

硕 士 研 究 生 读 书 报 告



题目 基于骨骼一致性的单目视频人体运动重建

作者姓名 王一平

作者学号 22251076

指导教师 李启雷

学科专业 电子信息-软件工程

所在学院 软件学院

提交日期 2022年12月

MOMotioNet: 3D Human Motion Reconstruction from Monocular Video with Skeleton Consistency

MINGYI SHI, Shandong University and AICFVE, Beijing Film Academy

KFIR ABERMAN, AICFVE, Beijing Film Academy and Tel-Aviv University

ANDREAS ARISTIDOU, University of Cyprus and RISE Research Centre

TAKU KOMURA, Edinburgh University

DANI LISCHINSKI, Shandong University, The Hebrew University of Jerusalem, and AICFVE, Beijing Film Academy

DANIEL COHEN-OR, Tel-Aviv University and AICFVE, Beijing Film Academy

BAOQUAN CHEN, CFCS, Peking University and AICFVE, Beijing Film Academy

摘要

本文介绍了MotioNet,一种深度神经网络,它可以直接从单目视频中重建 三维人体骨骼的运动.虽然以前的方法依赖于索具或反向运动学(IK)来 将一致的骨架与时间相干的关节旋转关联起来,但我们的方法是第一个直 接输出运动学骨架的数据驱动方法,这是一个完整的,常用的运动表示. 我们方法的关键在于一个嵌入运动学先验的深度神经网络,它将二维关节位置序列分解为两个独立的属性：单一,对称的骨骼,以及与全局根位置 和脚接触标签相关的3D关节旋转序列.这些属性被输入到一个集成的正向运动学(FK)层中,该层输出三维位置,并与地面真相进行比较.此外, 对恢复旋转的速度应用了对抗性损失,以确保它们位于自然关节旋转的流 形上.我们的方法的关键优点是,它学会了直接从训练数据中推断自然的 关节旋转,而不是假设一个底层模型,或者使用与数据无关的IK求解器从 关节位置推断它们.我们证明,强制执行单一一致的骨架以及时间相干关 节旋转约束解空间,导致自遮挡和深度模糊的更稳健处理.

**关键词**：计算方法→运动处理，神经网络，姿态估计,运动捕捉,运动分析

Abstract

We introduce MotioNet, a deep neural network that directly reconstructs the motion of a 3D human skeleton from monocular video. While previous methods rely on either rigging or inverse kinematics (IK) to associate a consistent skeleton with temporally coherent joint rotations, our method is the first data-driven approach that directly outputs a kinematic skeleton, which is a complete, commonly used, motion representation. At the crux of our approach lies a deep neural network with embedded kinematic priors, which decomposes sequences of 2D joint positions into two separate attributes: a single, symmetric, skeleton, encoded by bone lengths, and a sequence of

3D joint rotations associated with global root positions and foot contact labels. These attributes are fed into an integrated forward kinematics (FK) layer that outputs 3D positions, which are compared to a ground truth. In addition, an adversarial loss is applied to the velocities of the recovered rotations, to ensure that they lie on the manifold of natural joint rotations. The key advantage of our approach is that it learns to infer natural joint rotations directly from the training data, rather than assuming an underlying model, or inferring them from joint positions using a data-agnostic IK solver. We show that enforcing a single consistent skeleton along with temporally

coherent joint rotations constrains the solution space, leading to a more

robust handling of self-occlusions and depth ambiguities.

**Keywords：**Computing methodologies→Motion processing,Neural networks,Pose estimation, motion capturing, motion analysis

1 介绍

长期以来，捕捉人类的运动一直是一项基本任务，在数据驱动的计算机动画、特效、游戏、活动识别和行为分析中有着广泛的应用。运动在可控的设置中使用专门的硬件，如磁跟踪器、深度传感器或多相机光学系统。近年来被广泛研究的另一种方法是从普通的单眼RGB视频中进行姿态估计和三维运动重建。

从单目视频的运动捕获提供了许多优点，如一个简单的不受控制的设置，低成本，和一个非侵入性的捕获过程。虽然由于深度模糊和遮挡，三维人体姿态估计具有很高的挑战性，但近年来基于数据驱动学习的方法取得了重大进展。这些方法利用深度神经网络来学习关于预期运动的强先验，这可以显著地帮助消除歧义和完成缺失的数据。

给定一个人类运动的录像，我们的最终目标是在三维空间中重建运动。现有的一种方法从视频中提取一系列三维姿态序列，其中每个姿态由每个关节的三维位置指定。然而，虽然结果表示可能满足某些应用程序，但它是不完整的。特别是，它并不包含驱动一个被操纵和剥皮的虚拟3D角色所必需的所有信息，而且骨骼的骨骼长度的时间一致性也不能得到保证。虽然关节旋转可以通过逆运动学（IK）从关节位置恢复，但该解决方案通常不是唯一的，如图2所示。此外，对每帧姿态估计强制执行软时间相干性约束可能不能确保骨架几何形状在所有帧中保持不变，并可能导致不自然的运动。一个深度神经网络，训练后从一个普通的单眼视频中重建单个表演者的运动（图1）。我们的网络不是推断一系列三维关节位置序列，而是学习提取一个应用于单个三维骨骼的三维关节旋转序列。因此，IK被有效地集成在网络中，因此，是数据驱动的（学习的）。强制执行单一骨架和时间相干的关节旋转不仅限制了解空间，确保了一致性，而且还导致了对自遮挡和深度模糊的更稳健的处理。

为了训练我们的网络，我们利用现有的数据集，包含准确捕获的完整的3D人体运动。序列的三维姿态投影到2D，和网络学习分解产生的2d关节位置序列成两个独立的属性：一个，对称的，骨骼，编码的骨长度，定义一个几何不变沿着整个序列，和一系列3d关节旋转序列，捕捉运动的动态方面。将三维骨骼和关节旋转输入一个集成的正向运动学（FK）层，该层沿着骨骼层次连续应用旋转，重建原始的三维运动序列。除此之外，我们的网络还预测了根关节的全局位置序列，以及脚的接触标签，因为后者的感知重要性。网络损失是考虑三维骨骼的骨骼长度、根全局位置和脚接触标签以及FK层恢复的关节位置的术语的组合。当这些属性与地面真实的三维运动进行比较时，联合旋转是使用对抗性损失来学习的，这鼓励了它们的速度有一个自然旋转的分布。此外，为了减少脚滑冰制品的影响，我们增加了一个脚接触损失，以鼓励每只脚的速度在它应该与地面接触的帧中为零。我们的方法的一个关键优点是，它不需要一个IK步骤，这是与数据无关的，并假设了一个潜在的受约束的模型。相反，该任务被整合到网络中，网络学习直接从真实人体运动的训练数据中推断联合旋转，而不是解决它们。此外，由于我们的系统表示在时间卷积滤波器空间中的运动，学习到的运动自然是平滑的。所有这些都导致了通过FK层进行更多数据驱动的人体运动重建。

为了弥补训练数据和视频之间的差距，我们注入联合位置噪声训练输入序列和增加置信值的分布模拟置信值提取[曹et al. 2018]从各种真实的视频。这个增强步骤在解的空间中构成了一个正则化器，提高了结果的稳定性，并提高了对遮挡的鲁棒性。我们进行了一系列广泛的实验和消融研究，以研究我们的系统及其不同组件的性能，证明了我们的端到端、完全数据驱动的单眼运动提取方法的质量和稳定性。

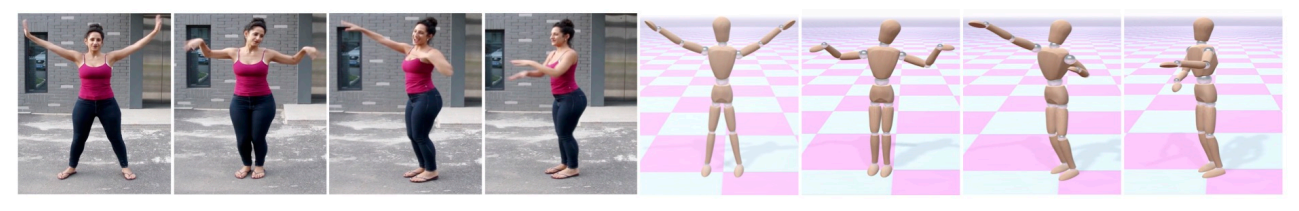


图1.给定一个表演者的单目视频，我们的方法，MotioNet，重建了一个完整的运动表示，包括一个单一的对称的骨架，以及一系列的全局根位置和3D关节旋转。因此，逆运动学被有效地集成在网络中，并且是由数据驱动的，而不是基于一个通用的先验。右边的图像是经过一个简单的操纵过程后从我们系统的输出渲染的。

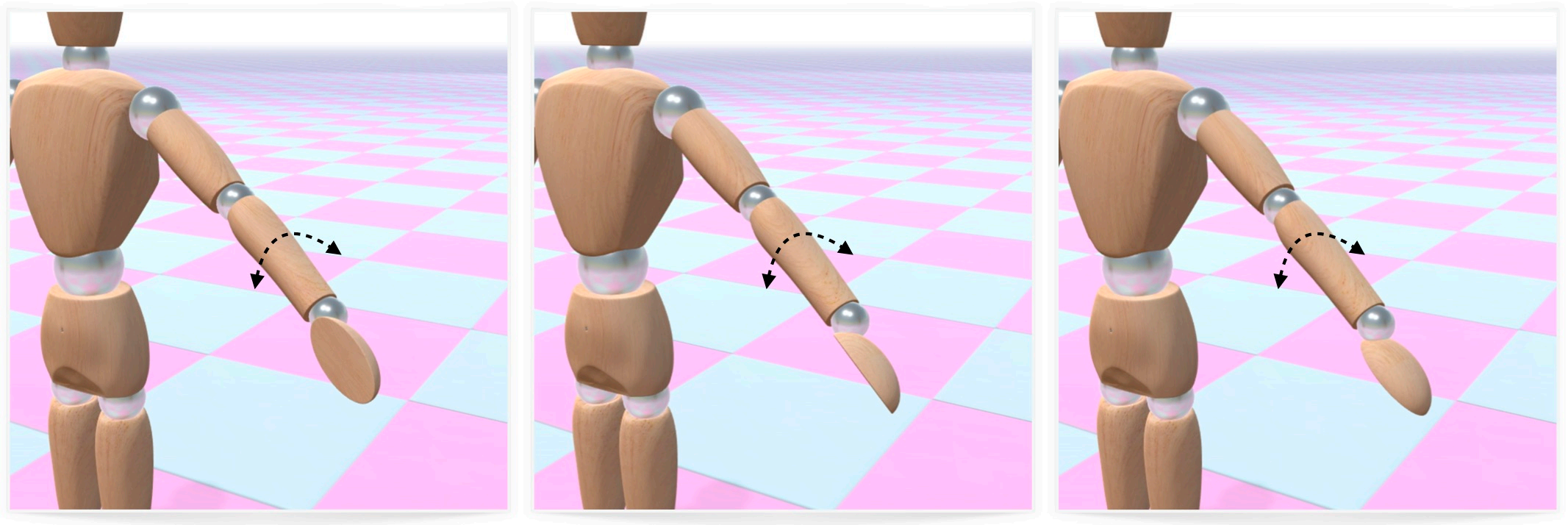


图2.关节旋转的模糊性。给定一组固定的三维关节位置，多个肢体旋转可以连接每一对连续的关节。因此，仅凭恢复的3D关节位置不足以驱动一个被操纵和剥皮的虚拟3D角色。

**2 相关工作**

深度学习时代之前，各种工作执行3D人体姿态估计的一组约束如2D关键点[Grochow等. 2004]，或通过拟合骨架基于体素的3D模型提取的视频，例如，[刘等2013；Vlasic等2008]。后一种方法大致分为生成方法，通过拟合模板模型重建人体姿态[魏等2012；Ye和Yang 2014]；判别方法，聚类像素假设身体关节位置，拟合模型，然后跟踪骨骼[夏普2012；Shotton等人2013]；混合方法，结合上述技术达到更高的精度[Baak等人2011]。关于三维姿态估计方法的详细概述，请参见Sarafianos等人[2016]。传统的方法大致可以分为基于深度学习的三维关节位置重建和三维形状恢复两种。

与这些方法不同的是，我们利用了一个单一的网络，支持三维关节旋转的直接回归，同时确保了骨架的时间一致性，最终解决了这两个问题。更具体地说，我们没有拟合骨架来检测三维人体姿态，而是将运动学集成到神经网络中，类似于[金泽等2018；Pavllo等2018；维legas等2018；Zhou等2016]，通过直接学习旋转数据来恢复关节角度。然而，周等人只在关节位置使用损失，关节角度估计为副产品，而金泽等人恢复关节旋转角度作为参数的一部分，SMPL身体模型使用非配对对抗性训练，为每个关节有一个单独的鉴别器。另一方面，Pavllo等人[2018]通过回归关节旋转，应用正向运动学，然后采用关节位置的损失来关注人体运动预测。相比之下，我们的网络学习使用可用的运动捕捉数据集的训练数据来输出具有自然速度的时间相干关节角，并在运动序列上进行训练，而不是单个帧。此外，我们的网络结构强制执行了一个单一的一致性骨架，它可以直接转换为动画文件，而不需要应用IK或时间平滑[Arnab等人，2019；Peng等人，2018]。

**3 运动重建**

我们的方法的关键在于一个深度神经网络，MotioNet，其独特的结构是受到了动画领域常见的工作方法的启发。MotioNet能够使用自然的、数据驱动的旋转信息来提取3D姿态，这可以直接转换为动画平台和游戏引擎中使用的运动文件格式。

在实践中，我们的网络学习将从输入视频中提取的二维关节位置映射成两个独立的组件： (i)一个单一的、对称的骨骼长度表示；（ii）一个动态的、骨骼独立的序列：关节旋转，全局根位置和脚接触标签。这两个部分构成了一个完整的、全局的、对运动的描述，不需要进一步的处理，或IK，以获得一个完整的3D动画。

为了训练我们的网络，我们利用运动捕捉数据集，该数据集包括由不同受试者执行的各种运动的时间序列，其中每个运动由单个骨架和一个三维关节旋转和全局位置的时间序列表示。在训练过程中，将三维运动从任意视角的角度投影到二维平面上，并训练网络从投影序列中恢复上述分量。通过检查足关节与地面的距离和速度，从捕获的运动中提取脚接触标签。在测试期间，预测的标签可以通过IK优化来约束关节的位置。

我们的主要技术贡献是一个新的2分支网络，它重建了运动的动态特性——关节旋转、全局根位置、脚接触标签，独立于静态的——一个单一的运动学骨架。我们的网络由各种组件组成，其中一些是以前由其他方法提出的，而另一些则是新的。我们使用的正向运动学层最初是由Villegas等人[2018]在运动重定向的背景下提出的。然而，我们是第一个使用它来进行姿态估计的人。此外，金泽等人[2018]已经对关节旋转角度应用了鉴别器，但以每帧的方式和使用绝对角度值。相比之下，我们的鉴别器判断了角速度的时间序列的实在性。最后，Pavllo等人[2019]也在三维姿态估计的环境中使用了一维时间卷积。然而，他们的网络被训练将二维关节位置提升到3D，而我们的网络被训练将关节位置转换为旋转。

3.1网络结构

我们的方法的一个高级关系图如图3所示。在训练过程中，每个数据样本C（Ps，q，r；c）∈RT×2J被并行输入两个网络，ES和EQ。ES的任务是估计单个骨架

1672409193553(1)

而EQ的目标是估计旋转，全局根位置和脚接触标签，

1672409243908(1)

我们训练一个单一的网络来估计所有的三个动态属性，因为它们之间的强相关性（例如，一个跑步的人的整体根位置，与四肢的旋转和与地面接触的时间时间相关）。由于旋转、全局位置和接触标签是动态的（每帧变化），我们设计EQ来保留时间信息。在实践中，它在时间轴上使用一维卷积层，产生一组关节旋转的时间集。相比之下，骨架是一个静态属性，ES使用自适应池来折叠时间轴，从而产生一个固定大小的骨长度向量，而不管输入的尺寸，如图5所示。

然后，在每个时间步t，将FK层应用于˜s，利用第t帧的联合旋转，得到

1672409415644

最后，给定三维骨架的局部位姿，我们加入全局平移，得到在帧t中完整估计的三维位姿，

1672409455142(1)

由于FK由微分算子组成，它可以作为一个层集成到网络中，通过它可以应用反向传播。给定时间步长的FK过程如图4所示。

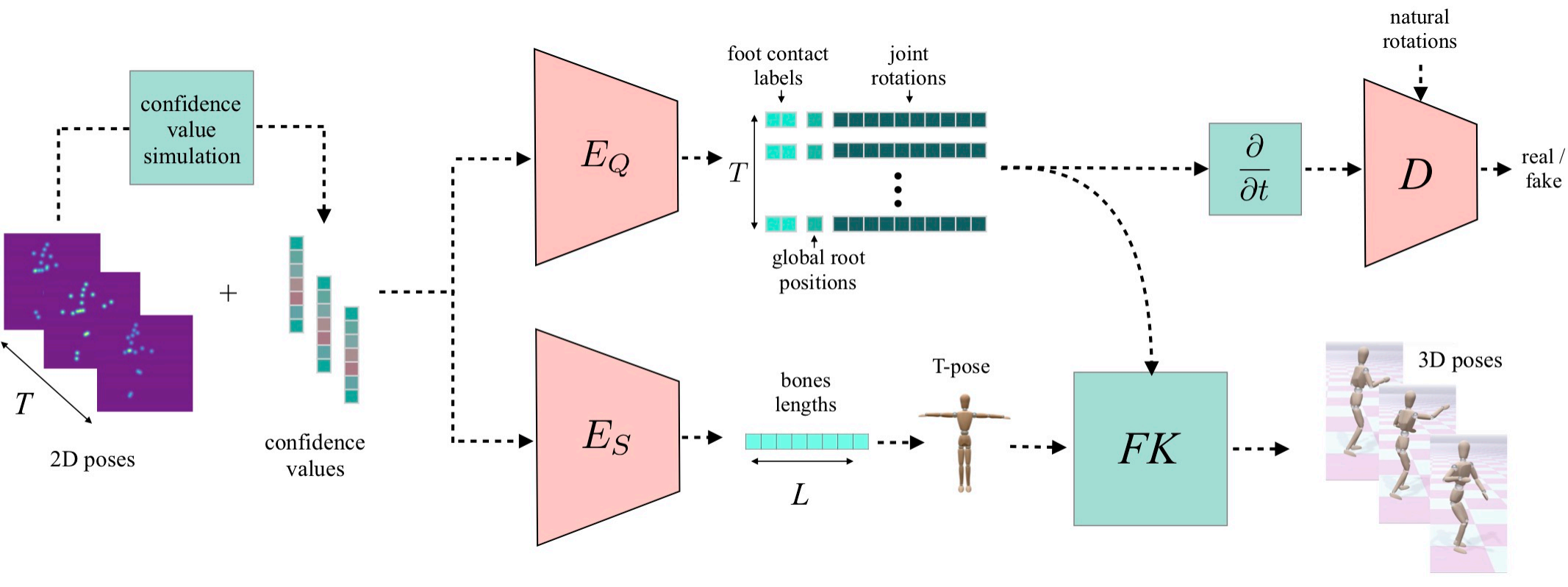


图3.我们的框架接收二维关节位置以及每个关节的置信值，这是基于真实视频的经验实验模拟的。它提取每帧的关节旋转和全局根位置，以及脚的接触标签，以及一个静态的（与持续时间无关的）骨架，使用两个编码器，EQ和ES。提取的旋转被输入一个鉴别器D，该鉴别器被训练来调整旋转角度的时间差异，以使用对抗性的训练来模拟自然旋转的分布。此外，旋转和转换为“t-pose”的静态特征被输入正向运动层FK，提取三维关节位置，并将其与地面真相进行比较。

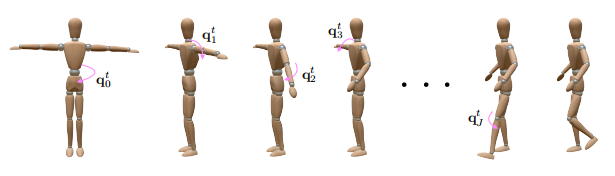


图4.我们的网络应用于“t-pose”骨架，通过连续旋转肢体从根到末端执行器。

3.2 训练与损失

为了训练我们的网络，我们利用了在第3.3节中更详细地描述的公开可用的运动捕捉数据集。我们使用了完全自我监督和部分自我监督训练的组合。对于完全监督训练，我们使用捕获的人体运动，其中完整的、旋转和位置三维关节数据可用，而部分监督训练只使用三维关节位置（位置表示）。在这两种情况下，二维输入都是通过投影三维训练数据自动生成的，因此成对的训练本质上是自我监督的。

我们使用的损失函数由四个组成部分组成：骨架损失LS、全局根位置损失LR、关节位置损失LP和旋转GAN损失LQ\_GAN。下面将更详细地描述这些组件。骨骼损失。设Ps，q，r∈P表示我们数据集中P中的三维运动序列，C（Ps，q，r，ci）表示使用摄像机的二维投影。骨架丢失LS确保编码器ES正确提取骨架∈：

1672409675738(1)

当使用透视投影时，三维骨架只能重建到一个比例因子。为了避免这种尺度模糊性，我们对数据中的所有三维骨骼进行了全局重新缩放，使它们的平均骨长度等于1。在推理时，如果执行者的真实脊柱长度已知，重建的三维姿态可以重新调整到一个真实的度量空间。

3.2 运动训练集

我们的训练数据来自CMU运动捕捉数据集[CMU 2019]和人类3.6M人体姿态数据集[Ionescu et al. 2014]。CMU数据集由144名受试者执行的2605个捕获动作组成。动作包括基本动作（走、跳、踢等），以及各种舞蹈动作（桑巴舞、萨尔萨舞等）。人类360万人的数据集包括超过360万个3D人体姿态和相应的图像，使用同步摄像机从四个不同的角度观看。这些动作由11位演员表演，包括17个日常场景，如吃饭、吸烟、打电话、散步等。这两个数据集中的参与者都具有不同的体型和骨骼比例。

为了训练网络以有监督的方式提取脚接触标签，我们必须首先从数据集中提取“地面真实”标签。具体来说，如果一个脚离地面的高度低于20毫米，并且它在周围的5帧窗口中的平均速度幅度低于1毫米/帧，我们认为它是接触的。地面的高度近似于序列中所有脚高度的最低20百分位数的平均值。

通过将三维关节位置序列（以摄像机坐标空间表示）投影到不同的摄像机视角上，生成输入到网络中的实际输入数据样本。因此，我们获得了一个丰富的具有地面真实值的数据集，它展示了不同角色、执行不同运动的骨骼是如何从不同的视角中出现的。

在测试时间中，当使用野外视频时，我们使用Cao等人[2018]的方法提取16个关节（根、颈、胸、头、肩、髋、膝盖、脚、肘部、手）。这些关节也存在于上述的数据集中。此外，人工添加了一个仅存在于三维数据集中的脊柱关节（计算为根关节和颈关节之间的二维空间平均值）。

为了使二维输入数据规范化，我们应用两个操作： (i)从每个关节位置减去根位置（在所有帧中）减去根位置，（ii）减去平均关节位置并除以标准偏差，而平均和标准偏差是整个数据集上每个关节的经验计算。由于从归一化的局部表示中丢弃了全局信息，因此我们在其中附加了全局二维速度（每帧）。

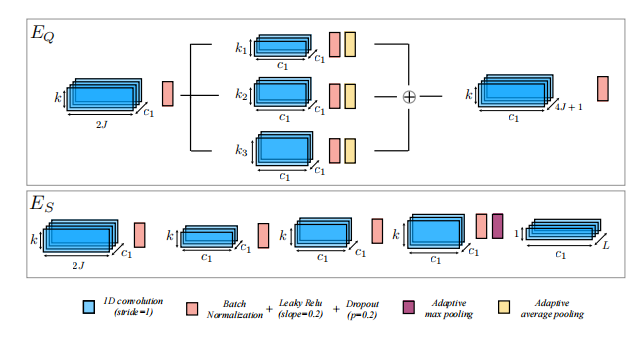


图5.我们的网络包含两个编码器，EQ使用并行卷积生成一组时间内的关节旋转、全局位置和脚接触标签，ES输出一个表示骨架的静态属性，使用一个自适应池化层来折叠时间轴。

3.4 实现细节

在实践中，我们的实现包括两个编码器，EQ和ES，一个正向运动学层FK和一个鉴别器d。两个编码器的层和尺寸如图5所示，而FK层由Villegas et al. [2018]描述。

EQ编码器中，有五个一维卷积层。第一层扩展了通道的数量，然后将结果并行输入三个不同大小的卷积层。这是受谷歌net的启发。在自适应池化后，将这三个结果相加，并输入最后一层，生成四元数和根深度的时间序列。ES编码器使用了一个简单的一维卷积层序列。同样，第一层扩展了通道的数量，最后一层输出一个骨长度的向量，这隐式地加强了对称性，因为每对对称的肢体都是由相同的坐标编码的。

鉴别器D也使用了一维卷积层序列。有一个每个关节的卷积层，它被总结为一个完全连接的层，它决定了旋转速度是真实的还是假的

**4 野外支撑视频**

在训练过程中，我们的网络的输入是通过将大多数干净的运动捕获的数据从三维投影到一个二维图像平面而产生的。然而，在测试时，网络应该能够使用OpenPose库处理从野外视频中提取的2D联合位置[Cao et al. 2018]。为了克服训练输入分布和测试时间分布之间的差距，我们采用置信值和增强。

*置信值。*在测试期间，当从野外视频中提取二维关节位置时，可能会由于遮挡而出现位置错误，甚至关节缺失。为了在训练过程中模拟这种输入特征，我们利用了最先进的二维姿态估计器提供的输出结构，并为每个关节附加了一个值cn∈[0,1]，它指定了该关节的二维位置估计的置信度。对于一个完全缺失的关节，置信度cn被设置为0。

*增强。*为了丰富观察到的样本，我们采用两种不同的方式进行数据增强： (1)剪辑长度：由于我们的模型是完全卷积的，不同长度的输入运动剪辑可以通过同一网络转发。然而，由于使用了批次，一个固定的时间长度可以提高训练时间。因此，我们随机选取批中样本的时间长度(T)，每次迭代，为60到200范围内的整数。这一增强步骤增强了静态参数的解纠缠和序列的时间长度。(2)摄像机增强：由于我们在三维摄像机坐标空间中工作，我们的投影通过修改三维特征的深度（沿z轴的全局平移）和它的方向（根的全局旋转）来增强。这个操作很重要，因为它训练网络在不同的二维尺度下将相似的姿态映射到相同的（局部）3D中参数（旋转和骨骼长度），从而扩展了训练数据输入的分布。

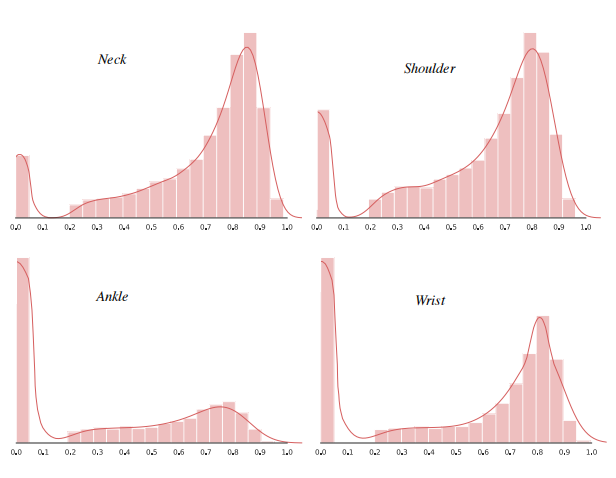


图6.使用从野外视频中提取的置信值（箱子）的经验分布对联合置信值（连续红线）对联合置信值的分布进行建模。

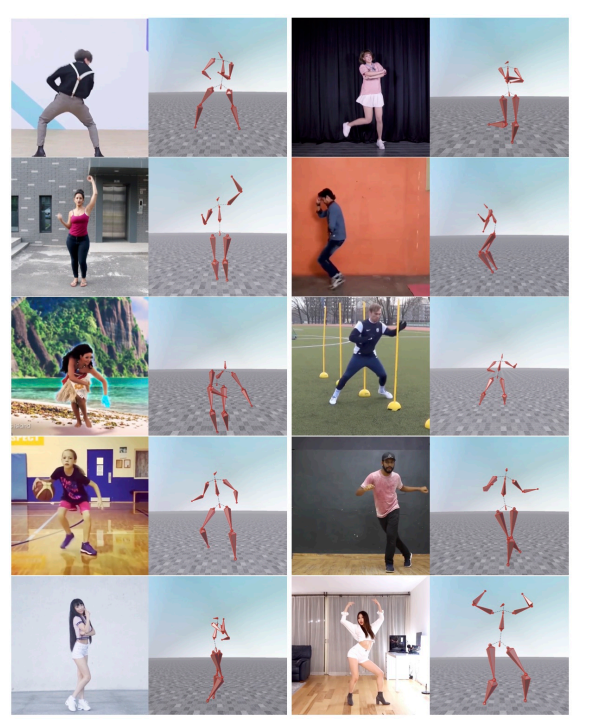


图7.通过野外视频重建运动的例子。MotioNet可以重建各种人体运动，同时具有稳健的部分闭塞和容忍不同的身体类型。视频中的帧在左边，重建的姿势在右边。

**5 结果及评价**

在本节中，我们展示了由MotioNet获得的三维人体运动重建结果，进行了实验和消融研究，以分析我们的框架中各种组件的性能，并与最先进的三维人体运动估计技术进行了比较。图7显示了一个结果库，展示了MotioNet能够在野外从视频中提取各种人类运动，同时对部分遮挡具有鲁棒性，对不同的身体类型具有耐受性。为了制作动画，我们将我们的网络产生的张量转换成BVH文件，并使用搅拌机软件来读取这些文件和渲染视频，而不需要任何进一步的后期处理。结果动画的质量可以在补充视频中进行检查。

5.1 Comparison to Other Methods

我们首先比较了用我们的方法与各种现有的方法重建关节位置的准确性。

我们在两个数据集上评估了我们的方法。第一个是人类360万数据集[Ionescu et al. 2014]，它采用了17个关节的骨骼表示。对于这个数据集，网络在5个受试者（S1、S5、S6、S7、S8）上进行训练，并在两个受试者（S9和S11）上进行测试，使用一个训练所有动作的单一模型。第二个数据集是HumanEva-I [Sigal et al. 2009]，它采用了15个关节的骨骼表示，包括三个不同的主题。我们通过训练一个模型，通过两个动作（行走、慢跑）来评估我们的系统。注意，与其他方法类似，该数据集的评估是在与地面真相在内的尺度、旋转和平移进行刚性对齐后计算的（a.k.a P-MPJPE）。

为了应用一个公平的比较，我们没有模拟置信图，而是在训练和测试期间使用2D检测到的关键点。我们跟踪Pavllo等人，[2019]，并使用级联金字塔网络（CPN）[Chen等人，2018b]来提取2d姿态。CPN所需要的边界框由Mask-RCNN提取，二维探测器由Human3.6M的二维地面真实样本进行微调。从表1和表2的结果可以看出，虽然我们的方法在关节位置上没有达到最高的精度，但它通常是前2-3种方法之一。尽管我们的方法针对的是一个略有不同的任务。此外，请注意，注意的是，由于估计的骨长度不准确（甚至是小的），强制执行单一的底层骨骼可能会在重建的关节位置引入一些错误。为了测量旋转分支的有效性，消除骨架重建中误差的影响，我们进行了一个使用地面真实（GT）骨架而不是重建骨架的实验。本实验只接受EQ训练，而FK接受GT骨架训练。表1中报告的结果表明，如果已知确切的骨架，姿态的精度可以更高。

为了证明我们的动态编码器EQ的有效性，我们用Pavllo等人[2019]提出的扩展的一维卷积体系结构取代了其结构，最初设计用于输出位置。表1中还描述了证明我们的架构更适合旋转预测的结果。



表1.与在人类3.6M数据集上使用MPJPE的其他关节位置重建方法的定量比较。图例：（）多帧作为输入，（+）额外的训练数据(Yang等人[2018]使用来自MPII数据集的2D注释。Pavlakos等人[2018a]使用了来自利兹体育姿势（LSP）数据集的额外数据)。

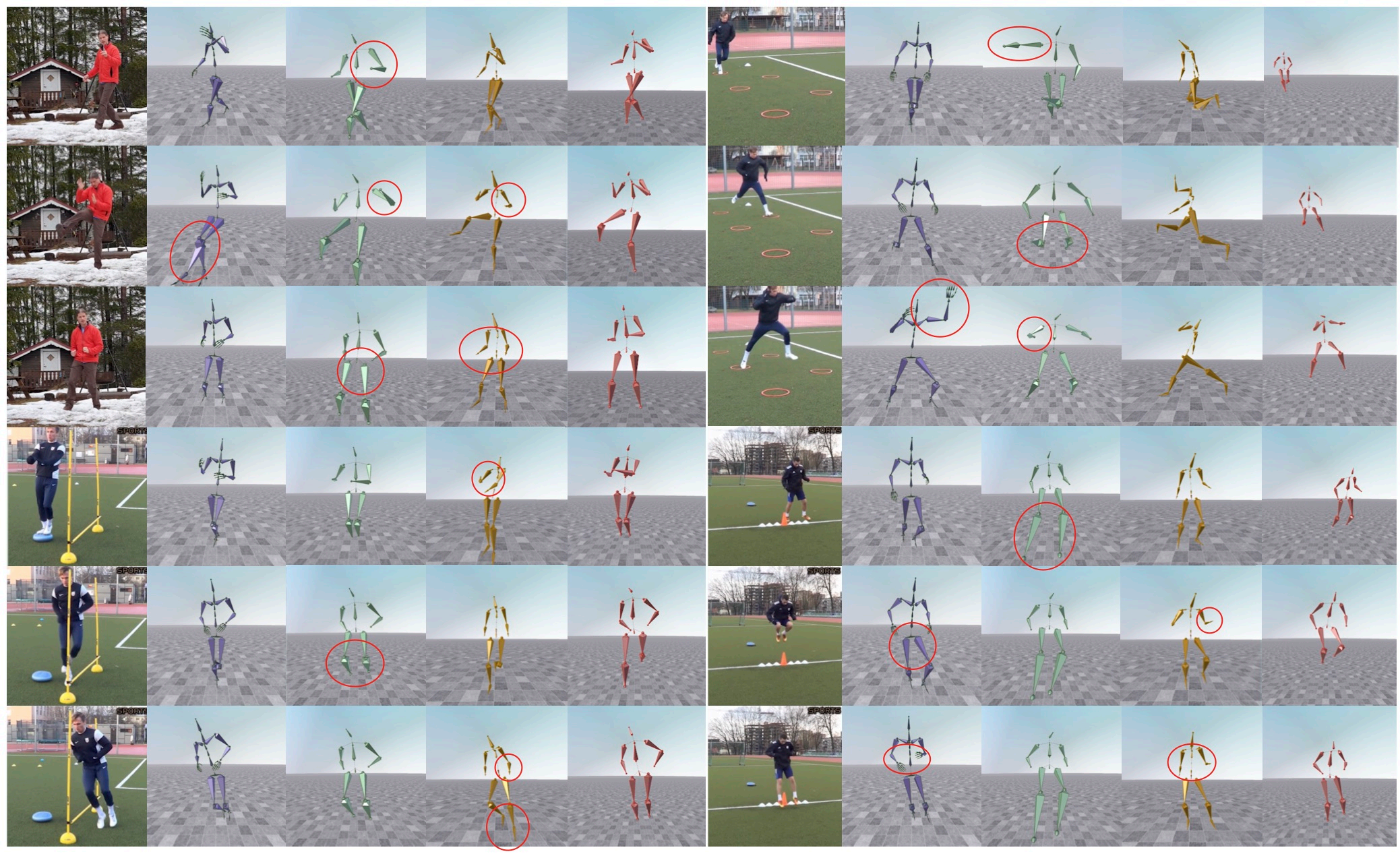


图8.与其他三种方法相比，我们对野外视频的研究结果进行了比较： (i)金泽等人[2018]（ii）Mehta等人[2017b][iii)Pavllo等人[2019] + Dong等人[2017]。请看补充视频，在那里可以看到，我们的方法恢复了合理的全局运动与平滑的关节旋转。

5.2 消融研究

在本节中，我们将评估和演示在我们的框架中的不同组件和损失的有效性。

*模拟置信度图和增强图。*在这个实验中，我们比较了两种不同的设置的结果：有和没有模拟置信图。为了测量结果，我们使用了人类3.6M数据集，该数据集还包含运动捕获的参与者的视频帧。这些框架能够模拟由二维位置提取（而不是二维投影）组成的完整管道，并根据地面真相来评估结果。我们从测试数据中使用239个原始视频，使用[Cao et al. 2018]提取2D关节位置，对GT评估3D输出，并将结果报告在表2中。可以看出，虽然置信图考虑降低了干净数据的结果，但它提高了从视频中提取的二维关节的结果。一个定性的比较可以在补充的视频中找到。可以观察到，在没有模拟置信图的情况下，对干净的投影数据进行训练，会导致轻微的抖动，这是由从视频帧中提取的噪声输入引起的，而没有出现在训练数据中。此外，我们进行了一个没有任何数据增强和一个没有t姿态损失的实验，并将结果报告在表2中，这展示了每一步的贡献。

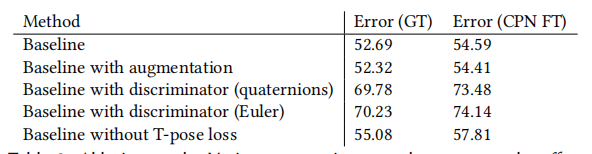


表2.消融研究。各种比较，以证明我们的框架中不同组件和损失与基线的有效性。第二列描述了由干净的二维输入（3D数据的投影）获得的误差，而第三列中的值来自于使用微调的CPN检测器从真实视频帧中提取的输入。

*旋转速度鉴别器。*在这里，我们使用了没有LQ\_GAN损失项的CMU测试数据集来评估运动重建的质量。一个定性的比较可以在补充的视频中找到。虽然由于时间卷积滤波器的存在，旋转在时间上是平滑的，但可以看出，在我们的框架中没有鉴别器可能会产生不自然的自旋转（特别是绕滚轴）和不一致的旋转速度。图11显示了在三种不同的设置中学习到的旋转的时间差异的分布：没有鉴别器，鉴别器应用于以单位四元数表示的旋转，鉴别器应用于以欧拉角表示的旋转。在最后的实验中，网络输出单位四元数，并对欧拉角实验进行了适当的转换。该图以欧拉角表示的方式描述了每轴的旋转速度的大小的数量。可以看出，对于这两种表示，速度的分布更接近原始数据的分布，即使输出旋转没有直接损失。此外，表3显示，基于四元数的鉴别器在定量上优于欧拉角实验，这是我们在其余实验中选择的表示。

*足部接触损失。*我们进行了一个实验，其中网络不考虑足接触预测和足接触损失，并在补充视频中显示了其结果。可以看出，当没有明确注意脚的接触时，可以很容易地观察到感知干扰的脚滑冰。特别是在我们的例子中，当FK以从根到末端执行器的连续顺序应用时，累积的误差使不稳定的脚更加明显。实验结果脚接触误差64.01与误差52.34脚接触损失，而误差计算之间的欧氏距离输出和GT脚位置只在帧有接触地面和平均随着时间。在补充视频中的定性评价表明，该预测有助于减轻足部滑动伪影。此外，如图9所示（对于我们的测试数据中的一些任意的运动槽），我们的实验表明，预测的脚标签（底部）的准确性与脚的位置（顶部）的误差之间有很高的相关性。尽管该网络没有使用预测的脚接触标签，但高接触预测精度可以产生更好的脚位置的精度。

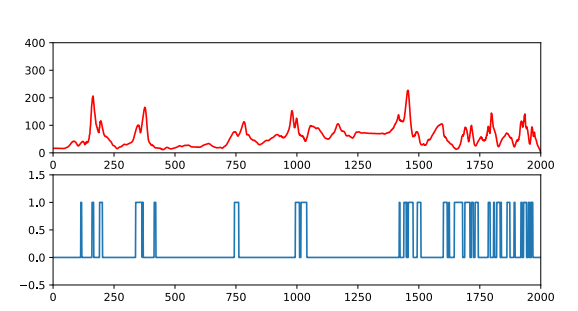


图9.足部接触标签的预测与足部位置误差相关。顶部：脚位置误差，底部：脚接触标签预测精度（0正确，1不正确）。

**6 结论、局限性和未来的工作**

我们引入了一种整体的方法，学习从单眼视频中提取一个完整的三维人体运动，而不是推断图12。足部接触标签的预测与足部位置误差相关。顶部：脚位置误差，底部：脚接触标签预测精度（0正确，1不正确）。一系列的人类姿势。使用标准的剥皮和后处理，类似于应用于在受控环境中捕获的数据，我们可以实现更高质量的动画，可以用于各种应用程序和游戏引擎，这可以在我们的补充视频的末尾看到。具体来说，虽然之前的工作只进行姿态估计，只重建三维关节位置，但我们的深度神经网络MotioNet被训练为恢复具有三维关节旋转序列的单个骨架，从而恢复关节位置。一般来说，关节的位置并不能完全捕捉到人的运动。仅由关节位置定义的人体姿势可能不满足固定的骨长度约束。即使使用了IK，歧义和时间上的不连续性仍然可能发生。相比之下，具有指定关节旋转的时间序列的固定骨架完全表征了运动；它是动画合成的原生表示，因为它通常组成角色装备空间的子空间。

我们的方法的一个关键优点是它不需要IK：关节位置和关节旋转都是从真实人体运动的训练数据中学习的，而不是解决。这导致了一个更自然的人类运动重建，通过图案网的FK层。此外，通过推断单个骨架，MotioNet保证保持骨骼长度沿整个运动序列固定。这提供了一个先验，促进运动重建存在明显的遮挡在输入视频。我们采用的另一种解决遮挡挑战的机制是向训练输入序列注入随机噪声，并增加其置信值，其分布类似于二维视频从野外姿态估计方法中提取的置信值。因此，该网络被训练来处理噪声和遮挡，类似于去噪自动编码器。最后，由于我们的系统是在运动空间中训练的，因此人类运动的固有平滑性从数据中学习出来，自然地实现了时间相干性。

根据设计，动机网只跟踪一个人类角色的运动。处理多角色的动作，特别是当角色经历亲密的互动时，显然是具有挑战性的，我们把它留给未来的工作。此外，我们目前的方法可能不能准确地恢复骨骼的全局定位，例如，当有显著的相机运动时。另一个限制是，我们的网络没有考虑任何人性的物理约束或性格之间的相互作用环境——运动重建纯粹是基于对视觉线索的分析和学习，这也可以被标榜为一种优势。因此，例如，在系统中没有接触约束的概念，因此由此产生的运动可能会受到足部滑动的影响。这可以通过预测足部接触条件和使用损失函数项来解决，如Lee等人[2018b]。

在其他可能的未来作品中，我们感兴趣的是从视频中学习人体运动的子空间，这样动画师就可以轻松地制作出物理上可信的动画。我们也有兴趣从目前公开的大量视频数据中学习运动风格之间的差异。带有运动样式注释的数据集对于创建生成模型非常有用，其中可以对新的运动样式进行采样。