2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW) | 978-1-6654-7218-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICMEW56448.2022.9859415

GOLFPOSE: GOLF SWING ANALYSES WITH A MONOCULAR CAMERA BASED HUMAN POSE ESTIMATION

Zhongyu Jiang[1⋆], *Haorui Ji*[2⋆], *Samuel Menaker*[2] *and Jenq-Neng Hwang*[1]

[1]Dept. Electrical & Computer Engineering , University of Washington
[2]SPORTSBOX.AI INC.
zyjiang@uw.edu, haoruij@sportsbox.ai, samm@sportsbox.ai, hwang@uw.edu

ABSTRACT

With the rapid developments of computer vision and deep learning technologies, artificial intelligence takes a more and more important role in sports analyses. In this paper, to attain the objective of automated golf swing analyses, we propose a lightweight temporal-based 2D human pose estimation (HPE) method, called GolfPose, which achieves improved performance than the state-of-the-art image-based HPE methods. Unlike traditional image-based methods, our temporal-based method, designed for efficient and effective golf swing analyses, takes advantage of the temporal information to improve the estimation accuracy of fast-moving and partially self-occluded keypoints. Furthermore, in order to make sure the golf swing analyses can run on mobile devices, we optimize the model architecture to achieve real-time inference. With around 10% of the parameters and half of the GFLOPs used in the state-of-the-art HRNet, our proposed GolfPose model can achieve 9.16 mean pixel error (MPE) in our golf swing dataset, compared with 9.20 MPE for HRNet. Furthermore, the proposed temporal-based method, facilitated with golf club detection(GCD), significantly improves the accuracy of keypoints on the golf club from 13.98 to 9.21 MPE.

Index Terms— Sports Analysis, Human Pose Estimation, Golf Swing, Line Segment Detection

1. INTRODUCTION

Sports are social-cultural activities and have already become important parts of our daily life. It allows people to interact with each other regardless of their social status, helps to improve the quality of people's lives, and also serves as a significant symbol for measuring the development and progress of a country and society.

A significant amount of resources have been allocated to the modern sports industry, which demands higher requirements not only on the athletes themselves but also in a lot of related supporting technologies. For example, tracking players' trajectories in the field[1] can improve the audiences' experience during game broadcasting, analyze and assess players' performance for better coaching[2], detect and prevent life-threatening situations to players, etc. These requirements call for accurate analyses of actions, conditions, and environments across different players, scenarios, and sports events.

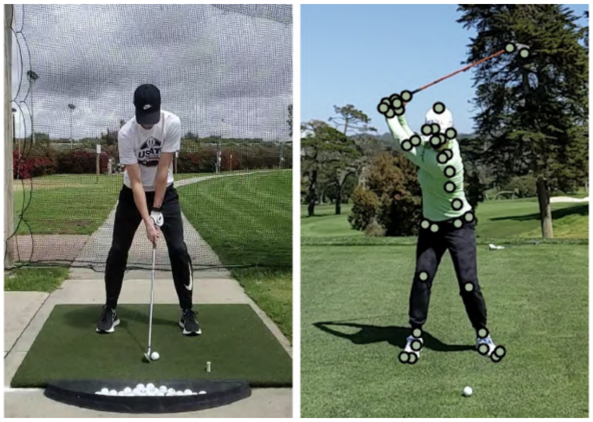

Fig. 1. Two sample images from our dataset. The right one is annotated with 38 keypoints.

Before the modern sports industry era, sports analytics could only use naked human eyes and their own experience to measure and analyze, which are unreliable, inefficient, and too subjective to generalize to different players, scenarios, and sports events. Therefore, manual analyses of sports are being gradually replaced by a combination of different sensors and algorithms that can automatically do all the cumbersome analyses. These capacities can help better assess the crucial sports event moments, resulting in more precise, efficient, and generalizable analysis results.

Among these newly emerging technologies, rapid development in computer vision communities combined with recent deep learning technologies have been highly appreciated in terms of efficiency and accuracy. Furthermore, thanks to the popularity of social media and online streaming, massive video and image data are generated and become available for researchers to utilize and improve the performance of their applications.

More than 24.8 million people played golf in the U.S. in

⋆ These two authors contribute equally

2020, and the demand for user-friendly automated golf swing analyses is rising. The crucial thing for sports analyses is how to understand and judge the motion of sports players. Therefore, an accurate and efficient human pose estimation (HPE) method is critical for reliable golf swing analyses. However, HPE for golf swing analyses is different from the other HPE tasks, because of the input format, motion blur and self-occlusion. Therefore, in this paper, we propose a temporal-based lightweight 2D HPE pipeline, called GolfPose, which can be running on mobile devices for golf swing analyses.

Our contributions include:

- A light-weight monocular temporal-based 2D human pose estimation model which provides accurate pose estimation results and can be deployed on mobile devices.
- Incorporating line segment based golf club detection(GCD) to further improve pose estimation accuracy.
- An annotated golf swing dataset with more than 500 videos of over 120 fps and 120,000 images.

2. RELATED WORKS

2.1. 2D Pose Estimation

Deep learning has proved its superior performance in many vision tasks, including pose estimation, which is widely applied in many areas[3, 4, 5]. Initially, the positions of keypoints can be regressed directly from images, while later on, estimating keypoint heatmaps[6] followed by choosing the locations with the highest values as the keypoint coordinates becomes the mainstream method.

Nowadays, human pose estimation (HPE) methods can be broadly classified into bottom-up and top-down approaches, depending on whether human body bounding boxes are detected first. Top-down approaches [7, 8] first detect human bounding boxes and then perform human keypoint detection within every bounding box's region. Bottom-up approaches [9, 10], on the other hand, first detect all keypoints on all humans in the image and then associate keypoints belonging to the same person individually. However, because of the dataset limitation and the difficulty of accurate human tracking in the wild videos, single-frame-based HPE tasks have been the main focus. In our case, the movement of golfers is limited, and the motion blur, as well as the self-occlusion, cannot be ignored. Therefore, we propose a temporal-based HPE model to utilize the temporal information and improve the HPE accuracy.

2.2. Sports Analyses

Action and trajectory analyses Performance assessment is critically needed when players are doing training. In some sports events like swimming, table tennis[11], soccer[12] and golf, ability of performing complex athletic motions matters the most. Therefore, in-depth analyses and assessments of captured actions from the players are critical to improve their skills and become more competitive in the game. In addition, besides the player itself, target tracking (like soccer balls and basketballs) is also important. Object trajectory analyses can provide another way of assessing players' performance and their interactions with the targets.

Player tracking In some team sports like basketball[13] and soccer[1], in addition to the highlight performance of the ace players, what matters more is the proper organization and consistent cooperation between every player in the team. Tracking the moving trajectories of players in both teams through video streams allows the coach to assess the interaction between teammates, and it is also invaluable to develop effective tactics and game strategies.

3. METHOD

GolfPose is targeted at golf-playing scenarios. Its goal is to generate accurate 3D pose estimation from a monocular swing video taken from mobile devices. The 2D GolfPose first generates reliable keypoints on both players' body and golf club, which are then systematically converted to 3D poses for further analyses. In this paper, due to the page limitation, we mainly focus on the innovations and performance improvements made to the 2D architecture of GolfPose. Our model is constructed to address this problem: First, we build a CNN-based temporal 2D HPE model based on an existing image-based HPE framework. Since our input is a short clip of a video sequence instead of a single image, we are able to utilize temporal information to increase the accuracy of keypoint prediction. Then, we implement a line segment algorithm, a traditional computer vision technique, to fix inaccurate predictions on golf club keypoints generated from the 2D HPE model. The overall pipeline of our 2D GolfPose architecture is illustrated in Figure 2.2. In the following sections, we will introduce the main components of our 2D GolfPose in detail.

3.1. Problem Formulation

Let $S \in \mathbb{R}^{L \times H \times W \times C}$ be the input video of L RGB frames ($C = 3$) with $H \times W$ size. Our goal of HPE is to predict a set of 2D keypoints $J \in \mathbb{R}^{N \times 2}$ for every frame in the video sequence, where N denotes the number of keypoints, which is $N = 38$ in our dataset. Our approach is sequence-based, i.e., in every time step t, it operates on a short video clip: $V = \{F_{t-n}, ..., F_t, ..., F_{t+n}\}$ and outputs the HPE result for the center frame F_t.

3.2. 2D Temporal based Model

To address self-occlusion and motion blur problems, temporal information is a good solution since the occluded keypoints may be visible in adjacent frames, and the blurry

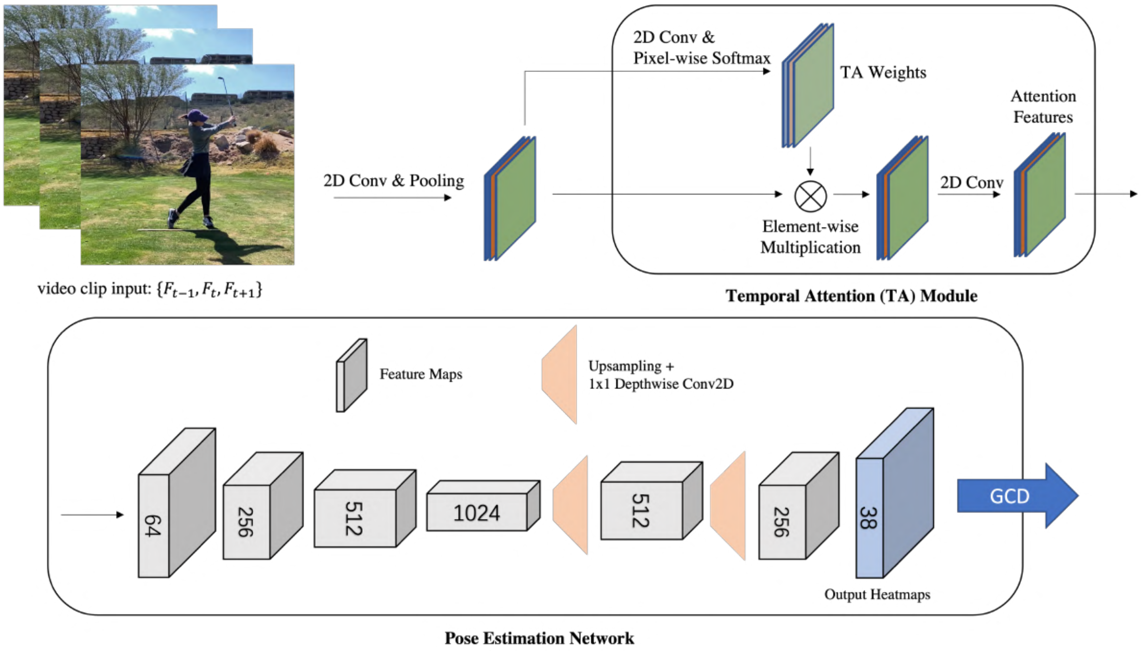

Fig. 2. Overall pipeline of our proposed GolfPose for human pose estimation. Three frames are used as the input, for example. First, there is a lightweight CNN-based 2D HPE model with temporal attention to utilize temporal information and increase the accuracy of occluded or fast-moving keypoint estimation. Then, golf club detection (GCD) is applied to fix inaccurate predictions on golf club keypoints generated from the 2D HPE model.

keypoints can be recovered based on multiple consecutive frames. Therefore, we propose a 2D temporal-based HPE model to generate accurate keypoints.

The temporal attention module is modified from [14], where the original temporal attention module is based on 3D convolutions, which incur high computational costs and cannot be performed real-time on mobile devices. In order to solve the latency issue, we replace the 3D convolutions with depthwise 2D convolutions and modify the architecture as well. As shown in Tables 1 and 4, this modification improves the inference speed significantly on mobile devices with only a small amount of accuracy degradation. The rest of the network is modified from LPN[15] by replacing the mobile incompatible operations.

To train the model, like other HPE model, we adopt the mean square error loss to minimize the difference between the predicted keypoint heatmaps H_{pred} and the ground truth keypoint heatmaps H_{GT}:

$$L_H = \|H_{pred} - H_{GT}\|_2^2 \tag{1}$$

3.3. Golf Club Detection

Fast movement of the golf club during the swing can cause issues in the 2D GolfPose model, such as missing detection of golf club due to motion blur. Moreover, golf club can move out of image boundary, which may also cause detection failure. These issues greatly hinder the performance of keypoints prediction, especially the points on the club hosel, which is crucial for analyzing golf swing. To address the problem of inaccurate hosel prediction, we resort to traditional computer vision techniques.

Algorithm 1 describes the proposed golf club detection (GCD) algorithm. Based on the prediction results from the 2D GolfPose keypoints model, we first determine the bounding box of the golf club and then apply the line segment detection (LSD) algorithm [16] to the cropped golf club region. The output of the LSD gives us lots of unconnected and short line segments pointing in various directions since LSD only depends on pixel information and can be easily influenced by other line-shape elements in the environment, like grass and ground.

We set the direction from the club top of the handle J_handle to the club middle hand J_md as the reference direction. With previously detected redundant and erroneous line segments and the reference direction, we can first remove those line segments which have large angles with respect to the reference direction and assign the vector directions of the remaining segments.

Next, we implement an iterative process to remove outliers that do not lie on the golf club and connect different vectors to form the whole golf club line. In every iteration step, we find a new vector in the search space to connect to current club line candidates and form a new line representing the golf club. The connected vector should have a consistent direction with the current club line with the minimum distance between these two vectors. The distance D between two vec-

Algorithm 1 Golf Club Detection

```
Input: Image $I_t$, Position of club mid hands $J_{md}$, top of handle $J_{handle}$, hosel $J_{hosel}$
Output: Updated hosel position $J_{hosel}$
  $\vec{D}_{init} \leftarrow J_{md} - J_{handle}$
  Compute bounding box covering whole golf club and crop out this region
  Apply Line Segment Detection from OpenCV and assign results as list of line segments, $L$
  for all $line$ in $L$ do
      if $\langle line, \vec{D}_{init} \rangle >= 15°$ then
          remove $line$
  Find $N$ line segments with the smallest distance to $J_{md}$
  Set searching space as all line segments that are not included in any of these $N$ line segments
  for all $line$ in $N$ line segments do
      while searching space $\neq \varnothing$ do
          $nline \leftarrow$ line segment with minimum distance $mdist$ to $line$
          if $mdist < thre$ then
              $line \leftarrow$ connect start of $line$ with end of $nline$.
              remove $nline$ in searching space
          else
              break
  $L_{club} \leftarrow$ longest $line$ from $N$ lines.
  if $\langle L_{club}, \vec{D}_{init} \rangle < 15°$ then
      $J_{hosel} \leftarrow$ projection of $J_{hosel}$ in the direction of $L_{club}$
  else
      $J_{hosel} \leftarrow$ end point of $L_{club}$
```

tors is the distance between the first one's end point and the second one's starting point.

4. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of our proposed GolfPose system. We mainly use the golf swing dataset we collected ourselves for benchmarking. Details of the dataset and training process as well as system performance will be discussed below:

4.1. Dataset and evaluation metrics

As far as we know, we are the first to apply deep learning based vision techniques to swing analyses in golf playing scenarios. In order to get reliable pose estimation results for golf swings, we use a subset of collected and annotated dataset for our performance evaluation. This subset dataset includes 120,000 images, where 100,000 images are used for training and 20,000 images for validation and testing which are not used in the training. Unlike traditional 2D human pose estimation datasets, e.g., COCO[17], our dataset is video based and is recorded with over 120fps to eliminate the motion blur.

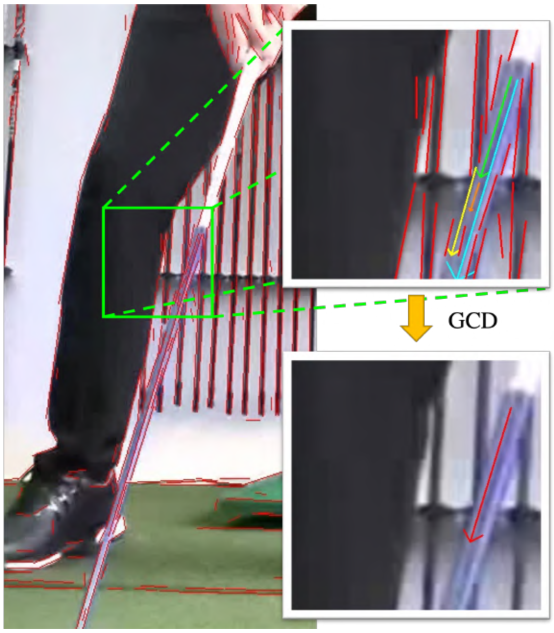

Fig. 3. Golf club detection (GCD). After the initial line segment detection, there are many noisy line segment detection results. The GCD first filters line segments with the help of J_{md} and J_{handle}, and then finds some potential starting line segments to form the golf club. As the figure shows, if the green segment is the starting segment and the yellow, orange and blue segments are candidates to be added to the search list, according to the distance between those segments, the orange segment is then closest to be connected the green segment, and they are merged to be the next starting segment.

There are in total 38 keypoints annotated in our dataset, as shown in the right image of Figure , including some keypoints on the golf club. Since video resolution and the size of recorded players in our dataset are relatively similar, we use 2D mean pixel error (MPE) as our evaluation metrics to evaluate the accuracy of the keypoint localization.

4.2. Training details

Since our GolfPose is designed for mobile devices, we implement the system using the publicly available TensorFlow framework and convert it to TensorFlow Lite model for mobile inference.

The pose estimation model is trained in an end-to-end manner. All parameters are initialized randomly from the zero-mean Gaussian distribution with $\sigma = 0.001$. We use Adam optimizer with a mini-batch size of 32 to update the parameters. The total number of training epochs is 150, and the initial learning rate is set to 0.001, reduced by a factor of 10 at the 90^{th} and 120^{th} epoch.

The detected and cropped human bounding boxes from the video frames are fixed to a certain aspect ratio (i.e., height: width = 4:3). The cropped bounding box is resized

Method	Input size	#Params↓	GFLOPs↓	MPE^A ↓	MPE^B ↓	MPE^C ↓
HRNet-W32[8]	256 × 192	28.5M	7.1	9.20	**8.48**	13.98
LPN[15]	256 × 192	2.9M	2.28	9.60	9.34	11.34
Ours (Conv3D)	256 × 192 × 3	2.8M	8.22	**9.08**	9.08	9.10
Ours (Conv2D)	256 × 192 × 3	3.2M	5.46	9.15	9.18	**8.95**
Ours (Depthwise Conv2D)	256 × 192 × 3	2.9M	3.42	9.16	9.15	9.21

Table 1. Experimental results on our golf swing test set. MPE^A stands for MPE of all keypoints. MPE^B for MPE of body keypoints. MPE^C for MPE of club keypoints

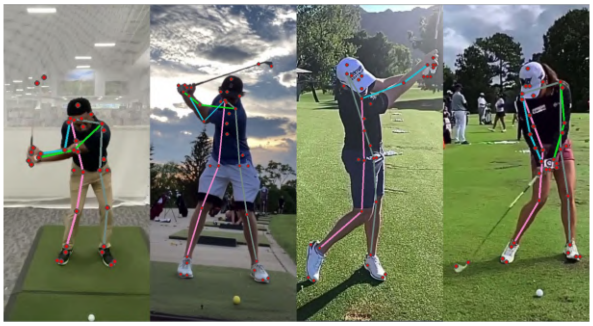

Fig. 4. Qualitative results on our dataset.

to 256 × 192, keeping the original aspect ratio and padding with the black background, and served as the input image. In addition to common data augmentation operations like random rotation, random scale, and flipping, we also add several additional augmentation operations, aiming at increasing the system robustness under specific conditions. For example, we randomly adjust the input image brightness to represent overexposed or underexposed environments. We also add random Gaussian noise and multi-frame averaging to the original image to simulate motion blur caused by camera shaking. Our training are performed on one NVIDIA 1080Ti GPU, and the training takes about 36 hours to complete.

4.3. Results

As shown in Table 1, we test our system performance on the 20,000 images in the validation and testing split. Compared to the HRNet[8] with the same input resolution, our GolfPose model can achieve better performance with much less number of parameters and fewer GFLOPs. Furthermore, for keypoints on the golf club, with the help of multi-frame temporal input and our temporal attention module, our performance is much better than the image-based state-of-the-art method, HRNet. Figure 4.2 shows some representative qualitative performance of the keypoints estimated by the proposed 2D GolfPose. In addition, our model is also favorable in terms of inference speed when deployed in mobile devices because of the smaller model size, and all the operations in the architecture can be accelerated by mobile GPUs.

Golf Club Detection (GCD) is incorporated to correct the failing cases from the HPE model and to improve golf club keypoint estimation accuracy. According to Table 2, GCD significantly decreases the golf club hosel's standard deviation of Pixel Error (Std. PE), from 33.89 to 23.19, which means GCD does correct the wrong estimation of the model output for failure cases.

Method	MPE^{hosel} ↓	Std. PE^{hosel} ↓
Model	11.93	33.89
GCD	**10.76**	**22.39**

Table 2. Performance improvement after GCD.

4.4. Ablation Study

We study the effect of each component in our methods, including the length of sequence input and different designs of the temporal attention module. All results are evaluated on our collected validation and testing set and with the same input size (256 x 192) and same training scheme.

Length of input sequence: Since our model is sequence-based instead of single image-based, we would like to explore the how input sequence length can affect the accuracy and inference speed of the model and eventually obtain a good trade-off between these two metrics. Table 3 shows the evaluation results in terms of accuracy and inference speed over input sequence lengths of 3, 5, 7 respectively. We can see that simply increasing the length is not a good choice, it can hurt both accuracy and inference speed.

Length	MPE^A ↓	Std. PE^A ↓	Inference Speed
3-frame	9.159	8.922	**50.3ms**
5-frame	**9.056**	**8.560**	50.7ms
7-frame	10.054	9.801	51.6ms

Table 3. Model performance with different input sequence length. Test on Samsung S20 Ultra.

Temporal attention module: In order to furthermore improve the inference speed of the temporal-based model, we also modify the original temporal attention module by replacing the 3D convolution with 2D convolution or depthwise 2D convolution. According to Table 1 and Table 4, the GFLOPs and inference time is significantly decreased with the modification while maintaining the similar performance, which shows that depthwise 2D convolution based temporal attention module is the best choice for mobile device inference, with a good balance between accuracy and inference speed.

Method	Inference Speed	GFLOPs
Single Frame	**27.6ms**	2.28
Conv3D	130.0ms	8.22
Conv2d	68.5ms	5.46
Depthwise Conv2D	**50.3ms**	3.42

Table 4. Inference speed per frame of different temporal-based models and the image-based model. Test on Samsung S20 Ultra.

5. CONCLUSION

In this paper, we propose a novel lightweight temporal-based 2D human pose estimation method, GolfPose, and a golf club detection method for further improving keypoint prediction accuracy on the golf club, which can be deployed on mobile devices for efficient and accurate golf swing analyses. The success of this pipeline is under the assumption that players are not moving,which is justified in golf swinging while not applicable to moving around players in other sports that require the human tracking mechanism to be added for temporal-based human pose estimation. In the future, we will work on migrating our method to other sports with the help of human tracking.

6. REFERENCES

[1] Lewis Bridgeman, Marco Volino, Jean-Yves Guillemaut, and Adrian Hilton, “Multi-person 3d pose estimation and tracking in sports,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[2] Erwin Wu, Takayuki Nozawa, Florian Perteneder, and Hideki Koike, “Vr alpine ski training augmentation using visual cues of leading skier,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 878–879.

[3] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R. Manmatha, and Mu Li, “A comprehensive study of deep video action recognition,” 2020.

[4] Yurui Ren, Ge Li, Shan Liu, and Thomas H. Li, “Deep spatial transformation for pose-guided person image generation and animation,” 2020.

[5] Jie Mei, Jenq-Neng Hwang, Suzanne Romain, Craig Rose, Braden Moore, and Kelsey Magrane, “Absolute 3d pose estimation and length measurement of severely deformed fish from monocular videos in longline fishing,” 2021.

[6] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, “Convolutional pose machines,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[8] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, “Deep high-resolution representation learning for human pose estimation,” in *CVPR*, 2019.

[9] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, “Openpose: Realtime multi-person 2d pose estimation using part affinity fields,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[10] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang, “Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5386–5395.

[11] Kaustubh Milind Kulkarni and Sucheth Shenoy, “Table tennis stroke recognition using two-dimensional human pose estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4576–4584.

[12] Anthony Cioppa, Adrien Deliege, Floriane Magera, Silvio Giancola, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck, “Camera calibration and player localization in soccernet-v2 and investigation of their representations for action spotting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4537–4546.

[13] Julian Quiroga, Henry Carrillo, Edisson Maldonado, John Ruiz, and Luis M Zapata, “As seen on tv: Automatic basketball video production using gaussian-based actionness and game states recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 894–895.

[14] Jiarui Cai, Yizhou Wang, Hung-Min Hsu, Haotian Zhang, and Jenq-Neng Hwang, “Dior: Distill observations to representations for multi-object tracking and segmentation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 520–529.

[15] Zhe Zhang, Jie Tang, and Gangshan Wu, “Simple and lightweight human pose estimation,” *arXiv preprint arXiv:1911.10346*, 2019.

[16] Jin Han Lee, Sehyung Lee, Guoxuan Zhang, Jongwoo Lim, Wan Kyun Chung, and Il Hong Suh, “Outdoor place recognition in urban environments using straight lines,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 5550–5557.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, “Microsoft coco: Common objects in context,” in *European conference on computer vision*. Springer, 2014, pp. 740–755.