

BlazePose

BlazePose，一个轻量级的**卷积神经网络(CNN)架构**，用于人体姿势估计，它是为移动设备上的实时推理量身定做的。在推理过程中，该网络为一个人产生33个身体关键点，并在Pixel 2手机上以每秒超过30帧的速度运行。因此它特别适合实时用例，如健身追踪和手语识别。在Pixel 3 上 GPU 运行，BlazePose 可以达到112 FPS(frames per second)。

论文 (BlazePose: On-device Real-time Body Pose tracking) 提出新颖的**身体姿态追踪解决方法和用热图和回归关键点的网路坐标技术搭建的轻巧身体姿态估计神经网络**。(a novel body pose tracking solution and a lightweight body pose estimation neural network that uses both heatmaps and regression to keypoint coordinates.)

BlazePose是从达芬奇的《维特鲁威人》中得到的启发，可预测人的臀部中点、外接整个人的圆的半径以及连接肩部和臀部中点的直线的倾斜角度，共计可预测33个人体关键点。根据手和脚的比例和方向信息，即使是非常复杂的情况，比如特定的瑜伽姿态，其也能得到一致的追踪。

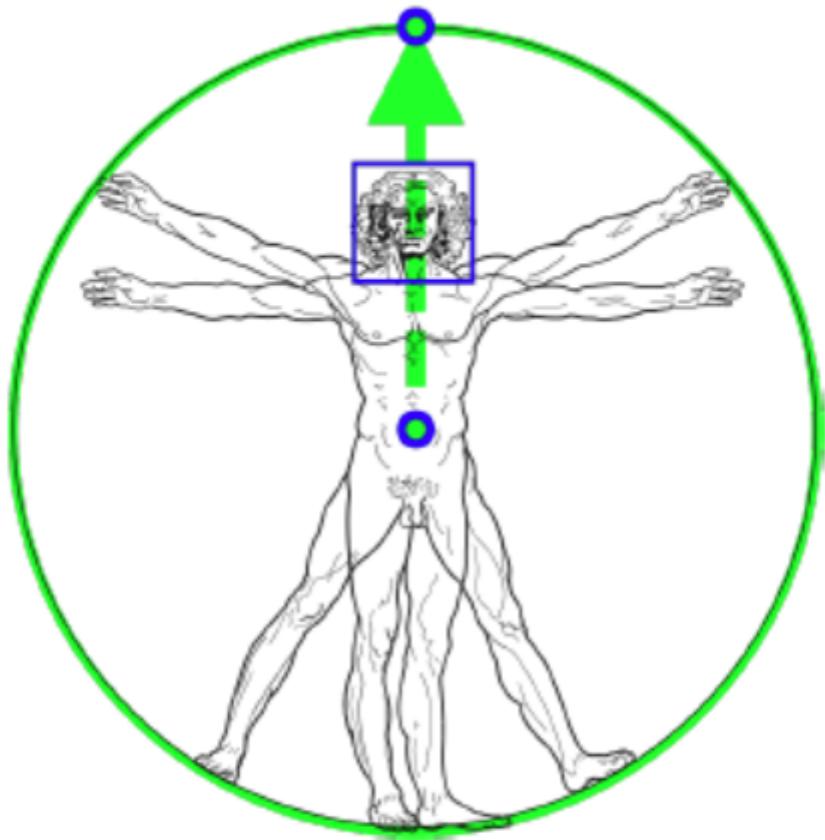


Figure 2. Vitruvian man aligned via our detector vs. face detection bounding box. See text for details.

[BlazePose Paper](#)

摘要

我们提出了BlazePose，这是一种用于人体姿势估计的轻量级卷积神经网络体系结构，专为在移动设备上进行实时推理而设计。在推理过程中，该网络为单个人生成33个人体关键点，并在Pixel 2手机上以每秒30帧的速度运行。这使其特别适合诸如健身跟踪和手语识别之类的实时用例。我们的主要贡献包括新颖的人体姿势跟踪解决方案和轻量级的人体姿势估计神经网络，该网络同时使用热图和回归关键点坐标。

1. Introduction

根据图像或视频进行人体姿势估计在各种应用（例如健康跟踪，手语识别和手势控制）中起着核心作用。

由于各种各样的姿势，众多的**自由度(degrees of freedom)**和**遮挡 (occlusions)**，这项任务具有挑战性。最近的工作在姿势估计方面显示出重大进展。常用的方法是为每个关节生成**热图(heatmaps for each joint)**以及**细化每个坐标的偏移量**。尽管这种热图选择可以以最小的开销将其扩展到多个人，但它使一个人的模型比适用于手机上的实时推断的模型大得多。在本文中，我们解决了这个特殊的用例，并演示了该模型的显著加速，几乎没有质量的下降。

与基于热图 (heatmaps) 的技术相反，基于**回归(regression)**的方法虽然对计算的要求较低且可扩展性较高，但它们试图预测平均坐标值，但通常无法解决潜在的歧义。已有一些研究表明，即使参数数量较少，**堆叠式沙漏架构(stacked hourglass architecture)**也可以大大提高预测的准确性。我们在工作中扩展了这个想法，并使用**编码器-解码器网络体系结构(encoder-decoder network architecture)**来预测所有关节的热图，随后是另一个直接回归到所有关节坐标的编码器。我们的工作的关键点是，热图分支(heatmap branch)可以在推理过程中丢弃，从而使其轻巧到可以在手机上运行。

2. Model Architecture and Pipeline Design(模型架构与管道设计)

2.1 Inference pipeline (推理管道)

在推论过程中，我们采用了**检测器-跟踪器(detector-tracker)**装置（见图1），该装置在各种任务（例如**手地标预测(hand landmark prediction)** 和 **密集人脸地标预测(dense face landmark prediction)**）上表现出出色的实时性能。我们的管道包括一个轻巧的**人体姿势检测器 (body pose detector)**，然后是一个**姿势跟踪器网络 (pose tracker network)**。跟踪器 (tracker) 预测关键点坐标 (keypoint coordinates)、当前帧(frame)上人物的存在状态(the presence of the person on the current frame)和当前帧的**姿态兴趣区域 (ROI)** (the refined region of interest of the current frame)。当跟踪器指示没有人在场时，我们在下一帧重新运行检测器网络。

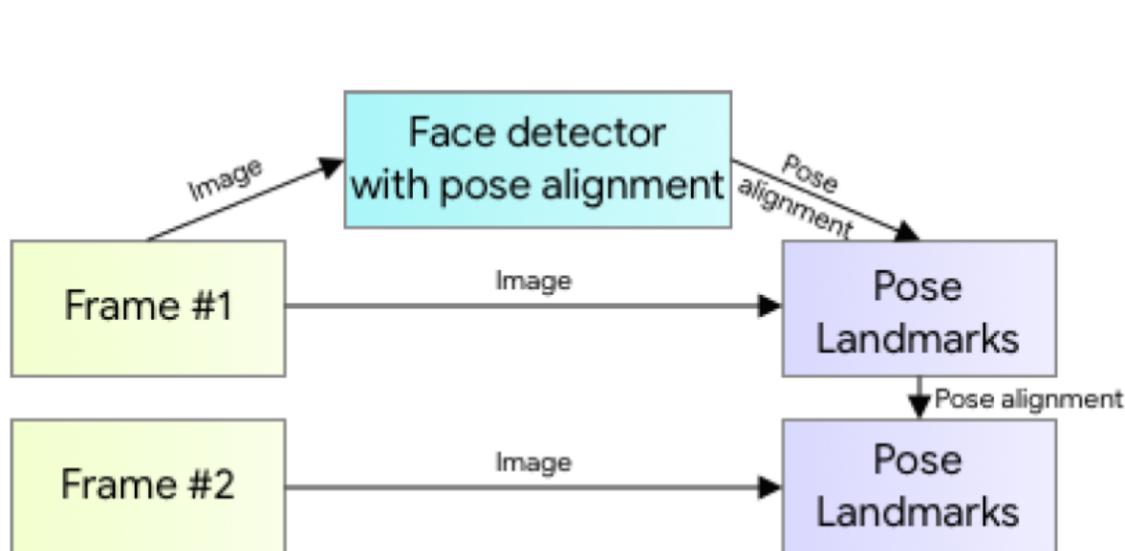


Figure 1. Inference pipeline. See text.

2.2 Person detector (人体检测器)

大多数现代对象检测解决方案在其最后的后处理(post-processing)步骤中都依赖于**非最大抑制 (Non-Maximum Suppression, NMS) 算法**。这对于具有很少自由度的刚性物体非常有效。然而，对于包括像人类那样的高度铰接的姿势(articulated poses)(连续动作、关节铰接)的场景，该算法容易崩溃，比如，人们挥舞或拥抱。这是因为多个模棱两可的框(boxes)满足了NMS算法的**联合阈值交集(IoU)**。

为了克服此限制，我们专注于检测相对刚性的身体部分（如人脸或躯干）的边界框(bounding box)。我们观察到，在许多情况下，神经网络中有关躯干位置的最强信号是人的脸部（因为它具有高对比度特征 (high-contrast features)，并且外观变化较少 (fewer variations in appearance)）。为了使这样的人体检测器快速，轻便，以及使AR应用程序有效，我们做了一个大胆的假设：在单人用例中始终应看到人员的头部。

因此，我们使用快速设备面部检测器(fast on-device face detector)作为人体检测器的替代品。该面部检测器可预测其他特定人对齐参数(additional person-specific alignment parameters)：人臀部的中点，整个人的外接圆大小，人体倾斜程度（两个中肩和臀部中心连接线的角度）。

2.3 Topology(拓扑结构)

通过获取BlazeFace，BlazePalm 和Coco 使用的那些点的超集(superset)，在人体上使用33个点。这使我们能够与各自的数据集和推理网络(inference network)保持一致。

与OpenPose 和Kinect 拓扑相反，我们仅使用最少数量的面部，手和脚上的关键点来估计姿态兴趣区域的旋转，大小和位置后续模型。拓扑图信息：

In contrast with the OpenPose and Kinect topologies, we use only a minimally sufficient number of keypoints on the face, hands, and feet to estimate rotation, size, and position of the region of interest for the subsequent model.

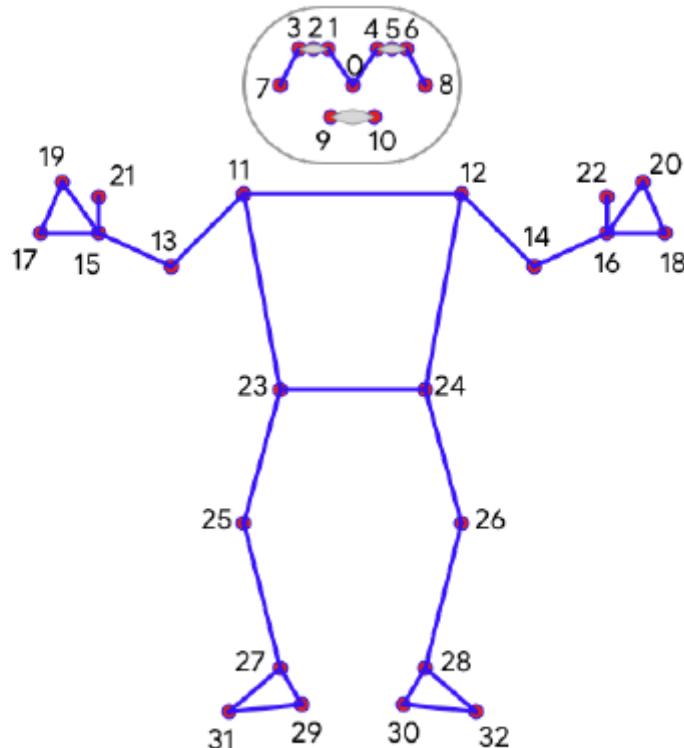


Figure 3. 33 keypoint topology.

Appendix A. BlazePose keypoint names

0. Nose
1. Left eye inner
2. Left eye
3. Left eye outer
4. Right eye inner
5. Right eye
6. Right eye outer
7. Left ear
8. Right ear
9. Mouth left
10. Mouth right
11. Left shoulder
12. Right shoulder
13. Left elbow
14. Right elbow
15. Left wrist
16. Right wrist
17. Left pinky #1 knuckle
18. Right pinky #1 knuckle
19. Left index #1 knuckle
20. Right index #1 knuckle
21. Left thumb #2 knuckle
22. Right thumb #2 knuckle
23. Left hip
24. Right hip
25. Left knee
26. Right knee
27. Left ankle
28. Right ankle
29. Left heel
30. Right heel
31. Left foot index
32. Right foot index

2.4 Dataset(数据集)

与大多数使用热图检测关键点的现有姿势估计解决方案相比，我们基于跟踪 (tracking-based) 的解决方案需要初始姿势对齐 (initial pose alignment)。我们将数据集限制在以下两种情形：要么可以看到整个人，要么可以清晰地标识臀部和肩部关键点。为了确保模型支持数据集中不存在的重度遮挡(heavy occlusions)情况，我们使用了大量的遮挡模拟增强(occlusion-simulating augmentation)。我们的训练数据集包括：一个或几个人的60K图像和一个人进行健身运动的25K图像。所有这些图像都是由人类进行标识。

2.5 Neural network architecture(神经网络架构)

我们系统的姿态估计组件 (the pose estimation component) 可预测所有33个人体关键点的位置，并使用管道(Section 2.1)第一阶段所提供的人员对齐建议 (person alignment proposal)。

我们采用热图(heatmap), 偏移量(offset)和回归(regression)方法相结合的方法。

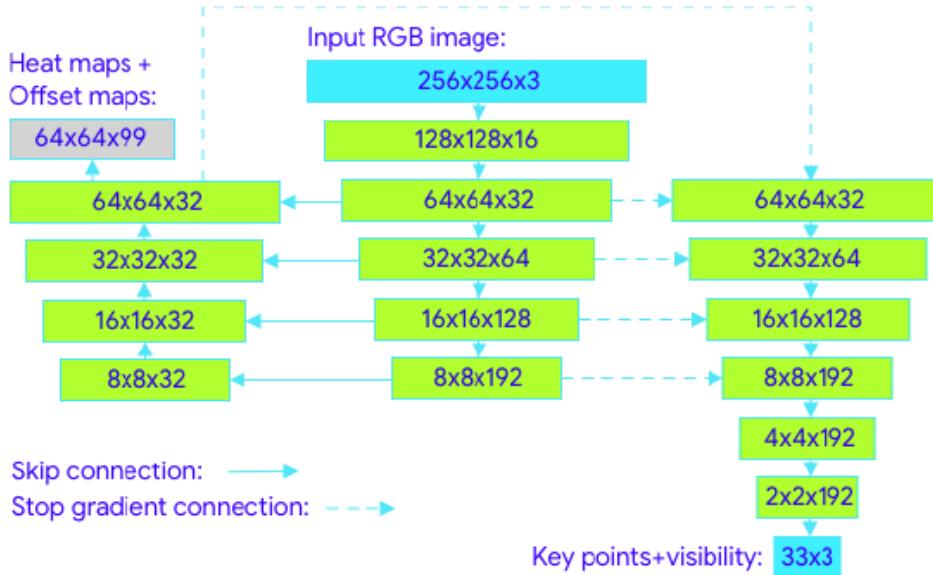


Figure 4. Network architecture. See text for details.

我们仅在训练阶段使用热图(heatmap)和偏移损失(offset loss)，并在运行推理(inference)之前从模型中删除相应的输出层(output layers)。因此，我们有效地使用了热图来监督轻量级嵌入(supervise the lightweight embedding)，然后再将其用于回归编码器网络(regression encoder network)。这种方法部分地受到Newell等人的Stacked Hourglass方法的启发。但在我们的案例中，**我们堆叠了一个很小的编码器-解码器基于热图的网络 (a tiny encoder-decoder heatmap-based network) 和一个随后的回归编码器网络 (a subsequent regression encoder network)。**

我们积极利用网络中所有阶段之间的跳连接(skip-connections)来实现高级功能和低级功能之间的平衡。但是，来自回归编码器 (regression encoder) 的梯度 (gradients) 不会传播回受热图训练 (heatmap-trained)的特征 (请注意图4中的梯度停止连接 (gradient-stopping connections))。我们发现这不仅改善了热图预测 (heatmap predictions)，而且还大大提高了坐标回归精度(coordinate regression accuracy)。

2.6 Alignment and occlusions augmentation (对齐与遮挡增强)

相关的姿态先验(a relevant pose)是所提出解决方案的重要组成部分。在训练过程中，我们故意在增强 (augmentation)和数据准备期间限制角度 (angle)，比例(scale)和平移(translation)的支持范围。这使我们能够降低网络容量 (network capacity)，使网络速度更快，同时在主机设备上所需的计算资源也更少。

基于检测阶段(detection stage)或先前的帧关键点(previous frame keypoints)，我们将人物对齐，以使臀部之间的点位于作为神经网络输入传递的正方形图像的中心。我们估计(estimate)旋转(rotation)作为臀中点与肩中点之间的直线L，并旋转图像，使L平行于y轴。估计比例，使所有的身体点都适合围绕身体的方形边界框，如图2所示。最重要的是，我们应用10%的比例(apply 10% scale)并进行移位增强 (shift augmentations)，以确保跟踪器(tracker)处理在框架和扭曲对齐之间的身体移动。

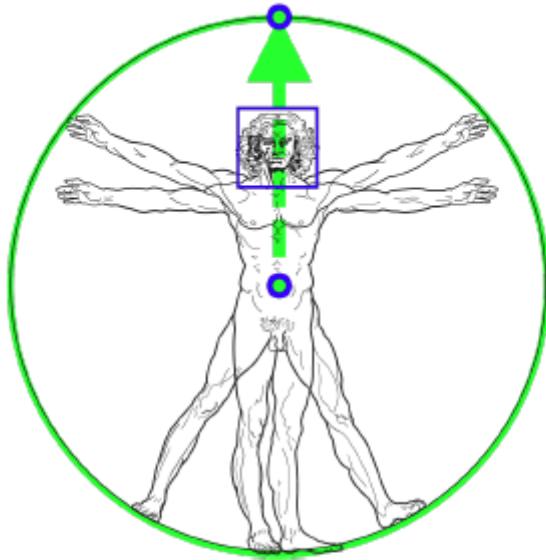


Figure 2. Vitruvian man aligned via our detector vs. face detection bounding box. See text for details.

为了支持对不可见点的预测，我们在训练过程中模拟了遮挡（填充各种颜色的随机矩形），并引入了**每点可见度分类器(per-point visibility classifier)**，该分类器指示是否遮挡了特定点和位置预测是否被认为是错误的。这样，即使在发生严重遮挡的情况下（例如仅上身）或大部分人的身体不在场景中时，也可以持续跟踪人。如图5所示。

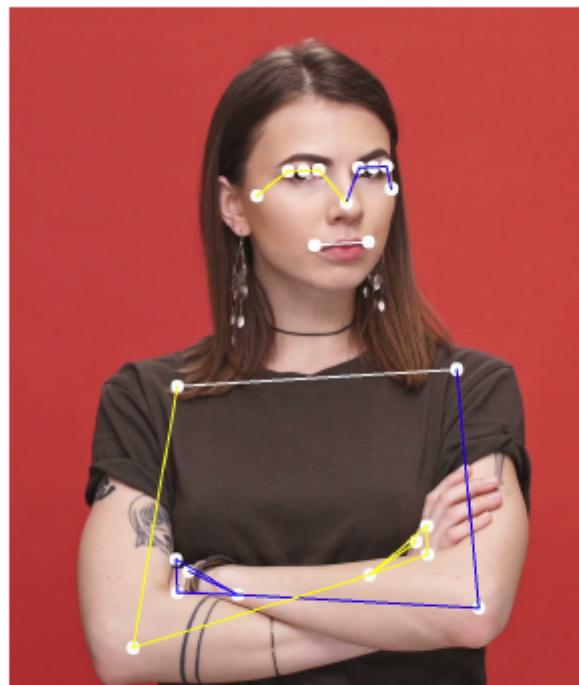


Figure 5. BlazePose results on upper-body case

3. Experiments (实验)

Model	FPS	AR Dataset, PCK@0.2	Yoga Dataset, PCK@0.2
OpenPose (body only)	0.4 ¹	87.8	83.4
BlazePose Full	10 ²	84.1	84.5
BlazePose Lite	31²	79.6	77.6

Table 1. BlazePose vs OpenPose



Figure 6. BlazePose results on yoga and fitness poses.

为了评估模型的质量，我们选择了OpenPose作为基准。为此，我们手动标识了两个包含1000个图像的数据集（inhouse datasets），每个数据集的场景中有1-2个人。第一个数据集称为**AR数据集(AR dataset)**，由各种野外人体姿势组成，而第二个仅由瑜伽/健身姿势组成。为了保持一致，我们仅使用MS Coco拓扑结构(MS Coco topology)（具有17个点）进行评估，这是OpenPose和BlazePose的共同子集。作为评估指标，我们使用公差为20%的正确点百分比（PCK@0.2）（Percent of Correct Points

with 20% tolerance) (在此假设，如果二维Euclidean误差 (2D Euclidean error) 小于相应人的躯干尺寸的20%，则可以正确检测到该点)。为了验证人类基线，我们要求两个作标识的人独立地重新标识AR数据集，并获得平均PCK@0.2为97.2。

我们训练了两个具有不同容量的模型：BlazePose Full (6.9 MFlop, 3.5M Params) 和BlazePose Lite (2.7 MFlop, 1.3M Params)。尽管我们的模型在AR数据集上显示的性能比OpenPose模型稍差，但在Yoga / Fitness用例上，BlazePose Full的性能优于OpenPose。同时，BlazePose在单个中层电话CPU(single mid-tier phone CPU)上的性能要比20核心台式机CPU(20 core desktop CPU)上的OpenPose快25-75倍，这取决于所要求的质量（效果，quality）。

4. Applications (应用领域)

我们开发了这种新的，易设备的，针对特定个人的人体姿势估计模型，以支持各种性能要求很高的用例，例如手语(Sign Language)，瑜伽/健身跟踪(Yoga/Fitness)和AR。该模型在移动CPU上几乎实时工作，并且可以在移动GPU上加快超实时延迟 (super-realtime latency)。由于其33个关键点拓扑结构与BlazeFace和BlazePalm 的保持一致，因此它可以成为后续**手势和面部几何估计模型的骨干(a backbone for subsequent hand pose and facial geometry estimation models)**。

我们的方法本身可扩展到更多的关键点，3D支持和其他关键点属性，因为它不是基于热图/偏移图(heatmaps/offset maps)的，因此不需要为每种新功能类型添加额外的全分辨率图层 (full-resolution layer)。

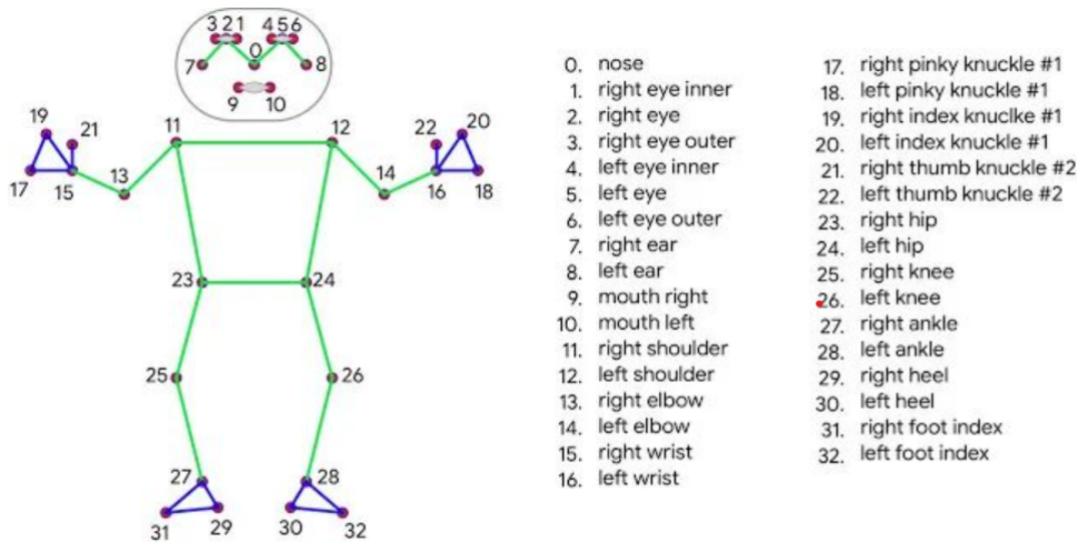
总结

方法

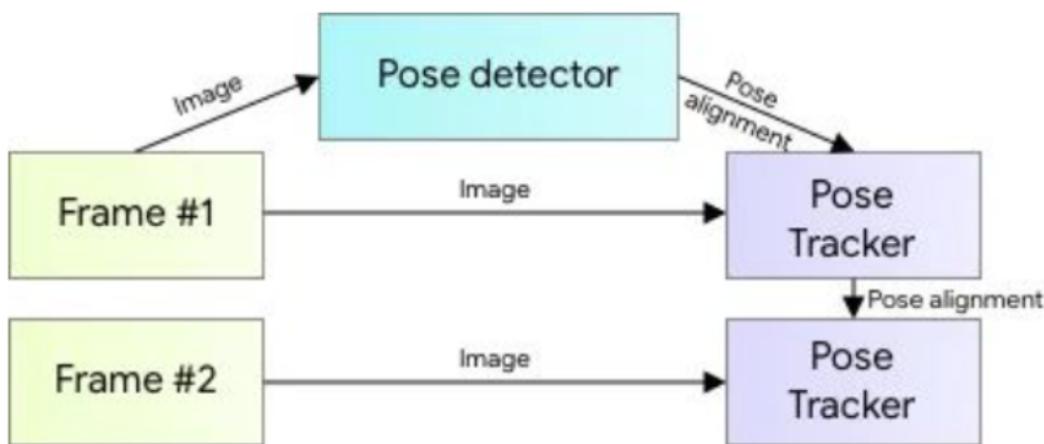
采用机器学习 (ML)，从单帧画面推断人体的 33 个 2D 关键点提供人体姿态追踪。相较于当前基于标准 COCO 拓扑的姿态模型，BlazePose 可以精确定位更多关键点，因此特别适合于健身应用。此外，当前最前沿 (SOTA) 的方法主要依靠强大的桌面环境进行推理，而我们的方法通过 CPU 推理在手机端实现了实时性能。如果利用 GPU 推理，BlazePose 可以实现超实时性能，从而运行后续的 ML 模型，如面部或手部追踪。

拓扑结构

人体姿态的当前标准是 COCO 拓扑，由横跨躯干、手臂、腿部和面部的 17 个关键点组成。不过，COCO 关键点只能定位脚踝和腕部的点，缺乏手和脚的比例和方向信息，而这些信息对健身和舞蹈等应用至关重要。因此，必须加入更多关键点，来用于手、脸或脚等特定域姿势预测模型的下游应用。在 BlazePose 中，我们提供了 33 个人体关键点的新拓扑，这是 COCO、BlazeFace 和 BlazePalm 拓扑的超集 (Superset)。这样一来，我们可以仅从姿势预测中确定与脸部和手部模型一致的身体语义信息。



姿态追踪的ML流水线



姿势预测利用了经过验证的两个步骤：检测器 - 追踪器 ML 流水线。流水线使用检测器首先定位帧内的姿态兴趣区域 (ROI)。追踪器随后根据此 ROI 预测所有 33 个姿态关键点。请注意，在视频用例中，检测器仅在第一帧上运行。后续帧将根据前一帧的姿态关键点得出 ROI。

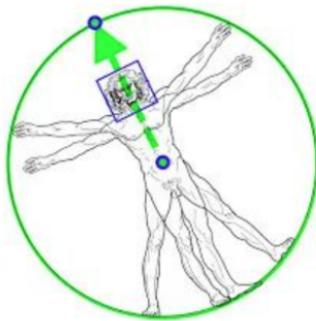
通过扩展 BlazeFace 进行姿态检测

为了实现由姿态检测和追踪模型组成的完整 ML 流水线的实时性能，每个组件都必须足够快，达到每帧仅使用几毫秒的程度。要实现此目标，对神经网络来说，我们发现关于躯干位置的最强信号是人的脸部（由于其高对比度特征和相对较小的外观变化）。因此，我们假设单人样例中头部应当可见，这个很强的先验假设（但对许多移动和网页应用有效）可以实现快速轻便的姿态检测器。

受亚毫秒级 BlazeFace 模型启发，我们训练了一个人脸检测器作为姿态检测器的代理。请注意，这个模型只能检测在一帧图像内人的位置，而不能用于识别个人。

与从预测关键点中得出 ROI 的 Face Mesh 和 MediaPipe 手部追踪流水线相比，对于人体姿态追踪，我们明确预测了两个额外的 虚拟 关键点，将人体中心、旋转和比例构建为一个圆。

受达芬奇的《维特鲁威人》所启发，我们预测了人的臀部中点、外接整个人的圆的半径以及连接肩部和臀部中点的直线的倾斜角度。这样一来，即使是非常复杂的情况，比如特定的瑜伽体式，也能得到一致的追踪。下图说明了这种方法。



维特鲁威人通过 BlazePose 检测器预测的两个虚拟关键点以及脸部边界框进行对齐

追踪模型

流水线姿势预测组件预测全部 33 个人体关键点的位置，每个关键点具有三个自由度（x、y 位置和可见度），额外加上上述两个虚拟对齐关键点。与当前采用计算密集型热力图预测的方法不同，我们的模型采用回归方法，由所有关键点的组合热力图/偏移预测监督。

具体来说，在训练过程中，首先采用热力图和偏移损失训练网络的中心和左塔。然后移除热力图输出并训练回归编码器（右塔），从而有效地利用热力图来监督轻量级嵌入向量。



追踪网络架构：热力图监督回归

结论

姿态识别的难点：多自由度(degrees of freedom)、遮挡

方法	优点	缺点
每个关节生成热图(produce heatmaps for each joint)以及细化每个坐标的偏移量	可以以最小的开销将其扩展到多个人	一个人的模型比适用于手机上的实时推断的模型大得多
基于回归(regression)的方法	对计算的要求较低且可扩展性较高	试图预测平均坐标值，但通常无法解决潜在的歧义
BlazePose 方法：即使参数数量较少，堆叠式沙漏架构(stacked hourglass architecture) 也可以大大提高预测的准确性。扩展该想法，并使用编码器-解码器网络体系结构(encoder-decoder network architecture) 来预测所有关节的热图，随后是另一个直接回归到所有关节坐标的编码器。	热图分支(heatmap branch)可以在推理过程中丢弃，从而使其轻巧到可以在手机上运行。 33 个关键点，使用面部特征数量	在单人用例中始终应看到人员的头部。

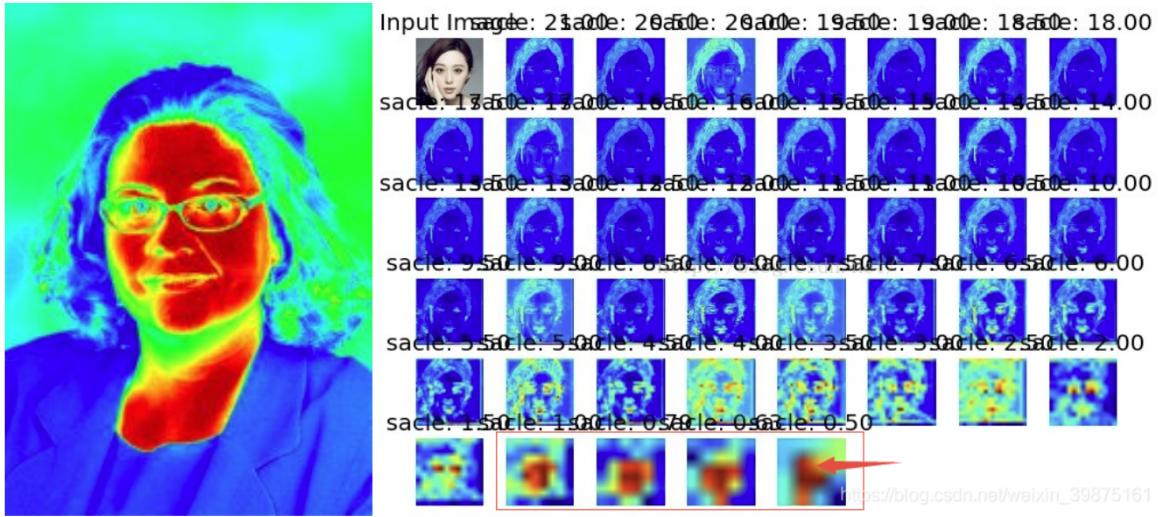
探究方向：扩展更多关键点、3D 支持、手势和面部几何估计模型

亮点：

1. 采用人脸检测器加上个别人体位置（肩膀，臀部中心）去做人体检测器。这些点位置比较稳定，变化比较少。
2. 训练用heatmap去做约束，之后finetune 和 前向 只考虑直接回归的分支
3. 人体关键点同时 预测出 置信度（可见度）。
4. 只做一次人脸检测，后一帧用上一帧人体关键点计算出的框。当点数少于某个阈值时，再用人脸检测器。

扩展：

图像的heatMap是什么，一副图片的heatmap可以帮助我们在上面检测到想要的object，如下左图所示：



可以直接的看到，人脸的区域有红色区域，然后用sliding window在图片上进行检测，对于每一个窗口里面的object进行识别，就是检测这个window里面的object是不是红色的区域，如果是就是检测到的人脸。这里我们首先看一下我们程序的结果：右上方的图片最后的那个区域就是我们想要的，那个红色的区域就是我们想要的。

具体的做法就是先在classification net上进行pre-train，之后去掉softmax层，改全部的fc层改成卷积层，之后把原来的那张图片输入网络，就会得到最终的heatmap，之所以有那么多的heatmap是因为对图片进行了不同程度的放缩。