3D Face Reconstruction From Radar Images

Valentin Braeutigam Vanessa Wirth Ingrid Ullmann Christian Schüßler

Martin Vossiek Matthias Berking Bernhard Egger

Friedrich-Alexander-Universität Erlangen-Nürnberg

Abstract

The 3D reconstruction of faces gains wide attention in computer vision and is used in many fields of application, for example, animation, virtual reality, and even forensics. This work is motivated by monitoring patients in sleep laboratories. Due to their unique characteristics, sensors from the radar domain have advantages compared to optical sensors, namely penetration of electrically non-conductive materials and independence of light. These advantages of radar signals unlock new applications and require adaptation of 3D reconstruction frameworks. We propose a novel model-based method for 3D reconstruction from radar images. We generate a dataset of synthetic radar images with a physics-based but non-differentiable radar renderer. This dataset is used to train a CNN-based encoder to estimate the parameters of a 3D morphable face model. Whilst the encoder alone already leads to strong reconstructions of synthetic data, we extend our reconstruction in an Analysis-by-Synthesis fashion to a model-based autoencoder. This is enabled by learning the rendering process in the decoder, which acts as an object-specific differentiable radar renderer. Subsequently, the combination of both network parts is trained to minimize both, the loss of the parameters and the loss of the resulting reconstructed radar image. This leads to the additional benefit, that at test time the parameters can be further optimized by finetuning the autoencoder unsupervised on the image loss. We evaluated our framework on generated synthetic face images as well as on real radar images with 3D ground truth of four individuals. The dataset is available at `https://doi.org/10.5281/zenodo.14264739`.

1. Introduction

Reconstruction of humans using alternative capturing systems other than optical sensors has become an increasingly studied field of research in the past few years [7, 33, 34, 36]. Among them, a category gaining increasing interest is millimeter wave (mmWave) radar, which offers a significant advantage over other methods. Due to its wavelength, it is capable of penetrating certain obstacles, for example, fabric [1] or even walls [36], that impede the view of the object of interest. The technology is already in widespread use at airports for security screening prior to flight. In 2019 172 airports in the United States of America used these scanners [30]. However, it has the potential for application in several other fields, where radar signals can be used for recognition and reconstruction tasks.

An example where it can be of significant benefit in the future is the monitoring of patients in clinics and sleep laboratories. Radar imaging offers the benefit compared to optical sensors in that it is not reliant on light and is able to monitor patients in their beds without removing any pillows or bed sheets. Consequently, this avoids the need for the patient to leave their bed for some examinations and allows their body to be monitored during night. This potentially fosters automatic over manual sleep monitoring in the future.

Preliminary work has already been conducted, for example, radar-based techniques to track the human body pose during sleep [33]. In this paper, we focus on the 3D reconstruction of human faces with their identity and expression. The face conveys a lot of information via facial expressions, which can be reconstructed, making it a key factor for assessing the state of patients at night. However, due to the variety of facial shapes, it is challenging to reconstruct faces accurately. One significant challenge that arises from radar reconstruction of faces is the dependency of the viewing angle of the radar system and the face that is captured since the reflected signals depend on the surface normals [31]. Therefore, not all parts of the face are visible in the radar images which makes them ambiguous, as we show in the Supplementary Material.

This paper presents a learning-based approach for reconstructing 3D faces from radar images, utilizing a 3D morphable face model (3DMM). To this end, we generate a synthetic radar image dataset from faces constructed from the Basel Face Model (BFM) 2019 [11]. Given the data, we train two architectures: an encoder that is trained fully-supervised and an autoencoder that combines our pre-trained encoder with a learned differentiable renderer, thereby imposing an additional form of supervision to our network and enabling optimization at test time.

Our contributions are as follows:

- a model-based 3D face reconstruction which is the first to only operate on images computed from radar signals
- an end-to-end training by approximating the physics-based radar renderer with a neural network which is differentiable and can generate synthetic radar images faster
- a publicly available dataset containing 10,000 synthetic radar images of faces generated with the BFM 2019 [11] and their corresponding parameters

2. Related Work

Face Reconstruction From RGB Images. In the case of conventional RGB images, the 3D reconstruction of human faces has already been widely investigated, with a multitude of approaches proposed to address this challenge as summarized by Egger et al. [10]. These methods employ either a face model as prior [9, 22] or directly estimate the points of the face [27]. Examples of commonly utilized publicly available face models are the BFM [11, 21] or, the FLAME face model [18], which are created from 3D scanned heads of a large group of people. The approaches can be further divided into two categories: learning-based approaches [9, 22, 27, 29], and learning-free approaches [3]. Some of these methods are based on landmarks or keypoints [3, 29], while others do not rely on landmarks [27], but utilize an image-to-image approach, which maps the input image to a depth and a face correspondence image.

In regard to our work, the learning-based method proposed by Chang et al. [5, 6] is of particular significance, as a part of our architecture is based on their findings on face reconstruction from RGB images. They combine two ResNet-101 models [12], one trained to predict shape parameters, another one to predict facial expression parameters, and an AlexNet model [17] which estimates pose parameters. Subsequently, they apply the parameters to the BFM 2017 [11] to construct a 3D mesh of the corresponding face.

Another method that motivated our work is presented by Tewari et al. [28]. They employ a convolutional neural network (CNN) encoder to predict the parameters of their parametric face model as well as camera and lighting parameters. A differentiable renderer is then utilized to generate an image from these parameters which completes the model-based autoencoder. Since there is no differentiable radar renderer available, and building one is highly non-trivial (if possible), we implement a differentiable autoencoder with a learned renderer. The core benefit of such a model-based autoencoder is that it can be trained unsupervised on the reconstruction task using a photometric loss. The work of Li et al. [8] utilizes the same autoencoder principle and focuses on reconstruction under occlusion.

Human Reconstruction From Radar Signals. In contrast to the reconstruction of objects in an optical setup, methods for reconstruction with radar signals rarely operate on images, but commonly on raw signals [7, 33–36]. A subfield that is investigated is human body reconstruction, where the body is obscured by obstacles or impaired vision due to darkness or weather influences [7, 35, 36].

Chen et al. [7] utilize radar signals in conjunction with RGB videos to reconstruct a full-body model of humans in visibility-impairing weather, for example, rain or fog, where optical sensors perform insufficiently. Other authors present methods that employ solid obstacles in their evaluation and utilize electromechanical signals within the WiFi frequency range to reconstruct the full body of humans hidden from view by walls or clothes [35, 36]. In order to achieve this, they utilize the SMPL body model [20] as a prior geometry and then utilize a network based on the concept of Mask R-CNN [13] to predict its shape. Yue et al. [33] reconstruct the pose of a human in the dark during sleep via WiFi signals. The objective of these works is to reconstruct the entire human body or to determine its overall pose, whereas we focus on human faces.

Zhang et al. [34] classify facial expressions with radar signals through the detection of facial muscle movements using a mmWave radar sytem. With their approach they are capable of detecting facial expressions in a constrained setting with an accuracy of 80.57%. Our work goes one step further and reconstructs the whole 3D face.

Xie et al. [32] are the first to perform 3D face reconstruction on radar signals. They use a learning approach based on the ConvNeXt-model [19] for facial landmark prediction and a FLAME [18] model to create a continuous face from the landmarks. In our approach, we predict face model parameters to predict the whole face at once, without using landmarks, and in addition propose an autoencoder framework enabling optimization in an Analysis-by-Synthesis fashion which further improves the reconstruction.

While all of the aforementioned methods directly operate on radar signals without converting them to an image, some radar-based approaches operate on images. Bräunig et al. [14] present an approach for 3D reconstruction that uses the theory of frequency shift keying to increase the reconstruction speed of hands compared to classical approaches. The authors utilize point clouds for reconstruction and show their results on human hands. Schüßler et al. [23] present a ResNet-based approach to classify hand poses, which are selected from the American sign language system.

In our work we combine image-based approaches proven useful on RGB images and radar imaging techniques.

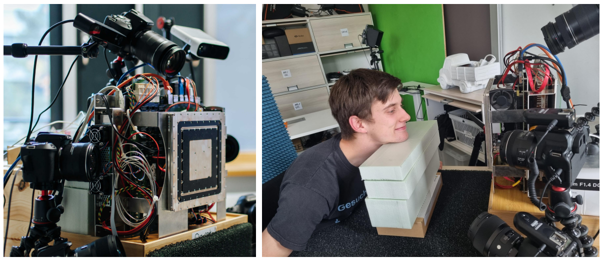

Figure 1. The real radar setup and RGB cameras for photogrammetry. The radar module consists of 94 transmitter and 94 receiver antennas in a square-shaped placement. Around the radar module five cameras are positioned to additionally reconstruct the captured face via photogrammetry. Four persons were captured in this setup, each showing five different facial expressions.

3. Methods

In the following, we introduce the individual components of our work, which include our radar capturing setup, synthetic radar image dataset, our proposed models for model-based 3D reconstruction and the used training parameters.

3.1. Radar Imaging

Radar Setup. We capture real radar images with a multiple-input multiple-output (MIMO) stepped-frequency continuous-wave (SFCW) radar system displayed in Figure 1. The system comprises 94 receiver antennas (RX) and 94 transmitter antennas (TX) in a square-shaped placement and is a submodule of an automotive radome tester [24, 25]. The radar system uses frequencies from 72 GHz to 82 GHz in 128 equidistant steps. The spatial resolution of the radar system is approximately 4 mm in x- and y-direction and 11 mm in z-direction. The face of the person being captured is located at a distance of 25 cm from the radar antennas with foam walls behind her to minimize the reflection of radar signals from there. The aperture extent is approximately 14 cm x 14 cm, with 3 mm spacing between the antennas.

Radar Reconstruction. The radar signals received by the radar system are reconstructed using a state-of-the-art approach for 3D mmWave image processing, namely back-projection [2, 14]. In the course of back-projection, for each RX and TX antenna combination, the correlation between the received signal and a signal hypothesis is summed in a 3D voxel grid [1]. Subsequently, a 2D array is extracted from this 3D voxel grid via maximum projection [15] along the z-axis and the values are scaled logarithmically since they range over several orders of magnitude. Henceafter, the dynamic range θ is restricted to remove noise induced by the radar signals. In our experiments we use a dynamic range of -15 dB for the decoder, but we sample from a range of values for encoder training. Afterwards, the range of the resulting data is scaled linearly to the range [0,1] to convert it to a radar image, which we call an amplitude image.

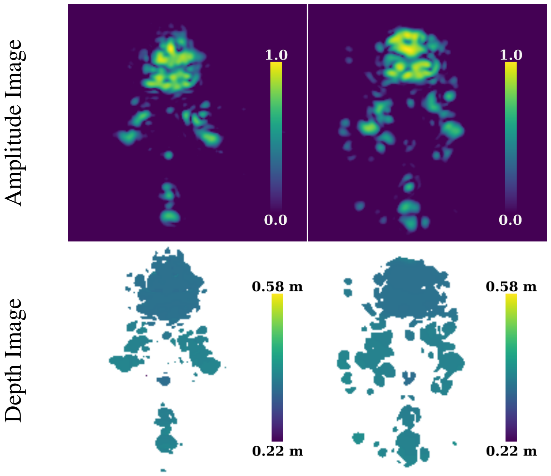

Figure 2. Examples for a real radar images (left) and synthetic-real radar images (right). The amplitude images have a dynamic range of -20 dB. The synthetic real images are generated from the mesh of the same person reconstructed via photogrammetry.

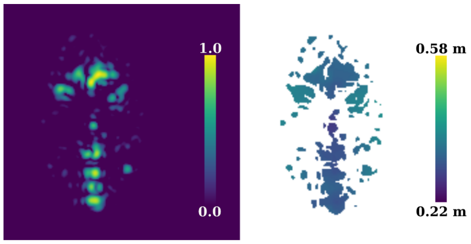

Figure 3. Examples of a synthetic amplitude image (left) with a dynamic range of -20 dB and a synthetic depth image in comparison (right).

Multimodal Capture. During capturing the radar signals, optical cameras are employed simultaneously to reconstruct the face via photogrammetry. Once the 3D reconstruction of the faces has been completed, they are used in the radar simulation, resulting in a synthetic version of the real face. The resulting images can be utilized for a quality comparison of the radar simulation. The capture setup configuration is illustrated in Figure 1. A comparison between the real amplitude image and the synthetic amplitude image generated from the photogrammetry mesh is shown in Figure 2.

In order to evaluate our approach on real data, we reconstructed real radar images of four male european individuals with fair skin and generated synthetic radar images from the corresponding photogrammetry mesh.

Depth Images. We employ depth images generated based on the radar signal as an additional input. The depth of each pixel is computed from the brightest scatterer along each depth slice of the voxel grid. In our experiments, these depth images are utilized once as an alternative to and once in conjunction with the amplitude images as input data.

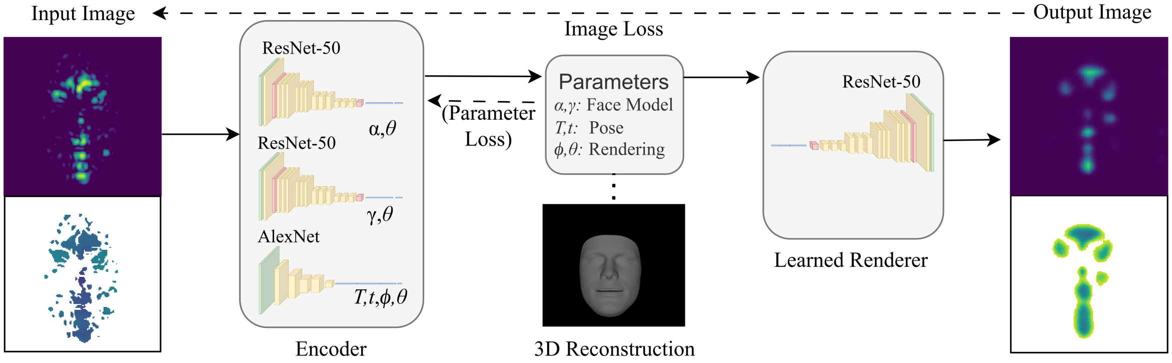

Figure 4. Overview of our method. The input image is fed to three encoder networks which predict the shape, expression, and pose of the face. These parameters are then fed to a differentiable renderer that reconstructs the input image. The encoder consists of two ResNet-50 models for predicting shape and expression and an AlexNet model for predicting the pose. The differentiable renderer is a ResNet-50 model that is ordered in the opposite direction. During training, both the parameter loss and image loss are applied. For inference, the encoder and decoder are frozen and only the image loss is optimized leading to the face model parameters holding the 3D face reconstruction. The idea is that the training with the decoder network leads to better results than training the encoder isolated, which is shown by [28] for a different task.

3.2. Training Dataset

We generated a synthetic dataset of 10,000 instances of face meshes with varying facial expressions. To this end, the face mask *face12* of the BFM 2019 [11] was utilized.

The generation of each face is based on a Gaussian-sampled shape vector α and an expression vector γ with a distribution of $\mathcal{N}(0, 1)$. The aforementioned vectors define the appearance of each face instance and the associated facial expression, as specified by the shape model. Additionally, the face is transformed by a uniformly sampled pose. The pose includes a translation t in each x- and y-direction between -5 cm and +5 cm, the z-value which is set to 0 cm, rotations T around the yaw axis within [-5, 5] degrees, and the pitch and roll axes within [-10, 10] degrees, respectively. For training and evaluation, the pose parameters are linearly scaled to values within [-1, 1].

Subsequently, the aforementioned meshes are employed to generate synthetic radar images utilizing the radar renderer by Schüssler et al. [26], which was adjusted to generate images of human hands by Bräunig et al. [4]. The renderer approximates the radar signals of a near-field MIMO radar system by raytracing and performs the back-projection algorithm [2, 14], analogously to real radar signals. To model the reflection property Schüssler et al. use a material factor to interpolate linearly between a diffuse and a specular material. Another attribute is the size of the simulated antennas, which differs from the real antenna size to assimilate the difference between real radar waves and simulated rays. For our dataset, we sample these two rendering parameters ϕ, the material factor within [0, 1], and the antenna size within [0.2, 0.3]. We then train the neural networks on these training images using the procedures outlined in the subsequent Sections 3.3 and 3.4.

3.3. Encoder

Two CNNs with the ResNet-50 [12] architecture are employed for the prediction of shape and expression parameters of the face model. Furthermore, an AlexNet model [17] is utilized to predict the pose. The outputs of the ResNet-50 models are scaled by applying a tanh-layer and multiplying the results with a scaling factor of three since this is the value range that contains 99.8 % of the values sampled from the dataset while still reducing the possible value range. For all networks, we utilize an additional fully connected layer that outputs the expected amount of parameters. The encoder architecture is based on Chang et al. [5, 6], while we use smaller networks that enable it to still run on a single GPU while adjusting to the later presented autoencoder architecture without decreasing the performance substantially. Additionally, we do minor adjustments like the tanh-layer. Experiments to combine all predictions into a single ResNet-50 model led to a decrease of performance.

The aforementioned models are trained on a training set of 8,500 synthetic radar images and evaluated on the remaining 1,500 images. During training, we apply randomly sampled dynamic ranges between -15 dB and -30 dB, for evaluation we apply a fixed dynamic range of -20 dB. The models predict the first 3DMM parameters $\alpha \in \mathbb{R}^{10}$ and $\gamma \in \mathbb{R}^{7}$, which cover approximately 85% of the shape variance and 76% of the expression variance of the BFM, and the pose parameters $t \in \mathbb{R}^3$ and $T \in \mathbb{R}^3$ applied to the face model mesh. Additionally, the networks estimate the applied dynamic range, the material, and antenna size properties. During training phase, the L2 loss between those parameters and the resulting parameters is employed as a loss function.

3.4. Learned Renderer and Autoencoder

Learned Renderer. As we like to achieve an unsupervised optimization, we need a renderer that generates a radar image based on our parameters. Since the renderer we used for creating our dataset is not fast enough for the training and also not differentiable, we approximate the physics-based renderer through training a decoder network on generated images. Therefore, we use a ResNet-50, with its layers ordered in the opposite direction. A fully connected layer is employed to map the input to the first convolutional layer of the ResNet-50. This renderer is trained on the parameters to generate amplitude images with an applied dynamic range of -15 dB, and depth images, respectively. It utilizes the face model, pose and rendering parameters as described in Section 3.2 to render synthetic radar images. We can then employ it to improve the training of our model.

Autoencoder. The learned renderer is combined with the encoder as displayed in Figure 4. In the case of the autoencoder training, the pre-trained models of the encoder and the decoder are utilized, with the weights of the latter being fixed. This ensures that inductive constraints of the parameters in the latent space do not experience structural changes induced by the autoencoder, but rather remain aligned with the original intention of face model parameters. Subsequently, the autoencoder is trained on the synthetic image set. The prediction of face model parameters is combined with the task of reconstructing the input image from the parameters.

The training loss is computed as follows:

$$\mathrm{L}_{train} = \mathrm{L}_{image} + \lambda \cdot \mathrm{L}_{params}, \tag{1}$$

where L_{image} denotes the L2 loss between output and input image, and L_{params} the L2 loss between the parameters computed by the encoder part (α, γ, T, t) and their ground truth. λ is a weighting factor to adjust the importance of the loss functions relative to each other. Since the results have not changed substantially for other values of λ, we use $\lambda = 1$. The network is trained on the same parameters as the encoder, described in the previous Section 3.3.

During evaluation, both the encoder and the decoder are fixed and the latent space variables are further optimized by the image loss.

3.5. Model Training

The Adam optimizer [16] is employed in the training of our network models, utilizing a Nvidia A100 graphics card. The encoder is trained with linearly scheduled learning rates in [0.01, 0.001], over the initial 150 epochs of the training period and over 200 epochs in total. To train the decoder and the autoencoder a learning rate of 0.001 is employed without learning rate scheduling over 300 epochs. We train the encoder and autoencoder in batches of 50 images, and the decoder in batches of 150 images. The batch sizes were selected to make use of the graphics card memory. The decoder has a higher batch size as the model is trained separately. The autoencoder takes approximately 4 hours of training time, the encoder 2 hours and the decoder 1.5 hours, the training with radar and depth images combined takes 0.75 to 1 hour longer due to the higher amount of data.

Runtime. In the mean it takes 58 ms to get the resulting image from the trained decoder models, while the physics-based renderer used for the creation of the dataset takes about 2 min to reconstruct an image. Both measurements were performed on a Nvidia GeForce RTX 4060 Ti.

4. Experiments & Results

In the following the encoder and autoencoder architectures are evaluated quantitatively and qualitatively. Both network types are trained and evaluated using three types of radar input data: *amplitude images*, *depth images* that are reconstructed from radar signals, and the combination of both image types in two channels of an image (*amplitude-depth images*). We further classify our data into three categories: synthetic images generated from synthetic meshes (*synth data*), synthetic images from photogrammetry reconstructions of real faces (*real-to-synth data*), and real images from real faces (*real data*).

Quantitative Results. The L2 error between the predicted and ground truth face model parameters of the faces in our validation set is presented in Table 1. Additionally, the mean Euclidean point distances between all corresponding mesh points of the face model meshes created with the result parameters are compared, without applying the pose. We compare these errors to the baseline, which is calculated by the error between the mean of the training set parameters and each face instance of the validation set. Furthermore, we use the error of randomly sampled parameter with the same distribution as the ground truth parameters as an additional baseline.

All variants of our methods demonstrate superior outcomes compared to these baselines. The results of the encoder and autoencoder demonstrate that the L2 error of the resulting shape and expression vectors and, since the pose error is similar, the total error for the autoencoder is smaller than for the encoder. Furthermore, it can be observed that the results for the shape vector for both, the encoder and the autoencoder version, demonstrate a reduction in L2 error and the point distance of the corresponding meshes when the reconstructed depth image is incorporated.

In the following evaluations, we compare the shape and expression parameter results for *real input*, *real-to-synth input* and additionally generated *synthetic input* which consists of

Method	L2 Shape ↓	L2 Expr ↓	L2 Translation ↓	L2 Rotation ↓	L2 Total ↓	Point Dist ↓
Baseline (Mean)	0.9905	0.9996	0.3316	0.2264	0.8077	4.42 mm
Baseline (Random)	2.0123	2.0593	0.6753	0.4467	1.6480	6.28 mm
Encoder (Amplitude)	0.8400	0.9212	0.0687	0.0063	0.6554	3.47 mm
Encoder (Depth)	0.7398	0.9103	0.0646	0.0053	0.6078	2.77 mm
Encoder (Amp.-Depth)	0.7350	0.8848	**0.0613**	**0.0051**	0.5975	2.82 mm
Autoencoder (Amplitude)	0.7705	**0.8151**	0.0794	0.0070	0.5943	3.29 mm
Autoencoder (Depth)	**0.6338**	0.8499	0.0657	0.0062	0.5436	**2.56 mm**
Autoencoder (Amp.-Depth)	0.6400	0.8202	0.0779	0.0071	**0.5390**	2.61 mm

Table 1. L2 error of the parameters computed by the encoder and the autoencoder evaluated on the synthetic validation set and the mean point distance of all corresponding mesh points. The mean L2 error for the evaluated parameters is given and split into the shape and expression parameters by the face model and the translation and rotation of the face mesh. The point distance is computed by the mean Euclidean distance of each point pair of the resulting mesh and the ground truth mesh, without applying the pose. The models are trained on synthetic amplitude images, depth images, and amplitude-depth images. The estimation of the pose has similar quality for all approaches, while the autoencoder performs better for shape and expression estimation and has a lower point distance, which are the core tasks of interest. The models trained on depth or amplitude-depth input perform better in the shape estimation and have a lower point distance.

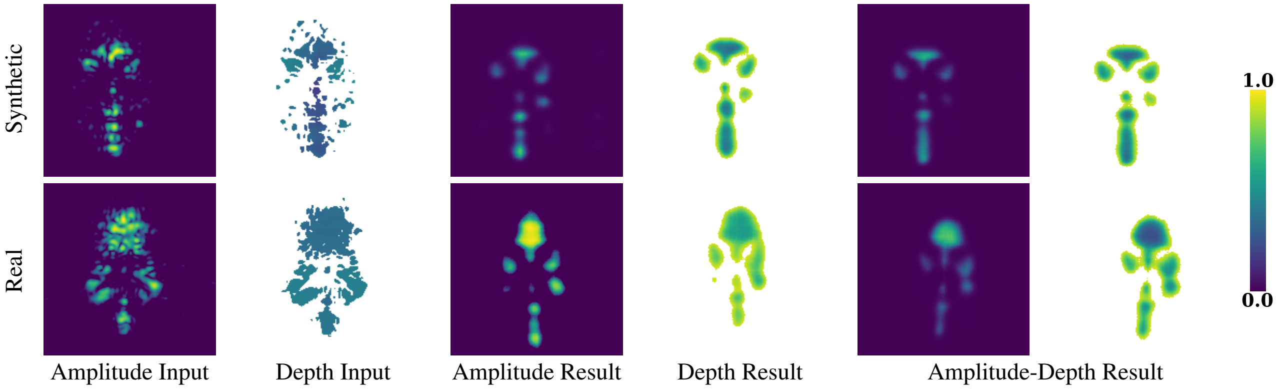

Figure 7. Reconstructions of the autoencoder models with the different settings. The first row contains the results for the synthetic radar image input for the differently trained models, while the second row contains the results for the real radar image input.

shape instances having the same five sampled expressions. The sampling distributions are the same as in Section 3.2. The evaluation shows correlations between the input and result parameters, in particular for the real images. The Figures 5 and 6 illustrate the comparison of the resulting faces with one another by computing the cosine similarity between the face model parameters and the parameters of the other face instances. Each cell contains the mean of face instances with the same shape or expression, respectively. Consequently, the similarity of faces can be compared to each other. The faces with the same ground truth shape, respectively expression, are anticipated to have the highest similarity. As a reference the cosine similarity of the ground truth parameters is presented in the Supplementary Material. Figure 5a shows the comparison of the *synthetic* shape vectors computed with an autoencoder trained on amplitude-depth images. The diagonal displays a high degree of similarity between vectors with the same ground truth shape, in comparison to other combinations of face vectors. While there are instances of two faces where the similarity between the shape vectors is also high, the values on the diagonal are the highest for most of the rows. Figure 5b illustrates the results for the same faces but without a pose. In this plot, the diagonal of high values is more prominent compared to the plot with a sampled pose.

The results for *real* amplitude and amplitude-depth images to the autoencoder are shown in Figures 6a and 6b. The shape results demonstrate higher values on the diagonal compared to the mean of the other values in the plot. With the exception of the first face in Figure 6b, these values represent the highest in each row/column. In Figure 6c the *real-to-synth* results are evaluated, which show a diagonal of high values for the shape evaluation as well. Of the three variants, only the autoencoder evaluated on real amplitude-depth images shows a diagonal when comparing the expression parameters. The encoder results and remaining autoencoder results can be found in the Supplementary Material.

To summarize, the comparison of the L2 loss shows that the autoencoder variants perform better in predicting the shape and expression parameters, while the use of the depth input

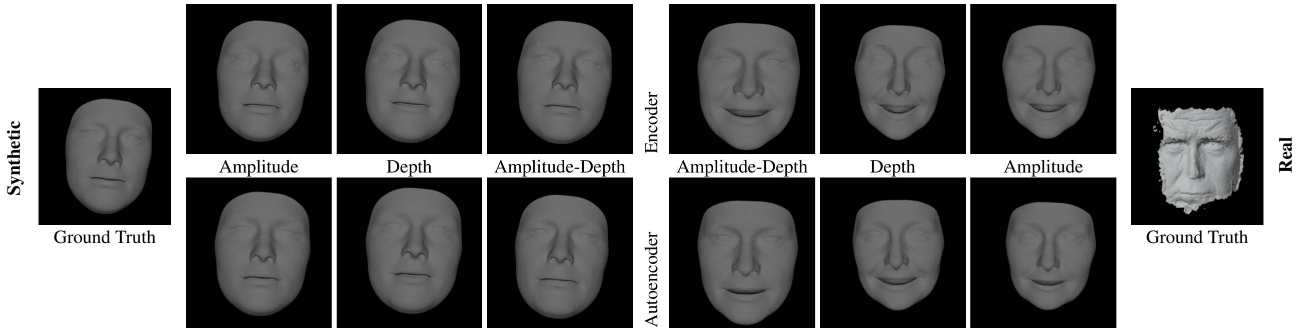

Figure 8. The resulting meshes generated with parameters of the different models. The left half depicts the results from the models evaluated on synthetic data, and the right half the results from the models evaluated on real data.

ids	0	1	2	3	4	5	6	7	8	9
0	0.583	0.489	0.214	-0.157	0.208	-0.383	0.039	0.093	-0.342	-0.114
1	0.489	0.717	0.551	-0.277	0.185	-0.197	-0.112	0.258	-0.199	-0.002
2	0.214	0.551	0.678	-0.213	0.257	-0.031	-0.045	0.196	-0.178	-0.128
3	-0.157	-0.277	-0.213	0.522	-0.121	-0.024	0.349	-0.137	0.060	-0.078
4	0.208	0.185	0.257	-0.121	0.133	0.039	-0.023	0.058	-0.058	-0.037
5	-0.383	-0.197	-0.031	-0.024	0.039	0.771	-0.502	0.165	0.464	0.371
6	0.039	-0.112	-0.045	0.349	-0.023	-0.502	0.599	-0.090	-0.026	-0.282
7	0.093	0.258	0.196	-0.137	0.058	0.165	-0.090	0.385	0.436	0.223
8	-0.342	-0.199	-0.178	0.060	-0.058	0.464	-0.026	0.436	0.835	0.412
9	-0.114	-0.002	-0.128	-0.078	-0.037	0.371	-0.282	0.223	0.412	0.299

Shape

ids	0	1	2	3	4
0	0.311	0.038	-0.000	-0.114	-0.132
1	0.038	0.635	0.387	-0.170	-0.165
2	-0.000	0.387	0.363	0.057	-0.026
3	-0.114	-0.170	0.057	0.297	0.210
4	-0.132	-0.165	-0.026	0.210	0.195

Expression

(a) Comparison of the shape parameters per shape instance (top) and expression parameters per expression instance (bottom) with an **uniformly sampled pose** as described in Section 3.2.

ids	0	1	2	3	4	5	6	7	8	9
0	0.741	0.180	0.047	0.320	0.446	0.037	0.053	0.362	0.135	0.121
1	0.180	0.297	0.273	0.206	0.366	0.213	0.027	0.140	0.135	0.004
2	0.047	0.273	0.865	0.195	0.408	0.275	0.198	0.237	0.368	0.141
3	0.320	0.206	0.195	0.685	0.181	0.363	0.468	0.628	0.179	0.222
4	0.446	0.366	0.408	0.181	0.650	0.032	0.208	0.141	0.285	0.153
5	0.037	0.213	0.275	0.363	0.032	0.669	0.178	0.553	0.095	0.049
6	0.053	0.027	0.198	0.468	0.208	0.178	0.666	0.245	0.106	0.536
7	0.362	0.140	0.237	0.628	0.141	0.553	0.245	0.802	0.270	0.075
8	0.135	0.135	0.368	0.179	0.285	0.095	0.106	0.270	0.701	0.135
9	0.121	0.004	0.141	0.222	0.153	0.049	0.536	0.075	0.135	0.470

Shape

ids	0	1	2	3	4
0	0.317	0.062	0.006	0.238	0.252
1	0.062	0.694	0.121	0.087	0.202
2	0.006	0.121	0.168	0.126	0.039
3	0.238	0.087	0.126	0.524	0.434
4	0.252	0.202	0.039	0.434	0.347

Expression

(b) Comparison of the shape parameters per shape instance (top) and expression parameters per expression instance (bottom) with a **neutral pose**.

Figure 5. Cosine similarity comparison between the face model parameters computed by the autoencoder evaluated on *synthetic* data input with a uniformly sampled pose (top) and a neutral pose (bottom).

ids	0	1	2	3
0	0.452	-0.357	0.258	-0.023
1	-0.357	0.431	-0.110	0.134
2	0.258	-0.110	0.601	0.130
3	-0.023	0.134	0.130	0.206

Shape

ids	0	1	2	3	4
0	0.361	0.251	0.149	-0.006	0.165
1	0.251	-0.001	0.102	-0.185	0.094
2	0.149	0.102	0.104	-0.024	0.024
3	-0.006	-0.185	-0.024	0.170	0.032
4	0.165	0.094	0.024	0.032	-0.212

Expression

(a) Comparison of the shape and expression parameters computed by the autoencoder evaluated on **real amplitude images**.

ids	0	1	2	3
0	0.372	0.274	0.574	0.144
1	0.274	0.935	0.347	-0.135
2	0.574	0.347	0.635	0.189
3	0.144	-0.135	0.189	0.346

Shape

ids	0	1	2	3	4
0	0.027	0.222	-0.097	0.020	0.260
1	0.222	0.413	0.050	-0.147	0.245
2	-0.097	0.050	0.217	-0.104	-0.101
3	0.020	-0.147	-0.104	0.479	0.074
4	0.260	0.245	-0.101	0.074	0.043

Expression

(b) Comparison of the shape and expression parameters computed by the autoencoder evaluated on **real amplitude-depth images**.

ids	0	1	2	3
0	0.460	0.357	0.065	0.149
1	0.357	0.600	-0.015	0.229
2	0.065	-0.015	0.471	0.423
3	0.149	0.229	0.423	0.286

Shape

ids	0	1	2	3	4
0	0.258	0.275	0.136	0.176	0.016
1	0.275	0.173	0.159	0.042	0.116
2	0.136	0.159	0.173	0.031	-0.022
3	0.176	0.042	0.031	-0.055	0.033
4	0.016	0.116	-0.022	0.033	-0.184

Expression

(c) Comparison of the shape and expression parameters computed by the autoencoder evaluated on the **real-to-synth amplitude images**.

Figure 6. Cosine similarity comparison between the shape and expression parameters from the autoencoder results derived from real data. The plots present the shape parameters (left) and expression parameters (right) grouped by instances of the same shape or expression, respectively.

improves the shape reconstruction. The cosine plots show that there is a correlation between the real face instance and the shape parameter output of the models, while there is no visible correlation for the expression vectors. We also showed that the pose influences the shape and expression parameter prediction.

Qualitative Results. Qualitative results are presented in Figures 7 and 8. For both, real and synthetic data, the image results appear to be a blurred version of the input, wherein smaller details, visible noise, and radar signal patterns have been removed. This is expected, as the physics-based renderer involves a random component.

Figure 8 compares meshes created with the resulting parameters across the different input types. In the case of synthetic images, the results appear similar across the different methods. However, in the case of real input, the meshes exhibit greater variance and appear to deviate more from the ground truth. The predicted pose is found to be in close alignment with the ground truth pose for all methods.

5. Discussion

Evaluation. The results of our experiments demonstrate that the autoencoder, as outlined in Section 3.4, exhibits superior performance in predicting face model shape and expression parameters compared to the fully-supervised trained encoder, as described in Section 3.3.

We thus conclude that the additional training with the decoder has a beneficial effect on the training process, due to the image reconstruction task and its role in regularizing the loss of the network. Furthermore, the decoder component, which has been trained as a learned renderer, can be utilized to generate images with an appearance similar to radar images. Since it is differentiable and computes the images at a significantly faster rate compared to the physics-based renderer, it can be further utilized to finetune the parameters by fixing the encoder and decoder and optimize the image loss.

In terms of predicting the shape parameters, the depth images appear to demonstrate superior performance compared to the amplitude images, across both the autoencoder and encoder models. We assume that this is because the face pixel values are more evenly distributed and there are less small region peaks with high values, since this is the major difference between the image types.

Limitations. Despite the promising outcomes revealed by our evaluation, our work is not without limitations, thereby suggesting avenues for future research. One challenge is the domain gap between synthetic and real images which decreases the resulting quality for real input. It appears in different patterns between the radar simulator and the real radar images, and in different scales between the synthetic meshes and the real faces. Ideally, we could, like Tewari et al. [28], train our autoencoder on real images, but we are restricted to a limited real test set, which is also limited in variance since all radar captured persons are male and from a similar geographical region. Another limitation is the approximation of the reflectance properties of skin by the physics-based renderer, which restricts our model in being physical correct.

Future Work. In future experiments, we propose collecting a larger dataset of real data incorporating a wider variety of human subjects for radar capture.

Another avenue is to evaluate each viewing angle by the expected output quality of the face reconstruction, which can be insightful and useful in practice. This topic is already under investigation as current publications show [31].

The method we present is generalizable to other applications than monitoring patients as long as it is a static environment and the person is within radar-capturing distance.

6. Conclusion

We presented an approach for face reconstruction based on radar images and the BFM 2019 [11]. We generated a dataset of 10,000 synthetic radar images from a physics-based radar simulator based on meshes created with the BFM 2019 and trained three CNN models fully-supervised to reproduce the ground truth parameters. Furthermore, a learned renderer was employed, trained to render radar images derived from the combination of the face model parameters and pose. This learned renderer is used to generate representative images close to the appearance of real data but significantly faster (more than 2000 times) and fully differentiable. While the reconstruction of faces only given radar data remains challenging, we demonstrated that the joint training process improves the reconstruction quality compared to the fully-supervised training approach.

On synthetic data, we achieve a mean Euclidean 3D point distance of 2.56 mm of the face meshes without applying the pose. Furthermore, the qualitative results appear visually similar to the ground truth faces. In the case of real data, we show that we can perform face recognition since instances of the same shape, but with a different pose and expression, have a higher similarity compared to other faces. However, the recognition of expressions is only possible with identity-specific trained models. The pose is consistent with the ground truth, however, for real data there is no discernible correlation between the ground truth and the resulting meshes concerning shape and expression in most of the results. The core benefit of our method is that it can be trained in an unsupervised fashion as model-based autoencoder on a large set of radar images without explicit 3D supervision which enables large scale training on real data. With our approach, we anticipate to guide future research towards higher-fidelity reconstructions.

7. Acknowledgements

This work was supported in part by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1483 – Project-ID 442419336, EmpkinS. This work was partially funded by the German Federal Ministry of Education and Research (BMBF), FKZ: 01IS22082 (IRRW). The authors are responsible for the content of this publication. The authors would like to thank the Rohde & Schwarz GmbH & Co. KG (Munich, Germany) for providing the radar imaging devices and technical support that made this study possible. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG). We thank Nikolai Hofmann and Paul Himmler for proofreading the paper.

References

[1] Sherif S. Ahmed. Microwave imaging in security — two decades of innovation. *IEEE Journal of Microwaves*, 1(1):191–201, 2021. 1, 3

[2] Sherif Sayed Ahmed, Andreas Schiessl, Frank Gumbmann, Marc Tiebout, Sebastian Methfessel, and Lorenz-Peter Schmidt. Advanced Microwave Imaging. *IEEE Microwave Magazine*, 13(6):26–43, 2012. 3, 4

[3] Brian Amberg and Thomas Vetter. Optimal landmark detection using shape models and branch and bound. In *2011 International Conference on Computer Vision*, pages 455–462, 2011. ISSN: 2380-7504. 2

[4] Johanna Bräunig, Christian Schüßler, Vanessa Wirth, Marc Stamminger, Ingrid Ullmann, and Martin Vossiek. A realistic radar ray tracing simulator for hand pose imaging. In *2023 20th European Radar Conference (EuRAD)*, pages 238–341, 2023. 4

[5] Fengju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. FacePoseNet: Making a Case for Landmark-Free Face Alignment. *arXiv: Computer Vision and Pattern Recognition*, 2017. arXiv:1708.07517 [cs]. 2, 4

[6] Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. ExpNet: Landmark-Free, Deep, 3D Facial Expressions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 122–129, 2018. 2, 4

[7] Anjun Chen, Xiangyu Wang, Kun Shi, Shaohao Zhu, Bin Fang, Yingfeng Chen, Jiming Chen, Yuchi Huo, and Qi Ye. ImmFusion: Robust mmWave-RGB Fusion for 3D Human Body Reconstruction in All Weather Conditions. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2752–2758, London, United Kingdom, 2023. IEEE. 1, 2

[8] Chunlu Li, Andreas Morel-Forster, T. Vetter, B. Egger, and Adam Kortylewski. Robust Model-based Face Reconstruction through Weakly-Supervised Outlier Segmentation. *arXiv: Computer Vision and Pattern Recognition*, 2021. 2

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D Face Reconstruction with Weakly-Supervised Learning: From Single Image to Image Set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 2

[10] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG)*, 39(5):1–38, 2020. 2

[11] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Lüthi, Sandro Schoenborn, and Thomas Vetter. Morphable Face Models - An Open Framework. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82, 2018. 2, 4, 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 4

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[14] J. Bräunig, V. Wirth, Christoph Kammel, Christian Schüßler, I. Ullmann, M. Stamminger, and M. Vossiek. An Ultra-Efficient Approach for High-Resolution MIMO Radar Imaging of Human Hand Poses. *IEEE Transactions on Radar Systems*, 2023. S2ID: aecd085eab99912652a22a413d103db54b91a327. 2, 3, 4

[15] P J Keller, B P Drayer, E K Fram, K D Williams, C L Dumoulin, and S P Souza. MR angiography with two-dimensional acquisition and three-dimensional display. work in progress. *Radiology*, 173(2):527–532, 1989. 3

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv: Computer Science - Machine Learning*, 2017. 5

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. 2, 4

[18] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, 36 (6):194, 2017. 2

[19] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16, 2015. 2

[21] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments*, 2009. 2

[22] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. Learning to Regress 3D Face Shape and Expression From an Image Without 3D Supervision. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7755–7764, Long Beach, CA, USA, 2019. IEEE. 2

[23] Christian Schuessler, Wenxuan Zhang, Johanna Bräunig, Marcel Hoffmann, Michael Stelzig, and Martin Vossiek. Radar-based recognition of static hand gestures in american sign language. In *2024 IEEE Radar Conference (RadarConf24)*, pages 1–6, 2024. 2

[24] Rohde & Schwarz. IMAGER, [online] available: https://www.rohde-schwarz.com/de/produkte/messtechnik/microwave-imaging/imager_256948.html/, Nov. 2024. 3

[25] Rohde & Schwarz. QAR50 automotive radome tester, [online] available: https://www.rohde-schwarz.com/qar50/, Oct. 2024. 3

[26] Christian Schüßler, Marcel Hoffmann, Johanna Bräunig, Ingrid Ullmann, Randolf Ebelt, and Martin Vossiek. A Realistic Radar Ray Tracing Simulator for Large MIMO-Arrays in Automotive Environments. *IEEE Journal of Microwaves*, 1 (4):962–974, 2021. Conference Name: IEEE Journal of Microwaves. 4

[27] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation, 2017. arXiv:1703.10131 [cs]. 2

[28] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Florian Bernard, Patrick Pérez, Patrick Pérez, and Christian Theobalt. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. *arXiv: Computer Vision and Pattern Recognition*, 2017. 2, 4, 8

[29] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. Regressing Robust and Discriminative 3D Morphable Models with a Very Deep Neural Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1502, Honolulu, HI, 2017. IEEE. 2

[30] tripsavvy. Which airports have full body scanners?, [online] available: https://www.tripsavvy.com/which-airports-have-full-body-scanners-3150257, Oct. 2024. 1

[31] Vanessa Wirth, Johanna Bräunig, Martin Vossiek, Tim Weyrich, and Marc Stamminger. Maroon: A framework for the joint characterization of near-field high-resolution radar and optical depth imaging techniques. *arxiv: Electrical Engineering and Systems Science - Image and Video Processing*, 2024. 1, 8

[32] Jiahong Xie, Hao Kong, Jiadi Yu, Yingying Chen, Linghe Kong, Yanmin Zhu, and Feilong Tang. mm3dface: Nonintrusive 3d facial reconstruction leveraging mmwave signals. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, page 462–474, New York, NY, USA, 2023. Association for Computing Machinery. 2

[33] Shichao Yue, Yuzhe Yang, Hao Wang, Hariharan Rahul, and Dina Katabi. BodyCompass: Monitoring Sleep Posture with Wireless Signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(2):1–25, 2020. 1, 2

[34] Xi Zhang, Yu Zhang, Zhenguo Shi, and Tao Gu. mmFER: Millimetre-wave Radar based Facial Expression Recognition for Multimedia IoT Applications. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–15, Madrid Spain, 2023. ACM. 1, 2

[35] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-Wall Human Pose Estimation Using Radio Signals. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, Salt Lake City, UT, 2018. IEEE. 2

[36] Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Hang Zhao, Tianhong Li, Antonio Torralba, and Dina Katabi. Through-Wall Human Mesh Recovery Using Radio Signals. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10112–10121, Seoul, Korea (South), 2019. IEEE. 1, 2