IPSJ Transactions on Computer Vision and Applications

RESEARCH PAPER Open Access

3D human pose estimation model using location-maps for distorted and disconnected images by a wearable omnidirectional camera

Teppei Miura* and Shinji Sako

Abstract

We address a 3D human pose estimation for equirectangular images taken by a wearable omnidirectional camera. The equirectangular image is distorted because the omnidirectional camera is attached closely in front of a person's neck. Furthermore, some parts of the body are disconnected on the image; for instance, when a hand goes out to an edge of the image, the hand comes in from another edge. The distortion and disconnection of images make 3D pose estimation challenging. To overcome this difficulty, we introduce the location-maps method proposed by Mehta et al.; however, the method was used to estimate 3D human poses only for regular images without distortion and disconnection. We focus on a characteristic of the location-maps that can extend 2D joint locations to 3D positions with respect to 2D-3D consistency without considering kinematic model restrictions and optical properties. In addition, we collect a new dataset that is composed of equirectangular images and synchronized 3D joint positions for training and evaluation. We validate the location-maps' capability to estimate 3D human poses for distorted and disconnected images. We propose a new location-maps-based model by replacing the backbone network with a state-of-the-art 2D human pose estimation model (HRNet). Our model is a simpler architecture than the reference model proposed by Mehta et al. Nevertheless, our model indicates better performance with respect to accuracy and computation complexity. Finally, we analyze the location-maps method from two perspectives: the map variance and the map scale. Therefore, some location-maps characteristics are revealed that (1) the map variance affects robustness to extend 2D joint locations to 3D positions for the 2D estimation error, and (2) the 3D position accuracy is related to the 2D locations relative accuracy to the map scale.

Keywords: 3D pose estimation, location-maps, Omnidirectional camera, Equirectangular image, Distortion, Disconnection

1 Introduction

Human pose motion capture is widely used in some applications, for example, computer graphics for movies and games, sports science, and sign language recognition. For this purpose, easy and low-cost methods are needed to capture the human pose motion. One of the main methods is human pose estimation. In recent years, human pose estimation has been actively researched, and deep neural network (DNN) has achieved considerable attention.

In human pose estimation research, RGB or RGB-D cameras are commonly used for input devices that take videos, images, or depth data. The input data are typically taken from the second-person perspective, and the data include approximate parts of the target person's body. DNN models estimate 2D or 3D joint positions from the input data.

*Correspondence: t.miura.288@nitech.jp
Department of Computer Science, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, 466-8555, Nagoya, Japan

Pose estimation methods for images taken from the first-person perspective is called the "egocentric-view pose estimation." In the egocentric-view setting, images or depth data are taken by body-attached devices. 3D pose estimation for the egocentric-view inputs is portable and trackable to a specific person. The 3D pose estimation, however, usually captures only limited joints of the human body because the input devices have a bounded angle of view. Therefore, it is extremely difficult for a body-attached camera to obtain images that include enough information for whole joint estimation. Additionally, the dynamic parts of the body (e.g., hands or feet) frequently move out of the camera angle view. These conditions make the estimation difficult.

We intend to apply 3D human pose estimation for sign language recognition and translation. Sign language is the visual communication method used by deaf people across the world; however, each region has different signs, as with oral language. Sign language is composed of some elements: handshapes, movements, positions, facial expressions, and peripheral information. For example, when the index finger points to something in sign language sentences, the meaning changes depending on what it is pointing to.

Some methods have been proposed for sign language recognition and translation from images [1, 2]. Most existing research handle regular images that are taken from the second-person perspective; thus, these approaches must install a camera in front of a signer in use scene. To overcome the restriction, we use a wearable camera for the input device because the system is available everytime, everywhere for signers.

An angle of the wearable camera view is not enough to obtain information for sign language recognition because sign language represents the meaning using a reachable space by hands and peripheral information. For example, a "head" of sign language is represented by pointing head part by index finger. However, if the index finger points to another person, the sign means "you" or "the person." For the reason above, the capability of capturing whole surrounding view is necessary for the wearable camera in our approach. Additionally, the tracking signer's pose is also important for sign language recognition because some signs represent the meaning by relative positional relation of body parts. For example, if the index finger points to around own face or chest, the sign means "myself".

We intend to propose a sign language recognition system using a wearable omnidirectional camera for the input device, which is portable for daily mobile use, and capable of obtaining enough elements for sign language recognition. As a first step for the system, we research 3D human pose estimation models for an RGB image taken by the wearable omnidirectional camera in this paper.

An omnidirectional camera can capture all surrounding information on a plane image, which is converted to an equirectangular image in our setting. We attach an omnidirectional camera to an area in front of a person's neck to obtain images including sign language elements: the face, hands, and the peripheral environment. Figure 1 shows the omnidirectional camera setting and the equirectangular image taken by the device. The omnidirectional camera closely attached to the human body captures images that have the following characteristics that are different from regular images.

Distortion: Objects that are placed around the polar points of the camera are displayed wider than the true image.

Disconnection: Objects that are placed on the border are divided into both edges of the image. Therefore, some parts of the human body often do not connect. For instance, when a hand goes out to an edge, the hand comes in from another edge.

Our approach is based on a convolutional neural network (CNN) similar to most of the recent monocular 3D human pose estimation methods. The existing methods, however, cannot apply well to our setting. First, their training data were captured with regular cameras placed at a position in which the cameras can capture almost the whole body from the second-person perspective. Second, most of the existing methods assume a skeletal structure on the image plane when the methods extend 2D joint

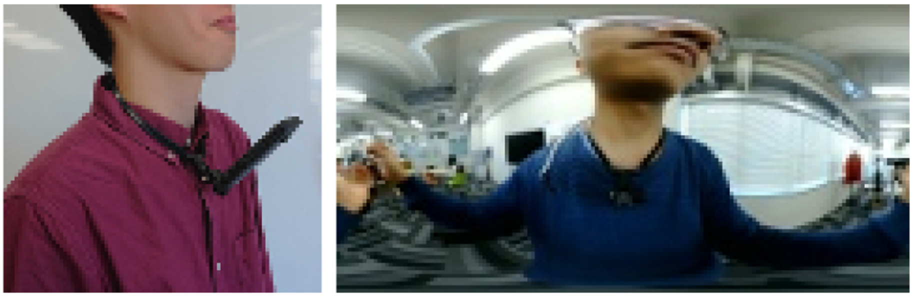

Fig. 1 A wearable omnidirectional camera is closely attached in front of the neck (left) , and an image is taken by the camera (right)

locations to 3D positions. For the reasons above, their methods fail not only 3D pose estimation but also 2D estimation, which is the basic step for 3D pose estimation, for our distorted and disconnected images. Figure 2 shows the estimation results of an existing 2D pose estimation method and our model.

To overcome these difficulties, we collect a new dataset captured by a wearable omnidirectional camera. More importantly, we introduce the location-maps method that is used to extend 2D joint locations to 3D positions in VNect, which is the 3D human pose estimation model proposed by Mehta et al. [4]. The method does not assume human body structures, and separately derives each x, y, and z position in 3D coordinates from 2D joint locations. Therefore, the location-maps method can reduce the impact of optical properties, which are the distortion and disconnection of equirectangular images. Xu et al. [5] indicated valid results of VNect for distortion images taken by a fish-eye camera placed at a position close to the body.

We validate that the location-maps method has the capability to estimate 3D joint positions with not only distortion but also disconnection caused by the wearable omnidirectional camera. Furthermore, we propose a new estimation model using the location-maps method by replacing the backbone network with a state-of-the-art 2D pose estimation model [6]. Our model is a simpler architecture than VNect, which proposed the location-maps method. In the Section 5, we evaluate our model and VNect in terms of accuracy and computation complexity. In the Section 6, we analyze the location-maps characteristics from two perspectives: the map variance and the map scale.

To the best of our knowledge, our work is the first approach to estimate 3D human poses from an omnidirectional camera closely attached to the body. Although the proposed method is a combination of existing techniques, our work is practical and useful from an application viewpoint. The contributions of this paper are summarized as follows:

- We collect a new sign language dataset that is composed of equirectangular images and synchronized 3D joint positions. The equirectangular images are taken by a wearable omnidirectional camera in our setting.
- We propose a new 3D human pose estimation model using the location-maps method for distortion and

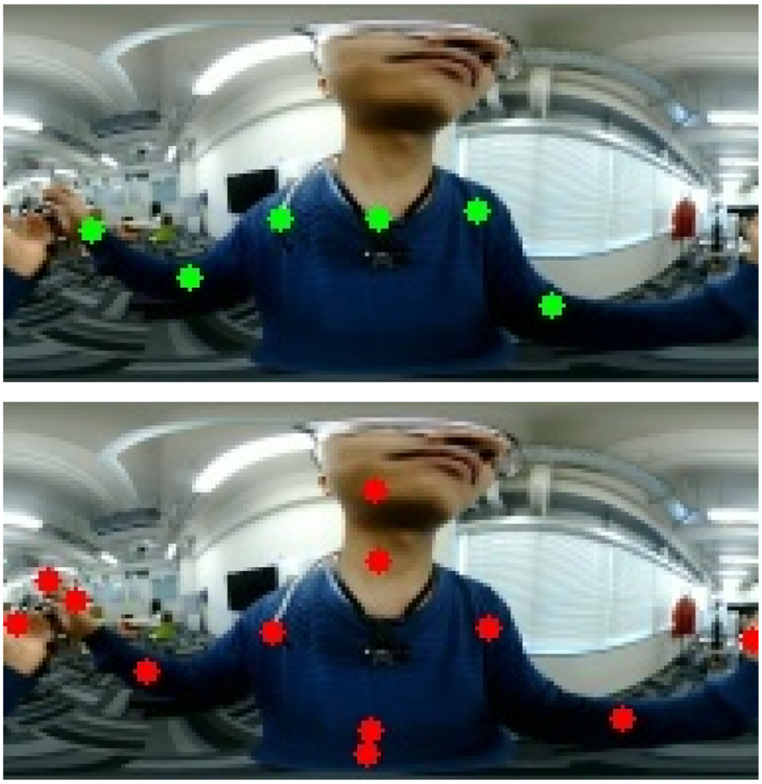

Fig. 2 Openpose [3], which is a 2D human pose estimator, fails the left wrist and nose on an image captured by our camera setting (top). Our model can estimate whole joints, including the left wrist and the left hand, which are disconnected on the image (bottom). This image shows the result of relocated 3D pose estimation on the 2D plane

disconnection images. The model is a simpler architecture than VNect, which is the reference model. Nevertheless, our model's performance is better with respect to accuracy and computation complexity.

- We reveal the location-maps characteristics that (1) the map variance affects robustness to extend 2D joint locations to 3D positions for the 2D estimation error, and (2) the 3D position accuracy is related to the 2D locations relative accuracy to the map scale.

2 Related works

Human pose estimation has been researched on learning-based and model-based approaches from various resources— RGB images, depth data, or MoCap data—and considerable recent progress has been achieved through CNN-based approaches. Our goal is to estimate the 3D human pose for RGB equirectangular images, which are distorted and disconnected, taken by a wearable omnidirectional camera. We discuss relevant approaches to estimate 3D human poses for RGB images.

One of the methods directly estimates 3D human poses for RGB images. Gerard et al. [7] and Catalin et al. [8] proposed models based on the relative positional relationship of human body parts. However, most direct estimation models are CNN-based to estimate 3D human poses [9–15], and then applied to subinformation to improve accuracy. Du et al. [11] proposed applying height maps generated with precalibrated monocular cameras. The height maps represent the length of human parts for input. Tekin et al. [10] fused confidence maps of 2D joint locations to 3D human pose estimation. Zhou et al. [15] fitted a kinematic model restriction to direct estimation.

However, the mainstream method extends 2D representations for RGB images to 3D human poses. The representations are extracted as silhouettes [16], body shapes [17, 18], or joint locations [4, 19–24]. In these methods, the quality of 2D representations directly affects the accuracy of 3D human pose estimation. CNN-based models [6, 25–28] proposed high-quality estimation methods of 2D joint locations for RGB images. These methods represent the confidence in joint locations on 2D planes as heatmaps.

The methods to extend 2D representations to 3D human poses generally use a kinematic model including restrictions of the human body [17, 20–23]. However, we introduce the location-maps method for our approach, which was proposed by Mehta et al. [4, 24] to extend 2D joint locations to 3D positions. The location-maps are generated to the x, y, and z positions as 2D planes without kinematic model restrictions and consideration of optical properties for the wearable omnidirectional camera.

As the most related research, Xu et al. [5] proposed a 3D human pose estimation model for single fish-eye camera images attached at a hat brim. The images are distorted because the fish-eye camera is closely attached to the body. The method extends 2D joint locations to 3D positions considering the optical properties for fish-eye camera settings. In their experiments, the validity of the location-maps method for distorted images was indicated.

3 Data collection

We collect a new dataset composed of RGB equirectangular images and synchronized 3D joint positions. We make the dataset following steps (1) and (2). (1) RGB equirectangular images are taken by an omnidirectional camera that is closely attached in front of a person's neck. The omnidirectional camera captures the complete surrounding information, including the face, hands, and other upper body parts on the RGB equirectangular image. (2) We capture 3D upper body joint positions simultaneously by skeleton estimation. The skeleton is obtained from RGB-D camera data for the person to whom the omnidirectional camera is attached.

3.1 Format

RGB equirectangular images: We collect equirectangular images with the fixed pole axis of an omnidirectional camera. At the setup of the camera with body attachment, we set the pole axis to be perpendicular to the ground. Therefore, the center joints (head, neck, torso, and waist) will always be around the centerline of the width on images, even if the body and the camera tilt. The omnidirectional camera captures RGB equirectangular images in the aspect ratio of 1:2.

3D joint positions: We collect the 12 joint positions listed below as the upper body pose in 3D coordinates.

Center: Head, neck, torso, waist
Left: Left shoulder, left elbow, left wrist, left hand
Right: Right shoulder, right elbow, right wrist, right hand

In the collection scheme, we normalize the joint positions by the steps outlined below to simplify data handling.

1 Move all joints' absolute coordinates to relative coordinates with the camera position as root, where the camera position is manually fixed in a position that moves forward (shoulder width multiplied by 0.4) and downward (shoulder width multiplied by 0.1) from the neck. We determined the position according to actual camera position.

2 Rotate all joints' coordinates around the root position according to the shoulder line to be

parallel to the x-axis. The foreside of skeleton faces the $-z$ orientation after the rotation.

3 Rotate all joints' coordinates around the root position according to the line connected neck and torso to be parallel to the y-axis.

4 Move all joints' coordinates to enlarge or shrink according to the Euclidean distance between shoulders (shoulder width) to be 1.0. The joint positions are determined by

$$w = ||P_{l\_shoulder} - P_{r\_shoulder}||,$$
$$P_j = P_j \div w,$$

where P_j indicates the position of joint j in 3D coordinates. P_j is composed of x_j, y_j, and z_j.

We show examples of an equirectangular image and normalized 3D joint positions collected simultaneously in Fig. 3. Note that 2D joint locations on the equirectangular image are derived from 3D joint positions relative to the omnidirectional camera position according to the equirectangular projection [29].

3.2 Collection system

The data collection system is made of the main components listed below.

- Omnidirectional camera: Ricoh R Development Kit
- RGB-D camera: Intel RealSense Depth Camera D435
- Skeleton estimation software: Nui Track ver 1.3.5 (on Windows x64)

We hang an attachment device for the omnidirectional camera around a person's neck and then put the camera on the attachment. Therefore, the camera lens is placed in front of the neck in our setting. We take the person who is attached to the omnidirectional camera with the RGB-D camera from the second-person perspective. 3D joint positions are collected by the skeleton estimation software from the RGB-D camera. We illustrate the entire hardware setting of the data collection system in Fig. 4.

The data collection system simultaneously obtains an equirectangular image and 3D joint positions, and then converts the data according to each format that is defined in Section 3.1. Note that the skeleton estimation software occasionally fails to obtain 3D joint positions if a part of the human body has occlusions or is hidden. To deal with the unstable estimation, the collection system records the data only when the software succeeds to obtain all 3D joint positions; therefore, the frame rate of the dataset is not constant. The inconstant frame rate, however, is not a serious problem because the 3D human pose is estimated independently for each frame, without consideration of sequentiality, in our model.

3.3 Dataset detail

We use Japanese sign language motions as collecting data. We compose 16 example sentences that cover all of the hand position and movement definitions in a dictionary. The examples cover all 22 classes of hand positions, all 7 classes of hand movements, and 38 classes of handshapes in 59 classes. We select as many classes of handshapes as possible when composing examples to cover the hand position and movement definitions. The example sentences have 75 sign language words without overlapping.

We record the motions for 7 actors according to the 16 example sentences with the data collection system. Actors perform the sentences 3 or 4 times for each example. We do not specify clothes and eye-glasses that actors should wear. We conduct a leave-one-person-out cross validation for Section 5. We make datasets for each actor: the test data is an actor's data; the training and validation data are other 6 actors' data. The other 6 actor's mixed data are allocated to 5:1 into training and validation. We show the collected datasets in Table 1.

We collected 27,077 frames as test dataset-A at first. We, however, reduced the volume to 18,050 because we concerned the impact of quantitative imbalance to calculate results according to evaluation metrics. In the reduction step, we uniformly sampled the frames following the time sequence. Other test datasets are used original volume of collection.

4 Approach

Our work is 3D human pose estimation for equirectangular images taken by a wearable omnidirectional camera closely attached in front of a person's neck. We intend to apply this method to a sign language recognition system because of portability and capability to obtain sign language elements on 2D images: handshapes, movements, positions, facial expressions, and peripheral information. The equirectangular images are distorted and disconnected caused by the omnidirectional camera and close distance to the body. The distortion and disconnection make the 3D human pose estimation challenging.

We address the challenge by introducing the location-maps method that was proposed by Mehta et al. [4, 24]: however, they only proposed the location-maps to estimate 3D human poses for regular images without distortion and disconnection. Furthermore, we apply a state-of-the-art 2D human pose estimation model [6] as the backbone network to the location-maps for improvement of simplicity and accuracy.

4.1 The location-maps for 3D human pose estimation

location-maps method is one of the methods for extending 2D joint locations to 3D positions with respect to 2D-3D consistency on an image plane. In the VNect [4] model, the network generates 4 maps: heatmap H, and

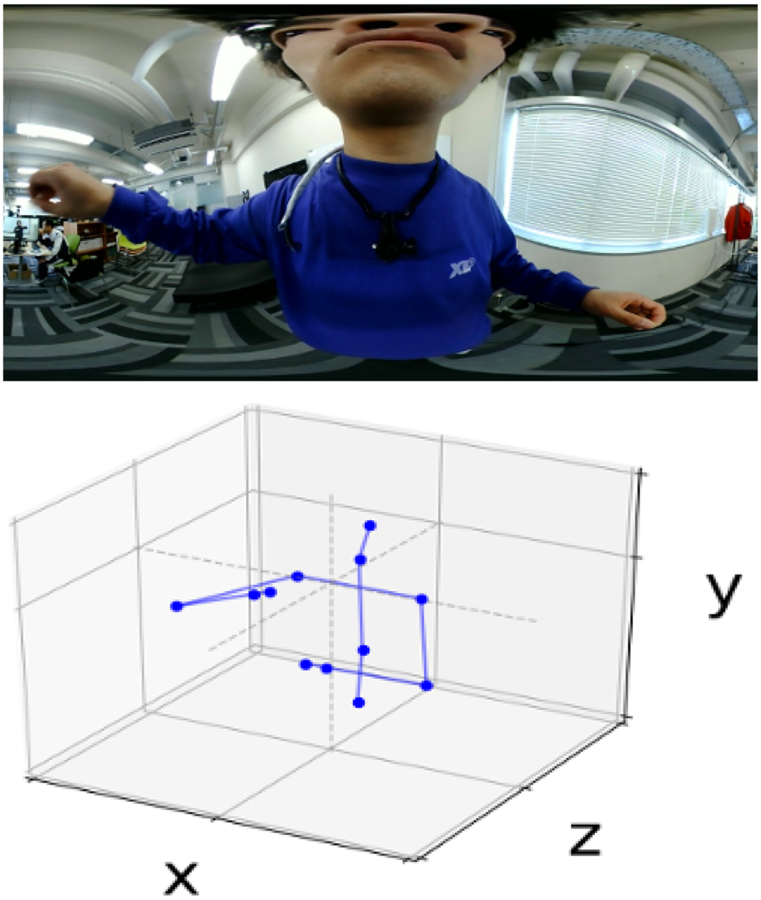

Fig. 3 An equirectangular image (top) and (bottom) normalized 3D joint positions

location-maps X, Y, and Z for each joint in the final stage. Then, 3D joint positions are derived from a combination of 4 maps.

The heatmap represents the confidence in each 2D joint location as a 2D probability distribution, and thus, a joint location is determined at the maximum cell in the heatmap. The 3D joint positions are derived according to the value of location-maps at the same 2D joint locations that were determined by the heatmap. To represent as a formulation, for each joint $j \in J$, where J is the number of joints, joint positions x_j in the 3D coordinates are

$$row_j, col_j = \text{argmax}\left(H_j\right), \tag{1}$$

$$x_j = X_j\left(row_j, col_j\right), \tag{2}$$

where H_j and X_j indicate the heatmap and the x location-map respectively, and argmax is the function that outputs indexes of row and column at the maximum cell in the

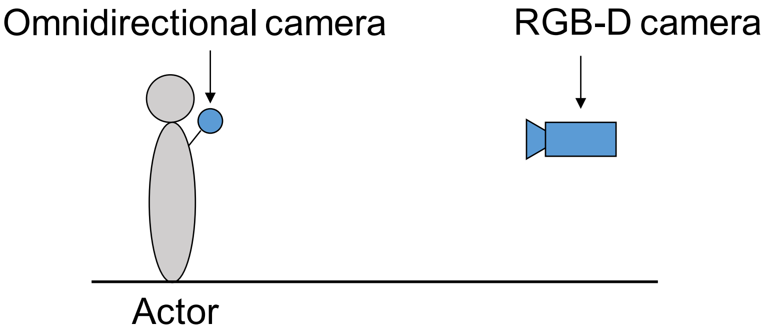

Fig. 4 Hardware setting of the data collection system

Table 1 The collected datasets. The number indicates the volume of data. The test data is an actor's data(dataset ID's actor). The training and validation data are comprised of other 6 actors' mixed data

Dataset	Test	Training	Validation
A	18,050	94,713	18,943
B	17,045	95,550	19,111
C	19,754	93,293	18,659
D	17,695	95,009	19,002
E	19,320	93,655	18,731
F	20,627	92,565	18,514
G	19,215	93,742	18,749

heatmap. We visualize the scheme for estimating 3D joint positions using heatmaps and location-maps in Fig. 5.

In the training session, the model studies to regress a heatmap H_j to a 2D Gaussian map that indicates the confidence in 2D joint location over the input image for each joint. location-maps X_j, Y_j, and Z_j are also studied. For 3D joint positions x_j, y_j, and z_j, the L2 loss formulation for x_j is

$$\mathrm{Loss}\left(x_j\right) = ||H_j^{\mathrm{GT}} \otimes \left(X_j - X_j^{\mathrm{GT}}\right)||_2, \tag{3}$$

where GT indicates the ground truth, ⊗ is the Hadamard product, and X_j^{GT} is the uniform distribution of x_j. The location-maps loss formulation indicates that the loss is weighted stronger around the 2D joint location by the ground truth heatmap H_j^{GT}.

The most important characteristic of the location-maps method is that the maps are studied according to only 2D-3D consistency that consists of 2D joint locations and 3D joint positions x_j, y_j, and z_j. Therefore, the location-maps can extend 2D joint locations to 3D positions without causing the effect of image distortion and disconnection because the method does not study the optical properties of input images and restrictions of the human body structure on the location-maps in the training session. We focus on the beneficial characteristics of the location-maps method, which did not mention the distortion and disconnection of images in the reference paper.

4.2 Proposed model

Mehta et al. [4] used the ResNet-based [30] architecture as the backbone network in the VNect model. ResNet has been widely used as the base network architecture for image recognition research in recent years. Mehta et al. replaced the layers of ResNet from *res5a* onwards with their own architecture including intermediate supervisions that represent the bone length for each joint.

We replace the backbone network with the high-resolution network (HRNet) of Ke et al. [6] for a 3D human pose estimation model using the location-maps method. HRNet is a state-of-the-art 2D human pose estimation model for images. The HRNet maintains high-resolution representations through the whole process, while conducting multi-scale fusions such that the high-to-low resolution representations of subnetworks that are produced in each stage.

Our model estimates heatmaps and location-maps that are potentially more accurate and spatially more precisely by replacing the backbone network with the HRNet. In addition, our model is a simpler architecture because it does not include intermediate supervisions of bone length as in the VNect.

5 Performance evaluation

We evaluate our model compared with the VNect of Mehta et al. [4], which is a reference 3D human pose estimation model using the location-maps method. In addition, we confirm that our model can estimate 3D human poses for distorted and disconnected images. Mehta et al. experimented with their model based on ResNet50 and ResNet100 as the backbone network. They proposed that the ResNet50-based model is more reasonable with respect to accuracy and computation complexity. We implement our model with the parameter size and the computation complexity according to the experiments.

5.1 Implementation and training for model

We apply HRNet-W24 and HRNet-W32 with our backbone network in our evaluation, where 24 and 32 represent the widths (channel) of the feature maps respectively. We implement 4 stages of the high-resolution network. Therefore, the widths of the other three parallel subnetworks are 48, 96, and 192 for the HRNet-W24-based model and 64, 128, and 256 for the HRNet-W32-based model. Figure 6 illustrates the architecture of our model based on HRNet-W24.

We train our model and the VNect with the training dataset. The schedule is set to a batch size of 64 and 50 epochs. The input image size is fixed at 96 (height) ×192 (width), and the variance in heatmaps are set to 1.0 in both models. We use the Adam optimizer [31], and set the initial learning rate at 0.001.

5.2 Results and evaluation

We use mean per joint position error (MPJPE) metrics and percentage of correct keypoints (PCK) metrics for evaluation. The error is the Euclidean distance between estimation and ground truth of joint position in 3D coordinates, where the shoulder width is 1.0 because of 3D joint position normalization when collecting the dataset. PCK metrics indicate a percentage of correct joint that the estimation error is below a threshold. Note that PCK is a more robust measure because MPJPE is heavily influenced by large outliers. In addition,

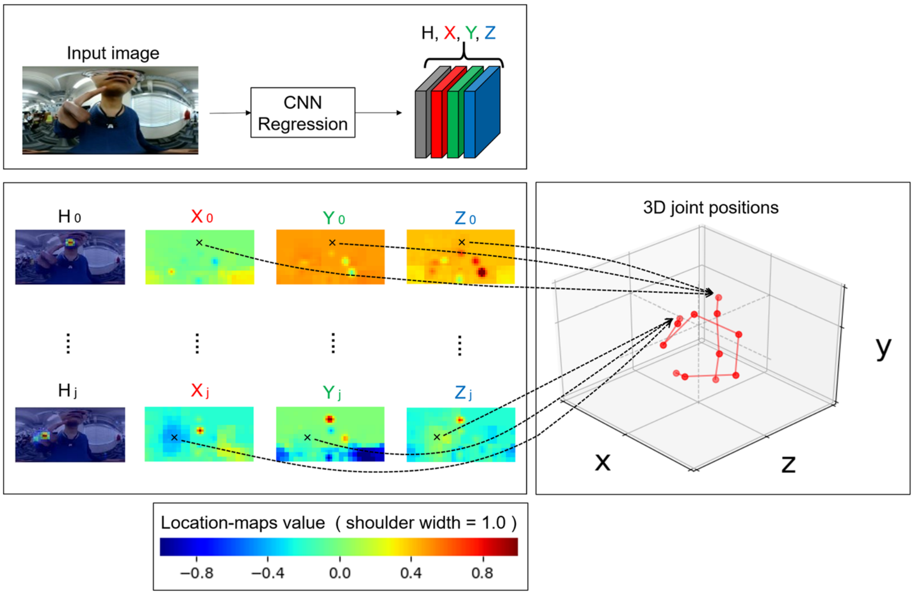

Fig. 5 Scheme of estimating 3D human poses using heatmaps and location-maps. The CNN regression model generates heatmaps and location-maps for each joint from an input RGB image. The 3D coordinate joint positions are estimated from their location-maps X_j, Y_j, and Z_j at the location of the maximum in the heatmap H_j

MPJPE has possibility to indicate better results for models that estimate shivering pose around average position. Therefore, PCK is more adequate metrics to evaluate 3D human pose estimation. Reference papers [4, 6] also mainly evaluate the model performance by PCK metrics.

Table 2 shows the MPJPE performance, and Table 3 shows the PCK performance on the validation and test datasets. In the validation dataset, the actors are the same persons as in the training dataset; however, the data are not included in the training scheme, and the actor of the test dataset is not included in the training dataset. For the reasons above, the test dataset is more difficult than the validation dataset.

The output map scale is double that of VNect in our model because of the HRNet-backbone's merit that generates high-resolution maps. We compare the models that output different scale maps from same scale inputs for evaluating the backbone characteristics.

Our model (HRNet-W24) is larger than VNect(ResNet50); however, our model is more efficient in terms of model size (#Params) and computation complexity (GFLOPs). In the MPJPE metrics (Table 2), our model is slightly better on the validation dataset; in

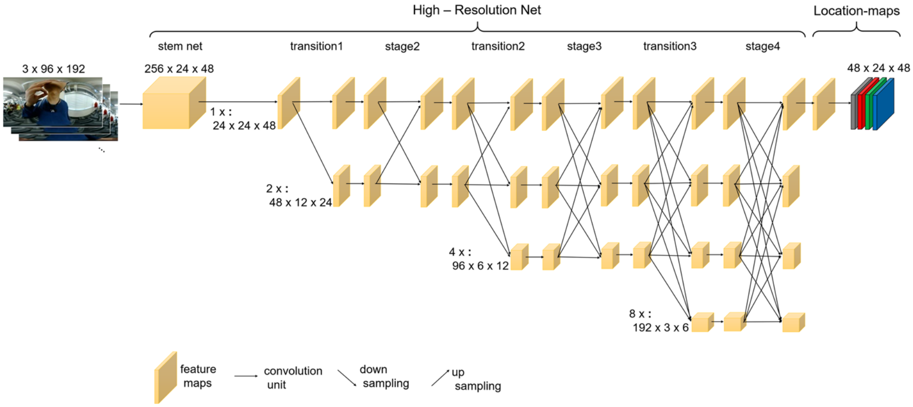

Fig. 6 The network architecture of our model based on HRNet-W24. The stem net convolutes input images to 256 (channel) ×24 (input height /4) ×48 (input width /4) regardless of the number of W. The network makes branches after the stem net according to W. The number of output maps is 48 because of 4 maps (H, X, Y, and Z) for each of the 12 joints in our setting

Table 2 MPJPE performance comparison on the validation and test dataset. The table shows some parts of joints in detail, however the All includes other joints (head, neck, torso, and waist). Lower value is better in this metrics

Model	Backbone	In-scale	Out-scale	Var.	#Params	GFLOPs	Shoulders	Elbows	Wrists	Hands	All
Validation dataset											
VNect	ResNet50	96 x 192	12 x 24	1.0	14.5 M	1.70	0.019	0.053	0.069	0.079	0.044
VNect	ResNet100	96 x 192	12 x 24	1.0	33.5 M	2.97	0.019	0.052	0.067	0.076	0.043
Ours	HRNet-W24	96 x 192	24 x 48	1.0	16.7 M	1.70	*0.015*	*0.044*	*0.060*	*0.069*	*0.037*
Ours	HRNet-W32	96 x 192	24 x 48	1.0	29.3 M	2.70	*0.015*	0.045	0.062	*0.069*	0.038
Test dataset											
VNect	ResNet50	96 x 192	12 x 24	1.0	14.5 M	1.70	0.040	0.218	0.387	0.439	0.201
VNect	ResNet100	96 x 192	12 x 24	1.0	33.5 M	2.97	0.037	0.215	*0.379*	*0.434*	*0.200*
Ours	HRNet-W24	96 x 192	24 x 48	1.0	16.7 M	1.70	0.040	0.215	0.395	0.467	0.204
Ours	HRNet-W32	96 x 192	24 x 48	1.0	29.3 M	2.70	*0.036*	*0.202*	0.393	0.463	0.201

The best performance estimations are italicized

contrast, VNect is slightly better on the test dataset. Turn into the PCK metrics (Table 3), our model indicates better performance on both of validation and test dataset. Thus, our model estimates 3D joint positions closer to ground truth; however, the model has possibility to obtain the larger outliers for unknown dataset. The outlier estimation causes the deterioration of MPJPE for our model. For the discussion above, our model based on HRNet performs better with respect to accuracy and computation complexity from the viewpoint of PCK metrics that is generally used to evaluate the performance of 2D/3D pose estimation.

We present some estimation results for distorted and disconnected images by our model (HRNet-W24) on the test dataset in Fig. 7. The blue line indicates the ground truth of the 3D human pose, and the red line indicates the estimation results. As shown in the figure, the location-maps method can estimate 3D human poses even if the images are distorted and disconnected.

Table 3 PCK performance comparison on the validation and test dataset. The table shows the thresholds of 0.1, 0.2, and 0.3. Higher value is better in this metrics

Model	Backbone	PCK @ 0.1	PCK @ 0.2	PCK @ 0.3
Validation dataset				
VNect	ResNet50	91.98	98.53	99.56
VNect	ResNet100	92.42	98.64	*99.60*
Ours	HRNet-W24	*94.02*	*98.79*	*99.60*
Ours	HRNet-W32	93.91	98.75	99.59
Test dataset				
VNect	ResNet50	48.11	65.74	77.63
VNect	ResNet100	49.49	67.01	78.11
Ours	HRNet-W24	49.65	67.07	77.89
Ours	HRNet-W32	*50.45*	*68.09*	*79.12*

The best performance estimations are italicized

5.3 Discussion for results and models

We discuss differences between proposed model and VNect in detail by estimation results. Figure 8 shows success and failure examples of models' estimation respectively in MPJPE metrics. The estimation deterioration has tendencies to occur when the input images has the occlusion and difficult view of hand by overlapping or distance from the camera. The deterioration tendencies, however, are found in common of both model's results. We cannot find clear differences and tendencies in the input images between proposed model and VNect.

For the further discussion, we indicate heatmaps and location-maps generated from the same success and failure cases in Fig. 9. The maps are shown of only worst joint in the failure examples, and the right hand joint maps are shown in the success. The ground truth of location-maps are uniform distributions of x, y, and z values of the joint position. Note that VNect maps are displayed on a double scale for easy comparison. From comparing heatmaps between success and failure cases, the poor heatmap estimation obviously cause to deteriorate the 3D joint position estimation. Turn into comparison of location-maps, we can find a tendency that VNect generates slight wider area close to ground truth in the location-maps. The wider area generation leads robustness to heatmaps deterioration when extending 2D locations to 3D positions by the location-maps. Therefore, the failure of 3D pose estimation is caused by heatmaps estimation deterioration in both model. VNect, however, reduces the impact by the more robust location-maps; in contrary, proposed model is influenced more by heatmaps deterioration. Relative heatmap variance to the map scale determines how wide

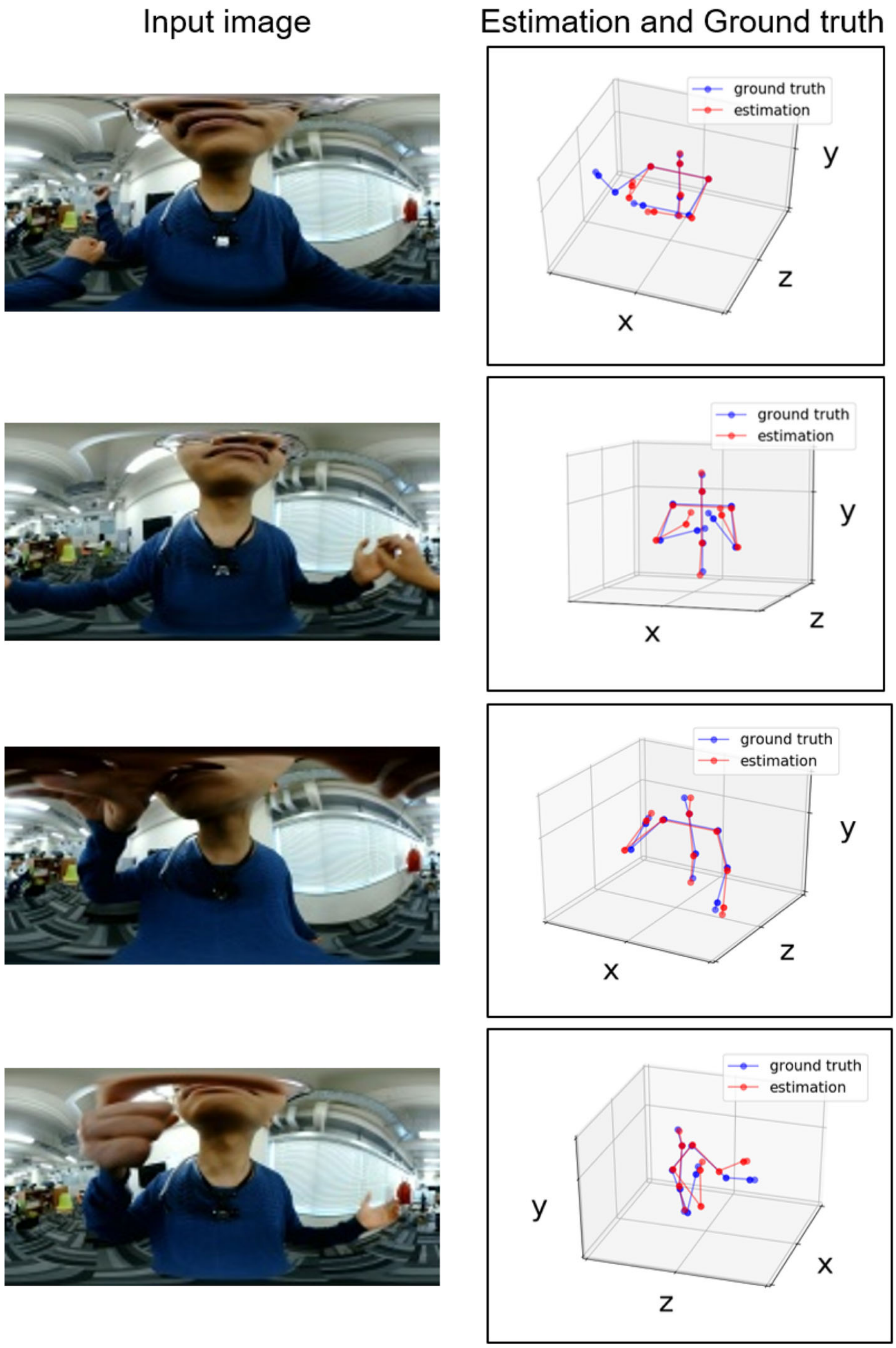

Fig. 7 Estimation results and ground truth for distorted and disconnected images. These results are estimated by our model(HRNet-W24) on the test dataset

the model generates area close to ground truth in the location-maps. We analyze the relationship of heatmap variance and location-maps robustness in Section 6.2.

From the discussion above and PCK evaluation, we believe that proposed model indicates better performance than VNect under the condition that the heatmaps estimation is success, because the backbone network (HRNet) is more rich architecture that repeatedly conducts multi-scale fusions such that the high-to-low resolution representations. In contrary, our model indicates poor performance than VNect under the condition that the heatmaps estimation is deteriorated, because the

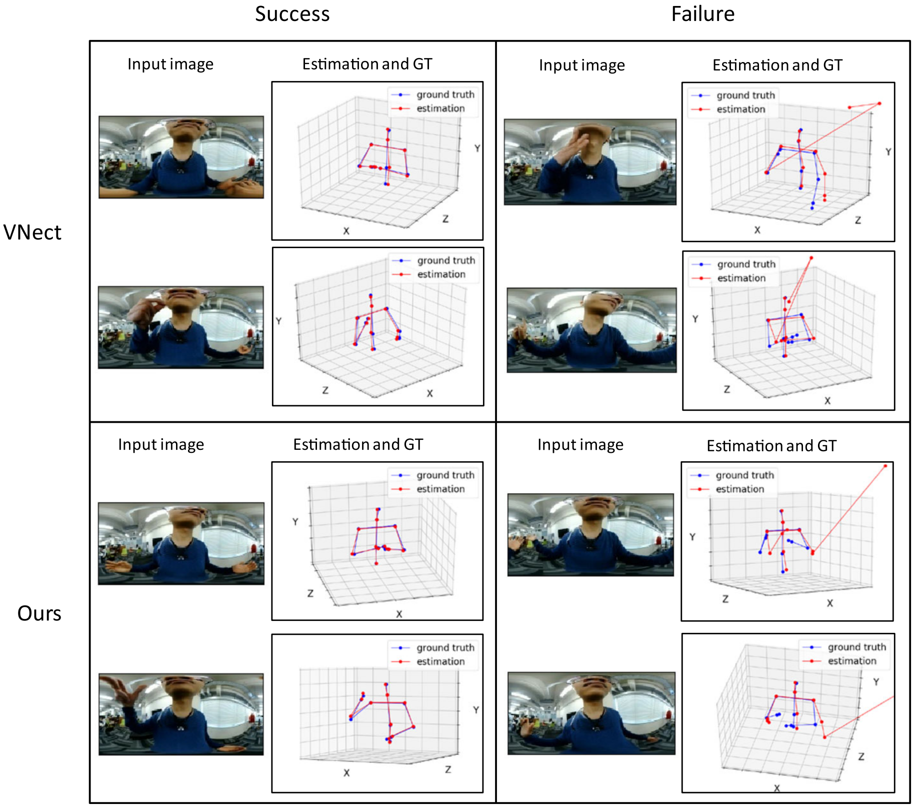

Fig. 8 Estimation and ground truth examples in success (left column) and failure (right column) cases. The upper row shows VNect (ResNet50) results, and the lower shows our model (HRNet-W24)

location-maps have lower robustness for heatmaps deterioration.

6 location-maps analysis

We conduct additional experiments to analyze the characteristics of the location-maps method from two perspectives. In the first perspective, we focus on the value of heatmap variance because our model is trained according to the loss formulation (3) that is composed of the location loss weighted by the heatmap. Therefore, the value of heatmap variance directly affects the generation of location-maps. Second, we change the output scale of our model according to the scale level of HRNet. The scale of maps affects the spatial precision of heatmaps and location-maps generated by our model.

6.1 Implementation and training

We modify the HRNet-W24 based model according to the analysis perspectives. The heatmap variance is changed to 4.0 and 8.0, while other parameters are the same as the base model. Additionally, we reimplement the location-maps section in Fig. 6 to 2x-scale and 4x-scale; thus, the output scale is 12 ×24 in the 2x-scale model, and the 4x-scale model is 6 ×12. Overall, we train and evaluate the base model and the 4 additional models that are 2 different variances and 2 different scales.

We use transfer learning and fine-tuning for these 5 models because of the reducing effect of the initial parameter for analysis. The trained parameter in the Section 5 is set to all models as the initial parameter, and then, fine-tuning is conducted with the dataset-A until 30 epochs with the batch size of 64. Other training parameters are the same as in the Section 5.

6.2 Analysis for the variance in maps

Table 4 shows the MPJPE results of the base model (var. 1.0) and the different variance models(var. 4.0 and 8.0) on the validation and test dataset-A. The error is nothing for the neck because the joint is the root position from our setting that the omnidirectional camera is set around the neck.

Considering the accuracy of each joint, the results indicate a tendency in which the smaller variance model estimates better for the static joints (e.g., head, torso, and waist); however, the larger variance model estimates better for the dynamic joints (e.g., wrists and hands). In the mean of all joints, the smaller variance's result is better on the validation dataset-A, and vice versa on the test dataset-

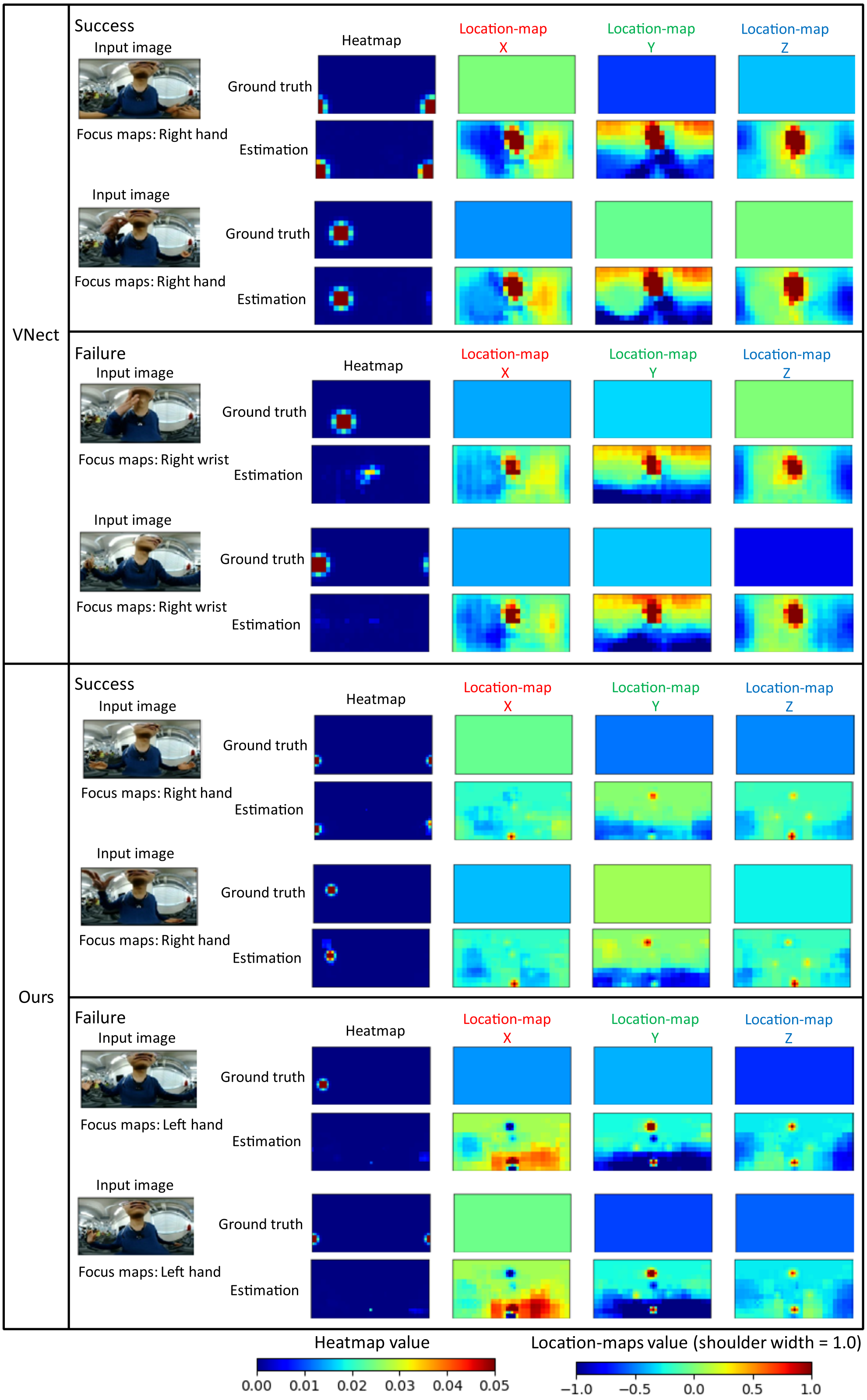

Fig. 9 Heatmaps and location-maps for input images in the same success and failure cases of Fig. 8. The maps are shown of only worst joint in the failure examples, and the right hand joint maps are shown in the success. The ground truth of location-maps are uniform distributions of x, y, and z values of the joint position. The measures of value for heatmaps and location-maps are shown below the table. Note that VNect maps are displayed on a double scale for easy comparison

Table 4 MPJPE performance comparison in different variances (1.0, 4.0, and 8.0) of heatmap on the validation and test datasets-A

Model	Backbone	Var.	Head	Neck	Torso	Waist	Shoulders	Elbows	Wrists	Hands	All
Validation dataset-A											
Ours (var. 1.0)	HRNet-W24	1.0	*0.018*	0.000	*0.021*	0.022	*0.012*	0.037	0.051	0.058	0.031
Ours (var. 4.0)	HRNet-W24	4.0	0.029	0.000	0.022	*0.018*	0.016	*0.034*	*0.045*	*0.051*	*0.030*
Ours (var. 8.0)	HRNet-W24	8.0	0.048	0.000	0.031	0.031	0.025	0.051	0.054	0.062	0.041
Test dataset-A											
Ours (var. 1.0)	HRNet-W24	1.0	0.052	0.000	0.053	0.066	0.029	0.160	0.296	0.349	0.153
Ours (var. 4.0)	HRNet-W24	4.0	*0.048*	0.000	*0.050*	0.066	0.027	*0.154*	*0.285*	*0.325*	*0.146*
Ours (var. 8.0)	HRNet-W24	8.0	0.052	0.000	0.051	*0.061*	*0.026*	0.157	0.289	0.329	0.147

The best performance estimations are italicized

A. According to the results above, the larger variance model has the robustness to estimate 3D joint positions for the difficult data that are the unknown or the dynamic joints.

In addition, we look at more detail of heatmaps and location-maps generated by the model. Figure 10 presents the maps of right hand that are generated for an image. Note that the measures of value for heatmaps are different scales in each variance model.

The location-maps in Fig. 10 indicate that the larger variance model generates smoother and more spacious location-maps than the smaller variance model. For example, the x position of the right hand for the input image is between −0.1 and 0.0 referred from the heatmap and x location-map in the model (var. 1.0). The larger models (var. 4.0 and 8.0) generate the x location-map in which the value range between −1.0 and 0.0 occupies a wider area on each x location-map. Therefore, changing the value of heatmap variance affects the robustness of location-maps for the estimation error of 2D location by the heatmaps.

In contrast, considering the heatmaps in Fig. 10, the heatmaps of the larger variance model are also generated more smoothly, which means a smaller value gap between each cell on the map. Therefore, the larger variance model potentially includes more error to estimate the 2D joint location by the heatmap, and thus, the value of heatmap variance affects the accuracy of 2D joint location estimation.

The location-maps method derives 3D joint positions with heatmaps and location-maps. If the model has a large variance, the model generates robust location-maps and estimates better than the small variance model for the dynamic joints such as hands and wrists. However, the excessive enlargement of variance causes the accuracy to decrease for 2D joint location estimation, which leads to a decrease in the accuracy of 3D human pose estimation.

6.3 Analysis for the scale of maps

Figure 11 indicates the L2 loss transition on the validation dataset-A for the different scale models(1x, 2x, and 4x) in the training session. In the graph, the 4x-scale and 2x-scale model fall into a local solution after approximately 9 and 10 epochs. We believe that the behavior caused by the reduction in representation capability because the map scale is shrinking. In this analysis section, we use the 4x-scale model at the 16 epoch and the 2x-scale model at the 28 epoch, which are the least L2 loss model during the training session.

Table 5 shows the MPJPE results for the base model (1x) and the different scale models (2x and 4x). On the validation dataset-A, the larger scale model estimates better than the smaller scale model. This tendency is clearly different from the different variance models. However, the smaller scale model indicates better results on the test dataset-A according to the mean of all joints; however, the clear tendency is not found for each joint.

Figure 12 illustrates the mean L2 distance error for the different scale models on the test dataset-A. The left graph indicates the 3D joint position error (3DJPE) for all joints, and the center is the 2D joint location error (2DJLE) that is estimated by the only heatmaps, where a cell on the heatmaps is 1.0. The right graph indicates the standardized 2D joint location error (std2DJLE) that applied a coefficient considering each map scale with the 2DJLE. In the 2x-scale model, the std2DJLE for each joint $j \in J$ is

$$
\begin{aligned}
row_j, col_j &= \operatorname{argmax}\left(H_j\right), \\
error_{row} &= \left|\left(row_j - row_j^{\mathrm{GT}}\right) \times 2\right|, \\
error_{col} &= \left|\left(col_j - col_j^{\mathrm{GT}}\right) \times 2\right|, \\
\mathrm{std2DJLE}(j) &= \sqrt{error_{row}^2 + error_{col}^2},
\end{aligned}
$$

where argmax is the function that outputs the indexes of row and column at the maximum cell in the heatmap, and GT indicates the ground truth.

The 2DJLE proportionally decreases according to the heatmap scale in the center graph; however, the 3DJPE is more related to the std2DJLE from the left and right graphs. Thus, shrinking the scale of maps reduces the

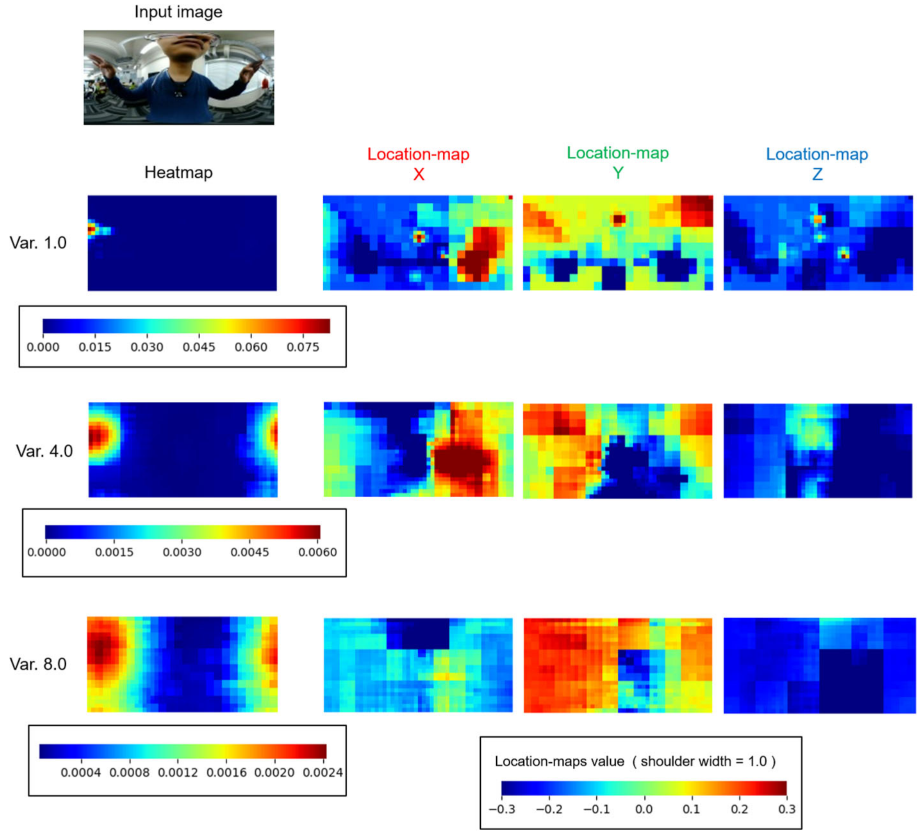

Fig. 10 Heatmaps and location-maps for an input image by each variance model (var.1.0, var.4.0, and var.8.0). Note that the measures of value for heatmaps are different scales, unlike location-maps measures

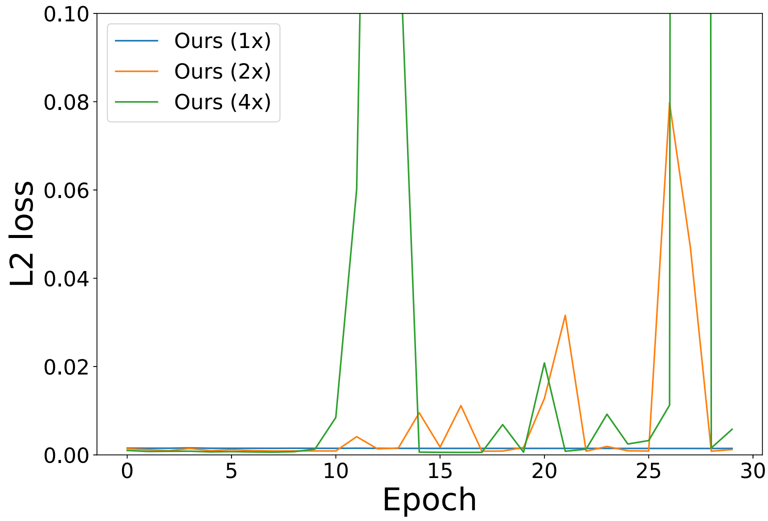

Fig. 11 The L2 loss transition on the validation dataset-A during fine-tuning for the different scale models(1x, 2x, and 4x). The 4x-scale and 2x-scale model fall into a local solution after approximately 9 and 10 epochs

Table 5 MPJPE performance comparison in different scales (1x, 2x, and 4x) of maps on the validation and test datasets-A

Model	Backbone	Out-scale	Head	Neck	Torso	Waist	Shoulders	Elbows	Wrists	Hands	All
Validation dataset-A											
Ours (1x)	HRNet-W24	24 x 48	*0.018*	*0.000*	0.021	0.022	*0.012*	*0.037*	*0.051*	*0.058*	*0.031*
Ours (2x)	HRNet-W24	12 x 24	0.019	*0.000*	*0.019*	*0.021*	0.013	0.039	0.052	0.059	0.032
Ours (4x)	HRNet-W24	6 x 12	0.023	0.001	0.022	0.024	0.017	0.043	0.055	0.064	0.036
Test dataset-A											
Ours (1x)	HRNet-W24	24 x 48	0.052	0.000	0.053	0.066	*0.029*	*0.160*	*0.296*	0.349	0.154
Ours (2x)	HRNet-W24	12 x 24	0.053	0.000	*0.051*	0.071	0.030	0.170	0.330	0.379	0.166
Ours (4x)	HRNet-W24	6 x 12	*0.050*	0.000	*0.051*	*0.065*	0.036	0.166	0.298	*0.338*	*0.153*

The best performance estimations are italicized

absolute value of the 2DJLE. However, the absolute error does not directly affect the 3D joint position estimation. In other words, the 2D joint location relative accuracy to the scale of maps affects the 3D human pose estimation.

7 Conclusion and future work

We collected a new dataset that is composed of equirectangular images taken by a wearable omnidirectional camera and synchronized 3D joint positions. The setting of device selection and attachment position was determined according to intention to apply to sign language recognition in the future.

We validated the capability of the location-maps method to estimate 3D human poses for equirectangular images including distortion and disconnection, of which the characteristics of the location-maps were not mentioned in Mehta et al. [4]. Furthermore, we proposed a new 3D human pose estimation model using the location-maps method by replacing the backbone network with a state-of-the-art 2D human pose estimation model (HRNet). Our model is a simpler architecture than the reference model. Nevertheless, our model indicates better performance with the aspect of accuracy and computation complexity.

We introduced MPJPE and PCK metrics to evaluate 3D human pose estimation. We think that PCK is better metrics in the point of view that can indicate whether an estimation result fits to a sign language pose under a threshold in contrast with MPJPE indicting how far distance. In the future, we intend to apply this tracking system to sign language recognition although we mainly studied 3D human pose estimation in this paper.

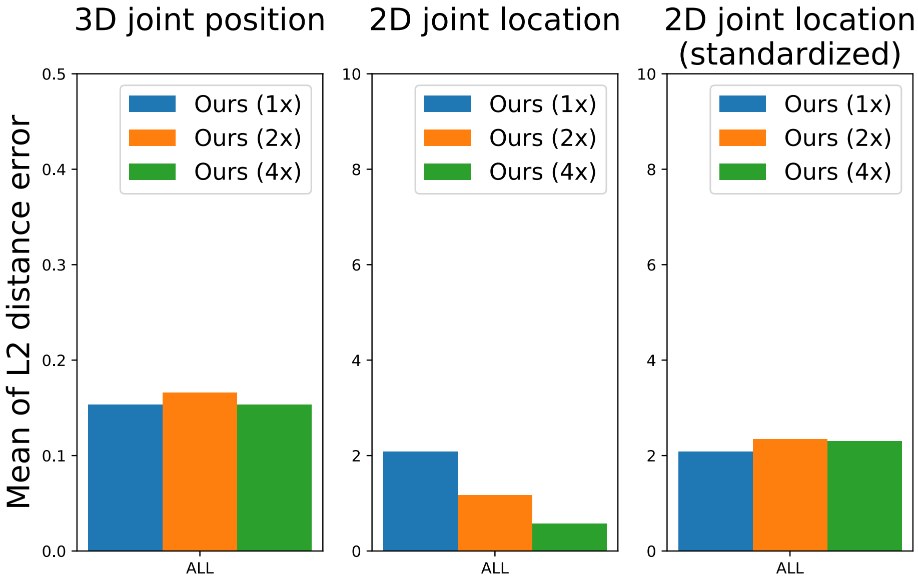

Fig. 12 The bar graphs indicate the mean L2 distance error for all joints on the test dataset-A. The 3D joint position errors are estimated by different scale models (left). The 2D joint location errors are estimated by only the heatmaps in each model, where a cell on the heatmaps is 1.0 (center). The 2D joint location errors are standardized considering the differences of each map scale (right)

In addition, we revealed the characteristics of the location-maps method. (1) The variance in heatmaps affects the robustness of the location-maps, and thus, enlargement of the variance leads to robustly extending 2D joint locations to 3D positions for the 2D estimation error by heatmaps. However, excessive enlargement causes to reduce the accuracy in 2D joint location estimation by heatmaps more than robustness of location-maps. (2) The shrinking scale of maps decreases the absolute value of the 2D joint location error; however, the reduction in absolute value does not directly affect 3D human pose estimation. The accuracy of 3D human pose estimation is related to the 2D joint location relative accuracy to the scale of maps.

To improve the model's performance, we introduce different variances for regression of heatmaps and location-maps respectively in the future work. Therefore, we set a distinctive variance for the heatmap mask H_j^{GT} in the location-maps loss formulation (3). The model generates more robust location-maps by setting a larger value to the location-maps' variance while keeping the accuracy in 2D joint location estimation by heatmaps.

Abbreviations

DNN: Deep neural network; CNN: Convolutional neural network; HRNet: High-resolution network; MPJPE: Mean per joint position error; PCK: Percentage of correct keypoints; 3DJPE: 3D joint position error; 2DJLE: 2D joint location error; std2DJLE: standardized 2D joint location error

Acknowledgements

Not applicable.

Authors' contributions

TM contributed to the proposing of the initial research idea, collected the dataset, conducted the experiments, and wrote the manuscript. SS supervised the work and advised the entire process of the research. All authors reviewed and approved the final manuscript.

Funding

This work was supported by JSPS KAKENHI Grant Number 17H06114.

Availability of data and materials

The dataset generated and/or analyzed during the current study are available under the license in the NIT-3DHP-OMNI repository, "https://drive.google.com/drive/folders/1XcQuS1dhSYggvb05CbQH-Qgd5OcBGyOe?usp=sharing".

Competing interests

The authors declare that they have no competing interests.

Received: 3 February 2020 Accepted: 12 August 2020
Published online: 31 August 2020

References

1. Pu J, Zhou W, Li H (2019) Iterative alignment network for continuous sign language recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, New York. pp 4160–4169
2. Koller O, Camgoz C, Ney H, Bowden R (2019) Weakly supervised learning with multi-stream CNN-LSTM-HMMs to discover sequential parallelism in sign language videos. IEEE Trans Pattern Anal Mach Intell 42:1–1
3. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2D pose estimation using part affinity fields. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, New York. pp 1302–1310
4. Mehta D, Sridhar S, Sotnychenko O, Rhodin H Shafiei M, Seidel HP, et al. (2017) VNect: real-time 3D human pose estimation with a single RGB camera. In: ACM Transactions on Graphics. vol. 36. ACM, New York, NY, USA. pp 1–14
5. Xu W, Chatterjee A, Zollhöfer M, Rhodin H, Fua P, Seidel HP, et al. (2019) Mo2Cap2: real-time mobile 3D motion capture with a cap-mounted fisheye camera. IEEE Trans Vis Comput Graph 25(5):2093–2101
6. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, New York. pp 5686–5696
7. Pons-Moll G, Fleet DJ, Rosenhahn B (2014) Posebits for monocular human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, New York. pp 2345–2352
8. Ionescu C, Carreira J, Sminchisescu C (2014) Iterated second-order label sensitive pooling for 3D human pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp 1661–1668
9. Tekin B, Rozantsev A, Lepetit V, Fua P (2016) Direct prediction of 3D body poses from motion compensated sequences. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, New York. pp 991–1000
10. Tekin B, Márquez-Neila P, Salzmann M, Fua P (2016) Fusing 2D uncertainty and 3D cues for monocular body pose estimation. ArXiv 1611.05708 abs/1611.05708
11. Du Y, Wong Y, Liu Y, Han F, Gui Y, Wang Z, et al. (2016) Marker-less 3D human motion capture with monocular image sequence and height-maps. In: European Conference on Computer Vision. Springer, New York. pp 20–36
12. Pavlakos G, Zhou X, Derpanis KG, Daniilidis K (2017) Coarse-to-fine volumetric prediction for single-image 3D human pose. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, New York. pp 1263–1272
13. Li S, Chan AB (2015) 3D human pose estimation from monocular images with deep convolutional neural network. In: Asian Conference on Computer Vision. Springer, New York. pp 332–347
14. Tekin B, Katircioglu I, Salzmann M, Lepetit V, Fua P (2016) Structured prediction of 3D human pose with deep neural networks. In: British Machine Vision Conference. BMVA Press, Durham. pp 130.1–130,11
15. Zhou X, Sun X, Zhang W, Liang S, Wei Y (2016) Deep kinematic pose regression. In: European Conference on Computer Vision. Springer, New York. pp 186–201
16. Agarwal A, Triggs B (2006) Recovering 3D human pose from monocular images. IEEE Trans Pattern Anal Mach Intell 28(1):44–58
17. Mori G, Malik J (2006) Recovering 3D human body configurations using shape contexts. IEEE Trans Pattern Anal Mach Intell 28(7):1052–1062
18. Rhodin H, Robertini N, Casas D, Richardt C, Seidel HP, Theobalt C (2016) General automatic human shape and motion capture using volumetric contour cues. In: European Conference on Computer Vision. vol. 9909. Springer, New York. pp 509–526
19. Mehta D, Rhodin H, Casas D, Fua P, Sotnychenko O, Xu W, et al. (2017) Monocular 3D human pose estimation in the wild using improved CNN supervision. In: International Conference on 3D Vision. IEEE, New York. pp 506–516
20. Chunyu W, Yizhou W, Zhouchen L, Alan LY, Wen G (2014) Robust estimation of 3D human poses from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, New York. pp 2369–2376
21. Edgar SS, Ariadna Q, Carme T, Francesc MN (2013) A joint model for 2D and 3D pose estimation from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, New York. pp 3634–3641
22. Edgar SS, Arnau R, Guillem A, Carme T, Francesc MN (2012) Single image 3D human pose estimation from noisy observations. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, New York. pp 2673–2680
23. Xiaowei Z, Menglong Z, Spyridon L, Kostas D (2017) Sparse representation for 3D shape estimation: a convex relaxation approach. IEEE Trans Pattern Anal Mach Intell 39(8):1648–1661
24. Mehta D, Sotnychenko O, Mueller F, Xu W, Sridhar S, Pons-Moll G, et al. (2018) Single-shot multi-person 3d pose estimation from monocular RGB. In: IEEE, New York. pp 120–130
25. Tompson J, Jain A, LeCun Y, Bregler C (2014) Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in Neural Information Processing Systems 27. Curran Associates, Inc. MIT Press, Cambridge. pp 1799–1807

26. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. Springer, New York. pp 483–499
27. Wei SE, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional pose machines. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE, New York. pp 4724–4732
28. Xiao B, Wu H, Wei Y (2018) Simple baselines for human pose estimation and tracking. In: European Conference on Computer Vision. Springer, New York. pp 472–487
29. Freeman TG (2002) Portraits of the Earth: A Mathematician Looks at Maps. American Mathematical Soc. https://doi.org/10.1111/j.1949-8535.2002.tb00041.x
30. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. Springer, New York. pp 770–778
31. Kingma D, Ba J (2014) Adam: a method for stochastic optimization. ArXiv 1412.6980 abs/1412.6980

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.