Coherent Reconstruction of Multiple Humans from a Single Image

Wen Jiang[2*], Nikos Kolotouros[1*], Georgios Pavlakos[1], Xiaowei Zhou[2†], Kostas Daniilidis[1]
[1] University of Pennsylvania [2] Zhejiang University

Input image Baseline Ours

Figure 1: **Coherent reconstruction of pose and shape for multiple people**. Typical top-down regression baselines (center) suffer from predicting people in overlapping positions, or in inconsistent depth orderings. Our approach (right) is trained to respect all these constraints and recover a coherent reconstruction of all the people in the scene in a feedforward manner.

Abstract

In this work, we address the problem of multi-person 3D pose estimation from a single image. A typical regression approach in the top-down setting of this problem would first detect all humans and then reconstruct each one of them independently. However, this type of prediction suffers from incoherent results, e.g., interpenetration and inconsistent depth ordering between the people in the scene. Our goal is to train a single network that learns to avoid these problems and generate a coherent 3D reconstruction of all the humans in the scene. To this end, a key design choice is the incorporation of the SMPL parametric body model in our top-down framework, which enables the use of two novel losses. First, a distance field-based collision loss penalizes interpenetration among the reconstructed people. Second, a depth ordering-aware loss reasons about occlusions and promotes a depth ordering of people that leads to a rendering which is consistent with the annotated instance segmentation. This provides depth supervision signals to the network, even if the image has no explicit 3D annotations. The experiments show that our approach outperforms previous methods on standard 3D pose benchmarks, while our proposed losses enable more coherent reconstruction in natural images. The project website with videos, results, and code can be found at: https://jiangwenpl.github.io/multiperson

1. Introduction

Recent work has achieved tremendous progress on the frontier of 3D human analysis tasks. Current approaches have established impressive performance for 3D keypoint estimation [35, 57], 3D shape reconstruction [11, 62], full-body 3D pose and shape recovery [15, 26, 28, 43], or even going beyond that and estimating more detailed and expressive reconstructions [42, 63]. However, as we progress towards more holistic understanding of scenes and people interacting in them, a crucial step is the coherent 3D reconstruction of multiple people from single images.

Regarding multi-person pose estimation, on one end of the spectrum, we have bottom-up approaches. The works following this paradigm, first detect all body joints in the scene and then group them, i.e., assigning them to the appropriate person. However, it is not straightforward how bottom-up processing can be extended beyond joints (e.g., use it for shape estimation, or mesh recovery). Different from bottom-up, top-down approaches first detect all people in the scene, and then estimate the pose for each one of them. Although they take a hard decision early on (person detection), they typically rely on state-of-the-art methods for person detection and pose estimation which allows them to achieve very compelling results, particularly in the 2D pose case, e.g., [9, 56, 64]. However, when reasoning about the pose of multiple people in 3D, the problems can be more complicated than in 2D. For example, the reconstructed people can overlap each other in the 3D space, or be estimated at depths that are inconsistent with the actual depth ordering, as is demonstrated in Figure 1. This means that it is crucial to go beyond just predicting a reasonable 3D pose for each person individually, and instead estimate a coherent reconstruction of all the people in the scene.

This coherency of the holistic scene is the primary goal of this work. We adopt the typical top-down paradigm, and our aim is to train a deep network that learns to estimate a coherent reconstruction of all the people in the scene. Start-

* Equal contribution.

† X. Zhou and W. Jiang are affiliated with the State Key Lab of CAD&CG, Zhejiang University. Email: xwzhou@zju.edu.cn.

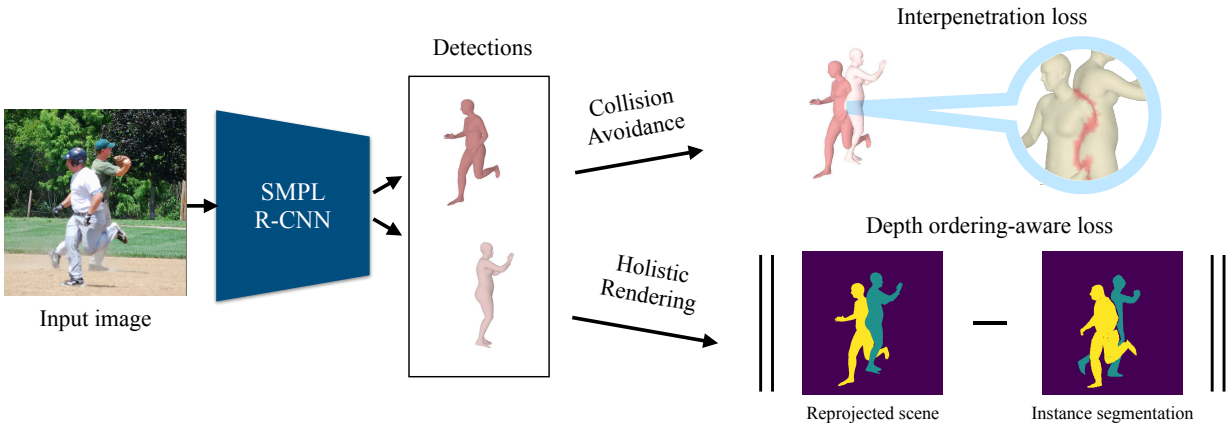

Figure 2: **Overview of the proposed approach.** We design an end-to-end framework for 3D pose and shape estimation of multiple people from a single image. An R-CNN-based architecture [19] detects all people in the image and estimates their SMPL parameters [34]. During training we incorporate constraints to promote a coherent reconstruction of all the people in the scene. First, we use an interpenetration loss to avoid people overlapping each other. Second, we apply a depth ordering-aware loss by rendering the meshes of all the people to the image and encouraging the rendered instance segmentation to match with the annotated instance masks.

ing with a framework that follows the R-CNN pipeline [48], a key decision we make is to use of the SMPL parametric model [34] as our representation, and add a SMPL estimation branch to the R-CNN. The mesh representation provided by SMPL allows us to reason about occlusions and interpenetrations enabling the incorporation of two novel losses towards coherent 3D reconstruction. First, a common problem of predictions from regression networks is that the reconstructed people often overlap each other, since the feedforward nature does not allow for holistic feedback on the potential intersections. To train a network that learns to avoid this type of collisions, we introduce an interpenetration loss that penalizes intersections among the reconstructed people. This term requires no annotations and relies on a simple property of natural scenes, i.e., that people cannot intersect each other. Besides collisions, another source of incoherency in the results is that the estimated depths of the meshes are not respecting the actual depth ordering of the humans in the scene. Equipped with a mesh representation, we render our holistic scene prediction on the 2D image plane and penalize discrepancies of this rendering from the annotated instance segmentation. This loss enables reasoning about occlusion, encouraging the depth ordering of the people in the scene to be consistent with the annotated instance masks. Our complete framework (Figure 2) is evaluated on various benchmarks and outperforms previous multi-person 3D pose and shape approaches, while the proposed losses improve coherency of the holistic result both qualitatively and quantitatively.

To summarize, our main contributions are:

- We present a complete framework for coherent regression of 3D pose and shape for multiple people.
- We train with an interpenetration loss to avoid regressing meshes that intersect each other.
- We train with a depth ordering-aware loss to promote reconstructions that respect the depth ordering of the people in the scene.
- We outperfrom previous approaches for multi-person 3D pose and shape, while recovering significantly more coherent results.

2. Related work

In this Section we provide a short description of prior works that are more relevant to ours.

Single-person 3D pose and shape: Many recent works estimate 3D pose in the form of a skeleton, e.g., [35, 39, 44, 47, 57, 59, 60, 67], or 3D shape in a non-parmetric way, e.g., [11, 53, 62]. However, here we focus on full-body pose and shape reconstruction in the form of a mesh, typically using a parametric model, like SMPL [34]. After the early works on the problem [14, 52], the first fully automatic approach, SMPLify, was proposed by Bogo *et al*. [4]. SMPLify iteratively fits SMPL on the 2D joints detected by a 2D pose estimation network [46]. This optimization approach was later extended in multiple ways; Lassner *et al*. [31] use silhouettes for the fitting, Varol *et al*. [62] use voxel occupancy grids, while Pavlakos *et al*. [42] fit a more expressive parametric model, SMPL-X.

Despite the success of the aforementioned fitting approaches, recently we have observed an increased interest in approaches that regress the pose and shape parameters directly from images, using a deep network for this task. Many works focus on first estimating some form of intermediate representation before regressing SMPL parameters. Pavlakos *et al*. [45] use keypoints and silhouettes,

Omran *et al*. [41] use semantic part segmentation, Tung *et al*. [61] append heatmaps for 2D joints to the RGB input, while Kolotouros *et al*. [29] regress the mesh vertices with a Graph CNN. Regressing SMPL parameters directly from RGB input is more challenging, but it avoids any hand-designed bottleneck. Kanazawa *et al*. [26] use an adversarial prior to penalize improbable 3D shapes during training. Arnab *et al*. [3] use temporal context to improve the regression network. Güler *et al*. [15] incorporate a test-time post-processing based on 2D/3D keypoints and DensePose [16].

Multi-person 3D pose: For the multi-person case, the top-down paradigm is quite popular for 3D pose estimation, since it capitalizes on the success of the R-CNN works [13, 48, 19]. The LCR-Net approaches [50, 51] first detect each person, then classify its pose in a pose cluster and finally regress an offset for each joint. Dabral *et al*. [10] first estimate 2D joints inside the bounding box and then regress 3D pose. Moon *et al*. [40] contribute a root network to give an estimate of the depth of the root joint. Zanfir *et al*. [65] rely on scene constraints to iteratively optimize the 3D pose and shape of the people in the scene. Alternatively, there are also approaches that follow the bottom-up paradigm. Mehta *et al*. [38] propose a formulation based on occlusion-robust pose-maps, where Part Affinity Fields [6] are used for the association problem. Follow-up work [37], improves, among others, the robustness of the system. Finally, Zanfir *et al*. [66] solve a binary integer linear program to perform skeleton grouping.

In the context of pose and shape estimation in particular, there is a limited number of works that estimate full-body 3D pose and shape for multiple people in the scene. Zanfir *et al*. [65] optimize the 3D shape of all the people in the image using multiple scene constraints. Our approach draws inspiration from this work and shares the same goal, in the sense of recovering a coherent 3D reconstruction. In contrast to them, instead of optimizing for this coherency at test-time, we train a feedforward regressor and use the scene constraints at training time to encourage it to produce coherent estimates at test-time. Using a feedforward network to estimate pose and shape for multiple people has been proposed by the work of Zanfir *et al*. [66]. However, in that case, 3D shape is regressed based on 3D joints, which are the output of a bottom-up system. In contrast, our approach is top-down, and SMPL parameters are regressed directly from pixels, instead of using an intermediate representation, like 3D joints. In fact, it is non-trivial to design a framework for SMPL parameter regression in a bottom-up manner.

Coherency constraints: An important aspect of our work is the incorporation of loss terms that promote coherent 3D reconstruction of the multiple humans. Regarding our interpenetration loss, Bogo *et al*. [4] and Pavlakos *et al*. [42] use a relevant objective to avoid self-interpenetrations of the human under consideration. In a more similar spirit to us, Zanfir *et al*. [65] use a volume occupancy loss to avoid humans intersecting each other. In different applications, Hasson *et al*. [18] penalize interpenetrations between the object and the hand that interacts with it, while Hassan *et al*. [17] penalize interpenetrations between humans and their environment. The majority of the above works uses the interpenetration penalty to iteratively refine estimates at test-time. With the exception of [18], our work is the only one that uses an interpenetration term to guide the training of a feedforward regressor and promote colliding-free reconstructions at test time.

Regarding our depth ordering-aware loss, we follow the formulation of Chen *et al*. [8], which was also used in the context of 3D human pose by Pavlakos *et al*. [44]. In contrast to them, we do not use explicit depth annotations, but instead, we leverage the instance segmentation masks to reason about occlusion and thus, depth ordering. The work of Rhodin *et al*. [49] is also relevant, where inferring depth ordering is used as an intermediate abstraction for scene decomposition from multiple views. Our work also aims to estimate a coherent depth ordering, but we do so from a single image with the guidance of instance segmentation, while we retain a more explicit human representation in terms of meshes. Finally, using instance segmentation via render and compare has also been proposed by Kundu *et al*. [30]. However, their multi-instance evaluation includes only rigid objects, specifically cars, whereas we investigate the, significantly more complex, non-rigid case.

3. Technical approach

In this Section, we describe the technical approach followed in this work. We start with providing some information about the SMPL model (Subsection 3.1) and the baseline regional architecture we use (Subsection 3.2). Then we describe in detail our proposed losses promoting interpenetration-free reconstruction (Subsection 3.3) and consistent depth ordering (Subsection 3.4). Finally, we provide more implementation details (Subsection 3.5).

3.1. SMPL parametric model

For the human body representation, we use the SMPL parametric model of the human body [34]. What makes SMPL very appropriate for our work, in comparison with other representations, is that it allows us to reason about occlusion and interpenetration enabling the use of the novel losses we incorporate in the training of our network. The SMPL model defines a function $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta})$ that takes as input the pose parameters $\boldsymbol{\theta}$, and the shape parameters $\boldsymbol{\beta}$, and outputs a mesh $M \in \mathbb{R}^{N_v \times 3}$, consisting of $N_v = 6890$ vertices. The model also offers a convenient mapping from mesh vertices to k body joints J, through a linear regressor W, such that joints can be expressed as a linear combination of mesh vertices, $J = WM$.

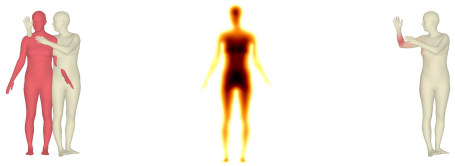

Figure 3: **Illustration of interpenetration loss.** Left: Collision between person i (red) and j (beige). Center: Distance field ϕ_i for person i, Right: Mesh M_j of person j. The vertices of M_j that collide with person i, i.e., located in non-zero areas of ϕ_i and visualized with soft red, are penalized by the interpenetration loss.

3.2. Baseline architecture

In terms of the architecture for our approach, we follow the familiar R-CNN framework [48], and use a structure that is most similar to the Mask R-CNN iteration [19]. Our network consists of a backbone (here ResNet50 [20]), a Region Proposal Network, as well as heads for detection and SMPL parameter regression (SMPL branch). Regarding the SMPL branch, its architecture is similar to the iterative regressor proposed by Kanazawa *et al*. [26], regressing pose and shape parameters, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ respectively, as well as camera parameters $\boldsymbol{\pi} = \{s, t_x, t_y\}$. The camera parameters are predicted per bounding box but we later update them based on the position of the bounding box in the full image (details in the Sup.Mat.). Although there is no explicit feedback among bbox predictions, the receptive field of each proposal includes the majority of the scene. Since each bounding box is aware of neighboring people and their poses, it can make an *informed pose prediction* that is consistent with them.

For our baseline network, the various components are trained jointly in an end-to-end manner. The detection task is trained according to the training procedure of [19], while for the SMPL branch, the training details are similar to the ones proposed by Kanazawa *et al*. [26]. In the rare cases that 3D ground truth is available, we apply a loss, L_{3D}, on the SMPL parameters and the 3D keypoints. In the most typical case that only 2D joints are available, we use a 2D reprojection loss, L_{2D}, to minimize the distance between the ground truth 2D keypoints and the projection of the 3D joints, J, to the image. Additionally, we also use a discriminator and apply an adversarial prior L_{adv} on regressed pose and shape parameters, to encourage the output bodies to lie on the manifold of human bodies. Each of the above losses is applied independently to each proposal, after assigning it to the corresponding ground truth bounding box. More details about the above loss terms and the training of the baseline model are included in the Sup.Mat.

3.3. Interpenetration loss

A critical barrier towards coherent reconstruction of multiple people from a single image is that the regression network can often predict the people to be in overlapping locations. To promote prediction of non-colliding people, we introduce a loss that penalizes interpenetrations among the reconstructed people. Our formulation draws inspiration from [17]. An important difference is that instead of a static scene and a single person, our scene includes multiple people and it is generated in a dynamic way during training.

Let ϕ be a modified Signed Distance Field (SDF) for the scene that is defined as follows:

$$\phi(x, y, z) = -\min\left(\text{SDF}(x, y, z), 0\right), \tag{1}$$

According to the above definition, inside each human, ϕ takes positive values, proportional to the distance from the surface, while it is simply 0 outside of the human. Typically, ϕ is defined on a voxel grid of dimensions $N_p \times N_p \times N_p$. The naïve generation of a single voxelized representation for the whole scene is definitely possible. However, we often require a very fine voxel grid, which depending on the extend of the scene, might make processing intractable in terms of memory and computation. One critical observation here is that we can compute a separate ϕ_i function for each person in the scene, by calculating a tight box around the person and voxelizing it. This allows us to ignore empty scene space that is not covered by any person and we can instead use a fine spatial resolution to get a detailed voxelization of the body. Using this formulation, the collision penalty of person j for colliding with person i is defined as:

$$\mathcal{P}_{ij} = \sum_{v \in M_j} \tilde{\phi}_i(v), \tag{2}$$

where $\tilde{\phi}_i(v)$ samples the ϕ_i value for each 3D vertex v in a differentiable way from the 3D grid using trilinear interpolation (Figure 3). The ϕ_i computation for person i is performed by a custom GPU implementation. This computation does not have to be differentiable; ϕ_i only defines a distance field from which we sample values in a differentiable way. By definition, $\mathcal{P}_{ij}$ is non-negative. It takes value 0 if there is no collision between person i and j and increases as the distance of the surface vertices for person j move farther from the surface of person i. In theory, $\mathcal{P}_{ij}$ can be used by itself as an optimization objective for interpenetration avoidance. However, in practice, we observed that it results in very large gradients for the person translation, leading to training instabilities when there are heavy collisions. Instead of the typical term, we use a robust version of this objective. More specifically, our final interpenetration loss for a scene with N people is defined as follows:

$$L_{\mathcal{P}} = \sum_{j=1}^{N} \rho\left(\sum_{i=1, i\neq j}^{N} \mathcal{P}_{ij}\right) \tag{3}$$

where ρ is the Geman-McClure robust error function [12]. To avoid penalizing intersections between boxes corre-

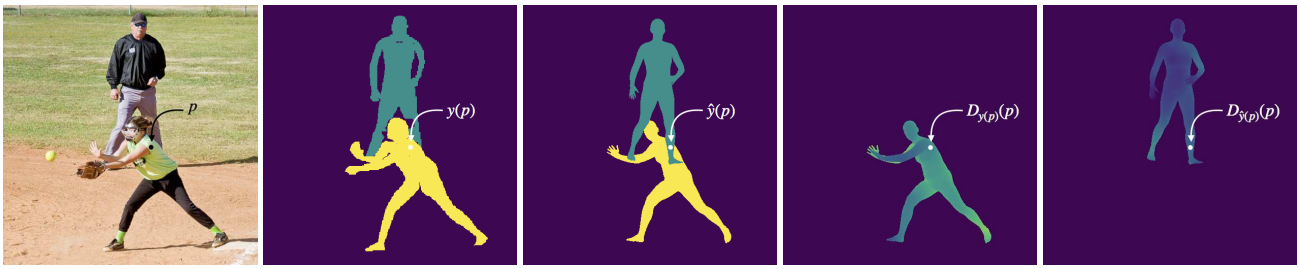

Figure 4: **Illustration of depth ordering-aware loss.** For an RGB image (first image), we consider the annotated instance segmentation (second image), and the instances based on the rendering of the estimated meshes on the image plane (third image). In case that there is a disagreement between the person index, e.g., for pixel p, where $y(p) \neq \hat{y}(p)$, we penalize the corresponding depth estimates at this pixel with an ordinal depth loss. The pixel depths $D_{y(p)}(p)$ and $D_{\hat{y}(p)}(p)$ are estimated by rendering the depth map independently for each person mesh (fourth and fifth image). This allows gradients to be backpropagated even to the non-visible vertices.

sponding to the same person, we use only the most confidence box proposal assigned to a ground truth box.

3.4. Depth ordering-aware loss

Besides interpenetration, another common problem in multi-person 3D reconstruction is that people are often estimated in incorrect depth order. This problem is more evident in cases where people overlap on the 2D image plane. Although it is obvious to the human eye which person is closer (due to the occlusion), the network predictions can still be incoherent. Fixing this depth ordering problem would be easy if we had access to pixel-level depth annotations. However, this type of annotations is rarely available. Our key idea here is that we can leverage the instance segmentation annotations that are often available, e.g., in the large scale COCO dataset [32]. Rendering the meshes of all the reconstructed people on the image plane can indicate the person corresponding to each pixel and optimize based on the agreement with the annotated instance annotation.

Although this idea sounds straightforward, its realization is more complicated. An obvious implementation would be to use a differentiable renderer, e.g., the Neural Mesh Renderer (NMR) [27], and penalize inconsistencies between the actual instance segmentation and the one produced by rendering the meshes to the image. The practical problem with [27] is that it backpropagates errors only to visible mesh vertices; if there is a depth ordering error, it will not promote the invisible vertices to move closer to the camera. In practice, we observed that this tends to move most people farther away, collapsing our training. Liu *et al*. [33] attempt to address this problem, but we observed that their softmax operation across the depths can result in vanishing gradients, while we also faced numerical instabilities.

Instead of rendering only the semantic segmentation of the scene, we also render the depth image D_i for each person independently using NMR [27]. Assuming the scene has N people, we assign a unique index $i \in \{1, 2, \ldots, N\}$ to each one of them. Let $y(p)$ be the person index at pixel location p in the ground truth segmentation, and $\hat{y}(p)$ be the predicted person index based on the rendering of the 3D meshes. We use 0 to indicate background pixels. If for a pixel p the two estimates indicate a person (no background) and disagree, i.e., $y(p) \neq \hat{y}(p)$, then we apply a loss to the depth values of both people for this pixel, $y(p)$ and $\hat{y}(p)$, to promote the correct depth ordering. The loss we apply is an ordinal depth loss, similar in spirit to [8]. More specifically, the complete loss expression is:

$$L_{\mathcal{D}} = \sum_{p \in \mathcal{S}} \log\left(1 + \exp\left(D_{y(p)}(p) - D_{\hat{y}(p)}(p)\right)\right) \quad (4)$$

where $\mathcal{S} = \{p \in I : y(p) > 0, \hat{y}(p) > 0, y(p) \neq \hat{y}(p)\}$ represents the set of pixels for image I where we have depth ordering mistakes (Figure 4). The key detail here is that the loss is backpropagated to the mesh (and eventually the model parameters) of both people, instead of backpropagating gradients only to the visible person, as a conventional differentiable renderer would do. This promotes a more symmetric nature to the loss (and the updates), and eventually makes this loss practical.

3.5. Implementation details

Our implementation is done using PyTorch and the publicly available mmdetection library [7]. We resize all input images to 512x832, keeping the same aspect ratio as in the original COCO training. For the baseline model we train only with the losses specified in Subsection 3.2, while for our full model we include in our training the losses proposed in Subsections 3.3 and 3.4. Our training uses 2 1080Ti GPUs and a batch size of 4 images per GPU.

For the SDF computation, we reimplemented [54, 55] in CUDA. Voxelizing a single mesh in a $32 \times 32 \times 32$ voxel grid requires about 45ms on an 1080Ti GPU. For efficiency, we perform 3D bounding box checks to detect overlapping 3D bounding boxes, and voxelize only the relevant meshes. Additionally, we reimplemented parts of NMR [27] to make rendering large images more efficient. This allowed us to have more than an order of magnitude of speedup since the forward pass complexity dropped from $O(Fwh)$ to $O(F +$

wh) on average, where F is the number of faces and w and h the image width and height respectively.

4. Experiments

In this Section, we present the empirical evaluation of our approach. First, we describe the datasets used for training and evaluation (Subsection 4.1). Then, we focus on the quantitative evaluation (Subsections 4.2 and 4.3), and finally we present more qualitative results (Subsection 4.4).

4.1. Datasets

Human3.6M [21]: It is an indoor dataset where a single person is visible in each frame. It provides 3D ground truth for training and evaluation. We use Protocol 2 of [26], where Subjects S1,S5,S6,S7 and S8 are used for training, while Subjects S9 and S11 are used for evaluation.
MuPoTS-3D [38]: It is a multi-person dataset providing 3D ground truth for all the people in the scene. We use this dataset for evaluation using the same protocol as [38].
Panoptic [24]: It is a dataset with multiple people captured in the Panoptic studio. We use this dataset for evaluation, following the protocol of [65].
MPI-INF-3DHP [36]: It is a single person dataset with 3D pose ground truth. We use subjects S1 to S8 for training.
PoseTrack [1]: In-the-wild dataset with 2D pose annotations. Includes multiple frames for each sequence. We use this dataset for training and evaluation.
LSP [22], **LSP Extended** [23], **MPII** [2]: In-the-wild datasets with annotations for 2D joints. We use the training sets of these datasets for training.
COCO [32]: In-the-wild dataset with 2D pose and instance segmentation annotations. We use the 2D joints for training as we do with the other in the-wild datasets, while the instance segmentation masks are employed for the computation of the depth ordering-aware loss.

4.2. Comparison with the state-of-the-art

For the comparison with the state-of-the-art, as a sanity check, we first evaluate the performance of our approach on a typical single person baseline. Our goal is always multi-person 3D pose and shape, but we expect our approach to achieve competitive results, even in easier settings, i.e., when only one person is in the image. More specifically, we evaluate the performance of our network on the popular Human3.6M dataset [21]. The most relevant approach here is HMR by Kanazawa *et al*. [26], since we share similar architectural choices (iterative regressor, regression target), training practices (adversarial prior) and training data. The results are presented in Table 1. Our approach outperforms HMR, as well as the approach of Arnab *et al*. [3], that uses the same network with HMR, but is trained with more data.

Having established that our approach is competitive in the single person setting, we continue the evaluation with multi-person baselines. In this case, we consider approaches that also estimate pose and shape for multiple people. The most relevant baselines are the works of Zanfir *et al*. [65, 66]. We compare with these approaches in the Panoptic dataset [24, 25], using their evaluation protocol (assuming no data from the Panoptic studio are used for training). The full results are reported in Table 2. Our initial network (baseline), trained without our proposed losses, achieves performance comparable with the results reported by the previous works of Zanfir *et al*. More importantly though, adding the two proposed losses (full), improves performance across all subsequences and overall, while we also outperform the previous baselines. These results demonstrate both the strong performance of our approach in the multi-person setting, as well as the benefit we get from the losses we propose in this work.

Another popular benchmark for multi-person 3D pose estimation is the MuPoTS-3D dataset [36]. Since no multi-person 3D pose and shape approach reports results on this benchmark, we implement two strong top-down baselines, based on state-of-the-art approaches for single-person 3D pose and shape. Specifically, we select a regression approach, HMR [26], and an optimization approach, SMPLify-X [42], and we apply them on detections provided by OpenPose [5] (as is suggested by their public repositories), or by Mask-RCNN [19] (for the case of HMR). The full results are reported in Table 3. As we can see, our baseline model performs comparably to the other approaches, while our full model trained with the proposed losses improves significantly over the baseline. Similarly with the previous results, this experiment further justifies the use of our coherency losses. Besides this, we also demonstrate that naïve baselines trained with a single person in mind are suboptimal for the multi-person setting of 3D pose. This is different from the 2D case, where a single-person network can perform particularly well in multi-person top-

Method	HMR [26]	Arnab *et al*. [3]	Ours
Reconst. Error	56.8	54.3	**52.7**

Table 1: **Results on Human3.6M**. The numbers are mean 3D joint errors in mm after Procrustes alignment (Protocol 2). The results of all approaches are obtained from the original papers.

Method	Haggling	Mafia	Ultim.	Pizza	Mean
Zanfir *et al*. [65]	140.0	165.9	150.7	**156.0**	153.4
Zanfir *et al*. [66]	141.4	152.3	**145.0**	162.5	150.3
Ours (baseline)	141.2	140.3	160.7	156.8	149.8
Ours (full)	**129.6**	**133.5**	153.0	156.7	**143.2**

Table 2: **Results on the Panoptic dataset**. The numbers are mean per joint position errors after centering the root joint. The results of all approaches are obtained from the original papers.

Method	All	Matched
OpenPose + SMPLify-X [42]	62.84	68.04
OpenPose + HMR [26]	66.09	70.90
Mask-RCNN + HMR [26]	65.57	68.57
Ours (baseline)	66.95	68.96
Ours (full)	**69.12**	**72.22**

Table 3: **Results on MuPoTS-3D**. The numbers are 3DPCK. We report the overall accuracy (All), and the accuracy only for person annotations matched to a prediction (Matched).

Method	MuPoTS-3D	PoseTrack
Our baseline	114	653
Our baseline + $L_{\mathcal{P}}$	34	202

Table 4: **Ablative for interpenetration loss**. The results indicate the number of collisions on MuPoTS-3D and PoseTrack.

down pipelines as well, e.g., [9, 56, 64]. For the 3D case though, when multiple people are involved, making the network aware of occlusions and interpenetrations during training, can actually be beneficial at test-time too.

4.3. Ablative studies

For this work, our interest in multi-person 3D pose estimation extends beyond just estimating poses that are accurate under the typical 3D pose metrics. Our goal is also to recover a coherent reconstruction of the scene. This is important, because in many cases we can improve the 3D pose metrics, e.g., get a better 3D pose for each detected person, but return incoherent results holistically. For example, the depth ordering of the people might be incorrect, or the reconstructed meshes might be positioned such that they overlap each other. To demonstrate how our proposed losses improve the network predictions under these coherency metrics even if they are only applied during training, we perform two ablative studies for more detailed evaluation.

First, we expect our interpenetration loss to naturally eliminate most of the overlapping people in our predictions. We evaluate this on MuPoTS-3D and PoseTrack, reporting the number of collisions with and without the interpenetration loss. The results are reported in Table 4. As we expected, we observe significant decrease in the number of collisions when we train the network with the $L_{\mathcal{P}}$ loss.

Moreover, our depth ordering-aware loss should improve the translation estimates for the people in the scene. Since for monocular methods it is not meaningful to evaluate metric translation estimates, we propose to evaluate only the returned depth ordering. More specifically, we consider all pairs of people in the scene, and we evaluate whether our method predicted the ordinal depth relation for this pair correctly. In the end, we report the percentage of correctly estimated ordinal relations in Table 5. As expected, the depth ordering-aware loss improves upon our baseline. In the same Table, we also report the results of the approach of Moon *et al*. [40] which is the state-of-the-art for 3D skeleton regression. Although [40] is skeleton-based and thus, not directly comparable to us, we want to highlight that even a state-of-the-art approach (under 3D pose metric evaluation) can still suffer from incoherency in the results. This provides evidence that we often might overlook the coherency of the holistic reconstruction, and we should also consider this aspect when we evaluate the quality of our results.

Method	Moon *et al*. [40]	Our baseline	Our baseline + $L_{\mathcal{D}}$
Accuracy	90.85%	92.17%	93.68%

Table 5: **Ablative for depth-ordering-aware loss.** Depth ordering results on MuPoTS-3D. We consider all pairs of people in the image, and we evaluate whether the approaches recovered the ordinal depth relation between the two people correctly. The numbers are percentages of correctly estimated ordinal depth relations.

Finally, we underline that we do not apply these coherency losses at test time. Instead, during training, our losses act as constraints to the reconstruction and ultimately provide better supervision to the network, for images that no explicit 3D annotations are available. The improved supervision leads to more coherent results *at test time too*.

4.4. Qualitative evaluation

In this Subsection, we present more qualitative results of our approach. In Figure 5 we compare our baseline with our full model trained with the proposed losses. As expected, our full model generates more coherent reconstructions, improving over the baseline as far as interpenetration and depth ordering mistakes are concerned. Errors can happen when there is significant scale difference among the people and there is no overlap on the image plane (last row of Figure 6). More results can be found in the Sup.Mat.

5. Summary

In this work, we present an end-to-end approach for multi-person 3D pose and shape estimation from a single image. Using the R-CNN framework, we design a top-down approach that regresses the SMPL model parameters for each detected person in the image. Our main contribution lies on assessing the problem from a more holistic view and aiming on estimating a coherent reconstruction of the scene instead of focusing only on independent pose estimation for each person. To this end, we incorporate two novel losses in our framework that train the network such that a) it avoids generating overlapping humans and b) it is encouraged to position the people in a consistent depth ordering. We evaluate our approach in various benchmarks, demonstrating very competitive performance in the traditional 3D

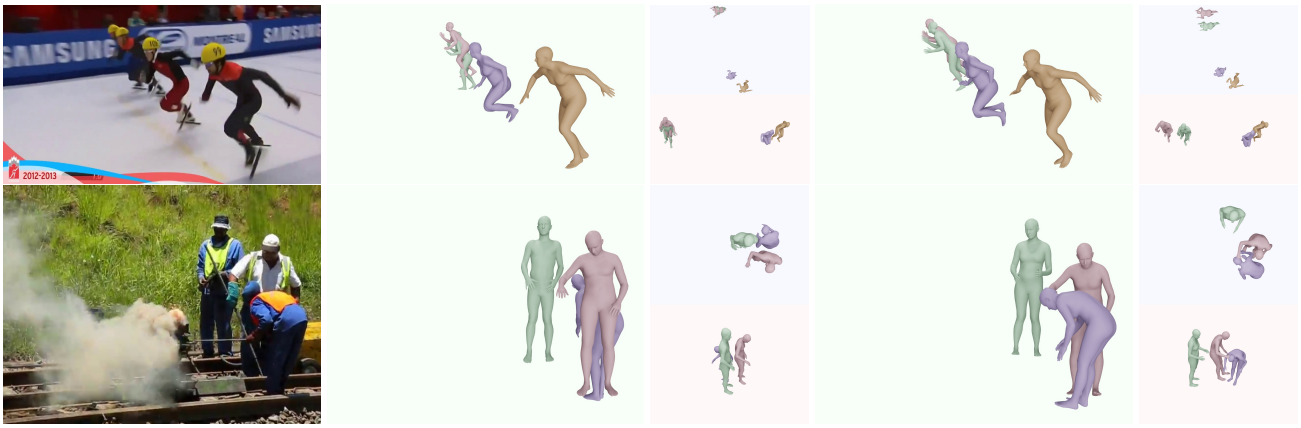

Figure 5: **Qualitative effect of proposed losses.** Results of our baseline model (center) and our full model trained with our proposed losses (right). As expected, we improve over our baseline in terms of coherency in the results (i.e., fewer interpenetrations, more consistent depth ordering for the reconstructed meshes).

Figure 6: **Qualitative evaluation.** We visualize the reconstructions of our approach from different viewpoints; front (green background), top (blue background) and side (red background). More qualitative results can be found in the Sup.Mat.

pose metrics, while also performing significantly better both qualitatively and quantitatively in terms of coherency of the reconstructed scene. In future work, we aim to more explicitly model interactions between people (besides the overlap avoidance), so that we can achieve a more accurate and detailed reconstruction of the scene at a finer level as well. In a similar vein, we can incorporate further information towards a holistic reconstruction of scenes. This can include constraints from the ground plane [65], background [17], or the objects that humans interact with [18, 58].

Acknowledgements: NK, GP and KD gratefully appreciate support through the following grants: NSF-IIP-1439681 (I/UCRC), NSF-IIS-1703319, NSF MRI 1626008, ARL RCTA W911NF-10-2-0016, ONR N00014-17-1-2093, ARL DCIST CRA W911NF-17-2-0181, the DARPA-SRC C-BRIC, by Honda Research Institute and a Google Daydream Research Award. XZ and WJ would like to acknowledge support from NSFC (No. 61806176) and Fundamental Research Funds for the Central Universities (2019QNA5022).

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 6

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 6

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *CVPR*, 2019. 3, 6

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 2, 3

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. *PAMI*, 2019. 6

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017. 3

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5

[8] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In *NIPS*, 2016. 3, 5

[9] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 1, 7

[10] Rishabh Dabral, Nitesh B Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. Multi-person 3D human pose estimation from monocular images. In *3DV*, 2019. 3

[11] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3D human shape estimation from single images. In *ICCV*, 2019. 1, 2

[12] Stuart Geman and Donald E McClure. Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute*, 4:5–21, 1987. 4

[13] Ross Girshick. Fast R-CNN. In *ICCV*, 2015. 3

[14] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *ICCV*, 2009. 2

[15] Rıza Alp Güler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *CVPR*, 2019. 1, 3

[16] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *CVPR*, 2018. 3

[17] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *ICCV*, 2019. 3, 4, 8

[18] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 3, 8

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 2, 3, 4, 6

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[21] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 36(7):1325–1339, 2013. 6

[22] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 6

[23] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011. 6

[24] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 6

[25] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *PAMI*, 41(1):190–204, 2017. 6

[26] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 3, 4, 6, 7

[27] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3D mesh renderer. In *CVPR*, 2018. 5

[28] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 1

[29] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 3

[30] Abhijit Kundu, Yin Li, and James M Rehg. 3D-RCNN: Instance-level 3D object reconstruction via render-and-compare. In *CVPR*, 2018. 3

[31] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 2

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 5, 6

[33] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3D reasoning. In *ICCV*, 2019. 5

[34] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015. 2, 3

[35] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 1, 2

[36] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*, 2017. 6

[37] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. XNect: Real-time multi-person 3D human pose estimation with a single RGB camera. *arXiv preprint arXiv:1907.00837*, 2019. 3

[38] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *3DV*, 2018. 3, 6

[39] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017. 2

[40] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In *ICCV*, 2019. 3, 7

[41] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018. 3

[42] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 1, 2, 3, 6, 7

[43] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 1

[44] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *CVPR*, 2018. 2, 3

[45] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 2

[46] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. DeepCut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016. 2

[47] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Deep multitask architecture for integrated 2D and 3D human sensing. In *CVPR*, 2017. 2

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2, 3, 4

[49] Helge Rhodin, Victor Constantin, Isinsu Katircioglu, Mathieu Salzmann, and Pascal Fua. Neural scene decomposition for multi-person motion capture. In *CVPR*, 2019. 3

[50] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net: Localization-classification-regression for human pose. In *CVPR*, 2017. 3

[51] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D pose detection in natural images. *PAMI*, 2019. 3

[52] Leonid Sigal, Alexandru Balan, and Michael J Black. Combined discriminative and generative articulated pose and non-rigid shape estimation. In *NIPS*, 2008. 2

[53] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. FACSIMILE: Fast and accurate scans from an image in less than a second. In *ICCV*, 2019. 2

[54] David Stutz. *Learning shape completion from bounding boxes with CAD shape priors*. PhD thesis, Masters thesis, RWTH Aachen University, 2017. 5

[55] David Stutz and Andreas Geiger. Learning 3D shape completion from laser scan data with weak supervision. In *CVPR*, 2018. 5

[56] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 1, 7

[57] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 1, 2

[58] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In *CVPR*, 2019. 8

[59] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2D and 3D image cues for monocular body pose estimation. In *ICCV*, 2017. 2

[60] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *CVPR*, 2017. 2

[61] Hsiao-Yu Tung, Hsiao-Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, 2017. 3

[62] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 1, 2

[63] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 1

[64] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 7

[65] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, 2018. 3, 6, 8

[66] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3D sensing of multiple people in natural images. In *NeurIPS*, 2018. 3, 6

[67] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3D human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017. 2