Generalizable Human Pose Triangulation
- Supplementary Appendix -

The main focus of the Supplementary Appendix is to demonstrate the application of the proposed model to novel camera arrangements and datasets that have unknown relative camera poses, i.e. extrinsic parameters $E_{ref,\,i} = [R_{ref,\,i}|t_{ref,\,i}]$, where *ref* is the reference camera, and i is each of the relative cameras. The camera poses are estimated based on the fundamental matrix estimation method described in the main paper. We further dissect relative camera pose estimation into the estimation of relative rotation, $R_{ref,\,i}$, and relative translation, $t_{ref,\,i}$, showing that the unknown translations have more significant impact on the performance than the unknown rotations (Appendix 1). Finally, we briefly discuss other works, implementation details, and the limitations of the model in more detail and propose future work, in addition to the main paper (Appendix 4). The ethical considerations are addressed in Appendix 5.

1. Performance with Estimated Camera Poses

We evaluate the performance of the generalizable human pose triangulation model in case when the camera poses are estimated using the proposed fundamental matrix estimation method, on Human3.6M. In particular, we compare the performances between the test sets with known extrinsics, estimated relative rotations $R_{ref,\,i}$, estimated relative translations, $t_{ref,\,i}$, and estimated extrinsics (both rotation and translation). Additionally, we compare the performances when Human3.6M is used as the training dataset (base-dataset experiment), and when CMU3, described in the main paper, is used for training (inter-dataset experiment).

The results are shown in Table 1. As expected, the performance on the base dataset is better than the performance on inter-dataset experiment. In overall, the performances on both the base experiment and inter-dataset experiment are satisfactory, taking into account that the rotations, translations, i.e. both, are unknown. Notably, the performance of the model significantly drops for unknown relative translations, while the unknown relative rotations only slightly affect the performance. We assume that the rotations are simply estimated more accurately than translations, hence the difference. To verify this assumption, we analyze 2D and 3D errors, defined in the main paper, for estimated rotations, i.e., translations separately.

Table 1: The evaluation of the model in case of unknown relative camera poses on Human3.6M [4]. We evaluate the model in base-dataset (same camera arrangement for training and testing) and inter-dataset (from CMU3 [6] to Human3.6M). We also dissect the analysis into the cases when rotation, i.e., translation only is unknown. Note that all Rs and ts shown in the table correspond to $R_{ref,\,i}$ and $r_{ref,\,i}$, but are abbreviated.

Base dataset (Human3.6M)			
Known $[R\|t]$	Estimated R	Estimated t	Estimated $[R\|t]$
29.1 mm	29.4 mm	36.7 mm	37.3 mm
Inter-dataset (CMU3 $\rightarrow$ Human3.6M)			
Known $[R\|t]$	Estimated R	Estimated t	Estimated $[R\|t]$
31.0 mm	33.6 mm	42.2 mm	44.5 mm

Ablative Analysis of Camera Pose Estimation. Table 2 shows the fundamental matrix estimation errors (E_{2D} and E_{3D}, described in the main paper) between the pairs of views, in case when only rotation is estimated and the translation is known, and vice versa. The errors in case of the estimated translations are always higher compared to the case of estimated rotations, therefore, this result might explain the performance drop shown in Table 1. The future work should focus on improving translation estimation.

Table 2: Dissecting the evaluation of fundamental matrix estimation on two cases — when the rotations, i.e., the translations are estimated, for all pairs of views on Human3.6M. The 2D errors, E_{2D} are shown in pixels, and 3D errors, E_{3D} are shown in millimeters.

	Estimated R		Estimated t	
Camera pair	E_{2D}	E_{3D}	E_{2D}	E_{3D}
(1, 3)	1.2	10.8	1.8	18.2
(2, 4)	0.9	9.7	1.6	15.3
(1, 4)	0.9	6.4	1.2	8.9
(2, 3)	0.6	3.9	1.0	4.5
(3, 4)	**0.4**	**1.2**	**0.7**	4.0
(1, 2)	**0.4**	1.8	**0.7**	**3.7**

Table 3: The Table of hyperparameters for the two tasks.

	3D pose	Camera
Learning rate	$5 * 10e^{-4}$	$10e^{-5}$
τ	1.5	1.2
α, β, γ	(1.0, 0.01, 0.02)	(1.0, 0.01, 0.0)
Network layer sizes	(1000, 900 900, 900, 700)	(1000, 900, 900)
# hypotheses in sample	200	100
Batch size	16	16

2. Other Works

The Epipolar Transformers [3] outperforms our method on Human3.6M (base dataset). However, note that our model outperforms their lightweight, transformer model on H36M (30.4mm, Table 6 [3], compared to our 29.1mm, Table 4, main paper). The difference in performances would most likely increase when evaluated on novel views, especially as the authors did not tackle the generalization problem at all. Further, their heavy-weight model might overfit even more on the base camera arrangement of the train dataset(s), so we can expect an increased performance drop on unseen views.

3. Implementation Details

The selected hyperparameters set is shown in Tab. 3. The two hyperparameters used specifically for pose triangulation, i.e., fundamental matrix estimation, are the number of joints in the pose model, $J = 17$, and the number of frames from which the camera hypotheses are sampled, $M = 80$.

The required number of training iterations is relatively small. We obtain our best results using only 500 iterations. In each iteration, we generate 200 hypotheses. This is a great advantage of the approach, especially when only small amount data annotations are required. In particular, 500 iterations correspond to 500 data samples, i.e., 500/16$\approx$32 batches (batch size 16, Table 3), meaning that the gradients were applied $\approx$32 times for the model to be fully trained. It takes about 3 minutes to train the model, but this can be further improved by more efficient implementation of the hypothesis generation on CPU. Moreover, the training time is shorter, which simplifies the optimal hyperparameter search. Finally, the current implementation fits into $\sim$1GB of GPU memory.

4. Limitations

The main limitation of our model is that it strongly depends on the performance of the 2D detector [7]. This is best seen in Table 4 that shows the difference in the performance on train, validation, and test[1]. The difference between the validation and test performance, in particular, can be explained by the fact that the 2D backbone has been fine-tuned on the whole training and validation splits, while it has never seen the test data. What this means is that we did not tackle the problem of train-to-test generalization; instead, we improved the between-test-sets generalization, which is a weaker result. The consequence of this train-test difference is that the performance on novel data will suffer mostly from the performance drop of the detector.

Table 4: The comparison between train, validation, and test performance on Human3.6M (in case of base-dataset configuration). There is a significant difference in the performance between train (validation) and test.

Human3.6M		
Train	Validation	Test
13.8 mm	14.2 mm	29.1 mm

Another limitation is that the current model does not learn end-to-end. The consequence is that the model, at best, learns to differentiate well between the poses. But once the poses are good enough, the network can't differentiate further and will simply assign the same scores, converging into an average of "good-enough" 3D poses[2]. Therefore, future work should definitely address this limitation by exploiting image features to obtain additional information about the keypoints. One way to use image features is through the confidence predictions, similar to previous works [5, 1, 2].

Finally, we assume that the intrinsic camera parameters and the scale are known.

5. Ethical Considerations

For all of our experiments, we use two well-known, public datasets — Human3.6M and CMU Panoptic Studio. From the information obtained from the corresponding websites, it is unclear whether the datasets have the IRB approvals. We verified with the authors of the Panoptic Studio that the dataset has the approval. We also contacted the authors of Human3.6M, but did not get the confirmation at the moment of writing.

References

[1] E. Brachmann and C. Rother. Neural-guided ransac: Learning where to sample model hypotheses. *2019 IEEE/CVF Interna-*

[1] Note that, for training, we use subjects 1, 5, 6, 7, for validation, we use subject 8, and the remaining subjects 9 and 11 are used for testing.

[2] The good poses should be the ones that are symmetric and the ones that have body part ratios consistent with the ratios of an average (training set) person. Note that the good poses should have high pose prior scores.

tional Conference on Computer Vision (ICCV), pages 4321–4330, 2019.

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1302–1310, 2017.

[3] Y. He, R. Yan, K. Fragkiadaki, and S.-I. Yu. Epipolar transformers. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7776–7785, 2020.

[4] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[5] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov. Learnable triangulation of human pose. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7717–7726, 2019.

[6] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[7] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018.