Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Multi-person 3D Pose Estimation from a Single Image Captured by a Fisheye Camera (*Supplementary Material*)

Yahui Zhang[a,**], Shaodi You[a], Sezer Karaoglu[a,b], Theo Gevers[a,b]

[a] *University of Amsterdam, Science Park 904, 1098XH Amsterdam, The Netherlands*
[b] *3DUniversum, Science Park 400, 1098XH Amsterdam, The Netherlands*

ABSTRACT

1. Overview

In this supplementary material, we provide additional experimental results and training details. Specifically, Section 2 gives training details on the modified CMU Panoptic and our newly collected 3DhUman datasets. In Section 3, we compare our method with existing methods to analyze the performance on different distortion strengths. In Section 4, we conduct experiments to analyze the effect of proposed re-projection module on sensitivity of different levels of image distortions. Additional visual results of applying our approach on modified CMU Panoptic and 3DhUman datasets are presented in Section 5.

2. Training Details

Please refer to Section 5.1 in our main paper for the details of CMU Panoptic (Joo et al., 2015) and 3DhUman datasets. Here, we provide additional training details on these two datasets.

Modified CMU Panoptic. The training set in our experiments includes: 160224_mafia1, 160224_mafia2, 160224_ultimatum1, 160224_ultimatum2 and 160226_mafia2. The test set contains: 160226_haggling1, 160422_haggling1, 160226_mafia1, 160422_mafia2, 160422_ ultimatum1 and 160906_pizza1.

3DhUman. Following CMU Panoptic dataset, our dataset consists of three commonly activities: posing, talking and walking. To avoid the negative impact of grayscales on Moon et al. (2019) and Lin and Lee (2020), all compared methods are first trained on modified CMU Panoptic dataset with grayscales and then finetuned on 3DhUman dataset.

For joint annotations, we manually labeled 2D poses in perspective images captured by pin-hole cameras and converted them to 3D poses based on the corresponding depths obtained by the lidar cameras. We then projected 3D poses by fisheye camera parameters to obtain the distorted 2D poses. Particularly, images were captured simultaneously by pin-hole, fisheye and lidar cameras.

3. Analysis of Distortion Strengths

The distortion strengths in a fisheye image are different according to the regions in one image. Therefore, we analyze the performance of the methods for each person with different distortion levels defined by θ_d/θ instead of image distortions for the whole image. Following the definition in Section 5.2.4 in the main paper, θ_d/θ is categorized into [0.46, 0.8), [0.8, 0.9), [0.9, 1) for all humans appearing in the testing set. Note that the larger the value of θ_d/θ, the small distortion strengths are. The experimental results are shown in Fig. 1.

It can be seen that the all methods provide the larger value of the two metrics with the larger the distortions, especially for the absolute depth estimation. The reason may be that large image distortions cause significant changes of persons on the image plane. Our method with re-projection module shows a promising solution and outperforms other methods.

4. Analysis of Re-projection Module on Image Distortion

Due to the existence of image distortions, persons at different locations on images may cause various distortion strengths. Therefore, even when persons express different 2D poses, they may be originated from the same 3D pose. In this section, we verify the robustness of our approach on images with various image distortions. Specifically, Fig. 2 shows 3D predictions performed by our method from the same image with two levels of image distortions in modified CMU Panoptic dataset, where

**Corresponding author.
e-mail: y.zhang5@uva.nl (Yahui Zhang)

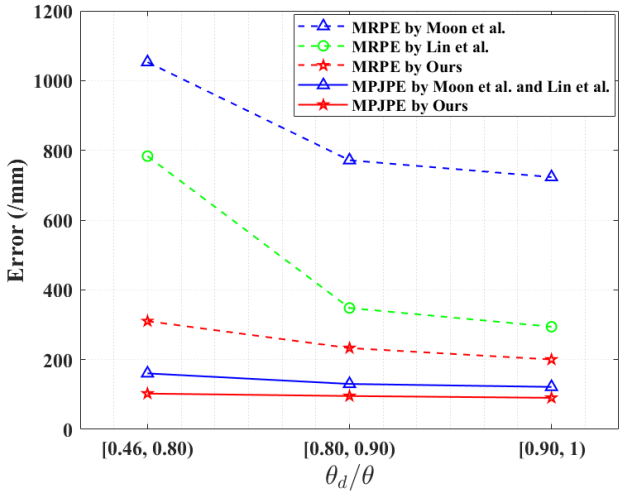

Fig. 1. Performance of the methods for different distortion levels on the modified CMU Panoptic dataset for HD cameras 4, 6, 9, 12 and 13.

three paired examples are provided. It can be seen that our method generates reasonable visual results and is not affected by different levels of image distortions, even the person scale at the edge of the image changes significantly.

5. Visual results

In this section, we present additional visual results of applying our method on modified CMU Panoptic and 3DhUman datasets shown in Fig. 3 and 4.

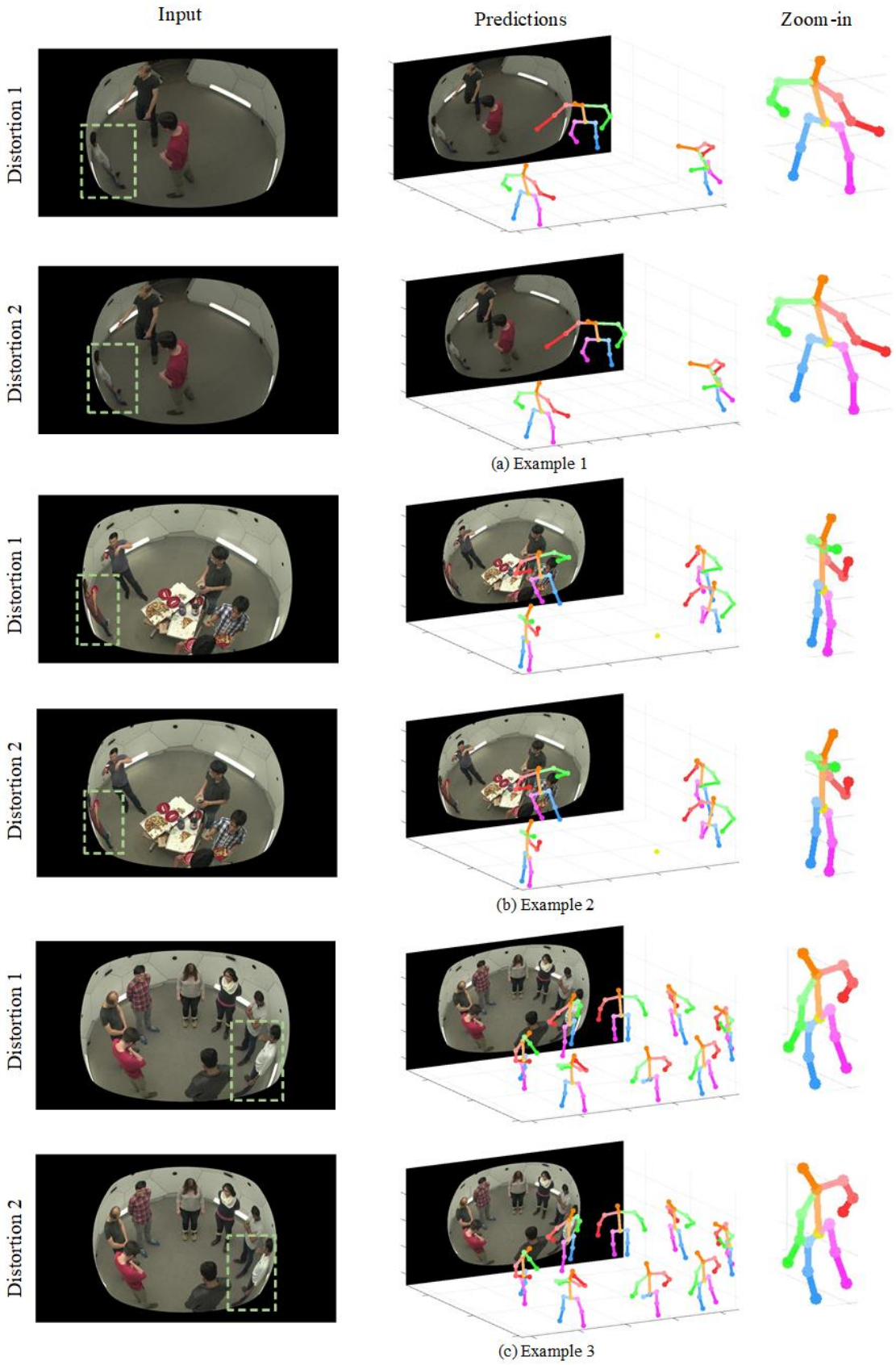

Fig. 2. The visual results of applying our method on images with different levels of image distortions, where Distortion 2 is larger than Distortion 1. Notice scales of the same person with different levels of image distortions changed. Particularly, human scales change the most at the border of the image.

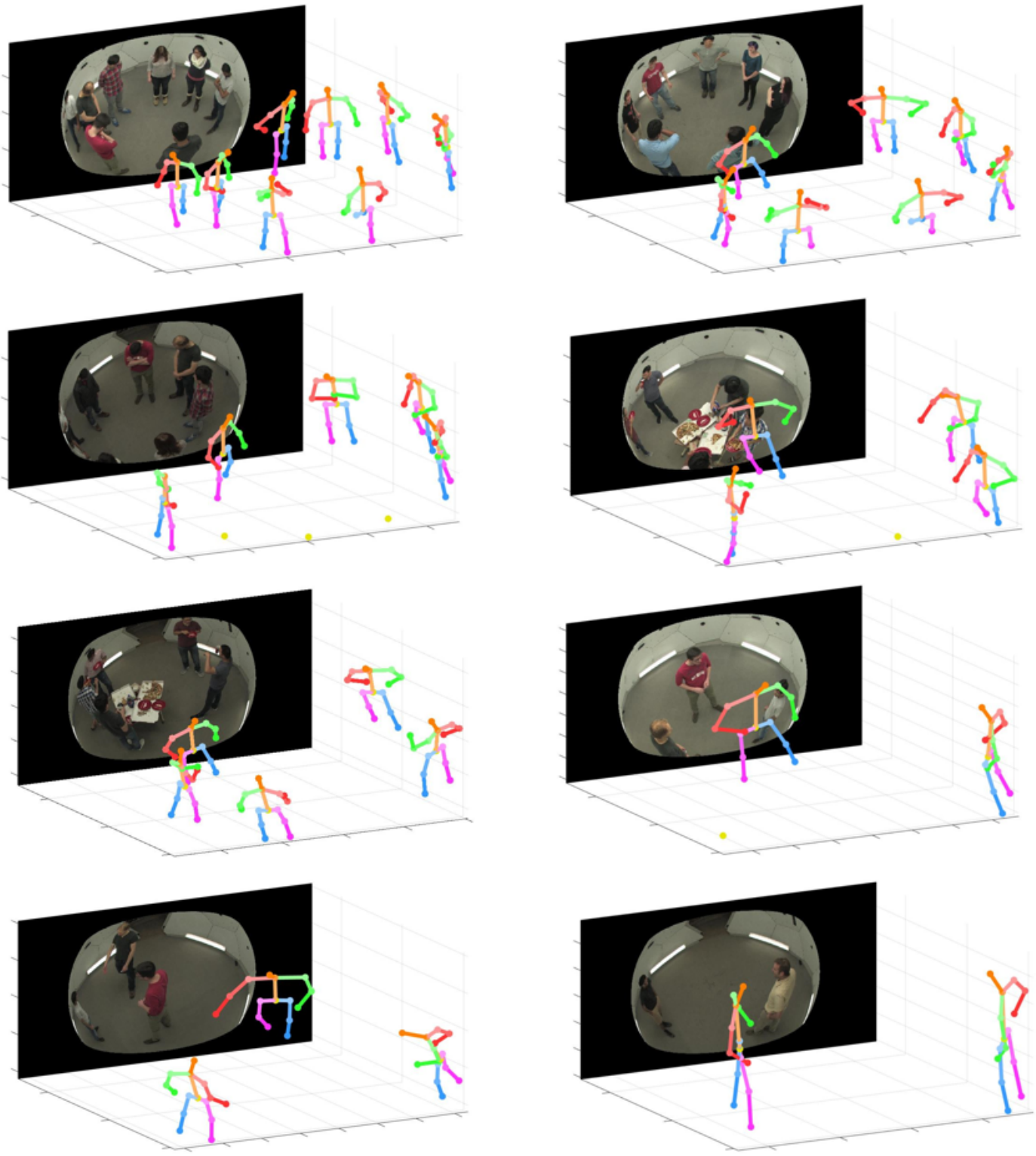

Fig. 3. Additional visual results of applying our method on the modified CMU Panoptic dataset.

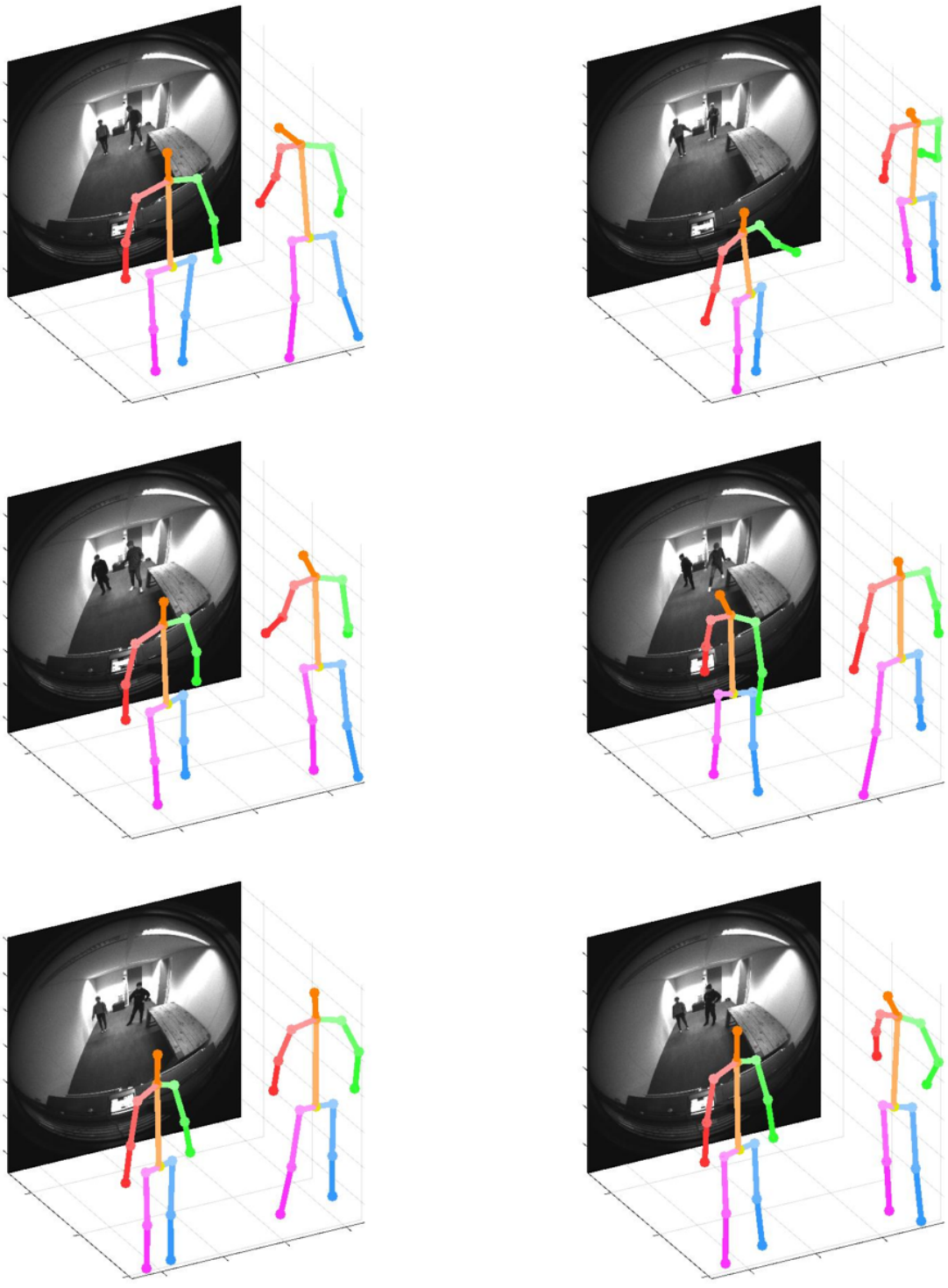

Fig. 4. Additional visual results of applying our method on the 3DhUman dataset.

References

Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y., 2015. Panoptic studio: A massively multiview system for social motion capture, in: ICCV, pp. 3334–3342.

Lin, J., Lee, G.H., 2020. Hdnet: Human depth estimation for multi-person camera-space localization, in: ECCV, pp. 633–648.

Moon, G., Chang, J.Y., Lee, K.M., 2019. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image, in: ICCV, pp. 10133–10142.