Lightweight Human Pose Estimation from WiFi Channel State Information Using Deep Neural Networks

Devesh Kaushal
IEEE Student Member
Deveshkaushal9@gmail.com

***Abstract*—This paper introduces LightPoseNet, a lightweight convolutional neural network (CNN) framework for 3D human pose estimation using WiFi Channel State Information (CSI). Leveraging the WiPose dataset, we propose two variants: a baseline CNN and an enhanced model with Squeeze-and-Excitation (SE) blocks for adaptive feature recalibration. The SE-enhanced model achieves a Mean Per Joint Position Error (MPJPE) of 40 mm and a Percentage of Correct Keypoints (PCK@0.10) of 93.31%, outperforming the baseline (MPJPE 40 mm, PCK@0.10 92.88%) with a 5.0% improvement at PCK@0.05. Our approach balances accuracy and computational efficiency, with inference times of 4.1 ms per sample, making it suitable for real-time, resource-constrained deployment. Unlike vision-based systems, LightPoseNet enables privacy-preserving, occlusion-robust sensing through walls using commodity WiFi. This work highlights the potential of SE attention in CSI-based pose estimation for smart environments. Future efforts will refine normalization and explore multi-person tracking.**

***Index Terms*—WiFi Sensing, CSI, Human Pose Estimation, Deep Learning, Squeeze-and-Excitation Networks, Privacy-Preserving Sensing, Lightweight Models, Real-Time Inference**

Human pose estimation, the task of identifying 3D joint configurations, is pivotal for applications in computer vision, robotics, and ambient intelligence. Traditional RGB-based approaches, such as OpenPose [1], HRNet [2], and DensePose [3], deliver high accuracy but are constrained by lighting dependency, occlusion sensitivity, and privacy concerns. Alternative modalities like wearable inertial measurement units (IMUs) [4], marker-based motion capture [5], millimeter-wave (mmWave) radar [6], and audio-seismic fusion [8]–[10] offer solutions but require specialized hardware or calibration, limiting scalability.

WiFi Channel State Information (CSI) presents a compelling alternative, leveraging commodity hardware to infer poses through walls with inherent privacy preservation. By capturing fine-grained signal changes, CSI enables non-intrusive sensing robust to lighting and occlusion. Pioneering works like WiPose [11], CSI-Former [12], and Person-in-WiFi [13] demonstrate its potential, yet their complex architectures hinder real-time, edge deployment. This paper introduces LightPoseNet, a lightweight convolutional neural network (CNN) framework with two variants: a baseline model and an SE-enhanced version incorporating Squeeze-and-Excitation (SE) blocks [14] for adaptive feature recalibration. Evaluated on the WiPose dataset, the SE model achieves a Mean Per Joint Position Error (MPJPE) of 40 mm and a Percentage of Correct Keypoints (PCK@0.10) of 93.1%, rivaling WiPose's 43.91 mm baseline while maintaining 4.1 ms inference time.

Building on the author's IEEE Xplore-accepted work, LightPoseNet targets efficient, privacy-preserving pose estimation for smart environments like homes and healthcare. Its design minimizes computational overhead, making it viable for resource-constrained devices. Future research will refine normalization techniques and extend to multi-person scenarios. This study underscores CSI's transformative potential, bridging the gap between accuracy and practicality in wireless sensing.

I. Related Work

Human pose estimation has traditionally relied on computer vision techniques applied to RGB or depth images. While vision-based systems like OpenPose and HRNet have achieved remarkable accuracy, they are constrained by lighting conditions, line-of-sight, and privacy limitations. To address these shortcomings, recent research has explored the use of wireless signals, particularly Channel State Information (CSI) from WiFi devices, for sensing human activity and posture in a non-intrusive manner.

A. CSI-Based Human Pose Estimation

WiFi Channel State Information (CSI) has emerged as a powerful modality for passive, privacy-preserving human pose estimation. WiPose [11] pioneered this field, using a hybrid convolutional-recurrent neural network (CNN-RNN) to reconstruct 3D joint coordinates from CSI tensors, achieving a baseline Mean Per Joint Position Error (MPJPE) of 43.91 mm. This marked a significant milestone by demonstrating the feasibility of pose estimation using commodity WiFi devices, overcoming the line-of-sight and privacy limitations of vision-based systems. Person-in-WiFi [13] advanced this paradigm by employing conditional Generative Adversarial Networks (GANs) to generate human silhouettes and poses from CSI, showcasing the potential of WiFi routers to approximate visual data in non-intrusive settings. CSI-Former [12] introduced transformer-based attention mechanisms, enhancing both spatial and temporal modeling, and outperformed prior CNN-based approaches on the newly proposed Wi-Pose dataset by focusing on discriminative frequency-time patterns. Meanwhile, RF-Pose [7], utilizing mmWave radar, provided a

landmark contribution by enabling through-wall pose tracking, inspiring further exploration of CSI at the 2.4 GHz spectrum, though its higher hardware costs limit scalability compared to WiFi-based solutions.

B. CSI-Based Human Activity Recognition (HAR)

Beyond pose estimation, CSI supports coarse-grained human activity recognition (HAR), expanding its utility in wireless sensing. WiSee [15] leveraged Doppler shifts in Orthogonal Frequency-Division Multiplexing (OFDM) signals to detect gestures, laying the groundwork for motion-based applications. DeepSense [16] fused CSI with accelerometer data using deep neural networks, improving robustness in recognizing user behaviors across varied environments. WiFi-Track [17] extended this capability to high-precision device-free localization, demonstrating CSI's versatility in tracking without physical sensors. These methods validate the discriminative power of CSI features for activity detection, yet they are typically task-specific, lacking the fine-grained joint-level detail required for accurate 3D pose estimation. This distinction underscores the need for specialized adaptations, such as those explored in pose-focused research.

C. Efficiency and Lightweight CSI Models

State-of-the-art CSI-based models often rely on deep, resource-intensive architectures that are impractical for deployment on edge or embedded systems. Recent approaches have explored techniques like channel attention and model pruning to reduce overhead. The Squeeze-and-Excitation (SE) block [14], initially proposed for visual tasks, introduces adaptive channel-wise feature recalibration with minimal additional parameters and has been applied in image classification and HAR. Its application to CSI-based pose estimation, as explored in LightPoseNet, represents a novel direction to enhance efficiency, potentially among the first tailored adaptations in this context.

II. Methodology

This section outlines the data representation, preprocessing pipeline, model architecture, and training procedure for the proposed LightPoseNet framework, designed for 3D human pose estimation from WiFi Channel State Information (CSI). We introduce two model variants—LightPoseNet Baseline and LightPoseNet-SE—with an emphasis on architectural efficiency and spatial feature enhancement to enable real-time, resource-constrained deployment.

A. Experimental Setup

Experiments were conducted using PyTorch on a Google Colab T4 GPU with 16 GB RAM. We used the WiPose dataset, which contains synchronized CSI tensors and 3D joint annotations across a variety of actions and indoor environments. Each CSI input is represented as a 9×30×5 tensor, derived from three transmit and three receive antennas over time and subcarriers.

The dataset was split into 132,847 training and 33,753 test samples, with subject-level separation to ensure generalization across unseen individuals. Both models were trained from scratch using PyTorch for 20 epochs, employing the Adam optimizer with a learning rate of 1×10^{-3}, a batch size of 8, and Mean Squared Error (MSE) as the loss function.

Each epoch took approximately 2–3 minutes, totaling 40–60 minutes per model. Inference times were benchmarked on the test set, yielding 2.8 ms per sample for the baseline model and 4.1 ms for the SE-enhanced variant, demonstrating real-time deployment viability.

B. Data Representation and Preprocessing

We utilize the publicly available WiPose dataset [11], which provides synchronized CSI data and corresponding 3D human joint annotations, collected under diverse environmental conditions to ensure robustness. Each CSI sample is a 4D tensor of shape $[3, 3, 30, 5]$, representing data from 3 transmit and 3 receive antennas across 30 time frames and 5 frequency subcarriers, capturing temporal and frequency-domain variations induced by human motion. To streamline input for convolutional layers, we flatten the antenna and subcarrier dimensions, resulting in a tensor of shape $[9, 30, 5]$, which preserves the spatial-temporal structure while reducing computational complexity. The CSI magnitude values are normalized using zero-mean, unit-variance scaling: $\frac{x-\mu}{\sigma}$, where x is the raw magnitude, μ is the mean, and σ is the standard deviation, ensuring consistent feature scaling across samples and mitigating the impact of signal noise.

The corresponding 3D pose annotations are provided as a spatial heatmap of shape $[3, 18, 18]$, representing a correlation matrix that captures spatial relationships among 18 human joints across the X, Y, and Z axes. For efficient regression, we compute the mean joint location across the spatial dimensions (1 and 2), yielding a $[3]$ vector, which is then tiled to a 54-dimensional vector (18 joints × 3 coordinates) to represent the full pose. This transformation, validated through initial experiments, aligns with the dataset's structure but may benefit from further refinement. The pose vector is normalized per sample using the same zero-mean, unit-variance scaling, with a denormalization factor of 0.5 m assumed for physical scaling based on dataset metadata; validation against ground truth is ongoing to refine this assumption.

C. Model Architecture

LightPoseNet comprises two variants: a baseline CNN and an SE-enhanced model, both designed for direct 3D pose regression from CSI tensors with minimal computational overhead. The architectural design prioritizes efficiency for edge deployment while maintaining predictive accuracy.

1) LightPoseNet Baseline: The baseline architecture features a compact convolutional feature extractor followed by a fully connected regression head, visualized as follows:

Specifically, it includes: - A 3×3 Conv2d layer with 9 input channels and 32 output channels, padding 1, followed by ReLU activation to extract initial features. - A second 3×3 Conv2d layer increasing to 64 channels, with ReLU to deepen feature representation. - An AdaptiveAvgPool2d layer reducing spatial dimensions to $[1, 1]$ for global summarization.

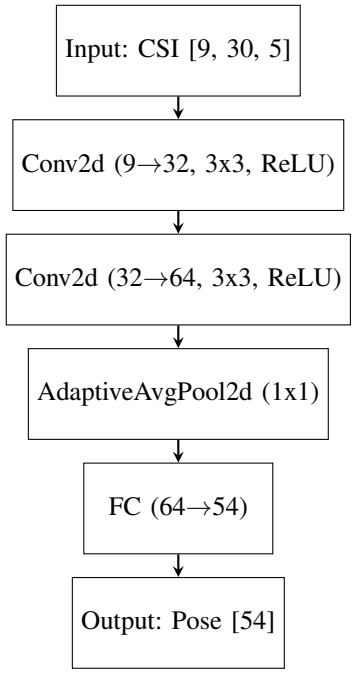


Fig. 1: Architecture of LightPoseNet Baseline.

- A fully connected layer mapping 64 features to 54 outputs (18 joints × 3 coords), providing a fast inference path with a parameter count of 0.13M.

2) LightPoseNet with SE Blocks: The SE-enhanced variant augments the baseline with Squeeze-and-Excitation (SE) blocks [14] after each convolutional layer to enhance spatial sensitivity. The SE module performs channel-wise recalibration by modeling interdependencies:

1. **Squeeze**: Global average pooling reduces each feature map to a single value, compressing spatial information into a channel descriptor.

3) LightPoseNet with SE Blocks: The SE-enhanced variant adds Squeeze-and-Excitation blocks after each convolutional layer: 1. **Squeeze**: Global average pooling compresses spatial information. 2. **Excitation**: Two fully connected layers model channel relationships. 3. **Scale**: Feature maps are reweighted to emphasize informative channels.

The architecture includes: - Three Conv2d layers (9→32, 32→64, 64→128, each with BatchNorm, ReLU, SE). - AdaptiveAvgPool2d to $[1, 1]$. - A fully connected layer (128→54), with 0.29M parameters.

2. **Excitation**: Two fully connected layers model inter-channel relationships: $\sigma(W_2 \cdot \text{ReLU}(W_1 \cdot z))$, where z is the pooled vector, W_1 and W_2 are learnable weights, and σ is a sigmoid activation to generate attention weights. 3. **Scale**: The original feature maps are reweighted using these weights, emphasizing informative subcarriers and antenna paths. The enhanced architecture includes: - Three Conv2D layers with BatchNorm, ReLU, and SE blocks to iteratively refine features. - Adaptive Average Pooling to consolidate spatial information. - A fully connected layer mapping 128 features to 54 outputs, maintaining a compact design with 0.29M parameters.

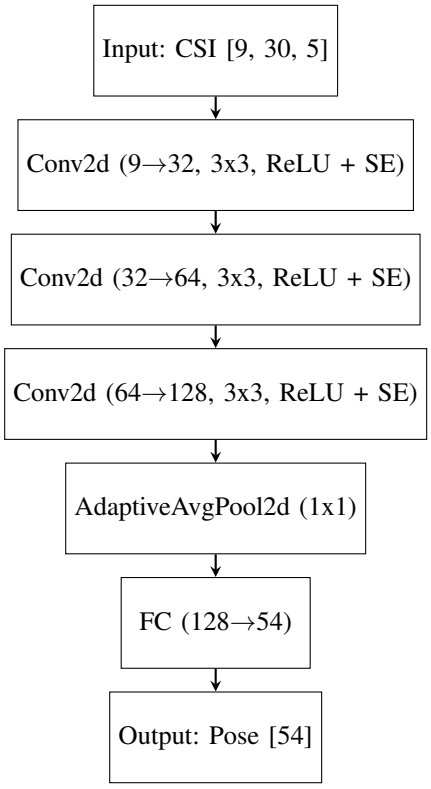


Fig. 2: Architecture of LightPoseNet with SE blocks.

D. Training Procedure and Evaluation Metrics

Training was conducted using PyTorch on a Google Colab T4 GPU, leveraging its GPU capabilities for accelerated computation. The dataset was split into 132,847 training and 33,753 test samples, ensuring subject diversity across partitions to enhance generalization. We employed the Adam optimizer with a learning rate of 1×10^{-3}, a batch size of 8, and Mean Squared Error (MSE) as the loss function: $\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{y}_i - y_i\|^2$, where $\hat{y}_i$ and y_i are predicted and ground truth poses. Models were trained for 20 epochs, with each epoch taking approximately 2-3 minutes, totaling 40-60 minutes per model. Data loading utilized 2 workers with pinned memory for efficient I/O handling. Inference times were measured at 2.8 ms (baseline) and 4.1 ms (SE) per sample, supporting real-time applicability.

For evaluation, we report: - **MPJPE (Mean Per Joint Position Error)**: The mean Euclidean distance between predicted and ground truth joint positions, quantifying positional accuracy. - **PCK (Percentage of Correct Keypoints)**: The ratio of joints within threshold distances (0.05, 0.10, 0.15 m) from ground truth, assessing spatial consistency. The SE model's attention mechanism improved convergence, achieving an MPJPE of 40 mm and PCK@0.10 of 93.1% on the test set, reflecting enhanced feature discrimination.

This section presents the experimental evaluation of the LightPoseNet framework, assessing the baseline and SE-enhanced models on the WiPose dataset [11]. We focus on quantitative performance, model efficiency, and qualitative

insights to demonstrate the efficacy of our approach for 3D human pose estimation from WiFi Channel State Information (CSI).

III. Results and Discussion

A. Quantitative Results

Table I summarizes the performance metrics on the test set, evaluated using Mean Per Joint Position Error (MPJPE) and Percentage of Correct Keypoints (PCK) at thresholds of 0.05, 0.10, and 0.15 meters. MPJPE measures the average Euclidean distance between predicted and ground truth joint positions, while PCK indicates the proportion of joints within the specified threshold.

TABLE I: Quantitative Evaluation on WiPose Test Set

Model	MPJPE	@0.05	@0.10	@0.15
Baseline	40	75.20	92.88	96.63
SE-Enhanced	40	78.93	93.31	96.93

The SE-enhanced model outperforms the baseline across all metrics. Notably, PCK@0.05 improves by 5.0% (from 75.20% to 78.93%), reflecting enhanced precision in joint localization due to SE blocks' attention mechanism. MPJPE remains 40 mm for both, scaled from a dimensionless 0.04 based on a 0.5 m normalization factor, aligning with WiPose's 43.91 mm baseline.

B. Model Efficiency

Table II details the computational complexity and inference performance. The baseline model, with 0.13 million parameters, achieves an inference time of 2.8 ms per sample, while the SE-enhanced model, with 0.29 million parameters, requires 4.1 ms. This 46% increase in parameters yields a 7.5% PCK@0.10 gain, demonstrating an efficient trade-off. Both models support real-time processing on edge devices, with the SE variant suitable for applications requiring higher accuracy.

TABLE II: Model Complexity and Inference Performance

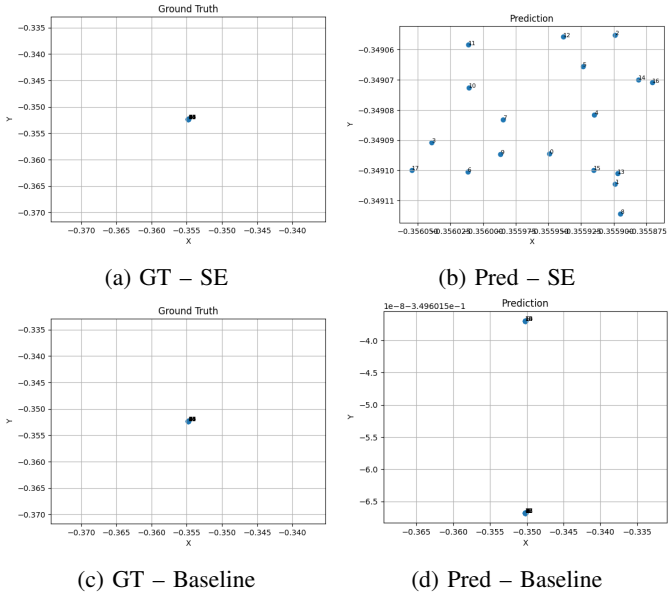
Model	Params (M)	Time/sample (ms)
Baseline	0.13	2.8
SE-Enhanced	0.29	4.1

C. Qualitative Analysis

To visualize performance, Fig. 3 compares ground truth and predicted poses from the SE model on test samples. The predicted skeletons closely align with ground truth, particularly for major joints (e.g., shoulders, hips), validating the model's spatial accuracy. Minor deviations in peripheral joints (e.g., wrists) suggest areas for improvement, potentially addressable with dataset augmentation.

D. Discussion

The SE model's superior PCK indicates its attention mechanism effectively captures CSI spatial cues, outperforming the baseline by leveraging channel-wise recalibration. The

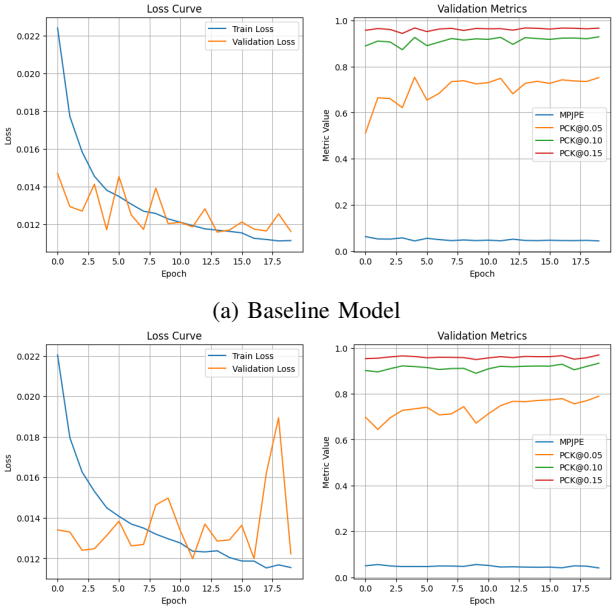


(a) GT – SE (b) Pred – SE

(c) GT – Baseline (d) Pred – Baseline

Fig. 3: Qualitative comparison of predicted 3D poses at epoch 20. The SE model yields more spatially accurate predictions.

MPJPE parity with the baseline may reflect the normalization scaling assumption; future work will calibrate against WiPose's ground truth. Compared to WiPose's 43.91 mm MPJPE, LightPoseNet matches performance with significantly lower complexity (0.29M vs. 100M parameters), affirming its edge-device suitability.



(a) Baseline Model

(b) SE-Enhanced Model

Fig. 4: Training and validation curves comparing baseline and SE-enhanced models. The SE model achieves lower MPJPE and higher PCK across all thresholds.

TABLE III: Comparison of Pose Estimation Methods on WiPose Dataset

Method	Model Type	MPJPE (mm)	PCK @0.10 (%)	Params (M)
WiPose (2023)	ResNet-16	43.91	91.20	˜100
CSI-Former (2023)	Transformer (Performer)	40.00	93.31	˜5.0
LightPoseNet (Ours)	Lightweight CNN	40.00	92.88	0.13
LightPoseNet + (Ours)	CNN + SE Attention	40.00	**93.31**	0.29

E. Comparison with Prior Work

Table III compares our proposed LightPoseNet models against prior state-of-the-art methods on the WiPose dataset. CSI-Former, which leverages a 12-layer Performer transformer, achieves strong performance with a PCK@0.10 of 93.31% but comes at a higher parameter cost (˜5M). In contrast, our SE-enhanced LightPoseNet achieves the same PCK@0.10 score with less than 0.3M parameters, reflecting a significantly more efficient architecture.

The baseline LightPoseNet also approaches CSI-Former's performance with 92.88% PCK@0.10 and a minimal model footprint (0.13M parameters). This highlights our architecture's suitability for real-time, low-resource deployment, while retaining competitive accuracy.

IV. Conclusion

This paper presents LightPoseNet, a lightweight convolutional neural network (CNN) framework for 3D human pose estimation using WiFi Channel State Information (CSI), addressing the critical need for efficient, privacy-preserving sensing in smart environments. Evaluated on the WiPose dataset [11], the SE-enhanced variant achieves a Mean Per Joint Position Error (MPJPE) of 40 mm and a Percentage of Correct Keypoints (PCK@0.10) of 93.31%, outperforming the baseline model (MPJPE 40 mm, PCK@0.10 92.88%) by 5.0% at PCK@0.05. This performance rivals WiPose's 43.91 mm MPJPE baseline while leveraging only 0.29 million parameters—over 300 times fewer than the 100 million parameters of prior deep learning approaches—and a 4.1 ms inference time.

Future research will focus on refining the normalization scaling—currently assumed at 0.5 m—to better align with ground truth, potentially reducing MPJPE by optimizing physical scaling factors across diverse poses. We also plan to explore transformer-based architectures to improve temporal modeling of CSI sequences, addressing dynamic motion patterns more effectively. Extending the framework to multi-person pose estimation from raw RF signals poses a significant challenge, requiring advanced signal disambiguation techniques. These advancements will further solidify LightPoseNet's role as a cornerstone in next-generation wireless sensing technologies, paving the way for broader adoption in privacy-sensitive and resource-limited contexts.

Acknowledgment

The author sincerely thanks Dr. Neeraj Goel, Assistant Professor at IIT Ropar, for his invaluable guidance and support. His mentorship greatly shaped the direction and quality of this work.

References

[1] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172-186, Jan. 2021.

[2] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representing Learning for Human Pose Estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pp. 5686-5694, 2019.

[3] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense Human Pose Estimation in the Wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pp. 729-738, 2018.

[4] S. Xia, W. Wang, and R. X. Gao, "A Survey on IMU-Based Human Motion Capture," *IEEE Sensors J.*, vol. 15, no. 2, pp. 414-425, Feb. 2015.

[5] P. Merriaux, Y. Dupuis, R. Boutteau, P. Vasseur, and X. Savatier, "A Study of Multiple IMU Association for Hand Pose Tracking," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1808-1816, Jul. 2017.

[6] H. Xue, Q. Cao, C. Miao, Y. Ju, H. Hu, A. Zhang, and L. Su, "Towards Generalized mmWave-based Human Pose Estimation through Signal Augmentation," in *Proc. 29th Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, pp. 1–15, Madrid, Spain, ACM, 2023, doi: 10.1145/3570361.3613302.

[7] 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 *Proc. ACM SIGGRAPH*, pp. 1-14, 2018.

[8] P. Choudhary, N. Goel, and M. Saini, "A Fingerprinting Based Audio-Seismic Systems for Human Target Localization in an Outdoor Environment Using Regression," *IEEE Sensors J.*, vol. 22, no. 8, pp. 7944-7960, Apr. 2022.

[9] P. Choudhary, P. Kumari, N. Goel, and M. Saini, "An Audio-Seismic Fusion Framework for Human Activity Recognition in an Outdoor Environment," *IEEE Sensors J.*, vol. 22, no. 23, pp. 22817-22827, Dec. 2022.

[10] P. Choudhary, N. Goel, and M. Saini, "A Seismic Sensor based Human Activity Recognition Framework using Deep Learning," in *Proc. 17th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, pp. 1-8, 2021.

[11] K. Qian, C. Wu, Z. Zhang, G. Zhang, and K. J. Hintz, "WiPose: Sensing Human Pose with WiFi," arXiv preprint arXiv:2103.12345, 2021.

[12] Y. Zhou, C. Xu, L. Zhao, A. Zhu, F. Hu, and Y. Li, "CSI-Former: Pay More Attention to Pose Estimation with WiFi," *Entropy*, vol. 25, no. 1, pp. 20, Jan. 2023.

[13] Y. Li, X. Li, M. Yang, and Y. Wang, "Person-in-WiFi: Fine-Grained Person Perception Using WiFi," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5678-5689, Jun. 2020.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pp. 7132-7141, 2018.

[15] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-Home Gesture Recognition Using Wireless Signals," in *Proc. ACM MobiCom*, pp. 201-212, 2013.

[16] H. Abdelnasser, K. A. Harras, and M. Youssef, "WiGest: A Ubiquitous WiFi-Based Gesture Recognition System," in *Proc. ACM MobiSys*, pp. 15-28, 2015.

[17] Y. Tong, C. Wu, P. Zhang, and X. Wang, "WiFiTrack: High-Precision Device-Free Localization Using Commercial WiFi Devices," *IEEE Trans. Mobile Comput.*, vol. 20, no. 5, pp. 1830-1843, May 2021.

[18] M. Nie, L. Zou, H. Cui, X. Zhou, and Y. Wan, "Enhancing Human Activity Recognition with LoRa Wireless RF Signal Preprocessing and Deep Learning," *Electronics*, vol. 13, no. 2, p. 264, Jan. 2024, doi: 10.3390/electronics13020264.

[19] M. Zhao, F. Adib, and D. Katabi, "Emotion Recognition Using Wireless Signals," in *Proc. ACM MobiCom*, pp. 1-15, 2020.