University of Glasgow | School of Computing Science

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

Stergious Aji (2546916A)

April 3, 2024

ABSTRACT

TODO

1. INTRODUCTION

Facial recognition is an evolving research domain within the field of computer vision, finding extensive use across areas including human-computer interaction, security surveillance, and forensic analysis. Its primary application revolves around biometric authentication, granting individuals access to their devices or restricted zones. This enables a non-intrusive, hands-free means of verifying identity, 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, facial 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. In recent years, the emergence of deep learning has reshaped human face classification, leveraging extensive online repositories of face images for improved performance and efficiency. However, 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 regulated. Furthermore, the transition to three-dimensional 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 gauge and delineate facial features. 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 limited by their inability to penetrate thin materials such as clothing and hair. In contrast, millimetre wave radar (mmWaves) can penetrate such materials and directly reach the skin's dermal layer [8]. This could enable greater performance in scenarios involving facial hair or adverse environmental conditions such as rain or fog.

Research into the efficacy of radar waves for 3D face recognition remains relatively limited, although recent studies indicate promising outcomes [9, 10, 11, 12, 13]. Radar sensors typically offer greater cost efficiency in terms of both acquisition and computation, as they consume less power compared to sensors NIR-based systems. Nevertheless, it is crucial to acknowledge the trade-off, as mmWaves often result in a sparser representation. This could impact recognition accuracy, where precision in detecting and mapping of facial features is paramount. Thus, we aim to address this limitation by integrating information from colour images, potentially paving the way for more resilient and versatile systems.

1.2 Research Contributions

Our work explores the effectiveness of using RGB cameras in conjunction with mmWave radar sensors for 3D facial recognition. Since there are no appropriate datasets available for this purpose, we have collated this data ourselves. We use the Intel RealSense L515 RGB-D camera [14] for photographing individuals' faces. Meanwhile, the Google Soli 60 GHz radar sensor [15] is employed to gather depth information by transmitting and measuring millimetre waves reflected from the target.

We have gathered face data of 21 participants under various conditions encompassing diverse poses, lighting settings, and common occlusion scenarios. This comprehensive approach led to a system that demonstrates resilience to environmental factors, in comparison, to systems relying solely on RGB data.

We developed a novel face recognition model using a deep convolutional neural network. This model was trained on the captured data to learn facial features from both the RGB and depth characteristics, simultaneously. We investigated different techniques in fusing these two modalities, aiming to pinpoint the most effective strategy that provides rich and distinctive representations for clean identity separation. The model's effectiveness is benchmarked against prior radar-based facial recognition systems, as well as, a comparison to using each modality independently.

The key contributions of this paper are summarised below:

- Compilation of a diverse face dataset comprising color images and mmWave signatures from 21 participants. The dataset encompasses five different poses, two lighting conditions, and two occlusion scenarios.
- We present `mmFace`, a novel face recognition model that harnesses both modalities yielding a robust system capable of handling common occluding materials and nullifying 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 in small-scale, 3D face identification using mmWave technology.

2. BACKGROUND

2.1 mmWave 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 have expanded its applicability to more small-scale domains [6]. mmWave sensing has particularly excelled in autonomous vehicles facilitating object detection for systems such as collision warnings and adaptive cruise control [16]. This is primarily due to its edge over traditional near-infrared waves employed by lidar cameras, specifically in its resilience to atmospheric conditions such as dust, smoke, fog, and rain [17]. This penetrative power of mmWaves make 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 [18]. However, the sensor's potential application to face recognition remains unexplored, presenting a unique research opportunity. Consequently, this is the sensor we used to capture mmWave face signatures during our data collection phase. 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 [19], a vital aspect for extracting accurate facial features. The Soli chip has a relatively low power consumption due to the fact that it sends 16 chirps every burst at a pulse-repetition frequency of 2 kHz, then stops transmitting until the next burst of chirps [20, 21]. Each burst is transmitted at 25 Hz giving an overall transmission duty cycle of 2% meaning the radar chip is turned off during the majority of its operation saving a lot of power for mobile applications.

2.2 Related Work

The exploration of millimetre waves for facial identification represents 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. [22]. While this paper primarily examines the classification of subjects based on their gait and body shape rather than facial features, 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.

Following this, Hof et al. [9] 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 [23]. While this collection encompasses a wide range of faces, including some individuals with spectacles and beards and, 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 smartphone integration. 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 neurons in their Neural Network and adding an extra hidden layer could compensate for this loss.

Lim et al. [10] proposes another deep neural network with a more traditional Multi-Layer Perceptron (MLP) architecture where every layer is fully connected to adjacent ones. The study utilised a small-scale, 61 GHz FMCW radar sensor developed by bitsensing Inc. [24], 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 subjects. The research also explored the impact of using a single receiving antenna, which resulted in a reduced accuracy of 73.7%. This finding is in line with Hof et al.'s [9] observation that an increased number of receiving antennas can enhance classification accuracy by the ability to capture more nuanced facial features.

During the same time frame, Kim et al. [11] 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 con-

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

nected 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. [12] 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 training and testing phases, a distance metric between the outputs of the networks is used to assess the similarity between 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. [10, 11], 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) [25] 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. [13] employs two different machine learning models on the dataset provided via the IEEE port [23]. 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 seconds, 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 [26], integrating these models with mmWave radar data presents a promising avenue. This could expedite face acquisition time while capitalising on the advantages of mmWaves in terms of their resilience to lighting variations and occlusions.

2.3 InsightFace

In the evolving field of face recognition, deep CNNs have emerged as a dominant approach due to their ability to 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. [26], alongside the robust Face Alignment model by Gho et al. [27]. 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 [28], 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) [29] and the YouTube Faces (YTF) datasets [30]. The integration of pretrained models offered by InsightFace into our system enables us to concentrate efforts on enhancing the performance of our model's depth and contour detection capabilities.

2.4 Multimodal Data Fusion Methods

Multimodality, as defined by Lahat et al. [31], refers to the use and analysis of multiple types of data, potentially arriving from multiple sensors. The idea is to extract and blend salient information gathered by each sensor. The integration of this diverse data lead 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 [32]. 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 effectiveness.

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 for the adjustment of the influence of specific modalities [33]. Late fusion is often simpler and more adaptable, proving effective when dealing with highly dissimilar data sources in terms of sampling rate, dimensionality, or unit of measurement. Furthermore, late fusion often yields superior 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. [34]. 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 challenges and 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. It would be challenging to attempt early fusion due to the substantial differences between the two modalities. Such integration would likely require heavy preprocessing of the radar data, potentially involving its conversion into a depth image.

3. METHODOLOGY

3.1 Data Acquisition

Following a comprehensive research of the field, the subsequent steps involved planning and executing 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 outlined in 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 on 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 penetrating 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 be not required as facial features could be captured regardless. This could 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. [11], 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 tight time frame of the project. Adhering to ethical standards regarding sensitive personal information, our participant pool consists of university students in the age range 18–35 years. While this results in an over-representation of individuals aged 20–25 years, it should not impact our study as age variance is not being explored. A total of 15 scenarios are captured for each participant at a distance of 20 cm from the sensors. Each time the sensors are run for 10 seconds, totalling 150 RGB frames and 3,750 mmWave frames per person. On top of this, scans of printed faces are also collected in order to train the model to detect the face liveness as well as its identity. This was restricted to the three frontal poses of all participants to mirror common spoofing strategies, providing another 30 RGB frames and 225 mmWave frames per fake counterpart.

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 marking the face distance and the five pose directions indicated by the yellow tape.

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

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 conditions are abbreviated as outlined in Table 1.

Moreover, the radar bursts acquired during the data collection phase undergo multiple FFT stages of preprocessing to convert the raw signals into discretised Complex Range-

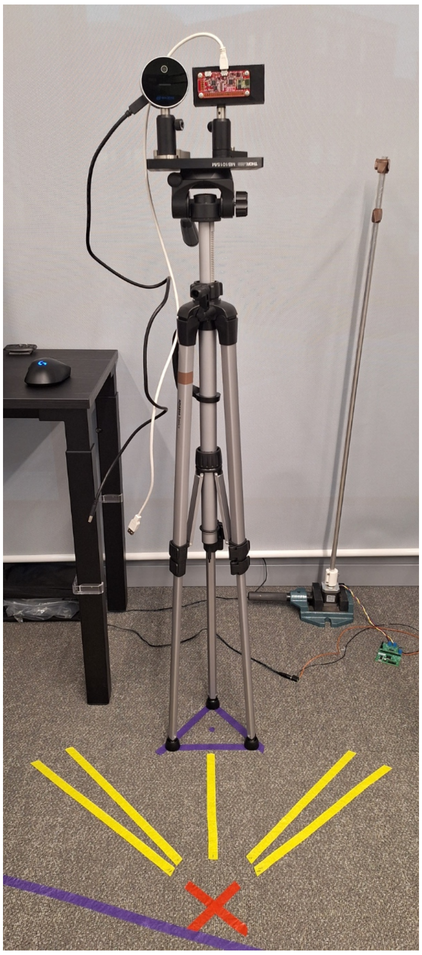

Figure 2: *Experiment setup used for data acquisition with the equipment mounted on a tripod. The red cross marks the* 20 cm *face distance and the yellow tape indicates the five pose directions.*

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

Table 1: *Table displaying full forms for abbreviations describing experiment conditions.*

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 [15, 20]. 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, and bursts sampled at a rate of 25 Hz. 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 of the same subject's face. The plot illustrates the discrete intensities of received signals across 16 Doppler bins along the x-axis and 32 Range bins along the y-axis.

3.2 mmFace

Building on the intuition from Section 2.3 of the Background 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 [35] and Google's FaceNet [36]. 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 model we are naming, `mmFace`. Furthermore, this facilitates exploration into the various methods in fusing the two modalities.

Figure 5 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 showing every layer. To summarise, `mmFace` takes two inputs: an mmWave face signature in an `ARD` format and an InightFace embedding extracted from the corresponding RGB frame. In order to simplify computation, the magnitudes of every complex range r and Doppler d bin are taken to produce an Absolute Range-Doppler (ARD) map using the absolute function as follows:

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

The inputs then go through three main stages: **mmWave Feature Extraction**, **Feature Fusion**, and finally, **Class Prediction** each described in detail below to finally produce a subject and liveness prediction. This is simply a binary classification: 0 for fake and 1 for real.

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 embedding vector.
2. **Feature Fusion**: The next phase involves fusing the extracted radar embedding with the RGB embedding input. We investigate seven different fusion strategies with the simplest being concatenation of the two features. This is then processed through a single fully connected layer to reduce the dimensionality before the final stage.
3. **Class Prediction**: Finally, the fused multimodal embedding is carried over through two separate fully connected layers to predict the identity of the face and its liveness property to counter against spoofing attacks.

The three-channel ARD format of the mmWave face signatures allows leveraging convolutional-based feature extraction due to its image-like format. Convolutional layers can detect spatial patterns among the range and Doppler profiles for specific faces, which can be enhanced with deeper layers. Due to the relatively small size of the `ARD` maps only being 32×16 bins/pixels, it was important that most of the information is kept reducing the need for any more max-poolings or strided convolutions. To simplify the fusion

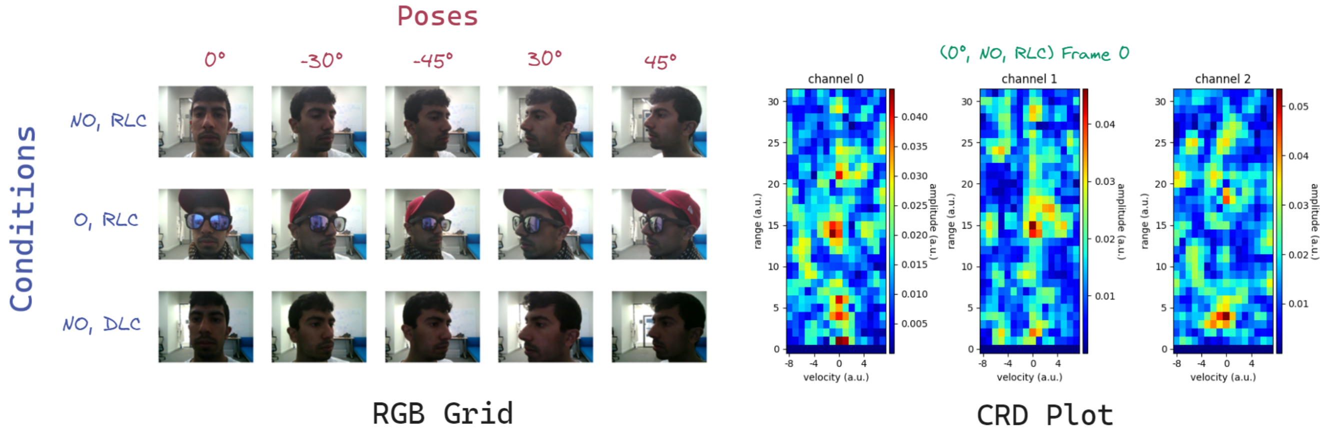

Figure 3: *Data samples collected for Subject 0. The left figure shows the RGB frames of all 15 scenarios organised by pose and condition. The right figure plots a single CRD frame showing amplitudes of reflected waves detected by the three receiving channels of the Soli, categorised into discrete Range-Doppler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

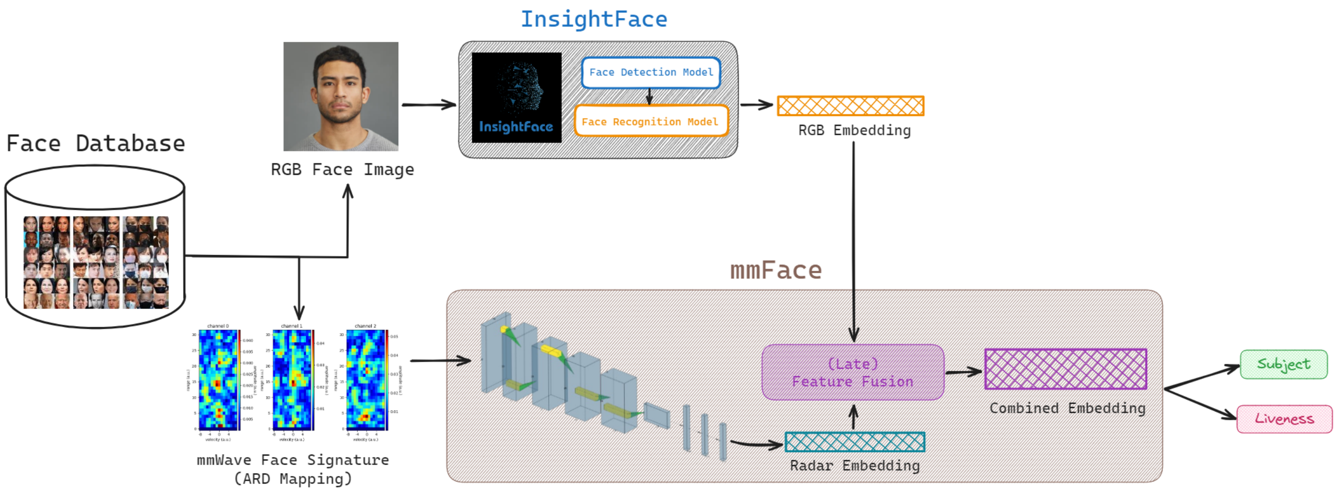

Figure 4: *High-level model workflow diagram of our proposed 3D face recognition system incorporating millimetre-wave radar and RGB images.*

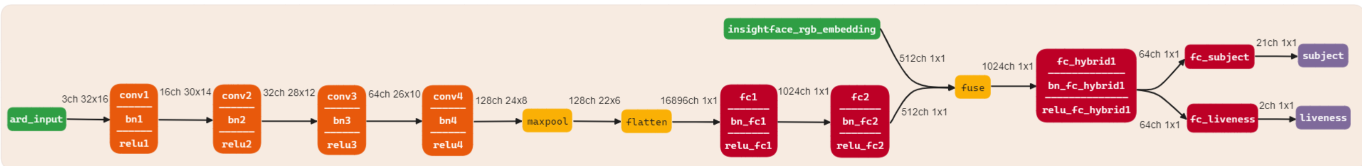

Figure 5: *Model architecture of our* `mmFace` *model displaying each layer, as well as its input and output sizes.*

stage, we decided to match the final radar embedding vector size with the 512-dimension InsightFace embedding using two fully connected layers. Four convolution layers with a kernel size of 3×3 with filter sizes increasing by powers of two starting from 16 to 128 were found to be sufficient with increasing layers affording diminishing returns. All linear transformations are followed by a batch normalisation layer to reduce overfitting and increase the generalisability of the model. All non-linear transformations were chosen to use the ReLU activation function to prevent vanishing weights.

We opted to focus on mixing the two modalities within the feature-space, specifically within the final layers of the neural network. This entails an easier fusion process since the data from both modalities are abstracted far enough into a compressed representation. Pure intermediate fusion is not feasible due to the black-box treatment of the InsightFace model making it difficult to integrate information from both modalities within its hidden layers. Early fusion presents significant challenges due to the dissimilarities in sampling rates and data formats. The `ARD` maps have to be synchronised and transformed into a depth image. This is an interesting avenue left to be explored possibly training a neural network to learn an idyllic pixel-wise depth representation of the radar bursts.

`mmFace` contains around 2.8 million parameters and takes on average 4.1 milliseconds (SD = 0.2 ms) to process a single (`ARD`, RGB-embedding) input pair on an NVIDIA GeForce GTX 1650 GPU.

3.2.1 Training

We created and trained our models using PyTorch version 2.1 for approximately 20 to 25 epochs, employing a Stochastic Gradient Descent optimiser minimising the Cross Entropy Loss (L_{CE}). We opted for a fixed learning rate of 0.01, an L2 regularization rate of 1e−3, and a momentum of 0.9. As our model generates two predictions, we merged the losses from each with equal weighting to ensure equitable learning of both attributes. This combined loss $\mathcal{L}$ is formalised 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, respectively, while $\hat{s}$ and $\hat{l}$ signify the model's corresponding predictions.

3.2.2 Testing

We evaluate our models through a zero-shot classification task in order to assess the generalisation ability at identifying unseen faces. We train the models on a random subset of 17 of the total 21 subjects, just over an 80% train split, keeping in mind this includes 17 fake counterparts totalling 38 subject instances. The remaining four subjects (or eight total instances) are left out for testing. During inference, the features from the final hybrid fully connected layer, `fc_hybrid1`, are used. The cosine similarity between each output embedding of the test data and eight pre-selected reference embeddings are calculated to find the maximal score and corresponding prediction for face identity and liveness. A decision threshold t is used such that the maximal score must be greater than it to qualify for a valid prediction. This ensures that predictions are not made for output embeddings that are mostly equidistant to all reference embeddings such that no practical decision can be made.

3.3 Feature-level Fusion

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

1. **Concatenate:** This is a concatenation of the two feature vectors, the most common strategy employed for its ease of implementation and effectiveness.

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

2. **Add:** This involves an element-wise vector addition of $\vec{\boldsymbol{x}}$ and $\vec{\boldsymbol{y}}$. This can be very effective if 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.

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

3. **Multiply:** This is an element-wise vector multiplication or the Hadamard product of $\vec{\boldsymbol{x}}$ and $\vec{\boldsymbol{y}}$. This preserves the original feature vector structure while emphasising relationships between corresponding feature elements.

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

4. **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 intuition behind this comes from the pairwise dot providing a more representative mixing of the feature elements since each radar feature is multiplied by every RGB feature which are all looked at during the pooling method to reduce the dimensionality of the matrix.

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

5. **Pairwise Dot Max:** This similarly involves the dot product followed by a column-wise max. Similar to the max-pooling strategy, bigger features are selected from the excessive mixing of the two modalities rather than an averaging which can squash certain feature correlations. $*$ denotes a selection of all rows of a matrix.

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

6. **Pairwise Dot Flatten:** This is the final pairwise dot fusion strategy now following the dot product with a flatten operation of the $n \times n$ matrix into an n^2-sized vector. This conversely keeps all of the correlations between the radar and RGB features giving the model to select which features are most useful through the final fully connected layers.

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

7. **Multi-Head Attention:** This involves using a self-attention mechanism popular in transformer architectures for natural language processing tasks. The key idea of self-attention is to isolate and mix the most important aspects from both feature vectors by transforming them into three separate representations, namely a query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$ [37]. Each embedding has a distinct role with the query capturing features that the model deems relevant for making predictions. $\boldsymbol{Q}$ may focus on facial features that are common or discriminative across both modalities such as facial landmarks and overall identity information. Meanwhile, $\boldsymbol{K}$ may focus on modality-specific elements such as the colour and texture information embedded within the RGB feature vector while structural information being offered by the radar embedding. Finally, $\boldsymbol{V}$ contains 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 affect, this mechanism is applied separately across multiple attention heads, decreasing the possibility of the model missing salient aspects of the feature. This is formalised below, where first the two inputs are stacked vertically, $\boldsymbol{X} = \begin{bmatrix} \vec{\boldsymbol{x}} \\ \vec{\boldsymbol{y}} \end{bmatrix}$, and copied through three separate linear transformations to obtain the $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ matrices. This is done for each of

the k attention heads using separate learnable weight matrices.

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

4. EVALUATION

We evaluate our model by examining its discriminative ability at representing four unseen faces. Various metrics such as the prediction accuracies, precision and recall are used to compare the different fusion strategies against each other as well showing the performance of individual modalities. Prior to the zero-shot task, we select and extract the final hidden representation of eight reference frame pairs that the test set will be compared against. These eight contain both the live and fake samples of the random subset of four unseen subjects.

4.1 Results

Firstly, the model accuracies at predicting both the subject identity and the liveness flag are measured. This is done by calculating the most similar reference embedding to the final feature vector of a particular test sample by the maximum cosine similarity. A decision threshold t of 0.5 is used to quantify a true prediction, however, all models achieved a high coverage with the lowest being the radar only model with a 79.8% coverage. Secondly, since

Table 2 displays the subject and liveness accuracies for all feature fusion strategies as well as the weighted F_β-measures, over all the respective classes in the test data. For completeness, the performance of the individual modalities are also listed in order to show the advantageous effect of the multi-modal feature fusions. A β of 0.5 was chosen for the F-scores since in face recognition systems falsely allowing the wrong identity through the system is more harmful than false negatives requiring twice as much emphasis for higher precision compared to recall.

The best and worst performers are highlighted in green and red respectively. Evidently, certain fusion strategies performed better at identifying subjects while others performed better at predicting the face liveness. The concatenation strategy was found to be best suited at representing faces for an accurate liveness check. Meanwhile, the pairwise dot then mean strategy outperformed within the subject category, attaining the highest accuracy and $F_{0.5}$ measure. However, it is clear that it is relatively poor at detecting the face liveness, obtaining an accuracy that is even lower than simply using the mmWave radar features alone. This is likely due to the mean pooling step squashing outlier feature correlations between the two modalities. The vector addition performed very poorly at predicting facial identities, being much better at predicting liveness in comparison.

Looking at the individual performance for the two prediction types is useful at judging the strategies' effectiveness at classifying unseen data to the respective classes.

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
Multiply	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.*

Fusion Strategy	Mean Accuracy (%)	Mean $F_{0.5}$ Score
Concatenate	91.7	0.915
Add	81.1	0.811
Multiply	91.9	0.916
Pairwise Dot Mean	84.8	0.844
Pairwise Dot Max	77.8	0.778
Pairwise Dot Flatten	90.7	0.903
Multi-Head Attention	91.3	0.900
Radar Only	67.4	0.643
RGB Only	77.4	0.778

Table 3: *TODO*

Fusion Strategy	Macro-Averaged AUC
Concatenate	0.743
Add	0.667
Multiply	0.728
Pairwise Dot Mean	0.920
Pairwise Dot Max	0.851
Pairwise Dot Flatten	0.856
Multi-Head Attention	0.823
Radar Only	0.538
RGB Only	0.912

Table 4: *Performance Metrics for Feature Fusion Techniques*

5. CONCLUSIONS

5.1 Future Work

Acknowledgments. This is optional; it is a location for you to thank people, most probably your family and your

supervisor.

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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

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

[17] 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`.

[18] 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`.

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

[20] 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.

[21] 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.

[22] 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.

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

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

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

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

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

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.