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

1. Introduction

In this supplemental document, we provide more implementation details and discuss the limitation of TRACE. Please refer to the **supplemental video** for more qualitative results and video sequences.

2. Implementation Details

In this section, we introduce the details of our image loss functions, temporal feature propagation module, inference and training details, and dataset details.

2.1. Loss Functions

Except for the temporal motion losses we introduced in the main paper, we also follow previous work [9, 15, 16] to supervise the estimated maps and SMPL parameters with standard image losses. $\mathcal{L}_{cm}$ is the focal loss [15] of the estimated 3D Center map. We also supervise the focal loss of the 2D Body Center heatmap, which is used for BEV-based 3D composition of the 3D Center map. The SMPL parameter loss $\mathcal{L}_{pm}$ consists of four parts, $\mathcal{L}_{\boldsymbol{\theta}}, \mathcal{L}_{\boldsymbol{\beta}}$, $\mathcal{L}_{prior}$, and $\mathcal{L}_{\boldsymbol{J}}$. $\mathcal{L}_{\boldsymbol{\theta}}$ and $\mathcal{L}_{\boldsymbol{\beta}}$ are L_2 losses on SMPL pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ parameters respectively. In the first 20 epochs of training, we also employ the Mixture of Gaussian pose prior [2, 11], $\mathcal{L}_{prior}$, to penalize the unreasonable poses $\boldsymbol{\theta}$.

The 3D body keypoint loss $\mathcal{L}_{\boldsymbol{J}}$ consists of 3 parts, $\mathcal{L}_{mpj}$, $\mathcal{L}_{pmpj}$, and $\mathcal{L}_{pj2d}$. $\mathcal{L}_{mpj}$ is the L_2 loss of 3D body keypoints $\boldsymbol{J}$. Following [15–17], we employ $\mathcal{L}_{pmpj}$ to alleviate the domain gap between training datasets, which is the L_2 loss of the predicted 3D body keypoints $\boldsymbol{J}$ after Procrustes alignment with the ground truth. We also employ $\mathcal{L}_{pj2d}$ to learn from 2D pose datasets, which is the L_2 loss of the 2D projection $\boldsymbol{J}_{2D}$ of 3D joints $\boldsymbol{J}$ using predicted 3D position $\widehat{\boldsymbol{t}_{\boldsymbol{i}}}$ in camera coordinates.

2.2. Temporal Feature Propagation Module

To extract long-term and short-term motion features, we construct a temporal feature propagation module by combining a ConvGRU [18] module, a Deformable convolution [23] module, and a residual connection. Inspired by [18], we employ a ConvGRU module to hold the long-term memory within a hidden state $\boldsymbol{H}_{i-1}$ and progressively update the memory with image feature map $\boldsymbol{F}_i$ to generate the new hidden state $\boldsymbol{H}_i$. Inspired by [1], to extract short-term motion features, we use $\boldsymbol{F}_i - \boldsymbol{F}_{i-1}$ to generate the feature sampling offset for Deformable convolution to process $\boldsymbol{F}_i$. With these modules, we progressively fuse the image feature maps ($\boldsymbol{F}_{i-1}$, $\boldsymbol{F}_i$) to generate a temporal image feature map $\boldsymbol{F}'_i$.

During inference, we sequentially process every clip of a video sequence. With a GRU-based temporal propagation module, the size of the video clips can be flexibly set.

2.3. Inference and Training Details

Inference details. The input images are resized to 512×512. The size of output maps is $D = 64, H = 128, W = 128, C = 128$. The hyper-parameters of the memory unit are set to $\lambda_c = 0.05, \lambda_s = 0.13, \lambda_d = 0.05, \lambda_m = 1, W = (1.2, 2.5, 25), \lambda_f = 100$. Following [10, 15, 16], we adopt the 6D representation [22] for SMPL pose parameters $\boldsymbol{\theta}$. To improve the temporal smoothness, we also employ the One-Euro filter [3] to process the estimated SMPL pose parameters and 3D trajectories $\widehat{\boldsymbol{t}}, \boldsymbol{T}$.

Training details. The confidence threshold of the Body Center heatmap is 0.12. The sampling ratios of different datasets are 18% Human3.6M [6], 20% MPI-INF-3DHP [13], 14% 3DPW [19], 12% PennAction [20], 16% PoseTrack [4], 20% DynaCam. $w_{(.)}$ denotes the corresponding weight of all loss items. We set these loss weights to $w_m = 300, w_W = 100, w_{mpj} = 300, w_{cm} = 100, w_{pmpj} = 260, w_{pj2d} = 300, w_{prior} = 1.6, w_\theta = 80, w_\beta = 60$.

2.4. Dataset Details

In this section, we introduce the details of used datasets. First, we introduce the 3D human pose datasets, 3DPW [19], Dyna3DPW [19], Human3.6M [6], and MPI-INF-3DHP [13].

3DPW is a multi-person in-the-wild video dataset. Videos in 3DPW are captured by tracking subjects with a dynamic camera to record their daily activities, which

Sequence Name	Frame Ranges
downtown_warmWelcome_00	0-588
downtown_weeklyMarket_00	0-1192
downtown_bus_00	0-800
office_phoneCall_00	0-879
downtown_walkBridge_01	156-1371
downtown_runForBus_01	300-485
downtown_sitOnStairs_00	84-918
downtown_bar_00	66-256
downtown_cafe_00	128-736
downtown_enterShop_00	0-1448
downtown_rampAndStairs_00	0-219
downtown_runForBus_00	88-350
downtown_windowShopping_00	0-1750
downtown_crossStreets_00	152-587
downtown_car_00	0-576
downtown_walking_00	0-1386

Table 1. Video sequences of Dyna3DPW.

matches our task. 3DPW contains rich human activities, such as shopping in market, running for the bus, walking on crowded bridge, and working in an office. 3DPW provide accurate 3D human pose and mesh annotations of the tracked subjects. Note that only the one or two subjects are labeled and all other people appearing in the images are not Since the global trajectory annotations are not accurate, we evaluate the 3D human pose and shape estimation on the test set of 3DPW.

Dyna3DPW is a subset of 3DPW test set that we constructed. Not all 3DPW videos have complete tracking annotations due to occlusion. Consequently, we select the 16 video sequences in Tab. 1 that are complete and call this subset Dyna3DPW. There are many challenging scenes in Dyna3DPW, such as pedestrians walking past the camera, occlusion between tracked subjects, object occlusion, truncation, and sudden changes of direction. We use this challenging subset to evaluate tracking and HPS accuracy in complex scenes with a moving camera.

Human3.6M [6] and **MPI-INF-3DHP** [12] are single-person 3D pose video datasets. They are captured in constrained experimental environments using static multi-view cameras. They provides 3D pose annotations for each frame. We sample every 5 frames to reduce redundancy. We use their training set for training.

We also use two 2D human pose datasets, PennAction [21] and PoseTrack [4]. **PennAction** contains 2326 video sequences of 15 different actions. Most videos in PennAction contain only a single person with 2D pose annotations. Part of videos containing multiple people while only a specific subject has 2D pose annotations. **PoseTrack** is an in-the-wild 2D video dataset for multi-person pose tracking. It contains 763 video sequences and provides 2D pose and ID annotations. We use these for training.

HMR [9]	CRMH [7]	ROMP [15]	TRACE
56.8	52.7	50.2	**42.0**

Table 2. Comparisons on Human3.6M using Protocol 2 of HMR [9] in PAMPJPE.

CRMH [7]	ROMP [15]	Pose2UV [5]	TRACE
102.6	72.0	69.5	**38.2**

Table 3. Comparisons on 3DMPB [5] in PAMPJPE.

ID switches	MOTA	IDF1	HOTA
1	96.1	98.0	64.8

Table 4. Tracking results on CMU Panoptic [8].

3. Experiments with Static Cameras

As a sanity check, we also evaluate the 3D human pose estimation and tracking with static cameras. On indoor Human3.6M, we follow the evaluation protocol 2 of HMR [9]. As results shown in Tab. 2, compared with previous SOTA methos, TRACE achieves comparable results. On in-the-wild 3DMPB [5], we follow Pose2UV [5] to report PAMPJPE in Tab. 3 and achieve new SOTA results. We evaluate tracking on severe occlusion scenes from CMU Panoptic [8], selected by CRMH [7]. As shown in Tab. 4, TRACE achieves promising results. These results testify the performance of TRACE with static cameras.

Runtime efficiency. TRACE runs at 32.2 FPS on MuPoTS while using 6.9GB of memory on a 4090 GPU. TRACE takes 174.1G FLOPs, which includes the image backbone, motion backbone, and heads, accounting for 41.0G, 66.4G, and 66.7G, respectively.

4. More Ablation Studies

Memory unit. Without the memory unit, on MuPoTS/Dyna3DPW, ID switches increase 31/94 and MOTA, IDF1, and HOTA decreases 6.5%/11.1%, 12.0%/49.1%, and 8.0%/39.6% respectively. These results demonstrate that the memory unit is an essential module for persistent tracking under occlusion. It enables that the predicted 3D motion offset can associate the occluded objects in memory with new detections.

Backbones. To enable end-to-end 5D Representation Learning (5DRL), we develop Temporal Feature Propagation (TFP) as a differentiable path for modeling temporal image features. On 3DPW, 5DRL with TFP only (w/o the optical flow from the motion backbone) outperforms BEV [16] by 18.1% in terms of PAMPJPE (38.4mm). Simultaneously learning temporal motion features from the motion backbone further reduces PAMPJPE by 1.6% (37.8mm).

5. Discussion of Limitations

As ours is the first method to perform one-stage global 3D human motion and trajectory estimation from videos captured by dynamic cameras, TRACE has some limitations that could be further explored in future work.

Due to the lack of training data for this task, we make restrictive assumptions to facilitate effective training. For example, we assume all videos are captured by a camera with fixed field of view (FOV=50 degree), without shot changes [14]. Additionally, due to the low diversity of human body shapes in the training datasets, 3D human body shape estimation is limited. Collecting or synthesizing a new dataset with sufficient diversity in camera and human body shape could potentially solve these problems. We also assume the camera coordinate of the first frame of a input video as the world coordinates. Future work should work on estimating the camera poses (e.g. pitch, yaw, and roll) from the monocular video to convert the results of TRACE to the real world coordinates. Here we focused on the problem of tracking specific subjects that are specified in the first frame. Our method, like BEV [16], actually finds all the people in the image but the tracking branch filters out all but the subjects. Future work should explore extending TRACE to track everyone in the scene.

References

[1] Gedas Bertasius, Christoph Feichtenhofer, Du Tran, Jianbo Shi, and Lorenzo Torresani. Learning temporal pose estimation from sparsely-labeled videos. *NeuIPS*, 2019.

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

[3] Géry Casiez, Nicolas Roussel, and Daniel Vogel. 1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2527–2530, 2012.

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

[5] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Pose2UV: Single-Shot Multiperson Mesh Recovery With Deep UV Prior. *TIP*, 2022.

[6] 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.

[7] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, pages 5579–5588, 2020.

[8] 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 *ICCV*, pages 3334–3342, 2015.

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

[10] 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.

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

[12] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved cnn supervision. In *3DV*, 2017.

[13] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, pages 120–130, 2018.

[14] Georgios Pavlakos, Ethan Weber, Matthew Tancik, and Angjoo Kanazawa. The one where they reconstructed 3d humans and environments in tv shows. In *ECCV*, 2022.

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

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

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

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

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

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

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

[22] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019.

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