Ray3D: ray-based 3D human pose estimation for monocular absolute 3D localization

Yu Zhan
Aibee Inc.

Fenghai Li
Beijing Technology and Business University

Renliang Weng
Aibee Inc.

Wongun Choi
Aibee Inc.

Abstract

In this paper, we propose a novel monocular ray-based 3D (Ray3D) absolute human pose estimation with calibrated camera. Accurate and generalizable absolute 3D human pose estimation from monocular 2D pose input is an ill-posed problem. To address this challenge, we convert the input from pixel space to 3D normalized rays. This conversion makes our approach robust to camera intrinsic parameter changes. To deal with the in-the-wild camera extrinsic parameter variations, Ray3D explicitly takes the camera extrinsic parameters as an input and jointly models the distribution between the 3D pose rays and camera extrinsic parameters. This novel network design is the key to the outstanding generalizability of Ray3D approach. To have a comprehensive understanding of how the camera intrinsic and extrinsic parameter variations affect the accuracy of absolute 3D key-point localization, we conduct in-depth systematic experiments on three single person 3D benchmarks as well as one synthetic benchmark. These experiments demonstrate that our method significantly outperforms existing state-of-the-art models. Our code and the synthetic dataset are available at https://github.com/YxZhxn/Ray3D.

1. Introduction

Accurate monocular 3D human pose estimation has found its wide applications in augmented reality [24], human-object interaction [6], and video action recognition [41]. While the problem has been extensively studied in recent years, it's a well-known ill-posed problem [23] with limited generalization capability. The problem becomes even more difficult when absolute 3D human pose estimation in a metric space is required, as knowing exactly where a human joint is in the World Coordinate System (WCS) is much more challenging than estimating the relative 3D offset of that joint from a reference point. While being more challenging, knowing absolute 3D poses is more desirable than the root-relative 3D poses in the real-world applications. For instance, unmanned store requires to detect the merchandise picked up by the customer, which relies on accurate hand localization in world coordinate system.

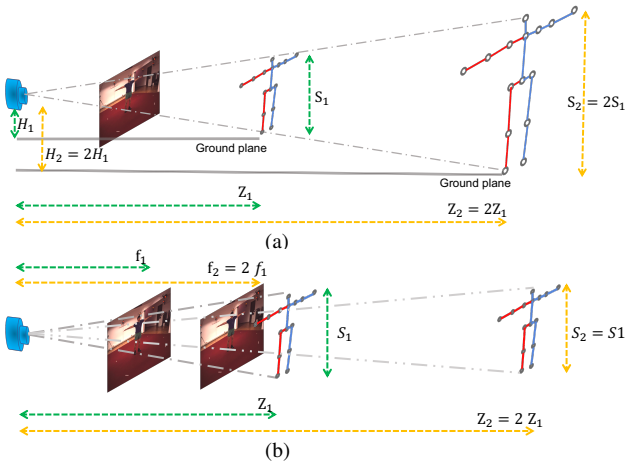

Figure 1. As shown in (a), if both the body size and the distance to camera are scaled up by twice, the projected 2D keypoints locations remain the same. Same phenomenon is observed in (b), where both the focal length and 3D distance are doubled. Z_1 and Z_2 refer to the distance from the person to the camera, H_1 and H_2 represent the height of camera from the ground plane. S_1 and S_2 are the scale of the person. f_1 and f_2 represent focal length of the camera. Such ambiguity is discussed in [4] as well.

A key-point's 2D pixel location is jointly determined by the scale of person's body figure, camera intrinsic parameters, camera extrinsic parameters and 3D position in the world coordinate system. These factors introduce ambiguities for 3D pose estimation. For instance, as shown in Figure 1 (a), if both the body size and the distance to camera are scaled up by twice, the projected 2D key-points locations remain the same. Similarly, if both the focal length and 3D distance are doubled, the 2D key-points keep the same, as illustrated in Figure 1 (b). Typically, there are more than one configuration of 3D key-points that can generate the same observation of 2D key-points in the image plane. Thus, naively learning a model to map from 2D pixel locations to 3D world locations is arguably prone to failure.

To resolve these ambiguities, a number of monocular 3D human estimation approaches have been proposed [4, 10, 29, 32, 44, 47]. These methods can be mainly categorized into two groups, *i.e.*, lifting methods and image based methods. Lifting methods [3, 8, 11, 25, 32, 46, 50] take the 2D human poses as input and lift the 2D pose to 3D pose. A few lifting methods normalize the input according to image resolution [32], and camera principal point [4]. While these normalization schemes improve the generalization ability to some extent, they fail to fully resolve the ambiguity due to variation in camera intrinsic parameters. On the other hand, image based approaches [2, 10, 16, 20, 22, 29, 43, 49] estimate the 3D root position based on the prior about the body size. In contrast, [29, 47] rely on image-based human depth estimation for absolute root-keypoint localization. The issue with these learning-based depth estimation approaches is lack of sufficient training data with viewpoint variations. For instance, the model trained with front-view viewpoint may not generalize well to cameras with large pitch value. Moreover, they fail to fully address the aforementioned ambiguities.

To address the challenges more effectively, we propose our Ray3D method. Firstly, in order to have an intrinsic-parameter-invariant representation, we convert the 2D keypoints in a pixel space to 3D rays in a normalized 3D space. With this simple design, our Ray3D approach achieves stable performance regardless of camera intrinsic parameter changes. Inspired by Videopose [32] and RIE [34], we fuse 3D rays from consecutive frames by using temporal convolution in order to further resolve the ambiguity introduced by occlusion and to improve accuracy. This temporal fusion mechanism stabilizes the output and generates more accurate 3D locations. Secondly, we jointly embed the camera extrinsic parameters into the network. Camera extrinsic parameters contain essential information for accurate 3D human pose estimation. Arguably, exploiting camera extrinsic parameters is the only way to resolve the human body part size ambiguity. For instance, in Fig. 1 (a), if the camera's height is known to be close to H_1, we can safely eliminate incompatible hypotheses like the 3D pose with S_2/H_2. Therefore, it's essential to incorporate the camera extrinsic parameters into the network for accurate absolute localization. To our best knowledge, none of the existing learning-based 3D human pose estimation approaches explicitly utilize these information. In contrast, we directly take camera height and camera pitch value as input, and learn an independent camera embedding through a Multi-Layer Perceptron (MLP). This camera embedding is then concatenated with temporally fused ray features for 3D pose estimation.

To understand and diagnose the absolute 3D pose estimators, we conduct a series of comprehensive and systematic experiments. Specifically, we explicitly benchmark the robustness of the approaches against focal length, principal point, camera pitch angle, camera height, camera yaw angle, body figure size variations on synthetic dataset. Furthermore, we evaluate generalization capability of these approaches on three single person benchmarks.

To summarize, the proposed method makes the following contributions,

- We convert the input space from 2D pixel space to 3D rays in a normalized coordinate system. This simple design effectively normalizes away the variations introduced by the camera intrinsic parameter changes as well as the camera pitch angle changes.
- We present a novel and simple network which learns a camera embedding using the camera extrinsic parameters, and jointly models the distribution of camera extrinsic parameters and 3D rays.
- We provide a comprehensive and systematic benchmarking of existing 3D approaches in terms of robustness against camera pose variations, as well as cross-dataset generalization.
- Experiments on three real benchmark datasets and one synthetic dataset clearly demonstrates the advantages of our Ray3D approach.

2. Related Work

2.1. Lifting based 3D human pose estimation

Lifting based 3D human pose estimation approaches learn a model to map from 2D human pose to 3D. While only using 2D coordinates without image features may appear sub-optimal, these methods enjoy good performance-cost trade-off.

The majority of lifting based approaches work on root-relative 3D human pose estimation [3, 8, 11, 25, 26, 32, 34, 46, 48, 50]. [26] is the pioneer work that introduces the lifting design. [3, 32, 34, 48] exploit temporal information to improve the 3D pose estimation accuracy, especially for the occluded cases. Pose ambiguities can be partially resolved by exploiting the temporal context. Shan *et al.* [34] propose to encode relative positional and temporal enhanced representations, and this approach achieves excellent root-relative 3D pose accuracy. Inspired by [34], we encode relative 3D normalized rays to improve the root-relative pose model.

Only a handful of lifting approaches can be applied for absolute 3D human pose estimation. Pavllo *et al.* [32] employ a trajectory model to estimate the 3D trajectory of the root joint. Chang *et al.* [4] normalize the input by subtracting principal point of the camera and reconstruct up to the canonical root depth. This root depth is further multiplied with focal length to generate the final absolute depth. While these approaches achieve promising results, they fail to fully normalize the input by using camera intrinsics. Meanwhile,

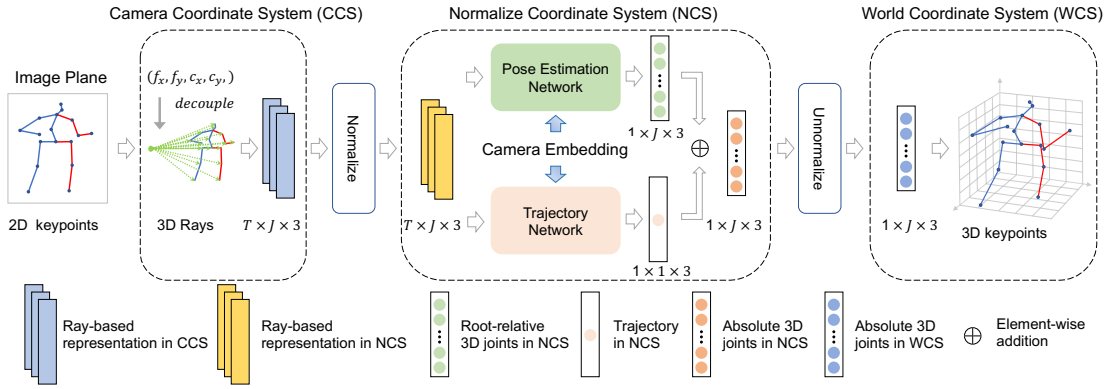

Figure 2. An exposition of our Ray3D architecture. In pre-processing, we convert 2D input to ray-based 3D representation. These 3D rays are transformed to NCS, which are subsequently fed to pose estimation network and trajectory network to predict the final absolute 3D pose. With unnormalization, the 3D pose under world coordinate system is obtained.

camera extrinsics are simply ignored. Thus, their performance is susceptible to camera intrinsics/extrinsics variations.

2.2. Image based 3D human pose estimation

Image based approaches aim to improve the 3D estimation accuracy by directly utilizing image features [1, 17, 18, 20, 21, 37–39, 45, 47, 49]. Rogez *et al.* [33] frame the estimation problem as pose proposal generation, proposal scoring and proposal refinement. The 3D location of the root joint is obtained by minimizing the distance between 2D pose and projected 3D pose. Moon *et al.* [29] devise a network to estimate the depth of root joint from cropped single person image, which inevitably loses contextual information of the subject. Alternatively, [37–39, 47] estimate the root depth from the full image up to a scale. The issue with this direction is the requirement of significant amount of training data with camera variations to make image-based depth estimation reliable. Meanwhile, camera extrinsics are not taken into account, which limits their generalizability. Additionally, [19] firstly proposes to estimate perspective camera for 3D body pose regression. [19] tries to estimate camera parameters along with the pose as well, which improves the generalizability.

2.3. Camera encoding

Few approaches have explicitly encoded camera extrinsics to assist the vision tasks. Nerf [5, 28, 42] is a popular 3D object reconstruction approach that directly takes 2D camera viewing angle into input. The 2D viewing angle (*i.e.*, pitch and yaw) is concatenated with 3D location of object point and then processed by a multilayer perceptron network. Differently, our approach learns an camera embedding specifically for camera pitch and camera height from the ground plane. This embedding largely resolves the ambiguities in the absolute 3D human pose estimation problem.

3. Proposed Method

Intuitively, accurate monocular 3D absolute pose estimation relies on sufficient ambiguity reduction. Our method is designed to resolve ambiguities introduced by camera intrinsic parameter variation, body occlusion and camera pose variations with normalized representation of key-points, temporal convolution and camera embedding correspondingly.

In Figure 2, we present the overview of the proposed Ray3D framework. To eliminate the impact of the intrinsic parameter variation, the 2D key-points are converted into 3D rays in Camera Coordinate System (CCS). To deal with the camera pitch angle variation, we further transform these 3D rays into (pitch) *Normalized* Coordinate System (NCS). Similarly, ground truth 3D poses are transformed to NCS as well. In this way, both the input and output of the model are aligned into the same coordinate system.

Temporal key-point motion information helps resolve 3D pose estimation ambiguity introduced by the occlusion [32, 34]. Following [34], we fuse 3D rays from consecutive frames temporally and encode the relative pose rays to capture motion information. Different from [34], our Ray3D approach learns a camera embedding which can (implicitly) provide strong constraints to eliminate ambiguities in absolute 3D pose estimation. Specifically, we employ an MLP network to learn a compact embedding for camera pose representation. This camera embedding is subsequently concatenated with latent 3D ray features for the pose prediction. This novel design largely improves the model's robustness against camera pose and body scale variations.

Following [32], we decompose the problem into two sub-problems, *i.e.*, root-relative pose estimation and root location estimation. These two sub-problems are solved by our pose network and trajectory network respectively. Specifically, the relative 3D pose generated by pose estimation network is added to the root joint coordinate predicted by trajectory network. Finally, the human pose are *unnor-*

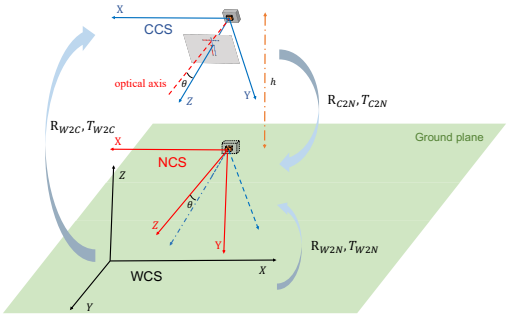

Figure 3. Normalized camera coordinate system is acquired by rotating the camera coordinate system along the x axis with degree of θ and translating along the z axis of the world coordinate system with distance of h. h is the height of camera in WCS. Such that the input and output of the lifting network are aligned in the same coordinate system.

malized into World Coordinate System (WCS).

3.1. Input pre-processing

Intrinsic parameter decoupling

Lifting based 3D pose estimation approaches manage to lift predicted 2D key-points $\{p_i\}_{i=1}^J$ to 3D key-points $\{P_i^C\}_{i=1}^J$ with deep neural network. $p_i = [x_i, y_i]$ stands for the location of the i th joint of the person in the input image coordinate system and $P_i^C = [X_i^C, Y_i^C, Z_i^C]$ represents the corresponding joint in CCS. J denotes the indices of joints. In order to achieve the invariance to the camera intrinsic parameter changes, we perform the following transformation to $\{p_i\}_{i=1}^J$ (camera un-distortion can be added if needed):

$$x_i^{ray} = \frac{x_i - c_x}{f_x}, y_i^{ray} = \frac{y_i - c_y}{f_y}, z_i^{ray} = 1. \quad (1)$$

Such that we have 3D rays $\{p_i^{ray}\}_{i=1}^J = \{[x_i^{ray}, y_i^{ray}, z_i^{ray}]\}_{i=1}^J$. In Eq. 1, c_x and c_y represent for camera center points, f_x and f_y denote the focal length. p_i^{ray} is a *ray* that points from optical center of the camera to the key-point i in the image plane.

Unlike [4], we completely eliminate the impact of focal length by explicitly normalize the ray representation with the calibrated focal length. Compared to [9], our 3D rays are converted to normalized rays in Normalized Coordinate System (NCS), which will be shortly discussed in the next subsection.

Extrinsic parameter decoupling with the Normalized Coordinate System (NCS)

Define a key-point in the world, camera, and *normalized* coordinate system as P_W, P_C and P_N respectively. With accurate calibration, one can acquire camera extrinsics including rotation matrix R_{W2C}, and translation vector T_{W2C}. And the transformation between P_W and P_C is as follows:

$$P_C = R_{W2C} \cdot P_W + T_{W2C}. \quad (2)$$

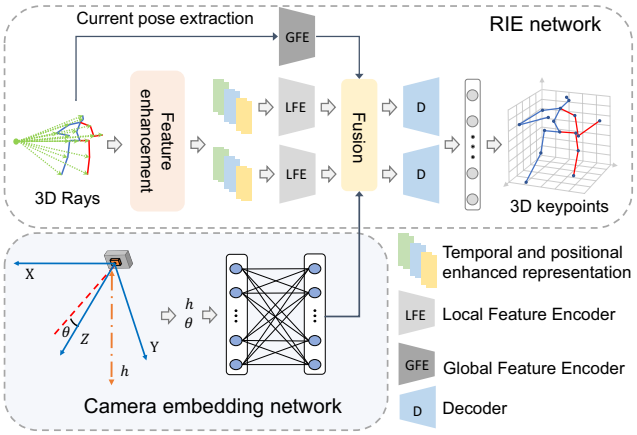

Figure 4. Overview of our lifting network. Our relative pose estimation and root joint estimation networks share the same RIE architectures. The network is equipped with positional and temporal enhanced representation. For more details, please refer to [34]. MLP based camera embedding works as a plug-in to generate embedding features, and then concatenate with latent ray features for final pose prediction.

In this paper, we aim to predict absolute 3D human pose in WCS. Camera pose in the 3D world can be determined by its 3D location, pitch, yaw and roll angles. Pitch, θ, describes the angle between the optical axis of the camera and the ground plane. With an assumption that camera yaw and roll values are close to 0, the camera pitch value and the camera height can uniquely specify the pose of a camera up to horizontal translation. In order to explicitly encode pitch for accurate pose estimation, we set up NCS as presented in Fig. 3. First, the CCS is rotated along the x axis to eliminate the pitch angle. Then, the coordinate system is translated along the z axis to the ground plane.

One can easily calculate the rotation matrix and translation vector between P_C and P_N:

$$R_{C2N} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}, \quad (3)$$

$$T_{C2N} = \begin{pmatrix} 0 & 0 & -h \end{pmatrix}. \quad (4)$$

According to the Eq. 2, 3 and 4, we have:

$$R_{W2N} = R_{C2N} \cdot R_{W2C}, \quad (5)$$

$$T_{W2N} = R_{C2N} \cdot T_{W2C} + T_{C2N}, \quad (6)$$

$$P_N = R_{C2N} \cdot P_C + T_{C2N}, \quad (7)$$

$$P_N = R_{W2N} \cdot P_W + T_{W2N}. \quad (8)$$

By applying the Eq. 7 to $\{p_i^{ray}\}_{i=1}^J$ and Eq. 8 to ground-truth 3D keypoints $\{P_i^W\}_{i=1}^J$, we can get normalized 3D rays $\{\hat{p}_i^{ray}\}_{i=1}^J$ and normalized 3D ground-truth $\{P_i^N\}_{i=1}^J$. As a result, our Ray3D network is trained to lift from $\{\hat{p}_i^{ray}\}_{i=1}^J$ to $\{P_i^N\}_{i=1}^J$ within the same coordinate system, which reduces the training difficulty and increases model robustness.

3.2. Lifting network

Absolute pose estimation

The task of estimating 3D absolute human poses is composed of two sub-problems, root location estimation (i.e., estimating the location of the center of mass of a person) and root-relative pose estimation (i.e., offset of each key point with respect to the center). We design our network to jointly learn to solve the two sub-problems with the trajectory network and the pose network, respectively (see Fig. 2). The outputs of these two networks are added to produce the absolute 3D poses.

Temporal motion information improves model's robustness against body occlusion. Inspired by [34], we adopt RIE architecture as our backbone network for both the root-relative pose network and trajectory network. As shown in Fig. 4, RIE network is enhanced with positional and temporal information. Relative positions of input key-points are encoded as positional information within the frame, and differences between the 2D pose of current frame and the one from neighbouring frames are treated as temporal information. Such enhanced input are divided into 5 groups (torso, left arm, right arm, left leg, and right leg) for local feature learning. In addition, global feature is extracted from current frame to maintain the consistency of overall posture. Feature fusion module aggregates all these features for 3D pose estimation. Using this architecture, we replace vanilla 2D human key-points with our intrinsic-invariant normalized 3D rays as input for both pose network and trajectory network to address ambiguities. We refer the reader to [34] for the details of RIE structure. Note that the contribution of this work is not on the specific network design, rather on the input representation and explicit embedding of camera extrinsic parameters. This novel design can be easily incorporated to the existing pose estimation framework.

Camera embedding

We argue that camera extrinsic parameters are critical for absolute pose estimation in WCS, and propose to explicitly utilize extrinsic parameters by learning an independent camera embedding through a Multi-Layer Perceptron with θ and h as inputs. Specifically, the camera embedding module is constructed with two fully connected layers followed with batch normalization [14], rectified linear unit [30] and Dropout [36]. As shown in Fig. 4, this camera embedding is concatenated with temporally fused latent 3D ray features in both relative pose prediction and trajectory network. Therefore, both networks exploit camera extrinsic parameters for robust and accurate pose estimation.

4. Experiments and results

The evaluation results of the proposed method with different experiment setups are reported in this section. First, datasets and evaluation metrics are introduced in Section 4.1, and details of implementation is described in Section 4.2. Section 4.3 showcases the comparison of our Ray3D and other state-of-the-arts on three public benchmarks. Then, generalization test result on a synthetic dataset is described in Section 4.4. Furthermore, the effectiveness of Ray3D's components is analyzed with ablation study in Section 4.5. Finally, the limitation of Ray3D is discussed in Section 4.6.

4.1. Datasets and Evaluation metrics

We evaluate our Ray3D on three public datasets captured with different camera poses and human poses. The camera intrinsics and extrinsics are provided by all datasets. *The following datasets contain personal identifiable information about human subjects. All the subjects in these datasets have granted their permission for the dataset creation.*

Human3.6M (H36M) [15] is a large-scale 3D human pose estimation dataset, which contains 3.6 million video frames recorded with four synchronized cameras. Following previous works [32, 48], five subjects (S1, S5, S6, S7, S8) and two subjects (S9, S11) with 17-key-point definition are used as training and testing data respectively for SOTA comparison.

Humaneva-I [35] is a much smaller dataset compared with H36M , which is captured in a controlled indoor environment with three cameras. The proposed Ray3D representation requires well-calibrated intrinsic and extrinsic, as a result, Camera 2 and Camera 3 are removed due to bad camera calibration.

MPI-INF-3DHP (3DHP) [27] consists of 1.3 million video frames, which covers more diverse human motions than Human3.6M. Following previous work [12], poses with 17 joints from Camera 0, 1, 2, 4, 5, 6, 7 and 8 are used for training. Sequence of TS1, TS3 and TS4 are adopted as test sets. TS2, TS5 and TS6 are excluded due to inaccurate or incomplete camera calibration.

In our experiments, we adopt following evaluation metrics: Mean Per Joint Position Error (MPJPE) in millimeters is used to evaluate root relative pose estimation results under CCS. To evaluate the performance of absolute pose under WCS, Absolute MPJPE (Abs-MPJPE) is adopted which calculates the difference between the prediction and GT pose in WCS. Mean of the Root Position Error (MRPE) proposed by [4] is used to evaluate the root joint's trajectory prediction.

4.2. Implementation Details

For our Ray3D approach, dimension of camera embedding is set as 64. Initial learning rate is 0.001. Adam op-

Table 1. Quantitative evaluation results under MPJPE on H36M using GT keypoints as input. (f = 9) means this approach utilizes 9 consecutive frames for pose estimation, and (f = 1) means the approach does not make use of temporal information. Best results are shown in **bold**.

MPJPE		Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Somke	Wait	WalkD.	Walk	WalkT.	Average
Hossain et al. [13]	ECCV'18	35.2	40.8	37.2	37.4	43.2	44.0	38.9	35.6	42.3	44.6	39.7	39.7	40.2	32.8	35.5	39.2
Liu et al. (f = 243) [25].	CVPR'20	34.5	37.1	33.6	34.2	32.9	37.1	39.6	35.8	40.7	41.4	33.0	33.8	33.0	26.6	26.9	34.7
Videopose. (f = 9) [32]	CVPR'19	37.0	40.7	35.2	37.4	38.4	44.2	42.3	37.1	46.5	48.8	38.9	40.1	38.5	29.9	32.6	39.2
PoseFormer (f = 9) [48]	ICCV'21	49.2	49.7	38.7	42.7	40.0	40.9	50.7	42.2	47.0	46.1	43.4	46.7	39.8	36.4	38.0	43.5
PoseAug (f = 1) [12]	CVPR'21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	38.2
RIE (f = 9) [34]	ACMMM'21	34.8	38.2	**31.1**	34.4	35.4	**37.2**	38.3	32.8	**39.5**	**41.3**	34.9	**35.6**	32.9	**27.1**	**28.0**	34.8
Ray3D (f = 9)		**31.2**	**35.7**	31.4	**33.6**	**35.0**	37.5	**37.2**	**30.9**	42.5	**41.3**	**34.6**	36.5	**32.0**	27.7	28.9	**34.4**

Table 2. Quantitative evaluation results under Abs-MPJPE and MRPE on H36M using CPN detected keypoints as 2D input. Best results are shown in **bold**.

Abs-MPJPE		Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch	Sit	SitD.	Somke	Wait	WalkD.	Walk	WalkT.	Average
Videopose (f = 9) [32]	CVPR'19	128.9	125.4	124.4	138.2	**108.2**	155.5	116.6	101.1	135.8	287.6	128.6	130.9	122.1	101.6	110.7	134.4
PoseLifter (f = 1) [4]	ICCV'19	140.9	113.2	139.9	148.2	122.0	155.3	121.5	121.1	170.0	267.6	139.2	142.9	146.4	132.1	135.2	146.4
PoseFormer (f = 9) [48]	ICCV'21	112.6	137.1	**117.6**	145.8	113.0	166.0	125.5	113.8	**128.8**	245.7	122.7	144.8	125.0	118.9	129.3	136.5
RIE (f = 9) [34]	ACMMM'21	143.2	133.2	143.9	142.7	110.9	151.4	125.9	98.4	136.4	273.4	127.5	138.9	126.8	107.3	116.0	138.4
Ray3D (f = 1)		**80.1**	100.8	123.8	125.5	110.7	111.8	96.1	99.3	129.4	176.3	**106.8**	129.2	120.4	109.1	106.6.	115.1
Ray3D (f = 9)		92.9	**97.4**	139.8	**118.6**	113.8	**105.9**	**84.5**	**74.9**	148.6	**165.7**	116.6	**113.9**	**98.2**	**83.6**	**87.9**	**109.5**
MRPE		Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Somke	Wait	WalkD.	Walk	WalkT.	Average
Videopose. (f = 9) [32]	CVPR'19	124.2	115.9	111.0	127.3	97.6	141.9	105.7	96.4	122.0	276.5	119.6	123.3	111.3	94.0	101.6	124.6
PoseLifter (f = 1) [4].	ICCV'19	134.7	102.3	126.9	135.7	109.9	138.5	110.7	110.9	170.0	252.4	128.4	133.9	139.4	121.6	124.4	135.1
PoseFormer (f = 9) [48]	ICCV'21	104.7	134.7	**103.9**	137.4	99.6	154.6	119.8	108.9	**108.2**	233.7	111.1	141.1	116.2	117.9	123.8	127.7
RIE (f = 9) [34]	ACMMM'21	139.1	124.5	129.9	133.1	**99.2**	141.4	116.3	93.5	124.0	265.9	118.4	131.3	117.1	100.4	109.2	129.6
Ray3D (f = 1)		**67.3**	91.7	113.6	111.8	104.5	96.3	85.8	94.6	124.4	161.7	**97.6**	119.5	110.9	100.9	94.8	105.0
Ray3D (f = 9)		83.7	**86.8**	128.9	**104.8**	109.3	**91.6**	**75.0**	**65.2**	143.9	**150.5**	108.6	**105.7**	**88.4**	**73.9**	**77.8**	**99.6**

timizer with exponential learning rate decay factor of 0.99 is employed. Horizontal flip augmentation is adopted both in training and testing. For H36M dataset, we adopt Cascaded Pyramid Network (CPN) [7] detected poses and GT 2D poses as input. As for Humaneva-I and 3DHP, only GT 2D poses are used.

4.3. Evaluation on public benchmarks

In this section, we first compare our Ray3D with state-of-the-art methods under all 15 action sequences on H36M, then generalizability of these methods is evaluated by cross-dataset testing. The comparing approaches include the latest PoseFormer [48], Videopose [32], PoseLifter [4] and RIE [34]. Note that PoseLifter was designed for absolute pose estimation. Different from Videopose and PoseLifter, both PoseFormer and RIE are incapable for absolute pose estimation. To test their capability for root-joint localization, we equip them with a trajectory model using their own network structure. For fair comparison, we carefully retrained PoseFormer, Videopose, PoseLifter and RIE with their provided source code under PyTorch [31].

H36M evaluation Table 1 shows the performance of the methods that focus on root-relative pose prediction where ground truth 2D keypoints are taken as input. From the table, we can observe that our Ray3D obtains comparable results compared to SOTA methods. Specifically, MPJPE surpasses RIE [34] by 0.4mm. Table 2 shows the results for absolute pose estimation in WCS using CPN [7] detected 2D poses on H36M dataset. It can be seen that Ray3D outperforms all SOTA methods for both Abs-MPJPE and MRPE with clear margin. Compared with RIE, our method reduces Abs-MPJPE by 28.9mm and MRPE by 30.0mm respectively. It is worth noting that Ray3D outperforms PoseLifter by 31.3mm under Abs-MPJPE when no temporal information is used. These results demonstrate that Ray3D is effective and generates more accurate absolute 3D locations.

Among these four baseline methods, PoseLifter using single frame performs the worst. And Ray3D working with 9 frames surpasses Ray3D using single frame. This verifies the benefits of using temporal features for 3D pose estimation. Another interesting observation is that these baseline methods perform similarly in terms of MRPE. This shows the network structure design is not the critical factor for absolute pose estimation. Rather the input representation and camera embedding are the keys for accurate keypoint localization.

Cross-dataset testing We train comparing models in 3DHP dataset, and evaluate them using H36M and Humaneva-I. 14-joint definition is applied for all datasets during cross-dataset testing. For H36M and 3DHP, we remove mid spine, neck and chin keypoints. As for Humaneva-I, the thorax key-point is removed out of original 15 joints. As shown in Table 3, none of the baselines work well in cross-scenario situations while the Ray3D shows good generalization performance in Humaneva-I and 3DHP dataset. This is because the camera intrinsic and extrinsic vary greatly across different scenes. Our Ray3D approach explicitly takes extrinsics(*i.e.*, camera pitch and camera height) as input to learn a camera embedding, which results in improved generalization ability.

4.4. Evaluation on synthetic dataset

In this section, we conduct in-depth systematic experiments to benchmark 3D pose estimators' robustness against camera intrinsic, camera rotation (yaw), camera pitch, camera translation and person scale variations on a carefully cu-

Table 3. Cross dataset evaluation. We adopt a 14-joint skeleton training on 3DHP, testing on H36M, Humaneva-I and 3DHP datasets. MPJPE, Abs-MPJPE and MRPE are reported. The unit of all numbers is mm. The best results are in **bold**.

method \datasets	H36M			HumanEva-I			3DHP		
	MPJPE	Abs-MPJPE	MRPE	MPJPE	Abs-MPJPE	MRPE	MPJPE	Abs-MPJPE	MRPE
Videopose (f = 9) [32]	**81.2**	1680.3	1686.6	86.2	1387.4	1387.1	58.2	149.1	143.0
PoseFormer (f = 9) [48]	97.8	1824.0	1818.9	104.7	1470.4	1452.0	47.3	207.5	211.3
PoseLifter (f = 1) [4]	92.9	573.3	570.5	240.2	1263.0	1129.3	76.7	147.8	133.6
RIE (f = 9) [34]	91.2	1679.9	1673.0	92.0	1375.8	1369.5	50.9	135.6	132.4
CDG (f = 1) [40]	95.6	-	-	-	-	-	90.3	-	-
Ray3D (f = 9)	84.4	**243.9**	**246.7**	**83.9**	**477.8**	**468.6**	**46.6**	**103.3**	**95.3**

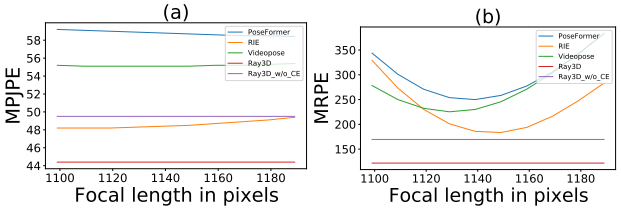

Figure 5. performance under MPJPE and MRPE in case of focal length changes are plotted in (a) and (b) respectively. The x-axis represents the focal length of the virtual camera in pixels.

rated synthetic benchmark. We use cameras from H36M to conduct camera augmentations. After the synthesis, the focal length of simulated cameras ranges from 1100 to 1180, where focal length in the training data ranges from 1143 to 1150. Camera rotation ranges from 0 to 360 degrees. Camera pitch ranges from 0 to 40 degrees. Camera translation ranges from 9 to 14 meters. The total length of human limbs ranges from 2.5 to 4.5 meters (roughly, the height of human ranges from 1 to 2 meters). Specifically, 100 virtual cameras are generated with fixed extrinsic for intrinsic generalization test, and 126 virtual cameras are generated with fixed intrinsic for extrinsic generalization test. We additionally simulated 324 cameras for training as well. Note that training and testing camera poses do not overlap. Five subjects (S1, S5, S6, S7, S8) and two subjects (S9, S11) with 14-key-point-definition are used for training and testing respectively. The detailed camera augmentation setting for training and testing are listed in the supplementary materials. To evaluate the effectiveness of camera embedding, we add a new baseline named Ray3D_w/o_CE where camera embedding branch is removed from Ray3D.

Intrinsic generalization To verify the robustness of methods to camera intrinsic change, we change the focal length of the cameras with fixed resolution. As shown in the Fig. 5 (a) and (b), focal length changes affect VideoPose, PoseFormer, RIE to varying degrees under MPJPE and MRPE metrics respectively. For instance, with only 4% variation on focal length, MRPE of the baseline approaches increase more than 50%. In contrast, both Ray3D and Ray3D_w/o_CE achieve stable result. This result clearly showcases the merits of our ray-based input representation.

Extrinsic generalization To evaluate the impact of the change of extrinsics on generalizability, we change the rotation, pitch angles and translation of the camera pose respectively. Note that the translation is measured by the euclidean distance between camera and subject. In addition, we design a new baseline approach for root joint localization. Specifically, we estimate the height of the root joint using the mean average height from subjects of H36M, *i.e.*, 93.95cm. Using this height assumption we may localize the root joint along its 3D ray. We term this approach as Ray Fixed Root Height (RFRH).

To measure the robustness of model against rotation variations, we rotate the camera around the scene center while keeping the camera height and camera pitch the same. As shown in Fig. 6 (a) and Fig. 7 (a), Ray3D outperforms baseline methods by large margin on both MPJPE and MRPE, which indicates Ray3D is not only able to accurately localize the joint in WCS, it's also able to estimate the root-relative pose robustly. Ray3D_w/o_CE shows better results than RIE, which indicates that normalized ray representation works better than vanilla 2D keypoints. Among the baseline approaches, the learning-based approaches achieve better result than RFRH. This demonstrates the learning based methods indeed manage to resolve the ambiguities to some extent through data-driven way. RFRH performs poorly due to the violation of root height assumption in the evaluation dataset. For instance, when the subject sits down on the floor, the root joint height can be close to 0.

To evaluate robustness against camera pitch variation, we change pitch angle of the cameras to various degrees while keeping the same distance between camera and the subject. As shown in Fig. 6 (b) and Fig. 7 (b), Ray3D consistently outperforms baselines at all pitch angles for MRPE and MPJPE.

Similarly, to evaluate model robustness against camera translation, we generate a batch of cameras with constant pitch angle and gradually changing distance from the subject. As shown in Fig. 6 (c) and Fig. 7 (c), baseline methods suffer from performance degradation as the camera moves away from subjects. In contrast, Ray3D achieves satisfactory results.

Person scale generalization To verify robustness against the person scale ambiguity, we change the bone length of

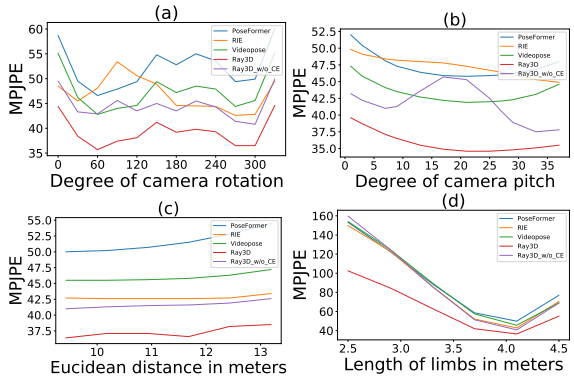

Figure 6. Figures (a), (b), (c) and (d) showcase the performance using MPJPE metric in case of rotation, camera pitch, translation and body scale variations correspondingly. The x-axis denotes the degree of camera rotation, the degree of camera pitch, euclidean distance between camera and subject in meters and the total length of human limbs in meters respectively.

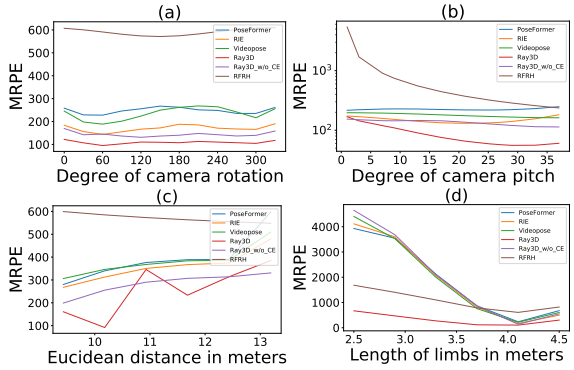

Figure 7. Figures (a), (b), (c) and (d) showcase the performance using MRPE metric in case of rotation, camera pitch, translation and body scale variations correspondingly. The x-axis denotes the degree of camera rotation, the degree of camera pitch, euclidean distance between camera and subject in meters and the total length of human limbs in meters respectively.

H36M body poses to 0.6-1.1 times as done in PoseAug [12]. The experimental results are shown in Fig. 6 (d) and 7 (d). Accuracy of all the comparing approaches degrades notably when the body size is small. For instance, for the smallest body figure, MRPE of PoseFormer, RIE, Videopose reach 4 meters, which is even higher than rule-based RFRH. MRPE of Ray3D increases to 800mm, which is still much better than baselines.

Throughout these systematic experiments on the synthetic dataset, we can safely arrive following conclusions. By converting 2D keypoints to normalized rays, both our Ray3D and Ray3D_w/o_CE achieve stable and accurate performance regardless of camera intrinsics variations. By adding camera embedding, Ray3D outperforms Ray3D_w/o_CE with clear margin for most of the testing cases except for a few camera settings in Fig. 7 (c). This verifies the effectiveness of camera embedding for both root-relative pose estimation and root joint absolute localization.

4.5. Ablation Studies

In this experiment, we further study the effectiveness of intrinsic decoupling, camera coordinate normalization and camera embedding using cross-dataset setting. We use RIE network as baseline model and add our modules gradually. All the models are trained on H36M dataset and tested on 3DHP dataset with 17-keypoint 2D poses as input. Table 4 summarizes the results. As we can see, MRPE keeps decreasing when adding proposed modules. Intrinsic decoupling improves the baseline model with a large margin. Camera normalization also drives the model to be more generalizable. With camera embedding, the trajectory estimation produces the best MRPE.

Table 4. Ablation study on 3DHP dataset with models trained with H36M dataset. IND denotes intrinsic decoupling, Nor. represents camera normalization. CE is camera embedding. We use RIE as base model. Best results are shown in **bold**.

Method	IND	Nor.	CE	MRPE
RIE	✗	✗	✗	1079.2
RIE_w_IND	✓	✗	✗	448.9
Ray3D_w/o_CE	✓	✓	✗	311.6
Ray3D	✓	✓	✓	**307.4**

4.6. Discussion

Our Ray3D approach outperforms the baseline methods significantly in terms of generalizability both on the three realistic benchmarks and on the synthetic dataset. This clearly showcases the robustness of the Ray3D approach. However, Ray3D's performance drops when the body figure size varies a lot, as shown in Fig. 7 (d). This is mainly due to the fact that all the training body poses are adult. Meanwhile, our approach assumes the subject is on the ground plane. The model may fail if the subject is off the ground for a long time (*e.g.*, climbing a ladder). Additionally, calibrated camera parameters need to be provided, which limits the use cases of Ray3D. Accurate 3D pose estimation might be misused for surveillance applications where skeleton configuration estimation can assist person identification. We advocate proper usage.

5. Conclusion

In this paper, we present an innovative monocular absolute 3D human pose estimation approach named Ray3D. This approach gradually resolves the inherent ambiguities through a series of novel designs: conversion from 2D keypoints to 3D normalized rays; temporal fusion of 3D rays; inclusion of camera embedding. As a result, Ray3D significantly outperforms SOTA methods on three realistic benchmarks and one synthetic benchmark.

References

[1] Thiemo Alldieck, Marcus A. Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *CVPR*, pages 1175–1186, 2019. 3

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter V. Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In *ECCV*, volume 9909, pages 561–578. Springer, 2016. 2

[3] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat-Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *ICCV*, pages 2272–2281, 2019. 2

[4] Ju Yong Chang, Gyeongsik Moon, and Kyoung Mu Lee. Absposelifter: Absolute 3d human pose lifting network from a single noisy 2d human pose. *ICCV*, 6(7):9, 2019. 1, 2, 4, 5, 6, 7

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. *CoRR*, 2021. 3

[6] Yixin Chen, Siyuan Huang, Tao Yuan, Yixin Zhu, Siyuan Qi, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *ICCV*, pages 8647–8656. IEEE, 2019. 1

[7] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, pages 7103–7112, 2018. 6

[8] Yu Cheng, Bo Yang, Bo Wang, Yan Wending, and Robby T. Tan. Occlusion-aware networks for 3d human pose estimation in video. In *ICCV*, pages 723–732. IEEE, 2019. 2

[9] Hanbyel Cho, Yooshin Cho, Jaemyung Yu, and Junmo Kim. Camera distortion-aware 3d human pose estimation in video with optimization-based meta-learning. In *ICCV*, pages 11169–11178, October 2021. 4

[10] Rishabh Dabral, Nitesh B. Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. Multi-person 3d human pose estimation from monocular images. In *3DV*, pages 405–414, 2019. 2

[11] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *ECCV*, volume 11213, pages 679–696, 2018. 2

[12] Kehong Gong, Jianfeng Zhang, and Jiashi Feng. Poseaug: A differentiable pose augmentation framework for 3d human pose estimation. In *CVPR*, pages 8575–8584, 2021. 5, 6, 8

[13] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *ECCV*, pages 68–84, 2018. 6

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, volume 37, pages 448–456, 2015. 5

[15] 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. 5

[16] Aaron S. Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. In *ECCVW*, volume 11132, pages 64–77, 2018. 2

[17] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *ECCV*, volume 12365, pages 18–35, 2020. 3

[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: video inference for human body pose and shape estimation. In *CVPR*, pages 5252–5262, 2020. 3

[19] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11035–11045, October 2021. 3

[20] 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

[21] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, pages 4501–4510, 2019. 3

[22] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *CVPR*, pages 4704–4713, 2017. 2

[23] Hsi-Jian Lee and Zen Chen. Determination of 3d human body postures from a single view. *Computer Vision Graph Image Process.*, 30(2):148–168, 1985. 1

[24] Huei-Yung Lin and Ting-Wen Chen. Augmented reality with human body interaction based on monocular 3d pose estimation. In *Advanced Concepts for Intelligent Vision Systems - 12th International Conference*, volume 6474, pages 321–331. Springer, 2010. 1

[25] Ruixu Liu, Ju Shen, He Wang, Chen Chen, Sen-Ching S. Cheung, and Vijayan K. Asari. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *CVPR*, pages 5063–5072, 2020. 2, 6

[26] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, pages 2659–2668, 2017. 2

[27] 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 *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 5

[28] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, volume 12346, pages 405–421. Springer, 2020. 3

[29] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single RGB image. In *ICCV*, pages 10132–10141, 2019. 2, 3

[30] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814, 2010. 5

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, pages 8024–8035, 2019. 6

[32] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, pages 7753–7762, 2019. 2, 3, 5, 6, 7

[33] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, pages 1216–1224, 2017. 3

[34] Wenkang Shan, Haopeng Lu, Shanshe Wang, Xinfeng Zhang, and Wen Gao. Improving robustness and accuracy via relative information encoding in 3d human pose estimation. In *ACMMM*, pages 3446–3454, 2021. 2, 3, 4, 5, 6, 7

[35] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4–27, 2010. 5

[36] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014. 5

[37] Márton Véges and András Lorincz. Absolute human pose estimation with depth prediction network. In *IJCNN*, pages 1–7. IEEE, 2019. 3

[38] Márton Véges and András Lorincz. Multi-person absolute 3d human pose estimation with weak depth supervision. In *International Conference on Artificial Neural Networks*, volume 12396, pages 258–270, 2020. 3

[39] Can Wang, Jiefeng Li, Wentao Liu, Chen Qian, and Cewu Lu. HMOR: hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *ECCV*, volume 12348, pages 242–259. Springer, 2020. 3

[40] Zhe Wang, Daeyun Shin, and Charless C. Fowlkes. Predicting camera viewpoint improves cross-dataset generalization for 3d human pose estimation. In *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part II*, volume 12536 of *Lecture Notes in Computer Science*, pages 523–540, 2020. 7

[41] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, pages 7444–7452. AAAI Press, 2018. 1

[42] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 3

[43] 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. 2

[44] 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 *NIPS*, pages 8420–8429, 2018. 2

[45] 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. 3

[46] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, pages 7374–7383, 2020. 2

[47] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. SMAP: single-shot multi-person absolute 3d pose estimation. In *ECCV*, volume 12360, pages 550–566, 2020. 2, 3

[48] Ce Zheng, Sijie Zhu, Matías Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. *ICCV*, 2021. 2, 5, 6, 7

[49] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *ICCV*, pages 7738–7748, 2019. 2, 3

[50] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *ICCV*, pages 398–407, 2017. 2