m2C2: mmWave-Camera Calibration via Human Pose

Jack Shilton[1], Mariko Isogawa[1], and Ken Sakurada[2]

***Abstract*— This paper presents m2C2, a markerless calibration method for estimating the 6-DoF extrinsic transformation between mmWave radar and RGB cameras using natural human joint poses. Despite the growing demand for robust multi-sensor perception systems, calibration remains challenging, not only because traditional calibration methods rely on controlled environments and manual, target-based setups, but because mmWave radar measurements are inherently sparse and noisy, making stable geometric correspondences difficult. m2C2 addresses these limitations by utilising natural human motion as a shared calibration domain, enabling automated recalibration without dedicated calibration targets. Our framework jointly optimises the extrinsic translation and rotation by minimising an objective function over 2D and 3D human keypoints. Our approach mitigates the noise and sparsity of mmWave data through a novel bone-length scaling mechanism, jointly optimised with the extrinsics, that absorbs systematic skeleton-proportion errors from the radar pose estimator, supported by a confidence-weighted reprojection loss. Verified through experiments on a real-world dataset, m2C2 achieves an average translation error of 13.23cm and a rotation error of 1.51°, relative to the ground truth, outperforming baseline PnP methods. Sensitivity analysis further demonstrates the method's robustness by displaying comparable results across different parameters, configurations, and subjects. These results demonstrate that m2C2 provides a robust and practical alternative to traditional marker-based calibration.**

I. Introduction

Multi-modal sensors are essential components for modern robotics and autonomous driving technologies, as reliance on a single sensing modality is insufficient to ensure safety, accuracy, and robustness in complex real-world environments. The usage of data from complementary devices reduces individual sensor inaccuracies and limitations, enhancing the overall system performance and reliability. However, to effectively leverage multi-modal sensors for downstream tasks such as object detection [1], [2], precise 6-DoF extrinsic transformations are required. This is achieved through extrinsic calibration, where known features are used to estimate the spatial alignment of individual sensors relative to a shared reference frame.

Traditional extrinsic calibration methods have largely relied on specialised, known targets to create shared reference data between sensors. The foundational method utilised a printed checkerboard pattern for intrinsic and extrinsic RGB camera calibration [3]. Since then, target-based approaches have been extensively studied for cross-modal sensor calibration, particularly for RGB–LiDAR systems [4], [5], [6], [7], [8]. These approaches rely on explicit calibration targets and controlled environments to establish geometric correspondences.

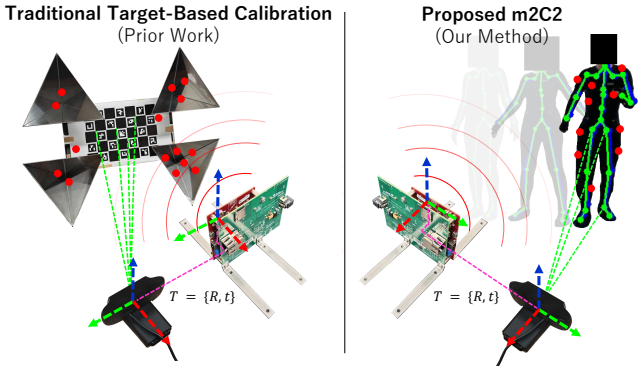

Fig. 1. Conventional target-based calibration (left) versus the proposed m2C2 framework (right). A specialised calibration board featuring a ChArUco pattern for the RGB camera and four corner reflectors for the mmWave radar are typically required in prior work. Our m2C2 method eliminates the need for these physical targets by using natural human joint poses as a shared domain for automated, target-free sensor alignment.

More recently, target-based calibration strategies have also been extended to emerging sensor modalities, such as event-based cameras [9], [10]. Alternative calibration board designs have been proposed to accommodate diverse sensing characteristics [11], [12]. However, despite these advances, such approaches generally require structured environments and manually prepared targets, limiting their applicability in dynamic, in-the-wild systems.

In this work, we address the challenge of extrinsic calibration between mmWave radar and RGB cameras. While calibration pipelines for LiDAR [13], [14] and event-based cameras [9], [15] are well-established, calibration involving mmWave radar remains a relatively underexplored subject. mmWave radars provide robust depth estimation in low-light environments and through partial occlusions (such as fog or smoke), complementing the higher spatial resolution and detailed colour information of RGB cameras. Although calibration with carefully designed radar reflectors or markers is feasible under controlled settings, such setups are often impractical in dynamic environments. More importantly, the sensing characteristics of mmWave radar itself pose a fundamental challenge for reliable mmWave–RGB alignment.

Unlike LiDAR, mmWave radar measurements are inherently sparse, noisy, and irregularly distributed, with limited spatial resolution and significant measurement uncertainty. RGB cameras capture dense 2D visual pixels, whereas mmWave radar produces extremely sparse 3D reflections that can be unstable across frames. This extreme sparsity and high noise level make stable cross-modal correspondence estimation difficult in natural scenes, and consequently con-

[1]Keio University, Japan shilton.jack@keio.jp; mariko.isogawa@keio.jp
[2]Kyoto University, Japan sakurada@i.kyoto-u.ac.jp

ventional target-based pipelines can become unreliable for accurate mmWave–RGB extrinsic calibration.

To address this, we propose **m2C2**, a novel, target-free 6-DoF extrinsic transformation estimation method that utilises natural human poses as a shared calibration domain. By leveraging 2D human joint positions and 3D Skinned Multi-Person Linear Model (SMPL) [16] keypoints extracted from sensor data, our approach removes the need for specialised calibration equipment and enables calibration in unconstrained environments. Specifically, we formulate calibration as an optimisation problem designed to minimise the reprojection error between corresponding human joint keypoints across modalities. This ensures a robust, accurate, and fully automated estimation of the extrinsic parameters between the mmWave radar and RGB camera.

Our primary contributions are summarised as follows:

1) We introduce the first human-joint-based extrinsic calibration method for mmWave–RGB sensor pairs, enabling target-free alignment in natural environments.
2) We propose a novel, jointly-optimised per-bone length scaling mechanism that absorbs the systematic skeleton-proportion errors of the radar pose estimator, which would otherwise be compensated by an inaccurate camera pose, supported by a confidence-weighted reprojection loss to reduce the influence of unreliable joints.
3) We evaluate m2C2 on real-world datasets, demonstrating its accuracy and establishing its viability as a practical alternative to traditional, marker-based calibration pipelines.

II. Related Work

This section reviews the related literature, across four categories: (1) Target-Based Calibration, (2) Target-Less Calibration, (3) Human-Pose-Driven Calibration, and (4) mmWave Calibration. We highlight key contributions within each category and position our method according to these contributions.

A. Target-Based Calibration

Target-based calibration relies on physical calibration objects with known geometry to determine correspondences between sensor measurements and real-world coordinates.

Zhang [3] established the standard checkerboard-based method for camera calibration, computing intrinsic and extrinsic parameters via homographies refined by maximum likelihood estimation. However, the method requires colour or intensity gradient information, limiting its use in multi-modal systems such as LiDAR and mmWave radar.

Zhang's work was further adapted for cross-modal calibration, most notably between RGB cameras and LiDAR. Early approaches by Zhang and Pless [4] and Unnikrishnan and Hebert [5] established the initial multi-modal methodology by introducing planar constraints to match LiDAR points to the checkerboard plane. Wang et al. [8] improved upon these methods by introducing the reflectance intensity of the printed checkerboard as additional information. Building upon geometric feature extraction, Zhou et al. [17] developed an automatic calibration method utilising line and plane correspondences. These methodologies were implemented across various other sensors, such as omnidirectional cameras [6], solid-state LiDAR [18], and event-based cameras [10].

Despite these advancements, standard printed checkerboards were insufficient for sensors without visual or intensity data. To resolve this, specialised multi-modal calibration boards with distinct geometric features were introduced. Methods by Beltrán et al. [11] and Yan et al. [12] utilise custom calibration boards with cutouts, such as circular holes, alongside the traditional checkerboard patterns. By segmenting these physical markers, these approaches create reliable reference points for calibration.

However, multi-modal calibration targets are typically large and require controlled environments for accurate calibration. This restricts their usage in practical, in-the-wild scenarios. To address these limitations, our method removes the need for physical calibration boards altogether. Instead, we utilise the human body as a natural calibration target, leveraging human joints as shared markers to achieve robust multi-modal sensor alignment from natural motion.

B. Target-Less Calibration

Target-less calibration methods eliminate the need for specialised physical targets by extracting correspondences directly from the environment. Early approaches estimated extrinsic parameters by maximising the mutual information between LiDAR reflectivity and camera image intensity [19], [20], providing a foundation for alignment without explicit geometric feature extraction and enabling early online camera–laser frameworks [21]. Subsequent methods improved accuracy by exploiting richer structural and semantic cues, such as semantic segmentation [22] and natural line features [23].

Learning-based approaches, building on self-supervised interest point detection [24], instead formulate calibration as a regression problem, with RegNet [25] and CalibNet [26] predicting extrinsic parameters directly via supervised 3D spatial transformer networks. More recent work adopts richer scene representations: Zhou et al. [27] align LiDAR poses by jointly learning neural density and colour fields, while Calib-Anything [28] removes environment-specific training by segmenting camera images and assigning LiDAR points to segmented classes via plane normal estimation. These methods target dense modalities, however, and do not transfer to the sparse, noisy returns of mmWave radar.

C. Human-Pose-Driven Calibration

Human-pose-driven calibration methods are a subfield of target-less calibration methods that perform calibration using human joints as calibration markers. These methods are particularly advantageous for multi-view calibration tasks due to being detectable from a large variety of viewpoints.

Takahashi et al. [29] first introduced this concept by detecting 14 distinct 2D keypoints from multiple unsynchronised, uncalibrated RGB cameras. These joints serve as common reference points across views, enabling synchronisation

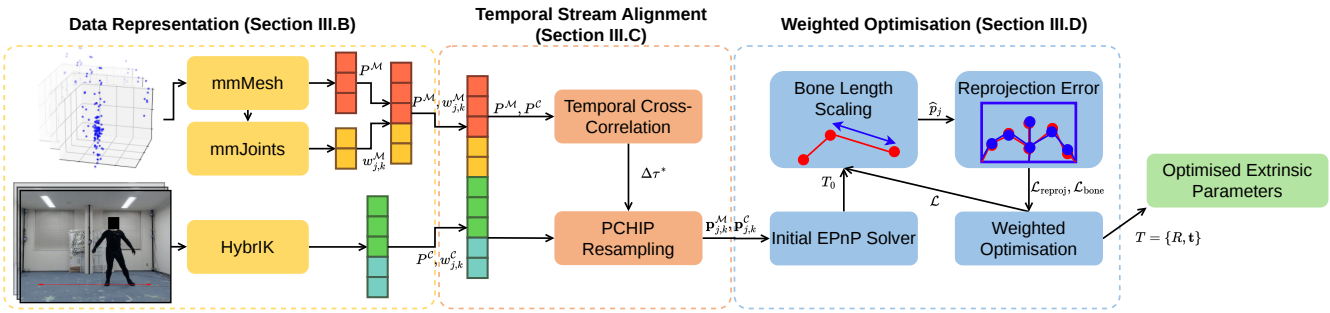

Fig. 2. Overview of the m2C2 framework. Raw data from mmWave radar and RGB cameras are processed independently to extract 3D and 2D SMPL keypoints. The streams are temporally aligned using motion signature cross-correlation and PCHIP Resampling. Finally, the extrinsic parameters (R,t) are estimated by minimising the confidence-weighted reprojection error, accounting for optimisable bone-length scaling.

and extrinsic calibration through bundle adjustment [30]. Their method optimises both camera parameters and joint positions by minimising joint reprojection errors, solving a Perspective-n-point (PnP) problem without the need for dedicated calibration objects.

Xu et al. [31] explore the feasibility of calculating point correspondences over time from the centroid of multiple segmented human subjects in a given scene. The method extracts person bounding boxes across multiple camera views and converts the bounding box correspondences to point correspondences, which act as input to a pose solver for each camera. This method allows for easier identification of human subjects at range in larger environments.

Liu et al. [32] extended joint-based calibration by proposing an automatic calibration method for both extrinsic and intrinsic parameters. Liu et al.'s approach introduces a novel loss function, based on the reprojection error between estimated 3D joints and detected 2D joint positions, allowing for the optimisation of multi-camera systems without traditional calibration targets.

Lee et al. [33] expanded the framework by Takahashi et al. [29], optimising with regards to the trajectories of orientated 3D joints, rather than static 2D joint positions alone. Their trajectory-based method significantly improved accuracy for both real and synthetic data over previous SOTA methods.

These methods demonstrate the effectiveness of using the human pose as a calibration target, establishing a strong basis for extending this methodology to multi-modal sensor systems.

D. mmWave Calibration

Calibration methods involving mmWave radar are a novel field compared to RGB or LiDAR calibration, largely due to the inherent sparsity and noise of radar point clouds. Early research by Sugimoto et al. [34] relied on the identification of corner reflector calibration equipment within the camera frame by correlating image coordinates with maximum reflection intensity peaks, allowing for calibration using a 2D homography matrix.

Natour et al. [35] extend this method to use three corner reflectors with a known euclidean distance for calculating the depth within the camera image to optimise the extrinsic parameters of the sensors.

Schöller et al. [36] presents a targetless implementation that utilises a two-stream convolution neural network to associate the location of moving vehicles in a motorway scene in both mmWave and camera data streams.

Wise et al. [37] presents a method that removes the need for moving vehicles by estimating the velocity of the radar, alongside known pose changes of a sensor with a rigid body connection to align these velocity estimates to estimate the extrinsic parameters. This approach removes the need for any calibration target to be present in the environment.

Most recently, Cheng and Cao [38] proposed an online target-less calibration network using classes detected using YOLO from radar data and camera images. These features are passed to a discriminator that identifies if both features correspond to the same object. The reprojection error of matching features is then minimised to solve the extrinsic parameters.

However, these methods rely on consistent detection of multiple features over extended periods of time, limiting their usability in real-world scenarios. In contrast, our proposed method requires only a single human target. By extracting 22 unique keypoints from the human body, our method provides a dense, reliable set of correspondences that can be used directly for parameter optimisation.

III. Methodology

A. Overview

We present a framework for estimating the extrinsic parameters T between sensors with small temporal offsets, leveraging SMPL joints as a common representation. Therefore, while this work focuses on mmWave radar and RGB camera, the methodology generalises to any sensor modality capable of human body pose estimation.

Let $T = \{R, \mathbf{t}\}$ denote the rigid-body transformation, consisting of a rotation matrix $R \in \mathrm{SO}(3)$ and translation vector $\mathbf{t} \in \mathbb{R}^3$. T represents the transformation from the mmWave radar frame $\{\mathcal{M}\}$ to the camera frame $\{\mathcal{C}\}$. Given a point $P_{\mathcal{M}}$ in the radar coordinate system, the transformation to map the point to $P_{\mathcal{C}}$ is calculated as:

$$P_{\mathcal{C}} = RP_{\mathcal{M}} + \mathbf{t} \tag{1}$$

B. Sensor Data Representation

To perform multi-modal calibration, we convert the raw input data from each sensor into a common SMPL model skeleton and estimate corresponding confidence scores for each joint. Each output is truncated to $N = 22$ to maintain a one-to-one mapping across all sensor types.

1) mmWave Radar: To estimate SMPL skeleton data from mmWave data, we leverage the *mmMesh*[39] architecture to estimate SMPL joint positions directly from the point cloud data. We extend the output using the *mmJoints*[40] module to provide a per-joint confidence score.

Let $P_k^{\mathcal{M}} = \{\mathbf{p}_{j,k}^{\mathcal{M}}\}_{j=1}^{N}$ be the set of 3D joint positions in the radar frame at time k. For each joint j, the mmJoints module estimates a normalised sensing score $\alpha_{j,k} \in [0, 1]$ and reliability score $r_{j,k} \in [0, 1]$. The combined confidence score is defined as:

$$w_{j,k}^{\mathcal{M}} = \alpha_{j,k} \cdot r_{j,k} \tag{2}$$

2) RGB Camera: For our work, intrinsic parameters of the RGB camera K are assumed to be known before calibration. This calibration is performed using a target-based approach [3] using a 11x8 checkerboard calibration pattern with 20mm squares.

We extract 2D SMPL joint positions within the camera frame using *HybrIK* [41], a deep-learning inverse kinematics method. During inference, the model produces a heatmap for each joint j at frame k, where the confidence score $w_{j,k}^{\mathcal{C}} \in [0, 1]$ is calculated based on the peak value of each heatmap.

C. Temporal Stream Alignment

Due to variances in latency between the mmWave radar and RGB camera data streams, a sub-frame temporal offset is present that must be estimated and corrected prior to calibration. Let $\Delta\tau$ denote the temporal offset (in seconds) between the data streams.

For a discrete sequence of frames indexed by k, the corresponding camera frame $f^{\mathcal{C}}$ from the radar frame $f^{\mathcal{M}}$ can be calculated as:

$$f^{\mathcal{C}}(\tau_k) = f^{\mathcal{M}}(\tau_k + \Delta\tau) \tag{3}$$

where τ_k represents the timestamp for the k-th frame. The goal of temporal alignment is to accurately estimate $\Delta\tau^*$ to resample the streams to a common time-frame.

1) Motion Signature Extraction: Direct comparison of joint positions between the radar and camera streams is unreliable due to the difference in dimensionality and coordinate systems. However, the velocity of robust joints displays a strong correlation when properly aligned. We first calculate the discrete motion signature $\sigma[k]$ for frame index k, defined as the median speed across a selection of robust joints $\mathcal{J} = \{0, 7, 8, 12, 16, 17\}$ (Pelvis, Left Ankle, Right Ankle, Neck, Left Shoulder, Right Shoulder):

$$\sigma^{\text{3D}}[k] = \text{median}_{j\in\mathcal{J}} \left\|\mathbf{p}_{j,k}^{\mathcal{M}} - \mathbf{p}_{j,k-1}^{\mathcal{M}}\right\|_2 \tag{4}$$

$$\sigma^{\text{2D}}[k] = \text{median}_{j\in\mathcal{J}} \left\|\mathbf{p}_{j,k}^{\mathcal{C}} - \mathbf{p}_{j,k-1}^{\mathcal{C}}\right\|_2 \tag{5}$$

where $\mathbf{p}_{j,k}^{\mathcal{M}}$ and $\mathbf{p}_{j,k}^{\mathcal{C}}$ denote the position of joint j at time k in the radar and camera streams, respectively. Z-score normalisation is applied using the median absolute deviation (MAD), calculated across all frames of each motion signature.

$$\tilde{\sigma}[k] = \frac{\sigma[k] - \text{median}(\sigma)}{1.4826 \cdot \text{MAD}(\sigma)} \tag{6}$$

2) Offset Estimation: Because the data streams may possess sub-frame temporal misalignment ($\Delta\tau \in \mathbb{R}$), cross-correlation cannot be performed with discrete indices. To evaluate the offset continuously, we define continuous motion signature functions $\tilde{\sigma}^{\text{3D}}(\tau)$ and $\tilde{\sigma}^{\text{2D}}(\tau)$ by linearly interpolating the discrete normalised sequences $\tilde{\sigma}^{\text{3D}}[k]$ and $\tilde{\sigma}^{\text{2D}}[k]$ over their respective timestamps $\tau_k^{\mathcal{M}}$ and $\tau_k^{\mathcal{C}}$.

Using these continuous representations, the optimal temporal offset $\Delta\tau^*$ is found by maximising the cross-correlation over a given continuous search window $[-w, w]$:

$$\Delta\tau^* = \underset{\Delta\tau\in[-w,w]}{\arg\max} \sum_k \tilde{\sigma}^{\text{3D}}(\tau_k) \cdot \tilde{\sigma}^{\text{2D}}(\tau_k + \Delta\tau) \tag{7}$$

Once the optimal continuous offset $\Delta\tau^*$ is estimated, the RGB frame timestamps are adjusted to match the radar timestamps.

$$\tau'_{\mathcal{C}} = \tau_{\mathcal{C}} + \Delta\tau^* \tag{8}$$

The radar data is then resampled at these timestamps $\tau'_{\mathcal{C}}$, using piecewise cubic Hermite interpolation (PCHIP) to produce temporally aligned streams.

D. Pose Optimisation

Following temporal alignment, the extrinsic parameters $T = \{R, \mathbf{t}\}$ are estimated via a confidence-aware Perspective-n-Point (PnP) optimisation framework, using the Efficient PnP (EPnP) solver to obtain an initial closed-form estimate of the extrinsic parameters[42]. Our approach addresses the inherent difference between radar and camera skeletons through two distinct factors: (1) confidence-weighted reprojection, and (2) bone length scaling.

1) Problem Formulation: Given temporally aligned 3D joint positions $P^{\mathcal{M}} = \{\mathbf{p}_{j,k}^{\mathcal{M}}\}$ from the radar and 2D joint positions $P^{\mathcal{C}} = \{\mathbf{p}_{j,k}^{\mathcal{C}}\}$ from the camera, with confidence scores $w_{j,k}^{\mathcal{M}}$ and $w_{j,k}^{\mathcal{C}}$ respectively, we formulate the calibration as a weighted reprojection error minimisation problem.

For each joint j at frame k, we define the reprojection error as:

$$e_{j,k} = \left\|\pi(T, \mathbf{p}_{j,k}^{\mathcal{M}}) - \mathbf{p}_{j,k}^{\mathcal{C}}\right\|_2 \tag{9}$$

where $\pi(\,)$ represents the camera projection function. The reprojection loss aggregates these errors over all joints and frames:

$$\mathcal{L}_{\text{reproj}} = \sum_k \sum_j w_{j,k} \cdot \rho(e_{j,k}), \tag{10}$$

where $w_{j,k}$ is the combined confidence weight of both sensors (Eq. 11), and $\rho(\,)$ is a robust loss function. Calibration is performed by minimising $\mathcal{L}_{\text{reproj}}$ with respect to T, extended in Sec. III-D.4 to jointly optimise the bone-length scale factors.

2) Confidence-Weighted Reprojection: To reduce the influence of inaccurate joint estimates within the optimisation, we combine confidence scores from both sensors. For each joint j at frame k, the combined weight is calculated as:

$$w_{j,k} = \sqrt{\bar{w}^{\mathcal{M}}_{j,k} \cdot \bar{w}^{\mathcal{C}}_{j,k}} \tag{11}$$

where $\bar{w}^{\mathcal{M}}_{j,k}$ and $\bar{w}^{\mathcal{C}}_{j,k}$ represent the min-max normalised confidence scores for the radar and camera sensors respectively:

$$\bar{w} = \frac{w - w_{\min}}{w_{\max} - w_{\min} + \epsilon} \tag{12}$$

This ensures that low confidence values in either modality are appropriately considered during optimisation.

3) Bone Length Scaling: Due to fundamental differences in how mmMesh and HybrIK estimate SMPL skeletons, an error in the bone length exists between the two representations. Rather than applying a 3D offset, we introduce an optimisable per-bone scale factor $\{s_j\}_{j=1}^{N}$ that preserves the joint angles while correcting bone lengths.

For a bone chain defined by parent indices $\text{pa}(j)$, the scaled joint position $\hat{\mathbf{p}}_j$ is computed recursively:

$$\hat{\mathbf{p}}_j = \hat{\mathbf{p}}_{\text{pa}(j)} + s_j \cdot \left(\mathbf{p}_j - \mathbf{p}_{\text{pa}(j)}\right) \tag{13}$$

where the root joint ($j = 0$) remains unchanged. To ensure positive scale factors, we define $s_j = \exp(\lambda_j)$ where λ_j are learnable parameters, initialised as zero.

4) Optimisation: Optimisation is defined in two stages. The initial stage is an estimate obtained using RANSAC-based EPnP [42] for each joint pair across all frames.

$$T_0 = \text{PnP}_{\text{RANSAC}}(P^{\mathcal{M}}, P^{\mathcal{C}}, K) \tag{14}$$

where K is the intrinsic parameters of the camera, provided through standard checkerboard calibration [3].

The robust kernel ρ is the Huber loss, with threshold δ set per sequence from the confidence-weighted median reprojection residual of the initialisation.

The second stage refines T_0 jointly with the bone-length scale factors, combining the reprojection loss of Eq. 10 with a regularisation term that penalises deviation of the log-scale factors from zero (from the estimator's original bone lengths):

$$\mathcal{L}_{\text{bone}} = \sum_{j=1}^{N} \lambda_j^2 \tag{15}$$

where λ_j are the log-scale parameters of Eq. 13. The refined parameters are obtained by minimising the combined objective over both the extrinsics and the scale factors:

$$T^*, \{\lambda_j^*\} = \underset{T,\{\lambda_j\}}{\arg\min}\, \mathcal{L}_{\text{reproj}} + \lambda_{\text{reg}} \mathcal{L}_{\text{bone}} \tag{16}$$

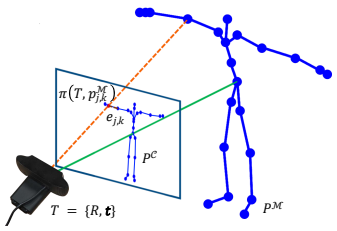

Fig. 3. Visualisation of the reprojection error minimisation. The 3D mmWave radar keypoints ($P^{\mathcal{M}}$, red) are projected onto the 2D image plane using the estimated extrinsic parameters T. The calibration framework optimises T by minimising the Euclidean distance (reprojection error $e_{j,k}$) between these projected points $\pi(T, p^{\mathcal{M}}_{j,k})$ and the corresponding ground truth 2D RGB camera keypoints ($P^{\mathcal{C}}$, blue).

where $\lambda_{\text{reg}} = 10^{-1}$ unless otherwise stated, controlling the strength of the regularisation to prevent preference of bone scaling over accurate reprojection.

IV. Evaluation

A. Dataset

To evaluate the effectiveness of the m2C2 framework, we collected a real-world dataset comprising five subjects (four male, one female) with diverse body types. To assess the robustness of the calibration, each subject performed two unique motion sequences across four distinct sensor configurations. These configurations involved varying the horizontal and vertical offsets, as well as the rotation between the mmWave radar and the RGB camera.

For each configuration, ground-truth camera poses were established using a 5×5 ChArUco board [43] ($\approx$45$\times$45 cm, 90 mm squares, `DICT_6X6_250`) captured at 1920$\times$1080 in a dedicated calibration take. The camera was static between the calibration take and the corresponding record take, so each sequence inherits its calibration take's ground-truth pose. Ground-truth pose accuracy, measured by Mocap-to-RGB joint PnP residual, has a median of 1.9 cm across all sequences.

Each recorded sequence consists of 3,000 frames ($\approx$5 min at 10 FPS), with the evaluation stream at 640$\times$480. Subjects were positioned 1.6–5.0 m from the sensors and performed lateral (side-to-side) and radial (forward-back) walking motions across two camera heights, where inter-stream temporal offsets were small ($|\Delta\tau| \leq 0.3$ s).

B. Evaluation Metrics

To evaluate our method, we compare the estimated transformation T^* against the ground truth T_{gt} obtained by the target-based calibration board. We employ the following metrics:

- Mean Translation Error (E_t): The Euclidean distance between the estimated translation vector t^* and the ground truth t_{gt}, measured in centimetres.

$$E_t = \|t^* - t_{gt}\|_2 \tag{17}$$

- Mean Rotation Error (E_R): The angular difference between the estimated rotation R^* and the ground truth

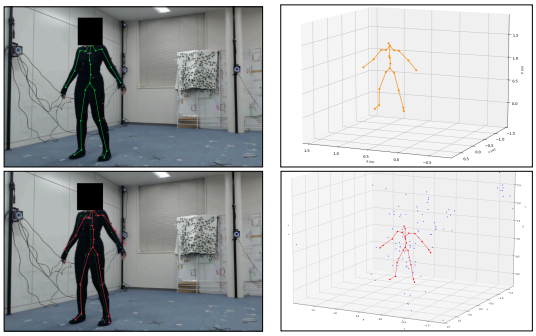

Fig. 4. Visualisation of the output from the m2C2 pipeline. Showing the HybrIK estimated pose (Top Left), mmMesh estimated pose (Top Right), calculated offset projection (Bottom Left) and raw mmWave data (Bottom Right).

R_{gt}, measured in degrees.

$$E_R = \arccos\left(\frac{Tr(R^* R_{gt}^T) - 1}{2}\right) \tag{18}$$

- Mean Reprojection Error (MRE): The average Euclidean distance in pixels between the projected 3D mmWave joints and the detected 2D RGB joints across all frames.

V. Results

Ablation study. Table I summarises the ablation results. The full m2C2 framework outperforms the baseline PnP method, reducing translation error from 23.72 cm to 13.23 cm and rotation error from 2.48° to 1.51°. Bone-length scaling is the dominant component, accounting for 4.90 cm of the translation error reduction. Temporal alignment contributes 0.98 cm to translation error but 2.88 px to reprojection error, reflecting the small inter-stream offsets of this dataset ($|\Delta\tau| \leq 0.3$ s). Confidence weighting contributes 0.08 cm.

Regularisation sensitivity. Table III reports accuracy across the bone-length regularisation weight λ_{reg}, alongside the mean deviation of the scaled bone lengths. Translation error remains within 13.04–13.23 cm for $\lambda_{reg} \leq 10^{-1}$, and degrades to 17.25 cm at $\lambda_{reg} = 10^{1}$, approaching the "w/o bone scaling" ablation as the scale factors are constrained towards unity.

Frame count sensitivity. Figure 5 shows the effect of the number of sampled frames on calibration accuracy. We vary the number of frames to evaluate robustness under shorter motion clips. The accuracy improves steeply up to 150 frames and continues to improve gradually until approximately 500 frames, after which it largely saturates. This suggests that a moderate number of frames is sufficient for reliable calibration.

Subject-wise evaluation. Table IV reports the calibration accuracy for each subject individually. The results are relatively consistent across subjects, although some variation remains depending on the individual motion and body shape.

TABLE I
ABLATION STUDY RESULTS AVERAGED OVER 20 SEQUENCES. E_t: TRANSLATION ERROR, E_R: ROTATION ERROR, MRE: MEAN REPROJECTION ERROR.
BOLD/UNDERLINE: BEST/SECOND BEST.

Method	E_t (cm) ↓	E_R (°) ↓	MRE (px) ↓
Baseline (PnP)	23.72	2.48	9.05
Baseline (PnP) + Time Shift	23.32	2.39	6.36
m2C2 (Full)	**13.23**	**1.51**	**5.41**
w/o Time Shift	14.21	1.54	8.29
w/o Conf. Weighting	13.31	1.55	**5.41**
w/o Bone Scaling	18.13	1.95	6.36
w/o Conf. + Bone	18.28	1.97	6.35
w/o Time + Bone	18.45	2.03	9.03

TABLE II
MEAN CALIBRATION ERROR ACROSS FOUR DISTINCT OFFSET CONFIGURATIONS.

Configuration	E_t (cm) ↓	E_R (°) ↓	MRE (px) ↓
Horizontal Offset (L/R)	15.14	1.74	7.60
Horizontal Offset (F/B)	11.27	0.58	4.28
Vertical Offset	11.31	0.42	4.21
Vertical Offset + Rotation	14.82	3.13	5.33

VI. Discussion

Calibration accuracy under mmWave sensing. The results demonstrate that m2C2 achieves competitive calibration accuracy without the need for specialised physical targets. Specifically, the observed translation error (E_t) of 13.23 cm and rotation error (E_R) of 1.51° are consistent with the inherent sensing limitations of mmWave radar. Compared with LiDAR or RGB sensors, mmWave radar measurements are sparse and irregular, with significant uncertainty in spatial resolution. In our setup, the radar provides an approximate range resolution of 4 cm, with angular resolutions of about 15° in the azimuth plane and 30° in the elevation plane. Considering these constraints, the obtained translation accuracy indicates that the proposed method can reliably estimate the relative sensor offset.

Effect of the shared SMPL representation. Our framework mitigates the sparsity and noise of mmWave measurements by transforming the observations into a shared representation based on SMPL keypoints. This representation allows geometric correspondence to be established between radar and camera modalities despite the large domain gap. The shared representation also proves insensitive to capture conditions: accuracy is essentially unchanged when subjects repeat takes in normal clothing rather than instrumented suits (mean E_t 11.27 cm vs. 11.83 cm over four matched pairs), indicating the method does not depend on specialised capture attire. However, the calibration accuracy also depends on the reliability of the upstream pose estimators. For example, Xue et al. [39] reports that the mmMesh estimator introduces an average joint position error of approximately 2.18 cm. Therefore, improvements in radar-based human pose estimation are expected to further enhance the performance of the proposed

TABLE III

SENSITIVITY ANALYSIS OF THE REGULARISATION PARAMETER λ_{reg} ON CALIBRATION ACCURACY. BONE DEV.: MEAN ABSOLUTE DEVIATION OF THE OPTIMISED BONE LENGTHS FROM THE ESTIMATOR'S ORIGINAL LENGTHS.

λ_{reg}	E_t (cm) ↓	E_R (°) ↓	MRE (px) ↓	Bone Dev. (%) ↓
10^{-4}	13.04	1.51	5.41	16.93
10^{-3}	13.04	1.51	5.41	16.92
10^{-2}	13.06	1.51	5.41	16.81
10^{-1}	13.23	1.51	5.41	15.96
10^{0}	14.29	1.58	5.45	11.95
10^{1}	17.25	1.77	5.74	5.81

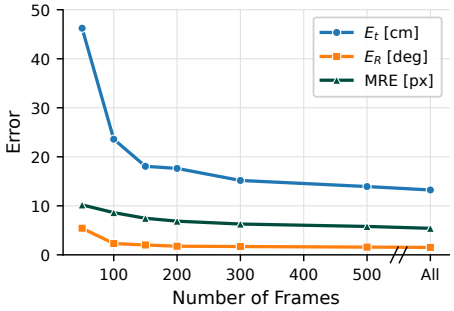

Fig. 5. Sensitivity analysis of the number of sampled frames on calibration accuracy. The plot illustrates the relationship between the amount of input data and the resulting mean translation error (E_t), mean rotation error (E_R), and mean reprojection error (MRE). All three metrics improve steeply up to approximately 150 frames and continue to improve gradually until around 500 frames, after which they largely saturate, indicating that the framework remains robust while benefiting from longer motion sequences.

calibration framework. The sensitivity analyses of Section V further show that the benefit of bone-length scaling derives from permitting scale adaptation at all, rather than from a precisely tuned prior.

Comparison with existing targetless calibration methods. Compared with previous approaches such as Schöller et al. [36] and Wise et al. [37], our method removes the requirement for sensor motion or vehicle trajectories, enabling calibration for static sensor installations. While recent work by Cheng and Cao [38] relies on YOLO-based feature matching across multiple objects over time, m2C2 instead utilises 22 human joint keypoints extracted from a single subject. Cheng and Cao report only a mean reprojection error (18.83 px); as reprojection error in pixels depends on image resolution and scene geometry, it is not directly comparable to ours. In contrast, we additionally report metric translation and rotation error against target-based ground truth, which prior mmWave–RGB targetless methods do not.

Limitations and failure modes. Our evaluation is limited to a controlled indoor space with fixed sensor placement, five subjects, and upright walking motions only, generalisation to other environments, ranges, motion types, and multiple simultaneous subjects is untested. The method assumes a single person, as the upstream estimators are single-subject, and calibration accuracy is bounded by their reliability.

The reprojection objective generally prefers a camera displaced in depth. This depth–scale ambiguity is only resolved when the subject's motion contains variation along the viewing angle. Purely lateral motion provides insufficient information, which is the main failure condition of the method and is consistent with the higher errors observed for lateral and rotated configurations (Table II).

TABLE IV

CALIBRATION ACCURACY BY SUBJECT.

Subject	E_t (cm) ↓	E_R (°) ↓	MRE (px) ↓
Subject A	13.77	1.14	4.16
Subject B	14.87	0.84	4.53
Subject C	17.46	2.81	5.35
Subject D	10.57	2.39	7.79
Subject E	10.55	0.72	5.21

VII. CONCLUSION

m2C2 introduces a novel target-free multi-modal sensor calibration framework between mmWave radar and RGB cameras. Our core contributions include the formulation of extrinsic calibration as a loss minimisation problem, utilising SMPL keypoints as a common domain, a confidence-weighted reprojection loss, and bone-length scaling.

Our framework solves three core problems in multi-modal calibration: (1) temporal misalignment between sensors, solved using motion-signature cross-correlation and sub-frame resampling using PCHIP, (2) varying joint reliability across modalities, addressed using confidence estimation and weighting, and (3) skeleton bone length variance between pose estimators, corrected with optimised bone scaling.

Through experimental analysis, we demonstrate the effectiveness of our approach, achieving an average translation error of 13.23 cm and a rotation error of 1.51°, without the need for specialised calibration markers. Ablation studies present the significance of each component in reducing each evaluation metric.

Future works should evaluate the feasibility of this methodology as a generalised method for multi-modal sensors, such as LiDAR and event cameras, replacing the current pose estimators with an SMPL estimator for the target sensor. Additionally, improving the framework to allow for online calibration should be considered.

A. Acknowledgements

This work was partially supported by JSPS KAKENHI (B) 26K02929.

REFERENCES

[1] J.-j. Lin, J.-i. Guo, V. M. Shivanna, and S.-y. Chang, "Deep learning derived object detection and tracking technology based on sensor fusion of millimeter-wave radar/video and its application on embedded systems," *Sensors*, vol. 23, no. 5, p. 2746, 2023.

[2] K. Shi, Z. Shi, A. Chen, Z. Xiong, J. Chen, and J. Luo, "Radar and camera fusion for object detection and tracking: a comprehensive survey," *IEEE Communications Surveys & Tutorials*, 2024.

[3] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 22, no. 11, pp. 1330–1334, 2000.

[4] Q. Zhang and R. Pless, "Extrinsic calibration of a camera and laser range finder (improves camera calibration)," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 3, 2004, pp. 2301–2306.

[5] R. Unnikrishnan and M. Hebert, "Fast extrinsic calibration of a laser rangefinder to a camera," Carnegie Mellon University, Tech. Rep., 2005.

[6] G. Pandey, J. Mcbride, S. Savarese, and R. Eustice, "Extrinsic calibration of a 3d laser scanner and an omnidirectional camera," *IFAC Proceedings Volumes*, vol. 43, no. 16, pp. 336–341, 2010.

[7] A. Geiger, F. Moosmann, O. Car, and B. Schuster, "Automatic camera and range sensor calibration using a single shot," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2012, pp. 3936–3943.

[8] W. Wang, K. Sakurada, and N. Kawaguchi, "Reflectance intensity assisted automatic and accurate extrinsic calibration of 3d lidar and panoramic camera using a printed chessboard," *Remote Sensing*, vol. 9, no. 8, 2017.

[9] M. Muglikar, M. Gehrig, D. Gehrig, and D. Scaramuzza, "How to calibrate your event camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021, pp. 1403–1409.

[10] J. Jiao, F. Chen, H. Wei, J. Wu, and M. Liu, "Lce-calib: Automatic lidar-frame/event camera extrinsic calibration with a globally optimal solution," *IEEE/ASME Transactions on Mechatronics*, vol. 28, pp. 2988–2999, 2023.

[11] J. Beltrán, C. Guindel, A. De La Escalera, and F. García, "Automatic extrinsic calibration method for lidar and camera sensor setups," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 17 677–17 689, 2022.

[12] G. Yan, F. He, C. Shi, P. Wei, X. Cai, and Y. Li, "Joint camera intrinsic and lidar-camera extrinsic calibration," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 11 446–11 452.

[13] Q. Herau, N. Piasco, M. Bennehar, L. Roldão, D. Tsishkou, C. Migniot, P. Vasseur, and C. Demonceaux, "Soac: Spatio-temporal overlap-aware multi-sensor calibration using neural radiance fields," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15 131–15 140, 2024.

[14] X. Lv, B. Wang, Z. Dou, D. Ye, and S. Wang, "Lccnet: Lidar and camera self-calibration using cost volume network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021, pp. 2888–2895.

[15] M. Cocheteux, J. Moreau, and F. Davoine, "Muli-ev: Maintaining unperturbed lidar-event calibration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024, pp. 4579–4586.

[16] M. Loper, N. Mahmood, J. Romero, G. Pons-moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 248:1–248:16, 2015.

[17] L. Zhou, Z. Li, and M. Kaess, "Automatic extrinsic calibration of a camera and a 3d lidar using line and plane correspondences," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 5562–5569.

[18] J. Cui, J. Niu, Z. Ouyang, Y. He, and D. Liu, "Acsc: Automatic calibration for non-repetitive scanning solid-state lidar and camera systems," in *arXiv preprint*, 2020.

[19] G. Pandey, J. R. Mcbride, S. Savarese, and R. M. Eustice, "Automatic targetless extrinsic calibration of a 3d lidar and camera by maximizing mutual information," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2012, p. 2053–2059.

[20] Z. W. Taylor and J. Nieto, "Automatic calibration of lidar and camera images using normalized mutual information," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2012.

[21] J. Levinson and S. Thrun, "Automatic online calibration of cameras and lasers," in *Robotics: Science and Systems (RSS)*, 2013.

[22] W. Wang, S. Nobuhara, R. Nakamura, and K. Sakurada, "Soic: Semantic online initialization and calibration for lidar and camera," *arXiv preprint*, 2020.

[23] T. Ma, Z. Liu, G. Yan, and Y. Li, "Crlf: Automatic calibration and refinement based on line feature for lidar and camera in road scenes," *arXiv preprint*, 2021.

[24] D. Detone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 337–33 712.

[25] N. Schneider, F. Piewak, C. Stiller, and U. Franke, "Regnet: Multimodal sensor registration using deep neural networks," in *IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 1803–1810.

[26] G. Iyer, R. K. Ram, J. K. Murthy, and K. M. Krishna, "Calibnet: Geometrically supervised extrinsic calibration using 3d spatial transformer networks," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1110–1117.

[27] S. Zhou, S. Xie, R. Ishikawa, K. Sakurada, M. Onishi, and T. Oishi, "Inf: Implicit neural fusion for lidar and camera," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 10 918–10 925.

[28] Z. Luo, G. Yan, and Y. Li, "Calib-anything: Zero-training lidar-camera extrinsic calibration method using segment anything," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[29] K. Takahashi, D. Mikami, M. Isogawa, and H. Kimata, "Human pose as calibration pattern; 3d human pose estimation with multiple unsynchronized and uncalibrated cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1775–1782.

[30] B. Triggs, P. F. Mclauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment - a modern synthesis," in *Vision Algorithms: Theory and Practice*, B. Triggs, A. Zisserman, and R. Szeliski, Eds., 2000, pp. 298–372.

[31] Y. Xu, Y. jhe Li, X. Weng, and K. Kitani, "Wide-baseline multi-camera calibration using person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 134–13 143.

[32] K. Liu, L. Chen, L. Xie, J. Yin, S. Gan, Y. Yan, and E. Yin, "Auto calibration of multi-camera system for human pose estimation," *IET Computer Vision*, vol. 16, no. 7, pp. 607—-618, 2022.

[33] S.-e. Lee, K. Shibata, S. Nonaka, S. Nobuhara, and K. Nishino, "Extrinsic camera calibration from a moving person," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 344–10 351, 2022.

[34] S. Sugimoto, H. Tateda, H. Takahashi, and M. Okutomi, "Obstacle detection using millimeter-wave radar and its visualization on image sequence," in *Proceedings of the International Conference on Pattern Recognition*, vol. 3, 2004, pp. 342–345.

[35] G. El Natour, O. Ait Aider, R. Rouveure, F. Berry, and P. Faure, "Radar and vision sensors calibration for outdoor 3d reconstruction," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 2084–2089.

[36] C. Schöller, M. Schnettler, A. Krämmer, G. Hinz, M. Bakovic, M. Güzet, and A. Knoll, "Targetless rotational auto-calibration of radar and camera for intelligent transportation systems," in *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 3934–3941.

[37] E. Wise, J. Peršić, C. Grebe, I. Petrović, and J. Kelly, "A continuous-time approach for 3d radar-to-camera extrinsic calibration," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 164–13 170.

[38] L. Cheng and S. Cao, "Online targetless radar-camera extrinsic calibration based on the common features of radar and camera," in *Proceedings of the IEEE National Aerospace and Electronics Conference*. Institute of Electrical and Electronics Engineers Inc., 2024, pp. 294–299.

[39] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang, and L. Su, "mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave," in *Proceedings of the ACM International Conference on Mobile Systems, Applications, and Services*, 2021, pp. 269–282.

[40] Z. Wang, M. Monjur, and S. Nirjon, "mmjoints: Expanding joint representations beyond (x,y,z) in mmwave-based 3d pose estimation," in *arXiv preprint*, 2025.

[41] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, "Hybrik: a hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3383–3393.

[42] V. Lepetit, F. Moreno-noguer, and P. Fua, "Epnp: an accurate o(n) solution to the pnp problem," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009.

[43] S. Garrido-jurado, R. Muñoz-salinas, F. Madrid-cuevas, and M. Marín-jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.