Journal of Computer and Communications, 2025, 13(6), 64-80
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Survey of Pedestrian Re-Identification Based on Millimeter Wave Radar and Vision Fusion

Qingyuan Yang, Zhipeng Quan, Jingxuan Li, Tingyv Jiang, Zhihao Deng, Xinran Qiu, Zhengjie Wang*

College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao, China
Email: 1790557735@qq.com, 371398728@qq.com, 2987349534@qq.com, 15853765837@163.com, 1020628268@qq.com, 3448288633@qq.com, *cieewangzj@163.com

How to cite this paper: Yang, Q.Y., Quan, Z.P., Li, J.X., Jiang, T.Y., Deng, Z.H., Qiu, X.R. and Wang, Z.J. (2025) A Survey of Pedestrian Re-Identification Based on Millimeter Wave Radar and Vision Fusion. *Journal of Computer and Communications*, **13**, 64-80.
https://doi.org/10.4236/jcc.2025.136005

Received: April 30, 2025
Accepted: June 16, 2025
Published: June 19, 2025

Abstract

With the advancement of technology and the growth of human demand, pedestrian re-identification is a key technology of intelligent systems and plays an important role in daily life. Traditional vision methods have certain limitations, but the millimeter-wave-vision fusion system, which complements the advantages of cameras and millimeter-wave radars, plays a greater role and is attracting widespread attention. This paper first introduces the effects of vision camera and millimeter-wave radar camera on human re-identification in a single mode, and then discusses the detailed processing required to fuse the data of the two modes, including the key technologies of sensor settings, data synchronization and sensor calibration. We also review the classification and evolution of millimeter-wave radar and visual fusion pedestrian re-identification methods, including data-level, feature-level and decision-level methods, and review the previous research methods, which will provide great inspiration for future research. Finally, this paper discusses the typical applications of millimeter-wave vision fusion systems, as well as the key technologies and potential challenges of fusing millimeter-wave radar and visual data, and looks forward to future research directions.

Keywords

Pedestrian Re-Identification, RGB Cameras, mmWave Radar, Feature Fusion

1. Introduction

Pedestrian re-identification (ReID) is a fundamental computer vision task that involves matching images or video sequences of the same person captured by different cameras at different times and locations [1]. With a wide range of promis-

ing prospects, it plays a crucial role in various applications, including intelligent surveillance systems for public safety, human-robot interaction, autonomous driving perception, and retail analytics [2].

Traditionally, ReID research has focused on visual data captured by RGB cameras [3]. Significant advancements have been made leveraging deep learning techniques, leading to impressive performance on benchmark datasets under ideal conditions [4]. However, visual ReID methods face inherent limitations in real-world scenarios. Their performance drastically degrades under challenging environmental conditions, such as: illumination variations, occlusion, viewpoint changes, low resolution, adverse weather, etc.

To overcome these limitations, researchers have increasingly explored multimodal approaches, integrating information from sensors beyond the visible spectrum [5]. Among various sensor modalities, millimeter-wave (mmWave) radar has emerged as a particularly promising candidate for enhancing pedestrian ReID robustness [6]. mmWave radar operates in the 30 - 300 GHz frequency range, offering unique advantages: robustness to environmental conditions, range and velocity information, penetration capability, privacy preservation, etc.

However, mmWave radar also has limitations, primarily its low spatial resolution compared to cameras and the sparsity of its point cloud data, which lacks rich texture and color information essential for appearance-based matching. This inherent complementarity between visual cameras and mmWave radar motivates their fusion for pedestrian ReID. By combining the strengths of both modalities, fused systems aim to achieve more reliable and robust performance across a wider range of operating conditions than is possible with either sensor alone [7]. **Figure 1** illustrates the conceptual framework of mmWave-visual fusion for enhanced pedestrian ReID.

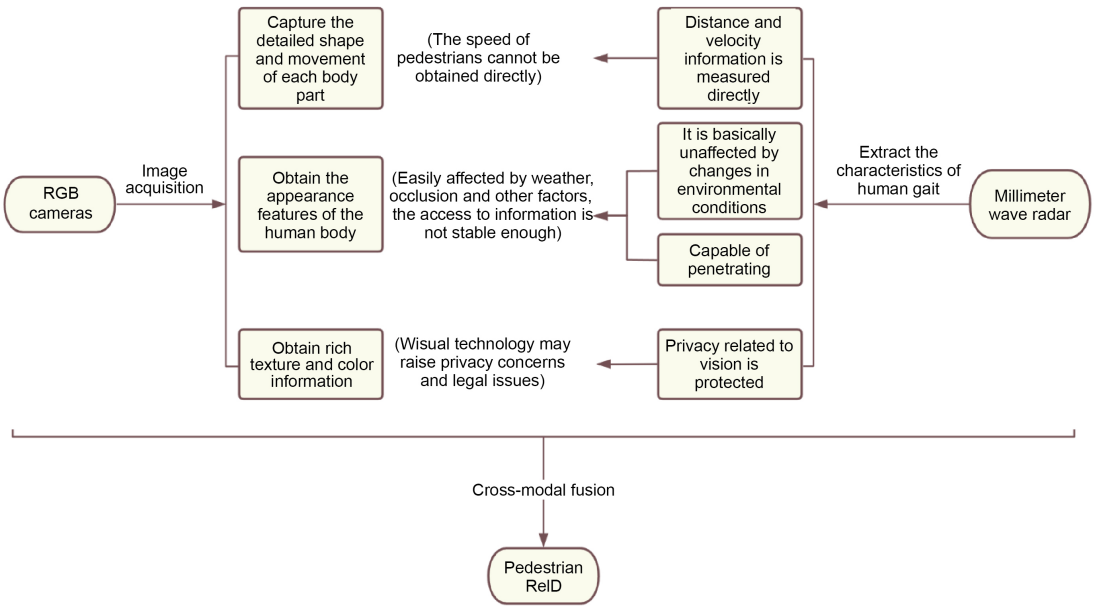

Figure 1. Conceptual diagram of mmWave radar and visual fusion for pedestrian re-identification.

The system integrates data from cameras (capturing appearance) and mmWave radar (capturing range, velocity, and operating in adverse conditions) to achieve more robust ReID performance compared to single-modality systems.

This review provides a comprehensive survey of the research landscape in mmWave-visual fused pedestrian ReID. The main contributions are:

1) A clear explanation of the fundamental concepts and the motivation behind fusing mmWave radar and visual data for ReID.

2) A detailed discussion and comparison of techniques for data acquisition, synchronization, calibration, and pre-processing specific to this multi-modal setup.

3) A systematic classification and critical analysis of existing fusion methods, highlighting their evolution and comparative performance.

4) An identification of key challenges and a discussion of promising future research directions.

This review aims to serve as a valuable resource for researchers entering the field and to stimulate further advancements in robust multi-modal perception systems. The subsequent sections delve into the fundamental theories (Section 2), data acquisition and processing (Section 3), fusion methodologies (Section 4), typical applications (Section 5), challenges and future directions (Section 6), and finally conclude the review (Section 7).

2. Fundamental Theories about Pedestrian Re-Identification

This section establishes the theoretical foundation necessary to understand mmWave-visual pedestrian ReID. We define pedestrian ReID and multi-modal ReID, followed by an analysis of the characteristics of visual and mmWave modalities and the rationale underpinning their fusion.

2.1. Pedestrian Re-Identification (ReID)

Pedestrian Re-identification is the task of associating observations of the same individual across a network of non-overlapping camera views [1]. Given a query image of a person of interest captured in one camera view, the objective is to retrieve all instances of the same person from a gallery set containing images from other camera views. It is fundamentally a matching problem, aiming to determine if two observations correspond to the same physical person despite variations in viewpoint, pose, illumination, occlusion, and background clutter. Early methods relied on hand-crafted features [8], while modern approaches predominantly utilize deep learning to learn discriminative feature representations directly from data [3] [4].

2.2. Multi-Modal Pedestrian ReID

Multi-Modal Pedestrian ReID extends the traditional ReID paradigm by leveraging information from multiple sensing modalities to improve matching accuracy and robustness [5]. While visual data (RGB) is the most common modality, it can be limited under challenging conditions. Integrating complementary sensor data,

such as depth maps (RGB-D), thermal imagery (RGB-T), or radio frequency signals (like mmWave radar), can provide additional cues to overcome the shortcomings of visual sensors alone. The core idea is that different modalities capture distinct aspects of a person or the environment, and their fusion can lead to a more comprehensive and resilient representation for matching [9].

2.3. Characteristics of Visual and mmWave Modalities

Understanding the individual strengths and weaknesses of visual cameras and mmWave radar is crucial for effective fusion.

- Visual Modality (Cameras): Cameras capture reflected light in the visible spectrum, providing rich information about the appearance of objects, including color, texture, and fine-grained shape details. This makes them highly effective for distinguishing individuals based on clothing, accessories, and general appearance under good conditions. However, they lack direct depth or velocity measurements and their performance is highly susceptible to environmental factors like illumination changes, shadows, glare, fog, rain, and occlusions.
- Millimeter-Wave (mmWave) Radar: mmWave radar emits radio waves and analyzes the reflected signals. Key characteristics include:
- Direct Range, Velocity Measurement and Angular Information: Provides precise measurements of the distance and radial velocity of detected points, offering valuable geometric and motion information. It also presents measures the angle of arrival of the reflected signals.
- Environmental Robustness: Largely unaffected by ambient light, weather conditions, or airborne particles.
- Susceptibility to Multipath Reflections: Radar signals can bounce off multiple surfaces, potentially creating ghost detections or inaccurate measurements in cluttered environments.
- Limited Resolution: Compared to cameras, radar offers lower spatial resolution, resulting in sparse point clouds that lack detailed shape or texture information. Distinguishing individuals based solely on radar data is challenging.

2.4. Necessity for Fusion

The complementary nature of visual and mmWave data forms the core rationale for their fusion in pedestrian ReID. Cameras excel at capturing appearance details crucial for identification in good conditions, while mmWave radar, which persists even when visual data is degraded, provides robust geometric and velocity information.

Fusion aims to leverage these complementary strengths:

- Enhanced Robustness: Radar data can compensate for visual failures caused by poor lighting, adverse weather, or occlusions. For instance, radar can still detect and track a pedestrian obscured by fog or rain where a camera would fail.
- Improved Discrimination: Combining appearance features with motion patterns and gait information can lead to more discriminative representations. Unique

walking patterns or body shape outlines from radar can help differentiate individuals who look similar in visual images.

- Richer Contextual Information: Fusing spatial information from both sensors can provide a more accurate 3D understanding of the scene and the pedestrian's position and trajectory within it.

Compared with systems that rely on a single modality, this combination ensures that the application system is more robust and reliable in complex real-world environments [10].

3. Data Acquisition and Processing

Successfully fusing mmWave radar and visual data for pedestrian ReID depends on data acquisition and processing. This section discusses the typical sensor setups, critical techniques for data synchronization and sensor calibration, and essential pre-processing steps for both modalities.

3.1. Sensor Setup and Hardware Configuration

A typical mmWave-visual ReID system involves one or more cameras and mmWave radar sensors deployed to cover the area of interest.

- Cameras: Standard RGB cameras are commonly used. Camera resolution, frame rate, lens type, and placement are crucial design parameters influencing visual data quality and coverage.
- mmWave Radar: Commercial off-the-shelf (COTS) mmWave radar sensors, often operating in the 77 - 81 GHz band (popular for automotive applications) or sometimes 24 GHz or 60 GHz bands, are frequently employed [6]. Key radar parameters include range resolution, velocity resolution, angular resolution, field of view, and update rate. Multiple radars might be used for wider coverage or improved localization accuracy through triangulation.
- Placement: Sensors are often co-located or placed with known geometric relationships. Optimal placement aims to maximize overlapping fields of view while minimizing mutual interference and ensuring pedestrians are adequately captured by both sensor types. **Figure 2** illustrates a common setup.

This configuration facilitates calibration and ensures that both sensors observe the same scene region, simplifying data association. The diagram should also present the subsequent processing steps.

3.2. Data Synchronization

Temporal alignment of data streams from cameras and radars is critical for accurate fusion. Mismatched timestamps can lead to incorrect associations between visual features and radar points corresponding to the same pedestrian at a given moment. Common synchronization techniques include:

- Hardware Triggering: A dedicated signal generator triggers both the camera exposure and radar chirp generation simultaneously. This offers the highest precision but requires specialized hardware and physical connections.

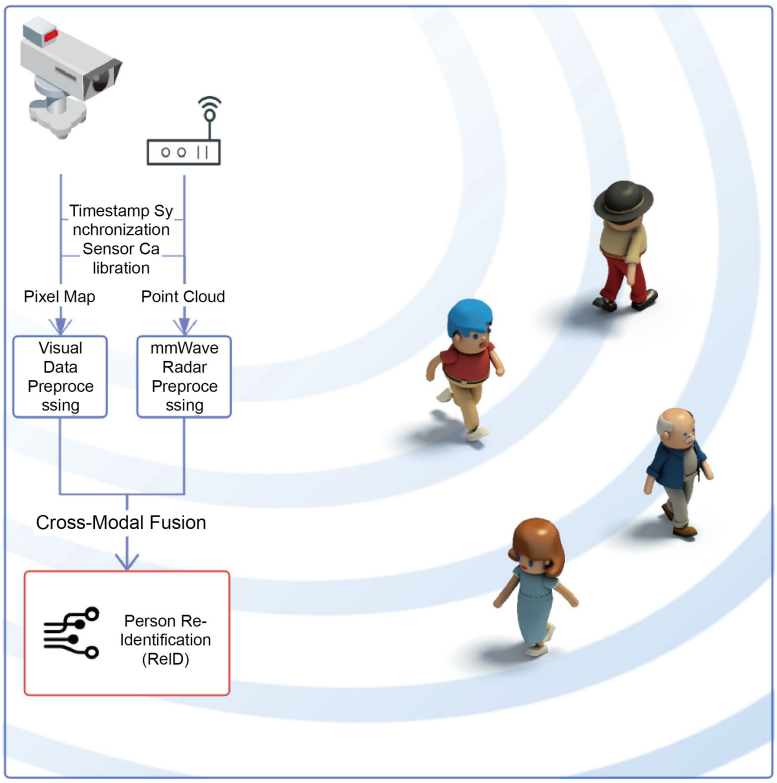

Figure 2. Example of a co-located mmWave radar and camera sensor setup.

- Software Synchronization: Sensors timestamp their data using their internal clocks. A central processing unit collects data and aligns streams based on these timestamps. Network time protocol (NTP) or precision time protocol (PTP) can be used to synchronize system clocks across a network, improving accuracy over unsynchronized clocks but potentially suffering from network latency and jitter.
- Data-Driven Synchronization: Analyzing correlations in the data itself (e.g. correlating motion patterns detected by both sensors) can sometimes be used for coarse alignment, often as a refinement step after initial clock-based synchronization.

3.3. Sensor Calibration

Spatial alignment, or calibration, is necessary to establish the geometric relationship between the coordinate systems of the camera(s) and the radar(s). This allows projecting radar points onto the image plane or transforming visual features into the radar's coordinate system, enabling data association and fusion.

- Camera Intrinsic Calibration: Determines the internal parameters of the camera. Standard methods like Zhang's chessboard pattern calibration [11] are widely used.
- Radar-Camera Extrinsic Calibration: Determines the rigid body transformation between the radar's coordinate system and the camera's coordinate system. Common approaches include:

- Target-Based: Using specific calibration objects (e.g. corner reflectors, spheres, calibration boards) visible to both sensors [12]. The 3D position of the target measured by radar is matched with its corresponding 2D projection in the camera image to solve for the extrinsic parameters.
- Targetless: Exploiting natural features or motion patterns in the scene that are detectable by both sensors, avoiding the need for dedicated calibration targets [13]. This is often more practical but can be less accurate.

In multi-sensor systems, the collaboration between millimeter-wave radar and RGB cameras faces three core challenges: calibration, synchronization, and data association. Hardware-level synchronization can achieve microsecond-level precision but is costly, while software synchronization (NTP/PTP) requires motion compensation algorithms to correct the data misalignment caused by millisecond-level delays. In terms of external parameter calibration, traditional checkerboard calibration has high accuracy but is limited to static environments, whereas targetless calibration is flexible but relies on rich natural feature matching. In data association, the modal differences between the radar's sparse point cloud and the camera's dense image can easily lead to target mismatches, necessitating the fusion of probabilistic models and deep learning features to enhance consistency. In practical use, it is essential to balance cost and performance: first, use a calibration board for precise calibration and automatically adjust during operation; prioritize hardware synchronization, but if the budget is insufficient, resort to software synchronization with algorithm compensation. Ultimately, success depends on the collaboration between hardware and algorithms to adapt to environmental changes in real-time.

3.4. Data Pre-Processing

Raw data from cameras and radar require significant pre-processing before fusion.

Visual Data Pre-processing includes the following context:

- Pedestrian Detection: Identify pedestrian locations in the image using object detectors (e.g. YOLO [14], Faster R-CNN [15]).
- Bounding Box Generation: Extract image patches corresponding to detected pedestrians.
- Background Subtraction: Isolate foreground pixels corresponding to moving objects.
- Normalization: Adjust image brightness, contrast, and size.

mmWave Radar Data Pre-processing including the following context:

- Point Cloud Generation: Radar signal processing (FFT, CFAR detection) generates a point cloud, where each point has coordinates (x, y, z), radial velocity (v).
- Noise Filtering: Remove spurious detections caused by noise or clutter using techniques like density-based spatial clustering of applications with noise (DBSCAN) [16] or statistical outlier removal.

- Clustering: Group radar points belonging to the same object. Algorithms like DBSCAN or Euclidean clustering are common.
- Tracking: Associate detections over time using filters like Kalman Filters [17] or particle filters to estimate pedestrian trajectories and smooth velocity measurements. This can help group points belonging to a single pedestrian over time.

Effective pre-processing cleans the raw data, extracts relevant information (e.g. pedestrian bounding boxes, clustered radar points associated with pedestrians), and prepares it for subsequent fusion stages.

4. Classification and Evolution of Pedestrian Re-Identification Methods Using Millimeter-Wave Radar and Visual Fusion

Pedestrian re-identification (ReID) is a critical task in computer vision and sensor fusion, aiming to recognize and track the same individual across different views or time instances. This section classifies fusion methods into data-level, feature-level, and decision-level approaches, reviews representative studies, and discusses their evolution, supported by comparative tables.

4.1. Data-Level Fusion

Data-level fusion involves merging raw or minimally processed data from mmWave radar and visual sensors before detection or identification. This approach leverages radar's precise localization to generate regions of interest (ROI) in images, which are then processed for re-identification. Early methods relied on straightforward data integration, but recent advancements incorporate deep learning for enhanced performance.

One of the earliest works, Milch and Behrens [18], proposed a method that uses a radar-generated target list to define the ROIs, and then verifies the pedestrian through visual profile analysis. This approach was computationally efficient but limited by the simplicity of visual features. Guo *et al*. [19] advanced this by introducing intra-frame clustering and inter-frame tracking, using radar data to filter noise and guide visual confirmation. Their method improved robustness in noisy environments.

More recently, Wang *et al*. [20] developed a three-layer fusion model integrating radar and monocular vision, using a visual attention mechanism to prioritize ROIs. This method enhanced detection in dynamic scenes. Similarly, Streubel and Yang [21] explored data-level fusion for indoor pedestrian tracking, projecting radar points onto stereo camera images. Their approach achieved high localization accuracy but required precise sensor calibration.

Data-level fusion excels in leveraging raw data complementarity but faces challenges in computational complexity and sensor synchronization. Recent studies, such as Plascencia *et al*. [22], have begun integrating deep learning into process fused data, improving scalability. **Table 1** compares key data-level fusion methods.

Table 1. Comparison of data-level fusion methods.

Study	Year	Technique	Application	Strengths	Weaknesses	Reference
Milch and Behrens	2001	Radar target lists, visual contour	Detection	Simple, low computation	Limited feature complexity	[18]
Guo *et al.*	2018	Clustering, visual confirmation	Detection/ Tracking	Noise reduction	Calibration sensitivity	[19]
Wang *et al.*	2011	Three-layer fusion, attention	Detection	Dynamic scene handling	High computation	[20]
Streubel and Yang	2016	Radar projection, stereo vision	Tracking	High accuracy	Calibration complexity	[21]
Plascencia *et al.*	2023	Deep learning fusion	Detection	Scalable	Data requirements	[22]

4.2. Feature-Level Fusion

Feature-level fusion extracts features from radar and visual data separately, and then combines them for re-identification. This approach has gained prominence with deep learning, enabling models to learn complex feature representations and fusion strategies dynamically.

Nobis *et al.* [23] proposed a deep learning architecture, which fuses projected radar data with camera images within network layers. This method improved 2D detection accuracy by learning optimal fusion levels. Plascencia *et al.* [22] extended this concept, transforming radar and lidar data into 2D grayscale images and fusing them with RGB images using a SegNet-based network. Their approach enhanced pedestrian detection in cluttered environments.

In addition to that, attention mechanisms have further refined feature-level fusion. Li *et al.* [24] introduced an attention-based network for pedestrian liveness detection, combining radar cross-section (RCS) features with visual data to distinguish real pedestrians from static images. Similarly, Yu *et al.* [25] proposed a dual cross-attention module (DCAM) for feature fusion, initially for vehicle detection but adaptable to pedestrians. Liu *et al.* [26] developed a multi-modal network integrating radar gait features with visual appearance, achieving robust re-identification in occluded scenes.

Feature-level fusion benefits from deep learning's ability to model complex relationships but requires substantial training data and computational resources. **Table 2** summarizes key methods.

Table 2. Comparison of feature-level fusion methods.

Study	Year	Technique	Application	Strengths	Weaknesses	Reference
Nobis *et al.*	2019	CRF-Net, deep fusion	Detection	Adaptive fusion	Data-intensive	[23]
Plascencia *et al.*	2023	SegNet, grayscale fusion	Detection	Clutter robustness	Computation-heavy	[22]
Li *et al.*	2022	Attention, RCS features	Liveness Detection	High specificity	Training complexity	[24]
Yu *et al.*	2025	DCAM fusion	Detection	Flexible	Limited pedestrian focus	[25]
Liu *et al.*	2024	Gait and visual fusion	ReID	Occlusion handling	Data requirements	[26]

4.3. Decision-Level Fusion

Decision-level fusion involves independent detection or identification by each sensor, followed by result integration. This modular approach is robust to sensor failures and suitable for real-time applications.

Yang *et al*. [27] proposed a decision-level fusion method using an unscented Kalman filter (UKF) for radar data and YOLOv5 with DeepSORT for visual tracking, matching targets in polar coordinates. This method achieved high tracking precision. An earlier study [28] introduced a multi-sensor tracking algorithm with back-projection and multi-hypothesis association, enhancing trajectory accuracy. Zhao *et al*. [29] focused on nighttime detection, combining infrared vision and radar with an improved YOLOv5 and extended Kalman filter.

Additional studies include Graves *et al*. [30], who used decision-level fusion for pedestrian collision warning, integrating radar localization with visual classification, and Zhu *et al*. [31], who developed a track-to-track fusion method for multi-pedestrian tracking. These methods prioritize simplicity and robustness but may miss early-stage data synergies. **Table 3** compares key approaches.

Table 3. Comparison of decision-level fusion methods.

Study	Year	Technique	Application	Strengths	Weaknesses	Reference
Yang *et al*.	2023	UKF, YOLOv5/DeepSORT	Tracking	High precision	Limited synergy	[27]
Cui *et al*.	2021	Back-projection, association	Tracking	Trajectory accuracy	Complexity	[28]
Zhao *et al*.	2023	YOLOv5, Kalman filter	Nighttime Detection	Low-light robustness	Data association	[29]
Graves *et al*.	2022	Radar-visual matching	Collision Warning	Simplicity	Limited integration	[30]
Zhu *et al*.	2022	Track-to-track fusion	Tracking	Multi-target handling	Synchronization	[31]

4.4. Evolution of Methods

The evolution of mmWave radar and visual fusion for pedestrian ReID reflects advancements in sensor technology and algorithms. Early methods (2000s) focused on data-level fusion, using radar to guide visual processing in simple scenarios. The 2010s saw decision-level fusion gain traction for its robustness, as seen in collision warning systems. Since 2015, feature-level fusion has dominated, driven by deep learning's ability to model complex multi-modal relationships. Future directions may involve end-to-end deep learning models and efficient handling of sparse radar data, addressing real-time constraints in autonomous systems.

4.5. Typical System Analysis

To better illustrate the fusion approach, we present a detailed explanation of several representative re-identification methods. This will facilitate a more comprehensive understanding of the advantages offered by millimeter-wave radar and vision fusion for re-identification tasks.

Zheng *et al*. [1] proposed the pedestrian alignment network (PAN) that utilizes two convolutional branches (the base branch and the alignment branch) and an affine estimation branch to simultaneously address pedestrian alignment and recognition issues. The base branch deploys a pre-trained ResNet-50 model on ImageNet and removes the final fully connected (FC) layer. The alignment branch consists of 3 ResBlocks and 1 average pooling layer, also adding an FC layer to predict multiclass probabilities. The affine estimation branch receives two activated input tensors from the base branch. The Res4 Feature Maps contain shallow feature maps of the original image, reflecting local pattern information; the Res2 Feature Maps are closer to the classification layer and encode attention and semantic cues for pedestrian recognition.

Zheng *et al*. [2] proposed a deep learning model that combines the advantages of verification models and recognition models. Establishing relationships through pairwise comparisons, such as partial matching and contrastive loss, is performed. Contrastive loss directly computes the Euclidean distance between two embeddings. In the recognition model, there exists an implicit relationship between the learned embeddings constructed using cross-entropy loss. The cross-entropy loss can be used. When the directions of the embedding vectors are similar, the network converges, thereby maintaining the similarity of the embeddings. The proposed model simultaneously utilizes both types of loss functions and benefits from pre-training on ImageNet, thereby overcoming the limitations of a single model.

In the performance evaluation of data-level fusion methods, computational complexity and real-time indicators have significant advantages. Regarding the issue of computational complexity, existing research mainly focuses on two dimensions: noise suppression and accuracy optimization. For example, Yu *et al*. [25] utilized modules to reduce technical complexity, providing a reference technical path for complexity control. In terms of real-time performance, most studies present a processing delay of less than 50 ms, which is largely attributed to the inherent characteristics of data-level fusion due to operating directly at the raw data layer, thus avoiding the time overhead associated with higher-level processing, such as feature extraction and decision reasoning.

Zheng *et al*. [1] adopted a multi-branch architecture (base branch, aligned branch, affine estimated branch), which has a higher complexity, resulting in a surge in memory and computation, and its multi-task joint optimization (recognition, alignment, and feature learning) further increases the training complexity. In addition, due to high-resolution feature map processing and multi-branch parallel computing, PAN has higher hardware requirements, resulting in large inference delays. The model proposed by Zheng *et al*. [2] fused the comparison loss of the verification model (based on Euclidean distance) and the cross-entropy loss of the recognition model, although it is manifested in the complexity of dual-objective optimization and high-dimensional embedding calculation. The transfer learning of the ImageNet pre-trained model (such as ResNet) reduces the training cost. In general, the model proposed by Zheng *et al*. [2] has made a breakthrough in com-

plexity and real-time, which can be used as a good reference for similar processing of complexity and real-time in this project to achieve more accurate pedestrian re-identification.

Although the constructed models of the two are different, both have improved the accuracy of pedestrian re-identification, providing certain insights and guidance for the fusion re-identification system.

5. Typical Applications of Pedestrian Re-Identification Methods Using Millimeter-Wave Radar and Visual Fusion

Multi-modal perception technology utilizing millimeter-wave (mmWave) radar has increasingly become a significant focus and an important field of study [32], which supports critical applications in autonomous driving, smart cities, and surveillance. Object detection and tracking based on radar-camera fusion have also gained growing attention [33]. We investigate these references, hoping to make a greater contribution to research on fusion-based pedestrian re-identification.

5.1. Nighttime Pedestrian Detection

Nighttime or low-light conditions challenge visual sensors, but radar's penetration ability ensures reliability. Zhao *et al*. [29] proposed a decision-level fusion framework using infrared vision and radar, achieving high accuracy in dark environments. Similarly, Zhang *et al*. [11] combined thermal imaging with radar for nighttime ReID, improving robustness.

5.2. Pedestrian Tracking

Real-time pedestrian tracking supports path planning and collision avoidance. Yang *et al*. [27] developed a decision-level fusion method using UKF and DeepSORT, ensuring precise trajectory tracking. Zhu *et al*. [31] introduced track-to-track fusion for multi-pedestrian scenarios, handling occlusions effectively. These methods enable vehicles to anticipate pedestrian movements.

5.3. Impact and Future Directions

Fusion-based ReID methods significantly enhance safety and reliability in autonomous systems. Studies like [28] demonstrate reduced false positives compared to single-sensor approaches. Future advancements may leverage datasets and focus on real-time processing and adverse weather performance.

6. Key Technical Challenges and Future Research Directions

Despite the promising potential of fusing mmWave radar and visual data for robust pedestrian ReID, several significant technical challenges hinder its widespread adoption and performance optimization. Addressing these challenges constitutes key future research directions.

6.1. Key Technical Challenges

1) Lack of Dedicated Benchmarks: The absence of large-scale, diverse, and publicly available datasets specifically designed for mmWave-visual pedestrian ReID (with ground-truth IDs across non-overlapping views under various conditions) is arguably the biggest obstacle. This impedes standardized evaluation, fair comparison of methods, and the training of data-hungry deep learning models.

2) Data Heterogeneity and Representation: Effectively fusing the sparse, geometric, and point cloud data from radar with the dense, semantic, and appearance-rich pixel data from cameras remains fundamentally challenging. Finding optimal representations for radar data that facilitate effective fusion with visual features is crucial.

3) Radar Data Quality and Interpretation: mmWave radar data can be noisy, suffer from multipath reflections (ghost targets), and have low angular resolution compared to cameras. Sparsity makes it difficult to infer detailed shapes or associate points reliably to specific body parts for fine-grained gait analysis or ReID, especially in crowds. Extracting discriminative features solely from sparse radar points is non-trivial.

4) Complexity vs. Real-Time Performance: Sophisticated hybrid fusion models (e.g. using transformers or complex attention mechanisms) often achieve better performance but come with high computational costs, making real-time deployment on resource-constrained platforms (like robots or edge devices) challenging.

5) Generalization and Domain Adaptation: Models trained on data from one specific sensor setup, environment, or weather condition may not generalize well to others (domain shift). Variability in radar hardware, camera types, environmental clutter, and pedestrian densities poses significant generalization challenges.

6.2. Future Research Directions

1) Benchmark Dataset Development: Creating large-scale, diverse mmWave-visual pedestrian ReID datasets covering various environments (indoor/outdoor), weather conditions, pedestrian densities, and sensor viewpoints is paramount. Including annotations for persistent IDs across non-overlapping views is essential.

2) Advanced Fusion Architectures: Exploring novel deep learning architectures tailored for heterogeneous sensor fusion. This includes investigating more advanced transformer variants, graph neural networks specifically designed for radar point cloud structure and radar-visual interaction, and perhaps integrating neural rendering techniques to bridge the modality gap.

3) Exploiting Richer Radar Information: Moving beyond basic point clouds (x, y, z, v). Research into effectively incorporating radar micro-Doppler signatures for gait recognition utilizing the full Range-Azimuth-Elevation-Doppler radar tensor, or learning discriminative features directly from raw radar ADC data holds significant promise.

4) Real-Time Optimization: Studying model compression, quantization, know-

ledge distillation, and efficient network architectures to enable real-time execution of complex fusion models on edge devices.

5) Cross-Modal Adaptation: Developing geometry-aware fusion models with 3D pose estimation to overcome the core challenge of view-invariant matching in non-overlapping sensor configurations, while addressing domain shift through adversarial feature representation.

Overcoming these challenges and pursuing these research directions will be crucial for unlocking the full potential of mmWave-visual fusion and realizing truly robust and reliable pedestrian ReID systems for real-world applications.

7. Conclusions

Pedestrian re-identification is a critical technology for intelligent systems, but traditional visual methods struggle in challenging real-world conditions. This review has surveyed the emerging field of mmWave-visual fusion for pedestrian ReID, motivated by the complementary strengths of cameras and mmWave radar.

We began by outlining the fundamental concepts of ReID, multi-modal ReID, and the characteristics of the visual and mmWave modalities, establishing the strong rationale for their fusion. We then discussed the essential practical aspects of data acquisition, including sensor setup, synchronization, calibration, and preprocessing techniques crucial for successful integration. The core of the review provided a systematic classification and analysis of fusion methodologies, tracing their evolution from early and late fusion approaches to the currently dominant intermediate/hybrid fusion strategies, particularly those leveraging deep learning, attention mechanisms, and graph networks. We highlighted the importance of experimental validation and explained the lack of dedicated ReID benchmarks. Finally, we identified key technical challenges, including data scarcity, heterogeneity, radar limitations, complexity, and generalization issues. Based on these challenges, we proposed promising future research directions, emphasizing benchmark creation, advanced fusion architectures, richer radar feature utilization, self-supervised learning, and real-time optimization.

In conclusion, fusing mmWave radar and visual data offers a compelling pathway towards achieving robust, all-weather, and reliable pedestrian re-identification systems. While significant challenges remain, the ongoing advancements in sensor technology, deep learning, and multi-modal fusion techniques promise substantial progress in this important research area, paving the way for more capable perception systems in surveillance, robotics, and autonomous driving.

Acknowledgements

The work is funded by the foundation of the Innovation and Entrepreneurship Training Program for College Students (202410424057).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Zheng, Z., Zheng, L. and Yang, Y. (2019) Pedestrian Alignment Network for Large-Scale Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*, **29**, 3037-3045. https://doi.org/10.1109/tcsvt.2018.2873599

[2] Zheng, Z., Zheng, L. and Yang, Y. (2017) A Discriminatively Learned CNN Embedding for Person Reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, **14**, 1-20. https://doi.org/10.1145/3159171

[3] Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L. and Hoi, S.C.H. (2022) Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 2872-2893. https://doi.org/10.1109/tpami.2021.3054775

[4] Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., *et al*. (2020) A Strong Baseline and Batch Normalization Neck for Deep Person Re-Identification. *IEEE Transactions on Multimedia*, **22**, 2597-2609. https://doi.org/10.1109/tmm.2019.2958756

[5] Uddin, M.K., Bhuiyan, A., Bappee, F.K., Islam, M.M. and Hasan, M. (2023) Person Re-Identification with RGB-D and RGB-IR Sensors: A Comprehensive Survey. *Sensors*, **23**, Article No. 1504. https://doi.org/10.3390/s23031504

[6] Bartsch, A., Fitzek, F. and Rasshofer, R.H. (2012) Pedestrian Recognition Using Automotive Radar Sensors. *Advances in Radio Science*, **10**, 45-55. https://doi.org/10.5194/ars-10-45-2012

[7] Cui, F., Zhang, Q., Wu, J., Song, Y., Xie, Z., Song, C., *et al*. (2023) Online Multipedestrian Tracking Based on Fused Detections of Millimeter Wave Radar and Vision. *IEEE Sensors Journal*, **23**, 15702-15712. https://doi.org/10.1109/jsen.2023.3255924

[8] Gray, D. and Tao, H. (2008) Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. 10*th European Conference on Computer Vision*, Marseille, 12-18 October 2008, 262-275. https://doi.org/10.1007/978-3-540-88682-2_21

[9] Ye, M., Wang, Z., Lan, X. and Yuen, P.C. (2018) Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking. *Proceedings of the* 27*th International Joint Conference on Artificial Intelligence*, Stockholm, 13-19 July 2018, 1092-1099. https://doi.org/10.24963/ijcai.2018/152

[10] Yao, S., Guan, R., Huang, X., Li, Z., Sha, X., Yue, Y., *et al*. (2024) Radar-Camera Fusion for Object Detection and Semantic Segmentation in Autonomous Driving: A Comprehensive Review. *IEEE Transactions on Intelligent Vehicles*, **9**, 2094-2128. https://doi.org/10.1109/tiv.2023.3307157

[11] Zhang, Z. (2000) A Flexible New Technique for Camera Calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 1330-1334. https://doi.org/10.1109/34.888718

[12] Oh, J., Kim, K., Park, M. and Kim, S. (2018) A Comparative Study on Camera-Radar Calibration Methods. 2018 15*th International Conference on Control, Automation, Robotics and Vision* (*ICARCV*), Singapore, 18-21 November 2018, 1057-1062. https://doi.org/10.1109/icarcv.2018.8581329

[13] Liu, X., Deng, Z. and Zhang, G. (2025) Targetless Radar-Camera Extrinsic Parameter Calibration Using Track-to-Track Association. *Sensors*, **25**, Article No. 949. https://doi.org/10.3390/s25030949

[14] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. arXiv: 1804.02767. https://arxiv.org/abs/1804.02767

[15] Ren, S., He, K., Girshick, R. and Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Anal-*

ysis and Machine Intelligence, **39**, 1137-1149. https://doi.org/10.1109/tpami.2016.2577031

[16] Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, 2-4 August 1996, 226-231. https://dl.acm.org/doi/10.5555/3001460.3001507

[17] Kalman, R.E. (1960) A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, **82**, 35-45. https://doi.org/10.1115/1.3662552

[18] Milch, S. and Behrens, M. (2001) Pedestrian Detection with Radar and Computer Vision. *Proceedings of PAL* 2001-*Progress in Automobile Lighting*, *Held Laboratory of Lighting Technology*, Vol. 9, 657-664.

[19] Guo, X., Du, J., Gao, J. and Wang, W. (2018) Pedestrian Detection Based on Fusion of Millimeter Wave Radar and Vision. *Proceedings of the* 2018 *International Conference on Artificial Intelligence and Pattern Recognition*, Beijing, 18-20 August 2018, 38-42. https://doi.org/10.1145/3268866.3268868

[20] Wang, T., Zheng, N., Xin, J. and Ma, Z. (2011) Integrating Millimeter Wave Radar with a Monocular Vision Sensor for On-Road Obstacle Detection Applications. *Sensors*, **11**, 8992-9008. https://doi.org/10.3390/s110908992

[21] Streubel, R. and Yang, B. (2016) Fusion of Stereo Camera and MIMO-FMCW Radar for Pedestrian Tracking in Indoor Environments. 2016 19*th International Conference on Information Fusion* (*FUSION*), Heidelberg, 5-8 July 2016, 565-572. https://ieeexplore.ieee.org/document/7527938

[22] Plascencia, A.C., García-Gómez, P., Perez, E.B., DeMas-Giménez, G., Casas, J.R. and Royo, S. (2023) A Preliminary Study of Deep Learning Sensor Fusion for Pedestrian Detection. *Sensors*, **23**, Article No. 4167. https://doi.org/10.3390/s23084167

[23] Nobis, F., Geisslinger, M., Weber, M., Betz, J. and Lienkamp, M. (2019) A Deep Learning-Based Radar and Camera Sensor Fusion Architecture for Object Detection. 2019 *Sensor Data Fusion*: *Trends*, *Solutions*, *Applications* (*SDF*), Bonn, 15-17 October 2019, 1-7. https://doi.org/10.1109/sdf.2019.8916629

[24] Li, H., Liu, R., Wang, S., Jiang, W. and Lu, C.X. (2022) Pedestrian Liveness Detection Based on mmWave Radar and Camera Fusion. 2022 19*th Annual IEEE International Conference on Sensing*, *Communication*, *and Networking* (*SECON*), Stockholm, 20-23 September 2022, 262-270. https://doi.org/10.1109/secon55815.2022.9918553

[25] Yu, X., Hu, T. and Zhu, H. (2025) Roadside Perception Applications Based on DCAM Fusion and Lightweight Millimeter-Wave Radar-Vision Integration. *Electronics*, **14**, Article No. 1576. https://doi.org/10.3390/electronics14081576

[26] Liu, R., Yao, T., Shi, R., Mei, L., Wang, S., Yin, Z., *et al*. (2024) Mission: mmWave Radar Person Identification with RGB Cameras. *Proceedings of the* 22*nd ACM Conference on Embedded Networked Sensor Systems*, Hangzhou, 4-7 November 2024, 309-321. https://doi.org/10.1145/3666025.3699340

[27] Yang, C., Huan, S., Wu, L., Weng, Q. and Xiong, W. (2023) Fusion of Millimeter-Wave Radar and Camera Vision for Pedestrian Tracking. 2023 5*th International Conference on Communications*, *Information System and Computer Engineering* (*CISCE*), Guangzhou, 14-16 April 2023, 317-321. https://doi.org/10.1109/cisce58541.2023.10142444

[28] Cui, F., Song, Y., Wu, J., Xie, Z., Song, C., Xu, Z., *et al*. (2021) Online Multi-Target Tracking for Pedestrian by Fusion of Millimeter Wave Radar and Vision. 2021 *IEEE Radar Conference* (*RadarConf*21), Atlanta, 7-14 May 2021, 1-6. https://doi.org/10.1109/radarconf2147009.2021.9455185

[29] Zhao, W., Wang, T., Tan, A. and Ren, C. (2023) Nighttime Pedestrian Detection Based

on a Fusion of Visual Information and Millimeter-Wave Radar. *IEEE Access*, **11**, 68439-68451. https://doi.org/10.1109/access.2023.3291398

[30] Graves, K., Kanwal, M., Yu, X. and Saniie, J. (2022) Design Flow of mmWave Radar and Machine Vision Fusion for Pedestrian Collision Warning. 2022 *IEEE International Conference on Electro Information Technology* (*eIT*), Mankato, 19-21 May 2022, 176-181. https://doi.org/10.1109/eit53891.2022.9813942

[31] Zhu, Y., Wang, T. and Zhu, S. (2022) Adaptive Multi-Pedestrian Tracking by Multi-Sensor: Track-to-Track Fusion Using Monocular 3D Detection and MMW Radar. *Remote Sensing*, **14**, Article No. 1837. https://doi.org/10.3390/rs14081837

[32] Wang, S., Mei, L., Liu, R., Jiang, W., Yin, Z., Deng, X., *et al*. (2025) Multi-Modal Fusion Sensing: A Comprehensive Review of Millimeter-Wave Radar and Its Integration with Other Modalities. *IEEE Communications Surveys & Tutorials*, **27**, 322-352. https://doi.org/10.1109/comst.2024.3398004

[33] Shi, K., *et al*. (2024) Radar and Camera Fusion for Object Detection and Tracking: A Comprehensive Survey. arXiv: 2410.19872. https://doi.org/10.48550/arXiv.2410.19872