# mmHand: Toward Pixel-Level-Accuracy Hand Localization Using a Single Commodity mmWave Device

Xiaoyu Zhang, Zhengxiong Li, *Member, IEEE*, Chenhan Xu, Luchuan Song, Huining Li, Hongfei Xue, Yingxiao Wu, and Wenyao Xu, *Senior Member, IEEE*

***Abstract*—The hand localization problem has been a long-standing focus due to its many applications. The task involves modeling the hand as a singular point and determining its position within a defined coordinate system. However, due to data modality limitations, existing hand localization technologies face several challenges. For example, vision-based localization raises privacy concerns, while wearable-based methods compromise user comfort. In this article, we introduce mmHand, a new device-free, privacy-preserving dynamic hand localization system with pixel-level accuracy, using a single commodity mmWave device. We first propose a mmImage generation tool to fully extract spatial information from raw mmWave data and introduce a novel 2-D image-format representation of mmWave data. Next, we design a framework that provides a new quality evaluation method and pixel space labeling for the mmWave data. Finally, we present a cross-modality spatial feature-enhanced model with high spatial feature extraction capabilities, which can accurately localize hand positions at the pixel level in the mmWave radar U-V pixel coordinate system. We evaluate the system with experiments on 12 subjects in three scenarios, and the results across four metrics demonstrate the effectiveness of our hand localization system.**

***Index Terms*—mmWave sensor, hand localization, cross-modality**

## I. Introduction

HAND localization has been widely applied in various fields, including AR/VR, robotics, and gaming. Many companies and organizations are actively developing accurate hand localization methods based on two primary types of human-computer interaction. The first is contact-based hand localization, such as Meta Quest [1], which typically requires a touch controller to determine the hand's position. However, contact devices are often seen as burdensome to users, compromising the overall experience. The second type is noncontact-based solutions, such as Ultraleap [2] and MediaPipe [3], which use cameras to provide high-resolution tracking in most environments. However, camera footage in the hand localization process raises privacy concerns. In contrast, RF sensors offer a privacy-preserving alternative as nonvisual devices. Despite this advantage, most RF sensors face challenges in achieving precise hand localization due to their relatively low resolution.

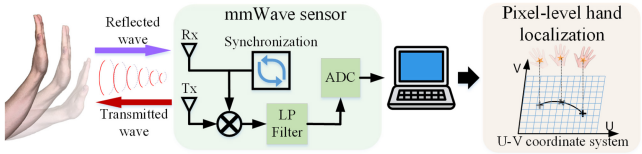



Fig. 1. mmWave sensor receives signals reflected from the hand and processes them for dynamic hand localization.

Given the current challenges, mmWave radar, a type of RF sensor, offers several advantages for hand localization products: (1) *high resolution:* operating in the GHz band, it provides millimeter-level accuracy; (2) *fast response:* it can respond as quickly as a 500 FPS camera, allowing for the localization of fast-moving hands; (3) *privacy-preserving:* it doesn't rely on vision-based signals; (4) *integratable:* its compact size enables easy integration into mobile devices for various applications; (5) *device-free:* utilizing frequency modulated continuous wave (FMCW) technology, it detects and localizes hands without requiring physical contact with the device.

There are several approaches to hand localization using mmWave sensors [4], [5]. However, existing methods struggle to effectively extract the full information from mmWave data and still rely on more complex and expensive mmWave devices to localize the hand.

In this article, we propose mmHand to achieve pixel-level accuracy in hand localization. As shown in Fig. 1, the mmHand system utilizes the mmWave sensor to emit signals

Received 1 October 2024; revised 3 January 2025 and 15 February 2025; accepted 19 February 2025. Date of publication 27 February 2025; date of current version 9 June 2025. This work was supported in part by the US NSF Award under Grant 2426470, and in part by the 2022 Meta Research Award. *(Corresponding author: Wenyao Xu.)*

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

Xiaoyu Zhang and Wenyao Xu are with the Department of Computer Science and Engineering, State University of New York at Buffalo, Amherst, NY 14068 USA (e-mail: zhang376@buffalo.edu; wenyaoxu@buffalo.edu).

Zhengxiong Li is with the Department of Computer Science and Engineering, University of Colorado at Denver, Denver, CO 80204 USA (e-mail: zhengxiong.li@ucdenver.edu).

Chenhan Xu and Huining Li are with the Department of Computer Science, North Carolina State University, Raleigh, NC 27695 USA (e-mail: cxu34@ncsu.edu; huiningl@buffalo.edu).

Luchuan Song is with the Department of Computer Science, University of Rochester, Rochester, NY 14627 USA (e-mail: lsong11@ur.rochester.edu).

Hongfei Xue is with the Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC 28223 USA (e-mail: hongfei.xue@charlotte.edu).

Yingxiao Wu is with the College of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, Zhejiang, China (e-mail: wuyingxiao@hdu.edu.cn).

Digital Object Identifier 10.1109/JIOT.2025.3546560

and capture their reflections from the hand, which are then processed for localization. The development of our system presented three key challenges: 1) the raw mmwave data are time-sequence energy signals, which are fundamentally different from image data, making it difficult to associate the time-series signal with pixels. To address this, we develop a mmImage generation tool to convert raw mmWave data into a new image format, correlating it with pixels; 2) existing methods cannot label the image-formatted mmWave data for hand localization model training, so we design a cross-modality-based labeling method to efficiently generate accurate pixel-space labels; 3) current hand localization models lack sufficient ability to extract high-quality spatial features from the image-formatted mmWave data. To enhance feature extraction, we connect the mmWave data with higher-resolution depth images, leveraging the spatial features of the depth images to guide mmWave hand localization model training and improve feature extraction. Extensive experiments demonstrate that the mmHand system performs well in hand localization, and the enhanced model effectively extracts high-quality spatial features.

In conclusion, our contributions are listed as follows.

1) We present a tool to generate a new image-format representation of mmWave data, called mmImage. This tool fully extracts the spatial information from the mmWave data and associates the time-series data with the pixel concept.
2) We design a new evaluation method for mmWave data quality that focuses on spatial information and effectively identifies low-quality data with indistinct spatial details caused by interference.
3) Leveraging the data format similarity between the mmImage and synchronized depth images, we propose a novel mmWave data labeling method that efficiently provides accurate pixel-space labels for training image-based task models.
4) We propose a cross-modality spatial-feature-enhanced hand detection model based on the mmImage. The model utilizes the spatial feature similarity between the mmImage and depth image to build an efficient cross-modality connection, enhancing the spatial feature extraction of the mmImage under the guidance of the higher-resolution depth image. Additionally, our spatial-feature-enhanced model can facilitate effective interaction between all image-format data modalities with similar spatial features.
5) We design an end-to-end hand localization system using a single commodity mmWave device. It requires no specific sensor placement and can localize hand positions from single-frame mmWave data. Extensive experiments involving 12 subjects across three scenarios, using four metrics, demonstrate the effectiveness of our proposed hand localization system.

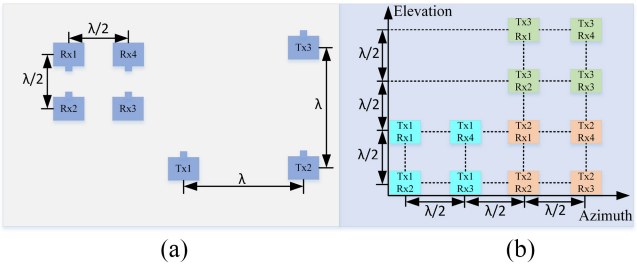
## II. Background

### A. Depth Image Modality

The depth camera is a widely used sensor capable of accurately measuring the distance between an object and the camera. Currently, three depth sensing technologies are commonly employed in the depth cameras, which are outlined below.

1) *Structured light (e.g., Kinect V1 and RealSense):* This type of depth camera emits invisible infrared lasers at a specific wavelength and captures depth information by analyzing distortions in the reflected encoded pattern.
2) *Stereo vision (e.g., Stereo IR 170 and ZED 2):* This depth camera captures two images of an object from different viewpoints and calculates depth by measuring the disparity between corresponding points in the two images.
3) *Time of Flight (e.g., Kinect V2 and SoftKinect):* This type of depth camera continuously emits laser pulses toward an object and measures depth by calculating the round-trip time of the reflected pulses.

Typically, the data format of a depth image is $(\mathcal{W}, \mathcal{H})$, where $\mathcal{W}$ and $\mathcal{H}$ represent the width and height of the image, respectively.




(a) (b)

Fig. 2. Example of a mmWave sensor with three transmitters (Tx) and four receivers (Rx). (a) shows the antenna placement and spacing. (b) shows the corresponding virtual.

### B. mmWave Data Modality

Fig. 2 illustrates an example of a mmWave sensor with three transmitting antennas (Tx) and four receiving antennas (Rx). During each frame period, the mmWave sensor samples $C$ chirps of data using multiple-input multiple-output (MIMO) technology, with each Tx sequentially emitting signals that are received by all Rx in one chirp. The mmWave sensor uniformly samples $N$ signal points from each received Tx signal by each Rx. Consequently, the format for one frame of mmWave data is represented as $(\mathcal{T}, \mathcal{R}, \mathcal{C}, \mathcal{N})$, where $\mathcal{T}$, $\mathcal{R}$, $\mathcal{C}$, and $\mathcal{N}$ denote the number of Tx, Rx, chirps, and samples, respectively.

Due to its high sampling rate, the mmWave sensor achieves high-frequency resolution. Using FMCW technology, differences in frequency are converted into distance information, so high-frequency resolution contributes to the sensor's range resolution. Furthermore, the presence of the 2-D virtual antenna array shown in Fig. 2 allows the mmWave radar to provide a 2-D spatial view for detecting object distances, akin to a depth camera. This capability makes it feasible to establish a cross-modality connection between mmWave data and depth images, as both modalities share spatial information characteristics. Details of the interaction between these two modalities are presented in Section V.

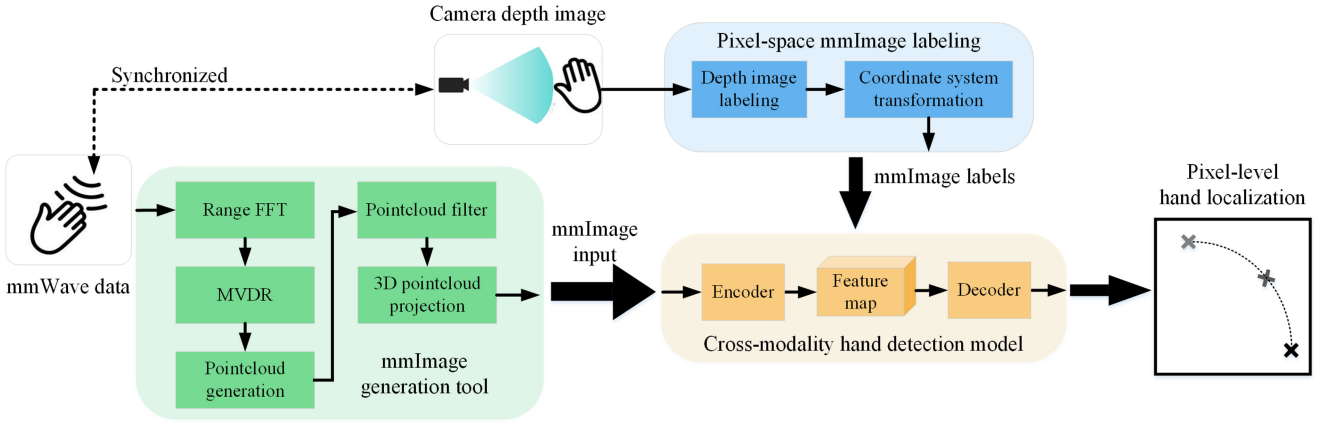



Fig. 3. mmHand system framework.

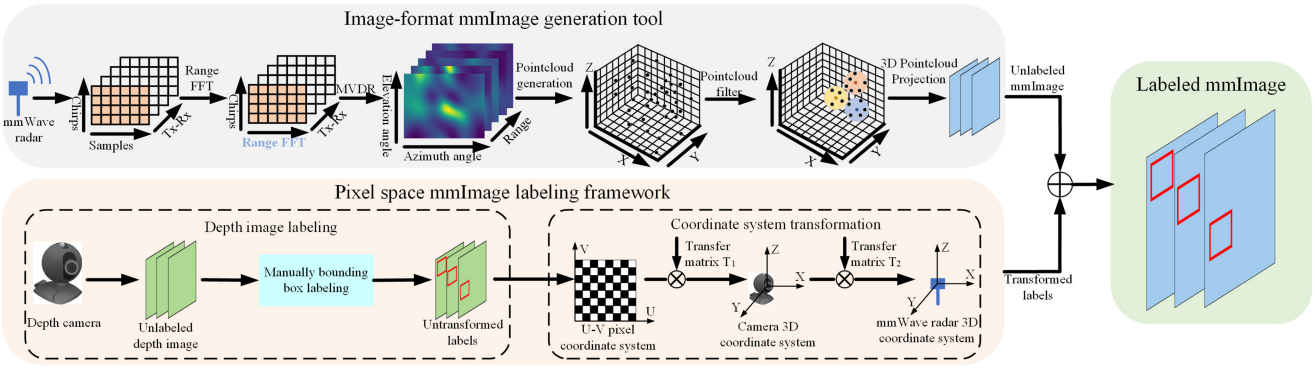



Fig. 4. mmImage generation tool for image-format conversion and the pixel space mmImage labeling framework.

## III. mmHand System Framework

Fig. 3 provides an overview of the mmWave-based hand localization system with pixel-level accuracy. The mmHand system consists of four components: the mmImage generation tool, pixel-space mmImage labeling and cross-modality hand detection model, and pixel-level hand localization. First, we introduce mmImage, a novel representation of mmWave data that fully captures spatial information and converts it into an image format. Second, we leverage synchronized depth camera images to generate accurate pixel-space hand location labels for the mmImage. Third, we establish a cross-modality connection between mmImages and depth images, enhancing spatial feature extraction for improved hand detection performance. Finally, we develop a context-aware algorithm that mitigates inaccurate predictions, ensuring precise and robust dynamic hand localization.

## IV. mmImage Generation Tool

In this section, we introduce the mmImage generation tool, which converts a frame of raw time-series mmWave data $(\mathcal{T}, \mathcal{R}, \mathcal{C}, \mathcal{N})$ into a new image-format representation called mmImage $(\mathcal{W}, \mathcal{H})$. As illustrated in Fig. 4, the mmImage generation process consists of five steps: 1) range FFT; 2) MVDR data processing; 3) point cloud generation; 4) point cloud filtering; and 5) 3-D point cloud projection. We will describe each of these steps in detail.

### A. Range FFT

When the mmWave sensor transmits an FMCW signal, it propagates to multiple objects at different positions, resulting in varying time delays in the received reflections. The range FFT processes these time delays by converting them into frequency peaks, effectively revealing the distance differences between objects. After applying the range FFT, the raw mmWave data $(\mathcal{T}, \mathcal{R}, \mathcal{C}, \mathcal{N})$ is transformed into the range bin format $(\mathcal{T}, \mathcal{R}, \mathcal{C}, \mathcal{D})$, where $\mathcal{D}$ retains the same value as $\mathcal{N}$ and represents the range dimension.

### B. Minimum Variance Distortionless Response (MVDR)

As shown in Fig. 2(a), the mmWave sensor is equipped with multiple transmitting and receiving antennas. Using MIMO (multiple-input, multiple-output) technology, different transmitting antennas operate in separate time slots, allowing the receiving antennas to form a virtual antenna array, as illustrated in Fig. 2(b). The 2-D spatial structure of this antenna array enables the mmWave sensor to detect objects in both the azimuth and elevation dimensions.

MVDR, also known as Capon Beamforming, is a valuable tool in mmWave data processing. After obtaining the range bins from the Range FFT step, we apply MVDR to combine the azimuth and elevation information, generating range-azimuth-elevation data $(\mathcal{D}, \Theta, \Phi)$ in the spherical coordinate system, where $\Theta$ and $\Phi$ represent the azimuth and elevation angles, respectively.

TABLE I
OVERVIEW OF MMIMAGE GENERATION

| Data type | Next step | Data format |
|---|---|---|
| Raw mmWave data | Range FFT | $(\mathcal{T}, \mathcal{R}, \mathcal{C}, \mathcal{N})$ |
| Range bin | MVDR | $(\mathcal{T}, \mathcal{R}, \mathcal{C}, \mathcal{D})$ |
| Range-azimuth-elevation data | Pointcloud generation | $(\mathcal{D}, \Theta, \Phi)$ |
| Pointcloud | Pointcloud filter | $(\mathcal{D} \cdot \Theta \cdot \Phi, (x, y, z, e))$ |
| Filtered pointcloud | 3D pointcloud projection | $(\mathcal{V} \cdot \mathcal{S}, (x, y, z, e))$ |
| mmImage | \ | $(\mathcal{W}, \mathcal{H})$ |

MVDR is a signal processing technique that enhances signals from the desired direction while suppressing interference from others. In 3-D point cloud generation, MVDR enables precise beamforming by dynamically adjusting the antenna array to focus on specific directions. By capturing signals from multiple angles, the system accumulates a dense set of data points, resulting in a more detailed and refined 3-D point cloud. This enhanced data density allows the MVDR beamformer to generate a comprehensive and high-resolution representation of hand regions and contours.

### C. Pointcloud Generation

Compared to the spherical coordinate system, the 3-D Cartesian coordinate system offers better spatial representation and can more directly express an object's spatial contour and precise location. Additionally, the 3-D Cartesian coordinate system has the advantage of being more easily associated with the U-V pixel coordinate system, facilitating pixel-level hand localization. Therefore, we convert the $\mathcal{D}\cdot\Theta\cdot\Phi$ points from the range-azimuth-elevation data $(\mathcal{D}, \Theta, \Phi)$ into the pointcloud $\mathcal{P} = (\mathcal{D} \cdot \Theta \cdot \Phi, (x, y, z, e))$ using (1). In this representation, the second dimension consists of the point's 3-D coordinates $(x, y, z)$ in the 3-D Cartesian coordinate system and the corresponding energy $e$, with the origin of the system located at the center of the mmWave sensor antenna array. However, many of the points are not related to the object, so in the next step, we apply a point cloud filter to refine the generated point cloud and retain only the object-related points

$$\begin{cases} x = r\cos\phi\cos\theta \\ y = r\cos\phi\sin\theta \\ z = r\sin\phi \end{cases} \quad r \in \mathcal{D}, \theta \in \Theta, \phi \in < \Phi\ . \tag{1}$$

### D. Pointcloud Filter

Due to signal attenuation, the received signal loses energy as it propagates over longer distances. In hand localization, since the hand is typically positioned close to the mmWave sensor, points corresponding to the hand generally exhibit higher energy. However, due to the resolution limitations of the sensor, high-energy points from different objects, such as the human trunk and hand–may become indistinguishably mixed, leading to challenges in accurate segmentation and localization.

To distinguish points from different objects, we use a clustering algorithm to filter out points with high discrimination. As outlined in Algorithm 1, we first sort the unprocessed pointcloud $\mathcal{P}$ in descending order based on the energy $e$. For each cycle $0 < i < \mathcal{V}$, the point with the highest energy is set as the cluster center $p_i$. In steps 4-6, we find the $\mathcal{S}$ nearest points to the cluster center $p_i$ (including $p_i$ itself), and these $\mathcal{S}$ points form a cluster with $p_i$ as the center. In steps 7-8, we add the $\mathcal{S}$ points to the processed point cloud set $\mathcal{G}$ and remove them from the unprocessed point cloud set $\mathcal{P}$ to ensure they are not considered in subsequent clustering cycles. After the clustering algorithm, we obtain a purified pointcloud $\mathcal{G} = (\mathcal{V} \cdot \mathcal{S}, (x, y, z, e))$ with $\mathcal{V} \cdot \mathcal{S}$ points.

**Algorithm 1** Pointcloud Clustering Algorithm

**Input:** unprocessed pointcloud set contains the 3D coordinate and energy value $\mathcal{P} = (\mathcal{D}\cdot\Theta\cdot\Phi, (x, y, z, e))$, cluster center number $\mathcal{V}$, cluster size $\mathcal{S}$
**Output:** processed pointcloud set $\mathcal{G}$
1: sort $\mathcal{P}$ in descending order based on the energy $e$
2: **for** $i = 1, ..., \mathcal{V}$ **do**
3: choose the first point $p_i$ in $\mathcal{P}$
4: calculate the distance of all the points to $p_i$ based on the 3D coordinate $(x, y, z)$ and get a distance set $D_i$
5: sort $D_i$ in ascending order
6: $D_i \longleftarrow D_i(1{:}\mathcal{S})$
7: $\mathcal{G} \longleftarrow \mathcal{G} \cup D_i$
8: remove $D_i$ from $\mathcal{P}$
9: **end for**
10: **return** $\mathcal{G}$

### E. 3-D Pointcloud Projection

As discussed in Section II-B, due to the spatial information similarity between mmWave data and depth images, we aim to establish a cross-modality connection between these two data modalities to improve hand localization. Since the 3-D point cloud differs from the 2-D depth image in data format, it is important to minimize the modality gap for an effective cross-modality connection. As shown in Fig. 2, the mmWave sensor's virtual antenna array provides a 2-D view on the antenna array plane, with the vertical axis representing the depth value. Following the data generation method used by depth cameras, we project the 3-D pointcloud onto the antenna array plane, creating a 2-D image-format representation of the mmWave data in the sensor's coordinate system. Finally, we discretize the data points in the image and crop it to a fixed size $(\mathcal{W}, \mathcal{H})$, which we refer to as mmImage.

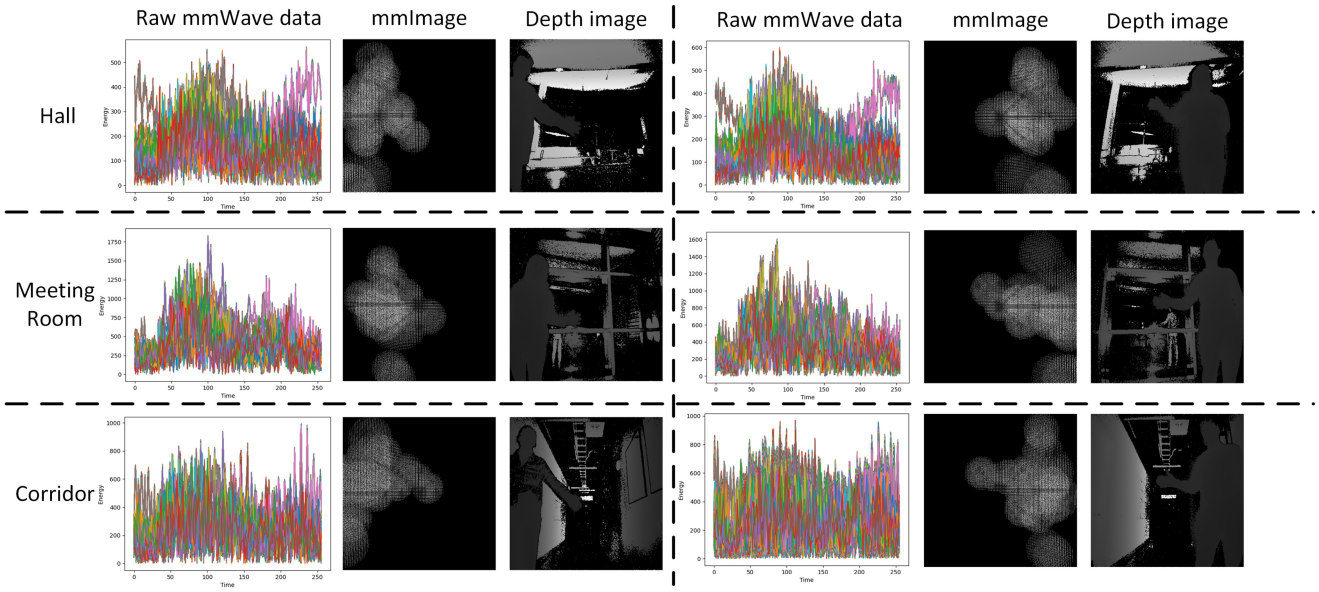



Fig. 5. Examples of the raw mmWave data, generated mmImage and synchronized depth image in three scenarios (Hall, meeting room and corridor).

TABLE II
SYMBOL DEFINITION AND DESCRIPTION

| Symbol | Definition description |
|---|---|
| $\mathcal{T}$ | The number of the transmitting antennas |
| $\mathcal{R}$ | The number of the receiving antennas |
| $\mathcal{C}$ | The number of chirps per frame |
| $\mathcal{N}$ | The number of samples per chirp |
| $\mathcal{D}$ | The number of ranges |
| $\Theta$ | The number of azimuth angles |
| $\Phi$ | The number of elevation angles |
| $\mathcal{V}$ | The number of clusters per frame |
| $\mathcal{S}$ | The number of points per cluster |
| $\mathcal{W}$ | The width of mmImage |
| $\mathcal{H}$ | The height of mmImage |

Table I details the steps involved in mmImage generation along with the corresponding data format at each stage, while Table II defines the symbols used in Table I. Fig. 5 presents a comparative analysis of the raw mmWave data, the generated mmImage, and the corresponding depth image across three distinct environments: a hall, a meeting room, and a corridor. Each set of images corresponds to a different subject. As illustrated in Fig. 5, our mmImage generation tool effectively extracts spatial information from raw mmWave data and maps it into pixel space, producing an mmImage that closely aligns with the depth image captured by the depth camera.

## V. MMIMAGE-BASED HAND LOCALIZATION

In this section, we introduce an mmImage-based hand localization system that fully leverages the spatial information similarity between the mmImage and the synchronized depth image. This system enables a new quality evaluation method and a 2-D pixel space labeling approach for time-series mmWave data, delivering high-accuracy, pixel-level hand localization.

### A. mmWave Data Quality Evaluation

Due to the sensitivity of mmWave signals to environmental interference and internal circuit deviations, the sampled raw mmWave data can vary significantly, leading to differences in the quality of the generated mmImages. To effectively evaluate the mmWave data and filter out poor-quality data, we propose a novel mmWave data quality evaluation method based on the structural similarity index measure (SSIM)

$$\mathrm{SSIM}(x,y)=\frac{\left(2\mu_x\mu_y+(k_1L)^2\right)\left(2\sigma_{xy}+(k_2L)^2\right))}{\left(\mu_x^2+\mu_y^2+(k_1L)^2\right)\left(\sigma_x^2+\sigma_y^2+(k_2L)^2\right)}. \quad (2)$$

The details of the SSIM calculation are based on (2), where $x$ and $y$ are the two images being compared to calculate the SSIM value. $\mu_x$ and $\mu_y$ represent the pixel sample means of the two images, $\sigma_x^2$ and $\sigma_y^2$ are the variances, and $\sigma_{xy}$ is the covariance. Additionally, $L$ represents the dynamic range of pixel values, and $k_1$ and $k_2$ are two constants, typically set to $k_1 = 0.01$ and $k_2 = 0.03$ by default. In situations where mmWave data are collected under poor conditions with significant interference, object signals may be obscured by environmental noise. In these cases, the mmImage generation tool may struggle to effectively extract spatial information from the noisy mmWave data, resulting in a low similarity between the generated mmImage and the synchronized depth image. This is captured by the SSIM metric. We compute the SSIM value for all mmImages against their corresponding synchronized depth images and classify mmImages with SSIM values above a predefined threshold as high-quality data.

### B. Pixel Space mmImage Labeling

Due to the inherent representation of mmWave data as a time-series energy signal, the signal at each time point is a mixture of reflections from both the object and the environment, making it impossible to label the object in the raw mmWave data. As shown in Fig. 4, we propose a novel

pixel-space mmImage labeling method that quickly provides accurate object labels for the mmWave data. This method can be broadly applied to image-based tasks, such as object detection and localization. The detailed labeling steps are as follows.

1) First, the mmWave sensor and depth camera are positioned at a fixed relative distance to collect data.
2) Second, the mmWave data and depth images are synchronized frame by frame using the timestamp.
3) Third, a pretrained object detection model is used to detect the object in pixel space from the collected depth images. The pretrained model can either be obtained by training on a manually labeled depth image dataset or by using a trained model from existing object detection works.
4) Fourth, since the detected labels are in the depth camera's U-V pixel coordinate system and the mmWave data are in the mmWave sensor's coordinate system, the detected labels must be transformed by multiplying them with two transfer matrices ($T_1$ and $T_2$) to align the coordinate systems.
5) Finally, each frame of the transformed label is used as the object label for the synchronized mmImage generated from the raw mmWave data.

Since the new mmWave data labeling method is based on the synchronized depth image, it is independent of different mmWave data processing methods. The primary factor affecting labeling accuracy is the synchronization deviation. The transfer matrix $T_1$, which converts the depth camera's U-V pixel coordinate system to its 3-D coordinate system, depends on the camera's internal fixed parameters. Similarly, the transfer matrix $T_2$, which converts from the depth camera's 3-D coordinate system to the mmWave sensor's 3-D coordinate system, depends on the relative distance between the two sensors. In the experiment, we maintain a fixed relative distance by placing the two sensors in specific locations on a wooden board. As a result, both transfer matrices are linear transformations with fixed parameters. We assume that these linear transformations can be learned by the deep learning model during training, enabling us to simplify the use of untransformed labels in the experimental evaluation.

### C. Hand Detection Module

After obtaining the labeled mmImage data, a deep learning model is required to detect the hand from the mmImage. Due to its fast response and strong performance in object detection [6], we use the Yolov3 model as our hand detection model structure. However, despite the mmImage generation tool effectively extracting spatial information from the raw mmWave data, the inherent lower resolution of the mmWave radar compared to cameras results in the mmImage having lower data quality than a depth image. This limitation hinders the mmImage-based model's ability to extract effective spatial features for hand detection. As shown in Fig. 6, we propose a cross-modality spatial-feature enhanced model, which establishes a guidance interaction from the depth-image-based model to the mmImage-based model. The depth-image-based model, a pretrained Yolov3 model with high hand detection accuracy, includes a feature encoder to extract effective spatial features from the depth image. By introducing a cross-modality loss function between the feature maps of the mmImage-based encoder and the depth-image-based encoder, the mmImage-based model is guided to extract more effective spatial features from the mmImage, leading to higher accuracy in hand detection.

*1) Cross-Modality Loss:* To enhance the feature extraction capability of the mmImage-based encoder, it is reasonable to constrain it to extract feature maps of similar quality to those produced by the depth-image-based encoder, considering their spatial information similarity. Therefore, we design a cross-modality loss function based on Cosine Similarity to guide the training of the mmImage-based hand detection model.

After feeding the mmImage and synchronized depth image into the mmImage-based encoder and the pretrained depth-image-based encoder, both encoders extract feature maps $f^j = R^{K\cdot W\cdot H}, j = 1, 2$, where $K$, $W$, and $H$ represent the number of channels, width, and height of the feature map, and $j = 1$ and $j = 2$ correspond to the feature maps of the mmImage-based encoder and the depth-image-based encoder, respectively. Next, we reshape the feature map $f^j$ into sets of row vectors $\mathcal{F}_r^j = (\overrightarrow{v_1^j}, \overrightarrow{v_2^j}, \ldots, \overrightarrow{v_{K\cdot H}^j})$ and column vectors $\mathcal{F}_c^j = (\overrightarrow{v_1^j}, \overrightarrow{v_2^j}, \ldots, \overrightarrow{v_{K\cdot W}^j})$ based on the width dimension $W$ and height dimension $H$, respectively. We then calculate the overall feature map Cosine Similarity using (3) to measure the spatial feature similarity between the two feature maps

$$\begin{aligned}
\text{similarity} &= \frac{\cos\left(\mathcal{F}_r^{j=1}, \mathcal{F}_r^{j=2}\right) + \cos\left(\mathcal{F}_c^{j=1}, \mathcal{F}_c^{j=2}\right)}{2} \\
\cos\left(\mathcal{F}_r^{j=1}, \mathcal{F}_r^{j=2}\right) &= \frac{\sum_{x=1}^{K\cdot H} \frac{\overrightarrow{v_x^{j=1}} \cdot \overrightarrow{v_x^{j=2}}}{\left\|\overrightarrow{v_x^{j=1}}\right\|_2 \left\|\overrightarrow{v_x^{j=2}}\right\|_2}}{K \cdot H} \\
\cos\left(\mathcal{F}_c^{j=1}, \mathcal{F}_c^{j=2}\right) &= \frac{\sum_{x=1}^{K\cdot W} \frac{\overrightarrow{v_x^{j=1}} \cdot \overrightarrow{v_x^{j=2}}}{\left\|\overrightarrow{v_x^{j=1}}\right\|_2 \left\|\overrightarrow{v_x^{j=2}}\right\|_2}}{K \cdot W}.
\end{aligned} \tag{3}$$

Finally, the cross-modality loss $\mathcal{L}_{\text{cross}}$ is calculated using (4).

$$\mathcal{L}_{\text{cross}} = 1 - \text{similarity}. \tag{4}$$

*2) Model Loss:* In addition to using the cross-modality loss $\mathcal{L}_{\text{cross}}$ to enhance feature extraction, we apply another loss function, $\mathcal{L}_{\text{label}}$, during the training of the mmImage-based model for accurate hand detection. The model output includes the location of a bounding box around the hand object and a confidence value. The label loss $\mathcal{L}_{\text{label}}$ consists of the bounding box location loss $\mathcal{L}_{\text{bbox}}$ and the confidence loss $\mathcal{L}_{\text{conf}}$, as defined in (5):

$$\mathcal{L}_{\text{label}} = \alpha_{\text{bbox}} \cdot \mathcal{L}_{\text{bbox}} + \alpha_{\text{conf}} \cdot \mathcal{L}_{\text{conf}} \tag{5}$$

while $\alpha_{\text{bbox}}$ and $\alpha_{\text{conf}}$ represent the weights of $\mathcal{L}_{\text{bbox}}$ and $\mathcal{L}_{\text{conf}}$, respectively. The details of the label loss function can be found in [7]. Thus, the total model loss function $\mathcal{L}$ is the weighted sum of the two loss functions, as shown in (6).

$$\mathcal{L} = \alpha_{\text{cross}} \cdot \mathcal{L}_{\text{cross}} + \alpha_{\text{label}} \cdot \mathcal{L}_{\text{label}} \tag{6}$$

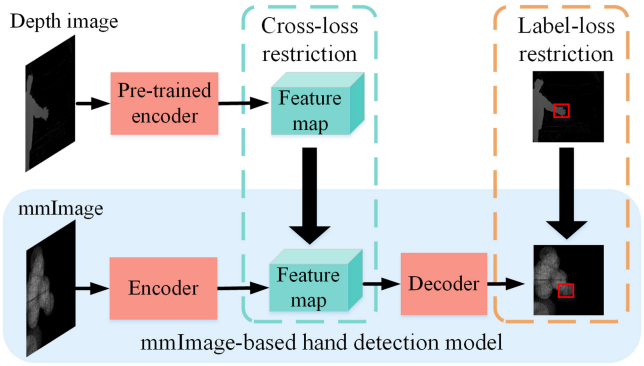



Fig. 6. Cross-modality spatial-feature enhanced hand detection model.

**Algorithm 2** Context-based dynamic hand localization algorithm.

**Input:** $\mathcal{U}$ consecutive frames of predicted hand locations $A_1, A_2, \ldots, A_{\mathcal{U}}$ and corresponding confidences $conf_1, conf_2, \ldots, conf_U$, the combined frame number $\mathcal{W}$

**Output:** $\mathcal{U}$ consecutive frames hand locations $\mathcal{O} = B_1, B_2, \ldots, B_{\mathcal{U}}$

1: initializes the output set $\mathcal{O} = \emptyset$
2: **for** $i = 1, ..., \mathcal{U}$ **do**
3: $B_i = \frac{A_i \cdot conf_i + \Sigma_{x=j}^{i-1} B_x \cdot conf_x \cdot \frac{\mathcal{W}-i+x}{\mathcal{W}}}{conf_i + \Sigma_{x=j}^{i-1} conf_x \cdot \frac{\mathcal{W}-i+x}{\mathcal{W}}}, j = \max(1, i-\mathcal{W})$
4: add $B_i$ in $\mathcal{O}$
5: **end for**
6: **return** $\mathcal{O}$

while $\alpha_{\text{cross}}$ and $\alpha_{\text{label}}$ represent the weights of $\mathcal{L}_{\text{cross}}$ and $\mathcal{L}_{\text{label}}$, respectively.

### D. Pixel-Level Dynamic Hand Localization

In hand localization, the hand is typically represented by its center, so pixel-level dynamic hand localization involves accurately identifying the hand center in pixel space. Since the bounding box is a rectangular box that tightly encloses the object, the center of the hand bounding box can be used as the hand center. By applying the trained mmImage-based hand detection model, we can localize the hand in pixel space with a confidence value. However, environmental disturbances and internal circuit deviations can reduce the quality of the mmWave data, making some frames unsuitable for hand localization. Additionally, due to model accuracy limitations, the trained mmImage-based model may fail to accurately localize the hand in certain frames, even when the data quality is good.

To address the challenges mentioned above, we designed a context-based algorithm for accurate dynamic hand localization, leveraging the continuity of hand movement. The details of the dynamic hand localization algorithm are provided in Algorithm 2. First, the mmImage-based hand detection model predicts the hand bounding box for $\mathcal{U}$ consecutive frames, where the center of each bounding box is set as the hand center. The $\mathcal{U}$ predicted hand centers and their corresponding confidence values are used as input for the algorithm. In step 1, we initialize an empty set to store the output hand centers.

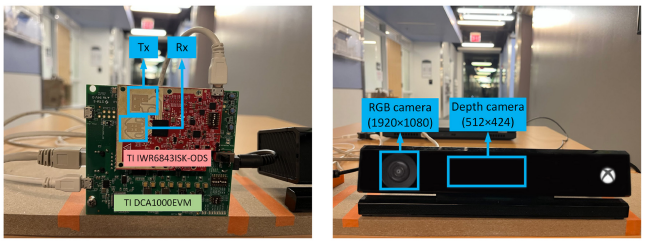



Fig. 7. TI 6843ISK-ODS and Kinect V2.

In steps 2-4, we calculate the hand center for each frame using a weighted summation. Specifically, for the $i_{th}$ frame, we compute a weighted sum of the $i_{th}$ frame's predicted hand center location along with the output hand center locations from the preceding $\mathcal{W}$ frames (if fewer than $\mathcal{W}$ frames are available, we sum all previous $i$ frames). We then divide this sum by the total weight of all combined frames to obtain the weighted average hand center location for the $i_{th}$ frame. Additionally, each preceding frame's weight in the $i_{th}$ frame's localization is determined by the product of its confidence and the time difference relative to the current frame, meaning the hand center location of the current frame depends more heavily on the most recent high-confidence frames. Finally, the algorithm outputs the hand center locations for the $\mathcal{U}$ consecutive frames.

In summary, the mmHand system is comprised of three key components: the mmImage generation tool, cross-modality hand detection, and pixel-level dynamic hand localization. Initially, the mmImage generation tool captures spatial information from raw mmWave data and converts it into pixel space. The generated mmImages are then used to train a hand detection model that benefits from improved feature extraction. Lastly, a context-aware algorithm is employed to overcome challenges related to inconsistent data quality and model detection bias, ultimately achieving precise, pixel-level dynamic hand localization.

## VI. Evaluation

### A. Experiment Preparation

*1) TestBeds:* As shown in Fig. 7(a), we use the TI IWR6843ISK-ODS mmWave radar for data sampling, connected to the TI DCA1000EVM board for data transmission. The radar has 3 transmitting antennas (Tx) and 4 receiving antennas (Rx). Each transmitting antenna emits its signal in an assigned time slot using MIMO technology, and the signal is received by the four receiving antennas for further processing. We set the number of chirps $C$ per mmWave frame to 64, with 256 samples per chirp. Additionally, as shown in Fig. 7(b), we use the Kinect V2 depth camera to collect depth image data. The camera operates at a 30 FPS frame rate, with a depth image resolution of 512 by 424 pixels.

*2) Data Collection and Labels:* As shown in Fig. 8, 12 volunteers (eight males and four females) are asked to move one hand freely in front of the mmWave sensor and the Kinect depth camera in three different scenarios: Hall, Meeting Room, and Corridor. These scenarios collectively represent a wide

Fig. 8. Three data collection scenarios: hall, meeting room and corridor (from left to right).

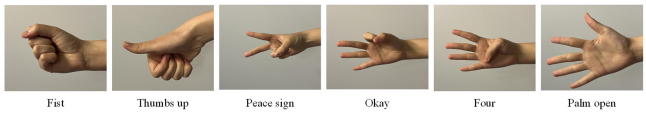



Fig. 9. Six basic hand gestures.

TABLE III
NUMBER OF SUBJECTS AND SYNCHRONIZED FRAMES IN EACH OF THE THREE SCENARIOS

| Scenario | Subjects | Frames number |
|---|---|---|
| Hall | 12 | 4089 |
| Meeting room | 12 | 4284 |
| Corridor | 12 | 3766 |

range of real-world applications for hand localization. The hall scenario covers large, open spaces, such as auditoriums, testing the system's ability to handle wide-area coverage. The meeting room represents medium-sized, enclosed environments like offices, where multipath interference from surrounding objects (e.g., chairs, tables and people) challenges the system's hand localization in cluttered settings. The corridor scenario simulates narrow, confined spaces like hallways, where more severe environmental interference occurs. Together, these scenarios ensure the system's robustness across diverse spatial conditions.

In each scenario, we collect five sets of left-hand data and 5 sets of right-hand data for each subject, with each set comprising 20 s of hand movement. To ensure a variety of hand gestures, volunteers are asked to perform six basic hand gestures. As shown in Fig. 9, these gestures include: fist, thumbs up, peace sign, okay, four, and palm open. During data collection, subjects stand at random distances from the radar (within a 2-m range) and move freely, adjusting their speed and transitioning between gestures. The combination of three scenarios, six gesture variations, and varying distances and movement speeds reflects the diversity and unpredictability of real-world environments, ensuring the system's robustness and effectiveness across different conditions. We preserve the timestamps of both the mmWave data and depth images during data collection and synchronize the two data modalities based on the timestamps. The synchronized mmWave data are then converted into the mmImage format for training the hand detection model. Additionally, a pretrained hand detection model [8] is used to label the mmImage in pixel space through the mmWave labeling method described in Section V-B. The number of subjects and synchronized labeled mmImages in each scenario are listed in Table III.

*3) Model Setting:* We use YOLOv3 as the base hand detection model due to its proven effectiveness in real-time object detection and its ability to extract robust spatial features from mmImages, outputting hand bounding boxes along with confidence values. To ensure high-quality data input, we set the SSIM threshold to 0.2. This threshold is determined based on the observation that compared to synchronized Kinect depth images, mmImages with an SSIM below 0.2 introduce stronger noise and degrade model performance. Only mmImages with SSIM values exceeding this threshold are included in the training process to enhance data reliability. For the loss function, the weights for cross-modality loss and label loss are both set to $\alpha_{\text{cross}} = 1$ and $\alpha_{\text{label}} = 1$, respectively. These values are chosen after testing various weight combinations, where equal weighting provided the best balance between maintaining consistency across modalities and ensuring accurate label supervision. We randomly select 80% of all subjects for training, with the remaining 20% reserved for testing. This split maximizes the training data while preserving a sufficient test set to evaluate generalization. The model is trained for 100 epochs with a batch size of 8, chosen to balance GPU memory constraints and training stability. The initial learning rate is set to $10^{-3}$, and reduced to $3 \times 10^{-6}$ using a cosine annealing strategy. This learning rate schedule is chosen after comparing it with step decay and exponential decay, with cosine annealing yielding smoother convergence and improved final accuracy. The mmHand system is implemented using PyTorch, and all training is conducted on a TITAN Xp GPU. This hardware setup allows for efficient model training while maintaining flexibility for hyperparameter tuning.

### B. System Comparison and Performance Metrics

*1) System Comparison:* In the experiments, we evaluate three different system structures.

*KinectSys (Benchmark):* KinectSys is a system that includes a hand detection model trained on Kinect depth images. Due to the high resolution of these images, KinectSys achieves high accuracy in hand detection and is robust across various scenarios. It serves as the benchmark for measuring the feature extraction capabilities of other systems.

*mmImageSys (Baseline):* mmImageSys has the same structure as KinectSys, but the hand detection model is trained on mmImage data. It serves as the baseline for evaluating the basic performance of mmImage data in hand localization.

*mmHand system:* The mmHand system is our proposed hand localization system, sharing the same structure as KinectSys and mmImageSys, but incorporating a cross-modality connection to KinectSys to enhance feature extraction in the mmImage. The system parameter settings are outlined in Section VI-A3.

*2) Performance Metrics:* We use the following four metrics to evaluate the performance of our proposed hand localization system.

$\boldsymbol{\mathcal{L}_{\text{cross}}}$ *(Cross-Modality Loss):* As introduced in Section V-C1, $\mathcal{L}_{\text{cross}}$ measures the inverse of the feature map similarity between two hand detection models trained by mmImage and Kinect depth images, respectively. Given

the high hand detection accuracy of the model trained by Kinect depth images, it excels at extracting high-resolution features. The feature map similarity evaluates the mmImage model's ability to extract similarly high-resolution features as the Kinect model.

***$\sigma^2$*** *(Attention Map Variance):* Variance $\sigma^2$ is a metric used to measure the concentration of a model's attention map [9], [10], with higher variance indicating a more scattered attention map. Given an attention map with dimensions $(W, H)$, it can be represented as a binary function $f(x, y)$, where $x$ and $y$ are the two dimensions, with $0 \leq x \leq W$ and $0 \leq y \leq H$. The variance $\sigma^2$ is calculated using (7)

$$\sigma^2 = \frac{\Sigma_{x=0}^{W}\Sigma_{y=0}^{H}\left((x-\bar{x})^2+(y-\bar{y})^2\right)\cdot f(x,y)}{\Sigma_{x=0}^{W}\Sigma_{y=0}^{H} f(x,y)} \tag{7}$$

where $\bar{x} = [\Sigma_{x=0}^{W}\Sigma_{y=0}^{H} x \cdot f(x,y)/\Sigma_{x=0}^{W}\Sigma_{y=0}^{H} f(x,y)]$ and $\bar{y} = [\Sigma_{x=0}^{W}\Sigma_{y=0}^{H} y \cdot f(x,y)/\Sigma_{x=0}^{W}\Sigma_{y=0}^{H} f(x,y)]$, and $(\bar{x}, \bar{y})$ represents the weighted center of the attention map.

*Average Precision (AP):* AP is a commonly used metric for measuring the accuracy of object detection [11], [12], [13]. It is defined as the area under the precision-recall curve, where precision and recall are calculated based on True Positive (TP), False Positive (FP), and False Negative (FN) values. To determine whether a detection is classified as TP, the IoU (Intersection Over Union) metric is used. IoU quantifies the overlap between the predicted hand bounding box and the labeled hand bounding box, and is mathematically defined as the ratio of the intersection area (overlapping region) to the union area (combined area of both boxes), given by $IoU = (Area\ of\ Overlap/Area\ of\ Union)$. An IoU value of 1.0 indicates a perfect match, while 0.0 means no overlap. A predefined IoU threshold is set, where detections with an IoU above the threshold are classified as TP, while those below are considered FP.

*Average Pixel Error (APE):* APE is a distance error metric used to measure the accuracy of hand localization [14]. We calculate APE by averaging the Euclidean distance between the predicted hand center and the labeled hand center over a continuous period. Given the high accuracy of the Kinect model, we use it to provide the groundtruth hand center locations.

### C. System Factor Analysis

*1) High-Resolution Feature Extraction:* As shown in Fig. 6, we apply the cross-modality loss on the feature map to guide the mmImage-based model in extracting high-resolution features similar to those of the pretrained hand detection model using Kinect depth images (i.e., KinectSys). In this section, we evaluate the effect of the cross-loss constraint in model training by comparing the average cross loss $\mathcal{L}_{\text{cross}}$ values on the testing dataset. As shown in Table IV, our mmHand system demonstrates a superior feature extraction ability compared to mmImageSys. The extracted feature map has a data format of $(K, W, H)$, where $K$, $W$, and $H$ represent the number of channels, width, and height, respectively. We average the feature map along the channel dimension $K$ and visualize the averaged feature map for comparison. As shown in Fig. 10, the feature map from the mmHand system closely resembles the feature map from the KinectSys, demonstrating the effectiveness of the cross-loss constraint in improving feature extraction.

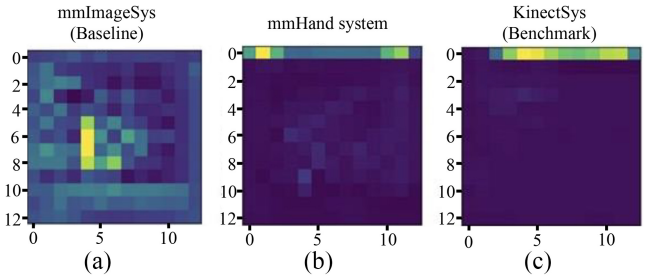



Fig. 10. Example of the feature map visualization results for the three systems: (a) mmImageSys, (b) our mmHand system, and (c) KinectSys.

TABLE IV
COMPARISON OF HIGH-RESOLUTION FEATURE EXTRACTION ABILITY BETWEEN MMHAND SYSTEM AND MMIMAGESYS (BASELINE) ($\mathbf{0 \leq \mathcal{L}_{cross} \leq 1}$, WHERE LOWER $\mathcal{L}_{cross}$ INDICATES BETTER PERFORMANCE)

| System Type | Scenario | $\mathcal{L}_{cross}$ |
|---|---|---|
| mmHand system | Hall | **0.335** |
| mmImageSys | Hall | 1.46 |
| mmHand system | Meeting room | **0.321** |
| mmImageSys | Meeting room | 1.73 |
| mmHand system | Corridor | **0.356** |
| mmImageSys | Corridor | 3.27 |

TABLE V
COMPARISON OF FEATURE ATTENTION ABILITY BETWEEN MMHAND SYSTEM AND MMIMAGESYS ($\mathbf{0 \leq \sigma^2 \leq 65025}$, WHERE LOWER $\sigma^2$ INDICATES BETTER PERFORMANCE)

| System Type | Scenario | $\sigma^2$ |
|---|---|---|
| mmHand system | Hall | **1092** |
| mmImageSys | Hall | 1580 |
| mmHand system | Meeting room | **546** |
| mmImageSys | Meeting room | 1837 |
| mmHand system | Corridor | **353** |
| mmImageSys | Corridor | 683 |

*2) Feature Attention:* Apart from improving high-resolution feature extraction, our cross-modality loss also guides the hand detection model to focus on specific feature areas. As shown in Table V, we compare the feature attention ability of hand detection models in two systems (i.e., mmHand system and mmImageSys) on the testing dataset. Our mmHand system has lower $\sigma^2$ (Attention Map Variance) compared to mmImageSys, indicating that the attention map from our mmHand system is more focused and demonstrates better feature attention ability. Additionally, we visualize the feature attention maps of the two systems and KinectSys. As shown in Fig. 11, the visualization results further confirm the effectiveness of our cross-modality loss in improving the system's feature attention ability.

*3) Hand Detection Results:* With improved high-resolution feature extraction and feature attention abilities, the mmImage-based model achieves more accurate hand detection. We compare the AP performance of the proposed mmHand system

TABLE VI
COMPARISON OF AP RESULTS BETWEEN MMHAND SYSTEM AND MMIMAGESYS ACROSS THREE SCENARIOS ($\mathbf{0 \leq AP \leq 100\%}$, WHERE HIGHER AP IS BETTER)

| System Type | Scenario | $AP_{0.5}$ | $AP_{0.4}$ | $AP_{0.3}$ | $AP_{0.2}$ | $AP_{0.1}$ |
|---|---|---|---|---|---|---|
| mmHand system | Hall | **29.96%** | **38.04%** | **46.97%** | **54.13%** | **59.51%** |
| mmImageSys | Hall | 15.35% | 21.13% | 26.93% | 34.69% | 43.03% |
| mmHand system | Meeting room | **21.77%** | **33.38%** | **43.70%** | **57.06%** | **67.75%** |
| mmImageSys | Meeting room | 15.77% | 26.85% | 35.18% | 43.87% | 56.74% |
| mmHand system | Corridor | **21.58%** | **30.87%** | **46.74%** | **57.06%** | **70.65%** |
| mmImageSys | Corridor | 20.11% | 28.10% | 38.70% | 49.28% | 60.88% |

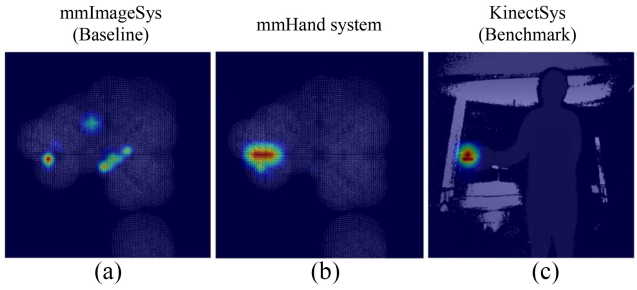



Fig. 11. Visualization of the feature attention maps for the three systems: (a) mmImageSys, (b) our mmHand system, and (c) KinectSys.

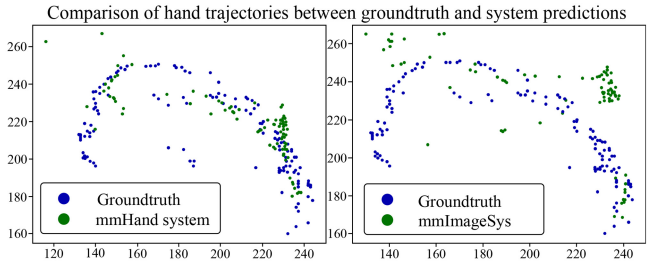



Fig. 12. Example of a pixel-level hand localization result comparison. Blue points represent the groundtruth. In the left figure, green points show the predicted hand centers from the proposed mmHand system, while in the right figure, green points represent the predicted hand centers from the mmImageSys system.

and mmImageSys across five different IoU thresholds (ranging from 0.5 to 0.1 in steps of 0.1). As shown in Table VI, the mmHand system outperforms mmImageSys in all three scenarios, demonstrating the superiority of our cross-modality spatial-feature enhanced model for hand detection.

*4) Pixel-Level Hand Localization:* To evaluate the accuracy of our hand localization system, we sample 200 consecutive frames during six hand movement periods from each of the 12 subjects and use our mmImage-based hand localization system to predict hand locations. As discussed in Section VI-B2, we use the predicted hand center from KinectSys as the groundtruth due to its high accuracy. We then calculate the APE for all subjects in each scenario. As shown in Table VII, the accuracy of our mmHand system is higher than that of mmImageSys. Additionally, Fig. 12 visualizes an example of a hand localization period, clearly demonstrating that the mmHand system achieves more accurate pixel-level dynamic hand localization.

TABLE VII
COMPARISON OF APE RESULTS BETWEEN MMHAND SYSTEM AND MMIMAGESYS ACROSS THREE SCENARIOS ($\mathbf{0 \leq APE \leq \sqrt{2}}$, WHERE LOWER APE IS BETTER)

| System Type | Scenario | APE |
|---|---|---|
| mmHand system | Hall | **0.265** |
| mmImageSys | Hall | 0.425 |
| mmHand system | Meeting room | **0.146** |
| mmImageSys | Meeting room | 0.584 |
| mmHand system | Corridor | **0.145** |
| mmImageSys | Corridor | 0.157 |

### D. System Impact Factors

*1) Radar-Hand Distance:* As the distance increases, the reflected signal strength weakens, reducing detection accuracy. In this section, we evaluate the impact of radar-hand distance on hand localization performance. Specifically, the subject stands at six different distances from the mmWave sensor, ranging from 0.6 to 1.6 m in 0.2 m increments. At each distance, we collect 200 consecutive frames over three hand movement periods, maintaining a constant movement speed of 0.2 m/s. Using the mmImage-based hand localization system to predict hand positions, we observe in Fig. 13(a) that the system's APE increases with distance, demonstrating that shorter radar-hand distances improve localization accuracy.

*2) Hand Movement Speed:* Hand movement speed can influence the hand localization system because faster movements can lead to motion blur in the radar signal, making it harder for the system to accurately track hand positions. To evaluate this impact on the proposed mmHand system, the subjects stand 0.6 m from the mmWave sensor and move their hand at three different speeds (0.2, 0.4, and 0.6 m/s). For each speed, we collect 200 consecutive frames over three hand movement periods to test the system's localization performance. As shown in Fig. 13(b), the APE increases with movement speed, demonstrating that faster hand movement reduces localization accuracy.

*3) Hand Size:* Theoretically, hand size can influence the hand localization system, as larger hands reflect more radar signals, resulting in a stronger and clearer signal for the system to process. In contrast, smaller hands may produce weaker reflections, making accurate detection and localization more challenging. However, since the range of hand size variation is relatively small [15], its overall impact is minimal. Typically, hand size is measured by hand length, defined as the distance

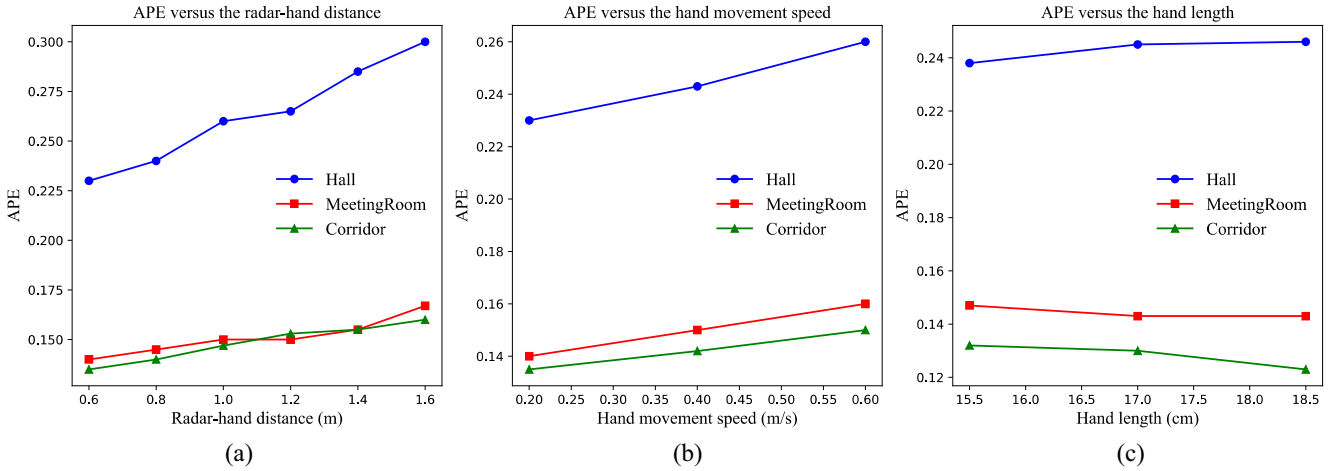



Fig. 13. Experimental results for system impact factors. (a) APE performance versus radar-hand distance. (b) APE performance versus hand movement speed. (c) APE performance versus hand size.

from the tip of the middle finger to the base of the hand [16]. To assess the impact of hand size on our mmHand system, we collect hand movement data from three subjects with different hand lengths (15.5, 17, and 18.5 cm). For each hand length, we record 200 consecutive frames over three hand movement periods to evaluate the system's localization performance. As shown in Fig. 13(c), the APE performance remains relatively consistent, indicating that hand size has little effect on the hand localization performance of the mmHand system.

Based on the experimental results of the three system impact factors, although radar-hand distance and hand movement speed influence system performance to some extent, the low APE ensures that the mmHand system remains viable for many real-world applications [17], [18], [19].

## VII. Discussions

### A. System Latency Analysis

We use the Intel Xeon CPU E5-1620 v4 @ 3.5 GHz to process the mmWave sensor data, while an Nvidia Titan XP is deployed for hand localization model predictions. The data format for each collected frame is set as $(\mathcal{T}, \mathcal{R}, \mathcal{C}, \mathcal{N}) = (3, 4, 64, 256)$. The time cost for each component of the mmHand system is presented in Table VIII, with the total system latency being approximately 0.1 s, which is sufficient for most real-time applications. Additionally, the data processing component of the system can be further accelerated in two areas.

First, the number of chirps per frame $\mathcal{C}$ is related to the velocity resolution. Reducing $\mathcal{C}$ will not affect the spatial resolution of the mmWave sensor, but it can decrease the mmWave data size and accelerate mmImage generation. However, too few chirps can lead to instability in mmImage generation, so balancing $\mathcal{C}$ is essential for achieving real-time performance in the mmHand system while maintaining localization accuracy. Second, in the mmImage generation tool, MVDR generates range-azimuth-elevation data in the format $(\mathcal{D}, \Theta, \Phi)$, where $\mathcal{D}$ ranges up to 10 m. In practice, if the human object is too far from the mmWave sensor, the reflected signal will have low energy and be heavily affected by environmental noise. Additionally, several works have demonstrated coarse-grained localization of objects from the range bin $(\mathcal{T}, \mathcal{R}, \mathcal{C}, \mathcal{D})$, which serves as the input to MVDR [20], [21], [22]. Therefore, it is feasible to add an object range detection module to filter out data outside the object's distance range, significantly reducing the time cost of MVDR.

TABLE VIII
mmHand System Latency Analysis

| Data collection | Data processing | Model detection |
|---|---|---|
| 0.03s | 0.05s | 0.02s |

### B. mmHand for Future Applications

*Multihand Localization:* Currently, our current mmHand system is capable of dynamically localizing a single hand with various gestures. However, with advancements in resolution and feature attention enabled by cross-modality learning, there is potential to classify and localize multiple hands from the generated mmImages [23], [24]. In future work, we plan to explore multihand localization using the mmHand system in simple scenarios, such as two hands from one or two individuals moving without overlap.

*Hand Pose Recognition:* Since the generated mmImage has high spatial information similarity with the synchronized depth image, it is possible to extract coarse-grained hand contour features from the mmImage, which can be used for hand pose recognition. To date, many super-resolution methods have been developed to improve object recognition accuracy [25], [26], [27]. In the future, we plan to integrate these methods into our mmHand system to achieve accurate hand pose recognition.

*Occlusions:* Due to the inherent penetration ability of mmWave signals [28], [29], mmWave sensor can detect objects behind occlusions, whereas vision-based cameras lose functionality in such conditions. However, occlusion causes significant signal attenuation, making the received signal

more vulnerable to environmental interference. Additionally, the material structure of the occlusion can induce phase shifts in the received signal, both of which negatively impact mmWave data quality. In the future, we will further investigate how different occlusions affect data quality and apply our mmHand system for hand localization in occluded environments.

### C. mmHand for Industry Applications

In addition to future research directions, the mmHand system has significant potential for industrial applications, particularly in environments where traditional camera-based systems face limitations. Below, we discuss three example use cases:

*Quality Inspection in Manufacturing:* Camera-based systems in the manufacturing operation line often fail when lenses are obscured by dust or grease. For instance, in food processing plants, workers frequently interact with machinery to adjust settings or sort items on a conveyor belt. Over time, airborne grease particles or fine food dust can accumulate on camera lenses, significantly blocking the camera from detecting and localizing hands accurately [30], [31]. This can lead to system failures in detecting wrong operations during production, thus making the product of bad quality. The deep learning approach can mitigate the impact of interfering substances to some extent through data-driven model training [32], [33], [34]. However, the complex and dynamic nature of the manufacturing environment, along with the diverse sources of interference, poses significant challenges for model training [35], [36], [37]. In contrast, the mmHand system operates using the mmWave signal to penetrate these contaminants, ensuring reliable hand localization. This robustness reduces downtime for cleaning and maintenance, making mmHand particularly suitable for continuous operation in contaminated environments [38], [39]. Additionally, the privacy-preserving nature of mmWave signals ensures that sensitive operation processes remain secure [40].

*Sterile Environments in Healthcare:* In healthcare, particularly in operating rooms or intensive care units, maintaining sterility is critical [41], [42]. For example, during a surgical procedure, surgeons often need to interact with medical devices without physically touching them to avoid contamination. Camera-based systems in such settings may fail due to condensation on the lens from the operating room's humid environment or blood splatters obstructing the camera's view [43], [44]. The mmHand system, however, operates effectively under these conditions because mmWave signals are unaffected by visual obstructions. Moreover, the mmHand system's penetration capability allows it to function even through thin surgical drapes or plastic covers, which are commonly used to maintain sterility [45], [46]. Additionally, mmWave radiation operates within the nonionizing spectrum, with energy levels significantly lower than those of X-rays or CT imaging systems, posing no known health risks to patients or medical staff [47]. These attributes make mmHand a safe and effective solution for enhancing precision and efficiency in sterile surgical environments.

*Harsh Conditions in Mining or Heavy Industries:* Mining operations or heavy industries present challenging conditions, including high levels of dust, vibration, and debris [48], [49], [50]. In these environments, the ability of mmHand to penetrate through obstructions and maintain accuracy makes it a reliable tool for remote control of heavy machinery [51]. With its compact form factor (millimeter-level antenna size) and low power consumption (within 5 mW [52]), the mmHand system is well-suited for integration into wearable edge devices that enable real-time control of machinery in mining environments. For instance, in underground mining, workers often operate machinery in low-visibility conditions due to airborne dust particles. The camera-based hand localization system doesn't work as the lens becomes obscured by dust [53]. In contrast, the mmHand system remains unaffected by such obstructions, ensuring accurate hand tracking and control of machinery.

## VIII. Related Work

*Cross-modality Learning:* Cross-modality learning is an effective method to combine the advantages of different data modalities, leading to improved performance. Currently, fusion and guidance are the two primary approaches in cross-modality learning technologies [54], [55], [56], [57], [58], [59]. In the fusion approach, features extracted from different data modalities are combined to enhance the performance of specific model tasks [60], [61], [62], [63]. For example, Xue et al. [64] introduced a DeepFusion model that integrates features from different sensor modalities along with cross-sensor correlations, improving performance in IoT applications. Shuai et al. [65] proposed a lightweight mmWave sensor and camera fusion system for more robust object detection, using mmWave sensor to address lighting issues and RGB images to enhance model accuracy. In contrast, the guidance approach enables a high-quality data modality to guide feature extraction from a lower-quality modality [66], [67], [68]. Cai et al. [69] developed a cross-modality interaction between depth images and RGB images, improving the encoder's ability to extract more precise depth information from RGB data. Zhao et al. [70] used RGB data to guide RF signal model training, effectively improving human pose estimation from RF signals. Unlike these works, our mmHand system explores a more efficient cross-modality learning approach by investigating a new image-format mmWave data representation for deeper integration between depth images and mmWave data.

*Hand Localization Sensor Systems:* Hand localization has been a widely studied problem for several years, with various technologies explored to achieve accurate localization using different data modalities [46], [71], [72], [73], [74], [75]. For example, Liu et al. [76] presented a dynamic hand localization and gesture recognition system based on RGBD video, which provides rich visual information but raises privacy concerns and requires sufficient and stable lighting conditions. Baldi et al. [77] developed a sensing glove with wearable sensors, such as IMU, ECG, and EMG,

TABLE IX
ADVANTAGES OF MMWAVE DATA COMPARED TO OTHER DATA MODALITIES

| Data modality | Technology | Advantages of the mmWave modality |
|---|---|---|
| Radio signal | WiFi, RFID and Bluetooth | Higher resolution |
| Acoustic signal | Microphone and Speaker | Not affected by sound pollution |
| Camera image | RGB and depth image | Privacy-preserving and light-free |
| Ultrasound | Ultrasonic | Lower price and smaller size |
| Wearable sensor data | IMU, ECG and EMG | No accumulated error, better user comfort and no battery capacity limitation |

which ensures accurate tracking but may cause user discomfort due to sensor placement and battery limitations. Wang et al. [78] proposed WiTrace, achieving centimeter-level passive hand localization using WiFi signals, though it necessitates dense WiFi deployments. Chen et al. [79] implemented a smartphone-based prototype using acoustic signals to locate hand positions, which avoids occlusion issues but is susceptible to environmental sound pollution. Additionally, McIntosh and Fraser [80] utilized ultrasonic waves for hand tracking, offering an alternative approach but facing challenges related to high costs and large device sizes. Table IX highlights the advantages of mmWave data compared to other data modalities [45], [81]. Unlike existing hand localization technologies that rely on specific sensor placements or complex data processing algorithms, our mmHand system provides privacy-preserving, device-free hand localization without these requirements.

## IX. CONCLUSION

In this article, we introduce mmHand, a novel hand localization system that achieves pixel-level accuracy using a single commodity mmWave radar. The system requires no specific sensor placement and can accurately predict dynamic hand locations in pixel space. A new mmImage generation tool is designed to fully extract spatial information from raw mmWave data and represent it in pixel space. Additionally, the system introduces innovative methods for quality evaluation and pixel space labeling of time-series mmWave data. Leveraging the spatial information similarity between mmWave data and camera depth images, the system also features a cross-modality spatial-feature enhanced model for more accurate pixel-level hand localization. Experiments with 12 subjects across three different scenarios, using four metrics, demonstrate the efficiency of our mmHand system in hand localization.

## REFERENCES


[1] "This is meta quest." MetaQuest. Accessed: Oct. 20, 2022. [Online]. Available: https://store.facebook.com/quest/products/quest-2/

[2] "Say hello to the second generation of our iconic hand tracking camera." Ultraleap. Oct. 2022. [Online]. Available: https://www.ultraleap.com/

[3] F. Zhang et al., "Mediapipe hands: On-device real-time hand tracking," 2020, *arXiv:2006.10214*.

[4] S. D. Regani, C. Wu, B. Wang, M. Wu, and K. R. Liu, "*mmWrite*: Passive handwriting tracking using a single millimeter-wave radio," *IEEE Internet Things J.*, vol. 8, no. 17, pp. 13291–13305, Sep. 2021.

[5] A. Ninos, J. Hasch, M. Heizmann, and T. Zwick, "Radar-based robust people tracking and consumer applications," *IEEE Sensors J.*, vol. 22, no. 4, pp. 3726–3735, Feb. 2022.

[6] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[8] L. Ondris. "LadaOndris/hand-recognition." 2020. [Online]. Available: https://github.com/LadaOndris/hand-recognition

[9] J. P. Robinson, Y. Li, N. Zhang, Y. Fu, and S. Tulyakov, "Laplace landmark localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 10103–10112.

[10] P. Domingos, "A unified bias-variance decomposition," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 231–238.

[11] R. Padilla, S. L. Netto, and E. A. Da Silva, "A survey on performance metrics for object-detection algorithms," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, 2020, pp. 237–242.

[12] S.-W. Kim, H.-K. Kook, J.-Y. Sun, M.-C. Kang, and S.-J. Ko, "Parallel feature pyramid network for object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 234–250.

[13] W. Fang, L. Wang, and P. Ren, "Tinier-YOLO: A real-time object detection method for constrained environments," *IEEE Access*, vol. 8, pp. 1935–1944, 2020.

[14] H. Fan et al., "LaSOT: A high-quality large-scale single object tracking benchmark," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 439–461, 2021.

[15] "What is my hand size?" Accessed: Sep. 1, 2024. [Online]. Available: https://choosehandsafety.org/choosing-hand-tools/hand-tool-size#ft1

[16] Ansell. "How to measure glove size." Accessed: Sep. 1, 2024. [Online]. Available: https://www.ansell.com/us/en/blogs/safety-briefing/na/na_how-to-measure-glove-size

[17] Y. Li, R. Reddy, C. Zhang, and R. Nandakumar, "Beyond-voice: Towards continuous 3D hand pose tracking on commercial home assistant devices," in *Proc. 23rd ACM/IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, 2024, pp. 151–162.

[18] N. Ettu, "Analysis of indoor localization methods for smart home automation systems," M.S. thesis, Dept. Inf. Technol. (Internet of Things), Savonia Univ. Appl. Sci., Kuopio, Finland, 2024.

[19] D. Schneider et al., "Accuracy evaluation of touch tasks in commodity virtual and augmented reality head-mounted displays," in *Proc. ACM Symp. Spatial User Interact.*, 2021, pp. 1–11.

[20] J. Bhatia et al., "Classification of targets using statistical features from range FFT of mmWave FMCW radars," *Electronics*, vol. 10, no. 16, p. 1965, 2021.

[21] S. Hamidi and S. Safavi-Naeini, "Single channel mmWave FMCW radar for 2D target localization," in *Proc. IEEE 19th Int. Symp. Antenna Technol. Appl. Electromagn. (ANTEM)*, 2021, pp. 1–2.

[22] Q. Zhao, G. Cui, S. Guo, W. Yi, L. Kong, and X. Yang, "Millimeter wave radar detection of moving targets behind a corner," in *Proc. 21st Int. Conf. Inf. Fusion (FUSION)*, 2018, pp. 2042–2046.

[23] H. Xue et al., "M$^4$esh: mmWave-based 3D human mesh construction for multiple subjects," in *Proc. 20th ACM Conf. Embed. Netw. Sensor Syst.*, 2022, pp. 391–406.

[24] C. Wu, F. Zhang, B. Wang, and K. R. Liu, "mmTrack: Passive multi-person localization using commodity millimeter wave radio," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2020, pp. 2400–2409.

[25] Y. Liu, S. Zhang, J. Xu, J. Yang, and Y.-W. Tai, "An accurate and lightweight method for human body image super-resolution," *IEEE Trans. Image Process.*, vol. 30, pp. 2888–2897, 2021.

[26] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[27] S. Anwar and N. Barnes, "Densely residual Laplacian super-resolution," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1192–1204, Mar. 2022.

[28] J. Pegoraro, M. Canil, A. Shastri, P. Casari, and M. Rossi, "ORACLE: Occlusion-resilient and self-calibrating mmWave radar network for people tracking," 2022, *arXiv:2208.14199*.

[29] H. Xue et al., "mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave," in *Proc. 19th Annu. Int. Conf. Mobile Syst., Appl., Services*, 2021, pp. 269–282.

[30] J. Kold and C. Silverman, "Conveyors used in the food industry," in *Handbook of Hygiene Control in the Food Industry*. Amsterdam, The Netherlands: Elsevier, 2016, pp. 367–382.

[31] R. Kohli, "Methods for monitoring and measuring cleanliness of surfaces," in *Developments in Surface Contamination and Cleaning*. Amsterdam, The Netherlands: Elsevier, 2012, pp. 107–178.

[32] P. Gutierrez, M. Luschkova, A. Cordier, M. Shukor, M. Schappert, and T. Dahmen, "Synthetic training data generation for deep learning based quality inspection," in *Proc. 15th Int. Conf. Qual. Control Artif. Vis.*, 2021, pp. 9–16.

[33] T.-H. Kim, H.-R. Kim, and Y.-J. Cho, "Product inspection methodology via deep learning: An overview," *Sensors*, vol. 21, no. 15, p. 5039, 2021.

[34] X. Zheng, S. Zheng, Y. Kong, and J. Chen, "Recent advances in surface defect inspection of industrial products using deep learning techniques," *Int. J. Adv. Manuf. Technol.*, vol. 113, pp. 35–58, Mar. 2021.

[35] J. Xu et al., "A review on AI for smart manufacturing: Deep learning challenges and solutions," *Appl. Sc.*, vol. 12, no. 16, p. 8239, 2022.

[36] J. Yang, S. Li, Z. Wang, H. Dong, J. Wang, and S. Tang, "Using deep learning to detect defects in manufacturing: A comprehensive survey and current challenges," *Materials*, vol. 13, no. 24, p. 5755, 2020.

[37] R. Malhan and S. K. Gupta, "The role of deep learning in manufacturing applications: Challenges and opportunities," *J. Comput. Inf. Sci. Eng.*, vol. 23, no. 6, 2023, Art. no. 60816.

[38] A. B. Zekri, R. Ajgou, A. Chemsa, and S. Ghendir, "Analysis of outdoor to indoor penetration loss for mmWave channels," in *Proc. 1st Int. Conf. Commun., Control Syst. Signal Process. (CCSSP)*, 2020, pp. 74–79.

[39] J. Ryan, G. R. MacCartney, and T. S. Rappaport, "Indoor office wideband penetration loss measurements at 73 GHz," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2017, pp. 228–233.

[40] Y. Sun, R. Hang, Z. Li, M. Jin, and K. Xu, "Privacy-preserving fall detection with deep learning on mmWave radar signal," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, 2019, pp. 1–4.

[41] S. N. Nazhat, A. M. Young, and J. Pratten, "Sterility and infection," *Biomed. Materials*. Cham, Switzerland: Springer Int. Publ., 2009, pp. 239–260.

[42] X. Ma, K. Yang, P. Reeves, and S. Yu, "RFID-based healthcare workflow management in sterile processing departments," in *Proc. IIE Annu. Conf. Proc.*, 2012, pp. 1–10.

[43] S. A. H. Mohsan, "Optical camera communications: Practical constraints, applications, potential challenges, and future directions," *J. Opt. Technol.*, vol. 88, no. 12, pp. 729–741, 2021.

[44] B. Scott, M. Seyres, F. Philp, E. K. Chadwick, and D. Blana, "Healthcare applications of single camera markerless motion capture: A scoping review," *PeerJ*, vol. 10, May 2022, Art. no. e13517.

[45] Z. Hussain, M. Sheng, and W. E. Zhang, "Different approaches for human activity recognition: A survey," 2019, *arXiv:1906.05074*.

[46] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3200–3225, Mar. 2023.

[47] R. Dilli, "Implications of mmWave radiation on human health: State of the art threshold levels," *IEEE Access*, vol. 9, pp. 13009–13021, 2021.

[48] A. B. Cecala, *Dust Control Handbook for Industrial Minerals Mining and Processing*. Scotts Valley, CA, USA: CreateSpace Independ. Publ. Platform, 2012.

[49] J. Colinet, C. N. Halldin, and J. Schall, *Best Practices for Dust Control in Coal Mining*. Scotts Valley, CA, USA: CreateSpace Independ. Publ. Platform, 2021.

[50] T. Liu and S. Liu, "The impacts of coal dust on miners' health: A review," *Environ. Res.*, vol. 190, Nov. 2020, Art. no. 109849.

[51] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proc. 3rd ACM Workshop Millim.-Wave Netw. Sens. Syst.*, 2019, pp. 51–56.

[52] "Iwrl6432: Single-chip low-power 57-Ghz to 64-Ghz industrial mmWave radar sensor." Accessed: May 2, 2025. [Online]. Available: https://www.ti.com/product/IWRL6432

[53] N. Yaghoobi Ershadi and J. M. Menéndez, "Vehicle tracking and counting system in dusty weather with vibrating camera conditions," *J. Sensors*, vol. 2017, no. 1, 2017, Art. no. 3812301.

[54] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "CrDoCo: Pixel-level domain transfer with cross-domain consistency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1791–1800.

[55] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Attention bridging network for knowledge transfer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5198–5207.

[56] L. Song, B. Liu, G. Yin, X. Dong, Y. Zhang, and J.-X. Bai, "TACR-Net: Editing on deep video and voice portraits," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 478–486.

[57] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3774–3783.

[58] F. M. Thoker and C. G. Snoek, "Feature-supervised action modality transfer," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 3751–3758.

[59] B. Sun, X. Ye, B. Li, H. Li, Z. Wang, and R. Xu, "Learning scene structure guidance via cross-task knowledge transfer for single depth super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7792–7801.

[60] X. Bruce, Y. Liu, and K. C. Chan, "Multimodal fusion via teacher-student network for indoor action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3199–3207.

[61] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image Vis. Comput.*, vol. 105, Jan. 2021, Art. no. 104042.

[62] E. Debie et al., "Multimodal fusion for objective assessment of cognitive workload: A review," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1542–1555, Mar. 2021.

[63] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," in *Proc. 34th Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 2351–2363.

[64] H. Xue et al., "Deepfusion: A deep learning framework for the fusion of heterogeneous sensory data," in *Proc. 20th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2019, pp. 151–160.

[65] X. Shuai, Y. Shen, Y. Tang, S. Shi, L. Ji, and G. Xing, "milliEye: A lightweight mmWave radar and camera fusion system for robust object detection," in *Proc. Int. Conf. Internet-Things Design Implement.*, 2021, pp. 145–157.

[66] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*.

[67] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021.

[68] L. Beyer, X. Zhai, A. Royer, L. Markeeva, R. Anil, and A. Kolesnikov, "Knowledge distillation: A good teacher is patient and consistent," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10925–10934.

[69] Y. Cai, L. Ge, J. Cai, N. M. Thalmann, and J. Yuan, "3D hand pose estimation using synthetic data and weakly labeled RGB images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 3739–3753, Nov. 2021.

[70] M. Zhao et al., "Through-wall human pose estimation using radio signals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7356–7365.

[71] S. Cui, R. Wang, J. Hu, J. Wei, S. Wang, and Z. Lou, "In-hand object localization using a novel high-resolution visuotactile sensor," *IEEE Trans. Ind. Electron.*, vol. 69, no. 6, pp. 6015–6025, Jun. 2022.

[72] T. Ohkawa, Y.-J. Li, Q. Fu, R. Furuta, K. M. Kitani, and Y. Sato, "Domain adaptive hand keypoint and pixel localization in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 68–87.

[73] Y. Che, Y. Song, and Y. Qi, "A novel framework of hand localization and hand pose estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2019, pp. 2222–2226.

[74] T. Shimizu, R. Hachiuma, H. Kajita, Y. Takatsume, and H. Saito, "Hand motion-aware surgical tool localization and classification from an egocentric camera," *J. Imag.*, vol. 7, no. 2, p. 15, 2021.

[75] S. Koitka, A. Demircioglu, M. S. Kim, C. M. Friedrich, and F. Nensa, "Ossification area localization in pediatric hand radiographs using deep neural networks for object detection," *PloS One*, vol. 13, no. 11, 2018, Art. no. e0207496.

[76] W. Liu, Y. Fan, Z. Li, and Z. Zhang, "RGBD video based human hand trajectory tracking and gesture recognition system," *Math. Problems Eng.*, 2015, to be published.
[77] T. L. Baldi, M. Mohammadi, S. Scheggi, and D. Prattichizzo, "Using inertial and magnetic sensors for hand tracking and rendering in wearable haptics," in *Proc. IEEE World Haptics Conf. (WHC)*, 2015, pp. 381–387.
[78] L. Wang, K. Sun, H. Dai, A. X. Liu, and X. Wang, "WiTrace: Centimeter-level passive gesture tracking using WiFi signals," in *Proc. 15th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, 2018, pp. 1–9.
[79] H. Chen, F. Li, and Y. Wang, "EchoTrack: Acoustic device-free hand tracking on smart phones," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, 2017, pp. 1–9.
[80] J. McIntosh and M. Fraser, "Improving the feasibility of ultrasonic hand tracking wearables," in *Proc. ACM Int. Conf. Interact. Surfaces Spaces*, 2017, pp. 342–347.
[81] Y. Zhao, H. Zhou, S. Lu, Y. Liu, X. An, and Q. Liu, "Human activity recognition based on non-contact radar data and improved PCA method," *Appl. Sci.*, vol. 12, no. 14, p. 7124, 2022.

**Xiaoyu Zhang** received the B.E. degree from Hefei University of Technology, Hefei, China, in 2017, and the M.E. degree from the University of Science and Technology of China, Hefei, in 2020. He is currently pursuing the Ph.D. degree with computer science and engineering with the State University of New York at Buffalo, Amherst, NY, USA.

His research interests include wireless sensing, Internet of Things, and smart health.

**Zhengxiong Li** (Member, IEEE) received the B.E. and M.E. degrees in computer science from Hangzhou Dianzi University, Hangzhou, China, in 2013 and 2016, respectively, and the Ph.D. degree from University at Buffalo, Buffalo, NY, USA, in 2021.

He is currently an Assistant Professor with Computer Science and Engineering Department, University of Colorado at Denver, Denver, CO, USA. His research interests focus on Internet of Things/mobile and cybersecurity.

**Chenhan Xu** received the B.E. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2017, and the Ph.D. degree from the State University of New York at Buffalo, Amherst, NY, USA, in 2023.

He is currently an Assistant Professor with the Department of Computer Science, North Carolina State University, Raleigh, NC, USA. His research focuses on the convergence of the Internet of Things, physiological science, and smart health.

Dr. Xu has received five conference best paper awards at IEEE ICHI in 2022, ACM MobiSys in 2020, ACM SenSys in 2019, IEEE ICC in 2019, and IEEE GLOBECOM in 2016.

**Luchuan Song** received the B.E. and M.E. degrees from the Department of Electronic Engineering and information science, University of Science and Technology of China, Hefei, China, in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree with the Department of Computer Science, University of Rochester, Rochester, NY, USA.

His research interest lies in human-related topic. (e.g. face animation and toonfication, 3-D face reconstruction, Deepfake detection etc.).

**Huining Li** received the Ph.D. degree in the Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA, in 2024.

She is an Assistant Professor with the Department of Computer Science, North Carolina State University, Raleigh, NC, USA. Her research interest lies broadly in Internet of Things, cybersecurity, and mobile computing.

**Hongfei Xue** received the B.Eng. degree from the University of Science and Technology of China, Hefei, China, in 2015, and the Ph.D. degree from the State University of New York at Buffalo, Amherst, NY, USA, in 2023.

He is an Assistant Professor with the Department of Computer Science, University of North Carolina at Charlotte, Charlotte, NC, USA. His research interests focus on building robust and intelligent wireless sensing systems.

**Yingxiao Wu** received the M.S. degree from Hangzhou Dianzi University, Hangzhou, China, in 2003, and the Ph.D. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2010.

Since 2021, she has been a Research Fellow with the Department of Computer, Hangzhou Dianzi University. Her research interests include wireless sensing, pervasive computing, and industrial Internet.

**Wenyao Xu** (Senior Member, IEEE) received the B.S. and M.S. degrees (Hons.) from Zhejiang University, Hangzhou, China, in 2006 and 2008, respectively, and the Ph.D. degree from the University of California at Los Angeles, Los Angeles, CA, USA, in 2013.

He is currently a Professor with the Computer Science and Engineering Department, State University of New York at Buffalo, Buffalo, NY, USA. His recent research foci include the Internet of Things, smart health, and cybersecurity.