编码器提取$\boldsymbol{X}_{1:N}$的动作特征，记为$\boldsymbol{F}^{\mathrm{mo}}\in\mathbb{R}^{J\times D^{\mathrm{mo}}}$。$D^{\mathrm{mo}}$表示动作特征维度。

2.2　基于人景互影响的动作特征更新

得到$\boldsymbol{F}^{\mathrm{sc}}$和$\boldsymbol{F}^{\mathrm{mo}}$后，VPHSI 循环执行两类计算单元，即场景信息捕获单元和场景适应度增强单元实现视觉感知下基于人景互影响的动作特征更新。

2.2.1　场景信息捕获单元

场景信息捕获单元旨在利用$\boldsymbol{F}^{\mathrm{mo}}$的全局−局部动态性捕获$\boldsymbol{F}^{\mathrm{sc}}$中存在人体行为影响关联的场景信息。场景信息捕获单元框架如图 2 所示，具体过程如下。

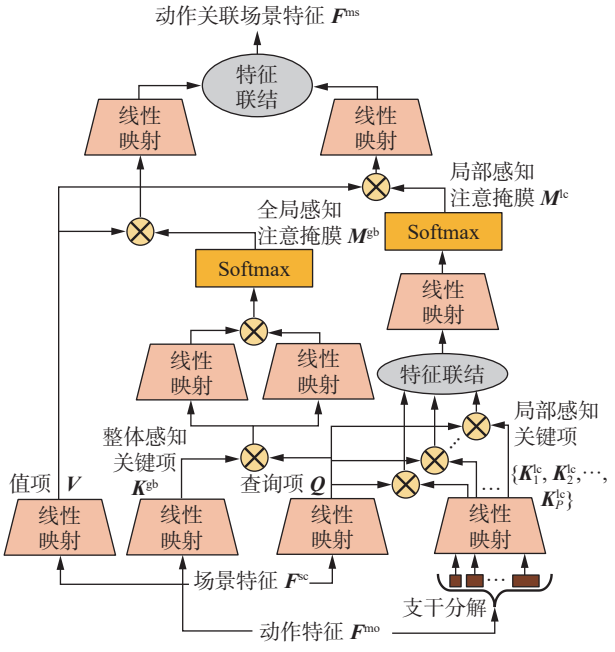

图 2　场景信息捕获单元框架

Fig. 2　Framework of scene information capture unit

1) 利用 $\boldsymbol{F}^{\mathrm{sc}}$ 计算值项 $\boldsymbol{V}\in\mathbb{R}^{S\times D}$ 和查询项 $\boldsymbol{Q}\in\mathbb{R}^{S\times D}$，同时利用 $\boldsymbol{F}^{\mathrm{mo}}$ 计算整体感知关键项 $\boldsymbol{K}^{\mathrm{gb}}\in\mathbb{R}^{J\times D}$。$D$ 表示线性计算模块的输出维度。

2) 计算全局感知注意掩膜：

$$\boldsymbol{R}_1^{\mathrm{gb}}=\boldsymbol{Q}\left(\boldsymbol{K}^{\mathrm{gb}}\right)^{\mathrm{T}}\boldsymbol{W}_1^{\mathrm{in}}+\boldsymbol{b}_1^{\mathrm{in}} \tag{1}$$

$$\boldsymbol{R}_2^{\mathrm{gb}}=\boldsymbol{Q}\left(\boldsymbol{K}^{\mathrm{gb}}\right)^{\mathrm{T}}\boldsymbol{W}_2^{\mathrm{in}}+\boldsymbol{b}_2^{\mathrm{in}} \tag{2}$$

$$\boldsymbol{M}^{\mathrm{gb}}=\sigma\left(\boldsymbol{R}_1^{\mathrm{gb}}\left(\boldsymbol{R}_2^{\mathrm{gb}}\right)^{\mathrm{T}}\right) \tag{3}$$

式中：$\boldsymbol{W}_1^{\mathrm{in}},\boldsymbol{W}_2^{\mathrm{in}}\in\mathbb{R}^{J\times J}$ 为可训练权重矩阵；$\boldsymbol{b}_1^{\mathrm{in}},\boldsymbol{b}_2^{\mathrm{in}}\in\mathbb{R}^{J}$ 是可训练偏置向量；$\boldsymbol{R}_1^{\mathrm{gb}},\boldsymbol{R}_2^{\mathrm{gb}}\in\mathbb{R}^{S\times J}$ 为计算得到的中间表示；$\sigma(\cdot)$ 表示 Softmax 函数；$\boldsymbol{M}^{\mathrm{gb}}\in\mathbb{R}^{S\times S}$ 表示全局感知注意掩膜。

3) 根据人体骨架分布结构，将 $\boldsymbol{F}^{\mathrm{mo}}$ 分解成 P 个不同身体支干对应的时−空动态特征，再基于线性计算模块计算不同身体部位时−空动态特征的局部感知关键项，用 $\{\boldsymbol{K}_1^{\mathrm{lc}},\boldsymbol{K}_2^{\mathrm{lc}},\cdots,\boldsymbol{K}_P^{\mathrm{lc}}\}$ 表示，其中 $\boldsymbol{K}_p^{\mathrm{lc}}\in\mathbb{R}^{J_p^{\mathrm{pa}}\times D}$。$J_p^{\mathrm{pa}}$ 表示第 p 个支干包含的关节点数量。需要说明的是，P 值的设定需要考虑动作局部多样性和计算成本。具体分析详见 3.3.2 节的消融实验。

4) 用 $\boldsymbol{Q}$ 和 $\{\boldsymbol{K}_1^{\mathrm{lc}},\boldsymbol{K}_2^{\mathrm{lc}},\cdots,\boldsymbol{K}_P^{\mathrm{lc}}\}$ 计算局部感知注意掩码：

$$\boldsymbol{M}^{\mathrm{lc}}=\sigma\left(\mathrm{Concat}\left(\boldsymbol{Q}\left(\boldsymbol{K}_1^{\mathrm{lc}}\right)^{\mathrm{T}},\boldsymbol{Q}\left(\boldsymbol{K}_2^{\mathrm{lc}}\right)^{\mathrm{T}},\cdots,\boldsymbol{Q}\left(\boldsymbol{K}_P^{\mathrm{lc}}\right)^{\mathrm{T}}\right)\boldsymbol{W}^{\mathrm{ma}}+\boldsymbol{b}^{\mathrm{ma}}\right) \tag{4}$$

式中：$\mathrm{Concat}(\cdot)$ 表示拼接操作符，$\boldsymbol{W}^{\mathrm{ma}}\in\mathbb{R}^{J\times S}$ 和 $\boldsymbol{b}^{\mathrm{ma}}\in\mathbb{R}^{S}$ 均为可训练的参数，$\boldsymbol{M}^{\mathrm{lc}}\in\mathbb{R}^{S\times S}$ 表示局部感知注意掩码。

5) 利用 $\boldsymbol{M}^{\mathrm{gb}}$ 和 $\boldsymbol{M}^{\mathrm{lc}}$ 分别提取 $\boldsymbol{V}$ 的全局感知场景特征和局部感知场景特征：

$$\boldsymbol{F}^{\mathrm{gb}}=\boldsymbol{M}^{\mathrm{gb}}\boldsymbol{V}\boldsymbol{W}^{\mathrm{gb}}+\boldsymbol{b}^{\mathrm{gb}} \tag{5}$$

$$\boldsymbol{F}^{\mathrm{lc}}=\boldsymbol{M}^{\mathrm{lc}}\boldsymbol{V}\boldsymbol{W}^{\mathrm{lc}}+\boldsymbol{b}^{\mathrm{lc}} \tag{6}$$

式中：$\boldsymbol{W}^{\mathrm{gb}},\boldsymbol{W}^{\mathrm{lc}}\in\mathbb{R}^{D\times D/2}$ 和 $\boldsymbol{b}^{\mathrm{gb}},\boldsymbol{b}^{\mathrm{lc}}\in\mathbb{R}^{D/2}$ 均为可训练参数；$\boldsymbol{F}^{\mathrm{gb}}\in\mathbb{R}^{S\times D/2}$ 和 $\boldsymbol{F}^{\mathrm{lc}}\in\mathbb{R}^{S\times D/2}$ 分别表示全局感知场景特征和局部感知场景特征。将这两类特征进行拼接，最终得到动作关联场景特征，记作 $\boldsymbol{F}^{\mathrm{ms}}\in\mathbb{R}^{S\times D}$。算法 1 对上述场景信息捕获过程进行总结。

算法 1 场景信息捕获

输入 场景特征 $\boldsymbol{F}^{\mathrm{sc}}$ 和动作特征 $\boldsymbol{F}^{\mathrm{mo}}$；身体支干数量 P。

输出 动作关联场景特征 $\boldsymbol{F}^{\mathrm{ms}}$。

1) 将 $\boldsymbol{F}^{\mathrm{sc}}$ 和 $\boldsymbol{F}^{\mathrm{mo}}$ 分别输入线性映射层，前者计算得到值项 $\boldsymbol{V}$ 和查询项 $\boldsymbol{Q}$，后者计算得到整体感知关键项 $\boldsymbol{K}^{\mathrm{gb}}$；

2) 根据式 (1)~(3) 计算全局感知注意掩膜 $\boldsymbol{M}^{\mathrm{gb}}$；

3) 将 $\boldsymbol{F}^{\mathrm{mo}}$ 分解为 P 个不同支干的时−空动态特征并依次输入线性映射层得到局部感知关键项 $\{\boldsymbol{K}_1^{\mathrm{lc}},\boldsymbol{K}_2^{\mathrm{lc}},\cdots,\boldsymbol{K}_P^{\mathrm{lc}}\}$；

4) 基于式 (4)，利用 $\boldsymbol{Q}$ 和 $\{\boldsymbol{K}_1^{\mathrm{lc}},\boldsymbol{K}_2^{\mathrm{lc}},\cdots,\boldsymbol{K}_P^{\mathrm{lc}}\}$ 计算局部感知注意掩码 $\boldsymbol{M}^{\mathrm{lc}}$；

5) 基于式 (5)~(6)，利用 $\boldsymbol{M}^{\mathrm{gb}}$ 和 $\boldsymbol{M}^{\mathrm{lc}}$ 分别提取 $\boldsymbol{V}$ 的全局感知场景特征和局部感知场景特征并进行融合，得到 $\boldsymbol{F}^{\mathrm{ms}}$。

场景信息捕获单元利用人体动作的全局特性 (整体运动模式) 和局部特性 (不同部位细节变化) 捕捉场景图像中与人体动作存在影响关联的场景信息。该信息能够反映人体动作在特定场景中所执行的任务内容，有助于准确理解人类行为意图。

2.2.2 场景适应度增强单元

场景适应度增强单元基于 $\boldsymbol{F}^{\mathrm{ms}}$ 探索 $\boldsymbol{F}^{\mathrm{mo}}$ 中适应场景信息影响的关节点空间表征，用于增强 $\boldsymbol{F}^{\mathrm{mo}}$ 的场景适应度。图 3 给出了场景适应度增强单元框架。

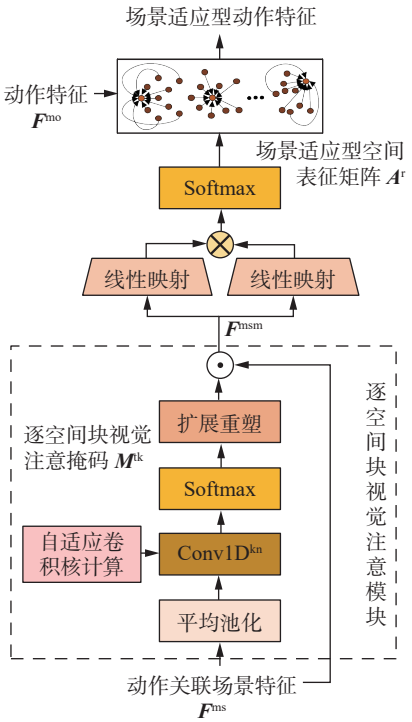

图 3 场景适应度增强单元框架

Fig. 3 Framework of scene correlation enhancement unit

1) 设计逐空间块视觉注意模块，捕获 $\boldsymbol{F}^{\mathrm{ms}}$ 中跨空间块的非线性交互关系并以此增强场景上下文信息。首先，对 $\boldsymbol{F}^{\mathrm{ms}}$ 中的所有空间块进行平均池化，以获得空间块特征，用 $\boldsymbol{F}^{\mathrm{tk}}\in\mathbb{R}^{S}$ 表示。然后，使用一维卷积层以及 Sigmoid 函数计算逐空间块视觉注意掩码：

$$\boldsymbol{M}^{\mathrm{tk}}=\sigma\left(\mathrm{Conv1D}^{\mathrm{kn}}\left(\boldsymbol{F}^{\mathrm{tk}}\right)\right)$$

式中：$\mathrm{Conv1D}^{\mathrm{kn}}(\cdot)$表示一维可变核卷积层，其卷积核尺寸$k$利用自适应卷积核计算方法[36]得到

$$k=\left|\frac{\log_2 S}{\delta}+\frac{b}{\delta}\right|_{\mathrm{odd}}$$

式中：$|\cdot|_{\mathrm{odd}}$表示获取最邻近奇数，δ和b是超参数。k能够适应场景特征空间分布，从而有效学习跨空间块的交互关系。基于上述过程，得到逐空间块视觉注意掩码，记为$\boldsymbol{M}^{\mathrm{tk}}\in\mathbb{R}^{1\times S}$。接着，计算跨空间块注意的全局−局部动作感知场景特征：

$$\boldsymbol{F}^{\mathrm{msm}}=\boldsymbol{F}^{\mathrm{ms}}\odot\mathrm{Transform}(\boldsymbol{M}^{\mathrm{tk}})$$

式中："⊙"为逐元素相乘操作符；$\mathrm{Transform}(\cdot)$表示扩展重塑操作符，用于调整$\boldsymbol{M}^{\mathrm{tk}}$的尺寸，使其与$\boldsymbol{F}^{\mathrm{ms}}$的尺寸相同；$\boldsymbol{F}^{\mathrm{msm}}\in\mathbb{R}^{S\times D}$表示特征计算结果。

2) 利用$\boldsymbol{F}^{msm}$计算场景适应型空间表征矩阵：

$$\boldsymbol{A}^{\mathrm{r}}=\sigma\left(\boldsymbol{B}_1^{\mathrm{T}}\boldsymbol{B}_2\right)\tag{7}$$

式中：$\boldsymbol{B}_1,\boldsymbol{B}_2\in\mathbb{R}^{S\times J}$均为$\boldsymbol{F}^{\mathrm{msm}}$输入线性映射模块得到的映射结果；$\boldsymbol{A}^{\mathrm{r}}\in\mathbb{R}^{J\times J}$表示计算得到的场景适应型空间表征矩阵。

3) 利用$\boldsymbol{A}^{\mathrm{r}}$对$\boldsymbol{F}^{\mathrm{mo}}$进行更新：

$$\boldsymbol{F}^{\mathrm{mou}}=\mathrm{Sig}(\boldsymbol{A}\odot\boldsymbol{A}^{\mathrm{r}}\boldsymbol{F}^{\mathrm{mo}}\boldsymbol{W}^{\mathrm{r}})$$

式中：$\boldsymbol{A}\in\{0,1\}^{J\times J}$是一个邻接矩阵，其中第$i$个和第$j$个关节点相连时$\boldsymbol{A}(i,j)=1$；$\boldsymbol{W}^{\mathrm{r}}\in\mathbb{R}^{D^{\mathrm{mo}}\times D}$表示可训练参数；$\mathrm{Sig}(\cdot)$表示 Sigmoid 函数；$\boldsymbol{F}^{\mathrm{mou}}\in\mathbb{R}^{S\times J}$表示得到的场景适应型动作特征，用于下一次的动作特征更新过程。

场景适应度增强单元利用动作关联场景特征探索场景信息影响下的关节点空间变化，然后计算场景适应型空间表征矩阵更新动作特征，使其更好地适应场景中存在影响关联的信息，提升场景适应能力，从而实现对复杂场景变化的自然响应，保障后续动作预测的准确性。

循环执行上述两类单元，最终完成基于人景互影响的动作特征更新，得到场景适应型动作特征，记为$\boldsymbol{F}^{\mathrm{sam}}\in\mathbb{R}^{J\times D}$。循环次数用$N^{\mathrm{r}}$表示。值得注意的是，上述两类单元主要通过线性映射计算注意力掩膜和表征矩阵。这是因为线性映射能有效保留输入信息，降低计算复杂度，保持模型可解释性，并避免非线性映射可能带来的复杂性和训练问题。

2.3　动作特征驱动噪声逆扩散

为了预测能够适应真实场景的人体动作，本文引入了扩散模型。这类模型受热力学第二定律的启发，学习从随机噪声到数据分布的逆向扩散过程，从而生成高质量、多样化的样本[31]。本文以$\boldsymbol{F}^{\mathrm{sam}}$作为驱动条件，引导扩散模型生成未来人体动作。

为了执行噪声逆扩散过程，本文设计了一类噪声预测网络，如图 4 所示。对于扩散时刻t的随机噪声样本（记作$\boldsymbol{Z}_t\in\mathbb{R}^{M\times 3J}$），噪声预测网络首先利用样本映射层将$\boldsymbol{Z}_t$映射至去噪域。$\boldsymbol{M}$表示待预测人体动作的时间维度。映射过程用公式表示为

$$\boldsymbol{Z}_t^{\mathrm{map}}=\mathrm{Add}(\boldsymbol{Z}_t,\mathrm{Flat}(\boldsymbol{x}_N))\boldsymbol{W}^{z}+\boldsymbol{b}^{z}$$

式中：$\boldsymbol{x}_N$表示已观测人体动作末尾帧；$\mathrm{Flat}(\boldsymbol{x}_N)$将$\boldsymbol{x}_N$的维度$(J,3)$调整为$(1,3J)$；$\mathrm{Add}(\boldsymbol{Z}_t,\mathrm{Flat}(\boldsymbol{x}_N))$将调整后的$\boldsymbol{x}_N$添加到$\boldsymbol{Z}_t$的每一行中；$\boldsymbol{W}^{z}\in\mathbb{R}^{3J\times D^{z}}$和$\boldsymbol{b}^{z}\in\mathbb{R}^{1\times D^{z}}$均为可训练参数；$\boldsymbol{Z}_t^{\mathrm{map}}\in\mathbb{R}^{M\times D^{z}}$表示计算得到的映射样本。然后，堆叠$N^{\mathrm{b}}$个噪声解析模块逐步提取$\boldsymbol{Z}_t^{\mathrm{map}}$的特征，同时使用残差连接稳定特征提取过程。

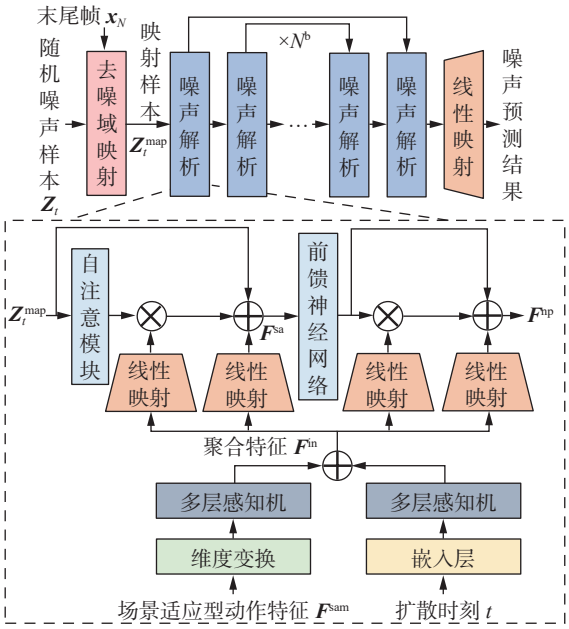

图 4　噪声预测网络框架

Fig. 4　Framework of noise prediction network

第一个噪声解析模块计算过程如下。

1) 对$\boldsymbol{F}^{\mathrm{sam}}$和$t$进行信息聚合：

$$\boldsymbol{F}^{\mathrm{in}}=\mathrm{MLP}(\mathrm{Flat}(\boldsymbol{F}^{\mathrm{sam}}))+\mathrm{MLP}(\psi(t))$$

式中：$\mathrm{MLP}(\cdot)$表示多层感知机；$\psi(t)$表示扩散时间嵌入层；$\mathrm{Flat}(\boldsymbol{F}^{\mathrm{sam}})$将$\boldsymbol{F}^{\mathrm{sam}}$的维度$(J,D)$变为$(1,JD)$；$\boldsymbol{F}^{\mathrm{in}}\in\mathbb{R}^{JD}$表示聚合特征。

2) 在$\boldsymbol{F}^{\mathrm{in}}$的引导下提取$\boldsymbol{Z}_t^{\mathrm{map}}$的自注意特征：

$$\boldsymbol{F}^{\mathrm{sa}}=\mathrm{SA}\left(\boldsymbol{Z}_t^{\mathrm{map}}\right)\phi_w\left(\boldsymbol{F}^{\mathrm{in}}\right)+\phi_b\left(\boldsymbol{F}^{\mathrm{in}}\right)+\boldsymbol{Z}_t^{\mathrm{map}}$$

式中：$\mathrm{SA}(\cdot)$表示自注意模块；$\phi_w(\cdot)$和$\phi_b(\cdot)$均表示输出维度为D^{z}的线性映射层；$\boldsymbol{F}^{\mathrm{sa}}\in\mathbb{R}^{M\times D^{z}}$表示提取的自注意特征。

3) 在$\boldsymbol{F}^{\mathrm{in}}$的引导下，使用前馈神经网络提取$\boldsymbol{F}^{\mathrm{sa}}$的特征：

$$\boldsymbol{F}^{\mathrm{np}}=\mathrm{FFN}(\boldsymbol{F}^{\mathrm{sa}})\phi_w\left(\boldsymbol{F}^{\mathrm{in}}\right)+\phi_b\left(\boldsymbol{F}^{\mathrm{in}}\right)+\boldsymbol{F}^{\mathrm{sa}}$$

式中：FFN(·)表示输出维度为D^{F}的前馈神经网络；$\boldsymbol{F}^{\mathrm{np}} \in \mathbb{R}^{M\times D^{\mathrm{F}}}$表示第一个噪声解析模块的输出特征。经过$N^b$个噪声解析模块，最终得到$\boldsymbol{Z}_t^{\mathrm{map}}$的特征，并将该特征输入线性映射层得到噪声预测结果。

得到噪声预测网络后，逐步执行从扩散时刻T到1的噪声预测及去噪过程，公式表示为

$$\boldsymbol{Z}_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{Z}_t-\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\mu(\boldsymbol{Z}_t,t,\boldsymbol{F}^{\mathrm{sam}},\boldsymbol{x}_N)\right)+\sqrt{\beta_t}\boldsymbol{\varepsilon}$$

式中：$\beta_t\in[0,1]$是预定义的方差参数；$\alpha_t=1-\beta_t$并且$\bar{\alpha}_t=\prod_{i=1}^{t}\alpha_i$；当$t=1$时，$\boldsymbol{\varepsilon}\sim N(0,\mathbf{I})$，否则$\boldsymbol{\varepsilon}=0$；$\boldsymbol{Z}_T\in\mathbb{R}^{M\times 3J}$是随机加噪样本，服从$N(0,\mathbf{I})$分布。基于上述去噪过程，最终得到$\boldsymbol{Z}_0\in\mathbb{R}^{M\times 3J}$。该结果即为未来动作预测结果。

2.4 训练

VPHSI通过前向扩散进行训练。首先，生成从扩散时刻T到1的带噪声未来人体动作：

$$\boldsymbol{Z}_t=\sqrt{\bar{\alpha}_t}\boldsymbol{Z}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{e}$$

式中：$\boldsymbol{Z}_0$是未来动作真值，$\boldsymbol{e}$为噪声。然后，噪声预测网络在每个扩散时刻预测$\boldsymbol{Z}_t$的噪声，并计算预测的噪声与$\boldsymbol{e}$的差值作为训练损失L：

$$L=\mathrm{E}_{e,t}\left[\|\boldsymbol{e}-\mu(\boldsymbol{Z}_t,t,\boldsymbol{F}^{\mathrm{sam}},\boldsymbol{x}_N)\|^2\right]$$

通过最小化L，最终得到最优的模型参数。

3 实验过程及结果分析

3.1 数据集及对比方法介绍

为了全面评估VPHSI的有效性，本文使用GTA-IM[27]、PROX[37]和BEHAVE[38]3个已公开数据集进行实验。

GTA-IM是一个合成的人景交互数据集，其包含100万个RGB-D视频以及三维人体动作骨架数据。该数据库涉及50个人类角色以及7个不同的场景。人体关节点数量为98。视频帧率为30 f/s。人类角色在不同场景中展示多类型的人体动作。每个场景有一个或多个楼层以及不同类型的房间，如客厅、卧室等。本文使用来自4个场景(r001、r002、r003、r006)的数据进行训练，并使用来自剩余3个场景(r010、r011、r013)的数据进行测试。同时，参考文献[28]，从98个关节点中选取21个主要关节点，即$J=21$。观测时间和预测时间分别设置为1 s(即$N=30$)和2 s(即$M=30$)。

PROX是一个大型人景交互数据集，提供了来自12个不同场景共计100 000 f的RGB-D场景图像帧和三维人体动作骨架数据。帧率为30 f/s。涉及的人类角色包括4名女性和16名男性。参考文献[28]，采用来自8个场景(即N3Library、MPH112、MPH11、MPH8、BSB、N0Sofa、N3Office和Werkraum)的数据用于训练，来自剩余4个场景(即MPH16、MPH1Library、N0SittingBooth和N3OpenArea)的数据用于测试。参考文献[28]，使用SMPL-X模型的22个主要关节(即$J=22$)。观测时间和预测时间与GTA-IM相同。

BEHAVE是一个用于研究人−物体交互的大型数据库，包含由4台Kinect RGB-D相机采集的多视角RGB-D帧以及相应的人体骨架数据和物体标记数据。帧率为30 f/s。该数据集涉及17种交互类型，包含8名参与者和20种物体。人体骨架数据包含67个关节点，本文选取其中25个主要关节点进行实验。参考官方文件，数据集分为训练集和测试集，分别包含231和90个样本。为了验证本文方法在提取场景信息方面的有效性，后续实验仅利用RGB帧和对应的人体骨架数据进行人体动作预测，未使用物体的标注数据。

本文选取7类基准方法用于对比实验，分别是：SCAFF[19]、LTD(learning trajectory dependency-based model)[18]、SIMLPE(simple yet effective MLP network)[12]、C-RNN(context-aware recurrent neural network)[25]、CAMP(contact-aware motion prediction model)[28]、STAG(staged contact-aware global human motion forecasting model)[29]和MCLD(multi-condition latent diffusion network)[31]。SCAFF、LTD和SIMLPE仅提取动作特征预测后续人体动作变化，没有考虑场景信息对动作预测的影响。C-RNN、CAMP、STAG和MCLD捕获与人体有接触关联或存在视觉反馈的场景信息来引导人体动作预测。

3.2 参数设置和评价标准

本文实验在Windows 10操作系统和Python 3.8环境下进行，采用PyTorch框架在4个NVIDIA RTX3090Ti GPU上进行模型训练和测试。对于VPHSI，在双模态特征提取阶段，使用STGCN[39]和ViT[40]分别作为图像编码器和动作编码器。在3类数据集上执行动作重建任务对STGCN进行预训练；在ImageNet-21k上执行图像分类任务对ViT进行预训练。D^{sc}和D^{mo}分别设置为512和256。在基于人景互影响的动作特征更新阶段，循环次数N^{r}设置为3，D设置为512。P设置为5，对应于5个身体支干：左臂、右臂、左腿、右腿和躯干。δ和b分别设置为2和1。在动作特征驱动噪声逆扩散阶段，T和N^{b}分别设置为1 000和8。

β_t 通过线性插值获得, 其中 $\beta_1 = 0.0001$ 且 $\beta_T = 0.05$。在噪声解析模块中, D^z 设置为 512。多层感知机包含 3 层网络层, 其输出维度分别为 128、256 和 512。在训练阶段, 批量尺寸设置为 64, 初始学习率设置为 0.001, 同时采用多步学习率调度器 (γ=0.9)。采用 Adam 优化器, 同时设置迭代次数为 500。

参考文献 [28], 本文采用使用平均关节点位置误差 (mean per joint position error, MPJPE) 来评价人体动作预测性能。

3.3　实验结果与分析

3.3.1　对比分析

表 1 ~ 3 给出了不同模型在 GTA-IM、PROX 和 BEHAVE 上获得的 MPJPE 结果。可以看出, SCAFF、LTD 和 SIMLPE 由于未考虑场景信息的影响, 性能相对较差; 而 C-RNN、CAMP、STAG 和 MCLD 加入了场景信息捕获来引导动作预测, 因此取得了较好的结果。不同于上述模型, VPHSI 充分探索了场景图像中与人体动作存在影响关联的场景信息, 有效模拟了视觉感知下人体动作与场景信息的互影响过程, 从而获得了最好的预测结果。并且, 表 1 还对比了不同模型的参数量。结果表明, VPHSI 在保持较少参数量的情况下得到了最优的预测结果, 展现了其优秀的计算效率和模型性能。这一优势在资源受限的实际应用场景中尤为突出。

表 1　各类方法在 GTA-IM 上的 MPJPE 和参数量对比

Table 1　Comparison of MPJPE and parameters of various models on GTA-IM

方法	时刻/s							参数量/×10^6
	0.3	0.5	0.7	1.0	1.5	1.7	2.0	
SCAFF[19]	49.3	67.9	73.9	94.2	98.2	101.2	102.9	7.5
LTD[18]	44.0	52.5	60.2	69.1	85.6	89.2	95.3	5.2
SIMLPE[12]	44.2	51.9	60.0	69.5	85.5	88.8	95.1	0.2
C-RNN[25]	43.4	51.1	59.5	68.9	79.1	84.2	88.8	9.4
CAMP[28]	42.5	50.8	59.2	67.5	75.5	82.3	86.9	9.1
STAG[29]	42.0	48.5	58.5	65.9	75.8	83.9	88.2	6.5
MCLD[31]	39.2	45.2	56.9	66.4	73.9	81.4	86.9	4.1
VPHSI	**38.2**	**43.1**	**56.3**	**65.2**	**72.0**	**78.2**	**82.3**	**3.8**

注: 加粗表示本列最优结果。

表 2　各类方法在 PROX 上的 MPJPE 对比

Table 2　Comparison of MPJPE of various models on PROX

方法	时刻/s						
	0.3	0.5	0.7	1.0	1.5	1.7	2.0
SCAFF[19]	93.9	122.0	158.3	171.6	283.5	310.3	352.7
LTD[18]	71.0	91.0	122.2	141.3	171.8	179.2	187.8
SIMLPE[12]	68.0	90.3	119.2	138.7	168.4	174.2	181.3
C-RNN[25]	65.9	89.5	109.4	133.5	158.4	163.8	177.3
CAMP[28]	62.0	89.9	104.7	127.5	149.3	152.4	167.5
STAG[29]	61.1	89.7	104.7	127.3	149.9	153.0	168.9
MCLD[31]	59.3	83.9	101.3	123.1	148.3	152.3	167.3
VPHSI	**57.2**	**82.4**	**99.3**	**121.4**	**144.8**	**149.3**	**164.9**

注: 加粗表示本列最优结果。

表 3　各类方法在 BEHAVE 上的 MPJPE 对比结果

Table 3　Comparison results of MPJPE of various models on BEHAVE

方法	时刻/s					
	0.5	0.7	1.0	1.5	1.7	2.0
SCAFF[19]	63.0	117.3	147.6	193.1	211.3	222.3
LTD[18]	49.2	107.8	132.8	178.3	183.2	191.4
SIMLPE[12]	49.3	105.1	130.9	175.9	183.0	189.5
C-RNN[25]	47.2	102.3	130.5	172.4	175.8	182.0

续表 3

方法	时刻/s					
	0.5	0.7	1.0	1.5	1.7	2.0
CAMP[28]	45.2	102.7	125.3	168.3	170.7	177.1
STAG[29]	44.7	101.4	123.0	165.2	168.9	171.9
MCLD[31]	42.8	99.3	120.1	158.2	163.0	166.8
VPHSI	**41.4**	**98.1**	**119.4**	**154.3**	**162.0**	**163.6**

注：加粗表示本列最优结果。

此外，图 5 给出了不同模型的可视化结果。可以看出，SCAFF 表现最差，甚至得到不合理的预测结果，例如在第 2 个场景中不上楼梯或在第 3 个场景中不坐在椅子上。CAMP 表现较好，但对于某些部位的关节点，如第 4 个场景中的右手腕部分，预测结果仍不准确。与上述模型相比，VPHSI 预测的未来动作最符合真实场景。例如，在第 2 个场景中，VPHSI 预测的右臂抬起幅度最接近真值；在第 3 个场景中，VPHSI 预测的手臂伸展和腿部弯曲最为自然。

值得说明的是，VPHSI 基于人景互影响关系动态更新动作特征，使得动作特征重点关注人体动作与场景之间的交互，弱化对场景其他信息的关注，从而有效降低外部干扰。为了验证这一点，本文比较了各个方法在 PROX 的 4 类不同的测试场景中的 2 s 动作预测结果 (如表 4 所示)。这 4 类场景在空间结构、物品属性、光照等方面均有较大差异。结果表明，VPHSI 在这 4 类场景中均取得了最好的预测结果，验证了该方法在面对复杂场景、光照等外部影响下仍具备较强的鲁棒性。

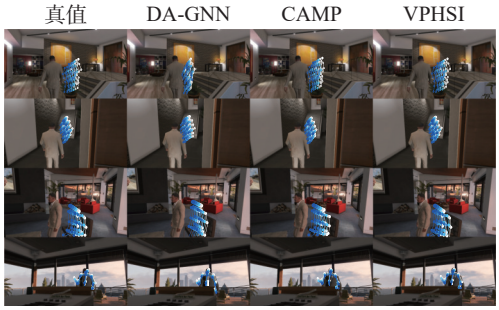

图 5　可视化结果对比

Fig. 5　Comparison of visualization results

表 4　各类模型在 PROX 上不同场景的 MPJPE 对比结果

Table 4　MPJPE comparison results of various models in different scenarios on PROX

模型	场景			
	MPH16	MPH1 Library	N0Sitting Booth	N3Open Area
SCAFF[19]	359.4	351.3	352.9	363.6
LTD[18]	189.0	187.5	191.5	200.8
SIMLPE[12]	161.3	180.9	185.2	193.3
C-RNN[25]	155.9	178.8	180.2	187.0
CAMP[28]	156.2	169.1	169.8	175.8
STAG[29]	154.9	171.1	172.5	175.9
MCLD[31]	155.8	168.9	170.9	175.2
VPHSI	**153.3**	**164.2**	**168.4**	**172.1**

注：加粗表示本列最优结果。

3.3.2　消融分析

本文在 GTA-IM 上对 VPHSI 的核心模块和参数进行了一系列消融研究。

1) 预训练编码器：为了探索不同预训练编码器对人体动作预测的影响，本实验增加了 4 个预训练模型：Res-sup(residual supervised model)[7]、UniGC(universal graph convolution model)[20]、ResNet50(residual network with 50 layers)[41] 和 EVA-CLIP(series of models of improving efficiency and effectiveness of CLIP training)[42]。Res-sup 和 UniGC 作为动作编码器，分别基于 RNNs 和 GCNs 设计；而 ResNet50 和 EVA-CLIP 则作为图像编码器，分别基于 CNNs 和 Transformer。表 5 给出了两类编码器的不同组合对应的 MPJPE 结果。可以看出，UniGC+EVA-CLIP 和 STGCN+ViT 的预测结果明显优于其他组合。这一结果表明，相较于 RNNs，GCNs 在学习关节点时−空变化特性方面具有更高的精确度。并且，Transformer 利用其自注意力

机制，能够更充分地关注图像中与动作相关的场景信息，因此相较于 CNNs 更具优势。值得注意的是，UniGC+EVA-CLIP 和 STGCN+ViT 的预测结果相差不大，这表明基于 GCNs 和 Transformer 设计的编码器在提取动作特征和场景特征方面的能力相近。

表 5　不同预训练编码器组合在 GTA-IM 上的 MPJPE 对比结果
Table 5　MPJPE comparison results of different combinations of pretrained encoders on GTA-IM

方法	时刻/s									平均
	0.1	0.3	0.5	0.7	1.0	1.3	1.5	1.7	2.0	
Res-sup+ResNet50	31.4	47.8	65.5	71.0	77.4	85.2	89.2	95.9	100.3	73.7
Res-sup+ViT	29.4	44.5	61.4	68.7	73.5	80.3	85.1	88.8	94.1	69.5
STGCN+ResNet50	30.1	46.1	63.1	66.3	75.2	79.9	85.5	91.4	96.9	70.5
UniGC+EVA-CLIP	25.6	**38.2**	45.1	56.5	**65.0**	70.3	72.1	**77.9**	**81.0**	60.8
STGCN+ViT	**25.3**	**38.2**	**43.1**	**56.3**	65.2	**69.9**	**72.0**	78.2	82.3	**59.5**

注：加粗表示本列最优结果。

2) 场景信息捕获单元中的P和D：考虑到P和D的值对场景信息捕获单元的性能影响较大，本文分别对P和D进行消融研究。

首先，设置不同的P值，对应不同的支干数量。表 6 给出了不同身体支干的图解描述及其相应的预测结果。可以发现，当P值较低，即为 2 和 3 时，相应的 MPJPE 较高。这可能是因为较小的P对应较少的身体支干数量，导致局部感知注意掩膜的空间多样性下降。并且，当P值从 5 增加到 9 时，MPJPE 的变化不大，说明过多的身体支干并不会明显提升掩膜的空间多样性。然而，较大的P会增加计算成本。综合考虑，P为 5 是最佳选择。

表 6　不同 P 值在 GTA-IM 上的 MPJPE 对比结果
Table 6　MPJPE comparison results of different values of P on GTA-IM

P	身体部位	图解	时刻/s							
			0.1	0.3	0.5	0.7	1.0	1.3	1.5	2.0
2	躯干、四肢		28.3	41.9	52.3	60.3	69.9	74.0	75.3	86.1
3	躯干、手臂、腿		27.4	40.3	49.8	58.5	68.3	73.3	73.5	85.0
5	躯干、左臂、右臂、左腿、右腿		**25.3**	38.2	**43.1**	**56.3**	**65.2**	**69.9**	72.0	**82.3**
9	躯干、左上臂、右上臂、左前臂、右前臂、左大腿、右大腿、左小腿、右小腿		25.5	**38.0**	**43.1**	56.5	**65.2**	**69.9**	**71.8**	**82.3**

注：加粗表示本列最优结果。

然后，本文比较了场景信息捕获单元中不同D对应的 MPJPE(见图 6)。可以看出，当D为 512 时，VPHSI 的性能最佳。

3) 逐空间块视觉注意模块：该模块旨在提升场景适应度增强单元对$\boldsymbol{F}^{\mathrm{ms}}$中场景上下文信息的理解能力，这对后续计算场景适应型空间表征矩

阵有很大影响。因此，本文首先对该模块进行消融实验以验证其有效性。如表 7 所示，带有逐空间块视觉注意模块的单元取得了较低的 MPJPE，这说明该模块能有效增强单元对场景上下文信息的理解能力。接着，本文分析了该模块的两个核心参数：δ和b，对人体动作预测的影响。表 7 给出了δ和b不同的值组合对应的 MPJPE。可以发现，当δ为 2 且b为 1 时，MPJPE 最低，意味着这种组合可以实现最有效的信息交互。

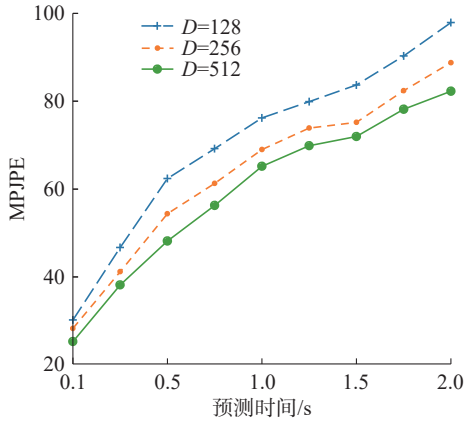

图 6　不同 D 值在 GTA-IM 上的 MPJPE 对比结果

Fig. 6　MPJPE comparison results of different values of D on GTA-IM

表 7　逐空间块视觉注意模块在 GTA-IM 上的消融结果

Table 7　Ablation results of the token-wise attention module on GTA-IM

消融设置	时刻/s				
	0.1	0.5	1.0	1.5	2.0
×	27.5	48.8	67.9	73.9	85.1
√	**25.3**	**43.1**	**65.2**	**72.0**	**82.3**
$\delta=1,b=1$	26.4	51.0	66.9	75.0	86.3
$\delta=1,b=2$	26.3	51.2	66.9	74.8	86.1
$\delta=2,b=0$	26.3	50.9	66.3	74.4	85.8
$\delta=2,b=1$	**25.3**	**43.1**	**65.2**	**72.0**	**82.3**

注：加粗表示本列最优结果。

4) 动作特征驱动噪声逆扩散：为了验证动作特征驱动噪声逆扩散对于人体动作预测的有效性，本文引入另外两个预测模型：RNNs 和 GCN-TCN(graph-temporal convolution network)[43]。比较结果如表 8 所示。可以发现，扩散模型的表现优于上述两类预测模型。这一结果表明，相比于传统神经网络，扩散模型通过多阶段采样和逐步精化的生成机制，更加有效地捕捉未来动作的复杂动态特性，从而生成更加真实、连贯的人体动作序列。

表 8　不同预测器在 GTA-IM 上的 MPJPE 对比结果

Table 8　MPJPE comparison results of different predictors on GTA-IM

预测器	时刻/s									平均
	0.1	0.3	0.5	0.7	1.0	1.3	1.5	1.7	2.0	
RNNs	34.6	51.6	69.5	76.4	83.1	90.4	96.2	101.6	105.3	78.7
GCN+TCN	30.5	40.8	56.3	60.3	77.9	80.5	83.9	89.2	95.1	68.3
扩散模型	**25.3**	**38.2**	**43.1**	**56.3**	**65.2**	**69.9**	**72.0**	**78.2**	**82.3**	**59.5**

注：加粗表示本列最优结果。

考虑到T对扩散模型影响重大，本实验设置了不同的T值并计算相应的 MPJPE，如表 9 所示。可以看到，当T为 1 000 时，MPJPE 最小，对应的 VPHSI 预测性能最好。$T=10$时得到的 MPJPE 最大，预测性能显著下降，这是因为较低的T值会降低动作生成结果的真实性。

此外，为了验证N^{b}对噪声预测网络性能的影响，本文比较了不同N^{b}值的 MPJPE，如表 10 所示。结果表明，当N^{b}为 8 时，MPJPE 最低，意味着堆叠 8 个噪声解析模块能够得到最好的预测效果。值得说明的是，当N^{b}超过 8 时，MPJPE 反而增加。这可能是因为过多的噪声解析模块增加了噪声预测网络的参数数量，导致训练过程中出现过拟合，使得训练性能下降。

表 9　不同T值在 GTA-IM 上的 MPJPE 对比结果

Table 9　MPJPE comparison results of different values of T on GTA-IM

T	时刻/s					
	0.5	0.7	1.0	1.3	1.5	2.0
10	62.5	77.3	84.2	92.4	98.3	109.4
100	55.2	63.6	72.3	77.3	82.9	95.2
1 000	**43.1**	**53.2**	**65.2**	**69.4**	**72.0**	**82.3**

注：加粗表示本列最优结果。

表 10　不同N^b值在 GTA-IM 上的 MPJPE 对比结果
Table 10　MPJPE comparison results of different values of N^b on GTA-IM

N^b	时刻/s					
	0.5	0.7	1.0	1.3	1.5	2.0
2	63.2	66.0	74.3	77.9	82.5	94.3
4	58.6	64.2	72.3	74.0	79.1	91.4
8	**43.1**	**53.2**	**65.2**	**69.4**	**72.0**	**82.3**
10	50.0	56.2	66.3	69.9	72.9	84.1

注: 加粗表示本列最优结果。

4　结束语

本文提出了一种基于视觉感知人景互影响的人体动作预测算法 VPHSI。对于输入的人体动作和场景图像, 该算法首先提取两类输入数据的特征, 即动作特征和场景特征, 然后基于这两类特征循环执行场景信息捕获单元和场景适应度增强单元。前者利用动作特征的全局−局部动作特性捕获场景特征中存在人体行为影响关联的场景信息; 而后者利用捕获的场景信息更新动作特征的空间表征, 以提升其场景适应度。通过循环执行上述两类单元, 可以有效实现基于人景互影响的动作特征更新, 得到场景适应型动作特征。接着, 将该特征作为驱动条件执行噪声逆扩散过程, 最终得到人体动作预测结果。实验结果验证了 VPHSI 在不同场景中优异的预测性能, 展现了其在多样复杂环境中的高精度和强鲁棒性。

值得说明的是, 本研究重点关注人与场景信息之间的互影响关系。特别地, 在涉及多个人物的复杂场景中, 除了人物与场景信息的交互外, 还需考虑人物间的社交互动。基于此, 未来计划在现有工作的基础上, 进一步探索不同人物间的交互模式, 并将其与已有的人景互影响机制相结合, 以实现对多人动作的精准预测。

参考文献:

[1] RENZ H, KRÄMER M, BERTRAM T. Comparing human motion forecasts in moving horizon trajectory planning of collaborative robots[C]//2023 IEEE International Conference on Robotics and Biomimetics. Koh Samui: IEEE, 2023: 1−6.

[2] LEE M L, LIU Wansong, BEHDAD S, et al. Robot-assisted disassembly sequence planning with real-time human motion prediction[J]. IEEE transactions on systems, man, and cybernetics: systems, 2023, 53(1): 438−450.

[3] ZHOU Xiaokang, LIANG Wei, WANG K I, et al. Deep-learning-enhanced human activity recognition for Internet of healthcare things[J]. IEEE internet of things journal, 2020, 7(7): 6429−6438.

[4] LI Qin, WANG Yong. Self-supervised pretraining based on noise-free motion reconstruction and semantic-aware contrastive learning for human motion prediction[J]. IEEE transactions on emerging topics in computational intelligence, 2024, 8(1): 738−751.

[5] URTASUN R, FLEET D J, LAWRENCE N D. Modeling human locomotion with topologically constrained latent variable models[C]//Workshop on Human Motion. Berlin: Springer Berlin Heidelberg, 2007: 104−118.

[6] LEHRMANN A M, GEHLER P V, NOWOZIN S. Efficient nonlinear markov models for human motion[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 1314−1321.

[7] MARTINEZ J, BLACK M J, ROMERO J. On human motion prediction using recurrent neural networks[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu: IEEE, 2017: 2891−2900.

[8] WOLTER M, YAO A. Complex gated recurrent neural networks[C]//32nd Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018: 1−11.

[9] 桑海峰, 陈紫珍, 何大阔. 基于双向 GRU 和注意力机制模型的人体动作预测[J]. 计算机辅助设计与图形学学报, 2019, 31(7): 1166−1174.
SANG Haifeng, CHEN Zizhen, HE Dakuo. Human motion prediction based on bidirectional-GRU and attention mechanism model[J]. Journal of computer-aided design & computer graphics, 2019, 31(7): 1166−1174.

[10] 王辉, 丁铂栩, 宋佳豪, 等. 基于 PointNet 和长短时记忆网络的三维人体动作预测[J]. 计算机应用, 2022, 42(S2): 60−66.
WANG Hui, DING Boxu, SONG Jiahao, et al. 3D human action prediction via PointNet and long short-term memory network[J]. Journal of computer applications, 2022, 42(S2): 60−66.

[11] WANG Hongsong, DONG Jian, CHENG Bin, et al. PVRED: a position-velocity recurrent encoder-decoder for human motion prediction[J]. IEEE transactions on image processing, 2021, 30: 6096−6106.

[12] GUO Wen, DU Yuming, SHEN Xi, et al. Back to MLP: a simple baseline for human motion prediction[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2023:

4809−4819.

[13] 张瑞鹏. 基于门控循环单元网络的人体动作预测方法研究[D]. 南京: 南京理工大学, 2021.
ZHANG Ruipeng. Research on human motion prediction method based on gated recurrent unit network[D]. Nanjing: Nanjing University of Science and Technology, 2021.

[14] LIU Xiaoli, YIN Jianqin, LIU Jin, et al. TrajectoryCNN: a new spatio-temporal feature learning network for human motion prediction[J]. IEEE transactions on circuits and systems for video technology, 2021, 31(6): 2133−2146.

[15] TANG Jin, ZHANG Jin, YIN Jianqin. Temporal consistency two-stream CNN for human motion prediction[J]. Neurocomputing, 2022, 468: 245−256.

[16] 张晋, 唐进, 尹建芹. 面向人体动作预测的对称残差网络[J]. 机器人, 2022, 44(3): 291−298.
ZHANG Jin, TANG Jin, YIN Jianqin. Symmetric residual network for human motion prediction[J]. Robot, 2022, 44(3): 291−298.

[17] 贺朵. 基于图卷积网络深度学习的人体动作识别与预测[D]. 西安: 西安理工大学, 2023.
HE Duo. Human action recognition and prediction based on deep learning of graph convolutional networks [D]. Xi'an: Xi'an University of Technology, 2023.

[18] MAO Wei, LIU Miaomiao, SALZMANN M, et al. Learning trajectory dependencies for human motion prediction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 9488−9496.

[19] LI Qin, WANG Yong, LYU Fanbing. Semantic correlation attention-based multiorder multiscale feature fusion network for human motion prediction[J]. IEEE transactions on cybernetics, 2024, 54(2): 825−838.

[20] WANG Xinshun, CUI Qiongjie, CHEN Chen, et al. GCNext: towards the unity of graph convolutions for human motion prediction[J]. Proceedings of the AAAI conference on artificial intelligence. 2024, 38(6): 5642−5650.

[21] 李沁. 基于三维骨架数据的人体动作预测及其应用研究[D]. 长沙: 中南大学, 2022.
LI Qin. Human motion prediction based on 3D skeleton data and its application[D]. Changsha: Central South University, 2022

[22] 代金利, 曹江涛, 姬晓飞. 交互关系超图卷积模型的双人交互行为识别[J]. 智能系统学报, 2024, 19(2): 316−324.
DAI Jinli, CAO Jiangtao, JI Xiaofei. Two-person interaction recognition based on the interactive relationship hypergraph convolution network model[J]. CAAI transactions on intelligent systems, 2024, 19(2): 316−324.

[23] 胡佳慧. 基于时空特征融合与动作序列补全的人体动作预测算法研究[D]. 长春: 吉林大学, 2024.
HU Jiahui. Research on human motion prediction algorithms based on spatiotemporal features fusion and motion sequence completion[D]. Changchun: Jilin University, 2024.

[24] BERMAN M G, JONIDES J, KAPLAN S. The cognitive benefits of interacting with nature[J]. Psychological science, 2008, 19(12): 1207−1212.

[25] CORONA E, PUMAROLA A, ALENYA G, et al. Context-aware human motion prediction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seattle: IEEE, 2020: 6992−7001.

[26] HASSAN M, CEYLAN D, VILLEGAS R, et al. Stochastic scene-aware motion prediction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 11374−11384.

[27] CAO Zhe, GAO Hang, MANGALAM K, et al. Long-term human motion prediction with scene context[C]// European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 387−404.

[28] MAO Wei, HARTLEY R I, SALZMANN M. Contact-aware human motion forecasting[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022: 7356−7367.

[29] SCOFANO L, SAMPIERI A, SCHIELE E, et al. Staged contact-aware global human motion forecasting[C]//The 34th British Machine Vision Conference. Aberdeen: BMVA Press, 2023: 589−594.

[30] XING Chaoyue, MAO Wei, LIU Miaomiao. Scene-aware human motion forecasting via mutual distance prediction[C]//Computer Vision–ECCV 2024. Cham: Springer, 2023: 128–144.

[31] GAO Xuehao, YANG Yang, WU Yang, et al. Multi-condition latent diffusion network for scene-aware neural human motion prediction[J]. IEEE transactions on image processing, 2024, 33: 3907−3920.

[32] LIU Zhenguang, LYU Kedi, WU Shuang, et al. Aggregated multi-GANs for controlled 3D human motion prediction[J]. Proceedings of the AAAI conference on artificial intelligence. 2021, 35(3): 2225–2232.

[33] ZHAO Mengyi, TANG Hao, XIE Pan, et al. Bidirectional Transformer GAN for long-term human motion prediction[J]. ACM transactions on multimedia computing, communications, and applications, 2023, 19(5): 1−19.

[34] BARQUERO G, ESCALERA S, PALMERO C. BeLFusion: latent diffusion for behavior-driven human motion

prediction[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 2317−2327.

[35] TIAN Sibo, ZHENG Minghui, LIANG Xiao. TransFusion: a practical and effective transformer-based diffusion model for 3D human motion prediction[J]. IEEE robotics and automation letters, 2024, 9(7): 6232−6239.

[36] WANG Qilong, WU Banggu, ZHU Pengfei, et al. ECA-net: efficient channel attention for deep convolutional neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11534–11542.

[37] HASSAN M, CHOUTAS V, TZIONAS D, et al. Resolving 3D human pose ambiguities with 3D scene constraints[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 2282–2292.

[38] BHATNAGAR B L, XIE Xianghui, PETROV I A, et al. BEHAVE: dataset and method for tracking human object interactions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 15914−15925.

[39] YAN Sijie, XIONG Yuanjun, LIN Dahua. Spatial temporal graph convolutional networks for skeleton-based action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 7444–7452.

[40] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]//International Conference on Learning Representations. NewOrleans: ICLR, 2021: 1−22.

[41] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770–778.

[42] SUN Quan, FANG Yuxin, WU L, et al. EVA-CLIP: improved training techniques for CLIP at scale[EB/OL]. (2023−03−27)[2024−11−15]. https://arxiv.org/abs/2303.15389.

[43] NARGUND A A, SRA M. SPOTR: spatio-temporal pose Transformers for human motion prediction[EB/OL]. (2023–03–11)[2024–11–15]. https://arxiv.org/abs/2303.06277.

作者简介：

李沁，讲师，博士，主要研究方向为人机交互和模式识别。E-mail：qinli@hutb.edu.cn。

陈飞扬，主要研究方向为计算机视觉和人机交互。E-mail：1689343195@qq.com。

刘利枚，教授，博士，主要研究方向为人工智能和智能决策。主持国家重点研发计划、国家社会科学基金等省部级以上项目 10 余项。发表学术论文 30 余篇，出版专著和教材 3 部。E-mail：seagullm@163.com。

[责任编辑：丁钰]