# 3D Ego-Pose Lift-Up Robustness Study for Fisheye Camera Perturbations

Teppei Miura[1,2], Shinji Sako[2] and Tsutomu Kimura[1]
[1]*Dep. of Information and Computer Engineering, National Institute of Technology Toyota College, Toyota, Aichi, Japan*
[2]*Dep. of Computer Science, Nagoya Institute of Technology, Nagoya, Aichi, Japan*

Keywords: 3D Ego-Pose Estimation, 3D Pose Lift-Up, Camera Perturbation, Robustness Study.

Abstract: 3D egocentric human pose estimations from a mounted fisheye camera have been developed following the advances in convolutional neural networks and synthetic data generations. The camera captures different images that are affected by the optical properties, the mounted position, and the camera perturbations caused by body motion. Therefore, data collecting and model training are main challenges to estimate 3D ego-pose from a mounted fisheye camera. Past works proposed synthetic data generations and two-step estimation model that consisted of 2D human pose estimation and subsequent 3D lift-up to overcome the tasks. However, the works insufficiently verify robustness for the camera perturbations. In this paper, we evaluate existing models for robustness using a synthetic dataset with the camera perturbations that increases in several steps. Our study provides useful knowledges to introduce 3D ego-pose estimation for a mounted fisheye camera in practical.

## 1 INTRODUCTION

Human motion capture is widely used in society, ex. virtual reality, augmented reality, and performance analysis in sports science. 3D human pose estimation in daily situations will be important to develop more services.

Researchers have proposed many 3D human pose estimation methods for external camera that is statically placed around the users. However, such a camera setup is impractical in daily life because of limitations, ex. portability, setup space and ground condition.

Wearable camera introduces 3D human pose estimation from the egocentric perspective in daily situations. However, the methods capture only parts of body due to limitation of the field of view and the proximate setup position.

Wider angle camera estimates 3D egocentric human pose (3D ego-pose) under wider variety of motions. Xu et al. and Tome et al. equipped a single fisheye camera around the user's head for 3D ego-pose estimation (Xu et al., 2019; Tome et al., 2019; Tome et al., 2020). Miura et al. introduced an omnidirectional camera mounted on the user's chest for wider field of view (Miura and Sako, 2022). We show their camera setups and captured images in Figure 1.

These unique camera optics and setups give rise to a shortage of the data in training dataset for deep neural networks. Additionally, acquiring a large number of the data, which are in-the-wild images with 2D / 3D pose annotations, for the egocentric perspective is a time-consuming task even if it is available in a professional motion capture system. To tackle the difficulties, past works generated vast synthetic datasets for each camera optical properties and mounted positions.

Generating synthetic data overcomes the shortage of the training dataset, however the 3D ego-pose estimation model must be trained for each camera setups. Tome et al. and Miura et al. proposed two-step estimation model that consists of 2D human pose estimation and subsequent 3D lift-up to reduce the training burden (Tome et al., 2019; Tome et al., 2020; Miura and Sako, 2022). In particular, Miura et al.'s model does not require to re-train the 3D lift-up model for changing camera optical properties by applying statically obtained camera parameters, however it still requires training for changing camera mounted position.

The mounted fisheye camera captures different images that are affected by the optical properties, the mounted positions, and the camera perturbations caused by body motion. Past works have tackled the data shortage and the re-training burden for changing the camera optics and positions due to propose syn-



Miura, T., Sako, S. and Kimura, T.
3D Ego-Pose Lift-Up Robustness Study for Fisheye Camera Perturbations.
DOI: 10.5220/0011661000003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 600-606
ISBN: 978-989-758-634-7; ISSN: 2184-4321

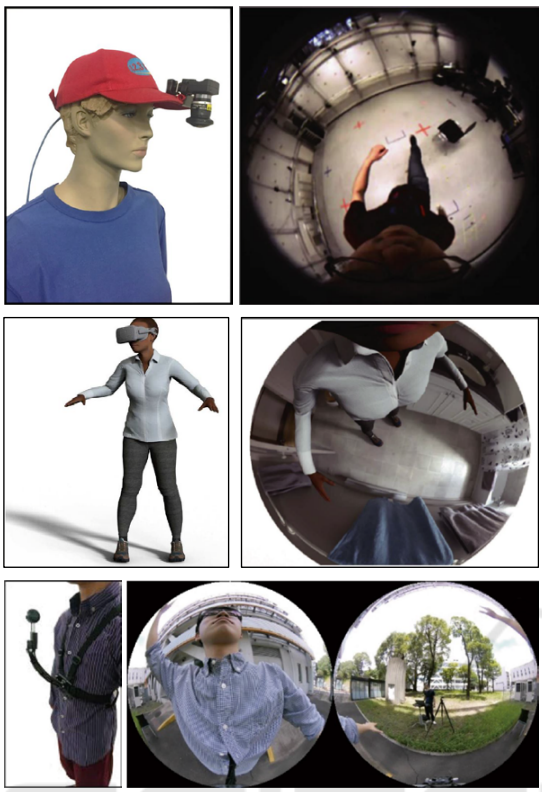

Figure 1: Camera setups and the captured images in Xu et al. (upper), Tome et al. (middle), and Miura et al. (bottom).

thetic data generation and two-step estimation model. However, the works insufficiently verify robustness of 3D ego-pose estimation model for the camera perturbations.

We evaluate robustness of two-step 3D ego-pose estimation models for the mounted camera perturbations in this paper. We generate synthetic dataset for training and evaluation, which increase the camera perturbations in several steps. We train and quantitatively evaluate the models using the synthetic dataset. Our study provides useful knowledge to introduce 3D ego-pose estimation for a mounted fisheye camera in practical. Our contributions are summarized as follows:

- We generate a synthetic dataset with incremental camera perturbations for 3D ego-pose estimation from a mounted fisheye camera, and it is publicly available.
- We evaluate the camera perturbation robustness of the 3D lift-up model in two-step 3D ego-pose estimations.

## 2 RELATED WORKS

We discuss monocular 3D human pose estimations focusing on camera setups: an external camera and a mounted fisheye cameras perspective.

### 2.1 3D Human Pose Estimation with an External Camera

Convolutional neural networks and large-scale 2D / 3D datasets have recently enabled advances in 3D human pose estimation from images. Two main approaches have emerged in monocular 3D human pose estimation: (1) direct regression approaches to 3D joint positions (Tekin et al., 2016; Pavlakos et al., 2017; Zhou et al., 2016; Mehta et al., 2017) and (2) two-step approaches that decouple the problem into tasks of 2D joint location estimation and subsequent 3D lift-up (Martinez et al., 2017; Xiaowei et al., 2017).

In direct regression approaches, the accuracy and generalization are severely affected by the availability of 3D pose annotations for in-the-wild images. Two-step decoupled approaches have two advantages: (1) the availability of high quality existing 2D joint location estimators that require only easy-to-harvest 2D pose annotations with images (Wei et al., 2016; Newell et al., 2016; Xiao et al., 2018; Sun et al., 2019) and (2) the possibility of training the 3D lift-up step using 3D mocap datasets and their ground truth 2D projections without images. Martinez et al. indicated that even simple architectures solve the 3D lift-up task with a low error rate (Martinez et al., 2017).

### 2.2 3D Ego-Pose Estimation with Mounted Fisheye Cameras

In recent years, researchers proposed 3D ego-pose estimation for lightweight monocular fisheye camera. Xu et al. proposed a direct regression model that estimates the unit vectors and the distances towards with 3D joint positions from the camera position (Xu et al., 2019). The 3D unit vectors are obtained by estimated 2D joint locations and an omnidirectional camera calibration toolbox (Scaramuzza et al., 2006).

Tome et al. proposed a two-step approach using a multibranch encoder-decoder model (*x*R-EgoPose) that estimates 3D joint positions from 2D joint location heatmaps (Tome et al., 2019; Tome et al., 2020). The model reduces the data collection and training burden because the 3D lift-up model is separately trainable by 3D joint positions and their ground truth 2D heatmaps without raw images. However, the 3D

Figure 2: Camera setup and a captured image in this paper.

lift-up model still requires data re-collection and re-training according to changes in the camera optics.

Miura et al. proposed a two-step model that deploys the 3D unit vectorization module between the 2D joint location estimator and the 3D lift-up (Miura and Sako, 2022). The 3D lift-up model does not require the data re-collection and re-training according to changes in the camera optics because the unit vectorization module confines the impact of optical properties by the camera intrinsic parameters. Additionally, the 3D lift-up model is trainable by 3D joint positions and their unit vectors from the camera position in publicly available.

Above works generated large-scale synthetic datasets to solve the problem of the shortage of the data because of unique camera positions and optical properties. Wang et al. proposed to estimate 3D ego-pose with weak supervision from an external view and collect a large in-the-wild dataset captured a mounted fisheye camera and an external camera (Wang et al., 2022).

Past works have tackled to reduce the data collection and training burden for the camera positions and optical properties. However, the works insufficiently verify robustness of 3D ego-pose estimation model for the camera perturbations caused by body motions.

## 3 APPROACH

We generate synthetic dataset for training and evaluation that increase the camera perturbations in several steps, to verify robustness of two-step 3D ego-pose estimation models. We introduce a single mounted omnidirectional camera that consists of back-to-back dual fisheye cameras in this paper. We show the camera position and a captured image in Figure 2.

### 3.1 Synthetic Data Generation

Acquiring a large quantity of in-the-wild images with 2D / 3D pose annotations for the egocentric perspective with the camera perturbations is a time-consuming task even if it is available in a professional motion capture system. To alleviate this problem, we

Table 1: Synthetic training and evaluation datasets with the camera perturbations.

| | background image | pos. ($\sigma^2$) | rot. ($\sigma^2$) | num. of data |
|---|---|---|---|---|
| train | indoor: 40<br>outdoor: 40 | 17.50 mm | 10° | 8,186 |
| eval. | indoor: 14<br>outdoor: 10 | 0.00 mm<br>8.75 mm<br>17.50 mm<br>26.25 mm<br>35.00 mm | 0°<br>5°<br>10°<br>15°<br>20° | 2,088 |

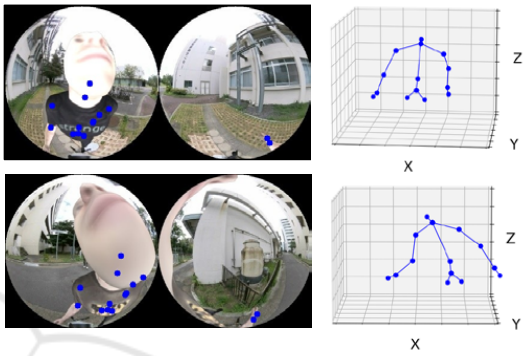


Figure 3: Synthetic image examples with ground truth 2D / 3D pose annotations. These examples are generated by identical CMU mocap data but applyed different perturbations. (upper) position $\sigma^2 = 0.00$ cm, rotation $\sigma^2 = 0°$. (bottom) position $\sigma^2 = 35.00$ cm, rotation $\sigma^2 = 20°$.

generate synthetic images with ground truth 2D / 3D pose annotations in our unique setup.

We render a synthetic human body model from a virtual mounted camera perspective. To acquire a large variety of motions, we build the dataset based on the large-scale synthetic human dataset SURREAL (Varol et al., 2017). We animate the human model using SMPL body model (Loper et al., 2015) with sampled motions from CMU MoCap dataset. Body textures are randomly chosen from the texture dataset provided by SURREAL.

To generate realistic images, we simulate the camera and background in a real-world scenario. The virtual camera is placed at a similar position as our setup. The camera randomly perturbates the position and rotation in each rendering. We apply the intrinsic camera parameters obtained by the real camera using the omnidirectional camera calibration toolbox (Scaramuzza et al., 2006). The rendered images are augmented with the backgrounds randomly chosen from 54 indoor and 50 outdoor images.

Our synthetic dataset contains ground truth 2D / 3D pose annotations that are easily generated to use the 3D joint positions and calibration toolbox. We acquire the following 14 body joints: head, neck, spine,

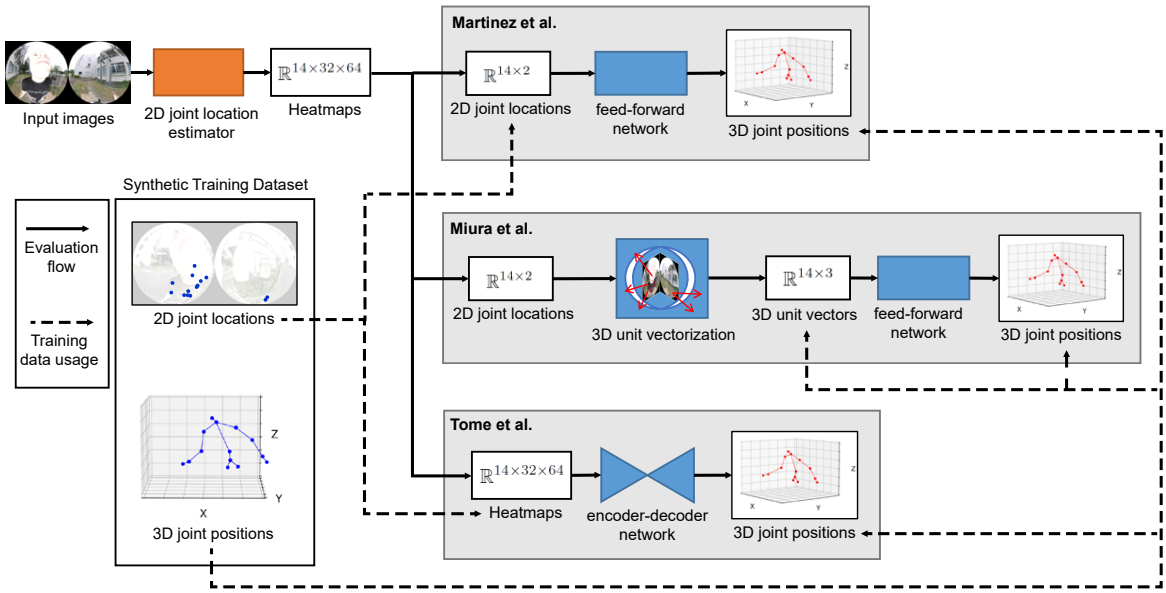


Figure 4: Whole process of two-step 3D ego-pose estimation models. The 3D lift-up models obtain each input from the first step 2D joint location estimator output. Note that Miura et al.'s model require only 3D joint positions for training data.

pelvis, hips, shoulders, elbows, wrists, and hands. The 3D joint positions are incorporated into the camera coordinate system, where we normalize the skeleton scale in shoulder width is 350 mm.

We apply different CMU MoCap data, body textures, background images, and the camera perturbations between the training and evaluation datasets. In the training dataset, the camera moves the position in 3D space according to a normal distribution $N(\sigma^2 = 17.50 \text{ mm})$ and rotates according to $N(\sigma^2 = 10^\circ)$, which is XYZ Euler angle. In the evaluation dataset, we change the camera perturbations of position ($\sigma^2 = 0.00$ mm to 35.00 mm) and rotation ($\sigma^2 = 0^\circ$ to $20^\circ$) in 5 steps. We collect 8,186 training data and 2,088 evaluation data for each perturbation. We indicate the synthetic dataset in Table 1 and Figure 3.

### 3.2 3D Ego-Pose Estimation Models

We verify the camera perturbation robustness for two-step 3D ego-pose estimation models (Martinez et al., 2017; Miura and Sako, 2022; Tome et al., 2019; Tome et al., 2020). The two-step models reduce the data collection and training burden because of the availability of existing 2D joint location estimators and the possibility of training the 3D lift-up model using 3D mocap datasets without images.

Martinez et al. proposed a simple feed-forward network model to estimate 3D joint positions from 2D joint locations. Miura et al. proposed to deploy a 3D unit vectorization module that converts 2D joint locations to 3D unit vectors between 2D joint location estimator and 3D lift-up. Therefore, the simple network estimates 3D joint positions from 3D unit vectors in the camera coordinate system. Tome et al. proposed a multibranch encoder-decoder model ($x$R-EgoPose) to estimate 3D joint positions from 2D joint location heatmaps. These 3D lift-up models require the use of a 2D joint location estimator in the first step. We show whole two-step models in Figure 4.

## 4 EVALUATION

We quantitatively evaluate the 3D lift-up models in two-step 3D ego-pose estimations. We use the mean joint position error (MJPE) as the evaluation metrics in 3D space. The error is the Euclidean distance between the estimation and the ground truth of a 3D joint position. Additionally, we also use pixelwise mean joint location error (MJLE) in 2D plane evaluation metrics.

### 4.1 Implementation and Training

Two step approaches require the use of a 2D joint location estimator in the first step. We build the 2D estimator based on MobileNet V2 with 3 deconvolution layers (Sandler et al., 2018). The 2D estimator outputs $32 \times 64$ pixel heatmaps from synthetic images with a resolution of $128 \times 256$ pixels. We train the 2D

Table 2: MJPE (mm) results on the evaluation dataset (position $\sigma^2 = 17.50$ mm, rotation $\sigma^2 = 10°$).

| model | head | neck | spine | pelvis | hips | shoulders | elbows | wrists | hands | all |
|---|---|---|---|---|---|---|---|---|---|---|
| Martinez et al. | 35.27 | 41.79 | 89.48 | 98.59 | 119.03 | 57.41 | 113.19 | 168.96 | 204.46 | 113.66 |
| Miura et al. (VD loss) | 81.44 | 61.56 | 74.60 | 86.72 | 102.39 | 66.42 | 109.93 | **163.47** | **197.66** | 113.15 |
| Miura et al. (L2 loss) | **26.55** | **30.82** | **70.83** | **82.84** | **96.58** | **43.36** | **99.19** | 163.91 | 199.44 | **101.14** |
| *x*R-EgoPose (p3d+hm) | 48.35 | 59.39 | 87.22 | 104.62 | 123.68 | 67.07 | 134.75 | 219.70 | 261.08 | 136.58 |

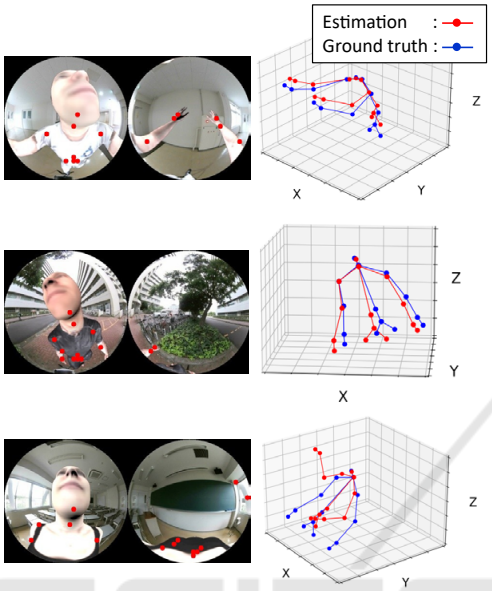


Figure 5: Examples of 3D pose estimation results on the evaluation dataset. (left) Input synthetic images and 2D joint location estimation results. (right) 3D joint position estimation results and ground truths.

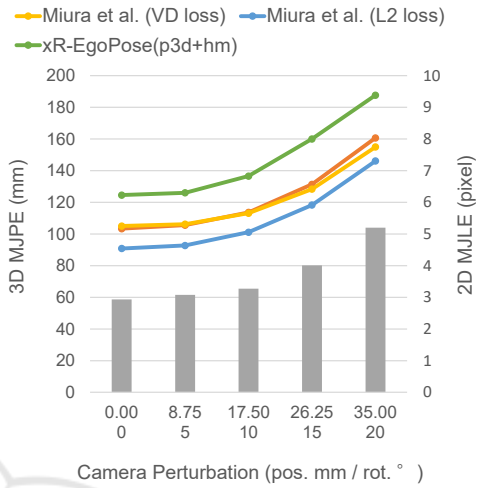


Figure 6: Evaluation results on evaluation datasets to perturb position and rotation.

estimator with synthetic images and 2D joint location heatmaps in training dataset. We use Adam optimizer, the initial learning rate of 0.001, batch size of 32, and 140 epochs.

We train Martinez et al.'s lift-up model with 3D joint positions and 2D joint locations in the training dataset with L2 loss function. We use Adam optimizer, the initial learning rate of 0.001, batch size of 32, and 140 epochs.

We train Miura et al.'s lift-up model with 3D joint positions and their unit vectors. Miura et al. proposed VD loss function but the proper coefficient parameters are difficulty found in grid search. We evaluate two trained models to apply each VD loss and L2 loss functions. The models are also trained under the same optimizer, learning rate, and other conditions with Martinez et al.'s model.

We train Tome et al.'s lift-up model (*x*R-EgoPose) with 3D joint positions and 2D heatmaps generated by 2D joint locations. In this paper, *x*R-EgoPose uses dual branch, 3D joint positions and heatmaps for loss function, because of our dataset limitations. The training conditions are same as other models.

## 4.2 Evaluation for Perturbations

We indicate 3D lift-up model results on the evaluation dataset (position $\sigma^2 = 17.50$ mm, rotation $\sigma^2 = 10°$) in Table 2. Miura et al.'s model (L2 loss) estimates 3D joint positions in best accuracy. We describe the estimation result examples of Miura et al.'s model in Figure 5. We find a failure of 3D lift-up when the 2D joint location estimator deteriorates in the bottom example (arms and hands).

We show the mean joint position error (MJPE) on evaluation datasets to perturb the positions and rotation in Figure 6. We also show the pixelwise mean joint location error (MJLE).

The 2D joint location estimator outputs worse estimation results following larger perturbations. In particular, the deterioration in accuracy is worse on evaluation datasets (positions $\sigma^2 \geq 26.25$ mm, rotation $\sigma^2 \geq 15°$) that are larger than training dataset. The accuracy of 3D joint position estimation is worse along with 2D estimation deterioration. Miura et al.'s lift-up model (L2 loss) indicates better performance in accuracy than other models on evaluation dataset for all perturbations.

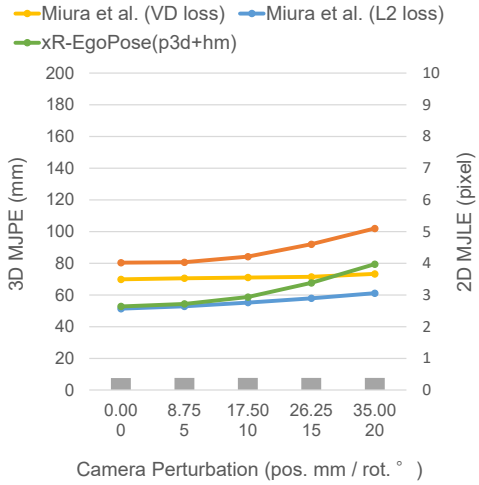


Figure 7: Evaluation results for ground truth 2D joint locations on evaluation dataset.

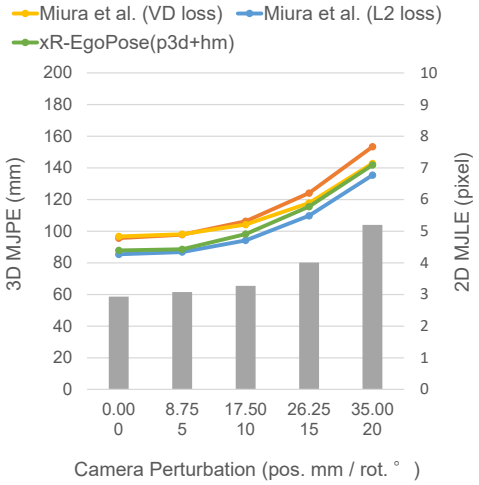


Figure 8: Evaluation results for the 3D lift-up model that are trained on ground truth 3D joint positions and estimated heatmaps or 2D joint locations.

### 4.3 Evaluation for Ground Truth 2D Joint Locations

We show the 3D lift-up model results on ground truth 2D joint locations of evaluation datasets instead of 2D joint location estimator outputs in Figure 7. Using ground truth 2D joint locations simulates a perfect 2D joint location estimator to evaluate just the 3D lift-up models. 2D joint location trivial error in the figure is raised by rounding up to convert the ground truth 2D joint locations to pixels on 2D plane.

Miura et al.'s model (L2 loss) and *x*R-EgoPose (p3d+hm) indicate similar performance in accuracy under trained perturbations (position $\sigma^2 \leq 17.50$ mm, rotation $\sigma^2 \leq 10^\circ$). However, Miura et al.'s model (L2 loss) restrains performance deterioration even on larger perturbations than the training dataset. Therefore, Miura et al's lift-up model has more robustness than other models to camera perturbations of position and rotation.

### 4.4 Evaluation for Training on 2D Estimator Outputs

*x*R-EgoPose can learn generalization and robustness to complex human poses by training with heatmaps that are obtained by the 2D joint location estimator (Tome et al., 2019; Tome et al., 2020). We train the 3D lift-up models on ground truth 3D joint positions and 2D estimator's outputs (heatmaps or 2D joint locations). We show the 3D lift-up model results and pixelwise 2D joint location error on evaluation datasets in Figure 8.

All models improve the performance compared with the ground truth trained model (compared with Figure 6) because of learning robustness for the 2D joint location estimator's failure. In particular, *x*R-EgoPose shows similar performance in accuracy to Miura et al.'s model (L2 loss). However, Miura et al.'s model still has more robustness on large perturbation.

## 5 CONCLUSION

We evaluated camera perturbation robustness of the 3D lift-up models in two-step 3D ego-pose estimations for a mounted fisheye camera. We first generated synthetic dataset for training and evaluation, which increase the camera perturbations in several steps. The dataset is publicly available.

In the ground truth 2D / 3D training, Miura et al.'s lift-up model estimated 3D ego-pose in high accuracy for incremental camera perturbations. Additionally, the possibility of ground truth training is great benefit to apply two-step approach for 3D ego-pose estimations in unique fisheye camera. *x*R-EgoPose indicated comparable accuracy and robustness in the training on 2D estimator outputs but Miura et al.'s model still has superiority on large camera perturbations.

In future work, we develop a 3D ego-pose estimation system for a mounted omnidirectional camera in practical. The two-step estimation model deploys 3D unit vectorization module proposed by Miura et al., which has robustness for the camera perturbations caused by body motions.

## ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP8K18517, JP22H00661, and JST SPRING Grant Number JPMJSP2112.

## AVAILABILITY OF DATA

The datasets generated and/or analyzed during the current study are available under the license in the NIT-UVEC-OMNI repository https://drive.google.com/drive/folders/1SbdaCIDhijvYdaFDdRiL_NTlRwk9xc00?usp=share_link.

## REFERENCES

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6).

Martinez, J., Hossain, R., Romero, J., and Little, J. J. (2017). A Simple Yet Effective Baseline for 3d Human Pose Estimation. In *IEEE International Conference on Computer Vision*, pages 2659–2668.

Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D., and Theobalt, C. (2017). VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM Transactions on Graphics*, 36:1–14.

Miura, T. and Sako, S. (2022). Simple yet effective 3D ego-pose lift-up based on vector and distance for a mounted omnidirectional camera. *Applied Intelligence*.

Newell, A., Yang, K., and Deng, J. (2016). Stacked Hourglass Networks for Human Pose Estimation. In *European Conference on Computer Vision*, pages 483–499.

Pavlakos, G., Zhou, X., Derpanis, K. G., and Daniilidis, K. (2017). Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1263–1272.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, New York, NY, USA. IEEE.

Scaramuzza, D., Martinelli, A., and Siegwart, R. (2006). A Toolbox for Easily Calibrating Omnidirectional Cameras. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5695–5701.

Sun, K., Xiao, B., Liu, D., and Wang, J. (2019). Deep High-Resolution Representation Learning for Human Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5686–5696.

Tekin, B., Katircioglu, I., Salzmann, M., Lepetit, V., and Fua, P. (2016). Structured Prediction of 3D Human Pose with Deep Neural Networks. In *British Machine Vision Conference*, pages 130.1–130.11.

Tome, D., Alldieck, T., Peluse, P., Pons-Moll, G., Agapito, L., Badino, H., and la Torre, F. D. (2020). SelfPose: 3D Egocentric Pose Estimation from a Headset Mounted Camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Tome, D., Peluse, P., Agapito, L., and Badino, H. (2019). xR-EgoPose: Egocentric 3D Human Pose From an HMD Camera. In *The IEEE International Conference on Computer Vision*, pages 7727–7737.

Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I., and Schmid, C. (2017). Learning from Synthetic Humans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4627–4635.

Wang, J., Liu, L., Xu, W., Sarkar, K., Luvizon, D., and Theobalt, C. (2022). Estimating Egocentric 3D Human Pose in the Wild With External Weak Supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13157–13166.

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional Pose Machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732.

Xiao, B., Wu, H., and Wei, Y. (2018). Simple Baselines for Human Pose Estimation and Tracking. In *European Conference on Computer Vision*, pages 472–487.

Xiaowei, Z., Menglong, Z., Spyridon, L., and Kostas, D. (2017). Sparse Representation for 3D Shape Estimation: A Convex Relaxation Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1648–1661.

Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Fua, P., Seidel, H.-P., and Theobalt, C. (2019). Mo2Cap2: Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2093–2101.

Zhou, X., Sun, X., Zhang, W., Liang, S., and Wei, Y. (2016). Deep Kinematic Pose Regression. In *European Conference on Computer Vision*, pages 186–201.