TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments

Yu Sun[1*] Qian Bao[2†] Wu Liu[2†] Tao Mei[4] Michael J. Black[3]
[1]Harbin Institute of Technology [2]Explore Academy of JD.com
[3]Max Planck Institute for Intelligent Systems [4]HiDream.ai Inc.
yusunhit@gmail.com, baoqian@jd.com, liuwu1@jd.com
tmei@hidream.ai, black@tuebingen.mpg.de

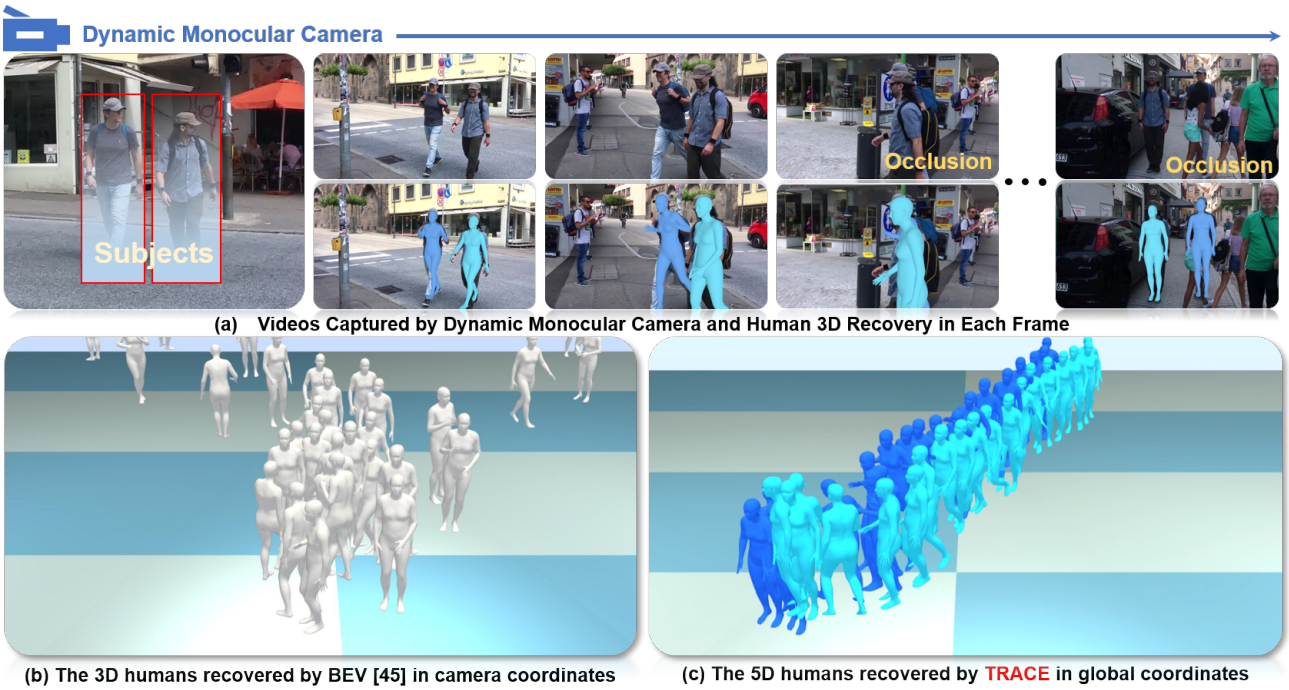

Figure 1. **5D temporal reconstruction of multiple 3D people in global coordinates with a dynamic camera.** We introduce TRACE, a monocular one-stage method with a holistic 5D representation that enables the network to explicitly reason about human motion across time. Unlike previous methods [44,45] **(b)** that regress all 3D people from each frame in camera coordinates, TRACE **(c)** tracks the subjects **(a)** presented in the first frame through time and recovers their global trajectories in global coordinates in one shot.

Abstract

Although the estimation of 3D human pose and shape (HPS) is rapidly progressing, current methods still cannot reliably estimate moving humans in global coordinates, which is critical for many applications. This is particularly challenging when the camera is also moving, entangling human and camera motion. To address these issues, we adopt a novel 5D representation (space, time, and identity) that enables end-to-end reasoning about people in scenes. Our method, called TRACE, introduces several novel architectural components. Most importantly, it uses two new "maps" to reason about the 3D trajectory of people over time in camera, and world, coordinates. An additional memory unit enables persistent tracking of people even during long occlusions. TRACE is the first one-stage method to jointly recover and track 3D humans in global coordinates from dynamic cameras. By training it end-to-end, and using full image information, TRACE achieves state-of-the-art performance on tracking and HPS benchmarks. The code[1] and dataset[2] are released for research purposes.

*This work was done when Yu Sun was an intern at JD.com.
†Corresponding author.

[1]https://www.yusun.work/TRACE/TRACE.html
[2]https://github.com/Arthur151/DynaCam

1. Introduction

The estimation of 3D human pose and shape (HPS) has many applications and there has been significant recent progress [6, 7, 21–24, 37, 38, 46, 49, 56, 57, 60]. Most methods, however, reason only about a single frame at a time and estimate humans in *camera coordinates*. Moreover, such methods do not *track* people and are unable to recover their global trajectories. The problem is even harder in typical hand-held videos, which are filmed with a dynamic, moving, camera. For many applications of HPS, single-frame estimates in camera coordinates are not sufficient. To capture human movement and then transfer it to a new 3D scene, we must have the movement in a coherent global coordinate system. This is a requirement for computer graphics, sports, video games, and extended reality (XR).

Our key insight is that most methods estimate humans in 3D, whereas the true problem is 5D. That is, a method needs to reason about 3D space, time, and subject identity. With a 5D representation, the problem becomes tractable, enabling a *holistic* solution that can exploit the full video to infer multiple people in a coherent global coordinate frame. As illustrated in Fig. 1, we develop a unified method to jointly regress the 3D pose, shape, identity, and global trajectory of the subjects in global coordinates from monocular videos captured by dynamic cameras (DC-videos).

To achieve this, we deal with two main challenges. First, DC-videos contain both human motion and camera motion and these must be disentangled to recover the human trajectory in global coordinates. One idea would be to recover the camera motion relative to the rigid scene using structure-from-motion (SfM) methods (e.g. [31]). In scenes containing many people and human motion, however, such methods can be unreliable. An alternative approach is taken by GLAMR [58], which infers global human trajectories from local 3D human poses, without taking into account the full scene. By ignoring evidence from the full image, GLAMR fails to capture the correct global motion in common scenarios, such as biking, skating, boating, running on a treadmill, etc. Moreover, GLAMR is a multi-stage method, with each stage dependent on accurate estimates from the preceding one. Such approaches are more brittle than our holistic, end-to-end, method.

The other challenge, as shown in the upper right corner of Fig. 1, is that severe occlusions are common in videos with multiple people. Currently, the most popular tracking strategy is to infer the association between 2D detections using a temporal prior (e.g. Kalman filter) [63]. However, in DC-videos, human motions are often irregular and can easily violate hand-crafted priors. PHALP [40] is one of the few methods to address this for 3D HPS. It uses a classical, multi-stage, detection-and-tracking formulation with heuristic temporal priors. It does not holistically reason about the sequence and is not trained end-to-end.

To address these issues, we reason about people using a 5D representation and capture information from the full image and the motion of the scene. This holistic reasoning enables the reliable recovery of global human trajectories and subject tracking using a single-shot method. This is more reliable than multi-stage methods because the network can exploit more information to solve the task and is trained end-to-end. No hand-crafted priors are needed and the network is able to share information among modules.

Specifically, we develop TRACE, a unified one-stage method for **T**emporal **R**egression of **A**vatars with dynamic **C**ameras in 3D **E**nvironments. The architecture is inspired by BEV [45], which directly estimates multiple people in depth from a single image using multiple 2D maps. BEV uses a 2D map representing an imaginary, "top down", view of the scene. This is combined with an image-centric 2D map to reason about people in 3D. Our key insight is that the idea of maps can be extended to represent how people *move in 3D*. With this idea, TRACE introduces three new modules to holistically model 5D human states, performing multi-person temporal association, and inferring human trajectories in global coordinates; see Fig. 2.

First, to construct a holistic 5D representation of the video, we extract temporal image features by fusing single-frame feature maps from the image backbone with a temporal feature propagation module. We also compute the optical flow between adjacent frames with a motion backbone. The optical flow provides short-term motion features that carry information about the motion of the scene and the people. Second, to explicitly track human motions, we introduce a novel *3D motion offset map* to establish the association of the same person across adjacent frames. This map contains a 3D offset vector at each position, which represents the difference between the 3D positions of the same subject from the previous frame to the current frame in camera coordinates. We also introduce a *memory unit* to keep track of subjects under long-term occlusion. Note that the 3D trajectories are built in camera space, and TRACE uses a novel *world motion map* that transfers the trajectories to global coordinates. At each position, this map contains a 6D vector to represent the difference between the 3D positions of the corresponding subject from the previous frame to the current frame and its 3D orientation in world coordinates. Taken together, this novel network architecture goes beyond prior work by taking information from the full video frames to address detection, pose estimation, tracking, and occlusion in a holistic network that is trained end-to-end.

To enable training and evaluation of global human trajectory estimation from in-the-wild DC-videos, we build a new dataset, DynaCam. Since collecting global human trajectories and camera poses with in-the-wild DC-videos is difficult, we simulate a moving camera using publicly available in-the-wild panoramic videos and regular videos cap-

tured by static cameras. In this way, we create more than 500 in-the-wild DC-videos with precise camera pose annotations. Then we generate pseudo-ground-truth 3D human annotations via fitting [20] SMPL [32] to detected 2D pose sequences [17,54,63]. With 2D/3D human pose and camera pose annotations, we can obtain the global human trajectories using the PnP algorithm [16]. This dataset is sufficient to train TRACE to deal with dynamic cameras.

We evaluate TRACE on two multi-person in-the-wild benchmarks (3DPW [50] and MuPoTS-3D [35]) and our DynaCam dataset. On 3DPW, TRACE achieves the state-of-the-art (SOTA) PA-MPJPE of 37.8mm, less the current best (42.7 [26]). On MuPoTS-3D, TRACE outperforms previous 3D-representation-based methods [39, 40] and tracking-by-detection methods [63] on tracking people under long-term occlusion. On DynaCam, TRACE outperforms GLAMR [58] in estimating the 3D human trajectory in global coordinates from DC-videos.

In summary, our main contributions are: (1) We introduce a 5D representation and use it to learn holistic temporal cues related to both 3D human motions and the scene. (2) We introduce two novel motion offset representations to explicitly model temporal multi-subject association and global human trajectories from temporal clues in an end-to-end manner. (3) We estimate long-term 3D human motions over time in global coordinates, achieving SOTA results. (4) We collect the DynaCam dataset of DC-videos with pseudo ground truth, which facilitates the training and evaluation of global human trajectory estimation. The code and dataset are publicly available for research purposes.

2. Related Work

Monocular 3D mesh regression with full images. Most existing methods [21–24, 26, 29, 37, 38, 46, 52, 53, 59, 60] take a multi-stage detection-based pipeline to estimate 3D HPS from cropped image patches, which exclude important cues, such as camera information and human-scene relationships. A few recent multi-stage [28] and one-stage [44, 45] methods have made steps towards using the full-image information. For instance, CLIFF [28] estimates 3D HPS by taking into account the bounding box locations, giving the method camera information and improving accuracy. To directly estimates multiple people at once from the full image, ROMP [44] introduces a 2D Center heatmap and a Mesh parameter map to represent 2D human locations and 3D human body meshes, respectively. BEV [45] goes beyond ROMP by introducing an imaginary bird's-eye-view map, which is combined with the front-view maps to construct a 3D view in camera coordinates. However, they only model 3D HPS in camera coordinates from a single image. Using "maps" like ROMP and BEV, TRACE also looks at the full image. We go further, however, by introducing novel maps that model human motions across a video sequence in global coordinates.

Tracking datasets. While there are many tracking datasets [8, 10, 11, 14, 18, 61] with 2D annotations, only a few [13, 15] capture the 3D trajectory of pedestrians. In both cases, the scene and human activities are limited. To address this, we use 3DPW [50] and MuPoTS-3D [35] for tracking evaluation. 3DPW is the most relevant dataset for our task and it provides a real-world test case. 3DPW contains videos that are captured by a moving camera that follows subjects to record their activities in many daily scenes. MuPoTS-3D contains rich multi-person interaction scenes with long-term occlusions for tracking evaluation.

Tracking 3D people through occlusions. Most existing methods [4, 42, 51, 63] perform tracking using 2D image cues. The classic tracking-by-detection paradigm focuses on associating the 2D detections using a temporal prior (e.g. Kalman filter). When applied to DC-videos containing rapid human and camera motions that violate the hand-crafted priors, such methods are brittle. Going beyond 2D, PHALP [40] separately extracts 3D human pose, appearance, and location with a multi-stage design from each video frame, and then assembles them for tracking. In contrast to these multi-stage methods, which are susceptible to errors in early stages, we explicitly learn the 3D human trajectory from temporal 5D cues in an end-to-end manner.

Monocular global 3D human trajectory reasoning. Most existing methods [64] that reason about global 3D human trajectories do so with static, calibrated, cameras in a multi-view setting. A few recent methods [31, 58] have addressed the ill-posed problem of extracting the global motions of humans from monocular video. Liu et al. [31] employ a structure-from-motion (SfM) method [43] to estimate the camera poses from monocular videos captured by a dynamic camera. However, when the input video contains the movement of multiple subjects, it is hard for SfM methods to extract sufficiently many stable keypoints for reliable camera estimation. GLAMR [58] adopts a multi-stage pipeline to infer the global human trajectory from root-relative local human 3D poses estimated from each frame. The per-frame human pose estimates make it vulnerable to occlusion. Additionally, GLAMR relies on bounding boxes, ignoring scene-related information. Consequently, GLAMR fails in common scenarios like riding a bike or skating. In concurrent work (in this proceedings), SLAHMR [55] uses a multi-stage optimization-based approach that combines structure from motion with human motion priors to estimate 4D human trajectories in global coordinates; this is very computationally expensive. In contrast to previous multi-stage methods, TRACE simultaneously combines scene information and 3D human motions with a novel 5D representation to holistically exploit all temporal cues and to enable end-to-end training.

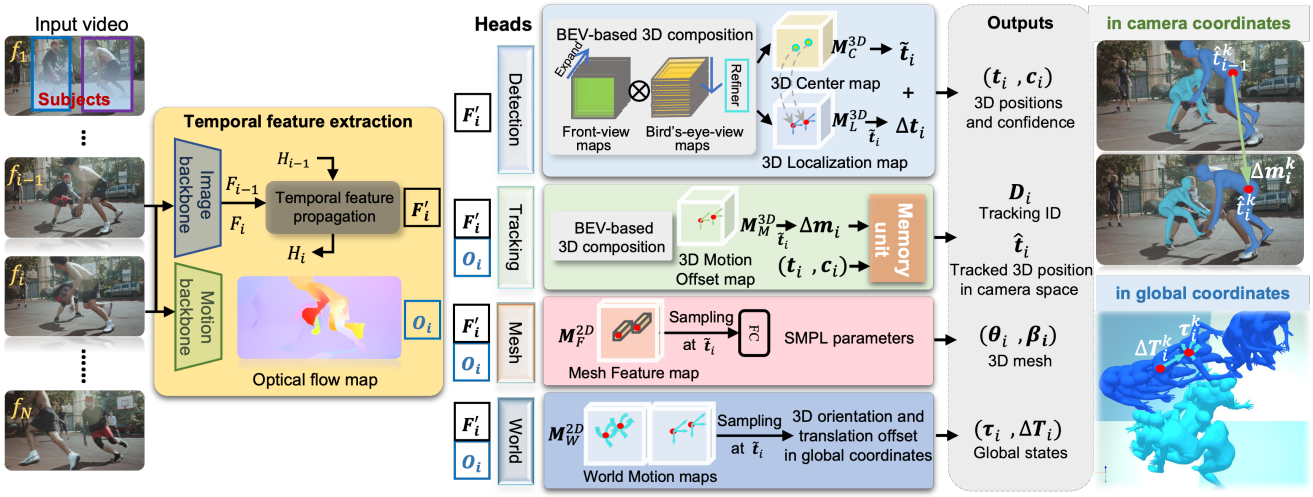

Figure 2. **TRACE Overview**. TRACE takes a video sequence and regions in the first frame corresponding to the subjects to be tracked. TRACE encodes the video and its motion with temporal features and optical flow. A novel 3D Motion Offset map reasons about human trajectories in camera coordinates. The World Motion map represents the trajectory in global coordinates. A memory unit deals with occlusions by encoding the subject identities. TRACE is trained end-to-end to estimate the 3D shape and pose of multiple people throughout a video in global coordinates. See Sec. 3.1 for details.

3. Method

3.1. Overview

The overall framework of TRACE is shown in Fig. 2. Given a video sequence captured with a dynamic camera $\{f_i, i = 1, \ldots, N\}$ with N frames, the user specifies K tracking subjects shown in the first frame. Our goal is to simultaneously recover the 3D pose, shape, identity, and trajectory of each subject in global coordinates. To achieve this, TRACE first extracts temporal features and then decodes each sub-task with a separate head network. First, via two parallel backbones, TRACE encodes the video and its motion into temporal image feature maps $\boldsymbol{F}'_i$ and motion feature maps $\boldsymbol{O}_i$.

The Detection and Tracking branches take these features and perform multi-subject tracking in *camera coordinates*. Unlike BEV [45], our detection method takes *temporal* image features $\boldsymbol{F}'_i$ as input. It uses the features to detect the 3D human positions $\boldsymbol{t}_i$ and their confidence $\boldsymbol{c}_i$ for all people in frame f_i. The Mesh branch regresses all the human mesh parameters $(\boldsymbol{\theta}_i, \boldsymbol{\beta}_i)$, in SMPL [32] format, from the input Feature maps. Unlike BEV, this branch takes both temporal image features and motion features.

The combined features $(\boldsymbol{F}'_i, \boldsymbol{O}_i)$ are fed to our novel Tracking branch to estimate the 3D Motion Offset map, indicating the 3D position change of each subject across frames. The new Memory Unit takes the 3D detection and its 3D motion offset as input. It then determines the subject identities and builds human trajectories $\widehat{\boldsymbol{t}}_i$ of the K subjects in camera coordinates. Note that, like BEV, our detection branch finds all the people in the video frames but our goal is to track only the K input subjects. Consequently, the memory unit filters out detected people who do not match the subject trajectories.

Finally, to estimate subject trajectories in global coordinates, the World branch estimates a world motion map, representing the 3D orientation $\boldsymbol{\tau}_i$ and 3D translation offset $\Delta \boldsymbol{T}_i$ of the K subjects in global coordinates. Accumulating $\Delta \boldsymbol{T}_i$, starting with the 3D position $\widehat{\boldsymbol{t}}_1$ of the tracked subjects in the first frame, gives their global 3D trajectory $\boldsymbol{T}$. Note that the global ("world") coordinates are defined relative to the camera coordinates of the first frame.

3.2. Holistic 5D Representation: Details

Rather than directly estimating camera poses from environment keypoints [31] or inferring global human trajectories from local body poses [58], we develop a 5D representation to directly reason about human states, perform multi-person temporal association, and infer human trajectories in global coordinates. Learning a holistic 5D representation is the foundation of our one-stage framework. The representation has five main parts.

i. Temporal feature maps. To construct the temporal feature maps encoding 5D human states and scenes information, we need to extract both single-frame image features and the motion features between adjacent frames. Therefore, given frame f_{i-1} and f_i, we adopt a parallel two-branch structure to extract temporal image feature maps $\boldsymbol{F}'_i$ and motion feature maps $\boldsymbol{O}_i$ for the current frame f_i. First, in the image branch, we extract single-frame feature maps $\boldsymbol{F}_{i-1}$ and $\boldsymbol{F}_i$ with an image backbone (HRNet-32 [9]). To extract long-term and short-term motion features, we construct a temporal feature propagation module by combining

a ConvGRU [47] module, a Deformable convolution [65] module, and a residual connection. With these, we fuse the image feature maps to generate a temporal image feature map $\boldsymbol{F}'_i$. See Sup. Mat. for details and experimental analysis. Additionally, in the motion branch, we estimate the optical flow map $\boldsymbol{O}_i$ between frames f_{i-1} and f_i with a motion backbone (RAFT [47]), to extract motion features of both people and scenes. From the combined temporal feature ($\boldsymbol{F}'_i$, $\boldsymbol{O}_i$), then we estimate five maps for the task.

ii. 3D detection maps. From the temporal image feature $\boldsymbol{F}'_i$, we estimate a 3D Center map $\boldsymbol{M}_C^{3D} \in \mathbb{R}^{1\times D\times H\times W}$ for coarse human detection and a 3D Localization map $\boldsymbol{M}_L^{3D} \in \mathbb{R}^{1\times D\times H\times W}$ for fine localization. The two 3D detection maps are composited from the front-view maps and the 2D bird's-eye-view maps following BEV [45]. For K subjects, we first parse out the detected 3D center positions $\widetilde{\boldsymbol{t}}_i$ and their detection confidences $\boldsymbol{c}_i$ from the 3D Center map. Then we sample the 3D Localization map at $\widetilde{\boldsymbol{t}}_i$ to obtain the fine 3D localization offset vectors $\Delta\boldsymbol{t}_i$. The predicted 3D translation of subjects to be tracked in camera coordinates is $\boldsymbol{t}_i = \widetilde{\boldsymbol{t}}_i + \Delta\boldsymbol{t}_i$.

iii. 3D Motion Offset map. To track subjects between adjacent frames using $\boldsymbol{F}'_i$ and $\boldsymbol{O}_i$, we estimate a 3D Motion Offset map $\boldsymbol{M}_M^{3D} \in \mathbb{R}^{3\times D\times H\times W}$ via a BEV-based 3D composition. We sample the $\boldsymbol{M}_M^{3D}$ at detected 3D center positions $\widetilde{\boldsymbol{t}}_i$ to obtain the 3D motion offset vectors $\Delta\boldsymbol{m}_i$, which represent the 3D position changes of the subjects from frame f_{i-1} to frame f_i in camera coordinates. With $\Delta\boldsymbol{m}_i, \boldsymbol{t}_i, \boldsymbol{c}_i$ as input, a memory unit (Sec. 3.3) predicts the tracked identities $\boldsymbol{D}_i$ and the optimized 3D position $\hat{\boldsymbol{t}}_i$ in camera coordinates, which filters out low-confidence detections while keeping track of subjects through long-term occlusion.

iv. Mesh Feature map. To regress the SMPL parameters of subjects, we first estimate a 2D Mesh Feature map $\boldsymbol{M}_F^{2D} \in \mathbb{R}^{C\times H\times W}$ from temporal features ($\boldsymbol{F}'_i$, $\boldsymbol{O}_i$). With the tracked 3D positions $\widehat{\boldsymbol{t}}_i = (\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{d}_i)$, we sample a mesh feature vector from $\boldsymbol{M}_F^{2D}$ at 2D positions $(\boldsymbol{u}_i, \boldsymbol{v}_i)$. To differentiate the features of people at different depths, we map the $\boldsymbol{d}_i$ to a 128-dim encoding vector via an embedding layer and add this with the mesh feature vector to regress SMPL [32] parameters $(\boldsymbol{\theta}_i, \boldsymbol{\beta}_i)$. The SMPL model maps the pose and shape, $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_i$, to a 3D human body mesh $\boldsymbol{B} \in \mathbb{R}^{6890\times 3}$. With a sparse weight matrix $\boldsymbol{R} \in \mathbb{R}^{Q\times 6890}$ that describes the mapping from $\boldsymbol{B}$ to Q body keypoints, we can obtain the 3D positions of body keypoints $\boldsymbol{J} \in \mathbb{R}^{Q\times 3}$ via $\boldsymbol{RB}$. With the estimated 3D positions $\widehat{\boldsymbol{t}}_i$ and a predefined camera projection matrix $\boldsymbol{P}$, the projected 2D keypoints are $\boldsymbol{J}_{2D} = \boldsymbol{P}(\boldsymbol{J} + \widehat{\boldsymbol{t}}_i)$.

v. World Motion map. $\boldsymbol{M}_W^{2D} \in \mathbb{R}^{6\times H\times W}$ contains the 6D vectors that describe the 3D orientation and the relative translation of the subjects in global coordinates. We sample $\boldsymbol{M}_W^{2D}$ at the 2D positions $(\boldsymbol{u}_i, \boldsymbol{v}_i) \in \widehat{\boldsymbol{t}}_i$ of the tracked subjects to obtain their 3D body orientation $\boldsymbol{\tau}_i$ and 3D body motion offset $\Delta\boldsymbol{T}_i$. For the tracked subject k, $\Delta\boldsymbol{T}_i^k$ represents their 3D position change from frame f_{i-1} to frame f_i in global coordinates. We take the camera coordinate system of the first frame as the global coordinate system for the video. The global 3D trajectories of all tracked subjects are obtained by accumulating $\Delta\boldsymbol{T}_i$ to their 3D position $\widehat{\boldsymbol{t}}_1$ in the first frame, $\boldsymbol{T} = \{\widehat{\boldsymbol{t}}_1, \widehat{\boldsymbol{t}}_1 + \Delta\boldsymbol{T}_2, \widehat{\boldsymbol{t}}_1 + \Delta\boldsymbol{T}_2 + \Delta\boldsymbol{T}_3, \dots\}$. Combining the 3D mesh, ID, and the global trajectories, the network reasons about 3D motions and trajectories of K tracked subjects in global coordinates.

3.3. Tracking with a Memory Unit

We construct the 3D trajectory of each subject by associating the 3D detections $\boldsymbol{t}_i$ over time with a 3D motion offset $\Delta\boldsymbol{m}_i$. To deal with long-term occlusions, we design the Memory Unit for persistent tracking, which will keep the memory for the full sequence. The memory unit stores the human states during inference and is not used for training. With predicted 3D positions $\boldsymbol{t}_i$, detection confidences $\boldsymbol{c}_i$, and 3D motion offsets $\Delta\boldsymbol{m}_i$ as inputs, the memory unit can track online. In each process, we have three stages.

i. Initialization. First, we discard predicted 3D positions $\boldsymbol{t}_i = (\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{z}_i)$ whose detection confidence is below a threshold λ_c. We observe that our input video is usually shot by tracking the subjects, therefore, we discard the detection whose $1/\boldsymbol{z}_i$ is below the scale threshold λ_s. To suppress duplicate detections, we find the detection pairs whose Euclidean distance is below a pre-defined threshold λ_d and discard the detection with lower detection confidence. In the first frame, we use the 3D positions $\boldsymbol{t}_1$ and detection confidences $\boldsymbol{c}_i$ of K subjects to initialize the K memory nodes.

ii. Memory node matching. The memory nodes store the 3D position $\boldsymbol{t}_{i-1}^k$ of subject k in previous frames. We match the memory node with the new filtered detections in the current frame. We select the predicted 3D translation $\boldsymbol{t}_i$ that best matches the 3D trajectory as the optimized 3D position $\hat{\boldsymbol{t}}_i$. Specifically, with filtered 3D positions $\boldsymbol{t}_i$ and 3D motion offset $\Delta\boldsymbol{m}_i$, we calculate the Euclidean distance matrix between $\boldsymbol{W}\boldsymbol{t}_{i-1}^k$ and $\boldsymbol{W}(\boldsymbol{t}_i - \Delta\boldsymbol{m}_i)$ where $\boldsymbol{W} \in \mathbb{R}^3$ is a distance weight vector. To avoid the effects of the nonlinear relationship between depth $\boldsymbol{z}_i$ and human scale in images, we convert the depth value z to $1/(1+z)$ when calculating the distance matrix. Using Hungarian matching, we keep the matched pairs whose matching distance is below a threshold λ_m and use them for memory update. We use ground truth tracks for training and only perform matching during inference.

iii. Memory update. We update the successfully matched memory nodes with the new 3D position and detection confidence. For the memory nodes without a matched detection, we accumulate the time since failure. Then we remove the memory nodes whose failure time is

above a threshold λ_f. Tracking can be done in two modes: on-line or off-line. The former does not allow looking back in time, while the latter does. In the off-line mode, if a new detection re-activates a non-matched memory node, the non-matched part of the 3D trajectory is linearly interpolated. Finally, the memory unit outputs the latest 3D positions $\widehat{\boldsymbol{t}_i}$ and tracking IDs $\boldsymbol{D}_i$ of all memory nodes.

3.4. DynaCam Dataset

Even with a powerful 5D representation, we still lack in-the-wild data for training and evaluation of global human trajectory estimation. However, collecting global human trajectory and camera poses for natural DC-videos is difficult. Therefore, we create a new dataset, DynaCam, by simulating camera motions to convert in-the-wild videos captured by static cameras to DC-videos.

We use over 1000 video clips captured by static regular cameras from the MPII Human Pose Database [3] as well as videos from the InterNet [2]. We also use over 200 panoramic video clips that are either recorded by us with an Insta360 RS panoramic camera or are downloaded from the InterNet [1]. We manually design the 3D rotation and field of view (FOV) of dynamic cameras to track the subjects in panoramic videos. With the designed camera motions, we can project the panoramic frames into perspective views. Also, to simulate the 3D translation of dynamic cameras, we crop the videos captured by static cameras with sliding windows. In this way, we can obtain abundant in-the-wild DC-videos with accurate camera pose annotations. Then we perform 2D human detection, tracking, and 2D pose estimation via YOLOX [17], ByteTrack [63], and ViT-Pose [54], respectively, to obtain 2D pose sequences of each subject. We estimate SMPL parameters by fitting the 2D poses using EFT [20] or ProHMR [25] and solve for their 3D positions in camera coordinates via the PnP algorithm (RANSAC [16]). Finally, we solve for the 3D human trajectories in the global coordinates with camera pose annotations. We manually filter out the failure cases. In this way, we generate more than 500 annotated DC-videos containing over 48K frames. More than half of video frames are generated from panoramic videos.

Limitations: The videos generated with our process only approximate real DC-videos shot in the wild since they lack perspective effects. Despite this they prove useful for training TRACE.

3.5. Loss Functions

TRACE is supervised by the weighted sum of 15 loss terms that fall into two groups: temporal motion losses and standard image losses. Here we focus on the novel temporal losses. Please refer to the Sup. Mat. for details of all losses.

To learn the temporal motion, we introduce a 3D motion offset loss $\mathcal{L}_m$ and a 6D world motion loss $\mathcal{L}_W$. $\mathcal{L}_m$ is the L_2 loss between the predicted 3D motion offset $\Delta \boldsymbol{m}_i$ and $(\boldsymbol{t}'_i - \boldsymbol{t}'_{i-1})$ where $\boldsymbol{t}'_i$ is the ground truth 3D human position at frame f_i in our pre-defined camera coordinates (FOV=50°), which is solved for via the PnP algorithm [16]. $\mathcal{L}_W$ consists of six parts, including an L_2 loss on the global 3D trajectory $\boldsymbol{T}$, an L_2 loss on the velocity/acceleration of 3D trajectory nodes $\dot{\boldsymbol{T}}/\ddot{\boldsymbol{T}}$, an L_2 loss of the velocity/acceleration of the 3D foot keypoints in global coordinates, and an L_2 loss on the global 3D body orientation $\boldsymbol{\tau}_i$.

4. Experiments

4.1. Implementation Details

Training details. During training, we directly use the ground truth trajectory of subjects to replace the estimated trajectory $\widehat{\boldsymbol{t}}$ for sampling the parameters. We use the pre-trained backbone of BEV [45] as the image backbone. We use RAFT [47] as the optical flow backbone. The training consists of two stages. In the first stage, we freeze the weights of the backbones and train the head network for 40 epochs with a learning rate of 5e-5. Then we train the image backbone and the head network together for 10 epochs with a learning rate of 1e-5. We use four V100-16GB GPUs for training. Limited by the GPU memory, we sample 4 video clips as a batch at each iteration; the clip length is 10 frames.

Training and evaluation datasets. For training, we use three 3D human pose datasets (Human3.6M [19], MPI-INF-3DHP [34, 35], and 3DPW [50]), two 2D human pose datasets (PennAction [62] and PoseTrack [12]), and our DynaCam dataset. We evaluate TRACE on two multi-person in-the-wild benchmarks, 3DPW [50] and MuPoTS-3D [35], and DynaCam. 3DPW videos are most consistent with our tracking scenario. Unfortunately, not all 3DPW videos have complete tracking annotations. We select the 16 videos that do and call this subset Dyna3DPW. We use this challenging subset to evaluate tracking and HPS accuracy in complex scenes with a moving camera.

Evaluation metrics. For global 3D trajectory estimation, we compute the average 3D position (**Traj**) and velocity (**Velocity**) error of the predicted global 3D trajectory $\boldsymbol{T}$. For multi-object tracking, we report the ID switch (**IDs**), Multi-Object Tracking Accuracy (**MOTA** [5]), Identification F1-score (**IDF1** [41]), and Higher Order Tracking Accuracy (**HOTA** [33]). To assess the accuracy of 3D human pose/shape estimation, we compute the Mean Per Joint Position Error (**MPJPE**), Procrustes-aligned MPJPE (**PMPJPE**), and Mean Vertex Error (**MVE**).

Please refer to Sup. Mat. for more details.

4.2. Comparisons to State-of-the-art Methods

Global 3D trajectory estimation. We aim to estimate the global human trajectory from dynamic cameras. We do not explicitly estimate the camera motion. Instead, we use a

Figure 3. Qualitative comparisons to SOTA methods (ByteTrack [63], PHALP [40], and GLAMR [58]) on 3DPW, MuPoTS-3D, and DynaCam. Red dash boxes highlight the ID switches. Yellow dot-dash boxes highlight the misses under occlusions.

Method	Translating Camera		Rotating Camera		FPS
	Traj↓	Velocity↓	Traj↓	Velocity↓	
BEV+DPVO	2.024	0.154	3.622	0.119	5.7
GLAMR [58]	1.745	0.105	1.759	0.055	1.5
TRACE	**1.433**	**0.082**	**1.038**	**0.045**	**21.2**

Table 1. Errors in estimated 3D human trajectory in world coordinates (in m) and runtime efficiency on the DynaCam dataset.

Methods	IDs↓	MOTA↑	IDF1↑	HOTA↑
Trackformer [36]	43	24.9	62.7	53.2
Tracktor [4]	53	51.5	70.9	50.3
FlowPose [51]	49	21.4	67.1	41.8
T3DP [39]	38	62.1	79.1	59.2
PHALP [40]	22	66.2	81.4	59.4
ByteTrack [63]	15	73.3	84.6	63.5
TRACE	**0**	**86.9**	**93.4**	**65.3**

Table 2. Tracking results on MuPoTS-3D [35]. Results of the compared methods [4, 36, 39, 40, 51] are from [40].

Methods	IDs↓	MOTA↑	IDF1↑	HOTA↑
PHALP [40]	5	97.9	93.7	74.2
YOLOX+ByteTrack	3	97.3	98.2	73.1
BEV+ByteTrack	37	93.6	79.1	59.3
TRACE w/o MO	4	95.8	97.3	72.7
TRACE	**1**	**99.3**	**99.7**	**74.7**

Table 3. Tracking results on Dyna3DPW [50]. w/o MO means ablating the estimated 3D motion offsets for tracking.

Method	PAMPJPE↓	MPJPE↓	PVE↓
HybrIK [27]	45.0	74.1	86.5
METRO [30]	47.9	77.1	88.2
GLAMR [58]	50.7	-	-
CLIFF [28]	43.0	**69.0**	**81.2**
D&D [26]	42.7	73.7	88.6
ROMP [44]	47.3	76.7	93.4
BEV [45]	46.9	78.5	92.3
TRACE	**37.8**	79.1	97.3

Table 4. Comparisons with SOTA methods for 3D human pose and shape estimation on the 3DPW test set.

world motion map to represent the global trajectory in world coordinates, which implies the camera motion. Therefore, we evaluate the global trajectory error, instead of the camera pose, in Tab. 3. First, we evaluate global 3D trajectory estimation on DynaCam. We compare TRACE with two baseline solutions. The first one uses BEV to estimate the subjects in camera coordinates and DPVO [48], a SLAM method, to estimate the camera and its motion; we call this BEV+DPVO. As shown in Tab. 1, TRACE significantly outperforms BEV+DPVO in the accuracy of global 3D trajectory estimation. The moving people in the scene make it hard for a SLAM method to extract stable corresponding keypoints. Additionally, our synthetic camera motions

differ from real camera motions and this may hurt DPVO's performance. A more direct comparison is to GLAMR [58]. TRACE outperforms GLAMR in both accuracy and efficiency. In Fig. 3(c), we also perform visual comparisons with GLAMR on DynaCam. These results demonstrate the benefit of estimating global human trajectory using a holistic 5D representation. We provide more results in the supplemental video.

Multi-subject tracking. To evaluate the performance of tracking subjects with dynamic cameras in real-world scenes, we compare TRACE with recent methods on Dyna3DPW. PHALP [40] is a recent (SOTA) method that uses 3D cues and appearance to track people using SMPL. YOLOX+ByteTrack [17, 63] is a recently proposed and popular tracking-by-detection solution. These methods are designed to track all the people in a scene. Therefore, we process their results to avoid them being penalized for tracking unlabeled passers-by; 3DPW has annotations for at most 2 people in a scene but some scenes contain many people. We first obtain their tracking results using their official code. We then select the tracking results that achieve maximum IoU with the labeled ground truth subjects; we use these tracks for evaluation. Note that, for a fair comparison with ByteTrack, TRACE runs in an on-line mode, without optimizing the past results. As shown in Tab. 3, TRACE outperforms PHALP, the previous 3D-representation-based method [40], and tracking-by-detection [17, 63] methods on Dyna3DPW. To evaluate the tracking robustness under long-term occlusion, we evaluate TRACE on MuPoTS-3D. Results of previous SOTA methods [4, 36, 39, 40, 51] are from [40]. Again, for a fair comparison, we filter out the unlabeled people in the ByteTrack results. As shown in Tab. 2, TRACE significantly outperforms previous SOTA methods. In particular, TRACE significantly reduces ID switches under long-term occlusions. These results illustrate the effectiveness and robustness of the proposed method for in-the-wild videos. The qualitative comparisons are presented in Fig. 3 and the supplemental video.

3D human pose and shape estimation. Finally, we evaluate 3D human regression performance in DC-video using the 3DPW test set. Because 3DPW does not provide ground-truth 3D translation in world coordinates, we evaluate root-relative 3D pose. We compare TRACE with the recent one-stage [44,45] and multi-stage [26–28,30,58] methods. Tab. 4 shows that TRACE significantly outperforms the previous method by 11.4% in terms of PMPJPE. These results demonstrate that learning a holistic 5D representation in an end-to-end manner is powerful. We adopt a fixed field of view (FOV=50°) to define our camera space, which is very different from that of 3DPW video. This results in a misalignment between the predicted 3D orientation of the body mesh and the ground truth, resulting in higher MPJPE and MVE. This illustrates the importance of also estimating the camera so that our estimated global coordinate system is consistent with the real world. This is future work.

4.3. Ablation Studies

Temporal 5D representation v.s. image-level 3D representation. We go beyond BEV's image-level 3D representation and build a temporal 5D representation. As shown in Tab. 1, 3, and 4, TRACE outperforms BEV or the multi-stage solutions using BEV on most metrics. This demonstrates the value of learning a holistic 5D representation.

3D Motion Offset map. We also evaluate the effect of using predicted 3D motion offsets for tracking. As shown in Tab. 3, 3D motion offsets improve performance by 3.5%, 2.4%, and 2.7% in terms of MOTA, IDF1, and HOTA.

5. Conclusions

Human pose and shape estimation is not an end to itself. Rather, estimating the 3D human in motion is useful for many tasks from behavior analysis to computer graphics. However, to be useful, it is important to know the motion of humans with respect to the 3D scene and other people. This means that HPS methods must estimate humans in a global coordinate system and provide consistent tracks of people across time. For generality, they also need to be able to do this from arbitrary moving cameras.

To tackle these challenging problems, we propose a novel 5D representation and a new neural architecture that reasons about people in 5D; that is, their 3D position, temporal trajectory, and identity. Moving to a 5D representation enables our method, TRACE, to take a holistic view of the video, processing full frames and incorporating temporal features. The core innovation of TRACE lies in its novel temporal representation in the form of new "maps" that represent the motion of people across time in the camera and global coordinates. These allow TRACE to be trained end-to-end, thus exploiting rich information from the video to solve the task. TRACE is the first such single-shot method for 3D HPS estimation from video and it achieves SOTA results on common benchmarks.

Future work should look at explicitly estimating the camera, using training data like BEDLAM [6], which contains complex human motion, 3D scenes, and camera motions. We believe that camera motion and human motion provide complimentary information that can be used to recover human motion in world coordinates with metric accuracy.

Acknowledgements: This work was partially supported by the National Key R&D Program of China under Grant (No. 2020AAA0103800) and Beijing Nova Program (No. 20220484063).

MJB Disclosure: https://files.is.tue.mpg.de/black/CoI_CVPR_2023.txt

References

[1] Bilibili. https://www.bilibili.com/. 6

[2] Pexels. https://www.pexels.com. 6

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014. 6

[4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *ICCV*, pages 941–951, 2019. 3, 7, 8

[5] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, pages 1–10, 2008. 6

[6] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of 3d human bodies exhibiting detailed lifelike animated motion. In *CVPR*, 2023. 2, 8

[7] 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*, pages 561–578, 2016. 2

[8] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. *arXiv:1905.00737*, 2019. 3

[9] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, pages 5386–5395, 2020. 4

[10] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *ECCV*, pages 436–454, 2020. 3

[11] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv:2003.09003*, 2020. 3

[12] Andreas Doering, Di Chen, Shanshan Zhang, Bernt Schiele, and Juergen Gall. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *CVPR*, pages 20963–20972, 2022. 6

[13] Mahsa Ehsanpour, Fatemeh Saleh, Silvio Savarese, Ian Reid, and Hamid Rezatofighi. Jrdb-act: A large-scale dataset for spatio-temporal action, social group and activity detection. In *CVPR*, pages 20983–20992, 2022. 3

[14] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, pages 1–8, 2008. 3

[15] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *CVPR*, pages 10849–10859, 2021. 3

[16] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3, 6

[17] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3, 6, 8

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 3

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

[20] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *ECCV*, 2020. 3, 6

[21] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 2, 3

[22] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 2, 3

[23] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *ICCV*, pages 11035–11045, 2021. 2, 3

[24] 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*, pages 2252–2261, 2019. 2, 3

[25] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *CVPR*, pages 11605–11614, 2021. 6

[26] Jiefeng Li, Siyuan Bian, Chao Xu, Gang Liu, Gang Yu, and Cewu Lu. D&d: Learning human dynamics from dynamic camera. In *ECCV*, 2022. 3, 7, 8

[27] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, pages 3383–3393, 2021. 7, 8

[28] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, pages 590–606, 2022. 3, 7, 8

[29] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-Fidelity Clothed Avatar Reconstruction from a Single Image. In *CVPR*, 2023. 3

[30] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 7, 8

[31] Miao Liu, Dexin Yang, Yan Zhang, Zhaopeng Cui, James M Rehg, and Siyu Tang. 4d human body capture from egocentric video via 3d scene grounding. In *3DV*, pages 930–939, 2021. 2, 3, 4

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *TOG*, 34(6):1–16, 2015. 3, 4, 5

[33] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, pages 548–578, 2021. 6

[34] 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

[35] 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*, pages 120–130, 2018. 3, 6, 7

[36] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, pages 8844–8854, 2022. 7, 8

[37] Gyeongsik Moon and Kyoung Mu Lee. Pose2Pose: 3d positional pose-guided 3d rotational pose prediction for expressive 3d human pose and mesh estimation. *arXiv*, 2020. 2, 3

[38] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *ICCV*, pages 803–812, 2019. 2, 3

[39] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3d representations. In *NeurIPS*, 2021. 3, 7, 8

[40] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people by predicting 3D appearance, location & pose. In *CVPR*, 2022. 2, 3, 7, 8

[41] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, pages 17–35, 2016. 6

[42] Weijian Ruan, Wu Liu, Qian Bao, Jun Chen, Yuhao Cheng, and Tao Mei. Poinet: pose-guided ovonic insight network for multi-person pose tracking. In *MM*, pages 284–292, 2019. 3

[43] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 3

[44] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, pages 11179–11188, 2021. 1, 3, 7, 8

[45] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, pages 13243–13252, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[46] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, pages 5348–5357, 2019. 2, 3

[47] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pages 402–419, 2020. 5, 6

[48] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *arXiv preprint arXiv:2208.04726*, 2022. 7

[49] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *CVPR*, 2023. 2

[50] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *ECCV*, pages 601–617, 2018. 3, 6, 7

[51] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018. 3, 7, 8

[52] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR*, 2023. 3

[53] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. ICON: Implicit Clothed humans Obtained from Normals. In *CVPR*, 2022. 3

[54] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *arXiv preprint arXiv:2204.12484*, 2022. 3, 6

[55] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *CVPR*, 2023. 3

[56] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *CVPR*, 2023. 2

[57] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *CVPR*, 2023. 2

[58] Ye Yuan, Umar Iqbal, Pavlo Molchanov, Kris Kitani, and Jan Kautz. GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras. In *CVPR*, 2022. 2, 3, 4, 7, 8

[59] 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*, pages 2148–2157, 2018. 3

[60] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, pages 11446–11456, 2021. 2, 3

[61] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*, 2017. 3

[62] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, pages 2248–2255, 2013. 6

[63] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, pages 1–21, 2022. 2, 3, 6, 7, 8

[64] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *ICCV*, pages 6239–6249, 2021. 3

[65] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 5