i-VisionGroup@Tsinghua

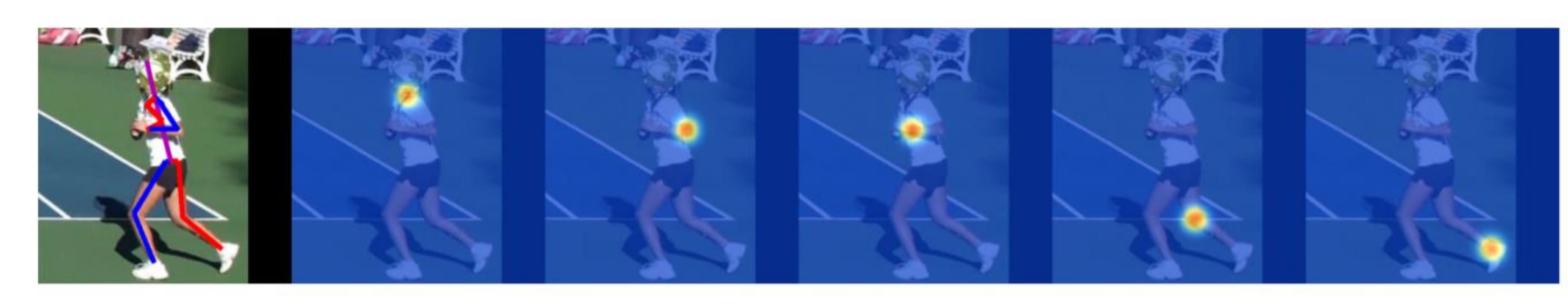
文献阅读

谭子萌

Pose estimation

- 单人姿态估计：检测关节点，常使用高斯热图回归方法
- 肢体变化、遮挡或背景中相似的物体 导致回归出准确的热图困难
- 主要考虑应用不同关键点之间的相互关系（结构先验）方面

1. 特征层面融合（显式利用）
2. GAN结构（隐式利用）



Structured Feature Learning for Pose Estimation

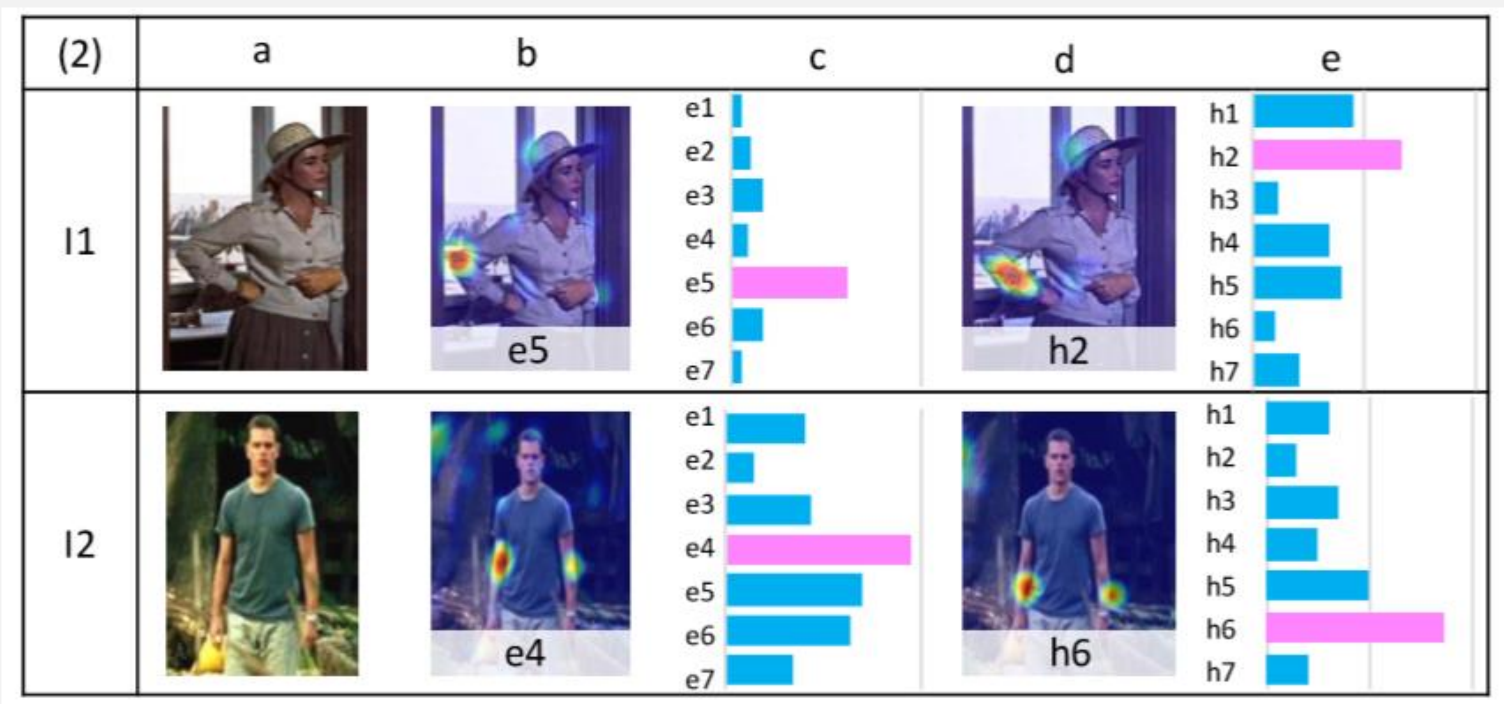
Xiao Chu   Wanli Ouyang   Hongsheng Li   Xiaogang Wang
Department of Electronic Engineering, The Chinese University of Hong Kong
xchu@ee.cuhk.edu.hk   wlouyang@ee.cuhk.edu.hk   hsli@ee.cuhk.edu.hk
xgwang@ee.cuhk.edu.hk

2016CVPR

从特征层面利用关节点间的结构相关性

（特征层面能够保留除了位置信息以外，其他更为丰富的信息）

不同关节点的不同特征通道上可能具有相关性或反相关性：



Method

Contribution:

1. 信息传递的方式：geometrical transform kernels
2. 信息传递的路径：Bi-directional tree model
3. 特征层面，实现端到端的训练，而非后处理

以VGG为骨架，共享fcn6层（4096channel）

其后fcn7层每个关节点分别得到128个特征图

对第k个关节点（x, y）像素位置

$$\mathbf{h}_{fcn7}^{k}(x,y) = f(\mathbf{h}_{fcn6}(x,y) \otimes \mathbf{w}_{fcn7}^{k} + \mathbf{b}_{fcn6}),$$

Geometrical transform kernels

希望用小臂特征图hm减少手肘en的错误响应、加强正确响应

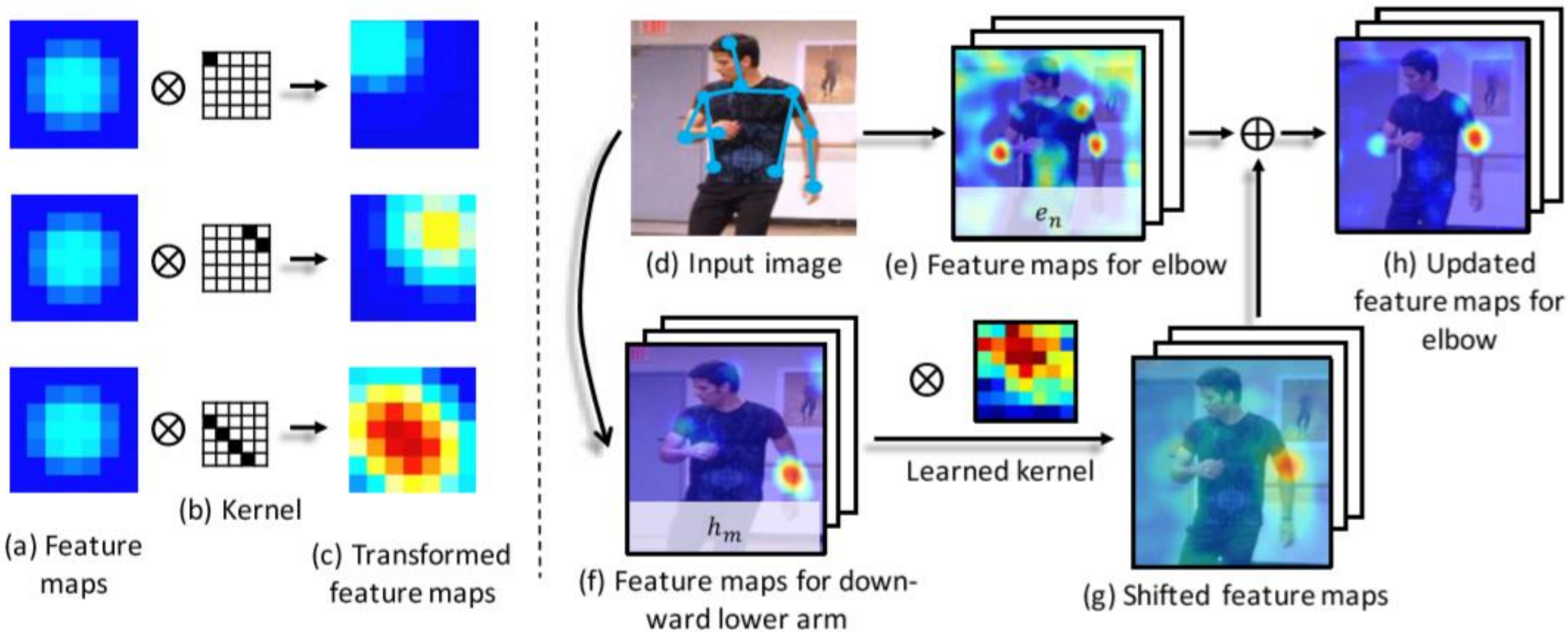
不能单纯地将en与hm相加：存在空间的不匹配性

→不对称的卷积核能够对特征图响应产生几何变换

相邻关节点间的相对空间位置固定，使得几何变换核容易学习到

此外对负相关的情况，使核为负值来抑制错误的响应

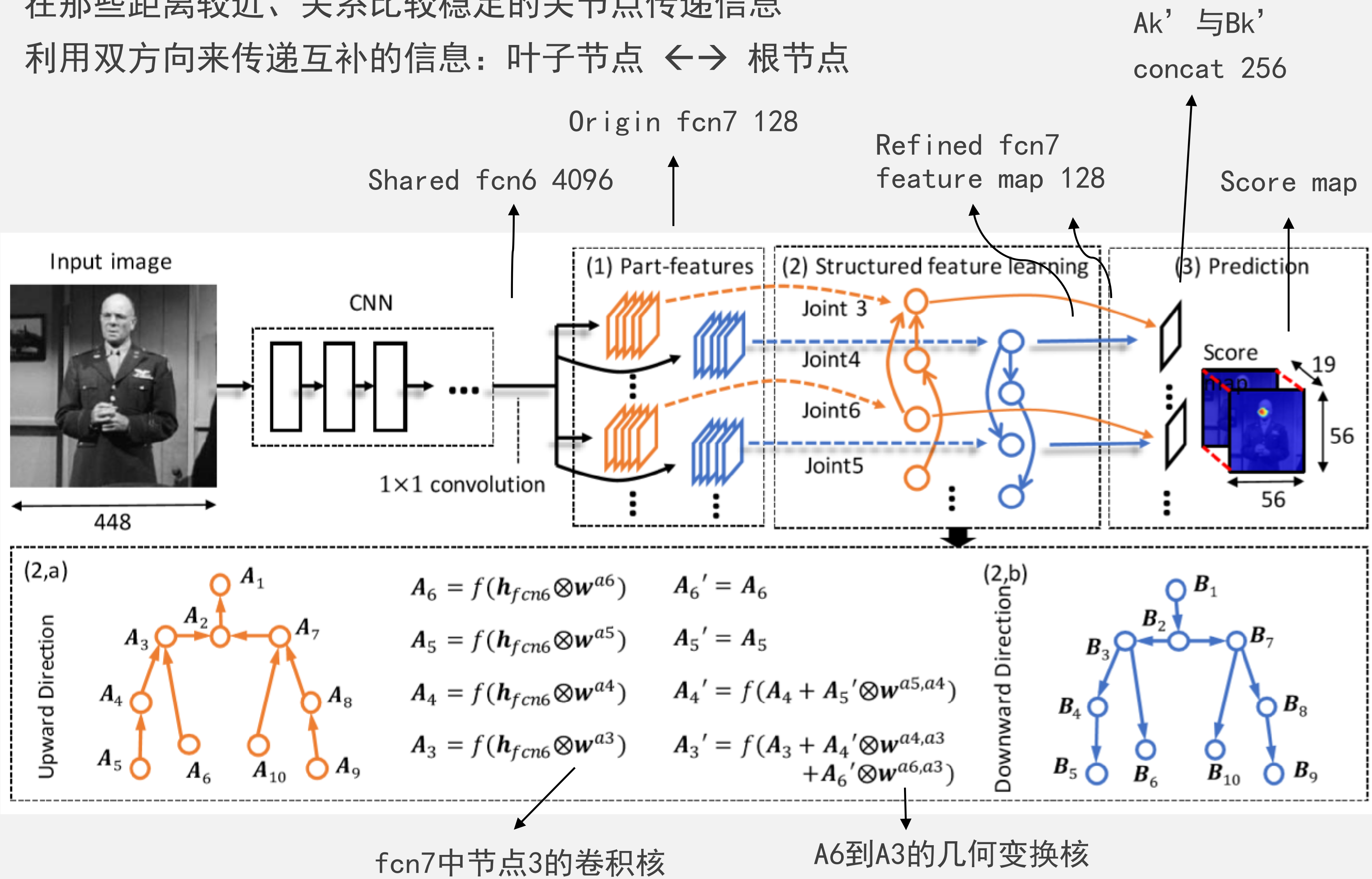
连续3个7×7的变换核 满足变换尺寸需求



Bi-directional tree model

在那些距离较近、关系比较稳定的关节点传递信息

利用双方向来传递互补的信息：叶子节点 <-> 根节点



Method

Score map：1×1卷积

$$\mathbf{z}_k = [\mathbf{A'}_k, \mathbf{B'}_k] \otimes \mathbf{w}^k_{pred}.$$

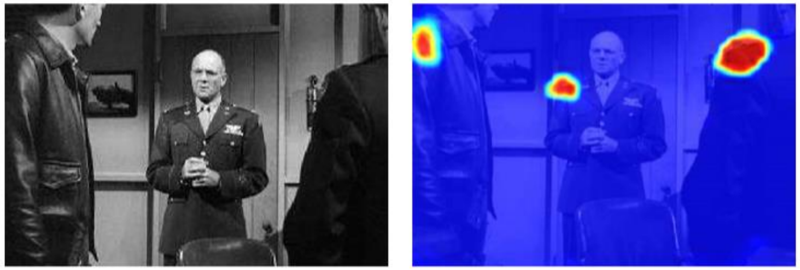
Loss： 分类问题（18个关节点+1个背景）

当(x, y)像素属于第k类时，tk(x, y)=1，否则=0

为解决类别不均衡问题，引入binary mask来随机采样0.05%的负样本

$$\sum_x \sum_y m(x,y) \sum_k t_k(x,y) log(\frac{e^{z_k(x,y)}}{\sum_{k'} e^{z_{k'}(x,y)}})$$

后处理：$[(dx)^2, (dy)^2]$
$$\begin{aligned} dx &= (x_i - x_j - x_r) \\ dy &= (y_i - y_j - y_r) \end{aligned}$$



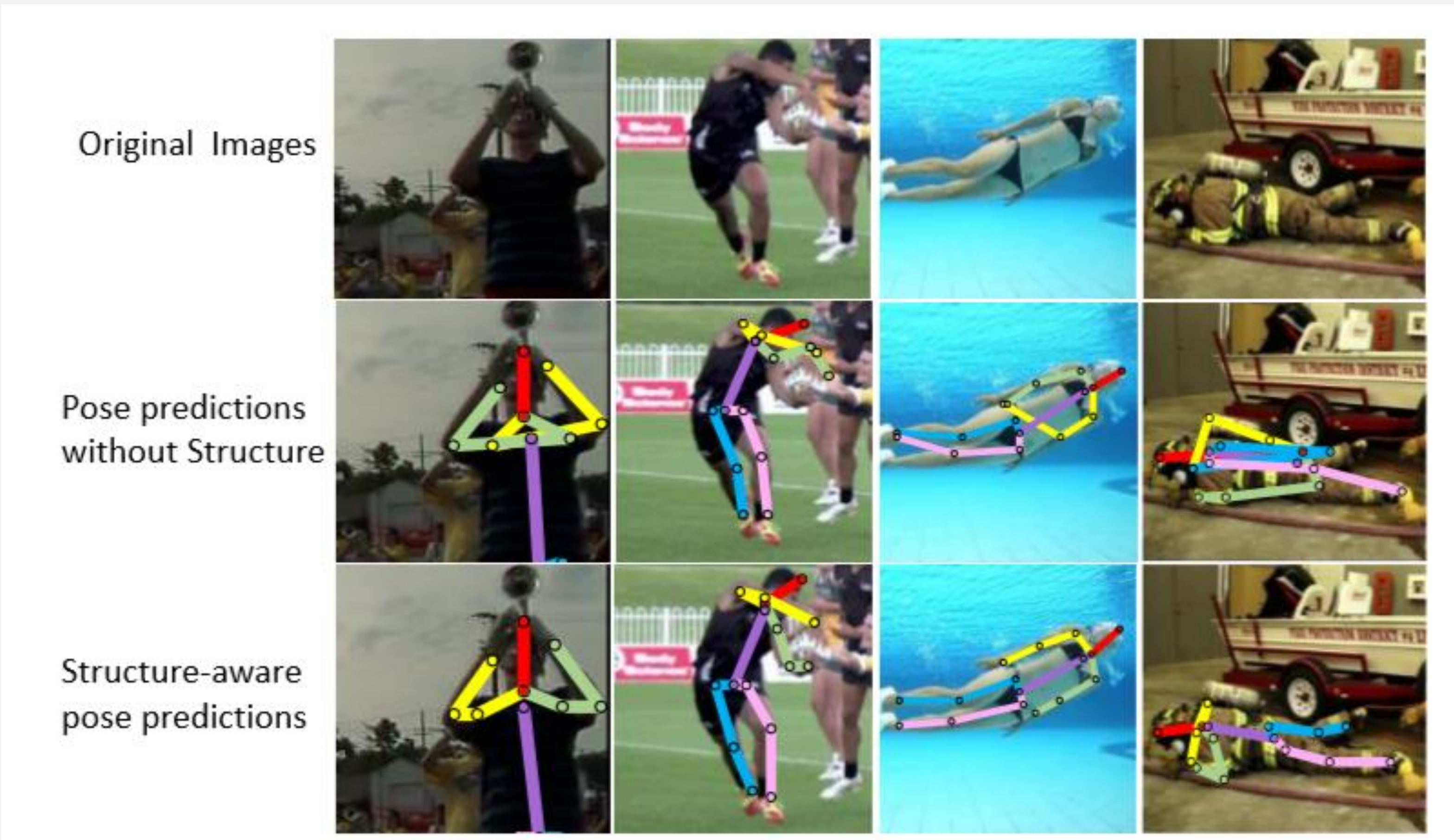
Test image Score map

Results

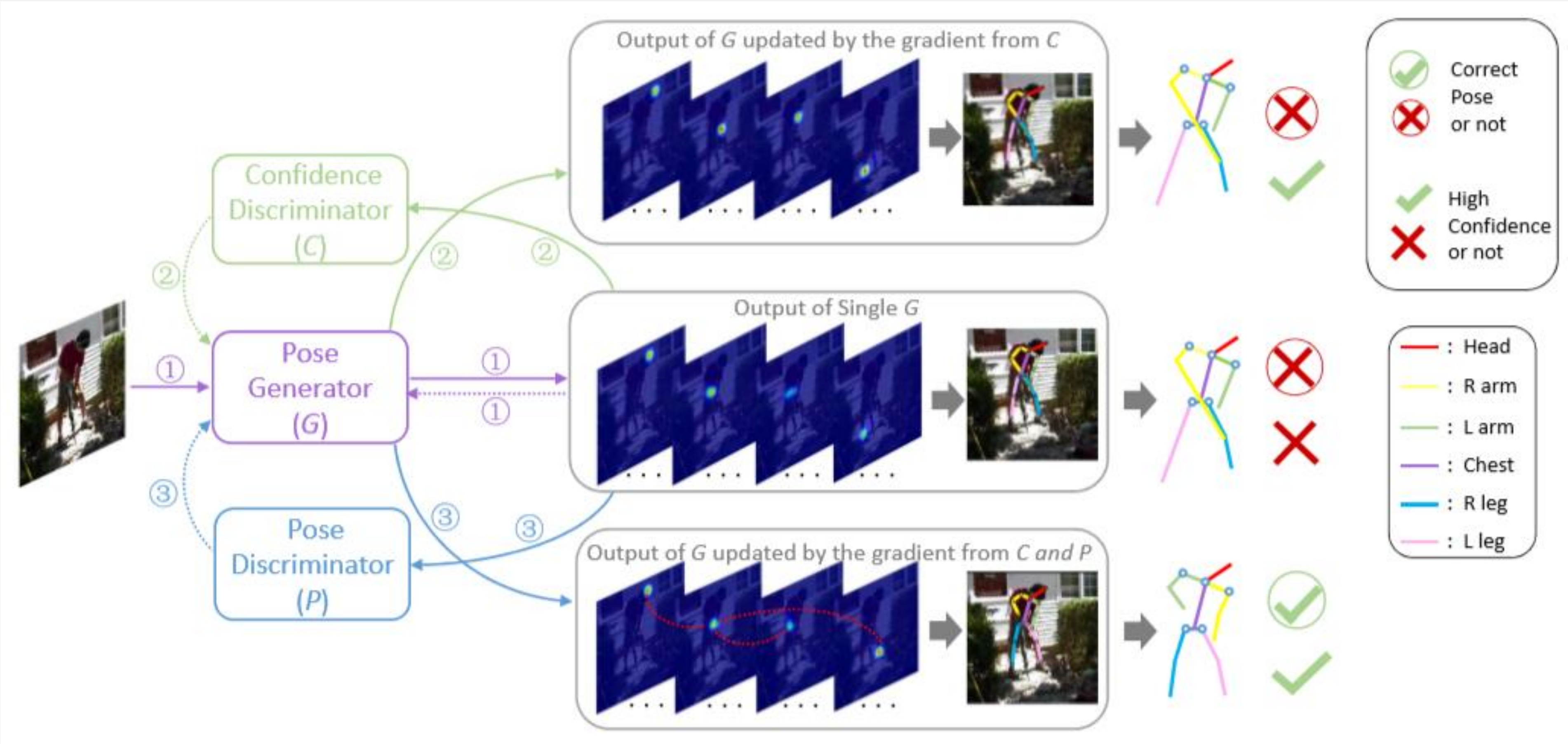
Adversarial PoseNet:
A Structure-aware Convolutional Network for Human Pose Estimation*

Yu Chen[1] Chunhua Shen[2] Xiu-Shen Wei[3] Lingqiao Liu[2] Jian Yang[1]
[1]Nanjing University of Science and Technology [2]University of Adelaide [3]Nanjing University 2017ICCV

单个关节点的热图估计可能导致出现生物学上不可能的姿态



网络遵循GAN结构，由一个生成器+2个判别器构成



Pose generator G

Multi-task: 输出32channel 16个pose heatmap + 对应的occlusion heatmap

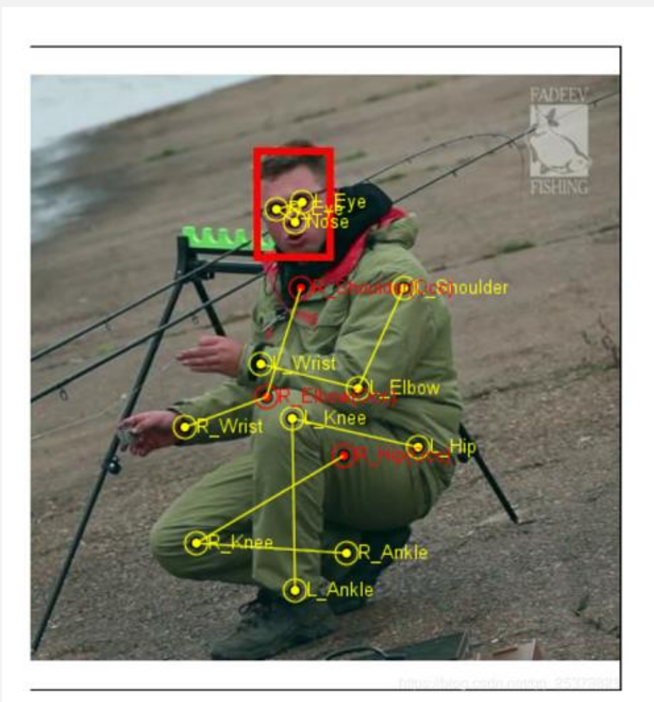
采用stack的方式 均为带有跳接结构的encoder-decoder网络

$$\begin{cases} \{\boldsymbol{Y}_n, \boldsymbol{Z}_n, \boldsymbol{X}\} = \mathcal{G}_n(\boldsymbol{Y}_{n-1}, \boldsymbol{Z}_{n-1}, \boldsymbol{X}) & \text{if } n \geqslant 2 \\ \{\boldsymbol{Y}_n, \boldsymbol{Z}_n, \boldsymbol{X}\} = \mathcal{G}_n(\boldsymbol{X}) & \text{if } n = 1 \end{cases}$$

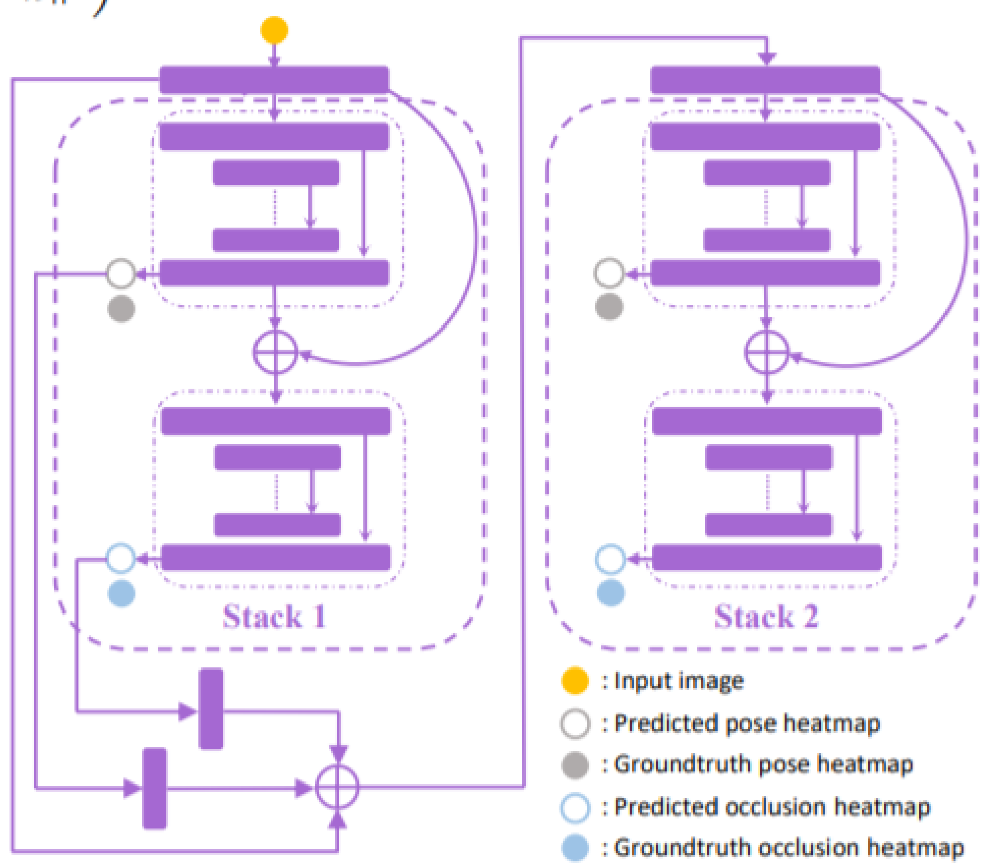
$$\mathcal{L}_G(\Theta) = \frac{1}{2MN} \sum_{n=1}^{N} \sum_{i=1}^{M} \left(\left\| \boldsymbol{y}^i - \hat{\boldsymbol{y}}_n^i \right\|^2 + \left\| \boldsymbol{z}^i - \hat{\boldsymbol{z}}_n^i \right\|^2 \right)$$

N:stack

M:data number



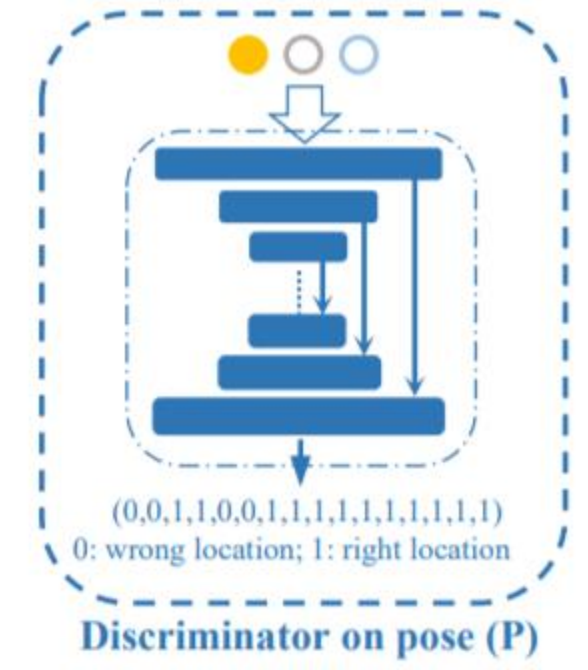
MPII数据集



Pose Discriminator P

区分那些不满足人体关节点约束的pose

为保证对输入图片姿态合理，需要同时送入输入图片



$$\begin{aligned}\mathcal{L}_P(G,P) = \mathbb{E}[\log P(\boldsymbol{y},\boldsymbol{z},\boldsymbol{x})] + \\ \mathbb{E}[\log(1-|P(G(\boldsymbol{x}),\boldsymbol{x})-\boldsymbol{p}_{\text{fake}}|)]\,.\end{aligned}$$

传统的gt=0/1会导致训练困难 → 对16个关节点分别分析

在传统GAN中pfake=0 → 可以认为那些和真实姿态接近的预测结果为真样本

(当一个关节点偏离真值很远时，会导致整个姿态出错)

$$\boldsymbol{p}^i_{\text{fake}} = \begin{cases} 1 & \text{if } d_i < \delta \\ 0 & \text{if } d_i \geqslant \delta \end{cases}$$

δ 为normalized distance

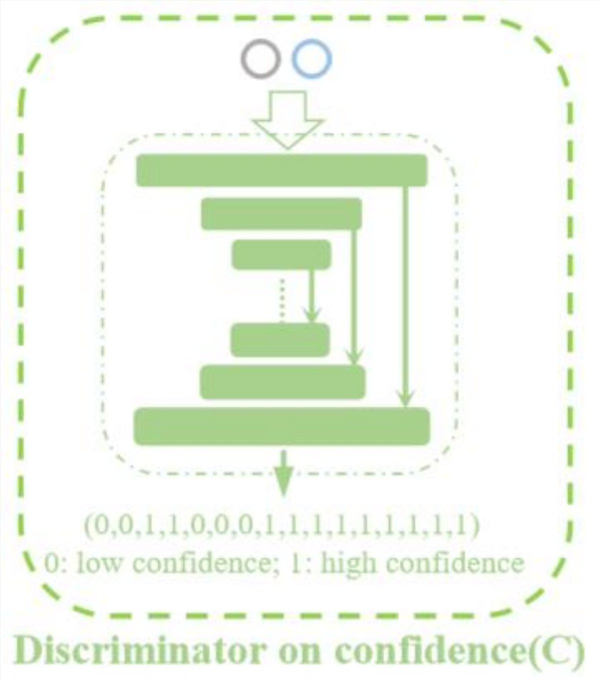
Confidence Discriminator C

区分输出热图的low-confidence与high-confidence(Gaussian centered)

即网络在所预测位置是否confident

输入为pose + occlusion heatmap

$$\mathcal{L}_C(G,C) = \mathbb{E}[\log C(\boldsymbol{y},\boldsymbol{z})] + \mathbb{E}[\log(1 - |C(G(\boldsymbol{x})) - \boldsymbol{c}_{\text{fake}}|)].$$



传统的gt=0/1会导致训练困难 → 对16个关节点分别分析

在传统GAN中cfake=0

→可以认为当prediction与gt heatmap较像时为真样本

$$\boldsymbol{c}_{\text{fake}}^{i} = \begin{cases} 1 & \text{if } \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\| < \varepsilon \\ 0 & \text{if } \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\| \geqslant \varepsilon \end{cases}$$

生成器G的训练

$$\mathcal{L}_G(\Theta) = \frac{1}{2MN}\sum_{n=1}^{N}\sum_{i=1}^{M}\left(\left\|\boldsymbol{y}^i - \hat{\boldsymbol{y}}_n^i\right\|^2 + \left\|\boldsymbol{z}^i - \hat{\boldsymbol{z}}_n^i\right\|^2\right)$$

$$\mathcal{L}_P(G,P) = \mathbb{E}[\log P(\boldsymbol{y},\boldsymbol{z},\boldsymbol{x})] + \mathbb{E}[\log(1 - |P(G(\boldsymbol{x}),\boldsymbol{x}) - \boldsymbol{p}_{\text{fake}}|)]\,.$$

$$\boldsymbol{p}_{\text{fake}}^i = \begin{cases} 1 & \text{if } d_i < \delta \\ 0 & \text{if } d_i \geqslant \delta \end{cases}$$

$$\mathcal{L}_C(G,C) = \mathbb{E}[\log C(\boldsymbol{y},\boldsymbol{z})] + \mathbb{E}[\log(1 - |C(G(\boldsymbol{x})) - \boldsymbol{c}_{\text{fake}}|)]\,.$$

$$\boldsymbol{c}_{\text{fake}}^i = \begin{cases} 1 & \text{if } \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\| < \varepsilon \\ 0 & \text{if } \|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\| \geqslant \varepsilon \end{cases}$$

总Loss：

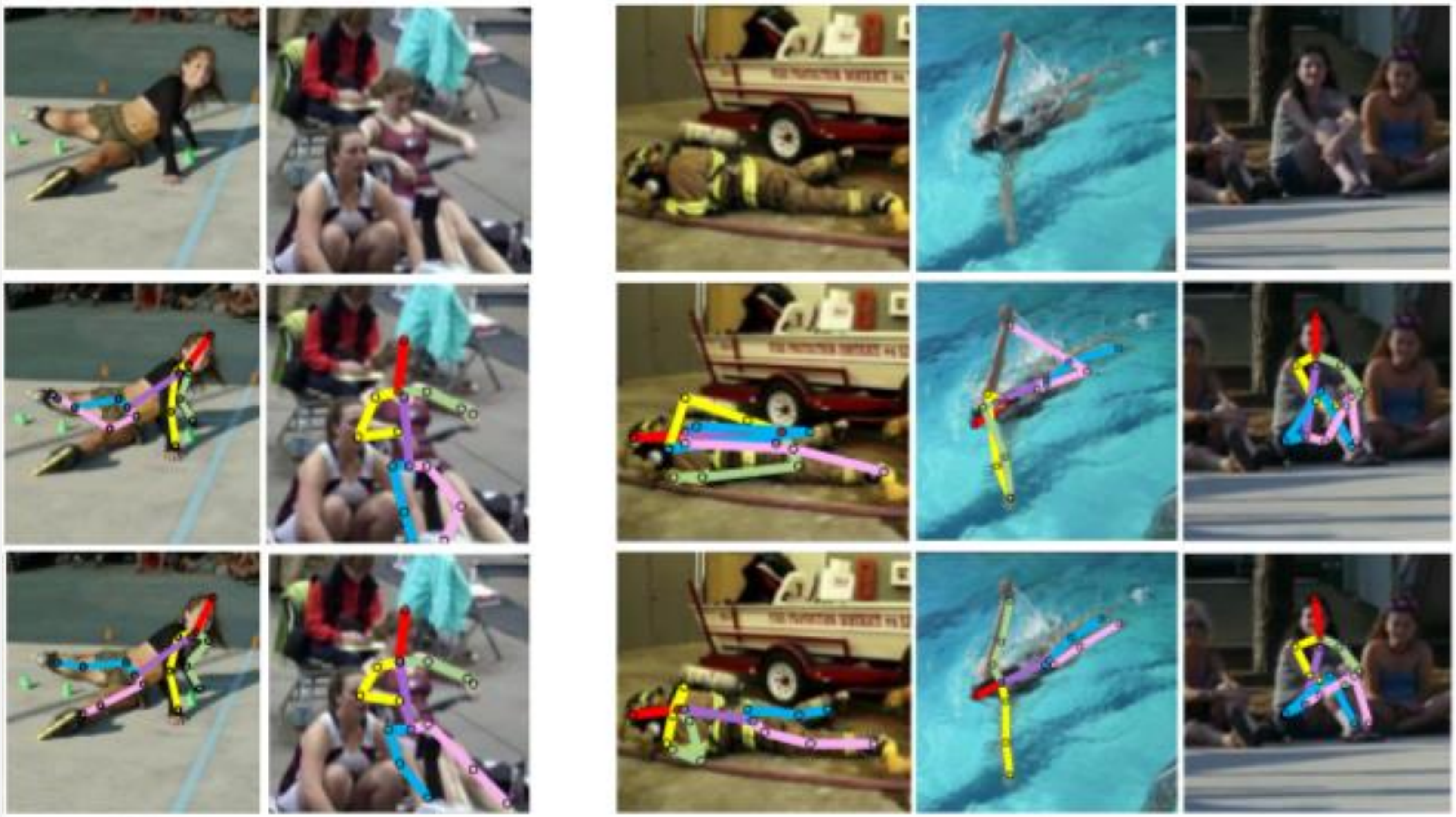
$$\arg\min_{G}\max_{P,C} \mathcal{L}_G(\Theta) + \alpha\mathcal{L}_C(G,C) + \beta\mathcal{L}_P(G,P)\,.$$

当cfake=creal时 α=0

当pfake=freal时 β=0

排除可能训练得不好的判别器的影响

Results



(b) (c)

谢谢大家!