Towards Forensic Height Estimation: Generating Digital Humans in CCTV Environments

A. Dörsch[1], C. Busch[1], C. Rathgeb[1]

[1]da/sec – Biometrics and Security Research Group, Hochschule Darmstadt, Germany

Abstract. Height estimation from surveillance footage is a common procedure in forensic identification. However, due to capture- and subject-related distortions, estimating human height in unconstrained environments remains error-prone. Although existing methods report promising performance, they are typically evaluated on datasets that do not reflect real-world surveillance conditions. In this work, we introduce HDA-Digital Humans in CCTV Environments (HDA-DH-CCTV), a synthetic dataset consisting of digital humans rendered across typical surveillance scenes. HDA-DH-CCTV comprises 2,250 images including variations in camera viewpoints, subject body-types and poses, as well as camera-to-subject distances. HDA-DH-CCTV is designed to support height estimation in surveillance settings and provides detailed annotations such as ground-truth height, camera parameters and reference object information. We conducted a series of experiments to evaluate the utility of synthetic data for height estimation in forensic settings. Our results on synthetic CCTV data show that incorporating estimated metric depth and reference object information significantly improves performance, achieving a mean absolute error (MAE) of up to 5 cm for distant subjects. These findings highlight the potential of synthetic data for scenarios in which real-world annotated surveillance data is difficult to obtain.

Keywords: Synthetic Surveillance Data · Digital Humans · Forensic Height Estimation

1 Introduction

Estimating the height of a subject from footage captured by closed-circuit television (CCTV) cameras is a common practice employed in forensic identification [11]. Such measurements are used to verify witness statements and to narrow down the list of potential suspects. However, accurately estimating a subject's height from in-the-wild surveillance images remains error-prone and has contributed to some wrongful convictions in the past [17].

Despite its apparent simplicity, estimating a subject's height from CCTV footage is subject to several sources of measurement errors. These can broadly be categorized into human-related and capture-related causes. Human-related measurement errors arise from variations in body posture and gait, as well as from

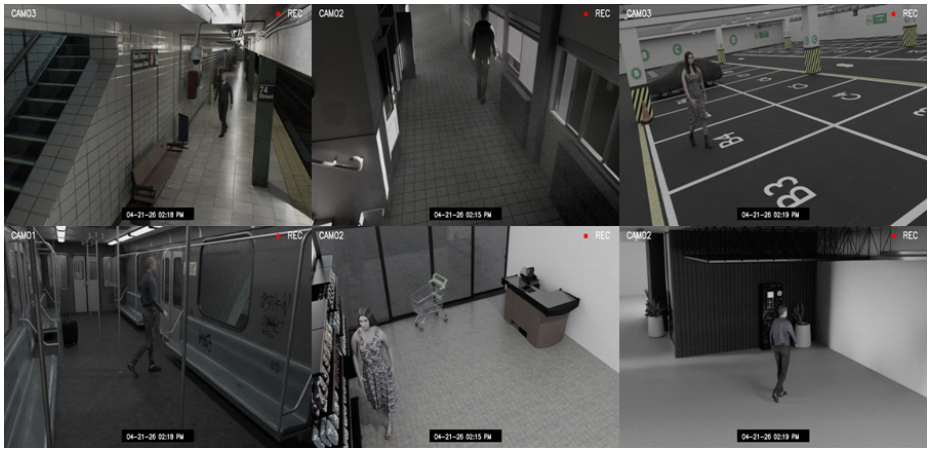

Fig. 1. Examples of rendered CCTV scenes from our proposed HDA-DH-CCTV dataset. Each scene contains a synthetic subject captured from the perspective of a CCTV-like camera in a typical surveillance environment.

clothing such as shoes or head coverings, which introduce additional height uncertainty [17]. Capture-related measurement errors arise from the projection of a recorded real-world scene onto a two-dimensional image [17], including variations in camera parameters [7], camera viewpoints and scale and perspective distortions.

Although recent height estimation models report strong performance under controlled conditions, their evaluation is often conducted on datasets that do not reflect typical CCTV scenarios. These include, for example, web-scraped datasets, which typically consist of celebrities (as their height is publicly known) captured during red-carpet events [3] or constrained in-house datasets with limited subject diversity [23]. Height estimation evaluations conducted using such datasets often exhibit limited variation in camera viewpoints, body shapes and height distributions, camera-to-subject distances and scene geometry. However, obtaining real-world CCTV footage is challenging, as it requires both (1) a captured CCTV scene and (2) ground-truth height annotations for the subject, which is rarely available in real-world settings.

In this work, we present the HDA-Digital Humans in CCTV Environments (HDA-DH-CCTV) dataset, designed to overcome the limitations of existing datasets and to investigate the potential of synthetic data in forensic CCTV scenarios. HDA-DH-CCTV comprises 2,250 images of digital humans rendered in typical surveillance-like environments (e.g., train stations, grocery stores, car parks and public buildings) from various CCTV-like perspectives as illustrated in Figure 1. Human subjects are generated using randomized body parameters (gender, height and body shape) and are combined with a variety of clothing assets to increase data diversity. Digital humans are depicted at varying

camera-to-subject distances, different walking poses and orientations, thereby introducing variations in posture and apparent height. Each rendered image is annotated with detailed ground-truth information such as height, camera parameters, reference object information, bounding-boxes and estimated metric depth information [22].

We conducted several experiments to investigate the impact of various cues on height estimation in CCTV scenarios. Our results show that relying solely on pixel-based subject measurements is highly unreliable. Incorporating both (1) estimated metric depths and (2) height information on reference objects (when available) significantly improves accuracy, highlighting the need for auxiliary cues in unconstrained surveillance scenarios.

In summary, this work makes the following contributions:

- We introduce HDA-Digital Humans in CCTV Environments (HDA-DH-CCTV), a synthetic dataset consisting of digital humans placed in CCTV-like environments across different camera viewpoints
- We provide detailed annotations, including ground-truth height, camera parameters, reference object information, bounding boxes and estimated metric depth
- We evaluate the impact of different cues on height estimation in rendered CCTV settings, demonstrating that combining estimated metric depth and reference object information significantly improves accuracy
- We demonstrate the utility of synthetic CCTV-like data for forensic height estimation research and highlight the importance of additional cues in unconstrained settings.

2 Related Work

2.1 Digital Humans in Computer Vision

The generation of synthetic data using digital humans and parametric 3D models has proven beneficial across a wide range of human-related computer vision tasks [21], [18], [16], [9], [20], [8], [1]. While 3D scene modelling provides full control over scene geometry and ground-truth metric annotations, verifying the accuracy of such attributes in prompt-generated outputs [13] remains challenging.

Particularly in security-related contexts, such as forensic CCTV analysis, reliable human-related annotations (e.g., ground-truth height measurements) are essential, instead of relying on the assumed metric correctness of text-to-image generation. However, despite the success of 3D rendering pipelines in complementing real-world training data, the use of digital humans for forensic height estimation in unconstrained surveillance settings remains largely unexplored.

2.2 Human Height Estimation

Existing datasets used to estimate human body height can be broadly divided into the following categories: (1) web-scraped datasets, (2) constrained datasets and (3) surveillance-oriented datasets. Datasets sourced from the web, such as IMDB-based collections [14] or Celeb-FBI [3], consist predominantly of celebrities, as their body heights are publicly known. However, these datasets are either limited in terms of environmental settings (e.g., red carpet events) or exhibiting a limited variety of body-types introduced by the celebrity population. Additional web-collected datasets, such as the 2-D image-to-BMI dataset [5] have been leveraged in recent height estimation work, such as [15]. However, such datasets do not reflect typical in-the-wild surveillance scenarios and therefore do not exhibit CCTV footage characteristics.

In addition, many height estimation methods are trained or evaluated by their own experimental datasets. However, these datasets are constrained by either (1) limited subject diversity [4], (2) are being recorded in controlled environments [2], or (3) their estimation methods are reliant on additional capture devices or sensors [19], [23].

While authentic surveillance footage is often restricted to forensic or law enforcement authorities, recent work by Ratthi et al. [10] introduced a surveillance-oriented dataset for human height estimation. However, such datasets remain limited in subject diversity and are additionally constrained by privacy, ethical and data-protection requirements. To address these limitations, we introduce HDA-DH-CCTV, a synthetic dataset for estimating body height in unconstrained CCTV environments. HDA-DH-CCTV combines realistic surveillance conditions with increased subject diversity while being compliant with data protection requirements through the use of synthetic data.

3 Dataset

HDA-DH-CCTV consists of 2,250 images of digital humans placed in diverse indoor and outdoor surveillance environments. All scenes were rendered using Blender (version 5.0.1) [1] and the Cycles rendering engine [2] at a resolution of 1280×720 pixels. Each image is captured from a CCTV-like perspective and is annotated with detailed ground-truth annotations. HDA-DH-CCTV introduces variations in camera viewpoints, camera-to-subject distances, human poses and body-types across typical surveillance settings. The dataset is designed to support height estimation in forensic settings and to investigate the effectiveness of synthetic data for real-world surveillance scenarios. A detailed summary of dataset statistics can be found in Table 1.

[1] `https://www.blender.org/`
[2] `https://www.cycles-renderer.org/`

Table 1. Summary statistics of the HDA-DH-CCTV dataset

Category	Property	Value
Dataset	Total Images	2,250
	Image Resolution	1280×720
	Number of scenes	9
	Images per scene	250
	Subjects per image	1
	Annotated reference objects per scene	1
Camera Configuration	Total Camera Configurations	27
	Cameras per scene	3
	Focal length range (mm)	19 – 59
	Sensor width range (mm)	28 – 40
Digital Humans	Male subjects	1,110
	Female subjects	1,140
	Unique poses/gait	8
	Body height range (m)	1.42 – 2.18
	Mean ± std height (m)	1.76 ± 0.18

3.1 Digital Human Generation

All subjects in HDA-DH-CCTV were generated using the open-source MPFB2[3] human character framework for Blender. For each rendered image, we used the parametric MPFB2 human body model and randomized body parameters such as gender, height, weight, muscle mass and body-type proportions within predefined ranges.

Each digital human was additionally assigned a randomized combination of assets from MPFB2 and user-contributed assets licensed under Creative Common Zero (CCO). These assets include hair, eye colour, eyebrows, eyelashes, skin texture, clothing, shoes and optional accessories. Each subject was assigned a random pose sampled from a set of eight handcrafted walking poses to introduce variations in pose and gait, as illustrated in Figure 2.

3.2 CCTV Scenes and Camera Setup

HDA-DH-CCTV consists of nine distinct surveillance environments, with each scene containing three camera configurations. For each scene, the cameras were manually positioned at elevated locations and angles to simulate typical CCTV perspectives. Scenes were either (1) manually created using publicly available 3D assets from BlenderKit[4] or (2) adapted from publicly available BlenderKit scenes. In each scene, we manually inserted and annotated a reference object (pixel height, real-world height, bounding box information, estimated metric

[3] `https://github.com/makehumancommunity/mpfb2`
[4] `https://www.blenderkit.com/`

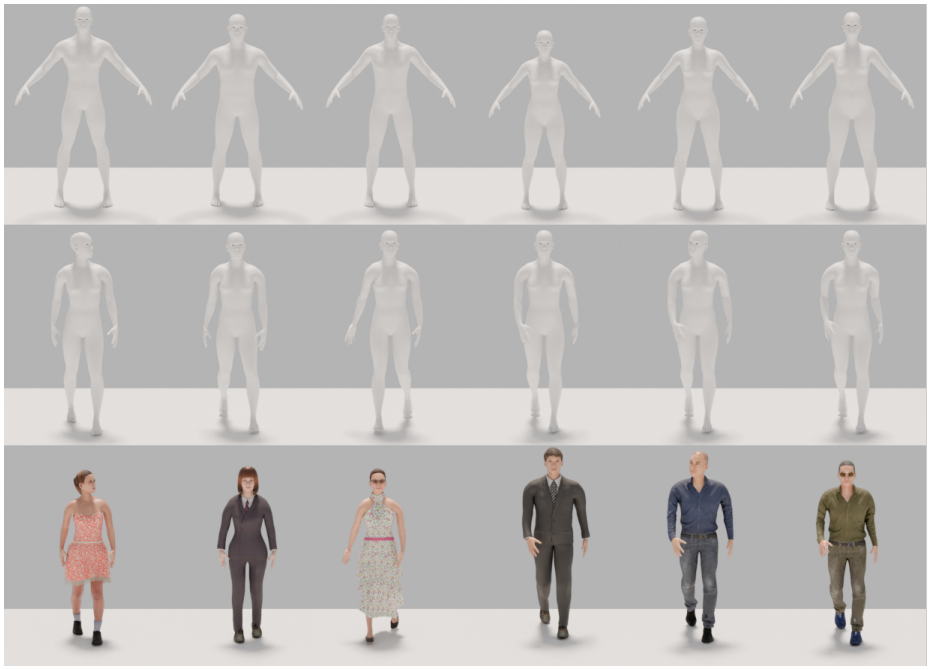

Fig. 2. Example of body variations in the generated subjects using the MPFB2 model. The upper row illustrates various body-type variations. The middle row demonstrates different walking poses applied to a fixed body mesh. The bottom row illustrates fully randomized combinations of body-type variations, applied poses and clothing assets.

depth), which can be used for downstream height estimation. For each rendered image, the subject is placed at random positions within a manually positioned walkable area to ensure subject visibility and to avoid placement collisions within the scene. Each rendered image contains exactly one randomized subject. Additionally, the orientation (yaw rotation), body parameters, pose and clothing of the subject were randomized for each rendered image. To simulate real CCTV camera effects, we post-processed the rendered images by applying varying degrees of desaturation, Gaussian noise and typical scanline effects. Additionally, we added visual information such as the underlying camera identifier along with a recording indicator and a timestamp. Figure 3 illustrates an example scene configuration captured from three different camera viewpoints.

4 Experiments

4.1 Depth Estimation

Recent work has demonstrated that depth information can be a valuable auxiliary cue for estimating human body height [6], [23]. However, surveillance footage

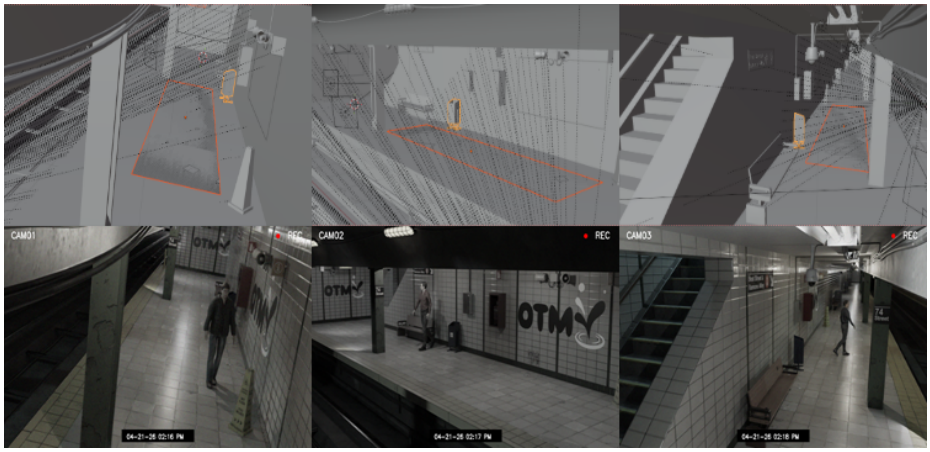

Fig. 3. Example of a surveillance environment captured from the three camera viewpoints. The top row shows the scene with basic surface shading, highlighting the walkable area (red) and the scene-specific reference object (trash can in yellow). The bottom row shows the corresponding rendered images post-processed with CCTV-style effects for each camera.

Algorithm 1 Extraction of depth map estimation in HDA-DH-CCTV

for each image I in HDA-DH-CCTV **do**
    Load image I and corresponding metadata M
    Estimate depth map $D \leftarrow \text{DepthAnything}(I)$
    Extract subject bounding box B_s from M
    Extract reference bounding box B_r from M
    Crop subject region $D_s \leftarrow D[B_s]$
    Crop reference region $D_r \leftarrow D[B_r]$
    Compute median depth $d_s \leftarrow \text{median}(D_s)$
    Compute median depth $d_r \leftarrow \text{median}(D_r)$
end for

is typically limited to RGB images rather than containing explicit depth measurements. To address this limitation, we estimate scene depth using Depth-Anything-v2 [22]. Specifically, we employ a metric depth estimation model pretrained on HyperSim[5] [12]. For each image in the HDA-DH-CCTV dataset, we estimated depth values for both the subject and the reference object, as described in Algorithm 1.

We extracted the median depth of the cropped bounding-box region instead of the mean depth to reduce outlier sensitivity. A visualization of Algorithm 1 is shown in Figure 4.

[5] `https://github.com/DepthAnything/Depth-Anything-V2/tree/main/metric_depth`

Fig. 4. Visualization of the depth estimation extraction. The left image shows a rendered RGB image from the HDA-DH-CCTV dataset with the reference object (yellow) and subject (red) bounding boxes. The corresponding image (right) visualizes the depth map (colour encodes relative scene depth) used to compute the median depth values based on the bounding box areas.

4.2 Height Estimation Using Synthetic Surveillance Data

In this work, the goal is to predict an individual's height solely from features derived from visual cues. This reflects real-world CCTV scenarios, where camera parameters or additional sensor information are typically unavailable. We evaluate the utility of synthetic data for height estimation and investigate how different visual cues derived from HDA-DH-CCTV contribute to height estimation in unconstrained surveillance settings.

To this end, we performed a randomized 80/20 train-test split on HDA-DH-CCTV. For all experiments, we trained a Random Forest regressor and evaluated three feature configurations:

- **Model 1** (Baseline): Subject pixel height (extracted from the bounding box)
- **Model 2**: Subject pixel height and estimated metric depth (see Sec. 4.1)
- **Model 3**: Subject and reference object pixel heights, estimated metric depth and known real-world height of the reference object

While configurations (1) and (2) are directly applicable to operational scenarios, configuration (3) assumes a reference object (with known height) within the captured CCTV scene. In practice, such measurements could be approximated using standardized objects (e.g. doors) if available.

In addition, we investigated the impact of different camera-to-subject distances on height estimation performance. To this end, we divided the test set into three distance ranges: near ($2.75 - 7.22$ m), mid ($7.22 - 10.61$ m) and far ($10.61 - 16.05$ m) using quantile-based splitting[6] based on precomputed camera-to-subject ground distances[7]. Height estimation performance is evaluated using the mean absolute error (MAE) in meters and summarized in Table 2:

[6] `https://pandas.pydata.org/docs/reference/api/pandas.qcut.html`

[7] Camera-to-subject ground distance is computed as the Euclidean distance in the ground plane between the camera and subject (x, y) coordinates.

Table 2. Model variations and underlying height estimation performance

Model	Near MAE (m) ↓	Mid MAE (m) ↓	Far MAE (m) ↓	Total MAE (m) ↓
1	0.15	0.14	0.13	0.14
2	0.12	0.12	0.11	0.12
3	0.07	0.07	0.05	0.07

The results in Table 2 show that height estimation performance remains relatively stable across different camera-to-subject distances for all evaluated models. A slight improvement in height estimation accuracy can be observed for subjects farther from the camera than subjects located closer to the camera. When relying solely on subject-specific features, the accuracy of the height estimation remains limited: When using only pixel-based subject information, the model achieves a total MAE of 0.14 m. Incorporating estimated subject metric depth, the MAE is slightly reduced to 0.12 m.

Additionally, incorporating reference object information significantly improves performance (MAE reduction by 50% compared to Model 1), resulting in a total MAE of 7 cm. The best performance can be observed for subjects at a far distance, where Model 3 achieves an MAE of 5 cm. These results highlight the importance of additional cues for reliable height estimation in surveillance settings, where scene context and perspective variation are more challenging than under controlled settings.

5 Limitations

The main contribution of this work is the introduction of the HDA-DH-CCTV dataset. Although HDA-DH-CCTV serves as a synthetic benchmark for unconstrained (forensic) height estimation, a domain gap between synthetic data and authentic CCTV footage cannot be excluded. However, the evaluation on authentic CCTV footage from operational forensic environments remains a major challenge, as this would require (1) access to suitable real imagery that may require collaboration with law enforcement agencies or forensic/police authorities, as well as (2) verified ground-truth height annotation of the suspect captured in the CCTV recording. The experiments conducted in Section 4 are therefore intended to demonstrate the potential of synthetic CCTV data and highlight the importance of auxiliary cues such as depth information and reference objects in unconstrained settings.

6 Conclusion

In this work, we introduced HDA-DH-CCTV, a dataset of digital humans rendered across typical surveillance scenes captured from CCTV-like cameras. HDA-DH-CCTV is designed to address the limitations of existing height estimation

datasets and to investigate the utility of synthetic data for height estimation in unconstrained forensic settings. Our dataset introduces variations in camera viewpoints, camera-to-subject distances, human body-types, clothing and poses, simulating real-world surveillance conditions across various scenes.

We conducted several experiments demonstrating that auxiliary cues such as metric depth and reference object information significantly improve height estimation performance in unconstrained settings. Although synthetic data provides detailed ground-truth annotations and can simulate operational scenarios, evaluating height estimation on real CCTV footage remains challenging due to the limited public availability of data. Overall, this work highlights the potential of synthetic data for forensic height estimation, particularly in scenarios where real-world data is scarce or unavailable. Although access to annotated real-world CCTV footage remains challenging, HDA-DH-CCTV provides a synthetic alternative for research on height estimation in unconstrained CCTV-like environments.

7 Acknowledgment

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

We thank the Blender community and the developers of MPFB2 for providing open-source 3D assets and digital human resources that supported the creation of the HDA-DH-CCTV dataset.

References

1. Bazavan, E.G., Zanfir, A., Zanfir, M., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: HSPACE: Synthetic Parametric Humans Animated in Complex Environments. arXiv (2022)
2. BenAbdelkader, C., Yacoob, Y.: Statistical body height estimation from a single image. In: 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. pp. 1–7 (2008)
3. Debnath, P., Rifa, U.A., Rafa, B.K., Akib, A.H.T., Rahman, M.A.: Celeb-FBI: A Benchmark Dataset on Human Full Body Images and Age, Gender, Height and Weight Estimation using Deep Learning Approach. arXiv (2024)
4. I. G. Andika Yudantara, R.W.C., Rohmawati, A.A.: Reference-Based Human Height Estimation. In: 2024 11th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI). pp. 571–578 (2024)
5. Jin, Z., Huang, J., Wang, W., Xiong, A., Tan, X.: Estimating human weight from a single image. IEEE Transactions on Multimedia **25**, 2515–2527 (2023)
6. Lee, D.S., Kim, J.S., Jeong, S.C., Kwon, S.K.: Human Height Estimation by Color Deep Learning and Depth 3D Conversion. Applied Sciences **10**(16) (2020)
7. Olver, A.M., Guryn, H., Liscio, E.: The effects of camera resolution and distance on suspect height analysis using PhotoModeler. Forensic Science International **318**, 110601 (2021)

8. Patel, P., Black, M.J.: CameraHMR: Aligning people with perspective. In: International Conference on 3D Vision (3DV) (2025)
9. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D., Tripathi, S., Black, M.: AGORA: Avatars in geography optimized for regression analysis. In: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (Jun 2021)
10. Ratthi, K.I., Yogameena, B., Perumaal, S.S.: Human height estimation using AI-assisted computer vision for intelligent video surveillance system. Measurement **236**, 115133 (2024)
11. Richter, S., Labudde, D.: Body Measurements for Digital Forensic Comparison of Individuals—An Overview of Current Developments. Applied Sciences **15**(23) (2025)
12. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: International Conference on Computer Vision (ICCV) 2021 (2021)
13. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10674–10685 (2021)
14. S. Günel, H.R., Fua, P.: What face and body shapes can tell us about height. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 1819–1827 (2019)
15. Sakina, M., Muhammad, I., Abdullahi, S.S.: A multi-factor approach for height estimation of an individual using 2d image. Procedia Computer Science **231**, 765–770 (2024), 14th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 13th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (EUSPN/ICTH 2023)
16. Tesch, J., Becherini, G., Achar, P., Yiannakidis, A., Kocabas, M., Patel, P., Black, M.J.: BEDLAM2.0: Synthetic humans and cameras in motion. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
17. Thakkar, N., Farid, H.: On the Feasibility of 3D Model-Based Forensic Height and Weight Estimation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 953–961 (2021)
18. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning From Synthetic Humans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
19. Velesaca, H.O., Vulgarin, J., Vintimilla, B.X.: Deep learning-based human height estimation from a stereo vision system. In: 2023 IEEE 13th International Conference on Pattern Recognition Systems (ICPRS). pp. 1–7 (2023)
20. Wang, S., Li, J., Li, T., Yuan, Y., Fuchs, H., Nagano, K., Mello, S.D., Stengel, M.: BLADE: Single-view Body Mesh Estimation through Accurate Depth Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
21. Wood, E., Baltrušaitis, T., Hewitt, C., Dziadzio, S., Cashman, T.J., Shotton, J.: Fake It Till You Make It: Face Analysis in the Wild Using Synthetic Data Alone. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3681–3691 (2021)
22. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: Advances in Neural Information Processing Systems. vol. 37, pp. 21875–21911. Curran Associates, Inc. (2024)

23. Yin, F., Zhou, S.: Accurate Estimation of Body Height From a Single Depth Image via a Four-Stage Developing Network. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8264–8273 (2020)