https://doi.org/10.53964/jmer.2024021

https://www.innovationforever.com

Journal of Modern Educational Research

ISSN 2790-3192 (Online)

Open Access

Research Article

A Human Pose Estimation Method Based on the BlazePose Model

Jialing Wu[1], Wanyi Li[1*], Yimin Chen[1], Rixiang Gu[1], Hanrui Weng[1], Jiaying Zheng[1], Qian Zhang[1], Yilin Wu[1]

[1]School of Computer Science, Guangdong University of Education, Guangzhou, Guangdong Province, China

[*]**Correspondence to: Wanyi Li, PhD, Associate Professor**, School of Computer Science, Guangdong University of Education, South of Guangzhou Avenue, Kecun, Haizhu District, Guangzhou, 510310, Guangdong Province, China; Email: luther1212@163.com

Received: September 27, 2024 **Revised:** November 21, 2024 **Accepted:** November 25, 2024 **Published:** November 26, 2024

Abstract

Objective: This study aims to evaluate the performance of the BlazePose model in human pose estimation for sports actions to improve the accuracy and computational efficiency of the model.

Methods: The research employed a variety of datasets for training and evaluating the model, investigating the generalization capability of the model under different experimental conditions. Detection and tracking were carried out through the two-stage mechanism of the BlazePose model, identifying the human figure contours and initially predicting the key points, and then refining the positions of the key points through the tracking module. Different parameters such as model complexity and key point smoothing were set, and the steps including image preprocessing, feature extraction, key point detection, naming, and skeleton construction were conducted.

Results: The research outcomes yielded the two-dimensional keypoints of single-frame human body images and the construction of 3D coordinate models, and obtained the visualization line charts of PCK experimental data. Through the comparison of PCK scores of different models, it was concluded that on the basketball Dataset, the performance of BlazePose was superior to that of OpenPose, particularly in terms of FPS performance. Although the PCK score of the lightweight version of BlazePose was slightly lower, its FPS performance was higher, making it suitable for scenarios with high requirements for speed. Furthermore, BlazePose effectively reduced the model complexity by offering models of different complexities and using occlusion simulation in training, without sacrificing too much accuracy.

Conclusion: This research presents the efficacy of lightweight neural networks in real-time human pose estimation and, through further experimental analyses, investigates the possibilities of enhancing their performance advantages and application effects.

Keywords: human pose estimation, lightweight convolutional neural network, blazepose

Citation: Wu J, Li W, Chen Y, Gu R, Weng H, Zheng J, Zhang Q,Wu Y. A Human Pose Estimation Method Based on the BlazePose Model. *J Mod Educ Res*, 2024; 3: 21. DOI: 10.53964/jmer.2024021.

https://doi.org/10.53964/jmer.2024021

1 INTRODUCTION

Human pose estimation[1], as a crucial branch of computer vision, holds significant importance for understanding and analyzing human actions. It provides potent technical support for the training and analysis of human actions by automatically identifying the key points of the human body in images or videos.

Although pose estimation techniques have demonstrated vast application potential in various aspects, there are still certain challenges in practical applications. Interference from complex backgrounds, occlusion among human body parts, as well as variable lighting and viewing angles can all have an impact on the performance of the algorithm. In rapidly changing motion scenarios, especially in cases of complex backgrounds or unfavorable lighting conditions, the accuracy and response speed of existing algorithms remain to be enhanced.

In recent years, scholars both domestically and internationally have carried out a considerable amount of research in this domain. For instance, Cao et al.[2] proposed a deep learning-based multi-person pose estimation approach that can effectively identify the key points of the human body in crowded scenes, enhancing the accuracy and robustness of pose estimation. In the field of sports, certain studies such as Zhang et al.[3]developed an athlete action analysis system based on pose estimation, which is capable of providing real-time feedback and guidance for athletes.

This paper investigates the accuracy of the lightweight convolutional upgraded network for human pose estimation by choosing the lightweight convolutional neural network BlazePose[4] architecture, while significantly reducing the computational complexity of the model and enabling its suitability for devices with limited resources.

2 MATERIALS AND METHODS

2.1 Adopt Lightweight Network Architectures

This research employed an enhanced architecture in the domain of human pose estimation. BlazePose is a lightweight convolutional neural network structure optimized for mobile devices, dedicated to real-time human pose estimation. With a relatively small number of parameters, this network structure effectively enhances the prediction quality of the model by drawing on the design of the stacked hourglass model[5]. It can rapidly identify and track 33 key points and predict their 3D positions. Additionally, BlazePose adopts the "BlazePicking" technology[6], which implements detection and tracking through a two-stage mechanism: Firstly, a lightweight CNN recognizes the human silhouette and preliminarily predicts the key points; Secondly, using this information, the tracking module refines the positions of the key points to enhance the stability and accuracy of positioning.

In this research, we constructed an efficient detector-tracker framework (as depicted in Figure 1, which has exhibited outstanding performance in multiple real-time tasks, including hand landmark prediction and dense facial landmark prediction. Our system comprises two main components: a streamlined human pose detector and a pose tracker network. The pose tracker is responsible for predicting the coordinates of the key points, assessing the presence of individuals in the current frame, and determining the region of interest for the pose in this frame. When the pose tracker detects no individuals in the current frame, we will restart the detector network in the next frame to re-perform human detection. Through this design, our framework is capable of effectively tracking human poses while maintaining real-time performance.

2.2 Model Design

This model adopts the fast downsampling[7] technique, which can rapidly decrease the image size and effectively lower the overall computational load, allowing the model to process image data promptly. To compensate for the possible information loss resulting from fast downsampling, a residual structure[8] is introduced in the model. This structure is capable of capturing shallow features and gradients, to a certain extent, making up for the information loss brought about by fast downsampling. Through strengthening the spatial information in the features, the model can learn more advanced semantic features and maintain performance even in the case of fast downsampling.

In the model design, the parameters are primarily concentrated on the backbone network to reduce the parameters and computational load on the residual branches. This design strategy is intended to focus the limited computing resources on the main branch to enhance the overall performance of the model. In terms of the output, the model generates multiple types of outputs(as depicted in Figure 2). Firstly, CenterHeatmap[B, 1, H, W] is employed to predict the geometric center of each individual, which is mainly utilized for existence detection and is equivalent to substituting the bounding box (bbox) in traditional object detection with an anchor point on the heatmap. Subsequently, KeypointRegression[B, 2K, H, W] regresses the coordinate values of the keypoints based on the center points, providing us with the precise location information of the keypoints.

2.3 The Training Process of Grid and Its Improvement

During the training process of the BlazePose network, a combined approach integrating keypoint detection and keypoint regression was employed. This training methodology encompasses the integration of heatmaps, offsets, and keypoint regression techniques, as depicted in Figure 3. The training procedure incorporates heatmaps, offsets, and keypoint regression techniques. This approach

https://doi.org/10.53964/jmer.2024021

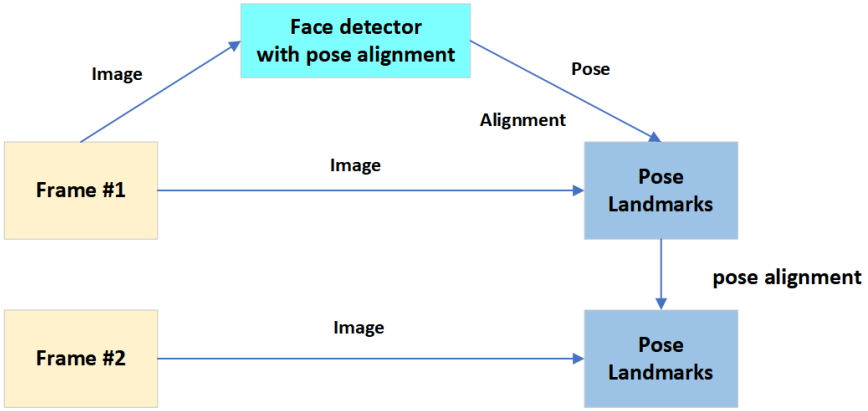

Figure 1. Flowchart of Human Prediction by Blazepose.

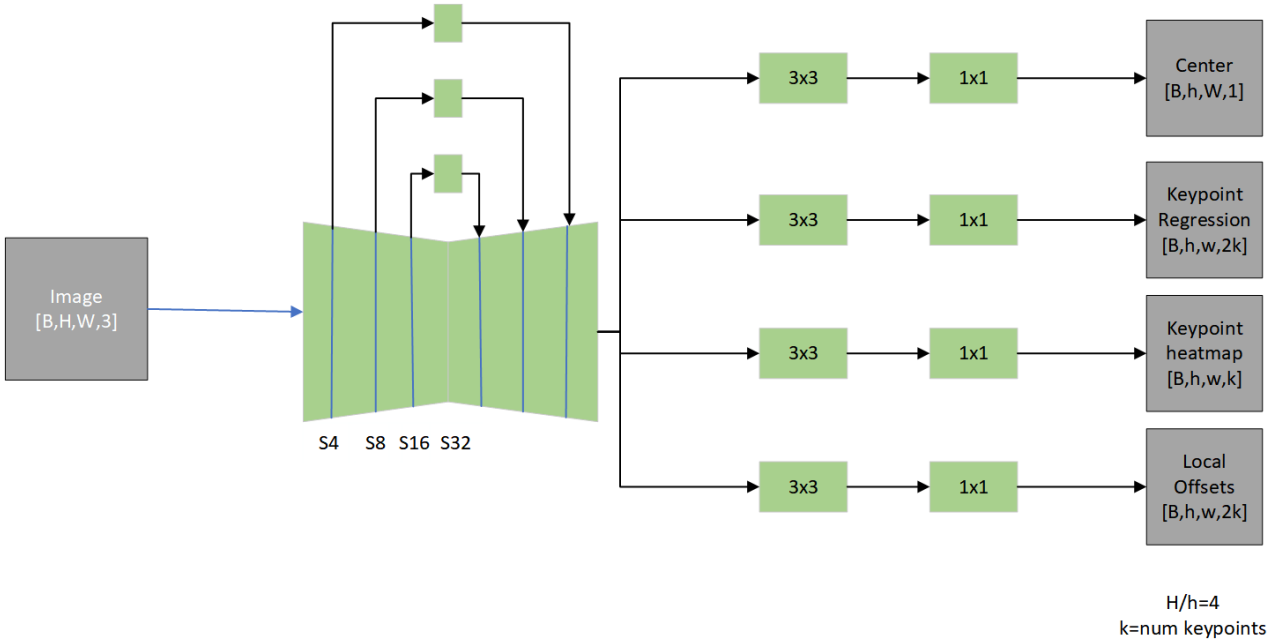

Figure 2. Model Design Structure Diagram.

corresponds to the heatmap (left), offset (middle), and regression output (right) in Figure 3.

In the initial stage of training, training was initially conducted using heatmaps and offsets. In the regression training phase[9], the output branch in Figure 3 (left) was truncated by the network. This truncation operation effectively utilized the heatmap to supervise the training process, achieving a notable enhancement in inference speed while maintaining the accuracy of keypoint detection. In contrast to the mere 17 human keypoints predicted by COCO Pose[10], the BlazePose network is capable of predicting 33 keypoints with confidence for each human body in every frame of the image.

This training approach effectively employs heatmaps to supervise lightweight embeddings[11] and applies them to the regression encoder network. In the training stage, we utilized heatmaps and offset losses, and removed the corresponding output layers from the model prior to inference. Such enhancements not only optimize heatmap prediction but also significantly enhance the accuracy of coordinate regression, enabling the BlazePose network to maintain efficient inference while attaining high precision.

2.4 Experiment

2.4.1 Experimental Illustrations

The input resolution of the image is set at 640x480 pixels; in the detection and tracking stage of the model, the confidence threshold is set at 0.5, signifying that only when the confidence of the model in the region containing the keypoint attains or exceeds 50%, will the region be regarded as a valid keypoint; the matplotlib library is employed for format conversion to ensure the correct processing and presentation of the data.

2.4.2 Method Design

To achieve a better balance between model complexity and accuracy, we introduced two versions of the BlazePose model with distinct complexities in our research to accommodate various application scenarios. The BlazePose Lite version reduces the number of model parameters and

https://doi.org/10.53964/jmer.2024021

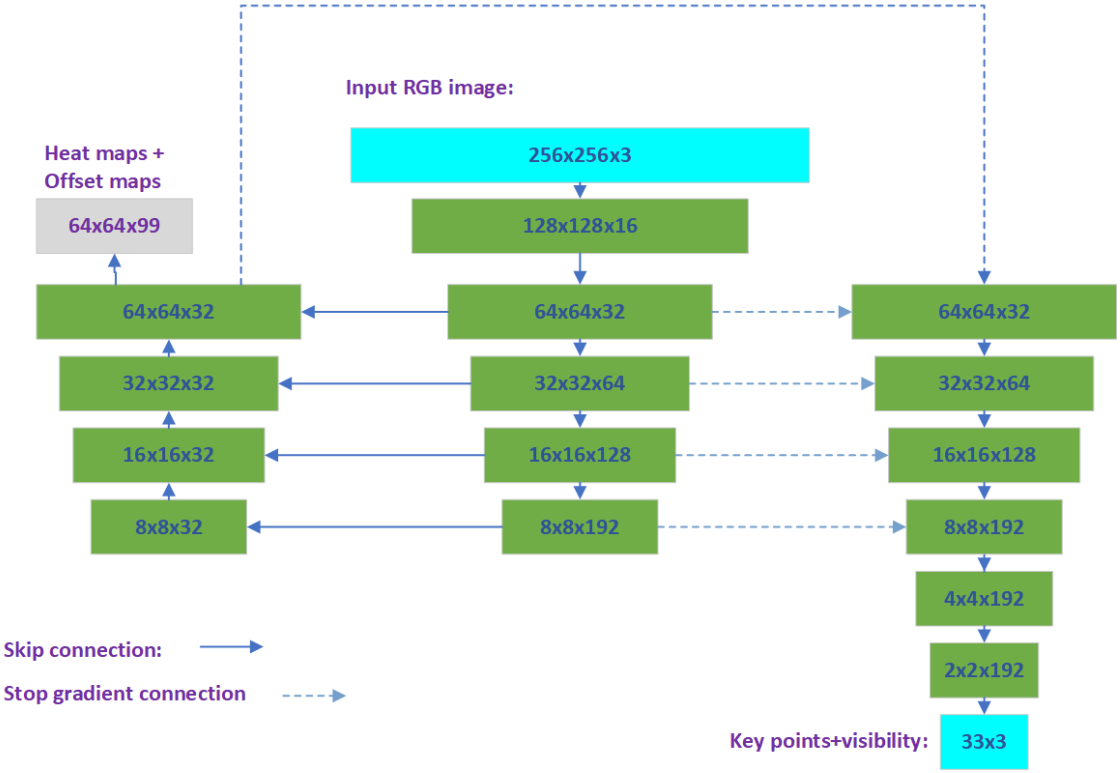

Figure 3. The Training Process of the BlazePose Network.

computational load, thereby attaining a faster inference speed. On the other hand, the BlazePose Full version offers higher accuracy and is applicable in circumstances where greater precision is demanded. This design enables a reduction in model complexity without sacrificing excessive accuracy, catering to diverse performance requirements and application settings.

In this study, the model was initialized first, with parameters such as complexity and keypoint smoothing being set. Once the image was input, it underwent four sequential processing steps: preprocessing, feature extraction, keypoint detection, and keypoint naming along with skeleton construction. During the image processing, the detected human keypoints were plotted on the image, and the information of each keypoint was printed. Simultaneously, the processing time was calculated to display the FPS.

Via the "get_coordinates" method, we extracted the coordinates (x, y) of each keypoint and stored them in a list for subsequent usage. All keypoint coordinates were normalized, with the x and y values ranging from 0 to 1, and the z value representing the depth information of the keypoint. The final output coordinates encompassed the pixel positions in the image and were adjusted through normalization, enabling the model to provide consistent outputs under different resolutions and viewing angles. The advantage of this approach lies in its ability to guarantee accuracy while reducing computational costs and enhancing processing efficiency, which is especially significant in environments with limited resources.

2.4.3 Experimental Data Set

In this study, the Ultralytics COCO8-pose public dataset, the basketball dataset, and the self-created dataset were selected to assess the model's generalization ability. All images were saved in the ".jpg" format and were divided into training sets and validation sets. Such a dataset combination facilitates the model's adaptation to different environments and enhances its generalization in various scenarios.

In view of the scarcity of a large-scale squat action dataset, a new dataset was constructed in this study (Figure 4), which was collected from multiple locations in the school. We precisely labeled the images using the LabelImg software and followed the PASCAL VOC format. Under controlled lighting and camera settings, images of diverse squat actions were captured at different times in the playground, gymnasium, and cafeteria, including occlusion and angle variations, thereby enhancing the practicality of the dataset and the reliability of the research.

4 RESULTS

4.1 Evaluation Indicator

The PCK (Percentage of Correct Keypoints) score is a widely employed metric for assessing the performance of pose estimation models. The PCK indicator measures the proportion of correctly predicted keypoints under a given normalized threshold α. The higher this proportion, the better the prediction accuracy of the model. Through PCK scores, we can visually observe the fitness performance of the model under different degrees of stringency. In this experiment, we adopted the PCK score for result analysis,

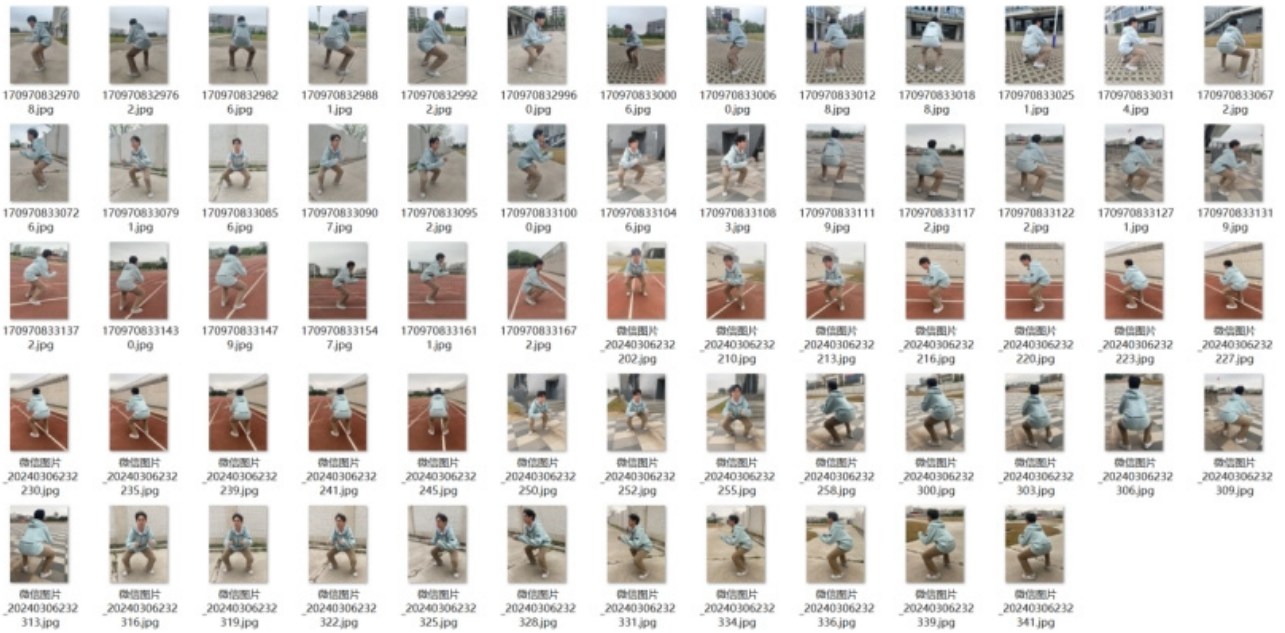

Figure 4. Self-made Dataset.

where the threshold was set as the minimum confidence threshold when tracking keypoints. The calculation formula of the PCK score is presented as Equation (1):

$$PCK(\alpha) = \frac{1}{N}\sum_{i=1}^{N} 1\left(\frac{d_{ref}}{d_i} \leq \alpha\right) \quad (1)$$

Specifically, we compute the distance between the actual keypoint coordinates in the label file and the predicted keypoint coordinates detected in the image, and utilize this distance to gauge the accuracy of the model. This approach offers an intuitive measurement of the matching degree between the predicted keypoints of the model and the actual annotations (ground-truth).

In comparison with other evaluation metrics, such as OKS[12], although it takes scale information into account, its calculation process is more complicated and less intuitive than PCK. Furthermore, AP (Average Precision)[13] and mAP (mean Average Precision)[14], while capable of providing comprehensive evaluations, are more frequently employed in object detection tasks rather than specifically for pose estimation. Hence, in this experiment, we select the PCK score as the primary evaluation metric to facilitate a more direct assessment and comparison of the performance of the models.

4.2 Experimental Results

According to the experimental data, we have ascertained that the average PCK (Percentage of Correct Keypoints) score for this set of data is 0.81. This value implies that in all test cases, approximately 81% of the keypoint predictions have reached the preset accuracy threshold. The fluctuation range of the PCK score is from 0.70 to 0.92, which indicates that the model can still maintain a relatively high accuracy in identifying keypoints when confronted with different test data. This outcome validates the effectiveness and reliability of the model in practical applications. The PCK score statistics for a certain group of data in the experiment are depicted in Figure 5.

The model conducts the recognition of human poses by analyzing individual frames within pictures and video streams, namely individual images or video frames, and simultaneously outputs the corresponding three-dimensional coordinate points. These points are utilized for constructing a three-dimensional data model. During this process, the input images are initially preprocessed to a size of 256x256 pixels to accommodate the input requirements of the deep neural network. Subsequently, these images are fed into a deep neural network encompassing 18 convolutional layers and multiple residual blocks for further processing. Eventually, the model is capable of outputting the three-dimensional coordinates of 33 key joints[14], precisely depicting the human pose. (as depicted in Figures 6-10)

According to the output results, it can be observed that BlazePose can precisely output the key coordinate points of basketball players under multiple circumstances. Whether the athletes are standing, running, jumping, striding or squatting, whether in an environment with or without occlusion, whether indoors or outdoors, BlazePose can exhibit relatively high accuracy. This suggests that BlazePose possesses excellent adaptability and accuracy when dealing with different actions and environmental conditions.

5 DISCUSSION

5.1 Experimental Analysis on the Generalization Ability of the Model

In this research, we assessed the generalization ability

https://doi.org/10.53964/jmer.2024021

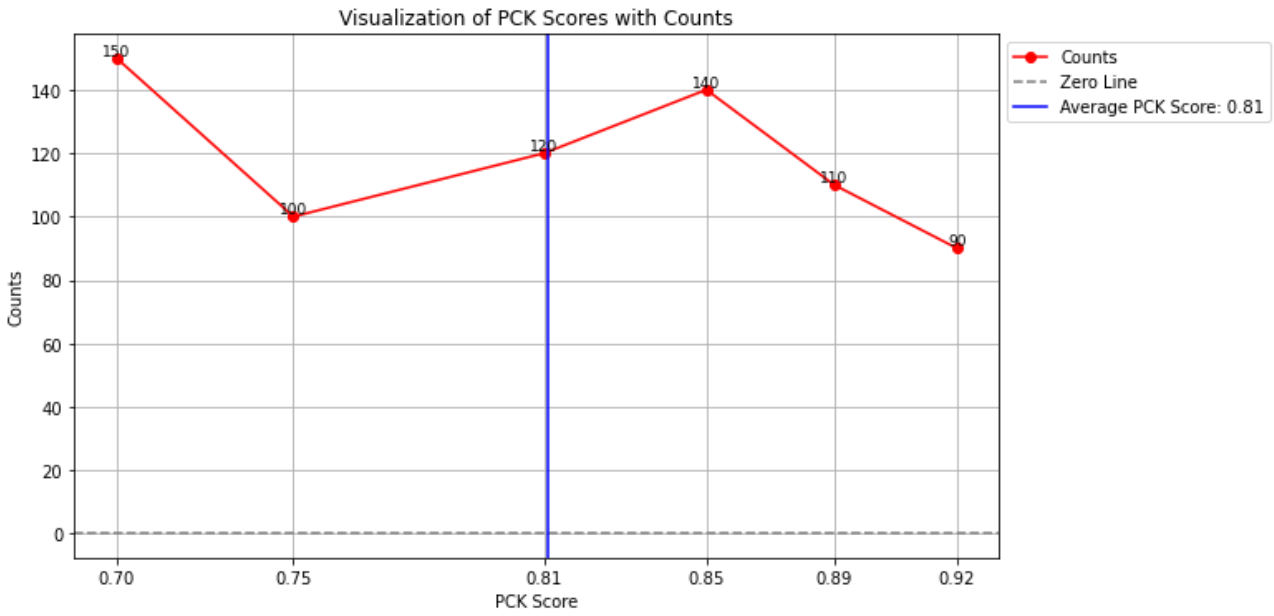

Figure 5. Line Chart of PCK Scores.

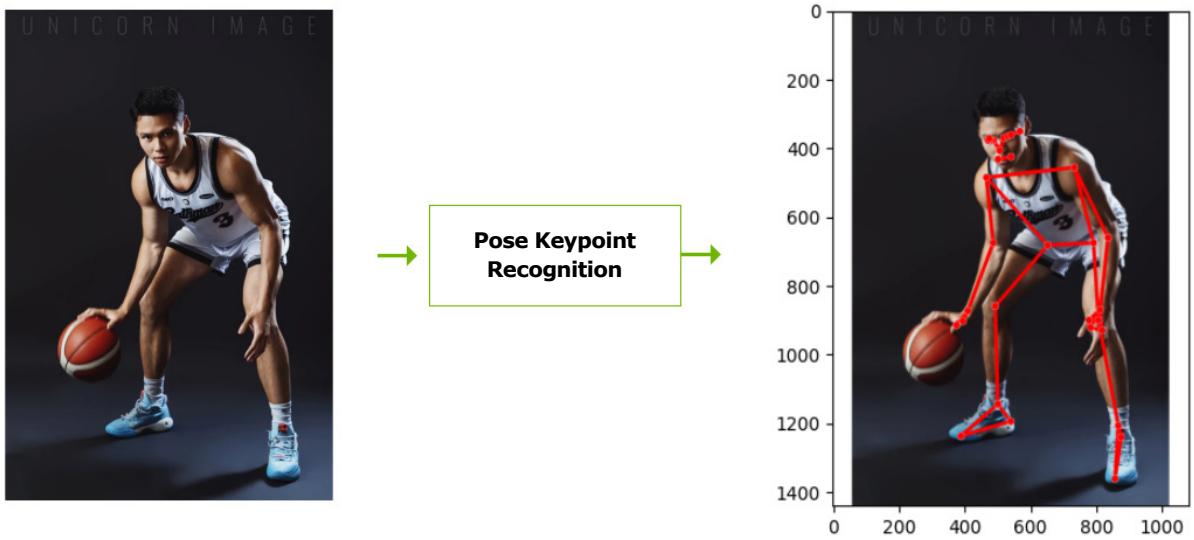

Figure 6. Comparison of Results of Human Pose Keypoint Recognition Before and After.

of the BlazePose model through conducting experiments on multiple datasets. We selected the public Ultralytics COCO8-pose dataset and a self-developed deep squat action dataset, which encompass diverse environments and human postures. The experimental outcomes indicate that the average PCK score of BlazePose on the COCO8-pose dataset is 0.81, and on the self-made dataset it is 0.78, demonstrating that the model exhibits relatively good consistency among different datasets. Furthermore, we tested the model's performance under various lighting conditions and background complexities, and discovered that the model performs superiorly in high lighting circumstances compared to low lighting ones, yet it can still maintain a relatively high level of accuracy in complex backgrounds.

The lightweight design of the BlazePose model enables real-time pose estimation on devices with limited resources, which theoretically contributes to the generalization of the model. Our model reduces the number of parameters and computational load, thereby minimizing the risk of overfitting and enhancing the model's performance on unseen data.

To enhance the model's generalization ability, we plan to explore integrating more contextual information and employing attention mechanisms in future studies. Additionally, we will consider strengthening the model's temporal analysis capability to better handle pose estimation issues in video streams.

5.2 Performance Comparative Evaluation (Quantitative Comparison of PCK Scores between OpenPose and BlazePose)

Judging from the comparison results in Table 1, the BlazePose model (comprising the full-featured version and

https://doi.org/10.53964/jmer.2024021

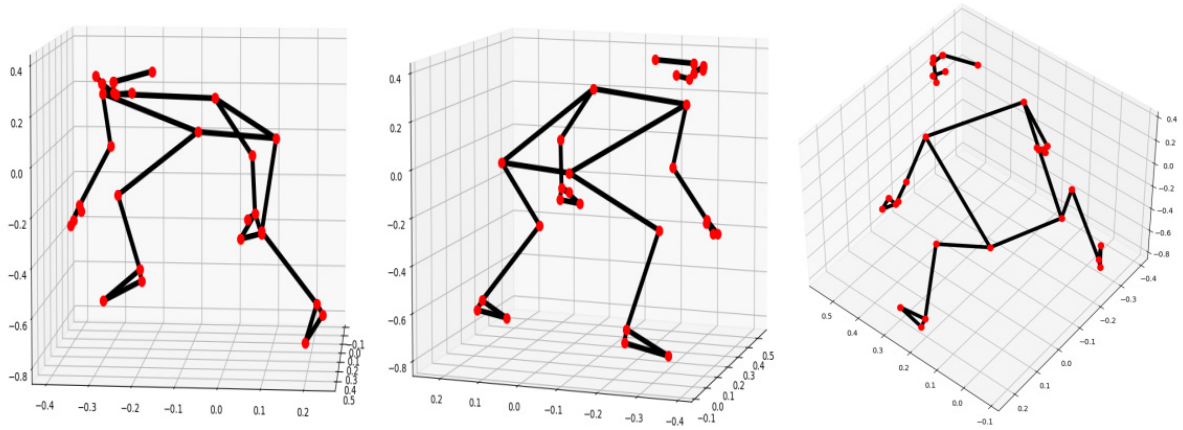

Figure 7. Construction of 3D Models with Frontal, Lateral and Rear Rotational Views.

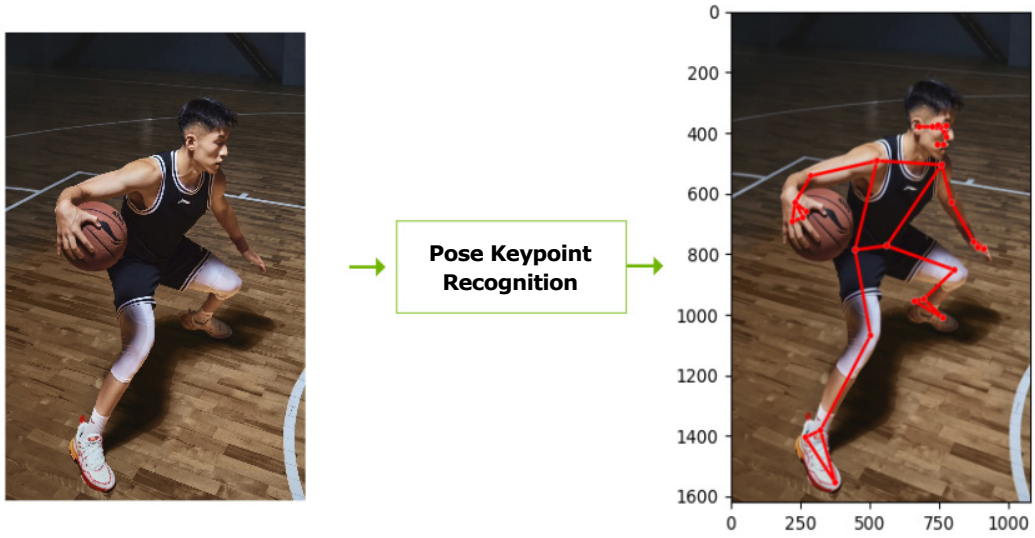

Figure 8. Comparison of Results of Human Pose Keypoint Recognition Before and After.

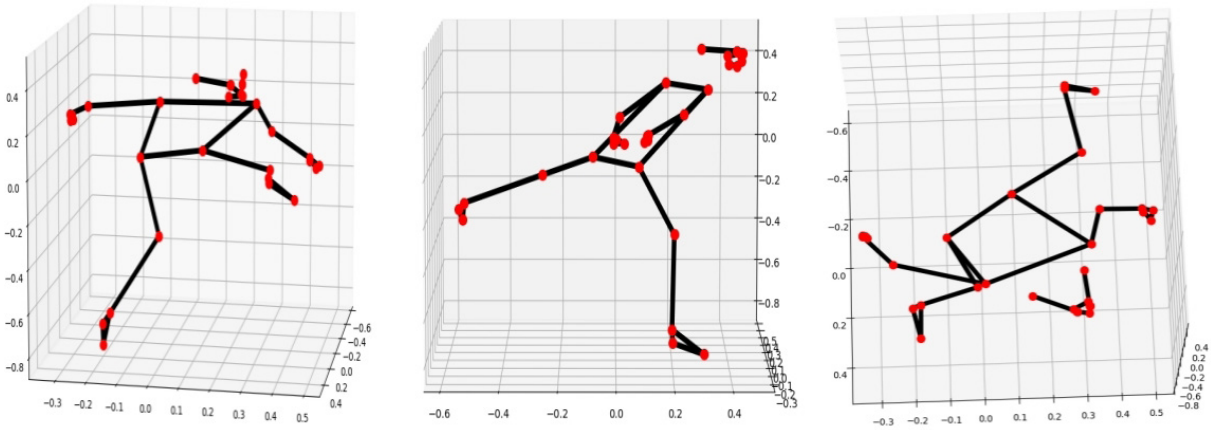

Figure 9. Construction of 3D Models with Frontal, Lateral and Rear Rotational Views.

the lightweight version) outperformed the OpenPose model on the basketball dataset, particularly in terms of FPS (frames per second) performance. Although the PCK score of the BlazePose lightweight version is marginally lower than that of the full-featured version, its FPS performance is higher. This indicates that in real-time application scenarios where rapid response is required, such as in sports analysis, the BlazePose lightweight version might be more appropriate.

Although the OpenPose model attains relatively high PCK scores on some datasets, its low FPS performance restricts its practicability in real-time applications. Nevertheless, on the basketball and fitness datasets, BlazePose Full outperforms OpenPose. These data offer the strengths and weaknesses of different models in diverse

https://doi.org/10.53964/jmer.2024021

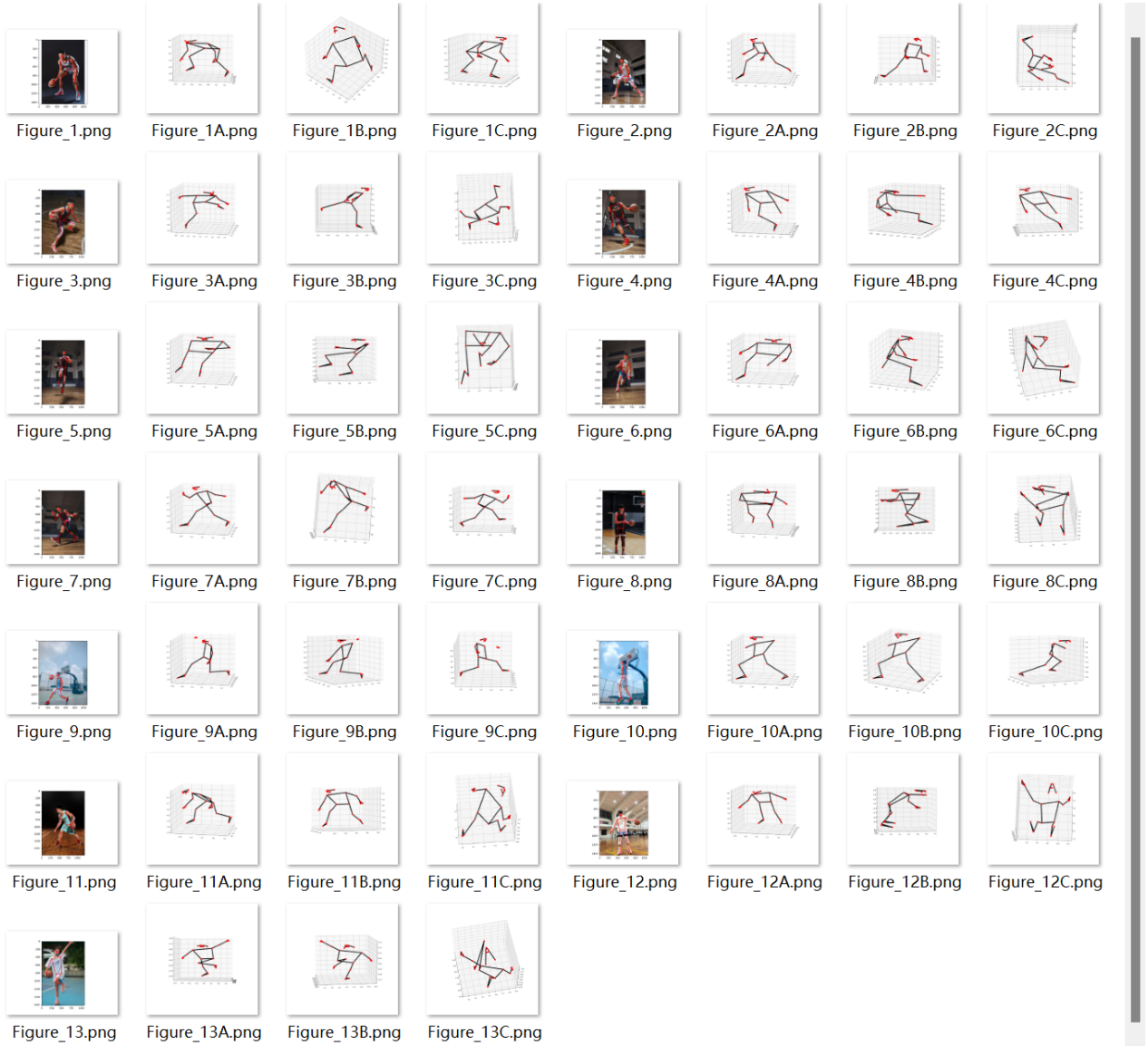

Figure 10. Output Partial Results of the Test Data Set.

Table 1. Data Table of Performance for Different Models

MODLE	FPS	Ultralytics COCO8-pose, PCK	17097083298XX Dataset, PCK	Basketball Dataset, PCK
OpenPose(Body Only)	0.5[1]	85.6	82.7	81.6
BlazePose Full Feature Edition	10[2]	82.5	84.5	85.2
BlazePose Lite Edition	32[2]	78.2	77.6	77.4

application scenarios, facilitating the selection of the appropriate model in accordance with specific requirements.

By designing two models with distinct levels of complexity: BlazePose Full (6.9MFlop, 3.5M Params) and BlazePose Lite (2.7MFlop, 1.3M Params), it becomes feasible to lower the model complexity without sacrificing excessive accuracy, so as to accommodate different performance requirements and application scenarios.

6 CONCLUSION

The real-time human pose estimation technology based on the BlazePose model holds significant application value in the sports domain, being capable of meticulously assessing the technical movements and training conditions of athletes. Nevertheless, traditional motion analysis approaches are frequently complex and time-consuming. To address this issue, this research puts forward a real-time human pose estimation method based on the BlazePose model.

The lightweight convolutional neural network and "BlazePicking" technology used in the study effectively improved the model's generalization and accuracy in real sports environments. The model's lightweight and fast response features make it more suitable for use on mobile devices, which provides the possibility for coaches to monitor athletes' movements and optimize training more

conveniently on mobile devices in real time.

Looking ahead, we will further optimize this technology. This can be achieved by expanding the dataset and increasing occlusion simulation to strengthen the model's robustness, facilitating its application in more sports scenarios and improving its performance when confronted with complex viewing angles and occlusions. Additionally, we plan to extend the experiment to larger datasets or longer movement sequences to evaluate its performance in more complex circumstances. To provide more consummate technical support for sports training. Future research will focus on the robustness, precision, and real-time performance of the algorithm, enhance its adaptability, and explore its application in a broader range of scenarios.

Acknowledgements

This work is supported by the National Undergraduate Innovation Training Project of China under Grant (No. 202414278015), the Guangdong Provincial Special Program in Key Areas for Higher Education Institutions (New Generation Electronic Information (Semiconductors)) (No. 2024ZDZX1040), the Collaborative Project for the Development of Guangzhou Philosophy and Social Science in 14th Five-Year Plan (No. 2023GZGJ171), the Educational Science Planning Project of Guangdong Province (No. 2022GXJK073, No. 2023GXJK125), the Special Support Program for Cultivating High-Level Talents of Guangdong University of Education (2022 Outstanding Young Teacher Cultivation Object: Wanyi Li).

Conflicts of Interest

The authors declared no conflict of interest.

Author Contribution

Wu J, Chen Y, and Gu R designed the experiment. Li W, Wu J, and Chen Y supervised the work. Gu R performed the data analysis. Weng H and Zheng J drafted the manuscript. All authors contributed to writing the article, read and approved its submission.

References

[1] Wang YF. Research on Human Pose Estimation and Action Recognition Based on Deep Learning [In Chinese]. Mianyang City, Southwest University of Science and Technology, 2024.

[2] Cao Z, Simon T, Wei SE et al. Real-time Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017; 7292-7300.[DOI]

[3] Zheng C, Zhu S, Mendieta M et al. 3D Human Pose Estimation with Spatial and Temporal Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021; 11656-11665.[DOI]

[4] Zhang HB, Liu DM, Fu NN et al. Research on Human Action Recognition Method Based on MediaPipe Pose [In Chinese]. *Ningxia Sci Technol*, 2024; 23: 79-91.

[5] Bazarevsky V, Grishchenko I, Raveendran K et al. BlazePose: On-device Real-time Body Pose tracking. *CoRR abs*.10204, 2020.

[6] Kong DZ. Research on the Application of MediaPipe-based Human Action Recognition Model in Y Balance Test [In Chinese]. Chengdu, Chengdu Sport University, 2024.

[7] Qin Z, Zhang Z, Chen X et al. FD-MobileNet: Improved MobileNet with a Fast Downsampling Strategy. arXiv preprint arXiv:1802.03750.

[8] Zhang Z, Wei H. Expression Recognition Method Based on Improved Residual Network Structure with Attention [In Chinese]. *Comput Appl Softw*, 2024; 41: 162-167.

[9] Huang Z, Wang C, Wei J, et al. Abnormal Gait Detection Based on BlazePose and Random Forest Algorithm. *Comput Technol Autom*, 2024; 43: 62-69.[DOI]

[10] Bazarevsky V, Grishchenko I, Raveendran K et al. BlazePose: On-device Real-time Body Pose tracking. *CoRR abs*, 2020; 10204.

[11] Yang LF, Zong ZW. Research on Human Joint Recognition in Tennis Serve Based on BlazePose [In Chinese]. *Ind Control Comput*, 2023; 36: 115-1120.

[12] Gao S. Research on Human-Machine Object Handover Based on Action Recognition [In Chinese]. China University of Mining and Technology, 2023.

[13] Song Y L, Wang C, Li D Y et al. Research on UAV Small Target Detection Algorithm Based on Improved YOLOv5s [J/OL]. *J Zhejiang Univ*,2024.

[14] Lv SJ, Qi YM, Deng SP et al. Research on Human Action Recognition Algorithm Based on 3D Skeletal Data [In Chinese]. *Eq Manuf Technol*, 2022; 9-30.