Supplementary material

1. Calculation of camera height and camera pitch

In this section, we introduce ways of calculation of camera height and pitch. We assume that the WCS is constructed with respect to the constraint that axis x and axis y of the WCS are placed on the plane ground.

Camera height

The 3D coordinates of the perspective camera $M = [X_W, Y_W, Z_W]^T$ in the WCS is calculated with its own rotation matrix R_{W2C}, and translation vector T_{W2C} as follows:

$$M = -R_{W2C}^{-1} \cdot T_{W2C}. \tag{1}$$

Such that the camera height equals to Z_W.

Camera pitch

Define the unit vector along the camera optical axis as $\boldsymbol{V_C} = [0, 0, 1]^T$ in the CCS. One can easily calculate its representation $\boldsymbol{V_W}$ in the WCS as:

$$\boldsymbol{V_W} = R_{C2W} \cdot \boldsymbol{V_C}. \tag{2}$$

Then it is straightforward to compute camera pitch θ that defines the angle between the optical axis of the camera and the ground plane in the WCS.

2. Synthetic dataset generation

To systematically evaluate the robustness of 3D human pose estimators against variations of camera intrinsic and extrinsic parameters, we create a synthetic dataset with intrinsic and extrinsic parameters augmentation from H36M dataset. In this section, we describe how we generate this synthetic dataset in details. First, the mathematics formulation of intrinsic/extrinsic augmentation is introduced. Then we describe the adopted augmentation parameters.

Intrinsic parameters augmentation

For camera intrinsic parameters augmentation, focal length f and principal points c are modified while keeping the projected 2D keypoints within the field of view. The augmentation is performed as follows:

$$\begin{aligned} &\hat{f} = f + \Delta f, \\ &\hat{c} = c + \Delta c, \\ &s.t. \quad 0 \leq \{x_i\}_{i=1}^{J} \leq W_I, \\ &\qquad 0 \leq \{y_i\}_{i=1}^{J} \leq H_I \end{aligned} \quad . \tag{3}$$

H_I and W_I as the height and width of the image frame. x_i and y_i are the projected 2D keypoint location. Different camera intrinsics result in different 2D keypoints in the image. Without 3D ray representation, the network struggles to predict the same 3D pose facing camera intrinsic variation.

Extrinsic parameters augmentation

For camera extrinsic parameters augmentation, we modify camera viewpoint β (camera rotation), the relative distance between subject and the camera γ (camera translation) and pitch of the camera θ (camera pitch). The augmentation is conducted as follows with constraint that the projected 2D keypoints are in the image frame:

$$\begin{aligned} &\hat{\beta} = \beta + \Delta\beta, \\ &\hat{\gamma} = \gamma + \Delta\gamma, \\ &\hat{\theta} = \theta + \Delta\theta, \\ &s.t. \quad 0 \leq \{x_i\}_{i=1}^{J} \leq W_I, \\ &\qquad 0 \leq \{y_i\}_{i=1}^{J} \leq H_I \end{aligned} \quad . \tag{4}$$

By changing camera rotation, camera translation and camera pitch, we may generate various located virtual cameras. Camera embedding is learned for every camera for generalisation of lifting network, which is helpful for accurate trajectory prediction. The specific augmentation parameters are detailed in the following section.

Augmentation parameters

Without loss of generality, we randomly select 1 camera whose id number is *55011271* from H36M to conduct camera augmentation. The overall summary of the dataset is shown in Table 1.

To evaluate model performance against camera intrinsic variations, we generate the intrinsic testing dataset. Specifically, the focal length and the coordinate of principal points[1] of simulated cameras are augmented. Note that for training dataset and extrinsic testing set, camera intrinsics are the same as original set-up in H36M. For instance, the focal length of simulated cameras in the intrinsic testing set ranges from 1100 to 1180, compared with the cameras from training set whose focal length is in the 1143-1150 range. Similarly, for the x coordinate of principal point of simulated cameras, it has a longer range of 450 to 550 in the intrinsic testing dataset, compared with a range of 508 to 514 in the training dataset. In total, we have 100 virtual cameras generated with fixed extrinsic for intrinsic generalization test.

For extrinsic generalization test, camera rotation, camera pitch and camera translation are augmented. Specifically, camera rotation ranges from 0 to 360 degrees at 30 degree interval such that extrinsic-testing cameras evenly

[1] We set the same value for x coordinate and y coordinate of principal point for simplicity.

Table 1. Technical summary of synthetic dataset constructed based on H36M.

Dataset	Num. of camera pose	focal length/pixel	x-coordinate of principal point/pixel	camera rotation/degree	camera pitch/degree	camera translation/meter	subjects
training	324	[1143:1150]	[508:514]	[60:300:120]	[2:38:2]	[9.05:11.70:0.76]	S1, S5, S6, S7, S8
extrinsic testing	126	[1143:1150]	[508:514]	[0:360:30]	[1:37:2]	[9.43:13.19:0.76]	S9, S11
intrinsic testing	100	[1100:1180]	[450:550]	0	12	4.5	S9, S11

revolve around the subjects. Camera pitch ranges from 0 to 40 degrees, which covers both frontal-view camera and large-pitch cameras. The interval of camera pitch is 2 degrees for both training dataset and extrinsic testing dataset. Camera translation ranges from 9 to 14 meters such that the relative distance between camera and subject is changing from the near to the distant. The interval of camera translation is 0.76 meter for training and extrinsic testing cameras. In total, 126 virtual cameras are generated with fixed intrinsic for extrinsic generalization test. And 324 cameras are generated for training, such that camera embedding module learns to cope with vast range of camera pose variations. We set the augmentation parameter to the range which are common in the real-world scenarios (*e.g.*, unmanned stores).

As for person scale generalization, the total length of augmented human limbs (bone length) ranges from 2.5 to 4.5 meters with the height of human ranging from 1 meter to 2 meters correspondingly. Note that the synthetic 3D human skeletons are only used for person scale generalization test, but excluded during model training stage.

3. Supplementary experiments

In this section, we report additional evaluation results to fully analyze the proposed Ray3D on public and synthetic datasets.

3.1. Evaluation on public benchmarks

H36M evaluation Table 2 shows the performance of the methods that focus on root-relative pose estimation where detected 2D keypoints are taken as input. When the number of video frames taken as input are similar, we can observe that our Ray3D obtains comparable results compared to SOTA methods under MPJPE metric in Camera Coordinate System (CCS). MPJPE of Ray3D surpasses PoseFormer [10] and Videopose [5] by 2.3mm and 0.9mm respectively, but Ray3D performs worse than RIE [6] by 1.1mm. We argue that Ray3D is designed for absolute 3D pose estimation in World Coordinate System (WCS), such performance of root-relative pose estimation in CCS is acceptable.

Table 3 shows the results for absolute pose estimation in WCS using GT 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 9.6mm and MRPE by 4.2mm respectively. These results demonstrate that Ray3D is effective and generates more accurate absolute 3D locations.

3DHP evaluation Table 4 shows the results for absolute pose estimation in WCS using GT 2D poses on 3DHP dataset. One can observe that Ray3D outperforms all SOTA methods for both Abs-MPJPE and MRPE with clear margin. For instance, compared with PoseFormer [10], our method reduces Abs-MPJPE by 44.4mm and MRPE by 51.7mm respectively.

Cross-dataset testing We train comparing models on H36M dataset, and evaluate them using H36M, Humaneva-I and 3DHP. 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 5, none of the baselines work well in cross-scenario situations while the Ray3D shows good generalization performance in H36M, Humaneva-I and 3DHP dataset. For instance, PoseFormer [10] is able to predict better root-relative pose than Ray3D, but it struggles to predict precise root joint. And PoseLifter [2] fails to generalize to cross datasets, achieving inferior MRPE performance.

Evaluation with noisy cameras To test the robustness of Ray3D when taking noisy cameras parameters as input, we add gaussian noise to intrinsic parameters (i.e., focal length and center points) and extrinsic parameters (i.e., rotation and translation) of H36M's cameras respectively.

The results of using noisy focal length and center points as input are shown in Fig. 1 and Fig. 2. As for the intrinsic parameters, Videopose [5], PoseFormer [10] and RIE [6] do not use focal length and center points as input, noisy intrinsic parameters has no impact on these methods. While Ray3D and Poselifter [2] explicitly decouple the intrinsic parameters from the input. Noisy intrinsic parameters cause inaccurate decoupling, which results in slight performance changes.

As for the extrinsic parameters, Videopose [5], PoseFormer [10] and RIE [6] use extrinsic parameters to convert final estimation from CCS to WCS, noisy extrinsic parameters cause performance degradation. Ray3D uses the well calibrated camera extrinsic parameters as an input, especially the camera height and camera pitch, which makes Ray3D sensitive to camera pitch change. As shown in the Fig. 3, after we added gaussian noise to camera pitch, the MPJPE increases from 81.2mm to 110.6mm. Fig. 4 shows the performance with noisy camera yaw, Ray3D does not decrease significantly. As shown in Fig. 5, all methods have

Table 2. Quantitative evaluation results under MPJPE on H36M using detected 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. * means this approach using 2D keypoints detected by CPN. Best results are shown in **bold**.

Detected keypoints as input		Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Somke	Wait	WalkD.	Walk	WalkT.	Average.
Dabral et al. [3]	ECCV'18	44.8	50.4	44.7	49.0	52.9	61.4	43.5	45.5	63.1	87.3	51.7	48.5	52.2	37.6	41.9	52.1
Cai et al. (f = 7) [1]	ICCV'19	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
Videopose. (f = 9)* [5]	CVPR'19	46.4	48.9	45.7	49.8	52.5	61.5	47.7	46.8	59.9	68.1	50.7	47.5	52.7	38.4	42.1	50.6
Yeh et al. [9]	NIPS'19	44.8	46.1	**43.3**	46.4	49.0	**55.2**	44.6	44.0	58.3	62.7	47.1	43.9	48.6	32.7	33.3	46.7
UGCN (f = 96) [7]	ECCV'20	**41.3**	**43.9**	44.0	**42.2**	**48.0**	57.1	**42.2**	**43.2**	57.3	**61.3**	**47.0**	**43.5**	**47.0**	**32.6**	**31.8**	**45.6**
PoseFormer (f = 9)* [10]	ICCV'21	47.9	51.5	49.3	50.8	53.7	58.6	49.5	46.6	62.0	70.3	52.6	49.3	53.8	40.5	43.0	52.0
PoseAug (f = 1)* [4]	CVPR'21	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	52.9
RIE (f=9)* [6]	ACMMM'21	44.8	47.9	46.1	47.4	50.4	57.6	45.7	44.6	**57.0**	64.2	49.5	45.7	50.9	36.6	39.8	48.6
Ray3D (f = 9)*		44.7	48.7	48.7	48.4	51.0	59.9	46.8	46.9	58.7	61.7	50.2	46.4	51.5	38.6	41.8	49.7

Table 3. Quantitative evaluation results under Abs-MPJPE and MRPE on H36M using GT as 2D 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**.

Abs-MPJPE		Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Somke	Wait	WalkD.	Walk	WalkT.	Average
Videopose (f = 9) [5]	CVPR'19	73.7	99.6	88.8	82.8	**81.7**	121.8	89.8	83.8	110.6	234.4	95.8	92.4	91.2	69.7	64.2	98.7
PoseLifter (f = 1) [2]	ICCV'19	65.5	86.9	103.9	81.4	95.2	109.2	80.1	107.3	152.4	245.0	106.2	95.6	115.5	87.1	69.8	106.8
PoseFormer (f = 9) [10]	ICCV'21	88.3	88.3	91.3	94.3	96.1	127.8	101.0	120.0	114.5	227.7	102.4	110.8	97.2	99.1	91.1	111.6
RIE (f = 9) [6]	ACMMM'21	75.4	90.7	**80.5**	80.9	75.3	100.4	85.9	92.2	**93.1**	200.9	**86.5**	87.9	88.5	67.8	58.6	91.0
Ray3D (f = 1)		**60.2**	75.2	102.0	70.6	92.5	**85.2**	71.7	**67.5**	123.9	129.5	87.0	77.6	92.7	74.0	67.7	85.2
Ray3D (f = 9)		65.6	**70.4**	100.1	**64.1**	92.0	86.6	**65.6**	73.2	119.2	**117.4**	92.9	**70.1**	**77.1**	**64.4**	**61.4**	**81.4**
MRPE		Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Somke	Wait	WalkD.	Walk	WalkT.	Average
Videopose (f = 9) [5]	CVPR'19	57.6	88.0	77.2	69.4	74.2	110.3	71.4	73.3	97.0	225.9	86.7	77.5	80.9	61.2	52.2	86.9
PoseLifter (f = 1) [2]	ICCV'19	51.3	75.6	87.8	67.9	83.0	96.3	63.8	100.0	138.6	231.6	93.5	83.8	108.4	73.1	51.1	93.7
PoseFormer (f = 9) [10]	ICCV'21	63.2	**63.2**	77.4	77.4	84.3	114.6	76.8	103.1	96.5	215.8	88.0	90.2	85.5	89.3	78.0	95.9
RIE (f = 9) [6]	ACMMM'21	60.6	78.3	**69.5**	69.5	**65.1**	90.6	68.3	81.5	**79.1**	192.1	**76.2**	73.6	80.2	59.5	**48.1**	79.5
Ray3D (f = 1)		**45.4**	63.4	97.7	57.5	88.0	**74.4**	**53.4**	59.4	116.9	119.1	79.8	60.9	85.5	64.8	56.1	**74.9**
Ray3D (f = 9)		59.3	65.4	99.8	**55.1**	93.1	80.5	55.2	**70.9**	116.4	**104.6**	89.9	**59.8**	**70.3**	**56.8**	52.4	75.3

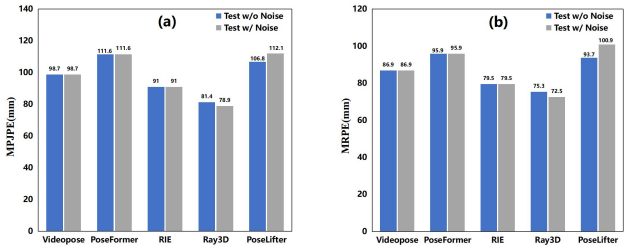

Figure 1. Performance under MPJPE and MRPE with noisy focal length are plotted in (a) and (b) respectively.

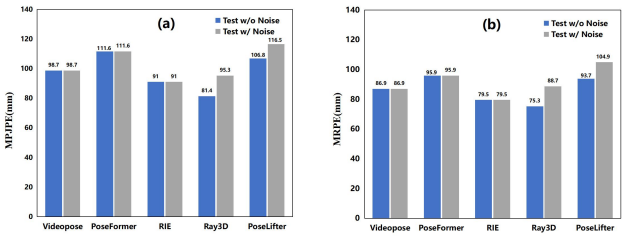

Figure 2. Performance under MPJPE and MRPE with noisy center points are plotted in (a) and (b) respectively.

the same performance drop when provided with noisy camera translation.

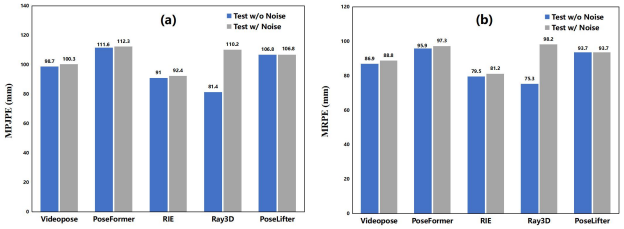

Figure 3. Performance under MPJPE and MRPE with noisy camera pitch are plotted in (a) and (b) respectively.

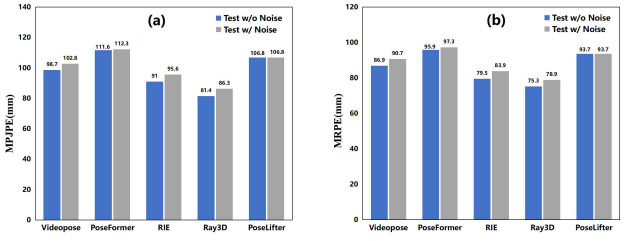

Figure 4. Performance under MPJPE and MRPE with noisy camera yaw are plotted in (a) and (b) respectively.

3.2. Evaluation on synthetic benchmarks

Integrate Videopose with Ray3D We integrate proposed Ray3D techniques with another baseline method Videopose [5]. The model is trained and evaluated on the pro-

Table 4. Quantitative evaluation results under MPJPE, Abs-MPJPE and MRPE on 3DHP using GT as 2D 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**.

method \metric	MPJPE	Abs-MPJPE	MRPE
Videopose (f = 9) [5]	52.5	148.4	145.8
PoseFormer (f = 9) [10]	**40.8**	147.8	147.5
PoseLifter (f = 1) [2]	78.2	143.6	129.1
RIE (f = 9) [6]	47.4	140.8	141.0
Ray3D (f = 1)	48.4	118.2	114.0
Ray3D (f = 9)	46.0	**103.4**	**95.8**

Table 5. Cross dataset evaluation. We adopt a 14-joint skeleton training on H36M, testing on H36M, Humaneva-I and 3DHP datasets. MPJPE, Abs-MPJPE and MRPE are adopted. (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. 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) [5]	46.1	133.4	120.9	85.1	284.6	283.1	104.6	1262.8	1266.1
PoseFormer (f = 9) [10]	50.0	146.6	129.8	**79.4**	260.7	253.9	**101.9**	1313.6	1320.1
PoseLifter (f = 1) [2]	56.9	147.4	135.1	3690.2	15170.6	16082.6	1180.9	6839.0	6899.4
RIE (f = 9) [6]	41.5	136.4	125.1	82.0	272.9	285.1	102.4	1185.0	1187.0
CDG (f = 1) [8]	52.0	-	-	-	-	-	111.9	-	-
Ray3D (f = 9)	**39.3**	**106.8**	**98.5**	81.5	**121.5**	**99.3**	108.1	**422.4**	**406.0**

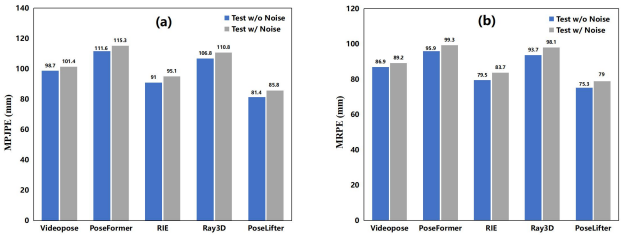

Figure 5. Performance under MPJPE and MRPE with noisy translation are plotted in (a) and (b) respectively.

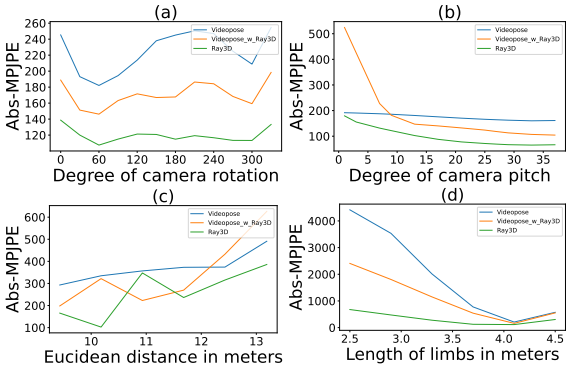

Figure 6. Figures (a), (b), (c) and (d) showcase the performance using Abs-MPJPE metric in case of rotation, camera pitch, translation and body scale variations correspondingly. The x-axis denotes the degree of camera rotation, the degree of camera pitch, euclidean distance between camera and subject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

posed synthetic dataset. As shown in the Fig. 6, Videopose [5] integrated with 3D ray representation and camera embedding techniques performs better than vanilla method under Abs-MPJPE metric. Same performance gain can be observed in the Fig. 7 under MRPE metric, which showcases that Ray3D incorporated to the different existing frameworks bring consistent improvement.

Intrinsic generalization As shown in the Fig. 8 (a) and (b), principal point changes affect VideoPose, PoseFormer, RIE to varying degrees under MPJPE and MRPE metrics respectively. In contrast, both Ray3D and Ray3D_w/o_CE achieve stable result. This result clearly showcases the merits of our ray-based input representation.

4. Qualitative results in WCS

In this section, we provide qualitative results generated by Ray3D and other state-of-the-arts on H36M and 3DHP datasets. Specifically, we visualize 3D keypoints in WCS generated by corresponding methods.

H36M Fig. 9 shows qualitative comparison of Ray3D with VideoPose, RIE and PoseFormer on H36M. We train all

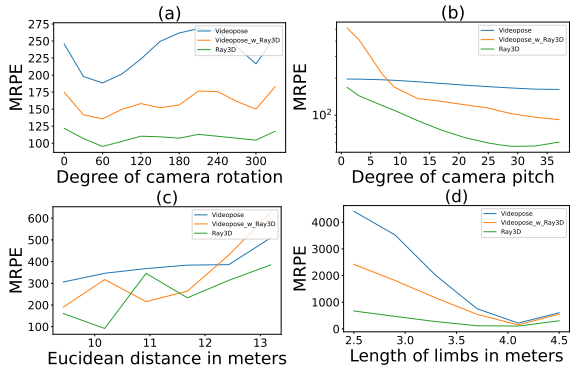

Figure 7. Figures (a), (b), (c) and (d) showcase the performance using MRPE metric in case of rotation, camera pitch, translation and body scale variations correspondingly. The x-axis denotes the degree of camera rotation, the degree of camera pitch, euclidean distance between camera and subject in meters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

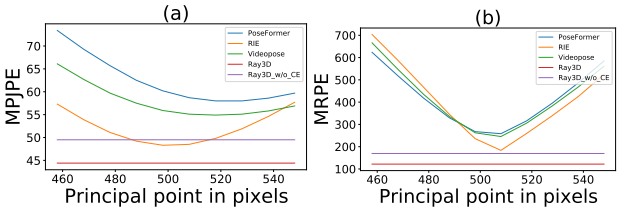

Figure 8. Performance under MPJPE and MRPE in case of principal point changes are plotted in (a) and (b) respectively. The x-axis represents x-coordinate of 2D principal point of the virtual camera in pixels.

four models on H36M with 17-joint definition. From the visualization, we can observe that Ray3D has superior ability to generate more precise location of root joint with comparable root-relative pose estimation results. Fig. 10 presents two examples of inferior estimation of Ray3D compared to baseline, yet the error is close among these methods.

3DHP In Fig. 11, we present the qualitative comparison of Ray3D with VideoPose, RIE and PoseFormer on 3DHP as well. The models are trained with 14-joint definition. Our Ray3D shows better performance than other state-of-the-arts clearly.

Cross-dataset In Fig. 12, we compare generalization of Ray3D with other state-of-the-arts on H36M. We train all four models on 3DHP and test them on H36M with 14-joint definition. One can clearly observe that Ray3D generates more accurate estimation results than other approaches, benefiting from our normalized ray representation and camera embedding design.

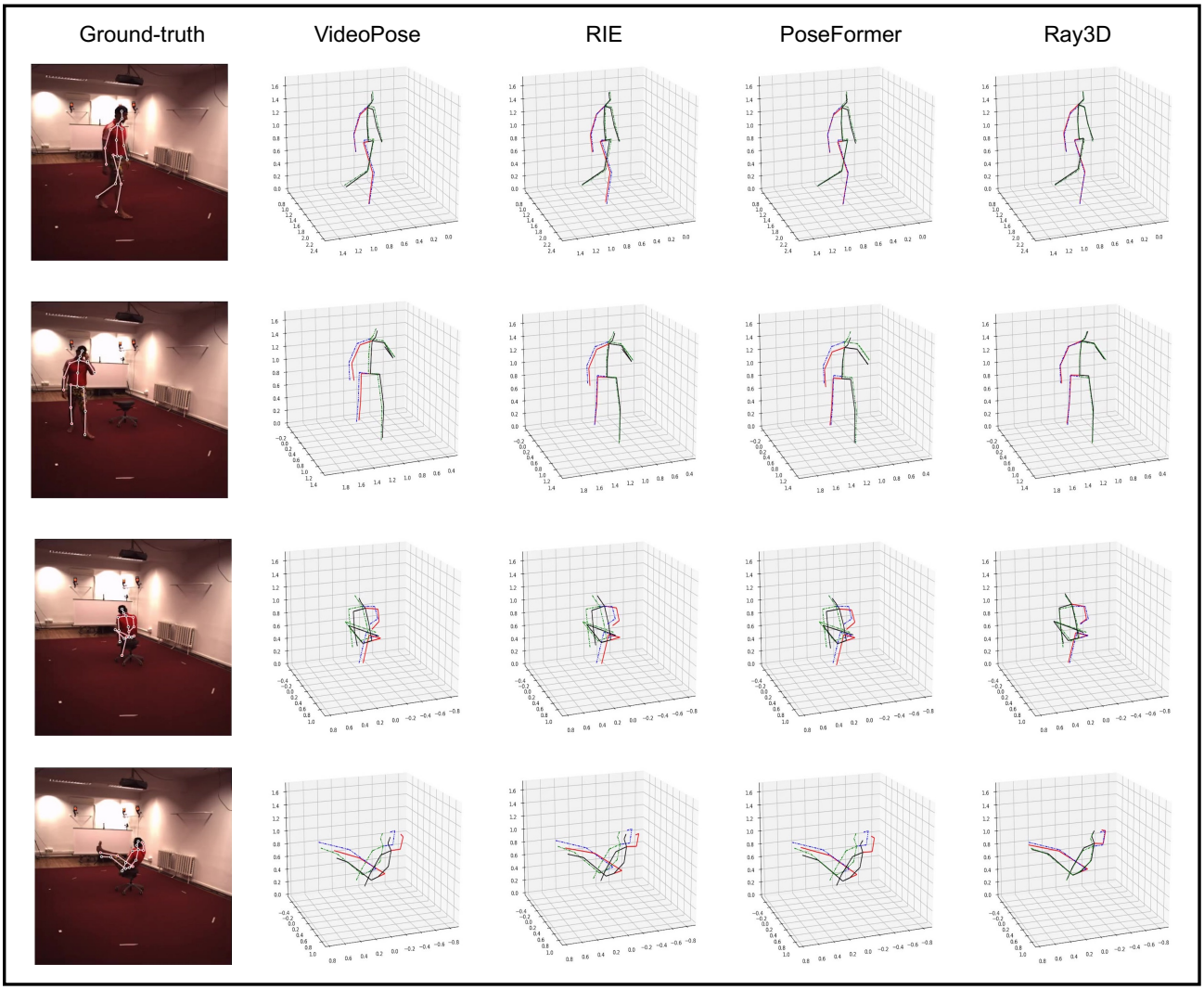

Figure 9. Qualitative comparison of Ray3D with VideoPose, RIE and PoseFormer on H36M. All four models are trained on H36M. First column shows 2D ground-truth poses. Black color denotes left part of person limbs, red color denotes right part of person limbs. 3D estimation results predicted by VideoPose, RIE, PoseFormer and Ray3D are shown in the second, third, forth and fifth column respectively. Dashed lines denote 3D ground-truth poses. Solid lines represent the poses estimated by corresponding approaches. Green and black color lines denotes left part of person limbs, blue and red lines denote right part of person limbs. 17-joint skeleton is visualised.

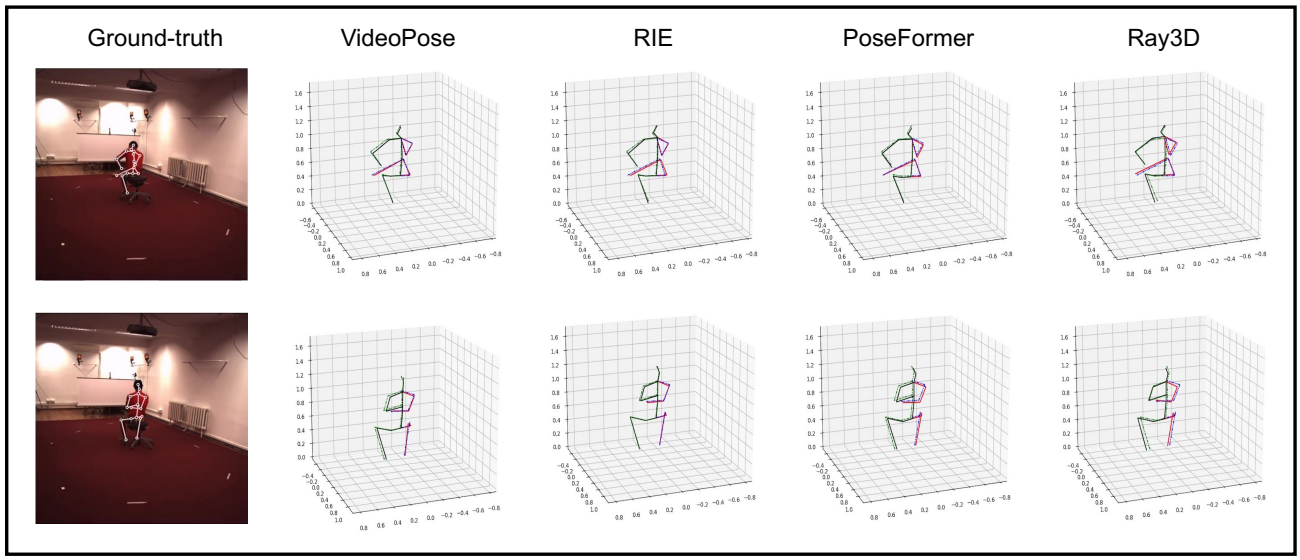

Figure 10. Visualization of inferior performance of Ray3D, compared with other state-of-the-arts on H36M. All four models are trained on H36M. First column shows 2D ground-truth poses. Black color denotes left part of person limbs, red color denotes right part of person limbs. 3D estimation results predicted by VideoPose, RIE, PoseFormer and Ray3D are shown in the second, third, forth and fifth column respectively. Dashed lines denote 3D ground-truth poses. Solid lines represent the poses estimated by corresponding approaches. Green and black color lines denotes left part of person limbs, blue and red lines denote right part of person limbs. 17-joint skeleton is visualised.

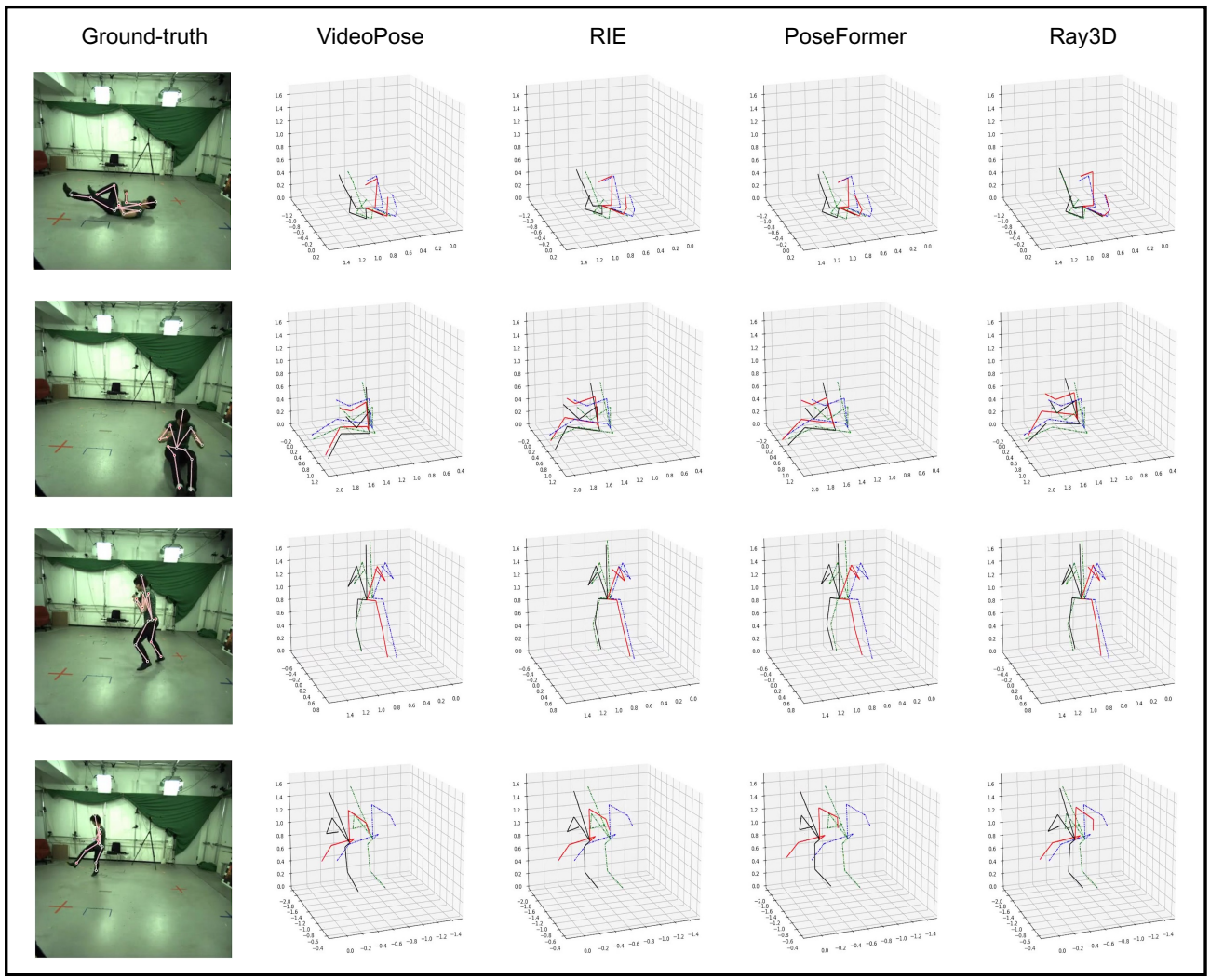

Figure 11. Qualitative comparison of Ray3D with VideoPose, RIE and PoseFormer on 3DHP. All four models are trained on 3DHP. First column shows 2D ground-truth poses. Black color denotes left part of person limbs, red color denotes right part of person limbs. 3D estimation results predicted by VideoPose, RIE, PoseFormer and Ray3D are shown in the second, third, forth and fifth column respectively. Dashed lines denote 3D ground-truth poses. Solid lines represent the poses estimated by corresponding approaches. Green and black color lines denotes left part of person limbs, blue and red lines denote right part of person limbs. 14-joint skeleton is visualised.

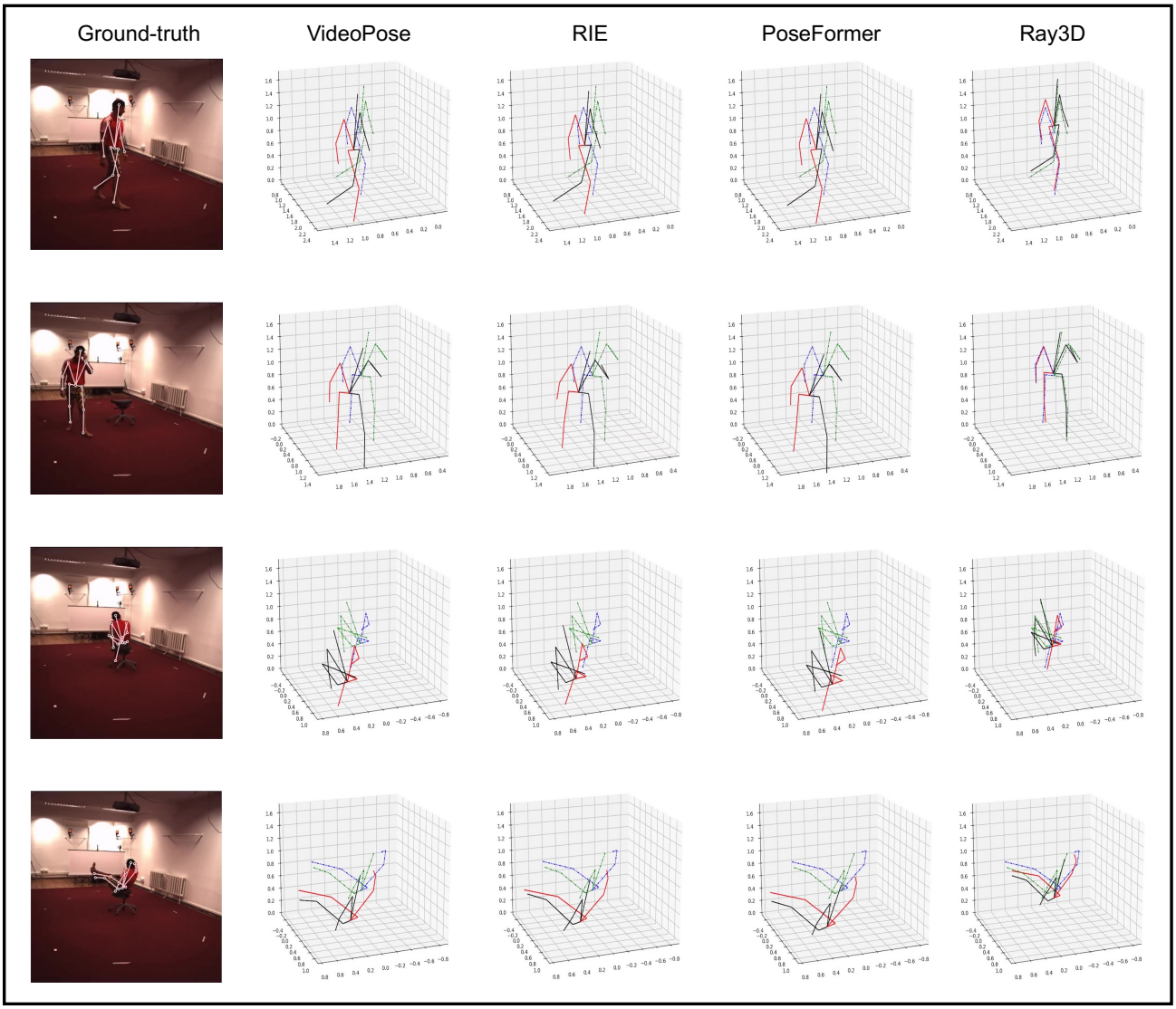

Figure 12. Visualization of generalization of Ray3D, compared with other state-of-the-arts on H36M. All four models are trained on 3DHP. First column shows 2D ground-truth poses. Black color denotes left part of person limbs, red color denotes right part of person limbs. 3D estimation results predicted by VideoPose, RIE, PoseFormer and Ray3D are shown in the second, third, forth and fifth column respectively. Dashed lines denote 3D ground-truth poses. Solid lines represent the poses estimated by corresponding approaches. Green and black color lines denotes left part of person limbs, blue and red lines denote right part of person limbs. 14-joint skeleton is visualised.

References

[1] 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. 3

[2] 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. 2, 3, 4

[3] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–683, 2018. 3

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

[5] 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, 4

[6] 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

[7] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision*, pages 764–780. Springer, 2020. 3

[8] 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. 4

[9] Raymond Yeh, Yuan-Ting Hu, and Alexander Schwing. Chirality nets for human pose regression. *Advances in Neural Information Processing Systems*, 32:8163–8173, 2019. 3

[10] 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, 3, 4