Journal of Biomechanics 203 (2026) 113341

Contents lists available at ScienceDirect

Journal of Biomechanics

journal homepage: www.elsevier.com/locate/jbiomech

Effect of markers in training dataset for markerless applications in biomechanics

Lucas Mercier[a, b, *], Thierry Cresson[a, b], Neila Mezghani[c], Sylvie Gervais[b], Carlos Vázquez[a, b]

[a] Laboratoire d'innovation ouverte en technologies de la santé, CRCHUM, 900 Rue Saint-Denis, Montréal, H2X 0A9, QC, Canada
[b] École de technologie supérieure, 1100 R. Notre Dame O, Montréal, QC H3C 1K3, Montréal, H3C 1K3, QC, Canada
[c] Centre de recherche LICEF, Université TÉLUQ, 5800 Rue Saint-Denis, Montréal, H2S 3L4, QC, Canada

HIGHLIGHTS

- Reflective markers introduce bias in pose estimation model training.
- Standard approaches will fail in marker-free clinical settings.
- Inpainting bridges the gap to markerless clinical analysis.

ARTICLE INFO

Keywords:
Markerless motion capture
Markers inpainting
Deep learning

ABSTRACT

The quality of dataset annotations used to train markerless motion capture models is crucial for obtaining reliable joint center estimations from videos. Because manually annotated datasets such as COCO are unsuitable for biomechanical applications, a common recommendation is to use videos synchronized with marker-based MoCap datasets. In these systems, reflective markers are placed on the skin surface, ideally on bony landmarks, and are then tracked by optical cameras to obtain highly accurate joint center annotations. While previous studies have suggested that visible reflective markers on images could bias model training, this effect has not been formally demonstrated. To address this, we used two MoCap datasets: one with 26 subjects each equipped with reflective markers mounted to rigid bodies, and the second, with 10 subjects with markers placed on bony landmarks. This allowed us to train pose estimation networks on images having visible markers. The models were then evaluated on test images with visible and inpainted markers. Our findings showed that when models were evaluated on images with markers inpainted, pixel position errors increased by +4.7% to +51.1% versus images with visible markers. This indicates that the presence of markers in training images can affect human pose estimation algorithms. To still utilize accurate annotations from MoCap, we recommend training on images with markers removed via inpainting. We also demonstrated that, in that case, the network does not rely on inpainted areas to estimate joint centers, thus making it a viable solution to the presence of markers in training images.

1. Introduction

Recent advances in deep learning indicate a promising future for markerless motion capture (Colyer et al., 2018; Uhlrich et al., 2023). Whereas traditional methods (referred to as MoCap in this study) rely on reflective markers placed on the patient to obtain joint center and/or bony landmark positions (Winter, 2009), markerless approaches rely solely on images from video cameras to estimate these positions (Avogaro et al., 2023). In this context, neural networks are trained to extract features such as edges, colors and textures from images (Zeiler et al., 2014) and to learn a mapping of these to joint positions. This training requires many examples with images and corresponding annotations for joint center positions (Sarandi et al., 2023).

Yet, collecting such a large volume of high quality data can be challenging (Neupane et al., 2024). This is especially true for specific populations, such as clinical patients with movement pathologies or elite athletes, where participant availability is restricted. Furthermore, the acquisition of such data relies on gold-standard motion capture (MoCap) systems, which require significant financial investment and specialized technical expertise to operate. For this reason, many authors (e.g.,

* Corresponding author at: Laboratoire d'innovation ouverte en technologies de la santé, CRCHUM, 900 Rue Saint-Denis, Montréal, H2X 0A9, QC, Canada.
Email address: lucas.mercier.1@ens.etsmtl.ca (L. Mercier).

https://doi.org/10.1016/j.jbiomech.2026.113341
Accepted 27 April 2026
Available online 29 April 2026

D'Antonio et al. (2020); Needham et al. (2021); Stenum et al. (2021)) use models such as HRNet (Sun et al., 2019), OpenPose (Cao et al., 2021) or RTMPose (Jiang et al., 2023) trained on manually annotated public datasets such as COCO (Lin et al., 2014).

However, these datasets are not suitable for biomechanical applications requiring specific movements, such as walking on a treadmill or performing squats, because they lack annotations for the movements (Seethapathi et al., 2019; Needham et al., 2021; Colyer et al., 2018; Wade et al., 2022). Some joint centers may be defined differently than in MoCap-based systems, or may simply not be provided (Seethapathi et al., 2019; Needham et al., 2021; Colyer et al., 2018; Wade et al., 2022). Moreover, as the datasets are manually annotated, it is impossible to quantify errors in their joint center definitions, unlike traditional marker-based MoCap systems. To train and validate pose estimation models with more suitable datasets, marker-based MoCap systems synchronized with video cameras can be used to collect such images with corresponding joint position annotations (Vafadar et al., 2021; Wang et al., 2021; Guo et al., 2025).

One limitation inherent in using these setups is that with them, markers are visible on training images (Hampali et al., 2020; Jatesiktat et al., 2024). As these markers have a very specific shape and color, features extracted from networks will most likely contain information about markers and the model could rely on these features to estimate joint centers (Hampali et al., 2020; Jatesiktat et al., 2024; Wu et al., 2023). In a clinical context, the absence of markers could then reduce the model's performance.

To circumvent this limitation, some authors (Hampali et al., 2020; Jatesiktat et al., 2024; Wu et al., 2023) have proposed removing markers from images when training human pose estimation models. However, the authors did not directly demonstrate the impact of markers. The study that has come closest to investigating this impact was conducted in the context of 6D object pose estimation (Rosskamp et al., 2024) where it was shown that models trained with visible markers perform worse on markerless images (images where markers were never present). To the best of our knowledge, no previous study has demonstrated this for human pose estimation.

The primary objectives of this study were twofold: first, to quantify the training bias induced by the visual presence of motion capture markers on human pose estimation; and second, to determine if inpainting techniques can serve as a viable way to prevent this bias and maintain performance when applied to images with no markers (referred to as markerless images). We hypothesized that (H1) the presence of markers in training images significantly degrades model performance when applied to markerless clinical images. We further hypothesized that (H2) training models on inpainted images effectively removes this bias, allowing for high-precision pose estimation in markerless contexts without suffering from prediction degradation.

In this study, we focus on 2D pose estimation because most 3D equivalent pipelines rely on predicting 2D joint centers (Neupane et al., 2024; Wang et al., 2021), adding either a triangulation step (Uhlrich et al., 2023) or incorporating a 2D-to-3D lifting model afterward (Peng et al., 2024). Since standard MoCap systems utilize either large rigid bodies with mounted markers (Fig. 1a) or groups of smaller markers (Fig. 1b), we anticipate that their varying visual footprints might induce distinct training biases. We thus conduct experiments on both types of marker setups.

2. Methods

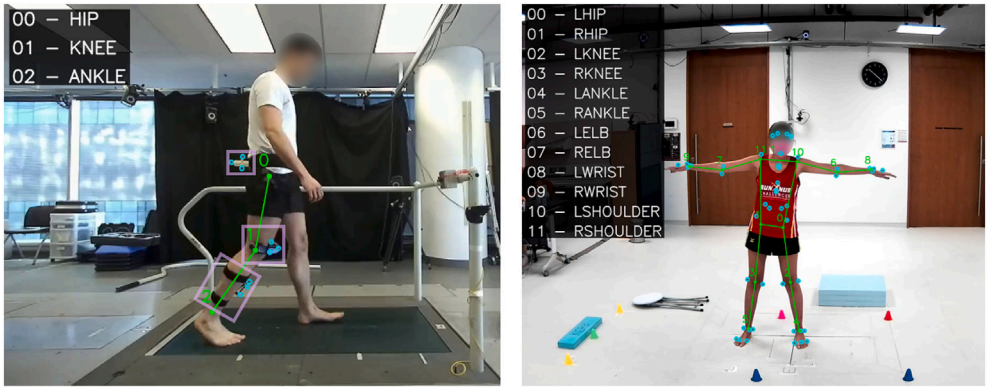
We first present the steps used to investigate both hypotheses, especially the dataset collection, the inpainting method and the pose estimation model selection (Fig. 2A). We then present the methodology used to evaluate the impact of markers in training images (H1) (Fig. 2B) and finally the approach used to assess whether predictions from models trained with inpainted images will deteriorate on markerless images (H2) (Fig. 2C).

2.1. Experimental framework

This section details the common experimental framework used to investigate both research hypotheses. First, we describe the two datasets (KKGVideo and RRIS40) which represent distinct marker configurations (rigid bodies and groups of small markers, respectively). Next, we present the inpainting process, which serves as a testing tool for the evaluation of H1 and as a training protocol for the validation of H2. Finally, we describe the HRNet pose estimation model and training configuration, which remain consistent across all experiments.

2.1.1. Dataset description

KKGVideo dataset – Reflective markers mounted to rigid bodies: Data from 26 healthy adults were collected for this study which was approved by the Ethics Committee of the Centre hospitalier de l'Université de Montréal Research Center. The subjects were equipped with the KneeKG^{TM} system (Emovi Inc., Canada), a reflective markers mounted to a rigid body system designed to reduce soft tissue artifacts for gait analysis (Lustig et al., 2012). These groups of markers define a technical axis system and, combined with a functional calibration (Hagemeister et al., 2005), allowed us to estimate the three-dimensional (3D) joint center positions of the hip, knee and ankle joints (Fig. 1a). A Zed2i^{TM} video camera (Stereolabs, San Francisco, USA) with a resolution of 1280×720 pixels at 60Hz was synchronized with the KneeKG



(a) (b)

Fig. 1. Position of the rigid bodies (in purple) and group of reflective markers (blue) with regard to 2D joint center annotations (green) for both, the KKGVideo (a) and RRIS40 (b) datasets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

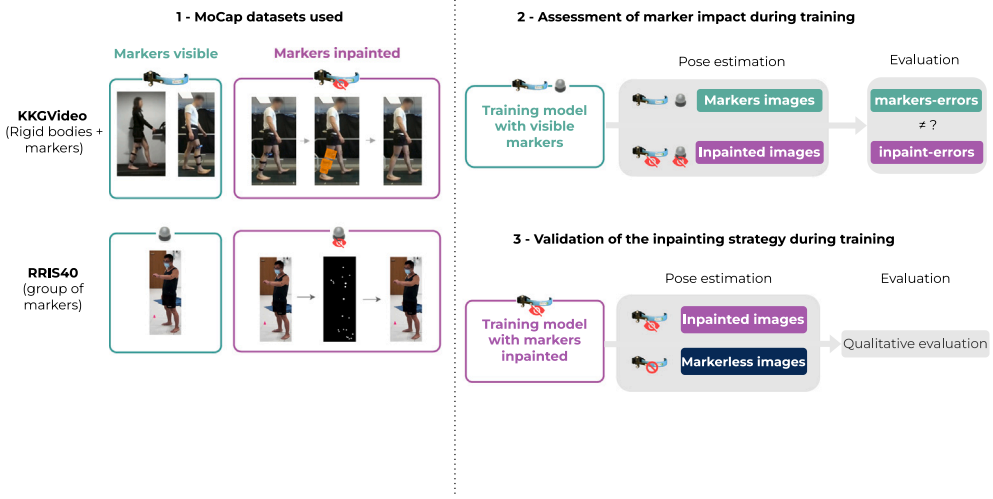


Fig. 2. Overview of the proposed method to answer both research questions. In (1), datasets used for each experiment. In (2), assessment of marker impact during training. In (3), validation of the inpainting strategy during training.

acquisition system to obtain corresponding images with 3D joint center positions projected into image space. We also recorded videos of participants during the warmup phase, during which they didn't wear any markers (referred to as markerless images).

RRIS40 dataset: The RRIS40 dataset (Jatesiktat et al., 2024) is a subset of the RRIS Asian-centric dataset (Liang et al., 2020). Only the test subset was made available, and consisted of ten participants, 5 males and 5 females. They were equipped with 40 small, 12.5 mm-diameter reflective markers (Qualisys, Göteborg, Sweden). The system was synchronized with 8 video cameras (See3CAM 24CUG e-con Systems India Pvt Ltd, Chennai, India) having a resolution of 1920×1200 pixels at 200 Hz. They performed different movements, including torso rotations and squats. To minimize cases where limb markers were occluded, we relied solely on the frontal camera for each subject. Based on 3D marker positions, we computed the 3D positions of 12 joint centers (hips, knees, ankles, shoulders, elbows and wrists) using the method proposed by (Dumas and Wojtusch, 2018) and projected them into image space (Fig. 1b).

2.1.2. Markers inpainting

To remove markers from the images, we employed a deep learning-based technique called inpainting. This generative approach takes an input image and a binary mask identifying the marker locations, and replaces these areas with realistic textures of skin and/or clothes such that the resulting image appears as though the markers were never present (Suvorov et al., 2022; Xiang et al., 2023). For smaller markers (RRIS40 dataset), their 3D positions from the MoCap system are projected into image space and a radius based on the markers' diameters defines the regions to inpaint (Fig. 3(A).a). For rigid bodies such as the KneeKG, one bounding box for each element of the system (femoral harness and both tibial straps) is placed automatically around them based on known joint center positions (Fig. 3(B).a). These regions to inpaint, represented by circles for small markers or bounding boxes for rigid bodies, are fed alongside the original images into the Large Mask (LaMa) inpainting model (Suvorov et al., 2022). LaMa relies on a convolutional network to capture the entire context of the image to remove markers from images (Rosskamp et al., 2024). The available LaMa model was provided already trained on two databases, namely, Places (Zhou et al., 2018) and CelebA (Karras et al., 2018), to generate realistic textures of humans from partially masked images.

2.1.3. Model implementation

Model selection: We chose the HRNet-w48 (Sun et al., 2019) for its performance on pose estimation benchmarks (Zheng et al., 2023) and its popularity in the literature (Gozlan et al., 2025). It follows a similar architecture to models such as OpenPose (Cao et al., 2021) or ViTPose (Xu et al., 2022) as it extracts features from images and then estimates joint centers in image space. Its main difference from previous approaches is that the latter tend to downscale the image resolution heavily during this process, while HRNet keeps it at a higher resolution, leading to better performance (Sun et al., 2019). This model outputs heatmaps, i.e., images where each pixel represents the probability of containing a joint center. The loss used is the mean-squared error loss between predicted $\hat{h}$ and ground truth heatmaps h (Eq. 1) over J joints. Heatmap annotations are created by placing a 2D Gaussian with $\sigma = 2$ pixels at each joint position in the image.

$$Loss = \frac{1}{J}\sum_{j=1}^{J}\left(h_j - \hat{h}_j\right)^2 \tag{1}$$

Weight initialization: The weights of the network (LeCun et al., 2015) are automatically updated during training to help the pose estimation model learn patterns from images and make optimal predictions. Two approaches can be found in the literature for model weight initialization. One of them, such as (Cronin et al., 2019) relies on weights obtained from ImageNet (Deng et al., 2009). The second one leverages weights from markerless pose estimation datasets such as COCO (Lin et al., 2014), with an example being that of (Guo et al., 2025). The two approaches differ primarily in their initialization. In the approach referred to as 'pose pretraining', the model is initially trained on a large-scale, markerless pose estimation dataset (such as COCO). Consequently, the network has already learned a foundational representation of human anatomy and joint locations from images without relying on markers. Conversely, if no pose pretraining is involved, the model starts the learning process from scratch and might learn to associate visual cues from markers with joint center positions. The impact of markers and inpainting can thus vary with the approach (with or without pose pretraining) used, so we used both approaches for the two datasets.

Hyperparameter search: Hyperparameters are used to control different aspects of the learning process (Bergstra and Bengio, 2012). To find the best set of hyperparameters yielding the most accurate and

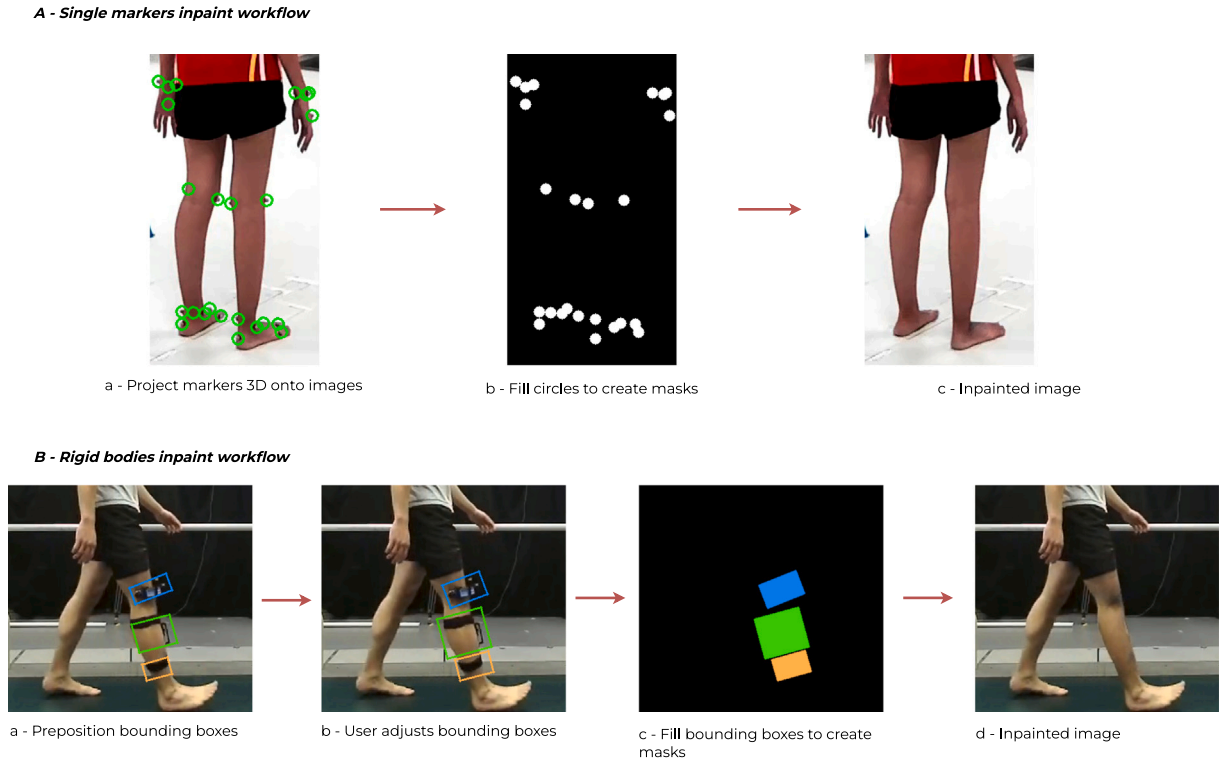


Fig. 3. Overview of the inpainting method depending on the marker type: single markers from RRIS40 (A) or markers mounted on rigid bodies from KKGVideo (B).

best-generalizing model (in our case, the learning rate, the scheduler and the model dropout), we performed an optimization using a random search over the hyperparameter space (Bergstra and Bengio, 2012), evaluating each configuration using a three-fold cross-validation (Kohavi, 1995). The selection criterion was defined as the average per-joint pixel error E_{avg} between the predicted $\hat{y}$ and the ground truth y pixel joint coordinates over J joints (Eq. 2). The configuration with the lowest error was then selected and used to train the final model. The hyperparameter search was performed for each subsequent model training.

$$E_{\text{avg}} = \frac{1}{J}\sum_{j=1}^{J}\left\|y_j - \hat{y}_j\right\|_2 \tag{2}$$

2.2. Marker impact during training (H1)

2.2.1. Model configuration and training

Dataset used: We trained the HRNet model on both the KKGVideo (with rigid bodies) and RRIS40 (with groups of markers) datasets to predict joint center positions from images with visible markers (Table 1). For KKGVideo, we used 16 subjects for training and validation and 10 for testing (with 2700 images per subject). As RRIS40 has only 10 subjects (1000 frames per subject), we employed a leave-one-out strategy (Kohavi, 1995) to ensure we would have sufficient subjects for evaluation and statistical analysis. Each of the 10 subjects was set aside in turn: the model was trained on the remaining 9 (with 6 used for training and 3 for validation), and testing was performed on the held-out subject.

Models configurations: As described in (Section 2.1.3), the weight initialization (with or without pose pretraining) can change the impact that markers have on training images. For each dataset, we trained two HRNet configurations on images with markers visible: with and without pose pretraining (Table 1). A hyperparameter search (Section 2.1.3) was performed for all configurations.

Table 1
Summary of all configurations used for training on both datasets with visible markers.

Config #	Alias	Dataset	Visible markers	Pose pretrain on COCO
1	hrnet-kkgvideo	kkgvideo	✓	
2	hrnet-kkgvideo-pretrain	kkgvideo	✓	✓
3	hrnet-rris40	rris40	✓	
4	hrnet-rris40-pretrain	rris40	✓	✓

2.2.2. Quantitative evaluation of markers impact

Error comparison: All the four model configurations of Table 1 were used to estimate joint centers from images of their respective test sets, which contained only subjects not seen during training. The estimation was performed on images with visible and inpainted markers. Model performance was measured using the average pixel error between the predicted and ground-truth 2D joints obtained with the marker-based MoCap system, over all test images. We also used the Object Keypoint Similarity (OKS) (Papandreou et al., 2017), which measures the normalized keypoint overlap, where 0 indicates no overlap between predictions and ground-truth, and 1 indicates full overlap; higher is better.

Statistical analysis: To assess the differences between errors of models trained with visible markers (Table 1) on images with visible and inpainted markers, a three-way repeated measures ANOVA was performed. The dependent variable was initially calculated as the mean pixel error between predicted and ground-truth joints per subject for each experimental condition. When residual diagnostics indicated non-normality, we instead aggregated the raw data by calculating the median pixel error. This robustified the dependent variable against skewed distributions and extreme values, allowing us to apply the same repeated-measures ANOVA framework while successfully meeting the parametric assumption of normality. The three within-subject factors were the timestamp (moment in time), the experimental condition (image with markers vs. inpainted images) and the joint. This analysis aimed to determine whether the error differed significantly across the different levels of these factors and their interactions. When

a significant three-way interaction was observed, a two-way repeated measures ANOVA was performed for each joint level, with the experimental condition and timestamp as factors. Where the timestamp effect was not significant, comparisons between the two experimental conditions were conducted using a Student's *t*-test when the normality assumption was met, or a Wilcoxon signed-rank test otherwise, to assess whether the error was statistically higher when markers were inpainted than when they were visible. Effect sizes for these comparisons were calculated using Cohen's d for t-tests and the rank-biserial correlation (r) for Wilcoxon tests. The significance level was set to $p < 0.05$ for all analyses.

2.2.3. Qualitative evaluation of markers impact

If the model trained with visible markers (Table 1) relies on marker information in images, then it should behave differently on images that don't contain markers. To visualize the areas of the image the trained networks (Table 1) used for their predictions, we employed a Class Activation Map (CAM) (Zhou et al., 2016) from the last layers of the network. We compared the CAMs produced on images with and without markers to assess any differences between these two configurations.

2.3. Inpainting strategy during training (H2)

To use the inpainting to remove markers during training, we needed to make sure that this process wouldn't create "hidden patterns" in images that would bias the model estimation. We thus investigate if a model trained on inpainted images would perform worse on markerless images.

2.3.1. Model training

As we only had access to markerless images for the KKGVideo dataset, we focused on it to investigate the impact of inpainting. We thus trained an instance of HRNet directly on inpainted images (Jatesiktat et al., 2024; Rosskamp et al., 2024) from the KKGVideo dataset, leveraging pose pretraining from COCO and selected the most generalizing dataset via hyperparameter optimization. We kept the same subject split as for the first hypothesis experiments (Section 2.2.1).

2.3.2. Qualitative evaluation of inpainting impact

We took markerless images (from the warmup phase of the KneeKG acquisition) from 5 out of the 12 test subjects (three males and two females). For each of these 5 subjects, we selected 4 images at different moments of their gait cycles and then manually created masks indicating where the KneeKG would be placed on the leg (Fig. 4). This gave an inpainted version of the original markerless image. We then used the model trained on inpainted images to obtain CAMs on both images and compared them to assess whether the model focused on different parts of the image.

3. Results

3.1. Assessment of marker impact during training (H1)

3.1.1. Quantitative assessment of the impact of markers in training images

Error comparison. Table 2 reports per-joint pixel errors of both configurations for the model trained on KKGVideo (Table 2a) and RRIS40 (Table 2b) dataset images with visible markers. The evaluation was performed both on images with visible markers and on those with markers removed via inpainting. Because the distribution of the average per-joint pixel error was notably skewed for both datasets, we provided the median and interquartile range (IQR). Since the resulting errors on the RRIS40 dataset (Table 2b) for left and right joints were close, we averaged both sides to improve the readability of the table. The last row presents the OKS for each dataset.

Statistical analysis. For all models trained on images with visible markers (Table 1), the repeated measures of ANOVA indicated a significant interaction with joint factors, while no significant effect of the timestamp was found. Because the normality assumption was not met, the Fig. 5 presents the results of the comparison between the experimental conditions using Wilcoxon sign rank test for *hrnet-kkgvideo* and *hrnet-kkgvideo-pretrain* configurations. Without pose pretraining (*hrnet-kkgvideo*), the median pixel error on inpainted images was higher than on images with visible markers for the hip ($p = 0.002$, $r = 0.96$) and the knee ($p < 0.001$, $r = 1.00$) (see Table 2 for full descriptive statistics; Fig. 5a). For the model with pose pretraining (*hrnet-kkgvideo-pretrain*), both the hip ($p = 0.002$, $r = 0.96$) and the knee ($p = 0.005$, $r = 0.89$) presented statistically higher pixel errors on images with markers inpainted (Fig. 5b). The difference between inpainted and marker images is especially visible on the knee joint center, with an error increase of +425% ($p < 0.001$, $r = 1.00$) without pretraining (*hrnet-kkgvideo*) and +40% ($p = 0.005$, $r = 0.89$) with the *hrnet-kkgvideo-pretrain* configuration.

For smaller markers (namely the RRIS40 dataset), we also report the Wilcoxon sign rank test because the normality requirements were not met. In that case, without pose pretraining (*hrnet-rris40*), we observed statistically higher pixel errors on all evaluated joints when markers were inpainted ($p \leq 0.010$, $r \geq 0.82$; Fig. 6a). This led to substantial error increases, ranging from +17.7% for the elbow to +178.6% for the ankle. Adding pose pretraining (*hrnet-rris40-pretrain*) mitigated these severe degradations, though errors on inpainted images

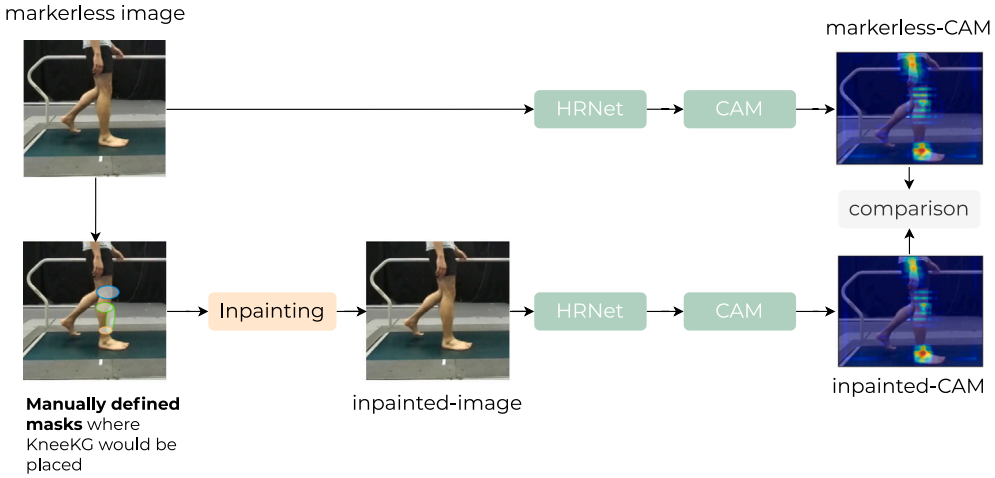


Fig. 4. Proposed process to assess whether the CAM of the *kkgvideo-inpaint-pretrain* model trained on inpainted images differs between markerless and inpainted images.

Table 2
Joint pixel error (median and interquartile range) and OKS (best in bold) on both KKGVideo and RRIS40 datasets.

	(a) KKGVideo dataset (markers and rigidbodies)				(b) RRIS40 dataset (small markers groups)			
Configuration	hrnet-kkgvideo		hrnet-kkgvideo-pretrain		hrnet-rris40		hrnet-rris40-pretrain	
Markers in test images	Visible	Inpaint	Visible	Inpaint	Visible	Inpaint	Visible	Inpaint
HIP	8.74 (8.39)	16.71 (14.78)	7.24 (7.22)	7.65 (7.24)	9.54 (9.05)	13.14 (9.73)	8.98 (7.73)	9.27 (7.76)
KNEE	3.26 (2.01)	17.27 (16.36)	2.45 (2.07)	3.42 (2.89)	2.48 (2.33)	4.21 (2.94)	3.64 (3.64)	4.12 (4.11)
ANKLE	4.01 (3.71)	4.22 (4.14)	3.36 (2.95)	3.77 (3.03)	1.49 (0.78)	3.44 (3.05)	2.06 (1.45)	2.88 (2.41)
SHOULDER	–	–	–	–	5.59 (3.76)	6.77 (4.92)	5.73 (3.23)	5.87 (3.43)
ELB	–	–	–	–	4.27 (3.16)	5.08 (4.08)	4.40 (2.71)	4.58 (2.85)
WRIST	–	–	–	–	2.50 (2.37)	9.24 (3.82)	2.56 (1.94)	3.13 (2.37)
Average	4.39 (4.78)	10.50 (15.82)	3.74 (4.16)	4.48 (4.36)	3.38 (4.51)	5.20 (5.86)	4.08 (4.16)	4.57 (4.16)
OKS	**0.943**	0.758	**0.959**	0.951	**0968**	0949	**0980**	0977

remained statistically higher for the elbows, wrists, knees, and ankles ($p < 0.05$, r ranging from 0.64 to 1.00; Fig. 6b). Notably, the error increase for the hip became non-significant with a small effect size ($p = 0.312$, $r = 0.20$). Finally, while the difference for the shoulder did not meet the strict statistical threshold ($p = 0.053$), it still exhibited a moderate-to-large effect size ($r = 0.60$).

3.1.2. Qualitative assessment of the impact of markers in training images

Establishing whether the network has learned intrinsic anatomical geometry or is merely exploiting markers as 'shortcuts' to estimate joint's center positions is fundamental to ensure the reliability of markerless motion capture in real-world settings. To investigate this, we employ class activation maps (CAM) to visualize the internal attention of the *hrnet-kkgvideo* and *hrnet-rris40* models, providing a direct assessment of the image parts used in the model's decision-making process (Fig. 7). Fig. 7(a) presents a comparison of CAM of the *hrnet-kkgvideo* model trained on KKGVideo dataset, with visible markers for the knee joint center prediction. A similar example is shown in Fig. 7(b) with the *hrnet-rris40* model for left wrist prediction. In that case, the model is supposed to look only at the desired joint, but the lack of markers makes it consider all other joints as represented by the high attention around them (Fig. 7b). Thus, regardless of whether the image contains groups of markers or reflective markers mounted to rigid bodies, the network's attention is focused on images with markers, but when they are inpainted, the activated regions are more broadly spread out and predictions (in blue) are further from the ground truth (green).

3.2. Validation of the inpainting strategy during training (H2)

Fig. 8 shows a comparison of the CAMs from the model trained on inpainted images from the KKGVideo dataset, tested on 5 subjects (three males and two females). In these examples, the CAMs are similar to both inpainted and markerless images. Similar results can be observed in the rest of the images.

4. Discussion

The primary goal of this study was to quantify the training bias induced by the visual footprint of motion capture markers in 2D pose estimation models. Specifically, we aimed to determine if the specific shapes and colors of markers provide artificial features on which networks rely instead of learning proper anatomical landmarks. Our results demonstrate that markers do indeed create a significant training bias, particularly for markers located near joint centers. Consequently, both study hypotheses were supported: the presence of markers during training adversely affected performance when models were applied to markerless images (H1), and the use of image inpainting to remove these markers proved to be an effective strategy for maintaining prediction accuracy in markerless contexts (H2). These findings provide the first direct evidence in human pose estimation that marker-based artifacts must be mitigated to ensure the clinical validity of markerless systems.

4.1. Assessment of marker impact during training (H1)

The results indicate a severe performance disparity for models trained on datasets with visible rigid bodies (such as KKGVideo). Specific position errors increased significantly on test images with inpainted markers compared to their visible-marker counterparts, with the knee joint exhibiting the most pronounced degradation (Table 2 and Fig. 5). This was expected because the femoral harness is very close to the joint center (Fig. 1a) and has a distinct blue color, providing an easy way for the model to estimate the knee. As the KneeKG is also composed of black elements, such as the tibial plate and both straps that maintain it, it creates a very distinct contrast, especially on Caucasian subjects, as in our case. Thus, even simple image processing filters like Sobel (Duda and Hart, 1973) can extract these elements from the images because they create strong gradients (examples in supplementary materials Figure A.9). Looking at the class activation map (Fig. 7a), it can be seen that when the model is trained and tested with the harness visible, the network's attention is focused on the knee, right below the rigid body. However, when this marker attachment is inpainted, the attention is more broadly spread out and spans across both knees, leading to predictions further from ground truths. This shows that the rigid body helps the network find the joint position. Adding pretraining from a database such as COCO helps because it contains many annotated images without markers, and therefore, the model has to learn to find joint center areas without relying on markers. Despite the significant reduction in the error provided by the inclusion of COCO pretraining, errors at the knee are still significantly higher on inpainted images compared to images with visible markers (Fig. 5b), indicating that even if the model was pretrained to identify joints from a very diverse dataset, the presence of markers during training still has a significant impact on it. For other joints, such as hips and ankles, small differences were found (respectively +6% and +4.7%) between marker and inpainted images because the joints are situated further from rigid bodies and are thus less impacted.

When the model was trained on the RRIS40 dataset (group of small markers), it was observed (Fig. 6) that errors in images with markers inpainted increase from +17.7% to +178.6% without pretraining (*hrnet-rris40* model; Fig. 6a) on a pose estimation dataset and from +0.9% to +51.1% when such pretraining is used (*hrnet-rris40-pretrain*; Fig. 6b). As for the KKGVideo dataset, the error differs across joints. For example, joints such as hips and shoulders produce smaller differences, respectively +0.9% and +4.8%, between both image types, than other joints (Fig. 6b). This can be explained by the fact that markers near these joints were often obscured by clothes. In those cases, the models could not always rely on features from markers, and thus had to find other ways to identify the joint centers with contextual information, reducing the gap between marker-based and inpainted images. The visualization of the network's CAMs for both inpainted and marker images (Fig. 7b) showed that some groups of markers, such as those on the wrist, help the network identify the joint center. When this group of markers is inpainted, the network fails to identify the region of the image containing the wrist. As these markers are smaller than rigid bodies, it could well

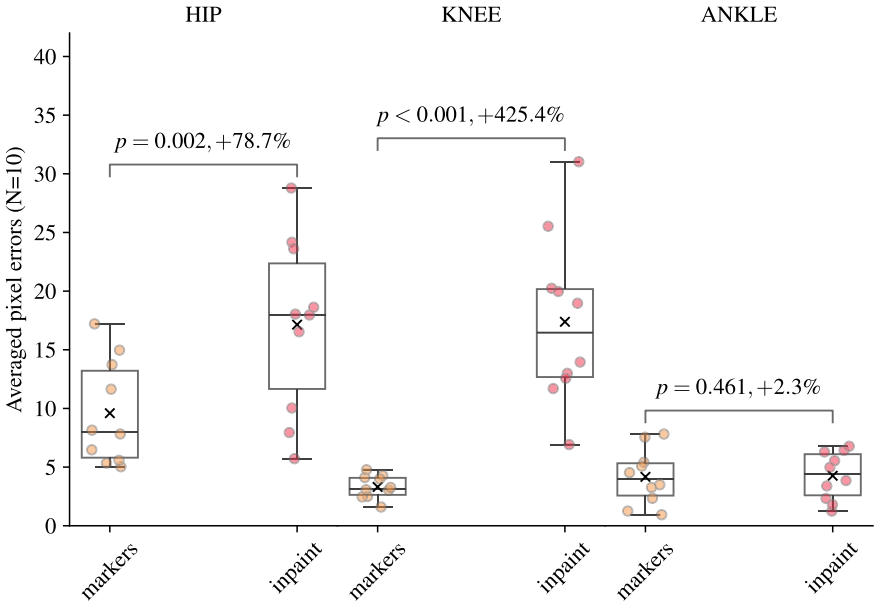
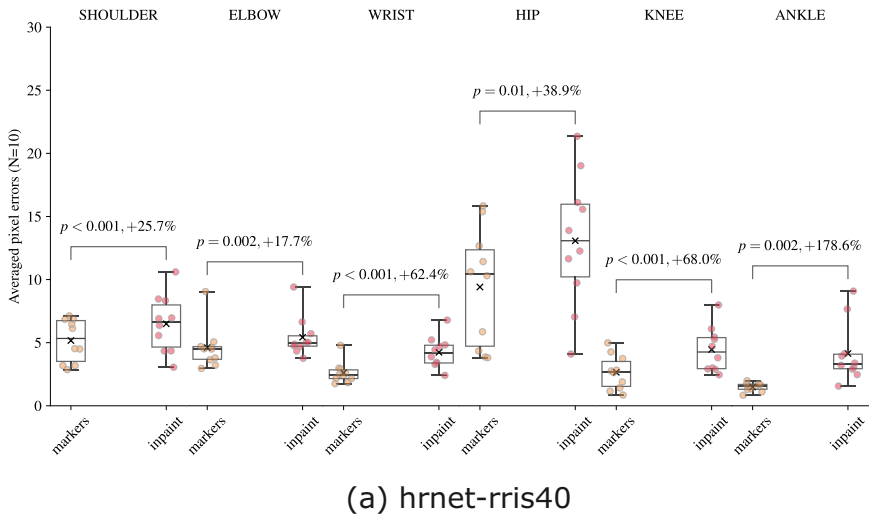


Fig. 5. This figure shows the distribution of the error of the hrnet-kkgvideo and *hrnet-kkgvideo-pretrain* models trained on images from the KKGVideo dataset with visible markers and tested on both images with (orange) and without (red) markers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

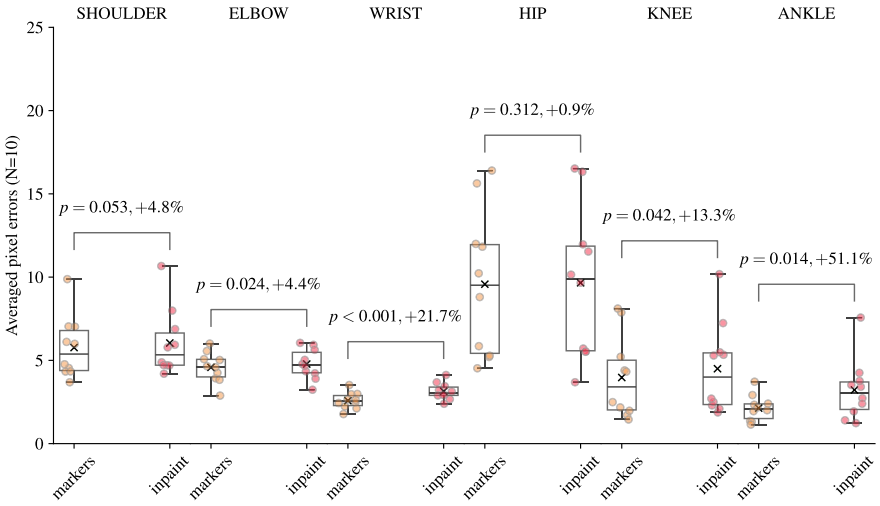
be assumed that they would not impact prediction, but this study proves the contrary.

4.2. Validation of the inpainting strategy during training (H2)

Learning directly on inpainted images could be a solution, as the model will not rely on markers to estimate the joint center position, as explored by authors like (Jatesiktat et al., 2024) and (Wu et al., 2023). One could argue that modifying the image and replacing some areas with artificial textures could bias the model, just like markers. The network's CAM illustration Fig. 8 showed little to no difference between inpainted and markerless images. Inpainted areas, such as the tibia, do not show up in CAMs from inpainted images, which means that the model trained on inpainted images does not rely on inpainted features to estimate joint centers. This suggests that, unlike images with visible markers, inpainted images can be used to train a pose estimation model to process markerless images, such as those found in clinical applications, without suffering any precision loss. We focused this specific



(a) hrnet-rris40



(b) hrnet-rris40-pretrain

Fig. 6. This figure shows the distribution of the error of the *hrnet-rris40* and *hrnet-rris40-pretrain* models trained on images from the RRIS40 dataset with visible markers and tested on both images with (orange) and without (red) markers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(a) hrnet-kkgvideo (b) hrnet-rris40

Fig. 7. Class activation map visualization for HRNet trained on KKGVideo (a) and RRIS40 (b) respectively with visible markers.

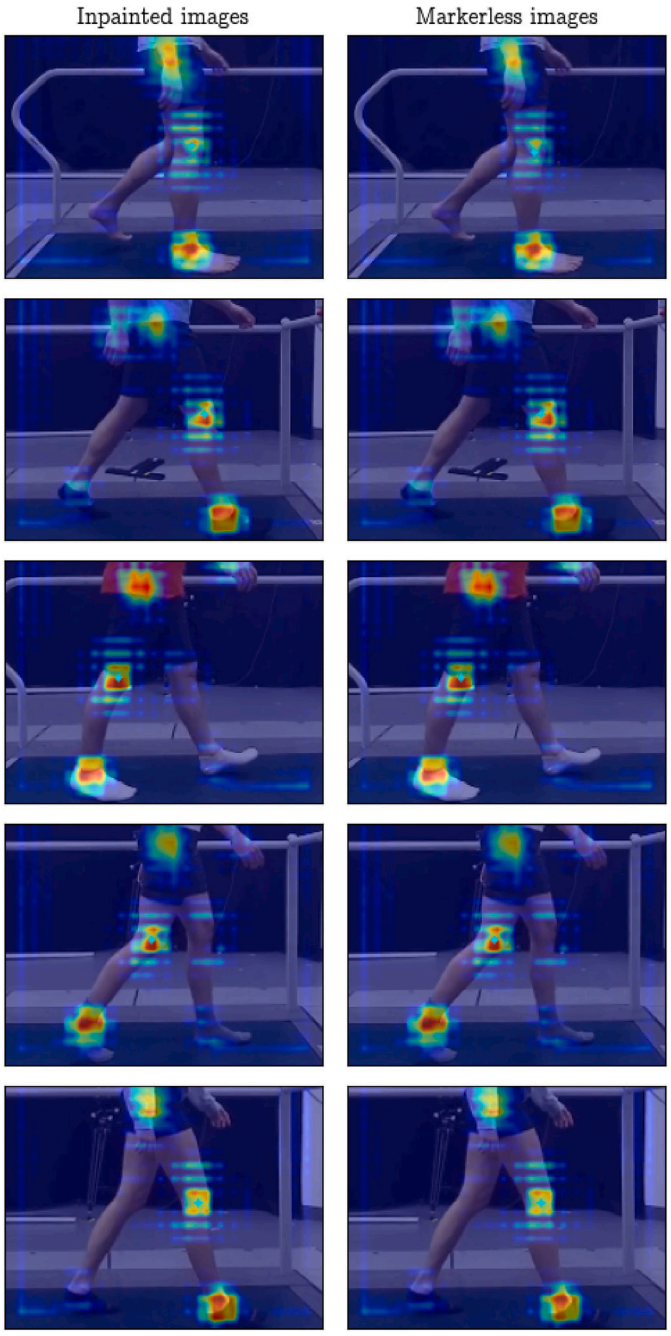


Fig. 8. Class activation map on images without markers (left) and the same image manually inpainted (right) produced by *kkgvdeo-inpaint-pretrain* model.

validation on the KKGVideo dataset because it offered the unique possibility of comparing inpainted images against markerless images from the same subjects. While we expect similar results for the smaller markers in the RRIS40 dataset, the lack of markerless reference images for that database prevented a direct quantitative validation of the inpainting strategy for those specific marker types. This also aligns with the conclusions by Rosskamp et al. (2024) which showed that predictions from pose estimation models trained on inpainted images do not degrade when applied to markerless images.

We only used HRNet (Sun et al., 2019) for the experiment herein because the vast majority of popular pose estimation models, such as OpenPose (Cao et al., 2021) and RTMPose (Jiang et al., 2023) follow a similar approach. First, a backbone is trained to extract features from images (Goldblum et al., 2025), and then a second part attempts to estimate joint center positions from these features. We showed that if markers are visible in images, then a network can extract these features and use them to estimate the pose (Table 2, Figs. 5 and 6). Because all these networks follow this *image to feature* extraction paradigm (Goldblum et al., 2025) we expect markers to have a similar impact during training on them as well. Moreover, as most 3D human pose estimation pipelines rely on the estimation of 2D keypoints from images (Neupane et al., 2024; Wang et al., 2021), the presence of markers in training images should most likely have an impact on 3D pose estimation. This could lead to the degradation of computed biomechanical parameters, as in gait analysis, for example (Sethi et al., 2022).

The present study has some limitations. Firstly, we had limited population diversity, as the KKGVideo dataset is mostly composed of Caucasian people and the RRIS40 is an Asian-centric database. Additionally, both datasets only contain healthy participants with relatively low body fat. Results were drawn from ten test subjects only, for each marker type, for both single markers with the RRIS40 dataset and rigid body-mounted markers with the KKGVideo dataset, which constitutes a small sample size. A limitation of this study is reporting spatial errors in pixels rather than millimeters. Establishing a reliable pixel-to-millimeter conversion was not feasible, as no physical calibration object such as a checkerboard was recorded in available videos. Furthermore, applying a static 2D calibration to monocular video would introduce severe perspective projection errors due to the dynamic, out-of-plane depth changes of joints during movement. Therefore, pixel error remains the most robust metric for evaluating 2D spatial degradation.

The study shows that, as some authors hypothesized (Jatesiktat et al., 2024; Hampali et al., 2020; Wu et al., 2023), markers in training images could have an impact on pose estimation models and lead to worse predictions on images without markers, which is typically the case in a clinical environment. To date, there is still no consensus on the usability of markerless systems for clinical exams because of prediction accuracy concerns and high variability. Further prediction degradation is thus to be avoided at all costs. We also showed that the use of inpainted images to train models allows maintaining high performance on markerless images while benefiting from highly accurate annotations from MoCap systems.

CRediT authorship contribution statement

Lucas Mercier: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Thierry Cresson:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Neila Mezghani:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition. **Sylvie Gervais:** Writing – review & editing, Formal analysis. **Carlos Vázquez:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Ethics statement

This research presents an accurate account of the work performed; all data presented are accurate and methodologies are detailed enough to permit others to replicate the work.

This manuscript represents entirely original work, and if work and/or words of others have been used, they have been appropriately cited or quoted and permission has been obtained where necessary.

This material has not been published in whole or in part elsewhere.

The manuscript is not currently being considered for publication in another journal.

That generative AI and AI-assisted technologies have not been utilized in the writing process or if used, the use of AI and AI-assisted technologies has been disclosed in the manuscript, and a statement will appear in the published work.

That generative AI and AI-assisted technologies have not been used to create or alter images unless specifically used as part of the research

design where such use must be described in a reproducible manner in the methods section.

All authors have been personally and actively involved in substantive work leading to the manuscript and will hold themselves jointly and individually responsible for its content.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Lucas Mercier reports that financial support was provided by Natural Sciences and Engineering Research Council of Canada. Lucas Mercier also reports that financial support as well as equipment, drugs, or supplies were provided by Emovi. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported by the NSERC grant RGPIN-2021-04293 and ALLRP 565212 – 21. We also would like to thank Bianca Marois for her help during data collection.

Appendix A. Supplementary data

Supplementary data for this article can be found online at doi:10.1016/j.jbiomech.2026.113341.

References

Avogaro, A., Cunico, F., Rosenhahn, B., Setti, F., 2023. Markerless human pose estimation for biomedical applications: a survey. Front. Comput. Sci. 5, https://doi.org/10.3389/fcomp.2023.1153160

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305. https://doi.org/10.5555/2188385.2188395

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., Sheikh, Y., 2021. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. 43, 172–186. https://doi.org/10.1109/TPAMI.2019.2929257

Colyer, S.L., Evans, M., Cosker, D.P., Salo, A.I.T., 2018. 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 4, 24. https://doi.org/10.1186/s40798-018-0139-y

Cronin, N.J., Rantalainen, T., Ahtiainen, J.P., Hynynen, E., Waller, B., 2019. Markerless 2D kinematic analysis of underwater running: a deep learning approach. J. Biomech. 87, 75–82. https://doi.org/10.1016/j.jbiomech.2019.02.021

D'Antonio, E., Taborri, J., Palermo, E., Rossi, S., Patanè, F., 2020. A markerless system for gait analysis based on OpenPose library. In: 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), pp. 1–6, https://doi.org/10.1109/I2MTC43012.2020.9128918 ISSN: 2642-2077.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848 ISSN: 1063-6919.

Duda, R.O., Hart, P.E., 1973. Pattern Classification and Scene Analysis: Theory and Practice. Wiley.

Dumas, R., Wojtusch, J., 2018. Estimation of the body segment inertial parameters for the rigid body biomechanical models used in motion analysis. In: Handbook of Human Motion. Springer, Cham, pp. 47–77. https://doi.org/10.1007/978-3-319-14418-4147

Goldblum, M., Souri, H., Ni, R., Shu, M., Prabhu, V., Somepalli, G., Chattopadhyay, P., Ibrahim, M., Bardes, A., Hoffman, J., Chellappa, R., Wilson, A.G., Goldstein, T., 2025. Battle of the backbones: a large-scale comparison of pretrained models across computer vision tasks. In: Proceedings of the 37th International Conference on Neural Information Processing Systems NIPS '23. Curran Associates Inc., pp. 29343–29371.

Gozlan, Y., Falisse, A., Uhlrich, S., Gatti, A., Black, M., Hicks, J., Delp, S., Chaudhari, A., 2025. OpenCapBench: a benchmark to bridge pose estimation and biomechanics. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 4056–4065, https://doi.org/10.1109/WACV61041.2025.00399 ISSN: 2642-9381.

Guo, L., Chang, R., Wang, J., Narayanan, A., Qian, P., Leong, M.C., Kundu, P.P., Senthilkumar, S., Garlapati, S.C., Yong, E.C.K., Pahwa, R.S., 2025. Artificial intelligence-enhanced 3D gait analysis with a single consumer-grade camera. J. Biomech. 187, 112738. https://doi.org/10.1016/j.jbiomech.2025.112738

Hagemeister, N., Parent, G., Van de Putte, M., St-Onge, N., Duval, N., de Guise, J., 2005. A reproducible method for studying three-dimensional knee kinematics. J. Biomech. 38, 1926–1931. https://doi.org/10.1016/j.jbiomech.2005.05.013

Hampali, S., Rad, M., Oberweger, M., Lepetit, V., 2020. HOnnotate: a method for 3D annotation of hand and object poses. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3193–3203, https://doi.org/10.1109/CVPR42600.2020.00326

Jatesiktat, P., Lim, G.M., Lim, W.S., Ang, W.T., 2024. Anatomical-marker-driven 3D markerless human motion capture. IEEE J. Biomed. Health Inform. 1–14. https://doi.org/10.1109/JBHI.2024.3424869

Jiang, T., Peng, P., Liao, R., Zhang, W., Yu, J., Bin, Z., Pan, S., Pang, G., Shao, W., Liu, Z., et al., 2023. Rtmpose: Real-time multi-person pose estimation based on mmpose. arXiv preprint arXiv:2303.07399.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive growing of GANS for improved quality, stability, and variation. In: International Conference on Learning Representations, Paper 447, pp. 1–26.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence – Volume 2 IJCAI'95. Morgan Kaufmann Publishers Inc., pp. 1137–1143.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Liang, P., Kwong, W.H., Sidarta, A., Yap, C.K., Tan, W.K., Lim, L.S., Chan, P.Y., Kuah, C.W.K., Wee, S.K., Chua, K., Quek, C., Ang, W.T., 2020. An asian-centric human movement database capturing activities of daily living. Sci. Data 7, https://doi.org/10.1038/s41597-020-00627-7

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014. Springer International Publishing, Cham, pp. 740–755.

Lustig, S., Magnussen, R.A., Cheze, L., Neyret, P., 2012. The KneeKG system: a review of the literature. Knee surg. sports traumatol. arthrosc.: off. j. ESSKA 20, 633–638. https://doi.org/10.1007/s00167-011-1867-4

Needham, L., Evans, M., Cosker, D.P., Colyer, S.L., 2021. Can markerless pose estimation algorithms estimate 3D mass centre positions and velocities during linear sprinting activities? Sensors 21, 2889. https://doi.org/10.3390/s21082889

Neupane, R.B., Li, K., Boka, T.F., 2024. A survey on deep 3D human pose estimation. Artif. Intell. Rev. 58, 24. https://doi.org/10.1007/s10462-024-11019-3

Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K., 2017. Towards accurate multi-person pose estimation in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Honolulu, HI, pp. 3711–3719, https://doi.org/10.1109/CVPR.2017.395

Peng, J., Zhou, Y., Mok, P.Y., 2024. KTPFormer: kinematics and trajectory prior knowledge-enhanced transformer for 3D human pose estimation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1123–1132, https://doi.org/10.1109/CVPR52733.2024.00113

Rosskamp, J., Weller, R., Zachmann, G., 2024. Effects of markers in training datasets on the accuracy of 6D pose estimation. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 4445–4454, https://doi.org/10.1109/WACV57701.2024.00440

Sarandi, I., Hermans, A., Leibe, B., 2023. Learning 3D human pose estimation from dozens of datasets using a Geometry-Aware autoencoder to bridge between skeleton formats. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, Waikoloa, HI, USA, pp. 2955–2965, https://doi.org/10.1109/WACV56688.2023.00297

Seethapathi, N., Wang, S., Saluja, R., Blohm, G., Kording, K.P., 2019. Movement science needs different pose tracking algorithms. arXiv:1907.10226.

Sethi, D., Bharti, S., Prakash, C., 2022. A comprehensive survey on gait analysis: history, parameters, approaches, pose estimation, and future work. Artif. Intell. Med. 129, 102314. https://doi.org/10.1016/j.artmed.2022.102314

Stenum, J., Rossi, C., Roemmich, R.T., 2021. Two-dimensional video-based analysis of human gait using pose estimation. PLOS Comput. Biol. 17, e1008935. https://doi.org/10.1371/journal.pcbi.1008935

Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep High-Resolution representation learning for human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Long Beach, CA, USA, pp. 5686–5696, https://doi.org/10.1109/CVPR.2019.00584

Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V., 2022. Resolution-robust large mask inpainting with Fourier convolutions. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, Waikoloa, HI, USA, pp. 3172–3182, https://doi.org/10.1109/WACV51458.2022.00323

Uhlrich, S.D., Falisse, A., Kidziński, Ł., Muccini, J., Ko, M., Chaudhari, A.S., Hicks, J.L., Delp, S.L., 2023. OpenCap: human movement dynamics from smartphone videos. PLOS Comput. Biol. 19, e1011462. https://doi.org/10.1371/journal.pcbi.1011462. Publisher: Public Library of Science.

Vafadar, S., Skalli, W., Bonnet-Lebrun, A., Khalifé, M., Renaudin, M., Hamza, A., Gajny, L., 2021. A novel dataset and deep learning-based approach for marker-less motion capture during gait. Gait Posture 86, 70–76. https://doi.org/10.1016/j.gaitpost.2021.03.003

Wade, L., Needham, L., McGuigan, P., Bilzon, J., 2022. Applications and limitations of current markerless motion capture methods for clinical gait biomechanics. PeerJ 10, e12995. https://doi.org/10.7717/peerj.12995

Wang, J., Tan, S., Zhen, X., Xu, S., Zheng, F., He, Z., Shao, L., 2021. Deep 3D human pose estimation: a review. Comput. Vis. Image Underst. 210, 103225. https://doi.org/10.1016/j.cviu.2021.103225

Winter, D.A., 2009. Three-dimensional kinematics and kinetics. In: Biomechanics and Motor Control of Human Movement. John Wiley & Sons, Ltd, pp. 176–199. https://doi.org/10.1002/9780470549148.ch7

Wu, E., Nishioka, H., Furuya, S., Koike, H., 2023. Marker-removal networks to collect precise 3D hand data for RGB-based estimation and its application in piano. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, pp. 2976–2985, https://doi.org/10.1109/WACV56688.2023.00299

Xiang, H., Zou, Q., Nawaz, M.A., Huang, X., Zhang, F., Yu, H., 2023. Deep learning for image inpainting: a survey. Pattern Recognit. 134, 109046. https://doi.org/10.1016/j.patcog.2022.109046

Xu, Y., Zhang, J., Zhang, Q., Tao, D., 2022. ViTPose: simple vision transformer baselines for human pose estimation. In: Proceedings of the 36th International Conference on Neural Information Processing Systems NIPS '22. Curran Associates Inc., Red Hook, NY, USA, pp. 38571–38584.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014. Springer International Publishing, pp. 818–833, https://doi.org/10.1007/978-3-319-10590-1-53

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., Shah, M., 2023. Deep learning-based human pose estimation: a survey. ACM Comput. Surv. 56, :11:1–:11:37. https://doi.org/10.1145/3603618

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, NV, USA, pp. 2921–2929, https://doi.org/10.1109/CVPR.2016.319

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2018. Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40, 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009