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

Abstract

Human sensing has greatly benefited from recent advances in deep learning, parametric human modeling, and large scale 2d and 3d datasets. However, existing 3d models make strong assumptions about the scene, considering either a single person per image, full views of the person, a simple background or many cameras. In this paper, we leverage state-of-the-art deep multi-task neural networks and parametric human and scene modeling, towards a fully automatic monocular visual sensing system for multiple interacting people, which (i) *infers the 2d and 3d pose and shape of multiple people from a single image, relying on detailed semantic representations at both model and image level, to guide a combined optimization with feedforward and feedback components,* (ii) *automatically integrates scene constraints including ground plane support and simultaneous volume occupancy by multiple people, and* (iii) *extends the single image model to video by optimally solving the temporal person assignment problem and imposing coherent temporal pose and motion reconstructions while preserving image alignment fidelity. We perform experiments on both single and multi-person datasets, and systematically evaluate each component of the model, showing improved performance and extensive multiple human sensing capability. We also apply our method to images with multiple people, severe occlusions and diverse backgrounds captured in challenging natural scenes, and obtain results of good perceptual quality.*

1. Introduction

Accurately detecting and reconstructing multiple people, possibly involved in interactions with each other and with the scene, based on images and video data, has extensive applications in areas as diverse as human-computer interaction, human behavioral modeling, assisted therapy, monitoring sports performances, protection and security, special effects, modeling and indexing archival footage, or self-driving cars.

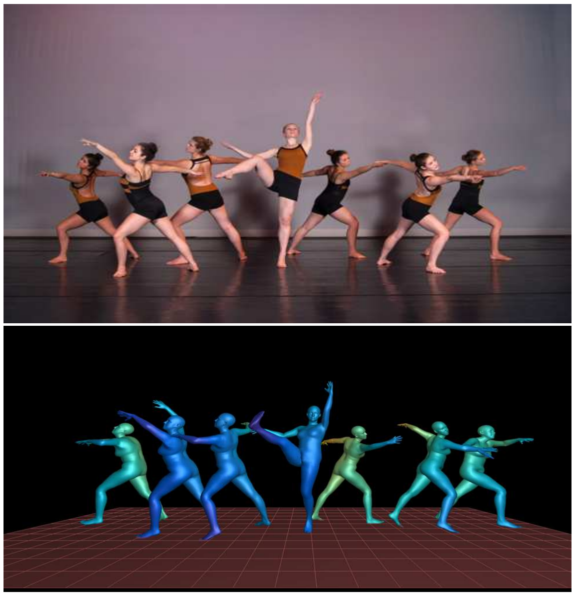

Figure 1: **Automatic 3d reconstruction** of the pose and shape of multiple people from a monocular image, as estimated by our model integrating person and scene constraints. We leverage feedforward and semantic feedback calculations for each person, with joint reasoning on ground plane and volume occupancy over all the people in the scene.

To support the level of modeling accuracy required by such applications, we ultimately need highly-detailed models able not just to detect people and their body joints in images, but also the spatial extent of body parts, as well as the three-dimensional pose, shape and motion for each person in the scene. For complex scenes, such demands would likely require a virtuous cycle between 2d and 3d reasoning, with feedback. One should further consider integrating anthropometry constraints, avoiding geometric collisions between the estimated models of multiple people, and reasoning about ground planes implicit in many scenes, as people rarely float,

[*]Authors contributed equally

unsupported in space – and if so, usually not for long. Reconstructions must also be temporally fluid and humanly plausible. Most importantly, constraints need to be enforced in the context of an image observation process which – even with many cameras pointed at the scene – remains incomplete and uncertain, especially in scenarios where multiple people interact. While the integration of such constraints appears challenging, their use provides the opportunity to considerably restrict the degrees of freedom of any natural human parameterization towards plausible solutions.

In this paper, we address the monocular inference problem for multiple interacting people, by providing a model for 2d and 3d pose and shape reconstruction over time. Our contributions include *(i)* a semantic feedforward-feedback module that combines 2d human joint detection, semantic segmentation, and 3d pose prediction of people, with pose and shape refinement based on a novel semantic cost that aligns the model body parts with their corresponding semantic images regions, producing solutions that explain the complete person layout while taking into account its estimation uncertainty, *(ii)* incorporation of scene consistency measures including automatic estimation and integration of ground plane constraints, as well as adaptively avoiding simultaneous volume occupancy by several people, and *(iii)* resolution of the temporal person assignment problem based on body shape, appearance and motion cues within a Hungarian matching method, then solving a joint multiple-person smoothing problem under both 2d projection and 3d pose temporal fluidity constraints. Our quantitative results on datasets like Panoptic [12] and Human3.6M [11] validate the importance of the ingredients in the proposed design. Qualitative results in complex monocular images and video show that the model is able to reconstruct multiple interacting people in challenging scenes in a perceptually plausible way. The model also supports the realistic synthesis of human clothing and appearance (human appearance transfer) as shown in our companion paper [39].

2. Related Work

Our work relates to recently developed deep architectures for 2d human pose estimation [4, 9, 21, 35, 36], 3d human pose estimation based on fitting volumetric models [2, 15], feedforward deep models for 3d prediction [18, 22, 40], as well as integrated deep models for 2d and 3d reasoning [23, 27, 34, 19]. Accurate shape and motion-capture systems, based on multiple cameras or simplified backgrounds, have also been proposed with impressive reconstruction results [3, 7, 13, 26]. Systems designed for the 3d reconstruction of multiple people are relatively rare and existing ones are based on multiple cameras [1, 6, 5, 12, 14]. In [6], the method uses an arguably low number of cameras (3-4) to reconstruct several people, with promising results, but the level of interaction is somewhat limited. The work of [12] proposes a multi-person tracking system (which we also use for our 'ground-truth' monocular evaluation), although the system relies on a massive number of RGB and RGB-D cameras for inference, and the capture dome offers inherently limited background variability. Our single person initialization relies on the Deep Multitask Human Sensing Network (DMHS) [23] for initial 2d and 3d pose inference (body joints, semantic segmentation, pose prediction), which is then refined based on our own implementation of the human body model SMPL [15], augmented with learned semantic vertex labeling information, and using a new semantic loss function, which represents one of our contributions. Systems based on discriminative-generative (feedforward-feedback) components for 3d human pose estimation date, in principle, back at least to [25, 28, 31] but our approach leverages considerably different image representations, body models, cost functions and optimization techniques. Our automatic ground plane and adaptive people volume occupancy exclusion constraints, as well as our multiple people assignment and smoothing costs are integrated in a novel and coherent way, although monocular single person costs based on simpler model formulations and/or multiple hypotheses tracking techniques exist in the literature [2, 20, 24, 29, 30, 38].

3. Multiple Persons in the Scene Model

Problem formulation. Without loss of generality, we consider N_p uniquely detected persons in a video with N_f frames. Our objective is to infer the best pose state variables $\mathbf{\Theta} = [\boldsymbol{\theta}_p^f] \in \mathbb{R}^{N_p \times N_f \times 72}$, shape parameters $\mathbf{B} = [\boldsymbol{\beta}_p^f] \in \mathbb{R}^{N_p \times N_f \times 10}$ and individual person translations $\mathbf{T} = [\mathbf{t}_p^f] \in \mathbb{R}^{N_p \times N_f \times 3}$, with $p \in N_p$ and $f \in N_f$. We start by first writing a per-frame, person-centric objective function $L_I^{p,f}(\mathbf{B}, \mathbf{\Theta}, \mathbf{T})$

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

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 a Gaussian mixture prior similar to [2]. The image cost for multiple people under all constraints can be written as

$$L_I^f = \sum_{p=1}^{N_p} L_I^{p,f} \quad (2)$$

If a monocular video is available, the static cost L^f is augmented with a trajectory model applicable to each person

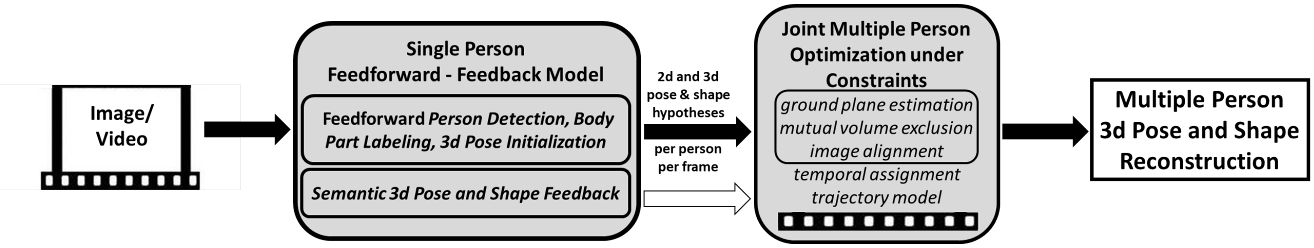
Image/
Video
Single Person
Feedforward - Feedback Model
Feedforward Person Detection, Body Part Labeling, 3d Pose Initialization
Semantic 3d Pose and Shape Feedback
2d and 3d pose & shape hypotheses
per person per frame
Joint Multiple Person Optimization under Constraints
ground plane estimation
mutual volume exclusion
image alignment
temporal assignment
trajectory model
Multiple Person 3d Pose and Shape Reconstruction

Figure 2: **Processing pipeline** of our monocular model for the estimation of 3d pose and body shape of multiple people. The system combines a single person model that incorporates feedforward initialization and semantic feedback, with additional constraints such as ground plane estimation, mutual volume exclusion, and joint inference for all people in the scene. For monocular video, the 3d temporal assignment of people is resolved using a Hungarian method, and trajectory optimization is performed jointly over all people and timesteps, under all constraints, including image consistency, for optimal results.

once the temporal assignment throughout the entire video has been resolved. The complete video loss writes

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_T = \sum_{p=1}^{N_p} \sum_{f=1}^{N_f} \left(L_I^{p,f} + L_T^{p,f} \right) \tag{3}$$

where L_T can incorporate prior knowledge on human motion, ranging from smoothness, assumptions of constant velocity or acceleration, or more sophisticated models learned from human motion capture data. In the next sections, we describe each cost function in detail.[1]

In order to infer the pose and 3d position of multiple people we rely on a parametric human representation, SMPL [15], with a state-of-the-art deep multitask neural network for human sensing, DMHS [23]. In practice, we cannot assume a constant number of people throughout a video and we first infer the parameters $\mathbf{B}$,$\mathbf{\Theta}$, $\mathbf{T}$ independently for each frame by minimizing the sum of the first two cost functions: L_S and L_C. Then, we temporally track the persons obtained in each frame by means of optimally solving an assignment problem, then re-optimize the objective, by adding the temporal and ground plane constraints, L_T and L_G. For those people detected in only some of the frames, optimization will proceed accordingly over the corresponding subset. An overview of the method is shown in fig. 2.

3.1. Single Person Feedforward-Feedback Model

SMPL [15] is a differentiable parametric human model – represented by template vertices $\mathbf{V_0}$ – and controlled by pose vectors $\boldsymbol{\theta} \in \mathbb{R}^{1\times72}$ and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{1\times10}$. The pose of the model is defined by a standard skeletal rig that has the main body joints. For each body part, the vectors controlling the pose are provided in axis-angle representations of the relative rotations w.r.t. their parents in the kinematic tree. The axis angle for every joint is transformed to a rotation matrix using the Rodrigues transformation. The shape parameters $\boldsymbol{\beta}$ impact limb size, height and weight and represent coefficients of a low dimensional shape space learned from registered meshes. SMPL provides matrix functions dependent on $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, namely $\mathbf{V}(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathbf{V}_0) \in \mathbb{R}^{n_V\times3}$, which gives the transformed vertex positions for the whole mesh, and $\mathbf{J}(\boldsymbol{\theta}, \boldsymbol{\beta}|V_0) \in \mathbb{R}^{n_J\times3}$, which outputs the joint positions for the associated kinematic tree. The total number of vertices in the SMPL model is $n_V = 6890$ and the total number of joints in the kinematic tree is $n_J = 24$. For simplicity of explanation, let $\mathbf{v}$ denote $\mathbf{V}(\boldsymbol{\theta}, \boldsymbol{\beta}|\mathbf{V}_0)$ and let $\mathbf{x}$ be $\mathbf{J}(\boldsymbol{\theta}, \boldsymbol{\beta}|V_0)$. We refer to the translation of the model in camera space as $\mathbf{t} \in \mathbb{R}^{1\times3}$.

DMHS [23] is a state-of-the-art feedforward multi-task deep neural network for human sensing that provides, for a given image $\mathbf{I} \in \mathbb{R}^{W\times H\times3}$, 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 $\mathbf{y}^{3D} \in \mathbb{R}^{m_J\times3}$, $\mathbf{y}^{2D} \in \mathbb{R}^{m_J\times2}$ 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 shown to perform well for both indoor images as well as outdoor. The challenges of integrating DMHS and SMPL stem from accurately fitting (transferring) the parametric SMPL model to the 3d joint positions predicted by DMHS, as well as designing semantic-based cost functions that allow to efficiently couple the model to the observations – perform 3d fitting in order to best explain the human layout in the image. In order to semantically assign model mesh components to corresponding image regions, one needs a consistent 'coloring' of their vertices according to the N_S human body part labels available e.g. in Human3.6M [11]. This can be achieved robustly, during a training process. We project and fit the SMPL model in multiple (4) views and for different ground truth poses from Human3.6M (we chose 100 different poses). Then each model vertex was associated the median image body part label, available in Human3.6M, transferred from images to the corresponding

[1] Whenever unambiguous, we drop the f and p super-scripts.

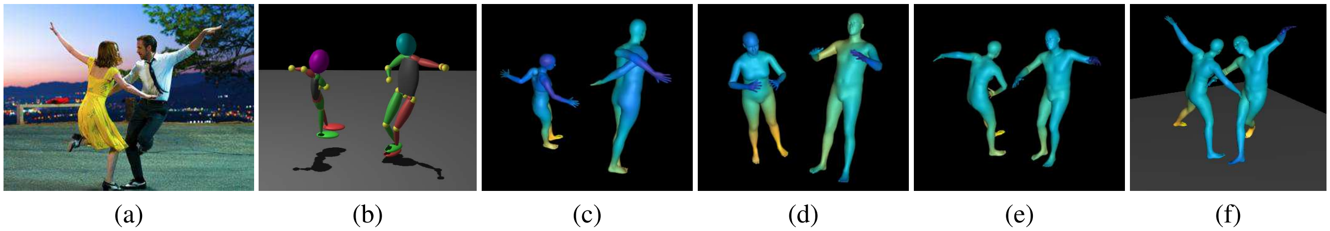

Figure 3: **3d pose transfer from DMHS to SMPL**. (a) input image. (b) 3d joints with links, as estimated by DMHS. (c) transfer after applying (4) directly minimizing Euclidean distances between common 3d joints in both representations. Notice unnatural body shape and weak perceptual resemblance with 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).

vertex projections. See fig. 4 for coloring examples.

3.1.1 Feedforward Prediction, Pose & Shape Transfer

We detail the transfer procedure for a single person and perform the same steps for all people in each frame of a video. To transfer the *feedforward prediction* of the configuration of joints $\mathbf{y}^{3D}$ obtained from DMHS to the SMPL model, we have to define a cost function $\Phi_{3d}(\boldsymbol{\theta}, \boldsymbol{\beta})$, and infer optimal $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ parameters. One such cost function is the Euclidean distance between joints shared in both representations (i.e. $i, j \in \mathcal{C}_J$, where $1 \leq i \leq m_J$ and $1 \leq j \leq n_J$ and $\mathcal{C}_J$ is the set of compatible joint indices)

$$\Phi_{3d}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{1}{|\mathcal{C}_J|} \sum_{i,j \in \mathcal{C}_J} \left\| \mathbf{y}^{3D}(i) - (\mathbf{x}(j) - \mathbf{x}(h)) \right\| \quad (4)$$

where h indicates the index of the pelvis and $\mathbf{x}(j) - \mathbf{x}(h)$ represents the centered 3d pose configuration with respect to the pelvis joint. Unless otherwise stated, we use $\|\cdot\|$ for the ℓ_2 norm, $\|\cdot\|_2$.

However, based on (4) the DMHS to SMPL transfer is unsatisfactory. This is because 1) the prediction made by DMHS is *not* necessarily a valid human shape, and 2) a configuration in the parameter space of $\boldsymbol{\beta}$ or even in the space of $\boldsymbol{\theta}$ does not necessarily represent an anatomically correct human pose. In [2], multiple regularizers were added: a norm penalty on $\boldsymbol{\beta}$ and a prior distribution on $\boldsymbol{\theta}$. However, these risk excessive bias.

We propose an alternative transfer equation, focusing on qualitatively modeling the pose predicted by DMHS so to preserve the 3d orientation of limbs. Our function $\Phi_{\cos}$ penalizes the cosine distance between limbs – or selected pairs of joints – that are shared in both representations (property denoted by $(i, j), (a, b) \in \mathcal{C}_L$ where $1 \leq i, j \leq m_J$ and $1 \leq k, l \leq n_J$). Given $\mathbf{a}_{ij} = \mathbf{y}^{3D}(i) - \mathbf{y}^{3D}(j)$ and $\mathbf{b}_{kl} = \mathbf{x}(k) - \mathbf{x}(l)$, the cost is

$$\Phi_{\cos}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \frac{1}{|\mathcal{C}_L|} \sum_{(i,j),(k,l) \in \mathcal{C}_L} 1 - \frac{\langle \mathbf{a}_{ij}, \mathbf{b}_{kl} \rangle}{\|\mathbf{a}_{ij}\| \, \|\mathbf{b}_{kl}\|} \quad (5)$$

While in practice the minimization of $\Phi_{\cos}$ converges quickly to a perfect solution (often close to 0) and the resulting pose is perceptually similar to DMHS, the implicit shape information provided by DMHS is lost. In situations where the original 3d joint prediction confidence is high (e.g. training and testing distributions are expected to be similar, as in Human3.6M), one can further optimize over $\boldsymbol{\beta}$, starting from solutions of (5)

$$\boldsymbol{\theta}^0 = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \Phi_{\cos}(\mathbf{y}^{3D}, \boldsymbol{\theta}, \boldsymbol{\beta}) \quad (6)$$

$$\boldsymbol{\beta}^0 = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, \Phi_{3d}(\mathbf{y}^{3D}, \boldsymbol{\theta}^0, \boldsymbol{\beta}) \quad (7)$$

Results of the proposed transfer variants are shown in fig. 3.

3.1.2 Semantic 3d Pose and Shape Feedback

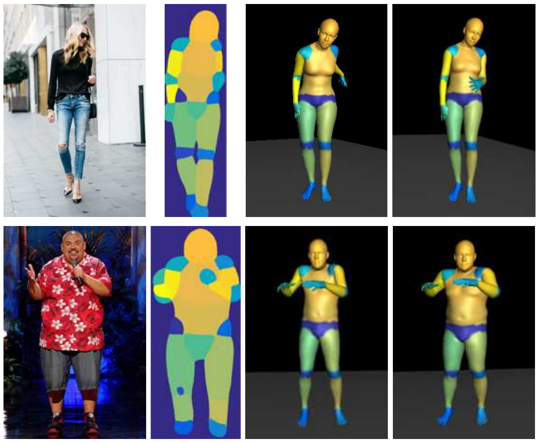

Figure 4: **Importance of semantic feedback** in capturing the correct 3d pose and body shape. From left to right: input image, semantic body labeling produced by DMHS, inferred body shape and pose without the semantic term (Φ_J only) and the semantically fitted model Φ_S.

After transferring the pose from DMHS to SMPL we obtain an initial set of parameters $\boldsymbol{\theta}^0$ and $\boldsymbol{\beta}^0$ and one can refine the initial DMHS estimate. One way to fit the 3d pose and shape model starting from an initialization [2, 32], is to minimize the projection error between the model joints and the corresponding detected image joint locations, $\mathbf{y}^{2d}$. We denote by $\mathcal{P}(\cdot)$ the image projection function, with fixed camera intrinsincs. One possible loss is the Euclidean distance, computed over sparse joint sets weighted by their detection confidence w (some may not be visible at all)

$$\Phi_J = \sum_{j=1}^{m} w_j \left\| \mathbf{y}^{2d}(j) - \mathcal{P}(\mathbf{x}(j) + \mathbf{t}) \right\| \tag{8}$$

The problem of minimizing $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ for monocular error functions, defined over distances between sparse sets of joints, is its ambiguity, as the system is clearly underdetermined, especially for depth related state space directions that couple along camera's ray of sight [33]. We propose a new error function based on projecting the mesh $\mathbf{v}$ in the image $\mathbf{I}$ and measuring the dense, pixel-wise semantic error between the semantic segmentation transferred by the model projection and a given DMHS semantic body part segmentation prediction $\mathbf{y}^S$.

We are given N_S semantic classes that describe body parts with $\mathbf{y}^S$ storing semantic confidence maps. We construct a function $f_S(\mathbf{p} = (x,y)^\top) = \text{argmax}_k\, \mathbf{y}^S(\mathbf{p},k)$ with $1 \leq x \leq W, 1 \leq y \leq H$ integers, that returns the body part label $1 \leq k \leq N_S$ of pixel location $\mathbf{p}$ in the image $\mathbf{I}$. Let $\mathbf{v}^k$ be vertices pertaining to the body part indexed in k and $\mathbf{p}^k = \mathcal{P}(\mathbf{v}^k(j) + \mathbf{t})$ their image projection.

We design a cost $\Phi_S(\boldsymbol{\Theta}, \mathbf{B}, \mathbf{T})$, where each point $\mathbf{p}$ from the semantic body part segmentation maps finds its nearest-neighbour in $\mathbf{p}^k$, and *drags* it in place. Appropriately using pixel label confidences (x,y) for a given class k as $\mathbf{y}^S$ is important for robust estimates in a cost that writes

$$\Phi_S = \sum_{k=1}^{N_S} \sum_{\substack{\mathbf{p} \\ f_S(\mathbf{p})=k}} \mathbf{y}^S(\mathbf{p},k) \min_{1 \leq j \leq N_S} \left\| \mathbf{p} - \mathbf{p}_j^k \right\| \tag{9}$$

In practice, our semantic cost is further weighted by a normalization factor $1/Z$, with $Z = \sum_{k=1}^{n_s} [\![f_S(\mathbf{p}) = k]\!]$ ensuring ϕ_S remains stable to scale transformations impacting the area of the semantic map (closer or further away, with larger or smaller number of pixels, respectively). Another desirable property of the semantic loss is that when confidences are small, Φ_S will have a lower weight in the total loss, emphasizing other qualitatively different terms in the cost. The total semantic loss can then be written

$$L_S = \Phi_J + \Phi_S \tag{10}$$

3.2. Simultaneous Volume Occupancy Exclusion

To ensure that estimated models of people in a scene are not inconsistent, by occupying the same 3d space volume simultaneously, we need additional processing. We design adaptive representations to first compute enclosing parallelepipeds for each person according to its current model estimates, rapidly test for intersections (far-range check), and only integrate detailed, close range collision avoidance into the loss when the far-range response is negative. For close-range volume occupancy exclusion, we use specialized terms obtained as follows: for each person model, we fit tapered superquadrics to each limb, and represent the limb by a series of N_b fitted spheres inside the superquadric, with centers $\mathbf{c}$ and radius r. For any two persons, p and p', we define the loss $L_C(p,p')$ based on distances between all spheres belonging, respectively, to the first and second person

$$\Phi_C(p,p') = \sum_{i=1}^{N_b} \sum_{j=1}^{N_b} \exp\Big[-\alpha d(\mathbf{c}(p,i), \mathbf{c}(p',j)) \Big] \tag{11}$$

$$d(\mathbf{c},\mathbf{c}') = \frac{\|\mathbf{c} - \mathbf{c}'\|^2}{r^2 + r'^2}. \tag{12}$$

The loss for N_p persons in a frame is defined as the sum over all pair-wise close-range losses $L_C(p,p')$ among people with negative far-range tests. People with positive far-range tests do not contribute to the volume occupancy loss. Notice how this cost potentially couples parameters from all people and requires access to their estimates. See fig. 5 for visual illustrations.

3.3. Ground Plane Estimation and Constraint

We include a prior that the scene has a ground-plane on which, on average, the subjects stand and perform actions. To build a correct hypothesis for the location and orientation of the plane, we design a cost that models interactions between the plane and all human subjects, but leaves room for outliers, including people who, temporarily or permanently, are not in contact with the ground. Specifically, we select the 3d ankle positions of all persons in all the frames of a video, be these $\mathbf{x}_i$, and fit a plane to their locations.

We assume that a point $\mathbf{z}$ is on a plane with a surface normal $\mathbf{n}$ if the following equation is satisfied $(\mathbf{z} - \mathbf{p})^\top \mathbf{n} = 0$, where $\mathbf{p}$ is any fixed point on the plane. Given that some of the ankles might be occluded, we use a confidence term to describe the impact they have on the fitting process. We use the confidence w_i from the DMHS 2d joint detector, with a two-folded purpose to 1) select the initial point $\mathbf{p}$ belonging to the plane as the weighted median of all ankle locations of the detected persons, and 2) weight measurements used in the robust L1 norm estimate of the plane hypothesis. Our plane estimation objective is

$$\mathbf{n}^* = \underset{\mathbf{n}}{\text{argmin}} \sum_i w_i \left| (\mathbf{x}_i - \mathbf{p})^\top \frac{\mathbf{n}}{\|\mathbf{n}\|} \right|_1 + \alpha \left| 1 - \mathbf{n}^\top \mathbf{n} \right|_1 \tag{13}$$

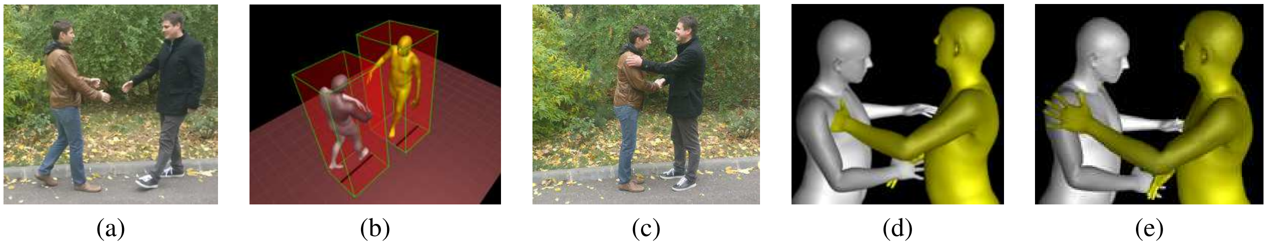

Figure 5: **Adaptive volume occupancy avoidance.** (a) input image where people are far apart, (b) visual representation for far-range collision check. (c) image where people are in contact. (d) inferred body shapes without and (e) with collision constraint, which ensures correct contact without simultaneous volume occupancy.

The estimates $(\mathbf{p}, \mathbf{n}^*)$ are then used in the ground-plane constraint term L_G to penalize configurations with 3d ankle joints estimates away from the plane

$$L_G^{p,f} = \left|(\mathbf{x}_l - \mathbf{p})^\top \mathbf{n}^*\right|_1 + \left|(\mathbf{x}_r - \mathbf{p})^\top \mathbf{n}^*\right|_1 \tag{14}$$

Where the subscripts l and r identify the left and the right ankles for a person p at time f. The weighting of the terms is performed adaptively based on confidences w_l, w_r of the associated ankle joints. If these are not visible, or are visible within some distance of the ground and not confident, constraints are applied. If the joints are visible and confident, or far from the ground, constraints are not applied.

3.4. Assignment and Trajectory Optimization

Independently performing 3d human body pose and shape optimization in a monocular video can lead to large translation variations along depth directions and movements that lack natural smoothness. For this reason, we propose a temporal constraint that ensures for each of the inferred models that estimates in adjacent frames are smooth. To achieve it, we first need to resolve the assignment problem over time (identify or track the same individual throughout the video), then perform temporal smoothing for each individual track.

To solve the person assignment problem, we use the Hungarian algorithm to optimally build tracks based on an inter-frame inter-person cost combining the appearance consistency (measured as distances between vectors containing the median colors of the different body parts, computed over the model vertices), the body shape similarity, and the distance between the appropriately translated 3d joints inferred for each person, at frame level.

Once the assignment has been resolved between every pair of estimated person models in every successive set of frames, and tracks are built, several motion priors can be used, ranging from a constant velocity model, to more sophisticated auto-regressive processes or deep recursive predictors learned from training data [17, 8, 37]. The integration of such motion representations in our framework is straightforward as long as they remain differentiable. Here we experiment with constant velocity priors on pose angles, $\mathbf{\Theta}$ as well as translation variables, $\mathbf{T}$. Our temporal loss function component at $L_T^f = L_T^{p,f}(\mathbf{\Theta}, \mathbf{T})$ frame $f \geq 2$ for a person (track) p is defined as

$$\begin{aligned} L_T^f = &\left\|(\boldsymbol{\theta}^{f+1} - \boldsymbol{\theta}^f) - (\boldsymbol{\theta}^f - \boldsymbol{\theta}^{f-1})\right\| + \\ &\left\|(\mathbf{t}^{f+1} - \mathbf{t}^f) - (\mathbf{t}^f - \mathbf{t}^{f-1})\right\| \end{aligned} \tag{15}$$

The shape parameters $\boldsymbol{\beta}_p$ are set as the median of $\boldsymbol{\beta}_p^f, \forall f$. Because smoothing axis-angle representations is difficult, the angle-related costs in (15) are represented using quaternions, which are easier to smooth. Gradients are propagated through the axis-angle to quaternion transformation during the optimization.

4. Experiments

We numerically test our inference method on two datasets, CMU Panoptic [12] and Human3.6M [11], as well as qualitatively on challenging natural scenes (see fig. 7). On Human3.6M we test different components of the model including semantic feedback, smoothing and the effect of multiview constraints. Panoptic in turn provides the real quantitative test-bed for the complete monocular system.

Given a video with multiple people, we first detect the persons in each frame and obtain initial feedforward DMHS estimates for their 2d body joints, semantic segmentation and 3d pose. Similarly to [16], we extend DMHS to partially visible people, by fine-tuning both the semantic and the 3d pose estimation components of DMHS on a partial view version of Human80K[10]. For each person we perform the transfer proposed in (5) that aligns the limb directions of 3d estimates predicted by DMSH with the limb directions of SMPL. The transfer gives an initialization for pose and shape. The initial translation of each person is set to 3 meters in front of the camera.

Human3.6M is a large-scale dataset that contains single person images recorded in a laboratory setup using a motion capture system. The dataset has been captured using 4 synchronized RGB cameras and contains videos of 11 actors performing different daily activities. We select 3 of the most difficult actions: *sitting*, *sitting down* and *walking dog* to test

	Haggling		Mafia		Ultimatum		Pizza		Mean	
Method	Pose	Translation	Pose	Translation	Pose	Translation	Pose	Translation	Pose	Translation
DMHS [23]	217.9	-	187.3	-	193.6	-	221.3	-	203.4	-
2d Loss	**135.1**	282.3	174.5	502.2	**143.6**	357.6	177.8	419.3	157.7	390.3
Semantic Loss	144.3	260.5	179.0	459.8	160.7	376.6	178.6	413.6	165.6	377.6
Smoothing	141.4	260.3	173.6	454.9	155.2	368.0	173.1	403.0	160.8	371.7
Smoothing Ground Plane	140.0	**257.8**	**165.9**	**409.5**	150.7	**301.1**	**156.0**	**294.0**	**153.4**	**315.5**

Table 1: Automatic 3d human pose and translation estimation errors (in mm) on the Panoptic dataset (9,600 frames, 21,404 people). Notice the value of each component and the impact of the ground-plane constraint on correct translation estimation.

Method	WalkingDog	Sitting	Sitting Down
DMHS [23]	78	119	106
Semantic Loss	75	109	101
Multi View	51	71	65
Smoothing	**48**	**68**	**64**

Table 2: Mean per joint 3d position error (in mm) on the Human3.6M dataset, *evaluated on the test set* of several very challenging actions. Notice the importance of various constraints in improving estimation error.

our single-person model. We use the official left-out test set from the selected actions, consisting of 160K examples. On this dataset we can only evaluate the pose inference under MPJPE error, but without the translation relative to the camera. We show results in table 2. We obtain an improvement over DMHS by using the proposed semantic 3d pose and shape feedback, cf. (10). On this dataset, we also experiment with multi-view inference and show a consistent improvement in 3d pose estimation. For multi-view inference, the loss function proposed in (10) is easily extended as a sum over measurements in all available cameras. Adding a temporal smoothness constraint further reduces the error. We also evaluated our method on all 15 actions from the official test set (911,744 configurations) and obtain an average error of 69 mm.[2]

CMU Panoptic Dataset. We selected data from 4 activities (*Haggling*, *Mafia*, *Ultimatum* and *Pizza*) which contain multiple people interacting with each other. For each activity we selected 2 sub-sequences, each lasting 20 seconds (i.e. 600 frames), from HD cameras indices 30 and 16[3]. In total, we obtain 9,600 frames that contain 21,404 people. We do not validate/train any part of our method on this data.

Evaluation Procedure. We evaluate both the inferred pose, centered in its hip joint, under mean per joint position error (MPJPE), and the estimated translation for each person under standard Euclidean distance. We perform the evaluation for each frame in a sequence, and average the results across persons and frames. We match each ground-truth person in the scene with an estimation of our model. For every ground-truth pose, we select the closest inferred model under the Euclidean distance, in camera space.

Ablation Studies. We systematically test the main components of the proposed monocular inference system and show the results detailed for each activity in table 1. Compared to DMHS, our complete method reduces the MPJPE error significantly, from 203.4 mm to 153.4 mm on average (-25%), while also computing the translation of each person in the scene. The translation error is, on average, 315.5 mm. The semantic projection term helps disambiguate the 3d position of persons and reduces the translation error compared to using only the 2d projection term. Temporally smoothing the pose estimates decreases the translation error further. Imposing the ground plane constraint makes the most significant contribution in this setup, decreasing the total translation error from 371 mm to 315 mm (-15%). Even though the total pose error also decreases when all constraints are imposed, on some sequences (e.g. *Haggling*) the error did not decrease when semantic terms are used. At a closer look, we noticed that the semantic maps and 3d initialization from DMHS were particularly noisy on those sequences of *Haggling*, camera index 30. Qualitative results in monocular images from the Panoptic dataset are shown in fig. 6. Our method produces perceptually plausible 3d reconstructions with good image alignment in scenes with many people, some only partially visible, and captured under non-conventional viewing angles.

5. Conclusions

We have presented a monocular model for the integrated 2d and 3d pose and shape estimation of multiple people, under multiple scene constraints. The model relies on feed-forward predictors for initialization and semantic fitting for feedback and precise refinement (shape adaption) to the observed person layout. It estimates and further integrates ground plane and volume occupancy constraints, as well as temporal priors for consistent, plausible estimates, within a single joint optimization problem over the combined representation of multiple people, in space and time. Our experimental evaluation, including ablation studies, is extensive, covers both single-person and multiple-person datasets and illustrates the importance of integrating multiple constraints.

[2] Detailed results can be seen at http://vision.imar.ro/human3.6m/ranking.php (Testset H36M_NOS10).

[3] For variability only, all testing is monocular.

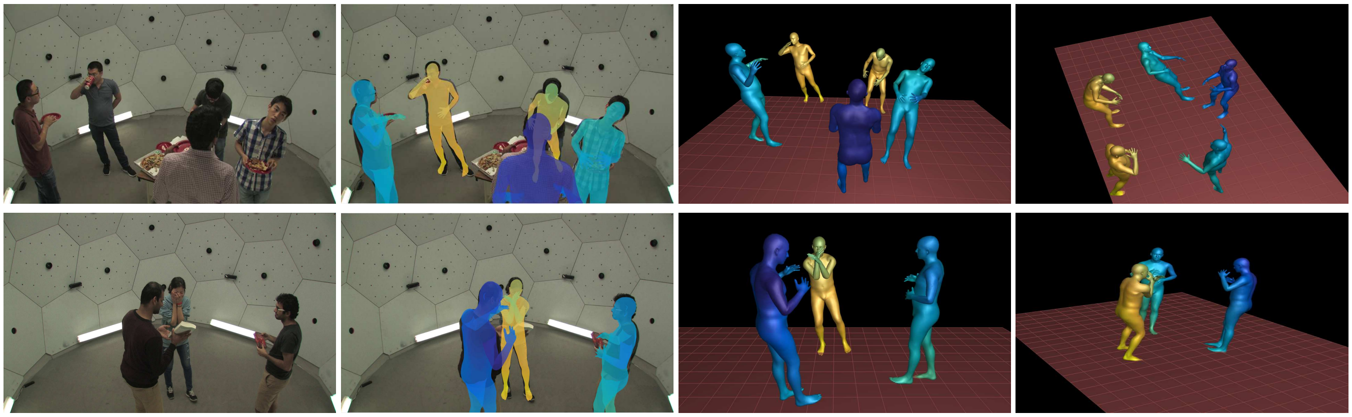

Figure 6: **Automatic monocular 3d reconstruction of multiple people in Panoptic videos**. Left to right: input image, inferred model overlaid to assess fitting quality, two different views of the 3d reconstruction. Unusual viewing angles, pose variability, partial views and occlusions, make monocular reconstruction challenging. Quantitative results are given in table 1.

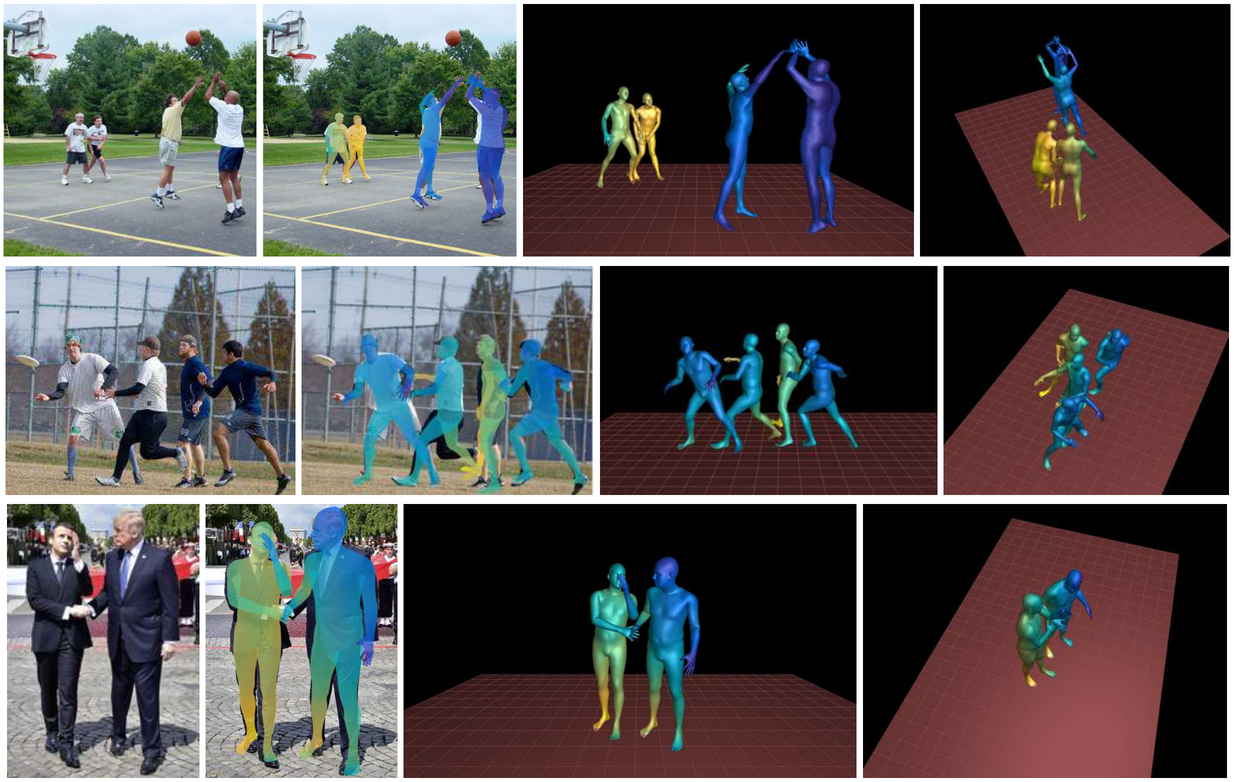

Figure 7: **Automatic 3d reconstruction of multiple people from monocular images of complex natural scenes**. Left to right: input image, inferred model overlaid, and two different views of 3d reconstructions obtained by our model (including ground plane). Challenging poses, occlusions, different scales and close interactions are correctly resolved in the reconstruction.

Moreover, we qualitatively show that the method produces 3d reconstructions with tight image alignment and good perceptual quality, in both monocular images and video filmed in complex scenes, with multiple people, severe occlusion and challenging backgrounds. To our knowledge, such a large-scale fully automatic monocular system for multiple person sensing under scene constraints has been presented here for the first time.

Acknowledgments: This work was supported in part by the European Research Council Consolidator grant SEED, CNCS-UEFISCDI PN-III-P4-ID-PCE-2016-0535, the EU Horizon 2020 Grant No. 688835 DE-ENIGMA, and SSF.

References

[1] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures for multiple human pose estimation. In *CVPR*, 2014.

[2] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016.

[3] A. Boukhayma, J.-S. Franco, and E. Boyer. Surface motion capture transfer with Gaussian process regression. In *CVPR*, 2017.

[4] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[5] A. Elhayek, E. Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In *CVPR*, 2015.

[6] A. Elhayek, E. de Aguiar, A. Jain, J. Thompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Marconi—convnet-based marker-less motion capture in outdoor and indoor scenes. *IEEE transactions on pattern analysis and machine intelligence*, 39(3):501–514, 2017.

[7] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. *Computational Studies of Human Motion: Tracking and Motion Synthesis*. NOW Publishers Inc, 2006.

[8] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics. In *ICCV*, pages 4346–4354, 2015.

[9] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016.

[10] C. Ionescu, J. Carreira, and C. Sminchisescu. Iterated second-order label sensitive pooling for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1661–1668, 2014.

[11] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 2014.

[12] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015.

[13] V. Leroy, J.-S. Franco, and E. Boyer. Multi-view dynamic shape refinement using local temporal integration. In *ICCV*, 2017.

[14] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt. Markerless motion capture of multiple characters using multiview image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2720–2735, 2013.

[15] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH*, 34(6):248:1–16, 2015.

[16] E. Marinoiu, M. Zanfir, V. Olaru, and C. Sminchisescu. 3D Human Sensing, Action and Emotion Recognition in Robot Assisted Therapy of Children with Autism. In *CVPR*, 2018.

[17] J. Martinez, M. J. Black, and J. Romero. On human motion prediction using recurrent neural networks. *CoRR*, abs/1705.02445, 2017.

[18] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

[19] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017.

[20] R. M. Neal, M. J. Beal, and S. T. Roweis. Inferring state sequences for non-linear systems with embedded hidden Markov models. In *NIPS*, 2004.

[21] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017.

[22] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 2017.

[23] A. Popa, M. Zanfir, and C. Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. In *CVPR*, 2017.

[24] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. *ECCV*, 2012.

[25] D. Ramanan and C. Sminchisescu. Training deformable models for localization. In *CVPR*, 2006.

[26] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In *ECCV*, 2016.

[27] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. In *NIPS*, 2016.

[28] L. Sigal, A. Balan, and M. J. Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*, 2007.

[29] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer. A joint model for 2d and 3d pose estimation from a single image. In *CVPR*, 2013.

[30] C. Sminchisescu and A. Jepson. Variational mixture smoothing for non-linear dynamical systems. In *CVPR*, 2004.

[31] C. Sminchisescu, A. Kanaujia, and D. Metaxas. Learning joint top-down and bottom-up processes for 3d visual inference. In *CVPR*, 2006.

[32] C. Sminchisescu and B. Triggs. Estimating Articulated Human Motion with Covariance Scaled Sampling. *IJRR*, 22(6):371–393, 2003.

[33] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *CVPR*, 2003.

[34] B. Tekin, P. Marquez Neila, M. Salzmann, and P. Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *ICCV*, 2017.

[35] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014.

[36] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014.

[37] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models. In *NIPS*, 2006.

[38] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall. A dual-source approach for 3d pose estimation from a single image. In *CVPR*, 2016.

[39] M. Zanfir, A. Popa, and C. Sminchisescu. Human appearance transfer. In *CVPR*, 2018.

[40] X. Zhou, M. Zhu, K. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, 2016.