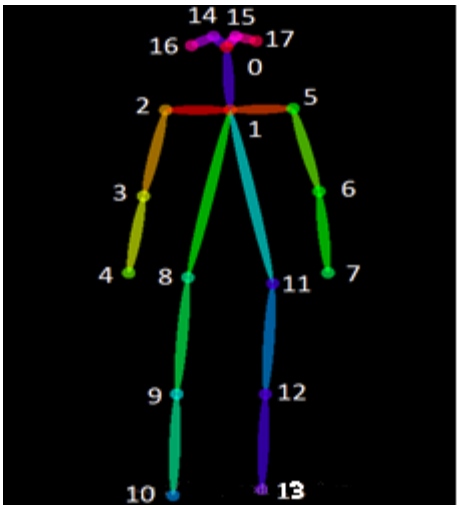
人体关键点检测

一、概要

人体关键点检测（Human Keypoints Detection）又称为人体姿态估计，是计算机视觉中一个相对基础的任务，是人体动作识别、行为分析、人机交互等的前置任务。一般情况下可以将人体关键点检测细分为单人/多人关键点检测、2D/3D 关键点检测，同时有算法在完成关键点检测之后还会进行关键点的跟踪，也被称为人体姿态跟踪。

目前 COCO keypoint track 是人体关键点检测的权威公开比赛之一，COCO 数据集中把人体关键点表示为 17 个关节，分别是鼻子，左右眼，左右耳，左右肩，左右肘，左右腕，左右臀，左右膝，左右脚踝。而人体关键点检测的任务就是从输入的图片中检测到人体及对应的关键点位置。



17 个人体关键点标注

二、标注格式

"keypoint_annotations": {
"human1": [261, 294, 1, 281, 328, 1, 0, 0, 0, 213, 295, 1, 208, 346, 1, 192, 335, 1, 245, 375, 1, 255, 432, 1, 244, 494, 1, 221, 379, 1, 219, 442, 1, 226, 491, 1, 226, 256, 1, 231, 284, 1],
"human2": [313, 301, 1, 305, 337, 1, 321, 345, 1, 0, 0, 0, 0, 0, 0, 0, 0, 313, 359, 1, 320, 409, 1, 311, 454, 1, 0, 0, 0, 330, 409, 1, 324, 446, 1, 337, 284, 1, 327, 302, 1],
"human3": [373, 304, 1, 346, 286, 1, 332, 263, 1, 0, 0, 0, 0, 0, 0, 345, 313, 1, 0, 0, 0, 0, 0, 0, 0, 0, 363, 386, 1, 361, 424, 1, 361, 475, 1, 365, 273, 1, 369, 297, 1]}
数列形式为：[x1,y1,v1,x2,y2,v2,· · ·,x14,y14,v14]，其中(xi,yi)为编号 i 的人体骨骼关节点的坐标位置，vi 为其状态（vi=1 可见，vi=2 不可见，vi=3 不在图内或不可推测）

三、评价指标

越大越好

目前最为常用的就是 OKS（Object Keypoint Similarity）指标，这个指标启发于目标检测中的 IoU 指标，目的就是为了计算真值和预测人体关键点的相似度。

OKS

$$OKS_p = \frac{\sum_i exp\{-d_{pi}^2/2S_p^2\sigma_i^2\}\delta(v_{pi}=1)}{\sum_i \delta(v_{pi}=1)}$$

其中：

p 表示groudtruth中，人的id

i 表示keypoint的id

d_{pi} 表示groudtruth中每个人和预测的每个人的关键点的欧氏距离

S_p 表示当前人的尺度因子，这个值等于此人在groundtruth中所占面积的平方根，即 $\sqrt{(x_2-x_1)(y_2-y_1)}$ 18个点的最小外接矩形

σ_i 表示第i个骨骼点的归一化因子，这个因此是通过对数据集中所有groundtruth计算的标准差而得到的，反映出当前骨骼点标注时候的标准差， σ 越大表示这个点越难标注。

v_{pi} 代表第p个人的第i个关键点是否可见

δ 用于将可见点选出来进行计算的函数

https://blog.csdn.net/m0_37163827/article/details/84887811

OKS矩阵

上面介绍的OKS是计算两个人之间的骨骼点相似度的，那一张图片中有很多的人时，该怎么计算呢？这时候就是构造一个OKS矩阵了。

假设一张图中，一共有M个人（groudtruth中），现在算法预测出了N个人，那么我们就构造一个M×N的矩阵，矩阵中的位置（i,j）代表groudtruth中的第i个人和算法预测出的第j个人的OKS相似度，找到矩阵中每一行的最大值，作为对应于第i个人的OKS相似度值。

AP（Average Precision）

根据前面的OKS矩阵，已经知道了某一张图像的所有人（groundtruth中出现的）的OKS分数，现在测试集中有很多图像，每张图像又有一些人，此时该如何衡量整个算法的好坏的。这个时候就用到了AP的概念，AP就是给定一个t，如果当前的OKS大于t，那就说明当前这个人的骨骼点成功检测出来了，并且检测对了，如果小于t，则说明检测失败或者误检漏检等，因此对于所有的OKS，统计其中大于t的个数，并计算其占所有OKS的比值。即假设OKS一共有100个，其中大于阈值t的共有30个，那么AP值就是30/100=0.3.

mAP（mean Average Precision）

顾名思义，AP的均值，具体计算方法就是给定不同的阈值t，计算不同阈值情况下对应的AP，然后求个均值就ok了。

单人姿态估计AP:

单人姿态估计，一次仅对一个行人进行估计，即在**oks**指标中$M=1$，因此一张图片中groundtruth为一个行人(GT)，对此行人进行关键点检测后会获得一组关键点(DT)，最后会计算出GT与DT的相似度**oks**为一个标量，然后人为的给定一个阈值**T**，然后可以通过所有图片的**oks**计算**AP**:

$$AP=\frac{\sum_p \delta(oks_p > T)}{\sum_p 1}$$

多人姿态估计AP:

多人姿态估计，如果采用的检测方法是自顶向下，先把所有的人找出来再检测关键点，那么其**AP**计算方法如同**单人姿态估计AP**； 如果采用的检测方法是自底向上，先把所有的关键点找出来然后再组成人，那么假设一张图片中共有**M**个人，预测出**N**个人，由于不知道预测出的**N**个人与groundtruth中的**M**个人的一一对应关系，因此需要计算groundtruth中每一个人与预测的**N**个人的**oks**，那么可以获得一个大小为$M \times N$的矩阵，矩阵的每一行为groundtruth中的一个人与预测结果的**N**个人的**oks**，然后找出每一行中**oks**最大的值作为当前**GT**的**oks**。最后每一个**GT**行人都有一个标量**oks**，然后人为的给定一个阈值**T**，然后可以通过所有图片中的所有行人计算**AP**:

$$AP=\frac{\sum_m \sum_p \delta(oks_p > T)}{\sum_m \sum_p 1}$$

如果是自底向上的话，是没有分人标注点，判断方法见12页

四、数据集

LSP（Leeds Sports Pose Dataset）：单人人体关键点检测数据集，关键点个数为 14，样本数 2K，在目前的研究中作为第二数据集使用。

FLIC（Frames Labeled In Cinema）：单人人体关键点检测数据集，关键点个数为 9，样本数 2W，在目前的研究中作为第二数据集使用。

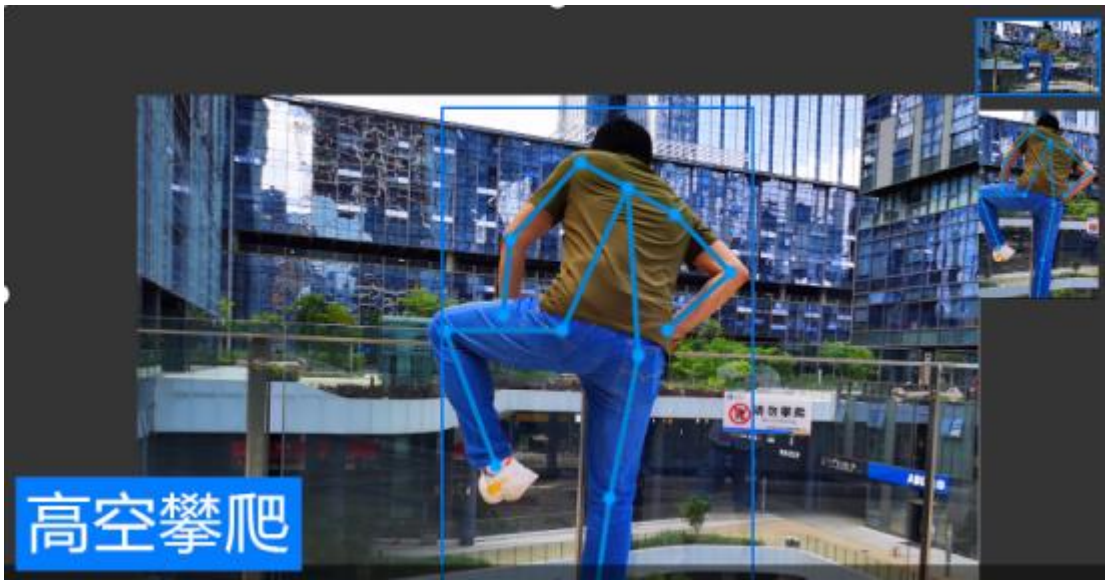
MPII（MPII Human Pose Dataset）：单人/多人人体关键点检测数据集，关键点个数为 16，样本数 25K，是单人人体关键点检测的主要数据集。

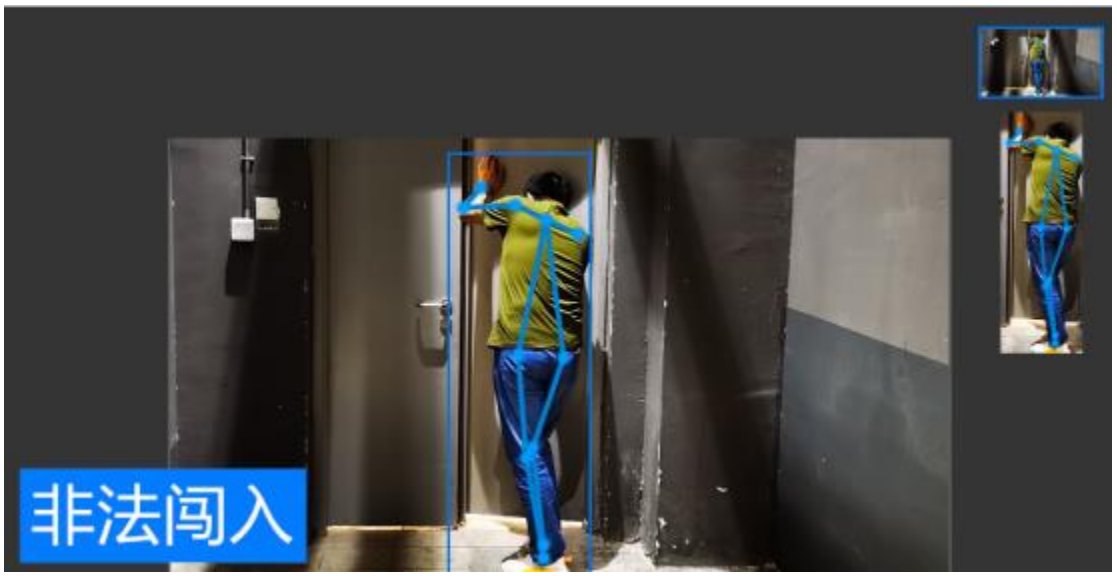
MSCOCO：多人人体关键点检测数据集，关键点个数为 17，样本数多于 30W，多人关键点检测的主要数据集，主流数据集；

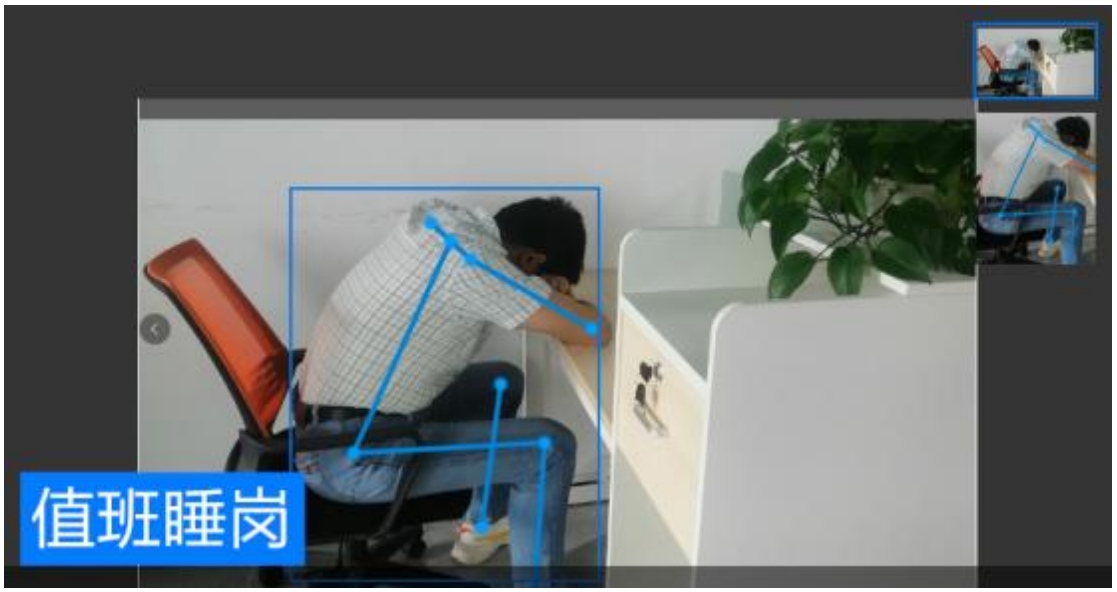
AI Challenger：多人人体关键点检测数据集，关键点个数为 14，样本数约 38W，竞赛数据集；

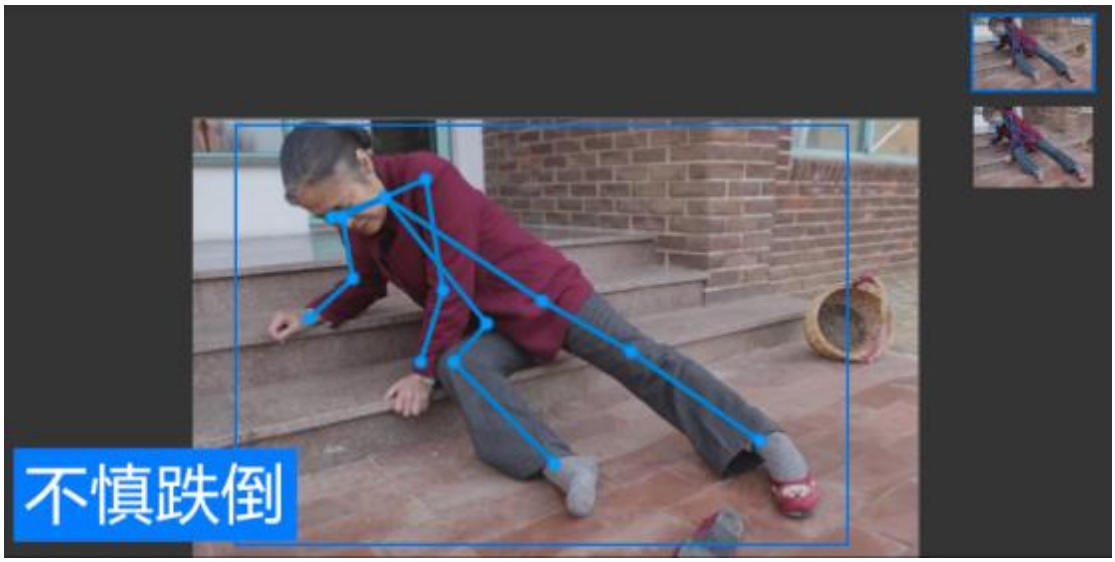
五、应用场景

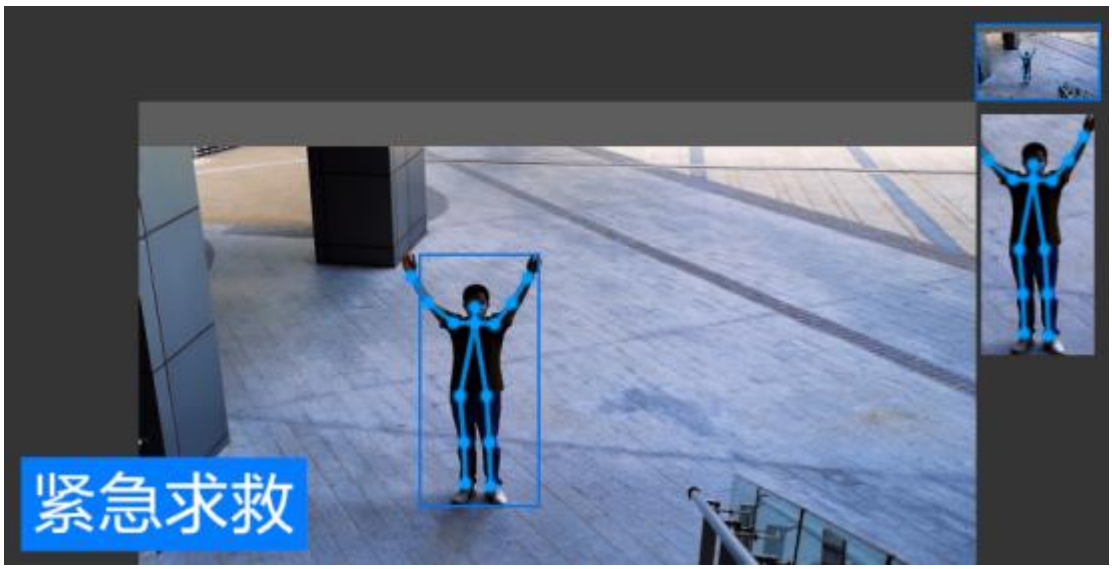
人体关键点检测可广泛应用 AI 人体行为分析、人物跟踪、视觉分析步态识别等相关领域。主要集中在智能视频监控，人机交互，人体动画，智能安防行业，运动员辅助训练等领域。

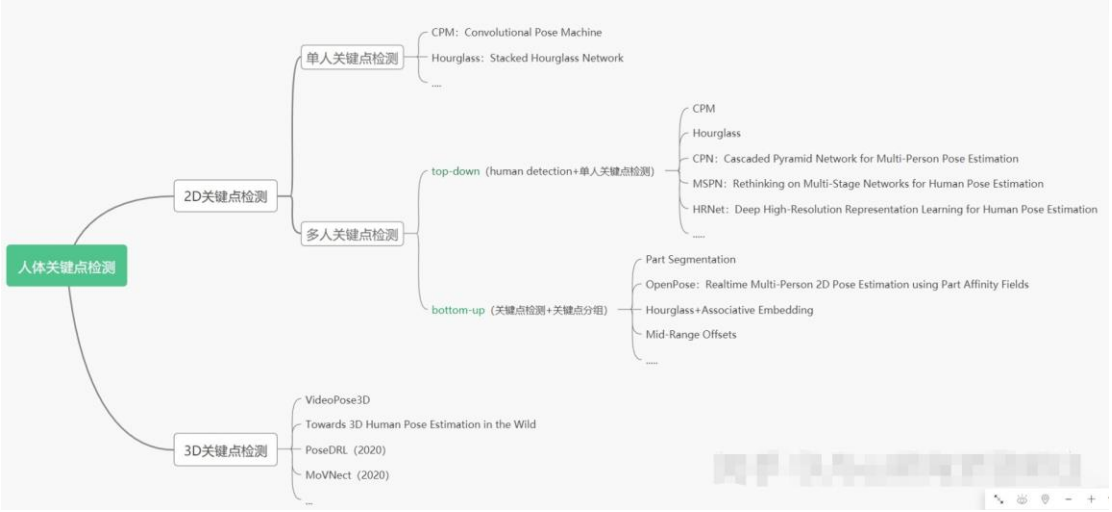












六、Ground Truth 的构建

Ground Truth 的构建问题，主要有两种思路，Coordinate 和 Heatmap，Coordinate 即直接将关键点坐标作为最后网络需要回归的目标，这种情况下可以直接得到每个坐标点的直接位置信息；Heatmap 即将每一类坐标用一个概率图来表示，对图片中的每个像素位置都给一个概率，表示该点属于对应类别关键点的概率，比较自然的是，距离关键点位置越近的像素点的概率越接近 1，距离关键点越远的像素点的概率越接近 0，具体可以通过相应函数进行模拟，如 Gaussian 等，如果同一个像素位置距离不同关键点的距离大小不同，即相对于不同关键点该位置的概率不一样，这时可以取 Max 或 Average。对于两种 Ground Truth 的差别，Coordinate 网络在本质上来说，需要回归的是每个关键点的一个相对于图片的 offset，而长距离 offset 在实际学习过程中是很难回归的，误差较大，同时在训练中的过程，提供的监督信息较少，整个网络的收敛速度较慢；Heatmap 网络直接回归出每一类关键点的概率，在一定程度上每一个点都提供了监督信息，网络能够较快的收敛，同时对每一个像素位置进行预测能够提高关键点的定位精度，在可视化方面，Heatmap 也要优于 Coordinate，除此之外，实践证明，Heatmap 确实要远优于 Coordinate。

七、跌倒识别

主要目的：检测画面中摔倒行人数量，主要用在空巢老人摔倒检测，大型地铁商场的自动扶梯摔倒检测以及地震火灾等自然灾害后搜救机器人搜索摔倒伤员。检测出行人的 21 个身体关键点，然后利用一些列精心设计的规则判断是否摔倒 规则： 根据测试集数据分析，设置了以下几种规则

- A. 所检测行人的 bbox 的宽高比限制（宽>0.7 * 高）
- B. 人体弯曲度：检测行人双肩中心，双胯中点，双膝中点，然后组成两个向量，分别是双胯中点指向双肩膀中点和双胯中点指向双膝中点，计算两者的夹角余弦作为人体弯曲度，大于 110 则很大可能有摔倒现象。

- C. 检测头和脚的 y 坐标进行比较

最终摔倒的判断规则是：（A & B）| C .这种方法使得大部分测试集图像都能准确判断摔倒行人

人体姿态估计知识点补充

AI 识别人的五个状态: 图像中是否有人，人在哪里，人是谁，人处于什么状态，人在做什么；

姿态估计: 通过将人体肢体关节联系起来进行人体姿态估计，通过姿态估计可以判断人的状态和行为；

本文提出一种实时检测多人 2D 姿态的方法；

采用非参数表征方法 Part Affinity Fields(PAFs 部分亲和度向量场)，去学习将身体部位和对应个体关联；

提出组合检测器可以减少推理时间，推出 OpenPose，多人 2D 姿态检测开源实时系统，包括身体，脚部，手部和面部关键点的检测；

引言

常用的姿态估计的方法:行人检测 + 担任姿态估计(自上而下的方法)，缺点:姿态估计受到行人检测的影响，运行时间和人数成比例；

自下而上的方法在前期的鲁棒性更稳健，还可能将运行时复杂性和图像中的人数进行分离；

OpenPose 提出多人姿态估计方法:通过 PAFs 来表征第一个自下而上的特征表示，PAFs 是 2D 适量场用于编码肢体在图像域的位置和方向；证明同时计算自下而上的检测和关联编码，能够为后续的解析过程提供足够的全局上下文，同时做到小部分计算成本和高质量结果；

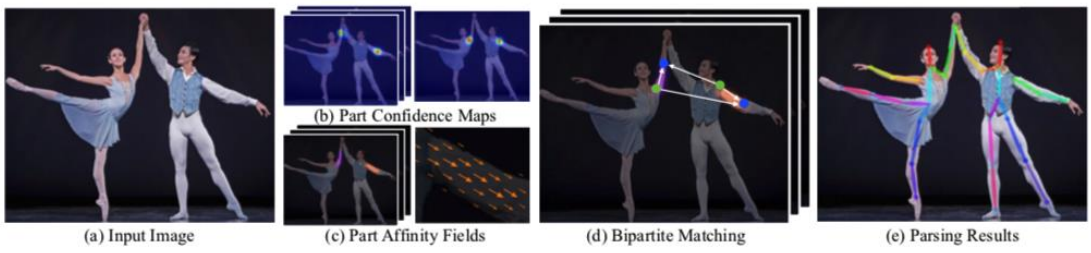


Fig. 2: Overall pipeline. (a) Our method takes the entire image as the input for a CNN to jointly predict (b) confidence maps for body part detection and (c) PAFs for part association. (d) The parsing step performs a set of bipartite matchings to associate body part candidates. (e) We finally assemble them into full body poses for all people in the image.

Fig. 1: **Top:** Multi-person pose estimation. Body parts belonging to the same person are linked, including foot keypoints (big toes, small toes, and heels). **Bottom left:** Part Affinity Fields (PAFs) corresponding to the limb connecting right elbow and wrist. The color encodes orientation. **Bottom right:** A 2D vector in each pixel of every PAF encodes the position and orientation of the limbs.

- 图2是本文方法整体的pipeline:
  - 输入$w \times h$图像a，为每一个人产生2D关键点定位e;
  - 实现CNN预测一组身体部位的2D置信图b S，和一组2D PAFs L对部位之间的联系进行编码图c;
  - 集合$\mathbf{S} = (\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_J)$有$J$个置信图，每个部分一置信图，其中$\mathbf{S}_j \in \mathbb{R}^{w \times h}, j \in \{1 \ldots J\}$;
  - 集合$\mathbf{L} = (\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_C)$有$C$个矢量场，每个肢体一矢量场，其中$\mathbf{L}_c \in \mathbb{R}^{w \times h \times 2}, c \in \{1 \ldots C\}$
  - 图像位置L_C是一个编码后的2D vector如图一所示(目前到这里对网络的输出信息理解还是不清晰...看之后的详细理解了)，最后通过贪心推理解析置信图和PAF输出所有人的2D关键点;
  - 总结一下整体流程: a输入图像，b预测关键点置信度&c关键点亲和度向量，关键点解析，人体骨骼搭建(连接关键点)

Simultaneous Detection and Association

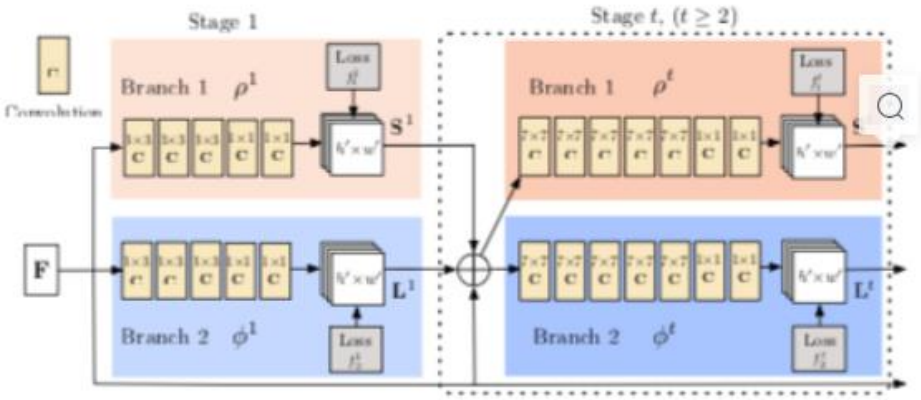


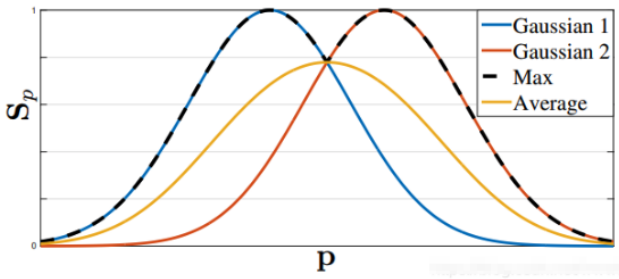
Figure 3. Architecture of the two-branch multi-stage CNN. Each stage in the first branch predicts confidence maps $\mathbf{S}^t$, and each stage in the second branch predicts PAFs $\mathbf{L}^t$. After each stage, the predictions from the two branches, along with the image features are concatenated for next stage.

- 从上图3可知，网络的backbone F 为VGG19的前10层，然后再将网络分为两个分支米色部分和蓝色部分，分别预测每个点的关键点置信度和亲和度向量，每个分支都是一个迭代预测架构；
- 首先由VGG19生成一组特征 F ,作为每个分支第一阶段的输入；
- 第一阶段网络产生一组检测置信度图 $\mathbf{S}^1 = \rho^1(\mathbf{F})$,和一组亲和度向量 $\mathbf{L}^1 = \phi^1(\mathbf{F})$, 其中 ρ^1 和 ϕ^1 第一阶段inference的CNN结构，之后的每一个阶段的输入都来自前一个阶段的预测结果和原始图像特征 F ，用以产生更精确的预测结果；
- ρ^t 和 ϕ^t 代表第t阶段的CNN结构: 其输出为 $\mathbf{S}^t = \rho^t\left(\mathbf{F}, \mathbf{S}^{t-1}, \mathbf{L}^{t-1}\right), \forall t \geq 2$ 和 $\mathbf{L}^t = \phi^t\left(\mathbf{F}, \mathbf{S}^{t-1}, \mathbf{L}^{t-1}\right), \forall t \geq 2$

Figure 4. Confidence maps of the right wrist (first row) and PAFs (second row) of right forearm across stages. Although there is confusion between left and right body parts and limbs in early stages, the estimates are increasingly refined through global inference in later stages, as shown in the highlighted areas.

- 上图4可以看出前期身体左右部分和四肢之间会比较混乱，但是通过后期各stage的迭代推理后，预测结果越来越精确；
- 每个阶段t对应两个损失函数$f_S^t = \sum_{j=1}^{J} \sum_{\mathbf{p}} \mathbf{W}(\mathbf{p}) \cdot \left\| \mathbf{S}_j^t(\mathbf{p}) - \mathbf{S}_j^*(\mathbf{p}) \right\|_2^2$和$f_L^t = \sum_{c=1}^{C} \sum_{\mathbf{p}} \mathbf{W}(\mathbf{p}) \cdot \left\| \mathbf{L}_c^t(\mathbf{p}) - \mathbf{L}_c^*(\mathbf{p}) \right\|_2^2$代表预测值和GT值的L2损失，其中$\mathbf{S}_j^*$代表真实的置信图,$\mathbf{L}_c^*$代表真实的身体关节联系向量，损失函数整体为$f = \sum_{t=1}^{T} (f_S^t + f_L^t)$，$j$代表关键点，$c$代表肢体，一个肢体对应两个关键点

Confidence Maps for Part Detection



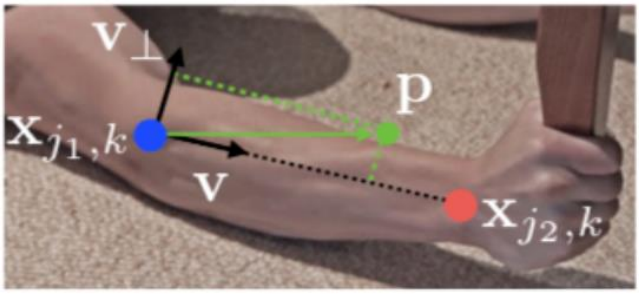
- 为了在训练阶段评估f_S,根据标注了2D的关键点的图生成GT置信图，每一个置信图表示身体的一个特定的部位在图像上某点发生的可能性，如果图像只有一个人每个置信图理论上只有一个峰值当图像中有多人时，对应每一个人k都对应一个可见的身体部位j的峰值；
- 首先对每一个人k生成个人的所有置信图$\mathbf{S}_{j,k}^*$,$\mathbf{x}_{j,k} \in \mathbb{R}^2$代表第$k$个人身体部位j对应的GT位置，在$p$点的值被定义为$\mathbf{S}_{j,k}^*(\mathbf{p}) = \exp\left(-\frac{\|\mathbf{p}-\mathbf{x}_{j,k}\|_2^2}{\sigma^2}\right)$,其中$\sigma$用于控制峰值的范围；
- 网络在位置P的预测值对应的GT值计算是如上图所示是取最大值$\mathbf{S}_j^*(\mathbf{p}) = \max_k \mathbf{S}_{j,k}^*(\mathbf{p})$，NMS思想，在预测阶段网络通过NMS获得最终的置信度；

Part Affinity Fields for Part Association

(a) (b) (c)

Figure 5. Part association strategies. (a) The body part detection candidates (red and blue dots) for two body part types and all connection candidates (grey lines). (b) The connection results using the midpoint (yellow dots) representation: correct connections (black lines) and incorrect connections (green lines) that also satisfy the incidence constraint. (c) The results using PAFs (yellow arrows). By encoding position and orientation over the support of the limb, PAFs eliminate false associations.

- 这小节解决了一个很重要的问题，在学习姿态估计论文之前可能会想到，人体的关键点检测出来后该怎么样建立彼此之间正确的联系呢，尤其是在多人姿态估计的时候，上图5a是关键点之间可能存在的关系，图b代表检测每对身体部分之前的肢体中间点但是缺点是只编码了点的位置信息缺少方向信息，同时它将肢体的支撑区域缩小到了一个点；
- 本文提出PAFs部分亲和场同时保持肢体区域之间的位置信息和方向信息，图c代表PAF是身体每个肢体的2D向量，图1d所示对于属于特定肢体的像素，2D向量编码了从肢体一个部分指向另一个部分的方向，每种类型的肢体都有连接其两个关联部位对应的亲和场；



- 通过一个简单的例子来分析这部分的内容，$\mathbf{x}_{j_1,k}$ and $\mathbf{x}_{j2}, k$分别代表k的肢体c两个对应的身体部位$j_1 and j_2$的GT位置，如果点p落在了肢体c上，$\mathbf{L}^*_{c,k}(\mathbf{p})$的值为$j_1$指向$j_2$的单位向量，不在这个肢体上的点值为0；
- 为了在训练阶段评估f_L,定义PAF在点p的GT值为: $\mathbf{L}^*_{c,k}(\mathbf{p}) = \begin{cases} \mathbf{v} & \text{if } \mathbf{p} \text{ on limb } c,k \\ \mathbf{0} & \text{otherwise} \end{cases}$,其中$\mathbf{v} = (\mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k}) / \|\mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k}\|_2$代表肢体方向的单位向量；
- 在$0 \le \mathbf{v} \cdot (\mathbf{p} - \mathbf{x}_{j_1,k}) \le l_{c,k}$ and $|\mathbf{v}_\perp \cdot (\mathbf{p} - \mathbf{x}_{j_1,k})| \le \sigma$范围内的点$p$被定义为在肢体$c$上，其中$\sigma$代表肢体的宽度，$l_{c,k} = \|\mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k}\|_2$代表肢体的长度；
- 点p的部分亲和力场GT值为所有人在此点上PAF平均值：$\mathbf{L}^*_c(\mathbf{p}) = \frac{1}{n_c(\mathbf{p})} \sum_k \mathbf{L}^*_{c,k}(\mathbf{p})$,其中$n_c(p)$代表非零向量的个数；
- 在预测阶段，对于两个候选的部位点d_{j1}和d_{j2}，我们沿着线段采样预测得到的PAF L_c，以测量两个部分之间的关联置信度，$E = \int_{u=0}^{u=1} \mathbf{L}_c(\mathbf{p}(u)) \cdot \frac{\mathbf{d}_{j_2} - \mathbf{d}_{j_1}}{\|\mathbf{d}_{j_2} - \mathbf{d}_{j_1}\|_2} du$,其中$p(u)$代表两个身体部位之间的位置:$\mathbf{p}(u) = (1-u)\mathbf{d}_{j_1} + u\mathbf{d}_{j_2}$，实际预测时对u区间进行均匀间隔采样求和来求解近似的积分值；

Multi-Person Parsing using PAFs

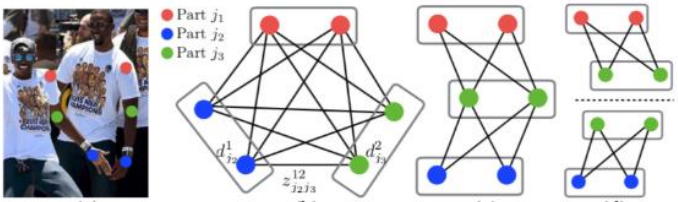


Figure 6. Graph matching. (a) Original image with part detections (b) K-partite graph (c) Tree structure (d) A set of bipartite graphs

- 对预测的置信度图进行nms操作后可以得到一组离散的候选身体部位，对于每一个部位存在多个候选，因为图像上有多个人或者存在FP的情况，从这些候选部位可以定义一个很大的可能肢体集合，通过上面的积分公式计算每一个候选肢体的分数；
- 本文提出greedy relaxation方法来产生高质量的匹配:
  - 首先根据预测置信图的得到离散的候选部位:$\mathcal{D}_\mathcal{J} = \{\mathbf{d}^m_j : j \in \{1 \dots J\}, m \in \{1 \dots N_j\}\}$，其中$d^m_j$代表第$j$类身体部位的第$m$个关键点的位置；
  - 我们的匹配目标是要求候选部位和同一个人的其他候选部位建立连接，定义变量$z^{mn}_{j_1 j_2} \in \{0,1\}$用来表示两个候选部位$\mathbf{d}^m_{j_1}$ and $\mathbf{d}^n_{j_2}$之间是否有连接,所有候选部位的连线集合为$\mathcal{Z} = \{z^{mn}_{j_1 j_2} : \text{for } j_1, j_2 \in \{1 \dots J\}, m \in \{1 \dots N_{j_1}\}, n \in \{1 \dots N_{j_2}\}\}$
  - 单独考虑肢体c所有对应的两个身体部位j_1和j_2，目的是找到总亲和值最高的图匹配方式，定义总亲和值为: $\max_{\mathcal{Z}_c} E_c = \max_{\mathcal{Z}_c} \sum_{m \in \mathcal{D}_{j_1}} \sum_{n \in \mathcal{D}_{j_2}} E_{mn} \cdot z^{mn}_{j_1 j_2}$，其中 $\begin{matrix} \forall m \in \mathcal{D}_{j_1}, \sum_{n \in \mathcal{D}_{j_2}} z^{mn}_{j_1 j_2} \le 1 \\ \forall n \in \mathcal{D}_{j_2}, \sum_{m \in \mathcal{D}_{j_1}} z^{mn}_{j_1 j_2} \le 1 \end{matrix}$，$E_{mn}$代表$d^m_{j1}$和$d^n_{j2}$之间的亲和度，注意:同类型的两个肢体没有公共点,匈牙利算法寻找最优匹配
  - 当考虑到多人的全身姿态估计时，就是一个K分图匹配问题，可以简化优化为$\max_\mathcal{Z} E = \sum_{c=1}^{C} \max_{\mathcal{Z}_c} E_c$，人体各肢体独立优化配对，然后将享有相同身体部分的链接组装成人体的全身姿态；

总结

姿态估计的意义：让机器能够理解图像中人的行为，让机器具有感知个人的行为的能力；

本文为解决 2D 图像中多人的姿态估计问题，提出一个非参数的通过编码人体四肢的位置和方向的关键点连接方法并设计了一个可以同时学习身体部位位置和联系的架构，并用高效率高质量的方法产生人体姿态检测的结果，最重要的是随着图像中人数的增加依然能保持算法的优势