# PISC: Physics-Informed Scene Constraints Method for Markerless 3D Human Localization for Safe Human-Robot Collaboration

Rafał Kozik[1,3], Szymon Buś[2], Aleksandra Pawlicka[1], Marek Pawlicki[1,3], and Michał Choraś[1,3]

[1] ITTI Sp. z.o.o, Poznań, Poland
[2] Sieć Badawcza Łukasiewicz – Przemysłowy Instytut Automatyki i Pomiarów PIAP, Warsaw, Poland
[3] Bydgoszcz University of Science and Technology, Bydgoszcz, Poland

**Abstract.** Accurate 3D human localization is important for safe human-robot collaboration in industrial environments. This problem is particularly challenging when occlusions, limited overlap between camera fields of view, and changing workspace configurations make reliable tracking process difficult. In this paper, we propose a markerless method for 3D human localization that combines multi-view geometry, scene physical constraints, and context-based sensor fusion. The method uses raycasting and epipolar constraints to estimate human body center positions from multiple calibrated cameras. To ensure robust operation under different sensing conditions, we formulate localization as a meta-policy learning problem that selects the most reliable estimation strategy based on the scene context. We evaluate the approach in a real human-robot workshop using the RoHuCAD dataset and additional experiments with tracker-based ground truth. The results show that the proposed method produces more accurate and more consistent localization than monocular pose estimation and depth-based baselines.

**Keywords:** human-robot collaboration · hybrid AI · artificial intelligence · computer vision

## 1 Introduction

Accurate 3D human localization is a key element for safe and effective human-machine collaboration, particularly in environments involving collaborative robots (cobots). Existing pose estimation approaches are often based on statistical models, such as SMPL [17]. However, under occlusions or ambiguous viewpoints, these techniques may produce physically infeasible body configurations. Voxel-based methods like VoxelPose or FasterVoxelPose [19] improve spatial accuracy but require retraining for each camera setup. This is limiting their flexibility in dynamic or reconfigurable workspaces. Wearable motion capture systems, while accurate, are intrusive and impractical for continuous use on the shop floor.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

In this work, we present a markerless method for 3D human localization based on scene constraints and multi-view geometry. The method uses epipolar constraints to combine information from multiple cameras and applies workspace constraints to limit incorrect estimates. It does not require additional hardware and can be used in reconfigurable environments. The approach is intended for human-robot collaboration in industrial workspaces.

## 2 Related Work

Pose estimation is a computer vision technique focused on identifying and monitoring the configuration of the human body parts in 2D (e.g., single image) or 3D space. Typically, this involves predicting the coordinates of specific keypoints, such as shoulders, elbows, wrists, hips, knees, and ankles, based on images or video sequences. Because the movement of specific body parts can often be correlated with specific human actions, interpretation of the human pose is useful for applications such as action recognition and video understanding.

In general, two- and three-dimensional pose estimation differ based on the source image/video input used for this task. Three-dimensional pose estimation requires additional (complementary to the 2D space) information, like additional depth-related data gathered by for example RGB cameras, 3D or LiDAR scanners, as well as by accelerometers or gyroscopes. In addition to 2D/3D differentiation, pose estimation techniques can be categorized also as the bottom-up and top-down methods in which individual body joints are initially identified and tracked to obtain a complete pose estimation (bottom-up techniques), or in which a person is initially identified and then, body pose is estimated within the detected bounding boxes (top-down methods) [13]. Pose estimation methods can also be classified based on the proposed model:

- Both 2D and 3D kinematic pose estimation, able to identify the relation between different human body parts based on joint and limb positions.
- 2D planar pose estimation, in which the human body is represented as a collection of planar structures (e.g., rectangles).
- 3D volumetric pose estimation, in which body joints are represented in a 3D voxel grid, capturing depth information.

### 2.1 Direct Methods

DeepCut [7][14] and DeeperCut [8] models that were one of the earliest bottom-up approaches developed to address the challenges of 2D multi-person pose estimation. These techniques first estimate the number of persons present in a scene, and then the models identify and distinguish specific parts of the bodies. Body parts detection is carried out using Fast R-CNN (Region-based Convolutional Neural Network).

HRNet (High-Resolution Network) [2] [15] is used for semantic segmentation, object detection, and image classification in 2D pictures using CNN (Convolutional Neural Network). One of the main strengths of this solution is the fact that

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

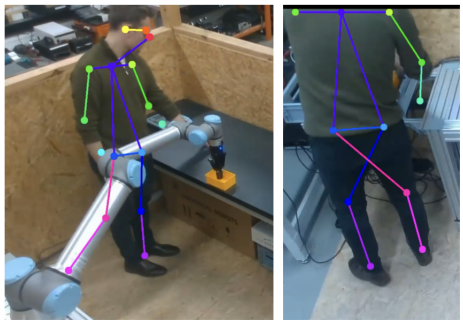

**Fig. 1.** Problem of occlusion impacting single image pose regression (PoseNet with ResNet50 backbone).

HRNet maintains high-resolution representations through the whole process. In contrast to the majority of pose estimation models it recovers high-resolution representations from low resolution representations produced by a high-to-low resolution network.

PoseNet [18][9] allows for real-time human pose estimation. It is carried out in two-stage manner. First, the input image or video is fed through a CNN. Next, the poses are decoded by a decoding algorithm and pose confidence scores, keypoint positions, and keypoint confidence scores are estimated. One of the problem with PoseNet is that it estimates body keypoints directly from the image and it does not check anatomical or physical correctness of the predicted pose. The example is shown in Fig. 1.

### 2.2 Regression with Human Body Priors

OpenPose [4] is a popular open-source multi-person pose detection library, providing real-time detection of the 2D (also 3D) orientation of up to 135 different body keypoints. Stacked Hourglass Networks [12] is an architecture using CNN and consisting of multiple stacked hourglass modules, allowing for successive pooling and unsampling of the input to generate the final prediction.

BlazePose [3] is a lightweight solution designed for mobile devices, able to track over 30 body keypoints in real-time. In this approach, an on-device face detector used as a proxy is combined with a lightweight body pose detector and the tracker predicting coordinates for body keypoints in the given video frame.

SMPL [10] is one of the most common 3D pose estimation models, constituting a family (SMPL, SMPL+H, SMPL-X) and applied in such solutions as ROMP or SMPLify. It consists of 3D body, face, hand and foot models, learned from a number of 3D scans. SMPL is able to capture statistical patterns and variations in body shape across the dataset.

ROMP [16] is an approach to a monocular, one-stage, regression of multiple 3D people from RGB videos. In this approach, the input image is used to estimate a body center heatmap (representing the probability of each position being

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

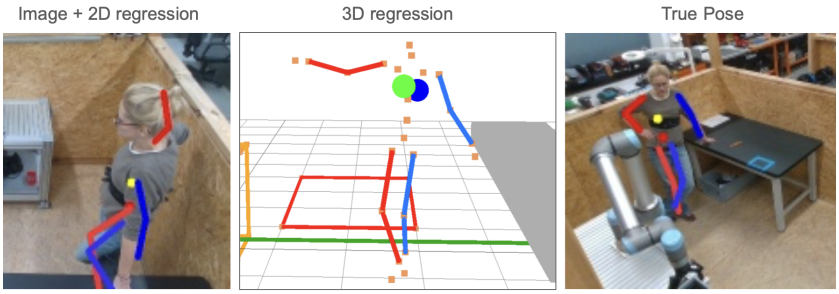


**Fig. 2.** The impact of missing information in single-view pose regression based on ROMP [17] (the right hand is not actually raised).

a body center), alongside specialized camera and SMPL maps. These maps are then integrated into a single Mesh Parameter map. From this combined representation, 3D body parameters are sampled at the 2D coordinates identified by the body center heatmap, which the SMPL model then uses to generate the final 3D body meshes.

SMPLify [5] is an approach to automatically estimate the 3D human pose and its shape from an image using DeepCut CNN. First, it predicts the 2D positions of body joints and then, based on this 2D prediction, a 3D mesh representing the human body is generated using SMPL.

Although the aforementioned methods achieve impressive performance in both 2D and 3D human pose estimation, they are often based on single-view input. As a consequence, the lack of multi-view information leads to an inherent loss of spatial and geometric details. This information loss often results in depth ambiguities and physically inaccurate poses, as illustrated in Fig. 2.

### 2.3 Multiple View Methods

V2V-PoseNet [11] is a different approach that converts 2D depth images to 3D volumetric form instead of direct regression of 3D keypoint coordinates from 2D depth images (pixel-to-coordinates approaches). This allows for taking 3D voxelized data as an input and for estimating the per-voxel likelihood for each keypoint (voxel-to-voxel approach).

Faster VoxelPose [20] is a solution addressing the problem of the high computational cost related to the re-projection of 2D depth images to voxelized forms. Volumetric feature aggregation is used in this solution instead of 3D convolutions to improve scalability and enable real-time performance.

MotionBert [1] is an approach to 3D pose estimation, mesh generation and skeleton-based action recognition from video sequences. The model uses a motion encoder based on Dual-stream Spatio-Temporal Transformer (DSTformer) neural network, which incorporates several dual-stream fusion modules. Each module contains a multilayer perceptron (MLP) that receives input from two

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

**Fig. 3.** Example of inaccuracies of 3D pose regression with VoxelPose (trained on considered workshop with 3 camera setup having small common field of view).

branches – for spatial and for temporal Multi-Head Self-Attention (MHSA). The spatial MHSA (S-MHSA) captures relationships between different joints, while the temporal MHSA (T-MHSA) models the motion dynamics of individual joints.

The problem with various multiple-view methods is that these typically require large-scale, high-quality training datasets to achieve robust generalization. Moreover, multi-view or volumetric approaches assume sufficiently wide fields of view to capture the subject from diverse angles. In practice, limited camera coverage or narrow viewing angles may reduce reconstruction accuracy and limit the applicability of such systems in constrained environments. For example, when using VoxelPose in the scenario considered in this research, we managed to track only one person, and with poor quality, as shown in Fig.3.

## 3 Conceptual overview of the proposed approach

The proposed method combines three key components: raycasting, epipolar constraints, and physical constraints with pose blending (Fig. 4). By combining these elements, the method improves geometric accuracy, multi-view consistency, and physical realism, making it more robust to noise and missing data.

In principle, raycasting is used to project and verify geometry in 3D space. Epipolar constraints ensure consistency between multiple camera views. Physical

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

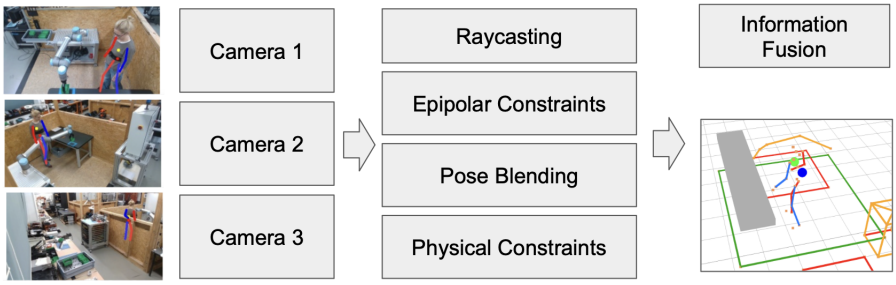


**Fig. 4.** Conceptual architecture of the proposed solution.

constraints guarantee that the estimated poses are physically correct, while pose blending provides the best limbs configurations across consecutive views.

It must be noted that in the following subsection, when discussing raycasting and 3D pose estimation, we refer specifically to the position of the human body center.

### 3.1 Raycasting Human Body Center

To bridge 2D image observations with their corresponding 3D spatial representations, we utilize a raycasting procedure that incorporates the intrinsic and extrinsic parameters of the camera. This process allows us to project a 2D point into a 3D ray in the world coordinate system, which is useful in tasks of 3D pose estimation. Particularly, in subsequent steps, this ray can be intersected with other rays to recover depth.

### 3.2 Epipolar Constraints

To recover the 3D point of human body centers, corresponding to a pair of image points $(\mathbf{u}_1, \mathbf{u}_2)$ (the yellow dots on the visualisation in Fig.6), we backproject each point into 3D space using the raycasting procedure. This yields two rays $\mathbf{r}_1(t_1) = \mathbf{o}_1 + t_1\mathbf{d}_1$ and $\mathbf{r}_2(t_2) = \mathbf{o}_2 + t_2\mathbf{d}_2$, originating from the two camera centers $\mathbf{o}_1$ and $\mathbf{o}_2$ and directed along $\mathbf{d}_1$ and $\mathbf{d}_2$, respectively.

Due to noise or imperfect calibration, these rays may not intersect exactly. Therefore, we compute the 3D point $\mathbf{P}$ as the point that minimizes the distance to both rays, often solved via the linear triangulation or midpoint method. One common approach is to solve the least-squares problem:

$$\min_{\mathbf{P}} \|\mathbf{P} - \mathbf{r}_1(t_1)\|^2 + \|\mathbf{P} - \mathbf{r}_2(t_2)\|^2$$

The solution gives an estimate of the most probable 3D location corresponding to the observed 2D projections.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

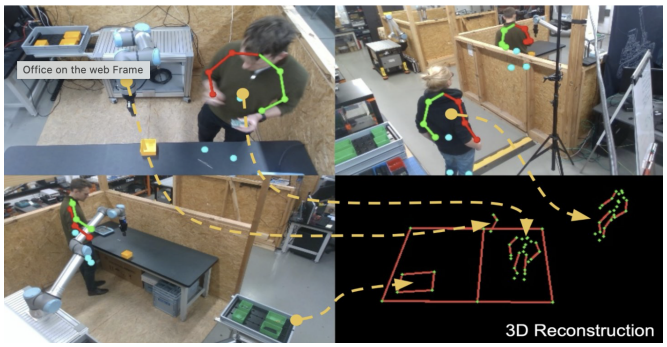


**Fig. 5.** The structure of the workshop.

### 3.3 Physical Constraints

In the approach we enforce the physical aspect of the workshop where the human-collaboration process takes place. In particular, cameras are used, each capturing footage that includes two to three people working alongside robots. The robots are labeled UR10 and AMR. Conceptually, this is shown in Fig. 5.

The recordings show stationary manipulators and human workers in one room. However, the full use case happens in a workshop with three workstations (two collaborative workstations for electrical and mechanical assembly, and one manual workstation for preparing containers), two transport areas, and a warehouse.

A human worker puts electrical and mechanical parts into containers at the manual workstation. Then, the AMR (Autonomous Mobile Robot) transports these containers to the warehouse. At the two collaborative workstations, the manipulator takes parts from a container and puts them on the assembly table. Then, the human worker assembles the parts in the required order. After the assembly is finished, the worker puts the final part or device into a container. The mobile robot then transports this container to the warehouse.

The camera and room setup creates some limits and assumptions for estimating the worker's 3D position. First, the floor is flat and level. Because of this, we can assume that the worker will not appear below this plane (or much above it). Second, the workshop area is observed by several cameras. This makes reliable triangulation possible with two or three camera views (depending on where the worker is). In contrast, the transportation area is often observed by a single camera, requiring monocular pose estimation techniques to infer the worker's position in 3D.

Addressing these constraints requires a seamless integration of multi-view and monocular tracking techniques. We must navigate the transition between high-redundancy zones and single-camera areas to ensure continuous and accurate 3D human-robot collaboration tracking.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

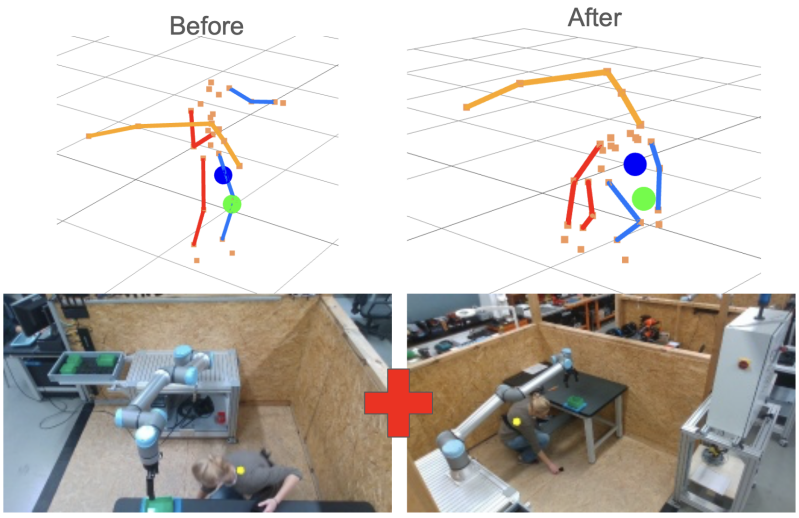


**Fig. 6.** An example of pose blending: lower limb data from the front view (left) is recovered using information from the side-view camera (right).

### 3.4 Pose blending

Besides the physical constraints of the setup (described in the previous section), the worker's lower body is often hidden by the workbench or other obstacles, such as the robotic arm. Therefore, we need to combine information from multiple camera views to realistically reconstruct the full-body pose, especially the position of the legs.

For pose blending, we replace selected SMPL model parameters $\boldsymbol{\theta}$ using values from different views of the same person. The vector $\boldsymbol{\theta}$ describes body articulation (the body pose configuration). By identifying joints that are occluded or unreliable (e.g., covered by the robotic arm or other obstacles), we selectively replace the corresponding joint parameters using estimates from another view where the joint is more clearly visible. This strategy improves the robustness and consistency of the reconstructed full-body pose. The example is shown in Fig. 6.

### 3.5 Information fusion

The task of human localization is formulated as an optimal policy selection problem. In this framework, the objective is to define a decision-making policy $\pi$ that, given a specific context $\mathcal{C}$, selects the most reliable estimation sensor (action) to minimize the expected human positioning error.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

The policy operates on a high-dimensional state space derived from the visual and geometric context of the scene. This context $\mathcal{C}$ is defined by a feature vector:

$$\mathcal{C} = \{v_0, \ldots, v_n, \text{sensor\_id}\} \tag{1}$$

where:

- $(v_0, \ldots, v_n)$ is the vector describing the context
- $sensor\_id$ specifies camera origin and detector type

The goal of the policy is to minimize the error ($L$), which represents the Euclidean distance between the estimated and actual ground-truth positions. We obtain the ground-truth from the trackers attached to the workers as described in the experiments section. Therefore, the optimal policy $\pi^*$ is defined as:

$$\pi^*(\mathcal{C}) = \arg\min_{a \in \mathcal{A}} \mathbb{E}[L \mid \mathcal{C}, a] \tag{2}$$

where $\mathcal{A}$ is the set of available estimation sensors (actions). As explained in previous section, the proposed method uses the following estimation techniques:

1. **Stereo-Triangulation Policy:** Optimal for multi-view overlap scenarios where geometric consistency is high.
2. **Monocular Image-to-World Policy:** Preferred when depth cues are ambiguous but camera intrinsics are well-calibrated.
3. **Depth-Map Integration Policy:** Preferred in scenarios with high-resolution depth information.

By training a regressor to predict the error for each method, the system effectively learns a meta-policy that switches between estimators in real-time, ensuring robust performance across diverse environmental conditions. In the experiments section, we evaluated different regression techniques.

## 4 Evaluation Dataset

For building the solution we created and used new RoHuCAD (Robots and Humans Collaborative Anomaly Detection) dataset published openly on Zenodo [6]. It is a dataset of human-robot collaboration in a robotic workshop. Two robots (collaborative manipulator - cobot, autonomous mobile robot - AMR) assist three human operators in assembly of electronic devices (Fig.7).

The dataset contains two recordings (each a few minutes long). They follow a similar scenario. The data is stored in ROS Noetic rosbag format. The dataset includes RGB-D recordings (color + depth) from three Intel RealSense D435i cameras. The data was recorded at 6 frames per second (for both color and depth). The dataset also provides camera calibration data (intrinsic parameters for each camera and extrinsic parameters with their positions and orientations in the workspace). In addition, the dataset includes spatial information about two robots used in the experiments: an AMR (Ez-Wheel SWD Starter Kit), and a collaborative robot arm (Universal Robots UR10e).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

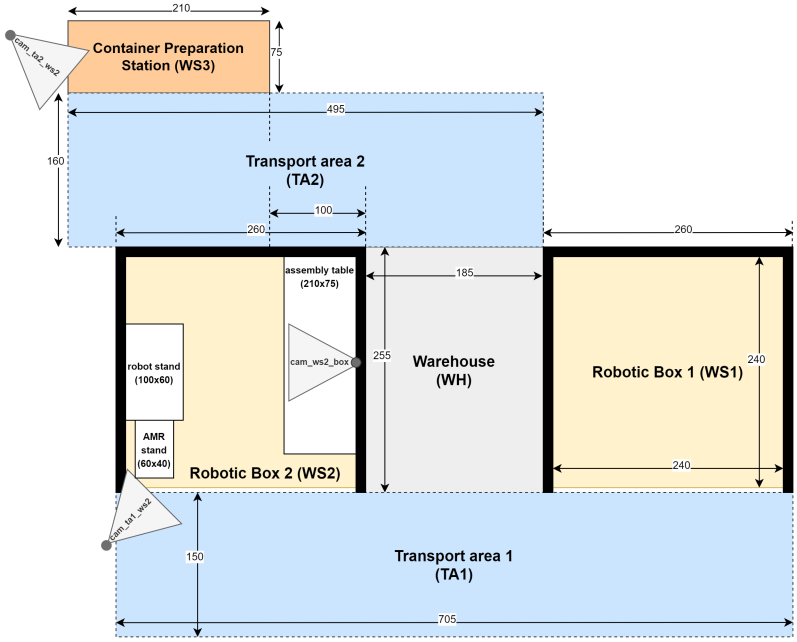


**Fig. 7.** The workshop layout.

## 5 Experiments

### 5.1 Tracking Accuracy Improvement Assessment

In the experiments, we focused on three alternative approaches for reliable pose estimation:

- The first approach, referred to as *estimated*, is based on monocular 3D pose estimation using ROMP.
- The second approach, referred to as *depth*, relies on a classical method based on depth data.
- The third approach, referred to as the *proposed* method, uses the proposed fusion algorithm reinforced with machine learning.

For technical reasons, we could not apply purely multi-view pose estimation methods such as VoxelPose, as the camera setup provides only a narrow overlapping field of view and some areas are covered by a single camera.

For the experiments, the overall layout of the workshop remained largely unchanged (when compared to RoHuCAD dataset). However, the process flow was modified (e.g. there are two people in the workshop area instead of one). Also slight adjustments were made to the camera positions to better accommodate the new tracking setup and capture improved interaction data.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

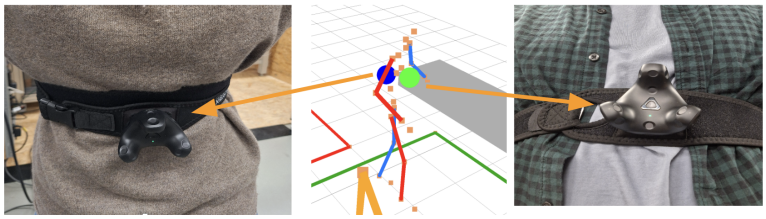

**Fig. 8.** Worker-mounted body trackers used for generating ground-truth spatial data.

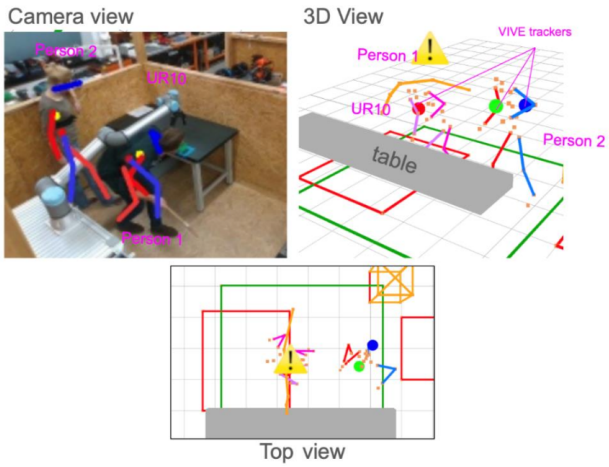


**Fig. 9.** Example of human tracking with an explanation of the symbol meanings. Person 1 has only one tracker mounted on the front (red sphere), while Person 2 has two trackers (one on the front and one on the back).

Moreover, for the experiments phase the workers were equipped with Vive [4] trackers to enable precise motion capture (Fig. 8). This setup enabled the acquisition of accurate ground-truth pose data for subsequent quantitative analysis.

During the experiment, we examined the differences between the values reported by the VIVE tracker and those returned by the proposed system. An illustrate example is shown in Fig. 9.

The three strategies for the 3D pose estimation, have been compared with the probability distribution (histogram) of the error (RMSE) for each solution. The results are shown in Fig. 10.

Ideally, the error distribution should be narrow and concentrated near zero, indicating low average error and limited variability. A left-compressed histogram reflects a system that consistently achieves high-precision predictions, with most errors falling well below 0.5 meter.

[4] https://www.vive.com/eu/accessory/tracker3/

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

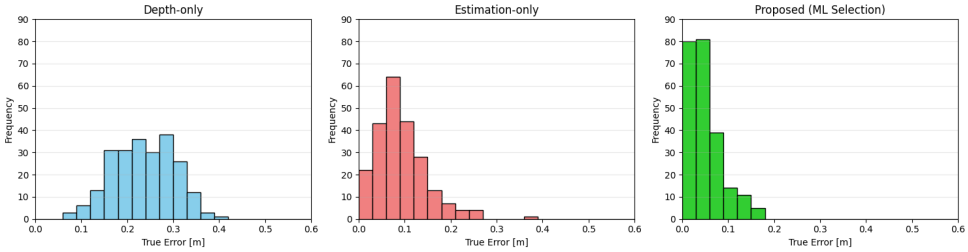


**Fig. 10.** Histogram of the prediction error between the estimated and ground-truth values (comparison of different techniques).

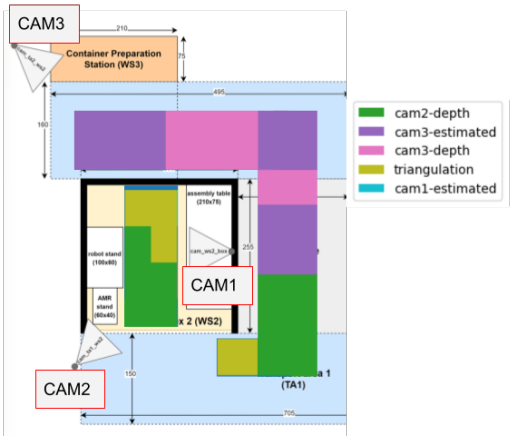


**Fig. 11.** Areas of competence for different position sensing methods overlaid on the workshop layout diagram.

The comparative diagrams illustrate that the error distributions for the proposed approach are significantly concentrated (squeezed) toward the lower end of the spectrum when compared to other techniques. As demonstrated in the results, the majority of errors for the proposed approach fall below 0.2 m, whereas the error distributions for other methods are notably more dispersed (stretched) across higher error regions.

Although the depth sensor shows many large errors, this does not mean the sensor is inaccurate. It is mainly due to the measurement conditions. As shown in the examples, the cameras are placed high above the worker. The distance is measured to the torso area (between the waist and shoulders). Because of this, large errors can appear due to following factors:

– Spatial occlusions: The target area is frequently obstructed by the worker's head, hands or equipment during tasks.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

– Varying profiles: The worker's orientation relative to the camera (standing back-to-camera or in a side profile) significantly impacts the visible surface area.
– Perspective effects: The steep downward angle of the camera increases the likelihood of depth noise when the subject's posture changes.

Moreover, Fig. 11 shows the areas where each position-sensing method works best (placed directly on the workshop layout diagram). This gives a spatial view of how the estimators perform in different work areas.

### 5.2 Meta-Policy Learning Assessment

As stated in the previous section, we formulate the tracking problem as a meta-decision task, where the system learns to predict the expected localization error for each available sensing strategy under a given context $\mathcal{C}$.

By validating the regression performance (MSE, MAE, and $R^2$), we indirectly assess the feasibility of optimal policy approximation. High predictive accuracy implies that sensor reliability is strongly context-dependent and that the learned meta-policy can effectively switch between estimation strategies in real time.

The experimental evaluation using 5×2 cross-validation demonstrates a clear distinction between linear and non-linear regression models in predicting the positioning error (Table 1). The linear baseline achieves an $R^2$ score of 0.4673, explaining less than half of the variance in the ground-truth error. Its relatively high MSE (0.2197) and MAE (0.3132) indicate that the relationship between contextual variables and localization error cannot be adequately captured by a linear model. This suggests the presence of strong non-linear dependencies and interactions between the contextual features and the reliability of the estimation sensors.

**Table 1.** 5x2 Cross-Validation Results

| Model | MSE (mean ± std) | MAE (mean ± std) | $R^2$ (mean ± std) |
|---|---|---|---|
| Linear | 0.2197 ± 0.0079 | 0.3132 ± 0.0036 | 0.4673 ± 0.0336 |
| RF | **0.0285 ± 0.0085** | **0.0829 ± 0.0053** | **0.9317 ± 0.0166** |
| GBM | 0.0310 ± 0.0062 | 0.1029 ± 0.0025 | 0.9255 ± 0.0116 |
| MLP | 0.0322 ± 0.0044 | 0.1010 ± 0.0037 | 0.9223 ± 0.0084 |

In contrast, all non-linear models significantly outperform the linear baseline. Random Forest achieves the best overall performance, with an $R^2$ of 0.9317, MSE of 0.0285, and MAE of 0.0829. This corresponds to approximately a four-fold reduction in MSE compared to the linear model, as well as an increase in explained variance. Gradient Boosting and MLP regressors achieve comparable performance, with $R^2$ scores of 0.9255 and 0.9223, respectively. The differences between these three models are marginal, indicating that the contextual feature space contains a strong and stable predictive signal.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

The relatively low standard deviations across folds further indicate that the models generalize consistently across different data splits. In particular, the Random Forest model combines high predictive accuracy with stable variance, making it a suitable candidate for deployment as the meta-policy approximator.

## 6 Conclusions

In this paper, we presented a markerless and physics-aware method for 3D human localization (designed for safe and robust human–robot collaboration with collaborative robots). By using constraints from multiple camera views, the method produces more physically correct 3D human poses. It also helps reduce problems such as unrealistic body positions and body configurations (especially under occlusion).

The proposed framework also eliminates the use of wearable sensors and naturally adapts to reconfigurable environments. Experimental results indicate improved robustness and realism of human localization, which are relevant for safety reasoning and interaction-aware robotic behaviors. The presented method provides a practical and scalable alternative for perception in human–machine collaboration scenarios.

## Acknowledgment

This work is funded under the ULTIMATE project, which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070162.

## References

1. MotionBERT: A Unified Perspective on Learning Human Motion Representations (Jul 2023), https://motionbert.github.io, [Online; accessed 9. May 2025]
2. Papers with Code - HRNet Explained (May 2025), https://paperswithcode.com/method/hrnet, [Online; accessed 9. May 2025]
3. Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., Grundmann, M.: Blazepose: On-device real-time body pose tracking. arXiv preprint arXiv:2006.10204 (2020)
4. Boesch, G.: The Complete Guide to OpenPose in 2025 - viso.ai. Viso (Oct 2024), https://viso.ai/deep-learning/openpose
5. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 561–578. Springer (2016)
6. Buś, S., Kaniuka, J., Świtlik, D., Główka, J., Kozik, R.: Rohucad: Robots and humans collaborative anomaly detection (2025). https://doi.org/10.5281/zenodo.14142968, https://doi.org/10.5281/zenodo.14142968

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39

7. eldar: deepcut (May 2025), https://github.com/eldar/deepcut, [Online; accessed 9. May 2025]
8. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. pp. 34–50. Springer (2016)
9. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: Proceedings of the IEEE international conference on computer vision. pp. 2938–2946 (2015)
10. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)
11. Moon, G., Chang, J.Y., Lee, K.M.: V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In: Proceedings of the IEEE conference on computer vision and pattern Recognition. pp. 5079–5088 (2018)
12. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. pp. 483–499. Springer (2016)
13. Odemakinde, E.: Human Pose Estimation - Ultimate Overview in 2025 - viso.ai. Viso (Dec 2024), https://viso.ai/deep-learning/pose-estimation-ultimate-overview
14. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4929–4937 (2016)
15. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5693–5703 (2019)
16. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11179–11188 (2021)
17. Sun, Y., Bao, Q., Liu, W., Fu, Y., Michael J., B., Mei, T.: Monocular, One-stage, Regression of Multiple 3D People. In: ICCV (2021)
18. tensorflow: tfjs-models (May 2025), https://github.com/tensorflow/tfjs-models/tree/master/posenet, [Online; accessed 9. May 2025]
19. Tu, H., Wang, C., Zeng, W.: Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In: European Conference on Computer Vision (ECCV) (2020)
20. Ye, H., Zhu, W., Wang, C., Wu, R., Wang, Y.: Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In: European Conference on Computer Vision. pp. 142–159. Springer (2022)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_39