Environment Adaptive 3D Pose Estimation Model and Learning Strategy

Yeseung Park*, Kyoungoh Lee*, and Sanghoon Lee*†
* Department of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea
† Department of Radiology, College of Medicine, Yonsei University, Seoul, South Korea
E-mail: {pys940617, kasinamooth, slee}@yonsei.ac.kr, Tel: +82-2-2123-2767

***Abstract*—Recently, 3D pose estimation models using deep learning structures have begun to show outstanding performance. However, the performance is guaranteed only for the general pose included in public databases. In other words, most estimation models sometimes show degraded results when a given video contains uncommon poses from specific situations such as exercise and dance. This problem arises from the limitation of the pose diversity of public databases. We propose a novel estimation model calibration (EMC) framework for environment adaptive 3D pose estimation to solve this problem. This framework aims to calibrate well-trained existing pose estimation models from public databases to suit the environment. To achieve this goal, the framework uses target data to analyze the problems of existing estimation models. Subsequently, the proposed ergonomic model handler generates a calibration dataset by directly correcting the problem caused by the target data. Using the generated calibration dataset, we calibrate the existing pose estimation model. In this paper, we provide various experimental results of pose estimation for verification of the proposed framework. Experimental results demonstrate performance improvements qualitatively and quantitatively in specific poses and show the efficiency of estimation model calibration.**

I. Introduction

The 3D human pose is a non-verbal communication method that has recently played an important role in various fields(*e.g.*, Sports performance scoring, virtual avatar creation, human-computer interaction, situational awareness, etc.). virtual training [1], [2], [3]. Also, due to the development of a vast database and deep learning technology, a pose estimation model can easily estimate a 3D pose from an image without additional equipment [4], [5], [6]. However, the high accuracy that conventional estimation models show in public databases does not always guarantee performance in various environments. In other words, existing estimation models tend not to estimate pose that has not been seen in the database well(*i.e.,* uncommon pose). The reason is that the public database used for training does not contain all the poses of a human being. Therefore, it is difficult to apply the existing estimation model without change to various fields requiring uncommon pose estimation. For example, as shown in Fig. 1, in sports fields involving many high-legged poses such as taekwondo or figure skating, there are cases with frequent estimation failures. For this reason, there are still high barriers to the application of the technique, unlike the visible advances in pose estimation model research. Therefore, to apply the existing successful pose estimation model to various fields, a method for correcting a conventional model suitable for the application field is required.

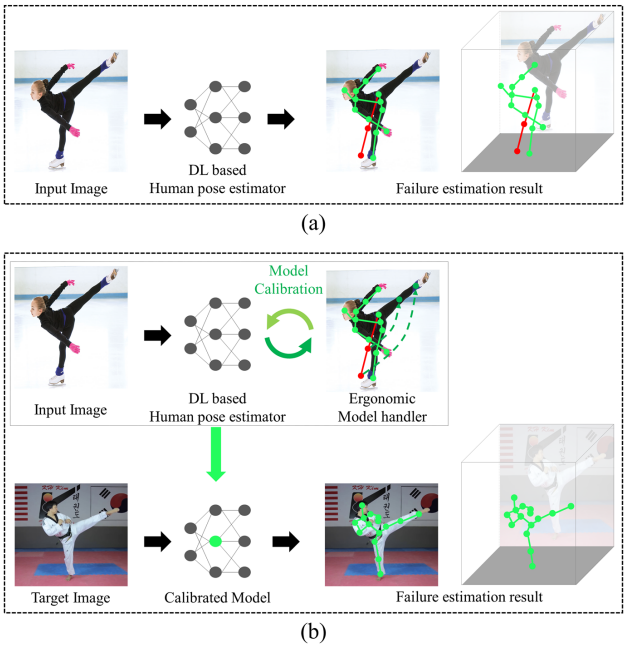


Fig. 1. Inference process of conventional model and Our proposed framework. (a) Inference example of the existing estimation model, (b) Example of the estimation result of our framework. (a) shows the inference process of the existing commercial attitude estimation model for an uncommon pose. (b) schematically shows our proposed Estimation Model Calibration Framework.

Most 3D pose estimation studies have adopted the two-step approach method. [4], [5], [6]. This two-step approach is an method for estimating a 3D pose by using the estimated 2D pose result obtained through 2D pose estimation models as a feature. Compared to a model that directly estimates a 3D pose from an image, this approach has recently become a big trend for 3D pose estimation because of its excellent performance and stability. However, despite improvement in performance and stability, existing approaches that rely on the 2D pose estimation model frequently shows large errors for difficult poses that are not in the data. To solve this problem, authors Zhou *et. al.* [7] and Rhodin *et. al.* [8] mainly used augmentation techniques of 3D pose data to alleviate the

bias of the database. Nevertheless, data augmentation cannot solve the problem without fundamental improvement of the 2D estimation model, which is the cause of the problem. The reason is that if the 2D pose estimation fails due to occlusion or a small amount of learning, it is very difficult to accurately estimate the 3D pose based on this. As shown in Fig. 1. (a), the uncommon pose estimation failure of the existing commercial pose estimation model shows poor results in both 2D and 3D pose estimation. As such, the accuracy of the two-dimensional pose estimation and the three-dimensional pose estimation are closely related to each other. Therefore, in order to improve the accuracy of 3D pose estimation in an inherently uncommon pose, direct improvement in the 2D pose estimation model is required.

To solve this problem, we propose the Estimation Model Calibration (EMC) Framework, a methodology that generates correct 2D keypoints for calibration model training. Estimation Model Calibration Framework consists of two important stages. The first stage is an Ergonomic Model Handler that receives an incorrect keypoint estimation through an existing commercial model as an input and corrects it. The second stage is a Model Calibration that corrects the commercial model to fit an uncommon field based on the calibration data. Ergonomic Model Handler is a user-friendly graphic user interface designed to process data required for model calibration easily and quickly. Model Calibration is a learning strategy that gradually changes the model so that an uncommon pose can be estimated using the refined 2D joint information generated through the ergonomic model handler. After these two stages, the 2D pose estimation model, which could not estimate uncommon poses, is gradually modified to fit the domain. The overall framework flow is shown in Fig. 1. (b). Meanwhile, for the model to learn properly, we determined the learning level that satisfies a small amount of learning and high estimation accuracy through various ablation tests related to this. And how the performance of the calibrated model affects the estimation of the 2D and 3D pose was evaluated in both qualitative and quantitative aspects.

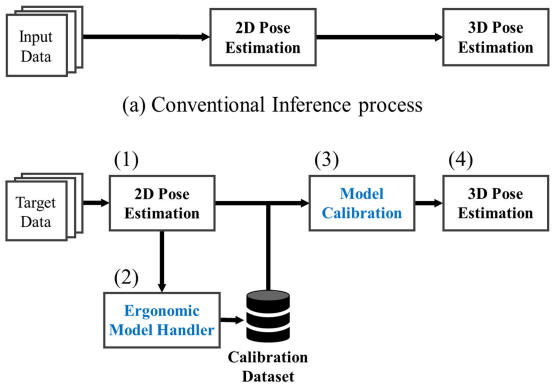
The contributions we propose in this paper are as follows.

- We proposed Estimation Model Calibration (EMC) Framework, a model calibration framework to apply a human pose to various fields.
- We proposed an Ergonomic model handler to easily and quickly fix noisy 2d joints.
- Through the proposed method, the calibrated model estimated a specific pose with a high level of accuracy qualitatively/quantitatively.

II. Main Method

A. Framework Overview

Most deep learning models [4], [5], [6] that estimate human pose from a given image proceed through the procedure shown in Fig. 2.(a). However, as mentioned above, the current model has difficulty in estimating a specific pose. Therefore, to solve this problem, we propose a fundamental correction method for



(a) Conventional Inference process

(b) Estimation Model Calibration Framework process

Fig. 2. Process of Estimation Model Calibration Framework

the pose estimation model. To solve this problem, this paper proposes Estimation Model Calibration(EMC) Framework. This is a methodology that extracts a complete 3D keypoint based on an uncommon pose image. The flowchart is shown in Fig. 2.(b). The process of the framework proceeds in the order of the numbers shown in the flowchart.

B. Target data

The uncommon pose estimation failure of existing deep learning models is due to lack of diversity of database. In the proposed framework, the target data aims to construct new pose data to solve the problem of the conventional estimation model. In other words, we newly construct images centered on uncommon poses that are not well detected in conventional 2D pose estimation models. In many studies, common data has been provided in large volumes,[9], but video data with special purposes is hard to obtain. Therefore, we extract images containing uncommon poses from youtube and center the person through image reconstruction. The uncommon pose data obtained through this process is represented as Fig. 1 (b). The target data composed of such uncommon pose detects the estimation problem of the existing model. And then improves the performance of the model according to the environment to which it is applied.

C. 2D pose estimation

In our proposed EMC Framework, 2D pose estimation is aimed at detecting estimation problems for the target environment of existing pose estimation models. Therefore, we obtain 2D poses on target data using existing learned 2D pose estimation models. In this case, we use detectron2 [10], the most commonly used public commercial model. The 2D pose estimation data estimated for a given target data are considered target 2D pose data sets. The target 2d pose dataset contains estimation error, which is paired with the target image. For model calibration, we visualize the direct problem with this estimated 2D pose dataset to the user via skeleton representations overlaid on images from the center console of

Fig. 3. Each person in a given target image is distinguished by a different color, and the joint of the skeleton is represented by a different marker depending on the confidence value of the estimation model. In this case, confidence is defined as the reliability of the estimated value of the pose estimation model. This indicates that joints with low confidence need modification. Therefore, when constructing data for calibration of a model, it is efficient to use joints with high values based on the confidence of the joints and to modify only those with low values. In this way, the user identifies intuitively represented problems of the existing 2D pose estimation model and corrects the model.

D. Ergonomic model handler

Ergonomic Model Handler directly provides users with estimation problem information of existing models through visualization and provides Ergonomic-based simple and easy relabeling capabilities for calibration of estimation models. Data augmentation is necessary for the diversity of learning in deep learning models, but methods such as simple flip do not play a direct role in solving the problem of the uncommon pose estimation. Therefore, we require human effort to correct the false results of the target data that fit the uncommon pose for accurate model calibration. To minimize the labor required, our proposed Ergonomic method has three characteristics:

- 1) Intuitively represent the problem by directly visualizing the results of the estimation model.
- 2) Provides an easy way to calibrate existing estimates in a short time through drag-and-drop methods.
- 3) Helps people judge quickly and intuitively by providing model occlusion information and confidence information about estimation results.

Fig. 3 gives a detailed picture of the proposed ergonomic method. Joint information generated by 2d pose estimation is visualized as shown in Fig. 3, allowing the user to intuitively grasp the joint information in which the problem occurred. The user selects and corrects the joint to be modified by the 'drag and drop' method, which saves time compared to directly inputting the joint position. Meanwhile, to reduce the time and effort of the user, the ergonomic model handler recommends the joint that needs to be modified based on the following two methods: **1. occlusion based alert, 2. confidence based alert**. The left side of Fig. 3 (b) is an example of occlusion-based alert, and the number of lost joints is displayed in the form of $(m, 0)$. Also, the figure on the right of Fig. 3 (b) is an example of confidence-based alert, and the number of joints with low confidence values is displayed in the form of $(0, n)$. Through these three ergonomic methods, human effort for GT data generation can be greatly reduced. Through this process, existing misestimated 2D joint information is reconstructed into Calibration Dataset to improve the model.

E. Model Calibration

In general, 2D pose estimation is learned by using the following equation as a loss. Let j_k^{GT} and j_k^{PRED} are the k^{th}

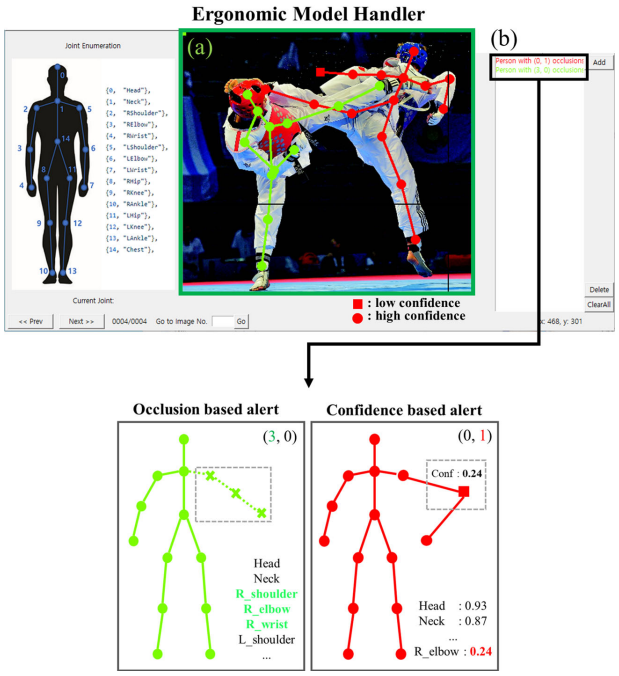


Fig. 3. Ergonomic Model Handler. (a) Joint information overlayed on target pose-centric images. (b) Occlusion/Confidence based alert method.

joint's ground-truth and the predicted 2D locations, respectively. Then, the 2D loss can be represented as

$$\mathcal{L}_{\mathrm{2D}} = \frac{1}{|K|} \sum^{k \in K} (j_k^{\mathrm{GT}} - j_k^{\mathrm{PRED}})^2, \tag{1}$$

where K and k are the set of joints and the joint index, respectively.

At this time, Model Calibration means a learning strategy of re-training the conventional 2D joint estimation model using this loss to estimate an uncommon pose. There are three training methods to calibrate the model; initial training, fine-tuning training, and model calibration. The initial training is the most basic learning method, and it is a method of newly learning the estimation model using only the target data. The fine-tuning training is a method of additionally learning only the target data to the baseline estimation model. Finally, the proposed model calibration is a method for additional learning using target data and common data together in the baseline estimation model. **The initial training method has the advantage of being the simplest to learn. However, due to the lack of information on the general poses, the model cannot estimate the general behavior well, so its versatility is very low.** On the other hand, in the case of the fine-tuning training and the proposed model calibration, the versatility and specificity of the model are high because an uncommon pose is estimated based on the general estimation ability of the existing model. Furthermore, since we use pre-trained models to find only additional data, the methodology has the advantage of fast learning convergence time. Meanwhile,

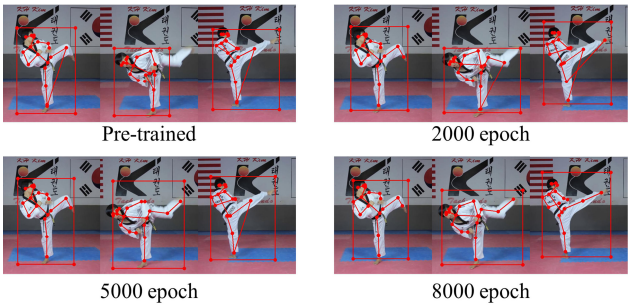


Fig. 4. 2d pose estimation result

TABLE I
POSE ESTIMATION ACCURACY (%) BY BODY PART

Method	Hip	Knee	Ankle	Leg	Total
Detectron2, baseline	95.5	94.1	93.0	94.2	96.1
Openpose, baseline	95.3	94.6	93.6	94.5	96.2
Ours, 2000 epochs	97.5	96.4	94.7	96.2	96.7
Ours, 5000 epochs	98.1	97.5	96.3	97.3	97.2
Ours, 8000 epochs	99.2	98.6	97.9	98.6	98.6
Ours, 10000 epochs	99.3	98.8	98.5	98.9	98.8

fine-tuning training proceeds with additional learning using only target data in the existing learning model. However, depending on the amount of learning, this method can cause an overfitting problem that estimates the pose of the target data well but cannot estimate other poses well. On the other hand, the calibration of the proposed model, which learns target data and general data in parallel, is relatively free from overfitting problems compared to fine-tuning training. Due to these advantages, this paper adopts the proposed model calibration, which learns data containing general and uncommon poses in parallel based on the existing model.

F. 3D Pose Estimation

3D Pose Estimation is the process of acquiring more precise 3D pose data using the calibrated 2D pose obtained through model calibration. Just by calibrating the 2D estimation model, the accuracy of both the 2D and 3D pose estimation models increases. We used facebook research's videopose3d model[11] as a 3d pose estimation model for inference. Also, we verified how the correction of the 2D model affects the 3D through experiments. The detailed experimental results are described in section 3.

III. EXPERIMENTAL RESULTS

A. Database

1) MS-COCO dataset: The Microsoft COCO Dataset (MS COCO) is an image dataset consisting of scenes with various objects and people in everyday situations. This dataset contains 328k images of 91 basic objects and humans. Due to this vast volume, this dataset is often used as a benchmark to compare different aspects of visual and human pose studies. We calibrate the model based on the composition of this dataset.

2) Calibration dataset: The data consists of 20 videos and a total of 3k frames. The generated data set is 3,000 images divided into 20 videos, and the ratio of target data to general data is 8:2 (2,400 target, 600 common). The common data included various everyday poses, not the poses of the target data. Meanwhile, to ensure diversity in learning, we designed the dataset to include various genders, clothes, and background colors as below.

- Gender : male 70%, female 30%
- Cloth type : uniformed 60%, casual 40%
- Background color : achromatic 60%, chromatic 40%

B. Efficiency of Ergonomic Model Handler

To measure the user convenience of the ergonomic model handler, we measured and compared the required time of the ergonomic model handler method with the existing method of directly checking and changing the position of the joint matched to the image.Ten subjects participated, And are six men ages range from 22 and 30 and four women ages range from 21 and 27.

The ten experimenters are divided into two groups. One group uses the classic joint modification method, and the other group uses the Ergonomic model handler. We asked two groups to process a total of 90 images (30 fps, 3 sec) with uncommon poses in each group's method and recorded the time taken for every ten frames. After the first experiment was over, the two groups changed their methodologies and conducted the second experiment. As shown in Fig. 5, our proposed ergonomic model handler took less time than the existing method for both groups. When the experimenters used the ergonomic model handler, the work efficiency increased about 3 times compared to the existing method. As the number of frames increased, we could observe a gradual increase in the interval. Considering that the amount of data required for additional training of the model is very large, the time efficiency of the experimenter can be maximized by using the ergonomic model handler.

C. Model Calibration Result

For the model to better estimate a specific motion, the 2D estimation model is calibrated through the 2D joint information. Such 2D joint information is reconstructed using the Ergonomic model handler. The most important point in the calibration of deep learning models is to ensure that the model learns certain behaviors to an appropriate level without overfitting. Additionally, we quantitatively/qualitatively evaluate the level of joint estimation for each corresponding branch to avoid overfitting the model to the additionally learned data. We selected Detectron2 as the baseline model and retrained 3k images with corrected joint information in batch size 2. Branching for learning is divided into baseline model (0 epoch), 2000 epoch, 5000 epoch, 8000 epoch, and 10000 epoch. The joint estimation accuracy was measured at

TABLE II
A SERIES OF AVERAGE PRECISIONS (APS) BASED ON OKS

Method	AP @0.5:0.95	AP @0.5	AP @0.75	AP medium	AP large
Openpose	63.2	85.5	67.7	58.2	68.7
Detectron2	67.1	88.1	73.0	62.4	75.7
Ours, 2000 epochs	67.8	88.8	73.6	62.9	76.1
Ours, 5000 epochs	68.5	89.3	73.9	63.3	76.6
Ours, 8000 epochs	69.1	89.7	74.3	63.7	77.0
Ours, 10000 epochs	68.7	89.3	74.1	63.4	76.7

each step.

We used the PCKh metric [12], which is the most used in the existing 2D pose estimation, to measure the accuracy of the calibrated model. At this time, we set the threshold α of the metric to 0.5. To check whether the modified model accurately estimates the target pose, kick pose, we expressed a subset consisting of the hip joint, knee joint, and ankle joint as leg joints and compared them with the overall PCKh value. As shown in Table. II-F, PCKh of baseline model of Detectron2 and Openpose model is lower than PCKh of total joint, PCKh of leg joint is lower. Meanwhile, as the learning continues, the PCKh of the leg joint gradually increases, suggesting that the revised data make the baseline model familiar to a specific pose. The PCKh score of the total joint and the PCKh score of the leg joint are similar in 5000 epoch and 8000 epoch models. However, in the case of 10000 epoch model, total PCKh is lower than leg PCKh. From this, we can infer that the model is learning the pose excessively. Meanwhile, to qualitatively evaluate the calibrated model, we output the results of each learning branch based on the same test video. This result can be confirmed in Fig. 4. As described above, the baseline model shows weak results for kicks in various poses. This tendency gradually disappears as the number of learning increases, and at 8000 epochs, the estimation is robust for all uncommon poses.

Meanwhile, to check whether the model is overfitted, we evaluated the MSCOCO database as shown in Table. III-C. MSCOCO uses Average precision(AP) based on OKS[9] as a metric to evaluate the accuracy of the model. The more successfully the model predicts the keypoint, the closer the OKS value is to 1.

As in the previous experimental results, our calibration model shows improved performance than the Detectron2 baseline model. However, in the evaluation test set, including many general poses, the 10000 epoch model shows a lower score than the 8000 epoch model. These results show that the 10000 epoch model is overfitted to the uncommon pose. Through such qualitative/quantitative evaluation, it is possible to determine the amount of learning required for model calibration.

D. Result of 3D Pose Estimation

To confirm the effect of the performance improvement of the 2D estimation model on the 3D estimation model, we checked

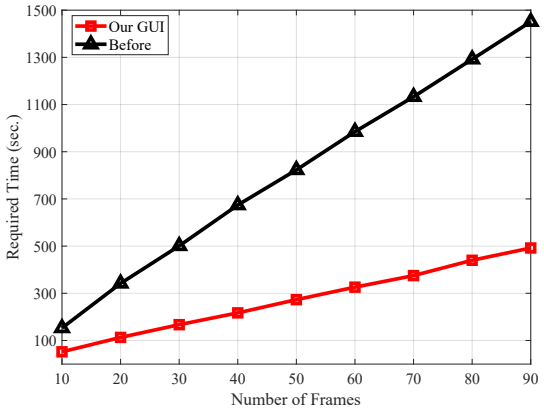


Fig. 5. Processing time of two joint modification methods

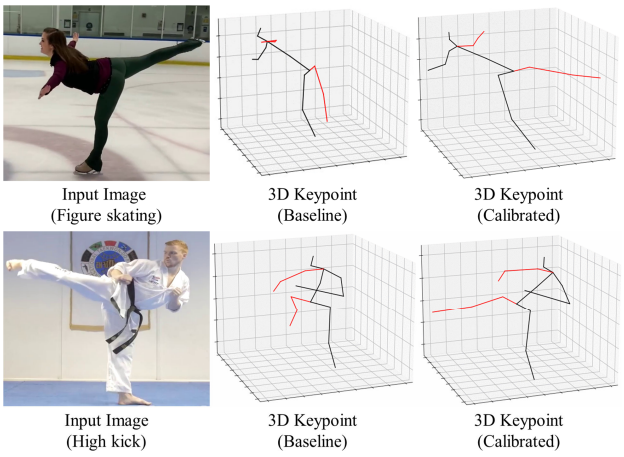


Fig. 6. Result of 3D baseline model(middle) and calibrated model(right) according to the input image

both evaluation protocols of Human 3.6M based on the improved 2D result. Our model showed a performance of 46.1 mm (Protocol 1) and 35.9 mm (Protocol 2) in two evaluations. This result shows an improvement over the existing baseline models of 46.8 mm (Protocol 1) and 36.5 mm (Protocol 2). In other words, the improvement of the 2D estimation result, the basis of the estimation, has a positive effect on the performance of the 3D estimation model. Qualitative results for pose not estimated by existing methods are shown in Fig. 6. The two rows of Fig. 6 are images and estimated results of figure skating and taekwondo. The middle column is the inference result using the 2D baseline model as an input, and the right column is the inference result using the corrected 2D model as an input. Even with the same input, the right 3D skeleton using the corrected model estimated the pose information of the image more accurately than the result of the baseline model.

IV. CONCLUSIONS AND FUTURE WORK

We found out that existing pose estimation models did not well estimate the unusual pose, and we confirmed that this

problem was caused by the lack of diversity in the database. To solve this problem, we proposed the Estimation Model Calibration Framework, which calibrates the existing model according to the target pose. Then we designed the calibrated model to show higher estimation accuracy for an uncommon pose than the existing model. Also, a user-recognition-oriented Ergonomic model handler was implemented so that the user can easily and quickly correct the wrongly estimated result. We experimentally confirmed that the ergonomic model handler is more effective than the classical method of directly modifying keypoints. Moreover, we confirmed that the calibrated model based on this successfully estimated an uncommon pose more qualitatively/quantitatively than the conventional results in the field of 2D and 3D pose estimation. Still, there are many difficulties in manually calibrating the model for a wide variety of special poses. Therefore, we will continue to designing an end-to-end system in which the EMC Framework analyzes vulnerable poses and automatically corrects them.

ACKNOWLEDGMENT

This work was supported by Institute for Information and communications Technology Promotion through the Korea Government (MSIP) (No. 2016-0-00204, Development of mobile GPU hardware for photo-realistic real-time virtual reality)

REFERENCES

[1] B. Kwon, J. Kim, K. Lee, Y. K. Lee, S. Park, and S. Lee, "Implementation of a virtual training simulator based on 360° multi-view human action recognition," *IEEE Access*, vol. 5, pp. 12 496–12 511, 2017.

[2] H. Joo, N. Neverova, and A. Vedaldi, "Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation," *arXiv preprint arXiv:2004.03686*, 2020.

[3] G. Abt, C. Boreham, G. Davison, R. Jackson, A. Nevill, E. Wallace, and M. Williams, "Power, precision, and sample size estimation in sport and exercise science research," 2020.

[4] K. Lee, I. Lee, and S. Lee, "Propagating lstm: 3d pose estimation based on joint interdependency," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 119–135.

[5] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann, "Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2272–2281.

[6] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D. N. Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3425–3435.

[7] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3d human pose estimation in the wild: a weakly-supervised approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 398–407.

[8] H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation for 3d human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 750–767.

[9] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 mscoco image captioning challenge," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, pp. 652–663, 2016.

[10] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[11] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7753–7762.

[12] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.