Scene-Aware 3D Multi-Human Motion Capture from a Single Camera

D. C. Luvizon[†] [ID] and M. Habermann [ID] and V. Golyanik [ID] and A. Kortylewski [ID] and C. Theobalt [ID]

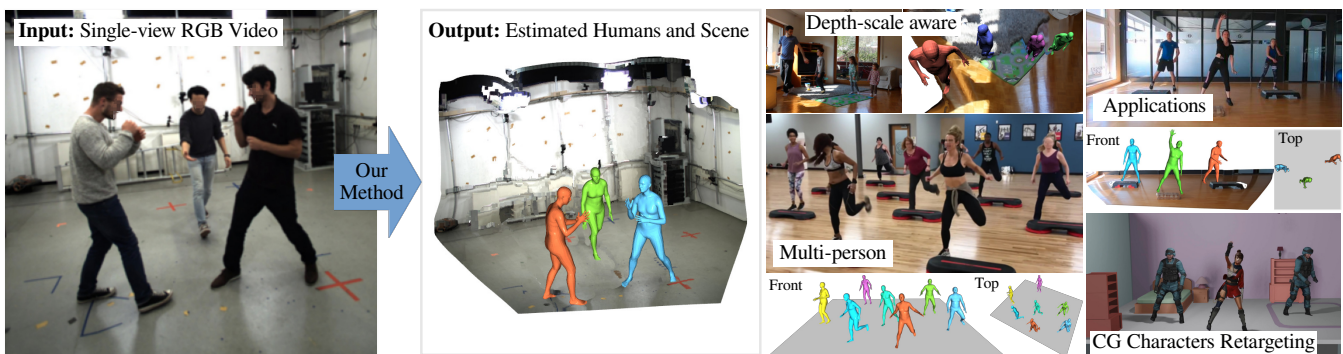MPI Informatics & Saarland Informatics Campus, Germany

Figure 1: *Our approach estimates absolute 3D positions of multiple humans in a scene, body shape and articulation in a globally and temporally coherent manner from a single monocular RGB video. It achieves higher 3D reconstruction accuracy than competing methods, allows motion re-targeting in the 3D space, and works exceptionally well even for in-the-wild videos.*

Abstract

In this work, we consider the problem of estimating the 3D position of multiple humans in a scene as well as their body shape and articulation from a single RGB video recorded with a static camera. In contrast to expensive marker-based or multi-view systems, our lightweight setup is ideal for private users as it enables an affordable 3D motion capture that is easy to install and does not require expert knowledge. To deal with this challenging setting, we leverage recent advances in computer vision using large-scale pre-trained models for a variety of modalities, including 2D body joints, joint angles, normalized disparity maps, and human segmentation masks. Thus, we introduce the first non-linear optimization-based approach that jointly solves for the 3D position of each human, their articulated pose, their individual shapes as well as the scale of the scene. In particular, we estimate the scene depth and person scale from normalized disparity predictions using the 2D body joints and joint angles. Given the per-frame scene depth, we reconstruct a point-cloud of the static scene in 3D space. Finally, given the per-frame 3D estimates of the humans and scene point-cloud, we perform a space-time coherent optimization over the video to ensure temporal, spatial and physical plausibility. We evaluate our method on established multi-person 3D human pose benchmarks where we consistently outperform previous methods and we qualitatively demonstrate that our method is robust to in-the-wild conditions including challenging scenes with people of different sizes. Code: `https://github.com/dluvizon/scene-aware-3d-multi-human`

CCS Concepts

• *Computing methodologies → Motion capture; Scene understanding;*

1. Introduction

Estimating the absolute 3D position, body shape, and articulation of multiple people in a scene is a fundamental research problem that has many applications in game development, VR/AR, and HCI.

Years of research went into developing sophisticated and expensive setups such as multi-view systems, motion capture suits, and manually or semi-automatically denoising of the tracked motions to then, for example, animate CG characters with these captured motions. However, one ideally would like to obtain such an absolute scene understanding from a capture setup that is easy to install, affordable, and that does not require expert knowledge, i.e. a *single RGB camera*. Such a lightweight setup would enable 3D motion capture for private users, e.g. avatar control via the smartphone, but it can also be applied for post production in the movie industry where, for example, one person should be replaced by another in a 3D consistent manner. At the same time, it has to be stated that performing motion capture given such limited data is exceptionally more difficult compared to multi-view systems. The major challenges for such a monocular setting, where only a single static video of the entire scene with moving persons is given, are the inherent depth ambiguity and occlusions, among many others.

Therefore, recent monocular approaches focus on a single human [MRC*17; PZDD17] or even assume an actor template is given [XCZ*18; HXZ*19; HXZ*20]. Recently, some works started to research the multi-person setting, but they either only learn a relative depth ordering of people in the scene [JKP*20] that is not 3D consistent over time or they directly predict absolute depth, which is prone to overfit to the settings shown in the training data [MCL19]. Most of those works leverage recent advances in Computer Vision and take as input several types of regressed data modalities obtained from models trained on large-scale data. This involves 1) 2D body joints [CHS*21; FXTL17], 2) joint angles [SBL*21], 3) normalized disparity maps [RBK21; LDC*19], and 4) human segmentation masks [CMS*21]. Interestingly, none of those works jointly considers *all* of those modalities.

To this end, this work investigates how each of those modalities can benefit the task of multi-person absolute 3D pose and shape estimation. A particular challenge, however, is that each individual modality has, of course, advantages, but also disadvantages. While 2D and 3D keypoint detections can help to infer the local 3D pose of a single person, they cannot ensure 3D consistency across humans and the scene. Joint angle estimates can be directly used to drive CG characters, but they are usually less accurate than the 3D keypoint detectors due to error accumulation along the kinematic chain. Normalized disparity maps provide global reasoning of the entire scene as well as the humans in terms of its scale-normalized depth, but they cannot provide absolute depth and scale of the scene. Finally, human segmentation masks can provide close to pixel-perfect and identity preserving segmentations of humans in the scene, but they lack a 3D understanding.

Now, to unite all the advantages of each of the modalities while compensating for their potential limitations, we propose the first optimization-based approach that jointly recovers the 3D position of all humans in the images, their articulated pose, their individual shapes, as well as a global scale of the scene relative the size of the humans from a single video recorded with a static camera; see Fig. 1. In particular, we propose a novel energy formulation, which infers the scene depth and a consistent person scale from scale-normalized disparity predictions by using the 2D and joint angle estimates of the humans in the scene as a prior. Once the

per-frame consistent depth is known, we reconstruct a dense point cloud of the static scene in 3D space by segmenting out the humans using the predicted segmentations and aggregating per-frame depth over time. Finally, we perform a coherent space-time optimization over the entire sequence to ensure temporal and spatial consistency as well as physical plausibility leveraging the aggregated scene estimate and the joint angle predictions. Note that in each of those steps, the *combination* of different data modalities is leveraged through our method and only this specific approach achieves the desired result in the considered setting, as extensively shown in our results. In summary, our primary technical contributions are as follows.

- The first monocular approach for multi-person 3D pose and scale estimation that jointly estimates multiple human poses and the 3D scene by combining data modalities in a novel optimization framework.
- A human body prior to disambiguate the scale of the scene, which allows us to perform a coherent space-time reasoning of the human motion in 3D and relative to a consistent scale.
- We show that the estimated 3D human bodies can be refined in 3D space and time by filtering body movements in 3D coordinates and by penalizing implausible poses w.r.t. the estimated scene, resulting in a more coherent final prediction.

Since our approach estimates joint angles, global positions and scale, the recovered 3D human poses can be directly applied to CG characters enabling exciting applications as shown in Section 4. Moreover, we demonstrate that the joint reasoning of the human body shape, pose, and the dense scene over the entire video sequence improves state of the art in terms of 3D localization, scene and person scale, as well as body pose compared to prior work, both, quantitatively and qualitatively. Finally, we show that several downstream applications can be directly derived from our method, such as monocular human motion capture and avatar control.

2. Related Work

3D human motion capture is an active research area, and many works have been proposed in the past [MSS*17; SSLW17; MHRL17; CR17; KAB20; ZHS*17; LPT22; TKS*16; SXW*18; WR19]. Since we target a monocular setting, we do not review multi-view- and depth-based methods. Instead, we review previous works that are most related to our method.

2.1. 3D Human Pose Estimation

2.1.1. Single Person Pose Estimation

Estimating the human body pose in 3D from a single image is a challenging problem that has been successfully handled by learning a human body prior from MoCap data [IPOS13]. To simplify the problem, previous methods usually predict 3D coordinates relative to the root joint, assuming a normalized human body size [MRC*17] and a fixed bounding box around the person in 3D space [MSS*17; PZDD17]. However, when multiple people are interacting with the environment, normalized and root-relative predictions are not enough to disambiguate the position and scale of individual persons in the scene. In addition, directly estimating

the 3D joint coordinates could result in implausible poses, which is a problem that can be mitigated by estimating joint angles instead [ZSZ*16].

Several works focus on estimating the full human mesh deformation from videos [XCZ*18; HXZ*19; HXZ*20], assuming that the actor mesh is provided in advance. Other works for single human estimation [KBJM18; KPBD19; PCG*19] rely on SMPL [LMR*15] as a proxy shape. Reconstructing shape proxies along with sparse 3D skeletons is desirable in many scenarios (*e.g.*, they can be used for body parts segmentation). Moreover, SMPL serves as a statistical prior on human body shapes and enables additional supervisory terms such as human silhouette overlays in 2D, which can result in higher accuracy [PCG*19].

2.1.2. Multiple Person Pose Estimation

Estimating positions of each person w.r.t. the others is crucial in multi-human pose estimation. Nonetheless, most of the existing multi-person methods are by design performing root-relative predictions [SBL*21; BCL*20; RWS17; RWS19]. Several techniques predict translations of each person in the camera reference frame. They either optimize the translation by projecting and fitting the estimated 3D poses into the image plane [MSM*20; ZZB*21a; DGM*19] or by directly regressing the distance of the root joint to the camera with a deep neural network [MCL19; LL20; WLL*20; ZFS*20]. The first case can be more robust to different camera setups, but is limited by the unknown height of each person in the scene. The second strategy is highly dependent on the training data and may not generalize to camera configurations not present in the training. Others explore human priors [LVC*21] to estimate a global trajectory [YIM*22], but still fail to recover the body size.

Recent methods performing human depth estimation are focused on penalizing depth ordering of multiple humans. For instance, Jiang *et al.* [JKP*20] uses instance segmentation masks to penalize depth inversion and Sun *et al.* [SLB*22] proposes to infer the depth of each person based on an imaginary bird's-eye-view representation and to estimate the person age as a proxy for the scale. Other approaches predict the relative depth among multiple persons by inferring some scene properties. A possible scene simplification is to assume a parametric planar floor, in such a way that each prediction can be positioned to respect a plausible human-floor contact [ZMS18; URA*21]. The common limitation of such approaches is the dependency on a simplified floor representation, which is often not the case in real applications. Contrarily, we estimate a scene point cloud that can represent an arbitrary ground floor.

The works from Jiang *et al.* [JKP*20] and Ugrinovic *et al.* [URA*21] are the most closely related to ours. Similarly to the former, we also render the estimated human models into the image plane to provide additional supervision in the depth dimension, and, related to the latter, we also disambiguate body size and depth for each person by constraining predictions with an estimated scene geometry. But differently from [JKP*20], that does not take the scene into account, and from [URA*21], that relies on a simplified scene representation and operates in a single frame, our method represents the scene as a frustum point cloud and performs optimization over the entire video sequence. In our work, we also rely on a human body proxy model [LMR*15] to estimate joint angles and we propose a new formulation to optimize the position of the humans and the scene in a joint optimization process. Therefore, our model improves the prediction of human positions by relying on an estimated proxy scene geometry that does not depend on a simplified parametric model.

2.2. Scene-aware Motion Capture

Predicting and understanding how humans interact in 3D has recently gained a lot of attention. Several current methods focus on positioning humans in a pre-scanned 3D scene [GMSP21; HYH*22; HCTB19] and on simultaneous estimation of human poses and objects humans interact with [WY21; CHY*19; YHT*22]. Different setups from ours assume an RGB-D sensor [ZZB*21b] or a moving camera [ZMZ*22; LYZ*21; HLL22; LBX*22] to estimating the scene geometry. Recent methods integrate physics-based constraints into monocular 3D human motion capture and mitigate foot-floor penetration and other severe artefacts [SGXT20; SGX*21]. Yu *et al.* [YPL21] also support composite scenes in the parcours and sports scenarios. Although there is a growing interest in investigating the interactions of humans and objects [DSJ*21; BXP*22], 3D motion capture of multiple humans with environmental awareness from a single monocular camera remains underexplored.

Determining the human scale in 3D is an ill-posed and challenging task. Bieler *et al.* [BGFR19] estimate the height of a single person from monocular videos by observing jumping people. Dabral *et al.* [DSJ*21] require an interaction with an object undergoing a free flight to resolve the absolute scene scale. Both methods assume motion influenced by the universal law of gravity near the surface of Earth, which allows them to relate the time spent in the air or the form of the observed trajectory with absolute distances in the metric units. The downside is that jumping humans or flying objects are restrictive assumptions. In contrast, we use a human body and 3D scene priors in 3D multi-human motion estimation and do not make strong assumptions about the observed human motions.

3. Method

The goal of our method is to estimate the absolute 3D position of each human in the scene, i.e., up to a unique and global scale that depends only on the size of the humans in the scene, their proxy shape and pose, as well as the scene scale solely from a monocular RGB video recorded with a static camera for which we know the intrinsics. To this end, we propose a unified approach that, for the first time, leverages *all* available data modalities, including 2D joint detections, regressed SMPL parameters, estimated disparity maps, and human segmentations. As illustrated in Figure 2, our method is divided into two stages. The first stage, *i.e. Image Modality Regression and Matching* (Section 3.1), extracts per-frame estimates and aggregates human-related predictions to individuals throughout the video sequence. The second stage, *i.e.* the proposed *Optimization Framework*, estimates the person and *per-frame* scene scale, the *global* 3D position of each person in the scene, as well as the refined articulated body pose in the form of joint angles per frame.

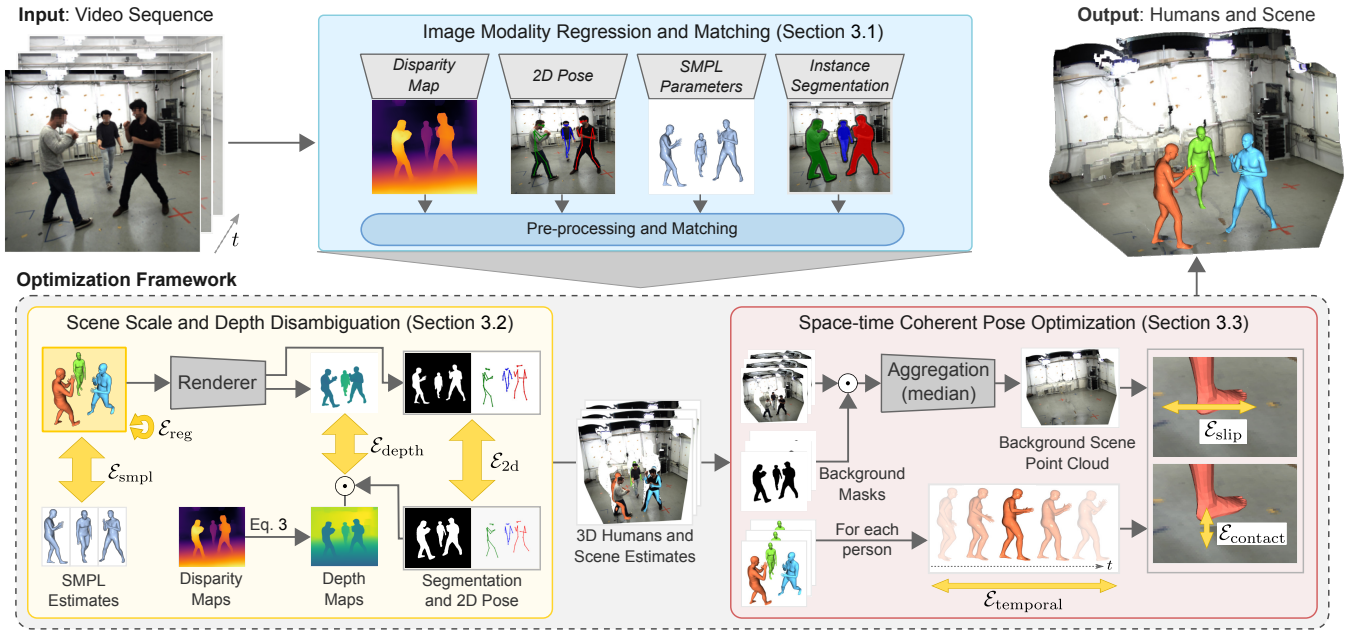The optimization framework is further subdivided into two parts.

Figure 2: *Overview of our method. For each frame in a monocular RGB video, we first estimate a normalized disparity map, 2D human poses, SMPL model parameters, and segmentation masks. These predictions are matched and tracked across frames to obtain per-person associations (blue box). The multi-modal estimates are then fed into our optimization framework. The first part of our optimization process estimates per-frame human models in global position and the scene geometry (yellow box). In the second part, the per-frame scene predictions are aggregated into a single point cloud representation and the human predictions are refined in a space-time coherent manner over the full video (red box). The yellow arrows indicate the energy terms minimized by our method. The output of our method is the 3D position of each human in the scene, their shape and pose as well as a consistent scene scale. $\odot$ is the Hadamard product.*

The *Scene Scale and Depth Disambiguation* part (Section 3.2) recovers a consistent and absolute 3D scene depth per frame, the human scales, and their absolute 3D position and body pose by jointly reasoning about multiple humans and the scene. The second part, referred to as *Space-time Coherent Pose Optimization* (Section 3.3), refines the pose and position of the estimated humans in a space-time coherent formulation, *i.e.* we enforce over the entire sequence the estimated poses to be temporally stable and physically plausible. For this, we leverage a rough scene geometry estimation, which is obtained by aggregating the absolute depth maps also estimated by our method. This final part significantly reduces artifacts, such as foot sliding, human-scene intersections, and jitter. Before we explain our method in more detail, we introduce relevant notations.

Notations. The input of our framework is a video sequence $\mathbf{I}_t$, with $t \in \{1, \ldots, T\}$, where T is the number of frames. We leverage the skinned multi-person linear model (SMPL) [LMR*15] to represent the humans in the scene. SMPL is a differentiable parametric human model that takes as input the pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{72}$, corresponding to the axis-angles of 24 body joints and the global body rotation, and PCA shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and produces a skinned human mesh

$$f_{\text{smpl}}(\boldsymbol{\theta}, \boldsymbol{\beta}) = \mathbf{V}, \tag{1}$$

where $\mathbf{V}$ are the posed and shaped vertices of the human body; for

more details we refer to their paper [LMR*15]. The mesh vertices regressed by SMPL can also be used to estimate a sparse 3D pose as $\mathcal{J}(\mathbf{V})$, where $\mathcal{J}(\cdot)$ is a linear regressor parameterized by a matrix $\mathbf{W} \in \mathbb{R}^{J \times 6890}$, and J denotes the total number of joints.

To account for translations in 3D space, we further add a translation $\boldsymbol{\Gamma}_{t,n} \in \mathbb{R}^3$ to the SMPL representation, where n is the person index. Furthermore, the 3D human pose models are overwhelmingly biased towards adult body sizes. Thus, we explicitly model the person scale by $s_n \in \mathbb{R}^+$ and our final human mesh can be defined as

$$\tilde{\mathbf{V}}_{t,n} = s_n \mathbf{V}_{t,n} + \boldsymbol{\Gamma}_{t,n}. \tag{2}$$

This human mesh for person n at time t is then fully determined by the parameters $\boldsymbol{\theta}_{t,n}$, $\boldsymbol{\Gamma}_{t,n}$, $\boldsymbol{\beta}_n$, and s_n, which we aim to recover in the following. Important to note is that the person scale s_n and shape $\boldsymbol{\beta}_n$ are unique for each person and consistent across the entire video sequence.

3.1. Input Modality Regression and Matching

To solve this underconstrained and challenging problem, our idea is to unite the strength of *all* data modalities, which recent state-of-the-art Computer Vision methods provide, in a single algorithm. More precisely, we leverage data-driven priors in the form of four

off-the-shelf methods for each frame of the input video sequence, as shown in Figure 2.

First, we obtain normalized disparity maps $\hat{\mathbf{d}}_t$ from the state-of-the-art DPT model [RBK21], which are then post-processed to enhance sharpness [WAT*20]. Note that these maps only encode *relative* and *normalized* depth and they are not consistent across frames, which becomes visible in the form of depth jitter.

Second, 2D pose tracking is obtained by AlphaPose [FXTL17], which coherently detects and tracks 2D joint positions $\hat{\mathbf{P}}_{t,n}^{2d} \in \mathbb{R}^{J \times 2}$ in image space and over time. Although this method is very robust due to training on large scale data, it falls short in predicting 3D.

Third, we predict the body shape $\boldsymbol{\beta}_{t,n}$ and joint angles $\hat{\boldsymbol{\theta}}_{t,n}$ for each person in each frame using ROMP [SBL*21]. Since ROMP predicts varying shapes for a single person across time, we average the predictions over the entire sequence to obtain a temporally consistent body shape. Thus, the vertices (Equation 2) are now only a function of the pose $\boldsymbol{\theta}_{t,n}$, translation $\boldsymbol{\Gamma}_{t,n}$, and scale s_n, which will be important in the next section. Moreover, to match the 2D AlphaPose and the SMPL detections, we leverage ROMPs projection model, compute the average Euclidean distance in image space, and pair detections with the lowest distance based on the Hungarian matching. It is worth mentioning that ROMP cannot account for out-of-distribution body sizes, e.g. small kids, neither it can predict the absolute 3D position of the humans with respect to the scene.

Fourth, we also leverage human segmentation masks, referred to as $\boldsymbol{\Omega}_{t,n} \in \mathbb{R}^{H \times W}$, which are obtained from Mask2Former [CMS*21]. Similarly, if we consider all the remaining pixels for frame t that do not belong to a person mask, we can also obtain a per-frame background segmentation mask $\mathbf{B}_t \in \mathbb{R}^{H \times W}$. To ensure that the 2D AlphaPose detections, the SMPL detections, and the foreground masks have a consistent person ID, we read the pixel values of the segmented masks at the 2D joint detections for each detected skeleton and apply a max-voting to retrieve the ID of the person.

In summary, the inputs to our algorithm now are:

- $\hat{\mathbf{d}}_t$: Normalized disparity maps
- $\hat{\mathbf{P}}_{t,n}^{2d}$: 2D joint predictions
- $\hat{\boldsymbol{\theta}}_{t,n}, \hat{\boldsymbol{\beta}}_n$: Pose angle and shape estimates
- $\boldsymbol{\Omega}_{t,n}, \mathbf{B}_t$: Human and background segmentations

Note that none of these predictions individually or by a trivial combination is discriminative enough to fully describe the entire scene, *i.e.* absolute 3D position, pose, and scale of the humans in the scene. Next, we demonstrate how our proposed method solves this problem.

3.2. Scene Scale and Depth Disambiguation

In the first part or our optimization process we focus on jointly obtaining the joint angles $\boldsymbol{\theta}_{t,n}$, shape parameters $\boldsymbol{\beta}_n$, global translation $\boldsymbol{\Gamma}_{t,n}$, and scale s_n of each person. Importantly, this step is performed jointly for the entire sequence, where the static camera is the global reference. However, estimating the height of a person *and* the scale given only a single RGB video is, by itself, an ill-posed problem as variations in scale can be compensated by a translation along the

depth and vice versa. As a result, infinitely many scale/translation combinations can lead to the same 2D image projections.

So far, we only considered individual humans without looking at the surrounding scene, although the scene itself can provide an important prior that helps to solve the above problem. Therefore, we leverage recent advances in monocular depth estimation [RBK21], which regress per-pixel normalized disparity maps $\hat{\mathbf{d}}_t$. It encodes the relative depth of each person in the scene, but obtaining the absolute depth values solely from $\hat{\mathbf{d}}_t$ is also an ill-posed problem, and further these predictions are not consistent across frames. The question remains, how the absolute scene depth or equivalently the human scales and translations can be recovered.

Our idea is to set the two entities, i.e., the scene and the humans, into a relation such that they constrain each other in an absolute 3D space. While the humans can already be represented in absolute space by means of their global translation $\boldsymbol{\Gamma}_{t,n}$ and scale s_n, we also require a per-frame conversion of temporally inconsistent normalized disparity maps to absolute depth maps, which can be defined as

$$\tilde{\mathbf{D}}_t = \frac{z_{\text{far},t} z_{\text{near},t}}{\hat{\mathbf{d}}_t (z_{\text{far},t} - z_{\text{near},t}) + z_{\text{near},t}} \tag{3}$$

where $z_{\text{near},t}$ and $z_{\text{far},t}$ are the near and far depth values, respectively. Intuitively, this operation shifts and scales the normalized disparity maps to convert them to absolute depth values. Importantly, these near and far values are optimized per-frame to compensate for the temporal inconsistencies in the disparity maps.

Once both humans and the scene can be represented in absolute 3D space, we now relate them to each other by jointly solving for $\boldsymbol{\kappa}_{t,n} \in \{z_{\text{near},t}, z_{\text{far},t}, \boldsymbol{\theta}_{t,n}, \boldsymbol{\beta}_n, \boldsymbol{\Gamma}_{t,n}, s_n\}$ by minimizing the energy

$$\underset{\forall t \in \{1,...,T\}, \forall n \in \{1,...,N\} : \boldsymbol{\kappa}_{t,n}}{\arg\min} E_{\text{I}}, \quad \text{with} \tag{4}$$

$$E_{\text{I}} = \mathcal{E}_{\text{depth}} + \mathcal{E}_{\text{2d}} + \mathcal{E}_{\text{smpl}} + \mathcal{E}_{\text{reg}}, \tag{5}$$

which is jointly optimized over the entire sequence. In particular, our energy is composed of a depth term $\mathcal{E}_{\text{depth}}$, a 2D image evidence term $\mathcal{E}_{\text{2d}}$, a joint angle and shape term $\mathcal{E}_{\text{smpl}}$, and additional regularization terms $\mathcal{E}_{\text{reg}}$. In the following, we explain each term in more detail.

3.2.1. Depth Consistency Energy

Most importantly, to ensure a coherent depth between the scene and all humans in the scene, we propose a depth consistency energy

$$\mathcal{E}_{\text{depth}} = \lambda_{\text{depth}} \sum_{t,n} \left(M(\Psi_d(\tilde{\mathbf{V}}_{t,n})) - M(\tilde{\mathbf{D}}_t) \right)^2, \tag{6}$$

$$M(\mathbf{D}) = \sum \frac{\boldsymbol{\Omega}_{t,n}}{|\boldsymbol{\Omega}_{t,n}|} \log(\mathbf{D}), \tag{7}$$

where $|\boldsymbol{\Omega}|$ denotes the number of foreground pixels, $M(\cdot)$ computes the average of the log-depth in the foreground, $\Psi_d(\cdot)$ is a differentiable rasterizer [RRN*20] that projects and converts a 3D mesh into a depth map in the image plane, and λ_{depth} is a hyperparameter. The vertices $\tilde{\mathbf{V}}_{t,n}$ refer to the estimated SMPL models of each person in global space, which are a function of the variables $\boldsymbol{\theta}_{t,n}, \boldsymbol{\beta}_n, \boldsymbol{\Gamma}_{t,n}$, and s_n (Equation 2). The rasterized human depths

are then compared to the estimated absolute depth map $\tilde{\mathbf{D}}_t$ of the scene (Equation 3), which are a function of the variables $z_{\text{near},t}$ and $z_{\text{far},t}$. Thus, this energy jointly optimizes the human *and* the scene parameters. However, since both sides of the penalty term contain free variables, this energy alone would not disambiguate the problem.

3.2.2. Image Projection Energy

We introduce an additional data term, which further constrains the human-related variables by enforcing the 3D bodies to project accurately into the image plane. More precisely, the data term

$$\mathcal{E}_{\text{2d}} = \mathcal{E}_{\text{joints}} + \mathcal{E}_{\text{silhouette}} \tag{8}$$

penalizes the error between the projected 3D body joints $\mathcal{J}(\tilde{\mathbf{V}}_{t,n})$ of the optimized SMPL models and the respective 2D body joints $\hat{\mathbf{P}}_{t,n}^{\text{2d}}$ regressed by AlphaPose with

$$\mathcal{E}_{\text{joints}} = \sum_{t,n} \left\| \Pi(\mathcal{J}(\tilde{\mathbf{V}}_{t,n})) - \hat{\mathbf{P}}_{t,n}^{\text{2d}} \right\|_2^2, \tag{9}$$

where $\Pi(\cdot)$ is the perspective camera projection operator. The right term of (8) penalizes the discrepancy between the SMPL silhouette and the instance segmentation masks:

$$\mathcal{E}_{\text{silhouette}} = \frac{\lambda_{silhouette}}{|\mathbf{\Omega}|} \sum_{t,n} \boldsymbol{\sigma}_{t,n} \left\| \Psi_s(\tilde{\mathbf{V}}_{t,n}) - \mathbf{\Omega}_{t,n} \right\|_2^2, \tag{10}$$

where $\Psi_s(\cdot)$ is a differentiable renderer [RRN*20] that projects and converts a 3D mesh into a silhouette image and $\boldsymbol{\sigma}_{t,n}$ is a visibility mask, so vertices hidden by other humans are not penalized.

3.2.3. Joint Angle and Shape Energy

Since (8) only constrains the parameters in 2D image space, we further add an additional data term that ensures that the optimized SMPL parameters are close the prediction of ROMP:

$$\mathcal{E}_{\text{smpl}} = \lambda_{\text{smpl}} \sum_{t,n} \left\| \boldsymbol{\theta}_{t,n} - \hat{\boldsymbol{\theta}}_{t,n} \right\|_1 + \left\| \boldsymbol{\beta}_n - \hat{\boldsymbol{\beta}}_n \right\|_1. \tag{11}$$

Here, $\|\cdot\|_1$ denotes the $L1$ norm.

3.2.4. Temporal and Human Priors

To further constrain the scale and position of a person, we leverage priors on the human body size and on the temporal information. This is achieved by our regularization term

$$\mathcal{E}_{\text{reg}} = \mathcal{E}_{\text{scale}} + \mathcal{E}_{\text{speed}}. \tag{12}$$

For the scale term $\mathcal{E}_{\text{scale}}$, our assumptions are two-fold: *i)* The scale of a person should not deviate too much from the standard person size, i.e., the standard SMPL size when $s_n = 1$, and *ii)* the average scale of multiple people in the scene should remain close to one. This dual assumption is enforced by

$$\mathcal{E}_{\text{scale}} = \lambda_{\text{scale}} \sum_n (s_n - 1)^2 + \left(\sum_n (s_n - 1) \right)^2, \tag{13}$$

where the first term accounts for the individual person scale and the second term accounts for the average scale of multiple persons.

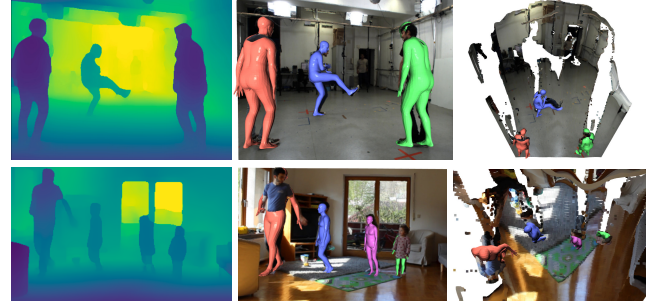In addition to the person scale, we also introduce an underlying

Figure 3: *Per-frame estimations of our method considering the first optimization part (Section 3.2) only. From left to right: Estimated depth map, frontal view of the scene and estimated humans, and top view. Note how the persons' absolute 3D location, articulated pose and shape as well as the scene scale can be recovered from a single input image, even with people of different sizes (bottom row).*

assumption that locomotion is rather smooth over time based on the physical limits of the human body, so we penalize large movements of the root joint by our energy

$$\mathcal{E}_{\text{speed}} = \lambda_{\text{speed}} \sum_{t,n} \left\| \boldsymbol{\Gamma}_{t,n} - \boldsymbol{\Gamma}_{t-1,n} \right\|_2^2. \tag{14}$$

In the optimization process described above, the *per frame* human parameters and the absolute scene depth are obtained by means of the optimized human $\forall t \in \{1,...,T\}, \forall n \in \{1,...,N\}$: $\boldsymbol{\theta}_{t,n}, \boldsymbol{\beta}_n, \boldsymbol{\Gamma}_{t,n}, s_n$, and scene $z_{\text{near},t}, z_{\text{far},t}$ parameters. Figure 3 shows our estimated scene and humans for example frames. Note that the estimated depth looks plausible, humans and the scene are coherent with each other, and the reprojection of humans into the input view looks accurate.

3.3. Space-time Coherent Pose Optimisation

Since we obtained absolute and per-frame human models and scene estimations, both information can be used together to further refine the human poses in a spatially and temporally coherent manner. Therefore, in the last part of our optimization method, we refine the estimated poses in 3D by enforcing physical plausibility between humans and the estimated scene, as well as by applying a temporal smoothness term. More precisely, we extend (4) by including a new energy term E_{II}:

$$\underset{\forall t \in \{1,...,T\}, \forall n \in \{1,...,N\}: \boldsymbol{\kappa}_{t,n}}{\arg\min} \quad E_{\text{I}} + E_{\text{II}}, \quad \text{with} \tag{15}$$

$$E_{\text{II}} = \mathcal{E}_{\text{contact}} + \mathcal{E}_{\text{slip}} + \mathcal{E}_{\text{temporal}}. \tag{16}$$

For implementing $\mathcal{E}_{\text{contact}}$ and $\mathcal{E}_{\text{slip}}$, we leverage the estimated scene geometry as a reference for enforcing foot contact and penalizing foot slipping. In the following, we first explain how the per-frame depth maps are aggregated into a static 3D scene representation, then we present the energy terms of E_{II} in more detail.

3.3.1. Scene Point Cloud Estimation

Our method relies on humans as anchors in the scene, *i.e.*, the estimated geometry around the humans tends to be coherent. However, mainly due to occlusions, the estimated per-frame absolute depth values are not yet temporally consistent for the whole scene. To obtain a static representation of the background, we rely on the segmentation masks to aggregate the depth values in the background from each frame into a single depth map. This static depth map representation is obtained by computing the per-pixel *median* for the entire video sequence, which is a metric robust to outlier depth values. We also experimented with more sophisticated aggregation strategies, such as aggregating values near the human anchors weighted by a Gaussian distribution—since the human positions are stable—but this strategy was significantly more expensive and resulted in marginal improvements. At the end of this aggregation process, we obtain a single depth map $\hat{\mathbf{D}}$ of the scene, which can be then converted to a point cloud representation $\mathbf{P} \in \mathbb{R}^{HW \times 3}$ in absolute 3D space.

3.3.2. Improving Physical Plausibility of Estimated Motions

Recently, a series of works highlighted the importance of physics awareness in monocular *single* person motion capture [SGXT20; SGX*21; RBH*21; LBX*22] with assumptions about the camera and floor plane positions. Inspired by them and the fact that we obtain a coherent and consistent scale estimation of the scene, we propose to model in our energy formulation the physical interaction between the humans and the environment. Here, the first term penalizes "floating" characters, *i.e.*, humans that are not in contact with the ground, and the second term penalizes foot sliding, *i.e.*, a foot that is in contact with the ground should not move.

More precisely, given the scene point cloud $\mathbf{P}$ and the estimated human meshes $\tilde{\mathbf{V}}_{t,n}$, floating characters are penalized by

$$\mathcal{E}_{\text{contact}} = \lambda_{\text{contact}} \sum_{t,n} \zeta \left(\left\| \min(\tilde{\mathbf{V}}_{t,n}^{y+} - \mathbf{P}) \right\|_1 \right) \qquad (17)$$

where $\tilde{\mathbf{V}}_{t,n}^{y+} \in \mathbb{R}^{1 \times 3}$ is the vertex of person n at time t with lower Y coordinate, considering that the Y-axis is the gravitational axis for our coordinate frame. In other words, the term $\mathcal{E}_{\text{contact}}$ minimizes the distance between the lower vertex $\tilde{\mathbf{V}}^{y+}$ of each prediction and its respective closest point in the scene point cloud. Here, $\zeta(\cdot)$ is a robust thresholding function, which only considers distances below $20cm$.

The term

$$\mathcal{E}_{\text{slip}} = \lambda_{\text{slip}} \sum_{t,n} \zeta \left\| \Delta(\tilde{V}_{t,n}^{y+}) \right\|_1 \qquad (18)$$

penalizes the movement of this lowest vertex in the time domain (Δ) when it is in contact with the scene. By applying those energy terms, we can now enforce that the humans interact more physically accurate with respect to the 3D scene.

3.3.3. Temporally Stable Pose

Furthermore, since the joint and absolute position optimized by E_{I} can still contain smaller jitter, we propose a temporal stability term

$$\mathcal{E}_{\text{temporal}} = \lambda_{\text{temporal}} \sum_{t,n} \left\| \Delta_t(\tilde{\mathbf{V}}_{t,n}) - \Delta_t(\bar{\mathbf{V}}_{t,n}) \right\|^2, \qquad (19)$$

based on the 1€ filter [CRV12], where $\Delta_t(\mathbf{V}_{t,n}) = \mathbf{V}_{t,n} - \mathbf{V}_{t-1,n}$ is the temporal variation of the human mesh vertices and $\bar{\mathbf{V}}_{t,n}$ are the estimated SMPL vertices after temporal filtering [CRV12]. This term allows us to obtain temporally more stable poses with significantly less jitter.

4. Experiments

In this section, we present an empirical evaluation of our method. We first briefly describe the datasets and metrics used in our experiments in Sections 4.1 and 4.2, followed by the implementation details in Section 4.3. Next, we compare our approach with the most related works to ours in Section 4.4. In Section 4.5, we perform a thorough ablation study of the main components of our method and show additional qualitative results in Section 4.6.

4.1. Datasets

MuPoTs-3D [MSM*18] is a test dataset composed of 20 video sequences with multiple people, including different types of cameras in indoor and outdoor environments. We followed the evaluation protocol from [MSM*18] in our experiments. This dataset is especially challenging due to the large amount of interactions between humans and the various types of scenes. Ground-truth 3D pose annotations are provided in absolute coordinates.

CMU Panoptic [JLT*15] is a dataset recorded in the Panoptic Studio with multiple people. As in preliminary work [JKP*20; ZMS18], we use this dataset for evaluation considering the sequences `haggling1`, `ultimatum1`, and `pizza1`, which are performed by several adults.

In addition to the previous datasets, we also evaluated our method quantitatively on Internet videos considering challenging cases with multiple people of different sizes, including adults and children.

4.2. Metrics

MRPE and AP. We quantitatively evaluate the prediction of the absolute 3D location of a human using the widely adopted mean root position error (MRPE), in millimeters, and the average precision of the human root joint (AP_{25}^{root}) [MCL19], considering the standard threshold of 25 cm.

3DPCK. The quality of the articulated 3D pose prediction is measured using root-relative 3DPCK [MRC*17], with the standard threshold of 15 cm. The 3DPCK metric enables measuring the correctness of the pose, independently of the prediction of the absolute 3D location of the human.

MPJPE. For a fair comparison with previous methods, we also report root-relative mean per-joint position error (MPJPE) in the CMU Panoptic dataset.

Jitter. Finally, since we are targeting high-quality temporal predictions in 3D coordinates, we also evaluate the amount of jitter of our estimations, which is a critical indicator for many downstream applications. For this evaluation, we adapted the *temporal smoothness error* e_{smooth} from [SGX*21] to evaluate the jitter in 3D coordinates.

Table 1: *Comparison of our method with previous approaches on MuPoTs-3D in the MRPE (lower is better), AP_{25}^{root}, and 3DPCK metrics (higher is better), considering the global 3D pose and the normalized (univ) ground truths. Our approach is superior to all compared methods on the absolute metrics (MRPE, AP_{25}^{root} and $3DPCK_{3d}$), i.e., the most expressive ones for 3D human motion capture. "†" evaluated on samples with IK only; "*" evaluated on root-relative predictions without IK; "‡" results only possible with an additional 2D fitting stage, implemented as our baseline.*

Method	Char. control	MRPE ↓	AP_{25}^{root}	$3DPCK_{3d}$	$3DPCK_{univ}$
LCR-Net [RWS17]	✗	–	–	–	53.8
LCR-Net++ [RWS19]	✗	–	–	–	70.6
3DMPPE [MCL19]	✗	–	31.0	–	**81.8**
SMAP [ZFS*20]	✗	–	45.5	–	80.3
XNect* [MSM*20]	✗	–	–	64.1	71.9
XNect† [MSM*20]	✔	639	31.6	56.5	60.1
CRMH [JKP*20]	✔	–	–	–	69.1
BEV [SLB*22]	✔	–	–	–	70.2
Baseline (ROMP+2D fitting)	✔	331‡	45.4‡	68.2‡	71.8
Ours	✔	**266**	**62.3**	**74.9**	<u>78.9</u>

4.3. Implementation Details

Our method is implemented in PyTorch [PGM*19] using Py-Torch3D [RRN*20] for the rasterization (6) and silhouette rendering (10). The camera intrinsics are used in the 3D joint projection (9), rasterization (6), and rendering (10) parts, and can be obtained from video metadata if not given. We apply the RM-Sprop [HSS12] optimizer with the parameters α and *momentum* set to 0.5 and 0.9, respectively, for all experiments. In the optimization process, we initially minimize the first part (4) only for 30 iterations, then perform the full optimization (15) for more 200 iterations. We use a learning rate initially set to 0.01 and exponentially decaying with factor 0.99. The weights $\lambda_{(.)}$ were empirically defined to balance the magnitude of the individual energy terms, and fixed in the method in all experiments, except when mentioned otherwise (ablation in Section 4.5). The values were defined as $\lambda_{depth} = \lambda_{speed} = 0.05$, $\lambda_{silhouette} = 0.1$, $\lambda_{smpl} = \lambda_{temporal} = 0.002$, $\lambda_{scale} = 0.0001$, $\lambda_{contact} = 0.001$, and $\lambda_{slip} = 0.01$. For numerical stability, we constrain the variables s_n, $z_{near,t}$, and $z_{far,t}$ to be non-zero and positive. Both human and background segmentation masks were post-processed with morphological erosion and dilation filters of size 3×3 and 5×5, respectively. For the sake of GPU memory efficiency, we use mini batches of ten images in the depth and silhouette losses. Our experiments run on a workstation with one Nvidia Titan V GPU with 12 GB of memory.

4.4. Comparison with Previous Methods

In Table 1, we compare our method to the most related prior work. We compare our method for human localization considering MRPE and AP_{25}^{root} metrics with the methods that are capable of providing such predictions. We use two protocols to evaluate the quality of the 3D pose. First, we compare against the global 3D pose without any normalization, which is a fairer protocol for our method, since we are capable of estimating the person scale (denoted by $3DPCK_{3d}$). In the second case, we compare against the *universal* 3D pose, which has all bone lengths normalized to a standard size,

Table 2: *Comparison of our method with previous approaches on the CMU Panoptic dataset for 3D pose estimation. Results reported in millimeters. Camera views capturing only the upper body parts were not used in our evaluation. † evaluated in all the sequences. Best results are **bold** on the standard sequences and <u>underlined</u> on the full-body visible sequences.*

Metric	Method	Haggling	Ultimatum	Pizza	Avg.
MPJPE	CRMH [JKP*20]†	129.6	153.0	156.7	146.4
	BEV [SLB*22]†	**90.7**	**113.1**	**125.2**	**109.6**
	Baseline	93.6	133.8	145.9	124.4
	Ours	<u>84.5</u>	<u>108.9</u>	<u>133.2</u>	<u>108.9</u>
MRPE	Baseline	235.2	269.6	356.4	287.0
	Ours	<u>213.7</u>	<u>208.0</u>	<u>229.7</u>	<u>217.1</u>

as described in [MSM*18] (denoted as $3DPCK_{univ}$). For this universal protocol, in our method, we assume person scale s_n equals to one for all predictions. Note how our method outperforms all prior work by a wide margin at 3D localization and also performs better at estimating the articulated pose compared to all other methods that allow for character control. As a baseline, we evaluate ROMP [SBL*21] predictions with an additional stage for fitting estimated SMPL models to AlphaPose 2D body joint detections, since this is the closest setup to our method without including our new energy formulation. For this, we assume a unitary person scale (w.r.t. the SMPL neutral model) and optimize only the global translation in 3D of each person. We also tried GLAMR [YIM*22] as our baseline, but it often failed in detecting and tracking humans and was not able to process long video sequences, resulting in a very low 3DPCK of 42.9% on MuPoTs-3D. In a similar manner, XNect [MSM*20] estimates the global position by fitting the predicted 3D poses into 2D body joints, assuming a universal and normalized human body size. The inverse kinematics (IK) stage from XNect allows this global estimation, however, since the optimized 3D human pose differs from the preliminary estimated pose, the accuracy after IK drops significantly. In summary, we observe that our approach outperforms previous methods for human position estimation by a significant margin, improving the average precision of the root joint from 45.4% to 62.3%. Our method also outperforms all other approaches for human pose estimation that are capable of driving a virtual character.

In Table 2, we compare our method with other approaches on the CMU Panoptic dataset. This dataset is specially challenging because in many sequences the persons are only partially visible, either due to occlusions, or because the camera is capturing only the upper body part of the actors. Even in this challenging scenario, our method performs on par with the recent BEV [SLB*22] method, which was trained on the CMU Panoptic dataset and, therefore, performs better in the cases of partial body visibility then our optimization approach. In order to evaluate the performance of our method on the more practical scenario of cameras recording the full body of the persons, we removed the few sequences capturing only the upper body parts. In this setup, we largely improve over other methods and over our baseline, as can be seen by the underlined numbers in Table 2.

For many downstream applications, such as gaming and charac-

Table 3: *Comparison of our method on MuPoTs-3D with previous approaches on temporal smoothness error e_{smooth}, that measures the amount of jitter in the predictions in millimeters. We also report the MRPE and 3DPCK$_{3d}$ metrics for completeness. Our method has a drastically lower jitter in the prediction compared to previous multi-person motion capture approaches.*

Method	Jitter ↓	MRPE ↓	3DPCK$_{3d}$ ↑
XNect [MSM*20]	136.4	639	56.5
ROMP [SBL*21]	59.6	331	68.2
Ours (E_I only)	17.5	281	73.5
Ours ($E_I + \mathcal{E}_{contact}$)	17.6	276	73.7
Ours ($E_I + \mathcal{E}_{contact} + \mathcal{E}_{slip}$)	17.1	273	73.8
Ours ($E_I + \mathcal{E}_{temporal}$)	7.8	272	74.8
Ours ($E_I + \mathcal{E}_{contact} + \mathcal{E}_{slip} + \mathcal{E}_{temporal}$)	**7.5**	**266**	**74.9**

ter control, jitter is a severe artifact that hinders usability. Therefore, we also evaluated our method by reporting the temporal smoothness error e_{smooth} in 3D coordinates. The results from our method, as well as from previous work in the literature related to ours, are shown in Table 3. In this experiment, we compared our approach with two methods from the literature, showing a significant improvement in reducing the jitter artifact. Furthermore, we also evaluated the contribution of different components of our method. For instance, the temporal energy term in our approach has a critical effect in reducing jitter. In addition, the contact and slip terms also contribute in a small proportion but consistently to all metrics, regardless the presence or absence of the temporal energy. When all terms are included, our approach is very stable, with an average jitter error below 1*cm*.

4.5. Ablation Study

In this section, we perform additional evaluations of the different components of our method. The results on MuPoTs-3D are shown in Tables 4 and 5. First, we evaluate the influence of the energy terms of the first part of our optimization framework. The energy term $\mathcal{E}_{depth}$ provides essential information to disambiguate depth and scale, which contributes to improving the position estimation. The flexibility provided by the person scale factor can be detrimental to the overall accuracy of the method if no constraints are imposed on it. This can be seen in the second row of Table 4, without $\mathcal{E}_{scale}$. By constraining our predictions to remain close to the original estimates from ROMP, our method enforces the final estimates to be valid and prevents them from collapsing, as shown in the results without $\mathcal{E}_{smpl}$. Finally, $\mathcal{E}_{speed}$ is relevant for reducing jitter and the silhouette term provides beneficial contributions to all the metrics. With all the energy terms, our method is stable and precise in estimating 3D position and pose. Since our method relies on off-the-shelf predictors as input, we also provide a concise evaluation considering two different 2D pose and three different depth estimation models from the recent literature. The results in Table 5 show that the influence of the depth estimation models is relatively small; however, the best performing model is the most recent transformer architecture, which suggests that our approach directly benefits from improved monocular depth estimations. Regarding 2D pose estimation, HRNet [WSC*20] performed worse

Table 4: *Ablation study for different energy terms. Without the proposed depth and scale terms, the global position in 3D cannot be precisely recovered, i.e., AP_{25}^{root} drops from 62.3 to 47.4% and to 22.2%, respectively. The SMPL term is critical for enforcing valid estimates, and the speed term contributes to reducing the jitter. The silhouette term provides consistent improvements in all the metrics.*

Experiment	Jitter ↓	MRPE ↓	AP_{25}^{root} ↑	3DPCK$_{3d}$ ↑
w/o $\mathcal{E}_{depth}$	7.8	284	47.4	**75.5**
w/o $\mathcal{E}_{scale}$	7.7	541	22.2	68.9
w/o $\mathcal{E}_{smpl}$	8.0	674	11.5	56.3
w/o $\mathcal{E}_{speed}$	8.9	<u>269</u>	**63.6**	74.8
w/o $\mathcal{E}_{silhouette}$	<u>7.6</u>	270	62.0	74.7
Ours (full)	**7.5**	**266**	<u>62.3</u>	<u>74.9</u>

Table 5: *Our results considering different models for 2D pose and monocular depth estimation. We observe that the human position estimation from our method benefits directly from advances in the monocular depth estimation when comparing MiDaS v2.1 [RLH*20] and DPT-Large [RBK21].*

2D Pose Model	Depth Model	MRPE ↓	AP_{25}^{root} ↑	3DPCK$_{3d}$ ↑
AlphaPose	MiDaS v2.1	278	55.8	**75.7**
AlphaPose	DPT-Hybrid	276	60.8	75.0
AlphaPose	DPT-Large	**266**	**62.3**	74.9
HRNet	DPT-Large	304	54.9	72.7

than AlphaPose, since HRNet relies on person detection as a first step, which makes it susceptible to detection failures.

4.6. Qualitative Results

Figure 4 provides additional qualitative results with predictions from our method in 3D coordinates, alongside the ground truth pose. We compare our method with XNect [MSM*20] and ROMP [SBL*21]. We can see that predictions from ROMP do often not correspond to the correct position of the humans in the scene, since it is not able to estimate the correct person scale. For XNect, we can observe that it also fails to recover the correct scale of the person, which can be observed from the top view. On the other hand, our approach can predict a 3D pose that corresponds to the ground truth human annotation and is coherently positioned in 3D coordinates. We also compare our method with GLAMR [YIM*22] and BEV [SLB*22] in Figure 6. GLAMR fails to track all the persons in the scene and BEV fails to predict coherent human positions. More qualitative comparisons are in the supplementary video.

Our method has the advantage of jointly estimating the humans and the scene point cloud, which can be further used to impose physical constrains in the estimated humans over time. The effect of these constraints can be visually seen in Figure 7, where we show a sequence of a person standing on the floor. In the top row, where no physical constraints were applied, we can observe that the right foot oscillates drastically from one frame to another. When the physical constraints are applied (the bottom row), this artifact is drastically reduced, and the right foot stays still in contact with the ground.
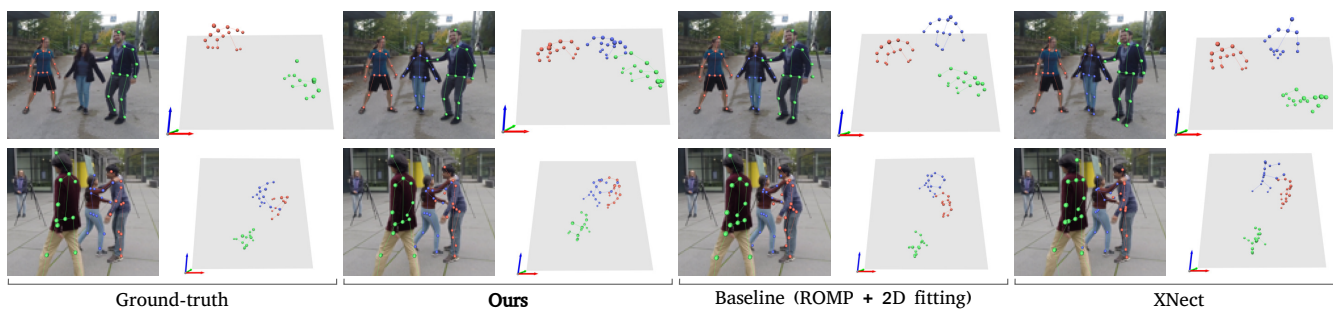
| Ground-truth | Ours | Baseline (ROMP + 2D fitting) | XNect |

Figure 4: *Comparisons of predictions from our method with other approaches. Compared to XNect and our baseline, our method is the only one that is able to estimate the person scale. Therefore, it predicts human positions in a more coherent way even for people of smaller height. 3D human poses are shown in the image plane (left) and top view (right). The ground-truth pose is not available for all the subjects in the dataset. Digital zoom is recommended.*
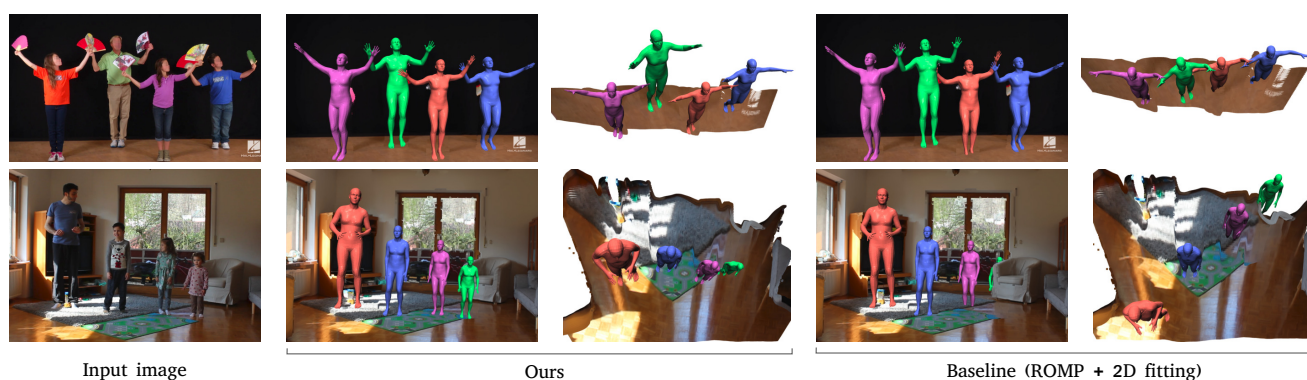

| Input image | Ours | Baseline (ROMP + 2D fitting) |

Figure 5: *3D Human poses estimated by our method from Internet videos. The baseline method can correctly localise the persons in the image plane, but fails drastically in positioning the characters in 3D. Note from our method the correct character order along the depth channel and the correctly estimated scale for each person. Digital zoom is recommended.*
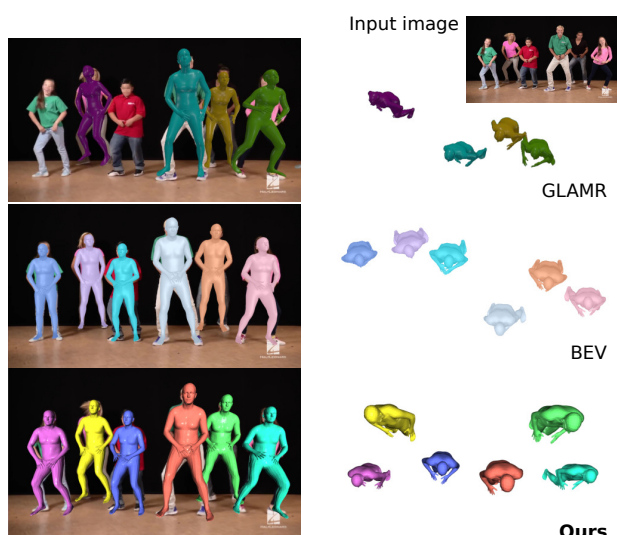

Figure 6: *Our results compared to BEV [SLB*22] and GLAMR [YIM*22] on a scene with people of different sizes.*

Since our method does not require any specific training procedure and rely on multiple predictions from models trained on a large corpus of data, our approach automatically generalizes well for in-the-wild and Internet videos, as can be seen in Figure 5 and can be directly used to drive virtual characters from monocular RGB videos; see Figure 8.

5. Discussion

Our method achieves low reconstruction errors, because it can successfully leverage multi-modal inputs to disambiguate the relative depths between humans and human scales better than previous works. Moreover, our results evince significantly less jitter and foot-floor penetrations than the evaluated baselines for multi-human 3D pose estimation and the ablative study confirms that all components of the method contribute to the final accuracy. We have demonstrated that the recovered 3D human motions can be applied for virtual character animation, as one potential application among the many others.

Limitations and Possible Extensions. Although our method outperforms competing methods and makes a step forward in monoc-
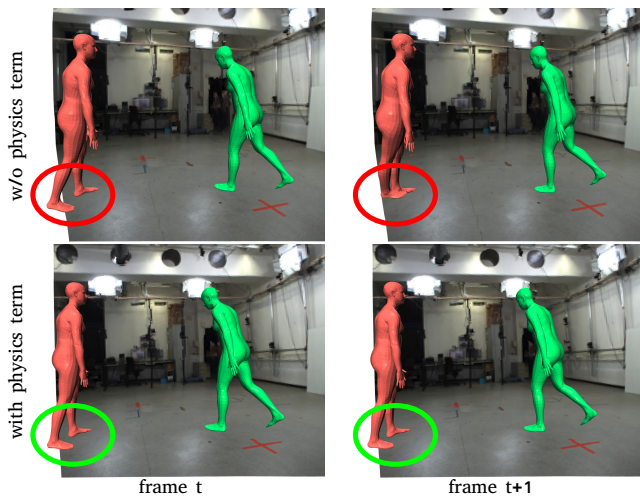
Figure 7: *The effect of the physical constrains imposed by the estimated geometry in our predictions. The results without $\mathcal{E}_{\text{contact}}$ and $\mathcal{E}_{\text{slip}}$ (top) contain more foot sliding artifacts than our results with physical constrains (bottom).*
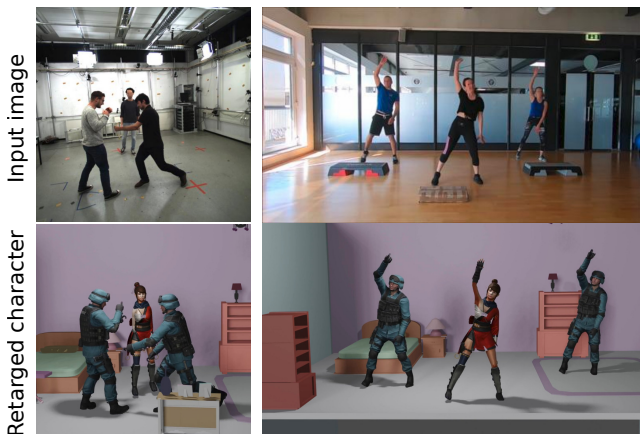

Figure 8: *Our method can be directly used to drive virtual characters or animate avatars in augmented reality applications (bottom row) from monocular RGB videos. Note the correct character order along the depth channel. Thanks to our physical plausibility constraints, barely any foot-floor penetrations or foot sliding are observed in the animations; see the video.*

ular multi-human 3D motion capture, it has several limitations caused by the severe ill-posedness of our monocular setting. All these limitations open possibilities for future extensions and follow-up works as described in the following.

First, our approach relies on multiple inputs from pre-trained models (depth maps and 2D body joints) and, therefore, could also be negatively affected by the output of those methods; for example if the estimated depth maps contain significant artefacts (*e.g.,* when

obtained on our-of-distribution environments). On the other hand, this implies that the performance of our approach has the potential to keep increasing in the future with the progress in related fields (*cf.* Table 5).

Our method also requires that people are entirely visible in most of the frames and move in the scene. Otherwise, the setting becomes degenerate, and we do not get enough cues for accurate reconstruction. Even though we mitigate artefacts that appear as violations of physical laws by geometric terms, some minor ones of this type remain. Further improvements can be attained by methods explicitly modelling physical laws as in single-human 3D motion capture [SGXT20; SGX*21; XWI*21].

Moreover, while the static camera assumption is practical, it is also very challenging, and a moving camera could provide additional 3D reconstruction cues. Finally, the proposed approach is an optimization method that can efficiently process an entire video sequence and extract relevant information about the scene from all frames globally. However, due to this characteristic, the method in its current version does not allow real-time applications.

6. Conclusion

We present a new holistic approach for multi-human 3D motion capture from a single static monocular RGB camera. Our core statement—that the synergy between multi-modal inputs and priors can significantly boost the 3D reconstruction accuracy in this challenging setting—is confirmed by extensive experiments in which we set a new state of the art on commonly used benchmarks. Moreover, as expected, we confirm that the constraints from the scene point clouds steadily boost the accuracy of the final 3D poses. Qualitatively, our reconstructions evince substantially fewer artefacts (such as jitter and foot-floor penetrations), enabling exciting downstream applications such as motion re-targeting for virtual characters. We believe that the proposed holistic approach for multi-human 3D motion capture can be extended in many useful ways, and we will be excited to see follow-ups.

7. Acknowledgements

References

[BCL*20] BENZINE, ABDALLAH, CHABOT, FLORIAN, LUVISON, BERTRAND, et al. "Pandanet: Anchor-based single-shot multi-person 3d pose estimation". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 6856–6865 3.

[BGFR19] BIELER, DIDIER, GUNEL, SEMIH, FUA, PASCAL, and RHODIN, HELGE. "Gravity as a Reference for Estimating a Person's Height From Video". *International Conference on Computer Vision (ICCV)*. 2019 3.

[BXP*22] BHATNAGAR, BHARAT LAL, XIE, XIANGHUI, PETROV, ILYA A, et al. "BEHAVE: Dataset and Method for Tracking Human Object Interactions". *arXiv preprint arXiv:2204.06950* (2022) 3.

[CHS*21] CAO, ZHE, HIDALGO, GINES, SIMON, TOMAS, et al. "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (2021), 172–186 2.

[CHY*19] CHEN, YIXIN, HUANG, SIYUAN, YUAN, TAO, et al. "Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, 8648–8657 3.

[CMS*21] CHENG, BOWEN, MISRA, ISHAN, SCHWING, ALEXANDER G, et al. "Masked-attention mask transformer for universal image segmentation". *arXiv preprint arXiv:2112.01527* (2021) 2, 5.

[CR17] CHEN, CHING-HANG and RAMANAN, DEVA. "3D Human Pose Estimation = 2D Pose Estimation + Matching". *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 2.

[CRV12] CASIEZ, GÉRY, ROUSSEL, NICOLAS, and VOGEL, DANIEL. "1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems". *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2012, 2527–2530 7.

[DGM*19] DABRAL, RISHABH, GUNDAVARAPU, NITESH B, MITRA, RAHUL, et al. "Multi-Person 3D Human Pose Estimation from Monocular Images". *2019 International Conference on 3D Vision (3DV)*. 2019, 405–414 3.

[DSJ*21] DABRAL, RISHABH, SHIMADA, SOSHI, JAIN, ARJUN, et al. "Gravity-Aware Monocular 3D Human-Object Reconstruction". *The IEEE Conference on International Conference on Computer Vision (ICCV)*. Oct. 2021, 12365–12374 3.

[FXTL17] FANG, HAO-SHU, XIE, SHUQIN, TAI, YU-WING, and LU, CEWU. "RMPE: Regional Multi-person Pose Estimation". *The IEEE Conference on International Conference on Computer Vision (ICCV)*. 2017 2, 5.

[GMSP21] GUZOV, VLADIMIR, MIR, AYMEN, SATTLER, TORSTEN, and PONS-MOLL, GERARD. "Human POSEitioning System (HPS): 3D Human Pose Estimation and Self-localization in Large Scenes from Body-Mounted Sensors". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. June 2021 3.

[HCTB19] HASSAN, MOHAMED, CHOUTAS, VASILEIOS, TZIONAS, DIMITRIOS, and BLACK, MICHAEL J. "Resolving 3D Human Pose Ambiguities with 3D Scene Constraints". *International Conference on Computer Vision (ICCV)*. Oct. 2019, 2282–2292 3.

[HLL22] HENNING, DORIAN F, LAIDLOW, TRISTAN, and LEUTENEGGER, STEFAN. "BodySLAM: Joint Camera Localisation, Mapping, and Human Motion Tracking". *European Conference on Computer Vision*. Springer. 2022, 656–673 3.

[HSS12] HINTON, GEOFFREY, SRIVASTAVA, NITISH, and SWERSKY, KEVIN. "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent". (2012) 8.

[HXZ*19] HABERMANN, MARC, XU, WEIPENG, ZOLLHÖFER, MICHAEL, et al. "LiveCap: Real-Time Human Performance Capture From Monocular Video". *ACM Transactions on Graphics (TOG)* 38.2 (2019), 14:1–14:17. ISSN: 0730-0301 2, 3.

[HXZ*20] HABERMANN, MARC, XU, WEIPENG, ZOLLHOEFER, MICHAEL, et al. "DeepCap: Monocular Human Performance Capture Using Weak Supervision". *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)* 1 (2020), 1 2, 3.

[HYH*22] HUANG, CHUN-HAO P, YI, HONGWEI, HÖSCHLE, MARKUS, et al. "Capturing and inferring dense full-body human-scene contact". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 13274–13285 3.

[IPOS13] IONESCU, CATALIN, PAPAVA, DRAGOS, OLARU, VLAD, and SMINCHISESCU, CRISTIAN. "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments". *IEEE transactions on pattern analysis and machine intelligence* 36.7 (2013), 1325–1339 2.

[JKP*20] JIANG, WEN, KOLOTOUROS, NIKOS, PAVLAKOS, GEORGIOS, et al. "Coherent reconstruction of multiple humans from a single image". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, 5579–5588 2, 3, 7, 8.

[JLT*15] JOO, HANBYUL, LIU, HAO, TAN, LEI, et al. "Panoptic Studio: A Massively Multiview System for Social Motion Capture". *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, 3334–3342 7.

[KAB20] KOCABAS, MUHAMMED, ATHANASIOU, NIKOS, and BLACK, MICHAEL J. "VIBE: Video Inference for Human Body Pose and Shape Estimation". *Computer Vision and Pattern Recognition (CVPR)*. 2020 2.

[KBJM18] KANAZAWA, ANGJOO, BLACK, MICHAEL J., JACOBS, DAVID W., and MALIK, JITENDRA. "End-to-end Recovery of Human Shape and Pose". *Computer Vision and Pattern Recognition (CVPR)*. 2018, 7122–7131 3.

[KPBD19] KOLOTOUROS, NIKOS, PAVLAKOS, GEORGIOS, BLACK, MICHAEL J., and DANIILIDIS, KOSTAS. "Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop". *International Conference on Computer Vision (ICCV)*. 2019, 2252–2261 3.

[LBX*22] LI, JIEFENG, BIAN, SIYUAN, XU, CHAO, et al. "D&D: Learning Human Dynamics from Dynamic Camera". *European Conference on Computer Vision*. Springer. 2022, 479–496 3, 7.

[LDC*19] LI, ZHENGQI, DEKEL, TALI, COLE, FORRESTER, et al. "Learning the depths of moving people by watching frozen people". *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, 4521–4530 2.

[LL20] LIN, JIAHAO and LEE, GIM HEE. "Hdnet: Human depth estimation for multi-person camera-space localization". *European Conference on Computer Vision*. Springer. 2020, 633–648 3.

[LMR*15] LOPER, MATTHEW, MAHMOOD, NAUREEN, ROMERO, JAVIER, et al. "SMPL: A skinned multi-person linear model". *ACM transactions on graphics (TOG)* 34.6 (2015), 1–16 3, 4.

[LPT22] LUVIZON, DIOGO C., PICARD, DAVID, and TABIA, HEDI. "Consensus-Based Optimization for 3D Human Pose Estimation in Camera Coordinates". *International Journal of Computer Vision* 130 (2022), 869–882 2.

[LVC*21] LI, JIAMAN, VILLEGAS, RUBEN, CEYLAN, DUYGU, et al. "Task-generic hierarchical human motion prior using vaes". *2021 International Conference on 3D Vision (3DV)*. IEEE. 2021, 771–781 3.

[LYZ*21] LIU, MIAO, YANG, DEXIN, ZHANG, YAN, et al. "4D Human Body Capture from Egocentric Video via 3D Scene Grounding". *2021 International Conference on 3D Vision (3DV)*. IEEE. 2021, 930–939 3.

[MCL19] MOON, GYEONGSIK, CHANG, JUYONG, and LEE, KYOUNG MU. "Camera Distance-aware Top-down Approach for 3D Multi-person Pose Estimation from a Single RGB Image". *The IEEE Conference on International Conference on Computer Vision (ICCV)*. 2019 2, 3, 7, 8.

[MHRL17] MARTINEZ, JULIETA, HOSSAIN, RAYAT, ROMERO, JAVIER, and LITTLE, JAMES J. "A simple yet effective baseline for 3d human pose estimation". *Proceedings of the IEEE international conference on computer vision*. 2017, 2640–2649 2.

[MRC*17] MEHTA, DUSHYANT, RHODIN, HELGE, CASAS, DAN, et al. "Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision". *International Conference on 3D vision (3DV)*. 2017 2, 7.

[MSM*18] MEHTA, DUSHYANT, SOTNYCHENKO, OLEKSANDR, MUELLER, FRANZISKA, et al. "Single-Shot Multi-Person 3D Pose Estimation From Monocular RGB". *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE. Sept. 2018 7, 8.

[MSM*20] MEHTA, DUSHYANT, SOTNYCHENKO, OLEKSANDR, MUELLER, FRANZISKA, et al. "XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera". Vol. 39. 4. July 2020 3, 8, 9.

[MSS*17] MEHTA, DUSHYANT, SRIDHAR, SRINATH, SOTNYCHENKO, OLEKSANDR, et al. "VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera". Vol. 36. 4. 2017 2.

[PCG*19] PAVLAKOS, GEORGIOS, CHOUTAS, VASILEIOS, GHORBANI, NIMA, et al. "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image". *Computer Vision and Pattern Recognition (CVPR)*. 2019, 10975–10985 3.

[PGM*19] PASZKE, ADAM, GROSS, SAM, MASSA, FRANCISCO, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". *Advances in Neural Information Processing Systems 32*. Ed. by WALLACH, H., LAROCHELLE, H., BEYGELZIMER, A., et al. Curran Associates, Inc., 2019, 8024–8035 8.

[PZDD17] PAVLAKOS, GEORGIOS, ZHOU, XIAOWEI, DERPANIS, KONSTANTINOS G, and DANIILIDIS, KOSTAS. "Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose". *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 2.

[RBH*21] REMPE, DAVIS, BIRDAL, TOLGA, HERTZMANN, AARON, et al. "HuMoR: 3D Human Motion Model for Robust Pose Estimation". *International Conference on Computer Vision (ICCV)*. 2021 7.

[RBK21] RANFTL, RENÉ, BOCHKOVSKIY, ALEXEY, and KOLTUN, VLADLEN. "Vision transformers for dense prediction". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, 12179–12188 2, 5, 9.

[RLH*21] RANFTL, RENÉ, LASINGER, KATRIN, HAFNER, DAVID, et al. "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer". *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020) 9.

[RRN*20] RAVI, NIKHILA, REIZENSTEIN, JEREMY, NOVOTNY, DAVID, et al. "Accelerating 3D Deep Learning with PyTorch3D". *arXiv:2007.08501* (2020) 5, 6, 8.

[RWS17] ROGEZ, GREGORY, WEINZAEPFEL, PHILIPPE, and SCHMID, CORDELIA. "LCR-Net: Localization-Classification-Regression for Human Pose". *CVPR*. 2017 3, 8.

[RWS19] ROGEZ, GRÉGORY, WEINZAEPFEL, PHILIPPE, and SCHMID, CORDELIA. "LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images". *IEEE transactions on pattern analysis and machine intelligence* (2019) 3, 8.

[SBL*21] SUN, YU, BAO, QIAN, LIU, WU, et al. "Monocular, One-stage, Regression of Multiple 3D People". *The IEEE Conference on International Conference on Computer Vision (ICCV)*. 2021 2, 3, 5, 8, 9.

[SGX*21] SHIMADA, SOSHI, GOLYANIK, VLADISLAV, XU, WEIPENG, et al. "Neural Monocular 3D Human Motion Capture with Physical Awareness". *ACM Transactions on Graphics* 40.4 (2021) 3, 7, 11.

[SGXT20] SHIMADA, SOSHI, GOLYANIK, VLADISLAV, XU, WEIPENG, and THEOBALT, CHRISTIAN. "PhysCap: Physically Plausible Monocular 3D Motion Capture in Real Time". *ACM Transactions on Graphics* 39.6 (2020) 3, 7, 11.

[SLB*22] SUN, YU, LIU, WU, BAO, QIAN, et al. "Putting people in their place: Monocular regression of 3d people in depth". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 13243–13252 3, 8–10.

[SSLW17] SUN, XIAO, SHANG, JIAXIANG, LIANG, SHUANG, and WEI, YICHEN. "Compositional Human Pose Regression". *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017 2.

[SXW*18] SUN, XIAO, XIAO, BIN, WEI, FANGYIN, et al. "Integral Human Pose Regression". *The European Conference on Computer Vision (ECCV)*. Sept. 2018 2.

[TKS*16] TEKIN, BUGRA, KATIRCIOGLU, ISINSU, SALZMANN, MATHIEU, et al. "Structured Prediction of 3D Human Pose with Deep Neural Networks". *The British Machine Vision Conference (BMVC)*. 2016 2.

[URA*21] UGRINOVIC, NICOLAS, RUIZ, ADRIA, AGUDO, ANTONIO, et al. "Body Size and Depth Disambiguation in Multi-Person Reconstruction from Single Images". *2021 International Conference on 3D Vision (3DV)*. IEEE. 2021, 53–63 3.

[WAT*20] WATSON, JAMIE, AODHA, OISIN MAC, TURMUKHAMBETOV, DANIYAR, et al. "Learning Stereo from Single Images". *European Conference on Computer Vision (ECCV)*. 2020 5.

[WLL*20] WANG, CAN, LI, JIEFENG, LIU, WENTAO, et al. "Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation". *European Conference on Computer Vision*. Springer. 2020, 242–259 3.

[WR19] WANDT, BASTIAN and ROSENHAHN, BODO. "RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation". *Computer Vision and Pattern Recognition (CVPR)*. 2019 2.

[WSC*20] WANG, JINGDONG, SUN, KE, CHENG, TIANHENG, et al. "Deep high-resolution representation learning for visual recognition". *IEEE transactions on pattern analysis and machine intelligence* 43.10 (2020), 3349–3364 9.

[WY21] WENG, ZHENZHEN and YEUNG, SERENA. "Holistic 3d human and scene mesh estimation from single view images". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, 334–343 3.

[XCZ*18] XU, WEIPENG, CHATTERJEE, AVISHEK, ZOLLHÖFER, MICHAEL, et al. "MonoPerfCap: Human Performance Capture From Monocular Video". *ACM Trans. Graph.* 37.2 (May 2018), 27:1–27:15. ISSN: 0730-0301 2, 3.

[XWI*21] XIE, KEVIN, WANG, TINGWU, IQBAL, UMAR, et al. "Physics-Based Human Motion Estimation and Synthesis From Videos". *International Conference on Computer Vision (ICCV)*. 2021 11.

[YHT*22] YI, HONGWEI, HUANG, CHUN-HAO P, TZIONAS, DIMITRIOS, et al. "Human-aware object placement for visual environment reconstruction". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 3959–3970 3.

[YIM*22] YUAN, YE, IQBAL, UMAR, MOLCHANOV, PAVLO, et al. "GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 11038–11049 3, 8–10.

[YPL21] YU, RI, PARK, HWANGPIL, and LEE, JEHEE. "Human Dynamics from Monocular Video with Dynamic Camera Movements". *ACM Trans. Graph.* 40.6 (2021) 3.

[ZFS*20] ZHEN, JIANAN, FANG, QI, SUN, JIAMING, et al. "Smap: Single-shot multi-person absolute 3d pose estimation". *European Conference on Computer Vision*. Springer. 2020, 550–566 3, 8.

[ZHS*17] ZHOU, XINGYI, HUANG, QIXING, SUN, XIAO, et al. "Towards 3d human pose estimation in the wild: a weakly-supervised approach". *Proceedings of the IEEE International Conference on Computer Vision*. 2017, 398–407 2.

[ZMS18] ZANFIR, ANDREI, MARINOIU, ELISABETA, and SMINCHISESCU, CRISTIAN. "Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, 2148–2157 3, 7.

[ZMZ*22] ZHANG, SIWEI, MA, QIANLI, ZHANG, YAN, et al. "EgoBody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices". *European Conference on Computer Vision*. Springer. 2022, 180–200 3.

[ZSZ*16] ZHOU, XINGYI, SUN, XIAO, ZHANG, WEI, et al. "Deep kinematic pose regression". *European Conference on Computer Vision*. Springer. 2016, 186–201 3.

[ZZB*21a] ZANFIR, MIHAI, ZANFIR, ANDREI, BAZAVAN, EDUARD GABRIEL, et al. "Thundr: Transformer-based 3d human reconstruction with markers". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, 12971–12980 3.

[ZZB*21b] ZHANG, SIWEI, ZHANG, YAN, BOGO, FEDERICA, et al. "Learning motion priors for 4d human body capture in 3d scenes". *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, 11343–11353 3.