University of Glasgow | School of Computing Science

mmFace: 3D Face Recognition using RGB and Millimetre Wave Radar

Stergious Aji (2546916A)

April 12, 2024

ABSTRACT

Secure and compact face recognition systems often rely on expensive, non-commercial hardware that must be tailor-made for specific needs. Current systems typically employ near-infrared waves to capture dense 3D maps of human faces due to their high precision. However, this approach has its limitations, since they can be blocked by thin materials such as spectacles and facial hair. With the recent success of low-power millimetre wave radar sensors for hand gesture recognition, we investigate their potential for more resilient 3D face recognition. We propose `mmFace`*, a hybrid end-to-end system that harnesses both RGB and mmWave signatures to accurately recognise facial identities and verify their authenticity, effectively countering 2D spoofing attacks.*

1. INTRODUCTION

Facial recognition is a key area of research within the field of computer vision, finding extensive use across areas including human-computer interaction, security surveillance, and forensic analysis. Its primary application being for biometric authentication, granting individuals access to their devices or restricted zones. This enables a non-intrusive, hands-free means of identity verification, eliminating the need to memorise passwords. Additionally, facial biometrics are naturally more attainable than other modalities such as fingerprints, palm prints, or iris scans [1].

Since its inception in the 1960s, face recognition technology has undergone significant growth. Initially pioneered by Bledsoe [2], early systems distinguished faces by comparing manually annotated landmark features such as the nose, eyes, and mouth. More recently, the emergence of deep learning has amplified the performance of human face classification, benefitting from the vast online repositories of face images. Nonetheless, these systems predominantly rely on images captured by RGB cameras, leaving them vulnerable to variations in lighting and facial pose [3]. By incorporating depth data and drawing attention to the geometric details of the face, the impact of such environmental factors can be mitigated. Furthermore, the transition to 3D facial recognition not only increases accuracy but also bolsters the security of biometric systems against spoofing attacks [4].

1.1 Motivations

The popularity of 3D face recognition is on the rise, evidenced by its adoption in smartphones with the likes of Apple and their Face ID [5] technology. This growing demand has pushed the commercialisation of depth-sensing technology to smaller form factors, facilitating its efficient real-time operation on mobile devices [6]. Face ID, in particular, has garnered a level of security that enables payment authentication within services such as Apple Pay. However, Apple's use of costly proprietary hardware and restrictive patents make it harder for smaller companies to adopt an equally compact and secure face recognition system.

Depth cameras used in this context typically employ an active form of acquisition. This involves projecting non-visible light onto the face, which is then reflected back, allowing sensors to measure and map facial characteristics. Lidar cameras, emitting waves in the near-infrared (NIR) spectrum, are the most prevalent choice given their capacity to acquire a dense 3D map of the subject's face [7]. However, they are often limited by their inability to penetrate thin materials such as clothing and hair. For instance, iPhone users must submit separate facial scans for scenarios involving spectacles or face masks in order for Face ID to operate in all situations [8]. In contrast, millimetre wave radar (mmWaves) can permeate such materials to directly reach the skin's dermal layer [9]. This could enable greater performance in situations involving facial hair, or within adverse environmental conditions such as rain and fog.

1.2 Problem Statement

Research into the efficacy of radar waves for 3D face recognition remains relatively limited, although recent studies indicate promising outcomes [10, 11, 12, 13, 14]. Radar sensors typically offer greater cost efficiency in terms of both acquisition and computation, as they consume less power compared to NIR-based systems. Nevertheless, it is necessary to acknowledge the trade-off, as mmWaves often result in a sparser representation. This could impact recognition performance wherein the precision in detecting and mapping facial features is paramount. Thus, we aim to counterbalance this limitation with the textural information gained from colour images, potentially paving the way for more resilient and versatile systems.

The effectiveness of employing RGB cameras in tandem with mmWave radar sensors for identity verification presents an open research area. As a result, no suitable datasets are available to help address it, compelling us to curate our own. This presents several challenges as we must collect a wide variety of faces under controlled conditions to ensure the generalisability of our model. Additionally, our objective is to leverage the full potential of mmWaves to create a system that is secure against misconduct. This means our model must learn to diagnose face liveness on top of recognising identities, in order to combat 2D spoofing and face concealment. Following data acquisition, the captured radar data must be appropriately preprocessed to enable effective

extraction of facial features from the mmWave signatures. Lastly, there are various strategies that can be employed to fuse the data from both modalities. Each must be empirically analysed to identify the most optimal approach to producing the richest representations.

1.3 Research Contributions

Our work explores developing a compact and secure 3D face recognition system through mmWave sensing. For this, we collated a dataset of face scans from 21 participants under various conditions encompassing diverse poses, lighting settings, and common occlusion scenarios. We used the Intel RealSense L515 RGB-D camera [15] for photographing subject faces. Meanwhile, Google's Soli 60 GHz radar sensor [16] was employed to gather depth information by transmitting and measuring millimetre waves reflected from the target.

We developed a novel face recognition model using a convolutional-based neural network architecture. This model was trained on the captured data to learn facial features from both the RGB and depth information simultaneously. We investigated different strategies in blending the two modalities, aiming to pinpoint the most effective method that provides distinctive embeddings for clean identity separation. The best-performing model is benchmarked against prior radar-based facial recognition systems, as well as, a baseline comparison to using each modality independently. This comprehensive approach led to a system that demonstrates robustness to environmental factors, surpassing those reliant solely on RGB images.

The key contributions of this paper are summarised below:

- A compilation of a diverse face dataset comprising colour images and mmWave signatures from 21 participants. The dataset encompasses five different poses, two lighting conditions, and two occlusion scenarios.
- We present `mmFace`, a hybrid face recognition model that harnesses both modalities yielding a robust system capable of handling face concealment and nullifying 2D spoofing attempts. The model exhibits strong generalisation capabilities to unseen faces and discerns between live and fake faces effectively.
- An empirical analysis of seven feature-level fusion methods is conducted to determine the most optimal approach for blending RGB and mmWave facial features.
- Our models and evaluations are open sourced[1] to facilitate further research into small-scale, 3D face identification using mmWave technology.

2. RELATED WORK

2.1 Millimetre Wave Radar Technology

Radio Detection and Ranging, or Radar, has been around for decades and plays an instrumental role in fields including space exploration, aviation, and maritime navigation. Recently, the miniaturisation of radar sensors to operate in the millimetre wave band has expanded its applicability to more small-scale domains [6]. mmWave sensing has particularly excelled in the autonomous vehicle domain, facilitating object detection for systems such as collision warnings and adaptive cruise control [17]. This is due to its edge over traditional lidar cameras, specifically in its resilience to atmospheric conditions such as dust, smoke, fog, and rain [18]. This penetrative power of mmWaves makes it a promising candidate for reliable facial recognition in uncertain, real-world scenarios.

Another notable example is Google's integration of their Soli sensor into the Pixel 4 smartphones for motion detection and gesture recognition [19]. However, the sensor's potential application to face identification remains unexplored, presenting a unique research opportunity. Consequently, this is the sensor we use to capture mmWave face signatures during our data collection procedure. A key driving factor for this choice is the Soli's miniature form factor of just 6.5 mm $\times$ 5.0 mm, and its use of Frequency Modulated Continuous Wave (FMCW) technology. This is proven to offer superior range resolution in comparison to other modulation techniques thanks to its high pulse compression [20]. Furthermore, the Soli chip has a relatively low power consumption. This is due to the fact that it sends 16 chirps every burst at a pulse-repetition frequency of 2 kHz, after which it halts transmission until the next burst cycle [21, 22]. Each burst is transmitted at a rate of 25 Hz, giving an overall transmission duty cycle of 2%. This effectively means that the radar chip remains inactive during the majority of its operation, saving a lot of power for small-scale mobile applications.

2.2 Face Recognition using mmWaves

The use of millimetre waves for face identification is a relatively new avenue of research, fuelled by the recent commercialisation of radar sensor technology. One of the earliest studies delving into human identification using mmWaves dates back to 2019, conducted by Zhao et al. [23]. While this paper primarily examines the classification of subjects based on their gait and body shape, it underscores the capacity of mmWaves to capture the subtle idiosyncrasies among individuals. These nuanced differences are crucial for machine learning models to accurately distinguish between unique subjects, thereby yielding high class separations.

The following year, Hof et al. [10] introduced an autoencoder capable of recognising human faces captured by an 802.11ad/y networking chipset operating at a 60 GHz centre frequency (f_c). This autoencoder effectively encodes mmWave face signatures with sufficient distinction to discriminate between positive and negative instances based on their Mean Squared Error (MSE) against reference embeddings. The study involved an extensive data collection effort, capturing face scans of 206 participants of varying genders and ages, across five different poses: frontal, as well as head rotations of 15° and 25° to the left and right. This dataset was subsequently made available through an IEEE Data Port [24]. While this collection encompasses a wide range of faces, including some individuals with spectacles and beards, it lacks representation of other common occlusion scenarios, such as head accessories, which our project aims to explore. Additionally, the study utilised a larger sensor with a total of 1024 transmit-receive antenna pairs, noted to capture redundant information. This contrasts with the compact Soli chip designed for integration within smartphones. The study simulated the impact of reducing the antenna count to 10, resulting in a significant decrease in the distinctiveness of facial signatures. Encouragingly, increasing the number of

[1] `https://github.com/StergiousAji/mmFace-3D-Face-Recognition-using-RGB-and-mmWave-Radar`

neurons in their network and adding an extra hidden layer could compensate for this loss.

Lim et al. [11] proposes a deep neural network with a more traditional Multi-Layer Perceptron (MLP) approach where every layer is fully connected to adjacent ones. The study utilised a small-scale, 61 GHz FMCW radar sensor developed by bitsensing Inc. [25], comparable to the Soli with a single transmit and three receiver antennas. The model attained a mean classification accuracy of 92% across eight subjects, surpassing the performance of both, a Support Vector Machine (SVM), and a tree-based Ensemble Learning approach trained on the same face signatures. It is important to note the relatively small-sized dataset used to train the model, raising concerns about potential overfitting as the data is not representative enough. The paper provides limited details on the data collection methodology used, only mentioning that the distances ranged from 30 cm to 50 cm. It can be assumed then that the study likely focussed on frontal poses without any occlusions for all eight participants.

During the same time frame, Kim et al. [12] conducted research utilising an identical sensor from bitsensing Inc., which boasted a range resolution of 2.5 cm. Their study introduces a Convolutional Neural Network (CNN) model consisting of three convolutional layers and three fully connected layers. The radar data underwent extensive preprocessing to convert it into a format more akin to images, suitable for the CNN. With a data split of 70%/15%/15% for training, validation, and testing, the model achieved an average classification accuracy of 98.7% on a limited dataset of only three individuals. Notably, the study also investigated the impact of wearing cotton masks. The results indicated a negligible decrease in average classification accuracy by 0.9%, which bodes well for the goals of our project. Nonetheless, it is important to approach these findings with caution due to the small dataset size. It remains uncertain whether this level of performance would hold consistently across a larger group of subjects with more varied occlusions.

Pho et al. [13] adopts a One-Shot Learning approach to the problem. This is where the model is trained with a single or only a few labelled instances, beneficial when there is a lack of training samples available. The proposed method constitutes a Siamese structure of two identical CNNs with shared parameters, mapping the input radar signals into latent space. During both the training and testing phases, a distance metric is used to evaluate the similarity between the outputs of the networks and the face inputs. Specifically trained for binary classification, the model receives pairs of face signatures from either the same or different individuals. The same bitsensing Inc. BTS60 chipset, used by Lim et al. and Kim et al. [11, 12], is employed to capture 500 frames of the faces of eight participants. An average classification of 97.6% was achieved, an improvement over the previous deep MLP model by Lim et al. involving the same number of people. t-Stochastic Neighbour Embedding (t-SNE) [26] is then applied for dimensionality reduction. The resulting visualisations demonstrate that the one-shot Siamese network effectively separates each individual's face into exclusive regions, simplifying the classification task. Although a small dataset is used, only encompassing frontal poses with no occlusion settings, the proposed method is well documented and is likely robust against larger datasets.

Challa et al. [14] employs two different machine learning models on the dataset provided via the IEEE port [24]. Their approach began with CNN-based autoencoders, followed by a Random Forest Ensemble Learning approach. A total of nine autoencoders were built, each tailored to different frame rates, focusing on compressing and reconstructing the original data from its latent form. The autoencoders were trained using randomly selected data samples from a subset of 186 face scans. The flattened and labelled outputs were then used to train and test nine discrete Random Forest models using identical hyperparameters, as recommended by the Sci-kit library. This methodology yields impressive results, achieving an average classification accuracy of 99.98% using all 1400 frames per individual. Even when reducing the number of frames to 70 per person, the model maintained a high accuracy of 97.1%. The paper presents an approach that is unique in comparison to the rest of the papers tackling this subject, showcasing an efficient model that is able to be deployed on mobile chips.

Research in this domain focuses exclusively on utilising data from radar sensors, largely driven by concerns surrounding privacy preservation. However, a significant drawback of this approach lies in the extended duration required to capture an accurate facial scan. The sensor typically needs to operate for several seconds, ranging between 10 and 15, to obtain a detailed scan. Such a time frame proves impractical in real-world scenarios, as it requires the subject to remain motionless for a prolonged period. Thus far, no study has explored the potential benefits of combining radar signatures with corresponding RGB data to enhance facial recognition capabilities. Given the high performance of existing deep learning models using RGB images alone, such as InsightFace [27], integrating them with mmWave radar data presents a promising avenue. This could expedite face acquisition time while capitalising on the advantages of mmWaves for environments where optical methods falter.

2.3 InsightFace

In the evolving field of face recognition, deep CNNs have emerged as a dominant approach due to their ability to automatically extract discriminative facial features from images. One significant advancement in this area is the InsightFace toolkit, implementing algorithms designed to address the intricacies of face analysis and recognition. Key works include the preliminary ArcFace model, introduced by Deng et al. [27], alongside the robust Face Alignment model by Gho et al. [28]. ArcFace employs a novel Additive Angular Margin Loss to maximise class separability, further enhancing the discriminative power in mapping face images to feature embeddings. However, this method was found to face challenges with label noise, requiring the "cleaning" of many real-world images sourced from the web. To address this, further progress was made with the Sub-center ArcFace model [29], introducing the concept of sub-classes to boost resilience against intra-class variations and label noise. It achieved state-of-the-art performance on many widely used benchmark datasets such as the Labeled Faces in the Wild (LFW) [30] and the YouTube Faces (YTF) datasets [31]. The integration of pretrained models offered by InsightFace into our system enables us to concentrate efforts on enhancing our model's ability to extract 3D structural information from mmWave face signatures.

2.4 Multimodal Data Fusion Methods

Multimodality, as defined by Lahat et al. [32], refers to the use and analysis of multiple types of data, potentially arriving from different sensors. The aim is to extract and blend salient information gathered by each sensor. The integration of this diverse data leads to outputs with richer representations than what could be achieved by the individual modalities alone.

One common strategy involves merging multiple data modalities before feeding them into a learning model, known as **Early Fusion** or **Data-level Fusion**. This technique entails combining data by eliminating correlations between sensors or fusing data in a common, lower-dimensional space [33]. Methods like Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) are frequently utilised for this purpose. However, a significant issue with early fusion is ensuring synchronisation between the RGB and radar frames, which is challenging due to their notably different sampling rates. Moreover, the continuous mmWave signals must be discretised to align with the format of the RGB data. An inherent drawback of early fusion is the potential to squash crucial information present within each individual modality, thereby impacting training efficacy.

Late Fusion, or **Decision-level Fusion**, operates by independently processing distinct data sources through separate models and then integrating them at the decision-making stage. A common approach involves calculating a weighted average of the separate predictions, allowing a way to regulate the influence of specific modalities [34]. Late fusion is often simpler and more adaptable, proving effective when dealing with highly dissimilar data sources in terms of sampling rate, dimensionality, or units of measurement. Furthermore, late fusion often yields better performance since errors from multiple models are managed independently.

Intermediate Fusion, or **Feature-level Fusion**, is rooted in neural network architectures and revolves around the concept of combining different modalities within the feature space where there is a higher level of abstraction of the raw data. This can range from a basic concatenation of the individual latent embeddings to employing autoencoders for non-linear feature fusion, as demonstrated by Charte et al. [35]. This approach offers greater versatility than early and late fusions since it allows for the integration of features at various depths within the neural network. However, it can pose challenges such as the risk of overfitting or difficulty in learning relationships between the different modalities.

Each data fusion technique presents its own set of considerations, necessitating experimentation to determine the most effective approach to merging RGB and mmWave signatures. A variant of late, feature-level fusion, where the embeddings from the final layers of each model are combined, was chosen as the most feasible.

3. METHODOLOGY

3.1 Data Acquisition

Following a thorough research of the field, the subsequent steps involved designing and conducting the data acquisition process required to train our proposed model. These experiments necessitated meticulous planning as the collected data directly determines the efficacy of the final model. As demonstrated by previous studies, it is crucial to compile multiple poses to enable the model to learn a complete 3D scan of the individual's face. Moreover, incorporating pose-invariance into the system is essential to accommodate real-world scenarios where individuals may not always present an exact frontal pose to the facial recognition system. Most studies focus on azimuth variations since individuals are less likely to tilt or pitch their heads by a significant amount. We similarly concentrated on head rotations around the yaw axis, deciding to capture facial poses at 0°, 30°, and 45° azimuth relative to the sensors.

Given that the experiment's objective is to explore the advantages of mmWave sensors for face recognition, we included two distinct lighting conditions in our data collection trials: standard and low-light environments. Our hypothesis is that mmWave face signatures remain unaffected by ambient lighting since the sensor employs its own active illumination of the target face, unlike the RGB camera. Hence, if the system can achieve higher accuracy by incorporating both modalities rather than relying solely on colour, it would strongly indicate that mmWaves provide robustness against diverse lighting conditions.

Finally, we delve into assessing the permeating capability of mmWaves to directly reach the skin through fabric and hair by injecting typical occlusion scenarios into our experiments. It is advantageous for facial recognition systems to inherently withstand common obstructions such as glasses, hats, masks, and so on. Presently, users often need to remove such accessories for systems to accurately identify and grant access to specific devices or areas. With mmWaves, we speculate that this may not be required as facial features could be captured regardless. This could greatly benefit security surveillance, especially in situations where individuals deliberately obscure their faces to conceal their identities. In our experiment, we capture scenarios both with and without occlusion. While cotton masks have been previously explored by Kim et al. [12], we incorporate other typical items such as sunglasses, hats, and scarves to mirror day-to-day use cases.

To ensure a diverse range of facial data, we recruited 21 participants within the limited time frame of the project. Adhering to ethical standards regarding sensitive personal information, our participant pool consists of male and female university students and faculty in the age range of 18–35 years. A total of 15 scenarios were captured for each participant at a distance of 20 cm from the sensors. Each time the sensors are run for 10 seconds with the participant being asked to maintain a neutral expression, totalling 150 RGB frames and 3,750 mmWave frames per person. On

Figure 1: *Equipment: Intel Realsense L515 RGB-D Camera (left) and Google's Soli* 60 GHz *radar sensor (right)*

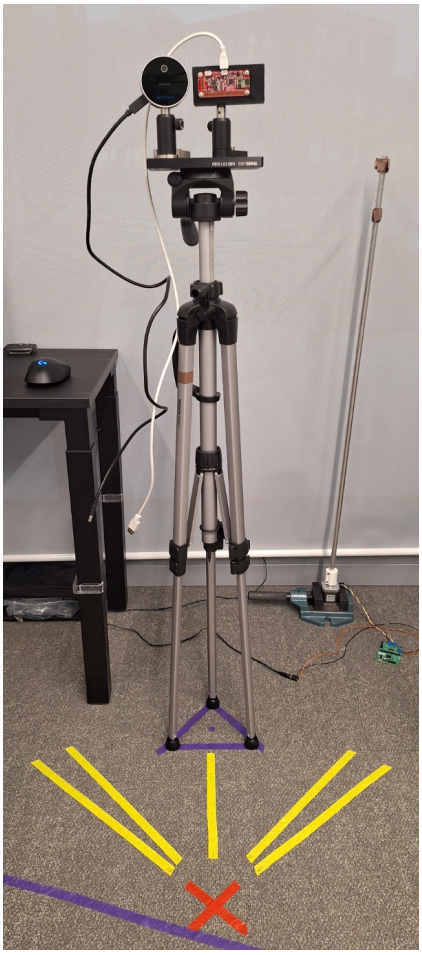


Figure 2: *Experiment setup used during data acquisition with the equipment mounted on a tripod. The red cross marks the* 20 cm *face distance and the yellow tape indicates the five pose directions.*

top of this, scans of printed faces are collected in order to train the model to detect the authenticity of the subject's face using the 3D information. This was restricted to the three frontal poses for each participant, mirroring common spoofing tactics, providing another 30 RGB frames and 225 mmWave frames per fake instance.

A close-up of the equipment setup used can be observed in Figure 1 showing the Intel Realsense RGB-D camera and the green Soli chip mounted side-by-side on a breadboard. The full experiment setup is photographed in Figure 2 with the red cross indicating the 20 cm face distance where subjects are positioned, and the five pose directions marked by the yellow tape.

To illustrate the results of the collection process, the left half of Figure 3 presents data samples from a single subject. This grid shows RGB captures from all 15 scenarios, with the three different conditions along the rows and the five pose variations along the columns. For brevity, the experiment conditions are abbreviated as outlined in Table 1.

Abbreviation	Expanded Form
NO	No Occlusion
O	Occlusion
RLC	Regular Lighting Condition
DLC	Dim Lighting Condition

Table 1: *Table displaying the full forms of abbreviations describing the experimental conditions*

3.2 Data Preprocessing

The radar bursts acquired during the data collection phase undergo multiple FFT stages of preprocessing to convert the raw signals into discretised Complex Range-Doppler (CRD) maps. These maps offer a two-dimensional representation of the reflected radar signal, where the range dimension corresponds to the distance from the Soli sensor, and the Doppler dimension corresponds to the radial velocity of the subject towards the sensor [16, 21]. Face scans are obtained using the Soli's short configuration, operating at an f_c of 60 GHz, with a maximum bandwidth B of 5.5 GHz. This configuration provides a range resolution Δr of:

$$\Delta r = \frac{c}{2B} = 2.7\,\text{cm}$$

where c denotes the speed of light. The Soli chip comprises a single transmit and three receiver antennas, each capturing a superposition of scattered reflections from the target. Given that the Intel RealSense captures RGB-D frames at a different sampling rate of 30 frames per second (FPS), timestamp information is also logged for the potential of synchronising the two modalities for early data fusion. The right half of Figure 3 depicts a plot of a single CRD frame across the three receiving channels for a subject's face. The plot illustrates the discretised intensities of received signals across 16 Doppler bins along the x-axis and 32 range bins along the y-axis.

In order to simplify computation, the magnitudes of each complex value encoding the range r and Doppler d amplitudes, are derived to generate an Absolute Range-Doppler (ARD) map as follows:

$$\texttt{ARD}_{r,d} = \text{abs}(\texttt{CRD}_{r,d})$$

Finally, data augmentations are applied to both the ARD and RGB frames, restricted to horizontal and vertical flips due to the distinct nature of both modalities. Augmentations were carefully selected to be semantically consistent across the layout of the mmWave face signatures and colour images. Rotational augmentations, for instance, cannot be seamlessly translated into the range and Doppler bins of the `ARD` as is the case with traditional images. Data augmentations aimed to inflate the small dataset size as well as induce positional equivariance for the model to learn facial features consistently across different instances.

3.3 mmFace

Building on the intuition from section 2.3 of the Related Work chapter, it is clear that the ArcFace model from the InsightFace toolkit emerges as the best choice for our project. It attains state-of-the-art classification results on accepted benchmark sets, outperforming the previous bests such as Facebook's DeepFace [36] and Google's FaceNet [37]. This selection allows us to treat the RGB data processing as a ***black-box*** framework, enabling us to concentrate efforts on perfecting the radar-based feature extraction model we are naming, `mmFace`. Furthermore, this facilitates exploration into the various methods in fusing the two modalities.

Figure 4 depicts a high-level diagram of the system workflow employed during training and inference. A more detailed architecture of our end-to-end `mmFace` model can be viewed in Figure 5 providing a comprehensive breakdown of each layer as well as their input and output channels. In summary, `mmFace` takes two inputs: an mmWave face

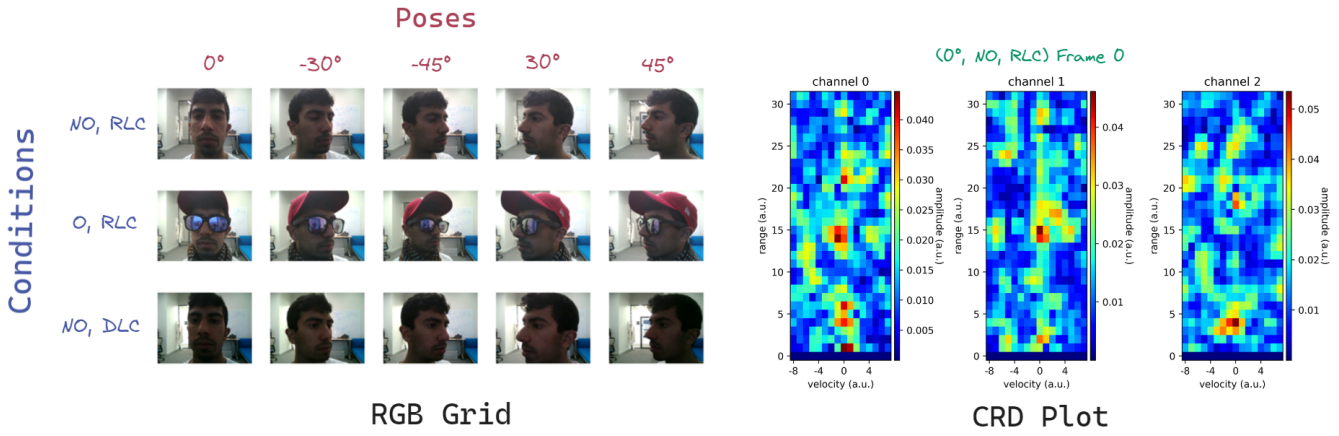


Figure 3: *Data samples collected for Subject 0. The left figure shows the RGB frames of all 15 scenarios organised by pose and condition. The right figure plots a single CRD frame showing amplitudes of reflected waves detected by the three receiving channels of the Soli, categorised into discrete Range-Doppler bins.*

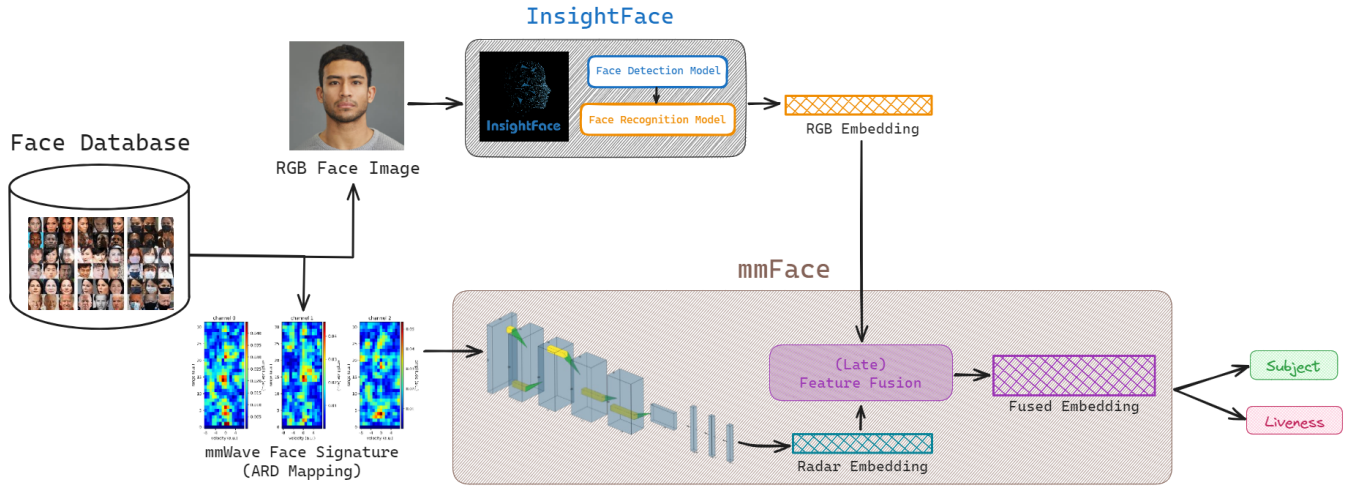


Figure 4: *High-level model workflow diagram of our proposed 3D face recognition system incorporating millimetre wave radar and RGB images.*

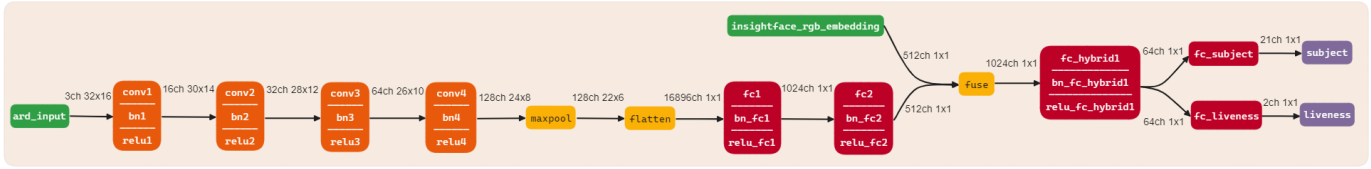


Figure 5: *Architecture diagram of our* `mmFace` *model displaying each layer, as well as their input and output channels.*

signature in an `ARD` format and an InightFace embedding (`rgb_emb`) extracted from the corresponding RGB frame. The model then uses both modes of information to ultimately yield a subject and liveness prediction. The liveness detection is a simple binary classification: 0 denoting a fake subject or 1 for real. The inputs undergo three main stages: **mmWave Feature Extraction**, **Feature Fusion**, and **Class Prediction**, each described in detail below.

3.3.1 mmWave Feature Extraction

Firstly, the `ARD` input is processed through four unstrided convolutions followed by a max-pooling then two fully connected layers to compress the radar embedding vector. The three-channel `ARD` format of the mmWave face signatures allows leveraging convolutional-based feature extraction due to its image-like structure. The convolutional layers are able to detect spatial patterns among the range and Doppler profiles specific to individual faces, with potential enhancement through deeper layers. Due to the relatively small size of the `ARD` maps, only being 32×16 bins per channel, it was imperative to preserve most of the information, minimising the need for additional max-poolings or strided convolutions. To streamline the fusion stage, we decided to match the final radar embedding size with the 512-dimensional InsightFace feature vector. Four convolution layers, each using a 3×3 kernel size and filter sizes sequentially increasing in powers of two starting from 16 to 128, were found to be sufficient, with additional layers affording diminishing returns. This stage is summarised as follows:

$$\text{feature\_extract}(\texttt{ARD}) = \texttt{radar\_emb}$$

3.3.2 Feature Fusion

The next phase involves fusing the extracted radar embedding with the RGB embedding input. This is modular in design in order to afford any compatible fusion strategy to be employed. This is then processed through a single fully connected layer to reduce its dimensionality before advancing to the final stage. This stage is summarised below:

$$\text{fuse}(\texttt{radar\_emb},\ \texttt{rgb\_emb}) = \texttt{fused\_emb}$$

We opted to focus on mixing the two modalities within the feature space, specifically within the network's final layers. This facilitates an easier fusion process since the data from both modalities are abstracted into a compressed representation. Pure intermediate fusion was not feasible due to the black-box treatment of the InsightFace model making it difficult to integrate information from within its hidden layers. Early fusion presents hurdles as well due to the dissimilarities in sampling rates and data formats. This is an interesting avenue left to be explored possibly training a neural network to transform the radar bursts into a pixel-wise 3D point cloud.

3.3.3 Class Prediction

Finally, the fused multimodal embedding vector is carried across two separate fully connected layers to predict the identity of the face and its authenticity which are subsequently outputted by the model. This is formalised below with $\hat{s}$ and $\hat{l}$ signifying the model's subject and liveness predictions respectively:

$$\text{classify}(\texttt{fused\_emb}) = (\hat{s}, \hat{l})$$

This classifier served as the primary approach to train the model, employing a Stochastic Gradient Descent (SGD) optimiser that was tasked with minimising the cross entropy loss (L_{CE}) of each prediction.

As our model generates two predictions, we merge the losses from each with equal weighting to ensure uniform learning of both attributes during backpropagation. This combined loss $\mathcal{L}$ is formulated as follows:

$$\mathcal{L} = L_{\text{CE}}(s, \hat{s}) + L_{\text{CE}}(l, \hat{l})$$

where s and l represent the true subject and liveness labels.

3.3.4 Training

We created and trained our models using PyTorch version 2.1 [38], running for 20 to 25 epochs. We opted for a fixed learning rate of 0.01, an L2 regularisation rate of $1e$–3, and a momentum of 0.9. Furthermore, we trained our models on a random subset of 17 out of the total 21 subjects, which accounts for just over an 80% training split. Note that this includes the 17 fake counterparts, giving 34 total subject instances. The remaining four subjects (or eight instances) are left out for testing.

All linear layers within the feedforward network are followed by batch normalisation to reduce overfitting and increase the generalisability of the model. ReLU activations are chosen for all non-linear transformations to prevent vanishing weights. Ultimately, the final `mmFace` model contains around 2.8 million parameters and takes, on average, 4.1 milliseconds (SD = 0.2 ms) to process a single (`ARD`, `rgb_emb`) input pair on an NVIDIA GeForce GTX 1650 GPU.

3.4 Feature-level Fusion Strategies

We investigate seven feature-level fusion strategies listed as follows in terms of two n-dimensional input feature vectors, $\vec{\boldsymbol{x}} = [x_1, x_2, \ldots, x_n]$ and $\vec{\boldsymbol{y}} = [y_1, y_2, \ldots, y_n]$:

i. **Concatenate:** This is a straightforward concatenation of the two feature vectors, the most common strategy employed for its ease of implementation and effectiveness as all information is preserved.

$$\texttt{concatenate}(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) = [\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}]$$

ii. **Add:** This involves an element-wise vector addition of $\vec{\boldsymbol{x}}$ and $\vec{\boldsymbol{y}}$. This can be highly effective when both feature vectors point in the same direction compounding their resulting summation. However, if both feature vectors point in opposite directions, then this would result in a more orthogonal resulting vector direction.

$$\texttt{add}(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) = [(x_i + y_i) \mid \forall i \in \{1, \ldots, n\}]$$

iii. **Hadamard Product:** This is the element-wise vector multiplication or the Hadamard product of $\vec{\boldsymbol{x}}$ and $\vec{\boldsymbol{y}}$. The idea here is to preserve the original feature vector structure while emphasising relationships between corresponding feature elements. It has been used successfully for multimodal residual learning by Kim et al. [39].

$$\texttt{hadamard\_product}(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) = [(x_i y_i) \mid \forall i \in \{1, \ldots, n\}]$$

iv. **Pairwise Dot Mean:** This involves a dot product of the transpose of vector $\vec{\boldsymbol{x}}$ with vector $\vec{\boldsymbol{y}}$, resulting in an $n \times n$ matrix followed by a column-wise mean operation to produce an n-dimensional fused feature vector. The rationale behind this stems from the pairwise dot providing a more comprehensive blending of the features. Each radar feature is multiplied by every RGB feature, and the resulting values are all attended to during the pooling process to reduce the dimensionality of the matrix.

$$\textit{Let } \boldsymbol{Z} = \vec{\boldsymbol{x}}^T \cdot \vec{\boldsymbol{y}} \textit{ in}$$
$$\texttt{pairwise\_dot\_mean}(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) = \left[\frac{1}{n} \sum_{j=1}^{n} \boldsymbol{Z}_{j,i} \;\middle|\; \forall i \in \{1, \ldots, n\} \right]$$

v. **Pairwise Dot Max:** This similarly involves the dot product followed by a column-wise max. In contrast to averaging, which could squash certain feature correlations, this approach isolates larger features resulting from the exhaustive mixing of both modalities, a similar intuition behind the max-pooling layers. Here, $*$ denotes the selection of all rows of a matrix.

$$\textit{Let } \boldsymbol{Z} = \vec{\boldsymbol{x}}^T \cdot \vec{\boldsymbol{y}} \textit{ in}$$
$$\texttt{pairwise\_dot\_max}(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) = [\max(\boldsymbol{Z}_{*,i}) \mid \forall i \in \{1, \ldots, n\}]$$

vi. **Pairwise Dot Flatten:** This is the final pairwise dot strategy now following the dot product with a flatten operation of the $n \times n$ matrix into an n^2-dimensional vector. This conversely retains all correlations between the radar and RGB features, leaving the subsequent fully connected layers of the model to determine which features are most relevant.

$$\textit{Let } \boldsymbol{Z} = \vec{\boldsymbol{x}}^T \cdot \vec{\boldsymbol{y}} \textit{ in}$$
$$\texttt{pairwise\_dot\_flatten}(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) = [\boldsymbol{Z}_{1,1}, \boldsymbol{Z}_{1,2}, \ldots, \boldsymbol{Z}_{n,n-1}, \boldsymbol{Z}_{n,n}]$$

vii. **Multi-Head Attention:** This involves using a self-attention mechanism prevalent in transformer architectures popularised by the seminal paper by Vaswani et al. [40]. It was initially designed for natural language processing tasks but has shown a lot of success in computer vision. The key idea of using self-attention is to isolate and mix the most important aspects from both feature vectors by converting them into three distinct representations: a query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$. Each embedding plays a unique role such as the query capturing features that the model deems relevant for making predictions. To do this $\boldsymbol{Q}$ may focus on features that are shared or discriminative across both modalities such as facial landmarks and overall identity information. Meanwhile, $\boldsymbol{K}$ might focus on modality-specific elements such as the colour and texture information embedded within the RGB feature vector, while 3D structural details being offered by the radar embedding. Finally, $\boldsymbol{V}$ encapsulates the actual fine-grained details captured by both modalities that will be attended to by the model based on the resulting query-key similarities. In order to maximise this effect, this mechanism is applied separately across multiple attention heads, reducing the risk of overlooking salient characteristics. This is expressed formally below, where first, the two inputs are stacked vertically, $\boldsymbol{X} = \begin{bmatrix}\vec{\boldsymbol{x}}\\ \vec{\boldsymbol{y}}\end{bmatrix}$, then copied through three separate linear transformations to obtain the $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ matrices. This is done for each of the k attention heads using separate learnable weight matrices.

$$\textit{Let } h_i = \text{Attention}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) \textit{ in}$$
$$\texttt{multihead\_attention}(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) = [h_1, \ldots, h_k]\boldsymbol{W}^O$$
$$\textit{where } \boldsymbol{Q}_i = \boldsymbol{W}_i^Q \boldsymbol{X},\ \boldsymbol{K}_i = \boldsymbol{W}_i^K \boldsymbol{X} \textit{ and } \boldsymbol{V}_i = \boldsymbol{W}_i^V \boldsymbol{X}$$

4. EVALUATION

We evaluate our models through a zero-shot classification task in order to assess their discriminative ability at representing the four unseen subjects. This also examines the extent at which the model can generalise to new faces, a task akin to real-world face recognition scenarios. Various metrics, including prediction accuracies, precision, recall, and ROC AUC are used to analyse the different fusion strategies. These metrics help to benchmark the feature fusion methods against solely using the individual modalities. Prior to the zero-shot task, we extract the features from the final hidden layer, `fc_hybrid1`, of eight pre-selected reference input pairs. These reference embeddings serve as the basis for comparison against the rest of the test set, encompassing both live and fake instances of the four unseen subjects. The reference frames were picked out of the frontal pose, non-occluding, regular lighting category (0°, NO, RLC) as this represents the most frequently encountered setting. During inference, the classifier component of the model is discarded such that the final latent vectors are used for comparison.

4.1 General Results

Firstly, we gauge the accuracies of the models in predicting both facial identity and authenticity. This involves determining the most similar reference embedding to each test sample using the maximal cosine similarity score. A decision threshold t of 0.5 is used such that the maximal score must be greater than it to qualify as a valid prediction, otherwise, it is marked as a failed prediction. This ensures that classifications are not made for output embeddings that are too far or equidistant to all reference embeddings such that no practical decision can be made.

All models demonstrate high test coverage, as illustrated in Table 3, with the radar-only model achieving the lowest coverage at 79.8%. Secondly, since the eight selected reference instances exhibit two properties – facial identity and liveness status – this allows for the calculation of separate metrics to assess each model's ability to predict identity independently of the liveness check.

4.1.1 Zero-shot Accuracies and F-measures

Table 2 presents the subject and liveness accuracies for all feature fusion strategies as well as the weighted F_β-measures, averaged over all the respective classes. For completeness, the performance of the non-hybrid models using the individual modalities are also listed for a baseline comparison. A β value of 0.5 was chosen for the F-scores to prioritise precision twice as highly over recall. Admitting incorrect identities is a lot more detrimental than missing true matches within secure face recognition systems.

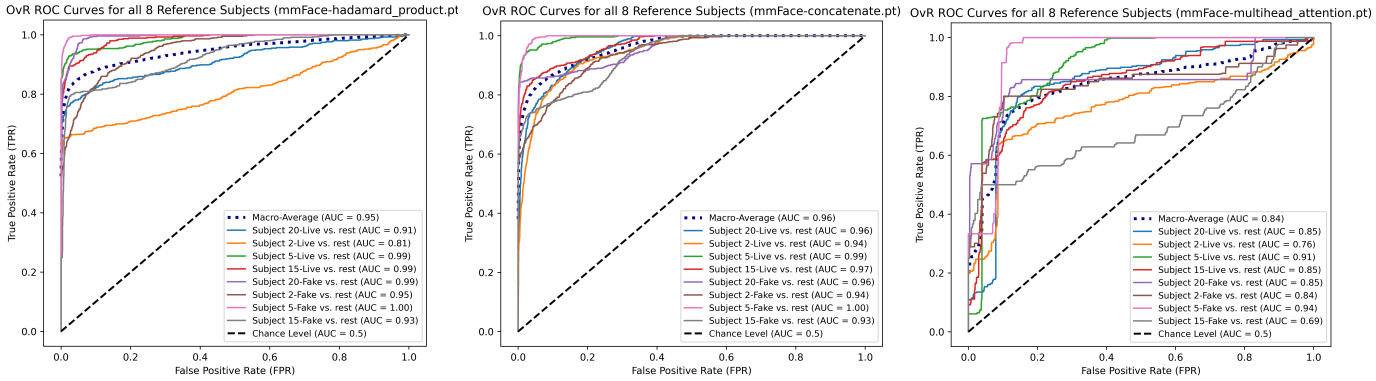
Fusion Strategy	Subject		Liveness	
	Accuracy (%)	$F_{0.5}$ Score	Accuracy (%)	$F_{0.5}$ Score
Concatenate	83.7	0.835	99.6	0.996
Add	63.0	0.629	99.2	0.992
Hadamard Product	87.1	0.869	96.7	0.963
Pairwise Dot Mean	88.8	0.880	80.8	0.808
Pairwise Dot Max	82.7	0.820	72.8	0.735
Pairwise Dot Flatten	86.7	0.862	94.7	0.944
Multi-Head Attention	86.3	0.851	96.4	0.950
Radar Only	38.2	0.370	96.6	0.916
RGB Only	85.5	0.855	69.3	0.701

Table 2: *Subject and liveness accuracies and weighted-averaged $F_{0.5}$ measures for the seven feature fusion strategies along with the individual modalities.*

The best and worst performers are highlighted in green and red respectively. Evidently, certain fusion strategies excel at discriminating subjects, while others are more adept at discerning face liveness. Among these strategies, the concatenation method emerges as the most effective for accurately verifying liveness. Meanwhile, the pairwise dot-then-mean strategy outperforms within the subject category, attaining the highest accuracy and $F_{0.5}$ measure. However, its ability to perceive face liveness is notably lacking, obtaining an accuracy that is even lower than simply using the mmWave radar features. This is likely due to the average-pooling step squashing outlier feature correlations between the two modalities that may be relevant for liveness detection. Lastly, the vector addition proves to be very ineffective in generating identity-specific embeddings, being much better at predicting liveness in comparison.

Examining the performance at predicting each category independently provides insight into each strategy and their effectiveness at classifying unseen data to their respective labels. However, it is just as important to determine the overall best strategy that performs equally well in both categories.

Table 3 showcases each of the nine models along with their average accuracy and $F_{0.5}$ measure, applying equal weighting to the subject and liveness results. It is apparent that



(a) *Hadamard Product* **(b)** *Concatenate* **(c)** *Multi-Head Attention*

Figure 6: *One-vs-the-Rest ROC curves plotted for each of the eight reference classes. Their respective AUC metrics are shown as well as the chance level and macro-averaged ROC curve. The sub-figures display curves for the top three ranking models.*

Fusion Strategy	Mean Accuracy (%)	Mean $F_{0.5}$ Score	Coverage (%)
Concatenate	91.7	0.915	99.7
Add	81.1	0.811	99.9
Hadamard Product	91.9	0.916	98.1
Pairwise Dot Mean	84.8	0.844	95.3
Pairwise Dot Max	77.8	0.778	95.6
Pairwise Dot Flatten	90.7	0.903	97.8
Multi-Head Attention	91.3	0.900	92.5
Radar Only	67.4	0.643	79.8
RGB Only	77.4	0.778	97.5

Table 3: *Averaged accuracy and $F_{0.5}$ score for the seven fusion strategies and non-hybrid models, applying equal weighting to subject and liveness predictions.*

the Hadamard product of the two feature vectors achieved the highest mean accuracy and F-measure. Furthermore, the table illustrates the fact that all fusion strategies successfully improve on the results of the non-hybrid models, which only utilise one of the modalities. On average, the fusion strategies exhibit a 9.6% improvement in mean accuracy, relative to the RGB-only model which obtained the highest accuracy out of the two non-hybrid models.

4.1.2 ROC Curves and AUC

Despite the positive aggregated results, it is worth noting that the element-wise product method is closely followed by the concatenation and multi-head attention strategies. Hence, it can be more useful to assess the model's performance at different decision thresholds to provide deeper insights into which strategy yields the better classifier. This is the main idea behind the Receiver Operating Characteristic (ROC) Curve which analyses the sensitivity or true positive rate (TPR) against the false positive rate (FPR) at varying decision thresholds. Given the multi-classification nature of our task, the One-vs-the-Rest (OvR) ROC curve is the most appropriate. Here each class is evaluated against the rest using a one-hot strategy. This, therefore, requires each of the eight reference classes to be plotted separately.

Figures 6a, 6b, and 6c plot the one-vs-the-rest ROC curves for the three highest overall performing strategies – the Hadamard product, concatenation, and multi-head attention, respectively. The blue dotted curves indicate the macro-averaged trend among the individual OvR curves, while the dashed diagonal line denotes the chance level equivalent to a randomly guessing model. Curves above the chance level and reaching the top-left corner demonstrate a better classifier. This is often quantified by the area under the ROC curve (AUC) metric.

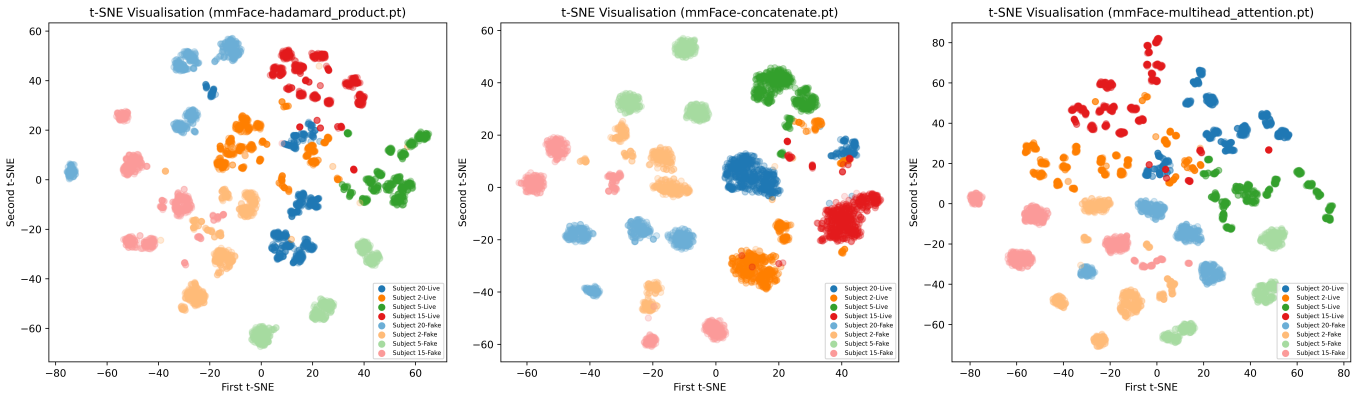
Table 4 lists the macro-averaged AUC metrics for all nine models. As made evident from both the table and plots, while the Hadamard product offers the highest accuracies, it does not maintain its predictive nature across all decision thresholds, attaining an AUC of 0.945. While this is reasonably high, it is overshadowed by the 0.961 AUC achieved by the concatenation strategy which ranked second-highest within the overall performance. Figure 6b shows that the concatenation method performs equally well at distinguishing the majority of the reference instances since all curves are tightly packed and in closer proximity to the top-left corner. In contrast, the ROC curves for the Hadamard product method are more dispersed, with an especially low distinguishing ability for the class *Subject 2-Live*. The disparate ROC curves for the multi-head attention strategy in Figure 6c suggest that it does not generate embeddings as robust as those of the other two strategies. It yields a proportionally higher number of false alarms compared to true positives for certain outlier classes.

Fusion Strategy	Macro-Averaged AUC
Concatenate	0.961
Add	0.918
Hadamard Product	0.945
Pairwise Dot Mean	0.914
Pairwise Dot Max	0.863
Pairwise Dot Flatten	0.925
Multi-Head Attention	0.913
Radar Only	0.735
RGB Only	0.901

Table 4: *Table listing the macro-averaged ROC AUC metrics for all fusion methods as well as the individual modalities.*

4.1.3 t-SNE Visualisations

Following this, it is important to visualise the final embedding vectors of each model to verify that each identity is pushed to distinct regions of the high-dimensional Euclidean space. t-distributed Stochastic Neighbour Embedding or t-SNE [26] is commonly employed to reduce the dimensionality of feature vectors to a lower space. Compared to Princi-



(a) *Hadamard Product* **(b)** *Concatenate* **(c)** *Multi-Head Attention*

Figure 7: *2D t-SNE projections of the final layer features of all test samples. Features are extracted from the* `fc_hybrid1` *layer from each of the top three ranking models.*

pal Component Analysis (PCA), *t*-SNE tends to preserve the local structure of the data, whereas PCA is better suited for retaining the global variance. This means that t-SNE can help identify meaningful patterns or groupings within the data since the similarities between data points are maintained providing interpretable visualisations.

Figures 7a, 7b, and 7c depict the final features of all test instances, extracted from the top three overall ranking strategies after t-SNE. Each data point is colour-coded by the subject identity, with those originating from fake instances shown in a lighter shade. It is clear that all three strategies group facial data from the same subject into similar regions of the multimodal space. It can be observed that fusion by concatenation results in a more singular, tighter clustering with greater separation between the distinct grouped islands of data points. Additionally, it can be seen that some data points overlap with non-similar clusters in each of the strategies, with the least confusion being observed within the Hadamard product visualisation. This explains its higher subject recognition accuracy in comparison to concatenation and multi-head attention.

Nonetheless, the concatenation strategy produces a clearer linear separation between fake and real faces, providing a rationale for it achieving the highest liveness detection rate of 99.6%. The other two strategies do not produce such a clear separation requiring more complex, non-linear boundaries between their representations of live and fake samples. It is evident that there is more noise among the features from the real dataset compared to the fake, producing more erratic clusterings and confusion within all three plots. This discrepancy can be attributed to the greater diversity of poses in the real dataset, as opposed to only frontal poses being included in the fake dataset.

Thus far, the models have been assessed from a general perspective, treating all test samples equally. However, it is just as important to examine the performance of the models under specific experimental conditions in isolation, such as dim lighting and occlusion settings.

4.2 Regular Lighting Condition

Firstly, we focus on the most frequently encountered case, faces under standard lighting with no facial obfuscation (NO, RLC). The same eight reference frames from the general experiment are used for this subset of the test collection. For brevity, Table 5 displays the mean accuracy and weighted F-measures for each of the nine models, assigning equal weight to subject and liveness predictions.

The results of the models under standard lighting mirror the main trends of the general findings, with the Hadamard product narrowly beating multi-head attention and concatenation in mean accuracy. Meanwhile, the pairwise dot max yields the lowest metrics. Notably, the concatenation strategy does yield a marginally higher $F_{0.5}$ measure by 0.02 compared to the element-wise product. The results from this setting serve as a baseline for comparison against the dim lighting and occlusion scenarios, which are subsequently analysed in isolation.

Fusion Strategy	Mean Accuracy (%)	Mean $F_{0.5}$ Score	Coverage (%)
Concatenate	91.7	0.918	99.6
Add	80.9	0.809	100.0
Hadamard Product	92.0	0.916	97.5
Pairwise Dot Mean	85.0	0.845	95.3
Pairwise Dot Max	77.6	0.778	95.3
Pairwise Dot Flatten	91.1	0.907	97.5
Multi-Head Attention	91.8	0.903	91.6
Radar Only	67.9	0.652	80.8
RGB Only	77.7	0.775	95.5

Table 5: *Averaged accuracy and $F_{0.5}$ score for the seven fusion strategies and individual modalities against the* ***non-occluding regular lighting*** *settings only. Equal weighting is applied to subject and liveness predictions.*

4.3 Dim Lighting Condition

Next, the dim lighting condition experiments are isolated and compared against the non-occluding captures under regular lighting (NO, RLC). The eight reference frames are now retaken from the (0°, NO, DLC) subset. Table 6 shows the aggregated metrics for each of the nine models, applying equal weight to subject and liveness measures. The findings suggest only a marginal decrease of 0.43% in subject accuracy, on average, in comparison to captures under regular light. Interestingly, the dim lighting setting produces a slight increase in liveness accuracy, averaging 0.2% across all fusion strategies. This is most likely due to the reduced sample size, amplifying the impact of outlier results.

This time the multi-head attention strategy emerges as the overall top performer, although it still exhibits a low coverage giving the pairwise dot-then-flatten method a slight

Fusion Strategy	Mean Accuracy (%)	Mean $F_{0.5}$ Score	Coverage (%)
Concatenate	90.4	0.903	99.9
Add	85.5	0.862	100.0
Hadamard Product	88.9	0.886	96.8
Pairwise Dot Mean	82.8	0.825	96.0
Pairwise Dot Max	79.8	0.792	93.2
Pairwise Dot Flatten	90.7	0.906	98.5
Multi-Head Attention	91.1	0.902	95.7
Radar Only	67.2	0.636	80.1
RGB Only	82.0	0.811	92.7

Table 6: *Averaged accuracy and $F_{0.5}$ score for the seven fusion strategies and individual modalities, isolated on the **dim lighting** settings. Equal weighting is applied to subject and liveness predictions.*

Fusion Strategy	Mean Accuracy (%)	Mean $F_{0.5}$ Score	Coverage (%)
Concatenate	91.4	0.911	99.7
Add	78.8	0.781	99.2
Hadamard Product	88.4	0.829	93.8
Pairwise Dot Mean	82.4	0.764	92.6
Pairwise Dot Max	77.6	0.710	91.5
Pairwise Dot Flatten	88.7	0.871	98.2
Multi-Head Attention	82.5	0.813	98.5
Radar Only	70.8	0.713	82.7
RGB Only	68.6	0.682	80.7

Table 7: *Averaged accuracy and $F_{0.5}$ score for the seven fusion strategies and individual modalities, isolated on the **occlusion** scenarios. Equal weighting is applied to subject and liveness predictions.*

edge. A more interesting observation arises with the RGB-only model, relying solely on the 2D InsightFace embedding. It achieves a relatively high $F_{0.5}$ measure of 0.821 for recognising subjects, contrary to our initial hypothesis. While it was anticipated that the RGB-only model would suffer under dim lighting, it still manages to outperform the radar-only model. Ignoring the small sample size, it is worth noting that the dim lighting environment was not controlled as rigorously such as with a measured lumen rating. Instead, it was achieved through the lowest light setting of our experiment room and closed shades. Additionally, variations in ambient lighting from the sun at different capture times were not accounted for. It is evident that a more extreme condition is necessary to fully assess the impact of ambient lighting on RGB captures and the robust InsightFace model.

4.4 Occlusion

Next, the captures incorporating day-to-day occlusion scenarios where participants were asked to wear hats, sunglasses and scarves were isolated. To minimise the risk of the model learning to segment specific colours or shapes, several types of hats were utilised including different coloured baseball caps and beanies. All sunglasses used included a mirrored lens such that the participants' eyes could not be seen by the RGB camera. Moreover, the participants were instructed to wear the accessories as they naturally would, providing more variations within the dataset. The eight reference frames are redrawn from the (0°, O, RLC) category and compared against the non-occluding captures under regular light (NO, RLC).

Table 7 presents the aggregated findings for all nine models, with a similar weighting scheme for the two predictions. Evidently, there is a 3.9% decrease in subject classification accuracy and a 1.7% decrease for liveness detection, averaged across all seven methods, compared to the non-occlusion scenarios. However, it can be asserted that this decline is due to the less discriminative RGB features, as evidenced by the lower performance of the RGB-only model compared to its general results. The radar-only model achieves an improved 42.3% accuracy at recognising unseen subjects with occlusion, and a liveness detection rate of 99.3%, beating the scores of the RGB-only model. This aligns with our hypothesis that mmWaves, with their ability to permeate through fabric to directly reach the dermal layer of the skin, could offer robustness against face concealment. However, this warrants verification with a bigger participant pool and even more extreme occluding accessories to say with confidence.

Concatenation arises as the outperforming strategy in the presence of occlusion, surpassing other fusion strategies by over 6.8% in mean accuracy. This is mainly attributed to its high liveness detection rate of 99.5%. Impressively, it is still able to cover a significant proportion of the test subset, achieving a 99.7% coverage rate, exceeding even the individual modalities by 18.0%.

4.5 Discussion

Taking all findings into account, it is safe to conclude that the concatenation strategy yields the best-performing model overall, especially excelling in determining facial authenticity. It attains second-highest in the aggregated general metrics, third-highest in the dim lighting setting, and top-ranking in the occlusion-only scenarios. On top of this, the concatenation fusion strategy produces the most distinctive final feature embeddings with clear separations between facial identities as well as liveness in its t-SNE plots. This is also evidenced by the method achieving the highest macro-averaged AUC showing a much greater sensitivity rate over false alarms across all decision thresholds.

There were three total candidates for the best fusion strategy, but the concatenation came out on top due to its higher coverage rate than both the Hadamard product and the multi-head attention mechanisms. The Hadamard product is a strong contender, achieving a slightly higher subject accuracy over all conditions. However, it falls short in its AUC, and therefore, classifying ability, yielding more false positives over certain thresholds than the concatenation strategy. Multi-head attention emerges as the weakest among the three, attaining a lower AUC metric and the lowest coverage rate of 92.5% out of all fusion strategies. Still, it was the top performer under dim light, by a small margin, showing promise in its capabilities.

Feature fusion by concatenation ensures that all extracted information from both modalities is kept, potentially explaining its higher performance over the others. The Hadamard product and multi-head attention strategies tend to mix different aspects from both modalities which may dilute certain characteristics that are relevant for classification. For instance, the element-wise multiplication assumes that the radar feature dimensions correspond exactly with the dimensions of the RGB features. This is most certainly never the case since both modalities are represented and learned within separate embedding spaces. Therefore, pairwise dot strategies, including multi-head attention as this essentially comprises a series of dot products, aimed to blend features in order to uncover all possible correlations between every radar and RGB feature for the following classification phase.

Nevertheless, it is clear that the subsequent pooling method has a tremendous impact in the model's effectiveness. For instance, the max-pooling strategy produced the poorest-performing model in all accounts, in many cases degrading

the performances of the individual modalities. The mean and flatten methods exhibit intriguing properties, with the average-pooling producing the highest subject accuracy but a relatively low liveness detection capability. Meanwhile, the flatten method performs relatively well within most tasks, narrowly missing the top three. This is interesting since the flatten method aims to preserve as much information, much like the concatenation strategy. It is important to note that following the flatten operation, the resulting feature vector expands to n^2 in size, meaning that the fused vector comprises $512^2 = 262,144$ dimensions. This demands additional fully connected layers to effectively compress it into the final 64-dimensional embedding vector. A single layer was chosen to ensure fair consistency across all models. Yet, it is apparent that this is insufficient to abstract all the information regarding how each radar and RGB feature relate to one another following the flatten.

It is also crucial to highlight that the AUC metric is invariant to specific decision thresholds. This has a profound impact whenever there is a disparity in the cost of false negatives over false positives. In secure facial biometric authentication, false positives indicate a larger flaw in the system, meaning that the underlying model should prioritise minimising them, even if that entails an increase of more tolerable false negatives. This is especially true for the binary liveness check, where the model should be rewarded for minimising the success rate of spoofing attacks. For this reason, the weighted $F_{0.5}$ scores are more indicative of the models' real-world applicability.

The results of the non-hybrid models, utilising the modalities independently, show that the RGB embeddings from the pretrained 2D InsightFace are rich enough to discern between real and fake faces at a moderately decent rate. This shows that the subtle pixel-to-pixel differences between the authentic and fake photographs are effectively extracted and used within the liveness prediction. This can be attributed to the lower quality of the printed faces on paper in comparison to the live face captures. Furthermore, the mmWave radar signatures obtained from the Soli lack the precision required for a clear separation between the different participants. However, the fusion models effectively leverage the 3D information from the radar features to boost the relatively modest liveness detection capabilities of InsightFace.

5. CONCLUSIONS

In this paper, we proposed `mmFace`, a novel 3D face recognition system designed to efficiently recognise and distinguish between unique identities, while also verifying their legitimacy. The system harnesses data gathered from RGB cameras, alongside compact mmWave radar sensors, to encode human faces, regardless of environmental conditions and the presence of face accessories. The system demonstrates robustness against 2D spoofing attacks, and attempts to conceal the face using common items such as scarves, sunglasses, and hats. Our models were trained on our own dataset featuring the faces of 21 subjects captured under specific controlled conditions. These include five facial poses: frontal, as well as, 30° and 45° left and right azimuth angles – two lighting settings: regular and dim – and finally, two occlusion scenarios: with and without face accessories.

Several strategies for blending the information from both modalities were analysed to determine the most optimal fusion method. We evaluated the strategies through a zero-shot classification task to assess the generalisability of our models using standard metrics such as prediction accuracy, F-measures and ROC curves. Furthermore, the fused feature vectors of all test samples were visually inspected in order to verify that specific subjects were pushed to exclusive regions of the multimodal space. Benchmarking the strategies against the non-hybrid models provided clear evidence that the hybrid models effectively improved on the performance of the individual modalities acting separately.

Next, the strategies were evaluated under the isolation of specific conditions such as dim lighting and occlusion settings. This ultimately resulted in the concatenation fusion strategy being found to be the best overall performing method. It produces high face recognition scores as well as extremely accurate liveness detection rates.

Our work investigated the application of mmWave sensing from a new perspective for compact, secure biometric authentication, not looked at by previous papers. However, even with the positive outcomes, there are a number of limitations that require attention.

5.1 Limitations

The primary limitation that needs to be addressed is that our curated dataset is much too small of a sample size to be representative of the general population. Moreover, a larger dataset would allow for a more extensive test set, providing an in-depth examination of the model's ability to generalise to unseen faces. We initially aimed to gather 50 participants, however, this goal was not met in time. Nonetheless, we believe our current dataset of 21 represents a notable improvement over previous work exploring mmWave sensing [11, 12, 13], often utilising only three to eight participants. This instils greater confidence in the performance of our proposed model.

The results of the dim lighting experiment exposed another flaw within the dataset: subtle inconsistencies between face scans. The dimly lit environment lacked strict control, as the experiment room allowed natural light that could not be fully blocked, resulting in variations in captures taken at different times of the day. This highlights the need for a more rigorous approach in future experiments, involving specific lumen measurements and a light-controlled room to ensure consistency across captures. This would also ensure higher-quality face scans to test whether the radar-based methods yield better classifications than RGB cameras in more extreme circumstances.

One of the major limitations uncovered was that the data collected by Google's Soli chip was too sparse to extract enough distinctive facial features. As a consequence, the radar-only model exhibited relatively low performance in subject prediction compared to the InsightFace pretrained model. Although this outcome was expected, it was noteworthy that the RGB-only model still achieved moderately high scores in the liveness detection category, which was anticipated to be dominated by the radar-only model.

5.2 Future Work

The identified limitations demonstrate areas of further scope for improving our proposed model. Expanding the dataset collection within a controlled environment not only enhances dataset quality but also opens up the possibility of employing vision transformers for the task. Vision trans-

formers have shown remarkable success in large-scale image classification tasks, leveraging self-attention mechanisms to learn complex visual patterns and relationships. However, transformer architectures are famously data-hungry, requiring millions of training samples to effectively learn from compared to CNNs [41]. The multi-head attention strategy showed promise within our model, achieving the third-highest ranking against our zero-shot task. However, it is important to note that it was only applied at a single layer for the sole purpose of integrating separate modalities. A possible avenue could be to use multi-head attention layers throughout the feedforward network, starting from patches of the base `ARD` input with added positional encoding to maintain spatial context. This approach would progressively extract relevant features through many layers of abstraction, attending to different aspects of the range and Doppler information.

Another area of interest was the pairwise dot-then-flatten strategy. It was decided to utilise only a single fully connected layer following the feature fusion, ensuring consistency across all fusion strategies. However, this flatten operation presents a unique case as it yields a feature vector spanning 260 thousand dimensions, requiring a series of hidden layers to reduce its embedding dimensions through a more gradual process. This presents an intriguing opportunity, especially considering the success of the concatenation strategy, which demonstrates that preserving all information from both modal features offers the most optimal compromise.

Finally, broadening the scope of experimental variables during data acquisition could prove beneficial. For instance, diversifying the clothing apparel to include more extreme cases of facial concealment through items such as ski masks or Halloween masks could offer valuable insights. Furthermore, a look into the effect of facial expressions in future data collection efforts could be interesting, training the model to be expression-invariant.

Acknowledgments. I would like to thank my supervisor, Dr Hang Dai, for his helpful comments and guidance throughout the year. I'd also like to thank Dr Chaitanya Kaul for providing me with the equipment, and his invaluable support with the data acquisition process which was critical to the project's success. Last but not least, a huge thanks to all who participated in our data collection study.

6. REFERENCES

[1] Song Zhou and Sheng Xiao. 3d face recognition: a survey. *Human-centric Computing and Information Sciences*, 8(1):1–27, 2018.

[2] Woodrow Wilson Bledsoe. The model method in facial recognition. *Panoramic Research Inc., Palo Alto, CA, Rep. PR1*, 15(47):2, 1966.

[3] Chenghua Xu, Yunhong Wang, Tieniu Tan, and Long Quan. Depth vs. intensity: Which is more important for face recognition? In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 342–345. IEEE, 2004.

[4] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security*, 10(4):746–761, 2015.

[5] Apple Inc. About Face ID advanced technology, 2023. Accessed: 2023-11-19 `https://support.apple.com/en-gb/102381`.

[6] A Soumya, C Krishna Mohan, and Linga Reddy Cenkeramaddi. Recent advances in mmwave-radar-based sensing, its applications, and machine learning techniques: A review. *Sensors*, 23(21):8901, 2023.

[7] Xin Wang, HuaZhi Pan, Kai Guo, Xinli Yang, and Sheng Luo. The evolution of lidar and its application in high precision measurement. In *IOP Conference Series: Earth and Environmental Science*, volume 502, page 012008. IOP Publishing, 2020.

[8] Apple Inc. Use Face ID while wearing a mask with iPhone 12 and later, 2024. Accessed: 2024-04-05 `https://support.apple.com/en-gb/102452`.

[9] David R Vizard and R Doyle. Advances in millimeter wave imaging and radar systems for civil applications. In *2006 IEEE MTT-S International Microwave Symposium Digest*, pages 94–97. IEEE, 2006.

[10] Eran Hof, Amichai Sanderovich, Mohammad Salama, and Evyatar Hemo. Face verification using mmwave radar sensor. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 320–324, 2020.

[11] Hae-Seung Lim, Jaehoon Jung, Jae-Eun Lee, Hyung-Min Park, and Seongwook Lee. Dnn-based human face classification using 61 ghz fmcw radar sensor. *IEEE Sensors Journal*, 20(20):12217–12224, 2020.

[12] J Kim, J-E Lee, H-S Lim, and S Lee. Face identification using millimetre-wave radar sensor data. *Electronics Letters*, 56(20):1077–1079, 2020.

[13] Ha-Anh Pho, Seongwook Lee, Vo-Nguyen Tuyet-Doan, and Yong-Hwa Kim. Radar-based face recognition: One-shot learning approach. *IEEE Sensors Journal*, 21(5):6335–6341, 2021.

[14] Muralidhar Reddy Challa, Abhinav Kumar, and Linga Reddy Cenkeramaddi. Face recognition using mmwave radar imaging. In *2021 IEEE International Symposium on Smart Electronic Systems (iSES)*, pages 319–322, 2021.

[15] Intel Corporation. Intel RealSense LiDAR Camera L515, 2023. Accessed: 2023-11-19 `https://www.intelrealsense.com/lidar-camera-l515/`.

[16] Jaime Lien, Nicholas Gillian, M Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. Soli: Ubiquitous gesture sensing with millimeter wave radar. *ACM Transactions on Graphics (TOG)*, 35(4):1–19, 2016.

[17] DF Robot. Eight Practical Applications of mmWave Radar Technology, 2023. Accessed: 2023-11-19 `https://www.dfrobot.com/blog-1650.html`.

[18] Cadence Design Systems. mmwave radar applications and advantages, 2022. Accessed: 2023-11-25 `https://resources.system-analysis.cadence.com/blog/msa2022-mmwave-radar-applications-and-advantages`.

[19] Nicholas Gillian Jaime Lien. Soli: Radar-based perception and interaction, 2020. Accessed: 2023-11-25 `https://blog.research.google/2020/03/soli-radar-based-perception-and.html`.

[20] Bassem R Mahafza. *Radar systems analysis and design using MATLAB*. Chapman and Hall/CRC, 2005.

[21] Eiji Hayashi, Jaime Lien, Nicholas Gillian, Leonardo Giusti, Dave Weber, Jin Yamanaka, Lauren Bedal, and Ivan Poupyrev. Radarnet: Efficient gesture recognition technique utilizing a miniature radar sensor. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[22] Kevin Mitchell, Khaled Kassem, Chaitanya Kaul, Valentin Kapitany, Philip Binner, Andrew Ramsay, Daniele Faccio, and Roderick Murray-Smith. mmsense: Detecting concealed weapons with a miniature radar sensor. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[23] Peijun Zhao, Chris Xiaoxuan Lu, Jianan Wang, Changhao Chen, Wei Wang, Niki Trigoni, and Andrew Markham. mid: Tracking and identifying people with millimeter wave radar. In *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 33–40. IEEE, 2019.

[24] Evyatar Hemo, Amichai Sanderovich, and Eran Hof. mmwave radar face signatures, 2018. `https://dx.doi.org/10.21227/wr67-kx23`.

[25] Bitsensing. BTS60 Technical Specification, May 2020. Accessed: 2023-12-01 `http://bitsensing.com/pdf/Technical_Specification_InCabinRadar_miniV.pdf`.

[26] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[27] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[28] Jia Guo, Jiankang Deng, Niannan Xue, and Stefanos Zafeiriou. Stacked dense u-nets with dual transformers for robust face alignment. In *BMVC*, 2018.

[29] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the IEEE Conference on European Conference on Computer Vision*, 2020.

[30] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[31] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011.

[32] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[33] Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saiedeh N Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1):28–44, 2013.

[34] Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. Effective techniques for multimodal data fusion: A comparative analysis. *Sensors*, 23(5):2381, 2023.

[35] Francisco Charte, David Charte, Salvador García, María J del Jesus, and Francisco Herrera. A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software, and guidelines. *arXiv preprint arXiv:1801.01586*, 2018.

[36] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[39] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. *Advances in neural information processing systems*, 29, 2016.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.