Journal of New Media
DOI: 10.32604/jnm.2022.027176
Article

Tech Science Press

Skeleton Keypoints Extraction Method Combined with Object Detection

Jiabao Shi[1], Zhao Qiu[1,*], Tao Chen[1], Jiale Lin[1], Hancheng Huang[2], Yunlong He[3] and Yu Yang[3]

[1]Hainan University, HaiKou, 570228, China
[2]University College London, Gower Street, London WC1E 6BT, UK
[3]Hainan Century Network Security Information Technology Co., Ltd., HaiKou, 570000, China
*Corresponding Author: Zhao Qiu. Email: qiuzhao@hainanu.edu.cn
Received: 12 January 2022; Accepted: 20 March 2022

Abstract: Big data is a comprehensive result of the development of the Internet of Things and information systems. Computer vision requires a lot of data as the basis for research. Because skeleton data can adapt well to dynamic environment and complex background, it is used in action recognition tasks. In recent years, skeleton-based action recognition has received more and more attention in the field of computer vision. Therefore, the keypoints of human skeletons are essential for describing the pose estimation of human and predicting the action recognition of the human. This paper proposes a skeleton point extraction method combined with object detection, which can focus on the extraction of skeleton keypoints. After a large number of experiments, our model can be combined with object detection for skeleton points extraction, and the detection efficiency is improved.

Keywords: Big data; object decetion; skeleton keypoints ; lightweight openpose

1 Introduction

Big data is the inevitable result of the development of the Internet and the Internet of Things. The core of big data technology is to realize the value of data. Artificial intelligence has undergone more than 60 years of development. Therefore, big data is the foundation of artificial intelligence, and artificial intelligence promotes the development of big data; together, big data and artificial intelligence form a new technological ecology. With the development of artificial intelligence and computer technology, the field of computer vision has received more and more attention. The continuous development of smart phones, laptops, cameras and other terminal devices makes images and videos used in more and more scenarios. In the analysis of the human body based on computer vision, human behavior recognition plays a very important role and has a wide range of application backgrounds. It is mainly used for human-computer interaction, intelligent video surveillance, patient monitoring systems, virtual reality, athlete-assisted training, robotics, etc.

Computer vision action recognition is currently roughly divided into three categories, based on RGB video data, based on skeleton data, and depth map sequences. The action recognition of the depth map sequence provides depth information that RGB does not have, but its large amount of data

contains flicker noise, which has been seldom studied at present. So scholars have made more attempts in action recognition based on RGB video and skeleton data. So to sum up, the detection of keypoints of human skeletons is one of the basic algorithms of computer vision, and it plays a fundamental role in the research of other related fields of computer vision.

Since the human body is quite flexible, there will be various postures and shapes. A small change in any part of the human body will produce a new posture. At the same time, the visibility of its keypoints is greatly affected by wearing, posture, and viewing angle. Facing the impact of occlusion, light, fog and other environments, in addition, 2D human skeleton keypoints and 3D human skeleton keypoints will have obvious visual differences, and different parts of the body will have a visual shortening effect, making human skeletons critical point detection has become a very challenging topic in the field of computer vision.

Object detection, also called object extraction, is another important field of computer vision. With the development of computer technology and the widespread application of computer vision principles, the use of computer image processing technology to track targets in real time is becoming more and more popular. The extraction of skeleton keypoints is also a hot spot in current research.

This paper proposes a skeleton point extraction method combined with object detection, which combines the object detection network with dynamic selection mechanism and an improved lightweight openpose algorithm to extract human skeleton points. The method in this paper is trained and analyzed on the object detection data set COCO and the human action recognition data set UCF101, and tested on the self-built data set.

2 Related Works

2.1 Object Detection

The task of target detection is to find objects of interest in images or videos, and to detect their positions and sizes at the same time. As one of the basic problems of computer vision, object detection forms the basis of many other vision tasks. Target detection can be divided into two stages, the traditional target detection algorithm period and the target detection algorithm period based on deep learning. Traditional target detection algorithms are mainly based on manually extracting features. The more representative algorithms are Viola Jones Detector [1] that uses a sliding window to check whether the target exists in the window, and calculates overlap on a dense grid of uniformly spaced cells. The HOG Detector [2] which normalizes the local contrast to improve the detection accuracy and the DPM algorithm [3] of the champion of the VOC object detection challenge.

After the birth and development of deep learning, computer vision has also been greatly improved on this basis, and object detection has a wide range of applications in many fields. Object detection methods based on deep learning can be divided into single-stage object detection and multi-stage object detection.

Single-stage object detection is to directly detect the entire image, so it has a relatively high detection speed. The more commonly used detection algorithms are YOLO algorithm and SSD algorithm. The YOLOv1 [4] algorithm was proposed by JosephRedmon et al. in 2016. It is the first single-stage deep learning detection algorithm. The algorithm divides the image into multiple grids, and then predicts the bounding box for each grid at the same time and gives the corresponding Probability. Compared with the two-stage algorithm, YOLOv1 has a very high efficiency, but the accuracy rate is relatively low. After continuous research and development and improvement, YOLOv2-YOLOv5 was launched one after another [5–7]. Due to the poor accuracy of the YOLO detection algorithm,

combined with the advantages of RCNN, the SSD algorithm [8] appeared. The SSD algorithm of Liu et al. uses a backbone network and four convolutional layers to perform target object features. At the same time, the object prediction mechanism is used to predict the object type and location information of different levels.

Two-stage target detection needs to first generate candidate regions from the image, and then perform target detection in the candidate regions. The RCNN algorithm proposed by Girshick et al. [9] in 2014 is to first select possible object frames through selective search, and then perform detection and classification. He K et al. proposed the spatial pyramid pooling layer SPPNet [10], which divides an image into image blocks of several scales, and then fuses the extracted features of each block together, thus taking into account the features of multiple scales. The Fast Rcnn [11] network is an improved version of RCNN and SPPNet. Under the same network configuration, a detector and a bounding box regressor can be trained at the same time. The Faster RCNN [12] of Ren et al. is the first end-to-end deep learning detection algorithm that is closest to real-time performance. The main innovation of this network is to propose a regional selection network for candidate frames, Can greatly improve the generation speed of the detection frame. The Cascade R-CNN proposed by Cai [13] in 2017 improves the accuracy of the two-stage target detection algorithm to a new level. It uses different IoU thresholds for training. By raising the IoU threshold to train the cascaded detector, the positioning accuracy of the detector can be higher. In order to make the network more focused to extract the skeleton keypoints, this paper combines the object detection model with the skeleton key point extraction model.

2.2 Keypoints Detection of Human Skeletons

The detection of keypoints of human skeletons, that is pose estimation, mainly detects some keypoints of the human body, such as joints, facial features, etc., through these keypoints, the description of human skeleton information and the construction of human skeleton topological maps.

The detection of keypoints of traditional human skeleton data is basically based on the idea of template matching. The core of the algorithm is how to use templates to represent the overall structure of the entire person, including the representation of keypoints, the representation of limb structure, etc.

The Pictorial Structure [14] proposed by Pishchulin et al. is a more classic algorithm in this idea. It proposes a spring deformation model to model the relative spatial position relationship between the component model and the overall model, and uses some spatial prior knowledge of the object. It not only reasonably restricts the spatial relative position of the overall model and the component model, but also maintains a certain degree of flexibility. In order to expand the range of matching poses, Yang & Ramanan proposed the concept of "mini parts", that is, each body structure part is divided into smaller parts to be able to simulate more pose changes, thereby improving the effect of template matching [15].

After the birth and continuous development of deep learning, the detection of keypoints of human skeletons has also entered a new stage. DeepPose [16] proposed by Toshev et al. is the first to apply deep neural networks to human pose estimation. It transforms the 2D human pose estimation problem from the original image processing and template matching problems into CNN image feature extraction and key point coordinate regression. Problem, and used some regression criteria to estimate the occluded or non-appearing human joint nodes. Also having a subversive status in the field of 2D human pose recognition is the stacked hourglass network for human pose estimation [17], which integrates and amplifies the multi-resolution feature ideas proposed by DeepPose [16]. The cascaded pyramid network proposed by face++ in 2017 for pose estimation [18] adopts a top-down detection strategy. Compared with the previous estimation of human body pose through complex network structures

such as feature scaling and restoration, Microsoft proposed a simple and effective simple baseline method for human body pose estimation and tracking in 2018, and achieved good performance using a simple network structure. The human body pose estimation open-source project openpose [19] of Neggie Mellon University uses VGG-19 as a skeleton feature, which has achieved good robustness, and the occluded part can also be estimated. Openpose's skeleton point extraction was subsequently used in multiple behavior recognition models. The HRNet [20] model jointly released by the University of Science and Technology of China and Microsoft Asia Research Institute can be used for high-resolution human pose estimation.

3 Method Design

In this paper, the key point recognition process of human skeletons combined with object detection is shown in Fig. 1. After preprocessing the image by Gaussian filter, we perform object detection on the processed image, and after filtering processing, we further manipulate the target person. Then an improved lightweight openpose algorithm is used to extract the key point coordinates of the skeletons of the target task.

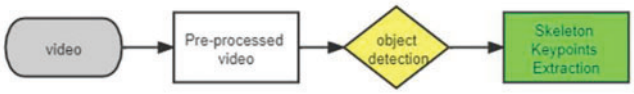


Figure 1: Skeleton point extraction combined with object detection

3.1 Object Detection Network with Sknet

For the detection of image characters in videos, this article uses a deep learning object detection toolbox mmdetection [21], which supports the mainstream object detection frameworks of Faster-RCNN, Mask-RCNN, Fast-RCNN, and Cascade-RCNN. This paper tested the three object detection architectures Faster-RCNN, Mask-RCNN, and Cascade-RCNN on the COCO data set. After comparing various indicators, it was decided to adopt the improved Cascade R-CNN method proposed by Cai Z et al. as object detection's frame.

The Cascade R-CNN method is mainly aimed at selecting the IOU threshold in object detection. It solves that the larger the threshold is, the easier it is to obtain high-quality samples, but at the same time, the problem of overfitting will also occur, which has certain advantages in accuracy. In order to make the network can better use the effective receptive field to capture information, this paper adds an adaptive dynamic selection mechanism SKNet on the basis of the original Cascade R-CNN convolutional network. After adding this mechanism, each of our neurons adaptively adjusts the size of its receptive field according to the multi-scale of the input information, and captures target objects of different scales.

SKNet [22] implements convolution through the three operations of Split, Fuse and Select. In Split, we convert the convolution sizes 3 and 5, and replace the 5∗5 conventional convolution kernel with a $3 * 3$ convolution kernel to improve efficiency. In Fuse, the result of fusing three branches through element summation is

$$U = U1 + U2 + U3 \tag{1}$$

Here we use global pooling to generate statistics,

$$S_c=F_{gp}(U_C) = \frac{1}{H * W} \sum_{i=1}^{H} \sum_{j=1}^{W} Uc\,(i,j) \tag{2}$$

Then use a fully connected layer to improve efficiency and reduce dimensionality,

$$z = F_{fc}\,(s) = \delta\,(B\,(Ws)) \tag{3}$$

In the select operation, we use the three weight matrices of a, b, and c to perform weighting operations on U1, U2, and U3 to find the final vector.

The Cascade R-CNN network structure integrated into the SKNet block is shown in Fig. 2.

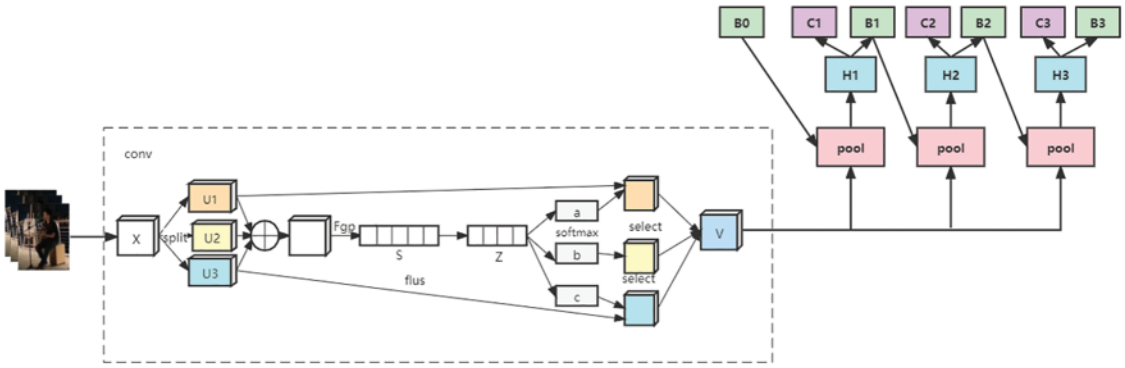


Figure 2: The Cascade R-CNN network structure integrated into the SKNet block

3.2 Extraction of Object Skeleton Points

After the object detection, through filtering and classification, this article only selects the object we want to extract the object skeleton points, that is only extract the skeleton points of the person. The lightweight openpose [23] is chosen as the framework for skeleton point detection. The lightweight openpose skeleton point detection comes from the fact that its parameters are only 15%. In the case of high efficiency, the performance is almost the same as the original openpose. On the basis of this framework, we have made improvements.

The lightweight openpose does not use a single VGG network [24] as the backbone architecture, but uses a combination of the MobileNet network and the VGG network. In this way, the model parameters for skeleton point detection can be reduced, and the calculation cost can be reduced. The external structure of the lightweight openpose is shown in Fig. 3. In order to increase the receptive field outside the lightweight openpose, this article discards the original single dilated convolution, and uses the ASPP [25] module to apply the dilated convolution to each branch. The ASPP block diagram is shown in Fig. 4. The extracted features are fused. The lightweight openpose internal network that joins ASPP is shown in Fig. 5.

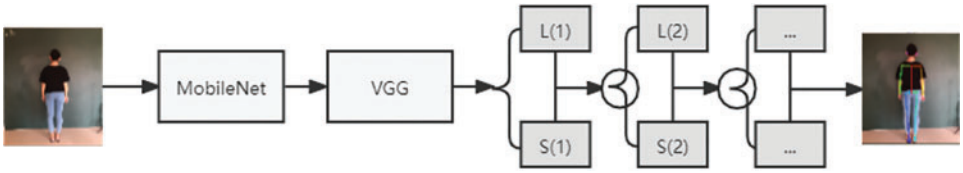


Figure 3: Openpose external structure diagram

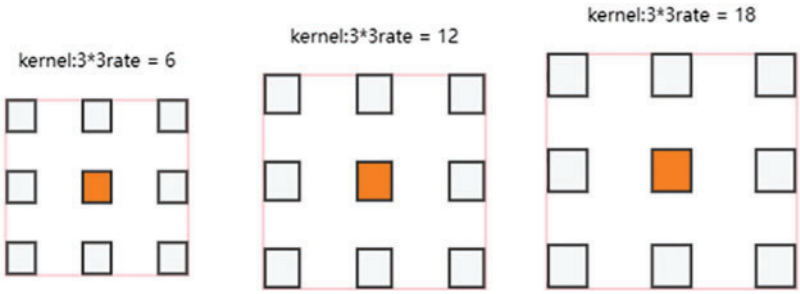


Figure 4: ASPP block diagram

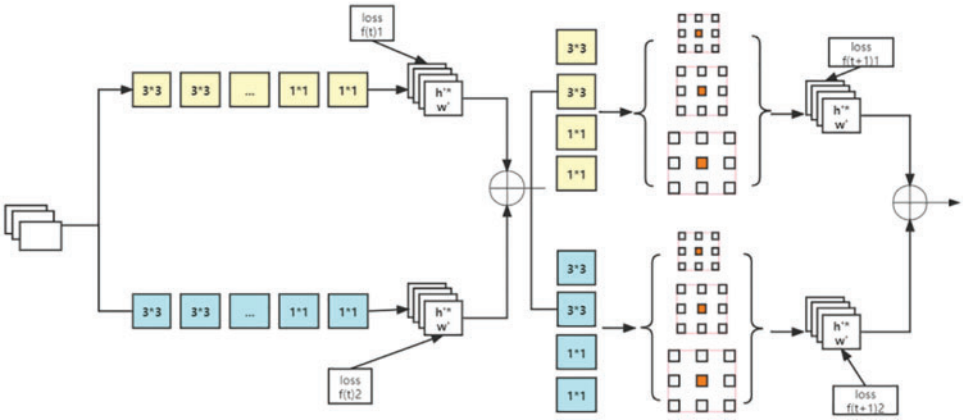


Figure 5: Add ASPP's lightweight openpose internal network diagram

There are two ways to output openpose skeleton points, namely 25 joint points and 18 joint points. The difference between the 7 joint points is mainly different from whether the foot joint points are recognized. In this paper, the ultimate goal of the skeleton points combined with object detection is to serve as the predecessor of action recognition and action classification. Therefore, the joint points of the foot do not have much influence on the final effect, so this article sets up the model in the model, and only selects the output format of 18 joint points. The output of the 18 bone points is shown in Fig. 6.

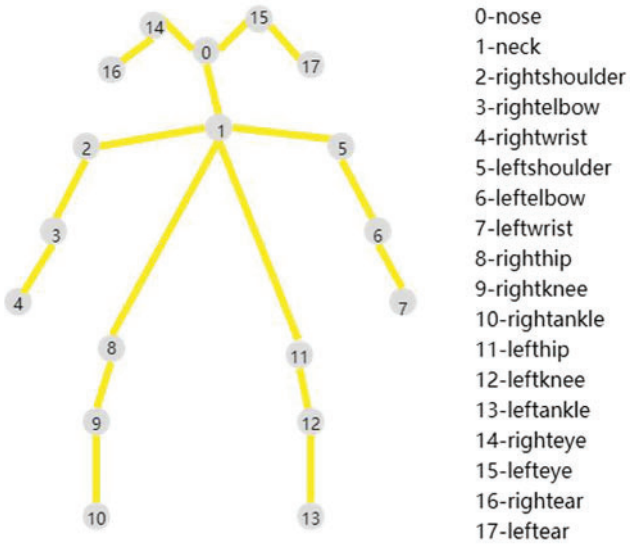


Figure 6: 18 skeleton keypoints map

4 Experimental Results and Analysis

4.1 Experimental Conditions and Data Sets

The experimental computer is configured as Intel Core i5-8500@3.0GHz, NVIDA GeForce 1080 GPU, running memory is 16GB, and the operating system is ubuntu 18.04. In the experiment, the deep learning framework used is Pytorch.

In object detection, the data set we use is the COCO data set [26]. The full name of the COCO data set is Common Objects in COntext, which is a large and rich object detection, segmentation and subtitle data set provided by the Microsoft team. This data set aims at scene understanding, which is mainly intercepted from complex daily scenes. The object in the image is calibrated through precise segmentation. The image includes 91 types of objects, 328,000 images and 2,500,000 labels. It is the largest data set with semantic segmentation so far. There are 80 categories and more than 330,000 images, 200,000 of which are labeled. The number of individuals in the entire data set exceeds 1.5 million.

In the skeleton key point extraction, that is, pose estimation, this paper uses the UCF101 data set. This dataset is a real-life action video action recognition dataset collected from YouTube, providing 13,320 videos from 101 action categories. UCF101 provides the greatest variety in motion, and has great changes in camera movement, object appearance and posture, object scale, viewpoint, cluttered background, lighting conditions, etc. UCF101 includes five categories: human-object interaction, pure body movements, human-human interaction, playing musical instruments, and sports. Videos in 101 action categories are divided into 25 groups, and each group can contain 4–7 videos of one action. Videos in the same group may have some common characteristics, such as similar backgrounds, similar views, and so on.

In the extraction test of skeleton keypoints combined with object detection, this paper also uses a self-built data set for testing. The self-built data set is randomly shot videos, each video is 15–45 s, and a total of 150 test videos are shot.

4.2 Evaluation Indicators and Experimental Results

Since mmdetection is compatible with the coco data set and the voc data set, in contrast, the training results of the coco data set will be better, so we chose to test the three mainstreams of Faster-RCNN, Mask-RCNN, and Cascade-RCNN on the coco data set. In the object detection framework, the object detection evaluation index used is the average accuracy. Generally speaking, the higher the AP value, the better its classification effect. The AP value is the area under the precision-recall curve. The formulas for precision and recall are as follows:

$$\text{Precision} = \frac{tp}{tp + fp} = \frac{tp}{n} \tag{4}$$

$$\text{Recall} = \frac{tp}{tp + fn} \tag{5}$$

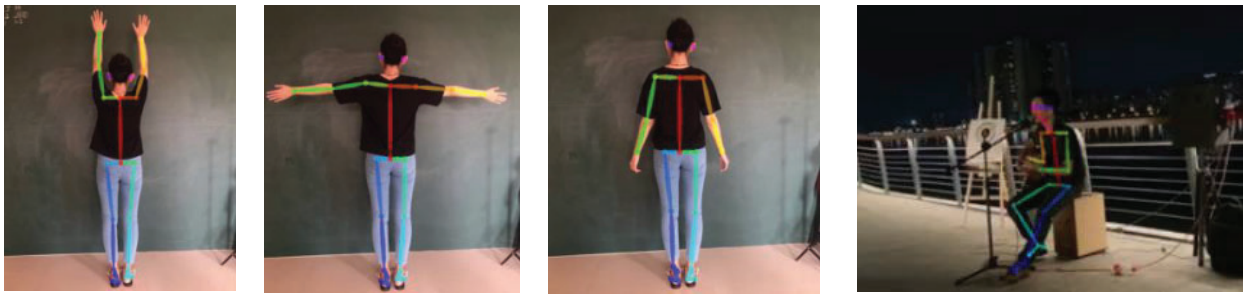
The AP values of the three object detection frameworks are shown in Tab. 1.

Table 1: 3 object detection frameworks AP on the coco dataset

	AP (average-precision)
Faster R-CNN	34.9
Mask R-CNN	38.2
Cascade R-CNN	42.8

After the following results are obtained, Cascade-RCNN is chosen as the initial architecture of object detection. We use the mmdetection toolbox, and the framework chooses Cascade-RCNN with dynamic selection mechanism as the detection architecture to get a good object detection effect.

When extracting the keypoints of skeletons, this article makes a comparison. The experimental results of using openpose to extract keypoints of skeletons are shown in Fig. 7. The following experimental images are all from video screenshots:



(a) (b) (c) (d)

Figure 7: Experimental results of openpose extracting keypoints of skeleton

Using lightweight openpose for skeleton point extraction, the extraction speed has been greatly improved. After the object detection combined with the dynamic selection mechanism, the key skeleton points can be extracted on the basis of the target detection. Some experimental diagrams of bone key point extraction combined with target detection are shown in Fig. 8:

(a) (b) (c) (d)

Figure 8: Experimental diagram of skeleton keypoints extraction combined with object detection

5 Conclusion

Big data and artificial intelligence complement each other and achieve each other. The development of artificial intelligence-computer vision is inseparable from the support of big data. This paper proposes a skeleton point extraction method combined with object detection. By combining the object detection network with the dynamic selection mechanism and the improved lightweight openpose, the skeleton keypoints can be extracted efficiently and accurately.

Acknowledgement: Thanks to Qiu Zhao for his comments on the paper and financial support.

Funding Statement: This work was supported by Hainan Provincial Key Research and Development Program (NO:ZDYF2020018), Hainan Provincial Natural Science Foundation of China (NO: 2019RC100), Haikou key research and development program (NO: 2020-049).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, HI, USA, pp. 1, 2001.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA, pp. 886–893, 2005.

[3] P. Felzenszwalb, D. Mcallester and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, pp. 779–788, 2008.

[4] J. Redmon, S. Divvala, R. Girshick R and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 779–788, 2016.

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 6517–6525, 2017.

[6] J. Redmon, A. Farhadi, "YOLOv3: An incremental improvement, 2018," Available: arXiv preprint arXiv:1804.02767, 2018.

[7] A. Bochkovskiy, C. Y. Wang and H. Liao H, "YOLOv4: Optimal speed and accuracy of object detection, 2020," Available: arXiv preprint arXiv:2004.10934.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "SSD single shot MultiBox detector," in *European Conference on Computer Vision*, Springer, Cham, pp. 21–37, 2016.
[9] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 580–587, 2014.
[10] K. He, X. Zhang, S. Ren and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *IEEE Transactions on Pattern Analysis & Machine Intelligence*, New York, NY, USA: IEEE Computer Society, pp. 1904–1916, 2015.
[11] R. Girshick, "Fast R-CNN," in *IEEE Int. Conf. on Computer Vision*, Santiago, CP, Chile, pp. 1440–1448, 2015.
[12] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *IEEE Transactions on Pattern Analysis & Machine Intelligence*, New York, NY, USA: IEEE Computer Society, pp. 1137–1149, 2017.
[13] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 6154–6162, 2018.
[14] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.
[15] E. Dogan, G. Eren, C. Wolf, E. Lombardi and A. Baskurt, "Multi-view pose estimation with flexible mixtures-oParts," in *Int. Conf. on Advanced Concepts for Intelligent Vision Systems*, Springer, Cham, pp. 180–190, 2017.
[16] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 1653–1660, 2014.
[17] A. Newell, K. Yang and D. Jia, "Stacked hourglass networks for human pose estimation," in *European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 483–499, 2016.
[18] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu *et al.,* "Cascaded pyramid network for multi-person pose estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7103–7112, 2018.
[19] Z. Cao, G. Hidalgo G, T. Simon, W. -E. and SheikhYaser, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 7291–7299, 2018.
[20] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *IEEE Conf. on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 5693–5703, 2019.
[21] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong *et al.,* "MMDetection: Open MMLab detection toolbox and benchmark," Available: arXiv preprint arXiv:1906.07155, 2019.
[22] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, pp. 510–519, 2019.
[23] D. Osokin, "*Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose*," Available: arXiv preprint arXiv:1811.12004, 2018.
[24] K. Simonyan, A. Zisserman, "*Very deep convolutional networks for large-scale image recognition*," Available: arXiv preprint arXiv:1409.1556, 2014.
[25] K. He, X. Zhang, S. Ren and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 97, no. 9, pp. 1904–1916, 2015.
[26] L. Yao, N. Ballas, K. Cho, J. R. Smith and Y. Bengio, "Microsoft COCO captions: Data collection and evaluation server," Available: arXiv preprint arXiv:1504.00325, 2015.