Supplementary Material for the paper: "Tracking People by Predicting 3D Appearance, Location and Pose"

Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Jitendra Malik
UC Berkeley

1. Introduction

In this document, we provide more details that were not included in the main manuscript, due to space constraints. We include more implementations details about our approach (Sections 2). We provide more details about the experiments of our paper (Sections 3). Finally, we extend the discussion about the failure cases of our system (Sections 4). Additionally, we encourage the readers to also watch the attached supplementary video, which is also available here: https://brjathu.github.io/PHALP.

2. Implementation details

Architecture: First, we provide some additional architectural details about the networks used in our pipeline. Regarding the HMR module, the architecture is similar to [8]. For the mask conditioning, we use the detections and masks from MaskRCNN [3] as a masking operation to mask out features that do not belong to the person of interest. This masking operation does not require any extra parameters, and it only acts on the last feature map of the convolutional part of ResNet. For the appearance head of HMAR, we use the same design as [9]. The only difference is that for the texture encoder, the input has four channels (RGB & mask), where the mask is used to indicate locations on the body that have not been visible during the video, thus invalid. Finally, for the HMMR part, we use a transformer similar to [8], with one layer and one head. The functionality is similar to the original HMMR [5], but for the future poses, instead of regressing a residual on the parameter space (θ, β), the residual is on the feature space.

Training: Next, we provide more details of the training procedure. First we train the HMR model, followed by the appearance head of HMAR and then the temporal head of HMMR model. For the HMR model, we follow the training details of [8], with the additional modification of using the mask conditioning part. Once, the HMR model is trained, we use it as the backbone for the training of HMAR model. For HMAR, we apply the estimated texture on the SMPL body and project it to the image. Then, we optimize the loss such that the texture of the projection matches the actual RGB values on the *segmented (MaskRCNN) pixels*. We train this model for 10000 iterations with an Adam optimizer with an initial learning rate of 0.001. HMAR is trained on a per frame basis. Finally, for the HMMR model, we use data from Human3.6M [4], InstaVariety [5] and AVA [2], where we use the pseudo-ground truth by [8]. We train the HMMR head for 10000 iterations with a learning rate of 0.0001. For the training and testing time of PHALP, we refer to Table 1. Note that our method can be optimized further for faster inference.

LMC: We observe that most of the failures of our tracking system are caused by missing detections or out-of-distribution poses in the videos. While, solving the detection or human pose reconstruction is not the scope of this paper, we propose a simple solution to overcome these failures at the tracking stage. For each new tracklet, we wait for seven frames for this tracklet to accumulate enough information. Then, we use the seven detected locations and their corresponding time-steps and regress k frames back into the past. The value k is determined by the old tracklets which have not been updated for $k + 7$ frames. Once we have the prediction of the past location for the new tracklet, we measure 3D location distance between old tracklets and predicted past location. Apart from location, we also measure appearance similarity between the old and new tracklets, considering the appearance does not change significantly over time. Finally, our cost function for LMC involves location distance and appearance distance and we solve it via Hungarian to assign new tracklets to old tracklets. For some qualitative examples of the effect that LMC has in tracking, please see Figure 1.

3. Experimental details

Comparison with baselines: For the evaluation on PoseTrack [1], MuPoTS [7] and AVA [2], we follow the test protocols of Rajasegaran *et al*. [9]. To evaulate the different methods, we only reject the detections if the IOU distance is zero. However, the non-rejected detections will have to compete for the ground-truth via a Hungarian matching algorithm. This is to avoid penalizing the methods based on

Figure 1. *LMC results:* We show two failure cases of our method. In the first example, MaskRCNN fails to detect the person (pointed in red) for a long period of time. This causes the error in location prediction and therefore when this person is detected again, a new tracklet is created. In the second example, HMR fails when the 3D pose is very different and causes a large cost in the pose distance. Due to this, a new tracklet is created for the same person. With LMC, we are able to look back and check whether these new tracklets can be connected with any old tracklets. If so, we connect them together as a single tracklet.

	Training time (days, single GPU)			**Test time (fps)**		
Model	HMR	HMAR	Temporal	Detection	3Dify	Track
[32]	5	3	0.5	∼7	∼26	∼2
PHALP	5	3	0.2	∼7	∼26	∼9

Table 1. **Computational budget for PHALP.** We report training and inference times of our method and previous 3D tracking method [32] on a single NVIDIA 2080 GPU. For a fair comparison we use the online setting of [9]. Our method has similar run times as [9].

their quality of detections.

Robustness: We evaluate the robustness of appearance aggregation, by adding pixel noise to the visibility masks. Adding noise even up to 90% of the pixels increases the IDs metric on PoseTrack only by 4%, meaning that appearance aggregation is quite robust. Moreoever, we are not so prone to drifting due to occlusions, because the appearance of a body part will not update when this body part is not visible. Specifically, ee observe that we are robust to the value choice of α_0, with only $\pm 2\%$ change in the IDs metric on PoseTrack for values of $\alpha_0 \in [0.1, 0.9]$.

Design choices: In Section 4 of the main manuscript, we included an investigation on some design choices of the proposed pipeline. The detailed results for those ablations are presented in Table 2.

Number of people: We also test the effect of the number of people in a video on the tracking metrics. Interestingly, we observe that, the IDs metric scales almost linearly with the number of people. This suggests that PHALP can work on crowded scenes. Please see Table 3.

Method	PoseTrack		
	IDs↓	MOTA↑	IDF1↑
PointRend Mask	558	58.9	76.2
PARE [6] Location	512	58.8	76.3
PHALP+LMC	520	58.9	76.3
PHALP	541	58.9	76.4

Table 2. **Ablation of different design choices for PHALP.** Due to the modular architecture of PHALP, we can replace different components of it at different stages. We evaluate PHALP with replacing MaskRCNN with PointRend, and HMAR location with PARE [6] location. We also evaluate PHALP+LMC, where we allow new tracklets to connect with old tracklets. This flexibility allows us to overcome the mistakes made at the detection stage and HMR stage.

# of people	1-2	3-4	5-6	7-8	9-10	≥10
Avg IDs↓	1.32	1.62	2.32	4.37	5.61	4.69
MOTA↑	19.24	52.81	47.79	62.98	66.08	65.96
IDF1↑	73.93	76.47	77.28	77.82	75.16	77.93

Table 3. **Effect of number of people in video.** We group the posetrack sequences based on the number of ground truth tracks and report report Avg IDs, MOTA, and IDF1 on each subgroup.

4. Failure cases

Our method, PHALP, relies on 1) MaskRCNN masks and 2) HMAR pose. Most of our failure cases can be attributed to mistakes in these two methods. For example,

non-maximum suppression for Mask RCNN is imperfect. This will create two detections for a single person or give a joint mask of two people. These masks, will hurt pose estimation, and then the location of the person followed by bad appearance representation. This will eventually affect the tracking performance too. On the other hand, for challenging cases, HMR can also give bad SMPL reconstructions. For example, when a person is heavily occluded, the number of visible pixels is very small and this will affect the pose prediction of HMR. Although PHALP depends on these two methods, the robust line fitting, averaging the appearance and pose smoothing with the HMMR model can recover from occasional failures of these methods.

References

[1] 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 *CVPR*, 2018. 1

[2] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 1

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 1

[4] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 2013. 1

[5] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. 1

[6] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 2

[7] 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*, 2018. 1

[8] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human mesh recovery from multiple shots. In *CVPR*, 2022. 1

[9] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3D representations. In *NeurIPS*, 2021. 1, 2