密集的传记：在野外的密集的人类姿势估计

图1：密集姿态估计旨在将RGB图像的所有人像素映射到人体的3D表面。 我们引入了DensePose-COCO，这是一个大型的地面真实数据集，包含在50K COCO图像上手动注释的图像到表面对应关系，并训练DensePose-RCNN，以每秒多帧的速度在每个人类区域内密集回归特定部位的UV坐标。 左图：DensePose-RCNN，Middle：DensePCOCOCO数据集注释的图像和回归对应关系，Right：身体表面的分区和UV参数化。

抽象

在这项工作中，我们建立了RGB图像和人体基于表面的表示之间的密集对应关系，我们称之为密集人体姿态估计。我们首先通过引入高效的注释流水线来收集出现在COCO数据集中的5万人的密集对应关系。然后，我们使用我们的数据集来训练基于CNN的系统，这些系统能够在野外提供密集的对应关系，即存在背景，遮挡和尺度变化。我们通过培训能够填补缺失的基础真值的“修补”网络来提高我们的培训集的有效性，并报告过去可以实现的最佳结果方面的明显改进。我们试验完全卷积网络和基于区域的模型，并观察后者的优越性;我们通过级联进一步提高准确性，获得实时提​​供高精度结果的系统。补充材料和视频提供在项目页面http：//densepose.org。

介绍

这项工作旨在通过建立从2D图像到3D人体表面的密集对应关系，进一步推动人们对图像理解的包络。 我们可以将此任务理解为涉及其他几个问题，例如对象检测，姿态估计，零件和实例分割，作为特殊情况或先决条件。 解决这个问题的应用程序需要超越平面标志性本地化（如图形，增强现实或人机交互）的问题，并且也可能是通用的基于3D的对象理解的垫脚石。

从图像到基于表面的模型建立密集对应的任务主要在深度传感器可用的环境中解决，如[41]的Vitruvian流形，度量回归森林[33]或更近期 密集点云对应[44]。 相反，在我们的例子中，我们将单个RGB图像视为输入，基于此我们建立了表面点和图像像素之间的对应关系

从图像到基于表面的模型建立密集对应的任务主要在深度传感器可用的环境中解决，如[41]的Vitruvian流形，度量回归森林[33]或更近期 密集点云对应[44]。 相反，在我们的例子中，我们将单个RGB图像视为输入，基于此我们建立了表面点和图像像素之间的对应关系

最近的几项工作旨在恢复密集对[3]或RGB图像集[48，10]之间的密切对应在无人监督的设置。 最近，[42]使用等变性原理来将图像集合对齐到一个共同的坐标系，同时遵循分组图像对准的一般思想，例如， [23,21]。

虽然这些作品是针对一般类别的，但我们的作品集中在可以说是最重要的视觉类别 - 人类。 对于人来说，可以通过[2]的皮肤多人线性（SMPL）模型或通过仔细控制的三维表面采集获得的更近期的亚当模型[14]来简化任务。 谈到图像到曲面映射的任务，在文献[2]中，作者提出了一种两阶段方法，首先通过CNN检测人类地标，然后通过迭代最小化将参数可变形表面模型拟合到图像。 与我们的工作平行[20]，开发[2]以端到端方式运行的方法，将迭代重投影误差最小化作为深度网络的模块，其恢复3D相机姿态和低维身体参数化。

图2：通过让注释者将图像分割成语义区域，然后为任何渲染零件图像上的每个采样点定位相应的表面点，我们注释图像与3D表面模型之间的密集对应关系。 红十字表示当前注释的点。 渲染视图的表面坐标定位3D模型上收集的2D点。

我们的方法不同于所有这些作品，因为我们采用全面的监督式学习方法，并收集图像与详细，精确的人体参数曲面模型之间的地面真实对应关系[27]：而不是在测试中使用SMPL模型我们只用它作为定义培训期间问题的一种手段。我们的方法可以理解为[26，1，19，7，40，18，28]中延伸人类标准的工作路线的下一步。在Fashionista [46]，PASCAL-Parts [6]和Look-Into-People（LIP）[12]数据集中提供了人体部分分割蒙版;这些可以被理解为提供图像 - 表面对应的粗化版本，而不是连续坐标，其预测离散的部分标签。表面水平监督最近才引入了合成图像[43]，而[22]中的8515图像数据集使用关键点和3D自动拟合图像进行注释。在这项工作中，我们引入了一种新的注释流水线，使我们能够收集COCO数据集的50K图像的地面真实对应关系，而不是损害我们训练集的范围和真实性，从而产生我们新的DensePose-COCO数据集。

我们的工作与最近的DenseReg框架最接近[13]，在这个框架中，CNN被训练成功在3D模型和“野外”图像之间成功建立密集对应关系。这项工作主要集中在人脸上，并对具有中等姿势变化的数据集评估其结果。然而，在这里，由于人体更高的复杂性和灵活性，以及​​更大的姿势变化，我们正面临新的挑战。我们通过设计适当的体系结构来解决这些挑战，如第2节所述。 3，它比DenseReg型完全卷积体系结构产生了实质性的改进。通过将我们的方法与[15]中最近的Mask-RCNN系统相结合，我们证明了一个有区别地训练的模型可以为包含数十个人的复杂场景以实时速度恢复高精度对应场：在GTX 1080 GPU上，我们的系统运行对于240×320图像为每秒20-26帧或对于800×1100图像为每秒4-5帧。

我们的贡献可以总结为三点。 首先，如第二节所述。 2，我们通过收集SMPL模型[27]和出现在COCO数据集中的人之间的密集对应关系，介绍第一个手动收集的任务的地面实况数据集。 这是通过一种新颖的注释管道完成的，该管道在注释过程中利用3D表面信息。

其次，如第二节所述。 如图3所示，我们使用结果数据集来训练基于CNN的系统，通过回归任何图像像素处的身体表面坐标，在野外提供密集的对应关系。 我们依赖于Deeplab [4]以及基于区域的系统，依靠MaskRCNN [15]实验全卷积体系结构，观察基于区域的模型在完全卷积网络上的优越性。 我们还考虑了我们方法的级联变体，相对于现有体系结构产生了进一步的改进。

第三，我们探索利用我们构建的地面真实信息的不同方式。 我们的监督信号被定义在每个训练样本随机选择的图像像素子集上。 我们使用这些稀疏的对应关系来训练“教师”网络，该网络可以在图像域的其余部分“修复”监督信号。 与稀疏点或任何其他现有数据集相比，使用此修复后的信号可显着提高性能，如第2节中的实验所示。4。

我们的实验表明，密集的人体姿态估计在很大程度上是可行的，但仍有改进空间。 我们用一些定性结果和方向来总结我们的论文，说明该方法的潜力。 我们将从我们项目的网页http://densepose.org公开提供代码和数据。

2. COCO密集数据集

收集丰富，高质量的训练集已成为分类[38]，检测和分割[8,26]任务进展的催化剂。 目前，对于真实图像的密集人体姿态估计，不存在手动收集的地面实况。 [22]和[43]的作品可以作为替代品，但正如我们在第二节中展示的那样。 4提供更差的监督。

在本节中，我们将介绍我们的COCO-DensePose数据集，以及评估测量，使我们能够量化第二部分中的任务进度。 4.我们收集了5万人的注释，收集了500多万手动注释的信件。

我们首先介绍我们的注释管线，因为这需要几个设计选择，这对3D注释来说可能更为普遍。 然后，我们转而分析所收集的地面真相的准确性，以及用于评估不同方法的结果性能度量。

图3：用于收集每部分通信注释的用户界面：我们为注释者提供了六个身体部分的预渲染视图，以使整个部分的表面可见。 一旦目标点被注释，点就会同时显示在所有渲染图像上。

2.1。 注释系统

在这项工作中，我们使用人类注释器来建立从2D图像到人体表面表示的密集对应。如果天真地完成，这将需要通过旋转操纵表面来为每个2D图像点“狩猎顶点”，这可能令人沮丧地效率低下。相反，我们构建了一条注释管道，通过该管道我们可以高效地收集图像到表面对应的注释。

如图2所示，在第一阶段，我们要求注释者划定与可见的，语义定义的身体部位相对应的区域。这些包括头部，躯干，下/上臂，下/上腿，手和脚。为了简化UV参数化，我们将零件设计成与平面同构，将肢体和躯干分成上下部分和前后部分。

对于头部，手部和脚部，我们使用SMPL模型[27]中提供的手动获取的UV场。对于其余部分，我们通过应用于成对测地距离的多维缩放来获得解包。图1（右）显示了24个部件的UV场。

我们指示注释者估计衣服后面的身体部分，以便例如穿着大裙子不会使对随后的注释变得复杂。在第二阶段，我们用k-means获得的一组大致等距的点对每个部分区域进行采样，并要求注释者将这些点与表面对应。采样点的数量根据零件的大小而变化，每个零件的采样点的最大数量为14.为了简化此任务，我们通过提供六个同一身体部位的预渲染视图来“展开”零件表面并允许用户在他们中的任何一个上放置地标。这允许注释者通过从六个选项中选择一个而不是手动旋转表面来选择最方便的视点。

在这项工作中，我们使用人类注释器来建立从2D图像到人体表面表示的密集对应。如果天真地完成，这将需要通过旋转操纵表面来为每个2D图像点“狩猎顶点”，这可能令人沮丧地效率低下。相反，我们构建了一条注释管道，通过该管道我们可以高效地收集图像到表面对应的注释。

如图2所示，在第一阶段，我们要求注释者划定与可见的，语义定义的身体部位相对应的区域。这些包括头部，躯干，下/上臂，下/上腿，手和脚。为了简化UV参数化，我们将零件设计成与平面同构，将肢体和躯干分成上下部分和前后部分。

对于头部，手部和脚部，我们使用SMPL模型[27]中提供的手动获取的UV场。对于其余部分，我们通过应用于成对测地距离的多维缩放来获得解包。图1（右）显示了24个部件的UV场。

我们指示注释者估计衣服后面的身体部分，以便例如穿着大裙子不会使对随后的注释变得复杂。在第二阶段，我们用k-means获得的一组大致等距的点对每个部分区域进行采样，并要求注释者将这些点与表面对应。采样点的数量根据零件的大小而变化，每个零件的采样点的最大数量为14.为了简化此任务，我们通过提供六个同一身体部位的预渲染视图来“展开”零件表面并允许用户在他们中的任何一个上放置地标。这允许注释者通过从六个选项中选择一个而不是手动旋转表面来选择最方便的视点。

当用户在任何渲染零件视图上指示一个点时，其表面坐标将用于同时显示其在剩余视图上的位置 - 这将给出对应关系的全局概述。 图像点以水平/垂直连续的形式呈现给注释者，这使得通过避免表面的自交会更容易提供几何一致的注释。 这个两阶段注释过程使我们能够非常有效地收集高度准确的信件。 如果我们根据完成它所需的时间量化注释任务的复杂性，我们已经看到，部分分割和对应注释任务几乎在同一时间，这是令人惊讶的，因为后者任务更具挑战性。 所收集的注释的可视化在图4中提供，其中表面和U，V坐标的划分示于图1

图4：注释的可视化：收集点的图像（左），U（中）和V（右）值

我们评估人类注释员在性能方面的黄金标准。通常在姿势估计中，要求多个注释者标记相同的界标，然后将其用于评估位置的变化，例如， [26，36]。在我们的例子中，我们可以渲染图像，我们可以访问用于渲染像素的真实网格坐标。因此，我们直接比较渲染过程中使用的真实位置和注释器估计的位置，而不是首先估计多个人类注释器中的“共识”地标位置。

特别是，我们为注释者提供通过与我们在地面真实注释中使用的完全相同的表面模型生成的合成图像，利用[43]的渲染系统和纹理。然后，我们要求注释者使用我们的注释工具将合成图像与表面对应起来，并且对于每个图像k估计正确的表面点i与人类annotatorsik估计的点之间的测地距离di，k：

其中g（·，·）测量两个表面点之间的测地距离。

对于任何图像k，我们仅对随机采样的一组曲面点Sk进行注释和估计误差，并在曲面的其余部分插值误差。 最后，我们对用于评估注释器性能的所有K个示例中的错误进行平均。

如图5所示，小表面部分的注释误差明显较小，具有独特的特征，可以帮助定位（面部，手部，脚部），而在通常由衣服（躯干，背部，臀部）覆盖的较大的统一区域 注释器错误可能会变大

2.3。 评估措施

我们考虑两种不同的方式来总结整个人体的对应精度，包括逐点评估和实例评估。

点评估。该方法通过比例正确点（RCP）对应关系评估整个图像域的对应精度，其中如果测地距离低于特定阈值则对应关系被声明为正确。当阈值t变化时，我们得到一条曲线f（t），它的面积为我们提供了对应精度的标量总结。对于任何给定的图像，我们都有一组不同的点与地面真实信号。我们总结了这些点的整体表现，并聚集在图像中。我们评估曲线下面积（AUC），AUC分别为a = 10cm，30cm的两个不同值，分别产生AUC10和AUC30，其中AUC10被认为是更精确对应的准确度量度。此性能指标可轻松应用于单人和多人情景，并可直接提供可比较的值。在图6中，我们提供了对合成数据的人类注释器性能的每部分逐点评估，这可以看作是我们系统性能的上限。

按实例评估。 受用于COCO数据集[26,36]上的姿态评估的物体关键点相似性（OKS）度量的启发，我们引入测地点相似度（GPS）作为对应匹配得分：

其中Pj是在实例j上注释的地面真值点的集合，ip是由点p处的模型估计的顶点，ip是地面实体顶点p，并且κ是归一化参数。 我们设置κ= 0.255，使得单个点的GPS值为0.5，如果它与地面真实的测地距离等于身体段的平均一半尺寸（对应于大约30cm）。 直观地说，这意味着通过完美的零件分割模型可以实现GPS≈0.5的分数，而高于这个分数也需要更精确地定位表面上的点。

一旦匹配被执行，我们遵循COCO挑战协议[26,37]并且在从0.5到0.95的多个GPS阈值处评估平均精确度（AP）和平均回忆（AR），其对应于测地距离的范围 在0和30cm之间。 我们使用相同的距离范围来执行每个实例和每个点的评估。

3.学习密集的人体姿势估计

现在我们转向训练一个预测图像像素和表面点之间密集对应关系的深层网络的任务。 最近，通过完全卷积网络体系结构[4]在密集回归（DenseReg）系统[13]中解决了这样的任务。 在这项工作中，我们通过将DenseReg方法与Mask-RCNN架构相结合来引入改进的架构[15]，产生了我们的'DensePose-RCNN'系统。 我们开发了DensePose-RCNN的级联扩展，进一步提高了准确性，并描述了一种基于训练的插值方法，该方法允许我们将稀疏监督信号变成更密集和更有效的变体。

3.1。 全卷积密集回归

最简单的架构选择包括使用完全卷积网络（FCN），该网络结合了分类和回归任务，类似于DenseReg。 在第一步中，我们将一个像素归类为属于背景或者几个区域部分中的一个，它们提供表面坐标的粗略估计。 这相当于使用标准交叉熵损失进行训练的标签任务。 第二步，回归系统指出零件内像素的确切坐标。 由于人体结构复杂，因此我们将其分解为多个独立的部分，并使用局部二维坐标系对每个部分进行参数化，以识别该表面部分上任何节点的位置。

图6：不同身体部位内的人体注释错误分布。

图7：DensePose-RCNN架构：我们使用区域提案生成和特征池的级联，紧接着是一个完全卷积网络，它密集地预测离散部分标签和连续表面坐标。

直观地说，我们可以说，我们首先使用外观来粗略估计像素所属的位置，然后通过一些小规模校正将其与精确位置对齐。 具体而言，在图像位置i处的坐标回归可以如下公式化：

在第一阶段中，我们将位置i分配给具有最高后验概率的身体部分c \*，如分类分支所计算的那样，而在第二阶段中，我们使用回归器Rc \*将点i置于连续的U，V 零件c \*的坐标参数化。 在我们的例子中，c可以取25个值（一个是背景），这意味着Px是一个25路分类单元，我们训练24个回归函数Rc，每个函数在其各自的部分c中提供二维坐标。 在训练时，我们使用交叉熵损失来进行零件分类，并且使用平滑的L1损失来训练每个回归器。 如果像素在特定部分内，则只考虑回归损失。

3.2。 基于区域的稠密姿态回归

使用FCN使得该系统特别容易训练，但加载相同的深度网络和太多的任务，包括零件分割和像素定位，同时需要尺度不变性，这对于COCO中的人来说变得具有挑战性。 在这里，我们采用基于区域的方法[34,15]，其中包括提案区域（ROI）级联，通过ROI池提取区域适应特征[16,15]，并将结果特征提供给 一个区域特定的分支。 这样的体系结构将任务的复杂性分解为可控模块，并通过ROI池实现规模选择机制。 同时，他们也可以以端到端的方式联合训练[34]。