清華大學
Tsinghua University

i-VisionGroup

文献分享

范博昊

单目Mesh

- 基础
  - SMPL(ACMTOG15)
- 多人
  - MSMP(CVPR18)
- 单人
  - Graph CNN(arxiv 2020.2)
  - EllipBody(arixv 2020.3)
  - HEMlets(arixv 2020.3)

SMPL

SMPL: A Skinned Multi-Person Linear Model

Matthew Loper [*12] Naureen Mahmood[†1] Javier Romero[†1] Gerard Pons-Moll[†1] Michael J. Black[†1]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Industrial Light and Magic, San Francisco, CA

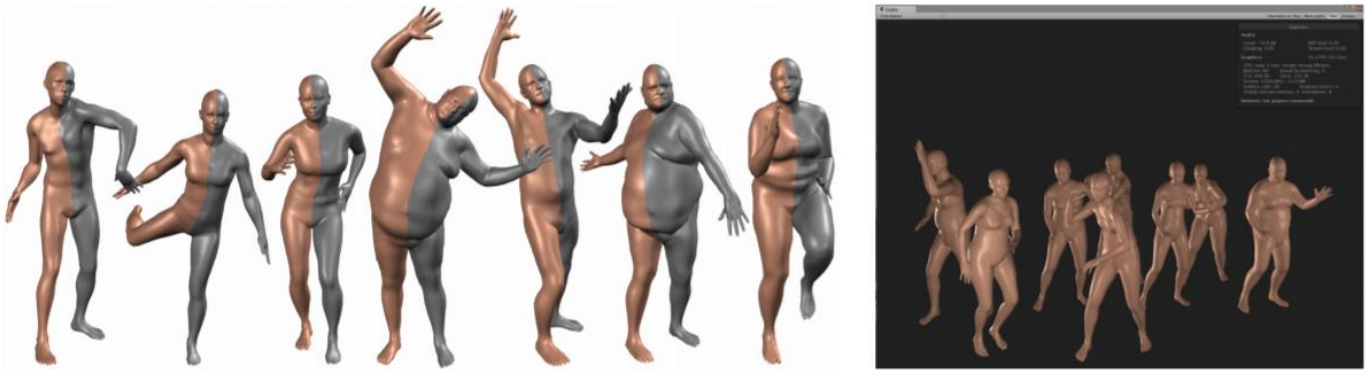


Figure 1: *SMPL is a realistic learned model of human body shape and pose that is compatible with existing rendering engines, allows animator control, and is available for research purposes. (left) SMPL model (orange) fit to ground truth 3D meshes (gray). (right) Unity 5.0 game engine screenshot showing bodies from the CAESAR dataset animated in real time.*

SMPL

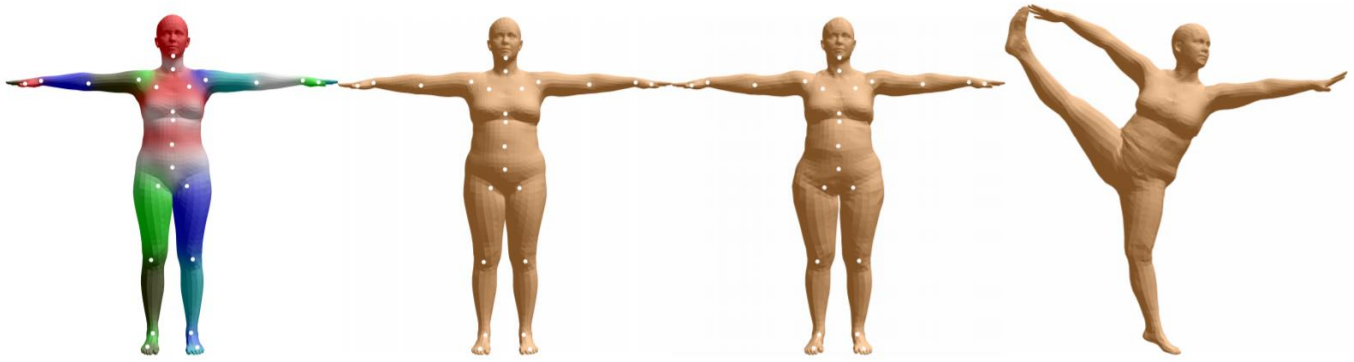
- N=6890 面片形式
- K=23 关节点
- 总共参数3x23+3+10+3=72个参数

总共描述为23个关节点的位置，整体的朝向3维

人的具体形状信息(高矮胖瘦)10个参数

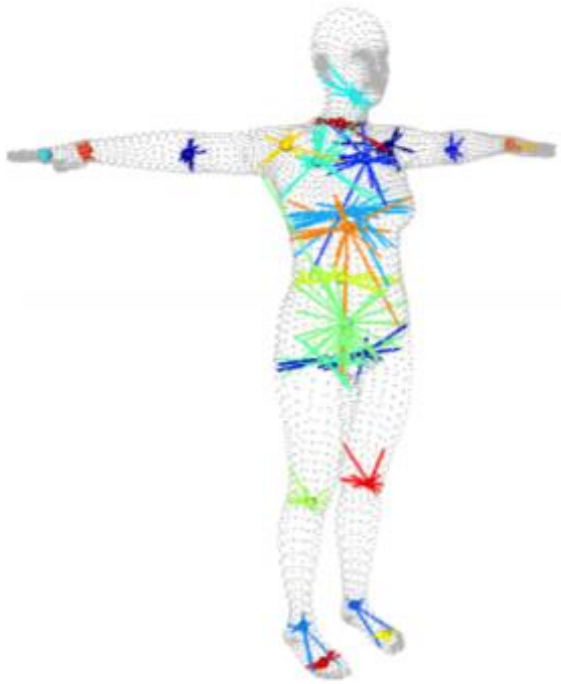
相机参数3维

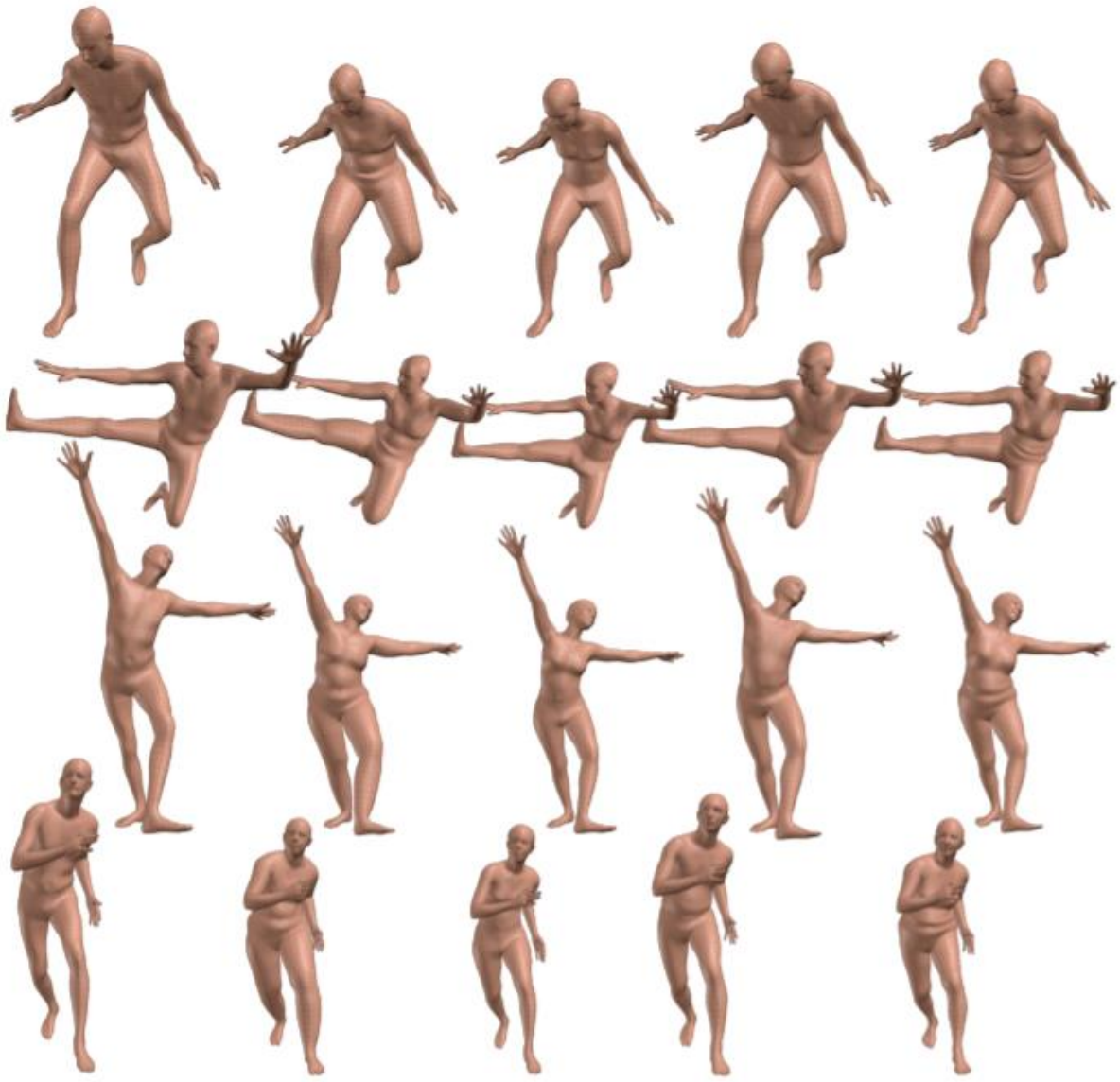
共85维数据



train

- 3d精细模型到smpl的85维特征的参数是训练出来的
- 所以smpl模型不是完全表示精细3d结果(e.g.深度图)





arxiv 2020

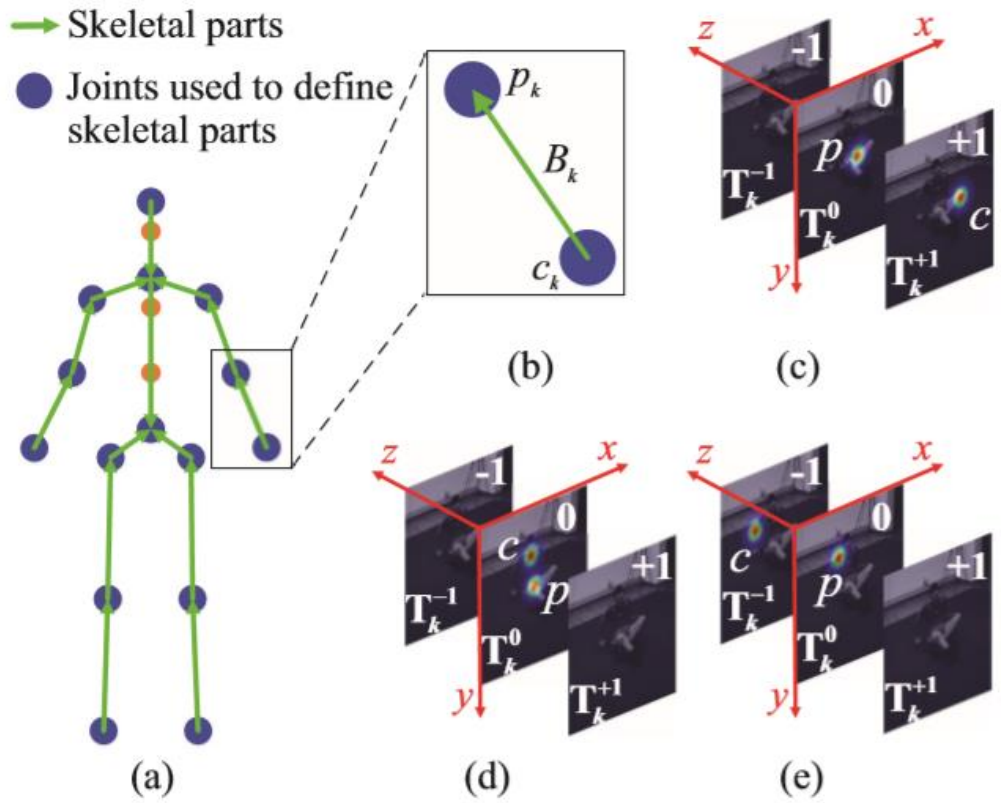
HEMlets PoSh: Learning Part-Centric Heatmap Triplets for 3D Human Pose and Shape Estimation

Kun Zhou, Xiaoguang Han, *Member, IEEE,* Nianjuan Jiang, *Member, IEEE,* Kui Jia, *Member, IEEE,* and Jiangbo Lu, *Senior Member, IEEE*

Abstract—Estimating 3D human pose from a single image is a challenging task. This work attempts to address the uncertainty of lifting the detected 2D joints to the 3D space by introducing an intermediate state - Part-Centric Heatmap Triplets (*HEMlets*), which shortens the gap between the 2D observation and the 3D interpretation. The HEMlets utilize three joint-heatmaps to represent the relative depth information of the end-joints for each skeletal body part. In our approach, a Convolutional Network (ConvNet) is first trained to predict HEMlets from the input image, followed by a volumetric joint-heatmap regression. We leverage on the integral operation to extract the joint locations from the volumetric heatmaps, guaranteeing end-to-end learning. Despite the simplicity of the network design, the quantitative comparisons show a significant performance improvement over the best-of-grade methods (e.g. 20% on Human3.6M). The proposed method naturally supports training with "in-the-wild" images, where only weakly-annotated relative depth information of skeletal joints is available. This further improves the generalization ability of our model, as validated by qualitative comparisons on outdoor images. Leveraging the strength of the HEMlets pose estimation, we further design and append a shallow yet effective network module to regress the SMPL parameters of the body pose and shape. We term the entire HEMlets-based human pose and shape recovery pipeline *HEMlets PoSh*. Extensive quantitative and qualitative experiments on the existing human body recovery benchmarks justify the state-of-the-art results obtained with our HEMlets PoSh approach.

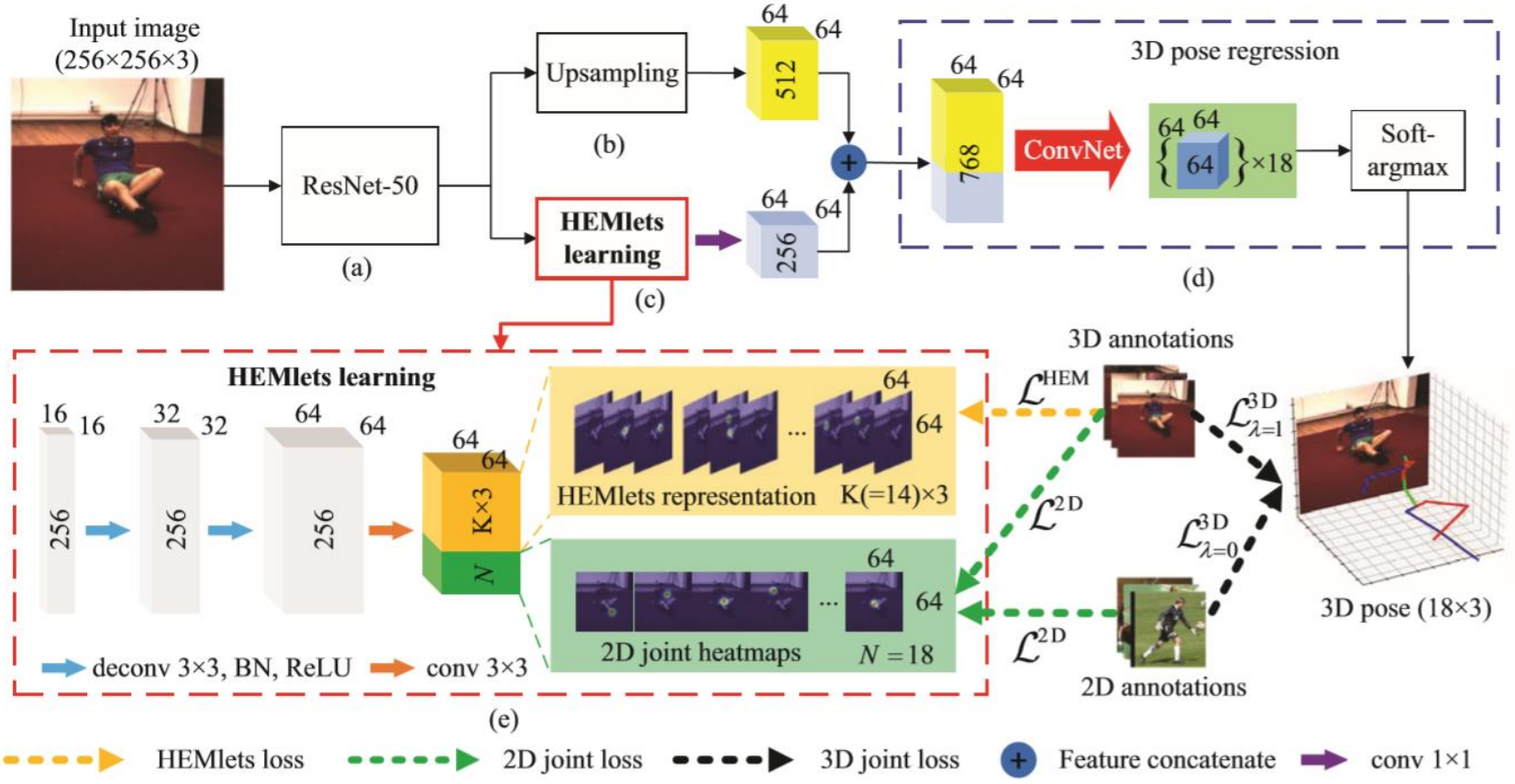
pose

- 2D pose+seg -> 3D pose -> 3D mesh



$$r(z_p, z_c) = \begin{cases} 1 & z_p - z_c > \epsilon \\ 0 & |z_p - z_c| < \epsilon \\ -1 & z_p - z_c < -\epsilon \end{cases},$$

结构



Loss

- loss

$$\mathcal{L}^{\mathrm{HEM}} = \|(\mathbf{T}^{\mathrm{gt}} - \hat{\mathbf{T}}) \odot \mathbf{\Lambda}\|_2^2,$$

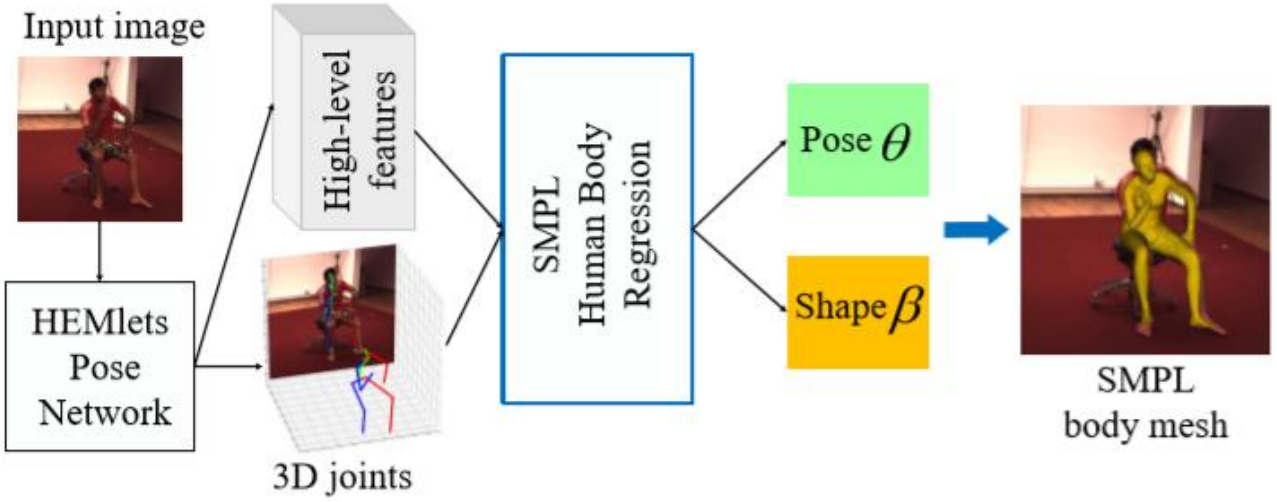
$$\mathcal{L}^{\mathrm{2D}} = \sum_{n=1}^{N} \|\mathbf{H}_n^{\mathrm{gt}} - \hat{\mathbf{H}}_n\|_2^2,$$

$$\mathcal{L}_{\lambda}^{\mathrm{3D}} = \sum_{n=1}^{N} (\left|x_n^{\mathrm{gt}} - \hat{x}_n\right| + \left|y_n^{\mathrm{gt}} - \hat{y}_n\right| + \lambda \left|z_n^{\mathrm{gt}} - \hat{z}_n\right|),$$

$$\mathcal{L}^{\mathrm{int}} = \mathcal{L}^{\mathrm{HEM}} + \mathcal{L}^{\mathrm{2D}}.$$

$$\mathcal{L}^{\mathrm{tot}} = \alpha * \mathcal{L}^{\mathrm{int}} + \mathcal{L}_{\lambda}^{\mathrm{3D}}, \qquad \alpha = 0.05$$

Mesh



$$\mathcal{L}_\theta = \sum_{i=1}^{24} \|R_i^{\text{gt}} - \hat{R}_i\|$$

$$\mathcal{L}_\beta = \sum_{i=1}^{10} \|\beta_i^{\text{gt}} - \hat{\beta}_i\|$$

$$\mathcal{L}_{\text{mesh}} = \mathcal{L}_\theta + \mathcal{L}_\beta + \mathcal{L}^{\text{tot}}$$

results

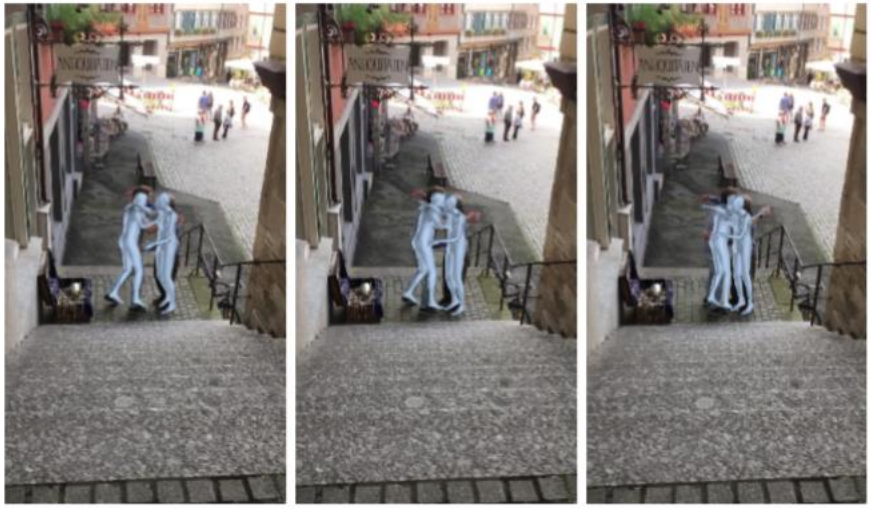
对比的方法分别是
CVPR18
ICCV19

Fig. 12. Qualitative comparisons of our method with some existing ones on human body model recovery. For each example, the input image is first shown, which is followed by the results of HMR [16], SPIN [18] and ours. For each resulting body mesh, two views are provided for visualization.

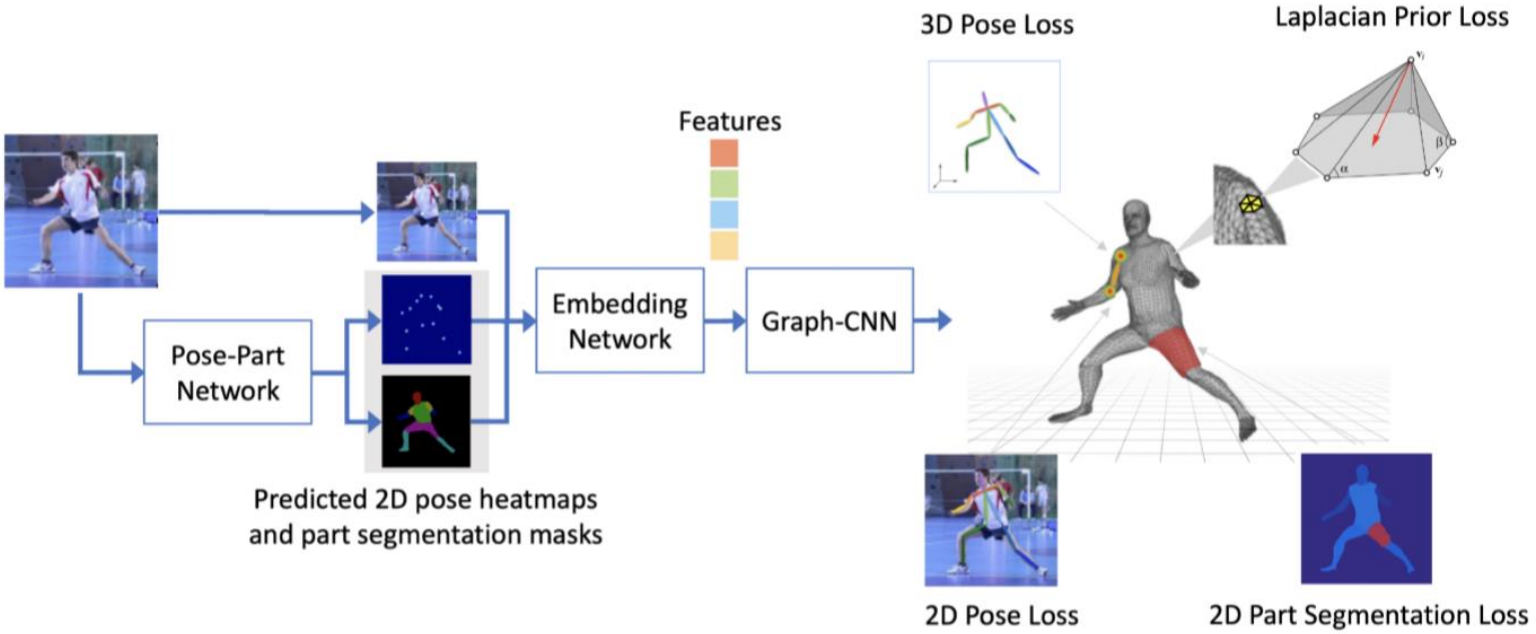
arixv 2020

Learning Nonparametric Human Mesh Reconstruction from a Single Image without Ground Truth Meshes

Kevin Lin[†] Lijuan Wang[‡] Ying Jin[†‡] Zicheng Liu[‡] Ming-Ting Sun[†]
[†]University of Washington [‡]Microsoft
{kvlin,mts}@uw.edu, {lijuanw,Ying.Jin,zliu}@microsoft.com



结构


3D Pose Loss
Laplacian Prior Loss
Features
Pose-Part Network
Embedding Network
Graph-CNN
Predicted 2D pose heatmaps and part segmentation masks
2D Pose Loss
2D Part Segmentation Loss

Graph CNN

$$Y = F(X; \bar{A}, W).$$

$$F(X; \bar{A}, W) = f_T(\cdots f_2(f_1(X; \bar{A}, W_1); \bar{A}, W_2) \cdots ; \bar{A}, W_T)$$

$$X_{t+1} = f_t(X_t; \bar{A}, W_t) = \sigma(\bar{A} X_t W_t)$$

$$\min_W \mathcal{L}(W) = \mathcal{L}_{Lap}(W) + \mathcal{L}_{3DPose}(W) + \mathcal{L}_{2DPose}(W) + \mathcal{L}_{2DPart}(W)$$

Laplacian prior

$$\delta_i = \sum_{\{i,j\}\in E} w_{ij}(v_i - v_j) = v_i - \left[\sum_{\{i,j\}\in E} w_{ij} v_j\right]$$

where $\sum_{\{i,j\}\in E} w_{ij} = 1$. To compute the Laplacians for the human body mesh, assume we have n vertices in the mesh, which means $V = [v_1, v_2, \cdots, v_n]^T$. We can use a $n \times n$ Laplacian matrix:

$$L_{i,j} = \begin{cases} w_{ij} & \text{if } \{i,j\} \in E \\ -1 & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

and compute the Laplacians $\Delta = [\delta_1, \delta_2, \cdot, \delta_n]^T$ using

$$\Delta = LV. \quad (7)$$

$$\mathcal{L}_{Lap}(W) = \sum_{d\in\{x,y,z\}} -\log \sum_{k=1}^{K} \hat{\phi}_{dk} \frac{\exp\left(-\frac{1}{2}(\Delta_d - \hat{\mu}_{dk})^T \hat{\Sigma}_{dk}^{-1}(\Delta_d - \hat{\mu}_{dk})\right)}{\sqrt{2\pi\hat{\Sigma}_{dk}}}$$

$$\mathcal{L}_{3DPose}(W) = \frac{1}{K}\sum_{i=1}^{K} \left|\left|J_{3D} - \bar{J}_{3D}\right|\right|_1$$

$$\mathcal{L}_{2DPose}(W) = \frac{1}{K}\sum_{i=1}^{K} \left|\left|J_{2D} - \bar{J}_{2D}\right|\right|_1$$

$$\mathcal{L}_{2DPart}(W) = \frac{1}{Z}\sum_{i=1}^{Z} \left|\left|B_{2D} - \bar{B}_{2D}\right|\right|_2^2$$

results

Laplacian prior	Part Seg Loss	mean Per-Vertex-Error
✗	✓	240.3
✓	✗	91.3
✓	✓	81.5

arxiv 2020

EllipBody: A Light-weight and Part-based Representation for Human Pose and Shape Recovery

Min Wang[1]
Shanghai Jiao Tong Unviersity
yinger650@sjtu.edu.cn

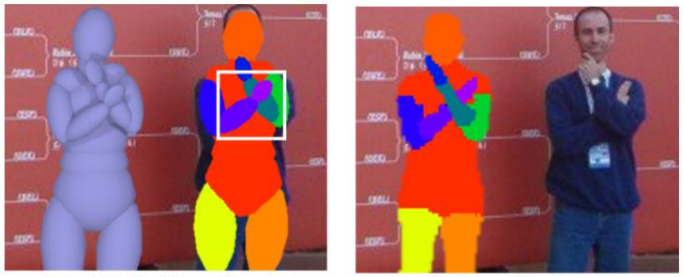
Qiu Feng[1]
Shanghai Jiao Tong Unviersity
phonic_chiu@sjtu.edu.cn

Wentao Liu[2]
Sensetime
liuwentao@sensetime.com

Chen Qian[2]
Sensetime
qianchen@sensetime.com

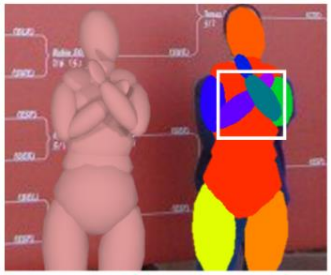
Xiaowei Zhou[3]
Zhengjiang University
xzhou@cad.zju.edu.cn

Lizhuang Ma[1]
Shanghai Jiao Tong University
ma-lz@cs.sjtu.edu.cn



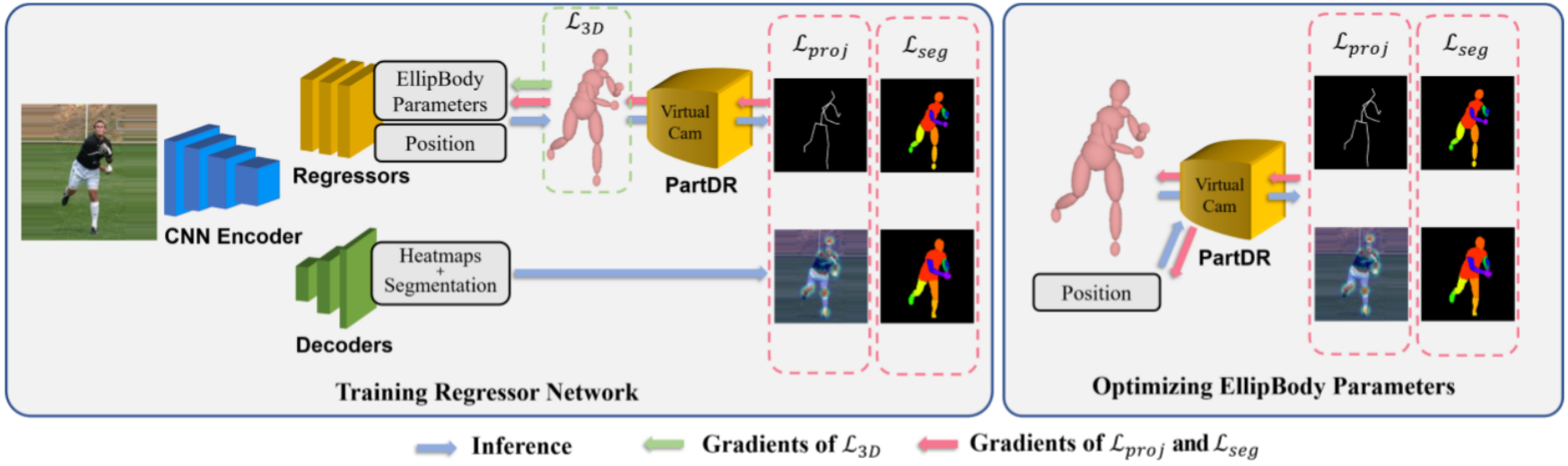
(a) 2D✓ Part Seg.×

(b) GT Part Seg.

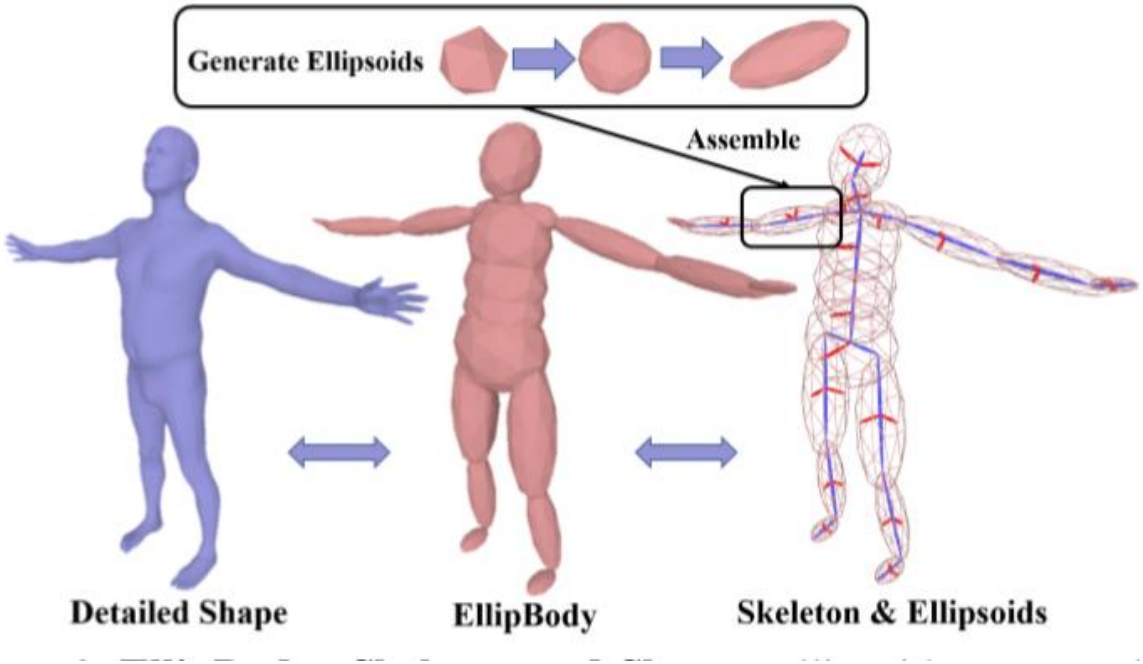


(c) 2D✓ Part Seg.✓

(d) Detailed model



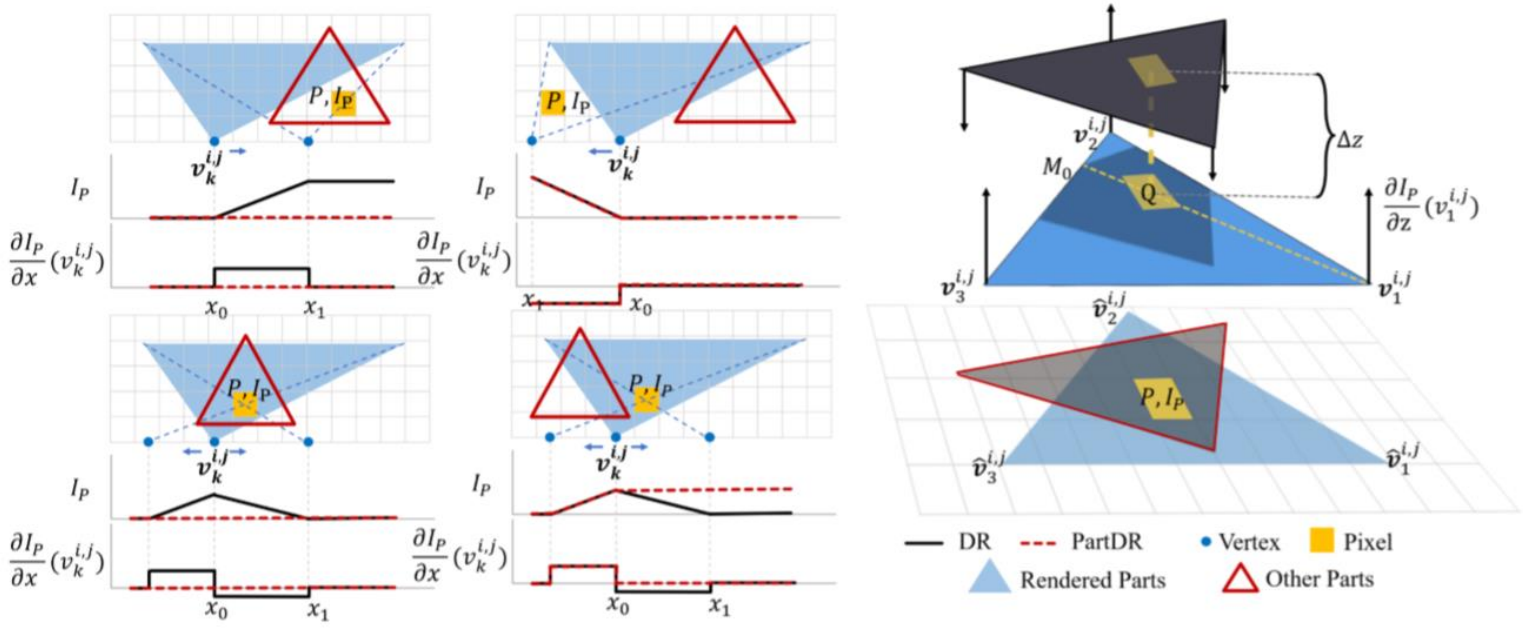
EllipBody



R 方向
C 中心位置
X 形状:
I+t*2

$$\mathbb{M} = \{\mathcal{E}_i | i = 1, ..., 20\}, \text{where}$$
$$\mathcal{E}_i = E(\boldsymbol{R_i}, \boldsymbol{C_i}, \boldsymbol{X_i}).$$

Segmentation/Heatmap loss backward



$$\frac{\partial I_P}{\partial z}(\boldsymbol{v}_k^{i,j}) = \lambda \cdot \delta I_P \cdot \log\left(\frac{\Delta(M_0, Q)}{\Delta(M_0, \boldsymbol{v}_k^{i,j}) \cdot \Delta z} + 1\right)$$

Loss

$$\mathcal{L} = \lambda_{3D}\mathcal{L}_{3D} + \lambda_{proj}\mathcal{L}_{proj} + \lambda_{seg}\mathcal{L}_{seg}.$$

$$\mathcal{L}_{3D} = \|\boldsymbol{S} - \hat{\boldsymbol{S}}\|_2^2$$

$$\mathcal{L}_{proj} = \|\boldsymbol{S}_{2D} - \hat{\boldsymbol{S}_{2D}}\|_1$$

$$\mathcal{L}_{seg} = \sum_{k=1}^{K}\sum_{i=1}^{w}\sum_{j=1}^{h}\|\mathcal{A}_{(i,j)}^{k} - \hat{\mathcal{A}}_{(i,j)}^{k}\|_2^2.$$

$$E(\theta, \beta) = \sum_{k=1}^{K} E_{ICP}(\{\boldsymbol{v}_k\}, \{\hat{\boldsymbol{v}}_k\}). \qquad (13)$$

E_{ICP} is the Iterative Closest Points (ICP) process [5] and $\{\boldsymbol{v}_k\}$ are vertices in k-th part of EllipBody, $\{\hat{\boldsymbol{v}}_k\}$ are vertices in corresponding part of SMPL model.

results


Image
GraphCMR
EllipBody Part Seg
EllipBody
Ours
Image
GraphCMR
EllipBody Part Seg
EllipBody
Ours

CVPR18

Monocular 3D Pose and Shape Estimation of Multiple People in Natural Scenes

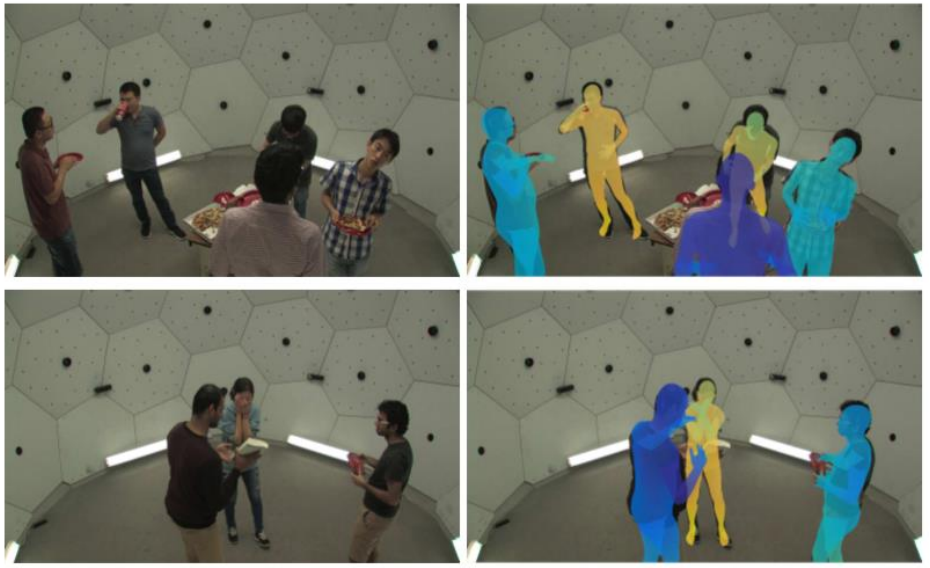
The Importance of Multiple Scene Constraints

Andrei Zanfir[2*] Elisabeta Marinoiu[2*] Cristian Sminchisescu[1,2]

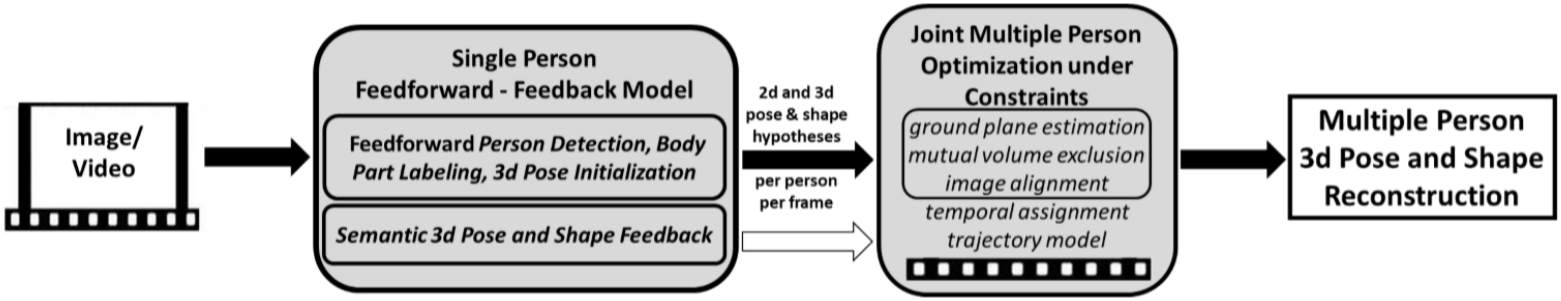
{andrei.zanfir, elisabeta.marinoiu}@imar.ro, cristian.sminchisescu@math.lth.se

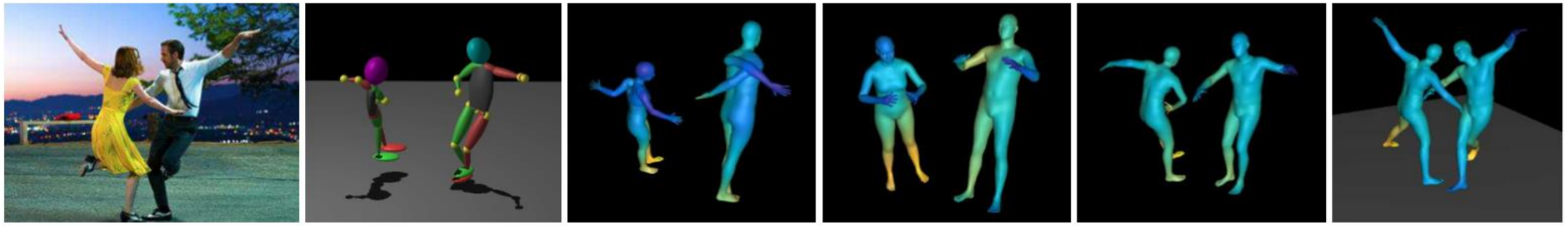
[1]Department of Mathematics, Faculty of Engineering, Lund University

[2]Institute of Mathematics of the Romanian Academy



结构





(a) (b) (c) (d) (e) (f)

Figure 3: **3d pose transfer from DMHS to SMPL**. (a) input image. (b) 3d joints with links, as estimated by DMHS. (c) transfer after applying (4) directly minimizing Euclidean distances between common 3d joints in both representations. Notice unnatural body shape and weak perceptual resemblance with the DMHS output. (d) is also obtained using (4) but with extra regularization on pose angles – offering plausible configurations but weak fits. (e) transfer results obtained using our proposed cost (5) which preserves limb orientation, and (f) inferred configurations after our semantic optimization, initialized by (e).

SMPL

DMHS [23] is a state-of-the-art feedforward multi-task deep neural network for human sensing that provides, for a given image $\mathbf{I} \in \mathbb{R}^{W \times H \times 3}$, the following estimates: the 2d and 3d joints of a single person as well as the semantic body parts at pixel level. We denote these 3 outputs by the matrices $\mathbf{y}^{3D} \in \mathbb{R}^{m_J \times 3}$, $\mathbf{y}^{2D} \in \mathbb{R}^{m_J \times 2}$ and $\mathbf{y}^{s} \in \mathbb{R}^{N_s \times W \times H}$, respectively. We denote by $m_J = 17$ the number of joints in the representation considered by the network and $N_s = 25$ the number of semantic body parts. The method has been

$$L_I^{p,f} = L_S^{p,f} + L_G^{p,f} + L_R^{p,f} + \sum_{\substack{p'=1 \\ p' \neq p}}^{N_p} L_C^f(p, p'), \quad (1)$$

where the cost L_S takes into account the visual evidence computed in every frame in the form of semantic body part labeling, L_C penalizes simultaneous (3d) volume occupancy between different people in the scene, and L_G incorporates the constraint that some of the people in the scene may have a common supporting plane. The term $L_R^{p,f} = L_R^{p,f}(\boldsymbol{\theta})$ is

$$\Phi_{3d}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{1}{|\mathcal{C}_J|} \sum_{i,j \in \mathcal{C}_J} \left\| \mathbf{y}^{3D}(i) - (\mathbf{x}(j) - \mathbf{x}(h)) \right\|$$

$$\Phi_{\cos}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{1}{|\mathcal{C}_L|} \sum_{(i,j),(k,l) \in \mathcal{C}_L} 1 - \frac{\langle \mathbf{a}_{ij}, \mathbf{b}_{kl} \rangle}{\|\mathbf{a}_{ij}\| \, \|\mathbf{b}_{kl}\|}$$

$$\boldsymbol{\theta}^0 = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \, \Phi_{\cos}(\mathbf{y}^{3D}, \boldsymbol{\theta}, \boldsymbol{\beta})$$

$$\boldsymbol{\beta}^0 = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, \Phi_{3d}(\mathbf{y}^{3D}, \boldsymbol{\theta}^0, \boldsymbol{\beta})$$

语义分割

$$\Phi_J = \sum_{j=1}^{m} w_j \left\| \mathbf{y}^{2d}(j) - \mathcal{P}(\mathbf{x}(j) + \mathbf{t}) \right\|$$

$$\Phi_S = \sum_{k=1}^{N_S} \sum_{\substack{\mathbf{p} \\ f_S(\mathbf{p})=k}} \mathbf{y}^S(\mathbf{p}, k) \min_{1 \le j \le N_S} \left\| \mathbf{p} - \mathbf{p}_j^k \right\|$$

$$\Phi_C(p, p') = \sum_{i=1}^{N_b} \sum_{j=1}^{N_b} \exp\left[- \alpha d(\mathbf{c}(p, i), \mathbf{c}(p', j)) \right]$$

$$d(\mathbf{c}, \mathbf{c}') = \frac{\|\mathbf{c} - \mathbf{c}'\|^2}{r^2 + r'^2}.$$

$$\mathbf{n}^* = \operatorname*{argmin}_{\mathbf{n}} \sum_i w_i \left| (\mathbf{x}_i - \mathbf{p})^\top \frac{\mathbf{n}}{\|\mathbf{n}\|} \right|_1 + \alpha \left| 1 - \mathbf{n}^\top \mathbf{n} \right|_1$$

results