基于时域卷积和半监督训练的视频三维人体姿态估计

Dario Pavllo* Christoph Feichtenhofer David Grangier* Michael Auli

ETH Zu¨rich Facebook AI Research Google Brain Facebook AI Research

摘要

在本文中，我们证明了基于二维关键点上扩展的时域卷积的全卷积模型可以有效地估计视频中的三维姿态。我们还介绍了反投影，一种简单有效的利用未标记视频数据的半监督训练方法。我们首先预测未标记视频的二维关键点，然后估计三维姿态，最后返回到输入的二维关键点。在监督环境下，我们的全卷积模型比文献中最好的结果在 Human3.6M 上平均每关节位置误差减少了 6 毫米，误差减少了 11 %，并且在 HumanEva-I 上也表现出了显著的改进。此外，反投影实验表明，在标记数据较少的半监督环境下，我们的全卷积模型比文献中最好的结果舒适。代码和模型可在 https://github.com/ facebook Research / VideoPose3D 上获得

1 引言

我们的工作集中在视频中的三维人体姿态估计。我们建立在最新方法的基础上，将问题描述为二维关键点检测和三维位姿估计[41，52，34，50，10，40，56，33]。虽然拆分问题可以说降低了任务的难度，但由于多个 3D 姿势可以映射到相同的 2D 关键点，这在本质上是模糊的。此前的工作通过使用递归神经网络建模时间信息来引导这种歧义[16，27]。另一方面，卷积网络在传统的 RNNs 引导的任务中对时间信息的建模非常成功，如神经机器翻译[11]、语言建模[7]、语音生成[55]、语音识别[6]等。

在本文中，我们提出了一种全卷积结构，它在二维关键点上进行时间卷积，以精确预测视频中的三维位姿(见图 1)。我们的方法兼容任何 2D 关键点检测器，可以通过扩展的卷积有效地处理大的上下文。相对于依赖 RNNs 的方法[16，27]，它在计算复杂度和参数个数方面都具有更高的精度、简单性和效率(§3)。

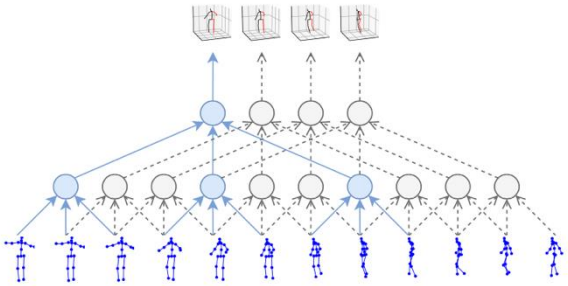


图 1：我们的时间卷积模型以二维关键点序列(下)为输入，以三维位姿估计为输出(上)，我们使用扩展的时间卷积来捕获长期信息。

配备了一个高精度和高效率的架构，我们转向有标记训练数据稀缺的设置，并引入一个新的方案来利用无标记视频数据进行半监督训练。对于需要大量标记训练数据的神经网络模型来说，低资源设置尤其具有挑战性，为 3D 人体姿态估计收集标签需要昂贵的运动捕捉设置以及冗长的记录会话。我们的方法受到了无监督机器翻译中循环一致性的启发，其中往返转译成中间语言和回到原语应该接近身份函数[46，26，9]。具体来说，我们用下架的 2D 关键点检测器预测未标记视频的 2D 关键点，预测 3D 姿态，然后将这些映射回 2D 空间(§4)。

综上所述，本文提供了两个主要贡献。首先，提出了一种简单有效的基于二维关键点轨迹上扩展时域卷积的视频三维人体姿态估计方法。我们表明，无论从计算复杂度还是模型参数个数方面，我们的模型都比基于 RNN 的模型在相同精度水平下的效率更高。

其次，我们引入了一种半监督的方法，它利用了未标记的视频，并且在标记数据稀缺时是有效的。与以往的半监督方法相比，我们只需要摄像机的内

在参数，而不需要地面真实的二维注释或带有外在摄像机参数的多视角图像。

与目前的研究状况相比，我们的方法在监督和半监督环境下都表现出了以前最好的表现方法。即使这些额外的标记数据用于训练，我们的监督模型的性能也优于其他模型。

2 相关工作

在深度学习取得成功之前，大多数三维位姿估计的方法都是基于特征工程和骨架和关节活动度的假设[48，42，20，18]。最早使用卷积神经网络(CNN)的神经方法主要是通过不需要中间监督的RGB 图像直接估计三维姿态来进行端到端的重构[28，53，51，41]。

两步位姿估计。一个新的 3D 位姿估计器族在 2D 位姿估计器的基础上，通过先预先确定图像空间中的 2D 联合位置(关键点)，然后将其提升到 3D[21，34，41，52，4，16]。这些方法优于端到端的方法，因为它们受益于中间监督。我们遵循这一方法。最近的研究表明，给定地面真实的 2D 关键点，预测 3D 位姿相对简单，其难点在于预测精确的 2D 位姿[34]，早期的方法[21，4]只需在可获得 3D 位姿的大量 2D 关键点上对预测的 2D 关键点进行 k 近邻搜索，然后简单输出相应的 3D 位姿。有些方法既利用了图像特征，又利用了二维的地面真实姿态[39，41，52]。或者，可以通过简单地预测 2D 关键点的深度来预测 3D 姿势[58]，有些工作强调骨骼长度和投影与 2D 地面真实度的一致性[2]。

视频位姿估计。大多数以前的工作都是在单帧环境下进行的，但最近人们努力利用视频中的时间信息，以产生更稳健的预测，并降低对噪声的敏感性。[53]从时空体的 HoG 特征(方向梯度直方图)推断三维姿态。LSTMs 已经被用于从单幅图像中预测的三维姿态的细化[30，24]。然而，最成功的方法是从 2D 关键点轨迹中学习。我们的工作属于这一类。

最近，LSTM 序列到序列学习模型被提出，它将视频中的 2D 姿态序列编码成一个固定大小的向量，然后解码成 3D 姿态序列[16]。然而，输入序列和输出序列的长度都是相同的，2D 姿态的确定性变换是更自然的选择。我们用 seq2seq 模型进行的实验表明，输出姿态往往会在长序列上漂移。[16]通过每 5 帧重新初始化编码器，以牺牲时间一致性来解决这个问题。还开展了 RNN 方法的工作，这些方法考虑了身体部位连接的先前因素[27]。

半监督训练。已有多任务网络[3]用于联合 2D 和 3D 位姿估计[36，33]以及动作识别[33]的研究，有的工作将 2D 位姿估计学习到的特征转移到 3D 任务中[35]，未标记的多视角记录被用于 3D 位姿估计的训练前表征[45]，但这些记录在无监督的环境中并不容易获得。生成对抗网络(GAN)能够在第二个只有二维标注的数据集上区分现实姿态和不现实姿态[56]，从而提供了一种有用的正则化形式。文献[54]利用 GANs 从未成对的 2D/3D 数据中学习，并包含 2D 投影一致性项。同样，[8]将生成的 3D 姿势随机投影到 2D 后，对其进行判别。[40]提出了一种基于有序深度标注的弱监督方法，利用深度比较增强的二维位姿数据集，例如“左腿在右腿后面”。

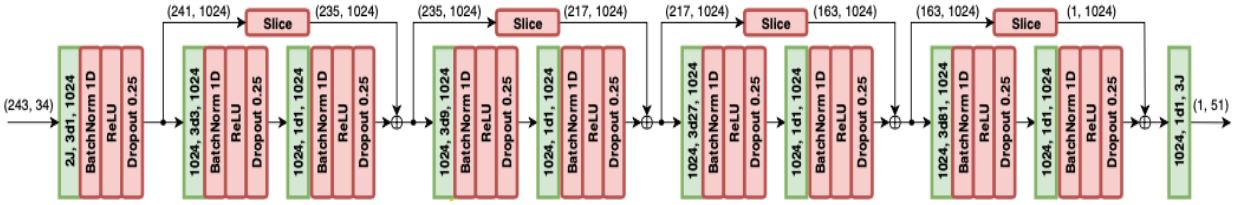
3D 形状恢复。虽然本文和所讨论的相关工作侧重于重建精确的三维姿态，但平行的研究路线旨在从图像中恢复人的完整三维形状[1，23]。这些方法是基于参数化三维网格的类型化方法，对位姿精度影响较小。

我们的工作。与[41，40]相比，我们不使用热图，而是使用检测到的关键点坐标来描述姿态。这允许在坐标时间序列上使用有效的一维卷积，而不是在单个热图上使用二维卷积(或在热图序列上使用三维卷积)。我们的方法也使得计算复杂度与关键点空间分辨率无关。我们的模型能以较少的参数达到较高的精度，并允许更快的训练和推理。与文献[34]提出的单帧基线和文献[16]提出的 LSTM 模型相比，我们通过在时间维上进行一维卷积来利用时间信息，并提出了一些优化方法，使得重建误差更低。与[16]不同，我们学习的是确定性映射，而不是 seq2seq 模型。最后，与本节提到的大多数两步模型(使用流行的堆叠沙漏网络[38]进行 2D 关键点检测)相反，我们表明 Mask R- CNN [12]和级联金字塔网络(CPN) [5]检测对 3D 人体位姿估计更加稳健。

3 时间扩展卷积模型

我们的模型是一个完全卷积结构，具有残馀连接，以 2D 姿态序列作为输入，通过时间卷积进行转换。卷积模型使批和时间维度上的并行化，而 RNNs 不能随时间而并行化。在卷积模型中，输

图 2：我们的全卷积三维位姿估计架构的实例化。输入由 243 帧(B = 4 块)、J = 17 个关节的重复场的二维关键点组成。卷积层为绿色，其中 2J，3d1，1024 表示 2·J 的输入通道，大小为 3 的核为膨胀 1，1024



表示输出通道。我们还给出了样本 1 帧预测的括号中的张量大小，其中(243，34)表示 243 帧和 34 个通道。由于有效卷积，我们对残差(左右，对称)进行切片，以匹配后续张量的形状。

出和输入之间的梯度路径具有固定的长度-较少的序列长度，这减少了影响 RNNs 的消失梯度和爆炸梯度。卷积结构还提供了对时间感知场的精确控制，我们发现它有利于为三维位姿估计任务建模时间依赖关系。此外，我们采用扩张型卷积[15]，在保持效率的同时，对长期支出进行建模。扩展卷积结构已经成功地用于音频生成[55]、语义分割[57]和机器翻译[22]。

输入层为每帧取 J 个关节的级联(x，y)坐标，并与核大小 W 和 C 输出通道进行时间卷积。随后，B ResNet 风格的块被跳过连接超圆[13]，每个块首先与核大小为 W，膨胀因子 D=WB 形成一维卷积，然后与核大小为 1 的卷积。卷积(最后一层除外)依次是批归一化[17]、校正线性单元[37]每个块的感受指数增加一个因子 W，而参数数目只线性增加。设置滤波器超参数 W 和 D，使任何输出帧的感受野形成复盖所有输入帧的树(见§1)。最后，最后一层利用过去和将来的数据输出对输入序列中所有帧的 3D 姿态的预测。

卷积图像模型通常应用零填充来获得与输入一样多的输出。但早期的实验在将输入序列与边界帧的副本填充到左边和右边时，只进行未填充卷积时，效果更好(见附录 A.5，图 9a)。

图 2 展示了我们的体系结构对于 B=4 块的 243 帧接受场大小的实例化。对于卷积层，我们设置 W=3，C=1024 个输出通道，我们使用一个漏码率 p=0.25。

4 半监督方法

我们引入半监督训练方法，在标记的三维地-真姿态数据的可用性有限的环境中，对精度进行 im 证明。我们利用 unla- beled 视频结合一个下架的 2D 关键点检测器，以一个反投影损失项来扩展监督损失函数。我们解决了一个无标记数据上的自编码问题：编码器(位姿估计器)从二维联合坐标执行三维位姿估计，解码器(投影层)将三维位姿投影回二维联合坐标。当来自解码器的二维联合坐标远离原始输入时，训练会受到惩罚。

图 3 表示我们的方法，它将我们的监督组件和作为正则化器的无监督组件结合起来。这两个目标是联合优化的，标记数据占据一个批次的前半部分，未标记数据占据后半部分。对于标记的数据，我们使用地面真实感 3D 姿态作为目标，训练一个监督的损失。未标记数据用于实现自编码器丢失，预测的 3D 姿态被投影回 2D，然后检查与输入的一致性。

轨迹模型。由于透视投影，屏幕上的 2D 位姿既取决于轨迹(即人体根关节的全局位置)，也取决于 3D 位姿(所有关节相对于根关节的位置)。如果没有全球位置，主题将总是以固定的比例重新投射在屏幕中央。因此，我们还对人的 3D 轨迹进行了回归，使反投影到 2D 能够正确进行。为此，我们优化了第二个网络，使摄像机空间中的全局轨迹回归。后者在将其投影回 2D 之前加入到位姿中。这两个网络具有相同的体系结构，但不像我们所观察到的那样共享任何权值，它们在以多任务方式训练时互相产生负面影响。当被测物体离相机较远时，精确轨迹的回归变得越来越困难，我们对轨迹的加权平均每关节位置误差(WMPJPE)损失函数进行了优化：

也就是说，我们在相机空间中利用地面真深度(yz)的逆来对每个样本进行加权。为了我们的目的，对于远场受试者来说，返回一个切前轨迹也是不必

$$E = \frac{1}{\mathbf{y}_z}\|f(\mathbf{x}) - \mathbf{y}\| \qquad (1)$$

要的，因为相应的 2D 关键点往往会围绕一个小区域集中。

骨长 L2 损失。我们希望激发对可能的 3D 姿

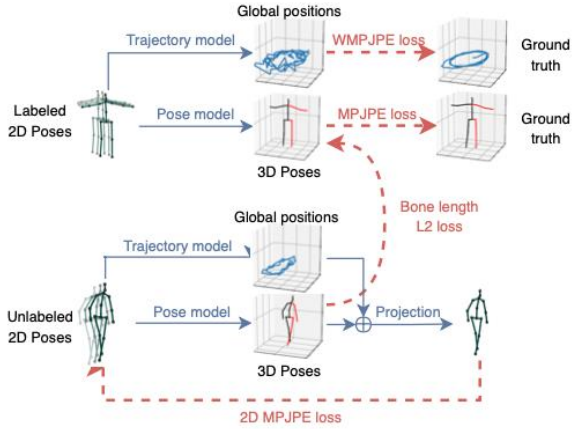


图 3：三维姿态模型的半监督训练，将可能预测的二维姿态序列作为输入。我们回归了人的三维轨迹，并添加了一个软约束来匹配未标记预测的平均骨长度。一切都是共同训练的。WMPJPE 代表“加权 MPJPE”。

势的预措辞，而不是仅仅复制插入。为此，我们发现加入一个软约束来近似地将未标记批中子弹头的平均骨长匹配到标记批的被试身上是有效的(图 3 中的‘骨长 L2 丢失’)。这个术语在自我监督中起着重要的作用，如我们在§6.2 中所示。

讨论。该方法只需要摄像机的固有参数，通常适用于商用凸轮机构。1 .该方法不依赖于特定的网络结构，适用于任何以二维关键点为输入的三维位姿检测。在我们的实验中我们使用了§3 中描述的体系结构，将 2D 姿态映射到 3D。为了将三维姿态投影到二维，我们使用了一个简单的投影层，它考虑了线性参数(焦距、主点)和非线性透镜畸变系数(切向和径向)。我们发现，Human3.6M 中使用的相机镜头畸变对位姿估计的影响微乎其微，但我们也包括了这些术语，因为它们总是能够更精确地模拟真实的相机投影。

5 实验装置

5.1 数据集与评价

我们在Hu-man3.6M[20，19]和Human Eva-I[47]两个运动捕获数据集上进行评估，Human3.6M包含 11 个受试者 360 万个视频帧，其中 7 个为 3D姿态标注。每个被试执行 15 个动作，使用 4 个 50 Hz的同步凸轮时代记录。在前期工作[41，52，34，50，10，40，56，33]的基础上，我们采用 17 关节骨架，对 5 个子项目(S1，S5，S6，S7，S8)进行训练，对 2 个被试(S9 和S11)进行测试。我们为所有的动作训练一个单一的模型。

HumanEva-I是一个小得多的数据集，在 60 Hz 时从 3 个相机视图记录 3 个子弹目。在[34，16]的基础上，我们对三个动作(Walk，Jog，Box)进行评估，每个动作训练一个不同的模型(单个动作- SA)。我们还报告了在为所有行动训练一个模型(多行动-MA)时的结果，如[41，27]。我们采用 15 关节骨架，并使用提供的列车/试验分隔。

在我们的实验中，我们考虑了三个评估准则：协议 1 是毫米级的平均每关节位置误差(MPJPE)，它是预测关节位置与地面真实关节位置之间的平均欧氏距离，如下[29，53，59，34，41]。协议 2 报告了在平移、旋转和缩放(P-MPJPE)[34，50，10，40，56，16]中与地面真值对齐后的错误。协议 3 在文献[45]的半监督实验中，仅在尺度上(N-MPJPE)将预测的姿态与地面真实的姿态一致。

5.2 2D位姿估计的实现细节

以往的工作[34，58，52]都是从地-真包围盒中提取主题，然后应用堆叠式沙漏探测器来预测地-真包围盒中的二维关键点[38]，我们的方法(§3 和§4)不依赖于任何特定的二维关键点探测器。因此，我们研究了几个不依赖于地面真实感盒子的二维探测器，这些盒子可以在野外使用我们的设置。除了堆叠式沙漏探测器外，我们还研究了具有ResNet -101-FPN[31] 骨 干 的 Mask R-CNN[12] ， 利 用 它 在 Detectron中的参考实现，以及级联金字塔网络工作(CPN)[5]，它代表了FPN的扩展。CPN实现需要外部提供包围盒(我们针对这种情况使用Mask R-CNN盒)。

对于Mask R- CNN和CPN，由于COCO的关键点不同于Human3.6M [20]，我们从COCO的预训练模型开始[32]，并对Human3.6M的二维投影上的探测器进行微调。

对于Mask R - CNN，我们采用了"拉伸 1x "计划训练的ResNet-101 骨干网[12]。在对Human3.6 M模型进行微调时，我们重新初始化关键点网络的最后一层，以及回归热图的deconv层来学习一组新的关键点。我们在 4 个GPU上以逐步衰减的学习速率进行训练：1e-3 进行 60k次迭代，然后 1e-4 进行 10k 次迭代，1e-5 进行 10k次迭代。推论时，我们在热图上施加一个softmax，并提取得到的二维分布的期望值(soft-argmax)，这样得到的预测比hard-argmax更加平滑和精确[33]。

对于CPN，我们使用了分辨率为 384×288 的ResNet-50 骨干。为了微调，我们重新初始化GlobalNet和RefineNet的最终各层(卷积权值和批归一化统计量)。接下来，我们在一个批 32 幅图像的GPU上以逐步衰减的学习速率进行训练：5e-5 (初始值的 1/10)进行 6k iter-ation，然后 5e-6 进行 4k次迭代，最后 5e-7 进行 2k次迭代。我们在微调的同时保持批归一化启用。我们用地面真实感包围盒进行训练，并用微调Mask R-CNN模型预测的包围盒进行测试。

5.3 3D位姿估计的实现细节

为了与其他工作的一致性[34，29，53，59，34，41]，我们根据摄像机变换对地面真实姿态进行旋转平移，不使用全局轨迹(半监督设置除外，§4)，对摄像机空间中的 3D姿态进行训练和评估。

作为优化器，我们使用Amsgrad [43]，训练 80 个历元。对于Human3.6M，我们采用指数递减的学习速率计划，从η=0.001 开始，每个历元施加收缩因子α=0.95。

所有的时间模型，即接受场大于一个的模型，都对姿态序列中样本的相关性很敏感(cf.§3)，这导致了假设独立样本的批归一化有偏统计量[17]。在初步的实验中，我们发现在训练过程中预先确定大量相邻帧的结果比没有时间信息的模型(批内样本具有很好的随机性)要差。我们通过从不同的视频片段中选择训练片段来减少训练样本中的相关性。剪辑集大小设置为我们架构的接受域的宽度，以便模型预测每个训练剪辑的单个 3D姿态。这对于概括来说很重要，我们在附录A.5 中对此进行了详细分析。

我们可以通过将步长设置为伸缩因子的展宽卷积替换为展宽卷积来极大地优化这个单帧设置(见Ap-pendixA.6)。这避免了从未使用的计算状态，我们只在训练期间应用此优化。在推理时，可以对整个序列进行处理，并重用其他 3D帧的术语状态，以便更快地进行推理。这是可能的，因为我们的模型在时间维度上不使用任何形式的池化。为了避免丢失帧到有效卷积，我们通过复制进行垫片，但只在一个序列的输入边界(附录A.5，图 9a所示为插图)。

我们观察到，批归一的默认超参数导致试验误差(±1mm)的大幅波动，以及用于推断的运行估计的波动。为了获得更稳定的运行统计量，我们使用了一个批归一化动量β的调度：从β=0.1 开始，按指数衰减，直到最后一个时刻达到β=0.001。

最后，在列车和测试时刻进行水平翻转增强。我们在附录A.4 中展示了这一效果。

对于HumanEva，使用N=128，α=0.996，使用 27 帧的感受野训练 1000 个历元。HumanEva中的一些帧由于传感器丢失而被损坏，我们将损坏的视频分割成有效的相邻块，并将其作为独立视频对待。

6 结论

我们介绍了一种简单的用于视频中三维人体姿态估计的全卷积模型。我们的架构在 2D关键点轨迹上使用扩展卷积来部署时间信息。本工作的第二个贡献是反投影，一种在标记数据稀缺时提高性能的半监督训练方法。该方法工作于未标记视频，只需要三角测量摄像机参数，在运动捕捉具有挑战性的场景(如户外运动)中具有实用性。

我们的全卷积结构将目前流行的Human3.6M数据集上最好的结果提高了 6mm的平均联合误差，相对减少了 11 %，同时也显示了对HumanEva-I的改进。当有 5K或更少的注释帧时，反投影可以在强基线上提高约 10mm的N-MPJPE (15mmMPJPE)位姿估计精度。