IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 16 July 2024, accepted 14 August 2024, date of publication 19 August 2024, date of current version 29 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3445465

RESEARCH ARTICLE

Human Pose Inference Using an Elevated mmWave FMCW Radar

STIRLING SCHOLES, ALICE RUGET, FENG ZHU, AND JONATHAN LEACH
School of Engineering and Physical Sciences, Heriot-Watt University, EH14 4AS Edinburgh, U.K.

Corresponding author: Jonathan Leach (j.leach@hw.ac.uk)

This work was supported in part by the Defence Science Technologies Laboratory through under Project Dstlx-1000147352, and in part by the Engineering and Physical Sciences Research Council (EPSRC) through under Grant EP/T00097X/1 and Grant EP/S026428/1.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Heriot-Watt University Ethics Committee, chaired by Prof. Jonathan D. Shepard.

ABSTRACT Human monitoring using radar systems operating in the GHz regime has generated significant interest as a result of the increasing availability of commercial radar systems. These sensors offer all weather performance, the ability to measure range and velocity, and the protection of anonymity. However, visually inferring activities present in radar data is often challenging without prior knowledge. Here, we address this by implementing a radar-to-pose system that converts the raw radar data into human poses, such that human forms can be identified and activities monitored. In comparison to prior works, we place our radar in an elevated position, more in line with the placement of existing real world monitoring systems e.g. cameras, or emerging systems, e.g. quadcopters. We present an ensemble predictor network and apply it to a number of human poses of increasing complexity, reporting accuracies in excess of 90%, and verify the generalizable nature of our approach with unseen validation data. We perform an in depth explainability analysis, exploiting the unique mappings created by our radar placement and network structure to confirm that the network is making rational predictions based on the true location of limbs.

INDEX TERMS Convolutional neural network, mmWave radar, human pose detection, explainable A.I., FMCW.

I. INTRODUCTION

The increasing availability of commercial radar systems operating in the GHz regime has recently generated significant interest in these sensors for human monitoring. Compared to optical sensors, radar based imaging techniques offer superior all weather performance and reduced privacy concerns. Additionally, the native ability of radars to record additional information, notably range and velocity, makes them appealing in several contexts. However, the often poor transverse resolution of commercially available radar systems means that the data they produce is often not directly human interpretable. This limitation is increasingly overcome with the use of machine learning approaches which convert the raw data into a more intuitive form, such as in the case of Human Activity Recognition (HAR).

Radar based HAR has seen significant recent progress with a common approach for recognition being the use of an activities micro-Doppler signature [1], [2], [3], [4]. Additionally, a large number of neural network architectures have been presented including; attention augmentation [5], [6], [7], [8], transformers [9], [10], convolutions [11], [12], [13], [14], [15], custom layers [16], and GANs to produce synthetic training data [17]. Whilst most HAR implementations feature a single radar, work has been presented showing the use of multi-radar distributed systems [18], [19], [20]. A complete analysis of radar based HAR is beyond the scope of this work and the reader is directed to relevant review sources, [21], [22], [23].

In comparison to HAR, the problem of converting radar data to human pose is less well investigated and

The associate editor coordinating the review of this manuscript and approving it for publication was Filbert Juwono.

presents unique challenges. Unlike HAR where activities can typically be recognised retrospectively, pose requires that the estimation takes place on roughly a frame-by-frame timescale. Additionally, whilst most HAR implementations only identify if a single action has taken place somewhere within the radar's Field-of-View (FoV), pose estimation requires the simultaneous spatial localization of several joints. This spatial localization is challenging given that many readily available radars operate as, or approximately as, 2D sensors. For instance, sensors which can determine the range and bearing (azimuth) to an object, but not its relative elevation. Despite these challenges, radar-to-pose has been achieved in a number of regimes notably, in looking through walls [24]. Single radar implementations of through wall systems have employed convolutional and transformer based networks [25], [26], [27], [28] and been used to identify the presence of multiple individuals [29], [30]. In this context, prior work has also examined reducing the complexity of the prediction networks [31] and performing HAR once a pose has been established [32].

Beyond through-wall applications, radar-to-pose has also been realised using micro-Doppler [33] and radar point cloud processing [34], [35] with other works focusing on shaping of input data [36] or optimizing predictions in cluttered environments [37]. In addition to single radar implementations, dual or multi-radar systems have also been reported. These multi sensor systems often overcome the 2D sensing nature of radars by rotating co-located sensors to orthogonal orientations such that they produce horizontal and vertical projections of the environment [38], [39], [40], [41]. Other implementations have explored distributed radar systems that view an area from multiple co-plannar perspectives [42], [43].

Inherent to the problem of radar-to-pose is a mechanism for capturing the ground truth position of limbs to create training labels for the radar data. This is most commonly achieved by widely available depth camera based image-to-pose systems such as the Microsoft Kinect [32], [44] or Intel RealSense [45] which have the benefit of being readily available. However, the accuracy of the ground truth positions output by these systems is often dependent on, occlusion, the operating environment, and the quality of the image-to-pose predictor being implemented. These limitations can be mitigated by using multiple cameras [40] or by using camera based motion capture technology which relies on optically tracking markers on a special garment [28], [46]. However, the use of multi-camera systems introduces challenges around cost, synchronisation, calibration, and cluttering the radar environment.

Existing radar-to-pose implementations focus on radars mounted on tripods co-planar with the ground. This is in contrast to most real world monitoring systems, e.g., security cameras, which are typically mounted in an elevated position. A limitation associated with the co-planar approach is that without the use of multiple or spatially separated radars, azimuthally symmetric human poses appear degenerate to 2D line sensors. Here, as shown conceptually in Fig. 1, we extend the prior work on single radar-to-pose detection by implementing a regime based on a single mmWave radar placed in an elevated position. In comparison to prior dual radar works using orthogonally rotated radars, our scheme allows a single radar with only range and azimuth sensitivity to measure a 3D space. This is achieved by projecting the elevation component into a range difference as shown in Fig. 1 R_0 and R_1. Additionally, this scheme improves the occlusion resistance of the radar for most human actions. We make use of a camera-free inertial sensor based motion capture suit to record ground truth limb positions without the aforementioned challenges and with cm scale accuracy.

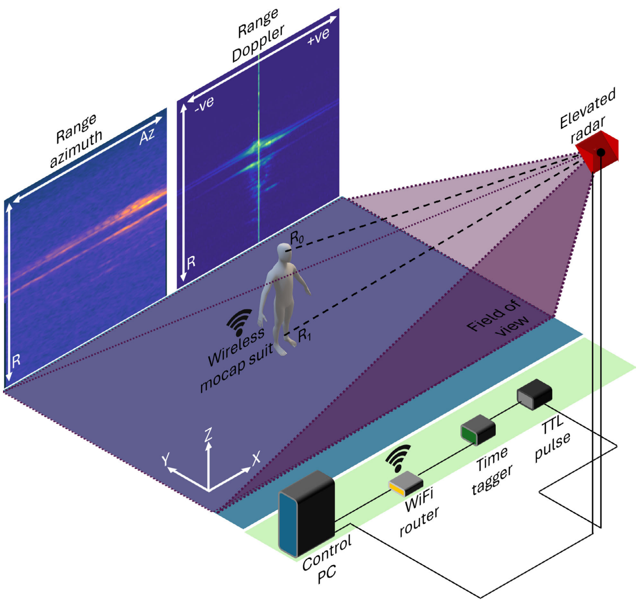


FIGURE 1. A conceptual representation of the experimental scenario. A mmWave radar is placed in an elevated position and declined at a 45° angle. This results in measurably distinct ranges R_0 and R_1 along the Z axis of objects within the Field-of-View. A markerless motion capture suit was worn by participants to record ground truth limb positions which were synchronised to the radar data by means of a time tagger system. The radar data was processed to output range-azimuth and range-Doppler depictions of the scene.

We present an ensemble predictor network which features 13 parallel independent networks for joint prediction and apply it to a number of human poses of increasing complexity. We experimentally evaluate the performance of this network and comment on the impacts of limb size and range of motion when making pose predictions. Additionally, we demonstrate the generalizable nature of our approach by applying a trained network to outdoor validation data featuring unseen individuals and backgrounds. Further, in contrast to prior works, we perform an explainability analysis on our network by using a novel ablation technique. This technique allows us to visualize the features of the input data which the network is using to make predictions. When combined with our unique radar placement, this analysis allows us to confirm that the network is making rational predictions based on the true location of limbs in the radar data.

II. EXPERIMENT AND NETWORK ARCHITECTURE

Our experimental setup is depicted conceptually in Fig. 1. We made use of a Texas Instruments TIDEP-01012 cascade Frequency Modulated Continous Wave (FMCW) radar. This radar consists of 4 AWR2243 radar modules each with 3 transmit and 4 receive channels monitoring an ≈ 70° azimuthal FoV. The modules are arrayed together in series to create a 12 × 16 Tx/Rx system. We operated our radar in a MIMO configuration summing over the elevation channels to improve the signal to noise ratio of the data. This configuration created a linear virtual antenna array with 86 elements having a boresight azimuthal resolution of 1.4° and no direct elevation resolution. The radar operated in the 77-81 GHz range with a bandwidth of 3.95 GHz corresponding to a range resolution of ≈ 3.6 cm. For the full radar specifications see supplementary materials S1.

At the start of each frame a trigger signal from the radar was conditioned into a TTL pulse by a Stanford Research Systems DG645 digital delay generator. This pulse was passed to the GPIO pins of a Raspberry Pi, which by means of a Python script, timestamped the pulse. Additionally, the first TTL pulse received by the Pi generated a UDP message that was passed through the local network of the WiFi router to the control PC. This UDP message was used to commence recording of the motion capture suit data in sync with the radar. The motion capture data was recorded using a marker free mocap system specifically, a Rokoko smartsuit pro 2 which communicated by WiFi with the control PC via the local network of the router.

Each frame of radar data was acquired in a 200 ms period and consisted of a 468 × 128 × 86 data structure with the format; samples-per-chirp, chirps-per-frame, virtual channels. By taking Fourier transforms along each axis of the data a range-azimuth-Doppler structure was created. This data structure was then used to create a two-dimensional range-azimuth representation of moving objects in the scene by summing over the non-zero Doppler components of the range-azimuth-Doppler data. Similarly, a range-Doppler representation was created by summing the azimuthal components. By representing the radar data as two-dimensional structures we were able to implement a neural network designed around convolution layers which are known for being effective in extracting features from image type data.

We present a network architecture built on an ensemble predictor. Specifically, in contrast to prior works where a single network is used to predict the position of all joints in the skeleton, we employ an ensemble of 13 identical independent networks, one for each joint. By maintaining independence between the networks, we could ensure that they learnt features in the radar data rather than becoming dependent on the inputs of companion networks. The structure of these networks is shown in Fig. 2. This ensemble approach means multiple joints can be identified simultaneously through network parallelization. Further, since each network is responsible for only a single joint, simpler networks can be used which reduces computation time when making estimates. Additionally, by discretizing the problem to individual coordinate outputs across multiple networks we are able to better isolate the features in the data used by the networks to make predictions, aiding in network explainability. The networks take in range-azimuth and range-Doppler depictions of the scene grouped into stacks of 3 consecutive frames which introduces a constant offset of 600 ms in the predictions. Using convolution and pooling the inputs are reduced to a common dimensiotabnality and concatenated. Additional convolution and pooling is applied to reduce the data to a dense latent space which is flattened and connected to a dense layer. The final dense layer contains 3 nodes corresponding to the X,Y, and Z coordinates of the joint.

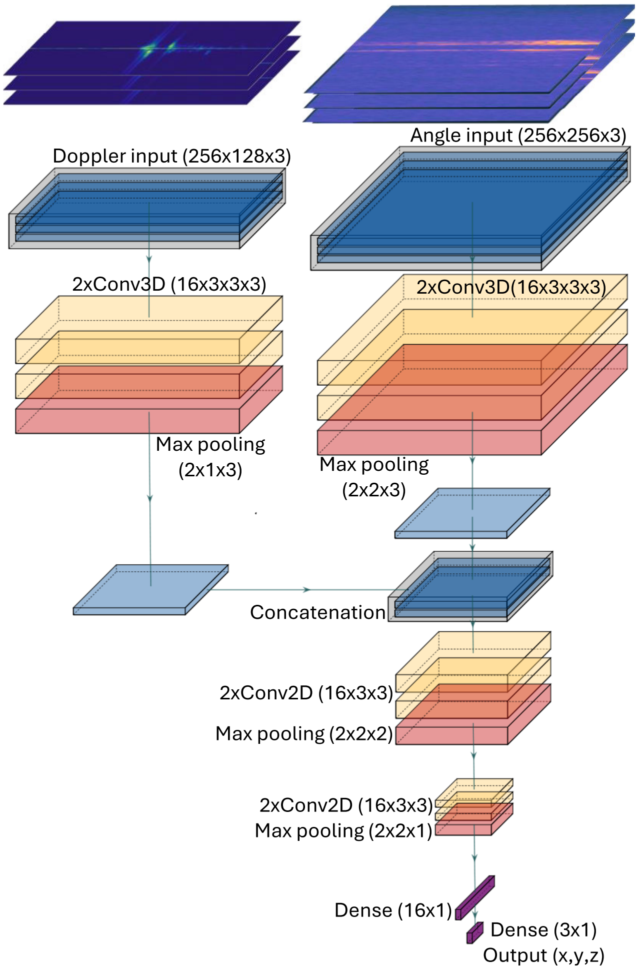


FIGURE 2. A summary of the network structure, here a single joint predicting network is shown. The networks take in range-azimuth and range-Doppler depictions of the scene grouped into stacks of 3 consecutive frames. Using convolution and pooling the inputs are reduced to a common dimensionality and concatenated. Additional convolution and pooling is applied to reduce the data to a dense latent space which is flattened and connected to a dense layer. The final dense layer contains 3 nodes corresponding to the X,Y, and Z coordinates of the joint.

To train the networks, the ground truth coordinates of all joints from the motion capture suit were normalized to values between -1 and 1 based on the FoV of the radar.

Further, the positions of all of the body joints were expressed in terms of their locations relative to the position of the head and re-normalized to span between −1 and 1. This was done to ensure that even small relative motions, such as arm movements during walking, are seen as significant changes in label values as opposed to small fluctuations around the global position of the joints [47]. The convolutional layers used ReLu activation whilst the final output layer used a tanh activation to match the −1 to 1 range of the training labels. Training loss was calculated as the root-mean-square error between the output of the network and the training label whilst using the Adam optimizer. The networks were monitored whilst being trained for at least 50 epochs and the best performing networks selected. A flowchart of this process is given in supplementary material S2.

III. RESULTS

Before conducting our experiments we verified that the motion capture suit would not provide an anomalously large radar return, see supplementary material S4. Additionally, informed consent was gained from all trial participants. To verify the functionality of our approach we first consider the case of arm motion in a single position. We select this action to explore the functionality of the network in characterising small features e.g. hands within a relatively local space. The participant was positioned in the centre of the 2.5×2×2.5 (X,Y,Z) m region of a squash court illuminated by the radar. A squash court was selected to mitigate the impact of clutter from the environment. The radar was placed in an elevated position using an observation balcony overlooking the squash court with an R_0 value of ≈ 4.8 m. The participant laterally raised an extended arm from a rest position at their side to over their head. This motion was repeated for both arms individually as well as both arms together. An initial dataset of 1296 frames was gathered for training the network, with an additional 432 frames gathered for validation.

Figure 3 shows a pose skeleton predicted by the network for a frame of validation data. The inset subpanel shows a reference photo of the pose from the perspective of the radar. The coloured dots in the skeleton correspond to the ground truth positions of the joints. The coloured lines indicate the predicted skeleton. The good overlap between the predicted and ground truth joint positions confirms the ability of the network to learn the relationship between the data from the elevated radar and the 3D positions of joints, even for cases of relatively little training data. A video showing the networks predictions on all validation data is available as supplementary material S5.

Table 1 quantifies the agreement between the predicted and ground truth skeletons on a per-joint-per-coordinate basis. The table shows overall errors of 2.0, 1.4, and 1.0 cm for the X,Y, and Z axes respectively, in line with previously reported single radar-to-pose systems [48]. A root mean squared error analysis is given in supplementary material S3. These values confirm the viability of elevated radars for human pose prediction. Table 1 also illustrates that accuracy is not necessarily uniform across all joints or axes despite all predictors sharing the same structure and training regime. Notably, the mean validation error and associated standard deviation is largest for the hands in the Z axis. This is consistent with both the small radar cross section of the hands [49], [50], [51] as well as hand motion in the Z axes having the largest range of motion in this case. Each network in the ensemble takes ≈ 9 μs (on an Nvidia 3090 GPU) to make a prediction on a single frame. Given that this

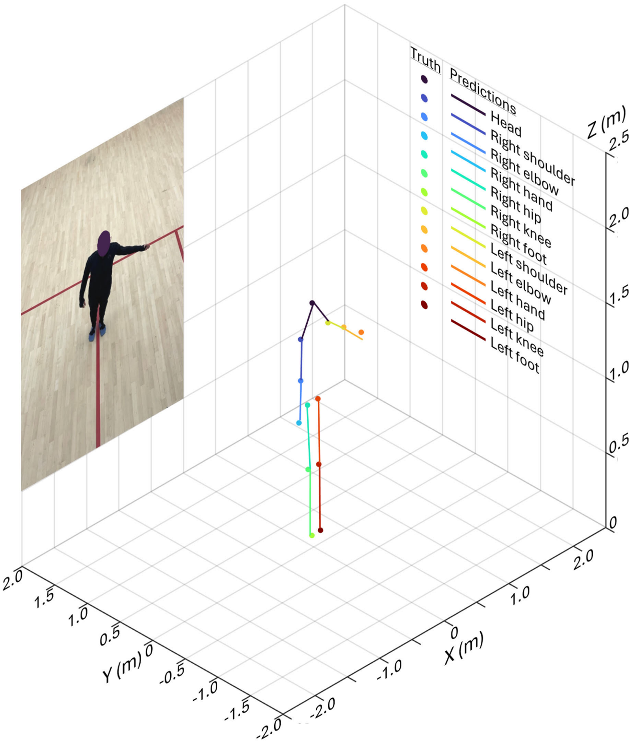


FIGURE 3. The skeleton predicted by the network for single location arm movements. The dots correspond to the ground truth positions of the joints, whilst the lines correspond to the network predictions. The subpanel shows a reference photo of the pose from the perspective of the radar. The figure has been plotted using the same perspective as Fig. 1.

TABLE 1. Per-coordinate-per-joint error for arm motion in a single position.

Joint	Mean validation error (cm) ± std		
	X	Y	Z
Head	2.0 ± 1.7	1.4 ± 0.9	0.7 ± 0.5
Right shoulder	1.5 ± 1.3	1.8 ± 1.2	0.8 ± 0.6
Right elbow	2.8 ± 2.4	1.5 ± 1.2	2.3 ± 3.0
Right hand	4.4 ± 4.2	2.8 ± 2.3	7.2 ± 10.5
Right hip	1.1 ± 0.9	0.9 ± 0.8	0.8 ± 0.6
Right knee	1.3 ± 1.1	1.1 ± 0.7	1.1 ± 2.5
Right foot	1.2 ± 0.8	0.6 ± 0.5	0.7 ± 0.6
Left shoulder	1.7 ± 1.6	1.1 ± 0.8	0.8 ± 0.6
Left elbow	2.5 ± 2.4	2.0 ± 1.4	3.5 ± 4.6
Left hand	4.6 ± 5.4	2.9 ± 2.1	7.1 ± 6.4
Left hip	1.1 ± 0.8	0.8 ± 0.6	0.8 ± 0.7
Left knee	1.1 ± 0.9	0.7 ± 0.5	0.7 ± 0.6
Left foot	1.1 ± 0.9	0.5 ± 0.4	0.7 ± 0.6
Overall average	2.0 ± 1.9	1.4 ± 1.0	2.1 ± 2.4

is less than the 200 ms frame acquisition time, real time processing would be feasible with sufficient optimization of the implementation.

Having confirmed the viability of our system we examined its ability to characterise an individual at arbitrary locations within the FoV. The network was retrained on a larger dataset of 10800 training frames (with 576 validation frames) of an individual walking within the FoV of the radar. In comparison to lateral arm movements, walking represents the movement of a large feature (the body) throughout a large range of motion. Figure 4 shows pose skeletons predicted by the network. Note that two separate frames of validation data for the same individual have been plotted on the same axes. The inset subpanel shows a reference photo of the pose from the perspective of the radar. The coloured dots in the skeleton correspond to the ground truth positions of the joints. The coloured lines indicate the predicted skeleton. The good agreement between the predicted and ground truth skeletons demonstrates the ability of our ensemble network architecture to generalize to multiple activities. A video showing the networks predictions on all validation data is available as supplementary material S6.

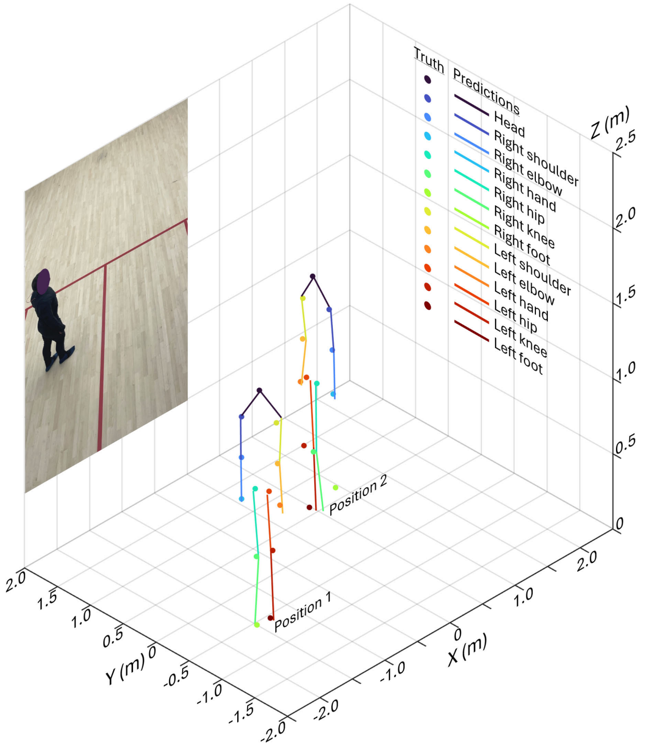


FIGURE 4. The skeleton predicted by the network for a single individual walking. The two skeletons correspond to a single individual at two different points in time. The dots correspond to the ground truth positions of the joints, whilst the lines correspond to the network predictions. The subpanel shows a reference photo of one of the poses from the perspective of the radar. The figure has been plotted using the same perspective as Fig. 1.

Figure 5 quantifies the error associated with the walking case. Specifically, each panel corresponds to an axis with the coloured lines and the bottom x-axis of each panel showing the fraction of validation predictions within a given absolute error. The upper x-axis maps the absolute error to a relative accuracy based on the field of view. The dashed lines show the average error and accuracy associated with an 80% threshold for each coordinate. Figure 5 visualises the distribution of errors associated with each joint and axes showing that for this case, on average, 80% of network predictions are within 36, 32, and 7 cm of their true values for the X,Y, and Z axes respectively. When considered relative to the total FoV of the radar, these errors represent network prediction accuracies of 93%, 92%, and 99% respectively. Comparing table 1 to Fig. 5 shows that adding the requirement of general localization to the network somewhat impacts performance, particularly in the X and Y axes where the change in the scope of the motion is the most significant. Further, in contrast to table 1, the errors associated with low radar cross section joints in the body, e.g. the hands, is not significantly different from the overall localisation error. This implies that for cases where the relative motion of body joints is small, the distribution of errors across the ensemble network will be approximately uniform.

Next we examine a case where the relative motion of body joints is large in the context of general localization. The participant was placed at 20 pseudo random locations within the FoV and performed a ‘clapping’ action by cyclically moving their extended arms coplanar with the floor from an outstretched position at their sides to in front of their chest. For each location within the FoV this motion was repeated in 4 directions facing; towards the radar, away from the radar, in the positive X direction, and, in the negative X direction. A dataset of 11376 training frames (with 576 validation frames) was used for training the network.

Figure 6 shows pose skeletons predicted by the network. Note that two separate frames of validation data for the same individual have been plotted on the same axes. The inset subpanel shows a reference photo of the pose from the perspective of the radar. The coloured dots in the skeleton correspond to the ground truth positions of the joints. The coloured lines indicate the predicted skeleton. The distinct poses of the predicted skeletons in Fig. 6 demonstrates the networks ability to reconstruct individuals at a variety of angles and positions within the FoV of the radar. A video showing the networks predictions on all validation data is available as supplementary material S7.

Figure 7 visualises the distribution of errors associated with each joint and axes for the clapping case in the same manner as Figure 5. For the clapping case, on average, 80% of network predictions are within 28, 23, and 3 cm of their true values corresponding to accuracies of 94%, 94%, and 99% for the X,Y, and Z axes respectively. Comparing Figs. 5 and 7 it can be seen that although the later case exhibits a small improvement in 80% average accuracy, the spread of errors across the joints is much larger. This spread of errors can be characterised by a set of networks predicting the global position of otherwise stationary joints, such as the head and body, and networks predicting the position of

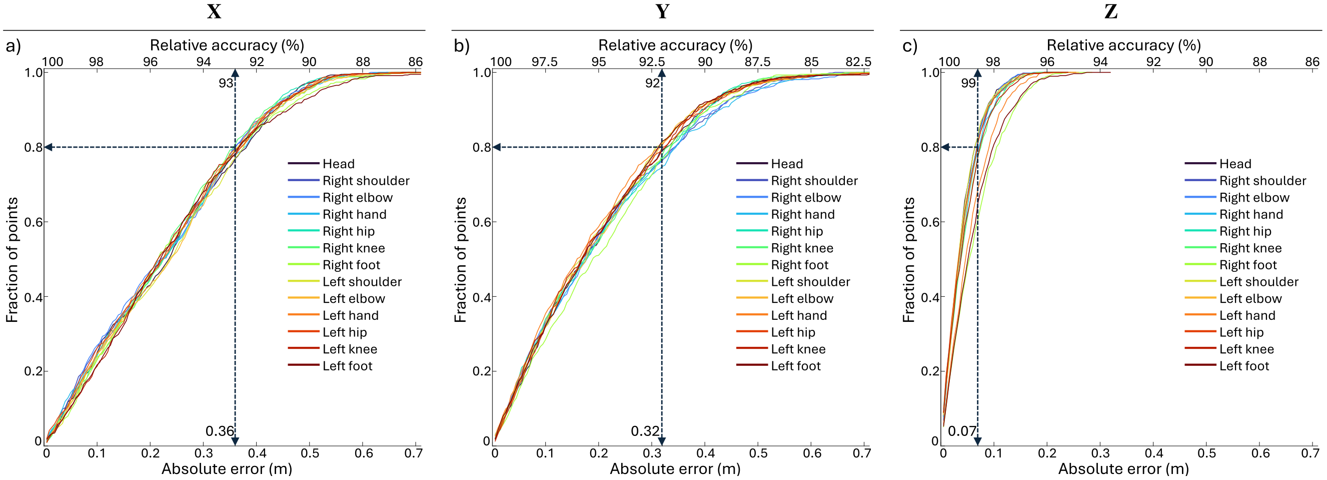


FIGURE 5. The fractional joint wise error associated with a single individual walking. Panels a), b), and c) show the error associated with the X,Y, and Z coordinates respectively. The coloured lines and the bottom x-axis of each panel show the fraction of validation predictions within a given absolute error. The upper x-axis maps the absolute error to a relative accuracy based on the field of view. The dashed lines show the average error and accuracy associated with an 80% threshold for each coordinate.

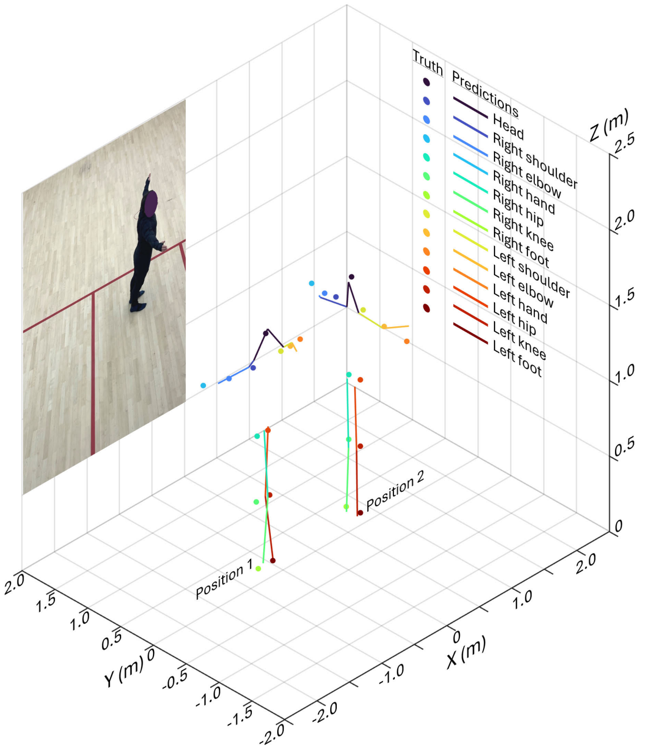


FIGURE 6. The skeleton predicted by the network for a single individual clapping in several directions. The two skeletons correspond to a single individual at two different points in time. The dots correspond to the ground truth positions of the joints, whilst the lines correspond to the network predictions. The subpanel shows a reference photo of one of the poses from the perspective of the radar. The figure has been plotted using the same perspective as Fig. 1.

joints with large motions relative to the head, in this case the joints in the arms. The set of networks predicting joints with small relative movements cluster together, exhibiting similar accuracies, and so they dominate the overall 80% average. By contrast, networks predicting joints with large relative motion, and in particular those joints with small radar cross sections, exhibit larger errors in prediction. This distribution of errors is closely related to the radar cross section of human limbs, since this directly impacts the visibility of the limbs as features in the training data and highlights one of the challenges of using relatively low power consumer radars for human pose estimation.

A. NETWORK ANALYSIS AND DISCUSSION

To validate our approach we perform two additional studies on our system. First, we apply the trained network to data featuring an unseen background and an unseen human subject. The new environment consisted of placing the radar outside of a first story window to view a street below. This configuration is analogous to both existing monitoring systems e.g. security cameras, and emerging systems, such as pose monitoring from drones [52], [53]. The height of the radar above ground, the acquisition time of the radar frames, and the absence of clutter from the scene was maintained from the training scenarios. Supplementary information S8 contains additional details on the new environment. An individual who was not present in the training data then walked within the FoV of the radar. Figure 8 shows pose skeletons predicted by the network. Note that two separate frames of validation data for the same individual have been plotted on the same axes. The inset subpanel shows a reference photo of the pose from the perspective of the radar. The coloured lines indicate the predicted skeleton. Note that no ground truth is available as the participant was not wearing the motion capture suit. This decision was made to confirm that the network had learnt inherent human features, as opposed to suit features, and that the network would be robust against real world metallic clutter such as keys, phones, and jewellery. Qualitative examination of Fig. 8 shows skeletons of comparable quality to those of Fig. 4

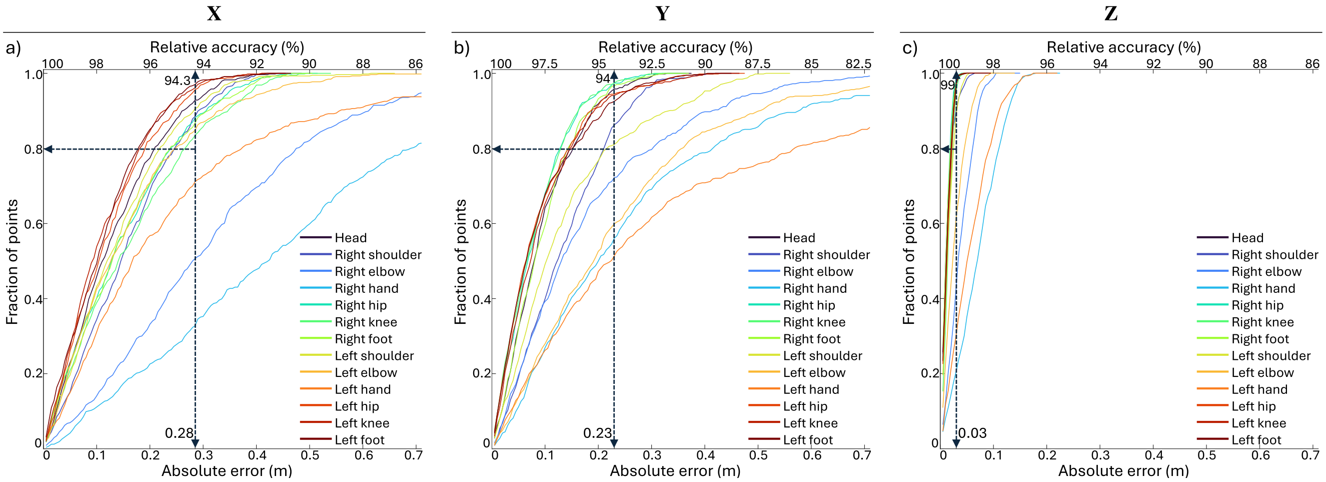


FIGURE 7. The fractional joint wise error associated with a single individual clapping. Panels a), b), and c) show the error associated with the X,Y, and Z coordinates respectively. The coloured lines and the bottom x-axis of each panel show the fraction of validation predictions within a given absolute error. The upper x-axis maps the absolute error to a relative accuracy based on the field of view. The dashed lines show the average error and accuracy associated with an 80% threshold for each coordinate.

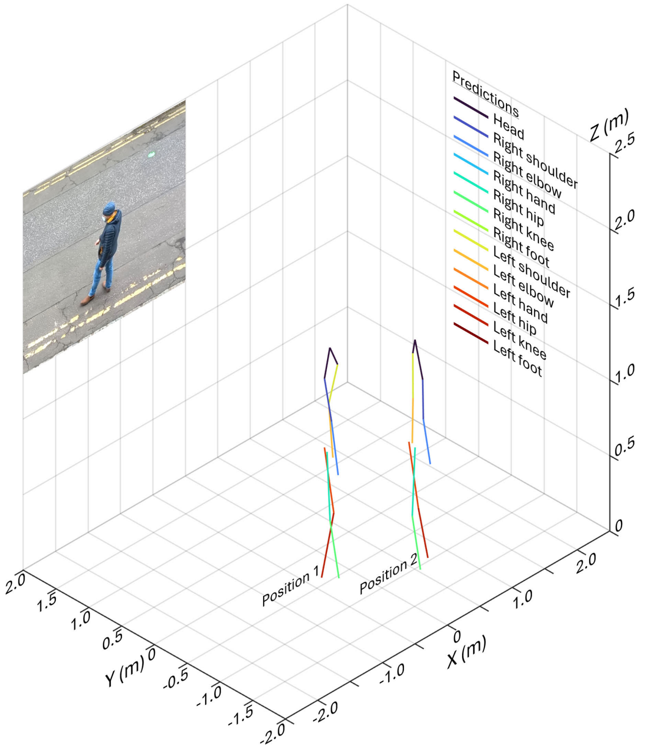


FIGURE 8. The skeletons predicted by the network for a single individual who was not in the training data whilst in an outdoor environment. The two skeletons correspond to a single individual walking outdoors at two different points in time. The lines correspond to the network predictions. No ground truth is available as the subject was not using the motion capture suit. The subpanel shows a reference photo from the perspective of the radar. The figure has been plotted using the same perspective as Fig. 1.

confirming that with sufficiently diverse training data and adherence to salient factors such as the height above ground of the radar and environmental clutter our approach could be generalized to a wide range of scenarios and individuals.

Second, we exploit the unique characteristics of our radar placement and ensemble network to conduct a novel ablation analysis. The ablation study we performed focused on validation data collected from the walking trial but could be applied to any of the cases we performed. The first phase of the ablation analysis was to determine which of the two network inputs is more significant in pose prediction. To compare the importance of the range-azimuth input to the range-Doppler input, each was sequentially set to zero and the performance of the pretrained network on the walking data evaluated. It was found that whilst removal of the range-Doppler data noticeably degraded the performance of the network, the impact of removing the range-azimuth input was more pronounced. This result is consistent with the observation that the range-Doppler maps do not natively encode azimuthal information about the global location of objects. Consequently, the range-Doppler input alone is insufficient to unambiguously determine the azimuthal location of objects moving freely within the FoV.

The next phase of the ablation study allows us to isolate and identify the specific regions within the range-azimuth input that are responsible for the networks predictions. This is achieved by unambiguously projecting the ground truth locations of the joints onto the range-azimuth map. This procedure is enabled by the known trigonometric relationships introduced by our novel radar placement and the occlusion proof motion capture system we use. The projection operation allows us to directly visualize the regions of the range-azimuth maps which correspond to real joint locations.

Figure 9 shows a ground truth skeleton projected onto the range-azimuth map corresponding to the reference inset seen in Fig. 4. Note how the height of the skeleton, i.e., the difference between the head and the feet appears as a distribution in range due to the placement of our radar (R_0 and R_1 in Fig. 1). Next, we apply the following operations:

i We generate a set of 2^{13} binary masks, these masks have a dimensionality of 256 × 256 to match the range-azimuth input and consist of randomly assigned 4 × 4 regions of either 1 or 0.

ii The single range-azimuth frame shown in Fig. 9 is multiplied by the set of masks to create 2^{13} modified frames.

iii The modified frames are passed through the pretrained network and the new predictions recorded.

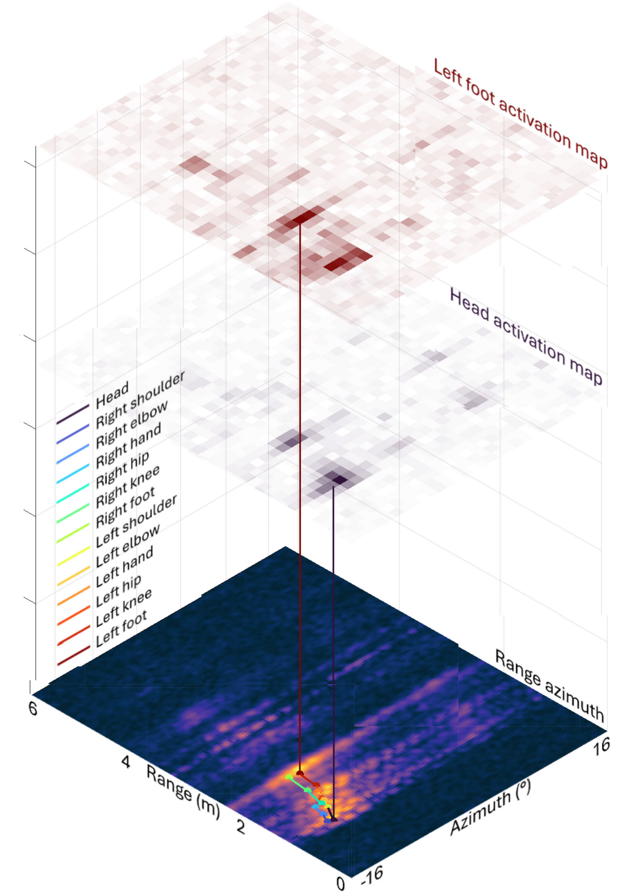


FIGURE 9. The activation maps for the head and left foot predicting networks. The 'floor' of the figure shows the range-azimuth representation of the scene. The skeleton is projected onto the range-azimuth using only the ground truth data. The coloured risers show the ground truth positions of the joints on the activation maps. The figure has been plotted using the same perspective as Fig. 1.

iv A 3×2^{13} element vector is created by calculating the difference between each component of the new predictions and the original prediction for the unmodified frame.

v This vector is then treated as a list of weights corresponding to how much a given mask impacts the networks prediction.

vi Each of the binary masks is then multiplied by its corresponding weight (minus the average of all weights) to produce a data structure with size $3 \times 256 \times 256 \times 2^{13}$.

vii The data structure is summed along its final axis to produce three activation maps with dimensionality 256 × 256.

viii This procedure is repeated for each network i.e., joint in the ensemble.

The procedure outlined above is conceptually similar to a simpler implementation of the LIME protocol [54] and is essentially a decomposition of the range-azimuth frame into a random pixel basis [55], [56]. By mapping the ground truth joint positions onto the radar data, we are able to associate distinct regions of pixels in the input with joints. By then creating the activation maps for each network we are able to determine which regions of pixels in the radar data the prediction of the network is most responsive to. Combined, these two observations allow us to infer that the network is most responsive to features in the radar data which correspond to the ground truth joint positions.

Figure 9 shows the total activation maps, i.e., the sum of the three component activation maps, for the head and left foot predicting networks for the corresponding range-azimuth frame. These activation maps show which regions of the range-azimuth frame are most influential in the networks prediction process. The coloured risers in Fig. 9 show the ground truth positions of the joints on the activation maps. From Fig. 9 it can be seen that the network responsible for predicting the location of the head is most sensitive to a region of the range-azimuth frame nearest the true position of the head. Similarly, the network responsible for predicting the location of the left foot is most sensitive to a region near the true position of the foot. Interestingly, both networks are somewhat aware of the positions of other features, the feet in the case of the head network and vice-verse for the foot network. This reciprocity occurs despite the networks being totally independent from one another but it is consistent with the observation that for the walking dataset the position of the feet and the position of the head are well correlated. Here, we have chosen to only show the activation maps for two networks for a single frame to aid in clarity of visualization and explanation. However, we stress that we have observed similar trends in activation across all networks for all tested validation frames in the walking dataset. The activation maps for the remaining networks and validation frames are available as a video in supplementary material S9. Additionally, we believe this type of mask based ablation study to be compatible with most image processing neural networks allowing for improved insight into the functioning of image-to-pose type classifiers.

IV. CONCLUSION

We extend the prior work on single radar-to-pose detection by implementing a regime based on a single elevated mmWave radar. We present an ensemble predictor network and apply it to a number of human poses of increasing complexity, reporting accuracies in excess of 90%. We demonstrate that the accuracy of network predictions is closely tied to the radar cross section of limbs and their relative range of motion in a scene. By applying a trained network to unseen validation data we demonstrate the generalizable nature of our approach. We perform an in depth explainability analysis, exploiting the unique mappings created by our radar placement and network structure to confirm that the network is making rational predictions based on the true location of limbs. We believe that this work demonstrates the viability of pose prediction using an elevated radar, a finding which could see radars augment existing elevated security sensors, i.e., cameras, or be implemented on new elevated platforms, such as quadcopter drones.

ACKNOWLEDGMENT

The authors also extend their thanks to the Oriam Sports Hall for the use of their facilities and to Mr. Willie Nel for insightful initial explanations of radar.

(Stirling Scholes and Alice Ruget contributed equally to this work.)

REFERENCES

[1] B. Çagliyan and S. Z. Gürbüz, "Micro-Doppler-based human activity classification using the mote-scale BumbleBee radar," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2135–2139, Oct. 2015.

[2] F. J. Abdu, Y. Zhang, and Z. Deng, "Activity classification based on feature fusion of FMCW radar human motion micro-Doppler signatures," *IEEE Sensors J.*, vol. 22, no. 9, pp. 8648–8662, May 2022.

[3] M. Chakraborty, H. C. Kumawat, S. V. Dhavale, and A. B. Raj, "Application of DNN for radar micro-Doppler signature-based human suspicious activity recognition," *Pattern Recognit. Lett.*, vol. 162, pp. 1–6, Oct. 2022.

[4] X. Li, Y. He, F. Fioranelli, and X. Jing, "Semisupervised human activity recognition with radar micro-Doppler signatures," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 9467531.

[5] C. Campbell and F. Ahmad, "Attention-augmented convolutional autoencoder for radar-based human activity recognition," in *Proc. IEEE Int. Radar Conf. (RADAR)*, Apr. 2020, pp. 990–995.

[6] L. Cao, S. Liang, Z. Zhao, D. Wang, C. Fu, and K. Du, "Human activity recognition method based on FMCW radar sensor with multi-domain feature attention fusion network," *Sensors*, vol. 23, no. 11, p. 5100, May 2023.

[7] S. Huan, L. Wu, M. Zhang, Z. Wang, and C. Yang, "Radar human activity recognition with an attention-based deep learning network," *Sensors*, vol. 23, no. 6, p. 3185, Mar. 2023.

[8] G. Lai, X. Lou, and W. Ye, "Radar-based human activity recognition with 1-D dense attention network," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[9] S. Huan, Z. Wang, X. Wang, L. Wu, X. Yang, H. Huang, and G. E. Dai, "A lightweight hybrid vision transformer network for radar-based human activity recognition," *Sci. Rep.*, vol. 13, no. 1, p. 12, Oct. 2023.

[10] X. Li, S. Chen, S. Zhang, L. Hou, Y. Zhu, and Z. Xiao, "Human activity recognition using IR-UWB radar: A lightweight transformer approach," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[11] X. Bai, Y. Hui, L. Wang, and F. Zhou, "Radar-based human gait recognition using dual-channel deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9767–9778, Dec. 2019.

[12] H. Du, T. Jin, Y. He, Y. Song, and Y. Dai, "Segmented convolutional gated recurrent neural networks for human activity recognition in ultra-wideband radar," *Neurocomputing*, vol. 396, pp. 451–464, Jul. 2020.

[13] W.-Y. Kim and D.-H. Seo, "Radar-based human activity recognition combining range–time–Doppler maps and range-distributed-convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 9743913.

[14] V. Lafontaine, K. Bouchard, J. Maítre, and S. Gaboury, "Denoising UWB radar data for human activity recognition using convolutional autoencoders," *IEEE Access*, vol. 11, pp. 81298–81309, 2023.

[15] A. Helen Victoria and G. Maragatham, "Activity recognition of FMCW radar human signatures using tower convolutional neural networks," *Wireless Netw.*, vol. 5, p. 17, Jun. 2021.

[16] Z. Sadeghi Adl and F. Ahmad, "Whitening-aided learning from radar micro-Doppler signatures for human activity recognition," *Sensors*, vol. 23, no. 17, p. 7486, Aug. 2023.

[17] B. Erol, S. Z. Gurbuz, and M. G. Amin, "GAN-based synthetic radar micro-Doppler augmentations for improved human activity recognition," in *Proc. IEEE Radar Conf. (RadarConf)*, Apr. 2019, pp. 1–5.

[18] G. Bhavanasi, L. Werthen-Brabants, T. Dhaene, and I. Couckuyt, "Patient activity recognition using radar sensors and machine learning," *Neural Comput. Appl.*, vol. 34, no. 18, pp. 16033–16048, Sep. 2022.

[19] R. G. Guendel, F. Fioranelli, and A. Yarovoy, "Distributed radar fusion and recurrent networks for classification of continuous human activities," *IET Radar, Sonar Navigat.*, vol. 16, no. 7, pp. 1144–1161, Jul. 2022.

[20] S. Waqar, M. Muaaz, and M. Pätzold, "Direction-independent human activity recognition using a distributed MIMO radar system and deep learning," *IEEE Sensors J.*, vol. 23, no. 20, pp. 24916–24929, Oct. 2023.

[21] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sens.*, vol. 11, no. 9, p. 1068, May 2019.

[22] K. Papadopoulos and M. Jelali, "A comparative study on recent progress of machine learning-based human activity recognition with radar," *Appl. Sci.*, vol. 13, no. 23, p. 12728, Nov. 2023.

[23] I. Ullmann, R. G. Guendel, N. C. Kruse, F. Fioranelli, and A. Yarovoy, "A survey on radar-based continuous human activity recognition," *IEEE J. Microw.*, vol. 3, no. 3, pp. 938–950, Jul. 2023.

[24] M. Zhao, T. Li, M. A. Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2018, pp. 7356–7365.

[25] Z. Zheng, D. Zhang, X. Liang, X. Liu, and G. Fang, "RadarFormer: End-to-end human perception with through-wall radar and transformers," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 22, 2023, doi: 10.1109/TNNLS.2023.3314031.

[26] Z. Zheng, J. Pan, D. Zhang, X. Liang, X. Liu, and G. Fang, "Through-wall human pose estimation by mutual information maximizing deeply supervised nets," *IEEE Internet Things J.*, vol. 11, no. 2, pp. 3190–3205, Jan. 2024.

[27] Z. Zheng, J. Pan, Z. Ni, C. Shi, D. Zhang, X. Liu, and G. Fang, "Recovering human pose and shape from through-the-wall radar images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 9741729.

[28] G. W. Kim, S. W. Lee, H. Y. Son, and K. W. Choi, "A study on 3D human pose estimation using through-wall IR-UWB radar and transformer," *IEEE Access*, vol. 11, pp. 15082–15095, 2023.

[29] Z. Zheng, D. Zhang, X. Liang, X. Liu, and G. Fang, "Through-wall human pose reconstruction based on cross-modal learning and self-supervised learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[30] Y. Song, T. Jin, Y. Dai, Y. Song, and X. Zhou, "Through-wall human pose reconstruction via UWB MIMO radar and 3D CNN," *Remote Sens.*, vol. 13, no. 2, p. 241, Jan. 2021.

[31] Y. Song, T. Jin, Y. Dai, and X. Zhou, "Efficient through-wall human pose reconstruction using UWB MIMO radar," *IEEE Antennas Wireless Propag. Lett.*, vol. 21, no. 3, pp. 571–575, Mar. 2022.

[32] Y. Song, Y. Dai, T. Jin, and Y. Song, "Dual-task human activity sensing for pose reconstruction and action recognition using 4-D imaging radar," *IEEE Sensors J.*, vol. 23, no. 19, pp. 23927–23940, Oct. 2023.

[33] X. Zhou, T. Jin, Y. Dai, Y. Song, and Z. Qiu, "MD-pose: Human pose estimation for single-channel UWB radar," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 5, no. 4, pp. 449–463, Oct. 2023.

[34] J. Zhong, L. Jin, and R. Wang, "Point-convolution-based human skeletal pose estimation on millimetre wave frequency modulated continuous wave multiple-input multiple-output radar," *IET Biometrics*, vol. 11, no. 4, pp. 333–342, Jul. 2022.

[35] S. Wang, D. Cao, R. Liu, W. Jiang, T. Yao, and C. X. Lu, "Human parsing with joint learning for dynamic mmWave radar point cloud," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 7, no. 1, pp. 1–22, Mar. 2023.

[36] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "Mm-pose: Real-time human skeletal posture estimation using mmWave radars and CNNs," *IEEE Sensors J.*, vol. 20, no. 17, pp. 10032–10044, Sep. 2020.

[37] Z. Cao, J. Zhang, R. Chen, X. Guo, and G. Wang, "Task-specific feature purifying in radar-based human pose estimation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 59, no. 6, pp. 9285–9298, Dec. 2023.

[38] G. Li, Z. Zhang, H. Yang, J. Pan, D. Chen, and J. Zhang, "Capturing human pose using mmWave radar," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, New York, NY, USA, Mar. 2020, pp. 1–6.

[39] Z. Zeng, X. Liang, Y. Li, and X. Dang, "Vulnerable road user skeletal pose estimation using mmWave radars," *Remote Sens.*, vol. 16, no. 4, p. 633, Feb. 2024.

[40] C. Xie, D. Zhang, Z. Wu, C. Yu, Y. Hu, and Y. Chen, "RPM: RF-based pose machines," *IEEE Trans. Multimedia*, vol. 26, pp. 637–649, 2024.

[41] C. Xie, D. Zhang, Z. Wu, C. Yu, Y. Hu, and Y. Chen, "RPM 2.0: RF-based pose machines for multi-person 3D pose estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 490–503, Jan. 2024.

[42] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3D human pose construction using WiFi," in *Proc. 26th Annu. Int. Conf. Mobile Comput. Netw.*, New York, NY, USA, Apr. 2020, pp. 295–308.

[43] L. Chen, X. Guo, and G. Wang, "MPTFormer: Toward robust arm gesture pose tracking using dual-view radar system," *IEEE Sensors J.*, vol. 24, no. 1, pp. 1051–1064, Jan. 2024.

[44] A. Ruget, M. Tyler, G. M. Martín, S. Scholes, F. Zhu, I. Gyongy, B. Hearn, S. McLaughlin, A. Halimi, and J. Leach, "Pixels2Pose: Super-resolution time-of-flight imaging for 3D pose estimation," *Sci. Adv.*, vol. 8, no. 48, Dec. 2022, Art. no. eade0123.

[45] W. Ding, Z. Cao, J. Zhang, R. Chen, X. Guo, and G. Wang, "Radar-based 3D human skeleton estimation by kinematic constrained learning," *IEEE Sensors J.*, vol. 21, no. 20, pp. 23174–23184, Oct. 2021.

[46] M. Mahbubur Rahman, D. Martelli, and S. Z. Gurbuz, "Radar-based human skeleton estimation with CNN-LSTM network trained with limited data," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Oct. 2023, pp. 1–4.

[47] I. Katircioglu, B. Tekin, M. Salzmann, V. Lepetit, and P. Fua, "Learning latent representations of 3D human pose with deep neural networks," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1326–1341, Dec. 2018.

[48] Z. Cao, W. Ding, R. Chen, J. Zhang, X. Guo, and G. Wang, "A joint global–local network for human pose estimation with millimeter wave radar," *IEEE Internet Things J.*, vol. 10, no. 1, pp. 434–446, Jan. 2023.

[49] Y. Deep, P. Held, S. S. Ram, D. Steinhauser, A. Gupta, F. Gruson, A. Koch, and A. Roy, "Radar cross-sections of pedestrians at automotive radar frequencies using ray tracing and point scatterer modelling," *IET Radar, Sonar Navigat.*, vol. 14, no. 6, pp. 833–844, Jun. 2020.

[50] P. Hügler, M. Geiger, and C. Waldschmidt, "RCS measurements of a human hand for radar-based gesture recognition at E-band," in *Proc. German Microw. Conf. (GeMiC)*, Mar. 2016, pp. 259–262.

[51] M. Yasugi, Y. Cao, K. Kobayashi, T. Morita, T. Kishigami, and Y. Nakagawa, "79 GHz-band radar cross section measurement for pedestrian detection," in *Proc. Asia–Pacific Microw. Conf. (APMC)*, Nov. 2013, pp. 576–578.

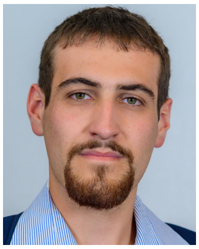
[52] Z. Marinov, S. Vasileva, Q. Wang, C. Seibold, J. Zhang, and R. Stiefelhagen, "Pose2Drone: A skeleton-pose-based framework for human-drone interaction," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 776–780.

[53] D. Palossi, N. Zimmerman, A. Burrello, F. Conti, H. Müller, L. M. Gambardella, L. Benini, A. Giusti, and J. Guzzi, "Fully onboard AI-powered human-drone pose estimation on ultralow-power autonomous flying nano-UAVs," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 1913–1929, Feb. 2022.

[54] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

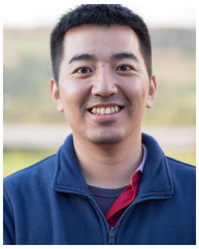
[55] B. I. Erkmen and J. H. Shapiro, "Ghost imaging: From quantum to classical to computational," *Adv. Opt. Photon.*, vol. 2, no. 4, pp. 405–450, 2010.

[56] M. J. Padgett and R. W. Boyd, "An introduction to ghost imaging: Quantum and classical," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 375, no. 2099, Aug. 2017, Art. no. 20160233.



STIRLING SCHOLES received the B.Sc. degree (Hons.) in physics and the M.Sc. degree (Hons.) degree in optics and photonics from the University of the Witwatersrand, Johannesburg, South Africa, in 2018 and 2020, respectively. He joined the Heriot-Watt University Quantum Optics and Computational Imaging (QOCI) group in 2020 to pursue a Ph.D. degree in applied imaging systems with a focus on data fusion approaches for high-speed tracking and identification of objects using 3D time-of-flight technology.

ALICE RUGET received the M.Sc. degree in electrical engineering from CentraleSupélec, Gif-sur-Yvette, France, in 2017, and the M.Sc. degree in biomedical engineering from ETH Zürich, Switzerland, in 2019. She is currently pursuing the Ph.D. degree with the Group HW Quantum, Heriot-Watt University, Edinburgh, U.K., with a focus on computational imaging for ultra-fast imaging in three dimensions. Her research interests include the development of machine learning algorithms to enhance the image quality for different imaging systems, such as single-photon sensitive detectors.



FENG ZHU received the B.Sc. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 2011 and 2017, respectively. Since 2017, he has been a Research Associate with the Department of Physics, Heriot-Watt University. His research interests include quantum and classical optics and imaging.

JONATHAN LEACH received the M.Sc. degree in physics and the Ph.D. degree from The University of Glasgow, in 2002 and 2004, respectively. He was a Senior Research Associate with the Quantum Photonics Group, University of Ottawa, Ottawa, ON, Canada. He joined Heriot-Watt University, Edinburgh, U.K., in 2012, to establish a research program in experimental quantum optics, where he is currently an Associate Professor. He has written more than 100 peer-reviewed articles in scientific journals. His research interests include applying classical and quantum optics techniques to solve problems in information and imaging science.