

Article

Unified End-to-End YOLOv5-HR-TCM Framework for Automatic 2D/3D Human Pose Estimation for Real-Time Applications

Hung-Cuong Nguyen ¹ , Thi-Hao Nguyen ¹, Rafal Scherer ²  and Van-Hung Le ^{3,*}

¹ Faculty of Engineering Technology, Hung Vuong University, Viet Tri City 35100, Vietnam; cuongnh@hvu.edu.vn (H.-C.N.); haont@hvu.edu.vn (T.-H.N.)

² Department of Intelligent Computer Systems, Czestochowa University of Technology, 42-218 Czestochowa, Poland; rafal.scherer@pcz.pl

³ Faculty of Basic science, Tan Trao University, Tuyen Quang City 22000, Vietnam

* Correspondence: van-hung.le@mica.edu.vn; Tel.: +84-973-512-275

Abstract: Three-dimensional human pose estimation is widely applied in sports, robotics, and healthcare. In the past five years, the number of CNN-based studies for 3D human pose estimation has been numerous and has yielded impressive results. However, studies often focus only on improving the accuracy of the estimation results. In this paper, we propose a fast, unified end-to-end model for estimating 3D human pose, called YOLOv5-HR-TCM (YOLOv5-HRet-Temporal Convolution Model). Our proposed model is based on the 2D to 3D lifting approach for 3D human pose estimation while taking care of each step in the estimation process, such as person detection, 2D human pose estimation, and 3D human pose estimation. The proposed model is a combination of best practices at each stage. Our proposed model is evaluated on the Human 3.6M dataset and compared with other methods at each step. The method achieves high accuracy, not sacrificing processing speed. The estimated time of the whole process is 3.146 FPS on a low-end computer. In particular, we propose a sports scoring application based on the deviation angle between the estimated 3D human posture and the standard (reference) origin. The average deviation angle evaluated on the Human 3.6M dataset (Protocol #1–Pro #1) is 8.2 degrees.

Keywords: YOLOv5; HR; 2D/3D human pose estimation; Convolutional Neural Network; Temporal Convolution Model; Pose-based Sports Application



Citation: Nguyen, H.-C.; Nguyen, T.-H.; Scherer, R.; Le, V.-H. Unified End-to-End YOLOv5-HR-TCM Framework for Automatic 2D/3D Human Pose Estimation for Real-Time Applications. *Sensors* **2022**, *22*, 5419. <https://doi.org/10.3390/s22145419>

Academic Editor: Biswanath Samanta

Received: 16 June 2022

Accepted: 16 July 2022

Published: 20 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Human pose estimation is regarded as one of the most interesting research areas in computer vision. It is applied to many fields such as healthcare, sports [1], activity recognition [2], motion capture and augmented reality, training robots, motion tracking for consoles [3], etc. Barla [4] presented seven applications of human pose estimation. In particular, sports have begun to use the results of human pose estimation in practice and competition [5]. Some applications are illustrated in Figure 1.

Human pose estimation is defined as the process of localizing human joints (also known as keypoints—elbows, wrists, etc.) in images or videos. There are two study directions for estimating human pose from images/videos: 2D human pose estimation and 3D human pose estimation. Two-dimensional human pose estimation is an intermediate result for the 3D human pose estimation. Based on the approach of Zhou et al. [6], 3D human pose estimation results are highly dependent on 2D human pose estimation. In the last five years, this task has been gaining much attention, and the research on 3D human pose estimation helps to build intuitive and important applications in robotics, for example training a robot to perform a certain task according to human activity. So far, many kinds of research have focused on improving the accuracy of human pose estimation in the

2D/3D space. Recently, there has been the research by Mehta et al. [7], who proposed an application to estimate 3D human pose in real-time (30 frames per second on a six-core Xeon CPU, 3.8 GHz, and a single Titan X (Pascal architecture) GPU) from RGB images. However, the accuracy of this 3D human pose estimation is not particularly high. To be able to apply the results of 3D human pose estimation in sports analysis, in scoring sports performances, a high-precision system of 3D human pose estimation and time is required. The fast computational time follows the speed of movement and performance of movements and tests in sports. Especially to reduce the processing time, the system is capable of performing 3D human pose estimation based on RGB images, which is common and easily collected data. Especially, the system is initially applied in a gym fitness center or to some non-competitive sports such as weightlifters, weightlifting competitions, single skating, and gymnastics. This means that in the setting, there is only one performer. Therefore, in this paper, we propose a unified end-to-end model for estimating 3D human pose from the RGB image of a monocular camera. The initial data for building the system are the RGB image and 3D human pose annotation. System run and test data include color images only. To obtain a model with high accuracy and fast computation time in 3D human pose estimation, the modeling steps must have high accuracy and a fast processing time: human detection, 2D human pose estimation, and 3D human pose estimation. The steps to build the system are described below.

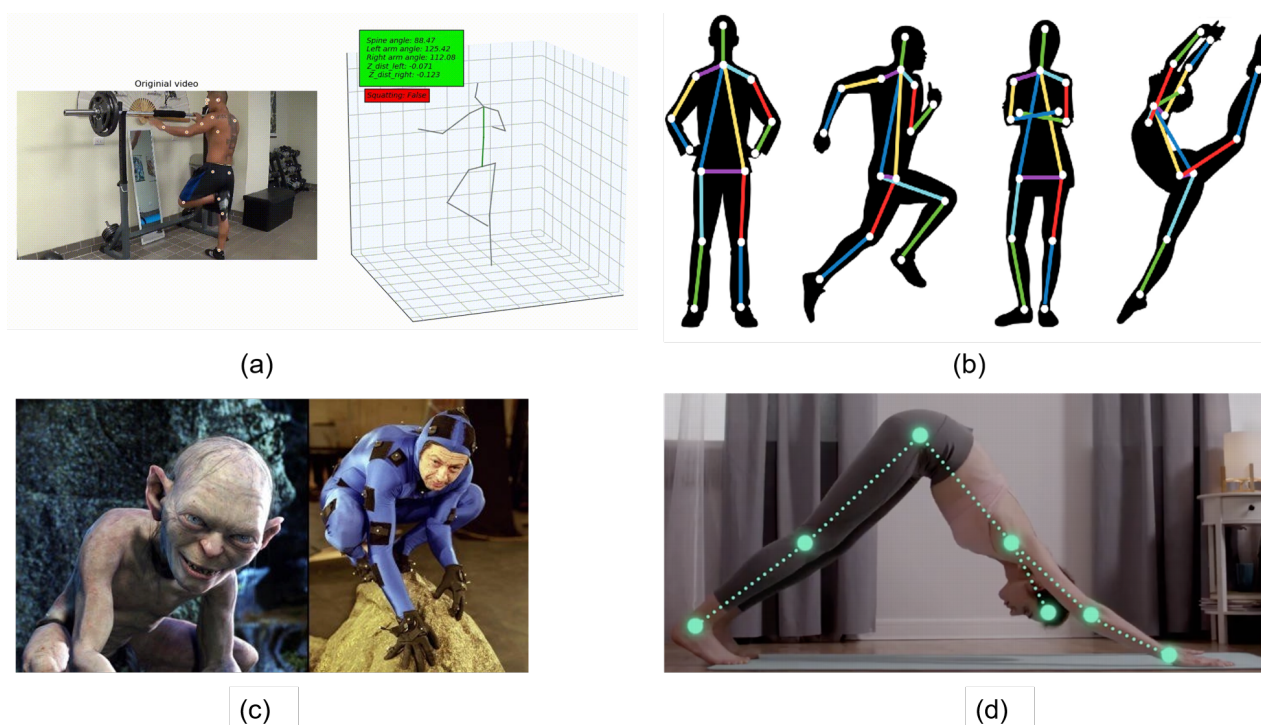


Figure 1. Illustrating of using human posture estimation in weightlifting practice (a) [5], healthcare, sports (b,d) [1], and robotics (c) [3].

In this paper, we are interested in both 2D human pose estimation and 3D human pose estimation problems from monocular RGB images or videos. Two deep-learning-based approaches can be used to estimate 2D human poses. The first is the regression method, which applies a deep neural network to learn a mapping from the input image to body joints or parameters of human body models to predict the keypoints on the human (keypoint-based). The second is the body part detection methods to predict the approximate locations of body parts and joints (bodypart-based). Deep learning (DL) networks have achieved remarkable results in estimation tasks. However, they still face many challenges such as heavy occlusion, a partially visible human body, and low image resolution. Sudharshan [8] presented some typical studies [9–16] on estimating 2D human posture in images or videos.

In Table 2 of [12], the authors showed the results obtained by the high-resolution network (HR) comparing the above methods for 2D human pose estimation on the COCO [17] dataset. HR is the most accurate across different configurations. Li et al. [18] used HR as a backbone for 2D human pose estimation on cropped human images of the Human 3.6M dataset [19]. As the Human 3.6M dataset contains 548,819 images of Pro #1 for testing, manually marking the data area of the person in the image would take a long time. This difficulty is very dependent on the person conducting the cropping and HR's estimated data area in the human data region, without regard for other regions in the image. This problem of detecting people in the image is considered to be 100% accurate. However, when applied to real problems, no approach is appropriate.

Three-dimensional human pose estimation is currently of great research interest in the field of computer vision. Recently, there have been many surveys on this issue [20–23]. According to the surveys, 3D human pose estimation from monocular RGB images or video is based on three methods: direct estimation method, 2D to 3D lifting method, and human mesh recovery method. Current studies on 3D human pose estimation have very impressive results. In the study of Li et al. [18], the average error of the 3D human pose estimation on the Human 3.6M dataset was 49.7 mm (Pro #1) and 37.7 mm (Pro #2). In [24], Chen et al. proposed a study with a mean error on the Human 3.6M dataset of 46.3 mm (Pro #1). However, like many other studies, this study is not interested in the 3D human pose estimation processing time.

We propose a unified end-to-end model, called YOLOv5-HR-TCM, for the real-time estimation of 3D human pose, as shown in Figure 2. The model that we propose is fully automatic from end-to-end in estimating 3D human pose from the monocular RGB images or video. The model we propose includes three stages: human detection, 2D human pose estimation, and 3D human pose estimation. In the first stage, we combine the advantages of processing speed and a contextual constraint (CC) into a pre-trained YOLOv5 network [25,26] for detecting a person in a crowd to detect the human in the RGB image. In the second stage, we use a pre-trained model of HR for estimating 2D keypoints/2D human pose in the RGB image. The third stage is the 3D human pose estimation by the temporal convolutions model (TCM) [27]. Unlike previous studies on estimating 3D human pose from a single-camera RGB image, our approach is a combination of CNNs that have the best performance currently in the tasks of person detection, 2D human pose estimation, and 3D human pose estimation. Finally, we apply a simple computational technique to compute the angle between the bone ground-truth and the estimated bone for scoring in a sports application. Our framework is fully automated and executes in real-time on a PC with a low configuration, with takes the input as the monocular RGB images or video and the ground-truth of the 3D human pose. The outputs are the estimated 3D human pose in 3D space and the average deviated angle of the bones.

The main contributions of the paper are as follows:

- We propose a unified end-to-end framework for automatic 3D human pose estimation. The framework is a combination of high-performance CNNs to perform sequential tasks: human detection, 2D human pose estimation, and 3D human pose estimation.
- We embedded efficient contextual constraints (CCs) into YOLOv5 for human detection and HR for 2D keypoint estimation/2D human pose estimation in images or video, called YOLOv5 + CC + HR combined. We also evaluated the results in detail at this stage on the Human 3.6M dataset.
- We applied the TCM and semi-supervised training method in our framework using the 2D human pose estimation results in the fine-tuning 3D human pose estimation model on the Human 3.6M dataset. The 3D human pose estimation results were also evaluated and compared with the baseline methods.
- We combined and integrated the proposed framework into a practical application for computing the angle of deviation of human poses in 3D space. This was applied for assessment and scoring in artistic gymnastics and training and dance assessment. Moreover, it operates in real-time on a PC with a low configuration.

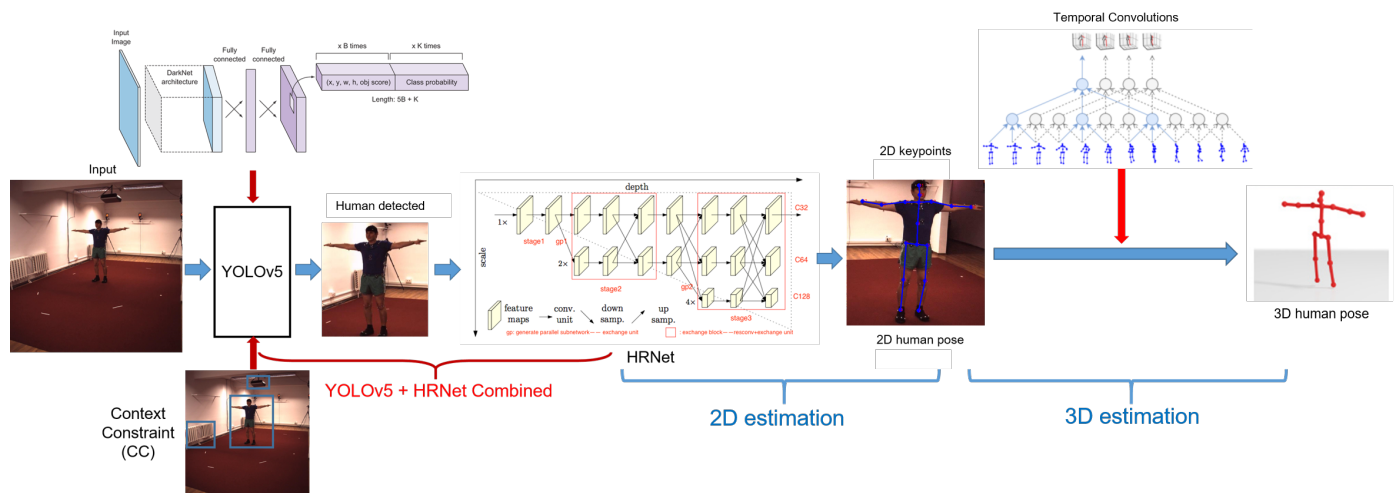


Figure 2. The unified end-to-end YOLOv5-HR-TCM framework for 3D human pose estimation from RGB images taken by a monocular camera.

The paper is organized as follows. Section 1 introduces human detection, 2D keypoint estimation, and 3D human pose estimation in the image and the applications. Section 2 discusses related works of the methods, the results of 2D keypoint estimation and 3D human pose estimation, and applications. Section 3 presents a combination of YOLOv5, context constraints, HR, TCM, and semi-supervised training methods for 3D keypoint estimation/3D human pose estimation. Section 4 introduces and presents the Human 3.6M dataset, evaluation metrics, implementations, results, discussions on 2D keypoint estimation, and 3D human pose estimation. Section 5 presents the application of computing the deviated angles on the 3D human skeleton. Section 6 concludes the paper and proposes some future work.

2. Related Works

Estimating human posture in 2D and 3D is of great research interest in computer vision. The results are applicable in many fields, especially in sports. In this paper, we are interested in estimating the human pose problem in 2D and 3D. The human pose estimation in 2D space is the human pose estimation in the color image obtained from monocular RGB images and videos. Three-dimensional human pose estimation determines the position of joints on the human skeleton, with each joint having coordinates (x, y, z) .

Estimating the 2D human pose of a single person can be divided into direct regression methods and heat-map-based methods [22,28]. Direct regression methods are the end-to-end use of a CNN to learn a mapping from the input image to estimate the joints/2D keypoints or parameters of human skeleton models. The heat-map-based methods predict the locations of body parts and joints/2D keypoints from the heat map probability [28]. In addition, the two survey studies [22,28] detailed the results of 2D human pose estimation from a single-view camera. Two-dimensional multi-person pose estimation is performed by top-down methods or bottom-up methods. The top-down methods detect and classify each human in the image, constrain them by bounding boxes, then estimate the pose of each detected person. The bottom-up method includes two main steps: extracting local features by predicting skeleton joint candidates and skeleton joint candidate assembling for individual bodies. All four methods of 2D human pose estimation are illustrated in Figures 3 and 4 of [28].

In this paper, we present seven outstanding studies on estimating 2D human pose from RGB images or videos. Toshev et al. [16] proposed CNN-based regression (DeepPose) to regress the skeleton joints/2D keypoints. DeepPose uses a cascade of such regressors to refine the pose estimates and obtain better estimates from the estimated candidates. DeepPose includes seven layers (five convolutional layers and two fully connected layers), as shown in Figure 2 of [16]. DeepPose's best results on the percentage of correct parts

(PCP) at 0.5 on LSP are 61%. Thompson et al. [9] proposed a new CNN architecture with multi-resolution that uses a sliding window detector to produce a coarse heat map output. The model includes the heat-map-based parts model for coarse localization, a module to obtain and crop the convolution features at the (x, y) location for each joint/keypoint prediction, and fine-tuning model prediction. The loss function used in training is the mean-squared error (MSE) distance. The best results on PCKh@0.5 of the MPII dataset [29] are 82% with all joints of the human pose. Wei et al. [10] proposed convolutional pose machines (CPM); this CNN is a multi-stage architecture to be end-to-end trained for predicting joints/2D keypoints on heat maps. Stage 1 computes image features, and Stage 2 and up make the actual prediction based on the heat maps. The result of a previous stage is the predictive input for the next stage. The best results on PCKh@0.5 of the MPII dataset are 87.95%, and on the ankle (the most challenging part), results on PCKh@0.5 are 78.28%. The best result on PCKh@0.5 of LSP is 84.32%. Carreira et al. [11] proposed a feedforward architecture called iterative error feedback (IEF). This architecture can learn rich representations from the hierarchical feature extractors of both input and output spaces, by using the top-down feedback strategy. That is, after each training step, the error value of the feature set will be the feedback. The input of each layer is $x_t = I + g(y_t - 1)$, where I is the image and $y_t - 1$ is the output of the previous layer. The best results on PCKh@0.5 of the MPII dataset are 81.3%. Newell et al. [14] proposed a CNN called stacked hourglass network (SHN). The model consists of several hourglass (HG) modules arranged in series. Each HG processes input information from high to low resolutions and then from low to high resolutions. Thus, a single HG is a sort of full convolutional network. Such stacked HGs are for the improved inference across scales.

This scheme takes advantage of the characteristics and relationships of the human body parts. The low resolution will learn the position of the joints of the limbs; the higher resolution will learn the position of the limb and the relationship between the parts. The estimated result of the SHN network is much higher than that of the previously proposed networks, the average results on PCKh@0.5 of the MPII dataset being 90.9%. Xiao et al. [13] proposed a simple and effective strategy, called simple baselines (SB) for 2D human pose estimation and tracking. This network is a combination of a ResNet and several transposed convolution layers. The HG network uses upscaling (low resolutions to high resolutions) to increase the feature map resolution and set the convolutional parameters in the next blocks. The SB forms skip connections for each resolution. The mean results on the *mAP* of the COCO dataset are 73.7% with ResNet-152, and the input size is 384×288 . Sun et al. [12] proposed the high-resolution network (HR) for predicting the 2D keypoints/joints of the human body. Unlike SHN, HR performs prediction based on a high resolution to low resolution to high resolution representation in parallel and connects the multiple resolutions. HR does not perform any heat map supervision. The mean results on the *mAP* of the COCO dataset are 77.0%.

Three-dimensional human pose estimation is usually performed based on two approaches [30]: the first is using DL networks, and the second is using the transformers (TransS) method.

Regarding methods based on DL, estimating the 3D human pose of a person from monocular RGB images/videos can be performed based on three methods [22], illustrated in Figure 3: the first is using the CNNs end-to-end to estimate the 3D human pose (M1 in Figure 3); the second is to use the CNNs to lift the 2D human pose to the 3D human pose (M2 in Figure 3); the third is to use the CNN to regress the 3D human pose from the 2D human pose (M3 in Figure 3). The taxonomy of 3D human pose estimation is shown in Figure 4.

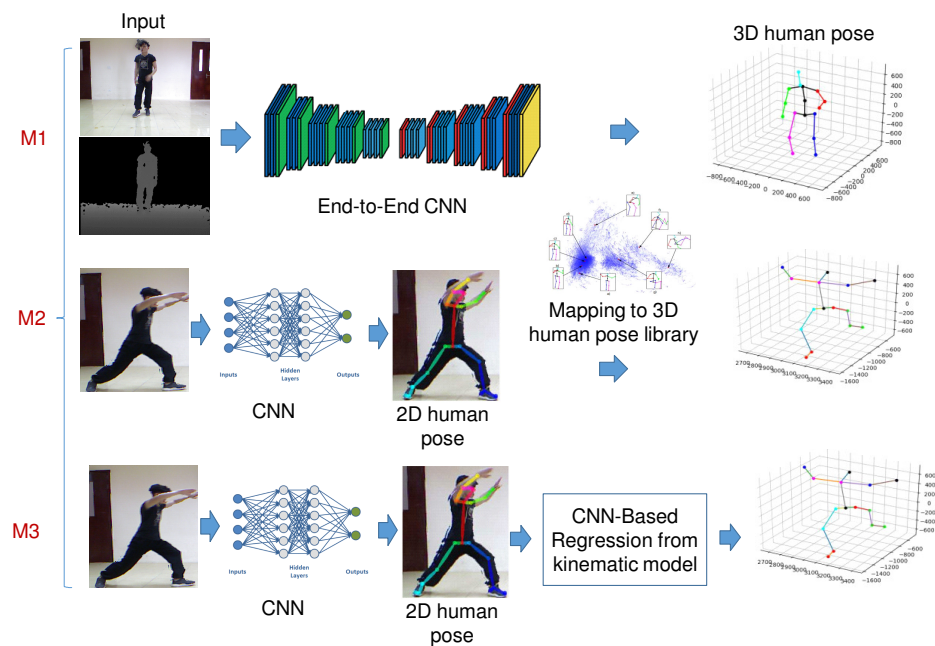


Figure 3. Illustrating three methods to estimate 3D human pose that use CNN-based from monocular RGB images/video.

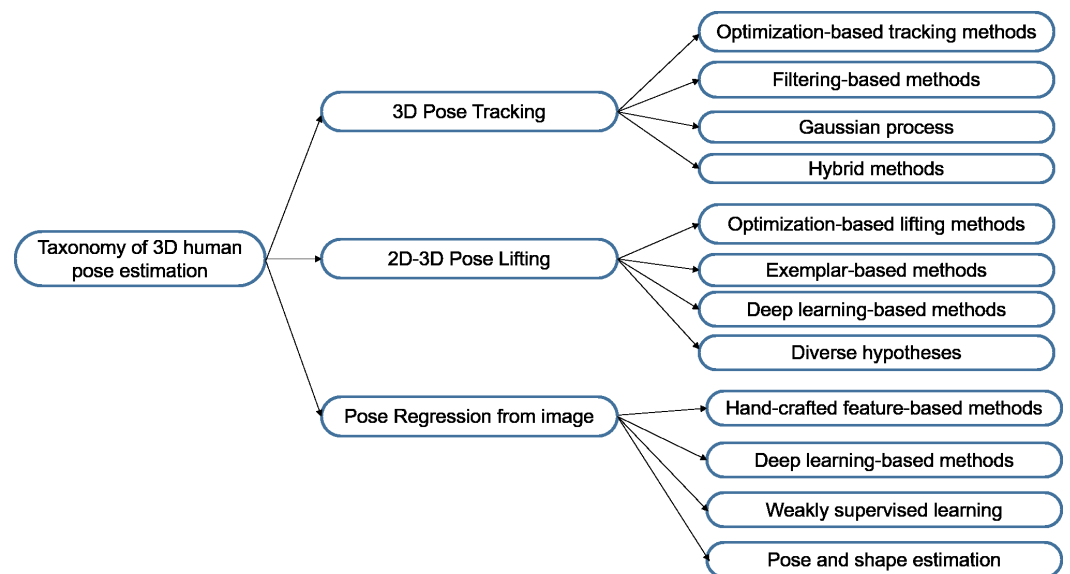


Figure 4. The taxonomy of 3D human pose estimation methods.

The results of 3D human pose estimation based on the two methods DL and TranS on the 3D human pose annotation of Human 3.6M is shown in Table 1.

Table 1. Statistics of the result of studies based on the MPJPE(mm) measurement on the ground-truth of the Human3.6M dataset for 3D human pose estimation.

Method	Model	Results of Mean per Joint Position Error (MPJPE) (mm)
Rhodin et al. [31]	DL	Protocol #1: 131.7
Tome et al. [32]	DL	Protocol #1: 88.39; Protocol #2: 70.4 Protocol #3: 79.6
Mehta et al. [7]	DL	ResNet 100: 82.5; ResNet 50: 80.5
Zhou et al. [6]	DL	Protocol #1: 64.9
Wang et al. [33]	DL	Protocol #1: 63.67
Veges et al. [34]	DL	Protocol #1: 61.1
Fang et al. [35]	DL	Protocol #1: 60.4; Protocol #2: 45.7 Protocol #3: 72.8
Omran et al. [36]	DL	Protocol #1: 59.9
Zhao et al. [37]	DL	Protocol #1: 57.6
Nibali et al. [38]	DL	Protocol #1: 57.0; Protocol #2: 40.4
Moon et al. [39]	DL	Protocol #1: 53.3; Protocol #2: 34.0
Lee et al. [40]	DL	Protocol #1: 52.8; Protocol #2: 43.4
Li and Lee et al. [41]	DL	Protocol #1: 52.7; Protocol #2: 42.6
Pavlakos et al. [42]	DL	Protocol #1: 51.9
Kocabas et al. [43]	DL	Protocol #1: 51.83
Bastian et al. [44]	DL	Protocol #1: 50.9
Tekin et al. [45]	DL	Protocol #1: 50.12
Karim et al. [46]	DL	Protocol 1: 49.9
Sun et al. [47]	DL	Protocol #1: 49.6
Rhodin et al. [48]	TranS	Protocol #1: 46.8
Chen et al. [24]	DL	Protocol #1: 46.3; Protocol #2: 37.7 Protocol #3: 50.3
Martinez et al. [49]	DL	Protocol #1: 45.5
Li et al. [50]	TranS	Protocol #1: 43.7; Protocol #2: 35.2
Zheng et al. [51]	TranS	Protocol #1: 44.3; Protocol #2: 34.6
Hossain et al. [52]	DL	Protocol #1: 39.2
Wang et al. [53]	DL	Protocol#1: 37.6
Pavlo et al. [27]	TranS	Protocol #1: 37.2; Protocol #2: 27.2
Pavlo et al. [54]	DL	Protocol #2: 36.0
Zhao et al. [55]	TranS	Protocol #1: 35.2
Zhao et al. [56]	TranS	Protocol #1: 35.2

The tree-dimensional HPE category has also received much research attention in the past decade. Wang et al. [20] conducted a full survey of 3D human pose estimation approaches, evaluation datasets, metrics, results, and applications. In this paper, we are only interested in 3D human pose estimation studies from monocular RGB images and videos. According to Song et al.'s study [57], the problem of the 3D human pose from

monocular RGB images and videos generally is solved by two families of methods: direct 3D human pose estimation and 2D to 3D human pose lifting. However, the paper of Wang et al. [20] solved the problem of estimating 3D human pose from monocular RGB images and videos by three methods: direct 3D human pose estimation, 2D to 3D human pose lifting methods, and SMPL-based methods. Direct 3D human pose estimation is performed by designing an end-to-end CNN to predict the 3D coordinates of the joints of the 3D human pose from the images. This method includes two classes: detection-based methods and regression-based methods. Here, we introduce some typical studies for 3D human pose estimation. Pavlakos et al. [42] proposed a CNN for the end-to-end learning paradigm, including two works: a convolutional network (ConvNet) to predict the 2D joint location and a subsequent optimization step to recover the 3D coordinate joints of the 3D human pose. The mean per joint error (MPJE) (mm) on the Human 3.6M dataset was 51.9 mm and on the HumanEva-I dataset was 24.3 mm. Chen et al. [58] proposed a method based on using a CNN for 2D human pose estimation and 2D human pose matching with a 3D human pose library. The MPJE on the Human 3.6M dataset (protocol #1) was 69.05 mm.

In the category of the transformer methods, Zheng et al. [51] recently proposed the PoseFormer method. The authors designed a spatial-temporal transformer structure to follow the 3D pose of the person and then modeled the human pose and the relationship between joints within a frame and between frames. This method has the lowest average estimation error ever: the MPJE on the Human 3.6M dataset was 44.3 mm (Protocol #1) and 34.6 mm (Protocol #2). The best 3D human pose estimation rate was 320 fps with the input 2D human pose detected on a single GeForce GTX 2080 Ti GPU. Although the estimation accuracy is very high in the 3D human pose estimation process, this approach only focuses on estimating 3D human posture, but does not pay attention to the accuracy and processing time of the whole 3D human pose estimation process.

The applications of human pose estimation include some areas such as activity recognition, motion capture and augmented reality, training robots, and motion tracking for consoles [3,57]. Stenum et al. [1] developed an application that evaluates human body performance over the lifespan based on human pose estimation. At the same time, the authors also analyzed the challenges and limitations of human-posture-based applications as the problems of hidden body parts, limited training data, limited capture errors, limited positional errors, and limited recording devices. Badiola et al. [59] surveyed the number of studies on posture estimation and its applications. This provides an overview of this area of research in computer vision.

3. The Unified End-to-End YOLOv5-HR-TCM Framework

In papers [18,27,37,43,49,60,61], particular emphasis was placed solely on improving the 2D to 3D lifting process, and the 2D keypoint estimation process only uses 2D keypoint detectors such as ResNet, Mask RCNN, SHN, etc. Our study is interested in the results of all of the steps of the 3D human pose estimation process. We present the steps as follows.

3.1. Human Detection

Detecting humans in images using CNN has been studied extensively and has achieved impressive results. Many CNNs such as R-FCN [62], Faster RCNN [63], SSD [64], YOLO [65–68], and Faster RCNN [63] are presented and compared in Jonathan's study [69]. An interesting model is Faster RCNN, which is an improvement of Fast RCNN [70]; it also integrates the region recommendation algorithm into the CNN model. Faster RCNN is based on two main ideas: building a single model consisting of a region proposal network (RPN) and Fast RCNN with a shared CNN. Inheriting Faster RCNN, He et al. [71] introduced the Mask RCNN based on Faster RCNN as the backbone for detecting and segmenting people in images. It achieves high accuracy, but the processing speed of Mask RCNN is relatively slow. To meet the requirement of fast computational time, YOLO appeared. YOLO is a CNN network with an average accuracy and very fast processing speed, up to 91 fps. Since the input is the input image, YOLO uses some simple steps of network convolution,

pooling, and fully connected layers to obtain the output. This architecture can be optimized to run on a GPU with a single forward pass, and thus achieves very high speeds. The main idea of YOLOv1 [65] is to divide the image into a grid cell with a size (7×7) . For each grid cell, the model will make predictions for a bounding box (B) of humans. Each box B includes five parameters (the coordinates of the center of the human (x, y) , width (w) of the human, the height of the human (h), and the confidence (cof_h) of the human prediction. Given the grid cells in the other (7×7) grid, the model also predicts the probability of each class of people. Confidence cof_h is defined by Equation (1):

$$cof_h = P(h) * IOU_{ground-truth}^{prediction} \quad (1)$$

where $P(h)$ is the probability that there is a human in the ce and $IOU_{ground-truth}^{prediction}$ is the intersection over union of the prediction region and the ground truth.

YOLOv1 [65] imposes spatial constraints on bounding boxes: each grid cell can predict only very few bounding boxes and only one class. During training, the loss function does not have a separate evaluation between the error of the small bounding box versus the error of the large bounding box.

To improve the disadvantages of YOLOv2, YOLOv2 and YOLO 9000 have come up with some strategies: batch normalization, using the anchor box architecture to make predictions, direct location prediction, adding fine-grained features, multi-scale training, and a light-weight backbone. YOLOv3 [67] has a similar architecture to YOLOv2, but it also brings some improvements: using logistic regression to predict the confidence of the bounding box; using Darknet-53 as the backbone; using the feature pyramid network (FPN) architecture to make predictions from various scales of feature maps; adding associations between prediction classes.

The object detection challenge is now more accessible to those who do not have powerful computer resources thanks to the architecture of YOLOv4 [68]. Using YOLOv4, we can train an object detection network with extremely high accuracy using only a 1080ti or 2080ti GPU. To bring computer vision applications into practice in the future, current networks will need to be re-optimized to tolerate weak computing resources or develop high parallelism on servers.

In this paper, we use a pre-trained model trained on the COCO dataset of YOLOv5 [26] for head and human detection in a crowd and a context constraint to obtain the bounding box of the detected human in the image. When using YOLOv5 to detect people in the image of the Human 3.6M dataset, many other objects are mistakenly detected as persons. In the image of the Human 3.6M dataset, the person has the largest bounding box in the image. Therefore, we propose that the bounding box of the person is the bounding box with the highest height among the bounding boxes detected and marked as the person.

We compared the proposed method with some studies on human detection (e.g. Mask RCNN, VGG, SSD, Mobilenet) in images combined with constraints (CC). The results are shown in Table 2.

People detection results are near 100%, and the processing time is 55 fps on our PC. This is a very impressive result; the output of this step is the bounding box of the person detected in the image.

Table 2. The results of human detection on the Human 3.6M dataset (Pro #1) evaluated on the CNNs.

Measurement/Methods	Number Testing Samples	Number Detected	AP_{50} (%)	AP_{55} (%)	AP_{60} (%)	AP_{65} (%)	AP_{70} (%)	Processing Time (fps)
YOLOv5 [26] + CC	548,819	548,346 (99.91%)	99.78	99.38	98.42	97.07	94.16	55
Mask RCNN [71,72] + CC	548,819	548,819 (100%)	97.17	96.93	96.61	96.12	95.51	2
MobilenetV1 SSD [73,74] + CC	548,819	507,991 (92.56%)	96.87	95.66	93.59	89.52	81.38	10
VGG SSD [74] + CC	548,819	536,496 (97.75%)	99.14	98.60	97.66	95.99	92.81	12
Mobilenet SSD [75] + CC	548,819	548,801 (99.99%)	77.04	75.93	73.43	68.34	59.99	4.34

3.2. Two-Dimensional Human Pose and 2D Keypoint Estimation

For human pose estimation and 2D keypoint estimation of people, one can use backbones such as ResNet [76], stacked hourglass networks (SHNs) [14], or some studies such as Openpose [77], 2D pose estimation using part affinity fields [78], convolutional pose machines (CPM) [10], cascaded pyramid network (CPN) [79], Simple Baselines [13], or DeeperCut [80]. The high-to-low and low-to-high frameworks performed with CNNs are stacked hourglass networks [14] (Figure 5a), cascaded pyramid network (CPN) [79] (Figure 5b), simple baselines [13] (Figure 5c), and DeeperCut [80] (Figure 5d), respectively, for estimating the human pose in the image.

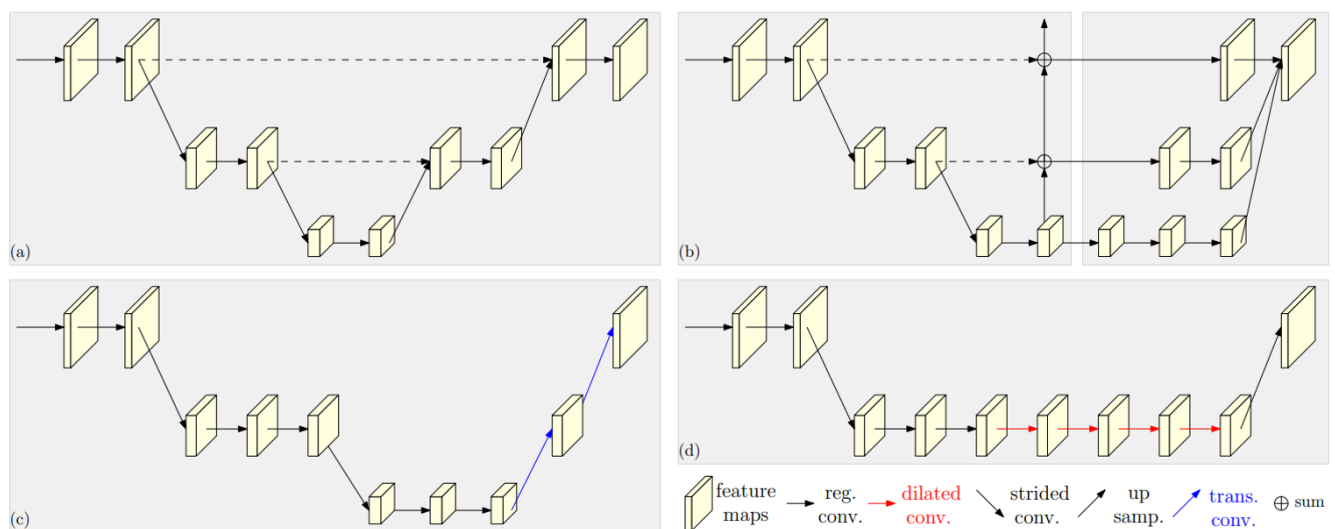
**Figure 5.** Illustration of the high-to-low and low-to-high processes [12] for 2D human pose and keypoint estimation of [13,14,79].

Figure 5 also shows that the high-to-low process of CNNs is sequential. While HR is presented in [12], it comes from the fact that when high-to-low convolutions are connected, the classification results at the region-level and pixel-level are low because this leads to enrichment of low-resolution representations, which means deterioration of high-resolution representations. HR implements parallel connections at the high-to-low resolution convolutions, which continuously strengthen multi-scale fusions across parallel convolutions of high-resolution representations, as illustrated in Figure 6. In particular, HR does not perform intermediate heat map supervision. Therefore, the accuracy of the keypoint detection and the computation time of HR is better than previous CNNs.

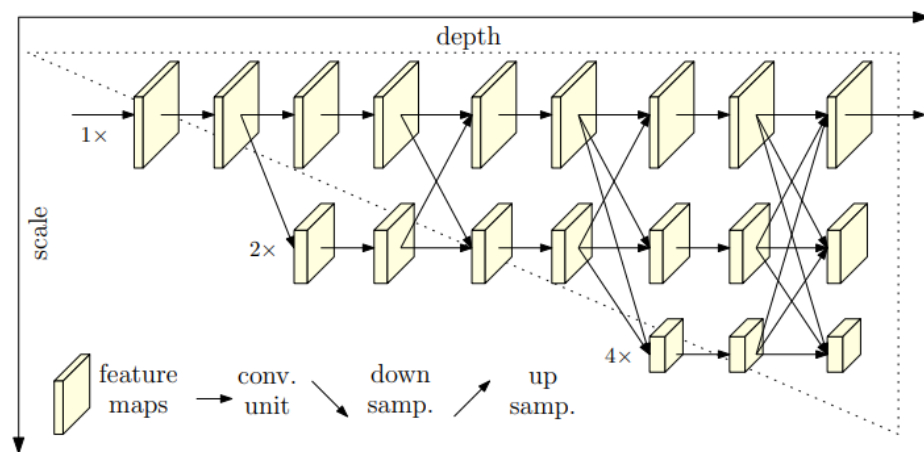


Figure 6. Illustration of HR architecture [12].

The aims of HR is to locate keypoints of the human pose in the image based on heat maps; the training model estimation process is the process of determining the value of the mean-squared error between the predicted heat maps and the ground-truth heat maps. The high-to-low network of HR includes four stages (HR_{sr} ; $s = 1 \dots 4$ is the stage number; $r = 1 \dots 4$ is the resolution at the s^{th} stage, its resolution is $\frac{1}{2^{r-1}}$ of the resolution of the first subnetwork), and the parallel processing of the subnetworks is shown as follows:

$$\begin{array}{ccccccc}
 HR_{11} & \rightarrow & HR_{21} & \rightarrow & HR_{31} & \rightarrow & HR_{41} \\
 & \searrow & & & & & \\
 & & HR_{22} & \rightarrow & HR_{32} & \rightarrow & HR_{42} \\
 & & & \searrow & & & \\
 & & & & HR_{33} & \rightarrow & HR_{43} \\
 & & & & & \searrow & \\
 & & & & & & HR_{44}
 \end{array}$$

HR performs exchanging the information across the parallel multi-resolution subnetworks by repeating multi-scale fusions, as illustrated in Figure 3 and Formula 3 of [12].

The results of the accuracy of the 2D human pose/2D keypoint estimation on the COCO and MPII datasets are shown in Tables 3 and 4, respectively. HR's results are the most accurate.

Based on the results presented in Table 1 of the paper by Li et al. [18], the 2D keypoint estimation results are very good, from 4.4 to 5.4 pixels on cropped human images using HR. In this paper, we propose the method of using a person detector in the image and then using the person detection results for 2D keypoint/2D pose estimation, as illustrated in Figure 2. Our approach, called YOLOv5 + HR Combined, combines the pre-trained human detection model of YOLOv5 on the CrowdHuman dataset and HR.

Table 3. Results of single-person keypoint detection [12] on the test set of the COCO dataset.

Method	Backbone-Size of Input	OSK (Object Keypoint Similarity) (%)		
		AP^{50}	AP^{75}	AP
OpenPose [77]	VGG-19	84.9	67.5	61.8
Mask-RCNN [71]	Faster RCNN	87.3	68.7	63.1
CPN [79]	ResNet-Inception— 384×288	91.4	80.0	72.1
Simple Baseline [13]	ResNet-152— 384×288	91.9	81.1	73.7
HR-W32 [12]	HR-W32— 384×288	92.5	82.8	74.9
HR-W48 [12]	HR-W48— 384×288	92.5	83.3	75.5
HR-W48 + extra data [12]	HR-W48— 384×288	92.7	84.5	77.0

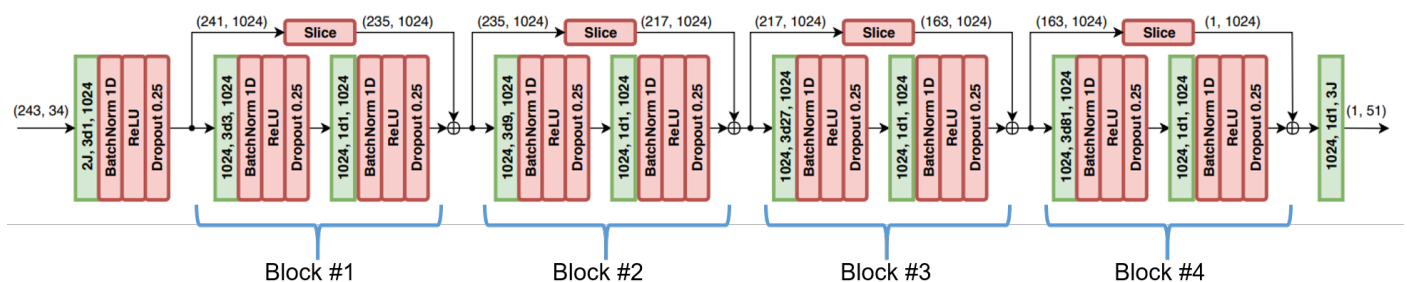
Table 4. Results of single-person keypoint detection [12] on the test set of the MPII datasets.

Method	Average of PCK@0.5(%)
DeeperCut [80]	88.5
SHNs [14]	90.9
Hourglass Residual Units (HRUs) [81]	91.5
Generative Adversarial Net.-SHNs [82]	91.8
Adversarial PoseNet [58]	91.9
Pyramid Residual Module (PRMs) [83]	92.0
Multi-scale structure-aware CNN [84]	92.1
Stacked U-Nets [85]	92.3
SimpleBaseline [13]	91.5
HR-W32 [12]	92.3

3.3. Three-Dimensional Human Pose Estimation From Estimated 2D Human Poses

As presented in the works of Chen et al. [28] and Zheng et al. [86], single-person 3D HPE is based on two main methods: using the CNNs to estimate directly from the images and using the CNNs to estimate from 2D human pose/2D keypoint data (2D to 3D lifting). We performed a small survey on 3D human posture estimation methods in the Human 3.6M database, and the statistical results are in Table 1. Currently, the results of the transformer (TranS) models show that the 2D to 3D lifting method obtains results better than the results of CNNs models on the Human 3.6M dataset, as shown in Table 1 of [27] and Table 1. Therefore, we chose the approach of using the TranS method for estimating 3D human pose.

Pavlo et al. [27] proposed the temporal convolutional model (TCM) with the input as a 2D keypoint sequence. The input layer uses a 2D human pose of each frame and applies it to a temporal convolution with kernel size $W = 3$, the output channels $C = 1024$, and a dropout rate $p = 0.25$; the number of blocks is 4; the tensor sizes are (243, 34); 243 frames is the receptive field and 34 channels (each frame is 17×2 ; 2 is the (x, y) dimensions), as illustrated in Figure 7.

**Figure 7.** Illustration of the TCM architecture [27].

In particular, the authors also proposed a semi-supervised training method by leveraging the unlabeled video for extending the supervised loss function with a back-projection loss term. There are two processes performed on the unlabeled video: the encoder implements 3D pose estimation from 2D joint coordinates, and the decoder is the back-projection of the estimated 3D pose to 2D joint coordinates.

4. Experimental Results

4.1. Data Collection, Implementations, and Evaluations

We used the benchmark Human 3.6M dataset [19] for evaluating the 2D human pose estimation/3D human pose estimation. Human3.6M is captured from 11 subjects/people (6 males and 5 females) in the Lab scene, which includes 16 daily activities (directions, discussion, greeting, posing, purchases, taking photos, waiting, walking, walking dog, walking pair, eating, phone talk, sitting, smoking, sitting down, miscellaneous). The frames were captured from time-of-flight (TOF) cameras, and the frame rate is from 25 to 50 Hz. Three-dimensional human pose annotations were marked by the MoCap system, and each pose includes 17 keypoints, as illustrated in Figure 8. In each human action, the camera's intrinsic parameters are provided.

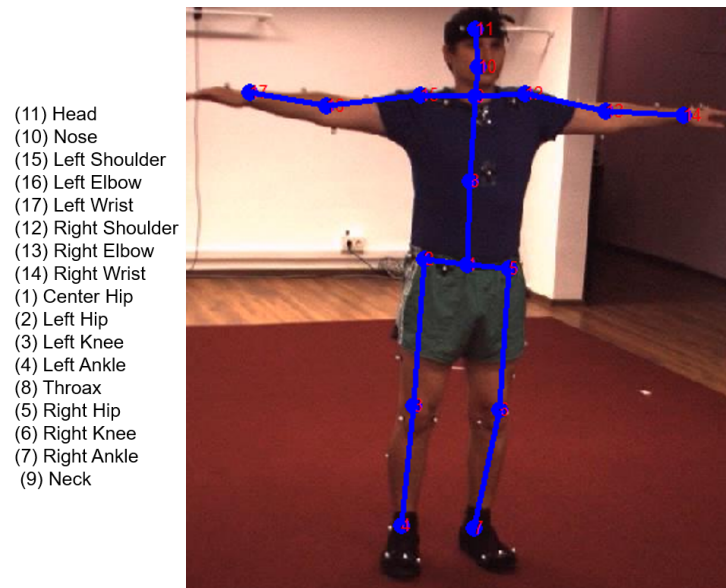


Figure 8. Example of 2D human pose/skeleton from the Human 3.6M dataset [19].

To evaluate the 2D human pose estimation, we used the camera's intrinsic parameter set to define the 2D human pose annotation on the image. The 2D human pose annotations are projected from 3D human pose annotation by Equation (2):

$$\begin{aligned} P2D.x &= \frac{P3D_c.x * fx}{P3D_c.z} + cx \\ P2D.y &= \frac{P3D_c.y * fy}{P3D_c.z} + cy \end{aligned} \quad (2)$$

where $P2D$ is the coordinate of the keypoint in the image. $P3D_c$ is the coordinate of the keypoint in the camera coordinate system, which is computed by (3) [87].

$$\begin{aligned} P3D_c.x &= \frac{(x_d - cx) * D(x_d, y_d)}{fx} \\ P3D_c.y &= \frac{(y_d - cy) * D(x_d, y_d)}{fy} \\ P3D_c.z &= depth(x_d, y_d) \end{aligned} \quad (3)$$

where fx , fy , cx , and cy are the intrinsic parameters of the camera. Before converting from 3D to 2D, the coordinates of the $P3D_c$ joints in the camera coordinate system need to be determined based on Formula (4).

$$P3D_c = (P3D_w - T) * R^{-1} \quad (4)$$

where R and T are the rotation and translation parameters to transform from the real-world coordinate system to the camera coordinate system. $P3D_w$ is the coordinate of the keypoint in the world coordinate system.

The training and testing data of the Human 3.6M dataset include three protocols: Pro #1 includes Subject #1, Subject #5, Subject #6, and Subject #7 for training and Subject #9 and Subject #11 for testing; Pro #2 is similar to Pro #1, but the predictions are further post-processed by a rigid transformation before comparing to the ground-truth; Pro #3 includes Subject #1, Subject #5, Subject #6, Subject #7, and Subject #9 for training and Subject #11 for testing.

In this paper, we used a PC with a Core I5 with GPU GTX 970, 4GB for fine-tuning, training, and testing 2D human pose estimation/3D human pose estimation. The programs were performed in the Python language (≥ 3.6 version) with the support of the CUDA 11.2/cuDNN 8.1.0 libraries. In addition, there are a number of other libraries such as Numpy, Scipy, Pillow, Cython, Matplotlib, Scikit-image, Tensorflow $\geq 1.3.0$, etc.

For 2D human pose estimation evaluation, we evaluated the average 2D keypoint localization errors (A2DLEs) of 2D keypoints/2D human pose annotation (P_g) and the estimated 2D keypoints/2D human pose (P_e) in pixels. This is defined as the Euclidean distance between the 2D keypoints annotation and the estimated 2D keypoints, as in Equation (5).

$$A2DLE = \frac{1}{N_{ac}} \sum_1^{N_{ac}} \frac{1}{N_f} \sum_1^{N_f} \frac{1}{17} \sum_1^{17} (P_e - P_g)^2 \quad (5)$$

where N_{ac} is the number of human actions, N_f is the number of frames in the human action, and 17 is the number of keypoints of the human pose.

For 3D human pose estimation evaluation, we used the mean per joint position error (MPJPE) measurement, which is the mean Euclidean distance between estimated 3D joint positions ($P3D_e$) and 3D joint positions' annotation ($P3D_g$), following Equation (6).

$$MPJPE = \frac{1}{17} \sum_1^{17} (P3D_e - P3D_g)^2 \quad (6)$$

The details of the protocols are as follows: Protocol (Pro) #1 uses the MPJPE measurement in millimeters (mm) to evaluate as [88]. Pro #2 uses the P-MPJPE (mm) [88,89]. Pro #3 uses the N-MPJPE [31] measurement.

4.2. Results and Discussions

The results of 2D keypoint estimation on Pro #1 of the Human 3.6M dataset are shown in Table 5. The results were evaluated on HR and its improved version, called Higher HR. The widths (w) of the high-resolution subnetworks in the last three stages were 32 ($w32$) or 48 ($w48$). The input image was resized to a fixed size $(256 \times 192)/(256 \times 192)$ or $(384 \times 288)/(384 \times 288)$ or $(512 \times 512)/(512)$ or $(640 \times 640)/(640)$. In Table 5, the error of the HR+U+S [18] methods is the highest, A2DLE = 4.4 pixels. The HR+U+S [18] and CPN [79] methods perform 2D keypoint estimation on the bounding box ground-truth of the human on the image. The results of the person detection are presented in Table 2. The results of the human detection step in the proposed method have an accuracy of close to 100%. The output of this step will be the input for 2D human pose estimation. The result in the next step of the proposed method (YOLOv5 + CC_HR_384_288) has an error of A2DLE = 5.14 pixels; this is the result of the estimated 2D keypoints on the bounding box of the YOLOv5 + CC detector. This result is better than the CPN+HR method [79] for human detection and 2D human pose estimation (A2DLE = 5.4). This shows that our proposed method for human detection is better with CPN [79].

The result of the method we proposed is very good; it is fully automatic with the input of the original image (1000×1002). The results of HR (HR_w48_384_288 [12], HR_w32_384_288 [12], HR_w32_256_192 [12], HR_w32_256_256 [12]) and Higher HR

(Higher_HR_w48_640 [90], Higher_HR_w32_640 [90], Higher_HR_w32_512 [90]) have a high error, and we used the pre-trained model that was fine-tuned on the COCO dataset. The processing time of human detection and 2D keypoint estimation was 3.15 fps.

Table 5. The 2D keypoint estimation/2D human pose estimation results on Pro #1 of the Human 3.6M dataset.

Method	Number of Parameters (Million(M))	Input Size of Image	Average Joint Localization Error (A2DLE) (pixels)	Processing Time (fps)
HR_w48_384_288 [12]	63.6M	Full size (1000 × 1002)	52.4	3.144
HR_w32_384_288 [12]	28.5M	Full size (1000 × 1002)	54.3	3.14
HR_w32_256_192 [12]	28.5M	Full size (1000 × 1002)	58.2	3.145
HR_w32_256_256 [12]	28.5M	Full size (1000 × 1002)	56.7	3.14
CPN [79]	-	Bounding-box of detected person	5.4	-
HR+ U + S [18]	63.6M	Bounding-box of detected person	4.4	-
Higher_HR_w48_640 [90]	63.6M	Full size (1000 × 1002)	40.0	2.5
Higher_HR_w32_640 [90]	28.6M	Full size (1000 × 1002)	40.2	2.6
Higher_HR_w32_512 [90]	28.6M	Full size (1000 × 1002)	40.8	2.88
Ours(YOLOv5+CC_HR_384_288)	63.6M	Full size (1000 × 1002)	5.14	3.15

The results of 3D keypoint estimation/3D human pose estimation on Pro #1, Pro #2, and Pro #3 of the Human 3.6M dataset are shown in Table 6. In Table 6, we compare the proposed method with the 3D human pose estimation methods that have the best results currently. At the same time, we also present information about boxes (bounding box of human detection) and 2D keypoint estimation (the method used to estimate 2D human poses). The method that we propose has an accuracy equivalent to the 3D human pose estimation methods based on the human bounding box data, which is the ground-truth. The error on the MPJPE, P-MPJPE, and N-MPJPE measures of our proposed method on Pro #1, Pro #2, and Pro #3 is 46.5 mm, 37.0 mm, and 46.4 mm, respectively. The method we propose is much more accurate than the VNect (ResNet-50) [7] method (the error is 80.2 mm of Pro #1). In particular, our proposed method is slightly better (MPJPE = 50.5 mm) than the GraFormer [55,56] methods (MPJPE = 58.7 mm [55] and MPJPE = 51.8 mm [56]) for estimating 3D human pose. The GraFormer [55,56] method is a method based on a recent proposal.

In this paper, we also compare the processing time of the proposed method with the VNect [7] method on the Human 3.6M dataset when performed on a computer with a low configuration, as presented in Table 7.

The results of 2D human pose estimation and 3D human pose estimation are illustrated in Figure 9.

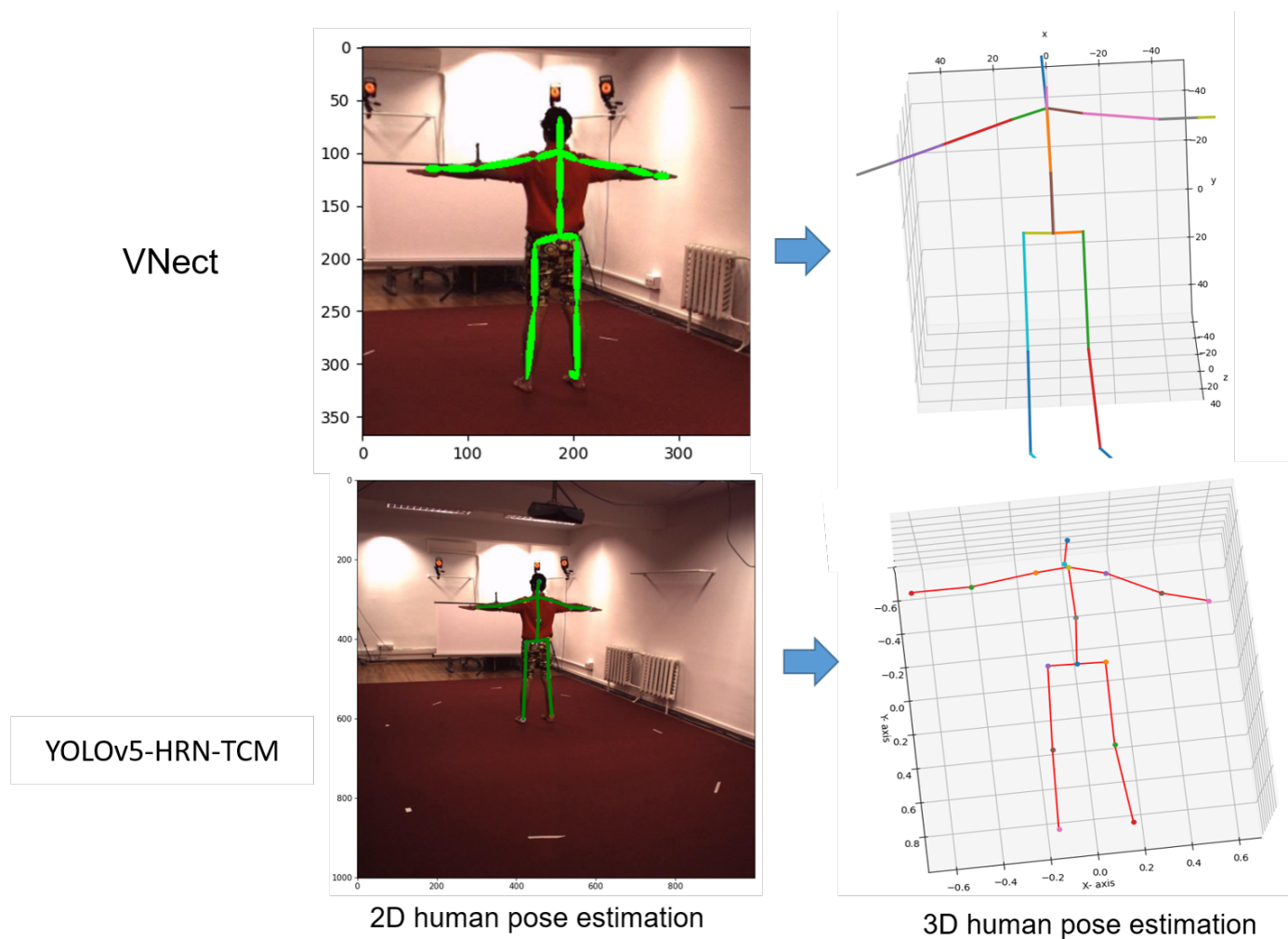


Figure 9. Illustrating of 2D and 3D human posture estimation. The VNect [7] method performs 2D human pose estimation on the resized image of (386×386) pixels. Our method proposes to estimate the human pose on a full-sized image (1000×1002) pixels.

Table 6. Illustration of 3D keypoint estimation/3D human pose estimation on Pro #1, Pro #2, and Pro #3 of the Human 3.6M dataset.

Method	2D Keypoints Estimation	BBoxes	Blocks	Receptive Field (frames)/ N. Epochs	MPJPE (mm) (Pro #1)	P-MPJPE (mm) (Pro #2)	N-MPJPE (mm) (Pro #3)
TCM + semi-sup. [27]	HR_w48_384_288	HR_w48_384_288	4	243/80	216.2	184.4	215.8
TCM + semi-sup. [27]	HR_w32_384_288	HR_w32_384_288	4	243/80	216.3	184.2	215.9
TCM + semi-sup. [27]	HR_w32_256_192	HR_w32_256_192	4	243/80	217.2	186.3	217.1
TCM + semi-sup. [27]	HR_w32_256_256	HR_w32_256_256	4	243/80	216.4	185.8	216.0
TCM + semi-sup. [27]	HR_w48_384_288	Ground-truth	4	243/80	124.6	104.3	123.8
TCM + semi-sup. [27]	HR_w32_384_288	Ground-truth	4	243/80	125.4	107.9	124.5
TCM + semi-sup. [27]	HR_w32_256_192	Ground-truth	4	243/80	124.22	105.7	123.8
TCM + semi-sup. [27]	HR_w32_256_256	Ground-truth	4	243/80	123.7	105.3	123.5
TCM + semi-sup. [27]	CPN	Mask RCNN	4	243/80	46.8	36.5	-
TCM + semi-sup. [27]	CPN	Ground-truth	4	243/80	47.1	36.8	-
TCM + semi-sup. [27]	CPN	Ground-truth	3	81	47.7	37.2	
TCM + semi-sup. [27]	CPN	Ground-truth	2	27	48.8	38.0	
TCM + semi-sup. [27]	Mask RCNN	Mask RCNN	4	243/80	51.6	40.3	
TCM + semi-sup. [27]	Ground-truth	Ground-truth	4	243/80	37.2	27.2	35.4
3d-pose-baseline [49]	SHN	SHN (cro. 440 × 440)	-	-/200	62.9	47.7	-
Cas.(Full-sup.) [18]	HR_w32_384_288	HR_w32_384_288	3	-/200	49.7	37.7	-
Adversarial Lea. [60]	SHN	SHN	-	-/90	58.6	37.7	-
SemGCN [37]	SHN	SHN	-	-/200	57.6	-	-
CVAE-based [61]	SHN + 2DPoseNet	SHN + 2DPoseNet	-	-/200	58.0	40.9	-
RootNet+PoseNet [39]	Mask RCNN	Mask RCNN	-	-/20	54.4	-	-
Multi-View Sup. [31]	ResNet-50	ResNet-50	-	-/-	-	64.6	-
Pose SS (PSS) [43]	ResNet-50	ResNet-50	-	-/140	65.3	57.2	-
Cas. (Weakly-sup.) [18]	HR_w32_384_288	-	2	-/200	60.8	46.2	-
VNect (ResNet-50) [7]	ResNet-50	ResNet-50	-	-/-	80.5	-	-
Vnect (ResNet-100) [7]	ResNet-100	ResNet-100	-	-/-	82.5	-	-
Boosting [91]	SHN	SHN	-	-/2	88.8	66.5	-
GraFormer [55]	SHN	SHN	-	-/100	58.7	-	-
GraFormer [55]	Ground-truth	Ground-truth	-	-/100	35.2	-	-
GraFormer [56]	SHN	SHN	-	-/100	51.8	-	-
GraFormer [56]	Ground-truth	Ground-truth	-	-/100	35.2	-	-
Our	YOLOv5_HR_384_288	YOLOv5_HR_384_288	4	243/80	50.5	37.0	45.7
Our	YOLOv5_HR_384_288	Ground-truth	4	243/80	46.5	35.0	44.4

Table 7. Processing time of 3D human pose estimation on the Human 3.6M dataset.

Method	Processing Time (FPS)
VNect [7]	1.36
Our (YOLOv5-HR-TCM)	3.146

5. Pose-Based Application

Figure 1 presents several applications based on human posture estimation. There have been studies using human posture to build applications in sports [59,92] and preserving and developing traditional martial arts [93,94]. Moreover, there is Zhang et al. [95], who published a dataset of human postures in martial arts, dancing, and sports. Scoring in sports competitions and martial arts performances have traditionally been based on the experts on the jury. The movements and actions of athletes are often very fast, so mistakes are inevitable. In particular, the assessment is based on the subjectivity and experience of the jury members. Therefore, having a system to support the process of assessing the accuracy of movements in sports competitions and martial arts performances has very high practical significance, as illustrated in Figure 10. Sports and martial arts competitions often take place in a larger space, so it is not reasonable to evaluate and score points based on the absolute coordinates of the person, bones, and joints. Therefore, we propose a rating and scoring system based on the deviation angle of the original shelf and important bones. Figure 10a,b illustrates calculating the angle $(\vec{a}, \vec{ox})$ between the straight line of two legs with the ox axis; the smaller this angle, the higher the score is.

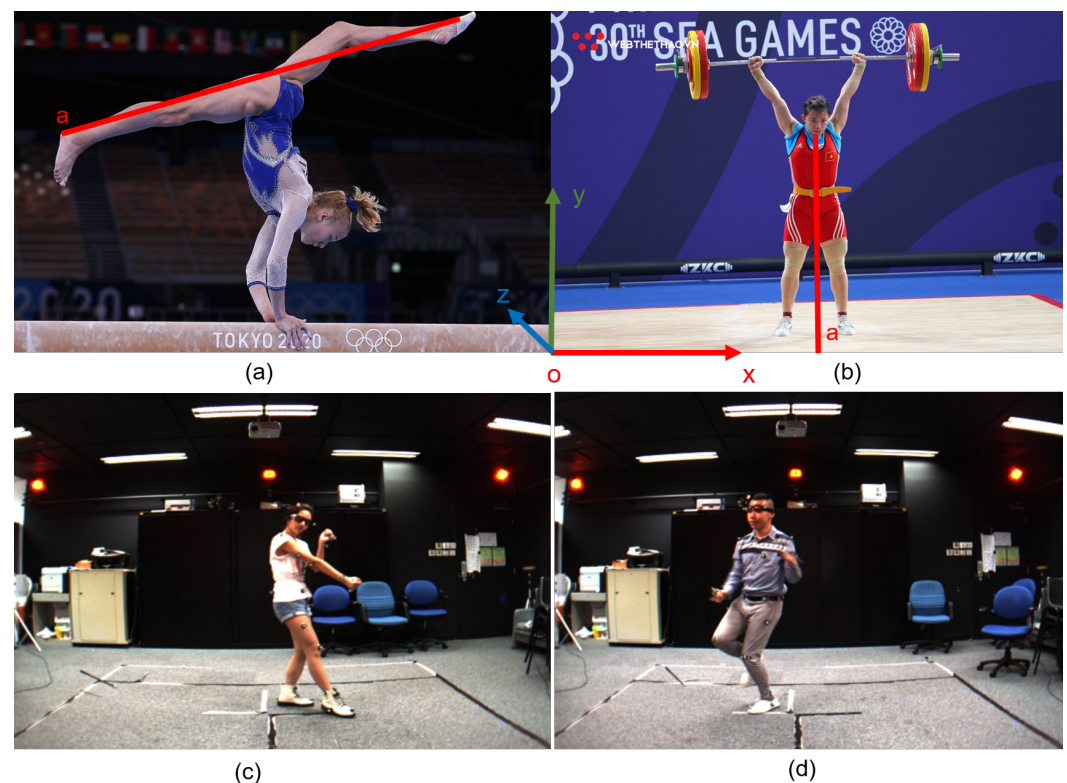


Figure 10. Illustration of computing movement scores in women's artistic gymnastics (a) and weightlifting (b). Images (c) and (d) are an illustration of a dancer performing jazz and hip hop dances [95,96].

In this paper, we propose an application based on the estimated human posture in 3D space. Our application is based on calculating the angle A_d of a pair of bones between the estimated human skeleton and the ground-truth of the human skeleton, as illustrated in

Figure 11. The deviation angle A_d (as Equation (7)) is then averaged over the pair of bones A_{av_d} (as Equation (8)).

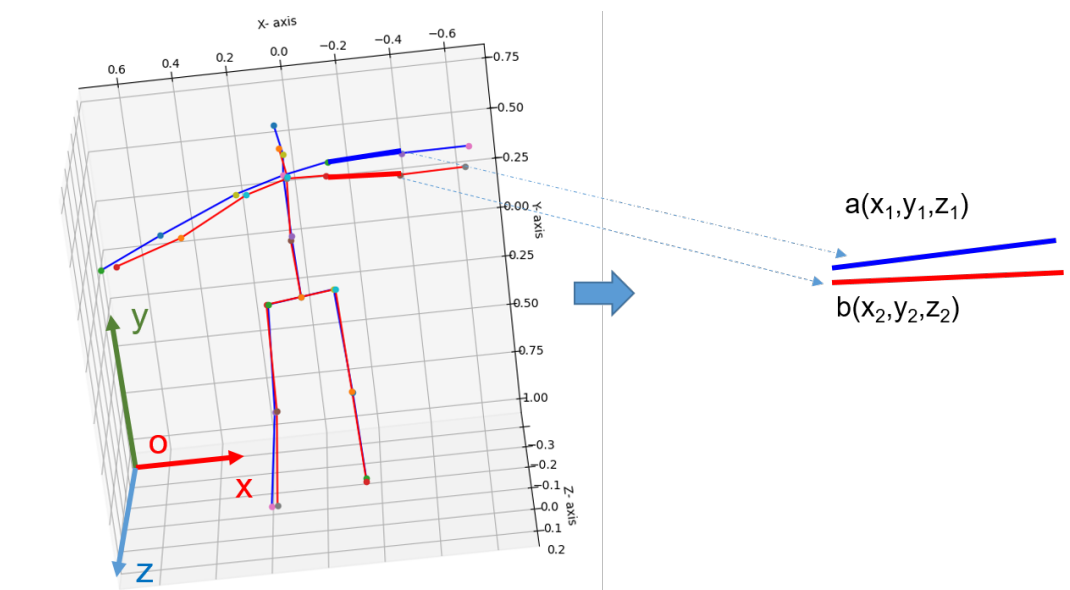


Figure 11. Illustration of the angle between a pair of bones between the estimated human skeleton and the ground-truth human skeleton in 3D space of the Human 3.6M dataset [19]. The left is the estimated human skeleton (red color) and the ground-truth human skeleton (blue color) in 3D space. The right shows the calculation of the angle between a pair of elbow bones.

$$A_d(\vec{a}, \vec{b}) = \arccos(\frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|}) = \arccos(\frac{x_1 y_1 z_1 + x_2 y_2 z_2}{\sqrt{x_1^2 + y_1^2 + z_1^2} \sqrt{x_2^2 + y_2^2 + z_2^2}}) \quad (7)$$

where vector $\vec{a}$ has coordinates (x_1, y_1, z_1) and vector $\vec{b}$ has coordinates (x_2, y_2, z_2) .

$$A_{av_d} = \frac{1}{16} \sum_{i=1}^{16} A_d \quad (8)$$

Based on the assessment and scoring of women's artistic gymnastics [97–99], as illustrated in Figure 10a, we propose how to evaluate and score the “Execution Score: execution, artistry, composition and technique” contest, as shown in Table 8. In Table 8, if the angle of deviation is 2 degrees, subtract 0.1 points.

Table 8. The method of evaluating and scoring women's artistic gymnastics with the contest “Execution Score”.

Error Angle (Degrees)	Score (Points)
0	10
2	9.9
4	9.8
...	...

As illustrated in Figure 10c,d, the human skeleton in dance teaching (hip hop, jazz) by experts and coaches is the original source of data for teaching and assessing the accuracy

of movements. In this paper, we propose a method of assessment and scoring in dance teaching based on the deviation angle between the experts' human skeleton (ground-truth) and the estimated human skeleton of the trainees. The details of the assessment and scoring are shown in Table 9. In Table 9, if the angle of deviation is 1 degree, subtract one point.

Table 9. The method of evaluating and scoring dance training.

Error Angle (Degrees)	Score (Points)
0	100
1	99
2	98
...	...

The results of the deflection angle between the pairs of bones on the Human 3.6M dataset are shown in Table 10. The average deviation angle between the estimated 3D human skeleton bones and the 3D human skeleton bones' ground-truth is 8.2 degrees; the scoring results based on Table 9 are illustrated in Figure 12. If the calculation is based on the average deviation angle of Table 9, the scoring system gives $100 - 8.2 = 91.8$ points.

The worst-case scenario is when the estimated angle error is 90 degrees, so the error rate of the current application is $\frac{8.2}{90} * 100 = 9.11\%$. This is a relatively large error, but is calculated on average over 16 human bones. However, in practical applications in sports, we are often only interested in some bones in the human body. As shown in Figure 10a, we are only interested in the angle between the legs and the shelf; the smaller the angle, the higher the score is. Figure 10b shows the results when we are interested in the angle between the "Thorax-Neck" and the ground floor; the closer the angle is to 90 degrees, the higher the score is.

Table 10. Deviation angle (A_Av_d - degrees) between estimated 3D human skeleton bones and 3D human skeleton ground-truth based on the Human 3.6M dataset (Pro #1).

Bone Pairs	Mean Deviation Angle (A_Av_d) (Degrees)
Center Hip-Right Hip	6.2
Right Hip-Right Knee	6.0
Right Knee-Right Ankle	6.9
Center Hip-Left Hip	6.2
Left Hip-Left Knee	5.1
Left Knee-Left Ankle	7.9
Center Hip-Thorax	6.8
Thorax-Neck	6.0
Neck-Nose	14.2
Nose -Head	10.1
Neck-Left Shoulder	9.0
Left Shoulder-Left Elbow	8.7
Left Elbow-Left Wrist	9.9
Neck-Right Shoulder	9.6
Right Shoulder-Right Elbow	8.8
Right Elbow-Right Wrist	10.1
Average	8.2

Based on the angular results in Table 10, we show the distribution of the deviation angle of the set "s_09_act_02_subact_01_ca_01" of the Human 3.6M dataset in Figure 13. Error distribution results are all concentrated in the range from 0 to 10 degrees.

Figure 12 shows the estimated results of 2D human pose and 3D human pose. Scores based on the rating in Table 9 are also shown.

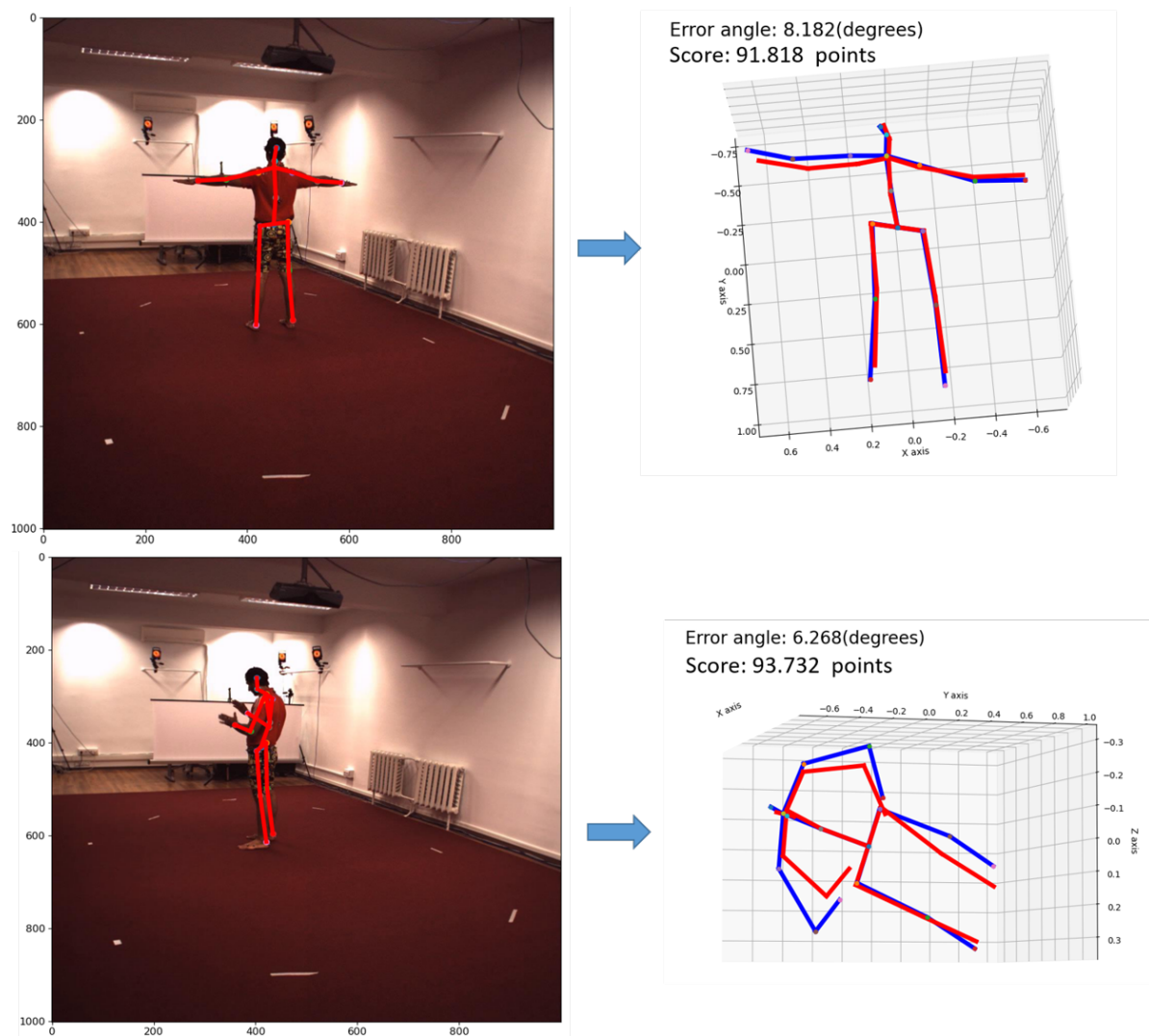


Figure 12. Illustration of the human posture estimation in color images (**left**) and the result of human posture estimation in 3D space (**right**). The blue skeleton is the ground-truth; the red skeleton is the estimated skeleton of the proposed YOLOv5-HR-TCM framework.

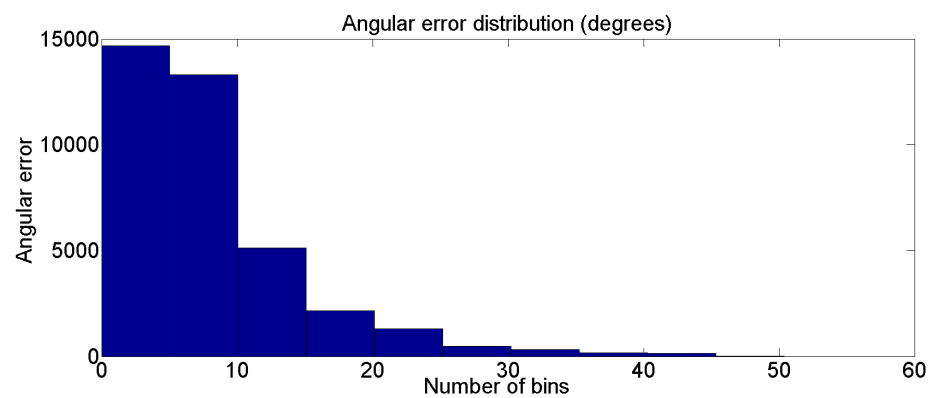


Figure 13. Deviation angle distribution between estimated 3D human skeleton bones and 3D human skeleton ground-truth on the dataset (*s_09_act_02_subact_01_ca_01*) of Human 3.6M dataset.

The entire source code of the sports scoring and estimation system is stored in link (https://drive.google.com/drive/folders/1WRr-L3IcH_lhSqMUJDaw1v23OBRdTXPC?usp=sharing, (accessed on 12 May 2022)).

Thus, our proposed method can perform end-to-end 3D human pose estimation at a rate of 3.146 fps, which can be improved on computers with higher frames for real-time speed response in a gym fitness center. However, the proposed model also has some limitations that currently only estimate the pose of a human in the image. Therefore, it usually applies only to certain non-competitive sports, but only to performance and scoring sports such as skating, gymnastics, weightlifting, etc.

6. Conclusions and Future Works

Estimation of 2D human pose and 3D human pose has been studied extensively in recent years. However, studies often focus on improving the accuracy of the estimation results. In terms of 2D human pose estimation and 3D human pose estimation processing time, especially in building applications based on 3D human pose estimation, there are still many limitations. This paper accomplished two main tasks: (1) We proposed a unified end-to-end framework for estimating the 3D human pose from color image input data, named YOLOv5-HR-TCM. The proposed framework is a combination of current best approaches at each step of the estimation process, such as human detection of color images, estimating the human pose on the bounding box of the detected human, and estimating the quantity of 3D human pose from 2D human pose (2D to 3D lifting method). (2) An application was built for assessment and scoring in artistic gymnastics, sports competitions, and assessment of teaching dance, traditional martial arts, and sports. In the near future, the survey and evaluation of the combination at each step such as human detection, 2D human posture estimation (e.g., EfficientHRet [100], YOLO-POSE [101]), and 3D human posture estimation (e.g., GraFormer [56]), to choose the best method in each step to build the best overall model, will be performed. Specifically, we will apply the results of 3D human posture estimation to many sports applications, human activity recognition, and sports analysis. More specifically, the evaluation test of the 3D human posture estimation model is based on color images, and scoring is based on the angle of deviation of the weightlifter in the gym fitness center.

Author Contributions: Conceptualization, H.-C.N. and V.-H.L.; Methodology, H.-C.N., T.-H.N. and V.-H.L.; Visualization, R.S.; Writing—original draft, R.S. and V.-H.L.; Writing—review & editing, V.-H.L., T.-H.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by Hung Vuong University under grant number 03/2022/HD-HV03.2022. The APC was paid under the project financed under the program of the Polish Minister of Science and Higher Education under the name "Regional Initiative of Excellence" in the years 2019–2022 project number 020/RID/2018/19. This research is funded by Tan Trao University in Tuyen Quang province, Vietnam.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The paper is our research, not related to any organization or individual. It is part of a series of studies by 2D, 3D human pose estimation.

References

1. Stenum, J.; Cherry-Allen, K.M.; Pyles, C.O.; Reetzke, R.D.; Vignos, M.F.; Roemmich, R.T. Applications of pose estimation in human health and performance across the lifespan. *Sensors* **2021**, *21*, 7315. [CrossRef] [PubMed]
2. Sawant, C. Human activity recognition with openpose and Long Short-Term Memory on real time images. EasyChair Preprint no. 2297, EasyChair, 2020. Available online: <https://www.semanticscholar.org/paper/Human-activity-recognition-with-openpose-and-Long-Sawant/e7503d2a381a4de534b9ece7d520435370ae517a> (accessed on 12 December 2021).
3. Minds, B. An Overview of Human Pose Estimation with Deep Learning. 2021. Available online: <https://beyondminds.ai/blog/an-overview-of-human-pose-estimation-with-deep-learning/> (accessed on 12 December 2021).
4. Barla, N. A Comprehensive Guide to Human Pose Estimation. 2021. Available online: <https://www.v7labs.com/blog/human-pose-estimation-guide> (accessed on 12 December 2021).

5. Tatariants, M. Human Pose Estimation Technology 2021 Guide. 2020. Available online: <https://mobidev.biz/blog/human-pose-estimation-ai-personal-fitness-coach> (accessed on 12 December 2021).
6. Zhou, X.; Huang, Q.; Sun, X.; Xue, X.; Wei, Y. Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 398–407. [CrossRef]
7. Mehta, D.; Sridhar, S.; Sotnychenko, O.; Rhodin, H.; Shafiei, M.; Seidel, H.P.; Xu, W.; Casas, D.; Theobalt, C. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM Trans. Graph.* **2017**, *26*, 44. [CrossRef]
8. Babu, S.C. A 2019 guide to Human Pose Estimation with Deep Learning. 2019. Available online: <https://nanonets.com/blog/human-pose-estimation-2d-guide/> (accessed on 5 December 2021).
9. Tompson, J.; Goroshin, R.; Jain, A.; LeCun, Y.; Bregler, C. Efficient object localization using Convolutional Networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 648–656.
10. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016.
11. Carreira, J.; Agrawal, P.; Fragkiadaki, K.; Malik, J. Human Pose Estimation with Iterative Error Feedback. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
12. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the CVPR, Long Beach, CA, USA, 16–20 June 2019.
13. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; pp. 1–16.
14. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In Proceedings of the 14th European Conference ECCV, Amsterdam, The Netherlands, 11–14 October 2016.
15. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. Available online: <https://arxiv.org/abs/1908.07919> (accessed on 5 December 2021).
16. Toshev, A.; Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the IEEE Conference on CVPR, Columbus, OH, USA, 23–28 June 2014.
17. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. *Microsoft COCO: Common Objects in Context*; Springer: Cham, Switzerland, 2014.
18. Li, S.; Ke, L.; Pratama, K.; Tai, Y.W.; Tang, C.K.; Cheng, K.T. Cascaded Deep Monocular 3D Human Pose Estimation with Evolutionary Training Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
19. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *TPAMI* **2014**, *36*, 1325–1339. [CrossRef] [PubMed]
20. Wang, J.; Tan, S.; Zhen, X.; Xu, S.; Zheng, F.; He, Z.; Shao, L. Deep 3D human pose estimation: A review. *Comput. Vis. Image Underst.* **2021**, *210*, 103225. [CrossRef]
21. Ji, X.; Fang, Q.; Dong, J.; Shuai, Q.; Jiang, W.; Zhou, X. A survey on monocular 3D human pose estimation. *Virtual Real. Intell. Hardw.* **2020**, *2*, 471–500. [CrossRef]
22. Dang, Q.; Yin, J.; Wang, B.; Zheng, W. Deep learning based 2D human pose estimation: A survey. *Tsinghua Sci. Technol.* **2019**, *24*, 663–676. Available online: <https://arxiv.org/abs/2012.13392> (accessed on 5 December 2021). [CrossRef]
23. Le, V.H.; Nguyen, H.C. A survey on 3D hand skeleton and pose estimation by convolutional neural network. *Adv. Sci. Technol. Eng. Syst.* **2020**, *5*, 144–159. [CrossRef]
24. Chen, X.; Lin, K.Y.; Liu, W.; Qian, C.; Lin, L. Weakly-supervised discovery of geometry-aware representation for 3D human pose estimation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10887–10896. [CrossRef]
25. Glenn Jocher. YOLOv5 Torials. 2021. Available online: <https://github.com/ultralytics/yolov5#tutorials> (accessed on 6 December 2021).
26. Jocher, G. Head and Person Detection Model. 2021. Available online: <https://github.com/deepakcrk/yolov5-crowdhuman> (accessed on 6 December 2021).
27. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
28. Chen, Y.; Tian, Y.; He, M. Monocular human pose estimation: A survey of deep learning-based methods. *Comput. Vis. Image Underst.* **2020**, *192*, 1–23. [CrossRef]
29. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
30. Tuli, S.; Dasgupta, I.; Grant, E.; Griffiths, T.L. Are Convolutional Neural Networks or Transformers More Like Human vision? Available online: <https://arxiv.org/abs/2105.07197> (accessed on 6 December 2021).

31. Rhodin, H.; Meyer, F.; Spörri, J. Learning Monocular 3D Human Pose Estimation from Multi-view Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8437–8446.
32. Tome, D.; Russell, C.; Agapito, L. Lifting from the deep: Convolutional 3D pose estimation from a single image. In Proceedings of the CVPR 2017: 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5689–5698. [\[CrossRef\]](#)
33. Wang, K.; Lin, L.; Jiang, C.; Qian, C.; Wei, P. 3D Human Pose Machines with Self-supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 1069–1082. [\[CrossRef\]](#)
34. Véges, M.; Varga, V.; Lőrincz, A. 3D Human Pose Estimation with Siamese Equivariant Embedding. *arXiv* **2018**, arXiv:1809.07217.
35. Fang, H.s.; Xu, Y.; Wang, W.; Liu, X.; Zhu, S.c. Learning Pose Grammar to Encode Human Body Configuration for 3D Pose Estimation. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
36. Omran, M.; Lassner, C.; Pons-Moll, G.; Gehler, P.; Schiele, B. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In Proceedings of the 2018 International Conference on 3D Vision, Verona, Italy, 5–8 September 2018; pp. 484–494. [\[CrossRef\]](#)
37. Zhao, L.; Peng, X.; Tian, Y.; Kapadia, M.; Metaxas, D.N. Semantic graph convolutional networks for 3D human pose regression. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3420–3430. [\[CrossRef\]](#)
38. Nibali, A.; He, Z.; Morgan, S.; Prendergast, L. 3D human pose estimation with 2D marginal heat maps. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1477–1485. [\[CrossRef\]](#)
39. Moon, G.; Chang, J.Y.; Lee, K.M. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 10132–10141. [\[CrossRef\]](#)
40. Lee, K.; Lee, I.; Lee, S. Propagating LSTM: 3D pose estimation based on joint interdependency. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; pp. 123–141. [\[CrossRef\]](#)
41. Li, C.; Lee, G.H. Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
42. Pavlakos, G.; Zhou, X.; Derpanis, K.G.; Daniilidis, K. Coarse-to-fine volumetric prediction for single-image 3D human pose. In Proceedings of the CVPR 2017: 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 1263–1272. [\[CrossRef\]](#)
43. Kocabas, M.; Karagoz, S.; Akbas, E. Self-Supervised Learning of 3D Human Pose using Multi-view Geometry. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1077–1086.
44. Wandt, B.; Rosenhahn, B. RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
45. Tekin, B.; Marquez-Neila, P.; Salzmann, M.; Fua, P. Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation. In Proceedings of the IEEE Conference on CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 3961–3970. [\[CrossRef\]](#)
46. Isakov, K.; Burkov, E.; Lempitsky, V.S.; Malkov, Y. Learnable Triangulation of Human Pose. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
47. Sun, X.; Li, C.; Lin, S. An Integral Pose Regression System for the ECCV2018 PoseTrack Challenge. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; pp. 1–5.
48. Rhodin, H.; Constantin, V.; Katircioglu, I.; Salzmann, M.; Fua, P. Neural scene decomposition for multi-person motion capture. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7695–7705. [\[CrossRef\]](#)
49. Martinez, J.; Hossain, R.; Romero, J.; Little, J.J. A Simple Yet Effective Baseline for 3d Human Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2659–2668. [\[CrossRef\]](#)
50. Li, W.; Liu, H.; Ding, R.; Liu, M.; Wang, P.; Yang, W. Exploiting Temporal Contexts with Strided Transformer for 3D Human Pose Estimation. *IEEE Transactions on Multimedia*, **2022**. Available online: <https://arxiv.org/abs/2103.14304> (accessed on 6 June 2022).
51. Zheng, C.; Zhu, S.; Mendieta, M.; Yang, T.; Chen, C.; Ding, Z. 3D Human Pose Estimation with Spatial and Temporal Transformers. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; Volume 1, pp. 11636–11645. [\[CrossRef\]](#)
52. Hossain, M.R.I.; Little, J.J. Exploiting temporal information for 3D human pose estimation. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; pp. 69–86. [\[CrossRef\]](#)
53. Wang, L.; Chen, Y.; Guo, Z.; Qian, K.; Lin, M.; Li, H.; Ren, J.S. Generalizing Monocular 3D Human Pose Estimation in the Wild. *arXiv* **2019**, arXiv:1904.05512.
54. Pavllo, D.; Grangier, D.; Auli, M. QuaterNet: A Quaternion-based Recurrent Model for Human Motion. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018.

55. Zhao, W.; Tian, Y.; Ye, Q.; Jiao, J.; Wang, W. GraFormer: Graph Convolution Transformer for 3D Pose Estimation. 2021, Volume 1, Available online: <https://arxiv.org/pdf/2109.08364.pdf> (accessed on 6 June 2022).
56. Zhao, W.; Wang, W.; Tian, Y. GraFormer: Graph-Oriented Transformer for 3D Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 20438–20447.
57. Song, L.; Yu, G.; Yuan, J.; Liu, Z. Journal of Visual Communication and Image Representation Human pose estimation and its application to action recognition : A survey. *J. Vis. Commun. Image Represent.* **2021**, *76*, 103055. [[CrossRef](#)]
58. Chen, C.H.; Ramanan, D. 3D human pose estimation = 2D pose estimation + matching. In Proceedings of the IEEE Conference on CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 5759–5767. [[CrossRef](#)]
59. Badiola-Bengoia, A.; Mendez-Zorrilla, A. A systematic review of the application of camera-based human pose estimation in the field of sport and physical exercise. *Sensors* **2021**, *21*, 5996. [[CrossRef](#)] [[PubMed](#)]
60. Yang, W.; Wang, X.; Ren, J.; Li, H. 3D Human Pose Estimation in the Wild by Adversarial Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
61. Sharma, S.; Varigonda, P.T.; Bindal, P.; Sharma, A.; Jain, A. Monocular 3D Human Pose Estimation by Generation and Ordinal Ranking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea 27 October–2 November 2019.
62. Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object Detection via Region-based Fully Convolutional Networks. In *Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 September 2016*; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29.
63. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster RCNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)]
64. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Proceedings of the ECCV (1), Amsterdam, The Netherlands, 11–14 October 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Lecture Notes in Computer Science; Springer: New York, NY, USA, 2016; Volume 9905, pp. 21–37.
65. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 779–788. [[CrossRef](#)]
66. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the CVPR 2017: 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016, 2017; pp. 6517–6525. [[CrossRef](#)]
67. Redmon, J.; Ali, F. YOLOv3: An Incremental Improvement. 2018. Available online: <http://arxiv.org/abs/1804.02767> (accessed on 18 April 2021).
68. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
69. Jonathan, H. Object Detection: Speed and Accuracy Comparison (Faster RCNN, R-FCN, SSD, FPN, RetinaNet and YOLOv3). 2018. Available online: <https://jonathan-hui.medium.com/object-detection-speed-and-accuracy-comparison-faster-r-cnn-r-fcn-ssd-and-yolo-5425656ae359> (accessed on 18 December 2021).
70. Girshick, R. Fast RCNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [[CrossRef](#)]
71. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask RCNN. In Proceedings of the ICCV, Venice, Italy, 22–29 October 2017.
72. Abdulla, W. Mask RCNN for Object Detection and Instance Segmentation on Keras and TensorFlow. 2017. Available online: https://github.com/matterport/Mask_RCNN (accessed on 12 December 2021).
73. SSD MobileNet V1 architecture. 2018. Available online: <https://iq.opengenus.org/ssd-mobilenet-v1-architecture/> (accessed on 22 December 2021).
74. gao hao. Single Shot MultiBox Detector Implementation in Pytorch. 2020. Available online: <https://github.com/qfgaohao/pytorch-ssd> (accessed on 12 December 2021).
75. Krishnan, S. Person-Detection. 2021. Available online: <https://github.com/SusmithKrishnan/person-detection> (accessed on 12 December 2021).
76. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
77. Openpose. Openpose. 2019 Available online: <https://github.com/CMU-Perceptual-Computing-Lab/openpose> (accessed on 23 April 2021).
78. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017.
79. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded Pyramid Network for Multi-person Pose Estimation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112. [[CrossRef](#)]
80. Insafutdinov, E.; Pishchulin, L.; Andres, B.; Andriluka, M.; Schiele, B. Deepcut: A deeper, stronger, and faster multi-person pose estimation model. In Proceedings of the European Conference on Computer Vision (ECCV 2016), Amsterdam, The Netherlands, 11–14 October 2018; Springer: New York, NY, USA; pp. 34–50. [[CrossRef](#)]

81. Chu, X.; Yang, W.; Ouyang, W.; Ma, C.; Yuille, A.L.; Wang, X. Multi-context attention for human pose estimation. In Proceedings of the CVPR 2017: 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5669–5678. [\[CrossRef\]](#)
82. Chou, C.J.; Chien, J.T.; Chen, H.T. Self Adversarial Training for Human Pose Estimation. In Proceedings of the APSIPA ASC 2018: 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Honolulu, HI, USA, 12–15 November 2018; pp. 17–30. [\[CrossRef\]](#)
83. Yang, W.; Li, S.; Ouyang, W.; Li, H.; Wang, X. Learning Feature Pyramids for Human Pose Estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1290–1299. [\[CrossRef\]](#)
84. Ke, L.; Chang, M.C.; Qi, H.; Lyu, S. Multi-Scale Structure-Aware Network for Human Pose Estimation. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; pp. 731–746. [\[CrossRef\]](#)
85. Tang, Z.; Peng, X.; Geng, S.; Wu, L.; Zhang, S.; Metaxas, D. Quantized Densely Connected U-Nets for Efficient Landmark Localization. In Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany, 8–14 September 2018; pp. 348–364. [\[CrossRef\]](#)
86. Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; Shah, M. Deep Learning-Based Human Pose Estimation: A Survey. *J. ACM* **2018**, *37*, 111.
87. Burrus, N. Kinect Calibration. 2014. Available online: <http://nicolas.burrus.name/index.php\Research/KinectCalibration> (accessed on 20 March 2022).
88. Li, S.; Zhang, W.; Chan, A.B. Maximum-Margin Structured Learning with Deep Networks for 3D Human Pose Estimation. *Int. J. Comput. Vis.* **2017**, *122*, 149–168. [\[CrossRef\]](#)
89. Liang, S.; Sun, X.; Wei, Y. Compositional Human Pose Regression. In Proceedings of the ICCV, Venice, Italy, 22–29 October 2017; Volume 176–177. pp. 1–8. [\[CrossRef\]](#)
90. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In Proceedings of the CVPR, Seattle, WA, USA, 13–19 June 2020.
91. Li, Z.; Wang, X.; Wang, F.; Jiang, P. On boosting single-frame 3D human pose estimation via monocular videos. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 2192–2201. [\[CrossRef\]](#)
92. Echeverria, J.; Santos, O.C. Toward modeling psychomotor performance in karate combats using computer vision pose estimation. *Sensors* **2021**, *21*, 8378. [\[CrossRef\]](#)
93. Thanh, N.T.; Hung, L.V.; Cong, P.T. An Evaluation of Pose Estimation in Video of Traditional Martial Arts Presentation. *J. Res. Dev. Inf. Commun. Technol.* **2019**, *2019*, 114–126. [\[CrossRef\]](#)
94. Nguyen, T.T.; Le, V.H.; Duong, D.L.; Pham, T.C.; Le, D. 3D Human Pose Estimation in Vietnamese Traditional Martial Art Videos. *J. Adv. Eng. Comput.* **2019**, *3*, 471. [\[CrossRef\]](#)
95. Zhang, W.; Liu, Z.; Zhou, L.; Leung, H.; Chan, A.B. Martial Arts, Dancing and Sports dataset: A challenging stereo and multi-view dataset for 3D human pose estimation. *Image Vis. Comput.* **2017**, *61*, 22–39. [\[CrossRef\]](#)
96. Le, V.H.; Sre, R. Human Segmentation and Tracking Survey on Masks for MADS dataset. *Sensors* **2021**, *21*, 8397. [\[CrossRef\]](#) [\[PubMed\]](#)
97. Australia, G. How Does Women’s Artistic Gymnastics Scoring Work? 2022. Available online: https://www.gymnastics.org.au/VIC/Posts/News_Articles/2018/August/How_does_Gymnastics_Scoring_Work_-_WAG_.aspx#:~:text=Each%20skill%20performed%20is%20given,to%20increase%20their%20start%20value (accessed on 20 March 2022).
98. Gymnastics, U. FIG Elite/International Scoring. 2022. Available online: https://usagym.org/pages/events/pages/fig_scoring.html (accessed on 20 March 2022).
99. Gymnastics, B. Scoring Guide. 2022. Available online: <https://www.british-gymnastics.org/scoring-guide> (accessed on 20 March 2022).
100. Neff, C.; Sheth, A.; Furgurson, S.; Tabkhi, H. EfficientHRNet: Efficient Scaling for Lightweight High-Resolution Multi-Person Pose Estimation. *arXiv* **2020**. arXiv:2007.08090.
101. Maji, D.; Nagori, S.; Mathew, M.; Poddar, D. YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, New Orleans, LA, USA, 19–21 June 2022; pp. 2637–2646. Available online: <https://arxiv.org/abs/2204.06806> (accessed on 20 June 2022).