Anatomical-Marker-Driven 3D Markerless Human Motion Capture

Prayook Jatesiktat, Guan Ming Lim, Wee Sen Lim, and Wei Tech Ang

***Abstract*— Marker-based motion capture (mocap) is a conventional method used in biomechanics research to precisely analyze human movement. However, the time-consuming marker placement process and extensive post-processing limit its wider adoption. Therefore, markerless mocap systems that use deep learning to estimate 2D keypoint from images have emerged as a promising alternative, but annotation errors in training datasets used by deep learning models can affect estimation accuracy. To improve the precision of 2D keypoint annotation, we present a method that uses anatomical landmarks based on marker-based mocap. Specifically, we use multiple RGB cameras synchronized and calibrated with a marker-based mocap system to create a high-quality dataset (RRIS40) of images annotated with surface anatomical landmarks. A deep neural network is then trained to estimate these 2D anatomical landmarks and a ray-distance-based triangulation is used to calculate the 3D marker positions. We conducted extensive evaluations on our RRIS40 test set, which consists of 10 subjects performing various movements. Compared against a marker-based system, our method achieves a mean Euclidean error of 13.23 mm in 3D marker position, which is comparable to the precision of marker placement itself. By learning directly to predict anatomical keypoints from images, our method outperforms OpenCap's augmentation of 3D anatomical landmarks from triangulated wild keypoints. This highlights the potential of facilitating wider integration of markerless mocap into biomechanics research. The RRIS40 test set is made publicly available for research purposes at koonyook.github.io/rris40.**

***Index Terms*— Anatomical landmarks, biomechanics, data collection, deep learning, markerless motion capture.**

I. Introduction

HUMAN motion capture (mocap) technology is a valuable tool for biomechanics research as it provides objective data and insights into human movement. For instance, accurate and reliable mocap data is important for sports analysis to improve athletic performance [1], as well as clinical assessment to evaluate movement disorders [2]–[4].

This work was supported by NTUitive Gap Fund (NGF-2022-11-003) and A*STAR under its National Robotics Programme (NRP) BAU grant, project titled - Assistive Robotics Programme (Award No: M22NBK0074).

Prayook Jatesiktat, Guan Ming Lim, Wee Sen Lim, and Wei Tech Ang are with the Rehabilitation Research Institute of Singapore, Nanyang Technological University, 11 Mandalay Road, Clinical Sciences Building, 308232, Singapore (e-mails: prayook001@e.ntu.edu.sg; guanming001@e.ntu.edu.sg; weesen.lim@ntu.edu.sg).

Wei Tech Ang is with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, N3-01A-01, 50 Nanyang Avenue, 639798, Singapore (e-mail: wtang@ntu.edu.sg).

Marker-based optical mocap systems are generally considered the state-of-the-art equipment [1], [3] or the "gold standard" [5]–[8] for assessing human movement. These systems require retro-reflective markers strategically placed on the subject's body to be tracked by infrared cameras equipped with active infrared light sources. The precise triangulation and tracking of these markers' 3D positions are facilitated by multiple synchronized and calibrated cameras.

However, marker-based mocap systems have several limitations due to the need for markers. First, the marker placement process is time-consuming and requires trained personnel to palpate specific anatomical landmarks on the subject's body. Second, the markers can affect the natural movement of the subject during data collection [7]. Third, extensive post-processing is required such as marker labeling, gap filling due to marker occlusion, and correcting mislabeled markers due to marker swap [9].

A promising alternative to marker-based mocap is markerless mocap, driven by recent advancements in deep learning models capable of directly identifying *virtual* markers from RGB image [5], [10]. This eliminates the need for *real* markers and overcomes various limitations associated with marker-based mocap by significantly reducing the time taken for subject preparation, data collection, and post-processing [1]–[3]. Additionally, deep learning models could learn a geometry-aware representation of the body and approximately infer the location of virtual markers in occluded regions based on other visible features [11]. This helps to reduce occlusion issues and minimize gaps in marker trajectories.

While markerless mocap has rapidly advanced with a model-centric focus on developing new architectures and training methods, a data-centric approach is also important as the performance of deep learning models depends on the quality and quantity of the training dataset [11], [12]. Some popular datasets for 2D human keypoint estimation include the MPII Human Pose [13] with 25K images and the COCO [14] or COCO-WholeBody [15] with 200K images. Although these datasets contain diverse human-in-the-wild images, the keypoints are manually annotated through crowdsourcing. This results in inconsistent annotation as different people may have different interpretations of a joint center [3] and potential mislabeling can affect estimation accuracy, especially for the hip and knee joint centers [16]. Furthermore, these datasets are annotated with around 16 to 23 body keypoints, providing only two joint centers for each lower and upper limb segment. Thus, it is difficult to evaluate rotation about a bone's longitudinal axis, as a minimum of three non-aligned markers are required

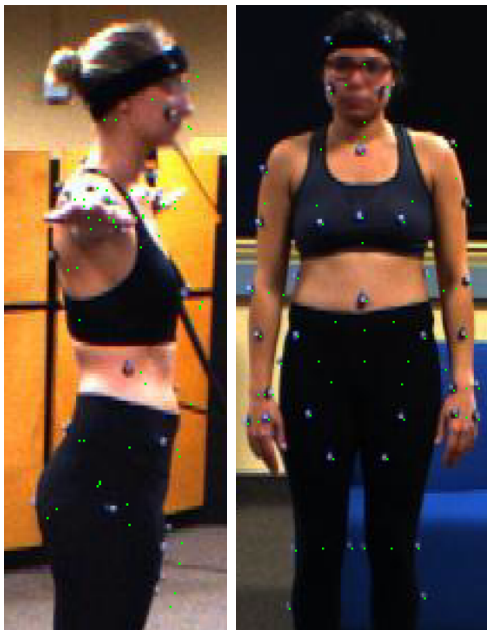

Fig. 1. Cropped images from the MoVi dataset [12] with 3D markers projected onto the 2D image plane. The inconsistency in camera calibration and synchronization caused misalignment between projected (green pixels) and actual markers. Best viewed on the digital version.

to fully reconstruct a rigid segment's six-degree-of-freedom transformation [3], [11].

To address the issues of inconsistent annotation and insufficient 2D keypoints, markerless mocap systems such as Theia3D use highly trained annotators with anatomical knowledge to manually annotate 51 salient features, including joint locations and other identifiable surface features, on images of over 500K humans in the wild [17]. Although quality control is done by additional expert annotators, it can be challenging for humans to accurately annotate occluded points. On the other hand, the MoVi dataset could automate the annotation process by collecting synchronized and calibrated marker-based mocap data with stationary video cameras to allow the overlay of the 3D skeletal pose in camera coordinates [12]. However, for some of the images, the inconsistent quality of the given camera calibration parameters and time synchronization results in significant errors during projection as shown in Fig. 1.

In this paper, we aim to integrate anatomical priors and data-driven methods to capture human movement from multi-view RGB video sequences. First, a high-quality dataset annotated with 40 surface anatomical bone landmarks is created to train a 2D keypoint detection model. This dataset, known as RRIS40, uses precise marker positions from marker-based mocap to generate pixel-accurate 2D anatomical landmarks. This allows us to scale the data collection to millions of images. Both the marker-based and markerless systems are carefully calibrated to ensure optimal spatial and temporal alignment.

Next, markers are removed from images to prevent the keypoint detection model from learning anatomical landmarks from the markers' distinct features [18]. A weighted triangulation method that considers confidence scores is proposed to improve the accuracy of 3D markers estimation from multiple 2D anatomical landmarks. Our system outputs the 3D positions of 40 virtual markers, which can be used to derive joint centers. This makes it compatible with existing biomechanical analysis workflows and provides orientation information for body segments.

Since the RRIS40 dataset involves human-in-mocap-lab images of subjects with similar clothing and barefoot conditions, the robustness and performance of the model are affected when subjects wear different clothing and footwear. To overcome this issue, the keypoint detection model is trained using a mixture of the RRIS40 dataset and diverse human-in-the-wild images [15]. This helps to improve the generalization ability of the model across different clothing and footwear scenarios, making it more useful in real-world applications.

To evaluate the effectiveness of our proposed method, extensive experiments are conducted using our RRIS40 test set and a publicly available dataset [19]. The performance of our method is validated against marker-based mocap and also compared with other markerless mocap methods [8], [20], [21]. The results demonstrate good accuracy and robustness in capturing human motion. Our RRIS40 test set is released publicly for research purposes at koonyook.github.io/rris40.[1]

II. Related Work

A. Markerless Motion Capture Methods

Markerless human motion capture (mocap) from multi-view videos generally involves detecting key features from the images and fitting an articulated human model to the features in 3D space [10]. In this section, we focus on the feature detection step, while more detailed discussions on both steps can be found in a review [22].

Early works on extracting image features relied on silhouette and edge-based techniques [19] to create a 3D visual hull for tracking a single subject [23], [24], but these techniques require good contrast between the captured subject and the background. Furthermore, it is difficult to track segments that are nearly rotationally symmetric around their local axes, as the silhouette of these segments barely changes when they rotate around these axes [25].

In recent years, significant progress has been made in computer vision and deep learning. Convolutional neural networks (CNNs) have played a key role in frameworks like OpenPose [26] and DensePose [27], enabling the detection of body joints and estimation of dense correspondence maps. Many methods rely on open-source frameworks like OpenPose, which can only detect around 20 body keypoints [8], [28], but these joint centers are insufficient to determine the orientation of body segments [3]. Hence, OpenCap uses two Long Short-Term Memory (LSTM) networks trained on a large mocap dataset to augment the 3D positions of the sparse 20 keypoints with a more detailed 43 surface anatomical markers [8]. Because those augmented anatomical positions are not extracted directly from the images, this process could introduce a bias towards average movement patterns and may not generalize to individuals with unique movement patterns such as rehabilitation patients [10].

[1]Use access code bX75G7sr to download.

Other methods that consider anatomical landmarks are PitchAI™ and Theia3D. PitchAI™ is a single-camera markerless mocap system for baseball pitching that estimates 53 3D markers (19 joint centers, 34 bony landmarks) from 19 2D joint centers [29]. Theia3D customized a dataset with manually annotated anatomical landmarks and an articulated multibody model is scaled to fit subject-specific landmark positions in 3D space [17]. However, Theia3D uses a proprietary list of anatomical landmarks, and their commercial software outputs joint centers, which makes it difficult to evaluate the accuracy of anatomical landmark detection. Instead of manual 2D annotation, our RRIS40 dataset is annotated with 40 anatomical landmarks from a precise marker-based mocap system.

B. Markerless Motion Capture Datasets

Datasets are important for training deep learning models and evaluating the performance of various methods. Most datasets for markerless human mocap contain diverse images of humans in the wild engaged in different activities and scenes [13]–[15]. While these datasets help in the training of robust models that can be applied to real-world scenarios, they may not be suitable for biomechanics applications. This is because manually annotated keypoints may be inaccurate and the sparse body keypoints are insufficient to determine the orientation of body segments [3].

To improve the precision of keypoint annotation, anatomical landmarks that are well-established in marker-based mocap can be used to train a deep learning model to output the marker positions. To integrate anatomical priors into markerless mocap, it requires synchronized and calibrated multi-view RGB videos that are paired with corresponding marker-based mocap data. Hence, Table I shows an overview of such datasets that could be used for training and benchmarking purposes.

The HumanEva dataset [30] is one of the earliest datasets that was made publicly available to develop and evaluate human pose estimation algorithms. The Human3.6M dataset [31] increased the size of the dataset significantly and included additional data from a depth camera and 3D body scans of all subjects. Both datasets were created to capture natural-looking image data that could be used to train realistic human sensing systems. Although 3D marker positions are not available for both datasets, the 3D joint locations from a 3D human model are provided. However, joint locations obtained from 3D human models differ significantly from those obtained from 3D anatomical landmarks as shown in Fig. 2 [32].

The GPJATK dataset [19] is mainly created for gait evaluation and identification, thus it only includes walking motions. Similarly, the ENSAM dataset [6], [33] is adapted for clinical gait analysis and includes the motion of pathological cases. To reduce errors related to marker misplacement for lower limbs, biplanar X-ray images were also acquired. The study found that training pose estimation method on the Human3.6M dataset and then fine-tuning it on the ENSAM training set significantly reduces joint position error compared to not fine-tuning on the ENSAM training set.

The MoVi dataset [12] is a large multimodal dataset with different combinations of optical mocap data, video data, and inertial measurement units (IMU). It includes 90 subjects who performed 21 everyday actions and sports movements. Besides generating a 3D human mesh model, it also releases 3D marker positions, but for some of the images, there are misalignments between the projected and actual markers as shown in Fig. 1.

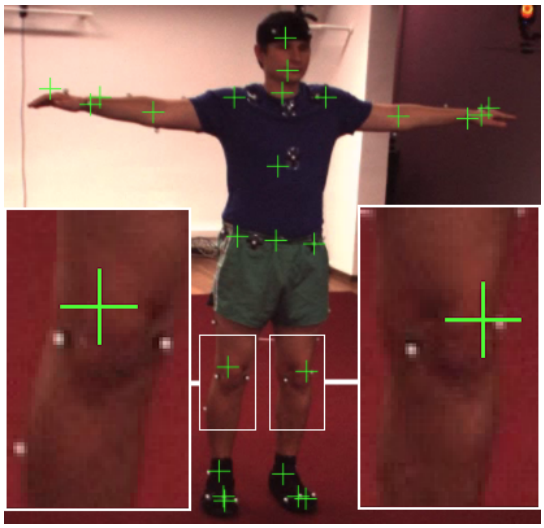

Fig. 2. A cropped image from the Human3.6M dataset [31] with 3D model-based keypoints projected onto the 2D image. The model fitting done in this dataset can introduce errors. In this case, the knee joint centers are shifted significantly from the midpoint of the two knee markers. Similarly, the hip joint centers are around the same level as LASIS and RASIS markers on the pelvis (Anterior Superior Iliac Spine) instead of being lower.

In this paper, we aim to create a high-quality dataset by precisely calibrating both the marker-based and markerless systems. Similar to the MoVi dataset, our subjects wore minimal clothing to minimize marker movement relative to the body [12]. Although this limits the variety of subjects' appearance, it is necessary for ensuring accurate marker data, as errors can arise from marker placement on regular clothing [30]. Our dataset includes over 200 subjects performing various activities. These subjects are part of an Asian-centric movement database [34].

C. Anatomical Landmarks

Marker-based mocap systems rely on anatomical landmarks such as segment endpoints and bony landmarks to estimate the position and orientation of the segment. A basic marker set uses 16 markers positioned on the skin near joint centers to define limb segment position and orientation with a straight line between markers, but it cannot calculate axial rotation of body segment [35]. For a more comprehensive marker set, the Vicon's Plug-in Gait uses around 38 markers for full-body modeling. The RRIS's Asian-centric movement database uses 84 markers based on a modified Calibrated Anatomical System Technique (CAST) [34]. In this work, we use a total of 40 surface anatomical landmarks as detailed in Section III-B.1.

Marker sets based on anatomical landmarks are advantageous due to their ease of placement and interpretation. However, their accuracy and reliability can be affected by marker placement variability and soft tissue artifacts. Between

TABLE I
COMPARISON OF DATASETS WITH SYNCHRONIZED MULTI-VIEW RGB VIDEOS AND MARKER-BASED MOTION CAPTURE (MOCAP) SYSTEM.

Dataset	HumanEva [30]	Human3.6M [31]	GPJATK [19]	MoVi [12]	ENSAM [6], [33]	RRIS40 (this work)
Dataset information						
No. of subjects	4	11	32	90	25(train) / 16(test)	209(train) / 10(test)
No. of frames or duration	$\sim$ 40,000	$\sim$ 3.6 million	18,764	$\sim$ 4 hours	$\sim$ 75,000 (train) $\sim$ 48,000 (test)	$\sim$ 2.5 million (train) $\sim$ 0.1 million (test)
Types of motion	Walking, jogging, throwing and catching a ball, gesturing, boxing	Typical human activities: walking, seated, etc.	Walking	20 predefined everyday actions, sports + one self-chosen movement	Walking	Typical human activities: walking, seated, self-chosen random movements, etc.
Appearance	Natural clothing with shoes	Moderately realistic clothing with shoes	Natural clothing with shoes	Tight-fitting minimal clothing with socks	Underwear, barefoot	Singlet or short sleeves shirt and shorts, barefoot
Accessible 3D marker position	No	No	Yes	Yes	No	Yes (test)
RGB video data						
No. of cameras	7 / 4	4	4	2	4	8
Camera model	UniQ UC685CL, Pulnix TM6710 / Basler A602fc	Basler piA1000	Basler piA1900-32gc	FLIR Grasshopper2	GoPro HERO7 Black	e-con Systems™ See3CAM_24CUG
Shutter type	Global	Global	Global	Global	Rolling	Global
Synchronization method	Software / hardware	Hardware	Frame-level hardware synchronization	Frame-level hardware synchronization	Blinking LED	Frame-level hardware synchronization
Extrinsic calibration	Single moving marker	30 reflective markers on a surface	Markers on the recorded subject (manual 2D annotation)	Single moving marker	Checkerboard with 6 reflective markers	Mocap-assisted calibration (refer to Section III-A)
Image resolution	659 $\times$ 494, 644 $\times$ 448 / 656 $\times$ 490	1000 $\times$ 1000	960 $\times$ 540	800 $\times$ 600	1920 $\times$ 1080	1920 $\times$ 1200
Frame rate (fps)	60	50	25	30	100	50
Marker-based motion capture						
No. of cameras	6 / 12	10	10	15	12	16
Camera model	ViconPeak	Vicon T40	Vicon MX-T40	Qualisys Oqus 300 & 310	Vicon	Qualisys Arqus A12 & Miqus M3
Image resolution	1 MP / 1.3 MP	4 MP	4 MP	1.3 MP	Unknown	12 MP(A12) / 2 MP(M3)
Frame rate (fps)	120	200	100	120	100	200
No. of markers	Unknown	Unknown	39	67	51	40

different skilled examiners, the marker placement root mean square (RMS) errors can be up to 24.8 mm [36]. Soft tissue artifacts occur because the skin and attached markers move relative to the bone during motion due to the deformation of tissues such as muscle or fat beneath the skin [3], [11]. This leads to unavoidable measurement errors of a few millimeters.

Although intracortical bone pins or biplanar videoradiography may provide more accurate motion data, the former is an invasive technique, while the latter involves exposure to radiation, making it impractical to collect large-scale data [3], [11]. Therefore, the use of anatomical landmarks from marker-based mocap is our preferred option as the errors from marker placement variability and soft tissue artifacts are relatively minor and more manageable when compared to the potential human errors involved in visually annotating joint centers or anatomical landmarks from 2D images.

III. METHOD

This section describes the components of our two main pipelines illustrated in Fig. 3. The first pipeline is for data collection, pre-processing, and model training. The second pipeline applies the trained model on multi-view videos to detect 2D keypoints and triangulate them to obtain 3D virtual marker trajectories.

A. Sensing Hardware and Camera Configuration

A marker-based optical mocap system and multiple RGB video cameras are used to collect the RRIS40 dataset. The mocap system consists of 16 Arqus A12 and Miqus M3 (Qualisys AB, Gothenburg, Sweden). The RGB video cameras' model is See3CAM_24CUG (e-con Systems India Pvt Ltd, Chennai, India) with a resolution of 1920 $\times$ 1200 pixels. Each camera is equipped with a varifocal lens, which allows for adjustable

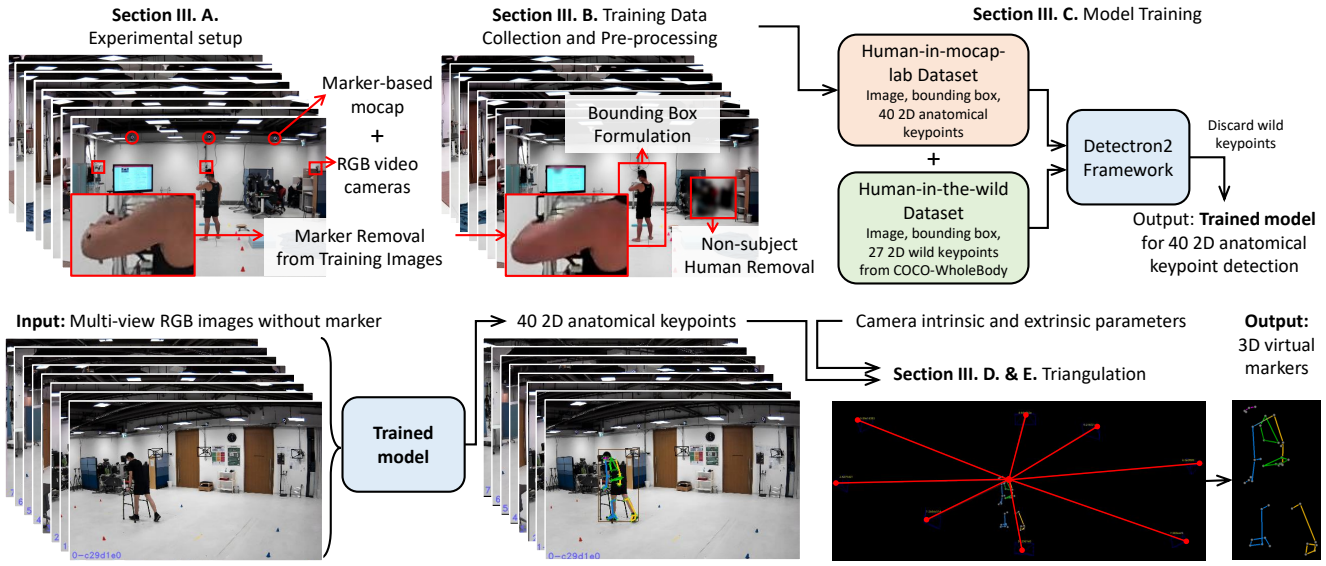

Fig. 3. **Top**: Overview of data collection and training pipeline. **Bottom**: Overview of 3D virtual marker tracking pipeline.

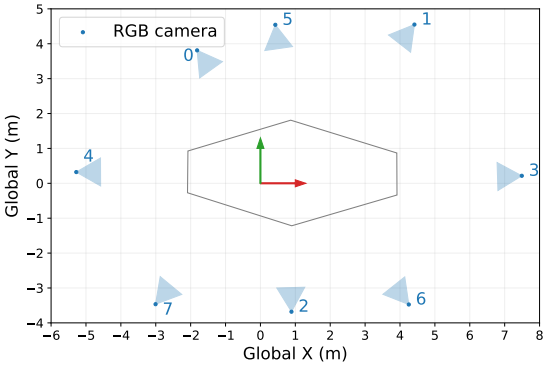

Fig. 4. A top-view map of all the camera placements and the capture volume used in the RRIS40 test set.

focal length, to change the size of field of view for optimal coverage of the capture area.

The mocap system captures data at a rate of 200 Hz. To ensure frame-level synchronization between the mocap system and the video cameras, an electronic circuit is used to receive synchronization pulses from the mocap system, generate synchronized pulses at a rate of 50 Hz, and transmit shutter pulses via copper wires to all the video cameras.

All the video cameras have been positioned at a height of 170 cm above the ground, facing towards a central capture area as shown in Fig. 4. This uniform height ensures that the training images are captured from a consistent range of perspectives. Additionally, most tripods can reach a height of 170 cm without requiring any additional framework for mounting the cameras, making deployment easier.

To ensure precise calibration, each video camera is fitted with three 1-watt white LEDs around the lens. These LEDs enable a regular video camera, which can only detect light in the visible spectrum, to see a round retro-reflective marker as a bright spot in the captured image. When the mocap system detects this marker in 3D space and the video camera simultaneously detects it in 2D on the image, a 2D-3D correspondence pair is formed. A sufficient number of these correspondence pairs collected throughout the capture volume can be used to accurately calculate camera pose (extrinsic parameters) and fine-tune intrinsic parameters.

An important camera setting to consider is the exposure time, which must be adjusted to minimize motion blur in the captured image. When recording video of human subjects, an exposure time of 3.888 ms is chosen to ensure that even during very fast movements, the edges of the human silhouette remain sharp. However, during calibration where the target object is a retro-reflective marker that can move faster than a human body, the exposure time is reduced to 0.996 ms. This results in a very dark capture environment, but the reflection from the retro-reflective marker is still bright enough to be detected.

B. Training Data Collection and Pre-processing

The training data contains images from the video cameras, annotated with 2D anatomical landmarks, and bounding box of the target subject. This section details how the data is collected and pre-processed before training a deep learning model.

This research involves human subjects. The data collection was approved by the Nanyang Technological University Institutional Review Board (IRB-2018-04-014). All the participants have signed the informed consent forms approved by the IRB. All methods were performed in accordance with the relevant guidelines and regulations.

1) RRIS40 Marker Set: A total of 40 markers are used in this work, which is a subset of the marker set for RRIS's Asian-centric movement database [34]. Marker clusters are not included as their placement is inconsistent across subjects and their large size can cause difficulty in the marker removal step later. The 40 markers consist of 4 on the head, 4 on the thorax, 4 on the pelvis, 7 on each upper limb, and 7 on each lower limb. The placement of these markers is carried out by trained personnel and standardized according to anatomical landmarks as shown in Fig. 5. During the quality control step,

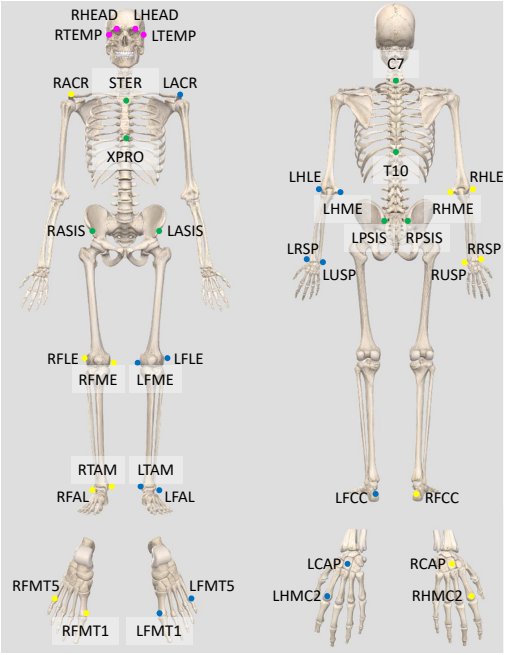

Fig. 5. RRIS40 marker placement.

Fig. 6. All the green crosses display projected 2D marker positions. Pixel-level precision during fast movements like jumping requires careful spatial and temporal alignment between the mocap system and the video cameras.

any marker that is found to be shifted or placed incorrectly will be removed and marked as unlabeled.

After post-processing of the mocap data, instead of directly using the nearest marker-based 3D sample, the 3D marker trajectories with a sampling rate of 200 Hz are linearly interpolated from the two nearest samples at the middle timing of the video camera exposure and projected to each video camera to obtain 2D positions with pixel-level accuracy. An example of this projection is shown in Fig. 6.

2) Marker Removal from Training Images: Training images captured from the video cameras always contain visible marker blobs. Thus, a deep learning model may learn to rely on the gray blobs of the visible markers and use them as a key feature to locate the marker itself. This overfitting can lead to a degradation in performance when there is no marker on the subject during actual markerless mocap.

Therefore, it is important to prepare the training data as if there is no marker on the subject. First, the pixels occupied by the marker must be identified. This is done automatically by taking the 2D projection and drawing a 2D radius around it. The radius size depends on the distance between the camera and the marker with an additional margin to cover the base and the shadow of the marker. Next, DeepFillv2 image inpainting technique [37], [38] that leverages a Generative Adversarial Network (GAN) is used to replace the pixel color in the target area by being aware of the surrounding context. An example of the result is shown in Fig. 7.

Recent work on hand tracking has found that raw image data containing hand markers can affect the training process as the markers provide extra features [18]. Thus, they proposed a marker removal network (MR-Net) comprising two stages: marker synthetization and marker removal. While the MR-Net works well for hand context, where hands are usually bare, training on humans with varying clothing may be challenging. Additionally, the work suggests that applying a CycleGAN-based method to the whole image may result in unnatural artifacts. Therefore, in this work, image inpainting is applied only to the pixels surrounding the marker region.

3) Non-subject Human Removal: When using multiple video cameras to capture a subject, it is common to have other non-subject humans present in the field of view. These individuals do not have markers and cannot be labeled, which may confuse the deep learning model. Therefore, non-subject humans are automatically detected using default human detection from Detectron2 [21] and blurred within a bounding box with smooth edges. It is important to note that all the pixels within the bounding box of the target subject are not blurred.

4) Bounding Box Formulation: Apart from 2D anatomical landmarks, 2D bounding box around each target subject is also required. This rectangular bounding box must cover not only the projected marker positions but also the full silhouette of all body parts. For example, even though there is no marker on the finger, the elbow, wrist, and hand markers are used to approximate the possible volume that the finger can reach. The 3D points on the surface of this volume are then projected onto each camera to approximate the bounding box.

C. Neural Network Architecture and Training

Our model uses the keypoint detection variant of Mask R-CNN [39] with a Feature Pyramid Network (FPN) as the feature extraction backbone. The network is based on Detectron2 implementation [21] as it is except for the number of output keypoints. The architecture is designed to produce a bounding box around each human subject and generates a 2D heatmap for each keypoint within its respective bounding box.

Fig. 7. An example of marker removal before (left) and after (right) using GAN-based context-aware image inpainting.

To allow robust detection of anatomical keypoints across various outfits and environments, the training data combines annotated images from two sources. The first source contains around 5.39 million human-in-mocap-lab unique images that are sampled from our RRIS40 training set. The annotation in this group includes 40 anatomical keypoints as described in Section III-B.1. The second source contains 56,599 human-in-the-wild images with annotations from the COCO-WholeBody dataset [15]. For this group, a subset of 27 wild keypoints is used for training. The keypoints are 12 joint centers from the hips, knees, ankles, shoulders, elbows, and wrists, 5 head keypoints from the nose, eyes, and ears, 6 foot keypoints from the big toes, little toes, and heel centers, and 4 hand keypoints from the index and little finger metacarpophalangeal.

Since none of the 40 anatomical keypoints overlap with the 27 wild keypoints, the keypoint head of the Mask R-CNN is adjusted to predict all 67 keypoints simultaneously. For the RRIS40 training set, 12 keypoints containing the joint centers from the hips, knees, ankles, shoulders, elbows, and wrists are augmented in the annotation if possible. For the keypoints that do not exist on each training data and cannot be augmented accurately, they are counted as unlabeled.

For model training, a batch size of 8 images is used and each epoch consumes a random mix of 1 round of the human-in-mocap-lab dataset and 24 rounds of the human-in-the-wild dataset, such that the sampling ratio between the human-in-mocap-lab dataset and the human-in-the-wild dataset is approximately 80:20. The learning rate starts from 0.02 and decreases to 0.002, 0.0002, and 0.00002 at iterations 425,100, 850,100, and 3,050,100 respectively. Before iteration 1,250,000, only the keypoint head is unfrozen, while all the weights and biases are unfrozen thereafter. After the training is completed at iteration 3,100,000 (which consumes about 4 epochs of sampled human-in-mocap-lab data and about 96 epochs of human-in-the-wild data), only the 40 anatomical keypoints are used in the subsequent triangulation step, while the 27 wild keypoints are discarded.

D. Strategic Triangulation

After obtaining the predicted 2D anatomical keypoints from multiple images of a subject without markers, the next step is to convert them into 3D virtual markers through triangulation. To do this, a 2D location on an image from a single camera can be represented as a 3D ray that points out from the camera's origin. The triangulation formula then calculates a virtual intersection point of all these rays, resulting in a 3D point. In an ideal scenario, the distance between the 3D point and each ray would be relatively small, less than 10 cm, making it easy to accept the solution.

However, in reality, some cameras may fail to capture certain points due to occlusion or confusion between the left and right sides. To ensure that the triangulation method remains robust even if individual camera predictions are inaccurate, our method rejects the data from cameras that deviate significantly from the consensus or fail to maintain continuity of marker trajectory.

The method proposed for triangulating one marker trajectory at a time is as follows.

1) For each frame, perform weighted ray-distance-based triangulation (refer to Section III-E) using all the possible combinations of cameras.
2) Among the triangulation combinations in a frame, for a combination with n rays, if at least one combination with greater than n rays with a smaller *maxRayDistance* exists, the combination with n rays will be eliminated. This will help to remove noisy data. The term *maxRayDistance* means the perpendicular distance from the 3D triangulated point to the farthest ray used in that triangulation combination.
3) From the remaining combinations, choose one combination in each frame to find the trajectory with the lowest cost. The cost is calculated using the following formula:

$$\sum_{f}(m_f d_f)^2 \tag{1}$$

where m_f is the distance multiplier with a default value of 1. If at least one of the selected triangulated points has only two contributed rays, m_f becomes 1.5 to penalize the selection of combinations with too few rays. d_f is the distance between the selected triangulated points from frame f and $f+1$.

By using this cost function design, unstable trajectories and triangulation combinations can be automatically rejected. Searching for the optimum trajectory can be done in polynomial time with dynamic programming.

E. Weighted Ray-distance-based Triangulation

It is common for a neural network that locates keypoints in a 2D space to provide a confidence score for each point.

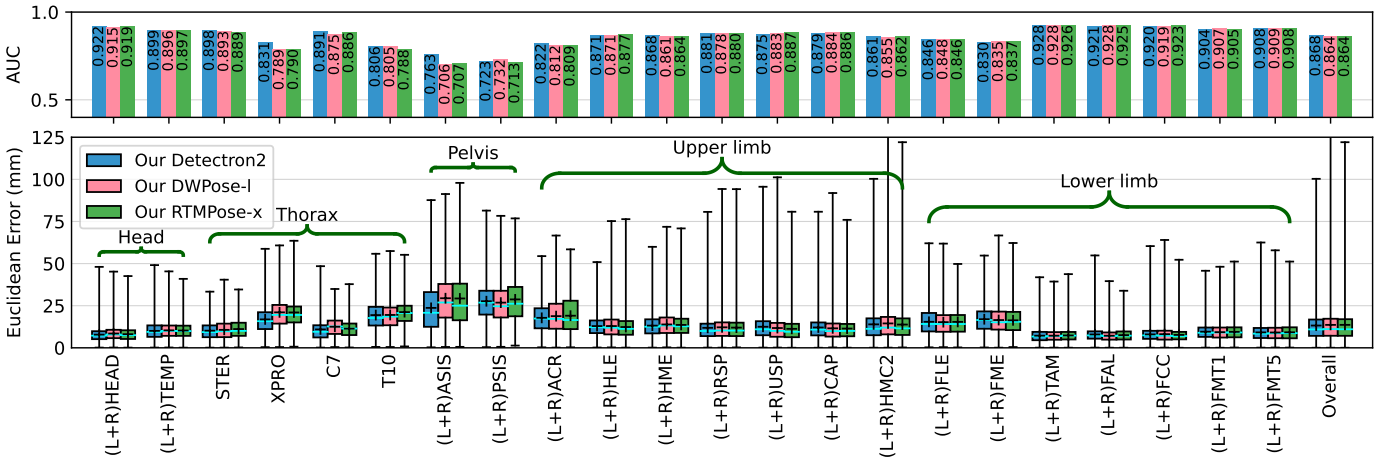

Fig. 8. Benchmarking results of our method with three different models (our DWPose-l and RTMPose-x are trained for ablation study in Section. IV-I) on the RRIS40 test set using ground truth from marker-based mocap. **Bottom**: Box plots of the Euclidean errors of virtual markers (lower is better). The whiskers of the box plots extend to the minimum and maximum errors without any other samples beyond the maximum boundary. **Top**: The ratio of area under the curve (AUC) for each virtual marker (higher is better).

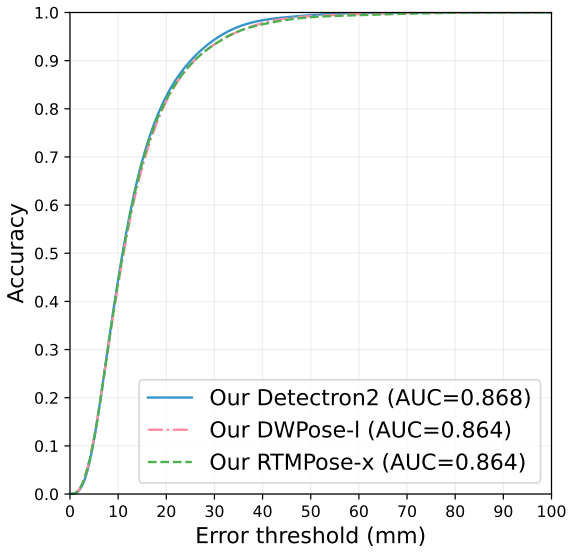

Fig. 9. Overall accuracy profile of our method with three different models from all the 40 virtual markers. Each curve represents a total of 4,081,976 Euclidean error samples from 10 test subjects from the RRIS40 test set. At every error threshold up to 100 mm, a ratio of error samples within that threshold is plotted.

For example, the keypoint detection version of Mask R-CNN [39] generates a heatmap of confidence within the bounding box for every keypoint. The 2D location with the highest confidence in the heatmap is selected as the answer. In this case, the confidence score at the peak corresponds to the score for that 2D keypoint prediction. Usually, this score is ignored in a normal triangulation. However, our weighted triangulation formula allows the utilization of the score as the triangulation weight to enhance the accuracy of the triangulation.

The triangulated 3D position (P) can be derived as

$$P = (\sum_i w_i Q_i)^{-1} (\sum_i w_i Q_i C_i), \quad \text{and } Q_i = I_3 - U_i U_i^\top \tag{2}$$

given that

- w_i is the weight or the confidence score of the i^{th} ray from the i^{th} camera.
- C_i is the 3D camera location associated with the i^{th} ray.
- U_i is the 3D unit vector that represents the back-projected direction associated with the i^{th} ray.
- I_3 is the 3-by-3 identity matrix.

The directional vector (U_i) of each back-projected ray is calculated by undistorting the 2D observation using *cv2.undistortPointsIter* to the normalized coordinate. Next, forms a 3D directional vector in the camera reference frame $[x\_undistorted, y\_undistorted, 1]^\top$ and rotates the direction to the global reference frame using the camera orientation. Lastly, normalizes the vector to get the unit directional vector (U_i).

As this formula is derived by minimizing the weighted sum of the square of the distance between the triangulated point and all the rays, the prediction with a lower confidence will have a lesser influence in the triangulation. This allows the triangulated point to be closer to the ray with a higher predictive confidence resulting in better overall accuracy.

IV. Evaluation

A. Evaluation on RRIS40 Test Set

A straightforward way to evaluate the accuracy of our system's virtual marker trajectories is to compare them against the actual marker trajectories from the marker-based mocap system. The RRIS40 test set of around 0.1 million frames from 10 subjects is processed for this comparison. The subjects (5 males, 5 females), were on average 29 years old (range: 21–63 years old), mean height was 1.67 m (range: 1.48–1.83 m), mean body mass was 63.9 kg (range: 47.0–91.2 kg), and mean BMI was 22.5 kg/m^2 (range: 18.4–29.3 kg/m^2). Our method outputs 40 virtual anatomical markers through strategic triangulation without using additional outlier removal, gap filling, low-pass filter, or inverse-kinematic model fitting.

A total of 4,081,976 Euclidean errors have been calculated and their cumulative distribution plot is shown in Fig. 9. This

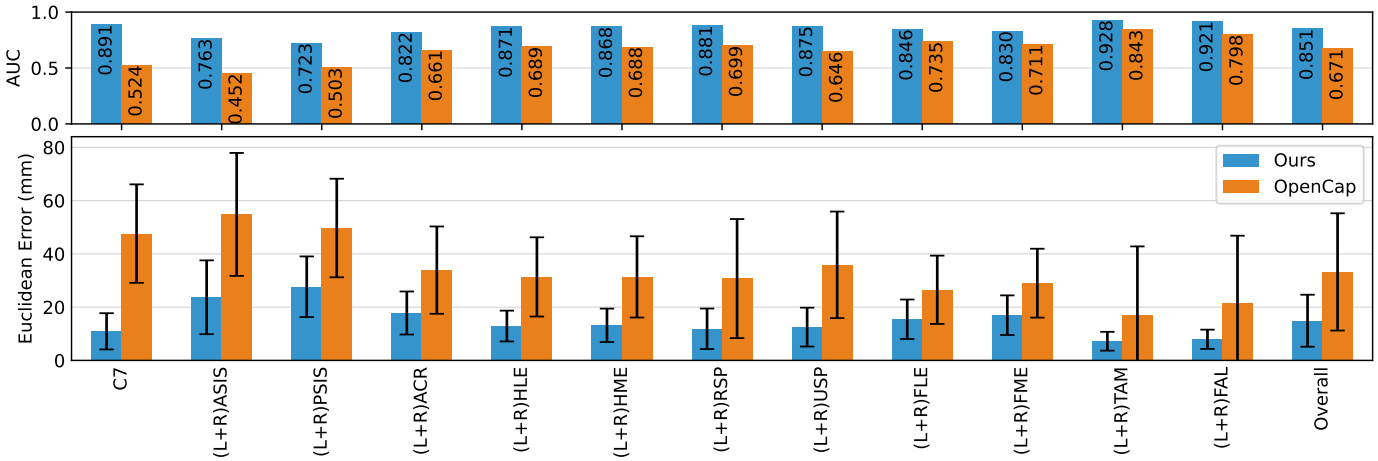

Fig. 10. Benchmarking results on the RRIS40 test set. **Bottom**: Error comparisons of overlapping virtual anatomical markers between our method and OpenCap (lower is better). Our method consistently produces significantly lower errors than OpenCap. **Top**: The ratio of area under the curve (AUC) for each virtual marker (higher is better).

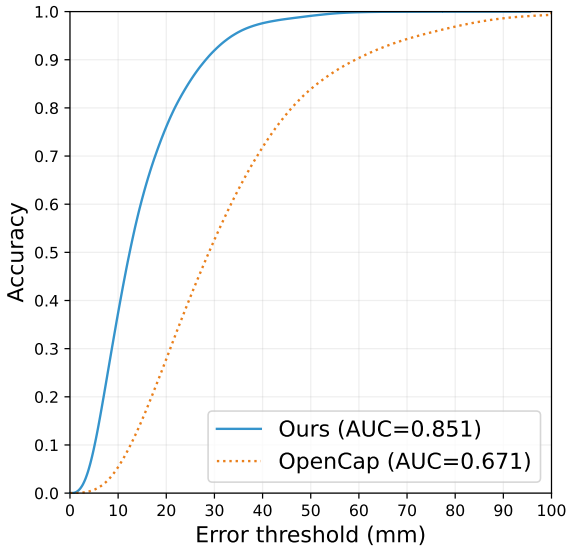

Fig. 11. Overall accuracy profile of our method and OpenCap from all the 23 overlapping virtual anatomical markers with a total of 2,337,253 Euclidean error samples from 10 test subjects from the RRIS40 test set.

plot is used to determine the quality of measurements in all evaluations, based on the ratio of the area under the curve (AUC) up to a cutoff of 100 mm. An AUC of 1.0 indicates perfect measurement without any error. The AUC and the error distribution for specific markers are shown in Fig. 8. Overall, the mean Euclidean error is 13.23 mm and the median Euclidean error is 10.92 mm.

Among all the anatomical keypoints, the pelvis markers which are the ASIS and PSIS, are the least accurate. This trend is consistent with a previous study which also showed that the error in marker placement on the pelvic bone landmarks is larger than any other body parts [36]. This suggests that the high levels of error are partially caused by the displacement of the markers in the train or test data.

B. Evaluation against an Anatomical-marker-based Motion Capture System on RRIS40 Test Set

Among the markerless multi-view mocap systems, OpenCap [8] is the only one that provides 3D anatomical marker positions in its outputs. However, unlike our method which trains a network to infer 2D anatomical landmarks from the image for direct triangulation, OpenCap uses OpenPose to infer 2D non-anatomical keypoints such as joint centers for triangulation and then uses another network to augment 3D anatomical landmarks from the triangulated 3D non-anatomical keypoints.

Since the anatomical keypoints of OpenCap do not completely overlap with those of RRIS40, only 23 keypoints are compared against the marker trajectories from the marker-based mocap system in the RRIS40 test set. To remove the variation of calibration methods, the same set of intrinsic and extrinsic camera parameters are used in this comparison.

Based on the results in Fig. 10 and Fig. 11, our method of direct 2D prediction and triangulation of anatomical landmarks is significantly better than OpenCap's 3D augmentation method in all comparisons. Our method displays an overall error of 14.90 ± 9.76 mm, which is lower than OpenCap's overall error of 33.22 ± 22.01 mm. The higher tracking stability with respect to OpenCap can be visually noticed in a supplementary video.

C. Evaluation against Joint-center-based Pose Estimation Tools

To compare against a wider range of methods, it is necessary to evaluate joint center locations. Although our system does not directly output 3D joint centers, these joint centers can be calculated from virtual anatomical markers based on existing anatomical studies.

The hip joint centers are calculated using the Bell and Brand hip joint in CODA pelvis coordinate system from C-Motion's guideline [40], [41]. The shoulder joint centers are calculated using the Rab Upper Extremity Model from C-Motion's guideline [42], [43]. For the knee joint, the midpoint between FLE and FME markers is used. For the ankle joint,

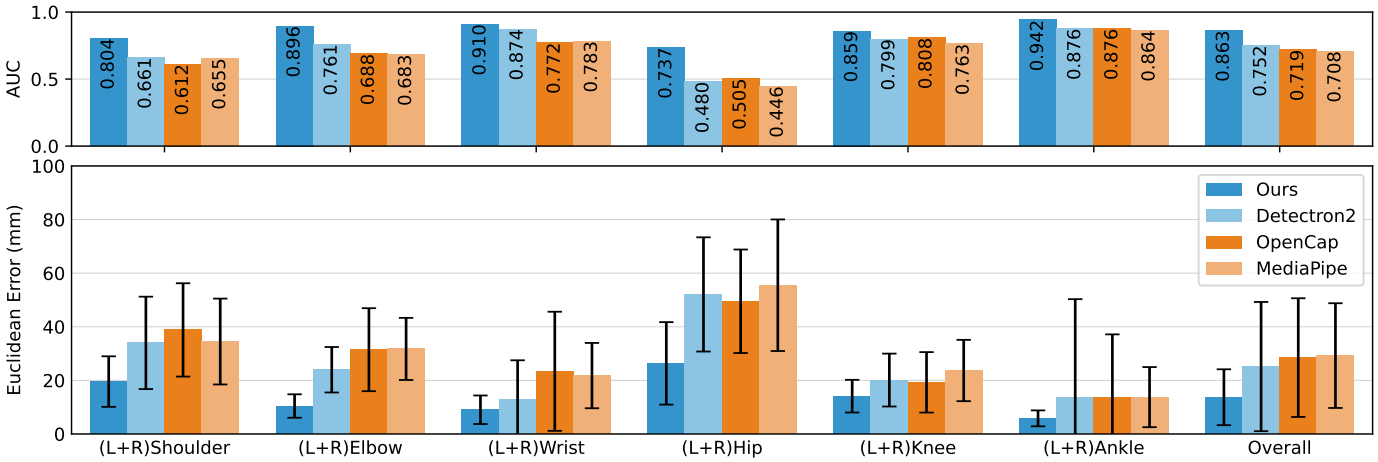

Fig. 12. Benchmarking results on the RRIS40 test set. **Bottom**: Error comparisons of different methods based on joint centers (lower is better). Our method has significantly lower errors than other methods in every key joint center. **Top**: The ratio of area under the curve (AUC) from each virtual marker (higher is better).

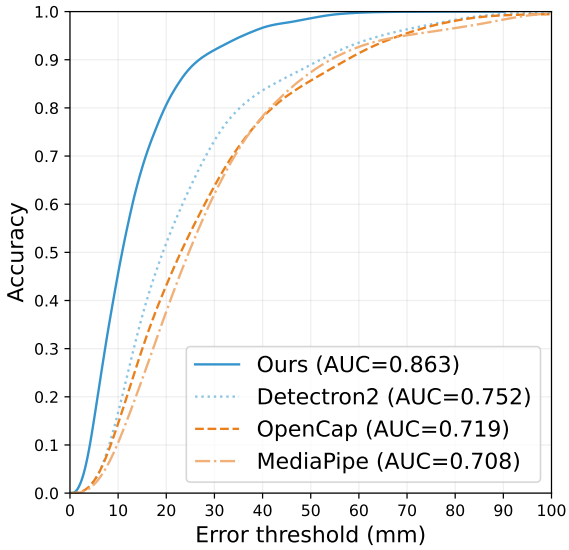

Fig. 13. Overall accuracy profile of our method, Detectron2, OpenCap, and MediaPipe from all the 12 key joint centers with a total of 1,169,671 Euclidean error samples from 10 test subjects from the RRIS40 test set.

the midpoint between FAL and TAM markers is used. For the elbow joint, the midpoint between HLE and HME markers is used. For the wrist joint, the midpoint between RSP and USP markers is used. The ground truth joint centers from marker-based mocap system are calculated using the same formula.

Our model is compared with three other tools: OpenCap, MediaPipe, and Detectron2. For OpenCap [8], its default OpenPose 2D keypoint detection, its own triangulation algorithm, and its default filter are used to obtain all the 3D joint center sequences. For MediaPipe [20], since it does not provide a multi-view pipeline, our 2D keypoint detection model is replaced by MediaPipe's pre-trained model. The subsequent step is performed on 2D joint centers using our triangulation algorithm. For Detectron2, our 2D keypoint detection model is replaced by the pre-trained model from Detectron2's repository [21]. This pre-trained model has the same neural network architecture as ours, which is the keypoint detection variant of Mask R-CNN. The differences are the set of keypoints, the source of training data, and annotation. Detectron2's pre-trained model is trained on Microsoft's COCO dataset [14] with 17 hand-annotated 2D keypoints.

The results are shown in Fig. 12 and Fig. 13. Our method has an overall joint center error of 13.71 ± 10.43 mm, which is significantly lower than Detectron2 (25.16 ± 24.11 mm), OpenCap (28.50 ± 22.14 mm), and MediaPipe (29.25 ± 19.53 mm). The results suggest that including anatomical marker-based annotations in the training data leads to improved accuracy.

D. Evaluation on GPJATK Dataset

Among all the datasets in Table I, GPJATK [19] is the only dataset that provides raw marker location data from a marker-based mocap system, along with synchronized RGB video data from more than two calibrated viewpoints. Therefore, this dataset is chosen to validate our method and check whether the learned anatomical keypoint can be transferred across different camera models, placements, and capture environments. Although our RRIS40 marker set shares 15 common markers with the GPJATK marker set, the placement position might differ slightly. For example, the shoulder marker for the RRIS40 marker set is positioned on the Acromion bone landmark, whereas the GPJATK marker set might place it on the Acromioclavicular joint, which is around one cm away.

In Fig. 14, it can be seen that our method outperforms OpenCap in all the overlapping markers. However, the performance is worse than the benchmark results on the RRIS40 test set. Several factors could explain these performance differences such as the placement of markers, a reduction in the number of cameras from 8 to 4, a decrease in camera resolution from 1920 × 1200 to 960 × 540, differences in the capture environment, or the use of a different calibration method. Additionally, the images captured around the head and shoulder area of the subjects are blurred to remove personal identifiers.

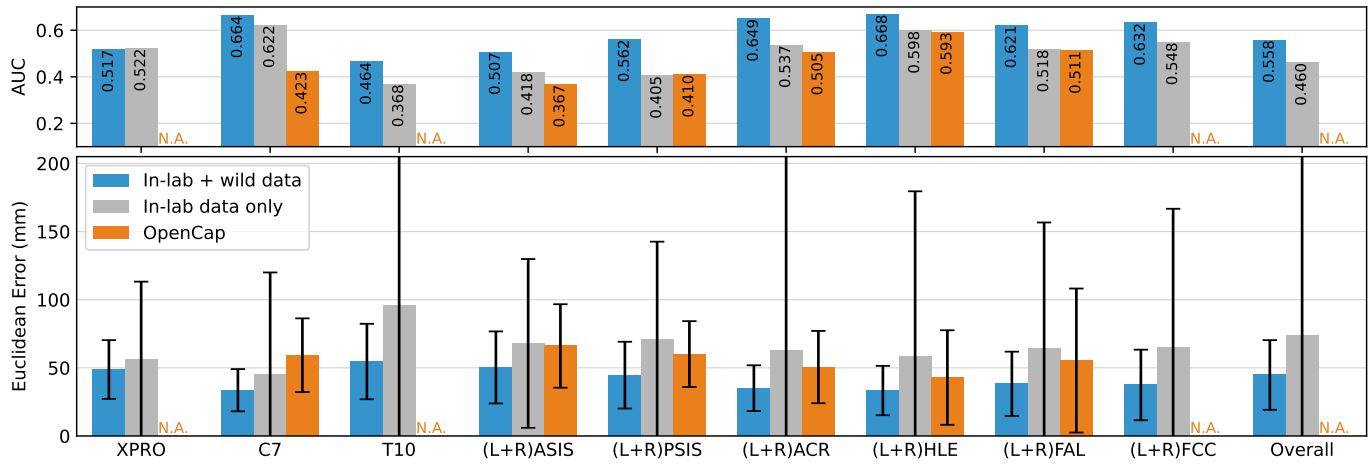

Fig. 14. Benchmarking results on the GPJATK dataset on 15 overlapping markers. **Bottom**: Error comparisons of different methods based on joint center (lower is better). The errors from our method are shown in two variants. The first variant is trained using a mixture of the RRIS40 dataset and the wild dataset. The second variant is trained using the RRIS40 dataset alone. Because the triangulated results from the second variant contain some gaps in marker trajectories, those gaps are always filled with linear interpolation. OpenCap results are included as references. **Top**: The ratio of area under the curve (AUC) from each virtual marker (higher is better). OpenCap does not provide XPRO, T10, LFCC, and RFCC to compare against those marker data in the GPJATK dataset.

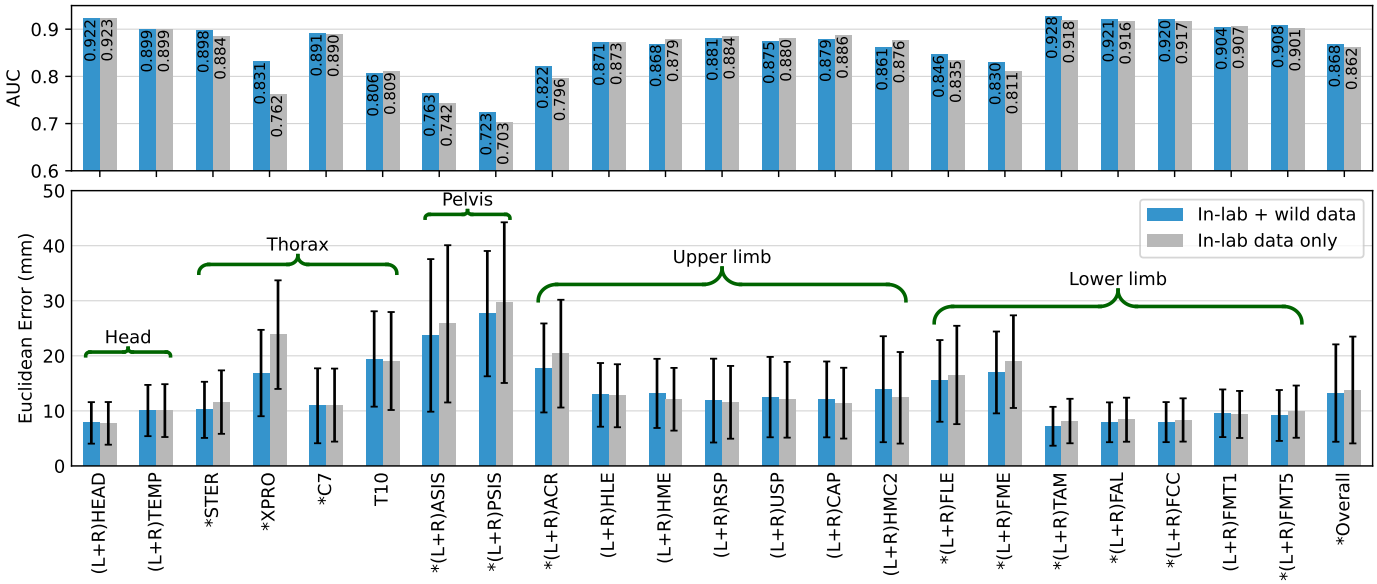

Fig. 15. Benchmarking results of two models on RRIS40 test set. **Bottom**: Error comparisons across all anatomical markers (lower is better). The errors from our method are shown in two variants. The first variant is trained using a mixture of the RRIS40 dataset and the wild dataset. The second variant is trained using the RRIS40 dataset alone. The * mark means that the errors from the first variant are significantly smaller than the errors from the second variant. **Top**: The ratio of area under the curve (AUC) from each virtual marker (higher is better).

E. Impact of Mixing Wild Images without Anatomical Keypoints in the Training Data

One common challenge of using training data from a single environment is the model's low transferability to a new environment. To determine whether the mixing of human-in-the-wild data can improve the model's transferability, another model is trained solely on the RRIS40 dataset and used as a baseline for comparison.

The comparison of both models on the RRIS40 test set is shown in Fig. 15. Although the model trained on mixed data shows a significantly smaller overall error, the difference is not very decisive as only 21 out of 40 markers display significantly lesser errors.

In contrast, when both models are compared on the GPJATK dataset which is an unseen environment, a clear difference is shown in Fig. 14. The model trained only on in-lab data produces much noisier results, as it struggles with human detection. In some frames, the model fails to detect the subject in one or more cameras, leading to a lower number of available triangulation rays. This ultimately results in either inaccurate triangulation or not enough rays to triangulate.

Incorporating wild images in the training data effectively eliminates overfitting from the human detection portion of the network, as demonstrated by the results obtained from the GPJATK dataset. Furthermore, this mixing of data does not appear to have any adverse impact on the accuracy of the

Fig. 16. **Top**: Snapshots of video records with 2D projections of 3D virtual markers while the subject is walking with different assistive outfits. **Bottom**: Qualitative results from the COCO dataset [14].

system in the seen environment like the RRIS40 test set.

F. Qualitative Evaluation on Assistive Outfits

Markerless mocap has a large potential for uses in rehabilitation research and assistive robotics. This is especially true in scenarios where marker-based mocap systems face challenges. For instance, when a subject is equipped with a walking aid, a robotic device, an exoskeleton, or a safety harness, marker attachment becomes difficult. This is because the markers cannot be attached to all the bone landmarks as they will be blocked by the obtrusive wearable pieces of equipment. Therefore, our markerless mocap system is tested to assess its stability and performance qualitatively when a subject wears unseen outfits.

In this experiment, a subject is recorded while walking with different types of walking aids which include a walking frame, an exoskeleton [44], a balance assistant robot [45], and an overhead body weight support system [46]. The results of a few frames together with some examples from the COCO dataset are shown in Fig. 16. Videos and additional images of these results are provided in the supplementary materials accessible through koonyook.github.io/rris40. Visually, all the triangulated 3D virtual markers are stable and well-aligned with the corresponding body parts.

G. Impact of Camera Reduction

To determine the impact of removing certain cameras, the entire RRIS40 test set is reprocessed 247 times with all possible camera combinations. The average error from each combination is plotted in Fig. 17. As expected, the range of average error increases as the number of cameras decreases. Since, the range of the average error is much larger when there are only two or three cameras, identifying the configurations

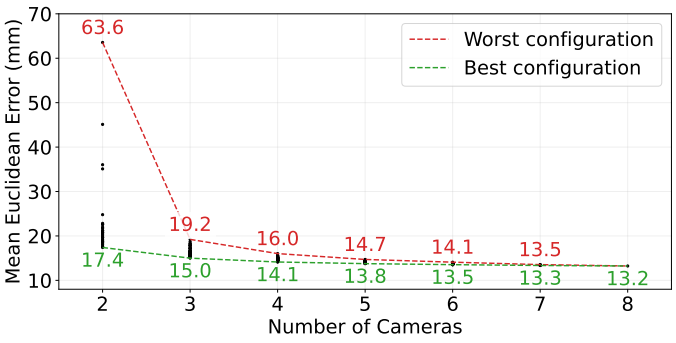

Fig. 17. The 247 average errors from 247 possible camera combinations. Each black point represents an average error from one combination of cameras when the results have been compared to marker-based mocap positions 4,081,976 times. When the number of cameras is limited, bad placement of the cameras can increase the overall measurement error up to 3.6 times.

TABLE II
THE BEST AND THE WORST CAMERA CONFIGURATIONS. THE POSITION OF EACH CAMERA INDEX IS SHOWN IN FIG. 4.

No. of cameras allowed	List of camera indexes in the best config.	in the worst config.
2	0, 7	3, 4
3	0, 3, 7	0, 5, 6
4	0, 2, 3, 7	0, 2, 5, 6
5	0, 2, 3, 4, 7	0, 1, 2, 5, 6
6	2, 3, 4, 5, 6, 7	0, 1, 2, 3, 5, 6
7	0, 2, 3, 4, 5, 6, 7	0, 1, 2, 3, 5, 6, 7
8	0, 1, 2, 3, 4, 5, 6, 7	

that contribute to the best and worst results can be helpful in optimizing the camera arrangement in many scenarios. Therefore, Table II is produced to analyze the characteristics of the best and worst camera placement combinations.

The worst configuration in the case of two cameras is when the cameras are positioned directly opposite each other. This arrangement is not ideal because the 2D keypoints extracted from cameras 3 and 4 provide very little information on the depth, that is, the global's X component. Even a slight 2D error in at least one of the cameras can significantly shift the triangulated position along the depth axis of both cameras. This issue can be avoided if the two cameras are positioned so that their depth axes are angled close to 90 degrees. That is why the best configuration in the case of two cameras is the one with cameras 0 and 7. These findings support the recommendation for optimal camera placement from Rahimian and Kearney [47].

H. Impact of Lower Image Resolution

The original video image with a resolution of RRIS40 test set is always scaled down to 1280 × 800 for the neural network's first layer input. To determine the impact of reduced image resolution, the image is scaled down to three additional resolutions: 320 × 200, 640 × 400, and 960 × 600 before scaling it back to 1280 × 800 for the network input. The results in Fig. 18 show a gradual increase in overall Euclidean error with lower image resolution. This decline in performance is likely due to the loss of spatial accuracy in 2D prediction,

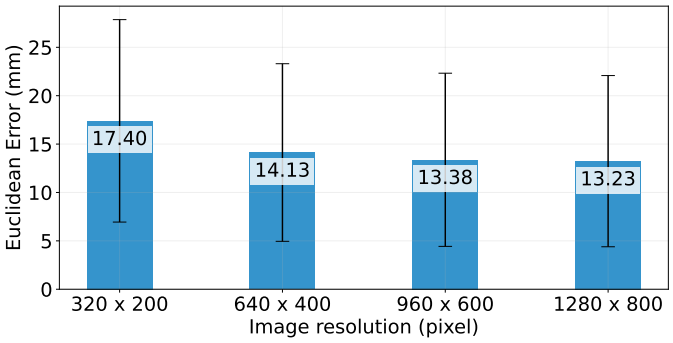

Fig. 18. Comparison of overall Euclidean errors with varying image resolutions.

which is expected as the back-projected ray from each camera can be sensitive to even a few pixels shift in 2D.

I. Impact of Neural Network Model Selection

In this ablation study, our Detectron2 is compared with two recent models: RTMPose-x [48] and DWPose-l [49]. The training is done on the MMPose platform [50]. Similarly to the original way of RTMPose-x training, the first and second stages of training with different augmentation pipelines run for 3 and 1 epochs respectively. Then, DWPose-l first-stage distillation training runs for 4 epochs using our trained RTMPose-x as the teacher and RTMPose-l as the student architecture. Then, the second-stage distillation is trained for 1 epoch. For all the training in this section, all the models are trained on one NVIDIA Titan RTX GPU using the same amount and mixture of datasets as our Detectron2 training but with a batch size of 72. To perform evaluation, RTMDet-m [51] is chosen for the human detection step before passing the bounding box to RTMPose-x and DWPose-l for 40 keypoint detection with the same detection input size of 384 $\times$ 288. The full inference pipeline (inclusive of RTMDet-m) for RTMPose-x and DWPose-l require computation of 56.35 and 48.44 GFlops/image respectively which are comparable to our Detectron2 with an input size of 1280 $\times$ 800 operating at 51.89 GFlops/image.

The results in Fig. 8 and Fig. 9 show that all three models are closely comparable. The mean Euclidean errors for our Detectron2, DWPose-l, and RTMPose-x are 13.23, 13.62, and 13.63 mm respectively. It is observed that Detectron2 may have falling-to-the-edge issues when the foot markers are close to an edge of the bounding box as shown in Fig. 16 e, f, and g. However, RTMPose-x and DWPose-l do not exhibit such issues, which could be due to their finer bin resolution. These findings show the potential of newer network architectures and offer opportunities for future exploration.

V. Limitations

Our system is designed for multi-view setups that capture the entire human body, which is commonly used in 3D mocap applications like gait analysis. However, as shown in the supplementary materials, its performance may be affected when dealing with heavily cropped human images or when tracking poses that are uncommon or absent in the training datasets. For example, actions like lying on the ground are absent from the RRIS40 dataset as the physical markers can be easily shifted or occluded during this action. Additionally, the use of the inpainting technique for marker removal may introduce unique artifacts that could bias the training dataset. Since the RRIS40 dataset mainly consists of Asian subjects, it may limit the adaptability of our model to other demographic groups. Although variations in height, gender, and body shape within a normal range do not significantly affect the model's generalizability, extra large body shapes may require further investigation. Future work could focus on expanding the diversity of datasets with more challenging poses and interactions among multiple subjects to enhance the system's capabilities.

VI. Conclusion

Our research work contributes to advancing the state-of-the-art in markerless motion capture (mocap), addressing the limitations of existing methods, and facilitating applications in clinical biomechanics and sports science. An important advantage of markerless mocap is that it reduces the time and manpower required for subject preparation and data post-processing. This makes the mocap workflow more efficient and practical for real-world applications. Future research may focus on more challenging poses and multi-subject tracking.

Acknowledgment

We wish to thank the ability data team at the Rehabilitation Research Institute of Singapore (RRIS) for their help in collecting and post-processing the mocap data. We are grateful to all the subjects who contributed to the RRIS40 dataset and support from the Guangdong Zhongxin Intelligent Rehabilitation Research Institute affiliated with Guangdong Jian Xiang Hospital Group We also thank Bharatha Selvaraj, Shijia Han, and Mark Nelson for their assistance and support in software development.

References

[1] S. L. Colyer, M. Evans, D. P. Cosker, and A. I. T. Salo, "A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system," *Sports Medicine - Open*, vol. 4, no. 24, pp. 1–15, Jun. 2018.

[2] W. W. T. Lam, Y. M. Tang, and K. N. K. Fong, "A systematic review of the applications of markerless motion capture (mmc) technology for clinical measurement in rehabilitation," *J. NeuroEng. Rehabil.*, vol. 20, no. 57, pp. 1–26, May 2023.

[3] L. Wade, L. Needham, P. McGuigan, and J. Bilzon, "Applications and limitations of current markerless motion capture methods for clinical gait biomechanics," *PeerJ*, vol. 10, pp. 1–27, Feb. 2022.

[4] S. Jeon, K. M. Lee, and S. Koo, "Anomalous gait feature classification from 3-d motion capture data," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 2, pp. 696–703, Feb. 2022.

[5] E. Ceseracciu, Z. Sawacha, and C. Cobelli, "Comparison of markerless and marker-based motion capture technologies through simultaneous data collection during gait: Proof of concept," *PLOS ONE*, vol. 9, no. 3, pp. 1–7, Mar. 2014.

[6] S. Vafadar, W. Skalli, A. Bonnet-Lebrun, M. Khalifé, M. Renaudin, A. Hamza, and L. Gajny, "A novel dataset and deep learning-based approach for marker-less motion capture during gait," *Gait & Posture*, vol. 86, pp. 70–76, May 2021.

[7] M. Moro, G. Marchesi, F. Hesse, F. Odone, and M. Casadio, "Markerless vs. marker-based gait analysis: A proof of concept study," *Sensors (Basel)*, vol. 22, no. 5, pp. 1–15, Mar. 2022.

[8] S. D. Uhlrich, A. Falisse, Ł. Kidziński, J. Muccini, M. Ko, A. S. Chaudhari, J. L. Hicks, and S. L. Delp, "Opencap: 3d human movement dynamics from smartphone videos," *Plos Computational Biology*, vol. 19, no. 10, pp. 1–26, Oct. 2023.

[9] M. Kitagawa and B. Windsor, "An overview and history of motion capture," in *MoCap for Artists Workflow and Techniques for Motion Capture*. Boston: Focal Press, 2008.

[10] R. J. Cotton, A. DeLillo, A. Cimorelli, K. Shah, J. Peiffer, S. Anarwala, K. Abdou, and T. Karakostas, "Markerless motion capture and biomechanical analysis pipeline," 2023. [Online]. Available: https://arxiv.org/abs/2303.10654

[11] A. Mathis, S. Schneider, J. Lauer, and M. W. Mathis, "A primer on motion capture with deep learning: Principles, pitfalls, and perspectives," *Neuron*, vol. 108, no. 1, pp. 44–65, Oct. 2020.

[12] S. Ghorbani, K. Mahdaviani, A. Thaler, K. Kording, D. J. Cook, G. Blohm, and N. F. Troje, "Movi: A large multi-purpose human motion and video dataset," *PLOS ONE*, vol. 16, no. 6, pp. 1–15, Jun. 2021.

[13] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conf. Comput. Vision Pattern Recognit.*, Columbus, USA, Jun. 2014, pp. 3686–3693.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Eur. Conf. Comput. Vision*, Zurich, Switzerland, Sep. 2014, pp. 740–755.

[15] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, and P. Luo, "Whole-body human pose estimation in the wild," in *Eur. Conf. Comput. Vision*, Glasgow, UK, Aug. 2020, pp. 196–214.

[16] L. Needham, M. Evans, D. P. Cosker, L. Wade, P. M. McGuigan, J. L. Bilzon, and S. L. Colyer, "The accuracy of several pose estimation methods for 3d joint centre localisation," *Scientific Reports*, vol. 11, pp. 1–11, Oct. 2021.

[17] R. M. Kanko, E. K. Laende, G. Strutzenberger, M. Brown, W. S. Selbie, V. DePaul, S. H. Scott, and K. J. Deluzio, "Assessment of spatiotemporal gait parameters using a deep learning algorithm-based markerless motion capture system," *J. Biomechanics*, vol. 122, pp. 1–7, Jun. 2021.

[18] E. Wu, H. Nishioka, S. Furuya, and H. Koike, "Marker-removal networks to collect precise 3d hand data for rgb-based estimation and its application in piano," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, Waikoloa, USA, Jan. 2023, pp. 2976–2985.

[19] B. Kwolek, A. Michalczuk, T. Krzeszowski, A. Switonski, H. Josinski, and K. Wojciechowski, "Calibrated and synchronized multi-view video and motion capture dataset for evaluation of gait recognition," *Multimedia Tools and Appl.*, vol. 78, pp. 32 437–32 465, Aug. 2019.

[20] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for perceiving and processing reality," in *3rd Workshop on Comput. Vision for AR/VR at IEEE Conf. Comput. Vision and Pattern Recognit.*, Long Beach, USA, Jun. 2019, pp. 1–9.

[21] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[22] Y. Desmarais, D. Mottet, P. Slangen, and P. Montesinos, "A review of 3d human pose estimation algorithms for markerless motion capture," *Comput. Vision Image Understanding*, vol. 212, pp. 1–19, Nov. 2021.

[23] L. Mündermann, S. Corazza, and T. P. Andriacchi, "The evolution of methods for the capture of human movement leading to markerless motion capture for biomechanical applications," *J. NeuroEng. Rehabil.*, vol. 3, no. 6, pp. 1–11, Mar. 2006.

[24] S. Corazza, L. Mündermann, E. Gambaretto, G. Ferrigno, and T. P. Andriacchi, "Markerless motion capture through visual hull, articulated icp and subject specific model generation," *Int. J. Comput. Vision*, vol. 87, pp. 156–169, Sep. 2010.

[25] H. Frühschütz, "Evaluation of markerless tracking for kinematics in tennis," Master's thesis, Technical University of Munich, Mar. 2017, available at https://mediatum.ub.tum.de/node?id=1362108.

[26] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[27] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *IEEE Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, USA, Jun. 2018, pp. 7297–7306.

[28] L. Needham, M. Evans, L. Wade, D. P. Cosker, M. P. McGuigan, J. L. Bilzon, and S. L. Colyer, "The development and evaluation of a fully automated markerless motion capture workflow," *J. Biomechanics*, vol. 144, pp. 1–9, Oct. 2022.

[29] T. J. Dobos, R. W. G. Bench, C. D. McKinnon, A. Brady, K. J. Boddy, M. W. R. Holmes, and M. W. L. Sonne, "Validation of pitchAI™ markerless motion capture using marker-based 3d motion capture," *Sports Biomechanics*, pp. 1–21, Oct. 2022.

[30] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vision*, vol. 87, pp. 4–27, Mar. 2010.

[31] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[32] M. Keller, K. Werling, S. Shin, S. Delp, S. Pujades, C. K. Liu, and M. J. Black, "From skin to skeleton: Towards biomechanically accurate 3d digital humans," *Trans. Graph.*, vol. 46, no. 6, pp. 1–12, Dec. 2023.

[33] S. Vafadar, W. Skalli, A. Bonnet-Lebrun, A. Assi, and L. Gajny, "Assessment of a novel deep learning-based marker-less motion capture system for gait study," *Gait & Posture*, vol. 94, pp. 138–143, May 2022.

[34] P. Liang, W. H. Kwong, A. Sidarta, C. K. Yap, W. K. Tan, L. S. Lim, P. Y. Chan, C. W. K. Kuah, S. K. Wee, K. Chua, C. Quek, and W. T. Ang, "An asian-centric human movement database capturing activities of daily living," *Scientific Data*, vol. 7, no. 1, pp. 1–13, Sep. 2020.

[35] J. D. Richards, *Biomechanics in Clinic and Research*. London: Churchill Livingstone, 2008, p. 265.

[36] U. D. Croce, A. Leardini, L. Chiari, and A. Cappozzo, "Human movement analysis using stereophotogrammetry: Part 4: assessment of anatomical landmark misplacement and its effects on joint kinematics," *Gait & Posture*, vol. 21, no. 2, pp. 226–237, Feb. 2005.

[37] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *IEEE Conf. Comput. Vision Pattern Recognit.*, Salt Lake City, USA, Jun. 2018, pp. 5505–5514.

[38] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *IEEE Int. Conf. Comput. Vision*, Seoul, South Korea, Oct. 2019, pp. 4470–4479.

[39] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE Int. Conf. Comput. Vision*, Venice, Italy, Oct. 2017, pp. 2980–2988.

[40] C-Motion, "Hip joint landmarks - visual3d wiki documentation," Nov. 2017. [Online]. Available: https://c-motion.com/v3dwiki/index.php/Hip_Joint_Landmarks

[41] A. L. Bell, D. R. Pedersen, and R. A. Brand, "A comparison of the accuracy of several hip center location prediction methods," *J. Biomechanics*, vol. 23, no. 6, pp. 617–621, 1990.

[42] C-Motion, "Tutorial: Rab upper extremity model - visual3d wiki documentation," Feb. 2014. [Online]. Available: https://c-motion.com/v3dwiki/index.php?title=Tutorial:_Rab_Upper_Extremity_Model

[43] G. Rab, K. Petuskey, and A. Bagley, "A method for determination of upper extremity kinematics," *Gait & Posture*, vol. 15, no. 2, pp. 113–119, Apr. 2002.

[44] L. Luo, M. J. Foo, M. Ramanathan, J. K. Er, C. H. Chiam, L. Li, W. Y. Yau, and W. T. Ang, "Trajectory generation and control of a lower limb exoskeleton for gait assistance," *J. Intell. Robotic Syst.*, vol. 106, no. 64, pp. 1–15, Nov. 2022.

[45] L. Li, M. J. Foo, J. Chen, K. Y. Tan, J. Cai, R. Swaminathan, K. S. G. Chua, S. K. Wee, C. W. K. Kuah, H. Zhuo, and W. T. Ang, "Mobile robotic balance assistant (mrba): a gait assistive and fall intervention robot for daily living," *J. NeuroEng. Rehabil.*, vol. 20, no. 29, pp. 1–17, Mar. 2023.

[46] M. Bannwart, M. Bolliger, P. Lutz, M. Gantner, and G. Rauter, "Systematic analysis of transparency in the gait rehabilitation device the float," in *14th Int. Conf. on Control, Automation, Robotics and Vision (ICARCV)*. Phuket, Thailand: IEEE, Nov. 2016, pp. 1–6.

[47] P. Rahimian and J. K. Kearney, "Optimal camera placement for motion capture systems," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 3, pp. 1209–1221, Mar. 2017.

[48] T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, "Rtmpose: Real-time multi-person pose estimation based on mmpose," 2023. [Online]. Available: https://arxiv.org/abs/2303.07399

[49] Z. Yang, A. Zeng, C. Yuan, and Y. Li, "Effective whole-body pose estimation with two-stages distillation," in *IEEE Int. Conf. Comput. Vision Workshop*, Paris, France, Oct. 2023, pp. 4212–4222.

[50] MMPose Contributors, "Openmmlab pose estimation toolbox and benchmark," 2020. [Online]. Available: https://github.com/open-mmlab/mmpose

[51] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen, "Rtmdet: An empirical study of designing real-time object detectors," 2022. [Online]. Available: https://arxiv.org/abs/2212.07784