Human Pose Estimation from Egocentric Social Interaction Videos

Qi Ma
D-MAVT, ETH Zurich
qimaqi@ethz.ch

Rui Wang
D-ITET, ETH Zurich
ruiwang46@ethz.ch

Yelan Tao
D-MAVT, ETH Zurich
yeltao@ethz.ch

Abstract

The pose estimation with an egocentric perspective is important for the autonomous robot, augmented reality and health care. However the dynamic movement, self-occlusion and self-motion in the first-person view cause poor performance for the standard pose and shape estimator like Easymocap. In this work, we integrate past and future information through PoseForecast module in TCMR. Moreover, we design our custom regressor for keypoints estimation and also do extensive ablation study about different Pose Initialization strategies. We achieved amazing performance compared to the YOU2ME original work, which formulates camera wearer pose estimation as a classification task. Lastly, we fit the SMPL model based on estimated keypoints and gain smooth and accurate results compared to running shape estimation directly.

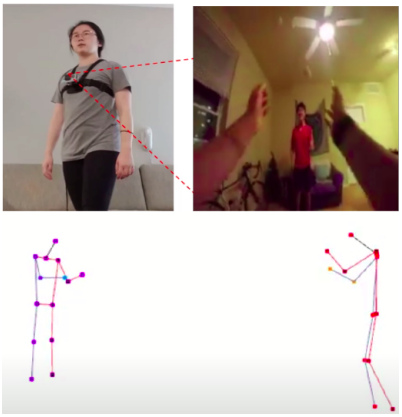

Figure 1. Visualization of YOU2ME dataset, top is camera location and frame, downside is groundtruth captured by panoptic Studio or Kinect

1. Introduction

Various approaches have been proposed in this area, including model-free and mode-dependent methods. For the first one most approach formulate the problem as simplistic skeleton estimation in 2D or 3D which is not trivial due to the depth ambiguity, limited data with ground truth and complexity of human articulations [5]. Model-based work like [13] regresses the model parameters from the input images. This method can also be used to regress shape parameters and camera intrinsics. To address the temporal inconsistency and unsmooth 3D motion along one sequence, video-based methods like [10] are proposed. However, it faces the trade-off between per-frame accuracy and temporal consistency [5]. Moreover, this method requires high-quality images and sometimes even multi-view data and it uses mostly the fixed camera which is not in line with many real-world applications like AR devices or autonomous mobile robots. The interaction between humans and robots is not always relatively static and it involves also self-occlusion by hand and missing target in the process.

Work [15] propose a way to utilize the interaction information and estimate the human poses. It will boost performance for certain interactions since the motion pattern can be exploited and inferred. For example if the camera wearer is shaking hands, then even though we have no information about the person we are interacting with, we can also infer that his pose is probably also shaking hands. The first-person view will improve the understanding of egocentric physical surroundings, making it possible for learning and interpreting interaction in a human context—attentive to high-level social behaviors. Nevertheless, the YOU2ME dataset is very challenging because of the low resolution (227x227). Thus the original paper formed it as a classification problem and their attempts to regress the human pose leads to the poor result. But 3 years later, can we use a more advanced method as TCMR [5] and easymocap [1] to do pose estimation, or even extend to SMPL model regression? Moreover, the TCMR is the current SOTA method in human pose estimation from video, where they take static feature sequence as input and build temporal features, integrating past, current and future features using attention after MLP. This approach overcomes the trade-off between per-frame accuracy and temporal consistency and is used to be the main part and baseline of our method.

Unlike the TCMR original work, we design our cus-

tom regressor for keypoints location regression since the YOU2ME dataset does not have shape parameter ground truth. We observed that encoding the egocentric pose information will significantly improve the performance, particularly for the cases when the interaction between camera wearer and interactee is highly related. Furthermore, we show that pose initialization is also highly effective in the custom regressor. The attention mechanism improves the per frame accuracy and also helps to stabilize the result. We show the detailed ablation of various 3D video benchmarks study with different input strategies and network structures in the custom regressor.

Our contributions can be summarized as follows. • We predict the interatee body pose and overcome challenge of YOU2ME dataset by encoding temporal feature and camera wearer's pose information.

• We conducted extensive experiments and ablation studies and show the influence of network architecture, temporal encoding module, input types.

• We predict accurate and consistent pose and shape compared to SOTA by implementing the easymocap [1] style tracking and enforce smoothness constrains between frame.

• We clean the YOU2ME raw dataset, resolving the issue of inverse images and interactive object, and missing ground truth. Make this data be better available in the future.

2. Related Work

Video-based 3D human pose and shape estimation Estimating human pose through video is a very popular area of research especially after boosting of deep learning. HMMR [9] learns temporal encoder which increases consistency between past and future frames. VIBE [10] proposed a bi-directional gated recurrent unit GRU unit and implement adversarial structure to leverage large dataset.

Egocentric video

The topic of egocentric frame plays an important role in robotic field and VR/AR. Recent work pays more attention to recognition tasks like object detection, human feature like face, arm hands recognition [16]. Similarly we focus on whole human body 3D pose estimation with additional input from camera wearer.

Learning from interactions Currently most work focus on generating smooth and realistic human motion but neglect the significance of corelations between human and humans. Work [18] achieve good improvements compared to previous approaches on generating natural and physically plausible human motion in a scene by bridging human motion synthesis and scene affordance reasoning. The interactive understanding between people and people is not only simple motion pattern combinations. It is also convolve to specific emotion, scenarios, social behaviors etc [8]. By exploiting the interaction information it offers widespread improvement in human-object interactions [3] [14], hand detection [12] [2], and also social interactions understanding [19] [7]. For body pose estimation the work [15] infer the first-person joints position by learning high-level cues from the interactions but it formulates the problem as pose classification due to dataset limitations. In contrast our work implement recursive regressor using both feature in the past and future.

3. Method

3.1. Problem formulation

The goal is to estimate human keypoints position given 2D image sequences. Input: N video frames from a chest-mounted camera. Output: N 3D poses and the output i $p_i \in R^{3J}$ where J is joints number for estimated human skeleton. In this work we choose J = 14 as same joints type between panoptic and kinect. Additionaly we have access to the pose information of camera wearer as 1 shows. This is in line with the actual situation because we can read pose information form AR device or a mobile robot.

3.2. Data preprocessing

YOU2ME dataset is challenging because of low resolution, occlusion, and target out of track. It contains usual interactions like a conversation, throwing and catching, sports, and patty. To ensure high performance we first go through all raw data and find out some flaws like reversed image, missing ground truth, and mislabeled ground truth between wearer and interactee. After dataset cleaning, we precompute the static features using pretrained ResNet from [11] based on the data normalization. Since the CMU 3D keypoints position uses unit meter while Kinect is already scaled from 0 to 1, we divide the CMU dataset by 1000 and normalize all data to canonical direction by computing transforming using the first frame.

3.3. Pipeline

Overview. We show the overall pipeline of our method in Figure **2**, which consists of four major components: temporal fusion encoder, pose initialization part, custom regressor and SMPL Model inference.

Temporal feature encoder. The temporal feature encoding part is highlighted as yellow in the figure, and has the same structure as the TCMR [4] temporal encoding module, taking 8 past frames and 8 future frames, outputting a temporal feature that holds an attention-based convex combination of past feature and future feature. This structure helps the model avoid a strong preference for the current static feature and improve the temporal consistency through frames. Our experiment shows that the indicated

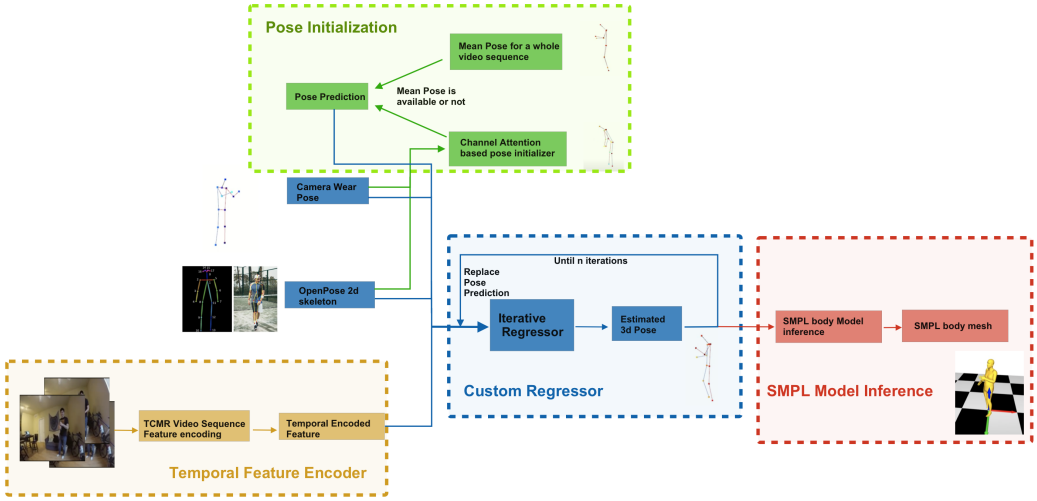

Figure 2. Overview of our method. Our pipeline consists of four part: pose initialization, temporal feature encoder, custom regressor and SMPL regressor, which are highlighted in green, yellow, blue, red, respectively. The pipeline takes the temporal feature obtained by temporal encoder with image sequences, egocentric feature and openpose 2d skeleton as input, outputing the 3d body joints of the interactee and the SMPL body mesh.

temporal feature could boost the accuracy of pose prediction.

Custom regressor. Our custom regressor, which is highlighted in the blue part, takes the temporal feature from the temporal encoder along with openpose feature and camera wearer pose as input and predicts the interactee body pose. The indicated component adopts an iterative regression method to better utilize the feature without making the network too complicated. At each iteration, we take the new pose prediction of the previous iteration as a new initialization. After reaching the desired number of iterations. The regressor itself consists of four fully connected layers with dropout and the ReLU activation function. In addition, as our iterative regressor features a residual connection between input and output, we find that the initialization would affect the overall accuracy and stability of our regressor, which leads to our third part, pose initialization.

Pose initialization. This part is highlighted in green, where we adopt two initialization methods. The first method is that we use the mean pose of each sequence to initialize the pose, the indicated method is effective in terms of boosting the accuracy and alleviating the large variance issue with the YOU2ME dataset. However,it is a quite strong assumption as the mean pose is not easily accessible. Therefore, we also implemented a channel-wise attention-based method taking the camera wearer pose and open pose 2d skeleton as an initialization method. w

SMPL body regressor. The last part is the SMPL body regressor, which adopts the method from [1] [6]. It filters the tracked 3d body pose jittering with smoothness constrains and interpolate between missing frames. It firstly estimate the body shape parameters globally towards each sequence, then optimize the body pose parameters with a LGBFS solver so that the SMPL body joints align with the 3d joints.

4. Experiment

Evaluation metrics. We evaluate our pipeline using the per-frame and temporal metrics. To report the per-frame accuracy we adopt per joint position error (MPJPE), Procrustes-aligned MPJPE (PA-MPJPE). Regarding the temporal evaluation we calculate the average difference between the groundtruth and predicted acceleration [9].

Dataset. YOU2ME dataset is captured by two method: Panoptic studio and Kinect captures. Our approach is evaluated on both: For Panoptic dataset we train on 11 sequences and evaluate on 3 sequences. For Kinect dataset we train on 17 sequences and evaluate on 4 sequences. We split the dataset in this way so that each set contains all basic interaction and the data balance is ensured. in test clips do not appear in the training set.

Pose Initialization During the experiment, we find that different pose initialization methods would have an impact on the final reconstruction precision. We show the comparison of four different initialization methods on Kinect validation dataset in Figure **3**: zero initialization, MLP initialization, attention-based initialization and per sequence mean pose initialization. As the YOU2ME groundtruth coordinate system varies from sequence to sequence, naively initializing a global mean pose would lead to a poor result. As we can see from the indicated figure, zero initialization leads to the highest error. And the attention-based initializa-

tion method, which consists of one channel-wise attention module and three fully connected layers with ReLU activation function, taking the egocentric pose and openpose keypoints as input, achieves a smaller error. As a comparison, we test an MLP of the same number of layers which yields a larger error compared to the attention-based method. It proves the effectiveness of the channel-wise attention module which holds a convex combination of the input features and amplifies the contribution of the important ones. Ultimately, we test using the per-sequence mean pose initialization, it could alleviate the inherent large variance issue with the YOU2ME groundtruth. However, it is a strong assumption that the mean pose of a video sequence is known.

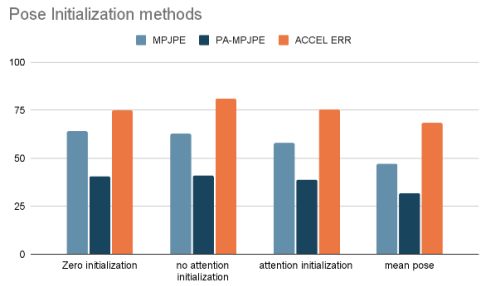

Figure 3. Investigation on different pose initialization methods, where we investigate zero initialization method, MLP based initialization method, attention based initialization method, and per-video sequence mean pose initialization method.

Network structure discussion As shown in Figure 4, we investigate the different network architectures. The baseline which is the original TCMR [4] structure, would fail completely on the YOU2ME dataset with large groundtruth variance during to limited expressiveness with SMPL model constrain. Both fully connected structure and temporal encoded network adopt our custom regressor and achieve better accuracy in all metrics, proving the effectiveness of our custom regressor on indicated dataset. Moreover, we further investigate the temporal encoder architecture. The performance increases with the usage of the temporal feature which proves the effectiveness of the temporal encoding module.

Ablation study on input type One of our key assumptions is that the egocentric feature of the camera wearer could boost the inference accuracy of the pose estimation for interactee in a social interactive setting. Therefore, We show the impact of different input types on both Kinect and Panoptic validation dataset in Table 1. We have tested three different input settings, only using the temporal encoded feature, temporal feature with the egocentric feature, and temporal feature, egocentric feature with openpose 2d keypoints. As our selection is performed with PA-MPJPE, we can infer from the table that the input with the temporal

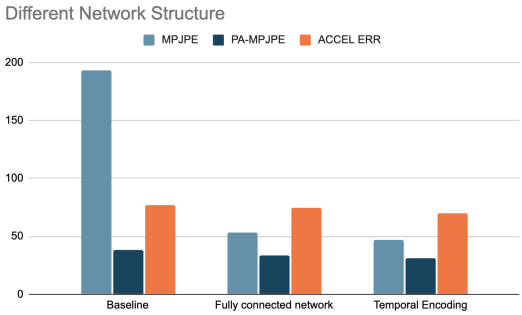

Figure 4. Investigation the performance of different architecture, the baseline is the original TCMR [4] method. The fully connected network is only using static feature and our custom regressor. The temporal encoding network stands for the usage of temporal encoding architecture and our custom regressor.

feature, egocentric feature and openpose keypoints achieve the best PA-MPJPE by a noticeable margin on both Kinect and Panoptic dataset. However, we do observe some performance drop in terms of acceleration error and MPJPE. The reason for this observation may be the growing dominance of static information with more added static features such as egocentric feature and openpose feature. Overall, the indicated issue could be alleviated in model selection, compromising PA-MPJPE to some extent for a more balanced overall model.

Dataset	Input	MPJPE↓	PA-MPJPE↓	AE↓
Kinect	S	**47.36**	32.02	69.27
	S+E	47.64	31.81	**65.62**
	S+E+2d	48.02	**31.58**	66.00
CMU	S	21.20	9.11	**13.69**
	S+E	**21.12**	8.78	14.07
	S+E+2d	22.61	**8.02**	16.70

Table 1. Investigation on different input types of the model on kinect and CMU dataset. Note that S stands for sequence feature encoded by TCMR style temporal encoder, E stands for egocentric body pose, 2d stands for the openpose keypoints feature input. MPJPE stands for the mean per join position error. PA-MPJPE stands for the procrustes aligned mean per joint position error. AE stands for the acceleration error.

Comparison of different interaction. In this part we analyze the per-frame pose accuracy over different interaction sequences. We show the result from both Kinect and Panoptic datasets. Additionally, we also compare the result between different input combinations: whether to take egocentric information as input. The positive impact on per-frame accuracy of adding egocentric pose can be shown in Table 2 3. However the acceleration error increase when taking the camera wearer pose as input. This means en-

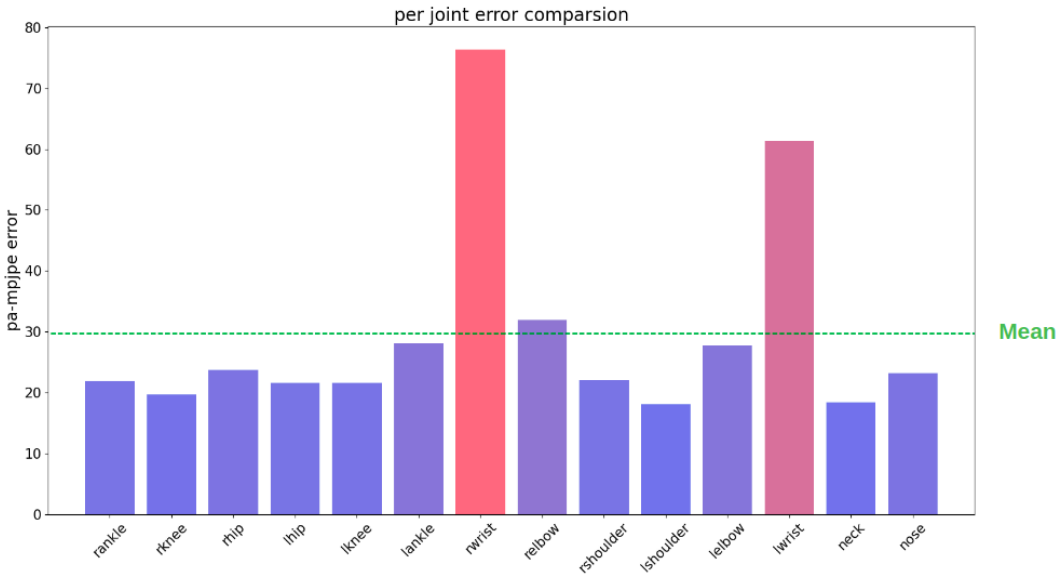

Figure 5. Performance comparison between different joints location

coding egocentric body pose into feature will consistently improve per-frame metrics but introduce more temporal inconsistency. The reason for this result may be the dominance of the current static feature, which is also in line with work [5]. Adding current frame pose information will potentially bring negative influence to the model and can prevent the temporal encoder from fully utilizing temporal information in the past and future frames [5]. Moreover, we can also see from the Table that the performance improvement by adding the pose is different from basic interaction. For motion patterns such as conversation and sports, the actions are not related and can be very different. But for the movement like throwing and catching or hand clapping the motion is highly related and similar. This explains the observation that the PA-MPJPE drops 15.5% for patty sequences but only 2.7% for conversation sequences. Lastly, the high-speed movement will bring more challenges to the pose estimation. The conversation sequence in 2 and sports sequence in 3 have larger action velocity and acceleration. Thus, it is difficult for the temporal encoding because of the large frame difference and contradiction that most human joints are close to stationary within the sequence.

Comparison of different joints. We visualize the PA-MPJPE from Kinect throwing and catch sequences between different joints, since it has relatively more movement and henceforth more challenging. As the figure 5 shows the PA-MPJPE error distribution is quite different over the joints location. The error in the wrist is much higher than in other parts. The same observation can be found in the elbow due to mostly hand movements in this sequence. The lower half body has lower errors in general since it is stationary most of the time, and because of the similarity between interactee and the camera wearer.

Kinect	Vel	Accel	PA-MPJPE↓	Accel-Error↓
catch*	6.08	75.43	29.70	86.09
catch			35.03	**61.41**
patty*	5.31	73.80	**23.02**	146.28
patty			27.24	65.64
convo*	**8.24**	**135.11**	28.61	112.9
convo			29.41	92.13

Table 2. Comparison of performance on different interaction with egocentric information using Kinect dataset, Note that * stands for sequence feature encoded by egocentric body pose.

CMU	Vel	Accel	PA-MPJPE↓	Accel-Error↓
convo*	3.15	29.14	8.00	14.58
convo			8.78	**18.41**
sports*	**6.74**	**54.96**	9.47	31.20
sports			9.52	39.06
hand*	6.60	40.85	**7.75**	18.55
hand			8.97	31.19

Table 3. Comparison of performance on different interaction with egocentric information using panoptic dataset, Note that * stands for sequence feature encoded by egocentric body pose.

Qualitative result. We provide the qualitative results in the Figure 6, 7, which consists of two parts. The upper part presents the result from the Kinect dataset, we randomly choose 3 frames from 4 basic interaction sequences: standing, throwing and catching, conversation, and hand motion.

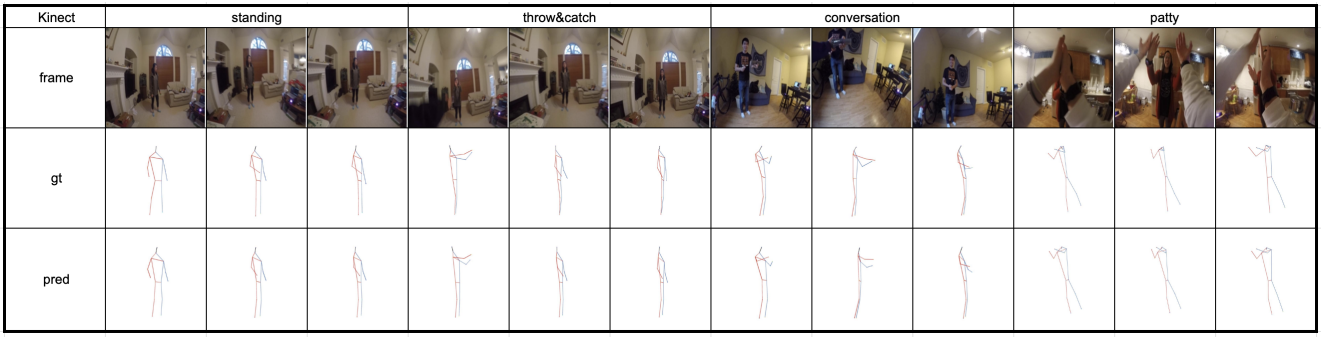

Figure 6. Qualitative result of Kinect dataset for different inter-person interactions

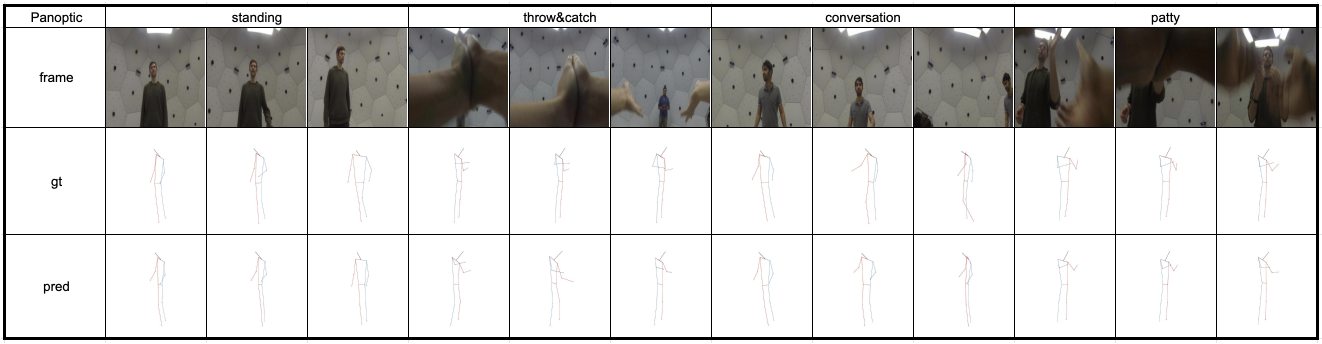

Figure 7. Qualitative result of Panoptic dataset for different inter-person interactions, You can find more result in the page dynamic demo sequence

The second part shows the result from CMU which is captured by the panoptic studio. Compared to the first one, the CMU dataset has less training data but is more accurate and clean. All result comes from the best performing model where 2d and egocentric pose is used. The results are rendered in open3d and the two sides of the body are colorized with red and blue. As both figures shows when standing still the estimation is accurate. Moreover, we can see that the information on the lower half of the body is beneficial for predicting the lower pose comparing between stand and patty in 6. However, the estimation is also quite good without clearly lower body input, this can be explained as the use of past and future information, and the input of egocentric pose information. This again underlines the importance of egocentric information because the model can learn how to infer the interactee pose from interactions. In addition, the motion pattern patty or clapping involve a lot of self-occlusion but we can see the prediction is also good. This is because that the action of clapping hands is fast so the future and past frames information is valuable in estimating current pose. Similar results can be observed in throwing movement. Unlike the other motion pattern, in conversation sequence the interactee movement is not related to egocentric movement too much and the hand movement is more complex than other basic motion. In the 7 we can observe more occlusion cases, such as hand occlusion and out of view cases. The model accurately learned how to extract information from previous and future frames and even though some prediction have larger difference compared to ground truth, action characteristics are successfully estimated. Moreover, a large misalignment can be seen in the last column of standing part. This is possible caused by the fact that the interactee have relative larger body shape and even though we normalize the data the larger width of human body trunk is not well learned.

Failure cases. As we discussed above that the YOU2ME dataset is challenging for low-resolution, self-occlusion, and missing targets. Here we visualize two of the failure cases in figure 8. As the top figure shows, the self-occlusion by hand movement completely obstructs the view, therefore the prediction is quite different from the ground truth. However, this happens only when the self-occlusion period is too long and when the temporal encoding of past and future are not in agreement with current poses. This is also the reason that the estimated pose remains the hand-raising pose before the occlusion happens. The bottom figure shows another potential case in real-world interaction such as when people bend down to pick something up or look away. In such cases, the person interacting will go out of sight for a while and there are no clues for estimating the pose at all.

Similarly, as long as this missing target situation does not last too long, our approach can learn relatively correct pose predictions from past and future features, although the input in this figure is only feet, the movement pattern of the upper body: raising hands, has been accurately estimated.

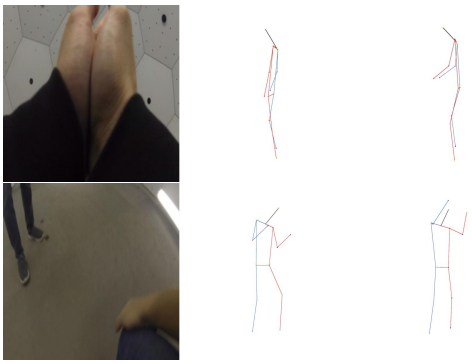

Figure 8. Failure cases of self-occlusion(top row) and missing target(bottom row), left column is input, middle is ground truth and right is estimated poses of interactee

Comparison with Frankmocap We show the side by side comparison of our method with frankmocap [17] in Figure 9. Our method with temporal encoder, egocentric feature and openpose 2d feature outperforms the frankmocap method by a margin on this social interactive dataset.

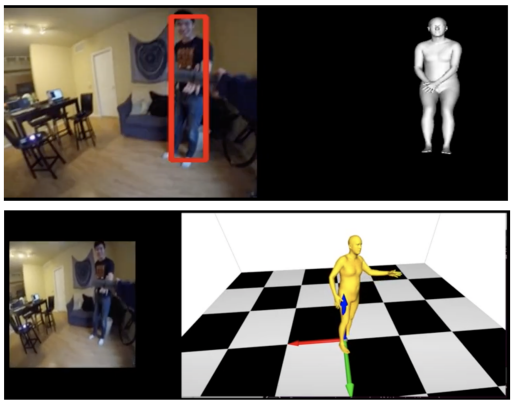

Figure 9. Side by side comparison of frankmocap [17] (top row) and ours (bottom row). You can find the demo videos in: fankmocap evaluation (https://www.youtube.com/watch?v=xwIF0trOvVs&t=29s and ours (https://www.youtube.com/watch?v=cr4gBjVR5lc.From the figures, we can see that the confusion caused by temporary occlusion is resolved by our tempral feature and could yield correct pose estimation (arm reaching out). Moreover, by enforcing the smoothness constraint, we get rid of the pose jittering, which is recommended to be viewed in the demo video .)

5. Summary

5.1. Conclusion

- We revisited You2Me dataset with "modern methods" as temporal encoding, SMPL body model and channel-wise attention methods. Compared to the original YOU2ME method, we aim to predict the interactee body pose with large variance, which was not feasible when YOU2ME was proposed.
- We performed detailed experiments and ablation studies on YOU2ME dataset in regard to network architecture, temporal encoding module, input type as egocentric feature and openpose. We also conducted a detailed analysis towards the prediction error of different actions.
- We finished the SMPL body mesh prediction pipeline, incorporating the easymocap [1] style tracking and smoothness constrains between frame. Our pipeline gives accurate, smooth and realistic prediction compared to popular repos as Frankmocap [17].

5.2. Limitations and future Work

- Due to the limitation of input resolution and lack of camera extrinsic information, we couldn't perform the SMPL mesh reprojection. It is possible to use more advanced social interation dataset to evaluate the performance of interatee pose estimation and SMPL mesh recovery.
- It is possible to investigate training a proper human pose and motion regularizer in social interaction scenarios to further constrain the predicted human pose and make it more realistic.

6. Contributions of team members

- **Rui Wang** was mainly in charge of the following tasks throughout this project: investigation on pose initialization methods, network structure modifications, network training, frankmocap evaludation, making demo videos, composition of slides and report.
- **Qi Ma** was mainly responsible for data cleaning and preprocessing, baseline modification and training, adding input combinations, making qualitative result demo and error distribution analysis in this work.
- **Yelan Tao** was mainly responsible for literature survey, data cleanning and demo preparation.

References

[1] Easymocap - make human motion capture easier. Github, 2021. 1, 2, 3, 7

[2] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE international conference on computer vision*, pages 1949–1957, 2015. 2

[3] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *Robotics: Science and Systems*, volume 3. Ann Arbor, Michigan;, 2016. 2

[4] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1964–1973, 2021. 2, 4

[5] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Beyond static features for temporally consistent 3d human pose and shape from a video. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 1, 5

[6] Junting Dong, Qi Fang, Wen Jiang, Yurou Yang, Hujun Bao, and Xiaowei Zhou. Fast and robust multi-person 3d pose estimation and tracking from multiple views. In *T-PAMI*, 2021. 3

[7] Alircza Fathi, Jessica K Hodgins, and James M Rehg. Social interactions: A first-person perspective. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1226–1233. IEEE, 2012. 2

[8] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2

[9] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 2, 3

[10] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 1, 2

[11] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2252–2261, 2019. 2

[12] Cheng Li and Kris M Kitani. Model recommendation with virtual probes for egocentric hand detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2624–2631, 2013. 2

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

[14] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. pages 8687–8696, 10 2019. 2

[15] Evonne Ng, Donglai Xiang, Hanbyul Joo, and Kristen Grauman. You2me: Inferring body pose in egocentric video via first and second person interactions. *CVPR*, 2020. 1, 2

[16] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. Egocentric future localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[17] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021. 7

[18] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9401–9411, 2021. 2

[19] Ryo Yonetani, Kris M Kitani, and Yoichi Sato. Recognizing micro-actions and reactions from paired egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2629–2638, 2016. 2