

Article

DUAL-Net: Joint Domain-Invariant and User-Adaptive Feature Learning for Gesture Recognition

Shuangjiao Zhai ^{1,*} , Bo Yang ^{2,*}, Zixin Dai ¹, Yujie Guo ¹ , Baojin Jing ¹, Jia Qin ¹ and Pinle Qin ¹

¹ School of Computer Science and Technology, North University of China, Taiyuan 030051, China; s202307006@st.nuc.edu.cn (Z.D.); s202407021@st.nuc.edu.cn (Y.G.); s202407026@st.nuc.edu.cn (B.J.); jqin@nuc.edu.cn (J.Q.); qpl@nuc.edu.cn (P.Q.)

² School of Mechanical Engineering, Taiyuan University of Technology, Taiyuan 030600, China

* Correspondence: sjzhai@nuc.edu.cn (S.Z.); yangbo0048@link.tyut.edu.cn (B.Y.)

Abstract

Human activity recognition has become an important component of human–computer interaction and ubiquitous computing. Among various sensing technologies, WiFi-based gesture recognition has attracted increasing attention due to its contactless nature and robustness to visual occlusion. However, environmental variations and user-specific differences often lead to significant performance degradation, particularly in cross-user scenarios. Existing methods primarily focus on learning domain-invariant representations, which may overlook user-specific characteristics that are essential for accurate recognition. To address this issue, we propose the Domain-invariant and User-Adaptive Learning Network (DUAL-Net), a dual-branch framework that jointly models domain-invariant and user-adaptive representations. Specifically, DUAL-Net incorporates a contrastive fusion learning (CFL) module with modality-specific encoders to learn complementary representations from WiFi and vision modalities. Furthermore, a spatial matrix difference (SMD)-guided cross-modal generation (CMG) module is introduced to generate user-adaptive WiFi features by incorporating structural priors derived from skeletal representations. To improve deployment efficiency, DUAL-Net adopts a two-stage learning framework, where adaptation is conducted offline to reduce online computational overhead. Experiments on the MM-Fi dataset and a self-collected dataset show that DUAL-Net achieves superior cross-user recognition performance compared with existing single-modality and multimodal methods. In addition, SMD-guided conditioning improves recognition accuracy by up to 8.79% over diffusion generation without structural guidance.

Keywords: WiFi CSI; WiFi–vision fusion; gesture recognition; domain-invariant learning; cross-user adaptation; diffusion generation



Academic Editor: Anthony Fleury

Received: 9 May 2026

Revised: 20 June 2026

Accepted: 26 June 2026

Published: 2 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Human activity recognition (HAR) is vital for intelligent interaction in fields such as smart homes [1–3], healthcare [4], and autonomous driving [5]. Accurate behavior recognition significantly enhances user experience, system efficiency, and safety. Recent advances have explored multiple sensing modalities, notably vision-based, sensor-based, and WiFi-based approaches [6–9]. Vision-based techniques excel in capturing fine-grained spatial information but remain vulnerable to illumination changes, occlusions, and viewpoint limitations. Sensor-based methods leveraging accelerometers and gyroscopes reliably record motion data but require users to wear devices, which reduces practicality and comfort.

In contrast, WiFi-based HAR has attracted increasing attention for its contactless sensing, affordability, and resilience to visual constraints. However, WiFi-based HAR approaches are still highly susceptible to environmental interference, such as background variations and user location [10]. Moreover, existing WiFi-based HAR methods typically emphasize domain-invariant features to improve robustness, yet they often overlook user-dependent variations in gesture execution style, speed, and amplitude. This omission limits their ability to generalize in real-world applications [11].

Motivated by these limitations, recent studies have investigated domain adaptation techniques and user-independent approaches, aiming to extract stable features across diverse environments and generalize across users [12–14]. However, these two lines of work typically address domain variations and user differences in isolation, overlooking the intricate interplay between environmental conditions and individual behavioral patterns. In addition, WiFi signals inherently lack the spatial resolution required to capture fine-grained, user-specific gesture details, which further constrains recognition performance in practical scenarios.

To address these limitations, multimodal learning has emerged as a promising direction that combines complementary sensing modalities to enhance feature robustness [15]. For instance, it has been successfully applied in domains such as audio–visual sentiment analysis [16], vision–language navigation [17], and WiFi–vision-based gait recognition [18]. In particular, WiFi–vision-based gesture recognition integrates the high-resolution spatial sensing of vision sensors with the robustness of WiFi signals against occlusion and illumination changes, effectively overcoming key limitations of single-modality systems [19,20]. However, WiFi–vision fusion still faces critical challenges similar to those in broader multimodal learning. In particular, structural and temporal mismatches between modalities, along with incomplete or missing data in real-world deployments, often lead to unstable fusion and degraded recognition performance [21,22]. These limitations are further compounded by the need to adapt to unseen users, where collecting personalized WiFi data is costly and impractical at scale.

To address these challenges, we propose the Domain-invariant and User-Adaptive Learning Network (DUAL-Net), a dual-branch framework that jointly models domain-invariant and user-adaptive representations. As illustrated in Figure 1, existing methods primarily emphasize domain-invariant representations, which may overlook user-specific characteristics that are essential for accurate recognition. DUAL-Net addresses this limitation by complementing domain-invariant representations with user-adaptive representations, enabling the model to capture both shared gesture semantics and individual user characteristics.

The first branch is a contrastive fusion learning (CFL) module designed to learn domain-invariant representations. Rather than directly enforcing cross-modal alignment, CFL employs modality-specific encoders to extract complementary information from WiFi and vision modalities and projects them into a shared semantic space. Contrastive learning is then applied to encourage representations of the same gesture class to remain close while separating those of different classes. This process enables the model to learn stable gesture semantics while reducing the influence of environmental variations.

However, domain-invariant representations alone are insufficient for modeling user-specific differences. To address this issue, DUAL-Net introduces a spatial matrix difference (SMD)-guided cross-modal generation (CMG) module. Specifically, skeletal keypoints extracted from vision data are transformed into spatial matrices to characterize structural differences between source and target users. The resulting SMD quantifies user-specific discrepancies and is incorporated as a conditioning signal to guide diffusion-based feature generation. By introducing structural priors into the generation process, CMG synthesizes

user-adaptive WiFi representations for unseen users without requiring additional WiFi data collection. Finally, the domain-invariant and user-adaptive representations are fused for gesture recognition.

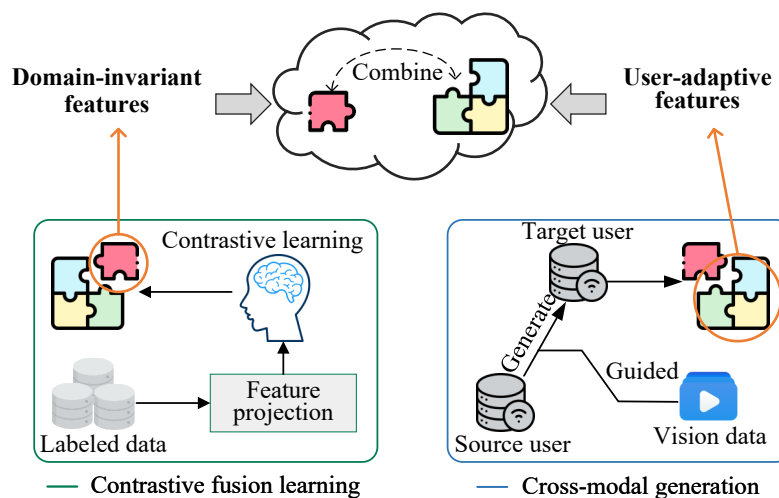


Figure 1. The workflow of DUAL-Net. Existing methods emphasize domain-invariant features, while DUAL-Net additionally learns user-adaptive features to capture individual differences.

To improve deployment efficiency, DUAL-Net adopts a two-stage learning framework. The CFL and CMG modules are performed during offline learning, while online inference avoids repeated multimodal adaptation and diffusion-based feature generation, thereby reducing computational overhead. We evaluate the proposed DUAL-Net framework on two datasets, namely the publicly available MM-Fi dataset and a self-collected multimodal gesture dataset. Experimental results demonstrate that DUAL-Net significantly outperforms single-modality baselines by 24.7% in recognition accuracy and surpasses existing multimodal approaches by 14.5%. Furthermore, it exhibits strong generalization, achieving over 90% accuracy in both cross-user and cross-domain scenarios.

The key contributions of this paper are summarized as follows:

- We propose DUAL-Net, a dual-branch WiFi–vision gesture recognition framework that jointly models domain-invariant and user-adaptive representations, addressing the limitations of domain-invariant feature learning in cross-user scenarios.
- We design a contrastive fusion learning (CFL) module with modality-specific encoders, which learns domain-invariant representations from WiFi and vision modalities while avoiding direct cross-modal alignment, producing stable and complementary gesture semantics.
- We introduce a spatial matrix difference (SMD)-guided cross-modal generation (CMG) module, which encodes skeletal information into spatial matrices to characterize structural differences across users and uses them as conditional signals for diffusion-based generation, enabling user-adaptive WiFi feature synthesis for unseen users.

2. Related Work

Robust gesture recognition has been widely studied from different perspectives. In this paper, we review related work along three major lines of research. The first concerns WiFi-based gesture recognition, which exploits wireless signals for contactless activity sensing. The second addresses domain-invariant and user-adaptive recognition, which aims to improve generalization across diverse environments and users. The third focuses on multimodal learning, with particular attention to WiFi–vision fusion that integrates

complementary sensing modalities. These three research directions form the basis of our proposed DUAL-Net framework.

2.1. WiFi-Based Gesture Recognition

With the widespread deployment of wireless infrastructure, WiFi signals have evolved beyond their role as communication carriers to serve as key enablers of wireless sensing [23,24]. In the context of gesture recognition, human movements influence the multipath propagation of WiFi signals, introducing variations in the phase and amplitude of Channel State Information (CSI). These perturbations encode motion-related patterns that can be leveraged to recognize user gestures in a contactless and device-free manner.

Early WiFi-based gesture recognition systems primarily relied on handcrafted time and frequency domain features such as amplitude extrema, mean, standard deviation, signal energy, and frequency peaks [25–28]. These features were typically classified using models such as Support Vector Machines (SVMs), k-Nearest Neighbors (KNNs), or Random Forests (RFs), and in some cases matched through dynamic time warping (DTW). While efficient and interpretable, these approaches often struggled to generalize due to fixed feature definitions and rigid decision boundaries. To overcome these limitations, subsequent research adopted deep learning techniques, employing CNNs and LSTMs to automatically capture spatial and temporal dependencies from raw CSI [29–31]. Other studies transformed CSI into spectrogram-like representations to exploit image-based architectures [32]. These models achieved higher accuracy and robustness but continued to face challenges when applied to new users or unseen environments. In addition, compared to visual signals, WiFi inherently provides limited semantic richness due to its coarse spatial resolution and indirect encoding of human motion. As a result, it remains difficult to capture subtle gesture variations across different users, which hinders generalization in cross-user and cross-domain scenarios. Related studies on scalable indoor sensing also show that WiFi-based systems remain affected by data quality, user participation, device burden, and environmental dynamics [33,34]. These findings further highlight the need to improve robustness and adaptability in practical WiFi sensing systems.

2.2. Domain-Invariant and User-Adaptive Gesture Recognition

Achieving robust gesture recognition across different environments and users remains a fundamental challenge in WiFi-based sensing. Spatial layout variations, signal interference, and user-specific gesture styles often lead to significant shifts in WiFi signal distributions, thereby compromising the generalization ability of recognition models. To address these issues, prior research has primarily explored two directions. One focuses on domain-invariant learning to mitigate environmental variations, while the other emphasizes user-adaptive learning to handle user-specific differences.

Domain-invariant techniques aim to reduce performance degradation caused by environmental changes. These approaches typically align feature distributions between source and target domains through adversarial learning or distance-based regularization. For example, adversarial training frameworks introduce a domain classifier that guides the feature extractor to learn domain-invariant representations [35,36]. Metric-based approaches such as Maximum Mean Discrepancy (MMD) have also been applied to reduce domain discrepancy in the feature space [37,38]. Recent domain generalization studies further indicate that robust representations should not only reduce domain-specific variations but also preserve class-discriminative information [39]. In WiFi-based gesture recognition, methods such as WiAG [40] and DPSense [41] generate virtual samples or apply signal enhancement to improve robustness under dynamic environmental conditions. However,

these works usually assume consistent gesture execution styles across different users and often fail to capture user-related variability.

In parallel, user-adaptive approaches seek to suppress irrelevant user-specific traits while retaining gesture semantics. For instance, CrossGR [14] employs a generative adversarial network to disentangle gesture-relevant information from user-dependent characteristics, thereby producing more generalized representations. Other works leverage few-shot learning and meta-learning strategies to improve recognition performance on unseen users using limited training data [42,43]. Although these methods enhance cross-user generalization, they often treat domain shifts and user variability as independent challenges without modeling their interdependence.

In practice, however, domain and user factors are often closely coupled. A user's gesture style may interact differently with WiFi signal patterns depending on environmental conditions such as room geometry, device orientation, or propagation paths. Recognition models that separately address domain-invariant learning and user-adaptive modeling may fail to capture these intertwined effects. This limitation highlights the need for integrated frameworks that can simultaneously extract stable domain-invariant features and adapt to user-specific variations. The proposed DUAL-Net framework addresses this gap by jointly incorporating domain-invariant and user-adaptive representations within a unified recognition pipeline, thereby improving performance under both cross-domain and cross-user conditions.

2.3. Multimodal Learning and WiFi-Vision Fusion

Multimodal learning has emerged as a promising direction for improving recognition performance by integrating complementary information from multiple sensing modalities [44]. In the context of human activity understanding, combining WiFi signals with vision data helps overcome the inherent limitations of each modality. WiFi signals are resilient to lighting changes and occlusions but suffer from low spatial resolution and limited semantic expressiveness, whereas vision signals provide rich spatial detail but are vulnerable to visual disturbances. Fusing these two modalities therefore offers the potential to improve both robustness and accuracy in complex and dynamic environments. However, multimodal adaptation remains challenging because heterogeneous domain shifts, cross-modal dependency preservation, modality imbalance, and missing inputs are difficult to address using unimodal alignment strategies alone [45].

Recent research has explored WiFi-vision fusion for a range of human sensing tasks, including activity recognition [46,47], gait recognition [48], and gesture recognition [20]. Zou et al. [46] employed decision-level fusion of WiFi and RGB video to enhance recognition robustness under degraded visual conditions. However, this strategy captures only limited cross-modal interactions, thereby restricting the benefits of fusion. Guo et al. [47] proposed a feature-level fusion framework that projects WiFi and vision data into a shared latent space to improve action recognition. While effective, this approach may not fully exploit the complementary characteristics of the two modalities. WiVi [48] developed a multi-path gait recognition system that matches WiFi-derived features with a visual gait library, but performance degrades in scenarios with closely spaced walking paths due to interference. More recently, WiVi-GR [20] introduced a gesture recognition framework based on channel superposition convolution and data-level fusion, incorporating comprehensive speed representations to improve classification accuracy. However, effectively modeling inter-modal correlations remains an open challenge. Many existing approaches rely on fixed fusion strategies and struggle with signal heterogeneity or incomplete modality data. These limitations underscore the need for more flexible and adaptive multimodal frame-

works that can dynamically align heterogeneous signals and maintain robustness when only partial modality inputs are available.

Although substantial progress has been made in these three research directions, each line of work still faces critical limitations. WiFi-based gesture recognition suffers from environmental sensitivity and limited semantic expressiveness. Domain adaptation and user-independent techniques improve robustness but often treat domain and user variability in isolation, overlooking their complex interplay. Multimodal learning, especially WiFi–vision fusion, mitigates some of these issues but still struggles with modality heterogeneity and incomplete data in practical deployments. These challenges highlight the need for a unified framework that can simultaneously capture stable domain-invariant features and preserve user-adaptive characteristics. DUAL-Net is proposed to address this gap by jointly modeling domain-invariant and user-adaptive representations for robust gesture recognition.

3. System Design

This section introduces the proposed DUAL-Net framework, which jointly models domain-invariant and user-adaptive representations within a unified WiFi–vision architecture. An overview of the framework is first provided, followed by detailed descriptions of its three core modules. To learn stable cross-modal representations, the contrastive fusion learning (CFL) module employs modality-specific encoders to extract complementary WiFi and vision features and optimizes a contrastive objective in a shared embedding space. Instead of directly enforcing feature alignment, this design encourages instance-level consistency while preserving modality-specific characteristics. To capture user-specific gesture variations, the spatial matrix difference (SMD)-guided cross-modal generation (CMG) module generates user-adaptive WiFi representations conditioned on skeletal structures extracted from vision data. The proposed SMD encodes structural differences between users and serves as a conditioning signal for diffusion-based generation, enabling adaptation to unseen users without additional WiFi data collection. Finally, the domain-invariant and user-adaptive representations are fused for gesture recognition. The entire framework follows a two-stage learning paradigm, where CFL and CMG are performed during offline training, while online inference is lightweight with reduced computational overhead.

3.1. Overview of the DUAL-Net Framework

As shown in Figure 2, the proposed DUAL-Net framework consists of three core modules: a contrastive fusion learning (CFL) module, a cross-modal generation (CMG) module, and a gesture recognition module. The CFL module first processes raw WiFi and vision inputs. WiFi signals are denoised using a Butterworth low-pass filter, while segmentation techniques are applied to vision data to isolate gesture-relevant regions. After temporal alignment, both modalities are encoded by modality-specific encoders and projected into a shared semantic space. Instead of enforcing direct cross-modal alignment, contrastive learning is applied to maximize intra-class similarity and increase inter-class separability, resulting in stable domain-invariant representations. However, domain-invariant features alone are insufficient to capture user-specific characteristics required for cross-user recognition. To address this limitation, the CMG module generates user-adaptive WiFi representations conditioned on visual inputs using a diffusion-based generative model. Specifically, skeletal keypoints are transformed into condition vectors via a spatial matrix difference (SMD) formulation, which provides structural guidance for the diffusion process and enables generation tailored to target users. This design eliminates the need for collecting WiFi data from target users, improving adaptability while reducing deployment cost.

Finally, the gesture recognition module fuses domain-invariant and user-adaptive features for classification.

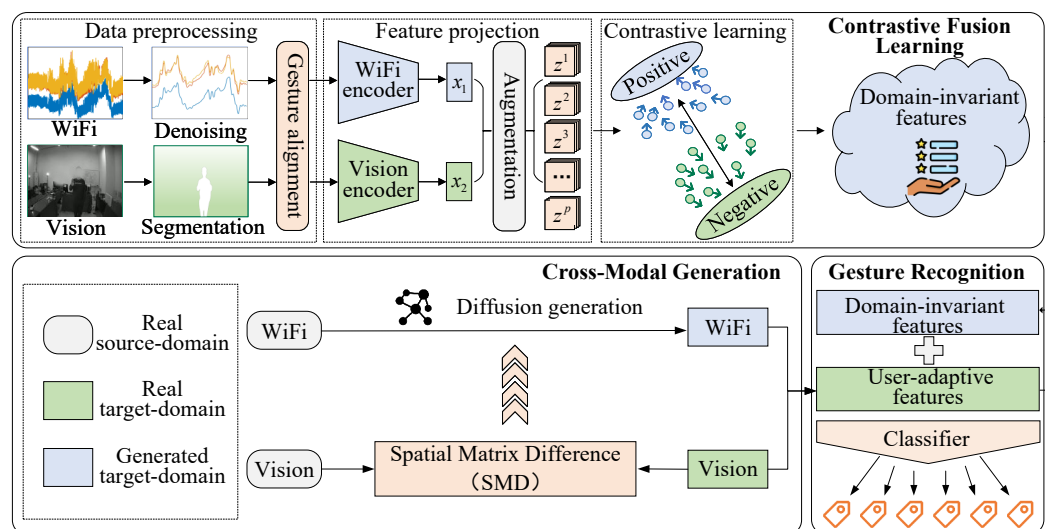


Figure 2. Overview of DUAL-Net. The framework learns domain-invariant features through contrastive WiFi–vision fusion, generates user-adaptive WiFi features via SMD-guided diffusion, and combines both feature types for cross-user and cross-environment gesture recognition.

Together, these modules form a unified framework rather than independent components. The overall design jointly learns domain-invariant representations and user-adaptive features to address the limitations of existing WiFi-based gesture recognition methods under environmental variations and inter-user differences. This integrated design improves recognition performance by capturing both shared gesture semantics and user-specific characteristics. In addition, it enables adaptation to unseen users without requiring additional WiFi data collection, thereby reducing data acquisition cost and supporting practical deployment scenarios.

3.2. Contrastive Fusion Learning Module

As depicted in Figure 2, the contrastive fusion learning module is composed of three main components, namely data preprocessing, feature projection, and contrastive learning. In the preprocessing stage, WiFi signals are denoised using a Butterworth low-pass filter to suppress high-frequency noise while preserving gesture-relevant temporal structures, and segmentation methods are applied to vision data to isolate gesture-relevant regions. After temporal alignment, the two modalities are encoded by modality-specific encoders and projected into a shared latent space. Finally, contrastive learning is applied to maximize intra-class similarity while minimizing inter-class discrepancy, enabling the model to extract gesture-relevant semantics that remain robust under varying environmental conditions.

3.2.1. Data Preprocessing

WiFi and vision data collected in daily environments are often corrupted by diverse types of noise, leading to degraded feature quality and reduced recognition performance. To address this challenge, a series of preprocessing steps are applied to denoise the input signals, isolate gesture-relevant regions, and ensure temporal alignment across modalities.

For the WiFi modality, Channel State Information (CSI) is extracted from the physical layer, providing fine-grained measurements of amplitude and phase across multiple subcarriers. However, CSI is highly vulnerable to ambient interference, which introduces instability and obscures gesture-related dynamics. To suppress such interference while preserving meaningful variations, a Butterworth filter is applied to the raw CSI sequences.

For the vision modality, RGB video data is first converted to grayscale to reduce computational overhead while retaining essential motion cues. Foreground segmentation is then performed using the U^2 -Net model [49], effectively isolating the subject from background clutter and highlighting gesture-relevant regions.

Finally, to ensure temporal consistency between WiFi and vision streams, a time-series alignment procedure is introduced. In the WiFi data, gesture execution is detected through variance-based thresholding of amplitude fluctuations, marking the onset and offset of each gesture. In parallel, average optical flow magnitudes are computed across video frames, and adaptive thresholds are applied to localize gesture boundaries in the vision data. This synchronization strategy ensures that both modalities are temporally aligned, providing a coherent basis for joint feature extraction and subsequent learning.

3.2.2. Feature Projection

After preprocessing, the WiFi and vision data are prepared for joint representation learning. A fundamental challenge, however, arises from the inherent heterogeneity between the two modalities. WiFi data are typically structured as two-dimensional temporal sequences, whereas vision data contain three-dimensional spatiotemporal information. Without proper transformation, this mismatch hinders effective fusion.

To address this issue, modality-specific encoders are employed to transform both WiFi and vision signals into a unified latent feature space. After encoding, the WiFi features are flattened into 29,232-dimensional vectors for DUAL-Net and 37,296-dimensional vectors for MM-Fi, while the vision features are flattened into 9152-dimensional vectors. These flattened features are projected into 128-dimensional embeddings through modality-specific projection heads. The architectural details of these encoders are summarized in Tables 1 and 2, which outline their respective layer configurations. Each encoder consists of convolutional blocks, optionally followed by pooling, and concludes with batch normalization and ReLU activation. This design ensures that both modalities are represented in a compatible format, providing a foundation for cross-modal alignment and joint learning.

Table 1. WiFi encoder configuration.

ID	Operation	Configuration
Input	CSI amplitude	DUAL-Net: $1 \times 200 \times 270$ MM-Fi: $1 \times 200 \times 342$
1	Conv2d MaxPooling2d BatchNorm2d	Kernel = 2×2 , FM = 64 Stride = 3 FM = 64, ReLU
2	Conv2d MaxPooling2d BatchNorm2d	Kernel = 2×2 , FM = 64 Stride = 1 FM = 64, ReLU
3	Conv2d BatchNorm2d	Kernel = 1×1 , FM = 32 FM = 32, ReLU
4	Conv2d BatchNorm2d	Kernel = 1×1 , FM = 16 FM = 16, ReLU

Within this shared space, a feature enhancement strategy is introduced to improve representation expressiveness and strengthen generalization. As formalized in Equation (1), the encoding function Enc maps the input features x into n enhanced features. Here, m denotes the number of modalities, α_{jk} are randomly sampled weights between zero and one, and z_k^i represents the enhanced features, where i indexes the i -th sample. This mechanism introduces controlled diversity into the enhanced features, enabling the network to capture richer semantic information beyond what raw encoders provide. By enriching the representation space in this way, it establishes a more discriminative and well-aligned

foundation, which is essential for the success of the contrastive learning process in the next stage.

$$z_k^i = \text{Enc}(x_1^i, \dots, x_m^i) = \sum_{j=1}^m \alpha_{jk} x_j^i, k = 1, \dots, n. \quad (1)$$

Table 2. Vision encoder configuration.

ID	Operation	Configuration
Input	RGB video clip	$1 \times 10 \times 160 \times 212$ 10 sampled frames
1	Conv3d	Kernel = $2 \times 5 \times 5$, FM = 64
	MaxPooling3d	Stride = 3
	BatchNorm3d	FM = 64, ReLU
2	Conv3d	Kernel = $2 \times 5 \times 5$, FM = 64
	MaxPooling3d	Stride = 3
	BatchNorm3d	FM = 64, ReLU
3	Conv3d	Kernel = $1 \times 5 \times 5$, FM = 32
	BatchNorm3d	FM = 32, ReLU
4	Conv3d	Kernel = $1 \times 2 \times 5$, FM = 16
	BatchNorm3d	FM = 16, ReLU

3.2.3. Contrastive Learning

To further refine the fused representations derived from WiFi and vision data, a contrastive learning strategy is employed to enhance feature discriminability across gesture categories. Building on the unified and semantically enriched feature space established during feature projection, contrastive learning strengthens decision boundaries by promoting semantically consistent clustering of features. While contrastive learning can be formulated by constructing positive pairs from different views of the same sample and negative pairs from different samples, such an instance-discrimination paradigm does not explicitly exploit semantic relationships among gesture categories. As a result, samples corresponding to the same gesture may remain dispersed in the representation space when variations in orientation, environment, or user-specific execution patterns are present. To learn gesture representations that are more robust to these variations, we employ supervised contrastive learning, where class labels are incorporated to guide feature alignment.

Specifically, fused features belonging to the same gesture category are encouraged to cluster together, whereas features from different categories are driven apart. This objective enhances intra-class consistency and inter-class discriminability in the shared latent space. The contrastive learning objective is defined as follows:

$$\mathcal{L}_{CL} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z^i \cdot z^p / \tau)}{\sum_{a \in A(i)} \exp(z^i \cdot z^a / \tau)},$$

here, z^i and z^p denote normalized features, $\cdot$ represents the inner product, τ is a temperature parameter, $P(i)$ denotes the set of indices of positive samples that belong to the same class as the anchor i , and $A(i)$ is the set of all other samples in the batch excluding i .

To evaluate the effectiveness of this process, we visualize the learned feature distributions using t-SNE. As shown in Figure 3, each point corresponds to a gesture feature, with different colors indicating different classes. After contrastive training, features belonging to the same class form compact clusters, while clusters from different classes are clearly separated. This improved organization provides strong evidence that contrastive learning enhances semantic alignment across modalities and increases the discriminative power of the fused feature space.

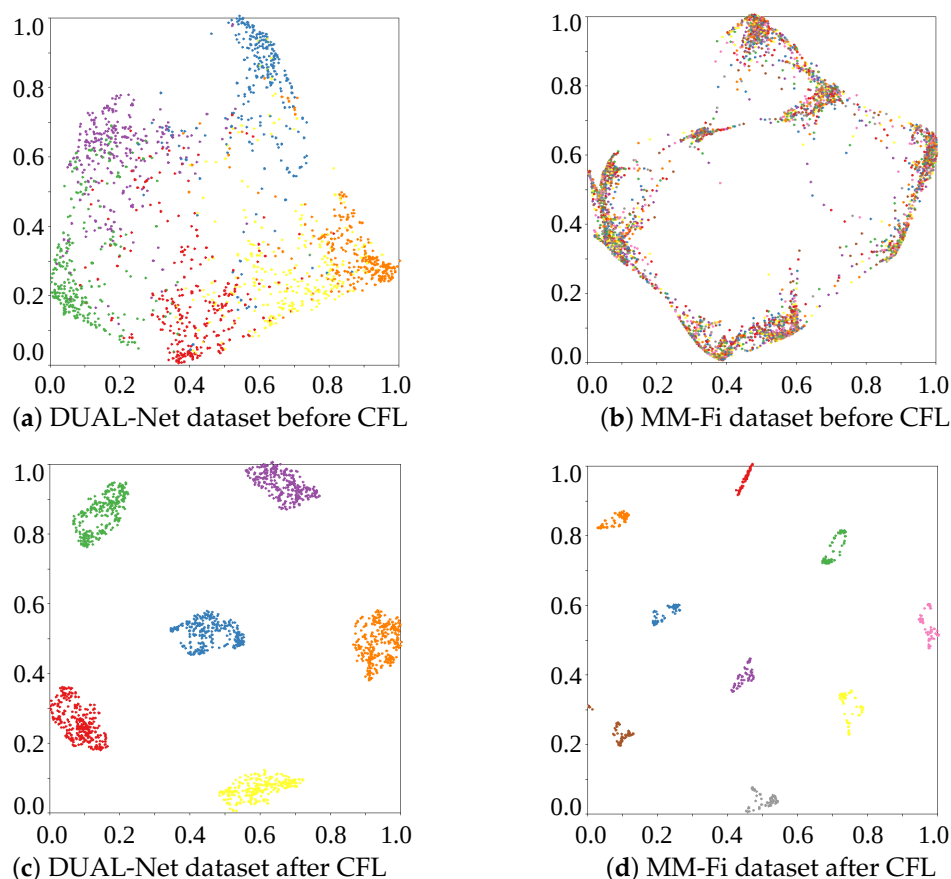


Figure 3. t-SNE visualization of fused gesture features before and after contrastive fusion learning (CFL) on the DUAL-Net and MM-Fi datasets. Different colors represent different gesture categories.

3.3. Cross-Modal Generation Module

As shown in Figure 4, the cross-modal generation (CMG) module aims to generate user-adaptive WiFi representations for unseen target users. The module is built upon a diffusion-based generative framework that progressively transforms source-user WiFi representations into target-user-adaptive features through a conditional denoising process. To ensure that the generated representations preserve user-specific characteristics, the proposed spatial matrix difference (SMD) is introduced as a structural conditioning signal. Specifically, skeletal information is encoded into spatial matrices to characterize structural variations between source and target users. The resulting SMD is then incorporated into the diffusion model to guide the generation process, enabling controllable synthesis of WiFi features that are consistent with the target user's motion patterns.

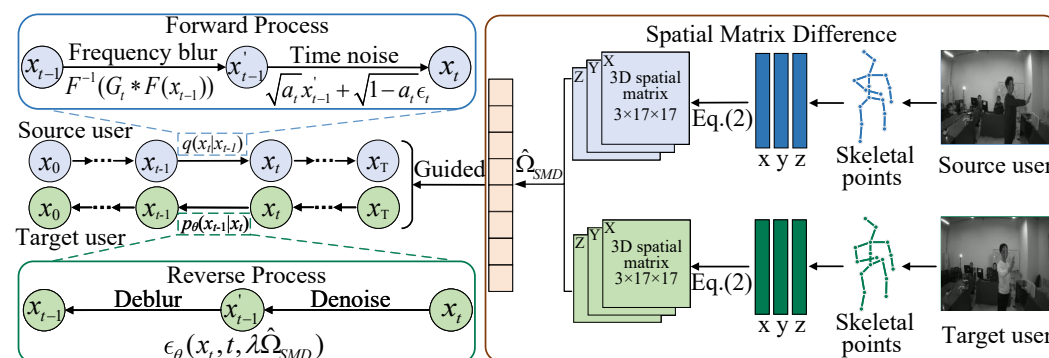


Figure 4. Illustration of the SMD-guided diffusion-based cross-modal generation module. The normalized SMD score is used as a structural condition to guide the reverse denoising process for user-adaptive WiFi generation.

3.3.1. Forward Process

The forward process in the diffusion model progressively corrupts the original WiFi signal by injecting noise, ultimately transforming it into a pure noise distribution. WiFi data can be represented as a two-dimensional tensor, where the horizontal axis corresponds to the temporal sequence and the vertical axis corresponds to subcarriers in the frequency domain. This structure encodes both time-domain dynamics and frequency-domain characteristics. Unlike image data, WiFi signals contain richer structural dependencies across these two domains, which necessitates a tailored noise design strategy to model their signal distribution more precisely.

To preserve this structure during noise injection, we incorporate perturbations in both the frequency and time domains. First, we apply Gaussian convolution in the frequency domain to suppress high-frequency components and achieve progressive spectral smoothing:

$$G_t * F(x_{t-1}),$$

where G_t is the Gaussian convolution kernel and $*$ denotes the convolution operation. Here, $F(\cdot)$ represents the Fourier transform, which maps the WiFi signal from the time domain to the frequency domain, enabling direct manipulation of its spectral components. The blurred frequency-domain signal is then transformed back to the time domain through the inverse Fourier transform:

$$x'_{t-1} = F^{-1}(G_t * F(x_{t-1})).$$

This operation ensures that high-frequency variations are effectively smoothed while preserving the overall temporal structure. Next, Gaussian noise is added in the time-domain to further degrade the signal and gradually push it toward a noise distribution:

$$x_t = \sqrt{\alpha_t} x'_{t-1} + \sqrt{1 - \alpha_t} \epsilon_t,$$

where α_t is a predefined noise schedule and ϵ_t denotes Gaussian noise. As t increases, the signal converges toward Gaussian noise while maintaining structured degradation.

Finally, after repeated application of frequency-domain blurring and time-domain noise injection, the resulting x_t follows a distribution formed by the mixture of the clean signal x_0 and Gaussian noise [50]:

$$q(x_t|x_0) \sim \mathcal{N}(x_0; \mu_t, \sigma_t^2 I).$$

This frequency–time forward process provides a structured noise schedule that preserves signal properties at early steps and enables controllable degradation, laying a principled foundation for the reverse denoising process.

3.3.2. Reverse Process

The reverse process, as the inverse of the forward corruption process, aims to progressively recover the clean WiFi data distribution from noisy observations through a parameterized denoising transformation. Whereas the forward process gradually corrupts the original signal with Gaussian perturbations in both the time and frequency domains, the reverse process learns the inverse mapping to reconstruct the underlying activity semantics and structural consistency. To formally realize this reconstruction, the learning objective is defined to align the model distribution with the true data distribution. Formally, the generative objective is to approximate the true data distribution $q(x_0)$ with a learnable model $p_\theta(x_0)$ by minimizing their Kullback–Leibler (KL) divergence:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} D_{KL}(q(x_0) \parallel p_{\theta}(x_0)).$$

Direct optimization of this divergence is intractable since $q(x_0)$ is unknown. To make the problem tractable, we adopt the diffusion probabilistic framework introduced by Ho et al. [51] and extended in DDIM [52]. This framework defines a forward diffusion process and a corresponding reverse denoising process, whose joint distributions can be factorized as:

$$q(x_{0:T}) = q(x_0) \prod_{t=1}^T q(x_t | x_{t-1}), p_{\theta}(x_{0:T}) = p(x_T) \prod_{t=1}^T p_{\theta}(x_{t-1} | x_t),$$

The forward process forms a linear Gaussian Markov chain that progressively injects noise into the data, while the reverse process aims to invert this transformation through a learnable denoising distribution.

To address the intractability of optimizing the global KL divergence, we apply the variational inference formulation [51], which decomposes the overall objective into a sum of stepwise KL divergences between the true posterior and the parameterized reverse transitions:

$$D_{KL}(q(x_0) \parallel p_{\theta}(x_0)) = E_{q(x_{0:T})} \left[\sum_{t=1}^T D_{KL}(q(x_{t-1} | x_t, x_0) \parallel p_{\theta}(x_{t-1} | x_t)) \right] + C.$$

This decomposition transforms a single intractable global objective into a set of manageable local objectives at each timestep, simplifying training and improving stability. Accurately modeling the conditional posterior $q(x_{t-1} | x_t, x_0)$ at each step is key to ensuring that the learned reverse process $p_{\theta}(x_{0:T})$ converges to the true data distribution.

Gaussian Posterior Approximation. Following the DDPM formulation, we approximate the reverse process using Gaussian transitions. Under this approximation, the posterior distribution at each diffusion step can be expressed as:

$$q(x_{t-1} | x_t, x_0) = N(x_{t-1}; \tilde{\mu}_{t-1}, \tilde{\sigma}_{t-1}^2 I),$$

where $\tilde{\mu}_{t-1} = \frac{1}{\tilde{\gamma}_t^2} (\gamma_t \tilde{\gamma}_{t-1}^2 x_t + \gamma_{t-1} \sigma_t^2 x_0)$, $\tilde{\sigma}_{t-1}^2 = \frac{\tilde{\gamma}_{t-1}}{\tilde{\gamma}_t} \sigma_t^2$.

The mean term $\tilde{\mu}_{t-1}$ represents an adaptive combination of the noisy sample x_t and the clean signal x_0 , weighted according to the noise variance σ_t^2 . This formulation yields a closed-form posterior that provides an optimal intermediate estimate of x_{t-1} conditioned on both observed and clean signals.

Learned Reverse Distribution. Building on the posterior formulation introduced earlier, the learned reverse process aims to approximate the true posterior during inference. To this end, the model defines a parameterized Gaussian transition:

$$p_{\theta}(x_{t-1} | x_t) = N(x_{t-1}; \mu_{\theta}(x_t), \sigma_{\theta}^2(x_t) I),$$

where both the mean $\mu_{\theta}(x_t)$ and the variance $\sigma_{\theta}(x_t)$ are predicted by a neural network. The training objective minimizes the KL divergence between the true posterior and the learned reverse distribution [50]:

$$D_{KL}(q(x_{t-1} | x_t, x_0) \parallel p_{\theta}(x_{t-1} | x_t)) = E_{q(x_0)} \left[\frac{1}{2\tilde{\sigma}_{t-1}^2} \|\tilde{\mu}_{t-1} - \mu_{\theta}(x_t)\|^2 \right].$$

When the model variance is fixed to match the posterior variance $\sigma_{\theta}(x_t) = \tilde{\sigma}_{t-1}$, this divergence simplifies to a weighted mean squared error (MSE) between the ground-truth and predicted means. This formulation not only reduces computational complexity but also preserves theoretical consistency with the DDPM framework [50,51]. Moreover, this

learned reverse distribution serves as the foundation for reconstructing clean RF signals in the generative denoising process, thereby ensuring accurate and stable signal restoration.

3.3.3. Conditional Guidance with SMD

Generating user-adaptive WiFi representations from RF signals alone is challenging, since WiFi measurements encode human motion implicitly through signal reflections. User-specific factors, including body proportions, limb lengths, gesture amplitudes, and execution patterns, are weakly encoded in the RF domain. However, they substantially affect propagation paths and multipath reflections. Therefore, the same gesture performed by different users can induce distinct CSI patterns. Modeling such variations requires additional structural guidance beyond RF signals.

To address this issue, we introduce the spatial matrix difference (SMD) as a cross-modal structural prior. The key idea is that variations in skeletal structure and motion trajectories are reflected in both skeletal representations and the corresponding WiFi signal patterns. By measuring structural differences across users, SMD provides a surrogate measure of user-specific variations in WiFi responses. This prior is incorporated into the diffusion process to guide the generation of user-adaptive WiFi representations.

During target-user adaptation, the proposed DUAL-Net framework does not require any WiFi samples from the target user. Instead, it utilizes a short RGB video of the target user performing a representative gesture, which is easily obtainable in practice and does not require dedicated WiFi data collection. The video provides visual observations for capturing user-specific motion characteristics while avoiding additional target-user WiFi sampling.

To obtain reliable skeletal representations, we employ VideoPose3D [53] to estimate 3D skeletal keypoints. Compared with conventional pose-estimation approaches, VideoPose3D preserves temporal consistency while maintaining computational efficiency. Each frame contains 17 keypoints, which are organized into skeletal matrices that encode pairwise structural relationships among body joints. For each spatial dimension (x, y, z) , we compute a 17×17 matrix in which each element captures the relative position between keypoints. As shown in Equation (2), $i, j \in [1, 2, \dots, 17]$ denote keypoint indices, and X represents the spatial matrix for the x -dimension (size $1 \times 17 \times 17$), with Y and Z defined analogously.

$$X = \begin{cases} x_i - x_j, & i \neq j; \\ x_i, & i = j. \end{cases} \quad Y = \begin{cases} y_i - y_j, & i \neq j; \\ y_i, & i = j. \end{cases} \quad Z = \begin{cases} z_i - z_j, & i \neq j; \\ z_i, & i = j. \end{cases} \quad (2)$$

To guide the cross-user generation process, it is crucial to quantitatively measure the discrepancy between the skeletal matrices of source and target users. We adopt the mean squared error (MSE) as the metric, since it provides a sensitive and differentiable measure of structural differences. Specifically, MSE captures the element-wise squared deviations between the spatial matrices of source and target users, effectively quantifying user-specific variations in body structure and gesture execution. The SMD is defined as:

$$\Omega_{SMD} = \frac{1}{n} (||X_t - X_s||^2 + ||Y_t - Y_s||^2 + ||Z_t - Z_s||^2),$$

where X, Y, Z represent the skeletal matrices in the three spatial dimensions, s and t refer to the source and target users, and $n = 3$ denotes the number of spatial dimensions. This formulation encourages the generated WiFi data to retain target user-specific spatial semantics derived from vision guidance.

Since the magnitude of Ω_{SMD} may vary across user pairs, the score is normalized to ensure numerical stability:

$$\hat{\Omega}_{SMD} = \frac{\Omega_{SMD} - \Omega_{\min}}{\Omega_{\max} - \Omega_{\min} + \epsilon},$$

where ϵ is a small constant introduced to avoid division by zero. The normalized SMD score lies within the range $[0, 1]$ and serves as the structural conditioning variable for the diffusion model.

To incorporate user-specific structural information into the generation process, the normalized SMD score is introduced as a structural condition for the diffusion model. The rationale is that skeletal discrepancies capture variations in body geometry and motion execution, which are closely associated with user-dependent WiFi responses. By conditioning the denoising process on this structural prior, the model is encouraged to generate WiFi representations that better reflect the characteristics of the target user.

Specifically, the normalized SMD score is first projected into a latent embedding and then incorporated into the denoising network as a conditioning signal. This projection maps the scalar SMD score into the same latent conditioning space as the diffusion timestep embedding, ensuring dimensional compatibility during denoising. The noise prediction network is therefore reformulated as:

$$\epsilon_{\theta}(x_t, t, \lambda \hat{\Omega}_{SMD}),$$

where x_t denotes the noisy WiFi representation at diffusion step t , $\hat{\Omega}_{SMD}$ denotes the normalized structural discrepancy derived from the source and target skeletal representations, and λ controls the contribution of the structural condition.

Accordingly, the conditional reverse distribution is defined as:

$$p_{\theta}(x_{t-1}|x_t, \hat{\Omega}_{SMD}) = \mathcal{N}(x_{t-1}; \mu_{\theta}(x_t, \hat{\Omega}_{SMD}), \sigma_t^2 I)$$

where the mean of the reverse transition is conditioned on the structural prior. The corresponding training objective is:

$$L_{CMG} = \mathbb{E} \left[\|\epsilon - \epsilon_{\theta}(x_t, t, \lambda \hat{\Omega}_{SMD})\|_2^2 \right]$$

Through this formulation, the SMD condition explicitly injects user-specific structural cues into the generation process, enabling the synthesized WiFi representations to better adapt to target-user characteristics. The coefficient λ is a standard guidance strength hyperparameter that controls the contribution of the SMD-based structural condition in the diffusion process. It is not a learned parameter but is selected based on a held-out validation set. In practice, we perform a lightweight grid search over a small range and observe that the model exhibits stable performance within a relatively wide interval, indicating that the proposed method is not sensitive to the exact choice of λ . This behavior is consistent with prior diffusion-based conditional generation methods, where similar guidance-scale parameters are typically tuned in the same manner.

3.4. Gesture Recognition Module

The gesture recognition module integrates domain-invariant and user-adaptive features for final classification. The domain-invariant representation h_{env} is obtained from the contrastive fusion learning module, which provides stable gesture semantics across different environments. The user-adaptive representation h_{usr} is derived from the WiFi features synthesized by the cross-modal generation module, which introduces user-specific structural and motion characteristics. These two representations provide complementary cues for recognition: h_{env} improves robustness against environmental changes, while h_{usr} compensates for inter-user variations.

The fused representation is computed as $c = g([h_{env}, h_{usr}])$, and the final prediction is $\hat{y} = \text{Softmax}(Wc)$, where W denotes the classification layer. In this formulation, the fusion function $g(\cdot)$ serves to combine the complementary information from the two branches before classification, allowing the classifier to exploit both environment-robust and user-adaptive cues.

We adopt a training strategy under the two-stage learning framework of DUAL-Net, separating offline representation learning from online inference.

Encoder Update. In this stage, both real and CMG-generated WiFi samples are used to update the encoders, while the classifier is kept fixed. This encourages the encoders to incorporate user-adaptive information from CMG-generated features while preserving domain-invariant representations learned from contrastive fusion learning.

Classifier Update. In this stage, the encoders are frozen and the classifier is optimized using both real and generated samples under a cross-entropy loss:

$$\mathcal{L}_{CE} = - \sum_i^L y_i \log \hat{y}_i.$$

This step refines the decision boundary by jointly considering stable gesture semantics and user-specific variations introduced by the generated samples. Overall, this alternating optimization strategy implements the two-stage learning paradigm of DUAL-Net, where CFL and CMG are performed during offline learning, while online inference remains lightweight with reduced computational overhead.

4. Experimental Setup

4.1. Datasets

While gesture recognition has achieved significant advances in both WiFi and vision sensing, publicly available datasets that provide synchronized multimodal data remain limited. This limitation presents a challenge for developing and evaluating robust models that require the integration of heterogeneous modalities. To facilitate a comprehensive evaluation of the proposed DUAL-Net framework, we construct a new multimodal dataset, referred to as the DUAL-Net dataset, which comprises synchronized WiFi Channel State Information (CSI) and RGB video data. The dataset is collected under varied environmental conditions and across a diverse set of users, with all participants providing informed consent. This makes it particularly suitable for studying cross-modal fusion, user-adaptive modeling, and domain-invariant representation learning.

In addition to our self-collected dataset, we employ the publicly available MM-Fi dataset [54], which includes five sensing modalities—WiFi CSI, RGB video, depth video, LiDAR, and mmWave radar—and covers a wide range of user behaviors in multiple environments. For this study, we focus on the WiFi CSI and RGB video components to ensure consistency in modality alignment. Evaluation on both the DUAL-Net and MM-Fi datasets demonstrates consistent performance across users and environments. A detailed summary of the two datasets used in our experiments is provided in Table 3.

Table 3. Summary of the datasets used for evaluation (# denotes the number of items).

Dataset Name	Modality	# Gestures	# Volunteers	# Envs	Env. IDs	# Samples
DUAL-Net	WiFi, RGB	6	10	5	A, B, C, D, E	1800
MM-Fi [54]	WiFi, RGB, Depth, LiDAR, mmWave	9	10	4	E01, E02, E03, E04	8640

DUAL-Net. As illustrated in Figure 5, the DUAL-Net dataset was designed to evaluate the proposed framework across different environments and users. It contains six gesture categories, as shown in Figure 6, namely *Push–pull*, *Sweep*, *Clap*, *Slide*, *Draw a circle*, and *Draw zigzag*. The data are collected from 10 volunteers (including 3 females) at five locations labeled A, B, C, D, and E. These gestures cover both simple directional movements and complex trajectory-based motions, while the five locations represent different spatial configurations that may affect gesture recognition. RGB video data was captured using a Sony FDR-AX60 camera (Sony Corporation, Tokyo, Japan) from different perspectives, and WiFi CSI was collected using two PCs equipped with Intel 5300 NICs (Intel Corporation, Santa Clara, CA, USA) as the transmitter and receiver. Each device had three antennas, and each antenna pair recorded 30 subcarriers, resulting in 270 subcarriers for each WiFi sample ($3 \text{ antennas} \times 3 \text{ antenna pairs} \times 30 \text{ subcarriers}$). For each location, six samples were collected per gesture category and per volunteer, resulting in 1800 samples in total ($6 \text{ gesture categories} \times 5 \text{ locations} \times 6 \text{ samples} \times 10 \text{ volunteers}$).

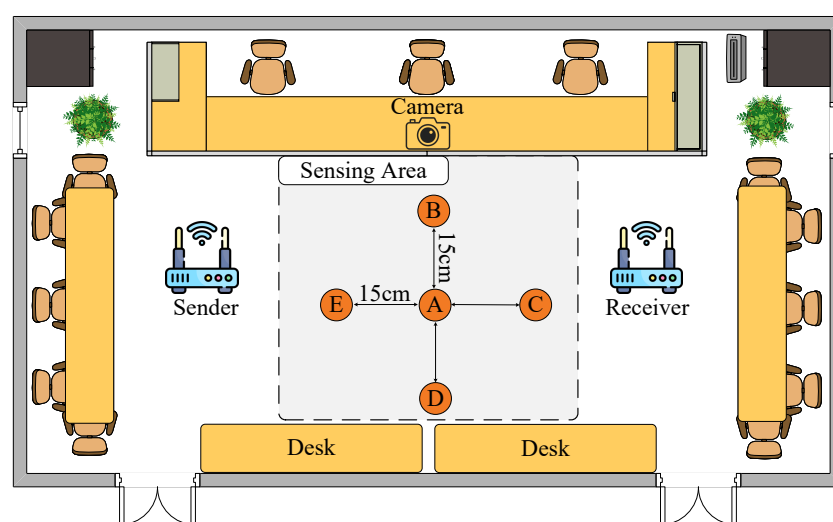


Figure 5. Multimodal data collection setup for the DUAL-Net dataset.

MM-Fi. The MM-Fi dataset provides a broad testbed for gesture recognition under varied environments and user profiles. It comprises five sensing modalities collected in four environments (E01–E04), with 10 participants in each environment performing 27 activities, nine of which are gesture-specific. We consider nine gesture categories covering arm stretching, arm extension, arm raising, and waving motions, including *horizontal arm stretch*, *vertical arm stretch*, *left arm extension*, *right arm extension*, *simultaneous arm extension*, *left arm raise*, *right arm raise*, *left-hand wave*, and *right-hand wave*. Each activity was recorded for 297 frames, and only RGB video and WiFi CSI are used in our experiments. RGB video was captured using a high-precision camera with a resolution of 640×480 pixels, providing detailed visual information for gesture recognition. WiFi CSI was captured using a TP-Link N750 router (TP-Link Technologies Co., Ltd., Shenzhen, China) and the Atheros CSI tool (Atheros 802.11n PCI/PCI-E chips) on the 5 GHz band with 40 MHz bandwidth. Each WiFi sample consists of 297 frames from 3 antenna pairs, with each pair recording 114 subcarriers, resulting in a CSI representation of $297 \times 3 \times 114$ per sample. In total, the selected MM-Fi gesture subset contains 8640 samples ($9 \text{ gestures} \times 4 \text{ environments} \times 24 \text{ samples} \times 10 \text{ volunteers}$).

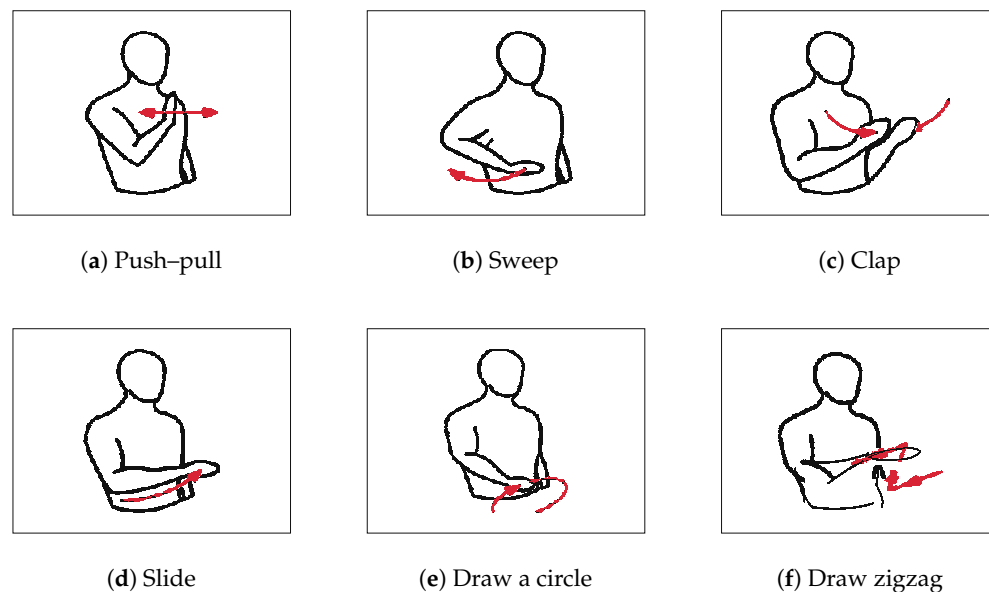


Figure 6. Illustrations of the six gesture categories in the DUAL-Net dataset, the arrows indicate the gesture directions.

4.2. Implementation

The proposed DUAL-Net framework was implemented in Python 3.8 using PyTorch v1.8 as the deep learning backend. PyTorch was selected for its flexibility and efficient support for multimodal data processing. All experiments were conducted on a workstation equipped with an Intel Xeon E5-2698 v4 CPU (Intel, Santa Clara, CA, USA), four NVIDIA Tesla V100 GPUs (Nvidia, Santa Clara, CA, USA), and 256 GB RAM, running the Ubuntu 20.04 operating system. For the cross-modal generation module, the diffusion-based generator was optimized using AdamW with a learning rate of 1×10^{-3} and a batch size of 16. The number of diffusion steps was set to 100. The noise schedule ranged from 1×10^{-4} to 0.003, and the blur variance was set to 1×10^{-10} . After training, the generated target-adaptive WiFi samples were saved offline and then used in the subsequent gesture recognition stage. For contrastive fusion learning, the model was trained using SGD with a learning rate of 0.01, momentum of 0.9, weight decay of 1×10^{-4} , and a batch size of 32. The training process lasted for 400 epochs, and the temperature parameter in the contrastive loss was set to 0.07. For gesture recognition, the recognition network was optimized using cross-entropy loss. The learning rate of the encoder part was set to 1×10^{-4} , while the learning rate of the classifier part was set to 1×10^{-3} . The batch size was set to 64, and the model was trained for 400 epochs. Each experiment was repeated ten times, and the average recognition accuracy was reported.

4.3. Experimental Notation

In the experiments, DUAL-Net denotes the full proposed framework. CFL denotes contrastive fusion learning, CMG denotes cross-modal generation, and SMD denotes spatial matrix difference. The abbreviations “w/” and “w/o” indicate “with” and “without”, respectively. In the ablation studies, w/o CFL removes the contrastive fusion learning module, w/o CMG removes the cross-modal generation module, and w/o SMD removes the SMD guidance from the generation process.

5. Experimental Results

This section presents a comprehensive evaluation of the proposed DUAL-Net framework. The experiments include overall performance assessment in cross-user and cross-environment scenarios, comparative analysis against multimodal and single-modality methods, ablation studies examining the roles of the contrastive fusion learning and cross-modal generation modules, evaluation of the recognition performance of generated WiFi data, and analysis of training dynamics, including the impact of synthesized sample quantity on cross-user accuracy and the convergence behavior of training loss. The results demonstrate that DUAL-Net achieves robust and scalable gesture recognition across diverse users and environments.

5.1. Overall Performance Evaluation of DUAL-Net

(1) Cross-user evaluation: In this experiment, we assess DUAL-Net's performance in cross-user scenarios, specifically evaluating its ability to generalize across different users. To rigorously examine the model's generalization capabilities, we conduct evaluations on both the DUAL-Net dataset and the MM-Fi dataset. For each dataset, the model is trained using gesture data from all but one user, who serves as the target for testing. This process is repeated for each user, ensuring comprehensive evaluation across all possible cross-user scenarios.

As shown in Figure 7, we evaluate the robustness and consistency of DUAL-Net using min-max bars, which indicate the maximum and minimum recognition accuracies obtained from ten randomly repeated experiments. To establish a fair comparison, we also compare with SenseFi [29], a widely adopted deep learning benchmark framework for WiFi human sensing. The results demonstrate that DUAL-Net achieves consistently high recognition accuracy with unseen users, reflecting its strong generalization ability in cross-user scenarios. While recognition accuracy typically exceeds 90%, a few cases fall below this threshold due to the inherent complexity of modeling diverse user characteristics. Nevertheless, DUAL-Net exhibits strong adaptability, achieving up to 100% accuracy in certain instances. Compared with SenseFi, DUAL-Net achieves consistent accuracy improvements, with average improvements of 9.52% and maximum improvements of 24.38% across different experimental settings, thereby confirming its effectiveness in enhancing cross-user gesture recognition.

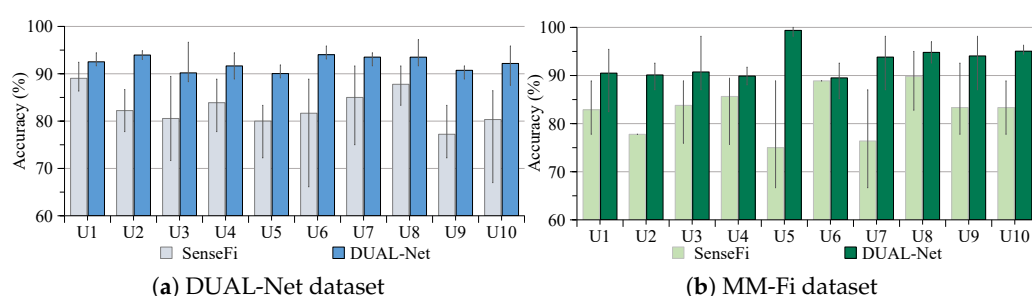


Figure 7. Cross-user evaluation results on the DUAL-Net and MM-Fi datasets, highlighting DUAL-Net's ability to generalize across unseen users.

(2) Cross-environment evaluation: In this experiment, we assess DUAL-Net's performance across varying environmental conditions, focusing on its ability to adapt to different locations and environmental setups. The evaluation is conducted on two datasets, the DUAL-Net dataset, which includes data collected from multiple locations, and the publicly available MM-Fi dataset, which includes data from various environmental deployments. For each dataset, we use one environment (or location) as the test set, while the remaining environments (or locations) are used for training. This approach ensures that

DUAL-Net is evaluated for its ability to generalize across different settings, a critical factor for real-world deployment.

As shown in Figure 8, the experimental results demonstrate that DUAL-Net consistently achieves over 90% gesture recognition accuracy on both the DUAL-Net and MM-Fi datasets. The relatively lower performance observed in the cross-location task on the DUAL-Net dataset is primarily due to the limited number of samples available at each location, which constrains the model's ability to generalize across spatial variations. Nevertheless, DUAL-Net maintains strong robustness across diverse environments, confirming its adaptability to spatial and environmental changes. Furthermore, when compared with SenseFi [29], a widely adopted deep learning benchmark framework for WiFi human sensing, DUAL-Net achieves average accuracy improvements of 11.93% on the DUAL-Net dataset and 14.35% on the MM-Fi dataset, thereby demonstrating its superior capability for robust cross-environment gesture recognition.

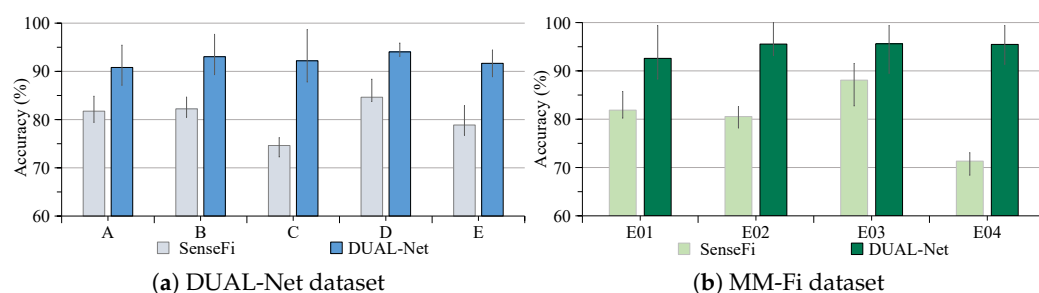


Figure 8. Gesture recognition accuracy across different locations on the DUAL-Net dataset and varying environmental conditions on the MM-Fi dataset.

5.2. Comparative Analysis of DUAL-Net

(1) Comparison with multimodal learning methods: To assess the effectiveness of the proposed DUAL-Net framework, we conducted an extensive comparative analysis against four state-of-the-art multimodal fusion methods: Attnsense [55], Deepsense [56], Concat [57], and Cosmo [58]. To ensure a fair comparison, all multimodal baselines are evaluated under the same dataset, train/test splits, and evaluation metrics. We strictly follow their original implementations and reproduce them under consistent software and hardware environments. For multimodal consistency, all methods use both WiFi and vision inputs when applicable, and identical preprocessing and augmentation strategies are applied where possible.

As shown in Figure 9, DUAL-Net consistently outperforms these methods on both the DUAL-Net and MM-Fi datasets. Specifically, DUAL-Net achieved recognition accuracy exceeding 90% across both datasets, with relative improvements of 18.52% and 3.08% over the strongest baseline, respectively. Moreover, DUAL-Net yields substantial F1-score gains (up to 30.57%), indicating its superior ability to differentiate between gesture categories and mitigate misclassification errors. These quantitative results highlight DUAL-Net's robust generalization capability across diverse users and environments, establishing it as a reliable and competitive framework for real-world gesture recognition applications.

(2) Comparison with single-modality methods: To comprehensively evaluate the effectiveness of DUAL-Net in heterogeneous multimodal fusion, we conducted comparative experiments on both the DUAL-Net and MM-Fi datasets against two representative single-modality gesture recognition methods, namely WiGr [59], a WiFi-based approach, and Snaapture [60], a vision-based approach. WiGr employs a two-path prototype network for WiFi-based cross-domain gesture recognition, while Snaapture integrates both static and dynamic gestures for vision-based recognition. These baselines represent state-of-the-art methods in their respective modalities, providing strong benchmarks for evaluation. For

single-modality baselines, we maintain the same evaluation protocol as multimodal methods, while restricting inputs to their corresponding modality (WiFi or vision) and ensuring that no cross-modal information is introduced.

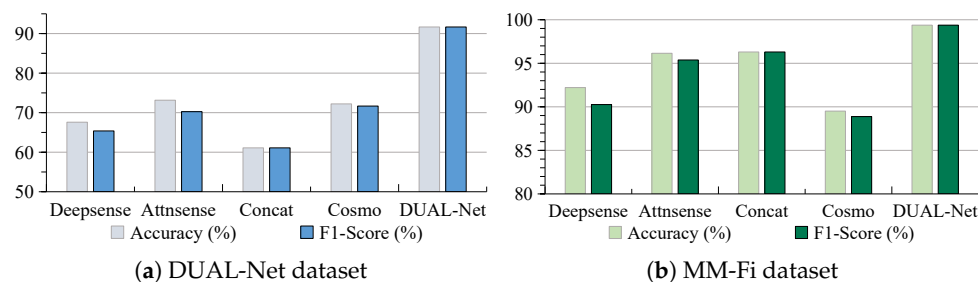


Figure 9. Comparison of DUAL-Net and multimodal baselines on the DUAL-Net and MM-Fi datasets.

As shown in Figure 10, DUAL-Net demonstrates clear superiority over both methods in terms of accuracy and F1-score across the two datasets. On the DUAL-Net dataset, DUAL-Net achieves relative gains of 27.74% in accuracy and 31% in F1-score compared to WiGr, and 18.21% in accuracy and 18.91% in F1-score compared to Snapture. On the MM-Fi dataset, similar improvements are observed, with accuracy and F1-score gains exceeding 45.51% and 47.63%, respectively. These results highlight the benefits of combining WiFi and vision signals, where DUAL-Net effectively leverages the complementary strengths of both modalities to enhance recognition performance. By jointly integrating WiFi's robustness to environmental occlusion with vision's fine-grained spatial representation, DUAL-Net overcomes the inherent limitations of single-modality approaches and achieves reliable performance across diverse sensing conditions.

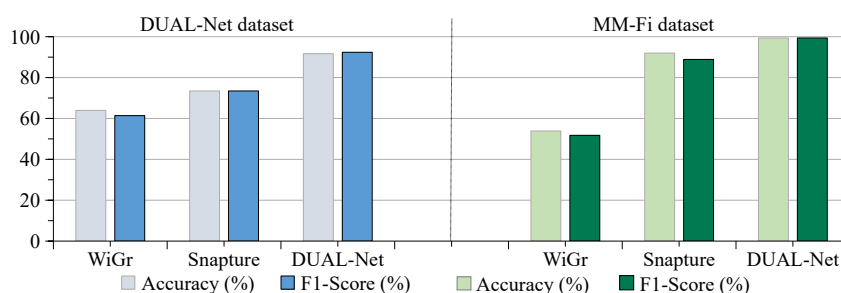


Figure 10. Comparison of DUAL-Net and single-modality baselines on the DUAL-Net and MM-Fi datasets.

5.3. Ablation Analysis of DUAL-Net

(1) Effectiveness of contrastive fusion learning: To evaluate the impact of the Contrastive fusion learning (CFL) module, we performed ablation experiments on both the DUAL-Net and MM-Fi datasets. As shown in Figure 11, the configuration without the contrastive fusion learning module (w/o CFL) exhibits lower gesture recognition accuracy, while the inclusion of the module (w/ CFL) substantially enhances performance. Specifically, integrating contrastive fusion learning improves the average accuracy by 22.53% on the DUAL-Net dataset and 7.52% on the MM-Fi dataset.

These results demonstrate that the contrastive fusion learning module effectively strengthens multimodal fusion by promoting consistent feature representations across WiFi and vision modalities. By maximizing intra-class similarity and enforcing inter-class separability, the module strengthens the discriminability of multimodal features, thereby enabling the model to more effectively distinguish fine-grained gesture categories. Consequently, DUAL-Net achieves a more coherent and unified feature space, which translates into higher accuracy and greater robustness in cross-user recognition scenarios.

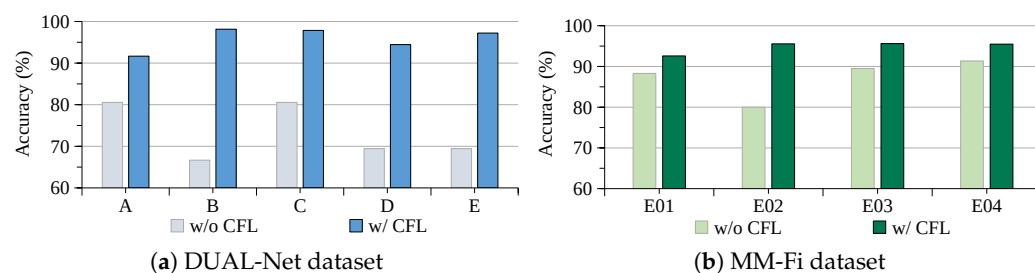


Figure 11. Recognition performance with and without CFL on the DUAL-Net and MM-Fi datasets.

(2) Effectiveness of cross-modal generation: To evaluate the contribution of the cross-modal generation (CMG) module, we conducted ablation studies using different configurations of DUAL-Net. As shown in Figure 12, the model without CMG (w/o CMG) achieves the lowest recognition accuracy, indicating that the absence of cross-modal generation significantly limits the model's ability to adapt to unseen users. In contrast, incorporating the full CMG module (w/ CMG), which generates WiFi representations guided by vision-based semantic priors, yields the highest performance. Specifically, the inclusion of CMG achieves up to 44.55% improvement on the DUAL-Net dataset and 17.82% improvement on the MM-Fi dataset, underscoring its effectiveness in producing user-adaptive features.

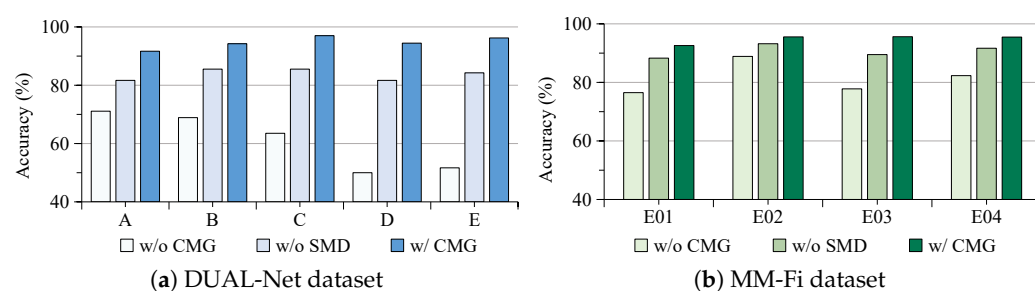


Figure 12. Recognition performance of the CMG module with and without SMD-guided conditioning on the DUAL-Net and MM-Fi datasets.

To further validate the role of cross-modal guidance, we compare the full CMG configuration with an alternative setting in which WiFi data are synthesized without structural conditioning (w/o SMD). The results show that removing SMD guidance leads to a performance degradation, demonstrating the importance of visual structural priors for user-adaptive WiFi generation.

5.4. Evaluation of Generated Data Performance

To evaluate the effectiveness of CMG, particularly the contribution of the proposed SMD-guided conditioning mechanism, we analyze generated WiFi representations from three complementary perspectives, including recognition performance, feature-level consistency, and signal-level consistency. First, we employ SenseFi [29] as an external gesture recognition model to compare WiFi samples generated without SMD guidance (w/o SMD), samples generated by the full CMG module (w/ CMG), and real WiFi samples used as a performance reference. Second, we visualize real, generated, and mixed WiFi features using t-SNE to examine feature distributions. Finally, we compare amplitude variations of real and generated WiFi signals to assess whether generated samples preserve gesture-related temporal dynamics. All evaluations are conducted on both the DUAL-Net and MM-Fi datasets.

(1) Recognition performance analysis: As shown in Figure 13, WiFi samples generated with SMD guidance consistently outperform those generated without SMD guidance on

both datasets. Specifically, recognition accuracy increases by up to 7.08% on the DUAL-Net dataset and 8.79% on the MM-Fi dataset. The results indicate that SMD guidance helps the diffusion model generate more useful WiFi features for unseen users. Although the accuracy is still lower than that obtained with real WiFi data, the generated samples preserve important gesture information and provide a practical alternative when target-user WiFi data are unavailable.

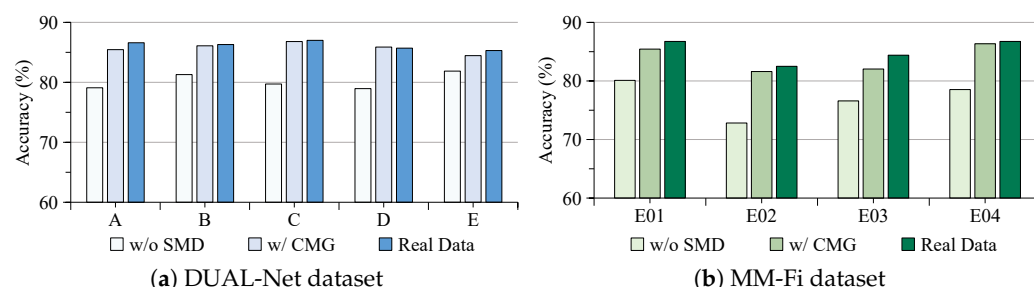


Figure 13. Accuracy comparison using WiFi data generated without SMD guidance (w/o SMD), generated by the complete CMG module (w/ CMG), and real data.

(2) Feature-level consistency analysis: To evaluate whether the generated WiFi data preserve gesture-related feature structures, we visualize the feature distributions of real, generated, and mixed WiFi samples using t-SNE. Figure 14 presents results on the DUAL-Net dataset, while Figure 15 reports results on the MM-Fi dataset. As shown in Figures 14a and 15a, real WiFi samples form clear class-wise clusters on both datasets, indicating that CSI features contain discriminative gesture information. Figures 14b and 15b further show that generated samples also exhibit separable class structures, suggesting that the CMG module preserves gesture-level semantic information during generation. When real and generated samples are jointly visualized in Figures 14c and 15c, the mixed distributions remain compact within classes while maintaining clear inter-class separability. This indicates that generated features are well aligned with real WiFi representations and can serve as complementary training samples.

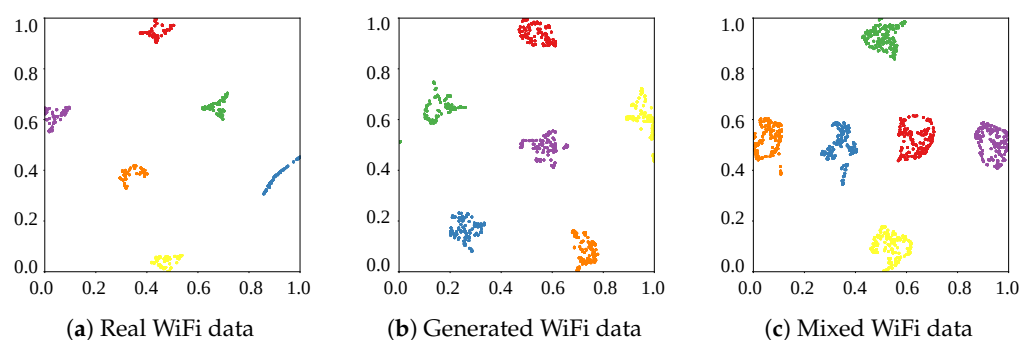


Figure 14. t-SNE visualization of WiFi feature embeddings for real, generated, and mixed data on the DUAL-Net dataset. Different colors represent different gesture categories.

(3) Signal-level consistency analysis: We further analyze whether the generated WiFi signals preserve gesture-related temporal dynamics by comparing the amplitude variations of real and generated samples. Representative real-generated pairs from the DUAL-Net and MM-Fi datasets are shown in Figures 16 and 17, respectively. As illustrated in Figure 16, the generated signals closely follow the dominant fluctuation patterns of the corresponding real signals, particularly in high-response regions. Similar trends can be observed in Figure 17, where the generated signals also preserve the main temporal variation patterns of real WiFi measurements. Although exact point-wise matching is not expected due to the stochastic nature of the generation process, the generated signals

consistently retain key amplitude variations and local response structures associated with gesture-induced CSI dynamics. These results indicate that the CMG module generates WiFi signals that are consistent with real data at the level of dominant temporal patterns across both datasets.

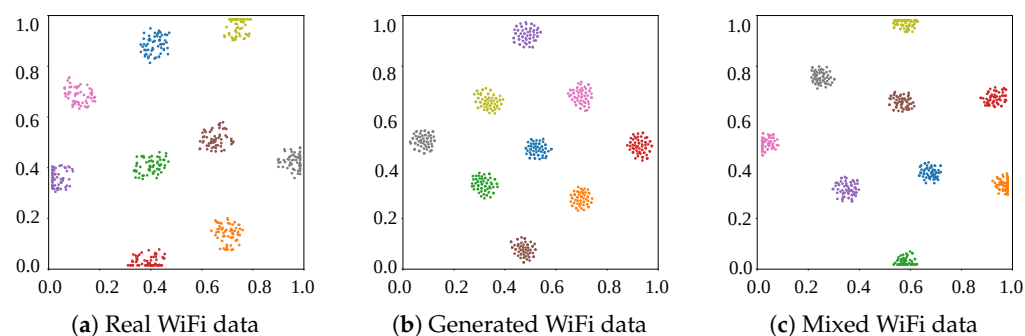


Figure 15. t-SNE visualization of WiFi feature embeddings for real, generated, and mixed data on the MM-Fi dataset. Different colors denote different gesture categories.

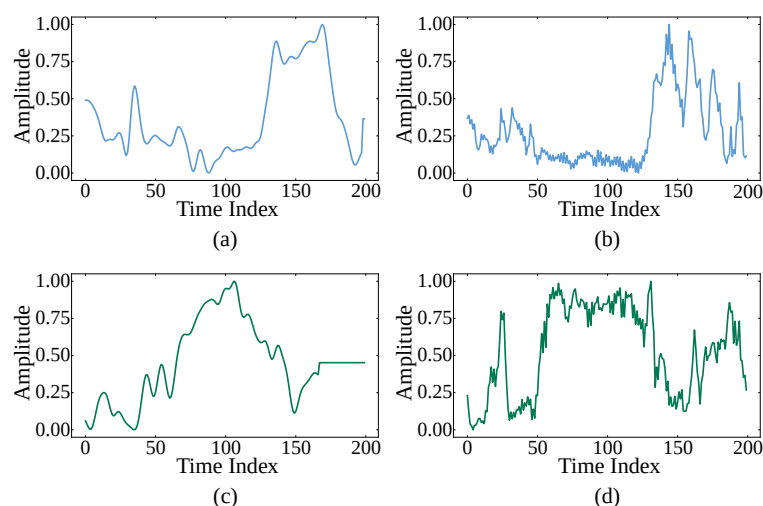


Figure 16. Amplitude variations of real and generated WiFi data on the DUAL-Net dataset. (a) Real WiFi data of user U1 for gesture G4. (b) Generated WiFi data of user U1 for gesture G4. (c) Real WiFi data of user U1 for gesture G1. (d) Generated WiFi data of user U1 for gesture G1.

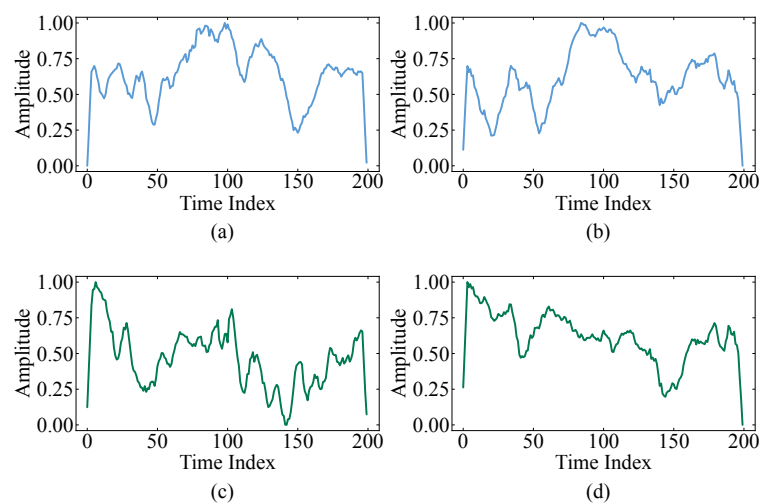


Figure 17. Amplitude variations of real and generated WiFi data on the MM-Fi dataset. (a) Real WiFi data of user U10 for gesture G2. (b) Generated WiFi data of user U10 for gesture G2. (c) Real WiFi data of user U10 for gesture G6. (d) Generated WiFi data of user U10 for gesture G6.

5.5. Additional Experimental Analysis

To further examine the behavior of DUAL-Net beyond the main recognition results, we analyze its adaptation performance, training process, confusion matrix analysis, and computational cost. These analyses are intended to clarify how the number of generated samples affects cross-user recognition, whether the training process converges stably, which gesture categories remain difficult to distinguish, and whether the model can support efficient inference.

(1) Adaptation and training behavior. Figure 18 shows the cross-user recognition accuracy obtained with different numbers of generated WiFi samples on the DUAL-Net and MM-Fi datasets. The accuracy generally increases as more generated samples are used, indicating that the generated data provide useful user-adaptive information when target-user WiFi samples are limited. The improvement becomes less pronounced as the number of generated samples increases, suggesting that a small number of generated samples already contributes a substantial part of the adaptation benefit.

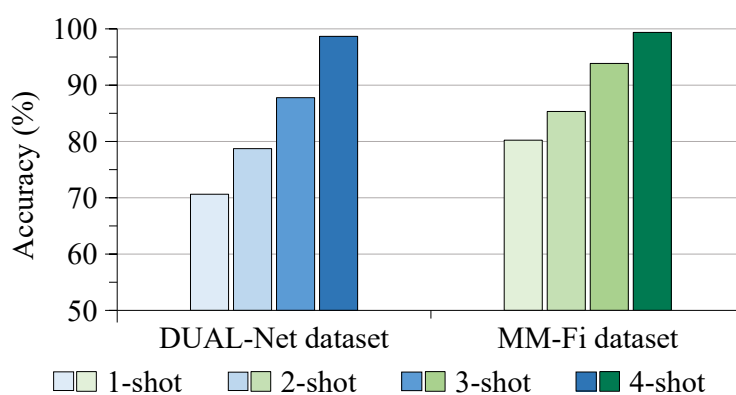


Figure 18. Cross-user accuracy with different numbers of generated WiFi samples.

Figure 19 presents the training loss curves on the two datasets. The loss decreases steadily during training and gradually approaches a stable range, indicating that the optimization process remains stable under the adopted training strategy. This result also suggests that jointly using real and generated samples does not introduce obvious instability during training.

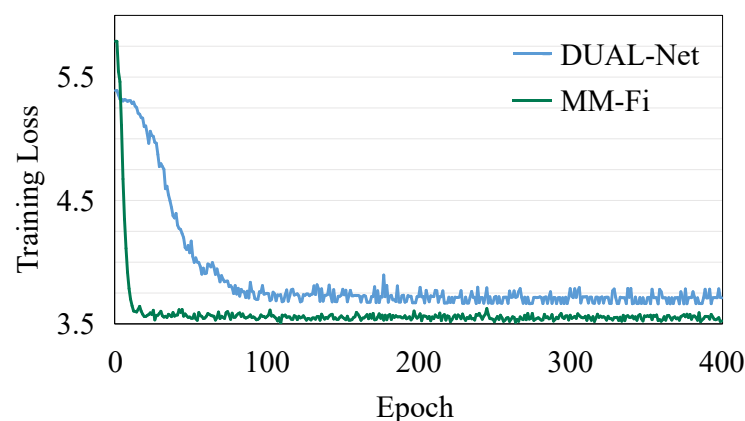


Figure 19. Training loss curves across different training epochs.

(2) Confusion matrix-based analysis: To further evaluate cross-target recognition behavior, we conduct a per-class analysis using the confusion matrix under target-user adaptation. In this experiment, the model is first trained on source-domain data and then adapted to the target user using CMG-generated WiFi samples. The corresponding results

on the DUAL-Net and MM-Fi datasets are shown in Figure 20, where (a) and (b) represent the two datasets, respectively. As shown in Figure 20a, most categories in the DUAL-Net dataset are correctly classified, with strong diagonal dominance. The main confusion occurs between G6 and G1, where G6 (*draw zigzag*) involves rapid multi-directional arm movements that partially overlap with the forward extension pattern of G1 (*push–pull*), leading to similar CSI temporal variations. A similar pattern is observed in Figure 20b. The primary misclassification occurs between G2 (*vertical arm stretch*) and G4 (*right arm extension*). Both gestures involve prominent single-arm raising or extension movements, which induce comparable amplitude fluctuations and similar energy distribution patterns in CSI signals, making them difficult to distinguish under target-user variations. Overall, the confusion matrices indicate that remaining errors mainly arise from gestures with similar motion structures and correlated CSI response patterns. While the proposed method effectively captures most class distinctions, it may still be limited in distinguishing gestures with highly similar motion trajectories and overlapping CSI variations.

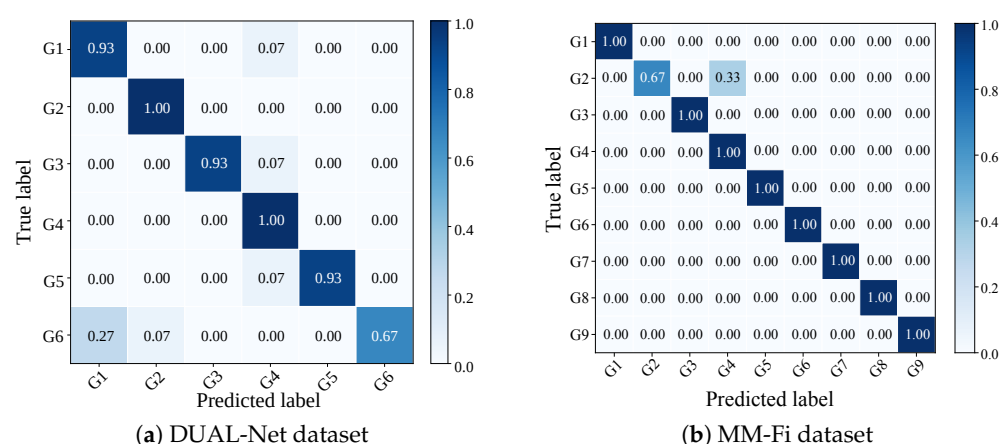


Figure 20. Confusion matrices for cross-target gesture recognition on the DUAL-Net and MM-Fi datasets. (a) Results on the DUAL-Net dataset. (b) Results on the MM-Fi dataset.

(3) Complexity and runtime analysis. DUAL-Net follows a two-stage learning paradigm consisting of offline representation learning and online user adaptation. The contrastive fusion module and the SMD-guided cross-modal generation module are employed only during offline training, which can be performed on cloud or server platforms. During deployment, gesture recognition relies on WiFi measurements and the trained recognition network. The visual-guided adaptation and diffusion-based feature generation processes are confined to the offline learning stage and are not required during online inference. Consequently, the additional computational cost introduced by these components does not affect deployment efficiency.

To quantify the computational overhead of the online stage, we evaluate the complexity and runtime of DUAL-Net under the target-user adaptation setting. The results are averaged over three profiling trials on CUDA with a batch size of 64. As shown in Table 4, DUAL-Net contains 107.40 M parameters, among which 57.93 M are trainable. The feature extraction model contains 49.75 M parameters, while only 0.28 M parameters are updated during adaptation. The classifier contains 57.65 M trainable parameters and serves as the primary component for decision boundary adaptation. The inference latency of DUAL-Net is 102.10 ms per batch, corresponding to 1.60 ms per sample and a throughput of 626.84 samples per second.

Table 4. Complexity and runtime statistics of DUAL-Net under the target-user adaptation setting.

Metric	Value
Total parameters	107.40 M
Trainable parameters	57.93 M
Feature extraction model parameters	49.75 M
Trainable feature extraction model parameters	0.28 M
Classifier parameters	57.65 M
Batch size	64
Latency	102.10 ms/batch
Per-sample latency	1.60 ms/sample
FPS	626.84 samples/s

6. Conclusions

This paper presents DUAL-Net, a domain-invariant and user-adaptive learning framework for WiFi-based gesture recognition. The core of the method lies in a contrastive fusion learning (CFL) module, which learns domain-invariant representations by aligning complementary information from WiFi and vision modalities, and a spatial matrix difference (SMD)-guided cross-modal generation (CMG) module, which generates user-adaptive WiFi features conditioned on structural cues to capture inter-user variations. Extensive experiments on the MM-Fi dataset and a self-collected dataset demonstrate that DUAL-Net consistently improves cross-user recognition performance compared with existing methods, highlighting the effectiveness of jointly modeling domain-invariant and user-adaptive representations. Overall, these results indicate that combining CFL-based cross-modal representation learning with SMD-guided CMG is an effective strategy for reliable WiFi-based gesture recognition. In addition, the proposed two-stage learning framework enables efficient deployment by separating offline multimodal learning from online inference. In future work, we will extend the framework to more complex multi-user and multi-scene settings while addressing deployment challenges such as computational overhead, latency, and scalability. We will also investigate robustness under missing modalities and explore additional sensing modalities such as RFID and mmWave to improve generalization.

Author Contributions: Conceptualization, P.Q. and B.Y.; methodology, P.Q. and B.Y.; software, J.Q. and B.J.; validation, S.Z., Z.D. and Y.G.; formal analysis, S.Z., Y.G. and B.J.; investigation, Z.D., Y.G. and B.J.; resources, P.Q. and B.Y.; data curation, S.Z. and J.Q.; writing—original draft preparation, S.Z. and Z.D.; writing—review and editing, S.Z., B.Y., Y.G., B.J., J.Q. and P.Q.; visualization, Z.D. and B.J.; supervision, P.Q. and B.Y.; project administration, B.Y. and J.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly supported by the Natural Science Foundation of China (62302466), Natural Science Basic Research Program of Shanxi Province (202303021212188, 202403021222182).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding authors.

Acknowledgments: This research was partly supported by Model Service Project for Personnel Safety Monitoring in Complex Confined Spaces.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yang, Z.; Zhen, Z.; Xu, H.; Zhang, Y.; Feng, X. RF-UI: Continuous User Identification Through Gaits Using RFID. *IEEE Trans. Cogn. Commun. Netw.* **2025**, *11*, 1682–1695.
2. Cai, Z.; Li, Z.; Chen, Z.; Zhuo, H.; Zheng, L.; Wu, X.; Liu, Y. Device-Free Wireless Sensing for Gesture Recognition Based on Complementary CSI Amplitude and Phase. *Sensors* **2024**, *24*, 3414. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Wang, H.; Wang, Y.; Yu, B.; Zhan, Y.; Yuan, C.; Yang, W. Attentional composition networks for long-tailed human action recognition. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *20*, 8. [\[CrossRef\]](#)
4. Solanas, A.; Patsakis, C.; Conti, M.; Vlachos, I.S.; Ramos, V.; Falcone, F.; Postolache, O.; Pérez-Martínez, P.A.; Di Pietro, R.; Perrea, D.N.; et al. Smart health: A context-aware health paradigm within smart cities. *IEEE Commun. Mag.* **2014**, *52*, 74–81. [\[CrossRef\]](#)
5. Ohn-Bar, E.; Trivedi, M.M. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 2368–2377. [\[CrossRef\]](#)
6. Dai, Z.; Zhai, S.; Qin, P.; Chai, R.; Zhao, P. WVGR: Gesture Recognition based on WiFi-Video Fusion. In Proceedings of the 2023 IEEE/CIC International Conference on Communications in China (ICCC), Dalian, China, 10–12 August 2023; pp. 1–6.
7. Miao, Q.; Li, Y.; Ouyang, W.; Ma, Z.; Xu, X.; Shi, W.; Cao, X. Multimodal gesture recognition based on the ResC3D network. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3047–3055.
8. Gammulle, H.; Denman, S.; Sridharan, S.; Fookes, C. TMMF: Temporal multi-modal fusion for single-stage continuous gesture recognition. *IEEE Trans. Image Process.* **2021**, *30*, 7689–7701. [\[PubMed\]](#)
9. Zhai, S.; Tang, Z.; Nurmi, P.; Fang, D.; Chen, X.; Wang, Z. RISE: Robust wireless sensing using probabilistic and statistical assessments. In Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, New Orleans, LA, USA, 25–29 October 2021; pp. 309–322.
10. Miao, F.; Huang, Y.; Lu, Z.; Ohtsuki, T.; Gui, G.; Sari, H. Wi-Fi sensing techniques for human activity recognition: Brief survey, potential challenges, and research directions. *ACM Comput. Surv.* **2025**, *57*, 107. [\[CrossRef\]](#)
11. Cao, C.; Ding, Y.; Dai, M.; Gong, W.; Zhao, X. Real-time Cross-Domain Gesture and User Identification via COTS WiFi. *IEEE Trans. Mob. Comput.* **2025**, *24*, 5124–5137.
12. Zhou, Z.; Wang, F.; Gong, W. i-Sample: Augment domain adversarial adaptation models for WiFi-Based HAR. *ACM Trans. Sens. Netw.* **2024**, *20*, 38.
13. Yan, H.; Zhang, X.; Huang, J.; Feng, Y.; Li, M.; Wang, A.; Ou, W.; Wang, H.; Liu, Z. Wi-SFDAGR: WiFi-Based Cross-Domain Gesture Recognition via Source-Free Domain Adaptation. *IEEE Internet Things J.* **2025**, *12*, 24159–24173.
14. Li, X.; Chang, L.; Song, F.; Wang, J.; Chen, X.; Tang, Z.; Wang, Z. CrossGR: Accurate and low-cost cross-target gesture recognition using Wi-Fi. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 21.
15. Manzoor, M.A.; Albarri, S.; Xian, Z.; Meng, Z.; Nakov, P.; Liang, S. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *20*, 74. [\[CrossRef\]](#)
16. He, L.; Wang, Z.; Wang, L.; Li, F. Multimodal mutual attention-based sentiment analysis framework adapted to complicated contexts. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 7131–7143. [\[CrossRef\]](#)
17. Zhu, F.; Zhu, Y.; Chang, X.; Liang, X. Vision-language navigation with self-supervised auxiliary reasoning tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10012–10022.
18. Deng, L.; Yang, J.; Yuan, S.; Zou, H.; Lu, C.X.; Xie, L. GaitFi: Robust device-free human identification via WiFi and vision multimodal learning. *IEEE Internet Things J.* **2022**, *10*, 625–636.
19. Gu, Y.; Zhang, X.; Yan, H.; Huang, J.; Liu, Z.; Dong, M.; Ren, F. WiFE: WiFi and vision Based Unobtrusive Emotion Recognition via Gesture and Facial Expression. *IEEE Trans. Affect. Comput.* **2023**, *14*, 2567–2581. [\[CrossRef\]](#)
20. Liu, X.; Tang, S.; Zhang, B.; Wu, J.; Ma, X.; Wang, J. WiVi-GR: Wireless-Visual Joint Representation Based Accurate Gesture Recognition. *IEEE Internet Things J.* **2024**, *11*, 2701–2711.
21. Zhang, C.; Yang, Z.; He, X.; Deng, L. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 478–493. [\[CrossRef\]](#)
22. Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; Liu, J. Human action recognition from various data modalities: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 3200–3225.
23. Ma, Y.; Zhou, G.; Wang, S. WiFi sensing with channel state information: A survey. *ACM Comput. Surv.* **2019**, *52*, 46.
24. Tan, S.; Ren, Y.; Yang, J.; Chen, Y. Commodity WiFi sensing in ten years: Status, challenges, and opportunities. *IEEE Internet Things J.* **2022**, *9*, 17832–17843. [\[CrossRef\]](#)
25. He, W.; Wu, K.; Zou, Y.; Ming, Z. WiG: WiFi-based gesture recognition system. In Proceedings of the 2015 24th International Conference on Computer Communication and Networks (ICCCN), Las Vegas, NV, USA, 3–6 August 2015; pp. 1–7.
26. Abdelnasser, H.; Youssef, M.; Harras, K.A. WiGest: A ubiquitous WiFi-based gesture recognition system. In Proceedings of the 2015 IEEE Conference on Computer Communications (INFOCOM), Hong Kong, China, 26 April–1 May 2015; pp. 1472–1480.

27. Li, H.; Yang, W.; Wang, J.; Xu, Y.; Huang, L. WiFinger: Talk to your smart devices with finger-grained gesture. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg Germany, 12–16 September 2016; pp. 250–261.
28. Zhang, O.; Srinivasan, K. Mudra: User-friendly fine-grained gesture recognition using WiFi signals. In Proceedings of the 12th International Conference on emerging Networking EXperiments and Technologies, Irvine, CA, USA, 12–15 December 2016; pp. 83–96.
29. Yang, J.; Chen, X.; Zou, H.; Lu, C.X.; Wang, D.; Sun, S.; Xie, L. SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing. *Patterns* **2023**, *4*, 100703.
30. Pan, X.; Jiang, T.; Li, X.; Ding, X.; Wang, Y.; Li, Y. Dynamic hand gesture detection and recognition with WiFi signal based on 1D-CNN. In Proceedings of the 2019 IEEE International Conference on Communications Workshops (ICC Workshops), Shanghai, China, 20–24 May 2019; pp. 1–6.
31. Yang, Z.; Zhang, Y.; Qian, K.; Wu, C. SLNet: A Spectrogram Learning Neural Network for Deep Wireless Sensing. In Proceedings of the 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23), Boston, WA, USA, 17–19 April 2023; pp. 1221–1236.
32. Zhang, Y.; Zheng, Y.; Qian, K.; Zhang, G.; Liu, Y.; Wu, C.; Yang, Z. Widar3.0: Zero-effort cross-domain gesture recognition with Wi-Fi. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 8671–8688.
33. Mansour, A.; Chen, W.; Ali, E.; Wang, J.; Weng, D. Towards Scalable Indoor Positioning Systems (IPS): User-Centric Challenges, Methods, and Recommendations for User-Friendly Crowd-Powered Framework. *Inform. Fusion* **2025**, *127*, 103718.
34. Mansour, A.; Ye, J.; Li, Y.; Luo, H.; Wang, J.; Weng, D.; Chen, W. Everywhere: A framework for ubiquitous indoor localization. *IEEE Internet Things J.* **2022**, *10*, 5095–5113. [[CrossRef](#)]
35. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 1–35.
36. Jiang, W.; Miao, C.; Ma, F.; Yao, S.; Wang, Y.; Yuan, Y.; Xue, H.; Song, C.; Ma, X.; Koutsonikolas, D.; et al. Towards environment independent device free human activity recognition. In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, New Delhi, India, 29 October–2 November 2018; pp. 289–304.
37. Li, L.; Wang, L.; Han, B.; Lu, X.; Zhou, Z.; Lu, B. Subdomain adaptive learning network for cross-domain human activities recognition using WiFi with CSI. In Proceedings of the 2021 IEEE 27th International Conference on Parallel and Distributed Systems (ICPADS), Beijing, China, 14–16 December 2021; pp. 1–7.
38. Hassan, M.; Kelsey, T.; Rahman, F. Adversarial AI applied to cross-user inter-domain and intra-domain adaptation in human activity recognition using wireless signals. *PLoS ONE* **2024**, *19*, e0298888.
39. Yu, X.; Yoo, S.; Lin, Y. Clipceil: Domain generalization through clip via channel refinement and image-text alignment. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 4267–4294.
40. Virmani, A.; Shahzad, M. Position and orientation agnostic gesture recognition using WiFi. In Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services, Niagara Falls, NY, USA, 19–23 June 2017; pp. 252–264.
41. Gao, R.; Li, W.; Xie, Y.; Yi, E.; Wang, L.; Wu, D.; Zhang, D. Towards robust gesture recognition by characterizing the sensing quality of WiFi signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2022**, *6*, 11. [[CrossRef](#)]
42. Yin, G.; Zhang, J.; Shen, G.; Chen, Y. FewSense, towards a scalable and cross-domain Wi-Fi sensing system using few-shot learning. *IEEE Trans. Mob. Comput.* **2022**, *23*, 453–468.
43. Wang, Z.; Li, J.; Wang, W.; Dong, Z.; Zhang, Q.; Guo, Y. Review of few-shot learning application in CSI human sensing. *Artif. Intell. Rev.* **2024**, *57*, 195. [[CrossRef](#)]
44. Jabeen, S.; Li, X.; Amin, M.S.; Bourahla, O.; Li, S.; Jabbar, A. A review on methods and applications in multimodal deep learning. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 76. [[CrossRef](#)]
45. Dong, H.; Liu, M.; Zhou, K.; Chatzi, E.; Kannala, J.; Stachniss, C.; Fink, O. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 5672–5691. [[CrossRef](#)] [[PubMed](#)]
46. Zou, H.; Yang, J.; Prasanna Das, H.; Liu, H.; Zhou, Y.; Spanos, C.J. WiFi and vision multimodal learning for accurate and robust device-free human activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 426–433.
47. Guo, J.; Shi, M.; Zhu, X.; Huang, W.; He, Y.; Zhang, W.; Tang, Z. Improving human action recognition by jointly exploiting video and WiFi clues. *Neurocomputing* **2021**, *458*, 14–23. [[CrossRef](#)]
48. Fan, J.; Zhou, H.; Zhou, F.; Wang, X.; Liu, Z.; Li, X.Y. WiVi: WiFi-Video Cross-Modal Fusion based Multi-Path Gait Recognition System. In Proceedings of the 2022 IEEE/ACM 30th International Symposium on Quality of Service (IWQoS), Oslo, Norway, 10–12 June 2022; pp. 1–10.
49. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404.

50. Chi, G.; Yang, Z.; Wu, C.; Xu, J.; Gao, Y.; Liu, Y.; Han, T.X. RF-diffusion: Radio signal generation via time-frequency diffusion. In Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, Washington, DC, USA, 18–22 November 2024; pp. 77–92.
51. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
52. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. In Proceedings of the International Conference on Learning Representations (ICLR 2021), Vienna, Austria, 4 May 2021; pp. 1–22.
53. Pavllo, D.; Feichtenhofer, C.; Grangier, D.; Auli, M. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7753–7762.
54. Yang, J.; Huang, H.; Zhou, Y.; Chen, X.; Xu, Y.; Yuan, S.; Zou, H.; Lu, C.X.; Xie, L. MM-Fi: Multi-Modal Non-Intrusive 4D Human Dataset for Versatile Wireless Sensing. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 18756–18768.
55. Ma, H.; Li, W.; Zhang, X.; Gao, S.; Lu, S. AttnSense: Multi-level attention mechanism for multimodal human activity recognition. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019), Macao, China, 10–16 August 2019; pp. 3109–3115.
56. Yao, S.; Hu, S.; Zhao, Y.; Zhang, A.; Abdelzaher, T. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In Proceedings of the 26th International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 351–360.
57. Cui, S.; Wang, R.; Wei, J.; Li, F.; Wang, S. Grasp state assessment of deformable objects using visual-tactile fusion perception. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 538–544.
58. Ouyang, X.; Shuai, X.; Zhou, J.; Shi, I.W.; Xie, Z.; Xing, G.; Huang, J. Cosmo: Contrastive fusion learning with small data for multimodal human activity recognition. In Proceedings of the 28th Annual International Conference on Mobile Computing and Networking, Sydney, NSW, Australia, 17–21 October 2022; pp. 324–337.
59. Zhang, X.; Tang, C.; Yin, K.; Ni, Q. WiFi-based cross-domain gesture recognition via modified prototypical networks. *IEEE Internet Things J.* **2021**, *9*, 8584–8596.
60. Ali, H.; Jirak, D.; Wermter, S. Snapure—A novel neural architecture for combined static and dynamic hand gesture recognition. *Cogn. Comput.* **2023**, *15*, 2014–2033. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.