一种基于 Transformer 的三维人体姿态估计方法

王玉萍[1]，曾毅[1]，李胜辉[2]，张磊[3]

(1. 郑州科技学院信息工程学院，河南 郑州 450064；
2. 河南机电职业学院大数据学院，河南 郑州 450064；
3. 郑州大学信息工程学院，河南 郑州 450001)

摘　　要：三维人体姿态估计是人类行为理解的基础，但是预测出合理的三维人体姿态序列仍然是具有挑战性的问题。为了解决这个问题，提出一种基于 Transformer 的三维人体姿态估计方法，利用多层长短期记忆(LSTM)单元和多尺度 Transformer 结构增强人体姿态序列预测的准确性。首先，设计基于时间序列的生成器，通过 ResNet 预训练神经网络提取图像特征；其次，采用多层 LSTM 单元学习时间连续性的图像序列中人体姿态之间的关系，输出合理的 SMPL 人体参数模型序列；最后，构建基于多尺度 Transformer 的判别器，利用多尺度 Transformer 结构对多个分割粒度进行细节特征学习，尤其是 Transformer block 对相对位置进行编码增强局部特征学习能力。实验结果表明，该方法相对于 VIBE 方法具有更好地预测精度，在 3DPW 数据集上比 VIBE 的平均(每)关节位置误差(MPJPE)低了 7.5%；在 MP-INF-3DHP 数据集上比 VIBE 的 MPJPE 降低了 1.8%。

关　键　词：多尺度 Transformer 结构；LSTM 单元；时间序列；注意力机制；三维姿态估计

中图分类号：TP 391　　**DOI**：10.11996/JGj.2095-302X.2023010139

文献标识码：A　　**文 章 编 号**：2095-302X(2023)01-0139-07

A Transformer-based 3D human pose estimation method

WANG Yu-ping[1], ZENG Yi[1], LI Sheng-hui[2], ZHANG Lei[3]

(1. School of Information Engineering, Zhengzhou University of Science and Technology, Zhengzhou Henan 450064, China;
2. College of Big Data, Henan Electromechanical Vocational College, Zhengzhou Henan 450064, China;
3. School of Information Engineering, Zhengzhou University, Zhengzhou Henan 450001, China)

Abstract: 3D human pose estimation is the foundation of human behavior understanding, but predicting reasonable 3D human pose sequences remains a challenging problem. To solve this problem, a Transformer-based 3D human pose estimation method was proposed, utilizing a multi-layer long short-term memory (LSTM) unit and a multi-scale Transformer structure to enhance the accuracy of human pose sequence prediction. First, a generator based on time series was designed to extract image features through the ResNet pre-trained neural network. Secondly, multi-layer LSTM units were used to learn the relationship between human poses in temporally continuous image sequences, thereby outputting a reasonable skinned multi-person linear (SMPL) human parameter model sequence. Finally, a multi-scale Transformer-based discriminator was constructed, and the multi-scale Transformer structure was employed to learn detailed features for multiple segmentation granularities, especially the Transformer block encoding the relative position to enhance the local feature learning ability. Experimental results show that the proposed method could yield better prediction accuracy than the VIBE method, which is 7.5% lower than the average (per) joint position error (MPJPE) of VIBE on the 3DPW dataset, and 1.8% lower than VIBE′s MPJPE on the MP-INF-3DHP

收稿日期：2022-04-07；**定稿日期**：2022-07-19

Received：7 April，2022；**Finalized**：19 July，2022

基金项目：河南省科技厅科技攻关项目(222102210174)

Foundation items：Henan Provincial Department of Science and Technology Science and Technology Project (222102210174)

第一作者：王玉萍(1979–)，女，教授，硕士。主要研究方向为机器视觉、虚拟现实与机器学习。E-mail：wangyupingpaper@163.com

First author：WANG Yu-ping (1979–), professor, master. Her main research interests cover machine vision, virtual reality and machine learning. E-mail：wangyupingpaper@163.com

dataset.

Keywords: multi-scale Transformer structure; LSTM unit; time series; attention mechanism; 3D pose estimation

三维人体姿态估计主要的目标是从三维空间中估计出人体关键点的位置。3D 姿态估计的应用非常广泛，包括人机交互、智能交通、视频合成、医疗监测等，同时也可以作为其他计算机视觉任务的基础，提供三维空间中人体关键点位置信息。

为了高效地预测出三维人体姿态序列，提高对长距离粒度特征提取的能力，获得更多空间信息，本文采用先获取 2D 信息然后提升到三维空间中进行姿态估计的方法，提出了基于 Transformer 的三维人体姿态估计方法。设计了基于时间序列的生成器，通过 ResNet 预训练神经网络提取图像特征，随后采用多层长短期记忆网络(long short-term memory，LSTM)单元学习时间连续性的图像序列中人体姿态之间的关系；同时还提出了基于多尺度 Transformer 的判别器，利用多尺度 Transformer 结构对多个分割粒度进行细节特征学习，其中 Transformer block 对相对位置进行编码增强局部特征学习能力。通过不断生成判别过程输出合理的三维人体运动序列，提高了三维人体姿态估计的预测精度。

1 相关工作

为了预测出合理的三维人体姿态序列，最近的工作提出了许多新颖的方法用于提高模型的预测精度。这些方法大致分为 2 类：①从 2D 图片中直接回归 3D 坐标，如文献[1-2]；②先获取 2D 信息，然后再提升到三维空间中估计人体姿态，如文献[3-4]。直接从 2D 图像中回归 3D 坐标的方法通过深度学习模型建立 RGB 图像与 3D 坐标的端对端地映射，能够充分利用丰富的图像信息，但是网络模型的计算量大，学习的特征过于复杂。先获取 2D 信息然后提升到三维空间中进行姿态估计方法的网络结构简单、实时性好、计算难度低，但该方法对于 2D 姿态估计网络的依赖性大，会造成误差放大和空间信息丢失的问题。而 Transformer 结构和基于时间序列的方法能够有效解决这种问题。

1.1 基于 Transformer 的神经网络

Transformer 是一种基于自注意力机制的深度神经网络，最初是在自然语言处理中应用，随后广泛应用到视觉任务中。DEVLIN 等[5]所基于 Transformer 的双向编码器表示(bidirectional encoder representation from transformers，BERT)引入了一种新语言表示模型，提出了双向 Transformer，通过限制左右上下文来预测未标记的文本词汇。在多个 NLP 任务上取得了 SOTA 的成果，但相较于传统语言模型，BERT 的每批次训练数据中只有 15%的标记被预测，这导致模型需要更多地训练步骤来收敛。DOSOVITSKIY 等[6]基于 Vision Transformer 的思想将图像分割成小块，并将 Transformer 直接用于图像块序列处理，在多个图像基准上实现了 SOTA 性能，但存在优化难，依赖大尺度数据的不足。LI 等[7]采用基于级联 Transformer 的人体姿态估计(pose recognition with cascade transformers，PRTR)模型，使用 2 个阶段级联的 Transformer 分别实现了人体检测和人体姿态估计，在人体检测阶段使用卷积提取 RGB 特征，再通过 Transformer 编码器提取上下文关系，以解码器预测人体框并进行裁剪；在姿态估计阶段，根据裁剪后的人体图像和位置编码，通过 Transformer 的编解码器得到最终坐标，最后以 Spatial transformer 端到端的序列化方式来实现人体姿态的估计。ZHENG 等[8]利用 PoseFormer 分别设计了空间和时间 Transformer 模块，在 3D 人体姿态估计上取得了最优的结果。其中，空间模块受 ViT 的启发，将每帧图片的 2D 关键点作为输入，通过图像块编码和空间位置编码得到高维特征，再输入空间 Transformer 编码器来提取关键点之间的联系。同样的，每一帧图片经过空间 Transformer 模块提取特征后，会被看作是时间 Transformer 的一个特征块，经过时间位置编码后，编码和时间 Transformer 编码器就能够捕捉到输入多帧之间的时间相关性。考虑到 Transformer 对于多尺度特征的学习效果，本文也引入了 Transformer 注意力机制模块，用来提高人体姿态估计的准确度。

1.2 基于时间序列的人体姿态估计方法

SHI 等[9]提出的 LSTM 是一种特殊的递归神经网络，可以有效地传递和表达长时间序列中的信息。其可用于解决一般递归神经网络中普遍存在的长期依赖问题，并避免梯度消失和爆炸，但由于 LSTM 的串行结构无法实现多帧并行处理。文献[4]以 2D 姿势序列作为模型的输入，通过时间卷积网络(temporal convolutional network，TCN)处理序列信息并输出预测的三维姿势。同时还采用了半监督

的训练方法，将相机坐标系下 3D 姿势投影回 2D 平面，引入了重投影损失，提高了模型的预测精度。文献[3]利用循环体系结构提出了时间编码器，即未来的人体姿态帧可以从过去的人体姿态帧中学习被遮挡或不明确的姿势。利用双层 GRU 单元(LSTM 变体)对提取的时间序列特征进行编码，有效学习了连续的图像帧中人体姿势的关系，提高了三维人体姿态序列的合理性，但缺少对长距离特征的提取。为了预测出合理的三维人体姿态序列，本文也采用基于 LSTM 的时间序列方法对提取的人体姿态特征进行学习。

1.3　基于 SMPL 参数化模型的姿态估计

LOPER 等[10]提出的蒙皮多人线性模型(skinned multi-person linear, SMPL)是一种参数化的人体模型，通过调整模型的形状参数和姿势参数可以驱动 SMPL 模型描述精准的人体运动姿态。SMPL 身体模型可以随着不同的姿势自然变形，并且还具有渲染速度快、适配性好、部署高效的特点。KANAZAWA 等[11]提出了一种端到端的从二维图像恢复三维人体模型的人体网格重建(human mesh recovery，HMR)框架。通过 kinematic tree 的形式表示每个关键点的 3D 旋转角矩阵，学习关键点的角度信息来预测合理的姿势和体型。文献[3]通过 ResNet-50 神经网络和 GRU 单元预测合理的人体运动的 SMPL 模型，并且使用 AMASS 数据集来进行对抗训练，使回归器产生更加逼真与合理的人体姿态。预测人体运动 SMPL 参数模型的方法比从图像中回归三维关键点位置的方法姿势预测更加精确，产生的人体运动更加合理。因此，本文也采用 SMPL 参数作为模型的输出参数，较为精确地刻画人物运动。

2　基于 Transformer 的三维人体姿态估计网络结构

如图 1 所示，整体网络结构主要分为生成器和判别器 2 个部分。对于生成器，则使用单人视频序列作为输入，在预训练好的卷积神经网络(convolutional neural networks，CNN)中提取每帧特征，再经过 3 层 LSTM 组成的时间编码器得到每帧的潜在变量，最后对每一帧的 SMPL 人体模型参数进行回归，估计姿态和形状，输出一个三维网格。对于判别器，将生成器的输出和 AMASS 的样本输出用于判断真假样本。

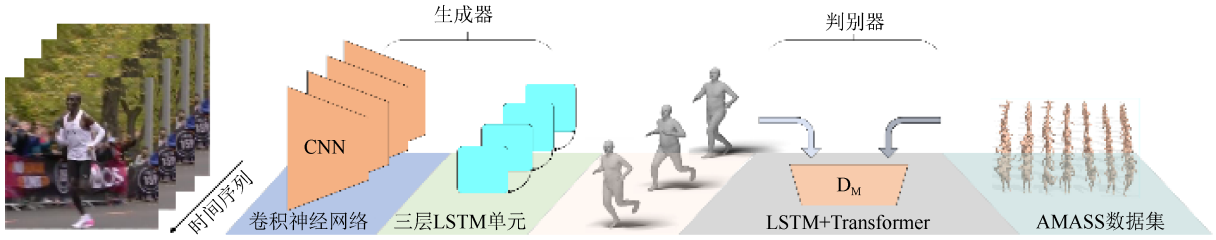


图 1　基于 Transformer 的三维人体姿态估计网络结构图

Fig. 1　Structural diagram of a three-dimensional human posture estimation network based on Transformer

2.1　生成器

当前图像中人体的姿势不确定或在视频的一些帧中身体被遮挡时，可根据过去的姿势信息来推测当前帧的信息，同时可以解决和约束姿势估计。生成器是由 2 层 ResNet-50 组成的卷积编码器、3 层 LSTM 组成的时序编码器及 2 层全连接层构成的回归器组成。对于给定的视频帧序列，输出每一帧对应的姿势和形状参数。序列中的 T 帧首先被输入到 CNN 中，该网络作为特征提取器，提取每一帧中人体姿势的空间特征，并为每帧输出一个特征向量。这些特征向量被输入到 LSTM 层中，并基于之前的帧为每一帧产生一个潜在的特征向量，学习到视频序列的时序特征，然后使用这些时序特征向量作为具有迭代反馈的回归器的输入。回归器用平均姿态初始化，得到 82 个 SMPL 参数。最后通过 SMPL 模型，得到包含 6 890 个顶点的高质量的人体 3D Mesh。

总体而言，所使用的时间编码器的损失由 2D(x)、3D(X)、姿势(θ)和形状(β)组成的 SMPL 参数模型的损失组成。这与对抗性 D_M 损失 L_{adv} 相结合，G 的总损失为

$$L_G = L_{3D} + L_{2D} + L_{SMPL} + L_{adv} \tag{1}$$

$$L_{3D} = \sum_{t=1}^{T} \left\| X_t - \hat{X}_{t2} \right\|_2 \tag{2}$$

$$L_{2D} = \sum_{t=1}^{T} \left\| x_t - \hat{x}_{t2} \right\|_2 \tag{3}$$

$$L_{SMPL} = \|\beta - \hat{\beta}\|_2 + \|\theta_t - \hat{\theta}_t\|_2 \tag{4}$$

其中，X_t为训练集中视频对应的真实 3D 人体关节位置；$\hat{X}(\Theta) = W\mathcal{M}(\theta,\beta)$是通过视频估计的人体 3D Mesh 的 3D 关节位置，及具有预训练线性回归器 W 计算得到；x_t为训练集中视频对应的真实 2D 人体关节位置；$\hat{x} \in \mathbb{R}^{j\times 2} = s\prod\left(R\hat{X}(\Theta)\right) + t$ 为使用弱透视相机模型后 3D 关节的 2D 投影，其中$[s,t]$，$t \in \mathbb{R}^2$，$\boldsymbol{R} \in \mathbb{R}^3$ 是全局旋转矩阵；β，$\hat{\beta}$ 分别为真实的和估计得到的 SMPL 形状参数；θ_t，$\hat{\theta}_t$ 分别为真实的和估计得到的 SMPL 姿势参数。

2.2 判别器

如图 2 所示，判别器结构由 3 层 LSTM 模型以及一个 Transformer 结构组合而成。3 层 LSTM 估计潜在特征，Transformer 结构聚合隐藏状态，放大重要帧的特征。

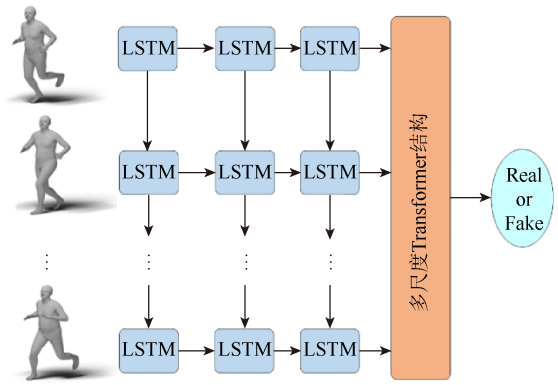


图 2 判别器网络结构图

Fig. 2 Discriminator network structure diagram

文献[11]中使用的判别器和重投影损失强制生成器生成与 2D 关节位置对齐的真实姿势。但是单个图像的识别结果不足以展示姿势序列。如果忽视时间上的连续性，多个不准确的姿势可能会被识别为有效的。为了解决该问题，本文使用运动判别器 D_M 来判断生成的姿势序列是否为真实的序列。生成器的输出 $\hat{\theta}$ 作为图 2 中描述的 LSTM 模型的输入，该模型在每个时间点处估计潜在的特征。为了捕获人体部位之间的全局依赖关系，本文使用了 Transformer 模块，最后一个线性层输出一个[0,1]之间的预测值，该值表示 $\hat{\theta}$ 属于合理的人类运动集合的概率。反向传播到 G 的对抗损失项为

$$L_{adv} = E_{\theta\sim pG}\left[(D_M(\hat{\theta}) - 1)^2\right] \tag{5}$$

判别器的目标函数为

$$L_{D_M} = E_{\theta\sim pR}[(D_M(\theta) - 1)^2] + E_{\theta\sim pG}\left[\left(D_M(\hat{\theta})\right)^2\right] \tag{6}$$

其中，pR 为来自 AMASS 数据集的真实运动序列；pG 为生成的运动序列。由于判别器 D_M 是根据有真实标注的姿势训练的，还可以学习合理的身体姿势配置，因此减少了对分离的单帧判别器的需要。

2.3 多尺度 Transformer 结构

与传统的专注于图像识别的分类器不同，判别器更加注重细节以区分生成的姿势是否真实。因此局部特征对判别器更加重要。但大的分割粒度会牺牲细节特征，而小的分割粒度会导致较长的序列，增大计算量。为了解决该问题，本文使用了 JIANG 等[12]提到的多尺度 Transformer 判别器。如图 3 所示，多尺度判别器在不同的阶段将不同大小的块作为输入，首先选择不同大小的块$(P,2P,4P)$将通过 LSTM 生成的潜在特征图分成 3 个不同的序列，最长的序列$\left(\frac{H}{P}\times\frac{W}{P}\right)\times\frac{C}{4}$结合可学习的位置编码作为第一阶段的输入。同样，第 2 个序列$\left(\frac{H}{2P}\times\frac{W}{2P}\right)\times\frac{C}{2}$，第 3 个序列$\left(\frac{H}{4P}\times\frac{W}{4P}\right)\times C$分别结合上一阶段的输出进行学习。这样 3 个不同的序列可以提取到不同尺度的语义结构和特征细节。在这个过程中，本文采用平均池化对特征图分辨率进行下采样。

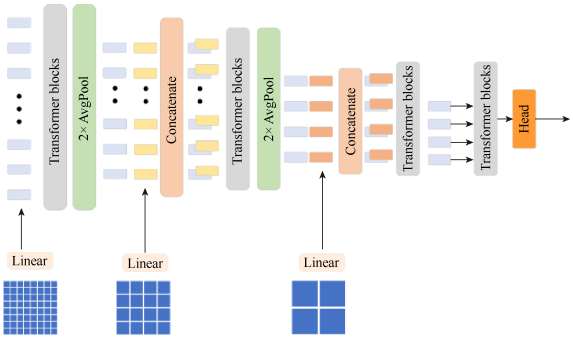


图 3 多尺度 Transformer 结构图

Fig. 3 Multiscale Transformer structure diagram

2.4 Transformer block 结构(图 4)

虽然经典 Transformer 使用确定性位置编码或可学习位置编码，如文献[13-14]。但 SHAW 等[15]提出的相对位置编码越来越受欢迎，如文献[16-19]。考虑单个自注意力层的头部的注意力机制公式为

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{7}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in R^{(H\times W)\times C}$ 分别为查询、键、值矩阵；H,W,C 分别为输入特征图的高度、宽度、维度。每个查询与 H 轴上的键之间的坐标差在$[-(H-1),H-1]$的范围内，同时由于 H 轴和 W 轴，相对位置可由参数化矩阵 $\boldsymbol{M}\in R^{(2H-1)\times(2W-1)}$ 表示。每个坐标轴的相对位置编码 E 取自矩阵 $\boldsymbol{M}$，并作为偏差项添加到注意图 QK^{T} 中，即

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_K}}+E\right)\boldsymbol{V} \quad (8)$$

相对位置编码在局部内容之间学习到了更强的特征，在大规模案例中带来了重要的性能提升，并得到了广泛地使用。因此，本文将其应用于判别器的可学习绝对位置编码之上。

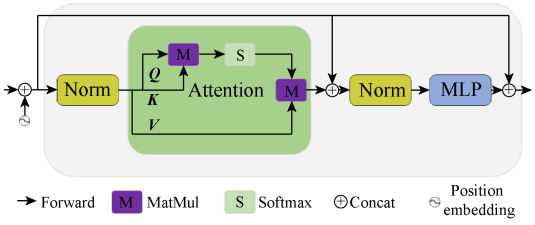


图 4　Transformer block 结构图

Fig. 4　Transformer block structure diagram

3　实验

3.1　实验环境与设置

3.1.1　实验环境(表 1)

表 1　实验环境

Table 1　Experimental environment

设备	参数
操作系统	Ubuntu20.04
深度学习框架	Pytorch1.10
CUDA 版本	11.5
开发软件	Pycharm
CPU	I7-12700KF
显卡	3 090(1 块)

3.1.2　实验细节

本文使用 HE 等[20]提出的 ResNet-50 网络作为图像编码器，在单帧姿势和形状估计任务上进行预训练，输出 $f_i\in R^{2048}$，如文献[11,21]。与 KANAZAWA 等[22]的方法类似，预先计算每个帧的 f_i 且不更新 ResNet-50 网络模型参数。本文使用 T=16 作为小批量大小为 32 的序列长度，这使得可以在单个 Nvidia RTX 3090 GPU 上训练本文的模型。对于时间编码器，本文使用隐藏大小为 1 024 的 3 层 LSTM。SMPL 回归器有 2 个全连接层，每个层有 1 024 个神经元，最后一层输出 $\hat{\theta}\in R^{85}$，包含姿势、形状和相机参数。生成器的输出作为假样本可作为判别器的输入，以及作为真实样本的真实标注的运动序列。最后的线性层预测每个样本的单个 fake/real 概率。本文还使用 KINGMA 和 BA[23]所提出的优化器，生成器和判别器的学习率分别为 0.000 05 和 0.000 1，以加快收敛速度。最后，损失函数中的每一项均有不同的权重系数，即 λ_{3D}=300，λ_{2D}=300，λ_{β}=0.06，λ_{θ}=60，λ_{adv}=0.5。

3.1.3　训练与测试

(1) 训练。本文使用批量混合的 2D 和 3D 数据集。PoseTrack 是本文使用的唯一真实 2D 视频数据集。对于 3D 注释，本文使用来自 MEHTA 等[24]提出的 MPI-INF-3DHP 数据集的 3D 关节标签。使用时，3DPW 数据集提供用来计算 LSMPL 的 SMPL 参数。MAHMOOD 等[25]提出的 AMASS 将生成器生成的样本和取自 AMASS 数据集的样本作为鉴别器的输入，训练其辨别真实动作和“伪”动作。AMASS 是一个大型开源经过 SMPL 参数标准化的三维人体动作捕捉数据集合，包含 40 h 的运动数据，超过 300 个主题及 11 000 个动作。本文还使用 VON MARCARD 等[26]3DPW 数据集进行消融实验，这表明了本文模型在室外数据上的性能。

(2) 测试。对于测试，本文使用 3DPW 和 MPI-INF-3DHP 数据集。使用 3DPW 数据集训练的结果，之前不使用 3DPW 数据集进行训练的工作进行直接比较。本文提供了对网络预测的关键点进行刚性变换后的平均每个关节位置误差(procrustes aligned mean per joint position error，PA-MPJPE)、平均每个关节位置误差(mean per joint position error，MPJPE)、加速度误差(acceleration error，Accel)和每个顶点误差(per vertex error，PVE)。本文将与最先进的单图像和时间方法进行比较。

3.2　实验结果对比

实验结果见表 2 和表 3。

表 2　3DPW 实验结果对比

Table 2　Comparison of 3DPW experimental results

Models	3DPW			
	PA-MPJPE	MPJPE	PVE	Accel
HMR [11]	76.7	130.0	-	37.4
SPIN [27]	59.2	96.9	116.4	29.8
VIB(direct)	58.7	100.0	118.5	28.7
VIBE	55.2	93.8	110.4	28.2
TR-VIBE(direct)	58.8	100.7	126.6	32.2
TR-VIBE	**53.5**	**86.3**	**101.8**	**25.5**

注：加粗数据为最优值

表 3 MPI-INF-3DHP 实验结果对比

Table 3 Comparison of MPI-INF-3DHP experimental results

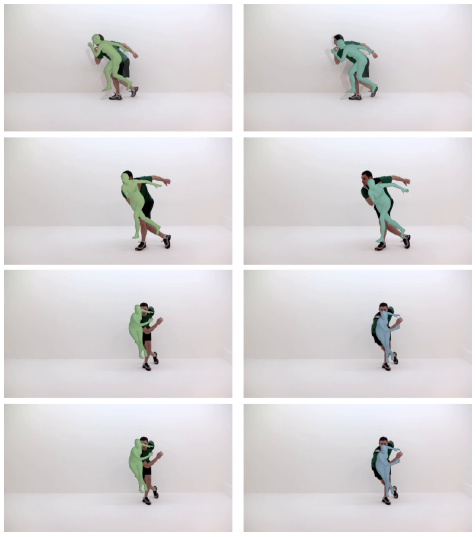
Models	MPI-INF-3DHP			
	PA-MPJPE	MPJPE	PVE	Accel
HMR [11]	89.8	124.2	-	-
SPIN [27]	67.5	105.2	-	-
VIB(direct)	66.8	103.2	916.8	33.2
VIBE	**64.3**	100.8	915.0	32.2
TR-VIBE(direct)	66.7	102.7	915.3	34.8
TR-VIBE	64.9	**99.0**	**907.9**	**30.1**

注：加粗数据为最优值

从表 2 和表 3 中可以看到，本文方法提高了 VIBE 的性能。在使用 3DPW 数据集进行训练时，在 MPI-INF-3DHP 数据集上的实验结果明显优于所有以前基于帧和时间的方法，而在 3DPW 数据集有一定下降，泛化能力有所降低。此外，在使用 3DPW 数据集进行训练时，在具有挑战性的 3DPW 和 MPI-INF-3DHP 数据集上均明显优于所有以前基于帧和时间的方法。其中“中”代表无。

3.3 室内和室外视频的定性比较结果

室内和室外视频的定性比较结果如图 5 和图 6 所示。



(a) (b)

图 5 室内定性比较图

Fig. 5 Indoors qualitative comparison diagram ((a) VIBE; (b) Ours)

3.4 消融实验

表 4 中添加了 Transformer，可以看到本文得到了平滑运动的同时精确单个姿势，改为 3 层 LSTM 后模型能力有所提升。其中添加 Transformer 对模型各项能力提升的效果较大。其中 VIBE-α 为在 VIBE 的基础上添加了 LSTM 结构的模型；VIBE-β 为在 VIBE 的基础上添加了 Transformer 结构的模型；TR-VIBE-α 为在本文方法上 Transformer 结构没有多尺度的模型；TR-VIBE-β 为本文方法上 Transformer 结构没有相对位置编码的模型。

(a) (b)

图 6 室外定性比较图

Fig. 6 Outdoor qualitative comparison diagram ((a) VIBE; (b) Ours)

表 4 LSTM 和 Transformer 的消融实验

Table 4 Ablation experiments for LSTM and Transformer

Models	3DPW			
	PA-MPJPE	MPJPE	PVE	Accel
VIBE	55.2	93.8	110.4	28.2
VIBE-α	55.1	94.2	110.2	28.5
VIBE-β	55.0	87.7	104.2	25.8
TR-VIBE	**53.5**	**86.3**	**101.8**	**25.5**
TR-VIBE-α	55.9	92.5	110.1	28.6
TR-VIBE-β	55.0	94.2	110.4	29.8

注：加粗数据为最优值

4 结束语

本文提出了一种基于 Transformer 的三维人体姿态估计方法，使用预训练的 ResNet-50 神经网络与 LSTM 单元组合成模型的生成器，用来生成合理的人体运动序列。通过多层 LSTM 单元与基于 Transformer 的判别器判断合理的人体运动姿态，在学习和对抗的过程中不断提高模型输出合理的 SMPL 参数模型序列的能力，实现较好的三维姿态估计效果，平均每个关节位置误差下降到 53.5%，

与 VIBE 相比，误差降低了 1.7%。在未来的工作中应该更加关注提升模型的预测性能，通过使用改进 Transformer 注意力机制代替 CNN 层进行 2D 关节的估计，提升模型对于多尺度特征的学习能力，并进一步改进 Transformer 结构，降低计算复杂度。

参考文献 (References)

[1] PAVLAKOS G, ZHOU X W, DERPANIS K G, et al. Coarse-to-fine volumetric prediction for single-image 3D human pose[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 1263-1272.

[2] LI S J, CHAN A B. 3D human pose estimation from monocular images with deep convolutional neural network[M]//Computer Vision - ACCV 2014. Cham: Springer International Publishing, 2015: 332-347.

[3] KOCABAS M, ATHANASIOU N, BLACK M J. VIBE: video inference for human body pose and shape estimation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 5252-5262.

[4] PAVLLO D, FEICHTENHOFER C, GRANGIER D, et al. 3D human pose estimation in video with temporal convolutions and semi-supervised training[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2019: 7745-7754.

[5] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2021-12-02]. https://arxiv.org/abs/1810.04805.

[6] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. [2021-12-05]. https://arxiv.org/abs/2010.11929.

[7] LI K, WANG S J, ZHANG X, et al. Pose recognition with cascade transformers[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2021: 1944-1953.

[8] ZHENG C, ZHU S J, MENDIETA M, et al. 3D human pose estimation with spatial and temporal transformers[C]//2021 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2021: 11636-11645.

[9] SHI X J, CHEN Z R, WANG H, et al. Convolutional LSTM Network: a machine learning approach for precipitation nowcasting[C]//The 28th International Conference on Neural Information Processing Systems - Volume 1. New York: ACM, 2015: 802-810.

[10] LOPER M, MAHMOOD N, ROMERO J, et al. SMPL: a skinned multi-person linear model[J]. ACM Transactions on Graphics, 2015, 34(6): 248.

[11] KANAZAWA A, BLACK M J, JACOBS D W, et al. End-to-end recovery of human shape and pose[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2018: 7122-7131.

[12] JIANG Y, CHANG S, WANG Z. Transgan: two pure transformers can make one strong gan, and that can scale up[J]. Advances in Neural Information Processing Systems, 2021, 34: 14745-14758.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all You need[C]//The 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000-6010.

[14] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. [2021-12-02]. https://arxiv.org/abs/2010.11929.

[15] SHAW P, USZKOREIT J, VASWANI A. Self-attention with relative position representations[EB/OL]. [2021-12-01]. https://arxiv.org/abs/1803.02155.

[16] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[EB/OL]. [2021-12-02]. https://arxiv.org/abs/1910.10683.

[17] HUANG C Z A, VASWANI A, USZKOREIT J, et al. Music transformer[EB/OL]. [2021-12-05]. https://arxiv.org/abs/1809.04281.

[18] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2021: 9992-10002.

[19] HU H, ZHANG Z, XIE Z D, et al. Local relation networks for image recognition[C]//2019 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2019: 3463-3472.

[20] HE K M, ZHANG X Y, REN S Q, et al. Identity mappings in deep residual networks[M]//Computer Vision - ECCV 2016. Cham: Springer International Publishing, 2016: 630-645.

[21] KOLOTOUROS N, PAVLAKOS G, BLACK M, et al. Learning to reconstruct 3D human pose and shape via model-fitting in the loop[C]//2019 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2019: 2252-2261.

[22] KANAZAWA A, ZHANG J Y, FELSEN P, et al. Learning 3D human dynamics from video[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2019: 5607-5616.

[23] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. [2022-01-03]. https://arxiv.org/abs/1412.6980.

[24] MEHTA D, RHODIN H, CASAS D, et al. Monocular 3D human pose estimation in the wild using improved CNN supervision[C]//2017 International Conference on 3D Vision. New York: IEEE Press, 2017: 506-516.

[25] MAHMOOD N, GHORBANI N, TROJE N F, et al. AMASS: archive of motion capture As surface shapes[C]//2019 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2019: 5441-5450.

[26] VON MARCARD T, HENSCHEL R, BLACK M J, et al. Recovering accurate 3D human pose in the wild using IMUs and a moving camera[M]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018: 614-631.

[27] KOLOTOUROS N, PAVLAKOS G, BLACK M, et al. Learning to reconstruct 3D human pose and shape via model-fitting in the loop[C]//2019 IEEE/CVF International Conference on Computer Vision. New York: IEEE Press, 2019: 2252-2261.