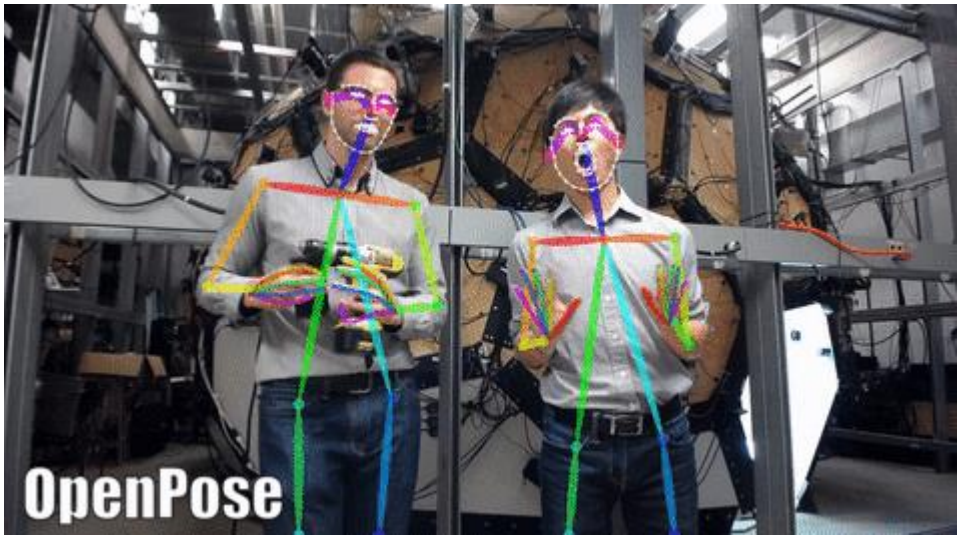
19.6.25 周报 张天意

Neural Body Fitting: Unifying Deep Learning and Model Based Human Pose and Shape Estimation（3DV 2018）

目前已经有一些成功的工作：生成人体关键点，棒状表示模型（个人感觉就是指 openpose）



本文作者提出更具挑战性的任务：estimating the parameters of a detailed statistical human body model from a single image

传统方法需要一个差不多初始化模型，然后把初值优化到最终结果（不需要 3d 训练数据——带 3d 动作标注的图片）

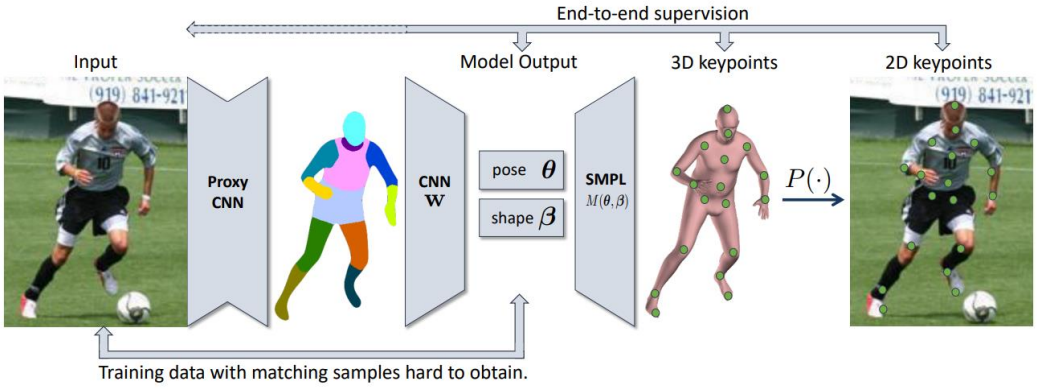
CNN 就是 forward prediction models，就不需要 initialization，但是需要 3d 姿态标注，不像 2d 标注好获得

他们近期的工作通过把重建出的模型投影回 2d 空间更新损失函数，就可以使用 2d 标注了

本文的**目的**：To analyze the importance of such components

components: image--(CNN,3d notation trained)-->smpl model(hybird params)-->image--(reproject)-->2d notation for CNN training

要形成闭环（loop）



NBF = 一个包含统计身体模型的 CNN
两种监督模式：full 3d supervision 和 weak 2d sup，bottom-up top-down 的方法，使得 NBF 既不需要初始化模型也不需要 3d 标注的训练数据
因为光照、衣服、杂乱的背景都不想要，专注于 pose 和 shape，所以用处理后的 image 代替原始 rgb image
结论：
1. 12-body-part 的分割就包含了足够的 shape 和 pose 信息
2. 这种处理后图像的方法比起用原图，效果有竞争力，更简单，训练数据利用率更高
3. 分割质量可以很大程度上预测三维重建（fit）的质量

Networks:
Segmentation Network:
RefineNet（基于 ResNet-101）
Fitting network:
repurpose a ResNet-50 network pretrained on ImageNet
replace the final pooling layer with a single fully-connected layer that outputs the 10 shape and 216 pose parameters

Loss func:
3D latent parameter loss
R ： the vectorized rotation matrices of the 24 parts of the body
I ： colour-coded part segmentation map
W ： CNN weights

$$\mathcal{L}_{lat}(w) = \sum_i^N |\mathbf{r}(\boldsymbol{\theta}(w, \mathbf{I}_i)) - \mathbf{r}(\boldsymbol{\theta}_i)| + |\boldsymbol{\beta}(w, \mathbf{I}_i) - \boldsymbol{\beta}_i|,$$

关节点坐标:

$$\mathcal{L}_{3D}(w) = \sum_i^N \|N_J(w, \mathbf{I}_i) - \mathbf{J}_i\|^2$$

投影后的 2d 关节点坐标

$$\mathcal{L}_{2D}(w) = \sum_i^N \|N_{2D}(w, \mathbf{I}_i) - \mathbf{J}_{2D,i}\|^2$$

D ： dataset with 2d or 3d notation

$$\mathcal{L}_{2D+3D}(w, \mathcal{D}) = \mathcal{L}_{2D}(w, \mathcal{D}_{2D}) + \mathcal{L}_{3D}(w, \mathcal{D}_{3D})$$

Evaluation & analysis

三个数据集 UP-3D，HumanEva-I，Human3.6M

type of input	*UP*	*H36M*
RGB	98.5	48.9
Segmentation (1 part)	95.5	43.0
Segmentation (3 parts)	36.5	37.5
Segmentation (6 parts)	29.4	36.2
Segmentation (12 parts)	27.8	33.5
Segmentation (24 parts)	28.8	31.8
Joints (14)	28.8	33.4
Joints (24)	27.7	33.4

Table 1: *Input Type vs. 3D error in millimeters*

We observe that explicit **part representations** (part segmentations or joint heatmaps) are more useful for 3D shape/pose estimation compared to **RGB images and plain silhouettes**.

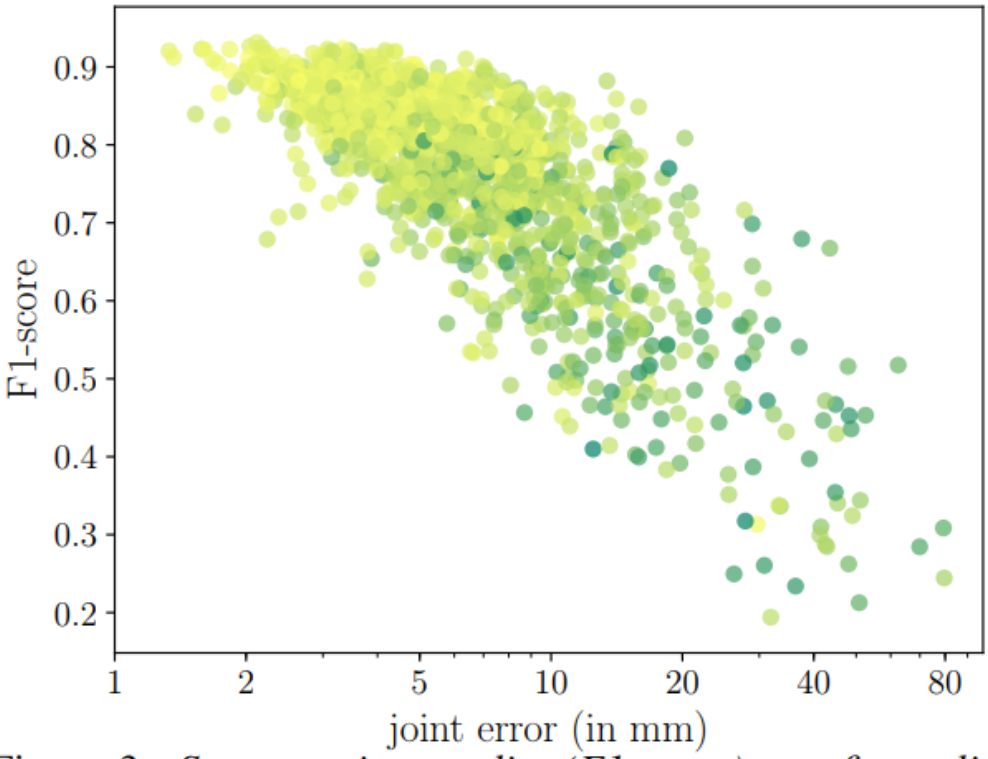


Figure 3: *Segmentation quality (F1-score) vs. fit quality (3D joint error).* The darkness indicates the difficulty of the pose, i.e. the distance from the upright pose with arms by the sides.

Val \ Train	VGG	ResNet	RefineNet	GT
VGG	107.2	119.9	135.5	140.7
ResNet	97.1	96.3	112.2	115.6
RefineNet	89.6	89.9	82.0	83.3
GT	62.3	60.5	35.7	27.8

Table 2: *Effect of segmentation quality on the quality of the 3D fit prediction modules ($err_{joints3D}$)*

the higher the F-1 score, the lower the 3D joint error.
F-1 score 代表了图像分割的质量

57.0，53.2，67.1，100

To determine the effect of segmentation quality on the results, we train three different part segmentation networks. Besides RefineNet, we also train two variants of DeepLab [9], based on VGG-16 [52] and ResNet-101 [14]. These networks result in **IoU scores of 67.1, 57.0, and**

53.2 respectively on the UP validation set.

we **then train four 3D prediction networks** - one for each of the part segmentation networks, andan additional one using the ground truth segmentations.

Loss	$\mathbf{err_{joints3D}}$	**PCKh**	$\mathbf{err_{quat}}$
L_{lat}	83.7	93.1	0.278
$L_{lat} + L_{3D}$	82.3	93.4	0.280
$L_{lat} + L_{2D}$	83.1	93.5	0.278
$L_{lat} + L_{3D} + L_{2D}$	82.0	93.5	0.279
L_{3D}	83.7	93.5	1.962
L_{2D}	198.0	94.0	1.971

Table 3: *Loss ablation study.* Results in 2D and 3D error metrics (*joints3D*: Euclidean 3D distance, *mesh*: average vertex to vertex distance, *quat*: average body part rotation error in radians).

(i) errjoints3D, the Euclidean distance between ground truth and predicted SMPL joints (in mm).

(ii) P CKh [4], the percentage of correct keypoints with the error threshold being 50% of head size, which we measure on a perexample basis.

(iii) errquat, quaternion distance error of the predicted joint rotations (in radians).

Error \ Ann.perc.	100	50	20	10	5	2	1	0
$\mathbf{err_{joints3D}}$	83.1	82.8	82.8	83.6	84.5	88.1	93.9	198
$\mathbf{err_{quat}}$	0.28	0.28	0.27	0.28	0.29	0.30	0.33	1.97

Table 4: *Effect of 3D labeled data.* We show the 3D as well as the estimated body part rotation error for varying ratios of data with 3D labels. For all of the data, we assume that 2D pose labels are available. Both errors saturate at 20% of 3D labeled training examples.

需要用多少 3d 标注

Method	Mean	Median
Ramakrishna et al. [45]	168.4	145.9
Zhou et al. [68]	110.0	98.9
SMPLify [6]	79.9	61.9
Random Forests [24]	93.5	77.6
SMPLify (Dense) [24]	74.5	59.6
Ours	64.0	49.4

Table 5: **HumanEva-I results.** 3D joint errors in mm.

Method	Mean	Median
Akhter & Black [1]	181.1	158.1
Ramakrishna et al. [45]	157.3	136.8
Zhou et al. [68]	106.7	90.0
SMPLify [6]	82.3	69.3
SMPLify (dense) [24]	80.7	70.0
SelfSup [62]	98.4	-
Pavlakos et al. [38]	75.9	-
HMR (H36M-trained)[22]	81.2	76.7
HMR [22]	**56.8**	-
Ours	59.9	52.3

Table 6: **Human 3.6M.** 3D joint errors in mm.

与其他方法的比较