mmFER: Millimetre-wave Radar based Facial Expression Recognition for Multimedia IoT Applications

Xi Zhang[†,1,2], Yu Zhang[†,1], Zhenguo Shi[1], Tao Gu[1]
[1]Macquarie University, [2]RMIT University
zaibuer@gmail.com, {y.zhang, zhenguo.shi, tao.gu}@mq.edu.au
[†]First authors with equal contribution

Abstract

Facial expression recognition plays a vital role to enable emotional awareness in multimedia Internet of Things applications. Traditional camera or wearable sensor based approaches may compromise user privacy or cause discomfort. Recent device-free approaches open a promising direction by exploring Wi-Fi or ultrasound signals reflected from facial muscle movements, but limitations exist such as poor performance in presence of body motions and not being able to detect multiple targets. To bridge the gap, we propose mmFER, a novel millimeter wave (mmWave) radar based system that extracts facial muscle movements associated with mmWave signals to recognize facial expressions. We propose a novel dual-locating approach based on MIMO that explores spatial information from raw mmWave signals for face localization in space, eliminating ambient noise. In addition, collecting mmWave training data can be very costly in practice, and insufficient training dataset may lead to low accuracy. To overcome, we design a cross-domain transfer pipeline to enable effective and safe model knowledge transformation from image to mmWave. Extensive evaluations demonstrate that mmFER achieves an accuracy of 80.57% on average within a detection range between 0.3m and 2.5m, and it is robust to various real-world settings.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

Keywords

mmWave, Facial Expression Recognition, Deep Learning

ACM Reference Format:
Xi Zhang[†,1,2], Yu Zhang[†,1], Zhenguo Shi[1], Tao Gu[1]. 2023. mmFER: Millimetre-wave Radar based Facial Expression Recognition for Multimedia IoT Applications . In *The 29th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '23), October 2–6, 2023, Madrid, Spain.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3570361.3592515

ACM MobiCom '23, October 2–6, 2023, Madrid, Spain

ACM ISBN 978-1-4503-9990-6/23/10…$15.00
https://doi.org/10.1145/3570361.3592515

1 INTRODUCTION

Multimedia Internet of Things (IoT) applications have become popular in recent years [41], ranging from digital entertainment [23, 45] to digital advertising [19]. To enable better service quality and user experience, emotional awareness has been advocated as a key factor in perception to "understand" audience [50, 57]. Facial expression recognition (FER) plays a vital role in emotional awareness [32, 46] since facial expressions are intuitive reflections of user's emotional states. A FER system infers facial expressions and delivers an assessment of audience's preference, interest level, engagement and reactions [9].

FER has been extensively studied over the last decade. Vision-based approaches [11, 40] achieve state-of-the-art accuracy, but their performance may be vulnerable to ambient lighting conditions [46], *e.g.*, watching movies with poor ambient light. The time-of-flight (ToF) camera (*e.g.*, depth camera [52]) may work better in low lighting, but it still suffers from natural illumination (*e.g.*, glare from glasses or exposed to sunlight) and fails with occlusion (*e.g.*, wearing masks). Most importantly, although security measurement can be put in place, cameras may raise serious privacy concerns (*e.g.*, from psychology aspect, people do not feel safe and comfortable with a camera constantly monitoring them [15]). Without compromising user privacy, wearable-based approaches have been advocated to recognize facial expressions using wearable sensors, *e.g.*, PPG [53], EEG [12], and earphones [51, 55], but long-time wearing may cause discomfort to users. Device-free approaches have been proposed

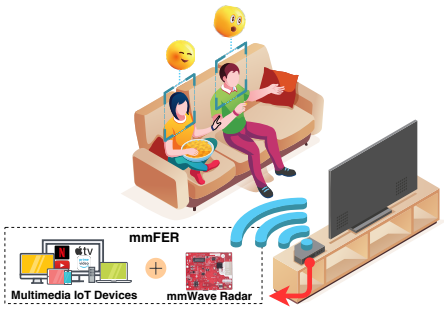

Fig. 1: mmFER use scenario: an example.

leveraging on Wi-Fi or ultrasound signals reflected from facial muscle movements. However, due to intrinsic limitations of Wi-Fi signals (*i.e.*, low bandwidth and multipath effect), the WiFi-based approach [7] may fail in presence of body motions (*e.g.*, leg shaking). The ultrasound-based approach [13] can reach 60cm at maximum but the distance of watching TV is typically in a range between 1m and 3m [4, 16]. Moreover, these device-free FER approaches typically do not support multi-user applications.

Millimeter wave (mmWave) sensing has recently been popular due to its high bandwidth and robustness. Compared with Wi-Fi and ultrasound sensing, mmWave sensing provides higher signal resolution, hence being capable of detecting subtle movements [30]. In addition, it can detect multiple targets due to its high range resolution [17]. mmWave sensing is also illumination free (*i.e.*, working in the dark) with fair penetration (*e.g.*, see-through glasses and masks) [54], hence being promising for a wide range of applications. Fig. 1 shows a typical scenario for multimedia IoT applications. Moving along this direction, in this paper we investigate several key challenges in designing an effective mmWave radar based FER system.

Facial expression usually triggers a series of facial muscle movements across multiple facial areas (*e.g.*, eyes, forehead, nose, cheek, lip and mouth) [8], representing as critical spatial information in mmWave signals. To acquire spatial information, mmWave radar integrates an antenna array with multiple transmitters (TX) and multiple receivers (RX) to enable MIMO (*i.e.*, multi-input and multi-output) in improving angular resolution [34]. A *de-facto* way is to acquire a set of point clouds generated by mmWave radar [36, 43]. However, due to the limited number of antennas available on an off-the-shelf mmWave radar (*e.g.*, TI IWR1843BOOST mmWave radar has 3 TX and 4 RX resulting up to 12 virtual antennas), the angular resolution is limited to 15-degree in azimuth only [1], resulting in limited number of point clouds. Also, to improve signal-to-noise ratio, point clouds generated from raw mmWave signals are intently merged by built-in algorithms [36], hence yielding notable sparsity. Our preliminary study in Section §2 shows that it is not feasible to detect facial muscle movements using sparse point clouds. Point clouds may be enhanced using advanced mmWave radar with high angular resolution [43]. However, advanced radar is usually bulky and more expensive (*e.g.*, 10x) compared to commercial grade radar and is typically used in high-end automotive. Supervised learning from cross-domain (*e.g.*, co-labeled LiDAR dataset) can also be used to generate dense point clouds [35], but its performance relies heavily on large-scale training datasets, which may be impractical for FER due to lack of mmWave datasets.

Turning away from sparse point clouds, we pay our attention to raw mmWave signals received from multiple antennas which contain rich Doppler information to generate spatial information. An immediate question is how to accurately extract rich spatial information representing subject's facial muscle movements from raw mmWave signals. Raw mmWave signals likely contain body motion information and ambient noise due to background moving objects, diffraction, and complex specular reflections [60]. Also, facial muscle movements are relatively subtle, *e.g.*, 5mm for cheek and 6mm for month by a "happiness" expression of adult [66], making it difficult to be distinguished from ambient noise. Beamforming [18, 59] uses narrow beams which focus on a relatively small area, hence avoiding ambient noise to some extent. However, it may not be able to detect multiple targets at a time, and ends up with much reduced spatial information since all antennas are phased to narrow the beam compromising on angular resolution [1].

Our idea is to convert the problem to a spatial localization problem. We first locate each subject by verifying his/her biometric information (*i.e.*, heart rate and respiration) to eliminate ambient noise of static/dynamic objects. We then essentially extract spatial facial information by filtering out irrelevant body motions. With spatial facial information obtained, we explore correlation between facial muscle movements and features extracted from spatial facial information. Deep Learning (DL) can potentially achieve high accuracy and better robustness due to its automated feature extraction capability [46, 64]. However, deep learning usually requires a large amount of training data to train an effective model, and small-scale training dataset may easily end up with poor performance. In reality, collecting large-scale mmWave dataset can be very costly due to labeling efforts and privacy concerns. Our idea is to leverage on rich image datasets available for FER [21, 29, 40, 62] and apply cross-domain transfer learning [27, 33] that enables model knowledge transformation from the image domain to the mmWave domain.

Our Approach To address the aforementioned challenges, in this paper we propose mmFER, a novel **mm**Wave radar

[1]The theoretical angular resolution in elevation is 58-degree on TI IWR1843BOOST, but it is larger than the 30-degree field of view in elevation, hence it is limited in practice.

based **FER** system that extracts and analyzes subtle facial muscle movements associated with raw mmWave signals to recognize facial expressions for multiple users. mmFER can recognize 7 standard facial expressions [10] as shown in Fig. 8, delivering valuable assessment of users' engagement and reactions to multimedia IoT applications.

To detect subtle facial muscle movements in raw mmWave signals, we propose a novel dual-locating approach based on MIMO to locate both subjects of interest and the facial areas of each subject. Specifically, we first locate subjects of interest (*e.g.*, who stay relatively stationary to watch TV) marked as anchor points in the azimuth detection range (*i.e.*, horizontal plane). This is achieved by sensing biometric information (*i.e.*, heart rate and respiration) of subject, filtering out background noise (*e.g.*, complex specular reflections by appliances and walls) and dynamic objects (*e.g.*, people moving around). We then propose a novel face-matching mechanism based on a Gaussian Mixture Model (GMM) to locate each subject's facial angle range in the elevation detection range (*i.e.*, vertical plane), avoiding body motions. To fully explore the transferability from image to mmWave, we design a novel cross-domain transfer pipeline, namely cross-transfer, using a pre-trained FER image model to train an mmWave model. Especially, we design a hybrid learning loss function that comprehensively fuses a set of loss functions to address the training over-fitting issue due to small-scale mmWave datasets. We also propose an autoencoder based feature alignment mechanism that learns the transition of latent features to progressively reshape complex mmWave data, eliminating the effect of data heterogeneity.

In a nutshell, enabled by dual-locating, mmFER can extract spatial information of facial muscle movements from raw mmWave signals. With cross-transfer, mmFER enables an effective and safe model knowledge transformation for mmWave-based FER. We fully implement mmFER using an off-the-shelf mmWave radar (*i.e.*, TI IWR1843BOOST) and conduct comprehensive evaluations with 10 subjects for a set of facial expressions in real-world settings. Results show that mmFER achieves an accuracy of 84.48% in a subject-to-radar distance between 0.3 and 1.5m and an accuracy of 80.57% when distance increases to 2.5m. In summary, our main contributions are as follows:

- A first-of-its-kind mmWave radar based FER system that detects subtle facial muscle movements associated with raw mmWave signals for multimedia IoT applications.
- A novel dual-locating approach to accurately locate on subjects' faces in space, and extract subtle facial muscle movements from noisy raw signals.
- A novel cross-domain transfer pipeline, *i.e.*, cross-transfer, to enable an effective and safe model knowledge transformation for mmWave-based FER with superior model performance.

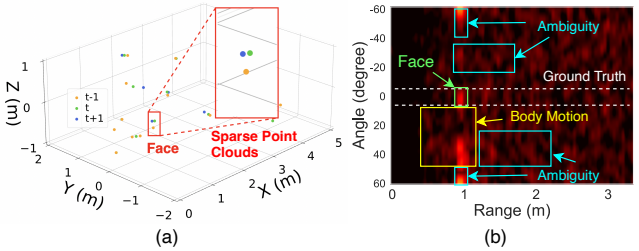

Fig. 2: Preliminary results: (a) Poor detection using highly sparse point clouds; (b) Challenges in noisy raw mmWave signals.

- An off-the-shelf mmWave radar based mmFER implementation with extensive experiments. Results show that mmFER is resilient to various real-world settings with high accuracy and robustness, outperforming the baselines.

Implication mmFER moves an important step towards the promising mmWave-based FER. The proposed mmWave radar approach easily removes privacy concern and illumination constraint, and robustly works in the scenario of wearing various accessories (*e.g.*, mask). It is also superior to Wi-Fi and ultrasonic approaches, providing a higher signal bandwidth with a longer detection range and multi-target capability. While this paper focuses on enabling mmWave-based FER for multimedia IoT applications, mmFER has opened possibilities to be widely applied in different domains such as recommendation systems, healthcare, augmented reality/virtual reality (AR/VR) systems, education, and more, providing valuable user feedback and survey. For instance, mmFER can be used in a wide range of recommendation systems to help sense users' preferences and reactions in a privacy-preserving manner. In healthcare systems, mmFER is able to offer medical professionals timely feedback about the mental state of patients (*e.g.*, with depression). In addition, mmFER can provide a robust way to understand users' attention and intent in an indoor or outdoor environment, improving the user experience of AR/VR systems.

2 PRELIMINARY

Principles of MIMO in mmWave Radar. mmWave radar has a linear RX antenna array which enables MIMO to estimate target's angle of arrival (AoA), hence providing spatial information for detecting multiple facial areas. In principle, since RX antenna array is designed with an interantenna distance (*e.g.*, 2.5mm on TI IWR1843BOOST), the reflected mmWave signals from a target travel different distances to reach each RX antenna, yielding phase difference which can be used to estimate the AoA of signals [34]. Theoretically, to enrich spatial information for detecting multiple facial areas, it is crucial to increase the angular resolution θ_{res} calculated by the number of RX antennas N_i, represented as $\theta_{res} = \frac{\lambda}{N_i d \cos(\bar{\theta})}$, where $\bar{\theta}$ equals 0 for a boresight view, λ denotes the wavelength of chirp, and d represents the interantenna distance. However, commercial mmWave

radars only integrate a limited number of antennas, *e.g.*, TI IWR1843BOOST has a 3x4 antenna array which achieves 12 virtual RX antennas with a 2-dimensional 2x8 antenna layout, yielding a limited 58-degree and 15-degree angular resolution in elevation and azimuth, respectively.

Set-up. To investigate the feasibility of mmWave sensing for FER, we conduct preliminary experiments using a commercial mmWave radar (TI IWR1843BOOST). We set up a watching movie scenario where we use a 27-inch screen and place the radar directly below the screen (1.1m height from the ground). We ask a subject to sit 1m away from the screen and watch movies. The radar configuration is shown in Table 1. We use "surprise" facial expression for this study since it triggers a relatively large range of facial muscle movements. Since the radar has a limited angular resolution especially in elevation (*i.e.*, 58-degree), the typical setup may cause loss of useful spatial information in elevation for detecting facial areas. Hence, we place the radar upright to switch the 15-degree angular resolution from azimuth to elevation.

Sparse Point Clouds. Point clouds represent all motions in space over time. We generate a set of point clouds based on raw mmWave signals using a *de-facto* approach [44]. Fig. 2(a) presents the changes of motion points in three consecutive time frames (*e.g.*, $t-1$ to $t+1$) with an interval of 0.5s. The results show that point clouds are sparse and largely contain irrelevant motions or ambient reflections. The red block marks the location of the subject's face as the ground truth. In particular, the zoom-in figure indicates that mmWave radar can detect the change of point clouds caused by facial muscle movements, but point clouds are highly sparse and only one point is available in each time frame. Thus, it is infeasible to leverage on sparse cloud points to obtain spatial facial information for FER.

Challenges in Raw mmWave Signals. We instead investigate the feasibility of using raw mmWave signals to extract spatial facial information. Given a pre-defined subject's position (*i.e.*, a set of range bins), we apply angle FFT [34] to transform time-domain raw data within a 1s time window to spatial data (*i.e.*, containing range and angle information). To understand the impact of body motion, we ask the subject to perform leg shaking in this experiment. To obtain the ground truth of facial angle range, we use vision-based face detection [63] to locate face in space. According to the ground truth lines shown in Fig. 2(b), it is feasible to locate the spatial facial information in angle-range heatmap (*e.g.*, marking with the green block). However, due to the limited 15-degree angular resolution, it is difficult to precisely locate facial areas and avoid irrelevant spatial information using angle FFT. Moreover, MIMO-enabled spatial data contain massive ambient noise represented as the ambiguity with the blue blocks, and the spatial information of body motion with the yellow block can be observed in the heatmap. Thus, it is challenging to extract subtle spatial information of facial expressions from raw mmWave signals.

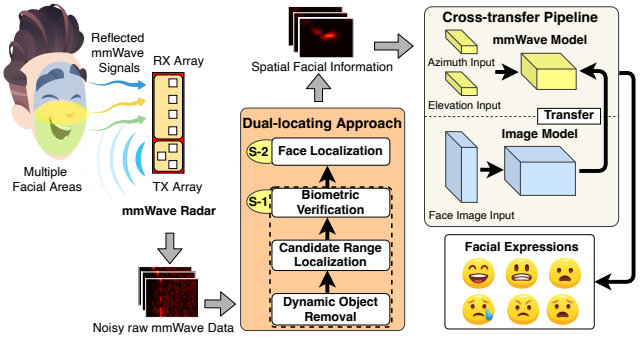

Fig. 3: Overview of mmFER system architecture.

3 mmFER DESIGN

Fig. 3 presents an overview of mmFER system architecture. Due to ambient noise existing in raw mmWave signals, we first transform the problem to spatial face localization, and propose a dual-locating approach to extracting spatial facial information. This approach takes raw mmWave signals as input and outputs the angle (A) range (R) heatmaps of multiple facial areas, as shown in Fig. 4(a). We then propose a cross-transfer pipeline and feed the outputs into an mmWave FER model for classification.

3.1 Dual-locating Approach

Challenge of Massive Ambient Noise. Facial muscle movements caused by facial expressions are subtle in millimeter-level [13]. mmWave radar is capable of detecting millimeter-level movements, but raw mmWave signals received indoors usually contain massive ambient noise due to dynamic and stationary objects, *e.g.*, people moving around, multi-path reflections by home appliances and walls. To further understand the impact of ambient noise, we conduct an experiment in the living room (*e.g.*, 2.9mx4.2m) with the same setup as in our preliminary study. We operate a fan 1.5 m behind the subject, and ask another subject to walk randomly in the room. Fig. 4(b) shows the range profile heatmap generated from raw mmWave signals, and it clearly shows that mmWave signals contain massive ambient noise. To address, based on the MIMO capability on mmWave radar, we propose the dual-locating approach which consists of a two-step process to first locate subjects, then locate the facial areas of each subject, as shown in Fig. 3.

3.1.1 *Subject Localization*

In *Step-1*, the dual-locating approach locates subjects of interest (*i.e.*, who stay stationary to watch TV), marked as anchor points in the azimuth detection range, and eliminates background noise and dynamic objects.

Dynamic Object Removal. Since dynamic objects (*e.g.*, people walking) have speed information, we can use range

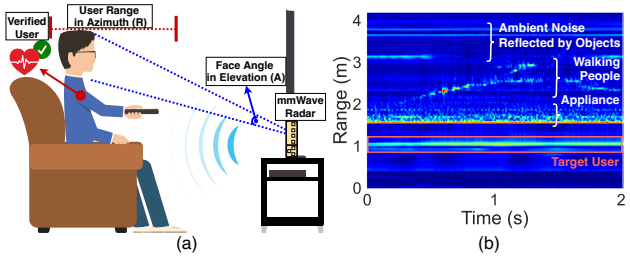

Fig. 4: (a) Face localization illustration; (b) Massive ambient noise in raw mmWave signals.

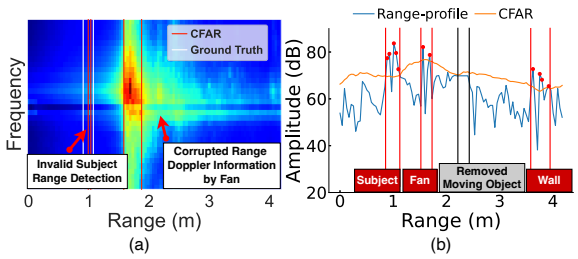

Fig. 5: (a) Invalid subject range detection in range Doppler heatmap; (b) Candidate object selection using adaptive CFAR.

Doppler information to distinguish the difference between stationary and dynamic objects in velocity for range detection [31, 61]. In a simple way, a fixed speed threshold (*e.g.*, 0.11m/s suggested in [31]) can be used to filter out dynamic objects which move faster than this threshold. However, we observe that applying range Doppler information for computing velocity is very sensitive to speed, often resulting in *unsafe* object removal (*i.e.*, subject removed due to his/her body motions). In reality, a subject may have body motions at different speeds, *e.g.*, 0.4-0.6m/s for leg shaking, and 0.9-1.2m/s for waving hands [25]. Due to radar's Doppler widening effect (*i.e.*, the area of Doppler spectrum is widened by the changes of object's speed and range) [22], the Doppler spectrum of body motions may largely overlap with that of facial movements (note that body movement may have a higher speed than facial movement). Due to the overlapping, removing body motion spectrum will result in loss of facial movement spectrum. Hence, the fixed threshold method may lead to subject localization failure. To enable a safe dynamic object removal, we use range profile information directly to estimate moving objects if they trigger range shift. Specifically, we define a time difference function as $RP_{td}(t) = RP_t - RP_{t-n}$, where RP_t is range profile on time t, and n is time slot. We also develop an adaptive velocity threshold formulated as $v_t = \frac{\Delta RES}{n \times F}$, where ΔRES is range resolution (*e.g.*, 4.2cm), F is frame periodicity (*e.g.*, 10ms). In addition, the range of dynamic object is in $[PK_t^i - B, PK_t^i + B]$, where $B = 1/2|PK_{t+1}^i - PK_t^i|$, PK denotes a find peak function based on $RP_{td}(t)$, and i denotes each dynamic object. Once the time difference function detects peaks in RP_{td}, it indicates in presence of moving objects whose velocity is greater than the threshold, and a set of range bins (*i.e.*, object positions) will be computed for accurate dynamic object removal. Conversely, the objects whose velocity is less than the threshold are recognized as stationary to safely protect subjects with body motions. Thus, dual-locating can safely remove dynamic objects and preserve subjects of interest based on the range shift over time.

Candidate Range Localization. After removing dynamic objects, we now locate subjects of interest (*i.e.*, stationary objects) marked by a set of range bins, as candidates for FER. Due to complex indoor reflections, raw mmWave signals contain massive ambient noise which impact candidate range localization. Existing works [31] obtain a resilient noise threshold by the cell-averaging constant false alarm rate (CFAR) algorithm [49] based on range Doppler information. When applied in our case, however, we reveal that range Doppler information may be easily corrupted by ambient motion noise (*e.g.*, regular vibration caused by fan or air conditioner) due to the Doppler widening effect. Fig. 5(a) shows that the range frequency caused by a running fan is largely widened in the range Doppler heatmap, and such enlarged noise frequency notably increases the noise threshold of CFAR. With the increased noise threshold, the selected range bins of subject using CFAR in red lines is reduced by 60% compared to the ground truth in white lines. Reduced spatial information may easily result in invalid removal of facial areas of subjects. Alternatively, we leverage the range profiles generated by the first FFT to detect the presence of objects with higher amplitudes than the noisy background. To accurately detect subject range (*i.e.*, no range bin reduction) and remove ambient noise, we propose an adaptive CFAR with a body knowledge based Gaussian distribution. In our work, accurately selecting subject's range (*i.e.*, covering subject's body) is the key to obtain spatial facial information. Our basic idea is to build a Gaussian distribution $G(\mu, \sigma)$ using the reported body shape statistics [20] to improve CFAR with a robust noise threshold function. Specifically, we define range profile RP as input of $C(\cdot)$ (*i.e.*, CFAR), and the output of CFAR denotes a set of range bins formulated as $C(RP_i) = RP_i - \hat{T}_i$, where $\hat{T}_i$ is an adaptive resilient noise threshold calculated from Eq. 1.

$$\hat{T}_i = \frac{1}{2(NC_n - NC_g)} \sum_{i+NC_n}^{i-NC_n} RP_i - \sum_{i+NC_g}^{i-NC_g} RP_i, NC_n > NC_g \quad (1)$$

where NC_n denotes number of noise cells. NC_g denotes number of guard cells formulated as $NC_g = \frac{1}{2}(\frac{\mu+3\sigma}{\Delta RES})$, where μ and σ obtained from $G(\mu, \sigma)$, and ΔRES is range resolution (*e.g.*, 4.2cm). Fig. 5(b) demonstrates that subject's range bins can be successfully located without information loss, and most of the ambient noise can be effectively removed using the adaptive CFAR.

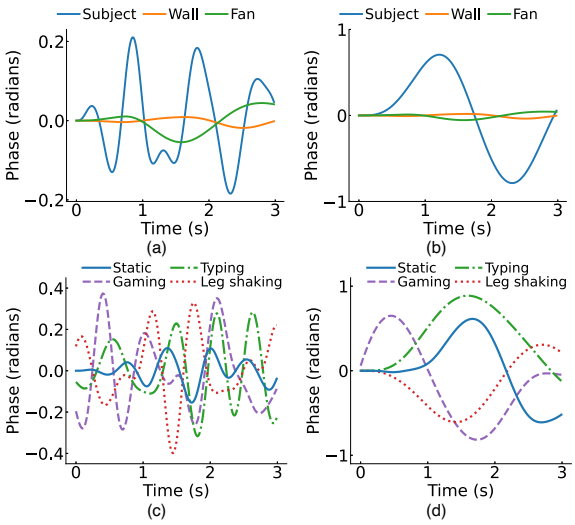

Fig. 6: (a) Heartbeat detection; (b) Respiration detection; (c) Heartbeat detection with body motions; (d) Respiration detection with body motions.

Biometric Verification. After candidate range localization, a number of stationary objects may be selected as candidates which may still contain ambient noise (*e.g.*, fan and wall) as shown in Fig. 5(b). To further validate subjects of interest in a list of candidates, we essentially leverage on biometric information (*i.e.*, heartbeat or respiration) to effectively distinguish human from other objects. Given a range bin of each candidate, we first access phase change defined as $\phi\,(bin, t)$ in mmWave signals over time. We then apply the 4th-order Butterworth band-pass filter [14] to extract heartbeat and breath signals from the phase change. To compare with the ground truth reported in existing vital sign detection works [38], we calculate chest displacement ΔR (mm) with phase change $\Delta\phi$ using $\Delta\phi = \frac{4\pi}{\lambda}\Delta R$, where λ is the mmWave wavelength. We can then verify the results with the chest displacement metrics (*e.g.*, 0.1-0.5mm for heartbeat and 1-12mm for respiration [38]). As shown in Fig. 6(a) and Fig. 6(b), the results of both fan (0.054mm for heartbeat and 0.109mm for respiration) and wall (0.018mm for heartbeat and 0.073mm for respiration) are invalid compared to the displacement metrics, while the subject is successfully verified by accessing both heartbeat (0.15mm) and respiration (1.22mm). Besides, since body motions may remain at this step, we test the impact of these motions (*e.g.*, playing games, typing, and leg shaking) on heartbeat and respiration detection. The results from Fig. 6(c) and 6(d) indicate that the phase change of both heartbeat and respiration can be accurately detected with no impact, *e.g.*, the verified displacement of 0.36mm on average for heartbeat with motions and 1.54mm on average for respiration with motions. After biometric verification, dual-locating is able to accurately locate subjects of interest with correct bins in azimuth.

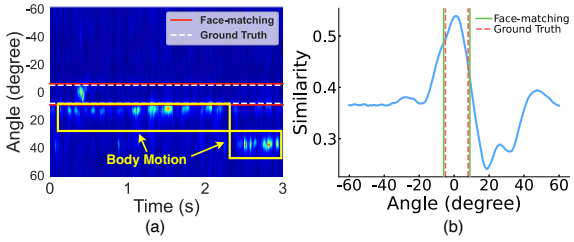

Fig. 7: Face-matching performance: (a) Face localization avoiding body motion; (b) Face matching similarity in elevation.

3.1.2 *Face Localization*

In *Step-2*, given a subject's location in azimuth (*i.e.*, anchor point), dual-locating locates subject's face with a specific angle range in the elevation detection range while avoiding irrelevant body motions. As aforementioned, due to the limited angular resolution of 15-degree on mmWave radar [36], it is difficult to locate small facial areas by angle range. To achieve high-resolution AoA estimation (*e.g.*, theoretically 1-degree in azimuth), we adopt CAPON [5] to generate enhanced angle range heatmaps. With enhanced angular information, we propose a face-matching mechanism that uses a greedy search based on GMM [48] for matching similarity of facial features in elevation. Once facial features are matched with angle range, dual-locating extracts the facial areas in the enhanced angle range heatmap (*i.e.*, spatial information) for the next process.

Face-matching Mechanism. Due to the advanced capability of approximating spectral features using GMM [31], we train a GMM model on pre-processed mmWave dataset x_{pre} with manual labels for generalization performance. Since the spectral features of facial muscle movements represent as a set of angle range heatmaps in a time slot (*e.g.*, 3s) with rich information in time-domain, we transform the angle range heatmaps to angle time heatmaps by collapsing the range information to enhance the features. We then define that x_θ as input data of GMM is obtained from the angle time heatmap on θ-degree. To measure the similarity between input x_θ and output by GMM, we use cosine similarity [42] as the loss function. Hence, our objective is to locate facial angle range θ by maximizing the similarity between the template distribution $GMM(x_{pre})$ and the searched distribution $GMM(x_\theta)$, as formulated in Eq. 2.

$$\theta_{max} = \arg \max_{(\theta,\psi)\in FOV} (loss(GMM(x_\theta), GMM(x_{pre}))) \quad (2)$$

where ψ denotes an experimental offset for angle range compensation, and θ_{max} is the optimized facial angle range. In particular, FOV is a 120-degree search grid in space. In short, Fig. 7(b) presents the greedy search process by maximizing similarity, and face-matching can locate subject's face with maximum similarity along with related angle range (in both azimuth and elevation). Fig. 7(a) shows that face-matching can accurately locate face with angle range, and effectively

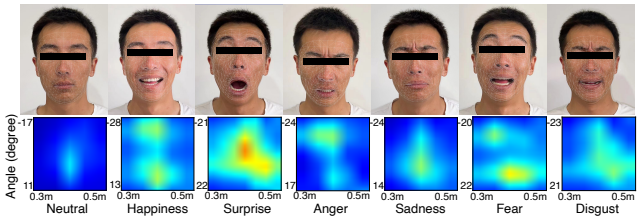

Fig. 8: Comparison between the facial expressions with landmarks and the located facial angle ranges in elevation.

avoid body motions (*i.e.*, playing games and leg shaking). In a nutshell, Fig. 8 demonstrates the outputs of dual-locating, comparing to different facial expressions.

3.2 Cross-transfer Pipeline

Challenges of Small-scale Dataset. Collecting a large-scale mmWave dataset to train a FER model can be very costly in reality due to high labor cost and user privacy, while small-scale training dataset may lead to poor performance. To fully understand this issue, we collect both mmWave and video datasets with 10 subjects (*e.g.*, 40 samples per expression per subject) in a watching-iPad scenario, as shown in Fig. 12(b). To obtain the ground truth, we use a pre-trained ResNet18 (*i.e.*, image model) [3], and fine tune it on the collected video dataset (face images cropped by [63]). We employ a conventional supervised approach with the manual truth labels (*i.e.*, T-label) to train an mmWave model after the dual-locating process. Fig. 9(a) reveals that the training accuracy of T-label severely degrades (*i.e.*, 45.71%) compared to the ground truth (*i.e.*, 95.28%). The underlying impediment is that T-label learns insufficient latent features based on a much limited data distribution provided by the small-scale training dataset. The key idea of cross-domain transfer learning [27, 33] is to transform domain-invariant latent features learned from a well-developed source domain (*e.g.*, image model) to a limited target domain (*e.g.*, mmWave model), enhancing latent feature extraction. Following this idea, we apply a Knowledge Distillation (KD) based cross-domain approach [56] (*i.e.*, teacher-student learning scheme) to both mmWave and video datasets. We use the image model as a teacher and the mmWave model as a student. As shown in Fig. 9(a), the KD based approach achieves a higher training accuracy (*i.e.*, 74.38%). However, we reveal that the KD based approach suffers from suboptimal training performance, yielding a critical accuracy gap between KD and the ground truth. The following two issues remain as impediments.

Training Over-fitting: Fig. 9(b) shows that while KD train loss descent performs well, KD test loss notably fails to decrease. The result demonstrates the training over-fitting issue (*i.e.*, unsafe training) that the model is too closely aligned to the small-scale training dataset, hence resulting poor generalization on testing dataset. To further understand the issue, we employ the t-distributed stochastic neighbor embedding

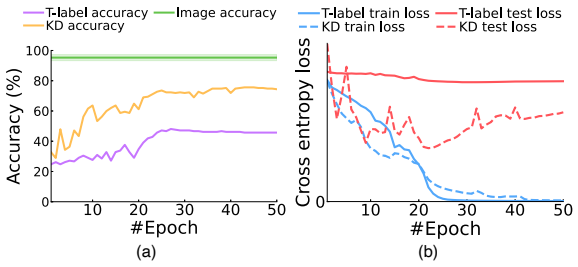

Fig. 9: Training performance comparison on small-scale mmWave dataset: (a) Training accuracy; (b) Loss descent.

(t-SNE) (*i.e.*, clustering analysis for loss optimization) based on the Euclidean distance (ED) function to visualize the distribution of the latent features for six expressions using both image and KD based mmWave models. As shown in Fig. 10(a) (*e.g.*, each color represents an expression), the feature distribution using a KD based mmWave model is messy and scattered compared to the ground truth (*i.e.*, clustered nicely using the image model), hence it implies that the underlying loss function of KD may perform poorly based on small-scale training dataset.

Domain Shift: Based on transferability metrics (*i.e.*, distance-based similarity analysis between domains) [28], we observe that there is a notable domain shift (*i.e.*, lack of similarity) between the feature distributions using both models shown in Fig. 10(a). Due to high data heterogeneity (*i.e.*, different raw data structures and shapes) between image and mmWave, model performance may easily degrade if the latent features between both domains are not well aligned (*i.e.*, domain shift). A keypoint-based approach has been proposed in [51, 65] to extract relevant key points (*i.e.*, less data heterogeneity) from images to achieve image-to-sensing transformation. To validate its effectiveness, we employ Mediapipe [37] to generate facial landmarks from images and use a keypoint-based approach to train an mmWave model, but the results reveal that this approach still suffers from training over-fitting caused by the small-scale dataset, severely degrading model accuracy (*i.e.*, 62.38%) shown in Fig. 10(b).

We hence propose a novel cross-transfer pipeline to enable effective and safe model knowledge transformation from image to mmWave. Specifically, to address training over-fitting, we design a hybrid learning loss function that comprehensively fuses a set of loss functions including a supervised loss by mmWave labels, a Kullback–Leibler (KL) divergence loss inspired by KD, and a contrastive loss based on *positive-negative correlation* to achieve superior model performance. To eliminate the impact of data heterogeneity, we also develop an autoencoder based feature alignment mechanism that learns the transition of latent features to progressively help align both domains, thus improving transferability.

3.2.1 ***Training Scheme Design*** Fig. 11 presents the detailed training scheme design of the cross-transfer pipeline.

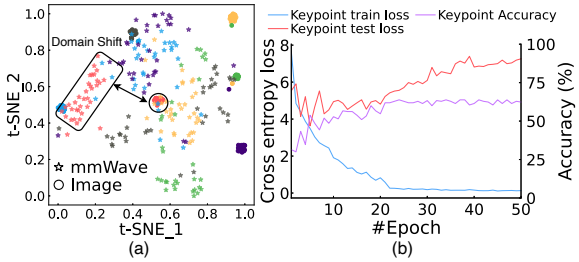

Fig. 10: (a) t-SNE result using KD; (b) Training over-fitting.

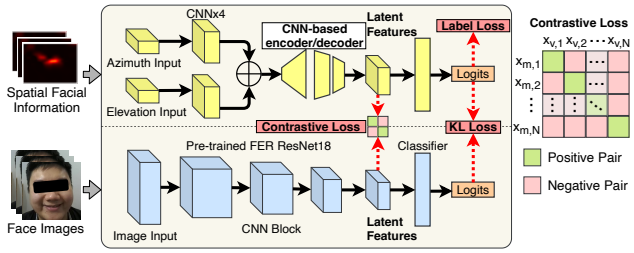

Fig. 11: Cross-transfer pipeline.

We keep the pre-trained ResNet18 as the image model, and fine-tune it on the collected face images for learning the latent features. We next freeze it (*i.e.*, model inference only) as a learning anchor point throughout the process of cross-transfer pipeline. Once mmWave data is processed by dual-locating, all of the spatial facial information will be used as the input of the mmWave model.

3.2.2 ***Feature Alignment*** Since mmWave data input consists of both azimuth and elevation heatmaps with different data shapes, simply combining them may result in loss of key features. Hence, we design two separated blocks with four convolutional neural network (CNN) layers for each to reshape and explore the key features, avoiding information loss. After this, both outputs of the heatmaps can be safely summed to prepare for the next process, as shown in Fig. 11.

To align the latent features with different data structures for image to mmWave, we design a CNN-based autoencoder to essentially learn from the transition of both latent feature distributions during training. In theory, autoencoder [2] is advocated to fully explore the latent information to improve training performance. In mmFER, due to data heterogeneity, the transition of feature distributions between both domains can be very complex, therefore fully exploring useful latent features is the key. We intuitively design an autoencoder to not only reshape the dimensionality of latent features (*i.e.*, data shape alignment), but also enhance transferability to contribute in loss optimization for training performance improvement. Since we set the image model as the learning anchor point, we place the proposed autoencoder in an mmWave model as a *middleware* to reshape the latent features from the mmWave model, aligning with the feature shape in the image model presented in Fig. 11.

3.2.3 ***Hybrid Learning Loss*** As aforementioned, existing cross-domain approaches may suffer from training over-fitting due to small-scale mmWave dataset. We reveal that the underlying impediment is due to an ineffective design of loss function. To address, we propose a hybrid learning loss function to improve the loss optimization for image to mmWave. Intuitively, according to the principle of loss optimization [26], if the feature distribution of each class is more tightly clustered (*i.e.*, smaller intra-class distance) and the distance between the feature distribution of each class pair is farther apart (*i.e.*, larger inter-class distance), it will result in much better training performance (*i.e.*, high accuracy and generalization). Inspired by this principle, we design our hybrid learning loss function with three key elements.

Cross Entropy Loss. Based on the T-label approach with manual truth labels, we define that cross entropy loss L_{gt} is calculated with mmWave data x_m and its ground truth label y_m using Eq. 3.

$$L_{gt} = H(y_m, x_m) = -\sum_{x \in X} y_m(x) \log(p(x)) \tag{3}$$

where H is the cross entropy loss function, $y_m(x)$ and $p(x)$ denote related probability computed by an mmWave model.

KL Loss. Referring to accuracy improvement using KD in Fig. 9(a), we utilize KL divergence as one of key metrics to measure the difference between the distribution of logits (*i.e.*, l_v and l_m) computed by both image and mmWave models in Fig. 3.2, hence the KL loss L_{kl} is formulated in Eq. 4.

$$L_{kl} = KL(l_m || l_v) = \sum_{x \in X} l_m(x) \log(\frac{l_m(x)}{l_v(x)}) \tag{4}$$

Contrastive Loss. Based on the principle of loss optimization, we design contrastive loss $L_{contrast}$ to essentially optimize two distance-based metrics in cross-domain. Our objective is to first *minimize* the distance between the latent feature distributions that have the same label (*i.e.*, intra-class distance) marked as *positive pair*, and then *maximize* the distance between the feature distributions that have different labels (*i.e.*, inter-class distance) marked as *negative pair*, in training image to mmWave. We visualize the *positive-negative correlation* as a 2-color matrix in Fig. 11, in which the green diagonal line denotes all positive pairs that each $x_{m,N}$ in the mmWave domain has the same label (*e.g.*, "happiness") with each $x_{v,N}$ in the image domain, where N denotes the number of classes, and the rest in the matrix denotes all negative pairs. Hence, we formulate $L_{contrast}$ in Eq. 5.

$$\begin{aligned} L_{contrast} =& L(x_m, x_v^+, \{x_{v,i}^-\}_{i=1}^{N-1}; f_m, f_v) \\ =& log(1 + \sum_{i=1}^{N-1} \sigma_i exp(f_m^T f_{v,i}^- - f_m^T f_v^+)) \end{aligned} \tag{5}$$

where x_v^+ and x_v^- denote the positive and negative inputs in an image model, respectively. Also, f_m, $f_{v,i}^-$, and f_v^+ denote the sets of latent features computed by input x_m, $x_{v,i}^-$, and

x_v^+, respectively. In particular, we design a regularizor σ_i as $S(D(\{x_v^+\}, \{x_{v,i}^-\}_{i=1}^{N-1})^{-1})$ to further contribute optimizing the negative pairs, where D denotes a standard ED function, and S denotes a default SoftMax function.

Combined with all elements, our hybrid learning loss function is formulated as $L_{hybrid} = L_{gt} + L_{kl} + L_{contrast}$. In short, cross-transfer pipeline enables an effective and safe model knowledge transformation for mmWave-based FER.

4 EVALUATION

4.1 System Implementation

Fig. 12(a) gives an overview of mmFER system setup. We implement mmFER using an off-the-shelf mmWave radar kit which consists of a TI IWR1843BOOST sensor board operating at 77-81GHz ($299) and a TI DCA1000EVM data capture board ($599). The IWR1843BOOST board has 58-degree and 15-degree angular resolution in elevation and azimuth, respectively. Thus, we place it upright to configure a 15-degree angular resolution in elevation for face localization. We use TI mmWave studio version 02.01.01 for radar processing and PyTorch version 1.4.0 for model training on a Windows PC with an AMD Ryzen 7 3700X CPU and a NVIDIA GeForce RTX 3090 GPU. Table 1 shows our radar configuration in detail. In particular, since we fully enable 3x4 MIMO on the IWR1843BOOST board, the chirps per frame generated are limited up to 32 [24]. Based on the default radar configuration in TI mmWave studio, we use the frame periodicity of 10ms (*i.e.*, frame rate of 100Hz). Since the frame periodicity limits the ramping time, we set the frequency slop of 100.0 MHz/us to maximize the radar bandwidth of 3.6GHz. In addition, we apply a coordinate calibration method to align both mmWave and image data configuration (*e.g.*, timeline and displacement between hardware).

In real-world deployment, we use the first 3s radar data for dual-locating, hence no additional privacy concern is raised by biometric verification. Our prototype is powered using a 5V power jack and connected to the PC for real-time processing. Specifically, the proposed face localization takes 400ms on average for processing including range localization of 180ms, biometric verification of 10ms, and face-matching of 210ms. We employ a 500ms sliding window with a 3s time window to feed data to the mmWave FER model, and the latency of each model inference is 4.5ms or 13ms on average in the GPU- or CPU-enabled mode, respectively. Besides, we employ the psutil version 5.9.0 to track resource usage. In watching a 10mins video, the CPU utilization of dual-locating is 6.2% on average, and the memory footprint is up to 368MB (the usage may vary on different platforms). For model inference, the CPU utilization is 49.8% on average and the memory footprint is up to 610MB in the CPU-enabled mode, while the GPU utilization is 44.5% on average and the peak memory footprint reaches 3890MB in the GPU-enabled mode. In practice, our prototype can be placed on top/bottom of or next to the multimedia device, as shown in Fig. 12(b)-(f). In addition, our prototype can be integrated with different embedded platforms (*e.g.*, Raspberry Pi or NVIDIA Jetson series) for portable deployment.

Table 1: mmWave radar configuration.

Frequency Slop	100.0 MHz/us	**Range Resolution**	4.2cm
ADC Samples/Second	7200K	**Idle Time**	7 μs
Chirp Cycle Time	46 μs	**Chirps/Frame**	32
Frame Periodicity	10.0ms	**Samples/Chirp**	256

Table 2: Devices with recommended viewing distances.

Device	Screen Size	Distance
iPad Pro 2021	11 inch	0.3 m
Lenovo ThinkBook Gen2	15.6 inch	0.7 m
Samsung CF390 Monitor	27 inch	1 m
LG C1 4K TV	55 inch	1-3 m

4.2 Evaluation Setup

We use a range of multimedia IoT devices which provide different viewing distances, as shown in Table 2. We set up various scenarios in a living room (2.9mx4.2m) of a residential apartment, as shown in Fig. 12(b)-(f). We recruit 10 subjects (3 females and 7 males, age ranging from 23 to 32, height ranging from 156cm to 189cm). We collect both mmWave and video datasets with 7 standard facial expressions ("neutral", "happiness", "sadness", "surprise", "fear", "disgust", and "anger") defined in [10] over a period of three weeks. A depth camera (*i.e.*, Zed 2 Camera operates at 30fps) is used to collect videos. Data collection has been approved by the Human Research Ethics Committee of our institute. During data collection, we ask subjects to perform different postures (*e.g.*, sit on sofa, sit on the ground, and stand) and body motions (*e.g.*, playing games, typing, and leg shaking) at a subject-to-radar distance ranging from 0.3m to 3m. We do not require subjects' faces strictly to the radar as this may affect their facial expressions. Subjects perform these activities in a natural way, hence a slight variation (*e.g.*, less than 10-degree) may be possible. As aforementioned, we use the pre-trained ResNet18 based on the FER Plus expression dataset [3], and fine tune it on our video dataset collected from the camera as the ground truth.

4.3 Dual-locating Performance

We first evaluate the performance of the dual-locating approach and its impact on face localization. We employ vision-based face detection in [63] to locate face and output angle ranges as the ground truth. We use the root mean squared error (RMSE) divided by the vision angle range as metric of face localization error drift, defined as $drift = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(C_i^{dl} - C_i^{v})^2}}{AR^v}$, where C_i^{dl} and C_i^{v} denote the center of facial angle range

Fig. 12: (a) mmFER implementation and system setup; (b)-(f) mmFER use scenarios.

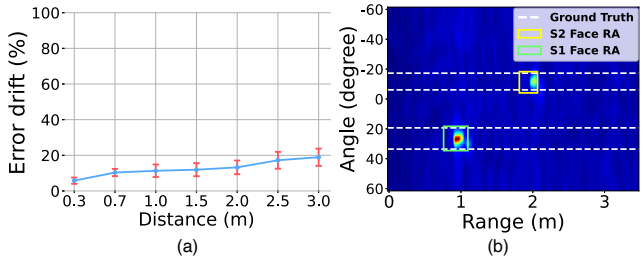

Fig. 13: Face localization performance: (a) Error drift by distances; (b) Multi-target face localization.

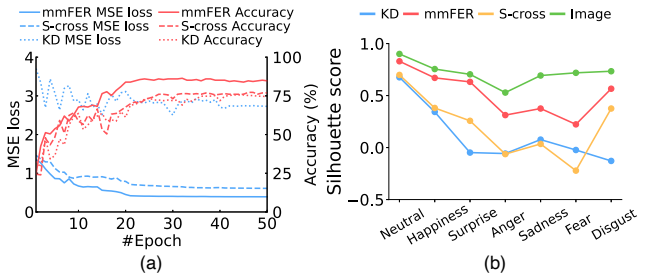

Fig. 14: (a) Training comparison; (b) Transferability performance.

located by dual-locating and vision, respectively, and AR^v denotes the ground truth angle range.

Impact of Different Distances. Facial angle range may inversely vary as the subject-to-radar distance increases, *e.g.*, a far distance results in a relatively small facial angle range. In this experiment, we evaluate the impact of different distances on face localization in dual-locating. We ask three subjects to sit at seven different viewing distances from a 27-inch screen (*i.e.*, the same experimental setup as in Fig. 12(d)) in turn and perform facial expressions following a recorded video. Fig. 13(a) demonstrates that the distance increase yields a slight increase of error drift on face localization, *e.g.*, the error drift rate is settled in the range from 5.7% at 0.3m to 18.9% at 3m. We also observe a little increase of error drift as distance increases. This may be due to a small angle range offset added in the process of face-matching. Moreover, due to physical variation, the standard deviation of error drift (*e.g.*, red error bar) has a minor increase as distance increases (*e.g.*, from 1.7% at 0.3m to 4.8% at 3m). Thus, the results indicate that dual-locating can effectively enable face localization at different subject-to-radar distances with minor error drift.

Multi-target Detection. We next evaluate the performance of dual-locating for locating multiple faces. We set up a scenario of watching TV with two subjects. Due to the upright radar setup, the azimuth FOV is limited, hence we ask one subject (S1) to sit on the ground 1m away and another subject (S2) to stand 2m away, as shown in Fig. 12(f). Fig. 13(b) shows that both subjects' faces can be successfully located in angle range heatmap when performing "surprise" facial expressions simultaneously. The results show that dual-locating can locate multiple subjects with accurate range bins and facial angle ranges marked by highlighted blocks. Also, compared to the ground truth of facial angle range, dual-locating achieves low error drift for both subjects, *e.g.*, 3.3% for S1 and 8.3% for S2. In addition, the visual results present that the heat face area (*i.e.*, spatial facial information) of S2 is relatively smaller than that of S1 due to the impact of subject-to-radar distance. Thus, dual-locating can locate face accurately for multiple targets.

4.4 Cross-transfer Performance

We now evaluate the performance of the proposed cross-transfer pipeline with the respect to performance of feature alignment and loss optimization compared to the state-of-the-arts. For a fair comparison, we select both small-scale image and mmWave datasets in a scenario of watching iPad to avoid the impact of viewing distance, as shown in Table 2. We collect 40 samples per expression per subject, and each sample is recorded in a 3s time window [7] by both camera and mmWave radar.

Baselines. We select the following cross-domain learning approaches as our baselines. All of approaches share with the same pre-trained ResNet18.

- **KD** [56] is a conventional cross-domain learning approach that uses knowledge distillation techniques without latent feature alignment support.
- **Keypoint** [51] is a facial landmark based image-to-sensing transformation approach. It uses Mediapipe [37] to generate facial landmarks.
- **S-cross** [65] is a standard unsupervised cross learning approach that leverages on a distance-based loss function to train a student model. It uses a vanilla autoencoder for latent feature alignment.
- **Image** [3] uses the fine-tuned image model as the ground truth.

Feature Alignment Performance. To quantify the effectiveness of the proposed feature alignment mechanism in the

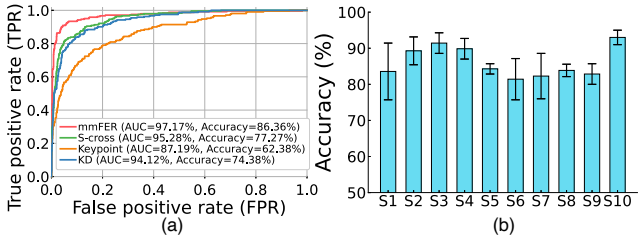

Fig. 15: (a) Training ROC curve; (b) FER accuracy by subjects.

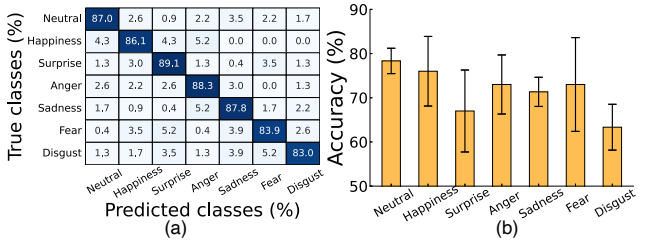

Fig. 16: (a) FER confusion matrix; (b) User-independent model performance.

cross-transfer pipeline, we first employ the mean squared error (MSE) as metric to measure the ED loss descent (*i.e.*, distance-based similarity) between the latent features of image and mmWave models. As shown in Fig. 14(a), mmFER remains the lowest MSE loss during training and achieves 0.39 at 50-epoch, outperforming both KD (2.73) and S-cross (0.61). Since we essentially design two separated CNN blocks to safely explore key latent features without information loss in the proposed mechanism, mmFER successfully achieves a lower MSE loss than S-cross that uses a vanilla autoencoder. The results demonstrate that mmFER outperforms both KD and S-cross in achieving the highest model accuracy of 86.36%, indicating that a lower MSE loss enabled by the cross-transfer pipeline leads to higher model accuracy.

Loss Optimization Performance. In this experiment, we compare the performance of the proposed hybrid learning loss function with the baselines. We follow a widely-used clustering benchmark [47], and employ the silhouette score (*i.e.*, ranging from -1 to 1 and being better if closing to 1) measured by both intra-class and inter-class distances [2]. We also compute the silhouette score of Image as the ground truth. Fig. 14(b) show that, mmFER achieves the best silhouette score of 0.52 on average, outperforming both KD (0.12) and S-cross (0.21). In particular, mmFER achieves the highest silhouette scores across all expressions over both baselines. Besides, the silhouette scores achieved by mmFER are closer to the ground truth, implying that the proposed hybrid learning loss can reduce intra-class distance and enlarge inter-class distance for improving loss optimization in training image to mmWave. Moreover, we use the receiver operating characteristic (ROC) curve to compare the training performance of mmFER with the baselines. Fig. 15(a) shows that mmFER achieves the highest AUC (*i.e.*, area under the curve) of 97.17 and the best training model accuracy of 86.36% on average, outperforming the baselines. This experiment demonstrates that cross-transfer pipeline can notably enhance model performance and transferability in cross-domain.

4.5 mmFER Performance

We now evaluate the overall performance of mmFER from different aspects.

[2]The silhouette score will be closer to 1 if intra-class distance is smaller while inter-class distance is larger.

FER Performance. We first evaluate the FER accuracy of mmFER in this experiment. We train an mmWave FER model based on all subjects' dataset with user-dependent setting. We randomly split the dataset to 80% for training and 20% for testing by 3 rounds and train a model per round. Fig. 15(b) shows that mmFER achieves a FER accuracy of 86.19% on average. The accuracy for each subject falls into the range between 81.43% to 93.00% with a small standard deviation of 3.74% on average. We also calculate the confusion matrix for each facial expression. Fig. 16(a) shows that the FER accuracy of mmFER for all expressions is settled in the range between 83.0% to 89.1% with a small standard deviation of 2.10%. In particular, "surprise" expression achieves the highest accuracy among all expressions. The experiment demonstrates that mmFER is effective to recognize different subjects' facial expressions with high accuracy.

User-independent Model. The physical variations among different subjects may challenge the generalization performance of mmWave FER model especially for unseen subjects. In this experiment, we evaluate the generalization of mmFER in user-independent condition. To quantify variations, we first select three subjects and visualize their facial mmWave data distribution by each expression using t-SNE shown in Fig. 17. The results demonstrate that different subjects lead to notable data pattern variations even for the same expression. We also apply the leave-one-subject-out validation (*i.e.*, leaving one subject independent as unseen and training model based on the remaining subjects' dataset). We train 10 user-independent models and test using 10-fold cross-validation. Fig. 16(b) shows that the FER accuracy by expressions drops, but it still remains at 71.70% on average with a standard deviation of 6.54%. In particular, the accuracy of "surprise" expression is decreased by 22.1%, implying that the variations may be enlarged due to the intrinsic large range of facial muscle movements (*e.g.*, data distribution is more scattered shown in Fig. 17). This experiment demonstrates that mmFER has good generalization performance to recognize unseen subjects' facial expressions.

Impact of Detection Distance. The subject-to-radar distance may impact on spatial facial information extraction (*i.e.*, a far distance may yield a small heat face area shown in Fig. 13(b)). In this experiment, we evaluate the impact of

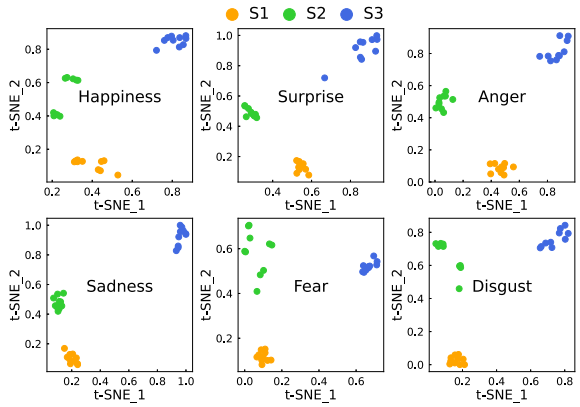

Fig. 17: Notable variations among different subjects.

distances on FER accuracy. We use the same experimental setup as in Section §4.3 and test mmFER in a range between 0.3m to 3m to cover use scenarios in Table 2. Fig. 18(b) plots that the FER accuracy decreases from 86.18% at 0.3m to 64.5% at 3m on average with a standard deviation of 3.65%. Since the angular resolution of mmFER is limited and the actual facial angle range becomes smaller as distance increases, the spatial facial information extracted from mmWave signals may be reduced, resulting in accuracy drop. However, the results prove that mmFER can still effectively achieve a good FER accuracy of 80.57% on average within 2.5m, which can be applied in a wide range of multimedia IoT applications.

Impact of Face Orientation. In this experiment, we evaluate the impact of face orientation (*i.e.*, the angle of head towards the radar) on FER accuracy as subject-to-radar distance increases. We set up 4 different angles of head in azimuth: 0-degree (*i.e.*, facing directly to the radar), 30-degree, 60-degree, and 90-degree. Subjects may turn their faces left or right randomly. We compare the performance of mmFER with the image model. Fig. 18(a) plots the FER accuracy heatmaps with different orientations and distances using image and mmWave models. We observe that the accuracy of the image model at 0-degree is decreased from 96.64% to 83.3% as distance increases, and mmFER achieves a similar result. The results also reveal that the accuracy drops for both models as angle of head increases at each distance, *e.g.*, down to 53.80% (image) and 46.94% (mmWave) from 0-degree to 30-degree. In particular, both models fail when angle of head is larger than 30-degree.

Impact of Body Motions. In this experiment, we evaluate the impact of body motions on performance. We consistently ask the same three subjects in Fig. 17 to perform different body motions (*e.g.*, playing games, typing, and leg shaking) in turn when watching videos (*e.g.*, 10 rounds each subject). We use the same experimental setup as in Fig. 12(d). Fig. 19(a) shows that mmFER achieves an accuracy of 82.2% on average when performing these body motions. In particular,

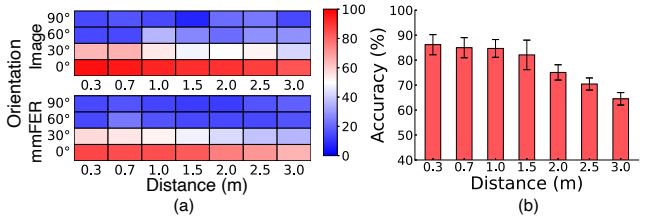

Fig. 18: (a) Face orientation comparison by distances; (b) Impact of different distances.

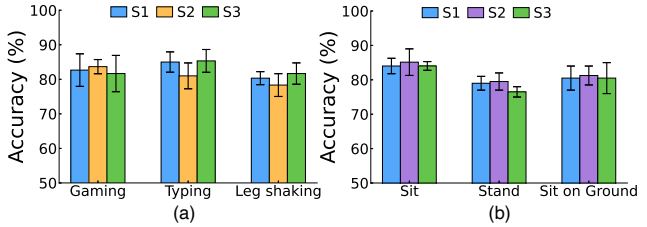

Fig. 19: Impact: (a) Body motions; (b) Different postures.

the result of leg shaking has a slightly lower accuracy of 80.06% compared to that of playing games and typing, implying that a larger range of body motions has more impact on face localization. Overall, the results prove that mmFER can avoid body motions to ensure system robustness.

Impact of Different Postures. In this experiment, we evaluate the impact of different postures on FER accuracy. We use the same experimental setup as above, and ask the three subjects to perform three regular body postures (*i.e.*, sit on chair, sit on ground, and stand). Fig. 19(b) shows the accuracy with three postures for three subjects. The result shows that mmFER achieves an accuracy of 81.15% on average with a standard deviation of 2.59%. We observe that the accuracy of both Stand and Sit on Ground drops slightly to 78.33% and 80.75%, respectively. This implies that spatial facial information extracted from mmWave signals may be affected by different postures due to the change of face orientation in elevation. Overall, this experiment demonstrates that mmFER is robust and resilient to a range of postures.

Impact of Wearable Accessories. Since wearing accessories on head or face is common in reality, different accessories may impact on FER performance due to materials (*e.g.*, plastic or metal frames) and occlusions (*e.g.*, partial or full face occlusions). We then investigate the impact of wearing accessories. We use the setup in a scenario of watching iPad in Table 2, and ask the subject to wear regular accessories in five different ways (*e.g.*, mask, glasses, cap, 3D glasses, and 3D glasses & mask) when reacting videos for 10 rounds. Fig. 20 plots that mmFER successfully achieves a FER accuracy of 80.47% on average with a standard deviation of 2.8% by wearing accessories. Especially, wearing a cap has less impact with an accuracy of 83.86% on average, while the accuracy of wearing glasses (81.52% on average) and 3D glasses (80%

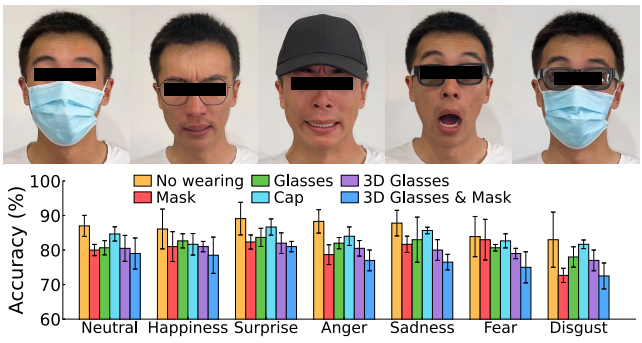

Fig. 20: Impact of regular wearable accessories.

on average) slightly drops as the frame materials may affect the reflection of mmWave signals. In addition, wearing mask yields an accuracy decrease, but mmFER still achieves an accuracy of 80% on average due to fair penetration. Even for wearing 3D glasses & mask, mmFER achieves a good accuracy of 77.07% on average. It implies that mmFER can work in a scenario with wearing mask regulations. Thus, the results indicate that mmFER is robust for wearing accessories.

5 DISCUSSION AND FUTURE WORK

Limited Field of View. As aforementioned in Section §2, due to hardware limitation, we place the IWR1843BOOST board upright, resulting in 58-degree of azimuth angular resolution and 30-degree of FOV. This setup may limit multi-target detection in azimuth in reality. Since the dual-locating approach is designed to locate multiple targets in azimuth for any angle resolution and FOV, this situation can be improved by using an mmWave radar with wider FOV or deploying multiple radars. We plan to further evaluate mmFER with the most recent mmWave radars from TI, *i.e.*, TI MMWCAS-RF-EVM (imaging radar equipped cascaded mmWave sensors) with 140-degree FOV.

Detection Range and Angular Resolution. In principle, the detection range of mmFER is subject to angular resolution which is limited by the number of RX antennas. Although we apply high-resolution AoA estimation to enhance angular resolution, the accuracy of mmFER may still drop when detection range increases (see Fig. 18). However, the proposed technical approach works with any mmWave radar device. To fundamentally improve angular resolution for a longer detection range, we plan to use TI MMWCAS-RF-EVM mmWave radar that provides an angular resolution of 1.4-degree by 86 virtual RX antennas in azimuth and 18-degree by 4 virtual RX antennas, and further evaluate mmFER.

mmWave Dataset Diversity. Our experiments show that cross-transfer pipeline significantly improves the performance of mmFER based on small-scale mmWave datasets. According to Fig. 17, we observe that different subjects may generate a large data distribution shift caused by individual face and expression variation, hence the performance of mmFER may vary with the diversity and scale of dataset. Since data collection can be very costly, achieving effective FER based on an extremely smaller dataset (*e.g.*, few-shot learning [58]) has its merits, which we leave for future work.

6 RELATED WORK

Device-free FER. Very few studies have been done in developing device-free FER systems that analyze Wi-Fi or ultrasound signals reflected from facial muscle movements. WiFace [7] exploits the unique changes of channel state information (CSI) in Wi-Fi signals caused by facial muscle movements. Although it has proven its effectiveness with fair accuracy in different settings, due to intrinsic limitations of Wi-Fi signals (*i.e.*, low bandwidth and multipath effect) [39], it may not be able to eliminate the impact of body motions. Also, it requires an extra antenna setup with manual placement configuration, hence leading to poor applicability. SonicFace [13] uses a customized microphone array that emits ultrasound signals for FER, but its detection range is limited to 60cm maximum due to the fact that ultrasound signals may be brittle by ambient noise [6], hence hindering the deployment of most multimedia IoT applications. Different from these works, mmFER leverages on mmWave radar to detect subtle facial muscle movements with higher signal resolution, capable of sensing multiple targets with a flexible detection range. mmFER is also robust and resilient to enable FER in different indoor scenarios.

Cross-domain Transfer Learning for FER. Due to limited FER datasets available in sensing domain, cross-domain transfer learning [27, 33] can be useful for image-to-sensing transformation. FaceListener [51] presents a teacher-student learning scheme based on knowledge distillation that learns the latent correlation between facial landmarks in images and reflected acoustic signals, achieving fair accuracy for FER. Similar idea can be applied to mmWave-based FER, however, our study shows that the performance may significantly degrade due to training over-fitting caused by small-scale mmWave training dataset. In mmFER, we propose the cross-transfer pipeline to address the issue, achieving effective model knowledge transformation for mmWave-based FER.

7 CONCLUSION

This paper presents a novel mmWave radar based FER system that detects subtle facial muscle movements caused by facial expressions for facial expression recognition. mmFER is a device-free solution that enables robust FER and delivers valuable assessment of users' engagement and reactions to multimedia IoT applications. Extensive evaluations demonstrate that mmFER is resilient to various real-world settings with high accuracy, hence it is capable of deployment in a wide range of scenarios with a flexible detection range for multiple users.

References

[1] Ahmed Abdelreheem, Ehab Mahmoud Mohamed, and Hamada Esmaiel. 2018. Location-Based Millimeter Wave Multi-Level Beamforming Using Compressive Sensing. *IEEE Communications Letters* (2018).

[2] Andrea Asperti, Davide Evangelista, and Elena Loli Piccolomini. 2021. A Survey on Variational Autoencoders from a Green AI Perspective. *SN Computer Science* (2021).

[3] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 279–283.

[4] Cédric Bertolus, Daniel Bailleul, and Marc Mersiol. 2017. Viewing distance requires large characters to ensure legibility on TV-set. In *Proceedings of the 29th Conference on l'Interaction Homme-Machine*. 147–155.

[5] J. Capon. 1969. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* 57, 8 (1969), 1408–1418. https://doi.org/10.1109/PROC.1969.7278

[6] Jyotismita Chaki. 2021. Pattern analysis based acoustic signal processing: a survey of the state-of-art. *Int J Speech Technol 24* (2021).

[7] Yanjiao Chen, Runmin Ou, Zhiyang Li, and Kaishun Wu. 2020. WiFace: facial expression recognition using Wi-Fi signals. *IEEE Transactions on Mobile Computing* 21, 1 (2020), 378–391.

[8] Han Cui, Weizheng Zhong, Zhuoxin Yang, Xuemei Cao, Shuangyan Dai, Xingxian Huang, Liyu Hu, Kai Lan, Guanglin Li, and Haibo Yu. 2021. Comparison of Facial Muscle Activation Patterns Between Healthy and Bell's Palsy Subjects Using High-Density Surface Electromyography. *Frontiers in Human Neuroscience* 14 (2021), 608.

[9] Zhiwei Deng, Rajitha Navarathna, Peter Carr, Stephan Mandt, Yisong Yue, Iain Matthews, and Greg Mori. 2017. Factorized variational autoencoders for modeling audience reactions to movies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2577–2586.

[10] Paul Ekman and Wallace V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology* 17, 2 (1971), 124.

[11] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In *Proceedings of the 18th ACM international conference on multimodal interaction*. 445–450.

[12] Sofien Gannouni, Arwa Aledaily, Kais Belwafi, and Hatim Aboalsamh. 2021. Emotion detection using electroencephalography signals and a zero-time windowing-based epoch estimation and relevant electrode identification. *Sci Rep 11* (2021).

[13] Yang Gao, Yincheng Jin, Seokmin Choi, Jiyang Li, Junjie Pan, Lin Shu, Chi Zhou, and Zhanpeng Jin. 2021. SonicFace: Tracking Facial Expressions Using a Commodity Microphone Array. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–33.

[14] Jian Gong, Xinyu Zhang, Kaixin Lin, Ju Ren, Yaoxue Zhang, and Wenxun Qiu. 2021. RF Vital Sign Sensing under Free Body Movement. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 3 (2021).

[15] Sam Gregory. 2010. Cameras Everywhere: Ubiquitous Video Documentation of Human Rights, New Forms of Video Advocacy, and Considerations of Safety, Security, Dignity and Consent. *Journal of Human Rights Practice* 2, 2 (05 2010), 191–207.

[16] Ke Gu, Min Liu, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. 2015. Quality Assessment Considering Viewing Distance and Image Resolution. *IEEE Transactions on Broadcasting* (2015).

[17] Tianbo Gu, Zheng Fang, Zhicheng Yang, Pengfei Hu, and Prasant Mohapatra. 2019. MmSense: Multi-Person Detection and Identification via MmWave Sensing *(mmNets'19)*.

[18] Unsoo Ha, Salah Assana, and Fadel Adib. 2020. Contactless Seismocardiography via Deep Learning Radars *(MobiCom'20)*.

[19] Arghavan Hadinejad, Brent D Moyle, Noel Scott, and Anna Kralj. 2019. Emotional responses to tourism advertisements: the application of FaceReader™. *Tourism Recreation Research* 44, 1 (2019), 131–135.

[20] Erik Hanson-Viana, Jorge Rojas-Ortiz, Marco A Rendón-Medina, Ricardo C Pacheco-López, Luciano R Ríos-Lara López, and Julio Palacios-Juárez. 2021. Influence of BMI, age, and gender on the thickness of most common thinned flaps. *Plastic and Reconstructive Surgery Global Open* 9, 3 (2021).

[21] Behzad Hasani and Mohammad H Mahoor. 2017. Facial expression recognition using enhanced deep 3D convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 30–40.

[22] Yuan He, Pascal Aubry, Francois Le Chevalier, and Alexander Yarovoy. 2014. Keystone transform based range-Doppler processing for human target in UWB radar. In *2014 IEEE Radar Conference*.

[23] Bita Houshmand and Naimul Mefraz Khan. 2020. Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*. IEEE, 70–75.

[24] Texas Instruments. 2020. *MMWAVE SDK User Guide*.

[25] Kota Irie and Kazunori Umeda. 2003. Detection of waving hands from images using time series of intensity values. *Journal of the Robotics Society of Japan* 21, 8 (2003), 923–931.

[26] Ashraful Islam, Chun-Fu Richard Chen, Rameswar Panda, Leonid Karlinsky, Richard Radke, and Rogerio Feris. 2021. A broad study on the transferability of visual representations with contrastive learning. In *ICCV'21*.

[27] Yanli Ji, Yuhan Hu, Yang Yang, Fumin Shen, and Heng Tao Shen. 2019. Cross-domain facial expression recognition via an intra-category common feature and inter-category distinction feature fusion network. *Neurocomputing* 333 (2019), 231–239.

[28] Junguang Jiang, Yang Shu, Jianmin Wang, and Mingsheng Long. 2022. Transferability in Deep Learning: A Survey. *ArXiv* (2022).

[29] Heechul Jung, Sihaeng Lee, Junho Yim, Sunjeong Park, and Junmo Kim. 2015. Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision*. 2983–2991.

[30] Julien Le Kernec, Francesco Fioranelli, Chuanwei Ding, Heng Zhao, Li Sun, Hong Hong, Jordane Lorandel, and Olivier Romain. 2019. Radar signal processing for sensing in assisted living: The challenges associated with real-time implementation of emerging algorithms. *IEEE Signal Processing Magazine* 36, 4 (2019), 29–41.

[31] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, and Wenyao Xu. 2020. VocalPrint: Exploring a Resilient and Secure Voice Authentication via MmWave Biometric Interrogation *(SenSys'20)*.

[32] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE transactions on affective computing* (2020).

[33] Shan Li and Weihong Deng. 2020. A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing* (2020).

[34] Xinrong Li, Xiaodong Wang, Qing Yang, and Song Fu. 2021. Signal Processing for TDM MIMO FMCW Millimeter-Wave Radar Sensors. *IEEE Access* 9 (2021), 167959–167971.

[35] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A. Stankovic, Niki Trigoni, and Andrew Markham. 2020. See through Smoke: Robust Indoor Mapping with Low-Cost MmWave Radar *(MobiSys '20)*.

[36] Chris Xiaoxuan Lu, Muhamad Risqi U. Saputra, Peijun Zhao, Yasin Almalioglu, Pedro P. B. de Gusmao, Changhao Chen, Ke Sun, Niki Trigoni, and Andrew Markham. 2020. MilliEgo: Single-Chip MmWave Radar

Aided Egomotion Estimation via Deep Sensor Fusion *(SenSys'20)*.

[37] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).

[38] Wenjie Lv, Wangdong He, Xipeng Lin, and Jungang Miao. 2021. Non-Contact Monitoring of Human Vital Signs Using FMCW Millimeter Wave Radar in the 120 GHz Band. *Sensors* 21, 8 (2021), 2732.

[39] Yongsen Ma, Gang Zhou, and Shuangquan Wang. 2019. WiFi Sensing with Channel State Information: A Survey. *ACM Comput. Surv.* (2019).

[40] Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. 2019. Frame attention networks for facial expression recognition in videos. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3866–3870.

[41] Ali Nauman, Yazdan Ahmad Qadri, Muhammad Amjad, Yousaf Bin Zikria, Muhammad Khalil Afzal, and Sung Won Kim. 2020. Multimedia Internet of Things: A Comprehensive Survey. *IEEE Access* (2020).

[42] Hieu V Nguyen and Li Bai. 2010. Cosine similarity metric learning for face verification. In *Asian conference on computer vision*.

[43] Kun Qian, Zhaoyuan He, and Xinyu Zhang. 2020. 3D Point Cloud Generation with Millimeter-Wave Radar. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* (2020).

[44] Kun Qian, Zhaoyuan He, and Xinyu Zhang. 2020. 3D point cloud generation with millimeter-wave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–23.

[45] Suman Saha, Rajitha Navarathna, Leonhard Helminger, and Romann M Weber. 2018. Unsupervised deep representations for learning audience facial behaviors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1132–1137.

[46] Najmeh Samadiani, Guangyan Huang, Borui Cai, Wei Luo, Chi-Hung Chi, Yong Xiang, and Jing He. 2019. A review on automatic facial expression recognition systems assisted by multimodal sensor data. *Sensors* (2019).

[47] Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*.

[48] Ismail Shahin, Ali Bou Nassif, and Shibani Hamsa. 2019. Emotion recognition using hybrid Gaussian mixture model and deep neural network. *IEEE access* 7 (2019).

[49] V Shrivathsa. 2018. Cell Averaging-Constant False Alarm Rate Detection in Radar. *International Research Journal of Engineering and Technology (IRJET)* 7 (2018), 2433–2438.

[50] Siddharth Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski. 2019. Impact of Affective Multimedia Content on the Electroencephalogram and Facial Expressions. *Sci Rep 9* (2019).

[51] Xingzhe Song, Kai Huang, and Wei Gao. 2022. FaceListener: Recognizing Human Facial Expressions via Acoustic Sensing on Commodity Headphones *(IPSN'22)*.

[52] Mariusz Szwoch and Paweł Pieniążek. 2015. Facial emotion recognition using depth data. In *HSI '15*. 271–277.

[53] Goran Udovičić, Jurica Đerek, Mladen Russo, and Marjan Sikora. 2017. Wearable Emotion Recognition System Based on GSR and PPG Signals *(MMHealth '17)*.

[54] Bram van Berlo, Amany Elkelany, Tanir Ozcelebi, and Nirvana Meratnia. 2021. Millimeter wave sensing: A review of application pipelines and building blocks. *IEEE Sensors Journal* (2021).

[55] Dhruv Verma, Sejal Bhalla, Dhruv Sahnan, Jainendra Shukla, and Aman Parnami. 2021. ExpressEar: Sensing Fine-Grained Facial Expressions with Earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–28.

[56] Lin Wang and Kuk-Jin Yoon. 2021. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[57] Qin Wang, Yanxiao Zhao, Wei Wang, Daniel Minoli, Kazem Sohraby, Hongbo Zhu, and Ben Occhiogrosso. 2017. Multimedia IoT systems and applications. In *2017 Global Internet of Things Summit (GIoTS)*.

[58] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. 2020. Generalizing from a Few Examples: A Survey on Few-Shot Learning. *ACM Comput. Surv.* (2020).

[59] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*. 14–26.

[60] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 2021. mmMesh: towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 269–282.

[61] Ao Zhang, Farzan Erlik Nowruzi, and Robert Laganiere. 2021. RADDet: Range-Azimuth-Doppler based Radar Object Detection for Dynamic Road Users. In *2021 18th Conference on Robots and Vision (CRV)*.

[62] Kaihao Zhang, Yongzhen Huang, Yong Du, and Liang Wang. 2017. Facial expression recognition based on deep evolutional spatial-temporal networks. *IEEE Transactions on Image Processing* 26, 9 (2017), 4193–4203.

[63] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.

[64] Yu Zhang, Tao Gu, and Xi Zhang. 2020. MDLdroidLite: a Release-and-Inhibit Control Approach to Resource-Efficient Deep Neural Networks on Mobile Devices *(SenSys '20)*.

[65] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7356–7365.

[66] Axel Zinkernagel, Rainer W. Alexandrowicz, Tanja Lischetzke, and Manfred Schmitt. 2019. The blenderFace method: video-based measurement of raw movement data during facial expressions of emotion using open-source software. *Behav Res 51* (2019).