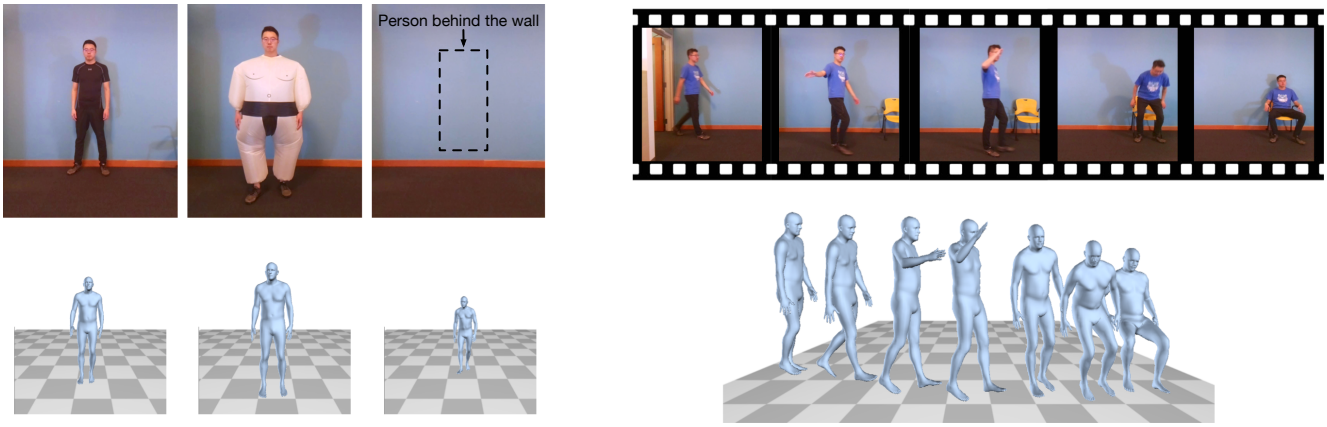
Through-Wall Human Mesh Recovery Using Radio Signals

Mingmin Zhao Yingcheng Liu Aniruddh Raghu Tianhong Li
Hang Zhao Antonio Torralba Dina Katabi
MIT CSAIL



(a) Human mesh recovery through occlusions

(b) Dynamic human mesh captures the body motion

Figure 1: **Dynamic human meshes estimated using radio signals.** Images captured by a camera co-located with the radio sensor are presented here for visual reference. (a) shows the estimated human meshes of the same person in sportswear, a baggy costume and when he is behind the wall. (b) shows the dynamic meshes that capture the motion when the person walks, waves his hand, and sits.

Abstract – *This paper presents RF-Avatar, a neural network model that can estimate 3D meshes of the human body in the presence of occlusions, baggy clothes, and bad lighting conditions. We leverage that radio frequency (RF) signals in the WiFi range traverse clothes and occlusions and bounce off the human body. Our model parses such radio signals and recovers 3D body meshes. Our meshes are dynamic and smoothly track the movements of the corresponding people. Further, our model works both in single and multi-person scenarios. Inferring body meshes from radio signals is a highly under-constrained problem. Our model deals with this challenge using: 1) a combination of strong and weak supervision, 2) a multi-headed self-attention mechanism that attends differently to temporal information in the radio signal, and 3) an adversarially trained temporal discriminator that imposes a prior on the dynamics of human motion. Our results show that RF-Avatar accurately recovers dynamic 3D meshes in the presence of occlusions, baggy clothes, bad lighting conditions, and even through walls.*

1. Introduction

Estimating a full 3D mesh of the human body, capturing both human pose and body shape, is a challenging task in computer vision. The community has achieved major advances in estimating 2D/3D human pose [15, 44], and more recent work has succeeded in recovering a full 3D mesh of the human body characterizing both pose and shape [9, 23]. However, as in any camera-based recognition task, human mesh recovery is still prone to errors when people wear baggy clothes, and in the presence of occlusions or under bad lighting conditions.

Recent research has proposed to use different sensing modalities that could augment vision systems and allow them to expand beyond the capabilities of cameras [46, 45, 12, 47, 50]. In particular, radio frequency (RF) based sensing systems have demonstrated through-wall human detection and pose estimation [48, 49]. These methods leverage the fact that RF signals in the WiFi range can traverse occlusions and reflect off the human body. The resulting systems are privacy-preserving as they do not record visual data, and can cover a large space with a single device, despite occlusions. However, RF signals have much lower spatial resolution than visual camera images, and therefore it remains an open question as to whether it is possible at all to capture dynamic 3D body meshes characterizing the human body and its motion with RF sensing.

In this paper, we demonstrate how to use RF sensing

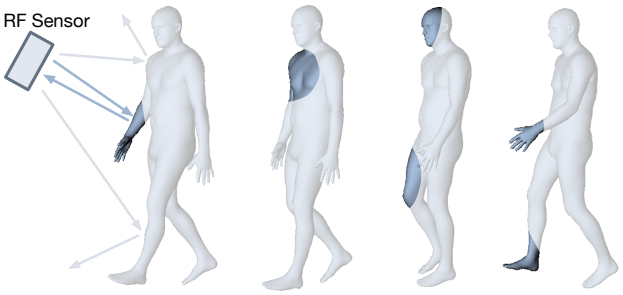


Figure 2: **Specularity of the human body with respect to RF.** The human body reflects RF signals as opposed to scattering them. A single RF snapshot can only capture a subset of limbs depending on the orientation of the surfaces.

to estimate dynamic 3D meshes for human bodies through walls and occlusions. We introduce RF-Avatar, a neural network framework that parses RF signals to infer dynamic 3D meshes. Our model can capture body meshes in the presence of significant, and even total, occlusion. It stays accurate in bad lighting conditions, and when people wear costumes or baggy clothes. Figure 1 shows RF-Avatar's performance on a few test examples. The left panel demonstrates that RF-Avatar can capture the 3D body mesh accurately even when the human body is obscured by a voluminous costume, or completely hidden behind a wall. Further, as shown in the right panel, RF-Avatar generates dynamic meshes that track the body movement. In Section 5.2, we show that RF-Avatar also works in dark settings and in scenarios with multiple individuals.

Inferring 3D body meshes solely from radio signals is a difficult task. The human body is specular with respect to RF signals in the WiFi range –i.e., the human body reflects RF signals, as opposed to scattering them. As illustrated in Figure 2, depending on the orientation of the surface of each limb, the RF signal may be reflected towards our radio or away from it. Thus, in contrast to camera systems where any snapshot shows all unoccluded body parts, in radio systems, a single snapshot has information only about a subset of the limbs. This problem is further complicated by the fact that there is no direct relationship between the reflected RF signals from a person and their underlying 3D body mesh. We do not know which part of the body actually reflected the signal back. This is different from camera images, which capture a 2D projection of the 3D body meshes (modulo clothing). The fact that the reflected RF signal at a point in time has information only about a unknown subset of the body parts means that using RF sensing to capture 3D meshes is a highly unconstrained problem – at a point in time, the reflected RF signal could be explained by many different 3D meshes, most of which are incorrect.

RF-Avatar tackles the above challenge as follows. We first develop a module that uses the RF signal to detect and track multiple people over time in 3D space, and create trajectories for each unique individual. Our detection pipeline extends the Mask-RCNN framework [21] to handle RF signals. RF-Avatar then uses each person's detected trajectory, which incorporates multiple RF snapshots over time, to estimate their body mesh. This strategy of combining information across successive snapshots of RF signals allows RF-Avatar to deal with the fact that different RF snapshots contain information about different body parts due to the specularity of the human body. We incorporate a multi-headed attention module that lets the neural network selectively focus on different RF snapshots at different times, depending on what body parts reflected RF signals back to the radio. RF-Avatar also learns a prior on human motion dynamics to help resolve ambiguity about human motion over time. We introduce a temporal adversarial training method to encode human pose and motion dynamics.

To train our RF-based model, we use vision to provide cross-modality supervision. We use various types of supervision, ranging from off-the-shelf 2D pose estimators (for pose supervision) to vision-based 3D body scanning (for shape supervision). We design a data collection protocol that scales to multiple environments, while also minimizing overhead and inconvenience to subjects.

We train and test RF-Avatar using data collected in public environments around our campus. Our experimental results show that in visible scenes, RF-Avatar has mean joint position error of 5.84 cm and mean vertex-to-vertex distance of 1.89 cm. For through-wall scenes and subjects wearing loose costumes, RF-Avatar has mean joint position error of 6.26 cm and mean vertex-to-vertex distance of 1.97 cm whereas the vision-based system fails completely. We conduct ablation studies to show the importance of our self-attention mechanism and the adversarially learned prior for human pose and motion dynamics.

2. Related Work

Shape representation. Compact and accurate representations for human body meshes have been studied in computer graphics, with many models proposed in prior work such as linear blend skinning (LBS), the pose space deformation model (PSD) [28], SCAPE [10], and others [8]. More recently, the Skinned Multi-Person Linear (SMPL) model was proposed by [33]. SMPL is a generative model that decomposes the 3D mesh into a shape vector (characterizing variation in height, body proportions, and weight) and a pose vector (modeling the deformation of the 3D mesh under motion). This model is highly realistic and can represent a wide variety of body shapes and poses; we therefore adopt the SMPL model as our shape representation.

Capturing human shapes. There are broadly two methods used to capture body shape in prior work. In scanning-based methods, several images of a subject are obtained,

typically in a canonical pose, and then optimization-based methods are used to recover the SCAPE or SMPL parameters representing the subject's shape. The authors of [14, 19, 20, 41, 6] used scanning approaches, incorporating silhouette information and correspondence cues to fit a SCAPE or SMPL model. However, scanning-based methods have the inherent limitation that they can be easily affected by clothing, so they only work well when subjects are in form-fitting clothes. They are also limited to indoor settings and do not properly capture motion dynamics. Thus, many recent works, including ours, use scanning methods only to provide supervision to learning-based methods.

In learning-based methods, models are trained to predict parameters of a shape model (e.g., SMPL). Such methods are challenging due to the lack of 3D human mesh dataset. Despite this, there has been significant success in this area. Bogo *et al*. [13] proposed a two-stage process to firstly predict joint locations and then fit SMPL parameters from a 2D image. Lassner *et al*. [27] developed on this approach, incorporating a semi-automatic annotation scheme to improve scalability. More recent work [23, 36] captured 3D meshes from 2D images using adversarial loss, and Kanazawa *et al*. [24] learned dynamic 3D meshes using videos as an additional data source. In this work, we adopt a learning-based approach, building on the above literature, and expanding it to deal with scenarios with occlusions and bad lighting.

Priors on human shape and motion. Capturing the prior of human shape and human motion dynamics is essential in order to generate accurate and realistic dynamic meshes. Supervision for training such systems is typically in the form of 2D/3D keypoints; often, there is no supervision for full 3D joint angles, so priors must be used for regularization. Bogo *et al*. [13] and Lassner *et al*. [27] used optimization methods to fit SMPL parameters and thus encode human shape; however, priors on human motion were not encoded when training their systems. Kanazawa *et al*. [23, 24] used an adversarial loss to provide a prior when considering shape estimation from 2D images and video but this method did not capture a prior on motion dynamics, as the discriminator operated on a per timestep basis. In this work, we introduce a new prior to capture motion dynamics. We also incorporate an attention module to selectively attend to different keypoints when producing shape estimates.

Wireless sensing to capture shape. Radar systems can use RF reflections to detect and track humans [5, 37, 29]. However, they typically only track location and movements and cannot generate accurate or dynamic body meshes. Radar systems that generate body meshes (e.g., airport security scanners) operate at very high frequencies [42, 7, 11]; such systems work only at short distances, cannot deal with occlusions such as furniture and walls, and do not generate dynamic meshes. In contrast, our system operates through walls and occlusions and generates dynamic meshes. There is also prior work utilizing RF signals to capture elements of human shape. RF-Capture [4] presented a system that can detect human body parts when a person is walking towards a radio transceiver. RF-Pose [48] presented a system to perform 2D pose estimation for multiple people, and RF-Pose3D [49] extended this result to enable multi-person 3D keypoint detection. Our work develops on these ideas by providing the ability to reconstruct a full 3D mesh capturing shape and motion, as opposed to only recovering limb and joint positions.

3. RF Signals and Convolutions

Much of the work on sensing people using radio signals uses a technology called FMCW (Frequency Modulated Continuous Wave) [40, 35]. An FMCW radio works by transmitting a low power radio signal and receiving its reflections from the environment. Different FMCW radios are available [2, 3] and RF-Avatar uses one similar to that used in [4] and can be ordered from [1]. Our model is not specific to a particular radio, and applies generally to such radar-based radios. In RF-Avatar , the reflected RF signal is transformed into a function of the 3D spatial location and time [49]. This results in a 4D tensor that forms the input to our neural network. It can be viewed as a sequence of 3D tensors at different points of time. Each 3D tensor is henceforth referred to as the *RF frame* at a specific time.

It is important to note that RF signals have intrinsically different properties from visual data, i.e., camera pixels: first, the human body is specular in the frequency range that traverse walls (see Figure 2). Each RF frame therefore only captures a subset of the human body parts. Also, in the frequency range of interest (in which RF can pass through walls), RF signals have low spatial resolution – our radio has a depth resolution about 10 cm, and angular resolution of 15 degrees. This is a much lower resolution than what is obtained with a camera. The above properties have implications for human mesh recovery, and need to be taken into account in designing our model.

CNN with RF Signals: Processing the 4D RF tensor with 4D convolutions has prohibitive computational and space complexity. We use a decomposition technique [49] to decompose both the RF tensor and the 4D convolution into 3D ones. The main idea is to represent each 3D RF frame as a summation of multiple 2D projections. As a result, the operation in the original dimension is equivalent to a combination of operations in lower-dimensions.

4. Method

We propose a neural network framework that parses RF signals and produces dynamic body meshes for multiple people. The design of our model is inspired by the Mask-RCNN framework [21]. Mask-RCNN is designed for

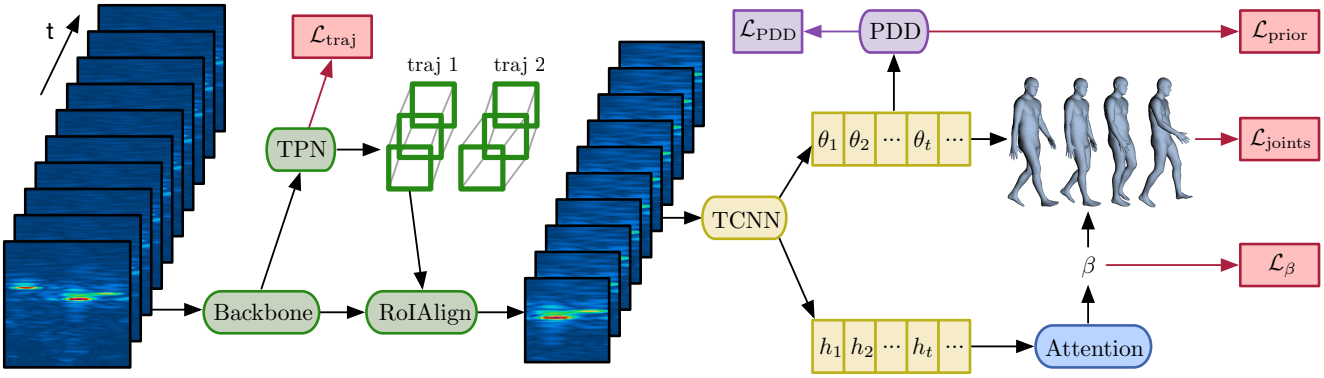


Figure 3: **Overview of the network model used in RF-Avatar.**

instance-level recognition tasks in 2D images; we extend it to handle 4D RF inputs and generate 3D body meshes over time. Figure 3 illustrates the 2-stage network architecture used in RF-Avatar. In the first stage of the model, we use a Trajectory Proposal Network (TPN) to detect and track each person in 3D space (Sec. 4.2). TPN outputs a trajectory (a sequence of bounding boxes over time) for each person, and we use this trajectory to crop the spatial regions in the RF tensor that contain this particular person.

The second stage of the model takes the cropped features as input and uses a Trajectory-CNN (TCNN) to estimate the sequence of body meshes of this person (Sec. 4.3). TCNN introduces an attention module to adaptively combine features from different RF frames when predicting the body shape (Sec. 4.3). TCNN also outputs a sequence of joint angles capturing the body motion. It uses a Pose and Dynamics Discriminator (PDD) to help resolve the ambiguities about human motion (Sec. 4.4). We describe how we use various forms of supervision to train RF-Avatar in Sec. 4.5.

4.1. Human Mesh Representation

We use the Skinned Multi-Person Linear (SMPL) model [33] to encode the 3D mesh of a human body. SMPL factors the human mesh into a person-dependent shape vector and pose-dependent 3D joint angles. The shape vector $\boldsymbol{\beta} \in \mathbb{R}^{10}$ corresponds to the first 10 coefficients of a PCA shape model. The joint angles $\boldsymbol{\theta} \in \mathbb{R}^{72}$ define the global rotation of the body and the 3D relative rotations of 23 joints. SMPL provides a differentiable function $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta})$ that outputs $N = 6890$ vertices of a triangular mesh given $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. A 3D mesh of a human body in the world coordinates is represented by 85 parameters including $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ (describing shape and pose via SMPL) and a global translation vector $\boldsymbol{\delta}$. Note that the 3D location of body joints, $\boldsymbol{J}$, can be computed via a linear combination of mesh vertices.

RF-Avatar recovers dynamic body meshes, i.e., a sequence of SMPL parameters including a time-invariant $\boldsymbol{\beta}$ characterizing the body, and a time-variant $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T)$ describing the joint angles, and a time-variant global translation vector $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_T)$ capturing the location.

4.2. Trajectory Proposal Network

The first stage in our 3D mesh estimation pipeline is to detect regions containing individuals and then track them over time to form trajectories. Our Trajectory Proposal Network (TPN) takes as input the 4D RF tensor. It first extracts features using a backbone with spatial-temporal convolutions, and then uses a recurrent region proposal network to propose candidate regions for each RF frame. After a further candidate selection stage with a box head, we perform a lightweight optimization to link the detections over time. We describe each TPN component in detail:

Backbone: This takes the raw sequence of RF frames as input and uses a set of decomposed 4D convolutional layers (see Sec. 3) with residual connections to produce features.

Recurrent Region Proposal Network (Recurrent-RPN): In contrast to prior work using RPN in detection and tracking [38, 21, 16], our recurrent-RPN has two major differences. First, we wish to detect individuals in the 3D world space instead of the 2D image space. Thus, our model uses 3D bounding boxes as anchors and learns to propose 3D regions by transforming these anchors. Proposing regions in 3D space removes scale-variation of regions due to perspective projection to image space [30]. For tractability, we choose 3D anchors to be those close to the ground plane. Second, our RPN works in a recurrent manner to propose regions for each RF frame sequentially. It uses recurrent layers on top of convolutional layers to predict object scores and regression outputs for all anchor regions. Non-maximal suppression (NMS) is used to remove duplicated proposals.

Box Head: To improve detection precision, we use a box head to further classify proposals into correct/incorrect detections. We use standard box head with RoIAlign [21].

Tracker: The tracker module receives proposals from the Box Head output at each timestep. It then associates together proposals that belong to the person, and stitches them

over time to form trajectory tubes. We use a lightweight optimization tracker based on bipartite matching [16].

4.3. Trajectory-CNN with Attention

Trajectory-CNN (TCNN) uses the cropped features from the TPN as input and estimates the body mesh parameters for each individual. To deal with the fact that different RF frames contain information about different body parts, we introduce a self-attention module to predict a temporally consistent shape $\boldsymbol{\beta}$. TCNN first extracts shape features at different timesteps as $\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_T)$. Our self-attention module uses a function f to attend to different frames and combine all the shape features into a fixed-length feature vector: $\widetilde{\boldsymbol{h}} = \frac{1}{\mathcal{C}(\boldsymbol{H})} \sum_t (f(\boldsymbol{h}_t) \cdot \boldsymbol{h}_t)$, where $\mathcal{C}(\boldsymbol{H}) = \sum_t (f(\boldsymbol{h}_t)$ is a normalization factor. We utilize multi-headed self-attention [31], allowing the neural network to attend to different aspects of the shape features differently. Feature vectors from different heads are concatenated together to produce the $\boldsymbol{\beta}$ prediction.

Empirical results show that this temporal self-attention leads to improved shape estimation and model interpretability. We further believe that the benefits of temporal attention extend to video-based 3D mesh models, since it allows the model to recognize that different frames may have different importance for estimating a particular mesh parameter. For example, height is better estimated from frames where the subject is standing as opposed to sitting.

4.4. Learning Pose and Dynamics Priors

We would like to learn a prior that encodes feasible human pose and motion dynamics in order to ensure that the 3D meshes it produces over time are realistic. Without such a prior, and especially given the weak supervision for the 3D joint angles (see Sec. 4.5), our model could produce arbitrary rotations of joints and/or temporally inconsistent meshes. This issue is exacerbated in the case of pose estimation from RF signals, as we only get sparse observations at each timestep, due to human body specularity.

We introduce an adversarial prior that regularizes both human body pose and motion dynamics and ensures realistic predictions; we call this the Pose and Dynamics Discriminator (PDD). PDD is a data-driven discriminator that takes our predicted sequence of 3D joint angles, and aims to distinguish it from real human poses and dynamics data. We use MoSh-ed data from the CMU MoCap dataset [26] as real dynamics data. It covers a diverse set of human subjects performing different poses and actions. In contrast to previous work, which uses a separate discriminator for each joint at a single time instance [23, 24], PDD considers all keypoints over a temporal window, which improves the estimated pose results.

The PDD is trained using a binary cross entropy loss and a gradient penalty term on the real data. Its objective function takes the following form:

$$\mathcal{L}_{\text{PDD}} = -\Big(\mathbb{E}_{\boldsymbol{\Phi}\sim p_{\text{data}}}[\log D(\boldsymbol{\Phi})] + \mathbb{E}_{\boldsymbol{\Theta}\sim p_E}\big[\log(1 - D(\boldsymbol{\Theta})\big]\Big) + \gamma \cdot \mathbb{E}_{\boldsymbol{\Phi}\sim p_{\text{data}}}[\|\nabla D(\boldsymbol{\Phi})\|^2], \tag{1}$$

where $\boldsymbol{\Theta}$ is the estimated joint angles from TCNN, and $D(\cdot)$ is our pose and dynamics discriminator.

Finally, we convert them to rotation matrices and feed to the discriminator. This technique allows for more stable training by bypassing the 2π wrapping nature of angle representations.

4.5. Training the Model

Past image-based solutions that recover 3D meshes use mostly weak supervision during training, in the form of the location of body joints. However, our empirical results (Sec. 5.3) show that weak supervision is insufficient for RF-based systems. Unfortunately, strong supervision that captures full information about 3D meshes is difficult to obtain, as it requires highly constrained setups involving a sophisticated multi-view camera setup, and minimally clothed subjects [32, 22]; such setups are not scalable.

To deal with this issue, we train our model using a combination of strong and weak supervision. The SMPL shape representation decomposes into a time-independent shape vector, $\boldsymbol{\beta}$, and time-dependent joint angles, $\boldsymbol{\theta}$. We obtain strong supervision for the time-independent shape vector by using an adapted version of the scanning/silhouette method from [6] once for each subject in our dataset, with each subject in a standard canonical pose. We need only perform this procedure once for each person, as the shape vector, $\boldsymbol{\beta}$, is constant for a given person. We adapt the procedure in [6] as follows. The original method solves an optimization problem to obtain both $\boldsymbol{\beta}$ and offsets for the N mesh vertices (to capture clothing and other small perturbations). We remove the optimization over the mesh vertices (as we wish to capture pure body shape, and do not wish to include clothing information) to obtain only $\boldsymbol{\beta}$. We henceforth refer to the mesh obtained from this method as a *VideoAvatar*.

Additionally, we use a system of 12 calibrated cameras and the AlphaPose algorithm [15, 44] to obtain ground truth information for 3D joint locations, obtained as subjects engage in activities (walking, standing up/sitting down, interacting with objects, etc). This serves as weak supervision for our system's joint angle predictions, $\boldsymbol{\theta}$.

Training TPN: We use standard anchor classification and regression losses [38, 21]. We compute ground truth 3D bounding boxes from the 3D poses reconstructed by 3D-AlphaPose. The total loss $\mathcal{L}_{\text{traj}}$ is the sum of losses from the RPN and the Box Head.

Training TCNN: As illustrated in Figure 3, TCNN has three different loss terms. We compute shape loss $\mathcal{L}_{\beta}$ and

3D joint loss $\mathcal{L}_{\text{joints}}$ by comparing our predictions with the ground truth provided by corresponding vision algorithms. We use the smooth L1 loss [17] for both of them. We note that in order to compute the joint locations in 3D world space, our model needs to predict the global translations $\mathbf{\Delta}$ as well. We use the bounding box centers and predicted local translations with respect to the box centers to obtain the global translations. Our TCNN also performs a gender classification and uses the SMPL model of the predicted gender to compute the vertex and the joint locations.

When training TCNN together with the PDD, we follow standard adversarial training schemes [18, 34] and use the following loss term for TCNN:

$$\mathcal{L}_{\text{prior}} = -\mathbb{E}_{\mathbf{\Theta}\sim p_E}\log(D(\mathbf{\Theta})), \tag{2}$$

where $D(\cdot)$ is our pose and dynamics discriminator.

The total loss for the TCNN is a sum of the terms:

$$\mathcal{L}_{\text{TCNN}} = \mathcal{L}_{\boldsymbol{\beta}} + \mathcal{L}_{\text{joints}} + \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{gender}}. \tag{3}$$

5. Experiments

We describe our dataset, implementation details, quantitative and qualitative results on shape and pose estimation, and analyze what is learned by the attention module.

5.1. Dataset and Implementation

Dataset: To train and test our model, we build a dataset containing 84 subjects (male and female). For each subject, we use an adapted version of the approach in [6] to obtain ground truth $\boldsymbol{\beta}$ vectors with the subjects in a canonical pose (Sec. 4.5) – we refer to this method as VideoAvatar. We obtain data for the subjects walking around and engaging in activities in 16 different environments around our campus, and use a co-located calibrated camera system to obtain ground truth keypoint locations for the subjects. Our camera system is mobile, allowing us to collect data in varied environments and build a representative dataset.

Implementation details: We use decomposed 4D convolutions (Sec. 3) with residual blocks. Each uses ReLu activation and Group Normalization [43]. We use 12, 3, 12 and 12 layers of convolution in our backbone, RPN, box head and TCNN, respectively. We also use 1 and 2 layers of spatially-distributed GRU for TPN and RCNN. Our self-attention module uses two fully connected layers with $tanh(\cdot)$ activation in the middle. Our PDD model uses 12 layers of 1D temporal convolution, followed by a fully connected layer. We implement our model in PyTorch. Our model is trained with the Adam [25] optimizer for 40000 iterations.

5.2. Qualitative Evaluation for Shape and Pose

RF-Avatar produces realistic meshes: Figure 4 shows the 3D meshes produced by our model for different poses and subjects, as compared to the RGB images captured by a co-located camera. As can be seen, qualitatively, the estimated meshes are realistic, and agree well with the body shapes of different subjects. Our model also handles different body shapes (for male and female subjects), poses, and multi-person scenarios effectively. In addition, considering the bottom row of images in Figure 4, our model can produce accurate meshes for partially occluded subjects, subjects behind a wall, and subjects in poor lighting conditions; a vision-based system cannot produce full meshes in these situations.

RF-Avatar effectively captures variation in body shape: To evaluate the quality of body shape predicted by RF-Avatar, we compare our prediction with the body shape captured by VideoAvatar [6], shown in Figure 6. VideoAvatar leverages a sequence of images to estimate a body mesh. The recovered mesh is overlaid on the RGB image of each person and is shown on the right side of each pair. To better compare the difference in body shape, we take the predicted shape of a subject (obtained by averaging predictions over a window of 10 seconds) from RF-Avatar and render the resulting mesh (in the same pose as VideoAvatar) and overlay it on the same background. This is shown on the left side of each pair. We see a close qualitative agreement between the ground truth and the output from RF-Avatar for male and female subjects with different body shapes.

RF-Avatar encodes human motion dynamics: Figure 5 demonstrates how our model can produce dynamic 3D meshes for different people over time, and how these meshes look realistic. We can see how the two subjects perform walking and lifting actions, and the produced meshes over time closely map to the performed actions.

5.3. Quantitative Evaluation for Shape and Pose

We now present quantitative results for our method, evaluating its performs on standard pose and body shape metrics. We also conduct ablation studies comparing with variants of our model that lack a particular component, namely variants that do not have supervision on the β parameters, do not use an attention mechanism, and use a frame-based discriminator (as in [23, 24]).

Metrics: We report the commonly used 3D joint metric Mean Per Joint Position Error (MPJPE). We also compute the per-vertex error as the average vertex to surface distance between the predicted mesh and the ground truth.

Table 1 shows the results for MPJPE and Per-vertex error respectively. As can be seen, for both MPJPE and per-vertex error, assessing recovered pose and shape quality respectively, the model that incorporates supervision for $\boldsymbol{\beta}$, self-attention, and the temporal discriminator, performs the best across all metrics. Of particular note is how the MPJPE drops from 6.05 cm to 6.88 cm when we do not use the temporal discriminator, demonstrating the value of the PDD in

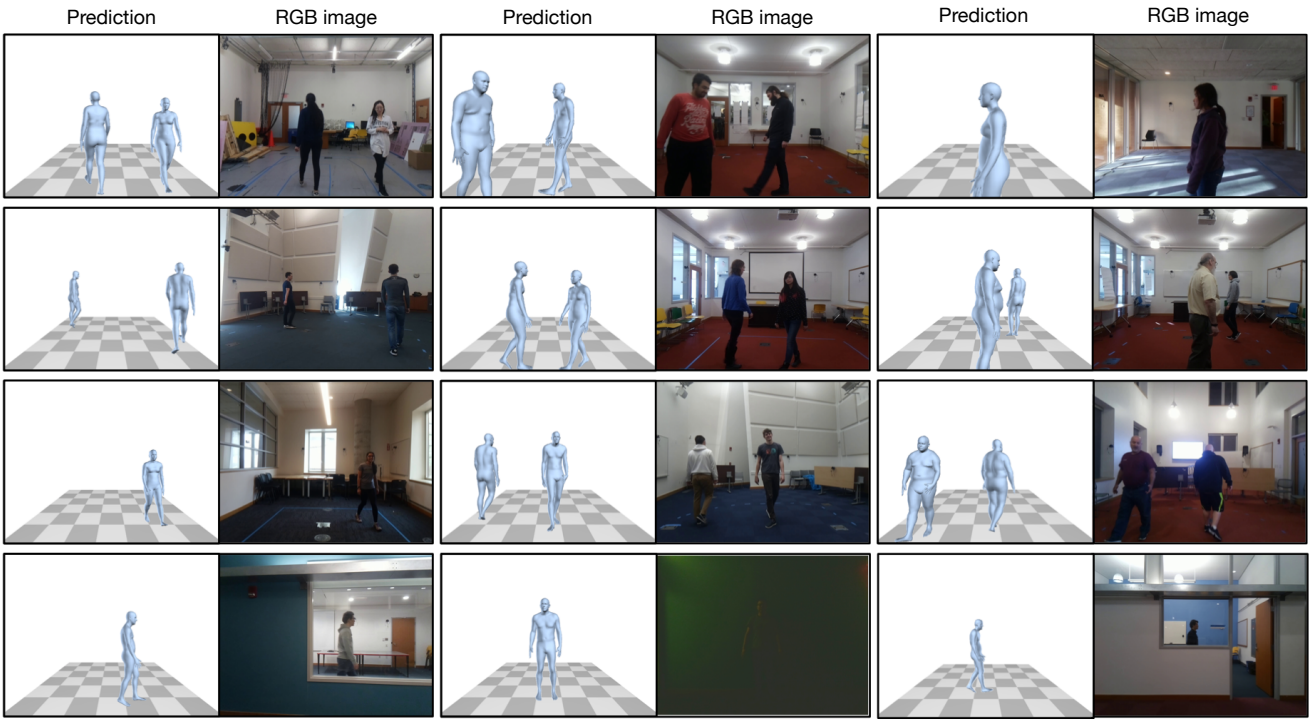


Figure 4: **Human mesh prediction from RF-Avatar.** We show images for visual reference. Our model captures different body shapes, poses, and multi-person scenarios effectively. The bottom row shows that RF-Avatar works despite occlusion and bad lighting conditions.

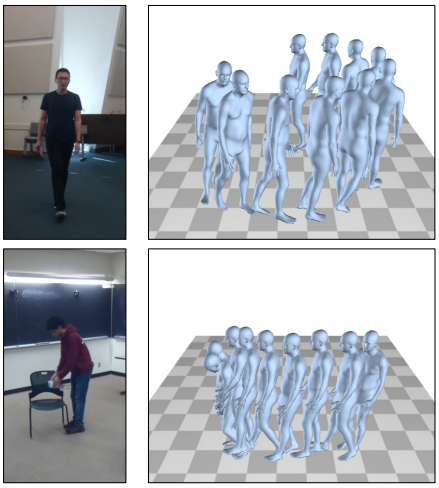


Figure 5: **Dynamic human meshes predicted from RF-Avatar.** RF-Avatar can capture dynamic mashes for different actions, including walking (top image) and lifting an object (bottom image).

learning motion dynamics to help resolve ambiguities. We also see the importance of adding strong supervision for β: the per-vertex error increases from 1.88 cm to 4.70 cm when it is removed. We also note here that the previous image-based mesh recovery methods have an MPJPE error around 8.8 cm [23] and a Per-vertex error around 11.8 cm [36]. Aside from the difference in datasets, we believe this difference in performance can be attributed to the fact that RF signals capture information about 3D space and our RF-based model is trained with stronger supervision than image-based methods.

We further see that the results using the TPN output (top row) are similar to the results using the ground truth bounding boxes (bottom row), illustrating the effectiveness of our entire detection, tracking, and shape estimation pipeline. This applies for both pose and shape metrics.

	MPJPE (cm)	Per-vertex error (cm)
RF-Avatar	**6.05**	**1.88**
No β loss	6.72	4.70
No attention	6.43	2.55
Frame-based disc.	6.88	2.24
With g.t. boxes	**5.75**	**1.65**

Table 1: Joint and vertex errors, assessing pose and body shape quality respectively.

Table 2 compares the results of our model for the shape and pose metrics for the total occlusion (through-wall) and line-of-sight scenarios. We see that our model performs well in the through-wall setting, even though it was never trained directly on through-wall data.

5.4. Analysis of Self-Attention

Table 1 shows that adding the self-attention module helps our quantitative results on shape and pose metrics.

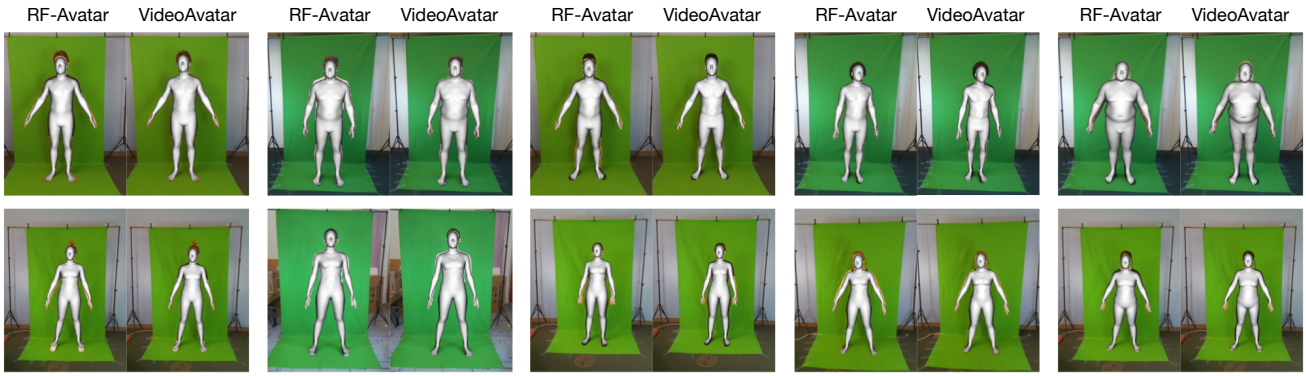


Figure 6: **Comparison of body shape recovered from RF-Avatar and VideoAvatar.** We render the mesh with the predicted shape estimated by RF-Avatar and the ground truth shape estimated by VideoAvatar and overlaid both on top of the corresponding RGB image.

	3D MPJPE (cm)	Per-vertex errors (cm)
Line-of-sight	5.84	1.79
Through-wall	6.26	1.97

Table 2: Results in the line-of-sight and through-wall settings.

Self-attention helps our model better combine information over time when estimating the shape vector. We visualize the learned multi-headed attention maps in Figure 7. Focusing on the second attention component first, we see that it has high activation for timesteps 11 and 12. The high activation at these times indicates that they may contain important shape information. When comparing with the RGB images around timesteps 11 and 12, we see that the subject is facing the radio and waving at these times, so these timesteps likely contain reflections from his arm and provide important information about his upper limbs.

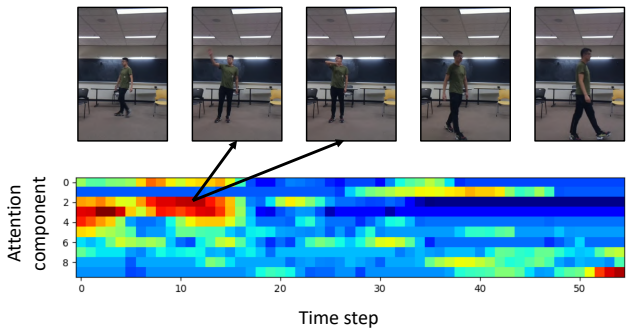
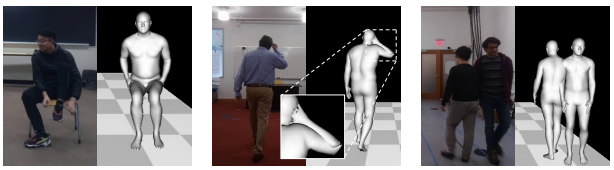


Figure 7: Learned attention maps over time for the different attention heads. We see that different attention components activate differently when the person is turning, waving hands and showing his side to the sensor.

5.5. Failure Modes

We analyze the failure cases of RF-Avatar. Typical failure examples are caused by (a) unusual body poses, (b) interpenetration of body meshes [13, 39], and (c) highly crowded scenes where people are very close to each other.



(a) Unusual body pose (b) Interpenetration (c) Crowded scene

Figure 8: **Typical failure cases of RF-Avatar.**

In Figure 8, we present examples of the typical failure cases. Figure 8(b) shows that RF-Avatar fails to handle unusual body poses (*e.g.* tying shoes). In Figure 8(b), interpenetration of estimated body meshes happens when the person raise his hand to hold glasses. In crowded scenes (*e.g.* Figure 8(c)) where people are very close to each other, RF-Avatar produces overlapped body meshes. Failure modes (a) and (b) are related to our choice of body mesh model, while failure mode (c) is due to the relatively low spatial resolution of RF signals in comparison to visible light.

6. Conclusion

This paper presented RF-Avatar a system that recovers dynamic 3D mesh models of the human body using RF signals. RF-Avatar is trained using cross-modality supervision from state-of-the-art vision algorithms, yet remains effective in situations that challenge vision systems, such as in poor lighting, and when subjects are occluded. We believe this work paves the way for many new applications in health monitoring, gaming, smart homes, etc. RF-Avatar significantly extends the capabilities of existing RF-based sensing systems, and the principles involved in its design could be utilized to improve the performance of existing computer vision methodologies.

Acknowledgments: We are grateful to all the human subjects for their contribution to our dataset. We thank the CSAIL members for their insightful comments.

References

[1] Emerald. https://www.emeraldinno.com/clinical. 3

[2] Texas instruments. http://www.ti.com/. 3

[3] Walabot. https://walabot.com/. 3

[4] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. Capturing the human figure through a wall. *ACM Transactions on Graphics*, 34(6):219, November 2015. 3

[5] Fadel Adib, Zachary Kabelac, Dina Katabi, and Robert C. Miller. 3D tracking via body radio reflections. In *Proceedings of the USENIX Conference on Networked Systems Design and Implementation*, NSDI, 2014. 3

[6] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3, 5, 6

[7] Brett Allen, Brian Curless, and Zoran Popović. The space of human body shapes: reconstruction and parameterization from range scans. In *ACM transactions on graphics (TOG)*, volume 22, pages 587–594. ACM, 2003. 3

[8] Brett Allen, Brian Curless, Zoran Popović, and Aaron Hertzmann. Learning a correlated model of identity and pose-dependent body shape variation for real-time synthesis. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 147–156. Eurographics Association, 2006. 2

[9] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 1

[10] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM transactions on graphics (TOG)*, volume 24, pages 408–416. ACM, 2005. 2

[11] Roger Appleby and Rupert N Anderton. Millimeter-wave and submillimeter-wave imaging for security and surveillance. *Proceedings of the IEEE*, 95:1683–1690, 2007. 3

[12] Amanda Berg, Jorgen Ahlberg, and Michael Felsberg. Generating visible spectrum images from thermal infrared. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1143–1152, 2018. 1

[13] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 3, 8

[14] Yu Chen, Tae-Kyun Kim, and Roberto Cipolla. Inferring 3d shapes and deformations from single views. In *European Conference on Computer Vision*, pages 300–313. Springer, 2010. 3

[15] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 1, 5

[16] Rohit Girdhar, Georgia Gkioxari, Lorenzo Torresani, Manohar Paluri, and Du Tran. Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 350–359, 2018. 4, 5

[17] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 6

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 6

[19] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1381–1388. IEEE, 2009. 3

[20] Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1823–1830. IEEE, 2010. 3

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, ICCV, 2017. 2, 3, 4, 5

[22] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate marker-less human shape and pose estimation over time. In *International Conference on 3D Vision (3DV)*, 2017. 5

[23] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. 1, 3, 5, 6, 7

[24] Angjoo Kanazawa, Jason Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. 2019. 3, 5, 6

[25] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR, 2014. 6

[26] Carnegie Mellon Graphics Lab. CMU Graphics Lab Motion Capture Database. 5

[27] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 3, 2017. 3

[28] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172. ACM Press/Addison-Wesley Publishing Co., 2000. 2

[29] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi. Making the invisible visible: Action recognition through walls and occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 4

[31] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *ICLR*, 2017. 5

[32] Matthew Loper, Naureen Mahmood, and Michael J Black. Mosh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics (TOG)*, 33(6):220, 2014. 5

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2, 4

[34] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018. 6

[35] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Kari Pulli. Short-range fmcw monopulse radar for hand-gesture sensing. In *2015 IEEE Radar Conference (RadarCon)*, pages 1491–1496. IEEE, 2015. 3

[36] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–468, 2018. 3, 7

[37] John Peabody Jr, Gregory L Charvat, Justin Goodwin, and Martin Tobias. Through-wall imaging radar. Technical report, Massachusetts Institute of Technology-Lincoln Laboratory Lexington United States, 2012. 3

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 4, 5

[39] Cristian Sminchisescu and Bill Triggs. Covariance scaled sampling for monocular 3d body tracking. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. 8

[40] Andrew G Stove. Linear fmcw radar techniques. In *IEE Proceedings F (Radar and Signal Processing)*, volume 139, pages 343–350. IET, 1992. 3

[41] Jun Kai Vince Tan, Ignas Budvytis, and Roberto Cipolla. Indirect deep structured learning for 3d human body shape and pose prediction. In *BMVC*, volume 3, page 6, 2017. 3

[42] Ruth M Woodward, Bryan E Cole, Vincent P Wallace, Richard J Pye, Donald D Arnone, Edmund H Linfield, and Michael Pepper. Terahertz pulse imaging in reflection geometry of human skin cancer and skin tissue. *Physics in Medicine & Biology*, 47:3853, 2002. 3

[43] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 6

[44] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. Pose Flow: Efficient online pose tracking. In *BMVC*, 2018. 1, 5

[45] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5580–5588, 2017. 1

[46] Zhoutong Zhang, Qiujia Li, Zhengjia Huang, Jiajun Wu, Josh Tenenbaum, and Bill Freeman. Shape and material from sound. In *Advances in Neural Information Processing Systems*, pages 1278–1288, 2017. 1

[47] Mingmin Zhao, Fadel Adib, and Dina Katabi. Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, MobiCom, 2016. 1

[48] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Dina Katabi, and Antonio Torralba. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 3

[49] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. Rf-based 3d skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 267–281. ACM, 2018. 1, 3

[50] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In *International Conference on Machine Learning*, ICML, 2017. 1