A Lightweight Framework for Human Height Estimation Using RGB Images and Planar Geometry

Sojeong Shin [*1] **ByungKon Kang**[† 1]

Abstract

This paper presents a monocular camera-based system for accurate height estimation. By combining precise camera calibration, semantic segmentation, and ray-plane intersection, the system enables contactless fitness assessments without depth sensors or manual intervention. Experiments show that height estimation is highly sensitive to the viewing angle, with minimized error occurring at optimal camera placements.

1. Introduction

Human height is a fundamental biometric implications across fitness, medicine, commerce, and public safety. Accurate height measurement supports personalized health assessments, clothing size recommendations, and even criminal investigations. In modern applications, especially those involving automation or remote access, the demand for non-contact, scalable, and infrastructure-light height estimation has grown significantly.

Traditional height measurement techniques require physical contact, stadiometers, and manpower, which are impractical in many real-world scenarios. In contrast, vision-based systems enable height estimation using existing camera infrastructure, such as surveillance CCTV or web cameras, without specialized hardware. A contactless approach allows for rapid assessments. An especially valuable feature in environments like clinics, gyms, smart mirrors, or CCTV-monitored public areas.

Additionally, estimating height from a single image is a proxy task for interpreting 3D geometry from 2D observations. It serves as both a practical objective and a technical challenge in computer vision.

[*]Equal contribution [1]Department of Computer Science, University of New York, StonyBrook, Korea. Correspondence to: Sojeong Shin <sojeong.shin@stonybrook.edu>, ByungKon Kang <byungkon.kang@sunykorea.ac.kr>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

Prior works have explored height estimation from single depth images using deep neural networks. For instance, Yin and Zhou (Yin & Zhou, 2020) reconstructs 3D human shape and regresses body height. However, such methods rely heavily on the availability of depth sensors and large-scale annotated datasets, which limits their applicability in monocular RGB scenarios.

Given these motivations, we aim to develop a robust, monocular camera-based system for height estimation. Our method combines classical computer vision techniques—camera calibration, semantic segmentation, and 3D reconstruction via ray-plane intersection—to achieve real-world usability and generalization without relying on depth sensors or multi-view setups.

Our main contributions are as follows:

- We introduce a fully vision-based pipeline that estimates human height using only a monocular camera and a planar reference mat, without requiring depth input or markers on the body.
- We show that accurate camera calibration significantly improves estimation robustness, especially under varying viewing angles.
- We demonstrate that the viewing angle (quantified by the ray's angle to the vertical axis) plays a key role in minimizing measurement error, and empirically validate optimal angles for accurate estimation.
- While we utilize a planar reference mat defined by existing floor geometry in our controlled experiments, our approach is generalizable to other environments by replacing the mat with alternative reference cues such as known object dimensions, floor markings, or structural priors.

2. Related Works

2.1. Traditional Height Measurement Techniques

Conventional height measurement techniques typically rely on contact-based devices such as stadiometers or ultrasonic sensors. These systems, although accurate, often require

physical contact with the user, causing discomfort, particularly in public or clinical settings. Furthermore, such devices require operator supervision and maintenance, limiting scalability in non-contact environments.

Depth cameras such as Microsoft Kinect or Intel RealSense offer an alternative by using infrared sensing to obtain 3D data (Zhang, 2012). However, their performance significantly degrades under outdoor lighting conditions. Sunlight interferes with the active infrared patterns used by these sensors, resulting in noisy or missing depth measurements (Han et al., 2013; Fang, 2021). Furthermore, these devices are relatively expensive and power-intensive. In contrast, we employs a single RGB webcam — readily available and cost-efficient — while achieving comparable levels of accuracy.

2.2. Monocular RGB Camera-Based Height Estimation

Several studies have explored human height estimation using only RGB images. Some methods apply statistical body models, such as SMPL (Bogo et al., 2016), to regress 3D human mesh from a single image, allowing height inference. These approaches, however, often require full-body visibility, clean backgrounds, and significant computational resources for optimization.

Others, like (Banerjee, 2018; Xu, 2022), regress height directly from 2D image cues using deep learning. While these methods are lightweight, they lack physical grounding and struggle with generalization across camera setups. Our method differs by explicitly reconstructing the 3D positions of the head and feet using calibrated geometry, improving interpretability and robustness.

2.3. Camera Calibration and 3D Reconstruction

Camera calibration is essential for precise 3D measurements. The widely-used technique by Zhang (Zhang, 2000), based on multiple checkerboard views, enables accurate intrinsic calibration. Tools like OpenCV (ope, 2024) have implemented this pipeline effectively. For extrinsic calibration, many works either assume ideal camera positioning or use simplified estimates of camera orientation (Battiato, 2007). Our study improves this by solving a Perspective-n-Point (PnP) problem based on a planar reference mat with known world coordinates, yielding accurate extrinsic parameters even with low-cost cameras.

The ray-plane intersection technique for depth inference has been used in robotics and AR (Hartley & Zisserman, 2003), but its adaptation for vertical human height measurement in monocular settings remains underexplored. By combining intrinsic and extrinsic calibration, we can reconstruct real-world 3D positions without relying on depth sensors.

Figure 1. Checkerboards for intrinsic calibration using OpenCV.

2.4. Human Segmentation for Height Estimation

Segmentation of human body parts is a prerequisite for localizing the head and foot pixels. Previous works have used bounding box-based detectors (e.g., YOLO, SSD), but such methods are coarse and prone to errors in cluttered scenes. Semantic segmentation models like Mask R-CNN (He et al., 2017) and HRNet (Sun, 2019) provide higher granularity, yet are computationally heavy.

In our study, we adopt DeepLabV3+ (Chen, 2018), which balances accuracy and speed. Even in a CPU-only environment, it produces reliable segmentation masks in approximately 0.8 seconds per frame. This enables real-time or near-real-time application without GPUs, which is an important consideration for deployment in public or embedded systems.

3. Proposed Method

Our height estimation pipeline consists of two primary components: (1) camera calibration and (2) 3D reconstruction via ray-plane intersection. We first calibrate both intrinsic and extrinsic camera parameters using checkerboard patterns and ground geometry. Then, we recover the 3D foot and head positions by projecting image rays into space and computing their intersections with the planar mat. This approach is inspired by CCTV-Calib (Rameau et al., 2023), which demonstrates how accurate 3D positioning and calibration can be achieved through geometric constraints.

3.1. Camera Calibration

3.1.1. Intrinsic Calibration

We compute the intrinsic parameters of the RGB camera using a Figure1. checkerboard pattern and OpenCV. From multiple views, the intrinsic matrix K and distortion coefficients d are computed. Our calibration resulted in a mean reprojection error of 0.1305 pixels, which corresponds to approximately 0.073 mm in real-world scale.

3.1.2. Extrinsic Calibration with Homography

We utilize a 200cm × 200cm reference mat marked with known 2D-3D correspondences to perform extrinsic calibration. As an initial step, we compute a homography H between the image plane and the planar mat using four man-

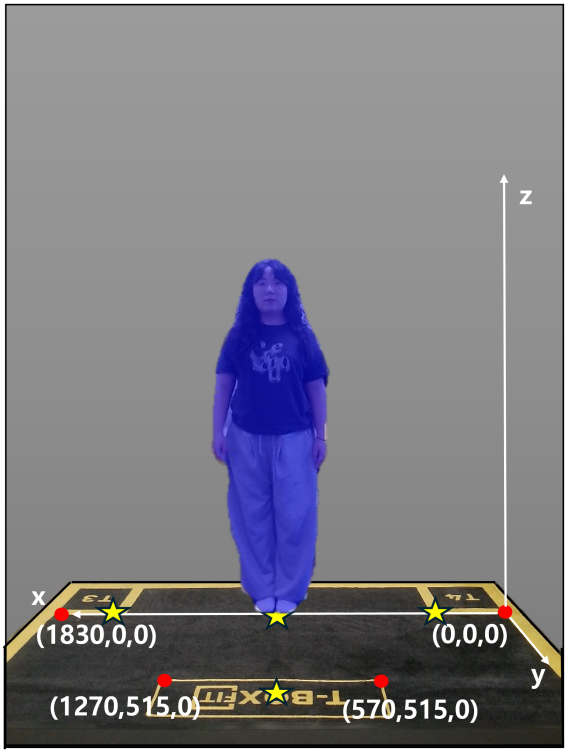

Figure 2. Visualization of 3D world axes and reference points(red) on the planar mat within the camera frame to compute the extrinsic matrix.

ually annotated corner points illustrated in Figure 2. The homography H relates points on the ground plane (X, Y, assuming $Z = 0$) to image coordinates (u, v) as:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = H \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}$$

where s is a scale factor. This homography provides a coarse approximation of the camera pose relative to the ground plane.

Inspired by the calibration strategy in (Rematas et al., 2024), where field markings are treated as fixed geometric constraints, we similarly utilize the mat's line-based structure used as a workout tool only during the initial calibration stage.

To obtain more precise camera positioning in 3D space, we solve the Perspective-n-Point (PnP) problem using OpenCV's `solvePnP()` function. This yields a rotation matrix R and a translation vector t, combining into the extrinsic matrix $[R|t]$

This matrix maps world coordinates on the mat to the camera coordinate frame. The resulting mean reprojection error for this stage was 0.6159 pixels.

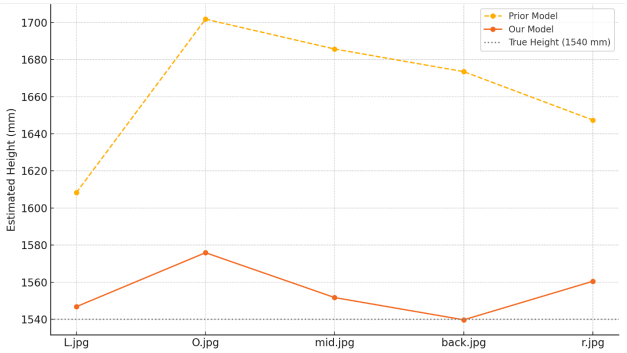

Figure 3. Plot for comparison between prior and proposed model.

3.2. 3D Reconstruction: Human segmentation

To isolate the human subject in each frame, we adopt DeepLabV3+ backbone for semantic segmentation. This model is pretrained on the COCO dataset and fine-tuned on person-specific masks to improve robustness under various clothing styles and lighting conditions.

For each frame, we extract the foreground person mask illustrated in Figure 2 and identify the topmost and bottom-most pixels along the vertical axis within the segmented region, which correspond to the head and foot positions, respectively.

Specifically, for the head, we find the topmost pixel in the segmentation mask. Then, by projecting vertically downward from the head pixel's (x, y) position, we locate the lowest connected pixel within the same column to define the foot position. Also, to reduce the impact of segmentation noise (e.g., hair artifacts), we apply shape filtering and enforce size constraints of the topmost pixels such as a minimum head size (e.g., 35 pixels in horizontal span) to ensure consistency.

These pixel coordinates are then passed to the 3D reconstruction stage, where they are back-projected as rays and intersected with the ground plane to compute real-world 3D locations.

This segmentation step is critical for locating anatomically relevant keypoints without relying on pose estimation models, which can be error-prone in monocular side views or occluded scenarios.

3.3. 3D Reconstruction: Ray-Plane Intersection

To estimate human height, we back-project 2D image pixels into 3D space using the camera intrinsic and extrinsic parameters. Each pixel is treated as a ray emanating from the camera center.

Given a pixel coordinate $\mathbf{p} = [u, v, 1]^T$, the corresponding

Figure 4. Height estimation comparison for subject SJ (GT: 1540) across four reference viewpoints. Our model (bottom) achieves more accurate predictions than the prior model (top). (unit: mm).

ray in the camera coordinate frame is calculated as:

$$\mathbf{r}_{\text{cam}} = K^{-1} \cdot \mathbf{p}$$

Let the normal of the ground plane be $\mathbf{n} = R \cdot [0, 0, 1]^T$ and a known point on the plane be $\mathbf{P}_0 = t$. The intersection scalar λ_{foot} for the foot ray is obtained by:

$$\lambda_{\text{foot}} = \frac{\mathbf{n}^T(\mathbf{P}_0)}{\mathbf{n}^T \cdot \mathbf{r}_{\text{foot}}}, \quad \text{where } \mathbf{r}_{\text{foot}} = K^{-1} \cdot \mathbf{p}_{\text{foot}}$$

Using this, the 3D foot position is:

$$\mathbf{X}_{\text{foot}} = \lambda_{\text{foot}} \cdot \mathbf{r}_{\text{foot}}$$

For the head, since it does not lie on the same plane, we scale the ray using the known height of the camera. Specifically, if the z-coordinate of the ground is $z_{\text{foot}} = -998$ mm, and the camera is located at the origin of the camera coordinate frame, the head point is computed as:

$$\mathbf{X}_{\text{head}} = \lambda_{\text{head}} \cdot \mathbf{r}_{\text{head}},$$
$$\text{where } \lambda_{\text{head}} \text{ is chosen such that } \mathbf{X}^z_{\text{head}} = 0$$

To compute height robustly along the real-world vertical direction (not necessarily aligned with the camera's z-axis), we define a unit upward vector $\mathbf{v}_{\text{up}} = R \cdot [0, 0, 1]^T$. Then, the final estimated height is the projection of the 3D displacement vector onto this vertical axis:

$$\text{height} = \left| (\mathbf{X}_{\text{head}} - \mathbf{X}_{\text{foot}})^T \cdot \mathbf{v}_{\text{up}} \right|$$

Using the known calibration parameters, we convert pixel-level measurements into real-world metric values (in mil-

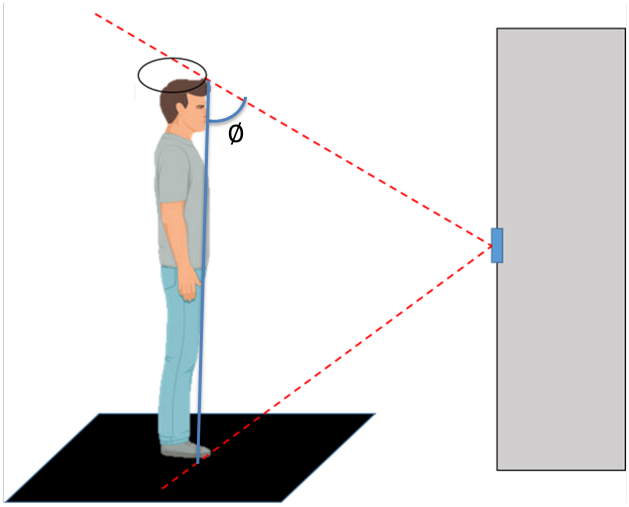

Figure 5. Geometric hypothesis: As the camera's view angle θ to the head ray becomes more parallel to the z-axis, small segmentation errors are amplified.

limeters), enabling accurate height estimation from monocular RGB images resulting in Figure6.

4. Experiments

4.1. Comparison with Prior Model

To evaluate the effectiveness of our proposed method, we conducted a controlled experiment using the same subject across four distinct camera angles, corresponding to the four star spots marked in Figure 2. Each position provides a different viewpoint to assess robustness under geometric variations.

Figure 4 shows a qualitative comparison between the prior model and our model for the subject SJ (ground truth height: 1540 mm). The top row visualizes the prior model's results, while the bottom row displays our model's.

As illustrated, the prior tends to overestimate the subject's height across all viewpoints, producing values as high as 1711.8 mm. In contrast, our model yields estimates much closer to the ground truth, with values ranging from 1542.6 mm to 1575.0 mm.

In Figure 3, we plot the estimated heights for both models across the four viewpoints. Our model demonstrates significantly improved accuracy, reducing the mean absolute error (MAE) by **87.4%** compared to the prior model.

4.2. Error Analysis and Hypothesis

Notably, an outlier appears at `O.jpg` in our method, where the estimated height slightly exceeds the ground truth.

Foot Position	MAE (cm)
O	2.31
L	0.83
Back	**0.44**
R	1.35

Table 1. Mean Absolute Error (MAE) by foot position across all subjects. Unit: cm.

To better understand this behavior, we set a geometric hypothesis regarding the head ray direction. Specifically, we observe that as the subject gets closer to the camera, they occupy a larger portion of the frame, resulting in a more parallel orientation of the head ray to the optical (z) axis. This geometric configuration amplifies small errors in pixel position, leading to larger reconstruction errors.

Experimental Result: As shown in Figure 5, we hypothesize that the estimation error increases when the head ray becomes more aligned with the z-axis. In other words, as θ approaches 0° or 180°, the projection of small pixel shifts into 3D space becomes more unstable.

Hypothesis 1: *If the theta value is less than 60°, or greater than 120°, the height estimation error tends to increase.*

To validate this hypothesis, we conduct the same experiment with 8 additional participants, each recorded at four different viewpoints illustrated in Figure6. For every position, we compute the ray angle θ and corresponding height estimation error, and analyze the trend across all cases.

5. Experimental Result – Effect of Viewpoint on Estimation Accuracy

To evaluate how the viewpoint affects height estimation accuracy, we conducted experiments with 8 participants, each positioned sequentially at four predefined foot positions: O, L, r, and back (as illustrated in Figure 6). These positions correspond to different viewing angles (θ) between the camera and the subject's vertical body axis.

For each individual, we measured the height estimation error at all four positions and plotted the results against their respective θ values (Figure 7). The x-axis represents the θ angle (in degrees), and the y-axis represents the height error (in cm).

Key Observations:

- Across all subjects, the height estimation error clearly varies depending on the viewpoint.
- A consistent trend emerges where the error is minimized when the subject is located at the **'back'** position.

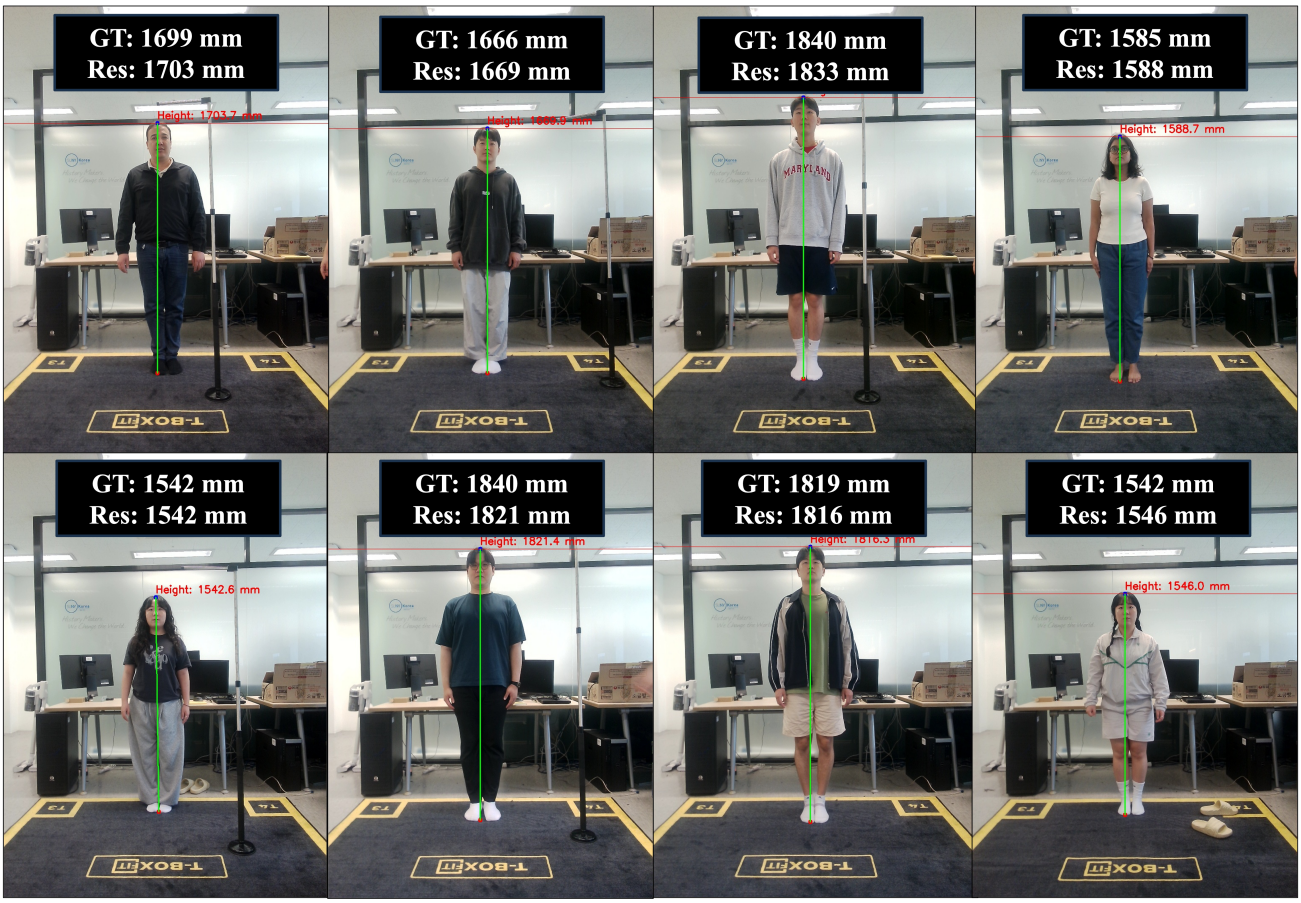

Figure 6. Height estimation comparison for 7 individuals with different body types as test subjects across four reference viewpoints.

- The mean absolute error (MAE) at each position, aggregated over all subjects, is summarized in Table 4.2, showing that the 'back' position yields the smallest error at **0.44 cm**.

These results support the hypothesis introduced in Figure 5, that a moderate viewing angle improves reconstruction accuracy by optimizing the head ray's intersection geometry.

6. Conclusion and Discussion

In this study, we proposed a monocular height estimation pipeline that leverages camera calibration and geometric ray-plane intersection to accurately infer human height in real-world units. Using a planar reference mat for extrinsic calibration and semantic segmentation for foot–head localization, we demonstrated that our system achieves centimeter-level accuracy under controlled settings.

Through extensive experiments involving eight participants across four predefined viewpoints (O, L, r, and back), we validated that the camera viewing angle θ significantly influences estimation accuracy. In particular, our results show that the **'back'** position—corresponding to a moderate downward viewing angle—consistently yields the lowest mean absolute error (MAE). A cross-subject analysis revealed that the MAE at the back position was as low as 0.44 cm, outperforming other views by a considerable margin.

Discussion. These findings support the hypothesis that height estimation error is minimized when the head ray is neither too steep nor too parallel to the optical axis. As the viewing angle becomes shallow (e.g., at the 'O' position), the head ray becomes more aligned with the z-axis, amplifying depth estimation errors. Conversely, an excessively oblique angle may also degrade performance due to segmentation inaccuracies or perspective distortion.

While our method assumes known calibration from a static mat, future work may explore calibration-free or self-supervised alternatives such as those used in surveillance settings (Rameau et al., 2023). Moreover, integrating temporal consistency across frames or fusing pose estimation cues could further improve robustness, particularly in real-time

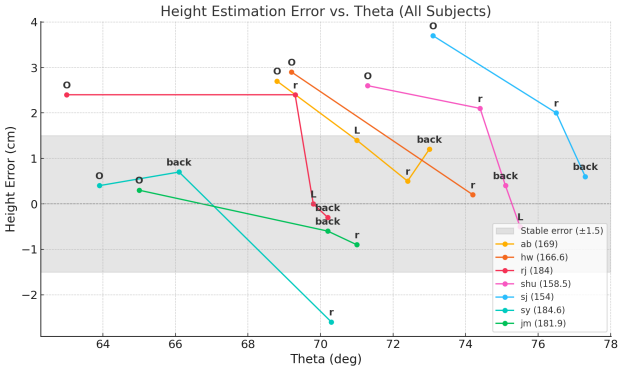

Figure 7. eight estimation error at all four positions and plotted the results against their respective θ values.

or outdoor applications.

In addition, we plan to fine-tune DeepLabV3+ to segment specific human body parts (e.g., torso or limbs), enabling partial-length measurements such as limb proportions or seated height. This approach can be extended beyond humans by incorporating generic object segmentation models, making the system **scalable** for estimating the physical dimensions of arbitrary objects in diverse environments.

Overall, this work provides a practical, lightweight solution for vision-based height estimation and offers geometric insights into viewpoint-aware performance optimization.

References

Opencv library, 2024. https://opencv.org.

Banerjee, S. e. a. Height estimation from a single image using convolutional neural networks. *arXiv preprint arXiv:1803.05813*, 2018.

Battiato, S. e. a. Sift features tracking for video stabilization. In *IEEE International Conference on Image Analysis and Processing*, 2007.

Bogo, F., Kanazawa, A., Lassner, C., Gehler, P. V., Romero, J., and Black, M. J. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016.

Chen, L.-C. e. a. Encoder-decoder with atrous separable convolution for semantic image segmentation. *ECCV*, 2018.

Fang, Y. e. a. Depth estimation from monocular images: A survey. *Pattern Recognition*, 2021.

Han, J., Shao, L., Xu, D., and Shotton, J. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Transactions on Cybernetics*, 2013.

Hartley, R. and Zisserman, A. *Multiple View Geometry in Computer Vision*. Cambridge university press, 2003.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *ICCV*, 2017.

Rameau, F., Choe, J., Pan, F., Lee, S., and Kweon, I. S. Cctv-calib: a toolbox to calibrate surveillance cameras around the globe. *Machine Vision and Applications*, 34 (2):28, 2023. doi: 10.1007/s00138-023-01476-1.

Rematas, K., Li, S., Wang, L., Xiang, T., Tuytelaars, T., and Vedaldi, A. A universal protocol to benchmark camera calibration for sports. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15009–15018, 2024.

Sun, K. e. a. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

Xu, W. e. a. Rethinking height estimation: A learning-based approach using scene and person context. *IEEE Access*, 2022.

Yin, F. and Zhou, S. Accurate estimation of body height from a single depth image via a four-stage developing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8267–8276, 2020.

Zhang, Z. A flexible new technique for camera calibration. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

Zhang, Z. Microsoft kinect sensor and its effect. *IEEE Multimedia*, 2012.