IdentityKD: Identity-wise Cross-modal Knowledge Distillation for Person Recognition via mmWave Radar Sensors

Liqun Shan
University of Louisiana at Lafayette
US
liqun.shan1@louisiana.edu

Rujun Zhang
Northeast Petroleum University
China
zhangrujun0214@gmail.com

Sai Venkatesh Chilukoti
University of Louisiana at Lafayette
US
sai-venkatesh.chilukoti1@louisiana.edu

Xingli Zhang
University of Louisiana at Lafayette
US
xingli.zhang@louisiana.edu

Insup Lee
University of Pennsylvania
US
lee@cis.upenn.edu

Xiali Hei
University of Louisiana at Lafayette
US
xiali.hei@louisiana.edu

Abstract

Recent advancements in person recognition have raised concerns about identity privacy leaks. Gait recognition through millimeter-wave radar provides a privacy-centric method. However, it is challenged by lower accuracy due to the sparse data these sensors capture. We are the first to investigate a cross-modal method, IdentityKD, to enhance gait-based person recognition with the assistance of facial data. IdentityKD involves a training process using both gait and facial data, while the inference stage is conducted exclusively with gait data. To effectively transfer facial knowledge to the gait model, we create a composite feature representation using contrastive learning. This method integrates facial and gait features into a unified embedding that captures the unique identity-specific information from both modalities. We employ two distinct contrastive learning losses. One minimizes the distance between embeddings of data pairs from the same person, enhancing intra-class compactness, while the other maximizes the distance between embeddings of data pairs from different individuals, improving inter-class separability. Additionally, we use an identity-wise distillation strategy, which tailors the training process for each individual, ensuring that the model learns to distinguish between different identities more effectively. Our experiments on a dataset of 36 subjects, each providing over 5000 face-gait pairs, demonstrate that IdentityKD improves identity recognition accuracy by 6.5% compared to baseline methods.

CCS Concepts

• **Security and privacy** → **Security services**; • **Computing methodologies** → **Machine learning**.

Keywords

Identity recognition, Cross-modality, Knowledge distillation, Contrast learning, Gait, mmWave radar

ACM Reference Format:
Liqun Shan, Rujun Zhang, Sai Venkatesh Chilukoti, Xingli Zhang, Insup Lee, and Xiali Hei. 2024. IdentityKD: Identity-wise Cross-modal Knowledge Distillation for Person Recognition via mmWave Radar Sensors. In *ACM Multimedia Asia (MMASIA '24), December 03–06, 2024, Auckland, New Zealand.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3696409.3700254

MMASIA '24, December 03–06, 2024, Auckland, New Zealand

ACM ISBN 979-8-4007-1273-9/24/12
https://doi.org/10.1145/3696409.3700254

1 Introduction

For decades, the problem of human authentication has been a persistent challenge. Biometric-based recognition was once considered a reliable human authentication method. Common biometrics involve technologies like retina scans [19], voice recognition [8], and facial recognition [32]. However, all these bio-signs necessitate extra effort from individuals for authentication. For instance, facial recognition demands positioning in front of a high-resolution camera to capture unique facial features, and fingerprint methods require one to place his/her fingers on a scanner to detect patterns. These methods are vulnerable to spoofing attacks, where unauthorized users deceive the system with counterfeit biometric traits of a legitimate user [2, 9, 20, 23]. To overcome these issues, researchers have turned to effortless and non-invasive authentication systems using behavioral biometrics, such as gait features. Notably, studies have shown that it is exceedingly difficult to spoof human gait [21], as the uniqueness of an individual's gait makes it challenging to imitate others effectively.

Gait recognition, which analyzes an individual's walking pattern, presents a promising alternative because it is non-invasive and difficult to spoof [15]. Imagine where your identity can be securely recognized without revealing sensitive information like your face. This vision drives our research. Our goal is to recognize individuals based on how they walk, using millimeter-wave (mmWave) radar while preserving privacy. These sensors can capture the movements of a person without showing their face, which is great for privacy. However, gait recognition using mmWave radar sensors faces challenges due to the sparse data these sensors capture, often leading to lower recognition accuracy [1, 25]. To address these challenges, we explore leveraging multimodal learning by incorporating facial data, which inherently contains a much richer and more detailed

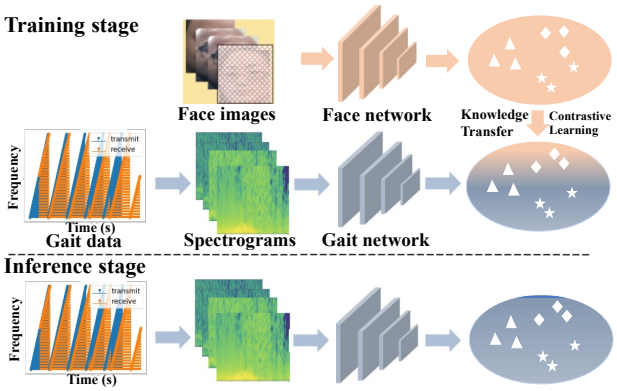

Figure 1: Cross-modal knowledge transfer. The face network (teacher) transfers the discriminative knowledge to the gait network (student) during training. During the inference, the gait network only conducts identity recognition. The same shape (triangle, star, diamond) represents the same identity.

set of identity-specific information, to significantly enhance the accuracy and reliability of gait-based identity recognition.

We propose an identity recognition method based on the knowledge distillation (KD) [11] via gait data from mmWave radar sensors, called IdentityKD. The overview of IdentityKD is shown in Figure. 1. We employ a KD strategy to facilitate the transfer of knowledge from the more informative facial modality to the less informative gait modality during training. However, we empirically find that directly applying conventional KD techniques to our cross-modal teacher-student learning framework does not yield the desired outcomes. We identify two primary factors contributing to this suboptimal performance: (1) a significant disparity in the latent space domains of the gait and face modalities, and (2) misalignment in the quality of the input data for faces and gaits. Specifically, traditional KD methods typically aim to minimize the difference between the teacher and student outputs, encouraging the student model to mirror the teacher model exactly. However, due to the considerable domain gap between the two modalities, this approach can result in overfitting. Additionally, when the quality of input data from the two sources is not comparable, the efficacy of the distillation process tends to diminish.

To solve these problems, we use contrastive learning to integrate facial and gait features into a unified embedding that captures the unique identity-specific information from both modalities shown in Figure. 1. During training, we use both the radar data (gait) and facial images. Specifically, we first convert the time-series gait data into spectrogram images, which can be processed similarly to images. Then, we use KD to transfer knowledge from the facial data to the gait data. The face network teaches the gait network to understand and learn from the detailed facial information, enhancing its ability to interpret the less detailed gait data. Meanwhile, we employ compositional contrastive learning, a technique that ensures features from the same individual are close together in the feature space (intra-class compactness) and those from different individuals are far apart (inter-class separability). This alignment is crucial for creating a unified representation that accurately captures identity-specific features from facial and gait modalities. Additionally, we use an identity-wise training strategy that tailors the learning process for each individual. This makes the gait network better at distinguishing between different persons, even if the data quality varies. After training, during the inference stage, the network can recognize persons using only the gait data from mmWave radar sensors.

Our method is tested on a dataset comprising 36 subjects, demonstrating that our approach significantly boosts identity recognition performance by approximately 5% to 9%. Our method ensures that in real-world use, only gait data is needed, which is less invasive and better for privacy. This approach addresses the limitations of radar data and provides a strong solution for security, surveillance, and access control where privacy is crucial. Our contributions are summarized as follows:

- We propose IdentityKD, a novel KD model for person recognition, featured by learnable compositional embeddings that bridge the semantic gap between gait and face modalities and utilize a distillation objective that simultaneously contrasts two modalities within a unified latent space.
- We present an identity-wise distillation strategy that adaptively tailors identity configurations for each subject during training, and refines the weight averaging process to enhance and stabilize the model across all individuals, particularly improving robustness in the least performing classes.
- We conduct experiments on a dataset of 36 subjects, and the results show that our method can effectively improve the accuracy of identity recognition.

2 Background

2.1 mmWave Radars for Person Identification

Principles of mmWave Radar. The mmWave radar operates on the frequency-modulated continuous wave (FMCW) principle, which enables it to measure the range, relative radial speed, and angle of a target [31]. The radar sends out a chirp signal that increases in frequency over time and processes the reflected signals from objects to determine their positions in three dimensions. For range measurement, the distance between the radar and the object is calculated using the formula involving the intermediate frequency (IF), the speed of light, the bandwidth of the chirp, and the chirp duration. A fast Fourier transform (FFT) is performed on the IF signal to measure the range of multiple objects at different distances. For angle estimation, the mmWave radar employs a linear antenna array. By emitting chirps with the same initial phase and sampling the reflected signals with multiple receiver antennas, the radar can determine the phase difference between signals received by consecutive antennas. The angle of arrival (AoA) is then calculated using the phase difference, the wavelength, and the distance between the antennas. This allows for the precise location of objects in a Cartesian coordinate system once the range and AoA are known. Based on the principle of mmWave radar, we can effectively capture relevant information about individuals entering the radar monitoring area, providing a foundation for subsequent data preprocessing.

Person Identification Using Gait. Gait, a biometric characteristic observable during walking, inherently includes identity-specific traits that can be used for human identification [3, 7, 13, 24]. Gait-based identification systems work by extracting walking features

from a data sequence, such as step length, arm swing amplitude, and walking pace, and then employing these characteristics as identifiers. As walking involves a sequence of movements, gait feature extraction typically spans multiple frames. In our study, we utilize such sequences to extract gait features from an mmWave radar sensor and to re-identify individuals.

2.2 Cross-modal Knowledge Distillation

Knowledge distillation (KD) is a model compression technique used to transfer knowledge from a complex model (teacher) to a lightweight model (student) to improve the overall performance of the small model [12]. KD techniques aim to minimize the discrepancy between the prediction scores (logits) of the teacher and student models. It has been suggested to broaden the feature information extracted from the model using cross-modal KD. This bridging technique establishes a connection across diverse data modalities to facilitate the transfer of knowledge, ultimately leading to improved performance. In cross-modal KD, most existing approaches focus on closely related modalities, including depth and optical flow images [10] and RGB and depth images [26]. Some researchers employed a multi-modal teacher combining face and speech to oversee the training of a single-modal student model, although they noted a significant disparity between the speech and face data, which posed challenges in enhancing performance [14, 16, 30]. Current work [5] transferred knowledge across image, audio, and video modalities uncovering richer multi-modal knowledge. Inspired by these studies, our paper targets the issue of cross-modal disparities, specifically between facial and gait data for person recognition.

2.3 Contrastive Learning

Contrastive self-supervised learning focuses on training an encoder to derive meaningful representations from unlabeled images by bringing together similar samples (positives) and separating dissimilar ones (negatives). In particular, instance discrimination treats augmented versions of the same image (e.g., through random cropping) as positives, while considering distinct images as negatives.

Contrastive learning techniques with unlabeled samples often struggle to capture category-specific information during the feature extraction process. However, when these self-supervised contrastive learning strategies are employed on annotated data, they have been shown to more effectively consolidate features from identical categories, resulting in denser clusters [17]. This validates the utility of supervised contrastive learning in the domain of identity recognition. As a result, we intend to integrate supervised techniques with the principles of contrastive learning to develop a model capable of discerning feature consistency within the same category as well as distinctions across different categories [6]. Our approach will specifically showcase contrastive learning through the use of dual perspectives of an individual's identity: one from a standard RGB image and the other from gait captured by mmWave radar sensors. We will construct contrastive pairs using mmWave radar data, consisting of anchor images and comparative samples. These pairs will then be processed by the contrastive learning network for advanced feature extraction, which we will elaborate on in the subsequent section.

3 Methodology

The overall structure of our IdentityKD method is shown in Fig. 2. We first establish a face feature extraction module to serve as the teacher model, extracting features from RGB imagery. Concurrently, we construct a gait feature extraction module that acts as the student model, tasked with gait feature extraction from data obtained via mmWave radar sensors. Subsequently, we employ KD methods to refine the student model through contrastive loss optimization. Additionally, the model is further enhanced through an identity-wise calibrated strategy, balancing the disparity of different classes.

3.1 Visual Feature Representation

As shown in Fig. 2, the visual feature extraction module is also the teacher feature extraction module. In this context, the superscripts t and s are associated with the teacher and student models, respectively. The input to this module is an RGB facial image sequence X_n, where each image x_i in X_n has dimension $\mathbb{R}^{3\times H\times W}$, with width W and height H. The facial dataset D_f can be expressed as

$$D_f = \{(x_i, y_i)\}_{i=1}^{n}, \quad n \in \{1, 2, 3, \ldots\} \tag{1}$$

where x_i refers to the i-th facial sample associated with the teacher model t. y_i is the label corresponding to the i-th sample.

The facial input dataset is subjected to random data augmentation for data augmentation to enhance the generalization capacity of the model. The preprocessed RGB data is then put into the teacher model to extract facial features. The teacher is a facial recognition system whose main goal is to learn the Euclidean distance between samples to reduce the distance between facial samples of the same person and increase the distance between facial samples of different people. We propose a feature fusion strategy (details in Sec. 3.3) for minimizing the effect of significant modality distinctions between cross-modal datasets. To avoid the inhomogeneity of the structure of the teacher-student feature capture networks across modalities, we harvest the outputs of the penultimate layer of the teacher and student models for cross-modal fusion. Note that for a classification task, the penultimate layer is the layer before the final classifier. We extract the penultimate feature embeddings before the classifier. In this study, x_i is fed into the FaceNet [28] encoder $F^t(\cdot)$ to generate the facial feature representation p_i^t:

$$p_i^t = F^t(x_i) \tag{2}$$

3.2 RF Feature Representation

We preprocess the raw gathered gait dataset from mmWave radar to create spectrogram images Z_n, where each image z_i in Z_n has dimension $\mathbb{R}^{3\times H\times W}$. We then feed these images to our student model $F^s(\cdot)$, labeling the gait dataset as

$$D_g = \{(z_i, y_i)\}_{i=1}^{n}, \quad n \in \{1, 2, 3, \ldots\} \tag{3}$$

where z_i refers to the i-th gait spectrogram image associated with the student model s.

Our thorough experiments found that the disparity between gait and facial image data modalities is considerable, which complicates the application of knowledge derived from facial features to gait feature enhancement. This leads to ineffective cross-modal knowledge transfer and subpar performance in identity recognition tasks. To address this, we suggest a contrastive learning approach that

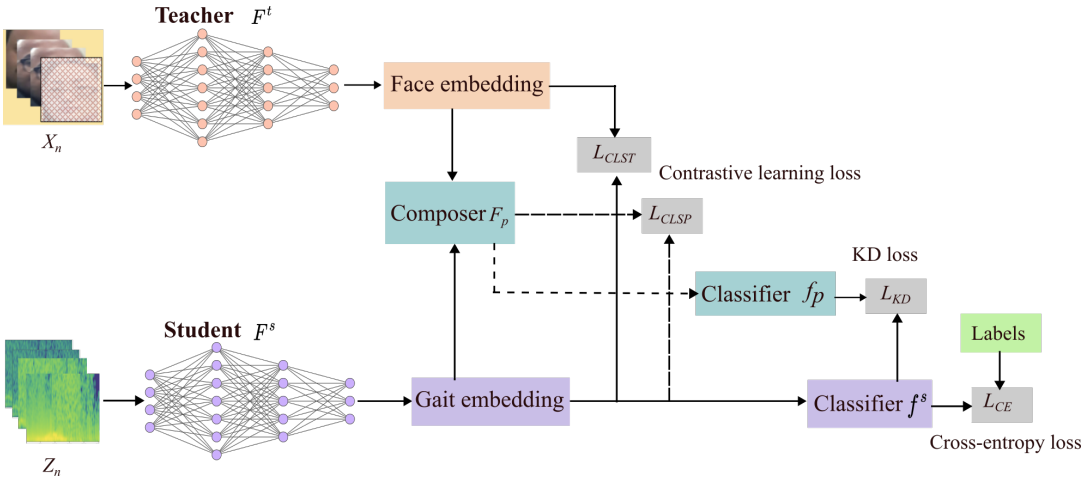

Figure 2: The framework of IdentityKD method where the teacher (face recognition model) transfers discriminative knowledge to the student (gait model). The Composer fuses gait and face embeddings. The framework utilizes contrastive learning loss, KD loss, and cross-entropy loss to optimize student model performance.

combines the embeddings from both the teacher and student models under a supervised framework. This is designed to bridge the semantic divide, causing dissimilar samples to diverge and similar samples to converge, thereby narrowing the domain and semantic gaps between different types of data. The penultimate feature embeddings, denoted by p_i^s, are acquired by passing spectrograms through the student model. Following this, a classifier $f^s(\cdot)$ is employed to produce the ultimate classification output, represented by y_i^s, which also signifies the student model's final logit outputs.

$$p_i^s = F^s(z_i) \tag{4}$$

$$y_i^s = f^s(P_i^s) \tag{5}$$

3.3 Fusion of Multi-Modal Representations

It has been previously noted that student and teacher model embeddings may not align semantically. To reconcile these potential semantic and domain discrepancies between modalities, our approach suggests a correction for the facial image embeddings. This is achieved by merging teacher and student embeddings and anchoring the combined embeddings to the specific aims of our task, thereby bridging the identified semantic gaps. Given that the network structures differ among modalities, the synthesis of cross-modal composition occurs at the penultimate layer.

Formally, the facial embedding p_i^t merges with the gait embedding p_i^s, through the induction of a residual on the teacher embeddings. The teacher embeddings are refined with a composition function $F_p(\cdot,\cdot)$, which introduces a residual function $f^{st}(\cdot,\cdot)$ that combines two modalities using normalization, concatenation, and a linear transformation:

$$p_i = F_p(p_i^t, p_i^s) = p_i^t + f^{st}(p_i^t, p_i^s), \tag{6}$$

where p_i is the fused embeddings. This procedure has connections to earlier studies that integrate features from multiple modalities [4], but our methodology focuses on modulating the teacher embedding through an adjustable residual. More critically, to guide outcomes of the classifier $f_p(\cdot)$, we fine-tune $F_p(\cdot,\cdot)$ with the objective function associated with the subject classification.

3.4 Identity-wise Cross-modal Distillation

Previous uni-modal approaches typically focus on transferring knowledge in the prediction space, where the student network is trained to mirror the output of the teacher network. However, this technique is not directly applicable to multimodal KD, as teacher networks are generally pre-trained with diverse task objectives and predict various classes. In light of this, we suggest employing contrastive learning within the latent feature space and then differentiating the class distributions in the predictive space.

Contrast loss. With uni-modal and compositional embeddings, we aim to distill knowledge by converging positive pairings and diverging negative ones across different modalities. For instance, positive pairings might consist of a facial image and the corresponding gait data of the same individual. Concretely, for a trio composed of a facial image, gait, and their combined embeddings, a contrastive loss can be calculated for each pairing to strengthen their association within the unified feature space. Formally, the contrastive loss (utilizing InfoNCE [22]) for a embedding pair is calculated as follows:

$$L_{CLST} = -\log \frac{\exp(\text{sim}(p_i^s, p_i^t)/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(p_j^s, p_j^t)/\tau)} = -\log p_i^{st} \tag{7}$$

$$L_{CLSP} = -\log \frac{\exp(\text{sim}(p_i^s, p_i)/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(p_j^s, p_j)/\tau)} = -\log p_i^{sp} \tag{8}$$

where L_{CLST} is the contrast loss between gait and face embeddings. L_{CLSP} is the contrast loss between gait and fusion embeddings. B refers to the size of the mini-batch. τ is the temperature factor to control the softness of logits. $\text{sim}(\cdot,\cdot)$ is the cosine similarity.

The two contrastive learning losses, L_{CLSP} and L_{CLST}, play complementary roles in optimizing the feature space for identity recognition. The L_{CLSP} loss is designed to minimize the distance between

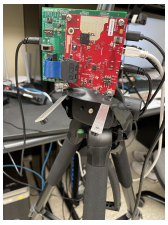

Figure 3: Data collection environment and mmWave device

embeddings of data pairs from the same person, thus enhancing intra-class compactness by clustering features of the same identity tightly together. On the other hand, L_{CLST} aims to maximize the distance between embeddings of data pairs from different individuals, thereby improving inter-class separability by pushing features of different identities further apart. These losses work synergistically: L_{CLSP} ensures that the model can accurately recognize and group features belonging to the same person, while L_{CLST} ensures clear boundaries between different identities. Together, they balance the feature space by simultaneously enforcing compact clusters for each identity and distinct separation between different identities. This dual approach harmonizes intra-class and inter-class relationships, ensuring that there is no contradiction between the two objectives and resulting in a robust and discriminative feature representation.

Considering the possibility of multiple positive face-gait pairings within a single batch pertaining to a particular subject, we introduce a multi-class contrastive loss that incorporates the class label k into its formulation:

$$L_{st}(p_i^s, p_i^t) = -\frac{1}{B_p}\sum_{i=k}\log p_i^{st} - \frac{1}{B_n}\sum_{i\neq k}\log(1-p_i^{st}) \tag{9}$$

$$L_{sp}(p_i^s, p_i) = -\frac{1}{B_p}\sum_{i=k}\log p_i^{sp} - \frac{1}{B_n}\sum_{i\neq k}\log(1-p_i^{sp}) \tag{10}$$

where B_p, B_n represent the counts of positive and negative pairs associated with the gait embedding P_i^s tagged with class k. These equations prompt the network to enhance the probabilities attributed to matching pairs and reduce those assigned to mismatched pairs. The multi-class loss synergizes the singular facial modality and the compounded multi-modal knowledge in a compositional manner:

$$L_{cs} = \frac{\gamma}{B}\sum_{i=1}^{B} L_{st}(p_i^s, p_i^t) + \frac{1-\gamma}{B}\sum_{i=1}^{B} L_{sp}(p_i^s, p_i) \tag{11}$$

with γ being a hyperparameter for tuning the contribution of each term and B being the batch size. In essence, the proposed loss L_{cs} acts as a similarity regulator to synchronize the embeddings within the multi-modal latent space.

Identity-wise KD loss. Empirically, we find that the model's performance varies across different identities, making KD loss unsuitable for all subjects. Inspired by [27], which tailors training to class performance, we introduce a margin m and use identity-specific training accuracy A_i to gauge subject difficulty. For a subject i with training accuracy A_i, we adjust the margin m_i for the next epoch based on A_i. Lower A_i suggests a higher risk of misclassification, so we reduce m_i. To prevent m_i from becoming too small, we introduce a scaling factor λ_1:

$$m_i \leftarrow (\lambda_1 + A_i)\cdot m \tag{12}$$

where m is typically $8/255$. This approach optimizes m_i during training. We also customize robustness regularization β_i for each subject:

$$\beta_i \leftarrow (\lambda_2 + A_i)\cdot \beta \tag{13}$$

where λ_2 is a scaling factor. The KD loss function incorporated by Kullback-Leibler divergence $KL(\cdot)$ and cross-entropy function $L_{CE}(\cdot)$ becomes:

$$L_{KD} = \frac{1}{B}\sum_{i=1}^{B}\frac{L_{CE}(y_i^s, y_i) + \beta_i(\mathrm{KL}(f_p(p_i), y_i) - m_i)}{1+\beta_i} \tag{14}$$

The normalization factor $1+\beta_i$ ensures balanced emphasis across subjects. Introducing m_i integrates the calibrated margin with subject-specific regularization. The total loss is:

$$L = \delta_1 L_{cs} + \delta_2 L_{KD} \tag{15}$$

where δ_1 and δ_2 balance each term's contributions. Note that, unlike [27], which did not apply this method to KD, our identity-wise approach transfers knowledge from facial data to gait data. This novel application of cross-modal KD for identity recognition demonstrates the method's effectiveness in a new context.

4 Experiment

4.1 Experimental Setup

Data collection. In this study, we employ a Texas Instruments (TI) 77 GHz FMCW radar IWR1843BOOST board in conjunction with a DCA1000EVM board to gather raw gait data within an office setting, as depicted in Fig. 3. The radar system operates from 77 to 81 Ghz, covering up to a 4 Ghz bandwidth. A Lenovo 740 laptop with TI mmWave studio software is used as a control system for our radar device to configure the FMCW wave parameters such as chirp width, repetition time, and chirp slope. The radar is positioned 1 meter from the 15 m by 20 m detection area, mounted on a tripod at a height of 1.5 meters. This setup aims to minimize occlusions among subjects by ensuring comprehensive monitoring of the entire detection zone. The specific radar settings used in our experiments are detailed in Table. 1. This configuration enables the FMCW radar to capture data at 200 frames per second, achieving a range resolution of 4 cm and a velocity resolution of 0.06 m/s. The superior range resolution significantly enhances our ability to distinguish between multiple targets, aiding in the effective detection and separation of subjects. Meanwhile, the fine velocity resolution allows for the collection of detailed gait data, increasing the likelihood of identifying distinctive gait characteristics.

Table 1: Radar configuration parameters

Parameter	Value	Parameter	Value
Start frequency	77 GHz	Bandwidth	900.9 MHz
No. of samples/chirp	230	Sampling rate	5000 ksps
No. of chirps/frame	200	ADC Samples	256
Frame periodicity	33 ms	Frequency slope	15 MHz/us

We recruit 36 volunteers to create our dataset. Participants are instructed to walk naturally within an office space. Walks directed towards and away from the radar are both considered separate instances. Each volunteer completes 100 walking trials, yielding

100 raw radar captures per person. The analog signal from mmWave radar is first digitized into a discrete beat signal. Then, a range-FFT is applied to produce a range-time map. We execute Short-Time Fourier Transforms (STFT) across the range bins, creating spectrograms that detail the micro-Doppler signatures of human movement. Finally, these spectrograms are merged to form the final mmWave gait biometric and structured into 256×256 pixels, aligning with the format of the facial data.

We use the Logitech-C920s Pro 1080 Webcam to collect the RGB facial image data while monitoring the gait data. We use MTCNN face detector [29] to conduct face detection and alignment, and then crop the face images to 160×160 pixels. During model training, random data augmentation is applied to RGB input images. Thirty-six people's faces are included in 5000×36 images.

Model. We select ResNet18 as the Baseline for our student model. For the teacher model, we utilize FaceNet [28] which is a model pre-trained for face recognition tasks with an accuracy of 99.08%.

Evaluation metrics. We use the accuracy (ACC) (%), Recall (%), and F1-score (%) as the evaluation indicators for models.

Training setup. During the training, FaceNet is frozen. Adam optimizer [18] is used to optimize the ResNet18 model. An epoch of 200 is run with 128 batches. The learning rate is initially set to 1e-4 and the MultiStepLR learning rate decay algorithm is used for adjusting the learning rate. γ and τ are set to 0.8 and 5.0, respectively.

4.2 Experimental Results

We compare IdentityKD with GaitSet [3], GaitPart [7], GaitGL [13], and CRF [24] under normal walking with view 0 in Table 2. IdentityKD demonstrates exceptional performance in identity recognition by achieving an accuracy of 98.58% using gait data, significantly surpassing other gait recognition methods such as GaitSet, GaitPart, GaitGL, and CRF, which record accuracies of 88.09%, 89.21%, 88.75%, and 93.95%, respectively. This remarkable performance highlights IdentityKD's effectiveness in utilizing cross-modal knowledge distillation, where facial data is leveraged during training to enhance the recognition capabilities of gait data during inference. The method nearly matches the accuracy of FaceNet (99.08%), illustrating its ability to integrate identity-specific information from both facial and gait modalities. This cross-modal integration enables IdentityKD to outperform traditional gait recognition approaches, showcasing its robustness and potential for real-world applications that prioritize privacy by using non-invasive gait data for identity verification.

Table 2: Comparisons of different methods using gait data.

Method	GaitSet	GaitPart	GaitGL	CRF	IdentityKD
ACC	88.09	89.21	88.75	93.95	**98.58**

We also assess IdentityKD, benchmarking it against the vanilla KD [12], the contrast fusion-based KD (CFKD) [5], the margin-based KD (MKD) [16], and the Baseline method. CFKD and MKD are cross-modal KD methods. We present the results in Table 3.

IdentityKD achieves the highest scores in all metrics (ACC: 98.58%, Recall: 99.01%, F1: 98.43%), suggesting it is superior in balancing precision and recall while maintaining high overall accuracy. The next best performer is MKD, followed by CFKD, then vanilla KD. The Baseline method exhibits the least performance, with its accuracy falling short by approximately 6.5% compared to that of IdentityKD.

Table 3: Comparison of IdentityKD against existing KD.

Method	ACC	Recall	F1
Baseline	92.17	92.77	93.13
vanilla KD	94.91	94.25	93.90
CFKD	96.12	95.97	96.45
MKD	96.38	96.16	96.18
IdentityKD	**98.58**	**99.01**	**98.43**

The outperformance of IdentityKD firmly attests to its robustness and efficacy in gait-based identity recognition tasks. It demonstrates the advantage conferred by incorporating facial data within the training paradigm, thereby substantially augmenting the discriminative capacity of mmWave radar for gait-based identity recognition. Notably, this advancement is realized while rigorously adhering to privacy considerations, given that the use of facial data is restricted solely to the training phase. During operational deployment, the inference relies exclusively on gait data, thereby ensuring a privacy-preserving recognition system.

Table 4: Person recognition results under different student model structures and metrics.

Method	ACC	Recall	F1
ResNet18	**98.58**	**99.01**	**98.43**
MobileNet	88.24	87.20	85.69
EfficientNet	95.88	92.20	94.98
Vgg16	87.62	84.25	87.50

4.3 Ablation Study

Impact of student model structure. Table 4 presents the identity recognition performance of various deep learning architectures when applied as student models in the IdentityKD framework.

ResNet18 shows the highest accuracy (94.71%) and Recall (94.15%), indicating its proficiency in correctly identifying positive instances. MobileNet and Vgg16 show improved performance compared to the 4-subject dataset but still lag behind the ResNet models. EfficientNet parallels ResNet18's 95.88% accuracy, making it a highly efficient and practical choice for deployment. Overall, the table illustrates that while all models benefit from KD, there is a clear hierarchy in performance, with ResNet models outperforming MobileNet, EfficientNet, and Vgg16, highlighting the importance of model selection in the context of scale and complexity.

Impact of contrastive learning loss. In IdentityKD, contrastive learning loss plays a crucial role, facilitating the learning of a compositional embedding that bridges the gap between cross-modal data while capturing semantics pertinent to the task. Our evaluation compares the efficacy of contrast losses derived from KL divergence and noise-contrastive estimation (NCE), as shown in Table 5. We can see that while both methods are effective for the identity task, the NCE method demonstrates a marginal advantage over KL in terms of these performance metrics.

Impact of identity-wise configuration. We investigate the influence of base scaling factor λ_1 and λ_2 by conducting experiments

Table 5: Comparison of contrast losses from KL and NCE.

Method	ACC	Recall	F1
KL	97.24	97.05	96.98
NCE	98.58	99.01	98.43

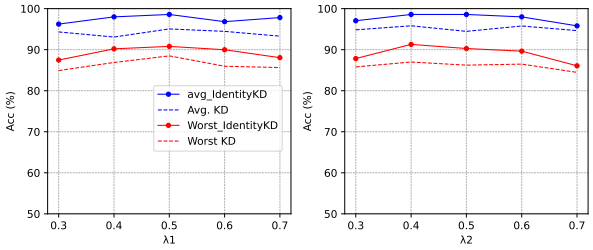

Figure 4: The average (solid) and the worst class robustness (dotted) of models trained with different λ.

of IdentityKD and identity-wise vanilla KD with λ varying from 0.3 to 0.7 shown in Fig 4. We can see that IdentityKD with different λ_1 and λ_2 show better overall and the worst class robustness than vanilla KD, among which $\lambda_1 = 0.5$ and $\lambda_2 = 0.4$ performs best.

5 Conclusion

We develop IdentityKD which uses gait data from mmWave radar with the assistance of facial images for person recognition. This technique mitigates the challenges posed by data misalignment through knowledge distillation methodology, augmented by a compositional embedding strategy. To bolster the model's robustness across various identities, we incorporate contrast learning along with identity-wise calibration into the knowledge transfer process. Our streamlined student model is designed to preserve individual privacy while still achieving precise identification through mmWave radar data. Comparative evaluations demonstrate that our method outperforms the baseline, offering exceptional performance improvements by 6.5%.

References

[1] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. 2015. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–13.

[2] Domna Bilika, Nikoletta Michopoulou, Efthimios Alepis, and Constantinos Patsakis. 2024. Hello me, meet the real me: Voice synthesis attacks on voice assistants. *Computers & Security* 137 (2024), 103617.

[3] Hanqing Chao, Kun Wang, Yiwei He, Junping Zhang, and Jianfeng Feng. 2021. GaitSet: Cross-view gait recognition through utilizing gait as a deep set. *IEEE transactions on pattern analysis and machine intelligence* 44, 7 (2021), 3467–3478.

[4] Yanbei Chen and Loris Bazzani. 2020. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer, 136–152.

[5] Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. 2021. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7016–7025.

[6] Fengkai Dong, Xiaoqiang Zou, Jiahui Wang, and Xiyao Liu. 2023. Contrastive learning-based general Deepfake detection with multi-scale RGB frequency clues. *Journal of King Saud University-Computer and Information Sciences* 35, 4 (2023), 90–99.

[7] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. 2020. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14225–14233.

[8] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. 343–355.

[9] Javier Galbally, Sébastien Marcel, and Julian Fierrez. 2014. Biometric antispoofing methods: A survey in face recognition. *IEEE Access* 2 (2014), 1530–1552.

[10] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2827–2836.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[12] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[13] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. 2020. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European conference on computer vision*. Springer, 382–398.

[14] Nakamasa Inoue. 2021. Teacher-assisted mini-batch sampling for blind distillation using metric learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4160–4164.

[15] Prabhu Janakaraj, Kalvik Jakkala, Arupjyoti Bhuyan, Zhi Sun, Pu Wang, and Minwoo Lee. 2019. STAR: Simultaneous tracking and recognition through millimeter waves and deep learning. In *2019 12th IFIP Wireless and Mobile Networking Conference (WMNC)*. IEEE, 211–218.

[16] Yufeng Jin, Guosheng Hu, Haonan Chen, Duoqian Miao, Liang Hu, and Cairong Zhao. 2023. Cross-modal distillation for speaker recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 12977–12985.

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33 (2020), 18661–18673.

[18] D Kinga, Jimmy Ba Adam, et al. 2015. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, Vol. 5. San Diego, California;, 6.

[19] Cástor Marino, Manuel G Penedo, Marta Penas, María J Carreira, and F Gonzalez. 2006. Personal authentication using digital retinal images. *Pattern Analysis and Applications* 9 (2006), 21–33.

[20] David Menotti, Giovani Chiachia, Allan Pinto, William Robson Schwartz, Helio Pedrini, Alexandre Xavier Falcao, and Anderson Rocha. 2015. Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Transactions on Information Forensics and Security* 10, 4 (2015), 864–879.

[21] Bendik B Mjaaland, Patrick Bours, and Danilo Gligoroski. 2011. Walk the walk: Attacking gait biometrics by imitation. In *Information Security: 13th International Conference, ISC 2010, Boca Raton, FL, USA, October 25-28, 2010, Revised Selected Papers 13*. Springer, 361–380.

[22] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[23] Nafiz Sadman, Kazi Amit Hasan, Elyas Rashno, Furkan Alaca, Yuan Tian, and Farhana Zulkernine. 2023. Vulnerability of Open-Source Face Recognition Systems to Blackbox Attacks: A Case Study with InsightFace. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 1164–1169.

[24] Yu Shi, Lan Du, Xiaoyang Chen, Xun Liao, Zengyu Yu, Zenghui Li, Chunxin Wang, and Shikun Xue. 2023. Robust gait recognition based on deep CNNs with camera and radar sensor fusion. *IEEE Internet of Things Journal* (2023).

[25] Dave Tahmoush and Jerry Silvious. 2009. Radar micro-Doppler for long range front-view gait recognition. In *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*. IEEE, 1–6.

[26] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699* (2019).

[27] Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. 2023. Cfa: Class-wise calibrated fair adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8193–8201.

[28] Ivan William, Eko Hari Rachmawanto, Heru Agus Santoso, Christy Atika Sari, et al. 2019. Face recognition using facenet (survey, performance test, and comparison). In *2019 fourth international conference on informatics and computing (ICIC)*. IEEE, 1–6.

[29] 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.

[30] Leying Zhang, Zhengyang Chen, and Yanmin Qian. 2021. Knowledge Distillation from Multi-Modality to Single-Modality for Person Verification. *Proc. Interspeech 2021* (2021), 1897–1901.

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

[32] Maheen Zulfiqar, Fatima Syed, Muhammad Jaleed Khan, and Khurram Khurshid. 2019. Deep face recognition for biometric authentication. In *2019 international conference on electrical, communication, and computer engineering (ICECCE)*. IEEE, 1–6.