Seeing the Invisible: Recovering Surveillance Video With COTS mmWave Radar

Mingda Han, Huanqi Yang, Mingda Jia, Weitao Xu, *Member, IEEE*, Yanni Yang, Zhijian Huang, Jun Luo, *Fellow, IEEE*, Xiuzhen Cheng, *Fellow, IEEE*, and Pengfei Hu

***Abstract*—Video surveillance systems play a crucial role in ensuring public safety and security by capturing and monitoring critical events in various areas. However, traditional surveillance cameras face limitations when it comes to malicious physical damage or obscuring by offenders. To overcome this limitation, we propose M^2 VISION, which is the first millimeter-wave (mmWave)-based video reconstruction system designed to enhance existing video surveillance cameras. M^2 VISION utilizes mmWave to sense the profile and motion signature of the target, integrating it with previously acquired visual data about the environment and the target's appearance, thereby facilitating the reconstruction of surveillance video. Specifically, our proposed system incorporates a dual-stage mmWave signal denoising algorithm to efficiently eliminate the noise and multiple-input multiple-output virtual antenna enhanced heatmap generation (MVAE-HG) method to obtain fine-grained mmWave heatmaps responsive to the target's profile and motion information. Moreover, we design the mm2Video generative network that first employs a multi-modal fusion module to fuse the mmWave and pre-acquired visual data, then use a conditional generative adversarial network (cGAN)-based video reconstruction module for surveillance video reconstruction. We conducted comprehensive experiments on M^2 VISION using a commercial mmWave radar and four surveillance cameras across various environments, with the participation of seven individuals. Evaluation results show that M^2 VISION can achieve an average structural similarity index measure (SSIM) of 0.93, demonstrating its effectiveness and potential.**

***Index Terms*—Deep generative network, video reconstruction, mmWave sensing.**

Manuscript received 28 February 2024; revised 17 June 2024; accepted 14 August 2024. Date of publication 19 August 2024; date of current version 5 November 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB3100400, in part by the National Natural Science Foundation of China under Grant 62202276 and Grant 62232010, in part by Shandong Science Fund for Excellent Young Scholars under Grant 2022HWYQ-038, in part by Shandong Science Fund under Grant 2023TSGC0105, in part by the National Natural Science Foundation of China under Grant 62302274, and in part by the Department of Science and Technology of Shandong Province under Grant 2024HWYQ-021 and Grant ZR2023QF113. Recommended for acceptance by H. Shen. *(Corresponding author: Yanni Yang.)*

Mingda Han, Yanni Yang, Xiuzhen Cheng, and Pengfei Hu are with the School of Computer Science and Technology, Shandong University, Qingdao 266237, China (e-mail: mingdhan@mail.sdu.edu.cn; yanniyang@sdu.edu.cn).

Huanqi Yang and Weitao Xu are with the Department of Computer Science, City University of Hong Kong Hong Kong, SAR 999077, China.

Mingda Jia is with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China.

Zhijian Huang is with the National Key Laboratory of Science and Technology on Information System Security, Beijing 101399, China.

Jun Luo is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798.

Digital Object Identifier 10.1109/TMC.2024.3445507

I. Introduction

With the escalating need for enhanced security in contemporary society, video surveillance systems have emerged as a critical element not only in industrial and commercial sectors but also within residential environments. These surveillance systems, underpinned by advanced technological components, offer myriad benefits, including the deterrence of unlawful activities, documentation of events, and provision of vital evidence, thereby contributing significantly to the overall security paradigm. Moreover, the video surveillance industry is undergoing rapid evolution, with the market revenue reaching $48.7 billion in 2022 and projected to reach $76.4 billion by 2027 [1].

Recognizing the critical role of cameras in security systems, offenders often target surveillance cameras as the initial attack vector, intending to incapacitate the entire video surveillance system. These attacks on surveillance cameras can be categorized into two main types: non-physical attacks and physical attacks. Non-physical attacks encompass the manipulation of surveillance systems without any physical intervention, typically exploiting system vulnerabilities or tampering with the camera's Ethernet cable to seize control of the signal from Ethernet surveillance cameras [2]. These intrusions have the capability to manipulate surveillance content, deceiving the surveillance system and generating a spoofing effect. Additionally, as intelligent surveillance systems [3] become more prevalent, adversarial attacks [4] have surfaced as a means to evade these advanced systems. On the other hand, the inherent limitations of camera line-of-sight (LoS) often leave surveillance cameras exposed, rendering them susceptible to physical attacks. Physical attacks involve direct tampering with the surveillance camera itself. This can entail using tools to obstruct the camera's view or causing physical damage to the camera, rendering it incapable of capturing visual information.

Efforts to combat **non-physical attacks** on surveillance systems include traditional watermarking-based [5], [6] and statistical features-based [7], [8] methods. Deep learning techniques [9], [10] have also been applied for this purpose. Furthermore, researchers are exploring the potential of using Wi-Fi signal [11] for video tampering detection, showcasing a broadened scope for countering non-physical attacks against surveillance systems through diverse technological avenues. Despite the substantial efforts in countering non-physical attacks, addressing **physical attacks** on surveillance systems remains a significant challenge,

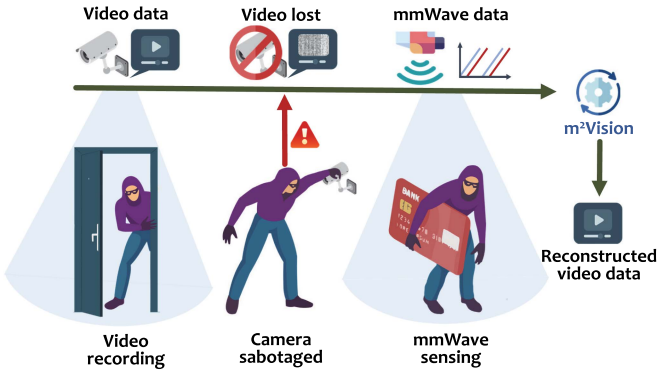


Fig. 1. **Scenario of M^2 VISION .** When the surveillance camera is vandalized or obscured, M^2 VISION uses mmWave devices to sense the target and reconstruct the surveillance video.

as existing software-level techniques are inherently inadequate to address such attacks. An effective countermeasure against physical attacks involves restoring compromised or missing video data through alternative data modalities. Radio frequency (RF) signals emerge as a promising option for this purpose, primarily attributed to their indifference to variations in light conditions and their ubiquity. One notable endeavor in this field is Wi2Vi [12], which leverages Wi-Fi signals for video reconstruction. However, Wi2Vi is limited to producing grayscale videos, and the resulting representation of human subjects is excessively indistinct. Further, although CSI2Video [13] offers the capability to generate color video frames, its efficacy is demonstrated in a singular scenario, leaving the performance in alternative scenarios unexplored.

Although Wi-Fi devices are ubiquitous, Wi-Fi-based sensing solutions encounter several notable limitations. First, schemes using Wi-Fi require modifications to the network interface cards (NICs), and Wi-Fi-based approaches must operate two devices (i.e., transmitter and receiver) simultaneously, thereby exacerbating user inconvenience and complicating deployment efforts. Furthermore, the inherent limitations and severe indoor multipath effects of Wi-Fi signals exacerbate the system's inability to achieve accurate sensing, which significantly hampers the system's ability to capture the intricate details and subtleties of the target.

In this paper, we introduce M^2 VISION, the first millimeter-wave (mmWave)-based surveillance video reconstruction system designed to enhance the functionality of surveillance cameras. The proposed system offers a novel solution to reconstruct lost surveillance video by leveraging the fine-grained sensing capabilities of commercial off-the-shelf (COTS) mmWave radar. M^2 VISION utilizes the fine-grained sensing capability of mmWave radar to sense the target's profile and motion information, and reconstruct the surveillance video by combining the target's appearance and environment information previously captured. The application scenario of M^2 VISION is illustrated in Fig. 1. When the target breaks into the surveillance area, he/she destroys or obscures the surveillance camera, resulting in the loss of surveillance video data. At this time, the mmWave radar senses the environmental intruder and reconstructs the surveillance video frames by combining the target's appearance information obtained before the camera was destroyed or obscured and the known environmental information. Two challenges need to be addressed to realize M^2 VISION:

Challenge 1. Fine-grained profile and motion feature extraction from single-chip COTS mmWave Radar: Though single-chip mmWave radar excels at detecting the presence or velocity of the target, the reconstruction of video demands micro-level detail, necessitating a novel approach to denoise the mmWave signal and extract fine-grained features of the target. These features encompass the directionality of movement, gait patterns, and three-dimensional positioning, collectively forming the target's detailed profile and motion signature. To address this, we introduce a dual-stage mmWave signal denoising algorithm to perform noise reduction before extracting the target's range and angular information, thus effectively mitigating the impact of ambient environmental noise. Meanwhile, to capture intricate features from the denoised mmWave data, we propose the Multiple-input multiple-output Virtual Antenna Enhanced Heatmap Generation (MVAE-HG) method to generate three kinds of fine-grained heatmaps reflecting the profile and motion signature of the target.

Challenge 2. Intrinsic difference between mmWave data and video data: Video data exhibits significant modality differences compared to mmWave data. While fine-grained mmWave features can be obtained, reconstructing the video data from these features poses a significant challenge. To tackle this challenge, we utilize the 3D human mesh as a medium and propose a theoretical model to explore the inherent correlation between video data and mmWave data. However, the intricate signal reflection characteristics of the human body make it challenging to derive corresponding video data from mmWave data through mathematical calculations. Thus, built on the strength of the deep generative model, we design the mm2Video network, a deep generative network with a multi-modal fusion module and a video reconstruction module, enabling effective conversion of mmWave features into video data.

The main contributions of this paper are as follows:

- We propose M^2 VISION, the first mmWave-based video reconstruction system, which overcomes the shortcomings of existing video surveillance systems by continuing to provide surveillance information via mmWave radar after the camera has been physically damaged or obscured.
- M^2 VISION proposes several approaches for reconstructing surveillance video from mmWave data with limited visual data, including a dual-stage mmWave signal denoising algorithm to remove pervasive noise, the MVAE-HG method to generate fine-grained heatmaps reflecting the target's profile and motion signature, and the mm2Video generative network to fuse different modalities features and generate surveillance video data. These methods address the key challenges presented above to recover surveillance video data.
- We conduct real-world experiments with seven participants in four different environments using different surveillance cameras. The evaluation results show that M^2 VISION

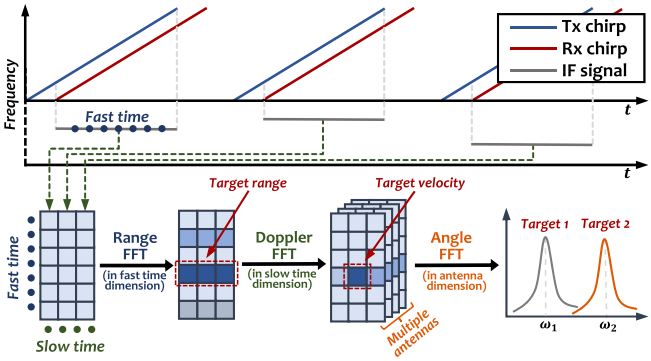


Fig. 2. Illustration of FMCW signal processing.

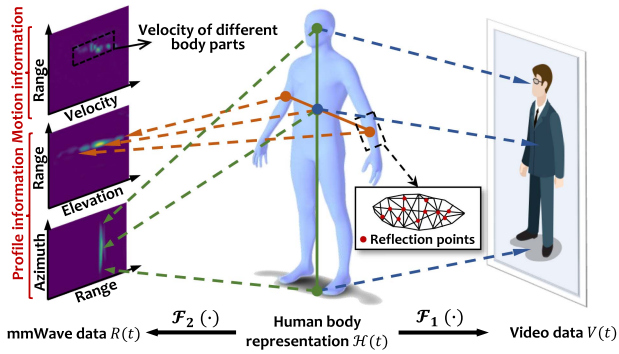


Fig. 3. **Correlation analysis.** Both the video data and mmWave data are functions ($\mathcal{F}_1$ and $\mathcal{F}_2$) of the human body. Therefore, a non-linear relationship (6) exists between them.

achieves an average structural similarity index measure (SSIM) of 0.93.

II. Preliminary

A. mmWave Sensing

The Frequency Modulated Continuous Wave (FMCW) mmWave radar employs the FMCW signal, commonly known as the chirp signal, as depicted in the upper part in Fig. 2. This chirp signal exhibits a linear increase in frequency over time t and can be described by $f = f_0 + Kt$, where f_0 represents the starting frequency and K is the frequency modulation slope. Given an amplitude A for the transmitted signal at time t, the mathematical expression for the transmitted FMCW signal $S_T(t)$ is

$$S_T(t) = Ae^{j(2\pi f_0 t + \pi K t^2)}. \tag{1}$$

Upon encountering an obstacle, such as a human, at a distance d, the mmWave radar captures a time-delayed version of the transmitted signal, identified as $S_R(t)$:

$$S_R(t) = \alpha S_T(t-\tau) = \alpha A e^{j(2\pi f_0(t-\tau)+\pi K(t-\tau)^2)}, \tag{2}$$

where α is the path loss, $\tau = 2d/c$ is the time delay, and c represents the speed of light. Subsequently, the transmitted signal $S_T(t)$ is mixed with the received signal $S_R(t)$, and the sum frequency component is filtered out by a Low Pass Filter (LPF) to obtain the Intermediate Frequency (IF) signal $S_I(t)$:

$$S_I(t) = \text{LPF}\{S_T(t) \cdot S_R(t)\} = \alpha A^2 e^{j4\pi(f_0+Kt)d/c}, \tag{3}$$

The FMCW mmWave radar offers the capability to extract essential information regarding the range, velocity, and angle properties of the target. Specifically, as shown in the lower part of Fig. 2, the range information of the target is determined by applying the Fast Fourier Transform (Range FFT) to multiple sampling points along the fast time dimension of the IF signal. Furthermore, the velocity information of the target is obtained by performing the Fast Fourier Transform (Doppler FFT) on multiple IF signals spanning the slow time dimension within a radar frame. Moreover, the angle information of the target is derived by subjecting the IF signals acquired from distinct receiving antennas to the Fast Fourier Transform (Angle FFT) operation. Collectively, these processes are commonly referred to as the 3D FFT, which responds to phase changes in different dimensions of the IF signal.

B. Correlation Analysis of Video Data and mmWave Data

Video Data: Given a video frame $V(t)$ depicting a person at time t, the 3D human mesh can be obtained by applying a human mesh recovery algorithm [14]. This algorithm represents the human as a collection of 3D points that capture both the shape (i.e., variations in height, weight, and body proportions) and the body's pose (i.e., articulation-induced deformations). To model the human body, we utilize the EllipBody representation [15], which employs L primitive ellipsoids to represent different body parts, as illustrated in Fig. 3. This representation offers a concise description of the underlying principles of mmWave signal reflections by the body, as described in the subsequent paragraph. The set of 3D points representing the l^{th} body part at time t can be denoted as $\mathcal{M}_l(t) = \{\mathbf{x}_n^l(t) \in \mathcal{R}^3, n = 1, \ldots, N_l\}$. Consequently, the complete human body can be expressed as $\mathcal{H}(t) = \sum_{l=1}^{L} \mathcal{M}_l(t)$. Since this 3D point set is derived from the video frame, there exists a mapping relationship denoted as $\boldsymbol{\mathcal{F}_1}(\cdot)$ between the video frame and the human body, which can be expressed as follows:

$$V(t) = \boldsymbol{\mathcal{F}_1}(\mathcal{H}(t) + \mathcal{A}(t)) = \boldsymbol{\mathcal{F}_1}\left(\sum_{l=1}^{L} \mathcal{M}_l(t) + \mathcal{A}(t)\right), \tag{4}$$

where $\mathcal{A}(t)$ represents additional information such as background and human appearance.

mmWave Data: When the transmitted chirp signal is reflected by the surface of the l^{th} body part, the received signal (or the mixed IF signal) carries information about that specific body part. This information is determined by two factors: the surface area and the orientation of the body part. For instance, the human torso exhibits higher reflectivity compared to other body parts due to its larger surface area, resulting in a larger radar cross section (RCS) captured by the scale parameter α in (2). The orientation of the body part determines the direction of signal reflection, which affects the phase of the received chirp signal and subsequently the phase of the IF signal. The mmWave data $R(t)$ that we receive can be considered as a synthesized signal of the reflected signals from all L body parts at time t. It encapsulates valuable information about the human profile, including the size and orientation of each body part, as well as the motion characteristics of different body parts, as depicted in Fig. 3. In real-world scenarios, mmWave radar signals are often

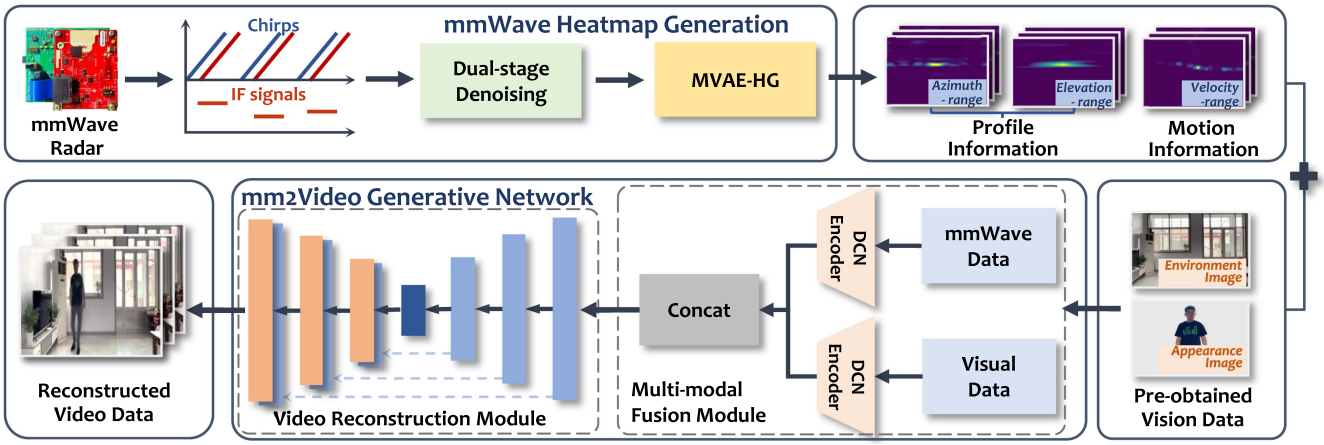


Fig. 4. **System overview.** M^2 VISION begins by using dual-stage denoising and MVAE-HG algorithm to generate detailed target profiles and motion heatmaps. Once these heatmaps are generated, they are combined with previously obtained target appearance and environmental vision information through a DCN-based multi-modal fusion module to ensure efficient fusion of different data types. Finally, a cGAN-based video reconstruction module is designed to generate the final surveillance video.

contaminated with noise from various environmental sources, as well as reflections originating from the human body. Consequently, the mmWave data can be mathematically modeled as a function of the 3D human mesh and ambient noise:

$$R(t) = \boldsymbol{\mathcal{F}_2}(\mathcal{H}(t) + \mathcal{N}(t)) = \boldsymbol{\mathcal{F}_2}\left(\sum_{l=1}^{L} \mathcal{M}_l(t) + \mathcal{N}(t)\right), \tag{5}$$

where $\boldsymbol{\mathcal{F}_2}(\cdot)$ represents the mapping relationship from the human body to the received mmWave data, and $\mathcal{N}(t)$ represents the additional noise interference.

Correlation: As indicated by (4) and (5), both the video data $V(t)$ and mmWave data $R(t)$ can be defined as functions of the 3D points of the human body. Therefore, the mmWave data, along with supplementary visual information such as background and human appearance details, can be transformed into video data through a non-linear function $\boldsymbol{\mathcal{G}}(\cdot)$:

$$V(t) = \boldsymbol{\mathcal{G}}\left(R(t), \mathcal{A}(t)\right). \tag{6}$$

Nonetheless, accurately determining the exact form of the function $\boldsymbol{\mathcal{G}}(\cdot)$ using traditional mathematical methods is challenging due to the inherent complexity and non-linearity associated with the transformation between the distinct modalities of mmWave and video data. In this work, we leverage the remarkable non-linear fitting capabilities of deep learning to train a deep generative model. This model establishes the mapping relationship between mmWave data and video data, enabling accurate transformation between these two modalities.

III. OVERVIEW

A. Problem Statement

In this paper, we consider the problem of recovering surveillance video using mmWave signals. Specifically, in scenarios where surveillance equipment is interfered with, maliciously destroyed, or obscured, we aim to utilize a single commercial single-chirp mmWave device to sense the profile and motion information of the target within a monitored area and reconstruct the surveillance video.

Let $R(t)$ represent the mmWave data captured when the camera is malfunctioning. Given that we have knowledge of the background information I_e and can capture the appearance information of the target I_a before the camera malfunction, our objective is to recover the video frame $V(t)$ using the following equation:

$$V(t) = \boldsymbol{f_\theta}(\boldsymbol{g}(R(t)), I_e, I_a), \tag{7}$$

where $\boldsymbol{f_\theta}(\cdot)$ is our proposed video reconstruction model with learnable parameters θ, and $\boldsymbol{g}(\cdot)$ is the designed mmWave feature extraction algorithm. Consequently, $\boldsymbol{g}(R(t))$ describes the relationship between extracted features and reflected mmWave signals from L body parts.

B. System Overview

As shown in Fig. 4, M^2 VISION has two parts: mmWave Heatmap Generation and mm2Video Generative Network.

The mmWave heatmap generation focuses on generating heatmaps that encapsulate the human profile and motion information of the target, which are subsequently utilized for video reconstruction. To achieve this, the raw mmWave data undergoes a two-stage denoising algorithm, meticulously designed to remove various types of noise effectively. Following that, the MIMO virtual antenna enhanced heatmap generation (MVAE-HG) method is employed to produce three types of fine-grained heatmaps: azimuth-range heatmap, elevation-range heatmap, and range-Doppler heatmap. The azimuth-range and elevation-range heatmaps efficiently capture the target's profile information, while the range-Doppler heatmap represents the target's motion information.

The mm2Video generative network consists of two key modules: the multi-modal fusion module and the video reconstruction module. The main purpose of the multi-modal fusion module is to encode and merge the mmWave data with pre-existing

Algorithm 1: Dual-stage mmWave Denoising.

Input: $\boldsymbol{M}$: raw IF matrix; N_F : frame number; N_I : IF signal number; W: window size
Output: $\boldsymbol{M''}$: dual-stage denoised IF matrix
1 **for** $i = 0; i < N_F - 1; i = i + 1$ **do**
2 $\quad \boldsymbol{n} \leftarrow \frac{1}{N_I} \sum_{k=0}^{N_I-1} \boldsymbol{M}(i, k, :)$ ▷ Calculate the noise vector
3 $\quad$ **for** $j = 0; j < N_I - 1; j = j + 1$ **do**
4 $\quad\quad \boldsymbol{t} \leftarrow \boldsymbol{M}(i, j, :)$ ▷ The j^{th} IF signal in the i^{th} radar frame
5 $\quad\quad \boldsymbol{M'}(i, j, :) \leftarrow \boldsymbol{t} - \boldsymbol{n}$ ▷ The initial denoised IF matrix
6 $\quad$ **end**
7 **end**
8 **while** Angle Estimation **do**
9 $\quad$ **for** $i = 0; i < \lceil N_F/W \rceil; i = i + 1$ **do**
10 $\quad\quad \boldsymbol{k} \leftarrow i * W : \mathbf{min}((i + 1) * W, N_F - 1)$
11 $\quad\quad \boldsymbol{N}(i, :, :) \leftarrow \mathbf{mean}(\boldsymbol{M'}(\boldsymbol{k}, :, :))$ ▷ Calculate the noise matrix
12 $\quad$ **end**
13 $\quad$ **for** $i = 0; i < N_F - 1; i = i + 1$ **do**
14 $\quad\quad j = \lfloor i/W \rfloor$ ▷ The noise matrix index
15 $\quad\quad \boldsymbol{M''}(i, :, :) \leftarrow \boldsymbol{M'}(i, :, :) - \boldsymbol{N}(j, :, :)$
16 $\quad$ **end**
17 **end**

environmental and profile data. This integration process effectively utilizes DCN-based encoders to perform multi-modal fusion. Subsequently, the cGAN-based video reconstruction module takes the fused feature map as input to facilitate the reconstruction of surveillance videos.

IV. mmWave Posture Feature Extraction

A. Data Preprocessing

The FMCW mmWave radar faces challenges in accurately extracting target information from its surroundings due to the presence of both dynamic and static targets. In addition to capturing signals from moving targets, the mmWave radar also captures reflections from stationary objects like tables, chairs, and walls. Consequently, the resulting heatmaps may contain significant noise, making it difficult to discern the relevant information about the target. To mitigate the impact of static noise, we design a dual-stage mmWave signal denoising algorithm, outlined in Algorithm 1. Specifically, in the initial stage, to mitigate the influence of static environmental reflections, which remain consistent over short durations, we compute the average of multiple IF signals within a single radar frame (line 2). This average is treated as the static noise vector and is subtracted from all IF signals to reduce static interference (line 5). The second stage of denoising, which is described in detail in the upcoming section, further refines the pre-denoised signal for more accurate angle estimation.

B. MIMO Virtual Antenna Enhanced Heatmap Generation

1) Profile Information Acquisition: The pre-denoised signal initially employs the range-FFT in the fast time dimension to compute the target's range information. This range information provides an indication of the target's distance from the mmWave device, facilitating the determination of the target's depth information within the reconstructed video frame. In addition, to enhance the robustness against other dynamic interference targets, we utilize mmWave radar's radial range resolution capability to eliminate the effects of distant interference targets after performing range FFT. Specifically, we keep the closest target to the radar as the main target and set the signals from other dynamic targets to zero.

Despite the initial denoising stages described in Section IV-A, some residual weak static noise may remain. Directly conducting angle estimation in such conditions could amplify the effects of these noise artifacts. We utilize the multiple signal classification (MUSIC) algorithm [16] for angle measurement, which relies on the orthogonality between the signal and noise subspaces. However, if the residual noise from the initial denoising is not effectively removed, this untargeted noise may be mistakenly classified as part of the target signal. This misclassification can alter the structure of the signal subspace, exacerbating the impact of interfering noise and potentially compromising the accurate identification of the target signal's true direction. To mitigate this issue, we have developed a sliding window-based approach for additional denoising, as detailed in lines 8-17 of Algorithm 1. Specifically, our method involves constructing a noise matrix by computing the mean value of W radar frames (line 11). Subsequently, the noise matrix is subtracted from each radar frame within the sliding window, enabling secondary noise elimination (line 15). Notably, the secondary stage denoising process operates on the slow time dimension (between radar frames). This is significantly longer compared to the initial stage of denoising, which occurs in the fast time dimension (between sampling points of a single IF signal). In this context, while the target remains stationary in the fast time dimension, there is relative movement in the slow time dimension. Consequently, our proposed dual-stage denoising algorithm can effectively distinguish the target from the background noise, ensuring it is not erroneously filtered out.

The estimation of angles is facilitated using the secondary denoised data. However, the employed single-chip mmWave radar is equipped with only three transmitting antennas and four receiving antennas, which restricts the acquisition of precise angle information. To overcome the limited number of physical antennas problem, we employ the Multiple Input Multiple Output (MIMO) virtual antenna technique, which enhances the granularity of the resulting heatmaps. As depicted in Fig. 5, the MIMO virtual antenna technique allows us to effectively address the hardware limitations by virtually expanding the three-transmitting four-receiving antenna array to behave as if it were a twelve-virtual-receiving antenna array. Consequently, eight virtual receiving antennas are utilized to calculate the azimuth angle, while two virtual antennas are employed to determine the elevation angle.

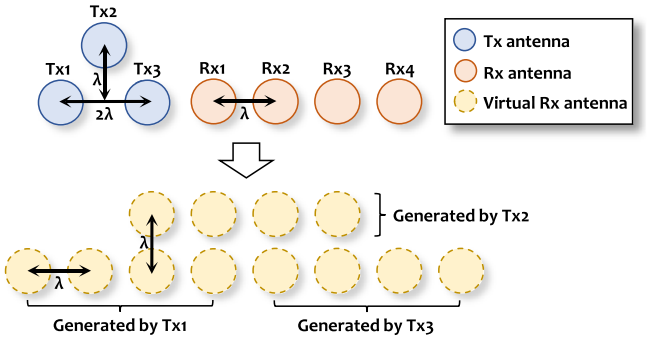
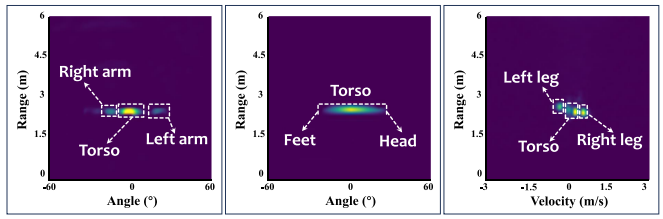


Fig. 5. Virtual antenna array leveraged in M^2 VISION.



(a) Range-azimuth. **(b)** Range-elevation. **(c)** Range-velocity.

Fig. 6. Heatmaps representing profile & motion information.

However, given that the target is in motion and the virtual antenna sets generated by Antenna 1 (Tx1) and Antenna 3 (Tx3) exhibit different air-port transmission times, the motion of the target relative to the mmWave radar induces a Doppler phase shift. This accumulated Doppler phase shift subsequently impacts the accuracy of the subsequent angle estimation. Therefore, it is imperative to execute Doppler phase compensation before proceeding with angle estimation. The Doppler frequency shift resulting from the target's motion is expressed as $\Delta f = \frac{2v}{\lambda}$, where λ denotes the wavelength, and v represents the target velocity, calculable through Doppler FFT (refer to Section IV-B-2). The Doppler phase shift can then be determined using $\Delta\varphi = 2\pi\Delta f T_c$, where T_c denotes the chirp duration Once the Doppler phase shift is computed, it is essential to apply Doppler phase compensation to the denoised data. For the data represented by $M(n,:)$ in the antenna dimension, where n denotes the virtual antenna index, Doppler compensation can be executed using the following equation:

$$M'(n,:) = M(n,:)e^{-jn\Delta\varphi}. \tag{8}$$

To derive the azimuth and elevation angle, we design a method based on the MUSIC, which is a high-resolution direction-finding algorithm designed for multiple antenna systems. Initially, we calculate the covariance matrix, denoted as R, of the received signal matrix X from K virtual antennas. This matrix is obtained by averaging the outer product of the received signals:

$$R = \frac{1}{K}\sum_{k=1}^{K} X(i)X^H(i), \tag{9}$$

where $(\cdot)^H$ represents the conjugate transpose operation. Subsequently, we perform the eigendecomposition of the covariance matrix:

$$R_x = \begin{bmatrix} U_s & U_n \end{bmatrix} \begin{bmatrix} \Lambda_s & \\ & \Lambda_n \end{bmatrix} \begin{bmatrix} U_s \\ U_n \end{bmatrix}, \tag{10}$$

where U_s and U_n are the signal space matrix and noise space matrix, respectively, and Λ_s and Λ_n are diagonal matrices comprising the eigenvalues in the signal space and noise space. With the eigendecomposition, the spatial spectrum can be expressed as:

$$P(\theta) = \frac{1}{\mathbf{a}^H(\theta)U_n U_n^H \mathbf{a}(\theta)}, \tag{11}$$

where $\mathbf{a}(\theta)$ is the steering vector associated with the desired angle θ. Finally, we compute the spatial spectrum corresponding to each angle. Combining this angle information with the previously obtained range information, we generate the heatmaps that effectively represent the target's profile information. The resulting two heatmaps are presented in Fig. 6(a) and (b), respectively.

2) Motion Information Acquisition: In addition to acquiring profile information, obtaining the velocity information of the target is crucial. This information plays a vital role in the mm2Video generative network, as it enables the differentiation of various body components based on their unique velocity characteristics, thereby enhancing the accuracy of reconstructed video frames. To achieve this, we directly utilize the first-stage denoised data and perform the Doppler FFT along the slow time dimension. This operation effectively extracts the target's velocity information. The resulting range-velocity heatmap, as depicted in Fig. 6(c), showcases the mmWave radar's capability to accurately discern distinct velocity components associated with different human body parts.

We leverage three heatmaps generated from mmWave signals, instead of using raw mmWave data, as inputs for the subsequent mm2Video generation network. This approach primarily aims to minimize noise and enhance the effectiveness of the subsequent learning process. The mmWave heatmaps filter out irrelevant noise, sharpening the focus on critical data patterns related to the target's profile and motion.

V. MM2VIDEO GENERATIVE NETWORK

Drawing on the formulae delineated in Section II-B, we can infer a correlation between video and mmWave sensing data for monitoring purposes. However, this relationship exhibits a high degree of complexity, rendering it difficult to fully encapsulate using traditional mathematical methods. Our primary objective is to transform mmWave data into video data. To accomplish this, we propose the mm2Video generative network as shown in Fig. 7. The designed mm2Video generative network comprises two key components: a multi-modal fusion module and a video reconstruction module.

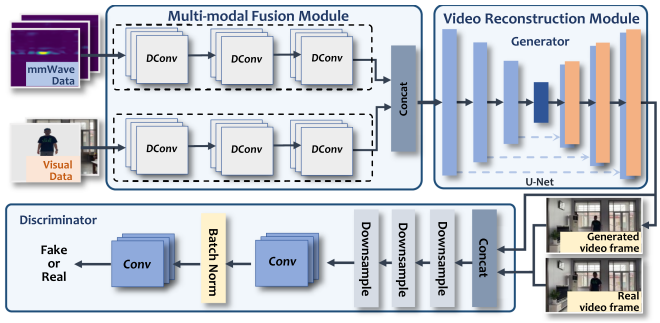


Fig. 7. mm2Video generative network.

A. Multi-Modal Fusion Module

Traditional image processing networks [17], [18] are structured to handle individual RGB images as their inputs, functioning on a per-image level. In contrast, the task of reconstructing video from surveillance footage necessitates the integration of features from disparate sources – specifically, optical and millimeter-wave (mmWave) data – to produce comprehensive video frames. Relying on unimodal features alone can lead to missing critical features, as data from both modalities are critical to recovering video frames. In order to effectively fuse data from different modalities, we propose a multi-modal fusion module that combines the three mmWave heatmaps and two visual images for a more effective representation of the surveillance video frame.

Standard Convolutional Neural Networks (CNNs) are commonly utilized for image fusion tasks, relying on the fundamental convolution operation:

$$\mathcal{Y}(p) = \sum_{p' \in \Omega} \mathcal{X}(p + p') * K(p'), \tag{12}$$

where $\mathcal{Y}(p)$ denotes the output feature map, $\mathcal{X}(p)$ represents the input feature map, $K(p')$ is the convolution kernel, and Ω defines the neighborhood around the pixel p.

The mmWave heatmaps capture distinct features such as the range, velocity, and angle of the target, the shape of these features in heatmap varies greatly from scene to scene or from relative position to position, even for the same target. Traditional CNNs, as mentioned above, employ a fixed geometry in their convolutional kernels, which can hinder their performance when dealing with the diverse scaling, rotation, and deformation characteristics present in mmWave heatmaps. These factors make it challenging to effectively model and extract features from mmWave heatmaps and fuse them with visual features using traditional CNNs.

To address the aforementioned challenge, we propose a multi-modal fusion module based on Deformable Convolutional Networks (DCNs) [19]. DCNs overcome this challenge by incorporating a learnable offset parameter within the convolutional filter. This parameter enables the network to dynamically modify the spatial sampling locations of the input mmWave heatmaps, adapting to its specific geometric distortions. This deformable mechanism significantly enhances the DCNs' capability to extract relevant features from mmWave heatmaps, unaffected by the typical distortions and anomalies, thereby improving target feature extraction efficiency. This process can be expressed as:

$$\mathcal{Y}(p) = \sum_{p' \in \Omega} \mathcal{X}(p + p' + \Delta p') * K(p'), \tag{13}$$

where $\Delta p'$ denotes the offset for the pixel p'. During training, the offsets are learned, allowing the model to adapt to geometric variations and focus on frequencies at different timescales. These offsets are predicted by another convolutional layer:

$$\Delta p' = \mathcal{F}_o(\mathcal{X}, K_o), \tag{14}$$

where $\mathcal{F}_o$ represents the offset layer, and K_o is the kernel for the offset layer. To handle irregular grid sampling locations, bilinear interpolation is employed:

$$\mathcal{I}(p) = \sum_{\epsilon \in N(p)} \mathcal{X}(\epsilon) * \max(0, 1 - |p_x - \epsilon_x|) * \max(0, 1 - |p_y - \epsilon_y|), \tag{15}$$

where $\mathcal{I}(p)$ denotes the interpolated value at location p, and $N(p)$ represents the set of nearest neighbor pixels around location p. Combining the deformable convolution with bilinear interpolation yields the final formula:

$$\mathcal{I}(p) = \sum_{p' \in \Omega} \mathcal{I}(p + p' + \Delta p') * K(p'). \tag{16}$$

The multi-modal fusion module employs three Deformable Convolutional (DConv) [19] layers to process the input multi-modal features. This design facilitates the fusion of three mmWave features and two visual features. DConv introduces an offset to standard grid sampling locations, thereby enabling more flexible and enhanced feature extraction. This approach optimally models the mmWave and visual features, making it well-suited for the extraction and fusion of multi-modal data. The fusion results serve as the condition for the subsequent video reconstruction module.

B. Video Reconstruction Module

The video reconstruction module employs a cGAN architecture [20]. cGANs have delivered impressive performance in image generation, restoration, and translation tasks [21], [22], [23]. They overcome issues such as mode collapse, lack of diversity, and instability that plague traditional GANs. By adding a conditional vector and random noise during image generation, cGANs offer better control and produce higher quality output. The video reconstruction module is mainly composed of two parts: the generator and the discriminator.

1) Generator: Traditional encoder-decoder networks used in generators require information flow to pass through all layers, which can increase computational and time costs, particularly for image-to-image translation problems where inputs and outputs share low-level information that does not need conversion [24]. To tackle this issue, we use the U-Net architecture [25] as our generator's network. Unlike traditional encoder-decoder

Algorithm 2: Training Process for mm2Video.

Input: $\{(c^1, r^1), \ldots, (c^n, r^n)\}$: n paired training data; μ: batch size
Output: θ_D: parameters of discriminator; θ_G: parameters of generator

for *each epoch* **do**
  for *each iteration* **do**
    Select μ paired instances from the input
    Select μ noise samples $\{n^1, \ldots, n^\mu\}$ from a distribution
    Produce synthetic data $\{\tilde{r}^1, \ldots, \tilde{r}^\mu\}$, $\tilde{r}^i = G(c^i|n^i)$
    Update discriminator parameter θ_D:

$$\theta_D \leftarrow \theta_D + \eta \nabla \tilde{V}(\theta_D)$$

    Select μ noise samples $\{n^1, \ldots, n^\mu\}$ from a distribution
    Select μ conditions $\{c^1, \ldots, c^\mu\}$ from input
    Update generator parameter θ_G:

$$\theta_G \leftarrow \theta_G + \eta \nabla \tilde{V}(\theta_G)$$

  end
end

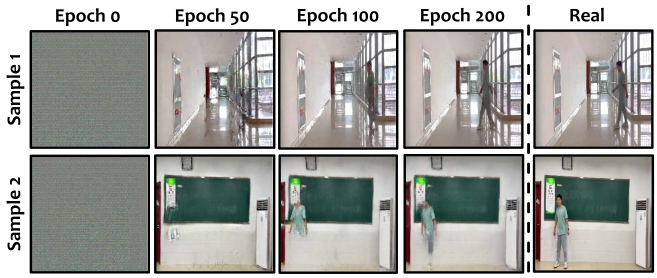
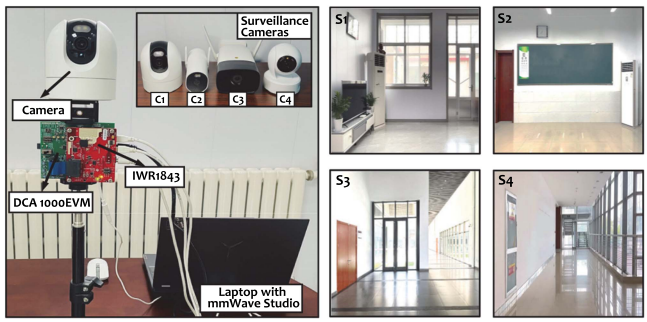


Fig. 8. Training process.



(a) Experimental devices. **(b)** Scenarios.

Fig. 9. Experimental setup.

networks, feature maps from each convolutional layer are concatenated with the corresponding upsampling layer, enabling efficient utilization of feature maps in subsequent calculations. This is referred to as skip connections, as shown by the blue dashed line in the upper panel of Fig. 7.

2) Discriminator: The discriminator is designed with three convolutional layers that work together to process the input spectrograms. In contrast to traditional discriminators that evaluate the entire video frame, our approach focuses on smaller, patch-level regions. The discriminator examines and classifies each patch within the video frame as either real or fake, providing a more detailed analysis of the generated output. The discriminator's convolutional layers are responsible for extracting and learning relevant features from these patches, enabling the model to make accurate decisions regarding their authenticity. This approach allows the model to provide a comprehensive assessment of the output, considering both local and global information within the video frame.

Training: The multi-modal fusion module has three DConv layers for mmWave and visual features, respectively, with a kernel size of 3×3 with padding of one. The video reconstruction module consists of a generator and a discriminator. The generator uses three stages: downsampling and upsampling, with concatenation. Its submodules apply a 4×4 kernel, stride of two, and padding of one. The discriminator employs three convolutional layers with a 1×1 kernel, leaky ReLU activation, and batch normalization. Models are trained for 300 epochs, using a 0.0002 learning rate for the first half and Adam optimizer for adaptive learning rate adjustments. To make the reconstructed video frames more similar to the ground truth, we use the Mean Squared Error (MSE) loss function for the magnitude of generated results and original frames.

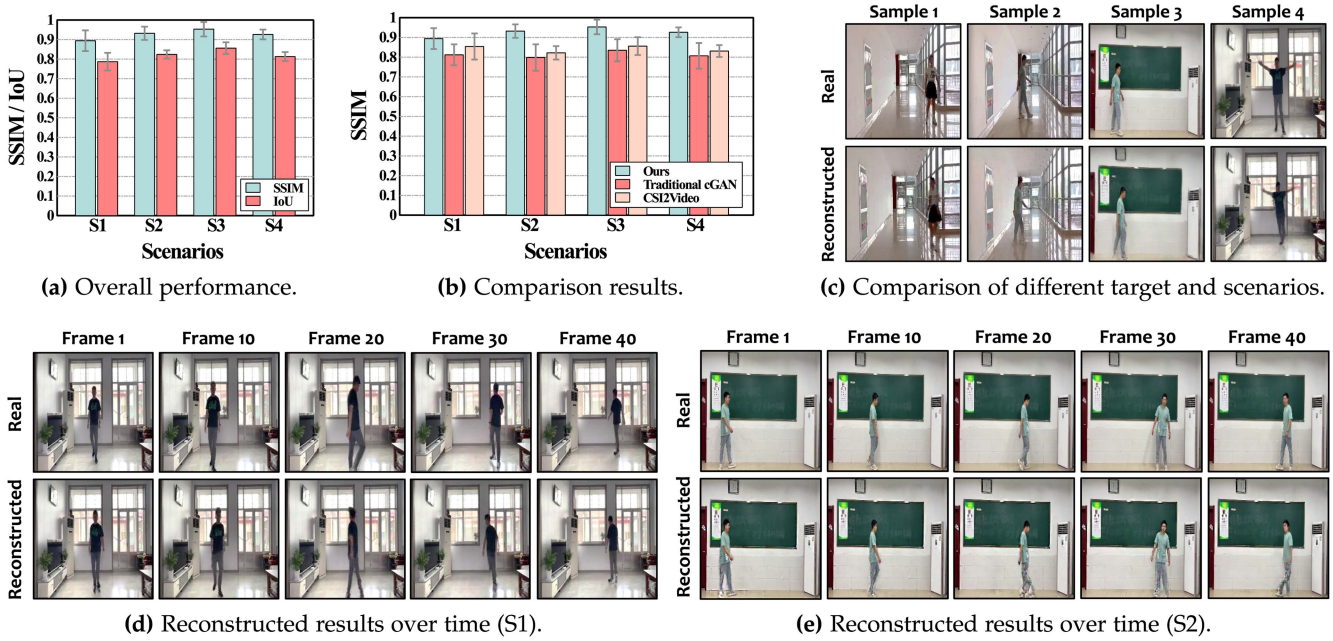
Algorithm 2 illustrates the training process, where θ_G and θ_D denote parameters of G (generator) and D (discriminator) respectively. V is the optimization objective function of cGAN [20]. Each epoch first updates θ_D with θ_G held fixed, then proceeds to update θ_G with θ_D fixed. The generator G combines the condition c with a noise vector n to generate a fake video frame $G(n|c)$. Moreover, the discriminator D receives another input that combines r and c to represent the real video frame under condition c. During training, D learns to differentiate between $G(n|c)$ and the ground truth $G(r|c)$, while G adjusts its parameters to produce a $G(n|c)$ that can deceive D. After the training, the generator G can correctly reconstruct a video frame using mmWave data. The mm2Video model's effectiveness is evident in Fig. 8, showing the initial 200 epochs of training.

VI. Evaluation

A. Experimental Setup

Experimental devices: The experimental devices used to evaluate M^2 Vision are shown in Fig. 9(a). Specifically, TI IWR1843 FMCW mmWave radar[1] and DCA1000EVM[2] are leveraged to collect mmWave data. The mmWave radar, operating at 77 GHz-81 GHz, employs three transmitter antennas to emit signals and four receiver antennas to capture signals. The default frame rate for mmWave radar is 20 FPS, with each frame containing 255 chirps and each chirp containing 256 ADC samples. Four commonly used surveillance cameras, namely Xiaomi CW400, Xiaomi AW300, EZVIZ H5, and TP-LINK IPC44AW are employed to acquire the video data. The cameras' specifications

[1]IWR1843: https://www.ti.com/product/IWR1843
[2]DCA1000EVM: https://www.ti.com/tool/DCA1000EVM



(a) Overall performance. **(b)** Comparison results. **(c)** Comparison of different target and scenarios.

(d) Reconstructed results over time (S1). **(e)** Reconstructed results over time (S2).

Fig. 10. Experimental results.

TABLE I
CAMERA SPECIFICATIONS

	Camera Type	Resolution	Frame Rate	Lens Resolution
C1	Xiaomi CW400	1440*2340	20 FPS	4 MP
C2	Xiaomi AW400	1296*1920	20 FPS	3 MP
C3	EZVIZ H5	1920*1080	15 FPS	2 MP
C4	TL IPC44AW	1920*1080	15 FPS	4 MP

are outlined in detail in Table I. The mmWave radar is located directly below the surveillance camera, and the position of the two is fixed during the experiment.

Data collection: We recruited seven participants (four males and three females)[3] to evaluate M^2 VISION. Their height range between 1.54 m and 1.83 m, and their weight range between 47.5 kg and 81 kg. The four scenarios for collecting data are shown in Fig. 9(b). We used Xiaomi CW400 as the default surveillance camera for video data collection. In each scenario, each participant walked randomly in front of the camera and mmWave radar for 20 minutes. We downsampled the frame rate of the mmWave radar and the camera to 10 FPS. Thus, each volunteer has a total of 12,000 video frames per scenario.

Metrics: We use SSIM [26] to evaluate M^2 VISION's performance, which compares the structural information of two images or video frames, such as luminance and contrast, rather than just comparing pixel values. SSIM can be calculated as

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{17}$$

where $\mu_{(\cdot)}$ is the average of $(\cdot)$, $\sigma^2_{(\cdot)}$ is the variance of $(\cdot)$, and σ_{xy} is the covariance of x and y. In addition, to accurately evaluate the discrepancies in localization and pose on the generated target person, we utilize the Mask R-CNN [27] to extract masks from both the real and the reconstructed video frames. Subsequently, we measure the intersection over union (IoU) between these masks. The IoU is defined by the equation:

$$IoU = \frac{|A \cap B|}{|A \cup B|}, \tag{18}$$

where A and B denote the areas covered by the masks of the real and reconstructed video frames, respectively.

B. Overall Performance

We evaluate the overall performance of M^2 VISION under four scenarios. As depicted in Fig. 10(a), the mean SSIM of the reconstructed video frames and real video frames in S1, S2, S3, and S4 are 0.89, 0.93, 0.95, and 0.93, respectively. Similarly, the mean IoU for the reconstructed and real video frames in S1, S2, S3, and S4 are indicated as 0.79, 0.82, 0.86, and 0.81. The slightly lower SSIM and IoU values observed in scenario S1 relative to S2 and S3 can likely be ascribed to the increased scene complexity caused by numerous background elements such as air conditioners, televisions, tables, and chairs, which present challenges to accurate video reconstruction. Additionally, unlike SSIM, which provides a broader assessment of the overall quality of the recovered video, IoU primarily evaluates the localization and pose accuracy of the target. Consequently, the lower elevation angle resolution contributes to a comparatively lower IoU value. Despite this minor limitation, M^2 VISION demonstrates robust performance across varying scenarios, emphasizing its potential as a reliable surveillance tool.

Fig. 10(c) presents a visual comparison of the reconstructed and original video samples across various participants and

[3]Ethical approval has been granted by the corresponding organization.

scenarios. Samples 1 and 2 exemplify how M^2 VISION can successfully reconstruct two distinct subjects within the same environment. Samples 3 and 4 demonstrate the system's capacity to handle different subjects in various environments. Notably, Sample 4 provides evidence that M^2 VISION can even reconstruct a subject's motion with high accuracy. Furthermore, Fig. 10(d) and (e) illustrate the video reconstruction results when a participant walks freely in two randomly selected scenarios. Evidently, M^2 VISION demonstrates exceptional accuracy in capturing the target's position and pose information. However, there is a slight degradation in the reconstruction quality below the knee, attributable to the constrained elevation angle sensing capability of the mmWave radar.

C. Performance Comparison

Comparison with cGAN: We evaluate the performance differences between our proposed mm2Video generative network and the traditional cGAN by directly concatenating visual data with mmWave heatmaps as inputs to the traditional cGAN. The result of this experimental comparison is illustrated in Fig. 10(b). Compared to the traditional cGAN, our approach offers enhanced performance by extracting and fusing features from different modalities, rather than simply concatenating all data. Our proposed multi-modal fusion module effectively captures the distinct characteristics of various modalities, enabling more accurate feature representation and fusion.

Comparison with baselines: We compare our system with the state-of-the-art Wi-Fi-based video recovery system CSI2Video [13]. We used two laptops (ThinkPad T400) with Ubuntu 14.04 LTS installed as transceivers, both equipped with Intel 5300 network interface cards (NICs) with the CSI tool [28] installed, operating in IEEE 802.11n monitor mode on Channel 120 at 5.6 GHz, for collecting Wi-Fi channel state information (CSI). The comparison result is shown in Fig. 10(b). Our mmWave-based scheme outperforms the Wi-Fi-based scheme, primarily because the multipath effect of Wi-Fi signal is more pronounced compared to mmWave signal in indoor environments. Additionally, the lower resolution of CSI further diminishes its performance.

D. Dependence on Visual Data

We evaluate how our system's performance varies with different degrees of visual data absence. Given that our network's inputs are fixed, we cannot directly remove the visual modal data. Instead, we substituted the visual data with 2D Gaussian white noise to simulate the absence of visual inputs. We conducted experiments where we substituted Gaussian white noise for 1) the environment image, 2) the appearance image, and 3) both the environment and appearance images. The results are presented in Table II.

There is a notable decline in performance after the removal of the visual modal data, particularly when environmental information is excluded. In our scheme, the reconstructed video frames inherently consist of visual data. This a priori visual data serves as the foundational basis for generating subsequent video frames. The target pose and position information, derived from mmWave data, are then superimposed on these frames to create complete video visuals. Without this initial visual data, the framework lacks a fundamental basis for video recovery, and thus a priori visual data is critical to recovery results.

TABLE II
DEPENDENCE ON VISUAL DATA

Scenario	w/. env. & app. info.	w/o. env. info.	w/o. app. info.	w/o. env. & app. info.
S1	0.89	0.54	0.74	0.40
S2	0.93	0.58	0.73	0.38
S3	0.95	0.60	0.72	0.35
S4	0.93	0.61	0.73	0.37

env. & app. info.: environment and appearance information

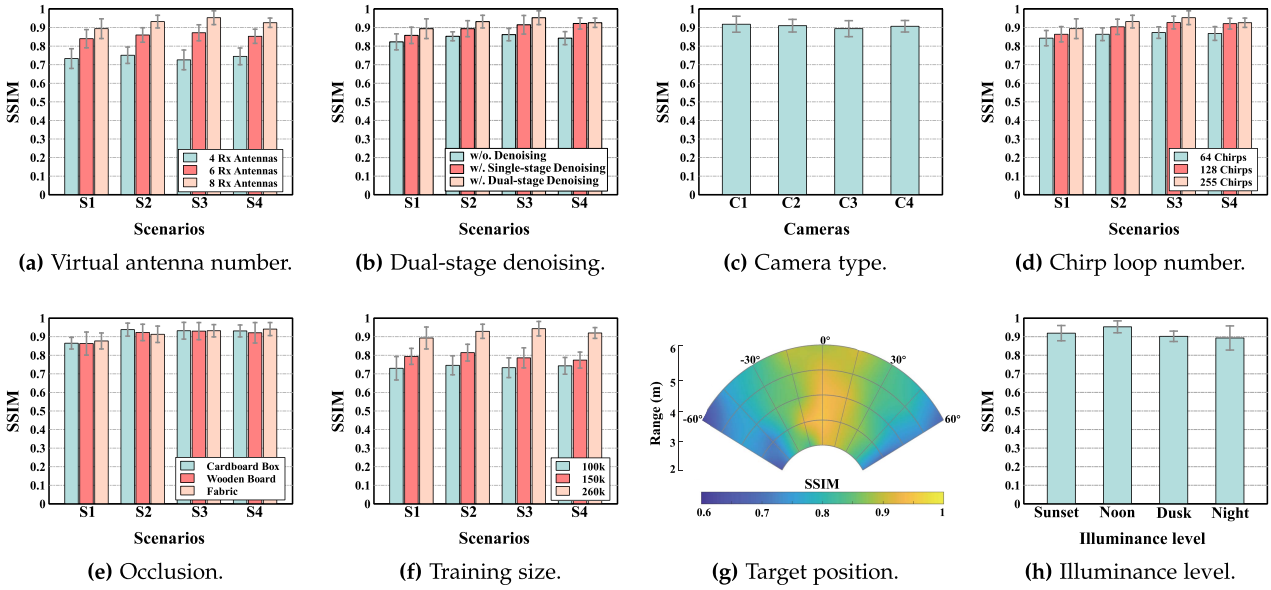
E. Parameter Evaluation

Impact of virtual antenna number: In this experiment, we investigate the impact of varying the number of virtual antennas on M^2 VISION's performance. Specifically, we conduct evaluations using 4, 6, and 8 virtual antennas. As shown in Fig. 11(a), using only four receiving antennas (no virtual antennas) results in an average SSIM of 0.73. Notably, the SSIM values increase to 0.86 and 0.93 when six and eight virtual antennas are employed, respectively. This improvement is due to the increased angle resolution of the mmWave radar. The angle resolution can be expressed as $\theta_{\text{res}} = 2/(N\cos(\theta))$, where N is the number of receiver antennas, and θ is the angle of arrival. The increased virtual antennas number enhances azimuth angle estimation and provides more detailed target profiles.

Impact of dual-stage denoising: We evaluate the effectiveness of our dual-stage denoising algorithm through experiments involving three denoising strategies: no denoising, single-stage denoising, and dual-stage denoising, applied to the mmWave data. As shown in Fig. 11(b), single-stage denoising increases the average SSIM value by 6.15%, while the dual-stage approach further elevates it by 3.23%. This improvement is due to the dual-stage algorithm's ability to effectively eliminate environmental noise initially, and further reduce residual noise impact on angle estimation subsequently, thus attesting to the proposed algorithm's efficacy.

Impact of camera type: We evaluate the effect of surveillance camera types on mm2Video training using four different cameras, as outlined in Table I. We collected thirty minutes of data for each participant in scenario S2. This data was partitioned, with 80% for training and the remaining 20% for testing. As shown in Fig. 11(a), the SSIM obtained from the test data are 0.92, 0.91, 0.89, and 0.91 using the video data collected from camera C1, C2, C3, and C4 for training. Given that the lowest resolution used is 1920×1080 (C3 and C4), which is uniformly downsampled to 256×256 for mm2Video network input, therefore the type of camera has no significant impact on system performance. This underscores M^2 VISION's potential to be widely deployed in future surveillance systems.

Impact of mmWave radar chirp loop: The number of chirp loops, as discussed in Section II-A, corresponds to the number of slow time samples in one radar frame and determines the



(a) Virtual antenna number. **(b)** Dual-stage denoising. **(c)** Camera type. **(d)** Chirp loop number.

(e) Occlusion. **(f)** Training size. **(g)** Target position. **(h)** Illuminance level.

Fig. 11. Parameter evaluation results.

velocity resolution of the mmWave radar. Fig. 11(b) reveals that an increase in chirp loops from 64 to 255 enhances the average SSIM of M^2 VISION from 0.862 to 0.93. The velocity resolution of the mmWave radar can be defined as $v_{\text{res}} = \lambda/(2N_cT_c)$, where λ is the wavelength, N_c is the number of chirp loops, and T_c is the chirp duration. As N_c increases, the radar can sense velocity components with finer granularity, leading to a more detailed motion heatmap. The enhanced granularity allows M^2 VISION to more accurately differentiate between body parts, leading to superior video reconstruction.

Impact of occlusion: To assess the impact of occlusion on the performance of M^2 VISION, we conducted experiments using different occluding materials, namely cardboard box, wooden board, and fabric, to mask the mmWave radar. The network trained with the data collected in Section VI-A is directly utilized for testing with the newly collected occluded data. The evaluation results, presented in Fig. 11(f), demonstrate the average SSIM under the cardboard box, wooden board, and fabric occlusions to be 0.92, 0.91, and 0.92, respectively. The results indicate that M^2 VISION is resilient to occlusions, as it achieves favorable video reconstruction results even when the radar is obstructed by different materials.

Impact of training size: In the experiment, we evaluate the impact of different training set sizes on the final reconstruction results. The training sets were comprised of 100k, 150k, and 260k video frames, respectively. The evaluation results, presented in Fig. 11(f), demonstrate that the mm2Video generative network achieves the best reconstruction performance when trained on a larger dataset with the 260k training set size achieving the highest SSIM of 0.92. Remarkably, even with a smaller training set size of 100k samples, the network exhibits a relatively high SSIM score of 0.74, indicating its strong generalization ability.

Impact of target position: In this experiment, we investigate the impact of different target positions on video reconstruction. Participants are instructed to march in place at various positions within a range of 2 m to 6 m and angles between -60° and 60° relative to the mmWave radar in S2. The collected data are then used to reconstruct the video data. As shown in Fig. 11(h), although the mmWave radar is not significantly affected by radial distance, as the target's radial distance increases, the occupied angle becomes smaller, resulting in a slight decrease in system performance. Furthermore, the radar angle resolution decreases as the angle of arrival increases. Consequently, M^2 VISION's performance is negatively impacted as the angle of the target increases.

Impact of illuminance level: To evaluate the influence of illuminance level on the performance of M^2 VISION, we collected visual and mmWave data at different times of the day in S3: 1) sunset, 2) noon, 3) dusk, and 4) night. Fig. 11(h) displays the mean SSIM values for these times, recorded as 0.94, 0.96, 0.91, and 0.89, respectively. M^2 VISION demonstrates enhanced performance under higher illuminance levels, which may be due to better acquisition to visual information of the target's appearance. However, the minimal variance among these values suggests that M^2 VISION is robust, maintaining consistent performance across a range of illuminance conditions.

Robustness against dynamic interference: We evaluated the robustness of M^2 VISION against multiple dynamic objects in S3. In addition to the main target, we introduced two other people as dynamic interference targets to perform the following actions: a) walk further away from the radar compared to the target, b) walk at the same radial distance as the target, and c) walk radially along the radar. The experimental results in Fig. 12 show that Case a performs best and Case b performs worst. This is because our system utilizes the radar's range-resolving capability described in Section IV-B-1 to effectively eliminates interference from more distant dynamic targets, as shown in the results of Case a. However, if the interference target is close and always within the same range bin as the main target, the algorithm struggles to distinguish them. This makes the

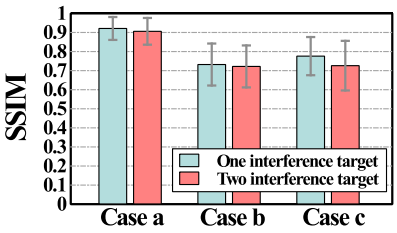


Fig. 12. Robustness against dynamic interference.

TABLE III
CROSS-PERSON STUDY

Scenario	S1	S2	S3	S4
SSIM	0.87	0.91	0.90	0.89

mmWave data becoming ineffective, leading to poor system performance, as shown in Cases b. We will further discuss the scenario involving multiple dynamic targets in Section VIII.

Cross-person study: In order to evaluate the impact of unseen targets, we iteratively selected one person for testing and the remaining six for training. We then averaged the results over seven iterations for each scenario. The evaluation results are shown in Table III. The experimental results indicate that the SSIM decreases by an average of 0.033 when encountering an unknown target. Although this represents a minor performance degradation, the overall impact on our system remains minimal.

VII. RELATED WORKS

A. RF Imaging

The field of RF sensing [29], [30], [31] is undergoing rapid evolution, witnessing the utilization of various RF signals, including Wi-Fi [32], mmWave [33], LoRa [34], and RFID [35], for various sensing tasks. RF imaging boasts the advantage of non-line of sight (NLoS) capabilities, garnering significant attention from researchers in recent years.

Wision [36] first explores the feasibility of achieving computational imaging using Wi-Fi signals, which leverage the designed algorithms to separate the multi-path Wi-Fi reflections from different objects into a coarse-grained image. WiSIA [37] leverages the COTS Wi-Fi devices to simultaneously detect objects and humans, segment their boundaries, and identify them within the image plane. The granularity of Wi-Fi imaging is constrained by the bandwidth limitations of 2.4/5 GHz, prompting researchers to explore mmWave technology for imaging purposes. For instance, Zhu et al. [38] proposed the construction of a large synthetic aperture radar (SAR) by relocating a 60 GHz mmWave device, employing RSS series analysis to profile and image objects. mmEye [39] introduces a super-resolution imaging algorithm that surpasses resolution constraints by jointly utilizing the transmit and receive arrays of mmWave radar to enhance spatial resolution.

As a subdomain of RF imaging, RF-based 3D human mesh recovery has garnered significant interest among researchers, which focuses more on the tracking and reconstruction of the human body. RF-Avatar [40] utilizes Wi-Fi signals to reconstruct a comprehensive 3D human mesh capturing both shape and motion. Wi-Mesh [41] proposes to leverage the 2D angle of arrival (AoA) estimation of the Wi-Fi signal reflections to visualize the shape and deformations of the human body for 3D mesh construction. Furthermore, mmWave signal, with its fine-grained sensing capabilities, has also been employed for 3D human mesh recovery. mmMesh [42] utilizes 3D point cloud data acquired from mmWave radar for real-time 3D human mesh estimation. M ^{4} esh [33] advances a step further by achieving 3D human mesh reconstruction of multiple targets using a single COTS mmWave radar. In addition, m ^{3} Track [43] specializes in using COST mmWave radar to track 3D posture across multiple individuals simultaneously, rather than recovering video frame data. The above mmWave-based sensing works motivate us to assist video recovery through the fine-grained sensing capabilities of mmWave.

B. Video Tampering Detection and Reconstruction

Video tampering detection is essential for the security of video surveillance systems. Traditional methods are typically based on watermarking [5], [44], which can be embedded directly into the video frames or metadata, providing a unique signature that can be used to detect tampering attempts. Meanwhile, the statistical features [7], [45] are often used to detect tampering attempts. Statistical features analysis provides a quantitative framework for identifying anomalies and deviations from expected patterns, which can indicate potential tampering. With the development of deep learning technology, various deep learning networks [9], [10] are designed to carry out video tampering detection. Furthermore, cross-modal techniques have also shown promising results in video tampering detection. For instance, Secure-Pose [11] leverages Wi-Fi signals to detect and localize video forgery attacks in video frames.

Nonetheless, while these solutions can identify or locate fake surveillance footage, they lack the capability to restore the original videos and offer little assistance when surveillance cameras are physically damaged or obstructed. Many works [46], [47] attempt to restore the video source data using meta-information recorded in the header of a file system, which is not possible with meta-information lost. Therefore, attempts [48], [49] have been made to restore the video data using the signature. In addition, cross-modal-based video reconstruction methods have been proposed, which help to recover the video by means of an auxiliary modality signal. For instance, Wi2Vi [12] is the pioneering work to reconstruct video data using Wi-Fi signals, but it generates only grayscale videos with blurry targets. Subsequent work, CSI2Video [13], improved upon this by generating color video, but its efficacy against complex backgrounds remains unverified due to its relativity simple and singular background. Moreover, the granularity of the Wi-Fi signal imposes limitations on the finesse and stability of the video reconstruction. It is noteworthy that the EM Eye [50] eavesdrops on video data via the electromagnetic leakage information from the camera. Although EM Eye can reconstruct high-quality grayscale video, its reliance on the camera's proper functioning renders it ineffective

in instances of physical damage to the camera. Consequently, we propose to reconstruct color surveillance video through the fine-grained sensing capability of mmWave signals and with the powerful generative ability of generative models.

VIII. Discussion

Scenario limitation: Our current study focuses exclusively on the video reconstruction of single-target and fixed-camera perspectives. Addressing multi-target scenarios involves two main challenges: 1) multiple targets segmentation and tracking; and 2) multi-target feature mapping. To tackle the first challenge, enhancing angular resolution through advanced hardware [51], [52] and using RF tracking algorithms [43], [53] can effectively segment and track multiple targets. Regarding the second challenge, we explore using the CustomVideo [54] to create identity-protected videos that capture the unique features of each target. Additionally, to differentiate multiple targets wearing identical masks, we plan to integrate behavioral features such as gait [55], [56] with appearance data, and employing clustering algorithms to distinguish the features of similarly masked individuals. Meanwhile, M^2 Vision focuses solely on surveillance cameras with fixed viewing angles. Consequently, environmental changes are typically minimal within this constraint. In our future plans, we aim to leverage the mmWave radar's capability to sense static environments. By integrating the sensed environmental features into a feature fusion module, we intend to enhance adaptability to dynamic scenarios, such as moving viewpoints or changing scenes.

Elevation angle resolution: As outlined in Section IV-B, the COTS single-chip mmWave radar's limited elevation resolution, attributed to the sparse vertical array of antennas, restricts its ability to accurately reconstruct the vertical profile information of the target. However, this constraint does not preclude the extraction of feature maps that encapsulate reflection data from different parts of the target's body, because the target's reflectance signals are consistently detected by the mmWave radar and represented in the mmWave heatmap. Like previous research [33], [42], [43], our approach leverages a deep learning model to perform further feature extraction from mmWave heatmaps. The deep learning model's powerful learning capabilities partially offset the antenna limitations. Moving forward, we plan to employ more sophisticated hardware, such as cascade radar [51] or 4D imaging radar [52], along with synthetic aperture radar (SAR) technology at the algorithmic level, to enhance the elevation resolution of our system.

Reconstructed video quality: The adopted approach currently entails setting the resolution of the recovered video at 256×256 to alleviate computational and time overhead. However, for subsequent investigations, we contemplate the possibility of augmenting the resolution of the generated video either through network structure modifications or the incorporation of a super-resolution technique [57]. Moreover, our present method relies on frame-by-frame reconstruction, leading to potential discontinuities between individual frames. To address this concern in forthcoming studies, we envision employing video generation models [58] to generate videos directly.

Occlusion limitation: In the current experiments, we verified the recoverability of M^2 Vision in the case of thinner, non-metallic obstacles due to the limited penetration capability of the high-frequency (77-81 GHz) mmWave radar we used. To improve the penetration through thicker obstructions such as walls in future studies, the use of lower frequency radars or ultra-wideband (UWB) radars could be considered. Moreover, integrating mmWave radar with additional sensors, such as acoustic sensors, through data fusion techniques could enhance our system's effectiveness against more substantial obstructions.

IX. Conclusion

In this paper, we propose M^2 Vision, the first mmWave-based surveillance video reconstruction system used to enhance existing video surveillance systems. We propose a series of signal processing algorithms to obtain mmWave heatmaps that respond to the target's profile and motion information. In addition, we design the mm2Video generative network, which efficiently fuses mmWave heatmaps with previously acquired vision data, and recovers the surveillance video based on the fused features through a cGAN-based generator. Extensive evaluations in various scenarios show that M^2 Vision achieves an average SSIM of 0.93, which demonstrates its potential to be widely deployed in future surveillance systems.

References

[1] MarketsandMarkets Research Private Ltd., "Video surveillance market –Analysis and forecast to 2026," 2022. [Online]. Available:https://www.marketsandmarkets.com/Market-Reports/video-surveillance-market-645.html
[2] C. Heffner, "Exploiting surveillance cameras like a Hollywood hacker," Tactical Netw. Solutions, Tech. Rep., 2013. [Online]. Available: https://media.blackhat.com/us-13/US-13-Heffner-Exploiting-Network-Surveillance-Cameras-Like-A-Hollywood-Hacker-WP.pdf
[3] H. Qian, X. Wu, and Y. Xu, *Intelligent Surveillance Systems*. Berlin, Germany: Springer, 2011.
[4] S. Thys, W. Van Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 49–55.
[5] F. Arab, S. M. Abdullah, S. Z. M. Hashim, A. A. Manaf, and M. Zamani, "A robust video watermarking technique for the tamper detection of surveillance systems," *Multimed. Tools. Appl.*, vol. 75, pp. 10855–10885, 2016.
[6] M. Fallahpour, S. Shirmohammadi, M. Semsarzadeh, and J. Zhao, "Tampering detection in compressed digital video using watermarking," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 5, pp. 1057–1072, May 2014.
[7] G. Singh and K. Singh, "Video frame and region duplication forgery detection based on correlation coefficient and coefficient of variation," *Multimed. Tools. Appl.*, vol. 78, pp. 11527–11562, 2019.
[8] M. A. Fayyaz, A. Anjum, S. Ziauddin, A. Khan, and A. Sarfaraz, "An improved surveillance video forgery detection technique using sensor pattern noise and correlation of noise residues," *Multimed. Tools. Appl.*, vol. 79, pp. 5767–5788, 2020.
[9] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2018, pp. 1–7.
[10] Q. Yang, D. Yu, Z. Zhang, Y. Yao, and L. Chen, "Spatiotemporal trident networks: Detection and localization of object removal tampering in video passive forensics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 4131–4144, Oct. 2021.
[11] Y. Huang, X. Li, W. Wang, T. Jiang, and Q. Zhang, "Towards cross-modal forgery detection and localization on live surveillance videos," in *Proc. IEEE Conf. Comput. Commun.*, 2021, pp. 1–10.

[12] M. H. Kefayati, V. Pourahmadi, and H. Aghaeinia, "Wi2Vi: Generating video frames from WiFi CSI samples," *IEEE Sens. J.*, vol. 20, no. 19, pp. 11463–11473, Oct. 2020.
[13] X. Li and R. Younes, "Recovering surveillance video using RF cues," 2022, *arXiv:2212.13340*.
[14] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7122–7131.
[15] M. Wang, F. Qiu, W. Liu, C. Qian, X. Zhou, and L. Ma, "Monocular human pose and shape reconstruction using part differentiable rendering," *Comput. Graph. Forum*, vol. 39, no. 7, 2020, pp. 351–362.
[16] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[19] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.
[20] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[21] X. Ding, Y. Wang, Z. Xu, W. J. Welch, and Z. J. Wang, "CcGAN: Continuous conditional generative adversarial networks for image generation," in *Proc. Int. Conf. Learn. Representations*, 2020, pp. 3012–2021.
[22] J. Gu, Y. Shen, and B. Zhou, "Image processing using multi-code GAN prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3012–3021.
[23] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2849–2857.
[24] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.
[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[27] W. Abdulla, "Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow," 2017. [Online]. Available: https://github.com/matterport/Mask_RCNN
[28] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11 n traces with channel state information," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, pp. 53–53, 2011.
[29] X. Wang, X. Wang, and S. Mao, "RF sensing in the Internet of Things: A general deep learning framework," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 62–67, Sep. 2018.
[30] R. Du et al., "An overview on IEEE 802.11 bf: WLAN sensing," 2023, *arXiv:2310.17661*.
[31] H. Hua, T. X. Han, and J. Xu, "MIMO integrated sensing and communication: CRB-rate tradeoff," *IEEE Trans. Wireless Commun.*, vol. 23, no. 4, pp. 2839–2854, Apr. 2024.
[32] H. Kong, L. Lu, J. Yu, Y. Chen, and F. Tang, "Continuous authentication through finger gesture interaction for smart homes using WiFi," *IEEE Trans. Mobile Comput.*, vol. 20, no. 11, pp. 3148–3162, Nov. 2021.
[33] H. Xue et al., "M^4 esh: MmWave-based 3D human mesh construction for multiple subjects," in *Proc. 20th ACM Conf. Embedded Netw. Sensor Syst.*, 2022, pp. 391–406.
[34] F. Zhang et al., "Exploring LoRa for long-range through-wall sensing," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 4, no. 2, pp. 1–27, 2020.
[35] Y. Chen, J. Yu, Y. Chen, L. Kong, Y. Zhu, and Y.-C. Chen, "RF-Spy: Eavesdropping on online conversations with out-of-vocabulary words by sensing metal coil vibration of headsets leveraging RFID," in *Proc. 22nd Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2024, pp. 169–182.
[36] D. Huang, R. Nandakumar, and S. Gollakota, "Feasibility and limits of Wi-Fi imaging," in *Proc. 20th ACM Conf. Embedded Netw. Sensor Syst.*, 2014, pp. 266–279.
[37] C. Li, Z. Liu, Y. Yao, Z. Cao, M. Zhang, and Y. Liu, "Wi-Fi see it all: Generative adversarial network-augmented versatile Wi-Fi imaging," in *Proc. 20th ACM Conf. Embedded Netw. Sensor Syst.*, 2020, pp. 436–448.
[38] Y. Zhu, Y. Zhu, B. Y. Zhao, and H. Zheng, "Reusing 60 GHz radios for mobile radar imaging," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 103–116.
[39] F. Zhang, C. Wu, B. Wang, and K. R. Liu, "mmEye: Super-resolution millimeter wave imaging," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6995–7008, Apr. 2021.
[40] M. Zhao et al., "Through-wall human mesh recovery using radio signals," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 10113–10122.
[41] Y. Wang, Y. Ren, Y. Chen, and J. Yang, "Wi-mesh: A WiFi vision-based approach for 3D human mesh construction," in *Proc. 20th ACM Conf. Embedded Netw. Sensor Syst.*, 2022, pp. 362–376.
[42] H. Xue et al., "mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave," in *Proc. 22nd Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2021, pp. 269–282.
[43] H. Kong et al., "M^3 Track: mmWave-based multi-user 3D posture tracking," in *Proc. 22nd Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2022, pp. 491–503.
[44] S. Chen and H. Leung, "Chaotic watermarking for video authentication in surveillance applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 5, pp. 704–709, May 2008.
[45] S. Chen, S. Tan, B. Li, and J. Huang, "Automatic detection of object-based forgery in advanced video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 11, pp. 2138–2151, Nov. 2016.
[46] L. Aronson and J. Van DenBos, "Towards an engineering approach to file carver construction," in *Proc. IEEE 35th Annu. Comput. Softw. Appl. Conf. Workshops*, 2011, pp. 368–373.
[47] B. Carrier, *File System Forensic Analysis*. Boston, MA, USA: Addison-Wesley, 2005.
[48] S. L. Garfinkel, "Carving contiguous and fragmented files with fast object validation," *Digit. Invest.*, vol. 4, pp. 2–12, 2007.
[49] G.-H. Na, K.-S. Shim, K.-W. Moon, S. G. Kong, E.-S. Kim, and J. Lee, "Frame-based recovery of corrupted video files using video codec specifications," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 517–526, Feb. 2014.
[50] Y. Long et al., "EM Eye: Characterizing electromagnetic side-channel eavesdropping on embedded cameras," in *Proc. Annu. Netw. Distrib. Syst. Secur. Symp.*, 2024. [Online]. Available: https://www.ndss-symposium.org/ndss-paper/em-eye-characterizing-electromagnetic-side-channel-eavesdropping-on-embedded-cameras/
[51] Texas Instruments, MMWCAS-RF-EVM: mmWave cascade imaging radar RF evaluation module. [Online]. Available: https://www.ti.com/tool/MMWCAS-RF-EVM
[52] Vayyar, "Vayyar imaging," [Online]. Available: https://vayyar.com
[53] C. Wu, F. Zhang, B. Wang, and K. R. Liu, "mmTrack: Passive multi-person localization using commodity millimeter wave radio," in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 2400–2409.
[54] Z. Wang et al., "Customvideo: Customizing text-to-video generation with multiple subjects," 2024, *arXiv:2401.09962*.
[55] H. Yang et al., "XGait: Cross-modal translation via deep generative sensing for RF-based gait recognition," in *Proc. 20th ACM Conf. Embedded Netw. Sensor Syst.*, 2023, pp. 43–55.
[56] X. Yang, J. Liu, Y. Chen, X. Guo, and Y. Xie, "MU-ID: Multi-user identification through gaits using millimeter wave radios," in *Proc. IEEE Conf. Comput. Commun.*, 2020, pp. 2589–2598.
[57] F. Kong et al., "Residual local feature network for efficient super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 766–776.
[58] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "MCVD-masked conditional video diffusion for prediction, generation, and interpolation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 23371–23385.

Mingda Han received the bachelor's degree from the School of Information Science and Engineering, Shandong Normal University, in 2016. He is currently working toward the PhD degree with the School of Computer Science and Technology, Shandong University under the supervision of Prof. Pengfei Hu. His research interests include smart sensing and AIoT.

Huanqi Yang received the bachelor's degree from the University of Electronic Science and Technology of China. He is currently working toward the PhD degree with the Department of Computer Science, City University of Hong Kong. He is supervised by Dr. Weitao Xu. His research interests include lay in smart sensing, IoT security, IoT+AI, and wireless network.

Mingda Jia received the bachelor's degree from the School of Artificial Intelligence of Xi'an Jiaotong University, in 2024. He is currently working toward the prospective master degree with the Institute of Automation, Chinese Academy of Sciences, under the supervision of Weiliang Meng, an associate researcher. His current research interests include focuses on multimodal learning and 3D object detection in computer vision.

Weitao Xu (Member, IEEE) received the PhD degree from the University of Queensland, in 2017 (advised by Prof. Neil Bergmann and Dr. Wen Hu). He is an assistant professor with the Department of Computer Science, City University of Hong Kong. Before that, he was a postdoctoral research associate with the School of Computer Science and Engineering (CSE), UNSW from 2017 to 2019. His research interests include mobile computing, sensor network, and IoT.

Yanni Yang received the BE and MSc degree from the Ocean University of China in Qingdao, in 2014 and 2017, respectively, and the PhD degree in computer science from the Hong Kong Polytechnic University, in 2021. She is currently an assistant professor with the School of Computing Science and Technology at Shandong University. She visited the Media Lab with MIT in 2019 as a visiting student. Her research interests include wireless human sensing, pervasive and mobile computing, and Internet of Things. She has published more than 20 papers in top academic conferences and journals.

Zhijian Huang received the PhD degree from the School of Computer, National University of Defense Technology, Changsha, China, in 2018. He is currently an assistant professor with the National Key Laboratory of Science and Technology on Information System Security, Beijing, China. His research interests include network and information security, and software testing.

Jun Luo (Fellow, IEEE) received the BS and MS degrees in electrical engineering from Tsinghua University, China, and the PhD degree in computer science from EPFL (Swiss Federal Institute of Technology in Lausanne), Lausanne, Switzerland. From 2006 to 2008, he has worked as a postdoctoral research fellow with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada. In 2008, he joined the faculty of the School Of Computer Science and Engineering, Nanyang Technological University in Singapore, where he is currently an associate professor. His research interests include mobile and pervasive computing, wireless networking, machine learning and computer vision, applied operations research, as well as security. More information can be found at http://www.ntu.edu.sg/home/junluo.

Xiuzhen Cheng (Fellow, IEEE) received the MS and PhD degrees in computer science from the University of Minnesota – Twin Cities, in 2000 and 2002, respectively. She was a faculty member with the Department of Computer Science, The George Washington University, from 2002 to 2020. Currently, she is a professor of computer science with Shandong University, Qingdao, China. Her research interests include focuses on blockchain computing, IoT security, and privacy-aware computing.

Pengfei Hu received the PhD degree in computer science from UC Davis. He is a professor with the School of Computer Science and Technology, Shandong University. His research interests include the areas of cyber security, data privacy, and mobile computing. He has published more than 40 papers in premier conferences and journals on these topics. He served as TPC for numerous prestigious conferences, and associate editors for IEEE TWC and IEEE IoT J. He is the recipient of 2022 ACM SIGBED China Rising Star Award.