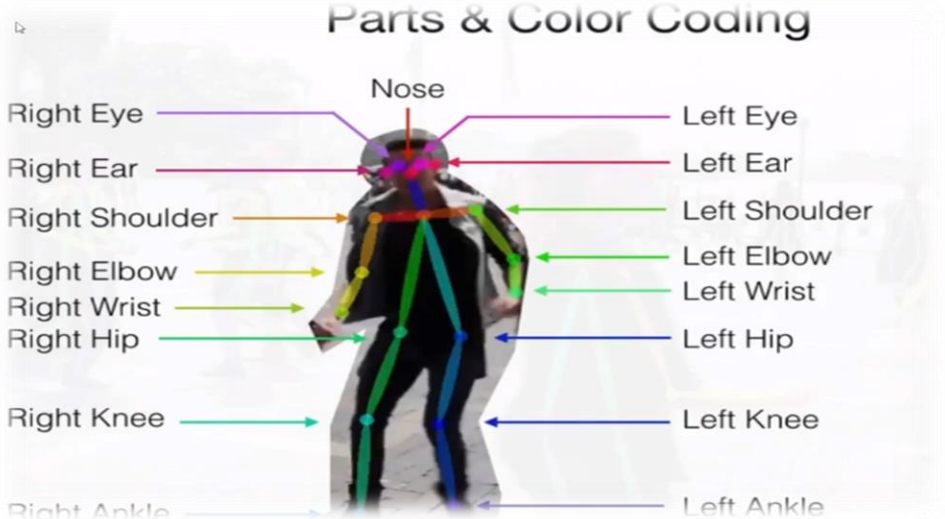
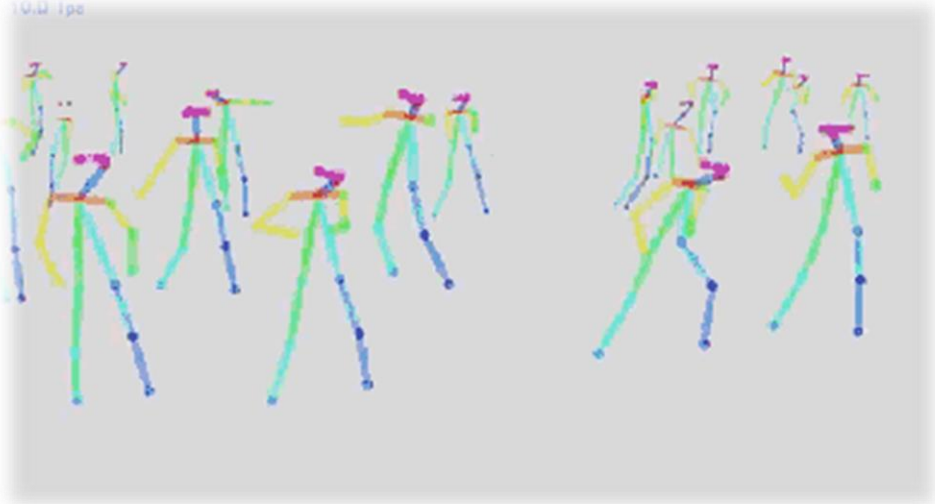
Realtime Multi-person 2D Pose estimation using Part Affinity Fields

实时多人人体姿态识别

研究目标

给定一张RGB图像，得到所有人体关键点的位置信息，同时知道每一个关键点是属于图片中具体哪个人的，即得到关键点之间的连接信息。

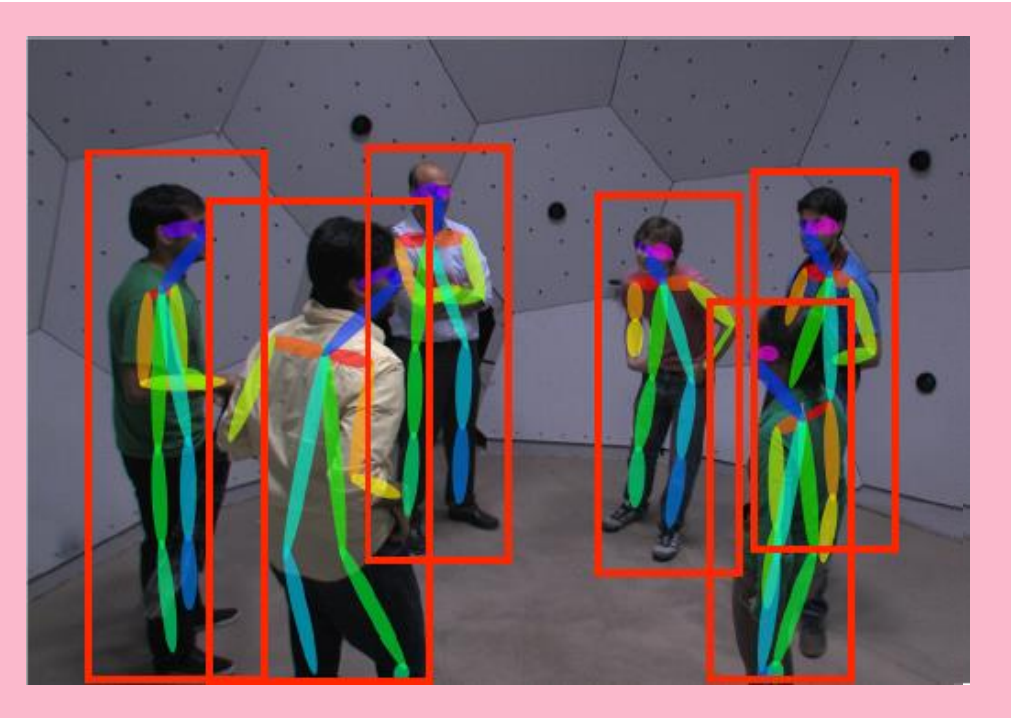




效果展示

实时的在视频上对每一帧图像进行检测的结果，并没有做任何的目标追踪，或其他利用时间上的连续性加速这个结果。

1 与传统多人人体姿态估计方法的对比



Top-down Approach:

Person Detection + Pose Estimation

缺陷:

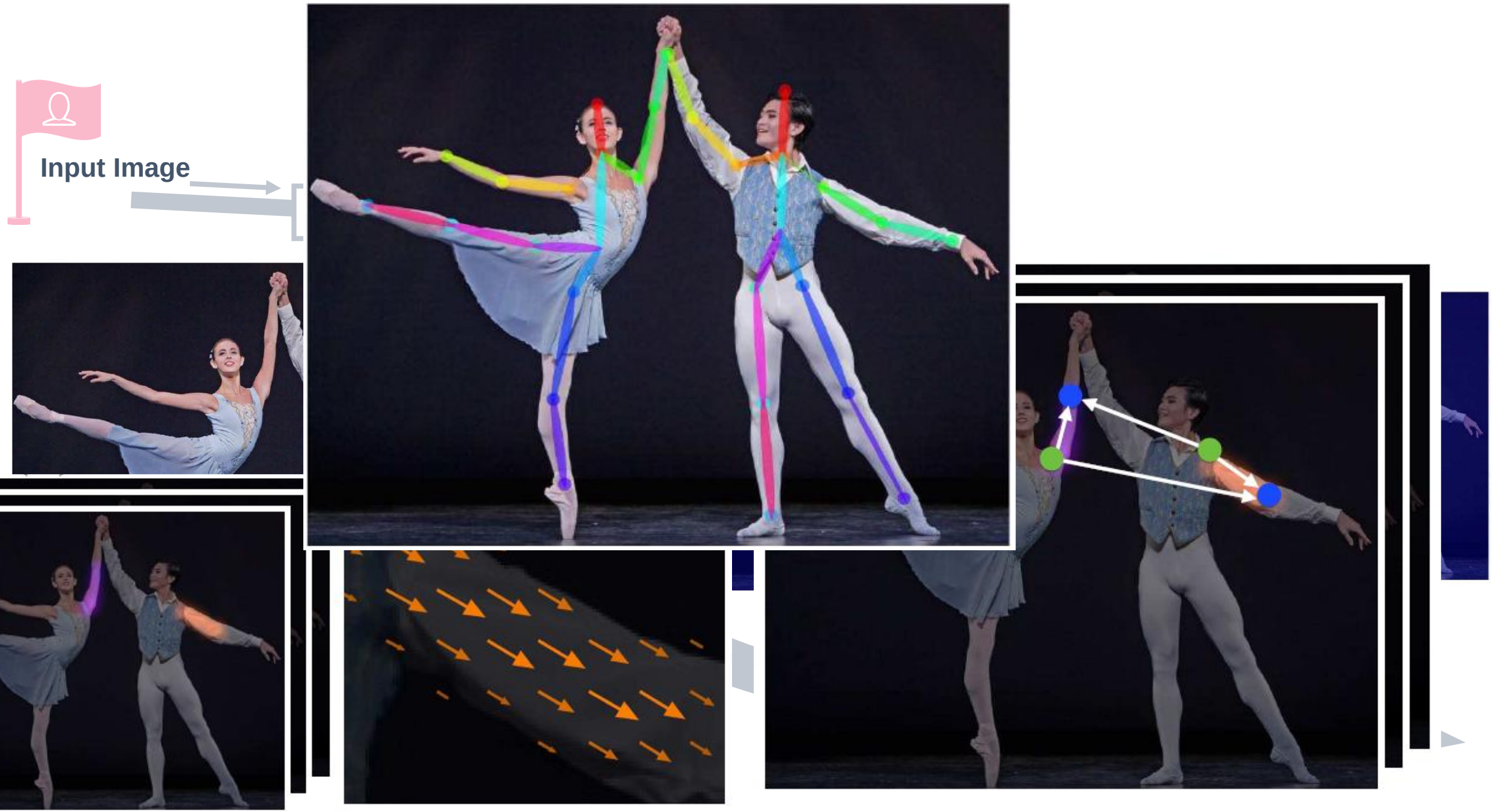
- 依赖人体姿态检测的结果
- 算法速度与图片中人的数目成正比

Down-top Approach:

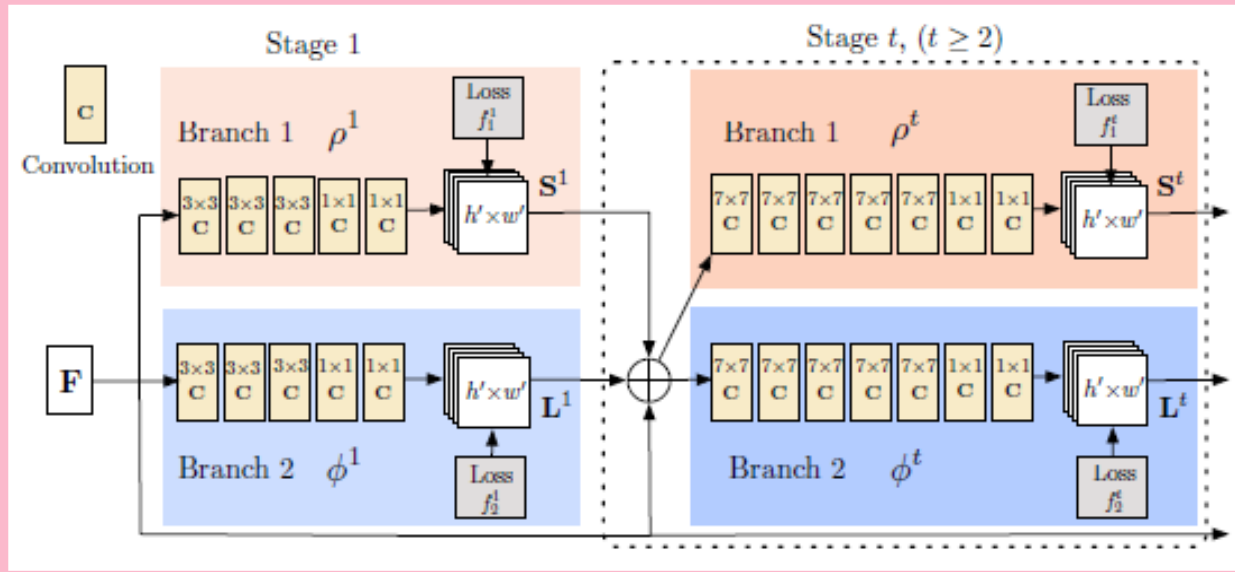
Parts Detection + Parts Association

- 检测：预测人体关键点的热点图
- 连接：人体关键点亲和场（PAF）

2 Jointly Learning Parts Detection and Parts Association



3 Simultaneous Detection and Association



两分支多阶段结构图

输出：

- 身体各部分位置信息的二维置信图集S

$S = (S_1, S_2, \ldots, S_J) \quad S_J \in R^{w \times h}, j \in \{1 \ldots J\}$

有J个置信图，每个关键点（点）对应一个S_J

- 身体各部分亲和力的二维矢量场L

$L = (L_1, L_2, \ldots, L_C) \quad L_C \in R^{w \times h \times 2}, c \in \{1 \ldots C\}$

有C个矢量场，每个肢体（线）对应一个L_C

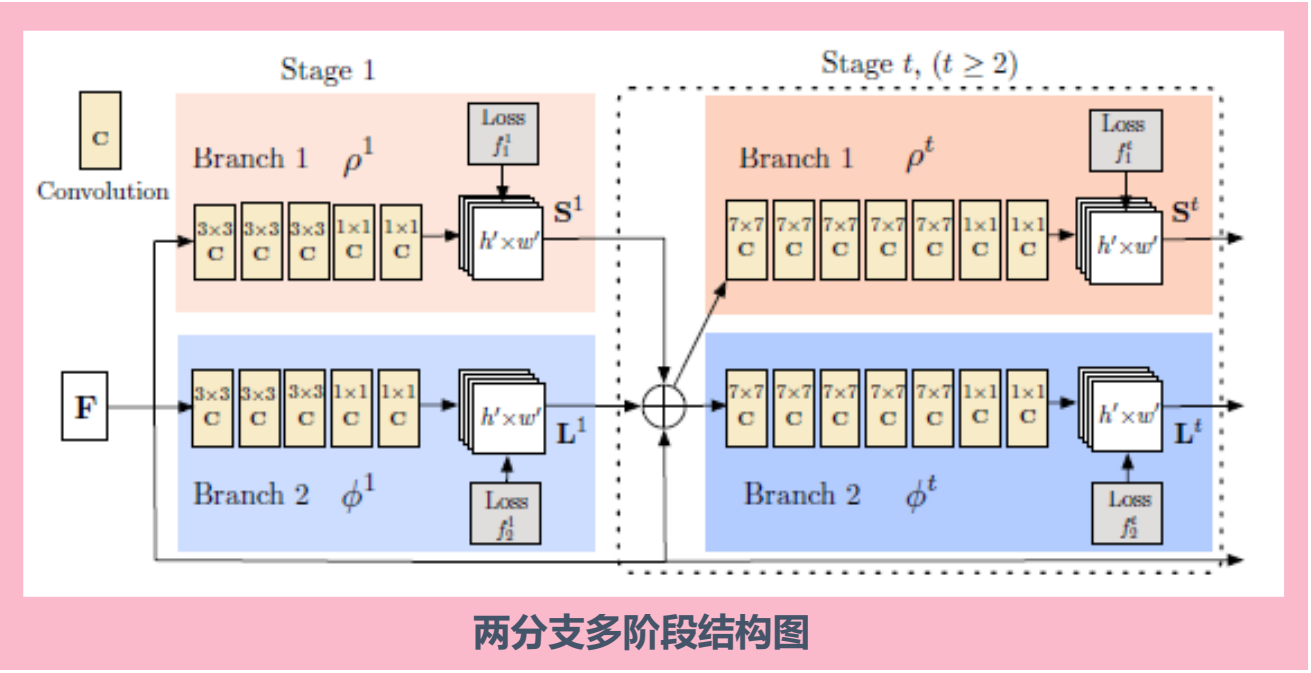
每个分支都是一个迭代的预测架构，在Convolutional Pose Machines（CPM）这篇论文的基础上，通过连续的阶段对预测结果进行优化，$t \in \{1, \ldots T\}$,并且在每个阶段都有中继监督（解决梯度消失的问题）。

图像首先被一个卷积神经网络处理后生成一个特征图集F（通过VGG-19的前10层进行初始化并微调）输入给每个分支的第一阶段。

第一阶段结果：$S^1 = \rho^1(F), t = 1, \quad L^1 = \varphi^1(F), t = 1$

第 t 阶段结果：$S^t = \rho^t(F, S^{t-1}, L^{t-1}), \forall t \geq 2, \quad L^t = \varphi^t(F, S^{t-1}, L^{t-1}), \forall t \geq 2$

3 Simultaneous Detection and Association



两分支多阶段结构图

输出：

- 身体各部分位置信息的二维置信图集S

$S = (S_1, S_2, \ldots, S_J) \quad S_J \in R^{w \times h}, j \in \{1 \ldots J\}$

有J个置信图，每个关键点（点）对应一个S_J

- 身体各部分亲和力的二维矢量场L

$L = (L_1, L_2, \ldots, L_C) \quad L_C \in R^{w \times h \times 2}, c \in \{1 \ldots C\}$

有C个矢量场，每个肢体（线）对应一个L_C

$$f = \sum_{t=1}^{T} (f_S^t + f_L^t)$$

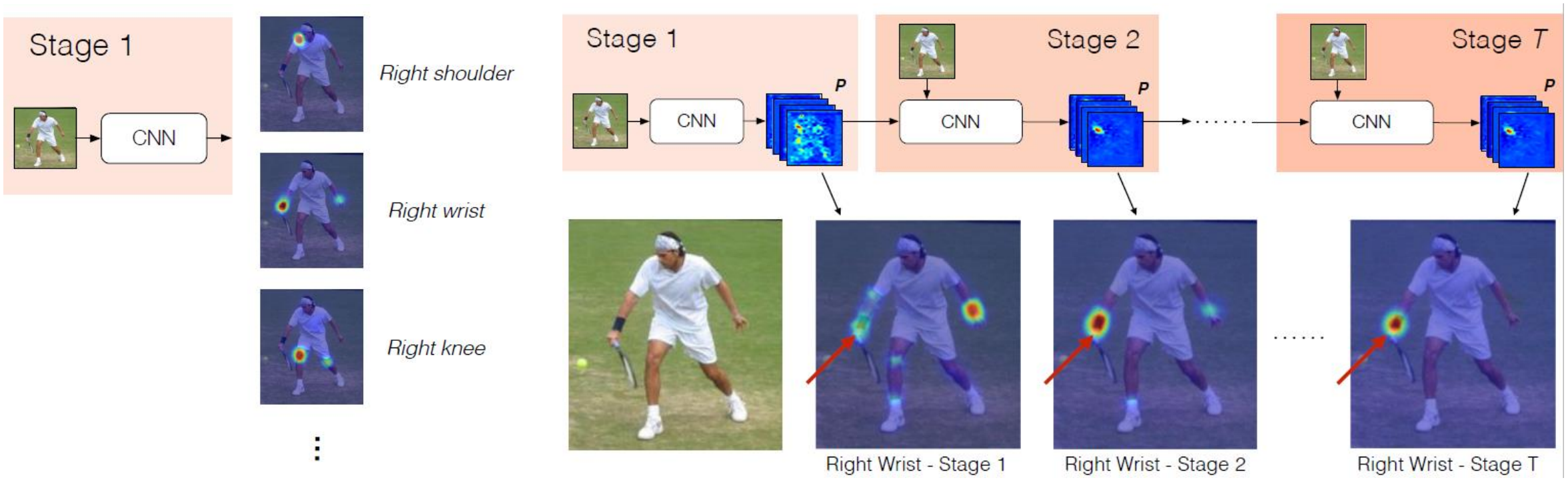
为了指导该神经网络，对两个分支各应用一个损失函数，并应用于每个阶段的结尾。

此处，为了解决某些数据集中不能完全标注所有人的情况，对损失函数进行了一个加权，每阶段的损失函数如下：

$$f_S^t = \sum_{j=1}^{J} \sum_{p} W(p) \cdot \left\| S_j^t(p) - S_j^*(p) \right\|_2^2 \qquad f_L^t = \sum_{c=1}^{C} \sum_{p} W(p) \cdot \| L_c^t(p) - L_c^*(p) \|_2^2$$

$S_j^*(p), L_c^*(p)$是真实的标注值，W非0即1，当图像中的p处没有标注时，则$W(p) = 0$，避免惩罚被模型预测为正的正样本

4 Confidence Maps for Part Detection



利用Convolutional Pose Machines预测人体关键点的位置。

4 Confidence Maps for Part Detection

- 为较好地得到损失函数f_S,通过标注二维的关键点生成了标注的真实置信图S^*（标定数据）。
- 每一个置信图都是对在每一个像素位置处出现的特定的身体关键点的信度的表达。

在测试时，通过非极大值抑制方法（搜索局部极大值，抑制非极大值元素）得到可能的身体关键点。

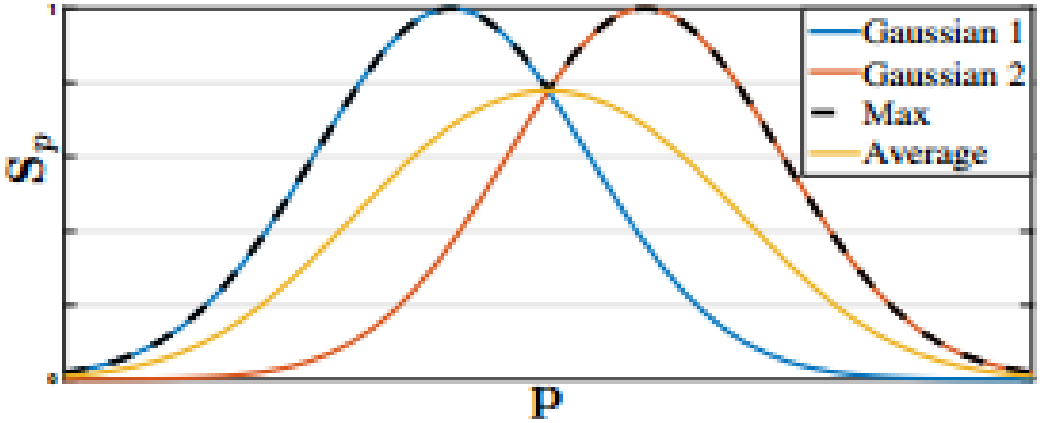
$S^*_{j,k}$在位置$p \in R^2$的值定义为：

$$S^*_{j,k}(\mathrm{p}) = e^{-\frac{\|p-x_{j,k}\|_2^2}{\sigma^2}} \qquad x_{j,k} \in R^2$$

$x_{j,k}$表示图像中第k个人的第j个身体关键点的标注真实值

产生多条曲线，进行一个取最大值的操作：

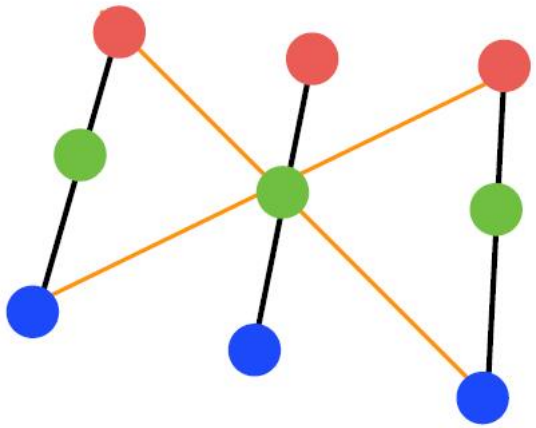
$$S^*_j(p) = \max_k S^*_{j,k}(p)$$



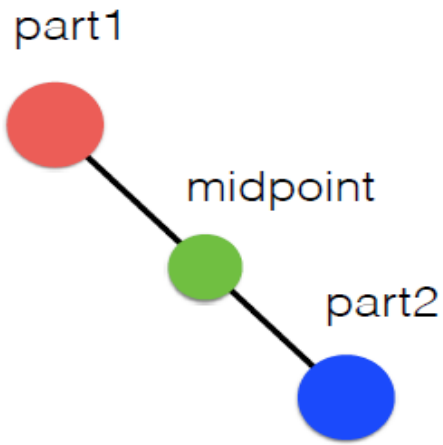
Q:为什么取最大值而不是取平均值?

5 Part Affinity Field for Part Association

如何把检测出的身体关键点组合成未知数目人的整体动作呢?



— Correct Connection
— Wrong Connection



Affinity score between part1 and part2
= confidence score of the midpoint

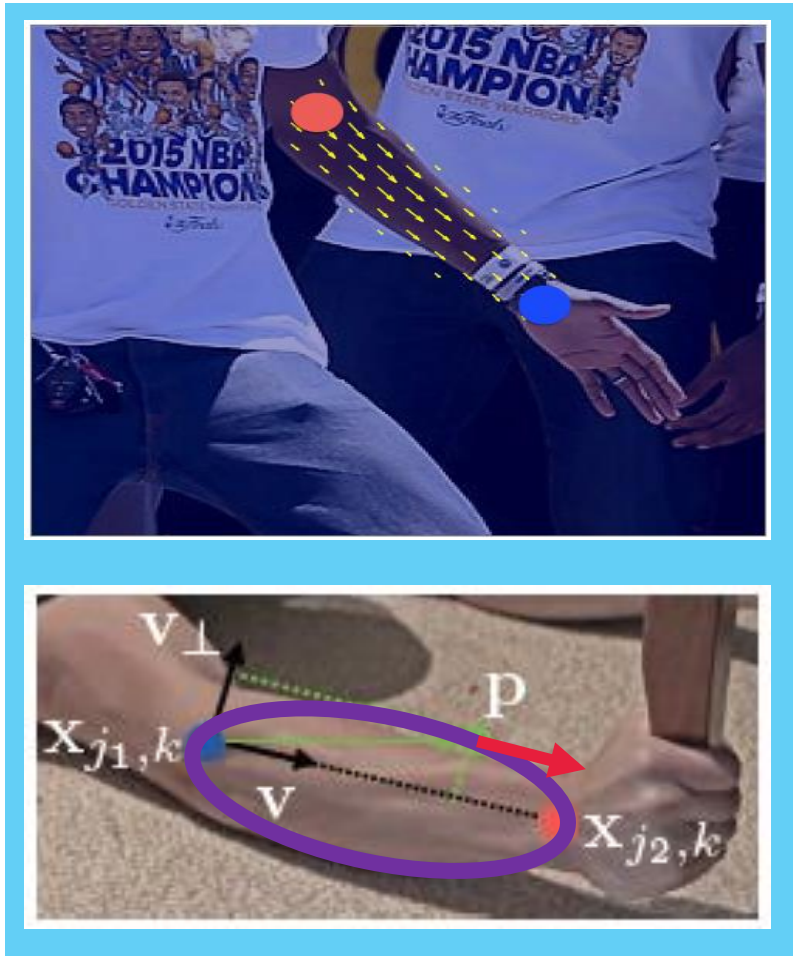
取中点连接法:

预测每个关键点之间的中点的热点图谱，假设有这个预测结果和两个关键点的位置，连接的中点在热点图谱对应的像素点的响应值作为这个连接的确信值。

局限性:

- 只考虑了每个肢体的位置信息，而没有肢体的旋转信息
- 将肢体的支撑范围缩小到了一个点上

5 Part Affinity Field for Part Association



为解决上述局限性，采用PAFs能够在整个肢体区域中保留位置和旋转信息。

人体关键部分亲和场是一个二维向量场：

对属于某个特定肢体区域上的像素点，都有一个向量从肢体的一个关键点指向另一个关键点。

为较好地得到损失函数f_L,通过定义人体关键部分亲和场（非手工标注），在图像p点处的$L^*_{C,\mathrm{k}}$为：

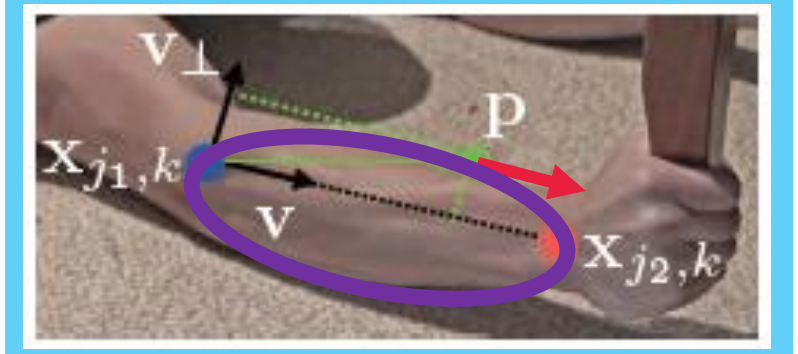
$$\mathbf{L}^*_{c,k}(\mathbf{p}) = \begin{cases} \mathbf{v} & \text{if p on limb } c,k \quad \mathbf{v} = (\mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k})/||\mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k}||_2 \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

理论上肢体上的点应该满足下列不等式：

$$0 \le \mathbf{v} \cdot (\mathbf{p} - \mathbf{x}_{j_1,k}) \le l_{c,k} \text{ and } |\mathbf{v}_\perp \cdot (\mathbf{p} - \mathbf{x}_{j_1,k})| \le \sigma_l$$

肢体长度$l_{c,k}$:$l_{c,k} = ||\mathbf{x}_{j_2,k} - \mathbf{x}_{j_1,k}||_2$ 肢体宽度σ_l

5 Part Affinity Field for Part Association



最后对得到的图片中所有人的亲和场取平均：

$$\mathbf{L}_c^*(\mathbf{p}) = \frac{1}{n_c(\mathbf{p})} \sum_k \mathbf{L}_{c,k}^*(\mathbf{p})$$

$n_c(p)$是所有k个人在p点处非零向量的个数。

在测试中，通过计算对应区域亲和场的积分来评估连接的好坏，两个可能的关键点位置d_{j1}, d_{j2}，L_c沿着两点间线段衡量连接的可靠性：

$$E = \int_{u=0}^{u=1} \mathbf{L}_c(\mathbf{p}(u)) \cdot \frac{\mathbf{d}_{j_2} - \mathbf{d}_{j_1}}{\|\mathbf{d}_{j_2} - \mathbf{d}_{j_1}\|_2} du$$

$$\mathbf{p}(u) = (1-u)\mathbf{d}_{j_1} + u\mathbf{d}_{j_2}$$

在实际计算中，通过对u进行等间隔采样再求和来逼近积分结果。

6 Multi-Person Parsing using PAFs

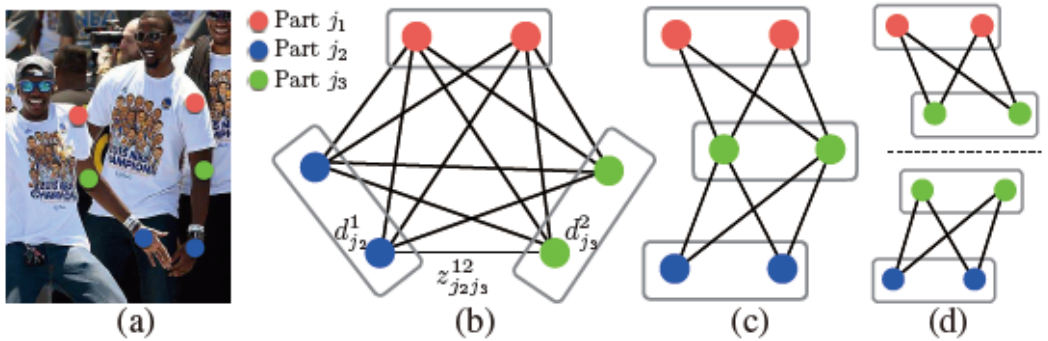


Figure 6. Graph matching. (a) Original image with part detections (b) K-partite graph (c) Tree structure (d) A set of bipartite graphs

得到的预测点有很多很多很多可能的连接，

现在就是在一个K维匹配问题中找到最优解（NP-Hard）。

小科普：NP-Hard

NP是指非确定性多项式（non-deterministic polynomial，缩写NP）。所谓的非确定性是指，可用一定数量的多项式运算去解决的问题。

- 例如，著名的推销员旅行问题（Travel Saleman Problem or TSP）：假设一个推销员需要从香港出发，经过广州，北京，上海，...，等 n 个城市， 最后返回香港。 任意两个城市之间都有飞机直达，但票价不等。假设公司只给报销 C 元钱，问是否存在一个行程安排，使得他能遍历所有城市，而且总的路费小于 C?
- 推销员旅行问题显然是 NP 的。因为如果你任意给出一个行程安排，可以很容易算出旅行总开销。但是，要想知道一条总路费小于 C 的行程是否存在，在最坏情况下，必须检查所有可能的旅行安排！ 这将是个天文数字。

6 Multi-Person Parsing using PAFs

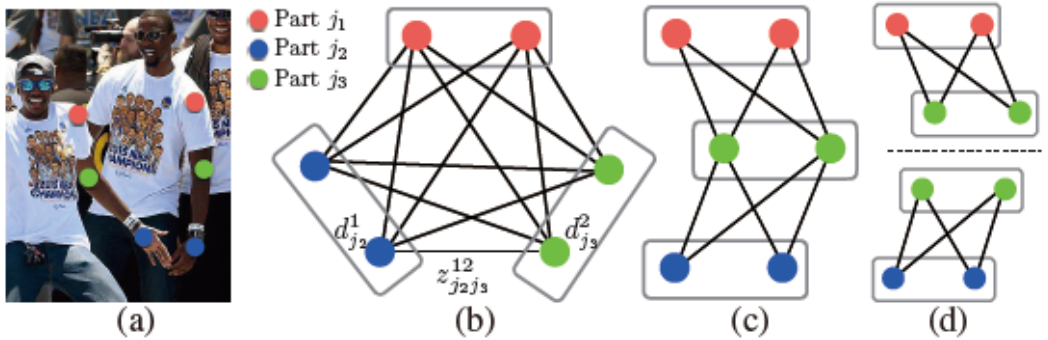


Figure 6. Graph matching. (a) Original image with part detections (b) K-partite graph (c) Tree structure (d) A set of bipartite graphs

得到的预测点有很多很多很多可能的连接，

现在就是在一个K维匹配问题中找到最优解（NP-Hard）。

小科普：二分图与匈牙利算法

基本概念：

无权二分图（unweighted bipartite graph）的最大匹配（maximum matching）和完美匹配（perfect matching），以及用于求解匹配的匈牙利算法（Hungarian Algorithm）（本篇文章是带权二分图）

二分图：

简单来说，如果图中点可以被分为两组，并且使得所有边都跨越组的边界，则这就是一个二分图。准确地说：把一个图的顶点划分为两个不相交集 UU 和VV ， 使得每一条边都分别连接UU、VV中的顶点。

6 Multi-Person Parsing using PAFs

小科普：二分图与匈牙利算法

匹配：在图论中，一个「匹配」（matching）是一个边的集合，其中任意两条边都没有公共顶点。例如，图 3、图 4 中红色的边就是图 2 的匹配。

Fig.1

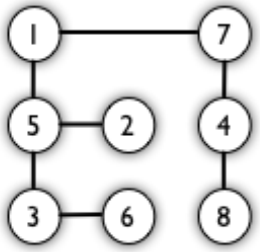


Fig.2

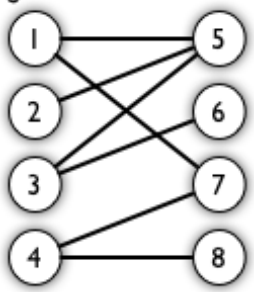


Fig.3

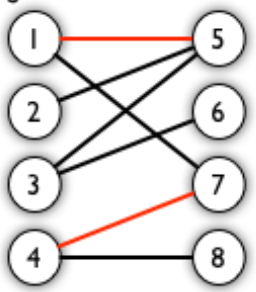


Fig.4

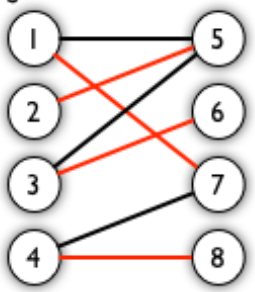
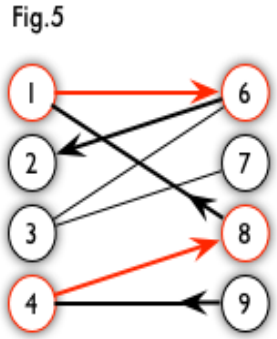


Fig.5

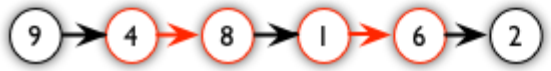


我们定义**匹配点**、**匹配边**、**未匹配点**、**非匹配边**，它们的含义非常显然。例如图 3 中 1、4、5、7 为匹配点，其他顶点为未匹配点；1-5、4-7为匹配边，其他边为非匹配边。

交替路：从一个未匹配点出发，依次经过非匹配边、匹配边、非匹配边...形成的路径叫交替路。

增广路：从一个未匹配点出发，走交替路，如果途径另一个未匹配点（出发的点不算），则这条交替路称为增广路（agumenting path）。例如，图 5 中的一条增广路如图 6 所示（图中的匹配点均用红色标出）：

Fig.6



最大匹配?

完美匹配?

增广路有一个重要特点：非匹配边比匹配边多一条。因此，研究增广路的意义是改进匹配。只要把增广路中的匹配边和非匹配边的身份交换即可。

通过不停地找增广路来增加匹配中的匹配边和匹配点。找不到增广路时，达到最大匹配（这是增广路定理）。匈牙利算法正是这么做的。

6 Multi-Person Parsing using PAFs

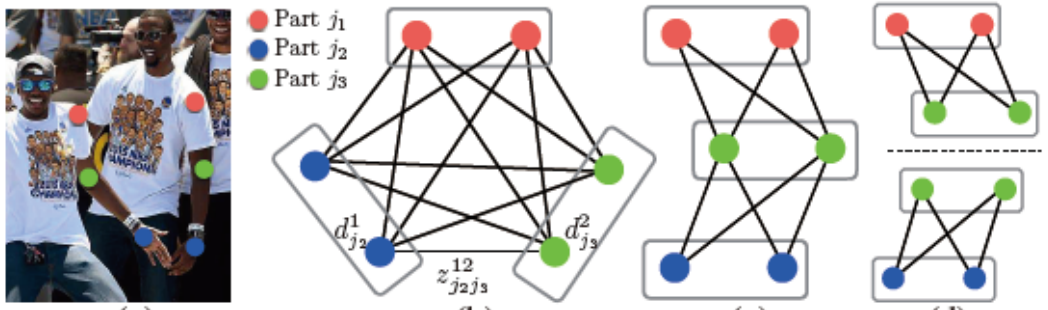


Figure 6. Graph matching. (a) Original image with part detections (b) K-partite graph (c) Tree structure (d) A set of bipartite graphs

针对这个问题，加入两个放松条件简化问题：

① 选择最小数量的边来获得人体姿态的生成树骨架，而不是使用完整的图形（Fig.6c）（只考虑相邻关键点的连接）

② 将匹配问题分解为一组二分图匹配的子问题，并独立确定相邻树节点的匹配（Fig.6d）（每次只考虑一个肢体的连接）

首先，得到多人可能的检测关键点，定义

$$D_J = \left\{ d_j^m : for\ j \in \{1 \dots J\}, m \in \left\{1 \dots N_j\right\} \right\}$$

N_j是一类可能关键点j的数目，$d_j^m \in R^2$是j类关键点的第m个点的位置。

定义一个变量$z_{j_1j_2}^{mn} \in \{0,1\}$，来判断两个关键点是否相连。

$$Z = \{ z_{j_1j_2}^{mn} : for j_1, j_2 \in \{1 \dots J\}, m \in \left\{1 \dots N_{j1}\right\}, n \in \{1 \dots N_{j2}\} \}$$

具体的执行：

从一个树的节点出发，连接另外一个节点，重复这个过程，直到把人体所有的树状结构里的连接都走一遍。

6 Multi-Person Parsing using PAFs

得到的预测点有很多很多很多可能的连接，通过积分给了每个可能的连接一个分数（score），对二分图可能的连接进行加权。

$$E = \int_{u=0}^{u=1} \mathbf{L}_c(\mathbf{p}(u)) \cdot \frac{\mathbf{d}_{j_2} - \mathbf{d}_{j_1}}{\|\mathbf{d}_{j_2} - \mathbf{d}_{j_1}\|_2} du$$

计算可能性最大的匹配结果：

$$\max_{\mathcal{Z}_c} E_c = \max_{\mathcal{Z}_c} \sum_{m \in \mathcal{D}_{j_1}} \sum_{n \in \mathcal{D}_{j_2}} E_{mn} \cdot z_{j_1 j_2}^{mn}$$

E_c是全部c类肢体连接的整体权重

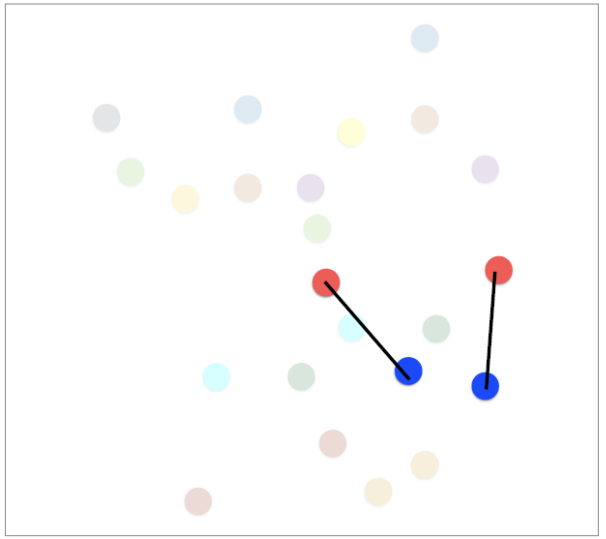
Z_c是c类肢体中Z的子集

约束条件：

$$\text{s.t.} \quad \forall m \in \mathcal{D}_{j_1}, \sum_{n \in \mathcal{D}_{j_2}} z_{j_1 j_2}^{mn} \leq 1$$

$$\forall n \in \mathcal{D}_{j_2}, \sum_{m \in \mathcal{D}_{j_1}} z_{j_1 j_2}^{mn} \leq 1$$

保证两边不共用一个顶点



$$\max_{\mathcal{Z}} E = \sum_{c=1}^{C} \max_{\mathcal{Z}_c} E_c.$$

之后我们就得到所有可能的肢体预测，然后把整个人都连起来啦！！！

THANKS