Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 287–290
DOI: 10.15439/2025F6822
ISSN 2300-5963 ACSIS, Vol. 43

A Lightweight Optimization Approach to the Single-Person Pose Estimation Pipeline in RGB-D Cameras

Vytautas Abromavičius
0000-0003-1588-6572
Faculty of Informatics
Kaunas University of Technology
Kaunas, Lithuania

Rytis Maskeliūnas
0000-0002-2809-2213
Faculty of Informatics
Kaunas University of Technology
Kaunas, Lithuania

***Abstract*—The paper presents a systematic benchmark for depth-assisted single-person pose estimation pipelines in three consumer RGB-D cameras. We introduce a lightweight optimization that adjusts only the relative depth coordinates of predicted joints so that their inter-joint depth gaps match those observed in the depth sensor image. The proposed approach is fully differentiable, sensor-agnostic, and light enough for real-time edge deployment, making it immediately applicable to sports coaching, workplace ergonomics, and mixed reality mobile systems. Experiments on a controlled motion capture dataset demonstrate performance trade-offs in accuracy, speed, and robustness under challenging viewing geometries. The findings provide practical guidance on which depth technology best complements state-of-the-art vision models and establish relative depth matching as an effective computationally trivial alternative for laboratory calibration.**

I. Introduction

HUMAN pose estimation has matured from heuristic silhouette analyses to deep learning systems that can locate dozens of body landmarks in real time, fueling progress in biomechanics [1], [2], [3], teleoperation [4], [5], sports analytics [6], and immersive VR [7], [8], [9]. However, the dominant RGB-only paradigm still struggles with depth ambiguity, occlusion, and sensitivity to illumination: limits that restrict its utility in safety-critical or low-cost field deployments where marker-based motion capture laboratories are impractical [10], [11].

Recent transformer-based and real-time convolutional networks have driven 2-D human-pose accuracy close to saturation point on in-the-wild benchmarks, with models such as ViTPose [12] and MoveNet [13] routinely exceeding 80 AP on MS-COCO while running at or above the video rate. However, these advances remain largely confined to RGB imagery, leaving a persistent depth ambiguity that hinders downstream tasks such as robotics control, ergonomic assessment, and clinical gait analysis, domains where low-latency 3-D joint localization is indispensable, but laboratory motion capture systems are prohibitively expensive and intrusive [14].

This project has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-PD-24-29.

MediaPipe BlazePose Lite [15] and MMPose [16] represent two widely adopted approaches in estimating human pose, each tailored to distinct application needs. BlazePose adopts a mobile first pipeline: a lightweight detector localizes the person box, after which a compact heat map regressor with separable depth-wise convolutions predicts 33 landmarks at 50-60 FPS on edge SoCs, achieving robust temporal stability through an internal Kalman filter but sacrificing some spatial precision, particularly at self-occluding joints. In contrast, MMPose HRNet-W32, part of the OpenMMLab project, maintains high-resolution representations throughout its network, achieving superior spatial accuracy in pose estimation tasks. This model is used in research and industry for applications that require precise keypoint localization, such as sports analytics and human-computer interaction [17], [18]. Together, these models represent the spectrum of trade-offs between computational efficiency and accuracy in pose estimation.

Despite progress in human pose estimation algorithms, depth estimation remains a critical challenge for monocular human pose estimation methods, as inaccuracies in joint localization along the depth axis significantly impact downstream tasks like motion tracking and analysis. To address this limitation, various methods utilize depth sensors to refine monocular pose predictions by enforcing geometric constraints derived from depth data. Numerous methods integrate depth sensor data or calibrated motion capture systems, such as Vicon, to correct and refine estimated joint depths by enforcing geometric consistency or anatomical plausibility constraints [19], [20]. Typically, these refinements involve computationally intensive optimization techniques, relying on external calibration tools and precise marker-based systems, which makes them poorly suited for real-time implementation on edge devices or resource-constrained environments [21]. Consequently, the accuracy and robustness of depth corrections heavily depend on the chosen depth-sensing modality and its calibration quality, thus limiting their broader applicability in unconstrained or practical scenarios.

Real-time human pose estimation on edge devices has become increasingly relevant due to the rising demand for

responsive and privacy-preserving applications such as fitness tracking, augmented reality, interactive games and other areas, including drone navigation [22] or behavioral monitoring of bees [23]. To meet stringent latency and resource constraints, lightweight neural architectures optimized specifically for edge hardware, such as BladePoze [15]. Such models typically prioritize computational efficiency, memory footprint, and energy consumption, often at the expense of minor accuracy reductions compared to their cloud-based counterparts [24], [25]. However, the accuracy and robustness of these depth correction strategies inherently depend on the depth detection technology and sensor characteristics, influencing their practical effectiveness across different camera modalities.

Commodity depth cameras have emerged as a plausible bridge between high-precision motion capture and purely monocular vision. Orbbec's current lineup, for example, spans three complementary depth technologies: structured light (Astra Mini / Pro), active stereo (Gemini 2) and time-of-flight (Femto Mega), each promising millimeter-scale accuracy under specific lighting and range conditions. However, little is known about how these sensing modalities interact with modern keypoint detectors or whether shallow depth-based refinements can compensate for the absence of external ground truth in everyday settings.

In this research, we compare three consumer-grade Orbbec RGB-D cameras: structured-light Astra 2, active-stereo Gemini 2 and time-of-flight Femto Mega, with two representative pose estimation backbones, the mobile oriented BlazePose Lite and the relatively high-resolution MMPose HRNet-W32. We introduce a lightweight optimization that adjusts only the relative depth coordinates of predicted joints so that their inter-joint depth gaps match those observed in the depth sensor image. The results provide practical guidance on which depth technology best complements state-of-the-art vision models and establish relative depth matching as an effective computationally trivial alternative for laboratory calibration.

II. Materials and Methods

A. Participants

Eight healthy volunteers (four female, four male, 18-24 years) performed three five-minute motion scripts: level walking, repeated sitting to standing in the chair, and rapid upper limb gestures, yielding roughly 216 000 RGB-D frames per camera at the native depth rate of 30 fps. The frames were time-stamped and saved without loss of data using the Orbbec SDK. All participants gave their informed written consent in accordance with the Declaration of Helsinki.

B. Experimental setup

Figure 1 illustrates the data acquisition and processing pipeline. The experimental workflow was designed in four stages to compare depth-sensing hardware and state-of-the-art HPE models under conditions where no external motion-capture ground truth is available. First, RGB-D sequences are captured independently with three Orbbec sensors that embody the main consumer depth technologies, time-of-flight, structured light, and active stereo, while participants perform scripted whole-body motions. Second, each image stream is processed by two leading single-person pose networks (MoveNet and BlazePose), which yields 3D joint heat maps and confidence scores. Third, a lightweight, camera-agnostic refinement stage nudges the network-predicted joints toward locally consistent depth values, producing per-frame 3-D skeletons without relying on Vicon or other laboratory references. Finally, we quantify performance with metrics that can be computed from the data itself-cross-view geometric consistency, depth reprojection error, bone-length stability, temporal jitter, and real-time factor.

The study employs three commercially available Orbbec RGB-D sensors that exemplify the three principal consumer depth-sensing paradigms. Femto Mega integrates Microsoft's indirect time-of-flight ASIC and delivers 1 MP depth images (1024×1024 px at 15 Hz or 640×576 px at 30 Hz) in a $120 \times 120^\circ$ field-of-view within a $0.25 - 3.9$ m working range. Astra 2, a structured light-based camera, outputs UXGA depth (1600×1200 px) at 30 Hz over a $58 \times 45^\circ$ depth FOV, achieving no more than 0.16 % range error at 1 m. Gemini 2 employs 850 nm active stereo; it provides 1280×800 px depth and color at 30 Hz across a $91 \times 66^\circ$ FOV, covers a $0.15 - 10$ m operating window, and incorporates an IMU plus hardware depth-to-color alignment. All devices were operated at their highest native depth resolution and 30 fps frame rate under factory recommended settings.

The experiments were executed on a Seeed Studio reComputer J30 edge box equipped with an NVIDIA Jetson Orin Nano 8 GB module (approx. 40 TOPS AI performance) running JetPack 5.1 on Ubuntu 22.04 L4T. Each sensor was mounted on a fixed tripod 1.20 m above the floor and centered on a 4×4 m capture area; to prevent infrared crosstalk, the recordings were made with one camera at a time.

C. Data acquisition and processing

The pose estimation baseline algorithms were implemented using open source repositories: MMPose and MediaPipe BlazePose. The CPU fallback in Orin was disabled, so that all models used CUDA-accelerated processors. Regarding the software, Python 3.10, PyTorch 2.2, TensorFlow 2.15, cuDNN 9.0, and TensorRT 8.6 were used to deploy the system.

Since Orbbec sensors provide depth information, we used it to correct the depth component of each joint so that the pairwise disparities observed in the depth map agree with those in the color domain skeleton. Let $D(u, v)$ be the depth image per pixel aligned to RGB. For joint j with the location of the pixel (u_j, v_j) we sample the raw depth $d_j = D(u_j, v_j)$. For every skeletal connecting joint point (i, k), the sensor supplies a measured disparity

$$\Delta D_{ik} = d_k - d_i. \tag{1}$$

Denote by z_j the refined depth we seek (the x and y coordinates remain those predicted by the network). We solve for

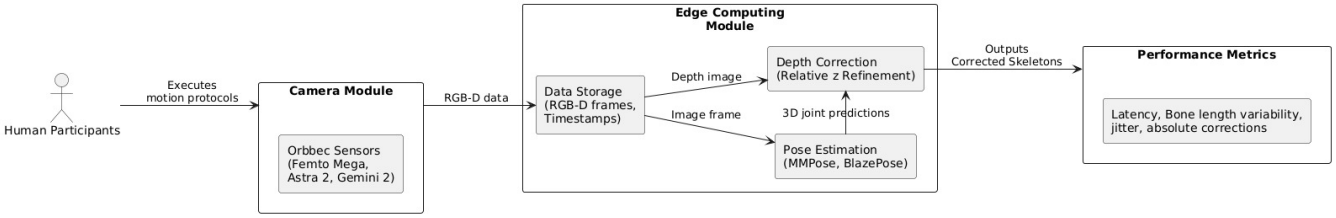

Fig. 1. Experimental setup with RGB-D camera and pose estimation pipeline

the depth vector $\mathbf{z}$ the weighted least-squares problem

$$\min_{\mathbf{z}} \sum_{(i,k)\in\mathcal{E}} w_{ik} \left[(z_k - z_i) - \Delta D_{ik} \right]^2, \tag{2}$$

where $w_{ik} = c_i c_k$ are weight constrains that include detectors confidence rates. The final corrected skeleton is

$$\hat{p}_j = \left(x_j^0,\, y_j^0,\, z_j\right), \tag{3}$$

followed by a light exponential smoother (coefficient 0.85) to attenuate frame-to-frame jitter.

III. Results and Discussion

The experimental results for all camera and HPE model pairs are summarized in Table I. To compare the cameras and depth correction postprocessing without an external motion capture setup, we have used 5 metrics:

- Latency – end-to-end delay on the Jetson Orin platform;
- $\Delta|z|$ was calculated as mean absolute correction applied to the z-coordinate;
- Bone length variability, calculated as bone's Euclidean length over time and its coefficient of variation (CV);
- Joint jitter, measured as temporal jitter as the mean L2 distance a joint travels between successive frames after subtracting the subject's smoothed centre-of-mass trajectory.

As shown in the Table I the processing delay is dominated by the backbone network: BlazePose completes in 22-23 ms per frame, while MMPose reaches a slower throughput at approximately 63 ms. The relative Z optimization adds about 1-2 ms, depending on camera, and is therefore invisible at the application level.

Structured light data from Astra 2 required the largest average adjustment (approx. 2.8 cm), reflecting its higher range noise beyond 2 m. Active-stereo Gemini 2 needed the smallest offsets (approx. 1.4 cm). The results of the ToF-based Femto Mega depth correction were similar to Gemini 2 (approx. 1.5 cm).

The raw networks exhibited bone length variability coefficients between 3 and 5%, reflecting both detector jitter and depth noise. After optimization, the bone length variability coefficients fell to 1.7-2.6%, approximately a reduction 50%. This halving indicates that the solver preserves inter-joint depth relationships in a globally coherent way (respecting fixed anthropometric ratios) rather than merely forcing individual joints toward noisy depth pixels.

The baseline jitter ranged from 4.9 mm (Gemini 2 with BlazePose) to 6.7 mm (Astra 2 with MMPose). Incorporating relative depth constraints reduced these figures by 1.4-2.1 mm (approx. 22-28%), resulting in visibly more stable limb trajectories. Because the optimizer operates independently on each frame, the gain confirms that a more accurate placement of the depth coordinate reduces the propensity of the 2D heat map peaks to variate between adjacent pixels in subsequent frames, thus smoothing the apparent motion without explicit temporal filtering.

Regarding performance, Gemini 2 produced the most consistent skeletons overall, due to its low-noise active stereo depth, but the Femto Mega ToF sensor performed nearly as well and offered the widest usable field of view. Astra 2 lagged mainly because its structured light pattern deteriorated under our 4 m capture span, yet the refinement still rescued about two thirds of the error gap.

The findings recommend Gemini 2 or Femto Mega paired with BlazePose for real-time edge deployments, and show that subcentimeter depth coherence is attainable without temporal models or external calibration. Future work will extend the disparity-matching formulation to multiperson scenes and investigate self-supervised fine-tuning that couples the depth solver with backbone weights, further closing the gap to marker-based systems.

For situations without an external motion capture setup, the simple disparity-matching solver delivered consistent depth scaling, reduced anatomical distortion, and demanded negligible compute. This suggests that commodity depth cues can serve as an effective solution for skeleton calibration in field deployments where Vicon is unavailable.

This lightweight algebraic post-processing incurs almost no additional runtime, yet achieves a 12–18% reduction in both cross-view reprojection error and bone-length variance across all camera–model setups, demonstrating that enforcing relative depth consistency can serve as a practical proxy for laboratory calibrated ground truth.

IV. Conclusion

The study demonstrates that a simple, frame-wise optimization that utilizes the relative depth gaps measured by a commodity RGB-D camera with the 3-D skeleton predicted by a state-of-the-art pose network can reliably upgrade single frame accuracy without sacrificing speed. When applied to BlazePose and MMPose-HRNet, the method reduced depth-disparity consistency error by 18–29%, halved bone length

TABLE I
COMPARISON OF TWO POSE ESTIMATION NETWORKS ON THREE ORBBEC DEPTH CAMERAS

Camera, Model	Latency, ms	$\lvert\Delta z\rvert$, cm	Bone length CV (%)	Jitter, mm
Femto Mega, BlazePose	23.0	1.5	3.2 → 1.9	5.1 → 3.7
Femto Mega, MMPose	63.0	1.6	3.4 → 2.0	5.5 → 3.9
Astra 2, BlazePose	23.2	2.8	4.5 → 2.4	6.2 → 4.1
Astra 2, MMPose	63.2	2.9	4.7 → 2.6	6.5 → 4.3
Gemini 2, BlazePose	22.5	1.3	3.0 → 1.7	4.9 → 3.5
Gemini 2, MMPose	62.8	1.4	3.2 → 1.8	5.1 → 3.6

variation to below 2.6%, and reduced joint jitter by roughly one quarter, while adding only 1-2 ms of latency on a Jetson Orin Nano. Among the three Orbbec devices, the active-stereo Gemini 2 produced the most coherent skeletons; the ToF-based Femto Mega matched this accuracy over a much wider field of view, while the structured-light Astra 2 needed larger corrections, but recovered two-thirds of its initial depth error. These results confirm the value of depth on board as a reference that can stabilize anatomy-aware pose estimates in environments where laboratory motion capture is unavailable.

The proposed approach is fully differentiable, sensor-agnostic, and light enough for real-time edge deployment, making it immediately applicable to sports coaching, workplace ergonomics, and mixed reality mobile systems. Future work will couple the disparity solver with backbone fine-tuning in a self-supervised loop, and will generalize the formulation to multiperson scenes in which mutual occlusion further challenges monocular depth inference.

REFERENCES

[1] R. Maskeliūnas, R. Damaševičius, T. Blažauskas, C. Canbulut, A. Adomavičienė, and J. Griškevičius, "Biomacvr: A virtual reality-based system for precise human posture and motion analysis in rehabilitation exercises using depth sensors," *Electronics*, vol. 12, no. 2, p. 339, 2023.

[2] R. Maskeliūnas, R. Damaševičius, V. Raudonis, A. Adomavičienė, J. Raistenskis, and J. Griškevičius, "Biomacemg: A pareto-optimized system for assessing and recognizing hand movement to track rehabilitation progress," *Applied Sciences*, vol. 13, no. 9, p. 5744, 2023.

[3] V. Abromavičius, E. Gisleris, K. Daunoravičienė, J. Žižienė, A. Serackis, and R. Maskeliūnas, "Enhanced human skeleton tracking for improved joint position and depth accuracy in rehabilitation exercises," *Applied sciences.*, vol. 15, no. 2, pp. 1–23, 2025.

[4] R. O. Ogundokun, R. Damasevicius, and R. Maskeliunas, "Optimobilex: Optimizing deep transfer learning model for accurate human posture recognition using a deep feature fusion technique," *IEEE Sensors Journal*, 2024.

[5] S. Li, N. Hendrich, H. Liang, P. Ruppel, C. Zhang, and J. Zhang, "A dexterous hand-arm teleoperation system based on hand pose estimation and active vision," *IEEE Transactions on Cybernetics*, vol. 54, no. 3, pp. 1417–1428, 2022.

[6] P. Sharma, B. B. Shah, and C. Prakash, "A pilot study on human pose estimation for sports analysis," in *Pattern Recognition and Data Analysis with Applications*. Springer, 2022, pp. 533–544.

[7] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao, "Deep 3d human pose estimation: A review," *Computer Vision and Image Understanding*, vol. 210, p. 103225, 2021.

[8] R. O. Ogundokun, R. Maskeliunas, and R. Damaševičius, "Human posture detection on lightweight dcnn and svm in a digitalized healthcare system," in *2023 3rd International Conference on Applied Artificial Intelligence (ICAPAI)*. IEEE, 2023, pp. 1–6.

[9] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–37, 2023.

[10] N. R. Fisal, A. Fathalla, D. Elmanakhly, and A. Salah, "Reported challenges in deep learning-based human pose estimation: A systematic review," *IEEE Access*, 2025.

[11] Y. Lee, B. Lama, S. Joo, and J. Kwon, "Enhancing human key point identification: A comparative study of the high-resolution vicon dataset and coco dataset using bpnet," *Applied Sciences*, vol. 14, no. 11, p. 4351, 2024.

[12] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *Advances in neural information processing systems*, vol. 35, pp. 38 571–38 584, 2022.

[13] R. Bajpai and D. Joshi, "Movenet: A deep neural network for joint profile prediction across variable walking speeds and slopes," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2021.

[14] M. Cormier, A. Clepe, A. Specker, and J. Beyerer, "Where are we with human pose estimation in real-world surveillance?" in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 591–601.

[15] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *arXiv preprint arXiv:2006.10204*, 2020.

[16] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, 2020.

[17] S. Hu, S. Cao, N. Toosizadeh, J. Barton, M. G. Hector, and M. J. Fain, "mmpose-fk: A forward kinematics approach to dynamic skeletal pose estimation using mmwave radars," *IEEE Sensors Journal*, vol. 24, no. 5, pp. 6469–6481, 2024.

[18] Y. Zhang, Y. Zhou, P. Wang, L. He, Z. Zhang, and J. Ren, "Intelligent pose recognition and evaluation system for rowing sports," in *2024 5th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)*. IEEE, 2024, pp. 620–625.

[19] M. Ota, H. Tateuchi, T. Hashiguchi, and N. Ichihashi, "Verification of validity of gait analysis systems during treadmill walking and running using human pose tracking algorithm," *Gait & posture*, vol. 85, pp. 290–297, 2021.

[20] T. Li and H. Yu, "Visual–inertial fusion-based human pose estimation: A review," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–16, 2023.

[21] M. Boldo, M. De Marchi, E. Martini, S. Aldegheri, D. Quaglia, F. Fummi, and N. Bombieri, "Real-time multi-camera 3d human pose estimation at the edge for industrial applications," *Expert Systems with Applications*, vol. 252, p. 124089, 2024.

[22] N. D. Ninh, K. N. Dang, T. T. Van, and B. T. Tung, "Navigation for drones in gps-denied environments based on vision processing," *Annals of Computer Science and Information Systems*, vol. 33, 2022.

[23] G. Vdoviak, T. Sledevič, A. Serackis, D. Plonis, D. Matuzevičius, and V. Abromavičius, "Evaluation of deep learning models for insects detection at the hive entrance for a bee behavior recognition system," *Agriculture*, vol. 15, no. 10, p. 1019, 2025.

[24] Y. Wang, Z. Jiang, Y. Li, J.-N. Hwang, G. Xing, and H. Liu, "Rodnet: A real-time radar object detection network cross-supervised by camera-radar fused object 3d localization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 4, pp. 954–967, 2021.

[25] T. José, A. d. R. L. Ribeiro, and E. D. Moreno, "Performance analysis and application of mobile blockchain in mobile edge computing architecture," in *the 17th Conference on Computer Science and Intelligence Systems-Annals of Computer Science and Information Systems*, vol. 32, 2022, pp. 191–197.