EgoPoseFormer: A Simple Baseline for Stereo Egocentric 3D Human Pose Estimation (Supplementary Material)

Chenhongyi Yang[1,2], Anastasia Tkach[2], Shreyas Hampali[2], Linguang Zhang[2],
Elliot J. Crowley[1], and Cem Keskin[2]

[1] University of Edinburgh
[2] Meta Reality Labs

1 Extra Experiments

1.1 2D Heatmap Pre-training.

Intuitively, 2D heatmap pre-training should tell the model what the appearance of a joint should look like, thus it can guide the deformable stereo attention to attend to relevant features, which can further help with accurately estimating the joints' 3D locations. In Tab. 1, we report how this pre-training influences the model performance on both the stereo UnrealEgo and the monocular SceneEgo datasets. On UnrealEgo, our pre-training improves the MPJPE of the pose proposal by 3.3mm and the final prediction by 3.1mm. However, on the SceneEgo dataset, such improvement becomes more significant, with 62.2mm for the pose proposal and 29.9mm for the final prediction. The reason is that as the SceneEgo dataset does not have stereo information, the appearance features become more important in localizing a joint. This experiment validates the effectiveness of our pre-training strategy.

Table 1: Ablation study on 2D heatmap pre-training on the UnrealEgo dataset.

		First Stage		Second Stage	
Dataset	Pretrain	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
UnrealEgo		48.8	40.1	36.5	35.1
	✓	45.5	38.3	33.4	32.7
SceneEgo		182.5	119.6	122.9	97.2
	✓	120.3	87.9	93.0	74.3

1.2 Qualitative Comparison with the baseline

In Fig. 1, we compare the qualitative failure cases of PPN (pose proposal), PRFormer (final prediction), and the baseline 'UnrealEgo' model. It shows that most errors in our model's final prediction are caused by the joint invisibility problem (mostly in lower-body). Compared with PRFormer, PPN's estimations are more inaccurate because they are computed using the coarse global features. On the other hand, the baseline's performance is far from satisfactory even when the joints are captured by both cameras.

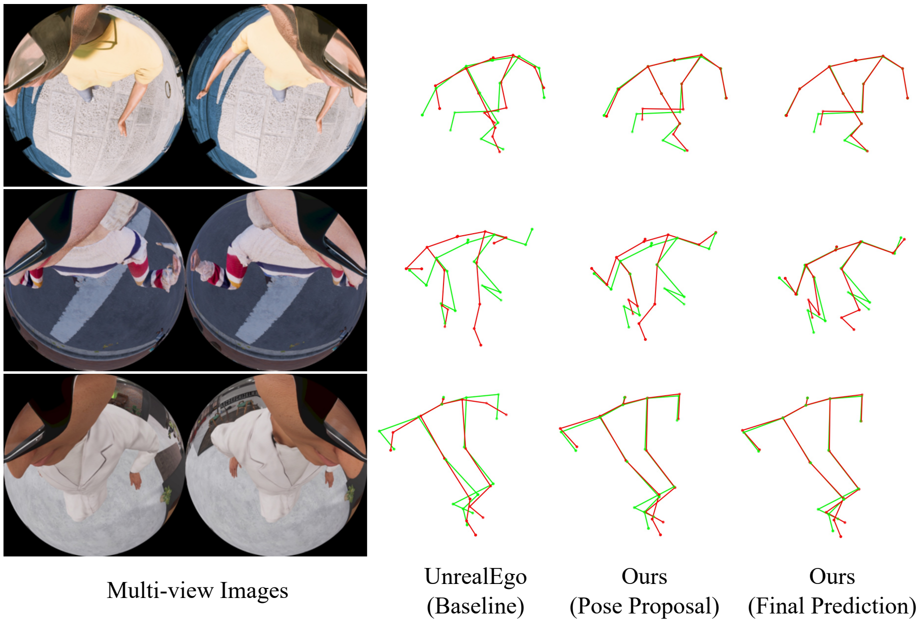

Fig. 1: Qualitative comparison between EgoPoseFormer and the UnrealEgo baseline model on the UnrealEgo dataset. Ground-truths are colored in green and predictions are colored in red.

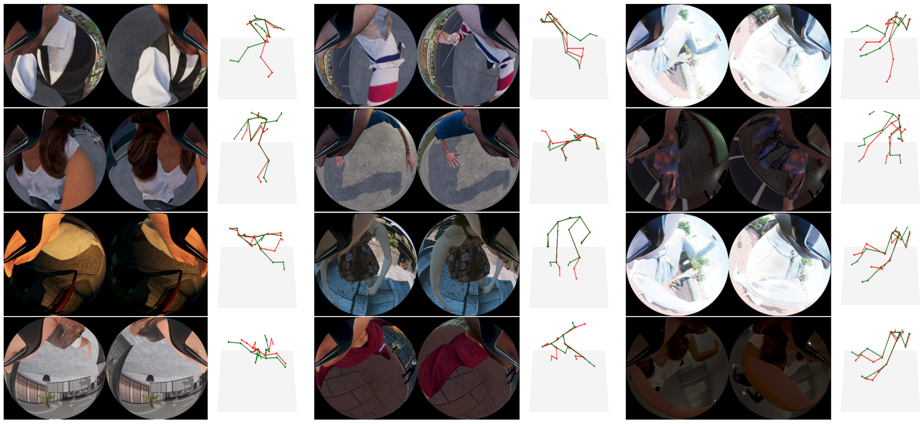

Fig. 2: Visualization of EgoPoseFormer's most inaccurate failure cases on the UnrealEgo dataset. Ground-truths are colored in green and predictions are colored in red.

1.3 Qualitative failure cases

In order to gain a more intuitive insight into the failure cases of our method, we present visualizations of some of the most inaccurate results in Fig. 2. A clear observation is that the majority of these failure cases are attributed to the problem of joint invisibility. For instance, in the first example in the first column, the lower body of the wearer is entirely occluded by his upper body, leading to a substantial discrepancy in the estimated locations of the lower body joints. In another instance, illustrated in the last example of the first column, nearly half of the wearer's body extends outside the FOV of both cameras, causing severe inaccuracy for the estimated 3D pose. These examples underscore the impact of joint invisibility in egocentric 3D pose estimation. Although our method can indeed estimate the locations of invisible joints in some cases, as we explained in the main paper, such an estimation is achieved by jointly looking at the visible joints and the background scene. However, when a large part of the wearer's body is invisible, the estimated pose is still far from accurate. Therefore, achieving accurate localization for such joints remains an important and valuable topic for future research endeavors.

1.4 Dependency of two stages

Here, we conduct an ablation experiment on the UnrealEgo dataset to check PRFormer's performance when the quality of the pose proposal varies. Specifically, we use perturbed ground truth, computed by adding Gaussian noises with different scales, to serve as pose proposals, based on which we use PRFormer to compute the refined pose estimation. We plot the dependency of the two stage's MPJPE in Fig. 3 Left. The result suggests that although PRFormer's performance is positively related to the accuracy of the pose proposal, the refined pose estimation is always more accurate than the initial estimation, validating the effectiveness of our PRFormer.

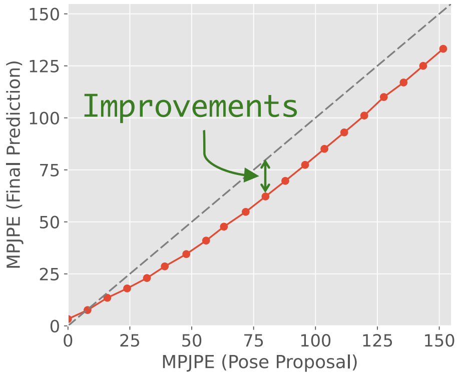

Fig. 3: Dependency of the accuracy of PPN and PRFormer.

2 Responsibility to human subjects

In alignment with the conference's ethical standards, this section addresses the ethical considerations and consent protocols relevant to our study, which incorporates real human data. Our research involves the evaluation of our method using two benchmarks: UnrealEgo [1] and SceneEgo [2], with the latter comprising images recorded by real humans. The SceneEgo dataset, as described in the foundational paper by [2], encompasses approximately 28,000 frames featuring two actors. Regrettably, the documentation provides no detailed information regarding these actors. We accessed the dataset through its open-source availability on the project's webpage[3] and GitHub repository[4]. We have ensured that our usage of this dataset meets the requirements in its license[5].

3 Potential ethical concerns

Technically, one limitation of our PPN and PRFormer is that they assume headset wearers to have a full body, resulting in poor support for estimating body poses for individuals with disabilities, particularly those who have lost parts of their bodies. We acknowledge this limitation and plan to address it in future work. Another potential negative impact of our model is related to user privacy. For example, malicious agents could misuse the technology to analyze a user's body pose without their permission. The widespread use of such technology could also potentially lead to increased surveillance and tracking of individuals, raising ethical concerns about its use in public and private spaces. Another concern could be the risk of reinforcing biases present in the training data, which could lead to inaccuracies in pose estimation for certain demographic groups. Finally, there is the potential for dependency on this technology. For example, one may first need to buy a headset before using our model, which might reduce users' ability to perform tasks without it, affecting their autonomy and skill development.

References

1. Akada, H., Wang, J., Shimada, S., Takahashi, M., Theobalt, C., Golyanik, V.: Unrealego: A new dataset for robust egocentric 3d human motion capture. In: European Conference on Computer Vision (ECCV) (2022) 3
2. Wang, J., Luvizon, D., Xu, W., Liu, L., Sarkar, K., Theobalt, C.: Scene-aware egocentric 3d human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13031–13040 (2023) 3

[3] https://people.mpi-inf.mpg.de/~jianwang/projects/sceneego/

[4] https://github.com/jianwang-mpi/SceneEgo/tree/main

[5] https://people.mpi-inf.mpg.de/~jianwang/projects/sceneego/data/LICENSE.txt