DOI:10.16136/j.joel.2022.12.0101

基于时空多特征融合网络的三维人体姿态估计

叶 俊[1],张 云[1,2*]

(1.云南省计算机应用重点实验室,云南 昆明 650500;2.昆明理工大学 信息工程与自动化学院,云南 昆明 650500)

摘要:目前,常见的三维(3D)人体姿态估计算法在表征学习上取得很好的效果,但是在人体骨架关节点处依然存在估计精度不佳等问题,因此,如何从单目 RGB 图像中利用冗余的二维(2D)姿态序列时空信息来估计人体姿态的有效方式是一个研究的难点。本文提出一种基于时空多特征融合网络的三维人体姿态估计算法,具体是结合一种图像外观信息和运动时序信息时空多特征融合层级方法,该方法利用一种紧凑的卷积神经网络(convolutional neural network,CNN)学习时空信息将二维关节点位置信息建模为三维关节点位置。实验结果表明,本文所提出的方法能实现较为先进的端对端姿态估计精度,而且不需要任何后处理阶段的姿态优化方法,本文得到的姿态估计在平均精度上得到有效的提升,证明本文方法能够有效提高人体姿态估计的准确性。

关键词:三维人体姿态估计;时空特征;运动补偿网络;特征融合网络
中图分类号:TP391.4 **文献标识码**:A **文章编号**:1005-0086(2022)12-1306-09

Three-dimensional human pose estimation based on spatio-temporal multi-feature fusion network

YE Jun[1], ZHANG Yun[1,2*]

(1. Key Laboratory of Applications of Computer Technology of Yunnan Province, Kunming, Yunnan 650500, China; 2. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China)

Abstract: At present, the common 3D human pose estimation algorithms have achieved good results in representation learning, but there are still problems such as poor estimation accuracy at the joint points of the human skeleton. Therefore, how to effectively estimate human pose from a monocular RGB image using the redundant 2D pose sequence sptatio-temporal information is a different point in the research. In this paper, a 3D human pose estimation algorithm based on spatio-temporal multi-feature fusion network is proposed. The method utilizes a compact convolutional neural network to learn spatio-temporal information to model the 2D joint position information as 3D joint position. The experimental results show that the method proposed in this paper can achieve relatively advanced end-to-end attitude estimation accuracy, and does not require any attitude optimization method in the post-processing stage. The pose estimation obtained in this paper can effectively improve the average accuracy, which proves that the method in this paper can effectively improve the accuracy of human pose estimation.

Key words: 3D human pose estimation; spatio-temporal features; motion compensation network; feature fusion network

1 引 言

三维(3D)姿态估计是计算机视觉中的一项关键技术,它用二维(2D)图像或视频来预测景物在3D空间中的相对位置,从而达到从 2D 图像或人类视觉系统来理解景物的 3D 姿态的目的。3D 姿

* **E-mail**: zhangyun92x@163.com
收稿日期:2022-02-22 **修订日期**:2022-04-08
基金项目:国家自然科学基金(61262043)、云南省科技计划项目(2011FZ029)和重点实验室开放基金项目(2020106)资助项目

态估计有广泛的应用,包括增强现实、人机交互、自动驾驶等方面。然而,将三维图像投影到二维图像上会导致深度信息丢失,使三维姿态估计的图像变得很模糊,解决这类问题常见的方法是使用多个摄像机系统,但是由于设置一个校准和同步多摄像机系统所带来的成本,使其变得不切实际。另外一个解决方案是使用深度传感器进行人体姿态估计,但是使用 RGB-D 传感器耗电且不普及,会使得基于 RGB 传感器的 3D 姿态估计变得很棘手,相较而言,目前以手机或网络摄像头获取的 RGB 视频为主的单目系统使用方便且普及,因此,解决在单目系统中使用立体计算机视觉,特别 3D,是姿态估计中面对的问题,具有重要的理论及实用价值。尽管经过多年的努力,人体姿态估计仍然是一个难题,因为视觉外观变化、视点变化、遮挡和姿态表示的高维性所带来巨大挑战。

面对这些挑战,近些年来许多相关研究者开始尝试用深度学习网络来解决以上存在的问题,文献[1]使用多视图自我监督的时间信息对多视图相机系统的 2D 身体姿势估计进行三角测量,对预测的 3D 身体骨架施加几何约束,以预测每个人的 3D 身体姿势。文献[2]对人体周围的 3D 空间进行精细离散化,并训练一个卷积神经网络(convolutional neural network,CNN)来预测每个关节的每个体素的可能性,采用了从粗到细的预测方法,并实现了图像特征的迭代细化和重复处理。文献[3]提出了一个多任务框架,用于从静止图像中联合 2D 和 3D 姿态估计和从视频序列中识别人类动作。文献[4]融合了上下文时空信息来解决自遮挡问题,从而让模型对不同程度的遮挡变得鲁棒,此外,更多地挖掘帧与帧之间的时空信息来进行特征融合,进而将关键点进行精确的定位。上面提及的方法虽然在单个图像和视频中取得目前比较优的人体姿态估计效果,但是在遇到频繁的遮挡以及存在大量的自由角度铰接关节点处时,依然在关节点存在估计效果差、成本高等情况。

因此,为了应对三维姿态估计当前仍然存在的问题,文中通过新的深度学习方法用来解决单视图三维姿态估计带来的挑战,具体而言,首先,提出了具有多级运动补偿的 CNN。它能学习到表征姿态的图像时空特征,并且对遮挡与运动模糊具有鲁棒性。其次,提出了一个从鲁棒的时空特征计算得到分离的 3D 姿态的参数化姿态映射函数。它能学习到系列 2D 关节点位置与 3D 关节点位置之间的关系,并且对姿态估计的优化无需任何后处理。最后,采用两个基准测试数据集对本方法进行了系统的实验评价。

2　算法框架

关于单目 3D 姿态估计,近年来已提出了两类方法:一种是用图像数据直接回归三维姿态的方法[5];另外一种是在姿态空间中找出与图像数据一致的姿态生成方法[6],但是随着数据集的增大,姿态估计开始向深度学习架构的方向发展,基于所采用的基本方法,大致可以分为直接图像回归三维姿态的架构和以关节点置信图来预测的 3D 姿态的架构,与第二种方法类似。本文提出新颖的基于时空特征融合的网络架构,如图 1 所示。

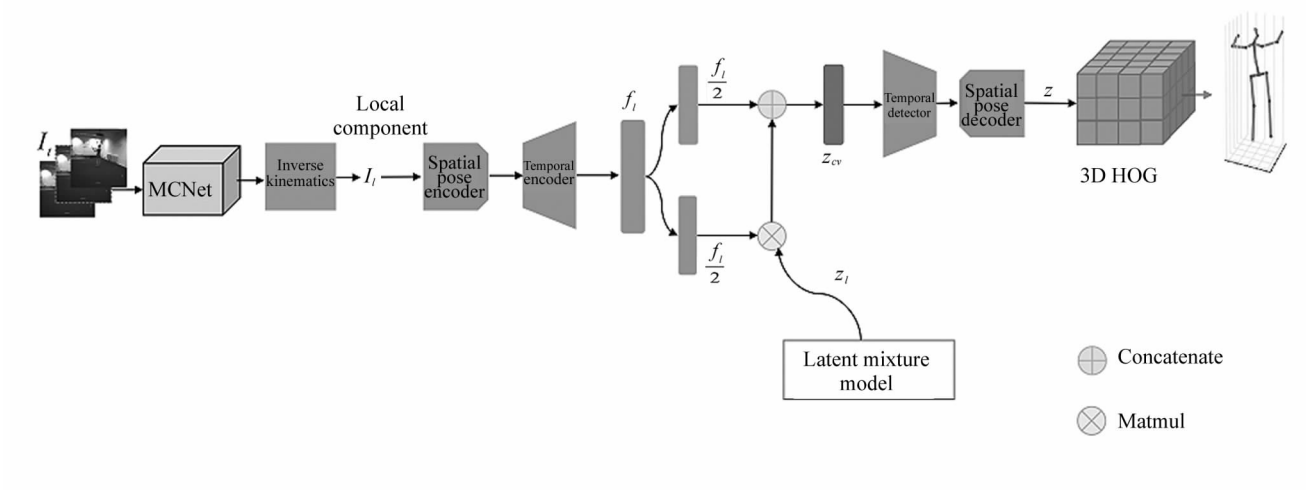

图 1　时空多特征融合三维姿态估计方法框架

Fig. 1　Method framework for 3D pose estimation based on spatio-temporal multi-features fusion

3 方　法

图 1 给出了本方法的处理流程。首先，整个网络将 n 帧内、当前帧 t 的邻域所有的 RGB 帧作为输入，定义在连续帧中查找出人的边界框。然后，进行补偿运动以形成时空特征。最后，学习从时空特征中到其中心视频帧之间的 3D 姿态映射关系，以得到用骨架表示的 3D 人体姿态输出。骨架中的各个关节点用相对于根节点（局部位姿结构序列）的 3D 位置来表示。之所以选择骨架表示法，是因为对其进行回归无需知道测试者的身体比例参数。虽然此骨架表示不具取向不变性，但是可以使用时空特征所提供的丰富信息克服该困难。

3.1　时空特征融合

让 I_i 包含一个主体序列的第 I 张图像，$Y_i \in \mathcal{R}^{3\cdot D}$ 是其相对应的三维关节点位置进行编码的向量。用回归判别方法来估计 Y_i 需要学习如下的映射函数 $f: X_i \rightarrow Y_i \approx f(X_i)$，其中 $X_i = \Omega(I_i; m_i)$ 是一个边界框中通过 CNN 获取的特征向量，m_i 是一个人体 I_i 的前景（人体）掩膜。模型参数通常是通过 N 组标签数据 (X_i, Y_i) 的训练集 $S = \{(X_i, Y_i)\}_{i=1}^{N}$ 就可以得到 f 的参数值，其中 Y_i 表示在视频帧中某一中心帧，然而，图像中的 3D 人体形貌时常会受到自遮挡或运动模糊固有因素的不利影响。因此，利用单帧图像 I_i 上的空间特征就想获得可靠的 3D 姿态估计通常是有难度的。此时，考虑引入从时序图像（视频段）上提取的时间特征作为一种新的信息源就显得很有必要了。据此，本方法用一个以 I_i 为中心帧、长度为 T 帧的视频段来组成一个 3D 时空数据集 $V_i = [I_{i-T/2+1}, \cdots\cdots, I_i, \cdots, I_{i+T/2}]$，相等地，$Z_i \rightarrow Y_i \approx f(Z_i)$，其中 $Z_i = S(V_i; m_{i-\frac{T}{2}+1}, \cdots\cdots, m_{i+T/2})$ 是根据训练数据容量 V_i 得到的特征向量，在训练中，训练集表示为 $S = \{(Z_i, Y_i)\}_{i=1}^{N}$，其中 Y_i 是出现在视频段 Z_i 中的姿势标签。传统方法是基于 3D 位置的联合表示的姿势树用于根节点表示姿势的一个经典方法，但是在时间过程中，会产生非约束骨骼引起的看起来不自然的运动长度以及关节运动超出正常关节范围，为了解决问题，使用固定的参考姿势并向前将姿势树的相对关节位移为逆运动学旋转矩阵。

为了确定 Y_i，文中根据逆向运动学的方法，直接将运动学描述的机器人状态作为输入函数来获取运动学积分以便获得某个时间 t 的人体的姿态模型，即基于 3D 关节的局部姿势序列映射到基于旋转姿势序列的对应关系，通过该过程可用确保整个序列骨架长度的一致性，即表示局部姿态序列 $I_l = \{[X_l^{(1)}, X_l^{(2)}, \cdots, X_l^{(p)}]\}$，其中 $[X_l^{(i)}]_t \in \mathcal{R}^{3\cdot D}$、$1 \leq p \leq P$、$1 \leq t \leq T$。$[X^{(i)}]_t$ 表示在某个时刻 t、第 i 个人体姿态的配置参数。图 1 中 Local component 部分可以表示为 $I_l = \{[X_l^{(1)}, X_l^{(2)}, \cdots, X_l^{(p)}]_t\}$，其中 $P_t^{(i)} \in R^{3D}$ 表示某时刻 t 的第 i 个人体姿态。逆运动学方法的输出尺寸为 $T \times P \times J \times 3$，通过重复实验总结规律，将动作信息中最大的对象数量 P 设置为 2，将关节点移动数量 J 设置为 17，时间步长 T 设置为 64。

3.2　运动补偿融合网络

对于视频运动中产生的运动模糊等情况，常见的方法是采用运动补偿方法，最近的一些学者提出运动补偿网络来解决超分辨率以及视频增强[7,8]，本文基于前人的研究工作，提出基于光流的金字塔多层级运动补偿融合网络合成不同尺寸、大小的运动补偿信息，运动补偿网络架构（motion compensation network，MCNet）如图 2 所示。

要改善受运动模糊（或空间分辨率降低）影响的姿态估计精度，就需要用去模糊处理来获得较清晰的视频帧。降低视频的运动模糊近期已成为一个活跃的研究邻域，CHO 等[5]考虑采用基于块窗口数倍的候选键去降低运动视频中的模糊问题，但是该方法没有运用相邻连续帧之间的空间连续性，因此，为了降低空间上的分辨率可能会造成在运动姿态估计上的精度损失以及需要获取更多的精确帧。本文提出的运动补偿网络模块能够学习视频图像帧块之间的表征信息，在图 2 中采用了金字塔多级融合在两邻接帧之间进行运动补偿，从而学习到更清晰的视频帧，运动补偿网络具体配置列在表 1 中。

首先，该模块中定义视频流中低分辨率（low-quality）在某一时刻 t 视频序列中某帧 I_t^L，具体而言，光流金字塔有 4 倍光流、2 倍光流和全光流 3 个尺度级。根据表 1 中第二栏配置的网络，4 倍光流补偿帧的生成过程描述如下。输入为两个视频帧，即目标帧 I_t^L 和其邻近帧 I_t^L，此外使用对称跳跃连接在解码器和编码器之间重建的图像编码特征，同时保留图像结构，在编码器和解码器之间增加 12 个残差块，以便细化模糊的特征图，然后通过将两帧连接用两个步长为 2 的卷积块进行向下采样得到一个光流 $\Delta_{t\rightarrow i}^{\times 4}$，对 $\Delta_{t\rightarrow i}^{\times 4}$ 与目标帧采用双线性插值的仿射变换（warping）Γ 得到经过 4 倍光流 $\Delta_{t\rightarrow i}^{\times 4}$ 补偿的对齐帧，$\hat{I}_{i,\times 4}^{L}$，其表达式如下所示：

$$\hat{I}_{i,\times 4}^{L} = \Gamma(I_i^L, \Delta_{t\rightarrow i}^{\times 4}), \tag{1}$$

式中,$\Gamma(\cdot)$表示 Warp 操作。$\hat{I}^{L}_{i,\times 4}$是粗光流 $\Delta^{\times 4}_{t\to i}$和 I^{L}_{t} 通过一个×4 运动估计得到。详细参数配置如表 1 所示。表中,k 表示核,s 表示步长,n 表示通道,例如 $k5$ 表示核大小为 5,$s2$ 表示步长为 2,$n32$ 表示通道数为 32。

根据表 1 中第三栏配置的网络,2 倍光流补偿帧的生成过程描述如下。输入为 I^{L}_{t}、I^{L}_{i}、$\Delta^{\times 4}_{t\to i}$、$\hat{I}^{L}_{i,\times 4}$。首先使用一个双层卷积和一个两倍向上的采样层去获取×2 的光流 $\Delta^{\times 2}_{t\to i}$(为了获得更多的多级光流表征信息),通过结合光流 $\Delta'^{\times 2}_{t\to i}$(=$\Delta^{\times 2}_{t\to i}+\Delta^{\times 4}_{t\to i}$)与 L^{L}_{t} 的基于双线性插值的仿射变换操作 Γ 得到经过 $\Delta'^{\times 2}_{t\to i}$ 补偿的对齐帧 $\hat{I}^{L}_{t,\times 2}$。具体如式(2)和(3)所示:

$$\Delta'^{\times 2}_{t\to i}=\Delta^{\times 2}_{t\to i}+\Delta^{\times 4}_{t\to i}, \tag{2}$$

$$\hat{I}^{L}_{t,\times 2}=\Gamma(I^{L}_{t},\Delta'^{\times 2}_{t\to i})。 \tag{3}$$

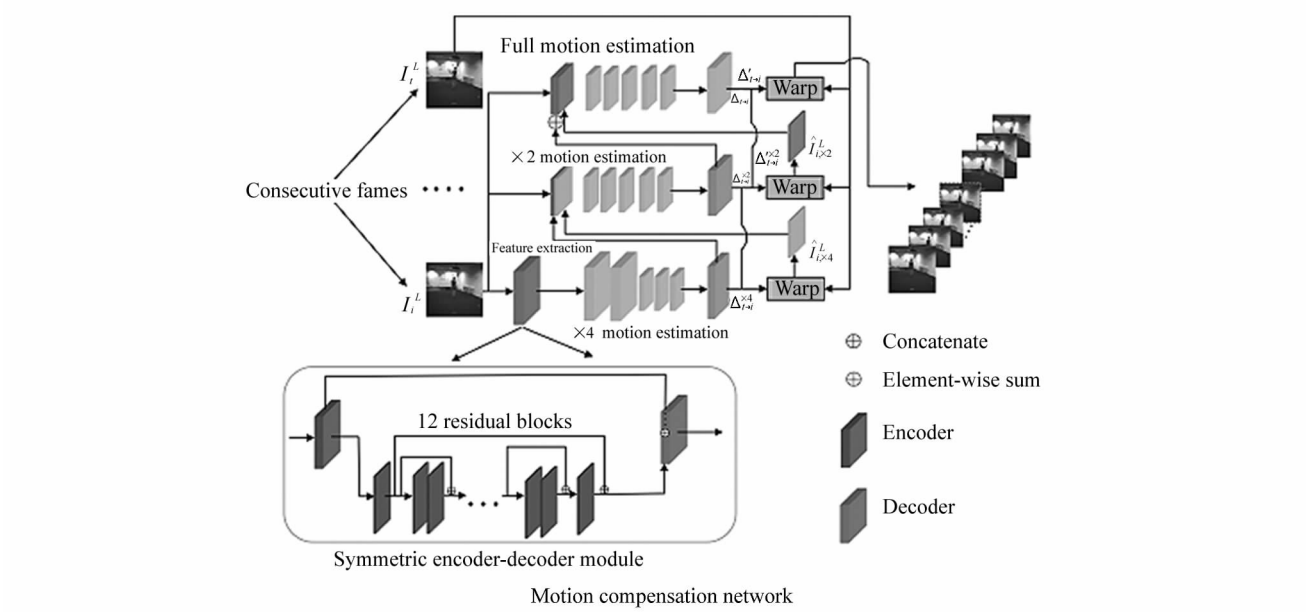

图 2　运动补偿网络

Fig. 2　Motion compensation network

表 1　多层结构运动估计配置表

Tab. 1　Multi-layer structure motion estimation configuration table

Layers	×4 flow	×2 flow	Full resolution flow
1	$k5s2n32$/ReLU	$k5s2n32$/ReLU	$k3s1n32$/ReLU
2	$k3s1n32$/ReLU	$k3s1n32$/ReLU	$k3s1n32$/ReLU
3	$k5s2n32$/ReLU	$k3s1n32$/ReLU	$k3s1n32$/ReLU
4	$k3s1n36$/ReLU	$k3s1n32$/ReLU	$k3s1n32$/ReLU
5	$k3s1n36$/tanh	$k3s1n8$/tanh	$k3s1n2$/tanh
6	Upscale ×4	Upscale ×2	—

根据表 1 中右栏配置的网络,全光流补偿帧的生成过程描述如下。输入 I^{L}_{t}、I^{L}_{i}、$\Delta^{\times 2}_{t\to i}$、$\hat{I}^{L}_{i,\times 2}$ 和 $\Delta'^{\times 2}_{t\to i}$(输入 $\Delta^{\times 2}_{t\to i}$ 和 $\hat{I}^{L}_{i,\times 2}$ 是为了充分利用额外的运动信息)。首先,将 I^{L}_{t}、I^{L}_{i}、$\Delta^{\times 2}_{t\to i}$和 $\hat{I}^{L}_{i,\times 2}$进行向量拼接操作。然后,全光流运动估计执行若干个下采样卷积层即可获取完整分辨率的光流表示,进而生成全光流 $\Delta_{t\to i}$。最后,通过混合光流 $\Delta'_{t\to i}$(=$\Delta'^{\times 2}_{t\to i}+\Delta_{t\to i}$)进行向量拼接操作与 I^{L}_{t} 的 Warping 操作 Γ 得到 $\Delta'_{t\to i}$补偿的对齐帧 $\hat{I}^{L}_{t,\times 2}$,形象化地,如图 2 最上面的全光流运动估计分支包含若干个没有去学习全光流表征信息下采样的卷积层,输出光流 $\Delta'^{\times 2}_{t\to i}$和相应的补偿帧 $\hat{I}^{L}_{i,\times 2}$和原帧 I^{L}_{t} 作为全分辨率模块的输入,然后生成全光流 $\Delta_{t\to i}$,最后将得到具有 $\Delta'_{t\to i}$全部流的补偿帧 $\hat{I}^{L}_{i}$。具体如式(4)和(5)所示:

$$\Delta'_{t\to i}=\Delta'^{\times 2}_{t\to i}+\Delta_{t\to i}, \tag{4}$$

$$\hat{I}^{L}_{i}=\Gamma(I^{L}_{t},\Delta'_{t\to i})。 \tag{5}$$

3.3　编码器与译码器

本文模型方法主要分为两个阶段,第 1 个阶段是空间姿态编码器阶段,该编码器包含 1 个 2D 残差卷积块并且学习每个 X_l 中的时间步长,第 2 阶段是

时间编码器应用多个 2D 残差 CNN 卷积来对时间步长 T 进行下采样操作，最后时间编码器输出被展平为产生时空局部特征的 f_l。在译码器结构之前，通常借已知部分的数据信息来估计全部数据总体，使用的方法是基于相对熵，又称为 KL 散度(kullback leibler，KL)，由于数据样本之间是基于动作状态的相似性，但是将会带来包含大量聚类等情况，为此采用高斯混合模型得到潜在先验知识，此外，使用条件策略将动作属性作为推理的一部分，首先从高斯混合模型采样得到一个潜在的向量 z_l 组件部分，将其应用到编码器中表示，潜在表征向量 z_l 以转换的视点为条件，视点在训练期间被设置为固定默认值。通过连接得到的结果再由线形层转换并且重新生成整型值以获得表征向量 z_{cv}。而序列解码器与编码阶段形成互补结构，本地空间与时间信息解码器执行时空上采样。获得潜在表征向量 z_{cv}，生成动作序列局部的位姿分量，进而获得每个关节点的位置表示，在本地空间解码器中使用转置卷积，根据实验，发现可以生成更好的序列。编码器和译码器详细的结构如图 3 所示。

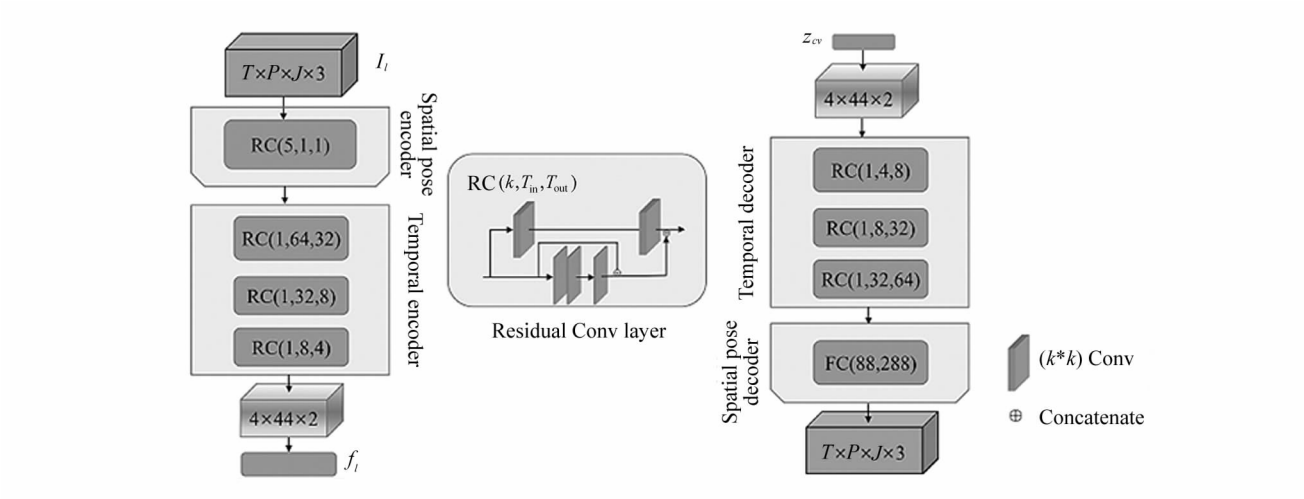

图 3　编码器和解码器模块

Fig. 3　Encoders and decoders modules

3.4　姿态回归

最后特征向量 Z 是基于 3D HOG(3D histogram of oriented gradients)描述符，它同时对外观和运动信息进行编码。首先将数据按体积细分等距单元格来计算。为了增加图像特征描述能力，使用多尺度方法。首先，使用不同核尺寸大小去计算 3D HOG 描述符，具体地，分别使用大小为 2×2 和 4×4 和 8×8 的 3 级多维度空间，同时将空间核尺寸大小设置为数值很小，去捕捉时序良好的时序信息，最终的特征向量 Z 通过将数倍分辨率聚合成单个向量。3D 姿态估计旨在找到映射关系，$Z\rightarrow f(Z)\approx Y$，其中 Z 是 3D HOG 在时空特征上的描述符，是其中心帧中 3D 姿态。本文为了更好地表示 f，在先前常见方法中采用基于核岭回归、支持向量机和核依赖估计。在核岭回归方法中，假定模型定义为 $y=f(x)+w$，试图去估计姿态回归函数 f，其中 w 为加性噪声，y 表示为在实数集 $\mathcal{R}$ 中一个相应变量值，x 表示 $\mathcal{R}^d$ 中的协变量或者预测变量，当观测一对 $(x^{(i)},y^{(i)})\in\mathcal{R}^d\times\mathcal{R}$ 时，$i=1,\cdots,n$，则 $y^i=f^*(x^i)+w^{(i)}$，基于再生核希尔伯特空间(RKHS)中核大小为 K 的 H，得到式(6)：

$$\hat{f}=\arg\min_{f\in H}\frac{1}{2}\sum_{i=1}^{n}(y^{(i)}-f(x^{(i)}))^2+\frac{\lambda}{2}\|f\|_H^2。\tag{6}$$

在最优化问题中，数据项是其中第一步，第二步是正则化项，来自于 RKHS 的惩罚函数 f 正则化项误差太大，因此，基于核逻辑回归表征定理[6]，提出 $\hat{f}$ 函数形式：

$$\hat{f}(\cdot)=\sum_{j=1}^{n}\alpha_j K(\cdot,x^{(j)}),\tag{7}$$

定义 $y=\begin{pmatrix}y^{(1)}\\ \vdots\\ y^{(n)}\end{pmatrix}\in\mathcal{R}^n$ 和 $K\in\mathcal{R}^{n\times n}$，$K_{ij}=k(x^{(i)},x^{(j)})$，然后简化式(7)对于最原始的姿态回归问题，推导出式(8)：

$$\hat{\alpha}=\arg\min\frac{1}{2}\|y-K\alpha\|_2^2+\frac{\lambda}{2}\alpha^{\mathrm{T}}K\alpha。\tag{8}$$

最后的推导公式表示为式(9)：

$$\frac{1}{2}\| \sum_{j=1}^{n}\alpha_j k(\cdot,x^{(j)}) \|_H^2 \Rightarrow$$

$$\frac{1}{2}\| \sum_{j=1}^{n}\alpha_j k(\cdot,x^{(j)}) \|_H^2 =$$

$$\sum_{i,j=1}^{n}\alpha_i\alpha_j\langle k(\cdot,x^{(i)}),k(\cdot,x^{(j)})\rangle_H =$$

$$\sum_{i,j=1}^{n}\alpha_i\alpha_j k(x^{(i)},x^{(j)}) = \sum_{i,j=1}^{n}\alpha_i\alpha_j k(x^{(i)},x^{(j)}) = \alpha^{\mathrm{T}}K\alpha。 \tag{9}$$

基于核岭回归方法,在训练模型过程中分别为每个维度训练模型,为了找到时空特征到 3D 姿态映射关系,一般使用的是核岭回归方法来解决正则化二乘问题,具体如式(10)所示:

$$\arg\min_W \sum_i \| Y_i - W\Phi_Z(Z_i) \|_2^2 + \| W \|_2^2, \tag{10}$$

式中,(Z_j,Y_j)是训练对,Φ_Z 表示为傅里叶近似指数 χ^2 核,W 可以通过 $W=(\Phi_Z(Z)^{\mathrm{T}}\Phi_Z(Z)+I)^{-1}\cdot\Phi_Z(Z)^{\mathrm{T}}Y$ 计算得到。

3.5　损失函数

本文提出可以提供准确、快速、端对端的可训练人体姿态估计网络,本文方法的损失函数由运动补偿网络和自动编码器两部分构成,其中,第一部分在训练运动补偿网络部分将损失函数设置为 L_1,将双线性插值仿射交换方法中次优化超参数设置为 λ,根据光流 $\Delta'_{t\to i}$,在给定的初始邻帧 I_i^L 与补偿帧 $\hat{I}_i^l$ 之间的平均绝对误差损失函数可以定义为:

$$L_1(\lambda) = \sum_{i=-N}^{N} \| I_i^L - \hat{I}_i^l \|_2^2 + \alpha \| \Delta'_{t\to i} \|_2^2。 \tag{11}$$

在编码器部分,通常优化基于编码器的网络架构需要在训练数据中对重建损失和潜在的特征空间之间进行平衡,序列解码器生成 3D 关节点表示。在训练优化阶段,使用均方误差损失函数结合三维空间 L_{3D}和局部运动重建损失 $L_{\mathrm{local}}^{\mathrm{rec}}$,定义该部分损失函数为 $L_2(\beta)$,β 为该部分的优化超参数。在该部分,采用高斯混合模型运用了 KL 散度,其损失函数定义为 L_{KL},因此,总损失函数定义为:

$$L_2(\beta) = L_{\mathrm{local}}^{\mathrm{rec}} + \lambda_{\mathrm{KL}}L_{\mathrm{KL}}, \tag{12}$$

式中,λ_{KL}表示优化的超参数。因此,在该模型中总损失函数为 L,如下式所示:

$$L = L_1(\lambda)+L_2(\beta) = \sum_{i=-N}^{N} \| I_i^L - \hat{I}_i^l \|_2^2 + \alpha \| \Delta'_{t\to i} \|_2^2 + L_{\mathrm{local}}^{\mathrm{rec}} + \lambda_{\mathrm{KL}}L_{\mathrm{KL}}。 \tag{13}$$

4　数据集与实验参数设置

本文选取了大规模人体姿态估计数据集 Human3.6m 和 KTH Multiview Football Ⅱ多视角足球数据集,其中 Human3.6m 包含最大规模运动捕捉数据集,大约有 360 万张图片和相应的复杂运动场景。具有 15 种不同的动作和 4 个不同视角的 11 个训练对象,在本文实验中测试 9 种不同的动作姿势。另外,为了验证本文模型在室外场景的鲁棒性,选取 KTH Multiview Football Ⅱ足球数据集作为室外场景数据集,该数据集是通过高速相机跟随一名在场地快速移动的足球运动员进行拍摄,视频是从 3 个不同的摄像机视点捕获得到,因此该数据集中足球员具有快速的运动,其输出姿态是 14 个三维关节坐标的矢量,在人体姿态部位中,分为骨盆(pelvis)、躯干(troso)、上臂(upper arms)、下臂(lower arms)、大腿(upper legs)以及小腿(lower legs)6 个部位。

4.1　性能评价指标

3D 人体姿态估计大多数方法采用标准是:MPJLE:"Mean Per Joint Localization Error",即表示每个联合定位误差的平均值,即表示每关节位置误差,指标越小,则可以认为 3D 人体姿态估计算法越好。因此,本文在数据集 Human3.6m 中,为了评价三维人体姿态估计准确率,通常将真实值和预测的联合位置之间的平均欧几里得距离(欧式距离度量,单位:mm)作为姿态估计精度评价指标,在测量关节点之间距离之前,首先要通过 Procrustes 变换将预测姿态估计的骨架和地面进行对齐,为了对比实验公平性,本文实验也采用同样的方法。

在 KTH Multiview Football Ⅱ足球数据集,通常的评价标准是使用 PCP(percentage of correctly estimated part)分数作为评价指标。在表 2 中对来自 20 个不同帧的真实图像姿态估计结果进行定量总结。PCP 分数百分比和设置 $\alpha=0.2$ 用于测量使用 2 个摄像头估计姿态估计性能。

4.2　实验结果

本文分别对比了前人在 Human3.6m 数据集和 KTH Multiview Football Ⅱ多视角足球数据集上的实验结果,在数据集 Human3.6m 上的实验结果如表 3 所示,本文方法主要对 10 种比较经典的三维姿态方法进行了对比实验,可以看出,本文提出的方法得到的身体各个部位估计平均欧式距离值均低于几个经典方法,平均欧式距离达到 41.6 mm,特别地,相比较 CHENG 等[4]提出的方法提升了 6%。

数据集 KTH Multiview Football Ⅱ提供同步图像和运动捕获数据,并且是三维人体姿态估计的标准基准集,为了测试本文方法,进行室外人体姿态估

计，如表 2 所示，本文的方法也要优于几个经典基准方法[9-13]，相较于最近人体姿态估计方法[9]，本文方法在人体所有姿势估计部位 PCP 分数提升了 3%，即使本文的算法是基于单 RGB 图像，而其他方法通常使用两台高速相机，由于该方法依赖于二维身体姿态检测器的三维骨架结构，当该方法在遇到人体外观信息较弱等情况时，造成深度信息丢失，相比之下，本文的方法能够在整流时空体积中收集图像外观时空信息和运动时序信息，因此，从总体来看，本文所用方法对平均欧式距离有比较大的提升，能够很好估计姿态特征，相比较下，总体性能良好。本实验提出的方法与最近的模型方法进行了对比实验，结果表明：即使在遇到运动模糊的情况下，本文方法也具有比较高的效率性和鲁棒性。

表 2　融合时空多特征方法与最先进的算法在数据集 KTH Multiview Football Ⅱ的比较

Tab. 2　Comparison results between the fusion spatio-temporal multi-feature method and the state-of-the-art algorithm in the dataset KTH Multiview Football Ⅱ

Method	Plevis	Troso	Upper arms	Lower arms	Upper legs	Lower legs	All parts
TEKIN, et al.[9]	99	100	74	49	98	77	79
BURENIUS, et al.[10]	97	90	53	28	88	82	69
AMIN, et al.[11]	92.4	91.1	75.4	72.9	76.7	47.7	80.5
PAVLAKOS, et al.[12]	—	—	80	64	85	79	77.0
BELAGIANNIS, et al.[13]	—	—	64	50	75	66	63.8
Ours	**101.2**	**98**	**84**	**54**	**95**	**76**	**81.4**

表 3　融合时空多特征方法与最先进的算法在数据集 Human3.6m 上的比较结果

Tab. 3　Comparison results between the fusion spatio-temporal multi-feature method and the state-of-the-art algorithm in the dataset Human3.6m

Method	Directions	Discussion	Eating	Greeting	Posing talk	Posing	Buying	Sitting	Sitting down	Avg
BOUAZIZI, et al.[1]	48.2	49.3	46.5	48.4	52.4	46.4	61.4	72.3	51.0	52.8
PAVLAKOS, et al.[2]	67.38	71.95	66.70	69.07	71.95	65.03	68.30	83.66	96.51	73.39
LUVIZON, et al.[3]	49.2	51.6	47.6	50.5	51.8	48.5	51.7	61.5	70.9	53.7
CHENG, et al.[4]	38.3	41.3	46.1	40.1	41.6	41.8	40.9	51.5	58.4	**44.4**
ZENG, et al.[14]	43.4	49.7	45.1	47.6	50.7	47.1	45.9	56.5	61.1	43.5
LI, et al.[15]	47.8	52.5	47.7	50.5	53.9	49.5	49.4	60.0	66.3	53.0
HU, et al.[16]	38.0	43.3	39.1	39.1	45.8	41.4	41.4	55.5	61.9	40.7
TOME, et al.[17]	64.98	73.47	76.82	86.43	86.28	110.67	68.93	110.19	173.91	94.63
TEKIN, et al.[18]	—	129.06	91.43	121.68	—	—	—	—	—	114.05
PAVLLO, et al.[19]	45.2	46.7	52.3	49.3	59.9	44.6	44.3	57.3	65.8	51.7
Ours	**37.2**	**39.3**	**45.4**	**38.5**	**41.3**	**40.3**	**38.9**	**39.2**	**54.4**	**41.6**

4.3　消融实验

为了评估本模型中运动补偿网络(MCNet)模块的有效性，本文在 Human3.6m 数据集上进行了消融实验，完整地移除运动补偿网络模块后测得每组的人体姿态估计的欧式距离以及平均欧式距离，实验结果如表 4 所示。结果表明，加入运动补偿网络模块之后，测得平均欧式距离降低了 2.9 个百分点，证明该模块确实能够对整体模型性能有提升结果。

表 4　消融 MCNet 模块对比表

Tab. 4　Ablation MCNet module comparison table

Pose	Remove McNet	Ours
Directions	37.2	**36.8**
Discussion	39.3	**38.4**
Eating	45.4	**44.5**
Greeting	38.5	**37.6**
Phone talk	41.3	**40.1**
Posing	40.3	**38.5**
Buying	38.9	**37.7**
Sitting	39.2	**37.2**
Sitting down	54.4	**52.9**
Average	**41.6**	**40.4**

4.4　消融实验对比可视化图

为了更好地对模型姿态估计预测方法进行解释,对模型定量分析之后,本文还使用类激活函数映射 CAM(class activation mapping)生成热力图,用于可视化分析模型中运动补偿模块所带来的运动模糊影响,如图 4 所示,图 4(a)中表示原帧带来的运动模糊,图 4(b)表示经过 MCNet 运动补偿网络后的热力图,通过热力图很清晰地看见在遇到相邻帧之间也能够很好学习光流表征信息,运动补偿网络充分利用了时空特征依赖,学习充分信息来提高表现能力。

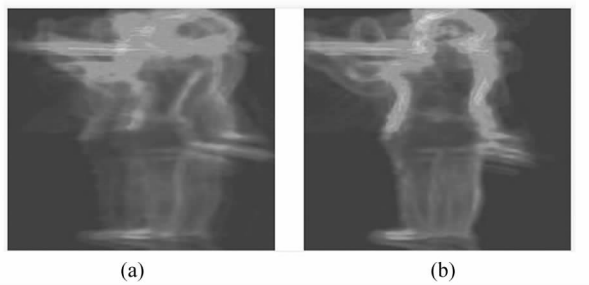

图 4　消融实验对比结果可视化热力图:
(a) 原帧运动模糊; (b) MCNet 运动补偿网络

Fig. 4　Ablation experiment comparison results visualization heatmap: (a) Original frame motion blur; (b) MCNet motion compensation network

4.5　实验结果图

本文的方法在 Human3.6m 数据集和 KTH-Multiview Football Ⅱ的结果分别如图 5、6 所示。

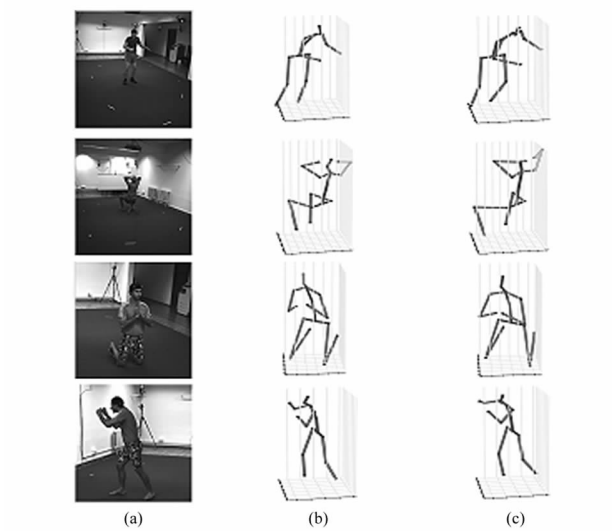

图 5　本文实验方法在数据集 Human3.6m 上进行实验得到三维姿态估计部分例子结果:
(a) 输入图片; (b) 预测结果; (c) 真实结果 GT

Fig. 5　The experimental method in this paper is tested on the dataset Human3.6m to obtain some example results of 3D pose estimation: (a) Input image; (b) Prediction result; (c) Real result GT

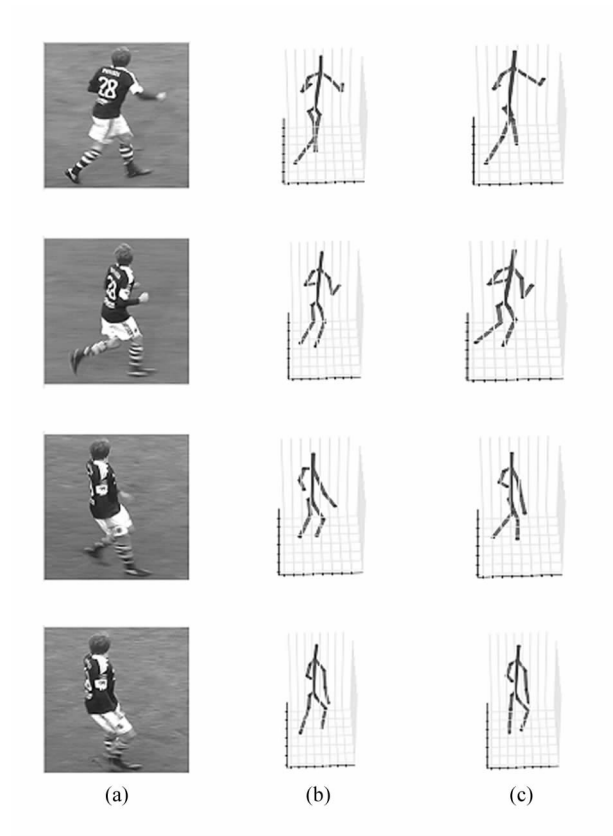

图 6　本文实验方法在数据集 KTH Multiview Football Ⅱ上进行实验得到三维姿态估计部分例子结果:
(a) 输入图片; (b) 预测结果; (c) 真实结果 GT

Fig. 6　The experimental method in this paper is tested on the dataset KTH Multiview Football Ⅱ to obtain some example results of 3D pose estimation: (a) Input image; (b) Prediction result; (c) Real result GT

5　结　论

针对目前三维人体姿态估计算法在图像和视频中表现在关节处效果不佳以及估计精度不高等问题,本文提出了一种利用时空信息融合运动补偿网络并且可以自动进行端对端学习架构,解决了准确性和实时性不平衡问题,用于单眼图像直接预测三维人体姿态估计。实验结果表明将时空多特征融合能够有效解决对人体姿态关节点位置之间的依赖关系,另外通过消融实验可以得出结论,本文方法在自遮挡带来的挑战性问题上取得较好的效果,在两个常见的三维人体姿态估计基准集,Human3.6m 数据集以及 KTH Multiview Football Ⅱ 数据集上与目前较为经典的方法相比较,本文方法能够较为明显提高准确性以及具有一定的通用性,可以用于其他类型的铰接式运动,因此该方法具有较好应用价值。在以后的工作中可以使用时序信息的经典方法,比

如 LSTM 和注意力机制应用到其他结构化时序视频位姿预测问题中，例如三维物体表面重建以及三维图像多视角渲染等工作。

参考文献：

[1] BOUAZIZI A, KRESSEL U, BELAGIANNIS V. Learning temporal 3D human pose estimation with pseudo-labels [C]//2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance, November 16-19, 2021, Virtual. New York: IEEE, 2021: 1-8.

[2] PAVLAKOS G, ZHOU X W, DERPANIS K G, et al. Coarse-to-fine volumetric prediction for single-image 3D human pose[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Hawaii, USA. New York: IEEE, 2017: 1263-1272.

[3] LUVIZON D C, PICARD D, TABIA H. 2D/3D pose estimation and action recognition using multitask deep learning [C]//IEEE Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake, USA. 2018, New York: IEEE, 2018: 5137-5146.

[4] CHENG Y, YANG B, WANG B, et al. Occlusion-aware networks for 3D human pose estimation in video[C]//IEEE Conference on Computer Vision and Pattern Recognition, October 27-November 02, 2019, Seoul, Korea (South). New York: IEEE, 2019: 723-732.

[5] CHO, S, LEE S. Fast motion deblurring[C]//ACM SIGGRAPH Asia 2009, December 16-19, 2009, Pacifico, Yokobama, Japan. New York: ACM, 2009: 1-8.

[6] SCHÖLKOPF B, HERBRICH R, SMOLA A J. A generalized representer theorem[C]//Computational Learning Theory, COLT 2011. Lecture Notes in Computer Science, vol 2111. Berlin: Springer, 2001: 416-426.

[7] BAO W, LAI W S, ZHANG X Y, et al. MEMC-Net: motion estimation and motion compensation driven neural network for video interpolation and enhancement[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43: 933-948

[8] LI F, BAI H H, ZHAO Y. Learning a deep dual attention network for video super-resolution[J]. IEEE Transactions on Image Processing, 2020, 29: 4474-4488.

[9] TEKIN B, MARQUEZ-NEILA P, SALZMANN M, et al. Learning to fuse 2D and 3D image cues for monocular body pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Hawaii, USA. New York: IEEE, 2017: 3941-3950.

[10] BURENIUS M, SULLIVAN J, CARLSSON S. 3D pictorial structures for multiple view articulated pose estimation [C]//IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, United States. New York: IEEE, 2013: 3618-3625.

[11] AMIN S, ANDRILU KA M, ROHRBACH M, et al. Multi-view pictorial structures for 3D human pose estimation[C]//Proceedings of the British Machine Vision Conference, September 9-13, 2013, Bristol, UK. York: BMVC Press, 2013: 45.1-45.12.

[12] PAVLAKOS G, ZHOU X W, DERPANIS K G, et al. Harvesting multiple views for marker-less 3D human pose annotations[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Hawaii, United States. New York: IEEE, 2017: 6988-6997.

[13] BELAGIANNIS V, AMIN S, ANDRILUKA M, et al. 3D pictorial structures for multiple human pose estimation[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 24-27, 2014, Ohio, United States. New York: IEEE, 2014: 1669-1676.

[14] ZENG A, SUN X, YANG L, et al. Learning skeletal graph neural networks for hard 3D pose estimation[C]//IEEE/CVF International Conference on Computer Vision, October 11-17, 2021, Montreal, Canada. New York: IEEE, 2021: 11436-11445.

[15] LI H, SHI B W, DAI W R, et al. Hierarchical graph networks for 3D human pose estimation[EB/OL]. (2021-11-23)[2022-02-22]. https: arxiv. org/abs/2111.11927.

[16] HU W, ZHANG C G, ZHAN F N, et al. Conditional directed graph convolution for 3d human pose estimation[C]//Proceedings of the 29th ACM International Conference on Multimedia, October 10-14, 2021, Lisboa, Portugal. New York: ACM, 2021: 602-611.

[17] TOME D, RUSSELL C, AGAPITO L. Lifting from the deep: convolutional 3D pose estimation from a single image [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Hawaii, USA. New York: IEEE, 2017: 2500-2509.

[18] TEKIN B, KATIRCIOGLU I, SALZMANN M, et al. Structured prediction of 3D human pose with deep neural networks[C]//Proceedings of the British Machine Vision Conference, September 19-22, 2016, York, UK. York: BMVC Press, 2016: 130.1-130.11.

[19] PAVLLO D, FEICHTENHOFER C, GRANGIER D, et al. 3D human pose estimation in video with temporal convolutions and semi-supervised training[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, California, United States. New York: IEEE, 2019: 7753-7762.

作者简介：

张　云　(1963—)，男，博士，教授，硕士生导师，主要从事智能系统开发以及图像处理方面的研究.