HaMeR: Hand Mesh Recovery for the EgoExo4D Hand Pose Challenge

Georgios Pavlakos
UT Austin
pavlakos@cs.utexas.edu

Dandan Shan
University of Michigan
dandans@umich.edu

Ilija Radosavovic
UC Berkeley
ilija@berkeley.edu

Angjoo Kanazawa
UC Berkeley
kanazawa@berkeley.edu

David Fouhey
New York University
david.fouhey@nyu.edu

Jitendra Malik
UC Berkeley
malik@eecs.berkeley.edu

Abstract

This report summarizes the approach of our team, HaMeR, for the EgoExo4D Hand Pose Challenge. The primary component of our approach is our recently introduced hand mesh recovery approach, HaMeR. We apply HaMeR out-of-the-box on the images of the EgoExo4D Challenge, and we observe very strong performance. Additionally, we further finetune the model, using the hand pose annotations from EgoExo4D. Finally, we experiment with an ensemble including the baseline approach of the EgoExo4D benchmark. Our overall submission placed 2nd in the EgoExo4D Hand Pose Competition.

1. Introduction

The EgoExo4D Hand Pose Challenge focuses on the task of 3D hand pose estimation from egocentric camera views. This is the first egocentric dataset that provides 3D ground truth and is captured in diverse environments, *i.e.*, in non-studio settings. The dataset of the challenge includes over 8k image examples with 3D hand pose ground truth coming from multi-view annotations (one egocentric and a number of exocentric cameras). The 3D pose ground truth is in the form of 3D keypoints for 21 hand joints.

For the challenge, we developed a solution based on our recent HaMeR method [12]. HaMeR focuses on the problem of 3D **Ha**nd **Me**sh **R**ecovery from a single image. We apply HaMeR out-of-the-box on EgoExo4D and we observe very strong performance. We further finetune HaMeR on data from EgoExo4D. Our final solution is based on an ensemble of these two models and the baseline POTTER model [16], which is also trained on EgoExo4D. Our submission is ranked 2nd in the leaderboard of the EgoExo4D Hand Pose Challenge. In this report, we describe our approach, we present an ablation for the different components we used and we provide extensive qualitative evaluation, including both successful reconstructions and failure cases.

2. Preliminaries

The main component of our approach is the HaMeR network [12]. HaMeR is a recent state-of-the-art model for Hand Mesh Recovery. It is a feedforward model that takes as input a single image of a hand and hand side (left or right), and estimates a 3D reconstruction of the hand in the form of the MANO parametric hand model [13]. For HaMeR, we adopt a fully transformerized architecture design using a ViT-H backbone [2, 15], followed by a transformer head. The transformer head regresses the parameters of the MANO model, *i.e.*, hand pose θ and and hand shape β, as well as the camera parameters π that allow us to project the hand to the image. We train HaMeR with a combination of 3D supervision losses (when 3D ground truth is available), and 2D supervision losses, using 2D keypoint annotations.

Besides using a large scale transformer model, the other big advantage of HaMeR comes from training on large scale datasets. More specifically, HaMeR is trained on over 4M images coming from a diverse range of datasets, including datasets with 3D annotations collected in a studio setting, *e.g.*, FreiHAND [17], InterHand2.6M [10], HO3D [6], DEX YCB [1], as well as datasets with 2D annotations on in-the-wild images, *e.g.*, COCO WholeBody [7], Halpe [3], MPII NZSL [14]. As we show in the original paper [12], the scale of the training data, along with the large scale architecture are the key contributors to HaMeR's performance.

3. Approach and Results

Here, we present our approach for the challenge submission. As already mentioned, our approach is based on the default HaMeR model, and we used the model that we have made publicly available. We observed (Table 1) that this model already outperforms the baseline model based on

POTTER [16], which is trained using data from EgoExo4D. One key difference of the baseline model is that it directly regresses 3D keypoint locations, while we regress MANO parameters. This can make it easier to overfit to specific camera settings (*i.e.*, particular set of intrinsics), while we train a model using a general focal length value to accommodate intrinsics from different datasets. As a result, we are capturing very accurately the local pose of the hand (PA-MPJPE metric), but the pose is not estimated accurately in the camera frame (which is what the MPJPE metric is capturing). This is a common observation for the parametric pose estimation methods (see also the discussion in the HMR paper [8]). In fact, we observed that with a very simple optimization of the orientation and the translation of the hand, we were able to get that error significantly decreased, yet not as low as the baseline. For this optimization, we used the 2D keypoints estimated by HaMeR, and we optimize the translation and rotation of the hand, such that it minimizes the reprojection error, measured under through the ground truth camera parameters.

Our next step was to create an ensemble of our model with the challenge baseline [16]. To transform both estimates in the same coordinate frame, we align the two sets of 3D keypoints using Procrustes Alignment. Then, we simply average the 3D coordinates for each joint. This ensemble gave a clear boost in the performance of both metrics.

Finally, we also considered training HaMeR using data from EgoExo4D. For simplicity, we only used the 2D keypoint labels, since from our initial experiments, it was not easy to use the 3D keypoint labels. These tend to be more noisy and not directly correspond to a valid hand geometry. An ideal setting would be to try fitting MANO to the 3D keypoints and using the MANO parameters as supervision, which is common strategy for methods on parametric pose estimation [4, 9, 11]. This preliminary experiment gave us a small boost for the PA-MPJPE metric which was the primary metric of the competition. The ensemble of these three models forms our final submission for the challenge (Table 2), where we ranked 2nd.

Qualitative results Next, we present a number of qualitative results of our approach on the EgoExo4D test set. Figure 1 shows a number of interesting successes, while Figure 2 shows some representative failures. In general, we observe that HaMeR is robust across a wide setting, including challenging hand poses, interactions with various objects, occlusions, truncations, different skin colors, different lighting conditions and examples where the hands wear gloves. However, there are still some limitations, particularly when the occlusions/truncations are very extreme (only a few visible fingers), the wrist location or hand orientation is ambiguous, and finally, if the original hand bounding box crop is not very accurate.

Method	MPJPE ↓	PA-MPJPE ↓
POTTER	**28.94**	11.07
HaMeR	76.95	10.36
HaMeR-align	36.75	10.36
HaMeR + POTTER (ens)	29.18	9.32
HaMeR + POTTER + HaMeR-ft (ens)	30.52	**9.30**

Table 1. **Ablation of our Model on the ExoExo4D Hand Pose Challenge test set (errors in mm).** We start with using our HaMeR model out-of-the-box that has never been trained on EgoExo4D. This outperforms the POTTER baseline on the PA-MPJPE metric, but the MPJPE metric is very high. This metric can be improved by optimizing the rotation and translation of the hand using the ground-truth image intrinsics (HaMeR-align). Since we do not update the hand pose, the PA-MPJPE metric stays the same. By creating an ensemble of HaMeR with the POTTER baseline, we achieve significant improvements for the PA-MPJPE metric (fourth row). Finally, by finetuning HaMeR on EgoExo4D and using this model in the ensemble, we get some minor improvements for the PA-MPJPE metric (fifth row). This last version corresponds to our Challenge submission.

Method	MPJPE ↓	PA-MPJPE ↓
PCIE_EgoHandPose	25.51	8.49
Ours	30.52	9.30
Death Knight	28.72	10.20
IRMV_sjtu	29.38	10.36
Baseline POTTER	28.94	11.07

Table 2. **Leaderboard of the EgoExo4D Hand Pose Challenge.** We present the top-5 approaches, ranked by the primary PA-MPJPE metric. Our approach based on HaMeR is ranked 2nd.

4. Conclusion

We presented our approach that achieved the 2nd place at the EgoExo4D EgoPose Hand Challenge. Our solution is based on HaMeR [12], our state-of-the-art model for hand pose estimation. We experiment with different versions, including finetuning on the EgoExo4D dataset [5] and creating an ensemble with the baseline POTTER model [5, 16]. As is common across the literature in this space, our parametric approach achieved very solid results on the local pose estimation (PA-MPJPE), but was inferior compared to the approaches regressing 3D keypoints (*e.g.*, POTTER baseline). Future work can consider the better integration of the EgoExo4D data (*e.g.*, creating pseudo ground-truth labels for training), as well as make better use of the intrinsics for pose inference (currently, the poses are estimated under a very extreme focal length).

References

[1] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl

Figure 1. **Successful reconstructions of HaMeR on EgoExo4D.** We observe that HaMeR is robust across a variety of settings found in EgoExo4D, including more challenging poses, interactions with different objects, heavy occlusions and truncations, different skin colors, challenging lighting conditions and hands covered by gloves.

Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

[3] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. *PAMI*, 2022.

[4] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *ICCV*, 2023.

[5] Kristen Grauman, Andrew Westbury, Lorenzo Torresani,

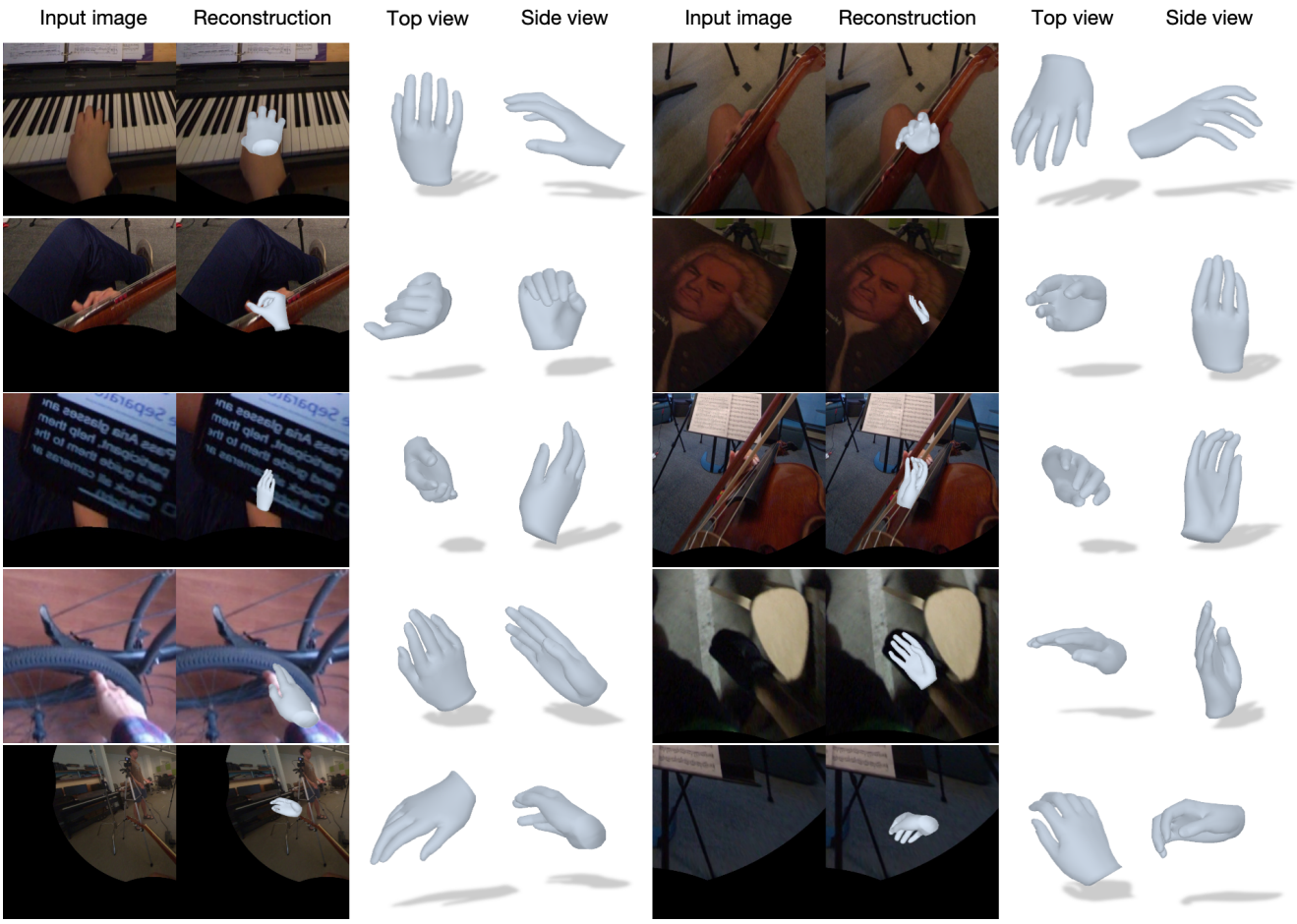

Figure 2. **Failure cases of HaMeR on the EgoExo4D test set.** We observed that despite its robustness, HaMeR can occasionally fail on the EgoExo4D examples. This can be often attributed to ambiguous wrist location, ambiguous orientation, extreme occlusions or truncations, as well as cases with non-centered or incorrect hand bounding box annotations.

Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-Exo4D: Understanding skilled human activity from first-and third-person perspectives. 2024.

[6] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, 2020.

[7] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020.

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

[9] 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*, 2019.

[10] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV*, 2020.

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

[12] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. 2024.

[13] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017.

[14] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[15] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 2022.

[16] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. POTTER: Pooling attention transformer for efficient human mesh recovery. In *CVPR*, 2023.

[17] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, 2019.