# TesseTrack: End-to-End Learnable Multi-Person Articulated 3D Pose Tracking

N Dinesh Reddy[1*] Laurent Guigues[2†] Leonid Pishchulin[2†] Jayan Eledath[2] Srinivasa G. Narasimhan[1]
[1]Carnegie Mellon University [2]Amazon

## Abstract

*We consider the task of 3D pose estimation and tracking of multiple people seen in an arbitrary number of camera feeds. We propose TesseTrack[1], a novel top-down approach that simultaneously reasons about multiple individuals' 3D body joint reconstructions and associations in space and time in a single end-to-end learnable framework. At the core of our approach is a novel spatio-temporal formulation that operates in a common voxelized feature space aggregated from single- or multiple camera views. After a person detection step, a 4D CNN produces short-term person-specific representations which are then linked across time by a differentiable matcher. The linked descriptions are then merged and deconvolved into 3D poses. This joint spatio-temporal formulation contrasts with previous piecewise strategies that treat 2D pose estimation, 2D-to-3D lifting, and 3D pose tracking as independent sub-problems that are error-prone when solved in isolation. Furthermore, unlike previous methods, TesseTrack is robust to changes in the number of camera views and achieves very good results even if a single view is available at inference time. Quantitative evaluation of 3D pose reconstruction accuracy on standard benchmarks shows significant improvements over the state of the art. Evaluation of multi-person articulated 3D pose tracking in our novel evaluation framework demonstrates the superiority of TesseTrack over strong baselines.*

## 1. Introduction

This paper addresses the problem of tracking and reconstructing in 3D articulated poses of multiple individuals seen in an arbitrary number of camera feeds. This task requires identifying the number of people in the scene, reconstructing their 3D body joints into consistent skeletons, and associating 3D body joints over time. We do not make any assumption on the number of available camera views and focus on real-world scenarios that often include multiple close-by interacting individuals, fast motions, self- and person-person occlusions. A key challenge in such scenarios is that people might strongly overlap and expose only a subset of body joints due to occlusions or truncations by image boundaries (*c.f*. Fig. 1), which makes it harder to reliably reconstruct and track articulated 3D human poses. Most multi-view strategies rely on multi-stage inference [9, 13, 20, 7, 8, 21, 15, 35] to first estimate 2D poses in each frame, cluster same person poses across views, reconstruct 3D poses from clusters based on triangulation, and finally link 3D poses over time [9, 8]. Solving each step in isolation is sub-optimal and prone to errors that cannot be recovered in later stages. This is even more true for monocular methods [4, 26, 33, 25, 42] where solving each step in isolation often represents an ill-posed problem.


**Figure 1:** We illustrate the output of Tessetrack on the Tagging sequence. The top two row potray the projections of keypoints on two views, while the bottom row shows the 3D pose tracking. Observe smooth tracking of people in the wild with moving cameras for long duration of time.

We propose TesseTrack, a top-down approach that simultaneously addresses 3D body joint reconstructions and associations in space and time of multiple persons. At the core of our approach is a novel spatio-temporal formulation that operates in a common voxelized feature space obtained by casting per-frame deep learning features from single or multiple views into a discretized 3D voxel volume. First, a 3D CNN is used to localize each person in

*Work done during DR internship at Amazon
†Equal Contribution
[1]Webpage can be found at http://www.cs.cmu.edu/~ILIM/projects/IM/TesseTrack/

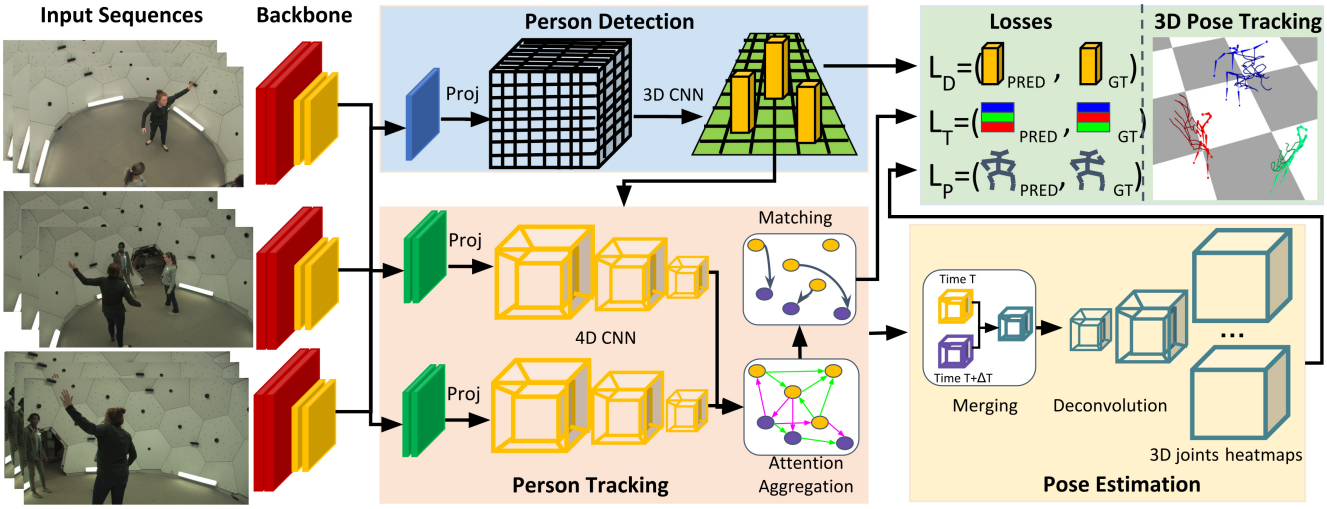


**Figure 2:** The complete pipeline of tessetrack has been illustrated. Initially, the video feed from multiple cameras is passed through shared HRNet to compute the features required for detection and 3D pose tracking. The final layer of the HRNet is passed through a 3D convolution to regress to the center of the human 3D bounding boxes. Each of the hypotheses is combined with the HRNet final layer to create a spatio-temporal Tube called tesseract. We use a learnable 3D tracking framework for a person association over time using spatio-temporal person descriptors. Finally, the associated descriptors are passed through deconvolution layers to infer the 3D pose. Note that the framework is end-to-end trainable except for the NMS layer in the detection network.

the voxel volume. Then, a fixed spatio-temporal volume around each person detection is processed by a 4D CNN to compute short-term person-specific representations. Overlapping representations at neighboring time steps are further scored based on attention aggregation and linked using a differentiable matcher. Finally, 3D body joints of the same person are consistently predicted at each time step based on merged person-specific representations. Notably, all components are implemented as layers in a single feed-forward neural network and are thus jointly learned end-to-end.

Our main contribution is a novel spatio-temporal formulation that allows simultaneous 3D body joint reconstruction and tracking of multiple individuals. In contrast to the multi-person 3D pose estimation approach of [46] who similarly aggregate per frame information in 3D voxel space, we address a more challenging problem of multi-person 3D pose tracking and propose end-to-end person-specific representation learning. TesseTrack does not make assumptions on the available number of camera views and performs reasonably well even in the purely monocular setting. Remarkably, using only a single view allows achieving similar MPJPE 3D joint localization error compared to the five-view setting of [46], while using the same five-view setting results in 2.4× reduction in MPJPE error (*c.f.* Sec. 4). In contrast to the multi-person 2D pose tracking method of [49] who rely on short-term spatio-temporal representation learning, our approach operates on the aggregated spatio-temporal voxel volume and provides a richer hypothesis comprising of tracked 3D skeletons.

Our second contribution is a novel learnable tracking formulation that allows extending person-specific spatio-temporal representation learning to arbitrary-long sequences. In contrast to [49] who use a heuristic pairwise tracking score based on pose distance and perform matching using the Hungarian method, we rely on an attention aggregation layer and a differentiable representation matching layer based on the Sinkhorn algorithm. Importantly, we match person-specific representations instead of the determined body pose tracklets, which allows to learn more expressive representations. In Sec. 4 we demonstrate that the proposed learnable tracking formulation not only improves tracking accuracy but also improves joint localization.

Our third contribution is a novel framework for the evaluation of multi-person articulated 3D pose tracking. Experimental evaluation on the Panoptic dataset [21] shows that TesseTrack achieves significant improvements in per-joint tracking accuracy compared to strong baselines.

Finally, our fourth contribution is an in-depth ablation study of the proposed approach and thorough comparisons to current methods on several standard benchmarks. In Sec. 4 we demonstrate that proposed design choices result in significant accuracy gains, thereby establishing a new state of the art on multiple datasets.

## 2. Related Work

**Single Person 3D Pose Estimation** methods can be subdivided into multi-view and monocular approaches. Multi-view approaches often rely on triangulation [18] of per view 2D poses to determine a 3D pose [9, 13, 21]. To improve robustness to 2D pose estimation errors, [1, 41] jointly reason over 2D poses seen from multiple viewpoints. Recent monocular approaches typically lean on powerful neural networks to mitigate the ambiguity of recovering 3D from

2D joint locations [34, 26, 41, 20, 37, 53, 11, 12]. [34, 26] directly regress 3D poses from 2D joint locations using deep networks. While being quite simple, they suffer from inaccuracies of 2D joint localization and the fact that appearance is not used during 3D pose prediction. [20, 37, 53, 17] intend to overcome these limitations by predicting a 3D volumetric representations from images: [17] augments 2D detection heatmaps with latent 3D pose features to predict 3D pose, [20] projects 2D feature maps to 3D volume and processes the volume to predict 3D joint locations. Similarly to [20, 37, 53, 17], we cast per-frame deep learning features from single or multiple views into a common discretized space. However, we address a more challenging problem of multi-person 3D pose tracking and process 4D spatio-temporal volumes to compute person-specific representations that allow to predict spatially and temporally consistent skeletons of multiple people. Our method is also related to [12, 11] who perform spatio-temporal representation learning optimized specifically for monocular case by introducing occlusion-aware training and spatio-temporal pose discriminator [11]. In contrast, our approach was not yet tuned to a monocular case and thus is expected to improve when using similar strategies.

**Multi-person 3D Pose Estimation** methods typically split the problem into 2D joint grouping in single frames and 3D pose reconstruction. 2D grouping is done using bottom-up [40, 10, 24, 36] or top-down [45, 50] strategies. In multi-view scenarios, recent approaches typically rely on triangulation of 2D poses of the same individual to reconstruct 3D poses [13, 15], while earlier methods extend pictorial structures model to deal with multiple views [6, 8, 7]. Independently solving 2D pose estimation, multi-view matching and triangulation are prone to errors. [46] project per view 2D joint heatmaps into a voxelized 3D space and directly detect people and predict their 3D poses in this space. Monocular approaches [30, 52] encode 2D and 3D pose features and jointly decode 3D poses of all individuals in the scene. Encoding the pose for all joints/limbs of the full-body, regardless of available image evidence, leads to potential encoding conflicts when similar body parts of different subjects overlap. Similar to [46] we cast per-frame feature maps into a voxelized 3D space and follow a top-down approach which starts with detecting people in this space. However, we address a more challenging problem of multi-person 3D pose tracking, which requires reasoning in spatio-temporal volumes extracted around person detections and merging extracted person-specific representations to reliably reconstruct and track 3D skeletons in arbitrarily long sequences. In contrast to [46] and similarly to [30, 52] our approach can operate in a purely monocular setting. However, unlike [30, 52] our approach does not suffer from encoding conflicts, since we cast feature maps into a common voxelized 3D space.

**Multi-person 3D Pose Tracking** was only addressed by few approaches [4, 9, 51, 29]. The multi-view approach of [9] follows a multi-stage inference where 2D poses are first predicted per frame, same person 2D poses are triangulated across views to recover 3D poses which are finally linked over time. In contrast, our formulation operates in a common spatio-temporal volume, is end-to-end learnable, and is not restricted to the multi-view setting only. An earlier monocular approach [4] relies on 2D tracking-by-detection and 2D-to-3D lifting to track 3D poses of walking pedestrians with a little degree of articulation. In contrast, we do make no assumptions about the type of body motions or people activities and address a harder problem of multi-person *articulated* 3D pose tracking. [51] compute per frame 2D and 3D pose and shape hypothesis and perform joint space-time optimization under scene constraints to reconstruct and track 3D poses. [29] encodes per frame 2D and 3D pose features and identities for all visible body joints of all people and employs a fully-connected deep network to decode features into complete 3D poses, followed by a spatio-temporal skeletal model fitting. In contrast, to [51, 29] who resort to a piece-wise trainable strategy, our approach is end-to-end trainable and thus can propagate people detection, tracking, and pose estimation errors back to input image pixels. Furthermore, our formulation seamlessly incorporates additional views, if available, to boost accuracy. We envision though that similar spatio-temporal model fitting strategies as in [51, 29] can be used to refine the output of our method.

## 3. TesseTrack: Multi-Person 3D Pose Tracking

To learn person tracking and pose estimation in 3D we build multiple differentiable layers with intermediate supervisions. Our network is made up of three main blocks, each one with an associated loss. The first block is a person detection network in 3D voxel space (3.1). Given person detections, a 4D CNN extracts a spatio-temporal representation of each detected person over a short period of time. In order to track people, we then solve an assignment problem between the set of descriptors for two frames $t$ and $t+\Delta t$ (3.2). All matched descriptors which overlap are then merged into a single descriptor which is finally deconvolved into a 3D pose for the person tracked at central frame (3.3).

### 3.1. Person Detection Network

Our approach starts with a multi-view person detection network (PDN) trained to detect people in 3D at a specific time instance. We use HRNet [45] as our backbone for extracting image-based features at each frame. We use the pre-final layer of the network and pass it through a single convolution layer to convert it into a feature map of size $R$. The feature maps coming from all the camera views are then aggregated into a 3D voxelized volume by an inverse

image projection method, similarly to [20], with the critical difference that we don't fuse the 2D joint heatmaps in 3D but the richer feature vectors picked from the pre-final layer of HRNet. The voxel grid is initialized to encompass the whole space observed by the cameras. Using the camera calibration data, each voxel center is projected into the camera views. We aggregate all the feature vectors picked in image space by concatenating them and passing through a shallow network with a softmax layer. This produces a unique feature vector of size $R$. We thus end up with a data structure of size $R \times W \times H \times D$ dimensions, where $W$, $H$, $D$ are the dimensions of the voxel grid and $R$ is the dimension of the feature maps. We then apply 3D Convolutions to this volume to generate detection proposals. For each person, we train the network to detect its "center", which is defined as the midpoint between neck and center of the hips. The loss at each time $t$ is expressed directly as a distance between the expected heatmap and the output heatmap, similarly to the CenterNet approach [14], except that our framework is in 3D instead of 2D:

$$L_D^t = \sum_{w=1}^{W} \sum_{h=1}^{H} \sum_{d=1}^{D} ||V_{Pred}^{w,h,d} - V_{GT}^{w,h,d}|| \qquad (1)$$

We apply non-maximum suppression (NMS) on the 3D heatmaps and only retain the detections with large score.

### 3.2. Spatio-Temporal Descriptors and Tracking

For each detected person we create a spatio-temporal volume of fixed dimension centered on the person and use a 4D CNN to produce a short time description of the person around the detection frame. We call this spatio-temporal volume a *tesseract* as it is a 4D volume of size R×T×X×Y×Z, where $T$ represents temporal window size and $X$,$Y$,$Z$ are the dimensions of the cuboid centered on the detected person. The goal of extending the volume in time around the detection frame is twofold. First, using a temporal context allows to better estimate the joint positions in the central frame, and especially to extrapolate/interpolate occluded joints or to handle pose or appearance ambiguities in a single frame. Second, extending a person's description in time generates a descriptor which overlaps with adjacent frames, hence producing descriptors that can be matched by similarity for tracking purposes.

**Tesseract Convolutions.** The input to this sub-network is still the output of the HRNet pre-final layer which is cast in 3D at each time stamp. We follow the same procedure as for the person detection network to generate the features for each time instance of the tesseract. The tesseract is then passed through multiple 4D convolutions and max pooling layers to produce a reduced size tesseract feature. These features represent a spatio-temporal descriptor of a person centered around a detection. This bottleneck descriptor is used in both the tracking and pose estimation modules.

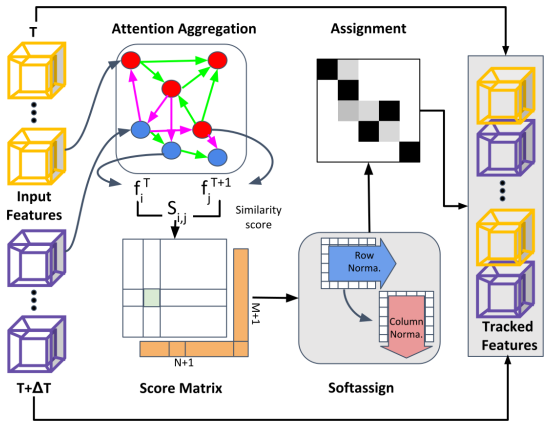


**Figure 3:** The learnable tracking framework. The input is the tesseract features for multiple detected humans at two different time instances. The output is an assignment matrix providing the correspondence between the detected persons at different times.

**Attention Aggregation.** Before temporal matching, as illustrated in Fig 3, we pass the features into a Graph Neural Network to integrate contextual cues and improve the features distinctiveness. We use two types of undirected edges: self edges, connecting features belonging to the same time instance and cross edges, connecting features from adjacent time instances. We use a learnable message passing formulation to propagate the information in the graph. The resulting multiplex network starts with a high-dimensional state for each node and computes at each layer an updated representation by simultaneously aggregating messages across all incident edges for all nodes.

Let ${}^{(l)}\mathrm{x}_i^t$ be the intermediate representation for element $i$ at time instance $t$ at layer $l$. The message $m_{\epsilon \to i}$ is the result of the aggregation from all features of persons $j : (i,j) \in \epsilon$, where $\epsilon \in \epsilon_{self}, \epsilon_{cross}$. Following [43, 5, 47] we pass the input through multiple message passing updates to get a final matching descriptors given as linear projections. They are given as $f_i^t = W_{\cdot}^{(L)} \mathrm{x}_i^t + b_{\cdot}$ for features at time $t$ and $f_i^{(t+\Delta t)} = W_{\cdot}^{(L)} \mathrm{x}_i^{t+\Delta t} + b_{\cdot}$ at time $t + \Delta t$, where $W$ are the weights learned for the GNN.

**Temporal Matching Layer.** The final features of the attention module are passed through a trained matching layer, which produces an assignment matrix. For a given time instance $t$, we consider the features of $N$ and $M$ persons at time $t$ and $t + \Delta t$ respectively. As in the standard bipartite graph matching formulation, an optimal assignment $P$ is a permutation matrix which maximizes the total score $\sum_{i,j} S_{i,j} P_{i,j}$ where $S \in R^{M \times N}$ is a score matrix. We compute the similarity $S_{i,j}$ between the descriptor $i$ at time $t$ and the descriptor $j$ at time $t + \Delta t$ using the inner product between descriptors $S_{i,j} =< f_i^t, f_j^{(t+\Delta t)} >$. As opposed to learned visual descriptors, the matching descriptors are not normalized, and their magnitude can change as per the feature during training to reflect the prediction confidence.

To let the network suppress some predicted persons (false detections) and to handle changes in the number of persons in the scene, we augment each set with a dustbin so that matching is always computed on a fixed length feature vectors. This leads to optimal assignments for each available detection and the rest unassigned dustbins always correspond one-to-one with the next time instance. Following recent end-to-end learning approaches which include an optimal assignment step, such as [31, 43], we use the Softassign algorithm [16] to solve the assignment problem by a differentiable operator. The Softassign algorithm is based on Sinkhorn iterative matrix balancing, which projects an initial score matrix into a doubly stochastic matrix by iteratively normalizing the matrix along rows and columns. When applied to the matrix $exp(S^{-}/\tau)$, it has been shown that Sinkhorn balancing corresponds to solving an entropy regularized problem which converges to the optimal assignment solution as $\tau$ goes to 0 [31]. The Softassign algorithm can be efficiently implemented on GPU by unrolling a fixed number of Sinkhorn iterations. After $T = 100$ iterations, we get a final score matrix $P$ and the association for the detection $i$ at time $t$ is then extracted as $\arg\max_j P_{i,j}$.

Since all of the above layers are differentiable, we can train the tracking module in a supervised manner with respect to the ground truth. Given ground truth associations $G$ between time $t$ and $t + \Delta t$, the objective function to be minimized is the log likelihood of the assignment $P$:

$$L_T^t = - \sum_{(i,j) \in G} \log P_{i,j} \tag{2}$$

### 3.3. 3D Pose Estimation

The last module of the network computes the persons' 3d poses using the persons descriptors and their tracking.

**Spatio-temporal descriptors merging.** If $T$ is the tesseract temporal window size, then after tracking a person for $T$ frames, we obtain $T$ spatio-temporal descriptors of this person which overlap at a common time and encode the person's pose and motion over a total time interval of length $2T - 1$. We thus merge all these descriptors to estimate the person's pose at their common time. As previously, we use a softmax-based merging strategy and the result is a single tesseract description for the central frame.

**Tesseract deconvolution.** The merged tesseract is finally passed through multiple 4D deconvolution layers to produce 3D heatmaps of person's joints at time $t$. If $T^q_{Pred}$ denotes the 3D heatmap obtained for the joint $q$, the predicted joint position $k^q_{Pred}$ is obtained by a soft-argmax operator, i.e. by a heatmap scores-weighted average of the voxel centers.

Similar to [20], we then combine two loss functions for the pose estimation task: a L1 distance computed on the keypoints positions and a loss on the response of the heatmap at the ground truth joint position:

$$L_P^{t,d} = \sum_{q=1}^{Q} \left[ ||k^q_{Pred} - k^q_{GT}||_1 - \beta . \log(T^q_{Pred}(k^q_{GT})) \right], \tag{3}$$

where $Q$ is the number of joints. In the end, we train our network end-to-end to minimize the sum of the three losses defined above over time, the person detection loss $L^t_D$, the tracking loss $L^t_T$ and the pose estimation loss $L^{t,p}_P$:

$$L = \sum_{t \in D} \left[ L^t_D + \alpha L^t_T + \gamma \sum_{p \in TP(t)} L^{t,p}_P \right], \tag{4}$$

where $D$ is the total duration of the sequence and $TP(t)$ represent the true positive detections at time $t$. The gradient is propagated back to the initial images, including through the HRNet backbone which is shared by the detection module and the tracking + pose estimation modules.

## 4. Experiments

### 4.1. Datasets and Metrics

We selected the following standard 3D human pose estimation datasets for experimental evaluation. All datasets provide calibrated camera poses.

**Human3.6M [19]** was captured from 4 cameras with a single human performing multiple actions. The dataset contains 8 actors performing 16 actions captured in controlled indoor settings. Motion capture was used to create ground truth 3D poses. We use 6 sequences to train and 2 sequences (S09, S11) to test our algorithm.

**TUM Shelf [6]** was captured indoors using 5 stationary cameras, with 4 people disassembling a shelf. The dataset provides sparse 3D pose annotations. Severe occlusions and random motion of the persons are the key challenges.

**TUM Campus [6]** was captured outdoors using 3 stationary cameras, with 3 people interacting on campus grounds. Similar to *Shelf*, it provides sparse 3D pose annotations. The dataset is challenging for 3D pose estimation due to a small number of cameras and wide baseline views.

**CMU Panoptic [21]** was built to understand human interactions in 3D. It contains 60 hours of data with 3D poses and tracking information captured by 500 cameras. We follow [46] and sample the same 5 cameras for evaluation, and use the same sequences for training. We split the training and testing sequences following [22].

**Tagging [48]** was captured in unconstrained environments where people are interacting in a social setting. There are no constraints on the motion of the cameras or the number of persons during the capture. This "in the wild" setting makes this dataset particularly interesting for 3D pose tracking. However, since no GT pose annotations are available, we only use this dataset for qualitative evaluation.

**Evaluation details.** Mean Per Joint Position Error (MPJPE) [44] evaluates 3D joint localization accuracy in mm and represents L2 distance between the GT and predicted joint locations. Percentage of Correct Keypoints (3D-PCK) [13] provides a more global view on the accuracy of 3D pose estimation and is computed similarly to its 2D PCK counterpart [3]. On Human3.6M we follow [20] and provide all comparisons using root-centered MPJPE metric. On Panoptic dataset, we follow [46] and provide all comparisons using non-root-centered MPJPE.

**Implementation Details.** We train TesseTrack on 8 $V100$ GPUs with 32 GB memory each. As model does not fit into a single GPU, we share the *tesseract* convolutions and the backbone across 2 GPUs. Each GPU has propagation weights of a single time instance. The tracking and the deconvolution modules are shared among both GPUs. During testing, the model can be computed on a single GPU using sequential processing. A learning rate of $0.01$ is used for all the modules. The Temporal Window ($T$) and the step size ($\Delta t$) used across the experiments is $5$ unless specified. The module was trained with $Q = 19$ keypoints with the voxel volumes size $64$. For all indoor experiments (*Panoptic*, *Human3.6M* and *Shelf*) we use a voxel volume of $12$m and for outdoor experiments (*Campus*, *Tagging*) the size is $50$m. For the tesseract a fixed volume size of $2.5$m is used across all datasets. We use panoptic [21] keypoint format in all the experiments except for *Human3.6M* evaluation. As *Shelf*, *Campus* and *Tagging* datasets have no training GT annotations we use multi-view triangulation to obtain auto-annotated 3D labels to finetune PDN module only. We use HRNet [45] for feature extraction with $R = 32$ and $\alpha = 1$ , $\beta = \gamma = 0.01$ in all experiments.

**TesseTrack variants**. We consider possible design choices for TesseTrack components: *F* - casting backbone's pre-final layer features into the voxelized space, *H* - using 2D joint detection heatmaps instead [46]; *T* - prediction using *tesseract* spatio-temporal module, *I* - instantaneous prediction per time instance instead; *D* - tracking using learned matcher, *G* - using heuristic matching using the Hungarian algorithm instead [49]; *L* - learned descriptor merging, *A* - simple heatmaps averaging instead [49]. This results into six TesseTrack variants: *HI*, *FI*, *FT*, *FTGA*, *FTGL*, *FTDL*. We also consider a simple tracking baseline that performs instantaneous prediction followed by the Hungarian matching of poses across time, which we denote as *FIG*.

## 4.2. Multi-Person 3D Pose Estimation

In this section, we evaluate TesseTrack on the task of multi-person 3D pose estimation. First, we demonstrate the improvements due to various design choices and show the robustness of TesseTrack to the number of available camera views on the *Panoptic* dataset. Then, we compare to the state of the art on *Panoptic*, *Shelf* and *Campus* datasets.

| Model | HI | FI | FT | FTGA | FTGL | FTDL |
|---|---|---|---|---|---|---|
| MPJPE (mm) | 16.3 | 13.8 | 8.0 | 8.1 | 7.5 | **7.3** |

**Table 1:** Ablation study of 3D pose reconstruction on the Panoptic dataset using non-root-centered MPJPE. We observe a clear increase in reconstruction accuracy with each additional improvement added to the model. Using the final layer of the backbone with a spatio-temporal descriptor-based network and learned matching and merging (FTDL) provides the best results in 3D reconstruction.

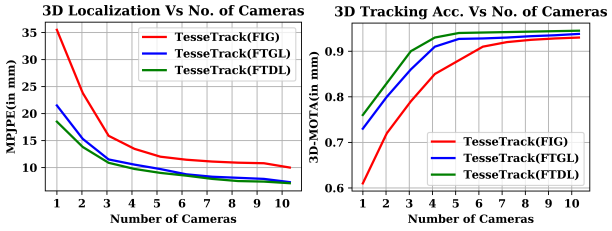


**Figure 4:** Impact of number of cameras on body joint localization error (MPJPE) (left) and pose tracking accuracy (3D MOTA) (right). Tessetrack (FTDL) shows the greatest advantage with lower number of cameras.

**Ablation analysis on Panoptic dataset.** MPJPE metric is used for comparison. Results are shown in Tab. 1.

*FI vs. HI.* We observe an improvement in reconstruction accuracy when using backbone features. This is because 2D heatmaps learned from 2D pose supervision might be missing out on crucial information required for accurate 3D joint reconstruction.

*FT vs. FI.* Most of the state-of-the-art methods use instantaneous 3D pose estimation and might struggle due to a lack of consistency of keypoints over time. TesseTrack enforces smoothness of the keypoints showing a clear improvement in 3D pose reconstruction.

*FTGL vs. FTGA.* Corresponding the human poses across time instances and merging them is generally a neglected problem. Most of the methods just average joint locations from different time instance inferences. We observe that relying on a learned merging framework at the descriptor level improves accuracy.

*FTDL vs. FTGL.* Differentiable matching module learns person-specific representations that are essential for reliable tracking. As expected, it improves over heuristic matching based on the Hungarian algorithm.

**Impact of Temporal Volumes.**Tessetrack can operate without temporal information, which leads to $-5.8$ mm MPJPE loss on Panoptic dataset (*c.f*. *FI* vs. *FT* in Tab.6).

**Robustness to number of cameras.** We evaluate the robustness of the best found *FTDL* architecture to the number of available camera views. To that end, we vary the number of cameras available at each time instance from one (monocular) to ten. Results are shown in Fig. 4 (left). First, we observe that *FTDL* can achieve a reasonable accuracy of $18.9$mm in the pure monocular scenario, although it was not specifically tuned for this setting. Intuitively, increasing the number of camera views results in a clear improvement in

| | Multi-View (5 views) | | Monocular | |
|---|---|---|---|---|
| Method | Tu et al. [46] | TesseTrack | Tu et al. [46] | TesseTrack |
| MPJPE (mm) | 17.7 | **7.3** | 51.1 | **18.9** |

**Table 2:** Comparison to the state of the art on the Panoptic dataset in multi-view and monocular settings. We show substantial improvement in reconstruction compared to the baseline method due to temporal consistency and end-to-end learnable framework.

| Method | Actor-1 | Actor-2 | Actor-3 | Total |
|---|---|---|---|---|
| Belagiannis et. [7] | 93.5 | 75.7 | 84.4 | 84.5 |
| Ershadi et. [15] | 94.2 | 92.9 | 84.6 | 90.6 |
| Dong et. [13] | 97.6 | 93.3 | 98.0 | 96.3 |
| Tu et al. [46] | 97.6 | 93.8 | 98.8 | 96.7 |
| TesseTrack | **97.9** | **95.2** | **99.1** | **97.4** |

**Table 3:** Evaluation of 3D-PCK accuracy on the Campus dataset. TesseTrack ourperforms baselines due to the temporal consistency constraints.

| Method | Actor-1 | Actor-2 | Actor-3 | Total |
|---|---|---|---|---|
| Belagiannis et. [7] | 75.3 | 69.7 | 87.6 | 77.5 |
| Ershadi et. [15] | 93.3 | 75.9 | 94.8 | 88.0 |
| Dong et. [13] | 98.8 | 94.1 | 97.8 | 96.9 |
| Tu et al. [46] | **99.3** | 94.1 | 97.6 | 97.0 |
| TesseTrack | 99.1 | **96.3** | **98.3** | **98.2** |

**Table 4:** Evaluation of 3D-PCK accuracy on the Shelf dataset. TesseTrack ourperforms baselines even in severe occlusions of the Shelf dataset.

joint localization accuracy. Compared to *FTGL* we observe noticeable improvements for fewer cameras, which underlines the advantages of differentiable matching. Compared to *FIG*, both *FTGL* and *FTDL* achieve dramatic improvements in localization accuracy, which demonstrates the importance of incorporating temporal information.

**Comparison to the State of the Art on Panoptic dataset.** We compare *FTDL* to the state-of-the-art approach of [46] in Tab. 2. TesseTrack achieves $2.4\times$ reduction in MPJPE in multi-view setting, and $2.7\times$ reduction in monocular scenario, which clearly shows the advantages of the proposed spatio-temporal formulation over [46].

**Comparison to the State of the Art on TUM datasets.** We use 3D-PCK metric and compare on TUM Campus in Tab. 3 and on TUM Shelf in Tab. 4. *FTDL* achieves significant improvements over the state of the art on both datasets.

### 4.3. Multi-Person Articulated 3D Pose Tracking

Most recent works on multi-person articulated 3D pose tracking [9, 51, 29] focus on evaluation of 3D pose reconstruction accuracy using MPJPE [44] or 3D-PCK [28]. However, this is not clear how existing methods advance actual body joint tracking accuracy in multi-person scenarios. We thus intend to fill in this gap and propose a set of novel evaluation metrics for multi-person articulated 3D pose tracking. To that end, we build on the popular Multiple Object Tracking (MOT) [32] and articulated 2D pose tracking metrics [2] and extend them to the 3D pose use case. The proposed metrics require predicted 3D body poses with track IDs. First, for each pair of (predicted pose, GT pose) 3D-PCK is computed. Predicted and GT poses are matched to each other by a global matching procedure that maximizes per pose 3D-PCK. Finally, Multiple Object Tracker Accuracy (MOTA), Multiple Object Tracker Precision (MOTP), Precision, and Recall metrics are computed.

| Method | Neck | Head | Shou. | Elbow | Wrist | Hip | Knee | Ankle | Avg |
|---|---|---|---|---|---|---|---|---|---|
| FIG | 89.7 | 87.4 | 90.8 | 88.0 | 82.2 | 92.7 | 89.1 | 92.4 | 87.6 |
| FTGL | 93.9 | 91.7 | 93.0 | 92.1 | 87.4 | 94.4 | **93.9** | 94.6 | 92.1 |
| FTDL | **94.6** | **93.6** | **93.4** | **92.7** | **88.2** | **94.7** | 93.8 | **95.0** | **94.1** |

**Table 5:** 3D MOTA evaluations on the Panoptic dataset. Using an end-to-end learnable framework (*FTDL*) systematically improves the accuracy of 3D pose tracking across all keypoints.

**Evaluation details.** Evaluation is performed on the Panoptic dataset using the proposed 3D MOTA metric. In the following we compare *FTDL* to *FTGL* and *FIG*.

**Impact of temporal representations on tracking.** Results are shown in Tab. 5. Using temporal person descriptors (*FTDL* and *FTGL*) significantly improves tracking accuracy compared to instantaneous person descriptor (*FIG*). Using a end-to-end learnable tracking framework (*FTDL*) instead of a Hungarian matching algorithm (*FTGL*) further improves tracking accuracy. This can be attributed to the fact that the learnable descriptors matching can distinguish interacting people much better than graph-based tracking methods.

**Robustness to number of cameras.** We analyze the accuracy of 3D pose tracking with respect to a varying number of cameras. Results are shown in Fig. 4 (right). While an increasing number of cameras allows improving the accuracy of all variants, we observe that relying on spatio-temporal representation learning results in significant tracking accuracy improvements specifically in the few cameras mode (*FTDL* and *FTGL* vs. *FIG*). Furthermore, using a learnable tracklet matcher (*FTDL*) results in consistent increase in tracking accuracy over a wide range of number camera views. Both observations underline the advantages of the proposed formulation when only a few cameras are available. Finally, in the pure monocular setting, *FTDL* achieves a reasonable 76% 3D MOTA accuracy, despite not being specifically tuned in this setting. We envision that incorporating scene constraints and performing spatio-temporal articulated model fitting [51, 29] should significantly boost the accuracy of TesseTrack in monocular setting.

### 4.4. Single Person 3D Pose Estimation

We compare to the state-of-the-art methods on Human 3.6M using the MPJPE metric under *Protocol #1*.

**Multi-View scenario.** Comparison to multi-view approaches is shown in Tab. 6 (bottom). TesseTrack clearly improves over the state of the art, which underlines the advantages of the proposed spatio-temporal formulation. Specifically, using temporal consistency improves the joint localization accuracy for ambiguous poses like sitting down and walking a dog. We conclude that temporal constraints

| *Protocol #1* | Dir | Disc | Eat | Greet | Phone | Photo | Pose | Purch | Sit | SitD | Smok | Wait | Walk | WalkD | WalkT | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Monocular methods, (MPJPE, mm) | | | | | | | | | | | | | | | | |
| Martinez et al. [27] | 51.8 | 56.2 | 58.1 | 59.0 | 69.5 | 78.4 | 55.2 | 58.1 | 74.0 | 94.6 | 62.3 | 59.1 | 65.1 | 49.5 | 52.4 | 62.9 |
| Iskakov et al. (monocular) [20] | 41.9 | 49.2 | 46.9 | 47.6 | 50.7 | 57.9 | 41.2 | 50.9 | 57.3 | 74.9 | 48.6 | 44.3 | 41.3 | 52.8 | 42.7 | 49.9 |
| Pavllo et al. [39] | 45.2 | 46.7 | 43.3 | 45.6 | 48.1 | 55.1 | 44.6 | 44.3 | 57.3 | 65.8 | 47.1 | 44.0 | 49.0 | 32.8 | 33.9 | 46.8 |
| Cheng et al. [12] | 38.3 | 41.3 | 46.1 | 40.1 | 41.6 | 51.9 | 41.8 | 40.9 | 51.5 | 58.4 | 42.2 | 44.6 | **41.7** | 33.7 | 30.1 | 42.9 |
| Cheng et al. [11] | **36.2** | **38.1** | **42.7** | **35.9** | **38.2** | **45.7** | **36.8** | 42.0 | 45.9 | **51.3** | **41.8** | **41.5** | 43.8 | **33.1** | **28.6** | **40.1** |
| TesseTrack | 38.4 | 46.2 | 44.3 | 43.2 | 44.8 | 48.3 | 52.9 | **36.7** | **45.3** | 54.5 | 63.4 | 44.4 | 41.9 | 46.2 | 39.9 | 44.6 |
| Multi-view methods, (MPJPE, mm) | | | | | | | | | | | | | | | | |
| Martinez et al. (multi-view) [27] | 46.5 | 48.6 | 54.0 | 51.5 | 67.5 | 70.7 | 48.5 | 49.1 | 69.8 | 79.4 | 57.8 | 53.1 | 56.7 | 42.2 | 45.4 | 57.0 |
| Pavlakos et al. [38] | 41.2 | 49.2 | 42.8 | 43.4 | 55.6 | 46.9 | 40.3 | 63.7 | 97.6 | 119.0 | 52.1 | 42.7 | 51.9 | 41.8 | 39.4 | 56.9 |
| Kadkhodamohammadi & Padoy [23] | 39.4 | 46.9 | 41.0 | 42.7 | 53.6 | 54.8 | 41.4 | 50.0 | 59.9 | 78.8 | 49.8 | 46.2 | 51.1 | 40.5 | 41.0 | 49.1 |
| Iskakov et al. [20] | 19.9 | 20.0 | 18.9 | 18.5 | 20.5 | 19.4 | 18.4 | 22.1 | 22.5 | 28.7 | 21.2 | 20.8 | 19.7 | 22.1 | 20.2 | 20.8 |
| TesseTrack (FI) | 18.0 | 19.8 | 19.9 | 19.0 | 20.1 | 17.6 | 21.1 | 23.7 | 26.8 | 20.6 | 20.0 | 19.5 | 19.2 | 21.7 | 18.6 | 20.4 |
| TesseTrack | **17.5** | **19.6** | **17.2** | **18.3** | **18.2** | **17.7** | **18.0** | **18.0** | **20.5** | **20.3** | **19.4** | **17.2** | **18.9** | **19.0** | **17.8** | **18.7** |

**Table 6:** 3D pose reconstruction accuracy of different methods on the Human3.6M dataset using root-centered MPJPE metric and *Protocol #1* from [20].

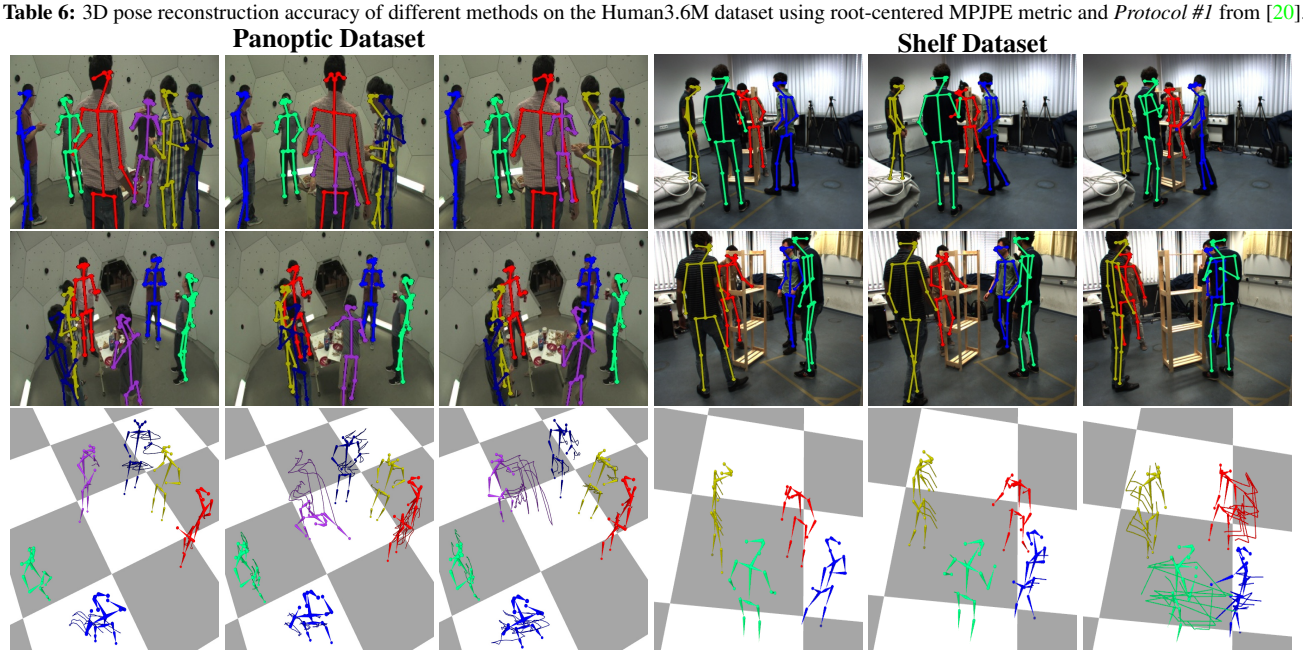


**Figure 5:** Qualitative results on Panoptic and Shelf datasets. TesseTrack can track people in the wild as well as when interacting in close proximity.

boost reconstruction accuracy in challenging actions.

**Monocular scenario.** Comparison to monocular methods is shown in Tab. 6(top). Despite not being specifically tuned for the monocular scenario, TesseTrack without bells and whistles outperforms most of the monocular approaches [12, 11]. Both [12, 11] also rely on spatio-temporal representation learning, but introduce occlusion-aware training which proved to be very useful specifically in monocular case, while [11] further reduce the error by adding a spatio-temporal discriminator to verify pose plausibility. Both improvements are orthogonal to our approach and thus can be incorporated to improve monocular case.

## 5. Conclusion

Reliably reconstructing and tracking the 3D poses of multiple persons in real-world scenarios using calibrated cameras is a challenging problem. In this work, we address it by proposing a novel formulation, TesseTrack, which jointly solves the tasks of tracking and 3D pose reconstruction within a single end-to-end learnable framework. In contrast to previous piece-wise strategies which first reconstruct 3D poses based on geometrical optimization algorithms and then subsequently linking the poses over time, TesseTrack infers the number of persons in a scene and jointly reconstructs and tracks their 3D poses using a novel 4D spatio-temporal CNN and a learnable tracking framework using differentiable matching. Experimental evaluation on five challenging datasets show significant improvements not only in multi-person 3D pose tracking but also in multi-person 3D pose reconstruction accuracy.

## 6. Acknowledgments

This paper was supported in parts by NSF Grants IIS-1900821 and CNS-2038612, DOT RITA Mobility-21 Grant 69A3551747111, and a PhD fellowship from Amazon.

## References


[1] Sikandar Amin, Mykhaylo Andriluka, Marcus Rohrbach, and Bernt Schiele. Multi-view pictorial structures for 3d human pose estimation. In *British Machine Vision Conference, BMVC*, 2013. 2

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 7

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 6

[4] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3d pose estimation and tracking by detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010. 1, 3

[5] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 4

[6] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures for multiple human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 3, 5

[7] Vasileios Belagiannis, Sikandar Amin, Mykhaylo Andriluka, Bernt Schiele, Nassir Navab, and Slobodan Ilic. 3d pictorial structures revisited: Multiple human pose estimation. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):1929–1942, 2015. 1, 3, 7

[8] Vasileios Belagiannis, Xinchao Wang, Bernt Schiele, Pascal Fua, Slobodan Ilic, and Nassir Navab. Multiple human pose estimation with temporally consistent 3d pictorial structures. In *ECCVw*, 2014. 1, 3

[9] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton. Multi-person 3d pose estimation and tracking in sports. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 1, 2, 3, 7

[10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[11] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. *arXiv preprint arXiv:2004.11822*, 2020. 3, 8

[12] Yu Cheng, Bo Yang, Bo Wang, Yan Wending, and Robby Tan. Occlusion-aware networks for 3d human pose estimation in video. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 723–732. IEEE. 3, 8

[13] Junting Dong, Wen Jiang, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation from multiple views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7792–7801, 2019. 1, 2, 3, 6, 7

[14] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019. 4

[15] Sara Ershadi-Nasab, Erfan Noury, Shohreh Kasaei, and Esmaeil Sanaei. Multiple human 3d pose estimation from multiview images. *Multimedia Tools and Applications*, 77(12):15573–15601, 2018. 1, 3, 7

[16] Steven Gold, Anand Rangarajan, et al. Softmax to softassign: Neural network algorithms for combinatorial optimization. *Journal of Artificial Neural Networks*, 2(4):381–399, 1996. 5

[17] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[18] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[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. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 5

[20] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *International Conference on Computer Vision (ICCV)*, October 2019. 1, 3, 4, 5, 6, 8

[21] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 1, 2, 5, 6

[22] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10873–10883, 2019. 5

[23] Abdolrahim Kadkhodamohammadi and Nicolas Padoy. A generalizable approach for multi-view 3d human pose regression. *Machine Vision and Applications*, 32(1):1–14, 2020. 8

[24] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[25] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

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

[27] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 8

[28] 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 *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 7

[29] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. XNect: Real-time multi-person 3D motion capture with a single RGB camera. volume 39, 2020. 3, 7

[30] 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 *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE, sep 2018. 3

[31] Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. *arXiv preprint arXiv:1802.08665*, 2018. 5

[32] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, 2016. 7

[33] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *The IEEE Conference on International Conference on Computer Vision (ICCV)*, 2019. 1

[34] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[35] Minh Vo N Dinesh Reddy and Srinivasa G. Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicle. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018*. IEEE, June 2018. 1

[36] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, 2017. 3

[37] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[38] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6988–6997, 2017. 8

[39] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 8

[40] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V. Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[41] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[42] N. Dinesh Reddy, Minh Vo, and Srinivasa G. Narasimhan. Occlusion-net: 2d/3d occluded keypoint localization using graph networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7326–7335, 2019. 1

[43] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020. 4, 5

[44] L. Sigal, A. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87(1):4–27, Mar. 2010. 6, 7

[45] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3, 6

[46] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. ECCV, 2020. 2, 3, 5, 6, 7

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[48] Minh Vo, Ersin Yumer, Kalyan Sunkavalli, Sunil Hadap, Yaser Sheikh, and Srinivasa Narasimhan. Automatic adaptation of person association for multiview tracking in group activities. *TPAMI*, 2020. 5

[49] Manchen Wang, Joseph Tighe, and Davide Modolo. Combining detection and tracking for human pose estimation in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 6

[50] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *arXiv preprint arXiv:1804.06208*, 2018. 3

[51] 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. 2018. 3, 7

[52] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 8410–8419. Curran Associates, Inc., 2018. 3

[53] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3