Electronic Research Archive

ERA, 31(3):1556–1569.
DOI: 10.3934/era.2023079
Received: 28 October 2022
Revised: 23 December 2022
Accepted: 29 December 2022
Published: 18 January 2023

http://www.aimspress.com/journal/ERA

Research article

Calibrated deep attention model for 3D pose estimation in the wild

Longkui Jiang[1], Yuru Wang[2,*] and Xinhe Ji[2]

[1] Technology School, Jilin Business and Technology College, Changchun, China
[2] School of Information Science and Technology, Northeast Normal University, Changchun, China

* **Correspondence:** Email: wangyr915@nenu.edu.cn; Tel: 08613604407105.

Abstract: Three-dimensional human pose estimation is a key technology in many computer vision tasks. Regressing a 3D pose from 2D images is a challenging task, especially for applications in natural scenes. Recovering the 3D pose from a monocular image is an ill-posed problem itself; moreover, most of the existing datasets have been captured in a laboratory environment, which means that the model trained by them cannot generalize well to in-the-wild data. In this work, we improve the 3D pose estimation performance by introducing the attention mechanism and a calibration network. The attention model will capture the channel-wise dependence, so as to enhance the depth analysis ability of the model. The multi-scale pose calibration network adaptively learns body structure and motion characteristics, and will therefore rectify the estimation results. We tested our model on the Human 3.6M dataset for quantitive evaluation, and the experimental results show the proposed methods with higher accuracy. In order to test the generalization capability for in-the-wild applications, we also report the qualitative results on the natural scene Leeds Sports Pose dataset; the visualization results show that the estimated results are more reasonable than the baseline model.

Keywords: computer vision; 3D human pose estimation; attention; 3D human pose calibration

1. Introduction

Taking a 2D monocular image as the input, the task of 3D pose estimation is to regress the human body joints' coordinates from a 2D image. The 3D body pose is important data for human-computer interaction, animation, virtual reality and video surveillance. For example, in the field of virtual reality, in order to realize the interaction between the user and the virtual environment, the device needs to capture the human image as input and estimate the 3D pose. Usually, the user is in a natural scene, so

a system to estimate the 3D pose for in-the-wild applications is a key technology. However, in-the-wild 3D pose estimation is challenged by the depth ambiguity associated with recovering the 3D pose from monocular 2D images. Monocular images lose depth information, which makes the transformation from 2D to 3D highly nonlinear. It becomes even more challenging for in-the-wild applications due to the lack of large-scale 3D human pose datasets. Most of the existing human pose datasets in natural scenes are captured as 2D images/videos; furthermore, the commonly used 3D human pose datasets, such as Human3.6M, were obtained in laboratory scenes. Therefore, the models trained on them usually show poor generalization ability and cannot be well applied to natural scenes with a complex and diverse background.

For the in-the-wild 3D pose estimation problem, a popular technology is the two-stage model [1–3]. The first stage performs 2D human pose estimation with a natural scene image as input, and it trains the model by using the 2D in-the-wild pose dataset. The second stage takes the output 2D pose and the original image as input and outputs the 3D human pose estimation result. This method succeeds in the 3D human pose estimation problem, but it still cannot generalize very well to in-the-wild images. The reasons lie in two aspects: 1) At the second stage, due to the 3D human pose dataset limitation, the model is trained on the dataset captured in the laboratory scenes. 2) There is no connection between the two sub-models, and they are forced to be trained separately so that they cannot share each other's feature information. This kind of 3D pose estimation methods show promising performance for single-person 3D pose estimation [4–6], but they rely considerably on human detection to localize each person prior to estimating the joints within the detected bounding boxes.

The famous Generative Adversarial Networks (GAN) [7] has also been used to solve the 3D in-the-wild human pose estimation problem. Yang et al. [8] used a generator to predict 3D pose samples and designed a multi-source discriminator to judge whether the samples are generated by the generator or from the original dataset. The generator and discriminator play against each other so that the generator can generate prediction results that are very close to the real samples. Wandt and Rosenhahn [9] built a module for predicting camera parameters on the basis of a GAN to predict 3D poses. They made use of the camera parameters to reproject the predicted 3D coordinates into 2D coordinates, and they updated the model parameters by performing backpropagation based on the loss between the estimated 2D coordinates and the ground truth. This model alleviates the problem of a lack of 3D human pose datasets for large-scale natural scenes, making the model more robust and generalizable. In addition, the popular self-supervised learning has also been introduced to the 3D pose estimation field. Wandt et al. [10] introduced the self-supervision idea to the CanonPose to regress the 3D pose from multi-view data; the model shows ability to alleviate the problem of dataset limitation for in-the-wild applications.

Because the body structure is significant for 3D pose estimation, many works have tried to reduce the estimation ambiguity by using a pictorial structure model (PSM) or graph neural network (GNN). PSMs had been widely used in both 2D and 3D pose estimation problems before deep networks began dominating the field. PSMs determine the optimal joint coordinates by simultaneously considering joints' appearance and spatial relations [11]. In a deep network model, PSMs are usually used to perform inference on the features extracted by deep models [12]. As compared with PSMs, GNN-based models [13,14] are powered by end-to-end learning.

Zhou et al. [15] split the 3D human pose estimation model into a 2D human pose estimation sub-model and a depth regression sub-model. Different from the two-stage method, the proposed two-stage cascaded architecture develops the connection between the two sub-models, and they share deep image

feature information and global semantic information. Since only the 2D in-the-wild human pose datasets are available, this method proposed a weakly supervised transfer learning model. It introduced a 3D geometric constraint based on the length ratio of human bones as the loss function. Therefore, this model can generalize well to in-the-wild scenes, and it achieved good performance on public datasets.

Weakly supervised models constitute a nice attempt at resolving the 3D pose regression problem with limited datasets, but the accuracy of 3D in-the-wild pose estimation is still not ideal. The performance is still associated with the difficulty of accurately regressing 3D coordinates from 2D images. In order to improve the model performance, the key is to improve the depth analysis ability. Attention [16,17] is a good idea to reach this goal. Chu et al. [18] introduced a multi-context attention mechanism to Convolution Neural Networks (CNN) architecture and realized an end-to-end model. In this model, hourglass networks are used to generate an attention map. Su et al. [19] imported the attention mechanism to capture the channel-wise information. In our work, we tried to import attention to the weakly supervised model.

Besides exploring the depth information, general knowledge is another way to resolve the depth ambiguity, because it is independent of data [20,21]. There is some work attempting to import geometry constraints into deep networks, such as the limb length [22] and the translation and rotation constraints [23]. These kinds of geometry knowledge are usually defined in the loss function to constrain the prediction to be congruent with body kinematics. Aiming to further explore kinematics knowledge, we propose the use of a network as constraints.

Taking the above model [15] as a baseline, this paper focuses on how to improve the accuracy of pose estimation for in-the-wild problems; the main contributions are as follows:

1) Introduce the attention mechanism to obtain depth information from monocular images so as to regress 3D information more accurately.

2) Propose the calibration of the estimation results by a multi-scale calibration network.

3) Experiments show the proposed model to have better estimation accuracy and generalization ability for in-the-wild 3D pose estimation problems.

2. 3D human pose estimation model based on attention mechanism

2.1. Baseline model

The baseline model [15] builds a 2D human pose estimation sub-model based on the stacked hourglass networks [24], and it is composed of multiple residual modules [25]. The residual module retains the original features while extracting the high-level features, and it can alleviate the gradient vanishing problem when the model becomes deeper. The residual module structure is shown in Figure 1, where "channel" represents the number of channels of the input image or feature map and k is the filter size. The residual module has two-path tasks. The first path is the convolution path. It is composed of three convolutional layers (white) with different kernel sizes. Between them is a batch normalization layer (blue) [26] and an activation function layer (orange) [27], which is used to extract the higher-level features for each channel in the 2D space. The second path is a skipping path, and it contains one convolutional layer with a kernel size of 1, which is used to retain the features of the original level. If the number of channels between the input and output of the residual module is inconsistent, the skipping path will adjust the number of input channels to the output channel number. At the end of the

module, the output results of the two paths will be combined as the final output. The execution flow of the residual module in the stacked hourglass network is shown in Figure 2.

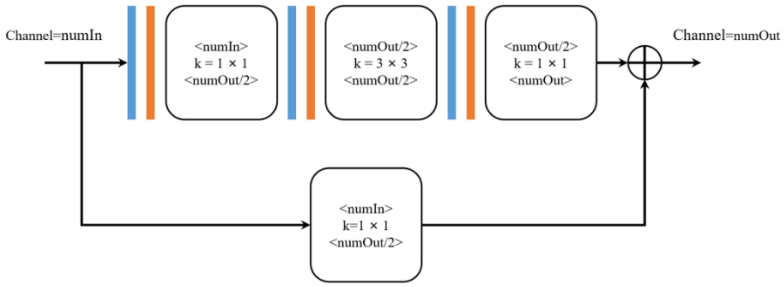

Figure 1. Structure of the residual module in the baseline model.

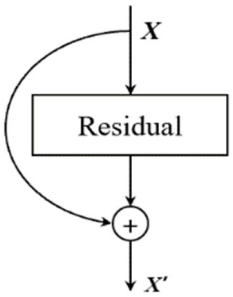

Figure 2. Execution process for the residual module in stacked hourglass networks.

2.2. Efficient channel attention module

For the 2D human pose estimation sub-model, the stacked hourglass network extracts each channel's features in their respective 2D space and does not capture their interdependence. In the following depth regression sub-model, it takes the output 2D heat maps and 2D pose as input to regress the depth coordinate; as a result, the depth information is hard extract well at the depth regression stage. In order to resolve this problem, we employed the efficient channel attention (ECA) module [28] as shown in Figure 3. This module uses the attention-based channel weight learning algorithm to improve the residual module by incorporating the ability to capture each channel's importance and analyze depth information. The details of the ECA module are as below.

The input to the module is a feature image of size $W \times H \times C$, where W, H and C are the width, height and channels of a single feature image, respectively. This module learns channel weights through three steps of operations:

1) Squeeze operation. Convert a 2D feature map into C real numbers by using the squeeze operation, and combine them into a real number list of size $1 \times 1 \times C$. Each real number has a global receptive field corresponding to the feature map to a certain extent. The squeeze operation employs global average pooling, as follows:

$$g(X) = \frac{1}{W \times H} \sum_{i=1,j=1}^{W,H} X_{ij} \tag{1}$$

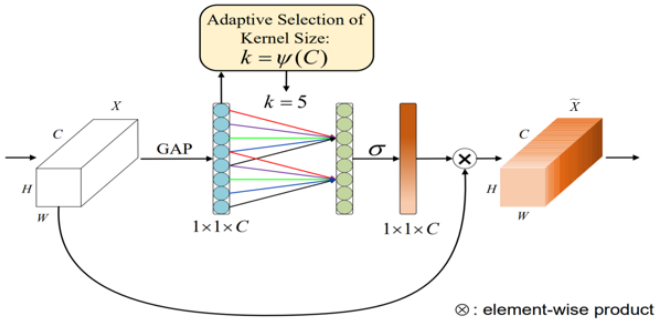

Figure 3. Structure of the ECA module.

2) Excitation operation. A band matrix W_k is used to capture the local cross-channel interaction, that is, only the interaction between each channel and its adjacent channels is considered so as to ensure the efficiency and performance. A real number list ω of size $1 \times 1 \times C$ is learned to generate weights for each feature channel and explicitly model the dependencies between feature channels. W_k is

$$\begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & w^{C,C-k+1} & \cdots & w^{C,C} \end{bmatrix} \tag{2}$$

Let $y = g(X)$ be the real number list, and the excitation operation is implemented as

$$w_i = \sigma\left(\sum_{j=1}^{k} w_i^j y_i^j\right), y_i^j \in \Omega_i^k \tag{3}$$

where Ω_i^k is the k adjacent channel set of y_i. In order to reduce the model complexity, all channels will share the same weights:

$$w_i = \sigma\left(\sum_{j=1}^{k} w_i^j y_i^j\right), y_i^j \in \Omega_i^k \tag{4}$$

This can be implemented by the fast 1D convolution with a kernel size of k, as follows:

$$w = \sigma\left(Conv1D_k(y)\right) \tag{5}$$

where σ is the sigmoid activation function, $Conv1D$ represents a 1D convolutional operation and the kernel size k is determined by the channel number C in an adaptive way, as follows:

$$k = \Psi(C) = \left|\frac{log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{odd} \tag{6}$$

where $|t|_{odd}$ is the odd number closest to t. The parameters γ and b are set to be 2 and 1, respectively. In this adaptive mapping function, the higher the channel dimension, the stronger the long-term interaction; and, the lower the channel dimension, the stronger the short-term interaction. In Eq (5), the weights of each channel are the final real number list.

3) Weighting operation on each channel of the input feature map X. As shown in Eq (7), the real number sequence w is the attention weight set of each feature channel, and the matrix dot product is used to weight each channel on the feature map.

$$X = X \cdot \omega \tag{7}$$

This operation performs the channel weighting operation based on the attention mechanism and weakens the features that are not very effective for the human pose estimation task; it also enhances the useful ones for both the 2D human pose estimation sub-model and depth regression sub-model.

2.3. Improved residual module

For 3D pose estimation, the extracted feature should include not only the 2D spatial information, but also the depth feature. However, the residual module in the baseline stacked hourglass network only extracts the spatial feature in each channel. As a result, it is insufficient in extracting depth information when regressing the 3D pose. As shown in Figure 4, we have improved the residual module (as shown in Figure 2) by incorporating the attention-based ECA module (as shown in Figure 3). The attention module learns the dependencies between the channels and enhances the depth regression ability of the 2D human pose estimation sub-model; therefore, it is expected to improve the estimation accuracy and robustness of the overall model.

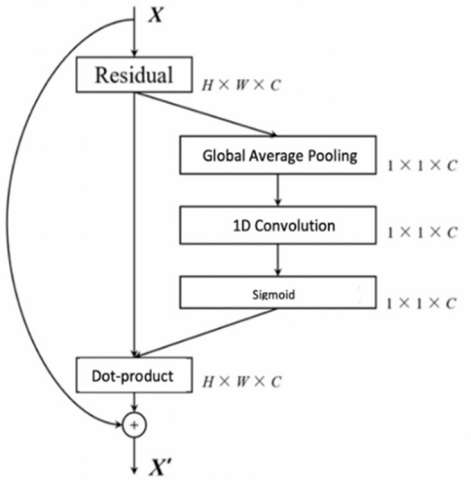

Figure 4. ECA-based execution process of the residual module.

3. Multi-scale 3D human pose calibration network

It is of great significance to use the 3D human pose calibration network to calibrate the results of the end-to-end benchmark model estimation. There are two main reasons for this. First, the model itself is inaccurate in recovering the 3D human pose from the monocular image. Second, the process of extracting joint coordinates from the model through the heat map is a quantitative process, and this process has certain inherent errors. The 3D human pose calibration network can not only calibrate the inaccurate joint coordinates of the model estimation, but it can also correct the error inherent to the process of thermal image quantization into joint coordinates.

3.1. Multi-scale human skeleton model

According to the knowledge of human physiology, the motion of each joint of the human body is relatively independent. For example, when the left wrist maintains a specified posture, the right wrist can still move freely without the restriction of the left wrist. However, the same set of joints of the human body are often related to each other. For example, the movement of the left wrist often drives the left elbow to move together. Therefore, this paper proposes a multi-scale human skeleton model based on the human joints and human structure at different scales, as shown in Figure 5.

The single-joint scale human skeleton model takes 16 single-joint points as the basic unit, the double-joint-scale human skeleton model takes the adjacent two joint points shown in Figure 5 as the basic unit and the four-joint-scale human skeleton model takes the four joint points as the basic units.

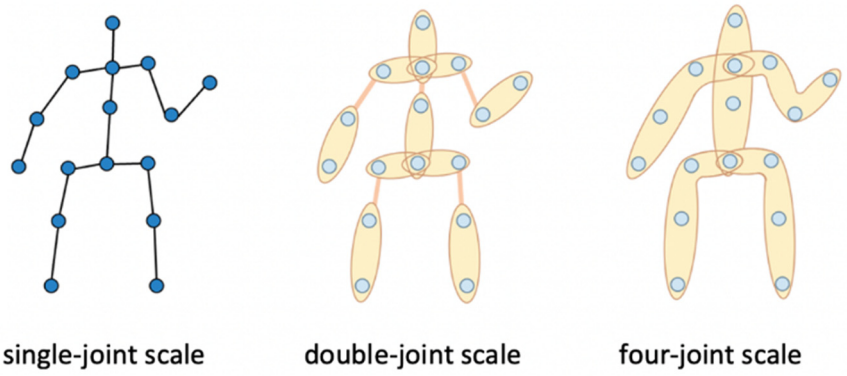

Figure 5. Multi-scale human skeleton model.

3.2. Multi-scale 3D human pose calibration network design

The above model gets the joint coordinates from the estimated heat map; it is a quantitative process, and there will be some inherent errors. Moreover, recovering the 3D pose from a monocular 2D image is inaccurate. Therefore, the predicted 3D pose output by the model is sometimes inconsistent with the human body kinematics. The movement of the human body will follow human dynamics. For example, in order to maintain balance during running, the human body will habitually swing the right arm forward while stepping out of the left leg. Therefore, we use a calibration network to reduce the inherent error caused by the quantification process of the joint's heat map and make the estimated 3D pose reasonable. In order to fully represent the joints and structures at different scales, this paper proposes a multi-scale human skeleton model according to the characteristics of human motion, as shown in the left part of Figure 6.

The human skeleton model used in this work is a kind of chain-like structure, and we employed the bidirectional recurrent neural network (BRNN) [29] to represent it. The BRNN is able to capture both temporal-dependent sequential data and chain-like structural data. Moreover, it also propagates the data information in both forward and backward modes, and this will solve the problem that the joint data input first cannot be sensed by the joint data input later. In this paper, combined with the multi-scale human skeleton model, a BRNN is used to implicitly construct a gesture semantic model. This model adaptively learns the structural and motion characteristics of the human body, according

to the human body joints and structures at different scales. This means that this model has the ability to effectively calibrate the 3D human pose.

The model takes the 3D human pose estimation results as input and groups them according to the multi-scale human skeleton model; it then employs a bidirectional gated recurrent unit (BiGRU) to construct a multi-scale 3D human pose calibration network. The structure of the network is shown in Figure 6. The BiGRU is a BRNN. Unlike the traditional BRNN, it can solve the long-term dependency problem between network nodes while maintaining high computational efficiency, as well as weaken the relationship between joints with less correlation. It is sensitive to the joints and structures with strong correlations.

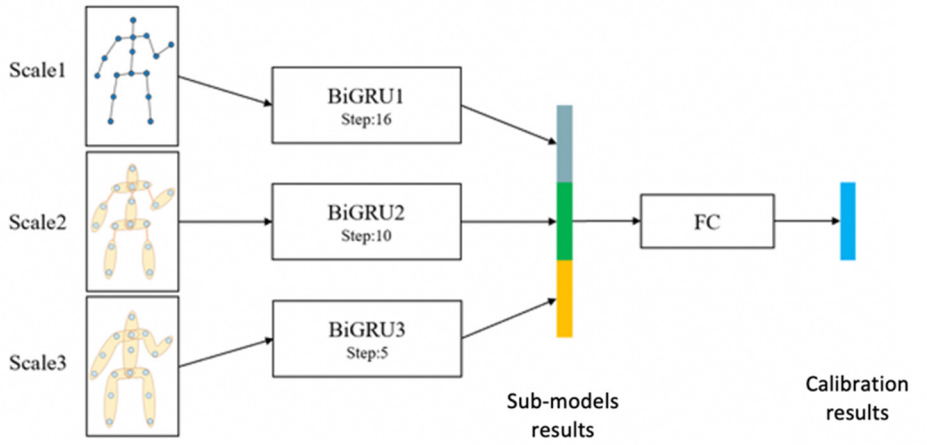

Figure 6. Structure of multi-scale 3D human pose calibration network.

As shown in Figure 6, this network includes three BiGRU sub-modules based on human skeleton models of different scales, and the hidden layer dimension (hidden_size) of each BiGRU sub-module was set to be 128. The BiGRU1 sub-module takes the single-joint-scale human skeleton model data as input, and the step size is 16; the BiGRU2 sub-module takes the double-joint-scale human skeleton model data as input, and the step size is 10; the BiGRU3 sub-module takes the four-joint-scale human skeleton model data as the input, and the step size is 5. In the calibration network, the output tensors of the three BiGRU sub-modules are concatenated and reduced to 48 dimensions (16 × 3) through a fully connected layer (FC) to obtain the 3D human joint coordinates. To train the model, we use the mean square error as the loss function.

4. Experiments and results

This section introduces the evaluation metrics for the 3D human pose estimation and the datasets used in the experiments; it then reports the experimental results tested on different public datasets.

4.1. Datasets and metrics

Three public datasets were used in this study: Human3.6M [30], Max Planck Institut Informatik (MPII) [31], Leeds Sports Pose dataset (LSP) [32]. Human3.6M is currently the most widely used 3D human pose dataset in laboratory scenarios. It contains about 3.6 million images, including 11

professional actors and 15 different daily movements. MPII and LSP are 2D datasets captured in natural scenes, where the MPII dataset provides about 25,000 images, including rich human poses and diverse complex environmental backgrounds. The LSP dataset provides 2,000 human motion pose images, which are often used to measure the generalization ability of the algorithm due to its small amount of data. In our experiments, the Human3.6M and MPII datasets were used to train the model, and the Human3.6M test set and LSP dataset were used to estimate the accuracy and generalization performance of the model, respectively.

In order to evaluate the model accuracy, Protocol 1 of Human3.6M was employed. According to this protocol, the model was trained on the 1st, 5th, 6th, 7th and 8th actors' data, and it was tested on the 9th and 11th actors. The mean per joint position error (MPJPE, with unit mm) for all estimated joint's coordinates was compared with the ground truth as shown in Eq (8), where N_s is the joint number and $m_{f,s}^{(f)}(i)$ is the estimation results for a certain joint.

$$E_{MPJPE}(f,s) = \frac{1}{N}\sum_{i=1}^{N_s} \left\| m_{f,s}^{(f)}(i) - m_{gt,s}^{(f)}(i) \right\|_2 \tag{8}$$

4.2. Experimental results of 3D human pose estimation model based on attention mechanism and pose calibration

We compared the proposed model with the state-of-the-art models on the Human3.6M test set, as shown in Table 1. The actions of 15 people were tested.

Table 1. MPJPE of each model on the Human3.6M dataset. The bold font represents the smallest error, and italics represents the second smallest error.

Protocol #1 (mm)	Dir.	Dis.	Eat	Greet	Phone	Photo	Pose.	Purch
Chen and Ramanan [2]	88.87	97.57	89.98	107.3	107.3	139.1	93.56	136.0
Tome et al. [33]	64.98	73.47	76.82	86.43	86.28	110.6	68.93	74.79
Mehta et al. [3]	59.69	69.74	60.55	68.77	76.36	85.42	59.05	75.04
Zhou et al. [15]	54.82	60.7	58.22	71.41	62.03	65.53	*53.83*	*55.58*
Martinez et al. [34]	*51.8*	**56.2**	*58.1*	**59.0**	69.5	78.4	55.2	58.1
Drover et al. [35]	58.4	59.4	58.7	*64.5*	**59.0**	*64.7*	60.9	57.0
Ours	**50.63**	*56.71*	**57.31**	66.15	*61.58*	**63.92**	**50.06**	**53.63**
Protocol #1 (mm)	Sit.	S.D.	Smoke	Wait	WalkD	Walk.	WalkT	Avg.
Chen and Ramanan [2]	133.1	240.1	106.6	106.2	87.03	114.0	90.55	114.1
Tome et al. [33]	110.1	172.9	84.95	85.78	86.26	71.36	73.14	88.39
Mehta et al. [3]	96.19	122.9	70.82	68.45	54.41	82.03	59.79	71.14
Zhou et al. [15]	75.2	111.59	64.15	66.05	**51.43**	63.22	55.33	64.9
Martinez et al. [34]	74.0	*94.6*	62.3	*59.1*	65.1	**49.5**	**52.4**	62.9
Drover et al. [35]	**61.6**	**85.8**	*60.4*	**57.4**	65.5	63.0	62.1	*62.3*
Ours	*71.88*	107.0	**59.60**	61.31	*53.05*	*58.09*	*53.29*	**61.62**

It can be seen from the comparison results in Table 1 that the average joint error (Avg. column) of the proposed model was lower than all similar models in the table, and that six of the 15 action categories showed the best accuracy. Compared with the baseline model [15], our model had better performance in almost all action classifications, indicating that our algorithm can make up for the

shortcomings of the benchmark model and greatly reduce the model's performance. Our method showed good performance for the actions "Dir.", "Eat.", "Photo", "Pose.", "Purch" and "Smoke", but it performed bad for the actions "Greet", "Sit.", "S.D." and "Wait". We analyzed the dataset for each action, and we found the reason for this to be that our model tends to be good at predicting the pose with small-scale movement. But, when the actor moves heavily, our prediction was not that good. Our model takes a single monocular 2D image as input, but it did not capture the temporal feature of the pose. Therefore, its performance when dealing with large-scale movement will be worse than the temporal model.

4.3. Ablation study

In order to better analyze the attention model and calibration network, we carried out cumulative comparison experiments; the experimental results are shown in Table 2.

Table 2. Cumulative contrast experiment.

Methods	MPJPE (mm)
Baseline	64.90
+ECA w/o calibration	63.84
+ECA +calibration_1	62.30
+ECA +calibration_all	61.62

In Table 2, "Baseline" is the baseline model, "+ECA" is the model adding the channel weight learning algorithm based on the attention mechanism and "w/o calibration" is the model without the multi-scale 3D pose calibration network. Relative to the baseline model, the error of this model was reduced to 63.84 mm, indicating that the channel-wise attention module enhances the model's ability to analyze depth information.

We also conducted two sets of comparative experiments on the proposed 3D human pose calibration network to test the effectiveness of the multi-scale 3D human pose calibration network. As shown in Table 2, for the 3D human pose calibration network, "calibration_1" indicates that only the single-joint-scale skeleton model was used as the input; "calibration_all" indicates that the entire multi-scale 3D pose calibration network was used for pose calibration. We can see from the results that, even if only the single-joint-scale human skeleton model is used as input, the calibration network can achieve performance improvement. When employing the multi-scale calibration method, the model showed an even higher estimation accuracy. The multi-scale 3D pose calibration network significantly improved the model performance.

4.4. Qualitative results

In addition to the above quantitative experimental results, we also visualized the estimation results in Figure 7. The results were tested on the 2D human pose LSP dataset with natural scenes. This dataset was unknown to the trained model, and the backgrounds are complex and diverse; moreover, it includes postures of various human movements, such as kicking a ball, horseback riding, running, etc., so it reflects the generalization ability of the model in natural scenes. In the experiments, this model

was compared with the baseline model. Following comparison, our model showed a more accurate 3D pose, and the visualization better fit to the human body structures. As shown in the 3rd row of Figure 7, the left leg was longer than the right in the baseline results, which is inconsistent with human body geometry. Our model is capable of adaptively learning the structure and motion of the human body; therefore, it achieved better performance. In the 1st row, the estimated 3D pose was more accurate when looking at the legs. In the right part of the last row, the output 3D pose was more accurate for the right leg. But, our model also had some failure cases; for example, in the right part of the 2nd row, the estimate pose is unreasonable.

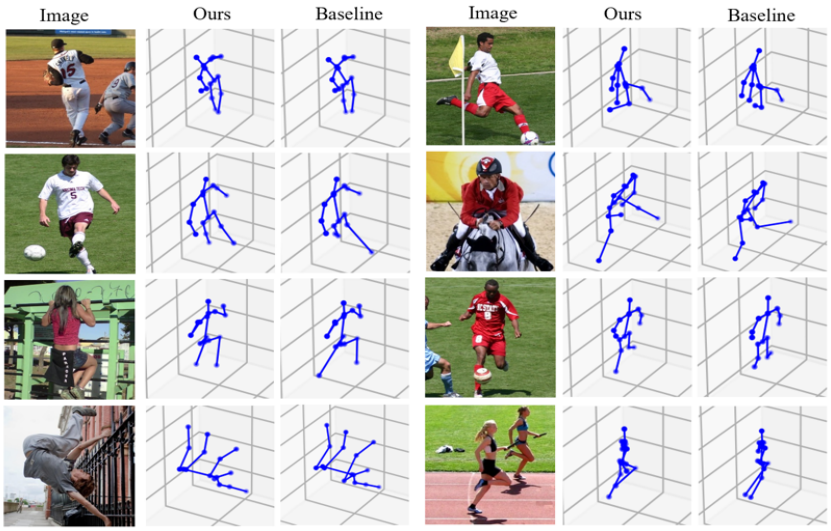

Figure 7. Visualization result comparison with the baseline model.

5. Conclusions

For the in-the-wild 3D pose estimation problem, we have proposed to improve the joints' regression accuracy in two ways. First, we introduced an attention mechanism to improve the depth regression ability of the model. It learns the dependencies between each channel, suppresses the weakly related channels and enhances the highly related ones. Second, we added a multi-scale calibration network to pose a kinematics constraint for model prediction. Through these two improvements, the model achieved good performance on public datasets. As seen from the experiments, our model takes a 2D image as input, so the model is not good at capturing the temporal dependencies. In our future work, we will try to import the attention module to a spatiotemporal framework, and also try do calibration to it. In order to further improve the model's generalizability to in-the-wild data, developing a weakly supervised model is also our future direction.

Acknowledgments

We would like to acknowledge the support of the Science and Technology Development Program of Jilin Province (20220101102JC) and Jilin Province Professional Degree Postgraduate Teaching Case Construction Project.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, et al., Single image 3D interpreter network, in *European Conference on Computer Vision*, (2016), 365–382. https://doi.org/10.1007/978-3-319-46466-4_22
2. C. H. Chen, D. Ramanan, 3D human pose estimation = 2D pose estimation+ matching, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 5759–5767. https://doi.org/10.1109/CVPR.2017.610
3. D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, et al., Monocular 3D human pose estimation in the wild using improved CNN supervision, in *2017 International Conference on 3D Vision (3DV)*, (2017), 506–516. https://doi.org/10.1109/3DV.2017.00064
4. D. Pavllo, C. Feichtenhofer, D. Grangier, M. Auli, 3D human pose estimation in video with temporal convolutions and semi-supervised training, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 7753–7762. https://doi.org/10.1109/CVPR.2019.00794
5. Y. Cheng, B. Yang, B. Wang, W. Yan, R. Tan, Occlusion-aware networks for 3D human pose estima- tion in video, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 723–732. https://doi.org/10.1109/ICCV.2019.00081
6. G. Moon, J. Y. Chang, K. M. Lee, Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019). https://doi.org/10.1109/ICCV.2019.01023
7. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. WardeFarley, S. Ozair, et al., Generative adversarial nets, in *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, **2** (2014), 2672−2680.
8. W. Yang, W. Ouyang, X. Wang, J. Ren, H. Li, X. Wang, 3D human pose estimation in the wild by adversarial learning, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 5255–5264. https://doi.org/10.1109/CVPR.2018.00551
9. B. Wandt, B. Rosenhahn, RepNet: Weakly supervised training of an adversarial reprojection network for 3D human pose estimation, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 7774–7783. https://doi.org/10.1109/CVPR.2019.00797
10. B. Wandt, M. Rudolph, P. Zell, H. Rhodin, B. Rosenhahn, CanonPose: Self-supervised monocular 3D human pose estimation in the wild, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 13289–13299. https://doi.org/10.1109/CVPR46437.2021.01309
11. S. Amin, M. Andriluka, M. Rohrbach, B. Schiele, Multi-view pictorial structures for 3D human pose estimation, in *British Machine Vision Conference*, (2013), 1–12.
12. H. Qiu, C. Wang, J. Wang, N. Wang, W. Zeng, Cross view fusion for 3D human pose estimation. in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 4342–4351. https://doi.org/10.1109/ICCV.2019.00444

13. H. Ci, C. Wang, X. Ma, Y. Wang, Optimizing network structure for 3D human pose estimation, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 2262–2271. https://doi.org/10.1109/ICCV.2019.00235
14. X. Ma, J. Su, C. Wang, H. Ci, Y. Wang, Context modeling in 3D human pose estimation: A unified perspective, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 6234–6243. https://doi.org/10.1109/CVPR46437.2021.00617
15. X. Zhou, Q. Huang, X. Sun, X. Xue, Y. Wei, Towards 3D human pose estimation in the wild: A weakly-supervised approach, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 398–407. https://doi.org/10.1109/ICCV.2017.51
16. Q. Zhou, B. Zhong, X. Liu, R. Ji, Attention-based neural architecture search for person re-identification, *IEEE Trans. Neural Networks Learn. Syst.*, **33** (2022), 6627–6639. https://doi.org/10.1109/TNNLS.2021.3082701
17. H. Guo, Z. Ren, Y. Wu, G. Hua, Q. Ji, Uncertainty-based spatial-temporal attention for online action detection, in *European Conference on Computer Vision (ECCV)*, (2022), 69–86. https://doi.org/10.1007/978-3-031-19772-7_5
18. X. Chu, W. Yang, W. Ouyang, C. Ma, A. L Yuille, X. Wang, Multi-context attention for human pose estimation, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 5669–5678. https://doi.org/10.1109/CVPR.2017.601
19. K. Su, D. Yu, Z. Xu, X. Geng, C. Wang, Multi-person pose estimation with enhanced channel-wise and spatial information, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 5667–5675. https://doi.org/10.1109/CVPR.2019.00582
20. Z. Cui, T. Song, Y. Wang, Q. Ji, Knowledge augmented deep neural networks for joint facial expression and action unit recognition, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, (2020), 14338–14349.
21. Z. Cui, P. Kapanipathi, K. Talamadupula, T. Gao, Q. Ji, Type-augmented relation prediction in knowledge graphs, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 7151–7159. https://doi.org/10.1609/aaai.v35i8.16879
22. M. R. Ronchi, O. M. Aodha, R. Eng, P. Perona, It's all relative: Monocular 3D Human Pose Estimation from weakly supervised data, preprint, arXiv:1805.06880.
23. V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, S. Ilic, 3D pictorial structures revisited: multiple human pose estimation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **38** (2016), 1929–1942. https://doi.org/10.1109/TPAMI.2015.2509986
24. A. Newell, K. Yang, J. Deng, Stacked Hourglass Networks for human pose estimation, in *European Conference on Computer Vision*, **9912** (2016), 483–499. https://doi.org/10.1007/978-3-319-46484-8_29
25. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90
26. S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, **37** (2015). 448–456.
27. R. Arora, A. Basu, P. Mianjy, A. Mukherjee, Understanding deep neural networks with rectified linear units, preprint, arXiv:1611.01491.

28. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155
29. M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, *IEEE Trans. Signal Process.*, **45** (1997), 2673–2681. https://doi.org/10.1109/78.650093
30. A. J. Haug, *Bayesian Estimation and Tracking: A Practical Guide*, Wiley, Hoboken, 2012.
31. J. Pearl, Fusion, propagation, and structuring in belief networks, *Artif. Intell.*, **29** (1986), 241–288. https://doi.org/10.1016/0004-3702(86)90072-X
32. H. T. Ma, Z. Yang, J. F. Griffith, P. C. Leung, R. Y. W. Lee, A new method for determining lumbar spine motion using bayesian belief network, *Med. Biol. Eng. Comput.*, **46** (2008), 333–340. https://doi.org/10.1007/s11517-008-0318-y
33. D. Tome, C. Russell, L. Agapito, Lifting from the deep: Convolutional 3D pose estimation from a single image, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 5689–5698. https://doi.org/10.1109/CVPR.2017.603
34. J. Martinez, R. Hossain, J. Romero, J. J. Little, A simple yet effective baseline for 3D human pose estimation, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 2659–2668. https://doi.org/10.1109/ICCV.2017.288
35. D. Drover, M. v Rohith, C. H. Chen, A. Agrawal, A. Tyagi, P. H. Cong, Can 3D pose be learned from 2D projections alone?, in *European Conference on Computer Vision*, (2018), 78–94. https://doi.org/10.1007/978-3-030-11018-5_7

AIMS Press