applied sciences

MDPI

Article

Enhancing Human Key Point Identification: A Comparative Study of the High-Resolution VICON Dataset and COCO Dataset Using BPNET

Yunju Lee [1,2,*], Bibash Lama [1], Sunghwan Joo [1] and Jaerock Kwon [3]

[1] School of Engineering, Grand Valley State University, Grand Rapids, MI 49503, USA; lamab@mail.gvsu.edu (B.L.); joos@gvsu.edu (S.J.)
[2] Department of Physical Therapy and Athletic Training, Grand Valley State University, Grand Rapids, MI 49503, USA
[3] Department of Electrical and Computer Engineering, University of Michigan, Dearborn, MI 48128, USA; jrkwon@umich.edu
* Correspondence: leeyun@gvsu.edu; Tel.: +1-616-331-6043

Abstract: Accurately identifying human key points is crucial for various applications, including activity recognition, pose estimation, and gait analysis. This study introduces a high-resolution dataset formed via the VICON motion capture system and three diverse 2D cameras. It facilitates the training of neural networks to estimate 2D key joint positions from images and videos. The study involved 25 healthy adults (17 males, 8 females), executing normal gait for 2 to 3 s. The VICON system captured 3D ground truth data, while the three 2D cameras collected images from different perspectives (0°, 45°, and 135°). The dataset was used to train the Body Pose Network (BPNET), a popular neural network model developed by NVIDIA TAO. Additionally, a comparison entails another BPNET model trained on the COCO 2017 dataset, featuring over 118,000 annotated images. Notably, the proposed dataset exhibited a higher level of accuracy (14.5%) than COCO 2017, despite comprising one-fourth of the image count (23,741 annotated image). This substantial reduction in data size translates to improvements in computational efficiency during model training. Furthermore, the unique dataset's emphasis on gait and precise prediction of key joint positions during normal gait movements distinguish it from existing alternatives. This study has implications ranging from gait-based person identification, and non-invasive concussion detection through sports temporal analysis, to pathologic gait pattern identification.

Keywords: human key point identification; high-resolution dataset; VICON motion capture system; Body Pose Net (BPNET); NVIDIA TAO; COCO2017 dataset; gait recognition; pathologic gait patterns

Citation: Lee, Y.; Lama, B.; Joo, S.; Kwon, J. Enhancing Human Key Point Identification: A Comparative Study of the High-Resolution VICON Dataset and COCO Dataset Using BPNET. *Appl. Sci.* **2024**, *14*, 4351. https://doi.org/10.3390/app14114351

Academic Editor: Lingfeng Shi

Received: 3 April 2024
Revised: 14 May 2024
Accepted: 16 May 2024
Published: 21 May 2024

1. Introduction

The estimation of human poses is a critical task in the field of computer vision, with a multitude of applications across diverse fields, including healthcare, entertainment, and robotics [1]. This process involves the analysis of images or videos in order to accurately determine the positions and orientations of key points or joints within the human body. One of the most crucial aspects of human pose estimation is the capacity to accurately identify the key joint positions of the human body during various movements. By monitoring the locations of these key points over time, researchers can analyze the kinematics of human movements, including joint angles, joint velocities, accelerations, and trajectories. This can provide valuable insights into a person's movement pattern, which can assist clinicians in diagnosing and treating a variety of musculoskeletal conditions [2]. This analysis yields valuable information for the study of human motion, with implications for areas such as sports performance, ergonomics, and rehabilitation. A variety of technologies have been developed to track human key points during different human movements. Optical motion

capture systems have emerged as the most popular, precise, reliable, and widely adopted technology for capturing human movements in clinical settings [3]. This technology is based on the attachment of reflective markers to the skin, a specialized camera system for tracking marker movement, and a kinematic model that converts marker positions into joint positions. However, the process of placing markers on subjects can be both laborious and time-consuming. Furthermore, the marker-based system requires subjects to operate within a confined environment, rendering it unsuitable for a multitude of other applications. In the context of sports applications, alternative systems such as inertial measurement units (IMUs) offer greater flexibility by not constraining subjects to closed environments. Nevertheless, the attachment of sensors to the subject's surface remains a necessary step, which can be impractical and might not accurately represent natural movement conditions. Consequently, the limitations of both marker-based and sensor-based technologies hinder their practicality in clinical and sports settings, as well as in forensic applications such as person identification, action classification, and anomaly detection in body movements during rehabilitation.

The current developments in key point identification focus on marker-less and sensor-less systems, with the objective of overcoming the limitations and challenges associated with these approaches. These advancements leverage computer vision and machine learning techniques, specifically deep learning algorithms, to identify key points accurately and robustly in the human body without the need for physical markers or sensors. These techniques typically use convolutional neural networks (CNNs) to learn features from 2D images and then predict the positions of body joints and landmarks. This approach has the potential to outperform traditional methods that rely on handcrafted features and model-based algorithms [4]. A number of libraries have been developed and proposed for the task of human pose estimation, including OpenPose [5], Dense Pose [6], AlphaPose [7], HRNet [8], and others. These libraries use deep learning frameworks [9] for developing, training, and deploying the model for the task. These frameworks provide ease to researchers trying to validate the estimations and explore the potential of their applicability in clinical or sports biomechanics. However, the accuracy of such trained models is limited by the quantity and quality of the data available for training the model. Only high-resolution and dedicated datasets are useful for developing the desired accurate inference engine, which can be potentially implemented in biomechanical applications.

Currently, there is a scarcity of dedicated datasets in the existing literature that has been specifically designed for the purpose of identifying the key joint positions of the human body during normal gait movements [4]. Therefore, the primary objective of this research was to address this gap by developing a comprehensive dataset that includes data from various subjects performing their normal gait. Furthermore, the proposed dataset was utilized to train a deep learning model, enabling the creation of a 2D estimator that can infer the 2D key joint positions of a person from the 2D images captured using a simple camera system. To achieve this objective, we utilized the VICON motion capture system [10] to collect motion data from a diverse range of subjects in a controlled motion capture environment while they performed normal gait. The objective of the dataset was to encompass a wide range of gait motions and to capture the intricate complexities of human gait in real-world scenarios.

The NVIDIA TAO's BPNET (Body Pose Network), a deep learning framework specifically designed for key joint position identification of a person, was utilized as a part of this research [11]. The selection of the NVIDIA TAO was driven by its extensive adoption and recognition in the computer vision community. This deep learning toolkit is well-regarded for its user-friendly interface and capability to facilitate rapid prototyping and deployment of deep learning models on NVIDIA GPUs [12]. The objective of training the model with our dataset was to validate the accuracy and robustness of the proposed dataset in identifying the key joint positions during different gait movements. The general workflow and entities of the proposed dataset have been demonstrated in Figure 1.

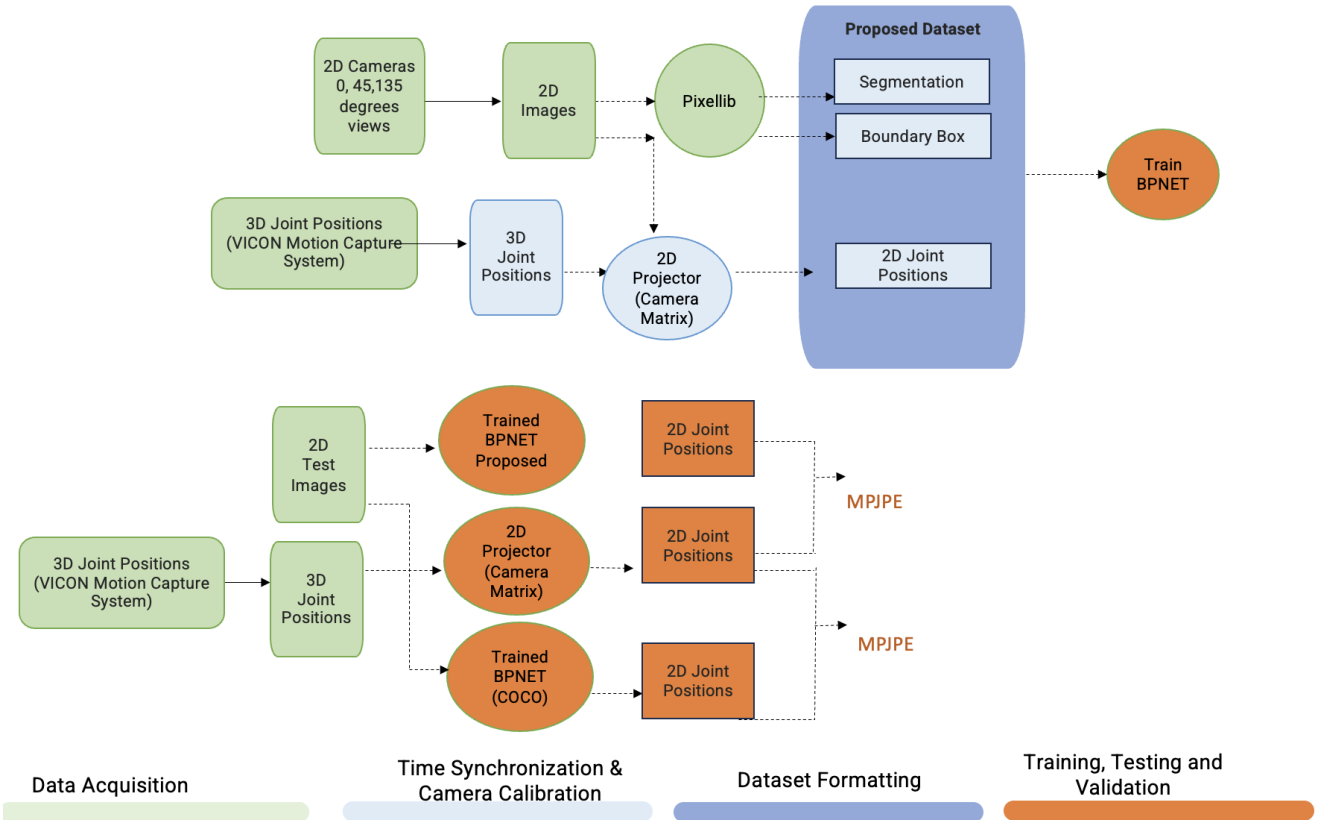


Figure 1. General workflow (the proposed dataset consists of the 2D images captured in different views and 2D key point position acquired from the transformation of VICON recordings, along with the segmentations and bounding box discussed in Section 3.1).

To assess the effectiveness of our approach, we conducted a comparative analysis between the performance of the BPNET trained on our dataset and a model trained on the widely used COCO 2017 (Common Objects in Context) [13] dataset. The COCO 2017 dataset has been extensively utilized in computer vision research for a variety of tasks, including object detection, segmentation, and human key feature identification tasks [5,6,14]. By contrasting the performance of our model with that of the COCO-based model, we evaluated the advantages and potential enhancements achieved through our dataset and training methodology.

This paper's contributions are as follows. Existing datasets for human pose estimation frequently prioritize quantity over quality. While large-scale datasets, such as COCO 2017, have been instrumental in advancing pose estimation models, their focus on vast image collections can introduce limitations. These limitations include inconsistencies in annotation detail and lower image resolution, which can hinder the accuracy of key point prediction for specific tasks such as gait analysis. Furthermore, existing datasets may not fully encompass the intricacies of human movement as effectively as dedicated motion capture systems. This study addresses these limitations by introducing a novel high-resolution dataset specifically designed for accurate key point identification during gait analysis. The dataset employs the precision of the VICON motion capture system to capture detailed three-dimensional ground truth data in conjunction with high-resolution images from multiple perspectives. This combination offers a more comprehensive and accurate representation of human movement than is possible with existing datasets.

2. Related Works

The field of determining human key joint positions has witnessed several notable trends and advancements. Marker- and sensor-based technologies have been widely used and established as the primary methods for capturing and analyzing human motion in fields such as clinical biomechanics, animation, sports, and research [3,15–17].

However, these technologies come with limitations like high operation costs and maintenance requirements, as well as the need for a controlled environment. In addition, potential issues like occlusions and drifts can arise from poor marker and sensor placements. These limitations have spurred researchers to develop more cost-effective and reliable motion capture technologies [18,19]. The advantages, challenges, and limitations of the marker-based and sensor-based motion capture systems have been discussed in [20]. To overcome these limitations, the latest advancements in the field implement machine learning and computer vision techniques, particularly deep learning algorithms, to achieve precise and reliable identification of key points in the human body [21]. This eliminates the need for the physical placement of markers and sensors and does not require the subject to be confined in a closed controlled environment.

Deep learning methods are at the forefront of the field of machine learning and have taken over most industries, including manufacturing [22], finance [23], construction [24,25], medicine [26], criminology [27], computer vision [28], and more. The application of these methods for human key point estimation has also been extensively studied in biomedical and computer-vision-based literature [29,30]. Zheng et al. provides a comprehensive review of recent deep learning-based solutions for 2D and 3D pose estimation, systematically analyzing and comparing them based on their input data and inference procedures [4]. These techniques have shown promising results and are being employed in various biomedical research studies to assess their clinical viability [31–34]. However, these methods have not yet been commercialized, and their implementation in human motion analysis continues to be an active area of research.

Several deep learning frameworks have been developed that provide researchers with sets of tools, algorithms, and abstractions to simplify the development, training, and deployment of deep learning models of interest. The current state-of-the-art frameworks include PyTorch [35], TensorFlow [36], Keras [37], etc. These frameworks can be trained with different forms of datasets based on the purpose intended for the model. Various libraries built on these frameworks intended for 2D and 3D human pose estimation have been proposed in the literature. Openpose, DensePose, Alphapose, and HRNET (High-Resolution Net) are among the most discussed libraries for the marker-less human pose estimation system discussed in the literature [5–7,14]. These libraries have been used to develop the prediction model by utilizing various publicly available datasets.

Various large-scale datasets for training such deep learning models have been proposed and discussed widely in recent years. The state-of-the-art dataset called COCO has proven to be significant for object detection, segmentation, and captioning tasks [13]. It has features like object segmentation, context recognition, and super pixel stuff segmentation. With over 330 K images, 1.5 million object instances, and 80 object categories, it provides a diverse training resource and is utilized to build various popular neural network models like OpenPose, DensePose, etc. Another popular dataset called MPII Human Pose Dataset [38] contains around 25,000 images of people in various everyday activities. It includes detailed annotations for body joint positions. Similarly, Human3.6m [39] stands as a widely used benchmark dataset for 3D human pose estimation in recent years. It provides accurate 3D joint positions for multiple subjects performing various activities. It includes synchronized video and motion capture data, enabling accurate ground truth annotations for human key point positions.

Another noteworthy dataset in the current literature that is specifically designed for human gait movements includes HumanEva [40], Gait3D [41], and CASIA [42]. HumanEva [40] consists of synchronized grayscale and color video sequences with corresponding 3D body poses captured from a motion capture system, featuring four subjects

performing six common actions, accompanied by error metrics for evaluating 2D and 3D pose accuracy, as well as separate training, validation, and testing sets. Gait3D comprises 4000 individuals and encompasses over 25,000 sequences obtained from 39 cameras capturing an unconstrained indoor environment [41]. CASIA, a collection of subsets, notably CASIA-A and CASIA-B, offers gait data from diverse walking scenarios and viewpoints [42]. It includes RGB videos and depth maps, providing visual and depth information for gait analysis. These extensive datasets have sparked significant advancements in 2D pose estimation, resulting in modern motion capture techniques achieving an average error of approximately 20 mm per joint [43].

However, many of these datasets consist of recordings of random images of human actions and may not accurately represent normal human gait. These factors can impact the accuracy of gait recognition models, highlighting the need for improvement in practical applications. Gait-specific datasets like GAIT3D and CASIA, although valuable, still leave room for enhancing the performance of trained frameworks. In a study by Zheng et al. [4], various deep learning models trained with these datasets were explored and discussed in detail. The study compiled several dedicated articles and summarized current advancements in deep learning-based 2D and 3D human pose estimation and concluded that the field suffers limitations of the dedicated database involving complex human movements. Therefore, there is a need to address this gap by developing a comprehensive dataset that includes high-resolution data from various subjects performing their normal gait. Thus, the current study provides a comprehensive high-resolution dataset that was specially prepared for gait analysis using high-resolution data captured using the VICON motion capture system. This proposed dataset can be used to train different deep-learning models for creating a 2D estimator capable of efficiently predicting the 2D key joint position of the person from 2D images.

Furthermore, it is also equally important to consider the choice of deep learning platforms for training the model with the dataset, as well as for model inference and deployment. The selection of the appropriate deep learning platform plays a crucial role in the success of biomedical research applications. These platforms provide flexibility to be fine-tuned with custom datasets to improve accuracy and adapt to specific domains or enhance performance. The results achieved by testing various deep learning models such as OpenPose, MediaPipe, AlphaPose, and HRNet trained with different state-of-the-art datasets have been widely discussed in the literature [5,13,20,34,41,42].

Recently, a study by W. Liu, et al. [44] has shown the mainstream and milestone approaches for these human body presentations since the year 2014 under unified frameworks. In particular, it provides insightful analyses for the intrinsic connections and methods evolution from 2D to 3D pose estimation and analyze the solutions for challenging cases, such as the lack of data, the inherent ambiguity between 2D and 3D, and the complex multi-person scenarios. Next, it summarizes the benchmarks, evaluation metrics, and the quantitative performance of popular approaches. Finally, it discusses the challenges and provides the deep thinking of promising directions for future research. We believe this survey will provide the readers (researchers, engineers, developers, etc.) with a deep and insightful understanding of monocular human pose estimation.

One notable framework for refining key point detection is BPNET [11], developed by NVIDIA TAO. Nvidia TAO (Train, Adapt, Optimize) is a platform for developing and deploying custom artificial intelligence (AI) models in various domains, including healthcare, retail, manufacturing, and robotics. It provides a suite of tools and libraries for end-to-end AI development, from data preparation and model training to deployment and inference. However, despite its popularity, there has been limited research on the potential of NVIDIA TAO for 2D and 3D human pose estimation and its application in identifying human key points and features and subsequent analysis. Furthermore, no evident literature can be found that tested or validated the NVIDIA TAO's BPNET model. In this study, the proposed dataset was implemented to train BPNET for the creation of the 2D inference engine capable to predicting the key joint positions of the person from the 2D images.

BPNET trained with the proposed dataset was then compared with another BPNET model that was trained using the COCO 2017 dataset for comparing the inference accuracy of the proposed dataset.

3. Material and Methods

3.1. Acquisition of 3D Joint Positions

The VICON motion capture system (Nexus 2.14, Vicon Motion Systems Ltd., Oxford, UK) [10] equipped with 16 high-speed, high-resolution cameras was used to acquire the key joint positions of all the participants performing normal gait. A total of forty-eight standard 14 mm reflective markers were placed and secured for all anatomical landmarks provided in the guidelines of the Plug-in Gait (PIG) Full body model [45] in the VICON system. A three-dimensional trajectory of major joint positions including (left and right) L/R ankles, L/R knees, L/R hips, L/R shoulders, L/R elbows, L/R wrists, L/R shoulders, and head was recorded and used for preparation of the dataset. Due to the limitation of the PIG full body model, the position of the L/R eye, L/R ear, and nose was kept the same as the center of mass of head positions.

Furthermore, the 3D trajectories obtained from Nexus 2.14, recorded at a sampling frequency of 120 Hz, were downsampled to 60 Hz to match the frequency of the 2D camera system. A MATLAB R2023a (MathWorks, Natick, MA, USA) spline interpolation function [46] was implemented for this task to match the frame capture rate of 2D cameras used in conjunction with the VICON motion capture system at 60 Hz.

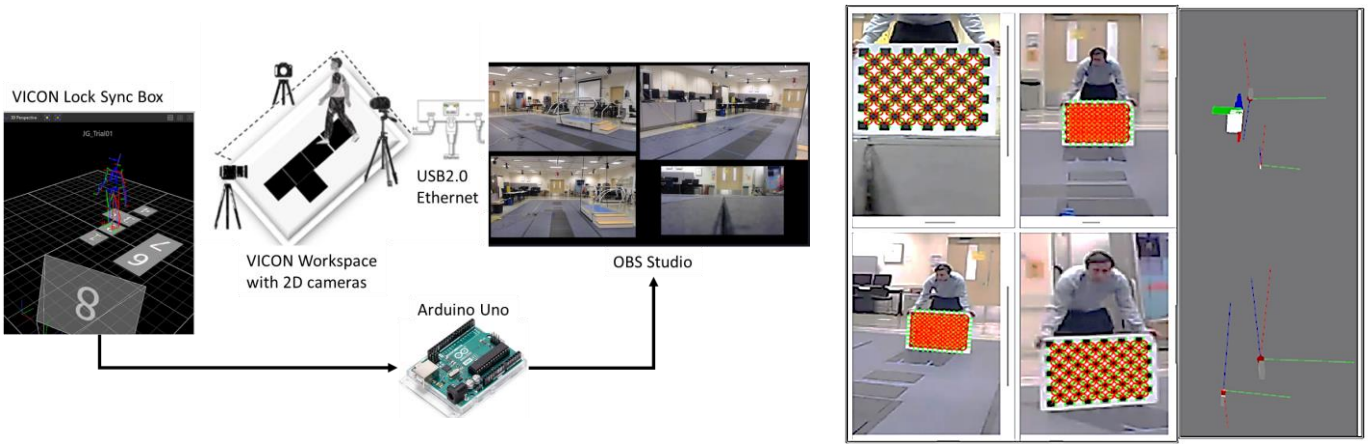
3.2. Acquisition of 2D Images

Three different 2D cameras (C922 Pro HD Stream Webcams, San Jose, CA, USA) were positioned at angles of 0, 45, and 135 degrees with respect to the subjects performing the gait. These three camera systems were first calibrated and synchronized with the VICON motion capture system before the recordings were captured. The detailed process of time-synchronization and calibration is discussed in Section 3.3. Three different camera systems recorded 2D video simultaneously at 60 Hz for every trial. The images were retrieved from the recorded videos for dataset preparation.

3.3. Time-Synchronization and Camera Calibration

Accurately projecting the 3D ground truth trajectories obtained from the VICON motion capture system into the simultaneously captured 2D images was a crucial requirement for this study. This task involved addressing two major challenges. The first challenge was synchronizing the images from the camera system with the VICON motion capture system. Open Broadcaster Software (OBS version 29.0.2) [47], an open-source software, was employed to synchronize the three differently oriented 2D cameras. OBS allowed for the simultaneous recording of multiple cameras using a single system, minimizing, or eliminating time lapses between different recordings.

To ensure accurate synchronization between OBS and the VICON motion capture systems, a trigger system was developed using Arduino Uno and VICON Lock Studio in Python [48]. The VICON Lock Studio, an integrated hardware for synchronizing analog or digital devices, delivered a digital signal as soon as the VICON system started recording. Arduino Uno detected this signal and triggered the hotkeys to initiate the OBS recordings. To further ensure precise synchronization, the code recorded the time stamp simultaneously for both the VICON motion capture system and the 2D camera recordings. A schematic of the trigger system is presented in Figure 2.



(**a**) (**b**)

Figure 2. Time synchronization and calibration of the VICON system and 2D cameras. (**a**) VICON Trigger System for time-synchronization. (**b**) 2D Cameras calibrated using checkerboard to project 3D ground truth data into the image 2D frame.

The second challenge involved formulating an accurate transformation matrix for converting 3D coordinates to 2D projection values. The calibration method suggested by Zhang [49] was employed to calculate the required transformation matrix for each respective synchronized 2D image. An 800 × 600 mm checkerboard (11 columns, 8 rows spaced at 60 mm) was used during the calibration process. The 2D projection for all images was determined by the following Equation (1):

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} fx & 0 & cx \\ 0 & fy & cy \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r11 & r12 & r13 & tx \\ r21 & r22 & r23 & ty \\ r31 & r32 & r33 & tz \end{bmatrix} \begin{bmatrix} Xw \\ Yw \\ Zw \\ 1 \end{bmatrix} \quad (1)$$

In the equation, (u, v) represents the projected points, (Xw, Yw, Zw) represents the 3D ground truth data from the VICON system, and s represents the scaling factor of the captured images. The first part of the transformation matrix comprising fx, fy, cx, and cy formulates the camera intrinsic matrix and denotes horizontal and vertical focal lengths and camera principal point coordinates, respectively. The second part of the equation comprised rij, and t formulates the camera extrinsic matrix. These parameters denoted the relative rotation and translation of the images with respect to the desired global origin.

To acquire the desired calibration matrix, 50–100 images were captured in different orientations and positions from all the synchronized three differently oriented (0, 45, and 135 degrees) 2D cameras and one additional camera centered at the global origin of the VICON motion capture system. These images were then optimized to obtain proper intrinsic and extrinsic matrices for 2D transformation. The optimization was carried out using the software Calib.io (version v1.6.2a. Calib.io ApS, Svendborg, Denmark) [50].

3.4. Model Training

BPNET, part of the TAO (Train, Adapt and Optimize) Toolkit by NVIDIA, was trained using the proposed dataset. BPNET, pre-trained on large-scale datasets, enables the model to capture general representations of a human pose from a vast amount of data, reducing the need to start training from scratch. Although it is unclear whether the network architecture (including the number of layers) is publicly available, it provided a strong foundation for further fine-tuning and adaptation to specific datasets or domains, saving time and computational resources. The images and the respective 2D projections of the key joint positions were used to train the images, and the new set of images and their respective projections were used for the validation of the trained model. Therefore, for the proposed

dataset, we used 23,741 as training images and 2224 as validation images, while for the commonly used dataset—the Common Objects in Context (COCO) 2017, COCO dataset—we used 118,287 as training images and 5012 as validation images, which included 80 object categories of everyday settings such as people, animals, furniture, etc. [13]. The proposed dataset consisted of approximately 24,000 images obtained from subjects' normal walking data. We excluded several images with occlusions for different key positions from the data analysis. The training was accomplished using NVIDIA TAO Toolkit v. 4.0 with RTX A6000 GPU (NVIDIA, Santa Clara, CA, USA).

3.5. Model Evaluation and Performance Metric

Two trained models were compared based on the model evaluation and mean per joint position error (MPJPE) values for their quantitative analysis. The model evaluation results were based on the inference results from the selected validation results. TensorRT (tensor runtime), a deep learning inference optimizer and runtime library developed by the NVIDIA engine, was used to acquire the optimized inference results. It helped in optimizing the inference process by applying various techniques such as layer fusion, precision calibration, and kernel autotuning, resulting in reduced memory consumption, increased throughput, and minimized latency during the inference [12]. The evaluation results were compared based on average precision computed over the different thresholds of IoU (intersection over union). These metrics are defined as

- Intersection over union (IOU): values range from 0 to 1, where a value of 1 indicates a perfect match between the predicted and ground truth bounding boxes, while a value of 0 indicates no overlap at all.
- Precision: measures the proportion of true positive detections out of all positive predictions made by the system.
- Recall: measures the proportion of true positive detections out of all relevant items present in the dataset.

Furthermore, mean per joint position error (MPJPE) values were computed for both models based on the inference results for the test data sample. The mean per joint position error (MPJPE) was calculated by using the formula [29].

$$\mathrm{E}_{\mathrm{MPJPE}}\ (f,\ S) = \frac{1}{Ns}\sum_{i=1}^{Ns} ||m_{f,S}^{(f)}(i) - m_{gt,S}^{(f)}(i)|| \quad (2)$$

where N_s is the number of joints in skeleton S, $m_{f,s}$ is the ground truth position of the ith joint, and $m_{gt,s}$ represent the estimated position of joint i. The inference results from the model trained with the proposed dataset were then validated for its accuracy by comparing with the results from the model trained with the COCO 2017 dataset. Two different trained models shown in Section 3.4 were tested with the approximately 2510 number of images to compute MPJPE that was selected from three subjects' data for both two models.

4. Experiment

4.1. Participants

The experiment involved the participation of 25 healthy individuals, comprising 17 males and 8 females. All participants fell within the age range of 20 to 30 years and provided written informed consent before the study commenced. The experimental protocol was approved by both the Institutional Review Board (IRB) and the university's Graduate Committee, ensuring adherence to ethical guidelines and research standards.

4.2. Data Collection

Prior to each trial, meticulous calibration and time synchronization of the three 2D cameras and the VICON motion capture system were carried out. This calibration process aimed to obtain an accurate transformation matrix, aligning the VICON coordinate system with the camera coordinate system. The detailed process of camera calibration and time

synchronization was discussed in Section 3.3. The images of the checkerboard were then used to acquire the optimized camera calibration matrix using Calib.io. The general workflow and experiment setup for the data collection process are demonstrated in Figure 3.

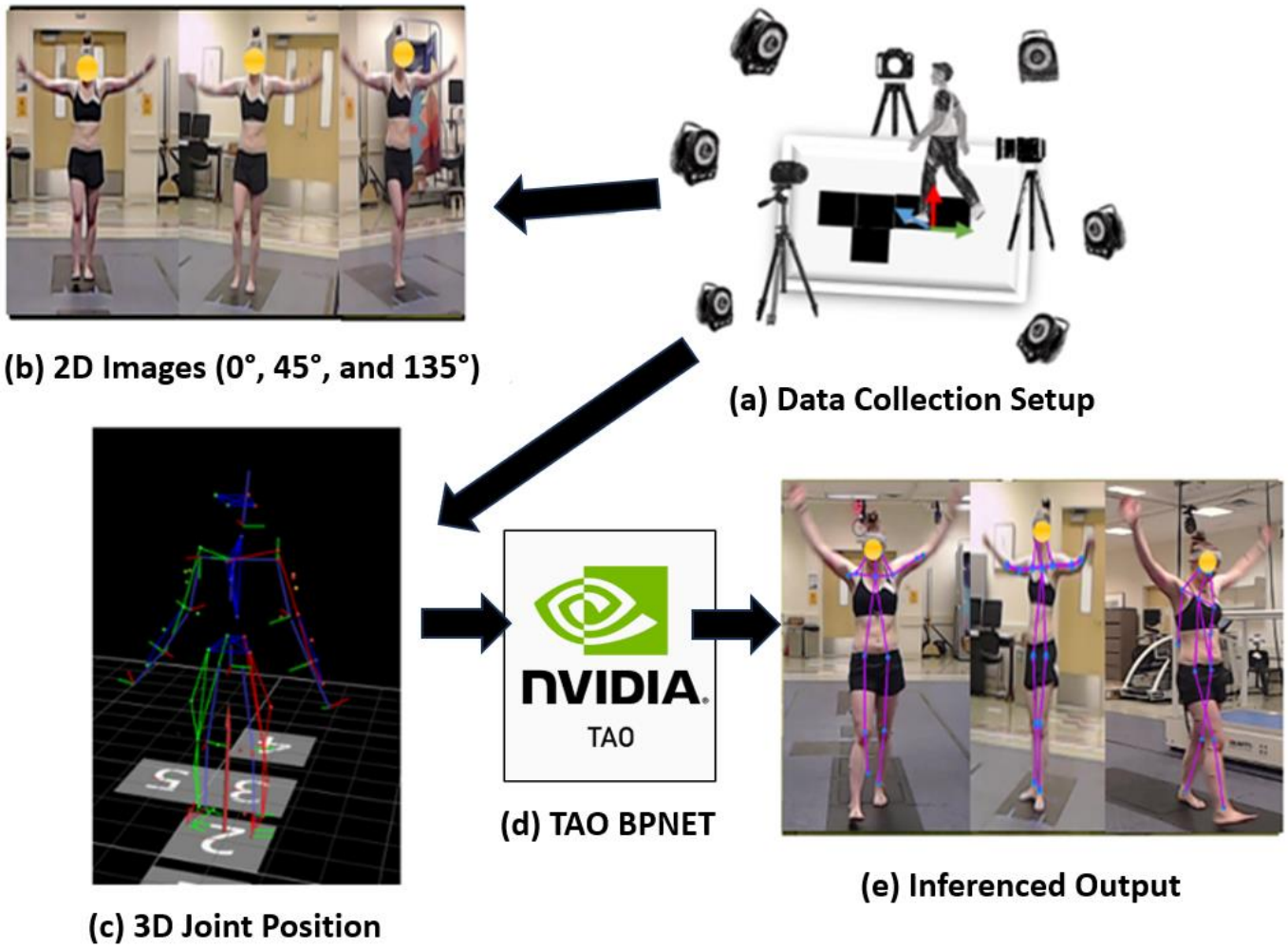


Figure 3. The general workflow and experiment setup for the data collection process (**a**) Data collection using VICON Motion Capture and three 2D cameras; (**b**) 2D images from different views; (**c**) 2D ground truth data from VICON system; (**d**) BPNET is trained with the collected data; and (**e**) BP inference outputs with 2D joint positions in 2D images.

The actual data collection process began after the completion and verification of the camera calibration process. Every experiment started with a collection of detailed information regarding general anatomical features such as age, body mass, height, and leg length, recorded for each participant, followed by the strategical placement of markers on all anatomical landmarks following the guidelines provided by the PIG full body model within the VICON system. These markers served as reference points for capturing precise movement and positioning data during the experiment.

Once the preparation phase was complete, the participants were instructed to walk at their usual speed within the designated workspace of the VICON system. Each participant performed five distinct trials, with each trial lasting between 2 to 3 s. As a result, a vast collection of 23,741 different view images was acquired throughout the experiment. To ensure an unbiased and random distribution of data, a subset of 2224 images was randomly selected for the purpose of model validation. This validation subset served as an independent evaluation set, enabling the assessment of the model's performance on unseen data, as well as validating its ability to generalize effectively.

4.3. Dataset Preparation and Model Training

The projections of 2D key joint positions were computed and collected systematically alongside their corresponding 2D images using the optimized calibration matrix discussed in Section 3.3. In addition to the key joint positions, to train the BPNET, a segmentation and boundary box were required as essential parts of the dataset. These entities were acquired by using an open-source application called Pixellib [51], which allowed for the retrieval of the segmentation and boundary box coordinates from the batched images recorded in the experiment. A sample of segmentation and boundary boxes obtained from Pixellib is demonstrated in Figure 4. Following the validation of segmentation and boundary box from the captured images, coordinates were acquired for every 2D image of the trial.

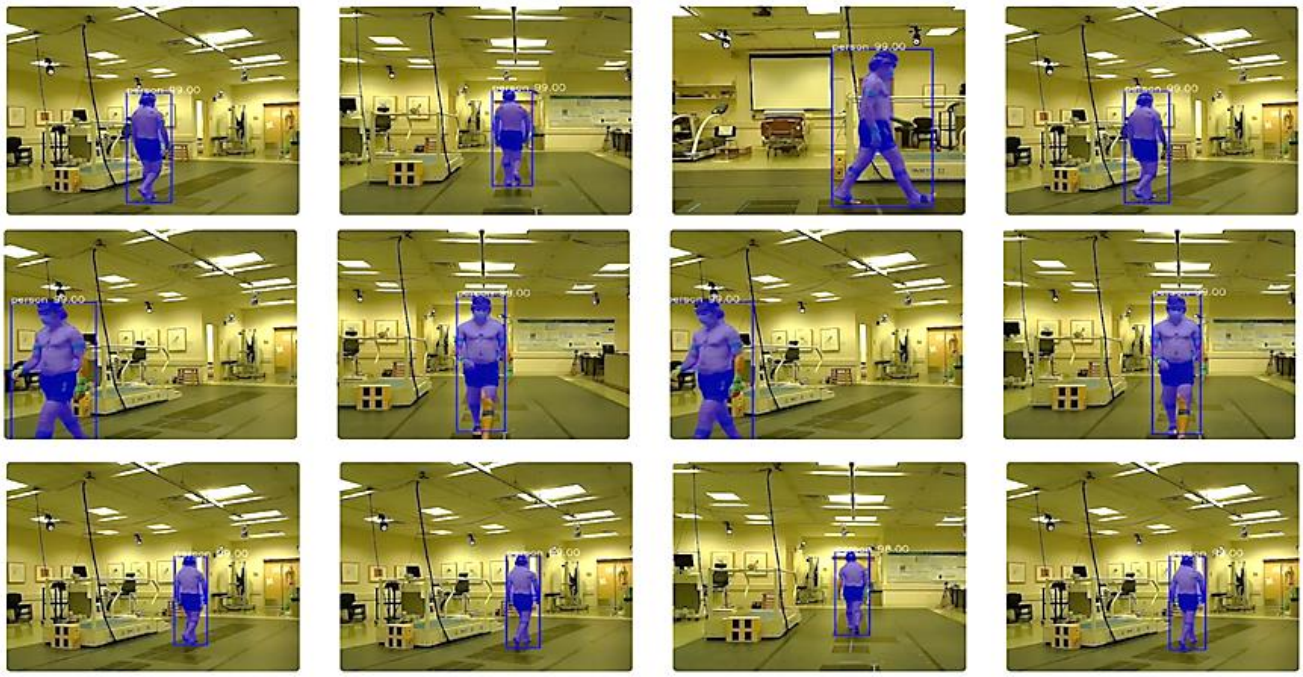


Figure 4. Segmentation and boundary box acquisition from Pixellib.

Another critical factor to consider within the present data collection methodology was the utilization of a different camera angle during various trials—specifically, a 135 degree angle instead of the customary 90 degree angle. This was because the 90 degree view camera system presented its own set of challenges, particularly in terms of camera calibration. Subsequently, this calibration difficulty resulted in the unavailability of VICON ground truth data for the 2D images captured with this camera configuration. Moreover, these experiments were initiated prior to the currently proposed methodology, which includes commercialized camera calibration using calib.io and time synchronization, as discussed in Section 3.3.

To acquire accurate ground truth data for these 2D images, a solution was pursued through the integration of an open-source code for human pose estimation called MediaPipe Pose [52]. The coordinates extracted from the MediaPipe Pose framework were utilized as a substitute for the absent VICON ground truth data, effectively providing an alternative reference for the 2D images captured under the 135 degree camera angle configuration. This strategic approach facilitated the continuation of the analysis despite the initial challenges posed by the camera calibration constraints.

5. Results

This study utilized three distinct datasets for training, validation, and testing. The training dataset consisted of 23,741 images for the proposed model, in comparison to 118,287 images from the COCO 2017 dataset. The validation dataset comprised a total of 2224 images across three angles (0, 45, and 135 degrees) for performance evaluation, with 5012 images from the COCO 2017 dataset serving as the comparative evaluation. Finally, in order to compute the MPJPE for these two models, we acquired the inference results with the additional trial of three subjects, which corresponded to the 2510 images.

The two trained models, one trained with the proposed dataset and the other trained with COCO 2017 dataset, were evaluated using the validation datasets that were set aside during the training of each respective model. The sample inference results from the model trained with the proposed dataset are depicted in Figure 5.

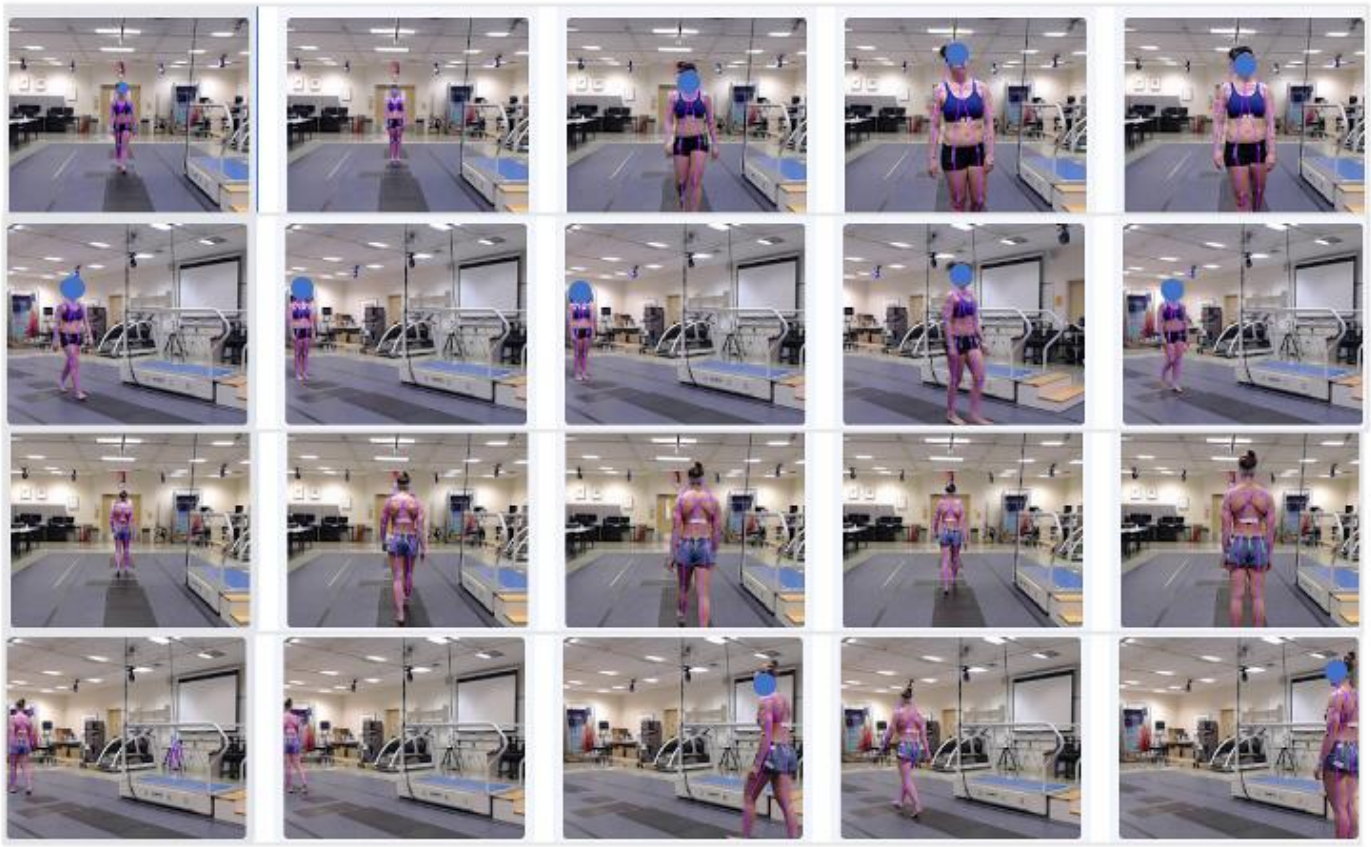


Figure 5. Inference results from the BPNET trained with the proposed dataset.

The model trained with the COCO 2017 dataset achieved an average precision (AP) of 0.547 and an average recall (AR) of 0.471 for all object sizes, as well as IoU (intersection over union) thresholds between 0.50 and 0.95. This resulting value was significantly lower than what was achieved with the model trained with the proposed dataset. The second model trained with a proposed dataset achieved a higher overall AP of 0.565 and AR of 0.617 for all object sizes and IoU thresholds between 0.50 and 0.95.

Furthermore, for individual IoUs at 0.5 and 0.75, the model trained with the proposed dataset recorded noticeably higher results than the COCO-trained dataset, except for the AP at the 0.75 IoU threshold, which was slightly higher for the COCO-trained dataset (0.692) than the proposed dataset (0.617). The evaluation results of two trained models are depicted in Table 1.

Table 1. Evaluation results between COCO (5012 images) and the proposed dataset (2224 images).

	COCO Dataset		Proposed Dataset	
IoU (Intersection Over Union) Threshold	**Average Precision (AP)**	**Average Recall (AR)**	**Average Precision (AP)**	**Average Recall (AR)**
0.5	0.668	0.458	0.873	0.915
0.75	0.413	0.692	0.527	0.582
0.5:0.95	0.547	0.471	0.565	0.617

Secondly, the two trained models were tested using data collected for additional three subjects performing normal gait inside the VICON workspace. The subjects performed normal gait for 2~3 s, while the three different views camera recorded images of the subject. The ground truth data were also simultaneously recorded using the VICON motion capture system for comparison. The inference results from the model trained with the proposed dataset are described in Figure 5.

For both the trained models, the Euclidean distance between the estimated joint position and the ground truth joint position was computed for twelve major joints, namely, R/L shoulder, R/L elbow, R/L wrist, R/L hip, R/L knee, and R/L ankle. A joint level MPJPE was computed by taking the means of the distances across all frames. Any undetected or falsely detected key joint position due to image quality and occlusions from the camera view were discarded from the results before computing MPJPE. Consequently, approximately 2510 images were selected from both datasets for the purpose of computing the MPJPE as the testing set.

The overall MPJPE values recorded for the COCO trained model were 3.759 mm (±1.795), while the overall MPJPE values for the proposed dataset were 3.211 mm (±1.730), as depicted in Figure 6.

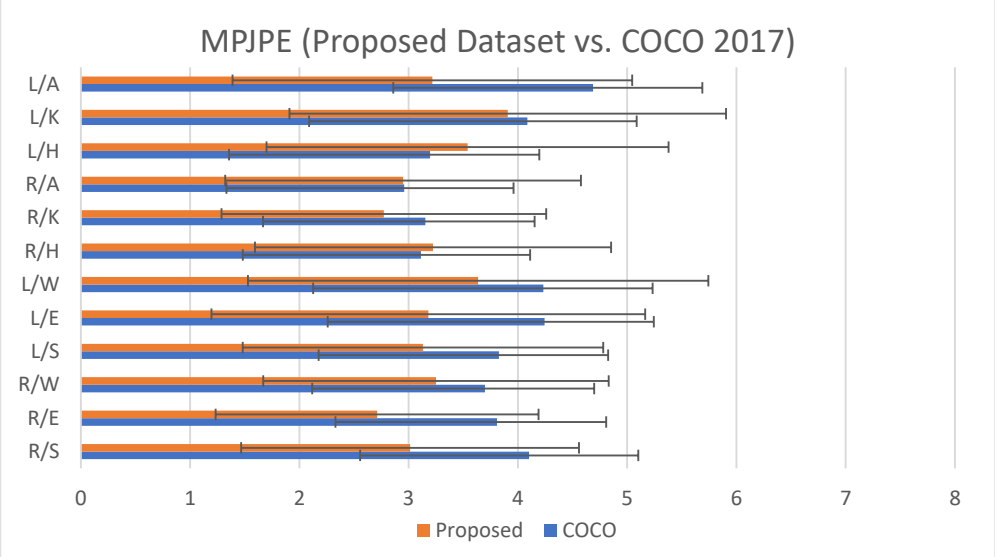


Figure 6. Comparison results for individual joints as MPJPE. Two different models trained with the proposed dataset (23,741 images) vs. COCO 2017 (118,287 images) were tested with approximately 2510 images for calculating MPJPE. The abbreviations are as follows: L/A: left ankle, L/K: left knee, L/H: left hip, R/A: right ankle, R/K: right knee, R/H: right hip, L/W: left wrist, L/E: left elbow, L/S: left shoulder, R/W: right wrist, R/E: right elbow, R/S: right shoulder.

On an individual joint level, the MPJPE values for the proposed dataset were comparatively lower than the COCO dataset, except for the left and right hip, with differences of 0.111 mm and 0.343 mm, respectively. Both models recorded comparatively higher MPJPE values for upper body joints, with the highest error recorded for the left wrist with a value of 5.364 mm for the model trained with the COCO dataset and 4.868 mm for the model trained with the proposed dataset. For lower body joints such as the hip, knee, and ankle, the proposed dataset was able to achieve an MPJPE value of 3.268 mm (±1.735), which was lower by 0.265 mm compared to the COCO-trained model. The overall MPJPE value for the lower and upper body is demonstrated in Table 2.

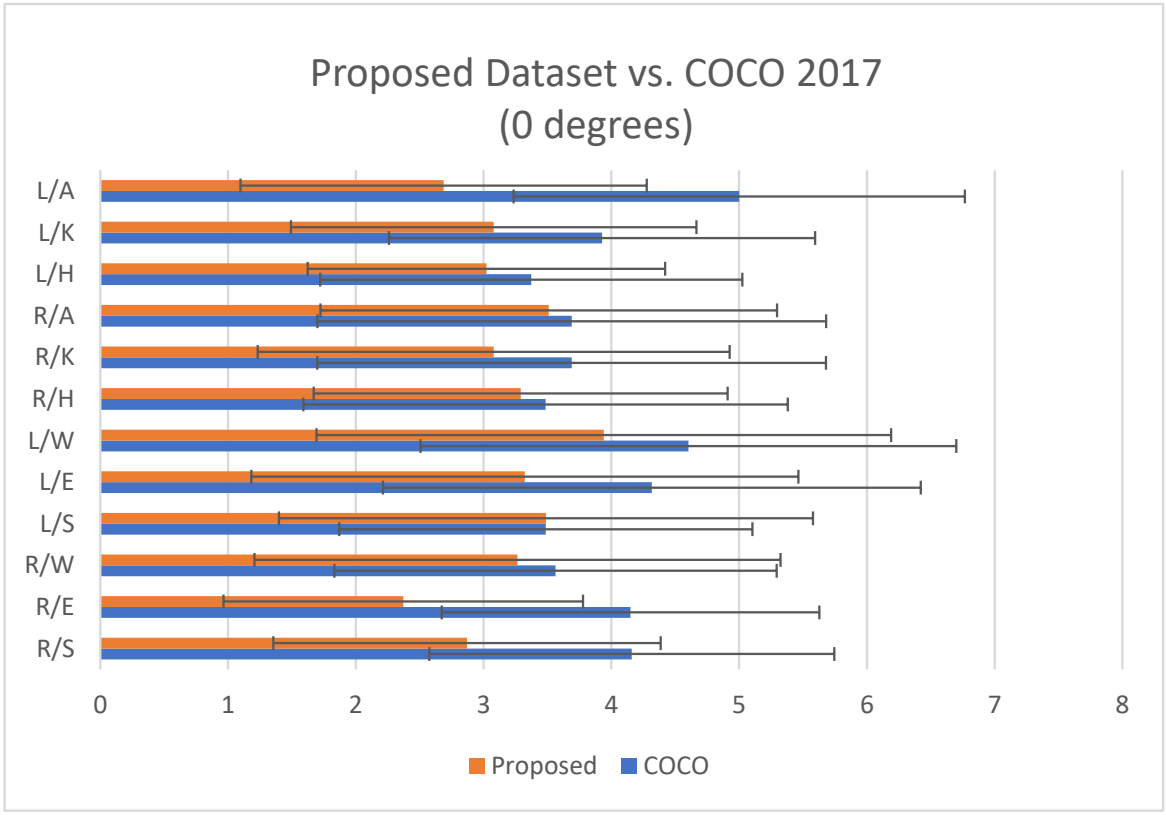
Table 2. Overall mean per joint position error (MPJPE). Two different models trained with the proposed dataset (23,741 images) vs. COCO (118,287 images) were tested with approximately 2510 images for calculating MPJPE.

	MPJPE (mm)		
Dataset	**Upper Body**	**Lower Body**	**Overall**
Proposed Dataset	3.154 ± 1.724	3.268 ± 1.735	3.211 ± 1.730
COCO Dataset	3.985 ± 1.823	3.533 ± 1.766	3.759 ± 1.795

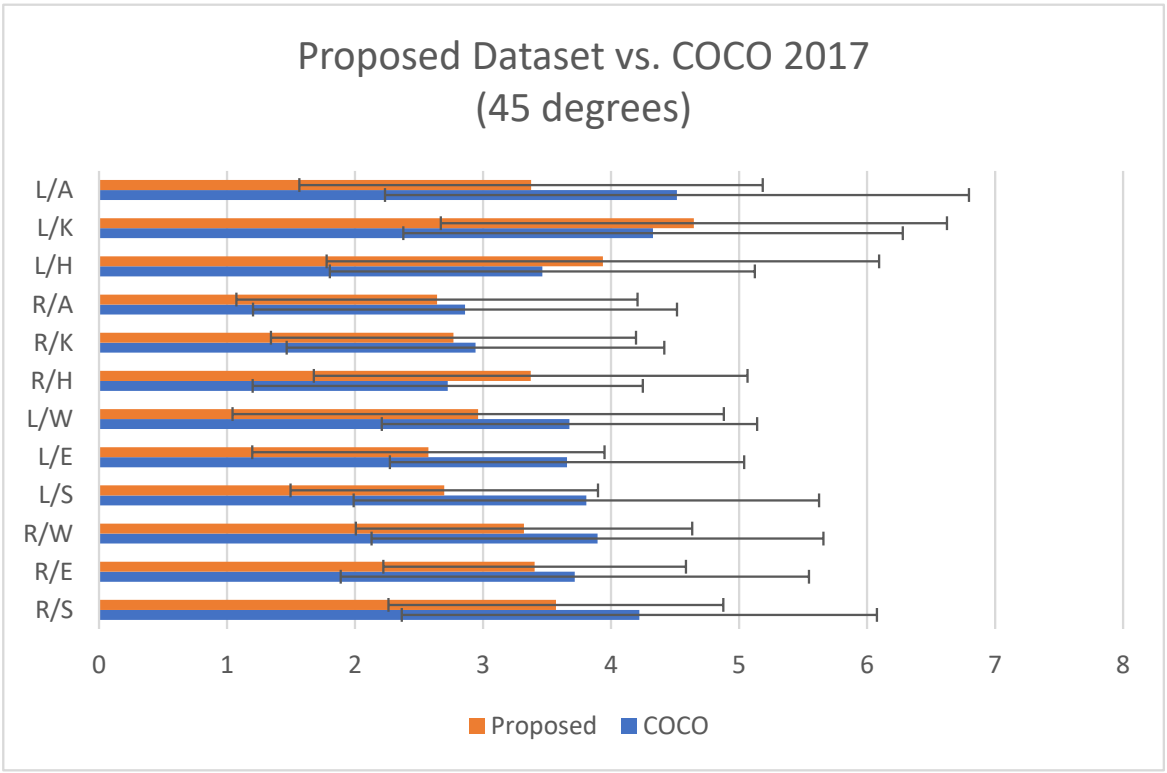
Furthermore, the MPJPE values for images captured in different views were also computed. The overall MPJPE values for the COCO-trained model were observed as being higher by approximate values of 0.793 mm, 0.378 mm, and 0.624 mm than the model trained with the proposed dataset for all 0, 45, and 135 degree views, respectively. However, at the individual joint level, slightly higher MPJPE values were recorded for the left and right hip at 45 and 135 degrees, respectively, and for the left knee at 45 degrees. The detailed MPJPE value recorded is represented in Table 3 and Figure 7a–c.

Table 3. Mean per joint position error (MPJPE) for different camera view images. The MPJPE was calculated by testing with approximately 2510 images for both models.

		MPJPE (mm)		
	Views	**Upper Body**	**Lower Body**	**Overall**
Proposed	0	3.210 ± 1.910	3.111 ± 1.638	3.160 ± 1.774
	45	3.272 ± 1.578	3.457 ± 1.772	3.272 ± 1.578
	135	3.211 ± 1.755	3.102 ± 1.584	3.156 ± 1.670
COCO	0	4.046 ± 1.769	3.861 ± 1.828	3.954 ± 1.798
	45	3.650 ± 1.722	3.472 ± 1.758	3.457 ± 1.722
	135	4.173 ± 1.911	3.387 ± 1.617	3.780 ± 1.764

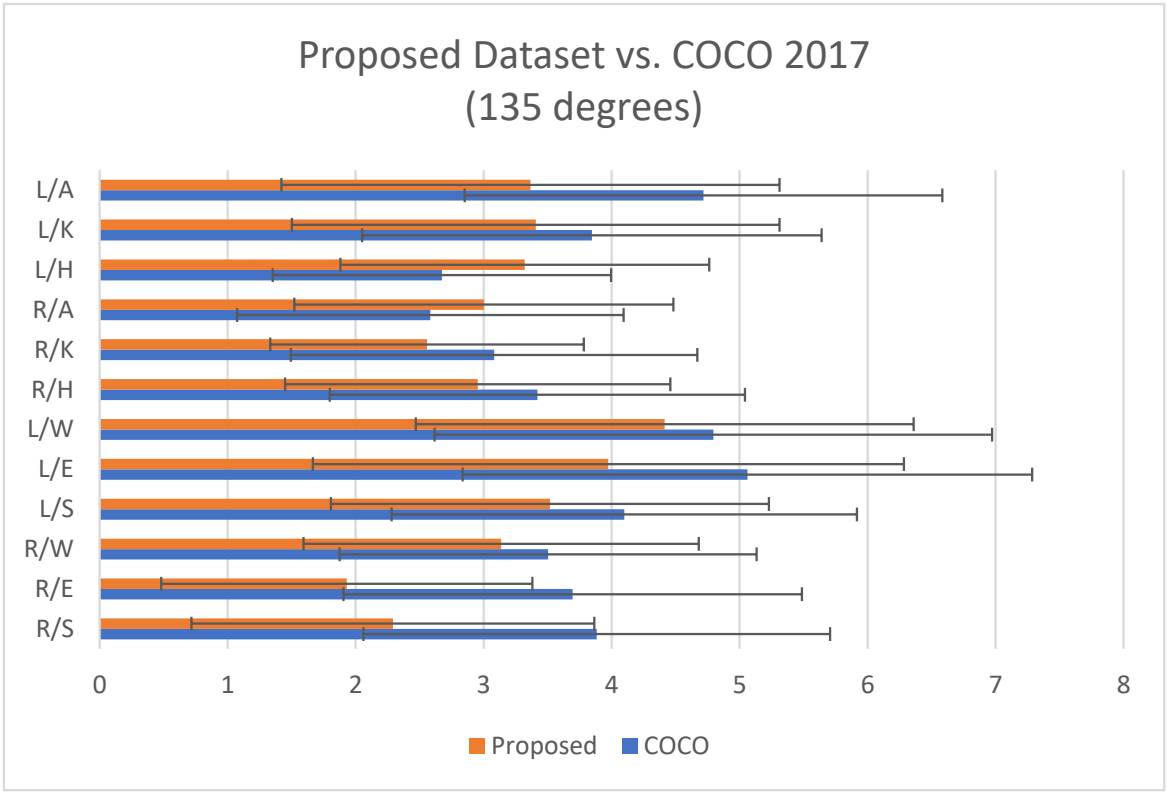


(**a**)



(**b**)

Figure 7. *Cont*.



(**c**)

Figure 7. (**a**) Comparison results of MPJPE for different views between the proposed test images at 0 degrees vs. COCO test images (both approximately, 830 images). (**b**) Comparison results of MPJPE for different views between the proposed test images at 45 degrees vs. COCO test images (both approximately, 830 images). (**c**) Comparison results of MPJPE for different views between the proposed test images at 135 degrees vs. COCO test images (both approximately, 830 images). The abbreviations are as follows: L/A: left ankle, L/K: left knee, L/H: left hip, R/A: right ankle, R/K: right knee, R/H: right hip, L/W: left wrist, L/E: left elbow, L/S: left shoulder, R/W: right wrist, R/E: right elbow, R/S: right shoulder.

6. Discussion

High-Resolution Gait Dataset vs. COCO in BPNET Training

This study compares the performance of two BPNET models trained on distinct datasets: the proposed high-resolution gait dataset and the widely used COCO 2017 dataset. The evaluation focused on the models' accuracy in identifying key body joint positions from 2D images. The findings revealed that the model trained on the proposed dataset consistently achieved superior performance compared to the COCO-trained model. This was evident in two key metrics, namely, average precision (AP) and average recall (AR), across various intersection over union (IoU) thresholds. Notably, the proposed dataset achieved higher AP and AR for most IoU thresholds, signifying its effectiveness in pinpointing key joint locations. Moreover, the model trained on the proposed dataset exhibited lower mean per joint position error (MPJPE) values for the majority of test subjects, further substantiating its accuracy.

Upon closer examination of the MPJPE results, however, it became evident that there were slightly higher error values for specific joints (left/right hips and left knee) at particular camera angles (45 and 135 degrees). These elevated error rates can be attributed to occlusions caused by the camera positioning at those angles. This observation highlights the necessity for further improvement in the model's and dataset's ability to handle occlusions effectively.

The proposed dataset demonstrably outperformed the COCO dataset despite having only a quarter of the image count. This highlights the significance of our research, as a gait-focused dataset leads to substantial reductions in computational cost and training time. The superiority of the proposed dataset stems from several factors:

- **High-resolution VICON data:** The proposed dataset leverages accurate 2D key joint positions captured using the high-resolution VICON motion capture system. This stands in stark contrast to the COCO dataset, where key point locations are manually annotated, potentially leading to lower accuracy.
- **Specificity for gait analysis:** The proposed dataset focuses on controlled normal gait patterns, minimizing external factors like occlusions, complex backgrounds, and lighting variations prevalent in COCO. The controlled environment allows for a more precise evaluation of BPNET, resulting in superior accuracy.
- **Alignment with study objectives:** The proposed dataset is specifically tailored to capture human motion and gait, aligning directly with the study's objectives of human key point identification. In contrast, the diverse object categories and poses in COCO make it less suitable for our specific research focus.

However, it is important to recognize the constraints inherent to the proposed dataset. The dataset was primarily designed for normal gait motions and may not perform optimally for other random and rapid human movements. Furthermore, the dataset's recordings are confined to controlled laboratory environments, which may result in the observed results being sensitive to occlusions encountered in real-world settings. It is therefore recommended that these limitations be taken into account when applying the dataset to scenarios involving non-standard gait movements or settings with significant occlusions.

It is also important to note that the current dataset consists of two-dimensional images and lacks actual ground truth data from the VICON motion capture system. To address this, MediaPipe Pose was employed, as discussed in Section 4.3, in an attempt to compensate for the lack of ground truth data. It is important to note that the absence of ground truth data might have influenced the results. The dataset obtained through the current methodology has the potential to enhance the accuracy of the model in future training and model development.

7. Conclusions

This study presents a novel dataset for human gait analysis, meticulously generated using the VICON motion capture system. The dataset's accuracy in capturing key joint positions during various gait movements unlocks valuable contributions in human pose estimation.

This research has significant implications for various domains. It offers a powerful tool for biomechanics research, enabling more detailed analyses of human movement and potentially informing advancements in areas like rehabilitation and prosthetics design. Furthermore, the capacity to accurately identify key points during gait paves the way for the exploitation of human biometrics for identification purposes, with potential applications in privacy-preserving security measures.

The proposed framework also holds promise for non-invasive concussion detection in sports by analyzing temporal changes in gait patterns. This could revolutionize athlete safety protocols. Moreover, the dataset's ability to capture subtle variations in gait patterns could be instrumental in identifying pathological gait patterns associated with various medical conditions, potentially improving patient outcomes.

In essence, this research provides a reliable and diverse dataset, opening the door to more accurate analyses in gait analysis and related fields. The potential for real-world applications, such as non-invasive concussion detection and privacy-preserving person identification, further highlights the significance of this work. This study demonstrates the power of deep learning frameworks for biomechanics research, paving the way for further exploration and advancements in the field.

Author Contributions: Conceptualization, J.K., S.J. and Y.L.; methodology, B.L., J.K. and Y.L.; software, B.L., J.K. and Y.L.; validation, B.L., J.K., S.J. and Y.L.; formal analysis, B.L.; investigation, B.L., J.K. and Y.L.; resources, J.K. and Y.L.; data curation, B.L., J.K. and Y.L.; writing—original draft preparation, B.L., J.K., S.J. and Y.L.; writing—review and editing, B.L., J.K., S.J. and Y.L.; visualization, B.L.; supervision, Y.L.; project administration, J.K. and Y.L.; funding acquisition, J.K. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the ADVANCE Grant Program by Michigan Economic Development Corporation (MEDC), grant number AGR2022-00097, and Grand Valley State University, the Center for Scholarly and Creative Excellence (CSCE), Research Grant (Lee 2021).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of Grand Valley State University (19-323-H-GVSU, approval on May 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: We thank the study participants for their exceptional cooperation.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-time human pose recognition in parts from single depth images. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 21–23 June 2011; pp. 1297–1304. [CrossRef]
2. Cooper, R.A.; Quatrano, L.A.; Stanhope, S.J.; Cavanagh, P.R.; Miller, F.; Kerrigan, D.C.; Esquenazi, A.; Harris, G.F.; Winters, J.M. Gait Analysis in Rehabilitation Medicine: A Brief Report: 1. *Am. J. Phys. Med. Rehabil.* **1999**, *78*, 278–280. [CrossRef] [PubMed]
3. Colyer, S.L.; Evans, M.; Cosker, D.P.; Salo, A.I.T. A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System. *Sports Med. Open* **2018**, *4*, 24. [CrossRef] [PubMed]
4. Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; Shah, M. Deep Learning-Based Human Pose Estimation: A Survey. *arXiv* **2022**, arXiv:2012.13392. Available online: http://arxiv.org/abs/2012.13392 (accessed on 16 April 2023). [CrossRef]
5. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv* **2019**, arXiv:1812.08008. Available online: http://arxiv.org/abs/1812.08008 (accessed on 29 March 2023). [CrossRef] [PubMed]
6. Güler, R.A.; Neverova, N.; Kokkinos, I. DensePose: Dense Human Pose Estimation in The Wild. In Proceedings of the CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018. [CrossRef]
7. Fang, H.S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.L.; Lu, C. AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 7157–7173. [CrossRef]
8. Sun, K.; Geng, Z.; Meng, D.; Xiao, B.; Liu, D.; Zhang, Z.; Wang, J. Bottom-Up Human Pose Estimation by Ranking Heatmap-Guided Adaptive Keypoint Estimates. *arXiv* **2020**, arXiv:2006.15480. Available online: http://arxiv.org/abs/2006.15480 (accessed on 14 June 2023).
9. Nguyen, G.; Dlugolinsky, S.; Bobák, M.; Tran, V.; López García, Á.; Heredia, I.; Malík, P.; Hluchý, L. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: A survey. *Artif. Intell. Rev.* **2019**, *52*, 77–124. [CrossRef]
10. Vicon Nexus 2.14. Vicon Motion Systems Ltd. Available online: https://www.vicon.com/software/nexus (accessed on 10 May 2022).
11. Nvidia Tao Toolkit (Bpnet). NVIDIA Corporation. 2021. Available online: https://developer.nvidia.com/tao-toolkit (accessed on 10 May 2022).
12. NVIDIA Corporation. NVIDIA TAO Toolkit. 2021. Available online: https://developer.nvidia.com/tao-toolkit (accessed on 10 May 2022).
13. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014. [CrossRef]
14. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5686–5696. [CrossRef]
15. Naeemabadi, M.; Dinesen, B.; Andersen, O.K.; Hansen, J. Influence of a Marker-Based Motion Capture System on the Performance of Microsoft Kinect v2 Skeleton Algorithm. *IEEE Sens. J.* **2019**, *19*, 171–179. [CrossRef]

16. Gaertner, S.; Do, M.; Asfour, T.; Dillmann, R.; Simonidis, C.; Seemann, W. Generation of Human-like Motion for Humanoid Robots Based on Marker-based Motion Capture Data. In Proceedings of the ISR 2010 (41st International Symposium on Robotics) and ROBOTIK 2010 (6th German Conference on Robotics), Munich, Germany, 7–9 June 2010; pp. 1–8.
17. Wirth, M.A.; Fischer, G.; Verdú, J.; Reissner, L.; Balocco, S.; Calcagni, M. Comparison of a New Inertial Sensor Based System with an Optoelectronic Motion Capture System for Motion Analysis of Healthy Human Wrist Joints. *Sensors* **2019**, *19*, 5297. [CrossRef]
18. Kadam, R.; Pawar, S.N. Development of Cost Effective Motion Capture System based on Arduino. In Proceedings of the 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 11–13 March 2020; pp. 1–6. [CrossRef]
19. Parks, M.T.; Wang, Z. Ka-Chun Siu, Current Low-Cost Video-Based Motion Analysis Options for Clinical Rehabilitation: A Systematic Review. *Phys. Ther.* **2019**, *99*, 1405–1425. [CrossRef]
20. Sharma, S.; Verma, S.; Kumar, M.; Sharma, L. Use of Motion Capture in 3D Animation: Motion Capture Systems, Challenges, and Recent Trends. In Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 14–16 February 2019; pp. 289–294. [CrossRef]
21. O'Mahony, N.; Campbell, S.; Carvalho, A.; Harapanahalli, S.; Hernandez, G.V.; Krpalkova, L.; Riordan, D.; Walsh, J. Deep Learning vs. Traditional Computer Vision. In *Advances in Computer Vision, Proceedings of the CVC 2019. Advances in Intelligent Systems and Computing, Las Vegas, NV, USA, 25–26 April 2019*; Arai, K., Kapoor, S., Eds.; Springer: Cham, Germany, 2020; Volume 943. [CrossRef]
22. Wang, J.; Ma, Y.; Zhang, L.; Gao, R.X.; Wu, D. Deep learning for smart manufacturing: Methods and applications. *J. Manuf. Syst.* **2018**, *48*, 144–156. [CrossRef]
23. Heaton, J.B.; Polson, N.G.; Witte, J.H. Deep learning for finance: Deep portfolios. *Appl. Stochastic Models Bus. Ind.* **2017**, *33*, 3–12. [CrossRef]
24. Akinosho, T.D.; Oyedele, L.O.; Bilal, M.; Ajayi, A.O.; Delgado, M.D.; Akinade, O.O.; Ahmed, A.A. Deep learning in the construction industry: A review of present status and future innovations. *J. Build. Eng.* **2020**, *32*, 101827. [CrossRef]
25. Huang, M.Q.; Ninić, J.; Zhang, Q.B. BIM, machine learning and computer vision techniques in underground construction: Current status and future perspectives. *Tunn. Undergr. Space Technol.* **2021**, *108*, 103677. [CrossRef]
26. Wang, F.; Casalino, L.P.; Khullar, D. Deep Learning in Medicine—Promise, Progress, and Challenges. *JAMA Intern. Med.* **2019**, *179*, 293. [CrossRef] [PubMed]
27. Campedelli, G.M. *Machine Learning for Criminology and Crime Research: At the Crossroads. N.p.*; Taylor & Francis: Abingdon, UK, 2022.
28. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
29. Zemouri, R.; Noureddine, Z.; Daniel, R. Deep Learning in the Biomedical Applications: Recent and Future Status. *Appl. Sci.* **2019**, *9*, 1526. [CrossRef]
30. Esteva, A.; Chou, K.; Yeung, S.; Naik, N.; Madani, A.; Mottaghi, A.; Liu, Y.; Topol, E.; Dean, J.; Socher, R. Deep learning-enabled medical computer vision. *NPJ Digit. Med.* **2021**, *4*, 5. [CrossRef] [PubMed]
31. Chen, K.-Y.; Zheng, W.-Z.; Lin, Y.-Y.; Tang, S.-T.; Chou, L.-W.; Lai, Y.-H. Deep-learning-based human motion tracking for rehabilitation applications using 3D image features. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 803–807. [CrossRef]
32. Vafadar, S.; Skalli, W.; Bonnet-Lebrun, A.; Assi, A.; Gajny, L. Assessment of a novel deep learning-based marker-less motion capture system for gait study. *Gait Posture* **2022**, *94*, 138–143. [CrossRef]
33. Khan, M.A.; Kadry, S.; Parwekar, P.; Damaševičius, R.; Mehmood, A.; Khan, J.A.; Naqvi, S.R. Human gait analysis for osteoarthritis prediction: A framework of deep learning and kernel extreme learning machine. *Complex Intell. Syst.* **2021**, *9*, 2665–2683. [CrossRef]
34. Cui, H.; Chang, C. Deep Learning Based Advanced Spatio-Temporal Extraction Model in Medical Sports Rehabilitation for Motion Analysis and Data Processing. *IEEE Access* **2020**, *8*, 115848–115856. [CrossRef]
35. PyTorch. Available online: https://pytorch.org/get-started/previous-versions/ (accessed on 15 June 2022).
36. TensorFlow. 2019. Available online: https://www.tensorflow.org/versions (accessed on 15 June 2022).
37. Keras. Available online: https://pypi.org/project/keras/ (accessed on 15 June 2022).
38. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693. [CrossRef]
39. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1325–1339. [CrossRef] [PubMed]
40. Sigal, L.; Balan, A.O.; Black, M.J. HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *Int. J. Comput. Vis.* **2010**, *87*, 4–27. [CrossRef]
41. Zheng, J.; Liu, X.; Liu, W.; He, L.; Yan, C.; Mei, T. Gait Recognition in the Wild with Dense 3D Representations and A Benchmark. *arXiv* **2022**, arXiv:2204.02569. Available online: http://arxiv.org/abs/2204.02569 (accessed on 1 June 2023).
42. Institute of Automation, Chinese Academy of Sciences, CASIA Gait Database. Available online: http://www.cbsr.ia.ac.cn/english/Gait%20Databases.asp (accessed on 10 May 2024).

43. Desmarais, Y.; Mottet, D.; Slangen, P.; Montesinos, P. A review of 3D human pose estimation algorithms for markerless motion capture. *Comput. Vis. Image Underst.* **2021**, *212*, 103275. [CrossRef]
44. Liu, W.; Bao, Q.; Sun, Y.; Mei, T. Recent Advances of Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective. *ACM Comput. Surv.* **2022**, *55*, 80. [CrossRef]
45. Vicon, Full body modeling with Plug-in Gait. Nexus Documentation. 2022. Available online: https://help.vicon.com/space/Nexus212/11247555/Plug-in+Gait+Reference+Guide (accessed on 10 May 2024).
46. Mathworks 2023. interp1.MATLAB Function. Available online: https://www.mathworks.com/help/matlab/interpolation.html?category=interpolation&s_tid=CRUX_topnav (accessed on 10 May 2024).
47. Lain. Open Broadcaster Software (OBS). OBS Studio. Available online: https://obsproject.com/ (accessed on 10 May 2024).
48. Kwon, J. OBS-Record. GitHub Repository. 2023. Available online: https://github.com/jrkwon/obs-record (accessed on 15 June 2022).
49. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Machine Intell.* **2000**, *22*, 1330–1334. [CrossRef]
50. Calib.io. Calib-Camera Calibrator. Calb.io. 2023. Available online: https://calib.io/ (accessed on 10 May 2022).
51. Ayoola, O. Simplifying Object Segmentation with PixelLib Library. 2021. Available online: https://pixellib.readthedocs.io/en/latest/ (accessed on 10 June 2022).
52. MediaPipe. Pose: A Multi-Platform Library for Pose Estimation. GitHub. 2022. Available online: https://github.com/google/mediapipe/blob/master/docs/solutions/pose.md (accessed on 20 July 2022).