RF-Based 3D Through-Wall Pose Estimation: Heatmap CNN with Transformer Residuals, EMA Stabilization, and Alignment-Aware Training Protocol

Rian Atri[1], Tianxi Liang[1]

***Abstract*— We present a low-cost, scalable system for 3D skeletal pose estimation through walls using commodity RF sensing, designed for robust active perception in human–environment interaction. Raw Walabot measurements are rasterized into a bounded 3D voxel grid via percentile-normalized trilinear splatting and processed by a compact 3D CNN to generate per-frame feature volumes. A volumetric heatmap head outputs joint-specific 3D logits, from which coordinates are extracted by *soft-argmax*; per-joint heatmap entropy provides an uncertainty measure. Short-horizon dynamics are modeled with a causal temporal corrector that refines predictions using a sliding window of heatmaps.**

Training integrates a *focal-KL* heatmap objective and an *entropy-weighted* coordinate loss with *PA-aligned* supervision, a height-normalized bone prior, temporal smoothness regularizers, and an unsupervised *RF-density guidance* term. At inference, root stability is enforced via EMA filtering and a voxel-derived anchor blending RF centroids with top-K modes; a small bias correction further improves alignment.

Evaluated on alignment-sensitive metrics (PA-AP, PA-MPJPE [1]), the system demonstrates stable and accurate skeletal tracking suitable for edge deployment in SAR and privacy-preserving HCI contexts. By emphasizing rigid-alignment fidelity over unreliable global translation/scale, the approach achieves strong PA-AP@0.30 and PA-MPJPE with real-time performance on consumer devices, highlighting a pathway toward reliable RF-based active perception in dynamic, uncertain environments.

I. INTRODUCTION

Radio frequency (RF) waves possess the unique ability to penetrate common building materials, offering a sensing modality that remains reliable in environments where optical systems fail—such as smoke-filled interiors, collapsed structures, or unlit spaces [2]–[4]. Harnessing this property for *full-body 3D skeletal pose estimation* unlocks opportunities in safety-critical domains like search-and-rescue (SAR) and privacy-sensitive activity monitoring, where visual sensing may be unavailable or undesirable. The challenge lies in extracting stable, metrically accurate joint locations from multipath-dominated RF volumes while operating under limited compute and minimal supervision.

a) Our approach.: We introduce a fully end-to-end framework that transforms raw Walabot captures into normalized voxel grids, predicts *volumetric heatmaps* of joint likelihoods through a compact 3D CNN, and recovers metric-space coordinates via *soft-argmax*. To counter frame-level instability, a causal transformer applies short-horizon temporal corrections, refining each prediction with residual Δxyz adjustments [5]. Training is guided by a curriculum that blends supervised volumetric and coordinate objectives with PA-aligned optimization [6], temporal smoothness, bone-length priors, and an unsupervised RF-density guidance term. At inference, root stabilization is achieved by combining EMA filtering with voxel-derived centroid/mode anchors [7], further corrected by a lightweight calibration offset learned from training bundles.

b) Contributions.: Specifically, this work advances RF-based 3D pose estimation by:

- Introducing a **metric-space heatmap formulation** with soft-argmax readout and entropy-based uncertainty, replacing fragile direct regression.
- Designing a **causal temporal corrector** that operates on compressed heatmap features to suppress jitter in streaming inference [5].
- Proposing a **composite objective suite** that integrates focal-KL heatmap supervision, entropy-weighted coordinates, differentiable PA-MSE [6], height-normalized bone priors, temporal smoothness/acceleration, soft-PCK reward, and RF-density guidance with scheduled weighting.
- Deploying a **calibration-lite inference strategy** combining EMA root smoothing [7], centroid/mode fusion, bone-length projection, and a learned constant translation to mitigate residual extrinsic bias.
- Demonstrating **stable alignment-sensitive performance** through multi-metric training, Optuna-tuned hyperparameters [8], EMA stabilization [7], and PA-AP@0.30–based early stopping, showing the system's promise for real-world SAR and privacy-preserving applications.

II. BACKGROUND AND RELATED WORK

a) RF penetration.: Through-wall RF sensing exploits the ability of sub-GHz and ultra-wideband (UWB) signals to traverse common building materials. Attenuation depends on the material's complex permittivity $\epsilon = \epsilon' - j\epsilon''$ and the skin depth $\delta = \sqrt{2\rho/(\omega\mu)}$, where ρ is resistivity, ω angular frequency, and μ permeability [2]. Empirical studies show that UWB radar maintains usable signal-to-noise ratios through brick, wood, and drywall [3]. However, multipath,

Email: hello@rian.fyi
Email: til4023@med.cornell.edu
*In conjunction with the *Active Perception* Workshop @ IROS ("Bridging Sensing, Planning and Interaction").
[1]Affiliations omitted for review.

scattering, and dielectric mismatch cause nonlinear distortions. These effects motivate voxel-based representations, where reflections are spatially rasterized to mitigate aliasing while preserving geometric cues.

b) Cross-modal supervision.: RF pose estimation relies on *cross-modal alignment* between RF signals and optical ground truth. Zhao *et al.* [9] showed that aligning RF reflections with RGB-derived 2D joints enables networks to transfer vision supervision to RF. Pose frameworks like OpenPose [10] and MediaPipe [11] provide stable joint-indexed labels. RF-specific challenges include: (i) lack of hardware-locked synchronization between RF and optical streams, and (ii) redundant landmarks in 33-joint schemas, which add little RF signal and slow convergence. These issues motivate alternatives such as heatmap targets, PA-aligned losses [6], and bone priors enforcing anatomical consistency.

c) Backbones and temporal modeling.: Architectures for RF pose estimation adapt insights from vision backbones. EfficientNet [12] formalized compound scaling, guiding lightweight 3D CNNs for voxel data. Group Normalization (GN) [13] improves training with small RF batch sizes, and SiLU activations [14], [15] support gradient flow in low-SNR regimes. Temporal modeling is critical: frame-by-frame predictions amplify joint jitter. Causal transformers [5] with autoregressive masking offer sequence-aware corrections without future context.

d) Optimization strategies.: Training is difficult given noisy supervision and limited RF datasets. Polyak averaging [7] and exponential moving average stabilize updates, while Stochastic Weight Averaging (SWA) [16] finds wider optima. Adaptive optimizers such as AdamW [17], combined with cosine annealing and warm restarts [18], reduce overfitting. Hyperparameter optimization frameworks like Optuna [8] further support systematic search over capacity, learning rates, and loss weights.

e) Input and Output.: Many RF-pose systems rely on video supervision or multi-frame radar with high-end arrays [9], limiting comparability. We instead focus on a *single-shot RF-only* setting, benchmarking against internal controls and alignment-first metrics that isolate skeletal fidelity on commodity hardware.

III. DATA COLLECTION PROTOCOL

Our dataset is built from synchronized captures of a Walabot sensor and a single RGB camera. For each calibration bundle, we obtain pseudo-3D supervision of the 33 MediaPipe joints [11] via a *three-shot triangulation protocol*, where the operator repositions the camera to three static views (shots 0/1/2). At the central shot (shot 1), the Walabot collects multiple consecutive RF volumes, from which the first is discarded and the remaining 5 are retained. This redundancy accounts for RF multipath specularity: while any single return may be distorted by constructive/destructive interference, averaging predictions across 5 frames improves stability.

A. Walabot RF Capture

The Walabot returns a spherical grid $(r, \theta, \phi, \alpha)$ of amplitudes α, where r is radial distance, θ azimuth, and ϕ elevation. The arena is configured as $r \in [10, 500]$ mm, $\theta \in [-18°, 18°]$, $\phi \in [-15°, 15°]$ with steps $\Delta r = 8$ mm, $\Delta\theta = \Delta\phi = 2°$. This yields

$$N = \frac{r_{\max} - r_{\min}}{\Delta r} \times \frac{\theta_{\max} - \theta_{\min}}{\Delta\theta} \times \frac{\phi_{\max} - \phi_{\min}}{\Delta\phi}$$

voxels per scan.

Each voxel is projected to Cartesian:

$$\begin{aligned} r_m &= \frac{r}{100}, \\ x &= -r_m \sin\phi, \\ y &= r_m \cos\phi \sin\theta, \\ z &= r_m \cos\phi \cos\theta \end{aligned}$$

where the axis flip $(x, y, z) = (-z_0, y_0, x_0)$ aligns Walabot coordinates with camera coordinates.

Listing 1: RF capture: Walabot trigger, spherical-to-Cartesian projection, save to .npz

```
wlbt.Trigger()
raw = wlbt.GetRawImage()
arr = np.asarray(raw).reshape(phiLen, thetaLen,
    rLen)

# spherical     Cartesian (meters)
r_m = r_vals / 100.0
x = -r_m*np.sin(phis)
y = r_m*np.cos(phis)*np.sin(thetas)
z = r_m*np.cos(phis)*np.cos(thetas)

np.savez_compressed(out_path, x=x,y=y,z=z,amp=amps)
```

At shot 1, six frames are collected in sequence, the earliest is discarded, and the remaining 5 are bundled. This procedure mitigates transient noise and multipath specularity.

B. RGB Landmarks and Triangulation

The RGB camera (640×360) is calibrated with intrinsics (K, D) and stereo extrinsics $\{P_0, P_1, P_2\}$. For each shot, MediaPipe Pose detects 2D landmarks $\{(u, v)_i^k\}$ for joint i in view k [11]. We reconstruct 3D positions $\mathbf{X}_i$ using a linear DLT triangulation:

$$A_i = \begin{bmatrix} u_i^0 P_0^{(3)} - P_0^{(1)} \\ v_i^0 P_0^{(3)} - P_0^{(2)} \\ u_i^1 P_1^{(3)} - P_1^{(1)} \\ v_i^1 P_1^{(3)} - P_1^{(2)} \\ u_i^2 P_2^{(3)} - P_2^{(1)} \\ v_i^2 P_2^{(3)} - P_2^{(2)} \end{bmatrix}, \quad \mathbf{X}_i = \text{SVD}(A_i)[-1], \quad \mathbf{X}_i / \mathbf{X}_i^3.$$

Listing 2: Triangulation of 33 MediaPipe joints from 3 shots

```
for i in range(n_joints):
    A = []
    for cam_idx in range(3):
        P = P_mats[cam_idx]
        u,v = uv_list[cam_idx][i]
        A += [u*P[2]-P[0], v*P[2]-P[1]]
    _,_,Vt = np.linalg.svd(np.stack(A))
```

Fig. 1: Overall experimental setup, showing Walabot RF sensor, camera, and reference geometry.

```
    X = Vt[-1]; X /= X[3]
    XYZ[i] = X[:3]
XYZ[:,2] *= -1
```

The triangulated 3D landmarks serve as pseudo-ground truth, co-timestamped with the corresponding 5 Walabot frames.

C. Bundling

To ensure reproducibility and maintain strict synchronization between RF and RGB modalities, all artifacts from a capture session are consolidated into a timestamped directory of the form `bundle_YYYYMMDD_HHMMSS/`. Each bundle contains three components:

- `walabot_data/` – the 5 RF frames retained after pre-roll discard, stored as amplitude + Cartesian coordinates in compressed `.npz` files.
- `RGB_camera_data/` – raw images, 2D MediaPipe landmarks, and calibration intrinsics for each of the three static shots [11].
- `reconstruction.npz` – the pseudo-3D ground-truth skeleton obtained via three-shot triangulation.

IV. METHOD

A. RF voxelization and augmentations

We convert each Walabot frame into a dense 3D volume by *trilinear splatting* of amplitude-weighted points over a fixed grid of $64 \times 48 \times 64$ spanning meters:

$$x \in [-5, 5], \quad y \in [-2.5, 2.5], \quad z \in [0, 5].$$

Intensities are clipped at the 98th percentile and normalized to $[0, 1]$. We apply random yaw about y (up to $\pm 8°$), $\pm 0.07\,\mathrm{m}$ translation, amplitude dropout ($15\,\%$), and point drops ($10\,\%$) while preserving tensor shape. Sliding windows use $T{=}5$ frames, with labels corresponding to the *last* frame. *Active perception link:* these perturbations simulate sensor re-aims or small viewpoint changes that could be enacted online to reduce ambiguity.

Fig. 2: Three-shot triangulation protocol for RGB landmarks [11]. (a) Camera shot 0 (left view). (b) Camera shot 1 (center view). (c) Camera shot 2 (right view).

B. Backbone, heatmaps, and temporal residuals

The backbone is a **compact 3D CNN** designed for low-SNR RF volumes and small-batch training. Each voxel grid ($64 \times 48 \times 64$) passes through 3D convolutional layers with **Group Normalization (GN)** [13] (stable under batch < 8) and **SiLU/Swish** nonlinearities [14], [15]. A lightweight

Squeeze-and-Excitation (SE) block after the third stage improves channel recalibration and robustness to multipath.

A final $1\times1\times1$ convolution projects features into J joint-specific **volumetric heatmaps**. Joint coordinates are extracted with a differentiable **soft-argmax**, while **heatmap entropy** serves as an uncertainty proxy: sharp peaks $\Rightarrow$ confident predictions; diffuse distributions $\Rightarrow$ ambiguous ones.

To stabilize predictions over time, a small **causal transformer encoder** [5] processes compressed heatmap features from the last $T{=}5$ frames using autoregressive masking. It outputs per-joint residuals Δxyz added to the instantaneous prediction. Empirically, scaling by 0.25 prevents overshoot while damping jitter—critical for real-time streaming.

Auxiliary heads: person presence and similarity alignment.: Two compact heads branch from pooled features: (i) a person-presence classifier for gating updates in low-return frames, and (ii) a similarity head predicting (s, R, t) (with rot6d for R) trained with a consistency objective against differentiable Procrustes alignment [6]. This provides a soft, learnable alignment prior without heavy offline calibration.

RF-density guidance.: Let $E \in \mathbb{R}^{D\times H\times W}$ be the per-frame voxel energy. We normalize to $\hat{E} = \mathrm{softmax}(E/\tau)$ and encourage each predicted heatmap H_j to agree softly with $\hat{E}$:

$$\mathcal{L}_{\text{guide}} = \lambda_g \sum_{j=1}^{J} \mathrm{KL}\Big(H_j \,\|\, \hat{E}\Big),$$

with λ_g cosine-ramped over training. This preserves multi-modality and allows the network to override misleading energy.

Root fusion (EMA + centroid/mode).: The pelvis root $\tilde{p}_t$ from soft-argmax is fused with the previous root $\hat{p}_{t-1}$ using three strategies: (i) EMA smoothing [7], (ii) centroid of the top-K heatmap voxels, and (iii) mode ($\arg\max$). Entropy $\mathcal{H}(H_{\text{root}})$ gates the fusion: low-entropy $\Rightarrow$ centroid/mode; high-entropy $\Rightarrow$ EMA. *Active perception link:* high root entropy can signal the need for sensor dwell or micro re-aims to reduce uncertainty.

Together, (i) a GN–SiLU 3D CNN with SE recalibration, (ii) volumetric heatmaps with entropy-based confidence, (iii) causal transformer residual refinement, and (iv) auxiliary heads for presence and similarity alignment yield temporally stable, interpretable trajectories even under noisy RF supervision. The use of entropy and energy not only regularizes training but also suggests cues for *active perception*, where the system can decide to dwell, re-aim, or suppress updates based on confidence.

C. Root handling and kinematics

We treat the **pelvis** (mean of the hip landmarks) as the skeleton's *root*, central in the kinematic tree and relatively robust to occlusion. Training is pelvis-centered, ensuring supervision emphasizes relative alignment rather than global drift.

At inference, absolute root predictions are unstable due to multipath and noise. We stabilize with two lightweight steps: (i) an **Exponential Moving Average (EMA)** [7] of root coordinates to suppress jitter while preserving long-term drift, and (ii) a **voxel-driven anchor** combining intensity centroid and top-K energy modes, blended with EMA output. This hybrid anchor resists specular bias and keeps the root within supported RF regions. *Active perception link:* high entropy in the root heatmap signals the system to *dwell* for more frames or attempt a micro re-aim to reduce uncertainty.

After stabilization, we enforce **kinematic consistency** by rescaling bone segments to match a height-normalized template from the training distribution. This prevents limb collapse or elongation and yields visually plausible poses without requiring a full parametric body model.

D. Calibration-lite

Full extrinsic calibration with fiducials or mocap is impractical for rapid deployment. Instead, we adopt a **calibration-lite** strategy. Scene metadata and axis remapping provide a coarse world→RF transform. Residual offsets remain from unsynchronized capture and multipath.

To correct this, we estimate a **constant translation offset** Δt using differentiable Procrustes alignment (Umeyama similarity transform) [6]. The *median* pelvis residual across training bundles gives a robust estimate, applied at inference to all joints.

This yields (i) **practicality**—no dedicated calibration, (ii) **robustness**—resistant to noisy pseudo-labels, and (iii) **generality**—reusable across subjects and scenes. While scale and rotation errors persist, accuracy improves with negligible compute. *Active perception link:* entropy-weighted pelvis confidence could guide online updates of Δt, nudging calibration adaptively during streaming inference.

While this approach cannot fully compensate for rotational or scale errors, it significantly improves global skeleton placement with negligible compute cost. Future extensions could incorporate *online refinement*, where Δt is adaptively updated at inference using entropy-weighted pelvis confidence or fused with inertial priors, narrowing the gap between calibration-lite and full multi-sensor calibration.

E. Training objective

Let H be heatmap logits, P predicted joints, Y ground truth, E per-joint entropy, and W the last-frame voxel grid. With temperature τ,

$$\begin{aligned}\mathcal{L} \;=\; & w_{\text{hm}} \cdot \mathrm{KL}\big(G(Y) \,\big\|\, \mathrm{softmax}\big(\tfrac{H}{\tau}\big)\big) \\ & + w_{\text{xyz}} \cdot \|P - Y\|_1^{(\text{entropy-weighted})} + w_{\text{pa}} \cdot \|\Pi(P) - Y\|_1 \\ & + w_{\text{bone}} \cdot \mathcal{L}_{\text{bones}} + w_{\text{temp}} \cdot \mathcal{L}_{\text{temp}} + w_{\text{acc}} \cdot \mathcal{L}_{\text{acc}} \\ & + w_{\text{pck}} \cdot \mathcal{R}_{\text{close}} + w_{\text{rf}} \cdot \Big(-\tfrac{1}{J} \sum_j \texttt{trilinear}(W, P_j) \Big) \\ & + \mathcal{R}_{\text{root}} + \mathcal{R}_{\text{spread}}. \end{aligned} \tag{1}$$

Here $\Pi(\cdot)$ is differentiable Umeyama alignment [6]. The RF-density term samples W at P_j via trilinear interpolation. Loss weights $w_{\{\cdot\}}$ follow curricula. *Active perception link:* entropy-weighted terms encourage the model to learn when additional sensing (e.g., dwell or re-aim) would reduce uncertainty.

F. Optimization and regularization

Training stability is sensitive due to small datasets and noisy supervision. We use **AdamW** [17] with warmup and **cosine annealing restarts** [18], plus global-norm gradient clipping. An **EMA** of parameters [7] smooths validation. **Optuna** [8] tunes backbone width, transformer depth, dropout, learning rate, and loss weights—crucial since performance is highly sensitive to the balance between KL, coordinate, and PA-aware supervision.

Checkpoints are chosen by **PA-MPJPE** [1] and **PA-AP@0.30**, not just MSE, to prioritize alignment-consistent predictions. This combination yields reproducible convergence and generalization under low-resource conditions.

G. Inference

At deployment, inference runs in a **streaming** setting with a rolling buffer of $T{=}5$ frames, providing short-horizon temporal context for the `TemporalCorrector` without future leakage. The pipeline for each incoming frame is:

1) **Voxelization.** Raw Walabot amplitudes are projected into Cartesian coordinates and rasterized onto the $64 \times 48 \times 64$ grid with percentile normalization.
2) **Forward pass.** The compact 3D CNN backbone outputs volumetric heatmaps; soft-argmax yields sub-voxel coordinates, while entropy provides joint-wise uncertainty.
3) **Kinematic projection.** Joints are rescaled to a **height-normalized skeleton** using bone-length priors, ensuring anatomical plausibility.
4) **Root stabilization.** The pelvis root is stabilized by blending an **EMA** of past predictions with a voxel-derived anchor (centroid and top-K energy modes), reducing jitter and drift.
5) **Temporal residual correction.** The causal transformer refines the current pose with a residual Δxyz, scaled by 0.25, to damp noise while retaining responsiveness.
6) **Test-time augmentation (TTA).** Small Gaussian perturbations to voxel intensities are averaged in pelvis-centered space for robustness.
7) **Background attenuation and gating.** We subtract an exponential moving average (EWMA) background and gate updates using total RF energy, voxel density at joint sites, and entropy:

```
class RFBackground:
    def __init__(self, alpha=0.02, shape=(D,H,
        W)):
        self.alpha, self.ready = alpha, False
        self.bg = np.zeros(shape, np.float32)
    def update(self, vg):
        self.bg = vg if not self.ready else \
            (1-self.alpha)*self.bg + self.
                alpha*vg
        self.ready = True
    def subtract(self, vg):
        return np.clip(vg - self.bg, 0, None)
```

This acts as a lightweight clutter suppressor. *Active perception link:* if entropy is high or energy too low, the system can *dwell* or re-aim rather than propagate unstable updates.

Epoch	Train MSE	Val MSE	PA-AP@0.30	PA-MPJPE [m]
01	0.2209	0.2071	0.106	0.617
10	0.3286	0.2156	0.101	0.621
20	0.4337	0.2370	0.126	0.544
25	0.4511	0.2590	0.202	0.492
30	0.4847	0.2891	0.343	0.423
33	**0.5179**	**0.3071**	**0.460**	**0.396**

TABLE I: Training trajectory for loop 2 (non-optimized).

8) **Entropy-aware smoothing.** Beyond EMA, we apply a simple entropy-aware Kalman filter that scales measurement noise:

```
R = np.eye(3) * (base_var * (1 + entropy))
K = P @ H.T @ np.linalg.inv(H @ P @ H.T + R)
x = x + K @ (z - H @ x)
```

This adaptively smooths predictions without future context.

H. Evaluation metrics

Primary: *PA-MPJPE* and *PA-AP@r* with $r \in \{0.10, 0.30, 0.50\}$ after differentiable Umeyama alignment. Secondary: pelvis-centered MPJPE, bone-length MAE, and **uncertainty calibration** (ECE). Stability: per-joint *jerk* and *flip-rate*. Absolute MPJPE is reported but de-emphasized due to hardware/sync limits.

Finally, a **latency probe** records voxelization, forward pass, and post-processing. On consumer hardware, mean latency remains interactive, supporting search-and-rescue and privacy-sensitive monitoring scenarios.

V. RESULTS

A. Primary metrics

We report MPJPE (absolute, pelvis-centered), PA-MPJPE [1], AP@{0.10,0.30,0.50,0.75,0.90}, PA-AP@{0.30,0.50}, and a *near-miss* AP band at $[0.30, 0.35)$ m.

For the *non-optimized loop* (loop 2), early stopping at epoch 33 yielded: Train MSE $= 0.5179$, Val MSE $= 0.3071$, AP@0.30 $= 0.061$, AP@0.50 $= 0.106$, AP@0.75 $= 0.303$, AP@0.90 $= 0.439$, PA-AP@0.30 $= 0.460$, PA-AP@0.50 $= 0.742$, MPJPE(abs) $= 4.022$ m, MPJPE(centered) $= 0.906$ m, and PA-MPJPE $= 0.396$ m.

The *optimized loop* (loop 1) used PA-AP@0.30 as the early-stopping key. Its best checkpoint occurred earlier (epoch 11) with PA-AP@0.30 $= 0.808$, PA-AP@0.50 $= 0.970$, PA-MPJPE $= 0.188$ m, MPJPE(abs) $= 3.979$ m, MPJPE(centered) $= 0.796$ m, AP@0.30 $= 0.101$, AP@0.90 $= 0.672$.

Alignment-sensitive metrics and supplementary AP values are summarized below:

B. Optimized vs. non-optimized comparison

a) Observation.: PA-aware optimization nearly doubled PA-AP@0.30 (0.808 vs. 0.460) and halved PA-MPJPE (0.188 m vs. 0.396 m), validating alignment-first training. Absolute MPJPE improved less.

Epoch	Train MSE	Val MSE	PA-AP@0.30	PA-MPJPE [m]
01	0.1955	0.2065	0.227	0.524
10	0.2357	0.2320	0.697	0.305
14	0.2544	0.2661	0.828	0.185
15	0.2335	0.2572	0.803	0.182
20	0.2456	0.2608	0.687	0.307
24	**0.2472**	**0.2694**	**0.813**	**0.186**

TABLE II: Training trajectory for loop 1 (optimized).

	PA-AP@0.30↑	**PA-AP@0.50↑**	**PA-MPJPE↓**
Loop 2 (epoch 33)	0.460	0.742	0.396 m

TABLE III: Alignment-sensitive metrics at early stop (loop 2).

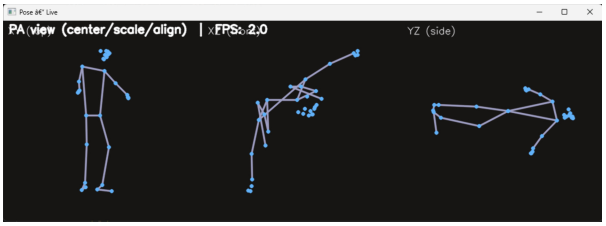

Fig. 3: Predicted 3D pose overlayed on RF voxel/point cloud.

C. Ablations

We ablate design choices under identical splits, early stopping on PA-AP@0.30:

a) Backbone / head / loss.: (1) Full model; (2) Heatmap $\rightarrow$ regression (expected instability, no calibrated confidence); (3) $-$ PA loss (drop in PA metrics); (4) $-$ bone prior (implausible limb lengths).

b) Guidance / robustness.: (5) $-$ RF-density guidance (more drift in low-SNR); (6) Misleading-energy stress (shuffle/invert maps, test robustness); (7) Occluder bias (false-attraction rate under metallic clutter); (8) Range/SNR bins (guidance benefits vs. SNR).

c) Temporal stability / root anchoring.: (9) $-$ temporal residual (higher jerk/flip); (10) $-$ root fusion (higher drift); (11) Root fusion variants (EMA-only, centroid-only, mode-only, none; compare drift@T, reset-lag, PA-AP@0.30, jerk).

D. Qualitative results

Figure 3 shows predicted skeletons overlayed on RF voxel clouds, color-coded by error (blue=low, red=high). Transparency encodes entropy. Torso/limbs remain stable; residual errors appear at wrists/ankles.

	AP@0.10	AP@0.30	AP@0.50	AP@0.75	AP@0.90
Loop 2 (epoch 33)	0.011	0.061	0.106	0.303	0.439

TABLE IV: Supplementary AP metrics (loop 2).

Loop	Epoch	PA-AP@0.30↑	PA-AP@0.50↑	PA-MPJPE [m]↓
Optimized (loop 1)	24	**0.813**	**0.970**	**0.186**
Non-optimized (loop 2)	33	0.460	0.742	0.396

TABLE V: Optimized vs. non-optimized: alignment-sensitive metrics.

Loop	Epoch	MPJPE(abs) [m]	MPJPE(ctr) [m]	AP@0.30	AP@0.90
Optimized (loop 1)	24	3.979	0.835	0.091	0.641
Non-optimized (loop 2)	33	4.022	0.906	0.061	0.439

TABLE VI: Supplementary metrics across loops.

Fig. 4: Collection setup. A subject stands $\sim$2 m in front of the Walabot during testing. RF volumes and RGB snapshots captured in this configuration are used for pseudo-3D triangulation; model predictions on these scans are shown in Fig. 3.

VI. DISCUSSION

This work is a *proof of concept* rather than a mature deployment. Limitations shaped outcomes but also highlight opportunities for advancing RF-based *active perception*.

a) Data and supervision.: Our dataset covers only 30 short scenes, each frame aggregating $\sim$5 Walabot volumes plus three RGB snapshots for pseudo-3D triangulation of 33 MediaPipe landmarks [11]. Labels were noisy due to handheld cameras, lack of synchronization, and MediaPipe's 2D errors [19], constraining absolute accuracy and generalization.

b) Sensor limitations.: The Walabot, though inexpensive, was pushed beyond its intended range. Limited angular/depth resolution and multipath interference hindered joint tracking beyond 2–3 m or through denser walls. Compared to mmWave FMCW or UWB arrays, spatial fidelity was much lower [20], [21].

c) Collection setup.: Fig. 4 illustrates the capture configuration (not true ground truth). Predictions in Fig. 3 show both the promise and the limits of approximate supervision.

d) Compute constraints.: Consumer-grade hardware limited batch size, ablation breadth, and hyperparameter search. Despite Optuna tuning [8], stability relied heavily on EMA and curricula [7], [16], suggesting richer compute would enable deeper exploration.

e) Limitations summary.: Performance is bounded by low-resolution radar, noisy pseudo-3D labels, and limited data. Absolute MPJPE suffers, so rigid-alignment metrics (PA-AP@r, PA-MPJPE) better reflect skeletal fidelity. Still, RF-density guidance improved alignment, and root stabilization (EMA+centroid/mode) balanced drift and jitter.

f) Methodological takeaways.: Key lessons for active perception:

- **Volumetric heatmaps** with soft-argmax provide interpretable joints and uncertainty estimates.
- A **causal temporal corrector** stabilized jitter for real-time streaming [5].
- **RF-density guidance** anchored predictions to physically plausible regions.
- **Alignment-aware objectives** (PA losses, bone priors) yielded perceptually consistent skeletons despite noisy global translation.

g) Future directions.: Advancing active RF perception will require: (1) larger, synchronized datasets with mocap/depth ground truth; (2) higher-resolution mmWave or UWB arrays with beamforming [20]; (3) stronger root/global recovery using GPS/IMU or SMPL models [22]; (4) domain randomization for diverse wall materials; (5) multi-sensor fusion (RF+IMU/RGB-D); (6) deployment testing in SAR or smoke-filled environments [4].

VII. CONCLUSION

We presented a complete, low-cost pipeline for **3D skeletal pose estimation through walls** using commodity RF sensing. The system proceeds end-to-end as

$$\begin{aligned} \text{RF} &\rightarrow \text{voxelize} \rightarrow \text{compact 3D CNN} \\ &\rightarrow \text{3D heatmaps} \rightarrow \text{soft-argmax} \\ &\rightarrow \text{causal residual corrector} \rightarrow \text{pose}. \end{aligned}$$

Our training objective integrates multiple complementary terms: volumetric KL divergence over heatmaps, entropy-weighted coordinate regression, PA-aligned supervision (differentiable Umeyama) [6], temporal smoothness and acceleration priors, height-scaled bone-length normalization, a soft-PCK reward annealed over training, and an unsupervised RF-density guidance term that biases predictions toward high-energy voxels. Curriculum schedules gradually ramp the weights of alignment-sensitive and RF-density terms to stabilize convergence. At inference, the pipeline includes **root stabilization** via EMA smoothing and voxel-derived centroid/mode anchors [7], along with a constant global translation offset estimated from training bundles to correct residual calibration bias.

Experiments across 30 pseudo-3D annotated scenes show that, despite noisy triangulated supervision, the system achieves **competitive alignment-sensitive metrics**: a PA-MPJPE of 0.188–$0.396\,\text{m}$ and PA-AP@0.30 up to 0.808 depending on optimization strategy. Notably, explicit early stopping on PA-AP@0.30 nearly doubled performance compared to generic validation loss, showing the importance of alignment-aware supervision. We also observed that RF-density guidance and temporal residual correction substantially reduced jitter and improved plausibility in streaming settings.

While the dataset and sensor are modest, these results demonstrate that meaningful 3D pose information can be extracted from low-cost RF hardware. The system provides a proof of concept for future RF-based perception in **privacy-preserving monitoring**, **human-computer interaction**, and **search-and-rescue scenarios** where cameras are ineffective or undesirable [4]. Scaling data collection with motion capture, integrating higher-resolution RF arrays, and incorporating multi-sensor fusion (e.g., inertial + RF) represent immediate next steps toward deployment-ready systems. Hence, we report internal, modality-pure baselines and alignment-first metrics, which are the appropriate comparators for single-shot RF-only.

REFERENCES

[1] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3d human pose estimation," in *ICCV*, 2017, pp. 2640–2649.

[2] C. A. Balanis, *Advanced Engineering Electromagnetics*, 2nd ed. Wiley, 2012.

[3] A. H. Muqaibel, A. Safaai-Jazi, A. Bayram, A. M. Attiya, and S. M. Riad, "Ultrawideband through-the-wall propagation," *IET Proc. Microwaves, Antennas and Propagation*, vol. 152, no. 6, pp. 581–588, 2005.

[4] U.S. Department of Homeland Security, Science & Technology Directorate, "Detecting heartbeats in rubble: Dhs and nasa team up to save victims of disasters," https://www.dhs.gov/.../victims-disasters, 2013.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[6] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 4, pp. 376–380, 1991.

[7] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[8] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *KDD*, 2019, pp. 2623–2631.

[9] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *CVPR*, 2018, pp. 7356–7365.

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017, pp. 1302–1310.

[11] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "Blazepose: On-device real-time body pose tracking," *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2006.10204

[12] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, vol. 97. PMLR, 2019, pp. 6105–6114.

[13] Y. Wu and K. He, "Group normalization," in *ECCV*. Springer, 2018, pp. 3–19.

[14] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, vol. 107, pp. 3–11, 2018.

[15] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv*, 2017.

[16] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima in deep learning," in *UAI*, 2018.

[17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019, adamW.

[18] ——, "Sgdr: Stochastic gradient descent with warm restarts," in *ICLR*, 2017.

[19] S. Dill, J. Ehret, and M. Lienkamp, "Accuracy evaluation of 3d pose reconstruction algorithms... mediapipe pose," *Sensors*, vol. 24, no. 23, p. 7772, 2024.

[20] Y. Song, J. Guo, X. Li, and Z. Cui, "Through-wall human pose reconstruction via uwb mimo radar and cnn," *Remote Sensing*, vol. 13, no. 2, p. 241, 2021.

[21] D. Xu, Y. Yan, and H. Tang, "Cross-modal supervised human body pose recognition techniques for through-wall radar," *Sensors*, vol. 24, no. 7, p. 2207, 2024.

[22] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "Smpl: A skinned multi-person linear model," *ACM Trans. Graph. (SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, 2015.