Environment Reconstruction based on Multi-User Selection and Multi-Modal Fusion in ISAC

Bo Lin, Chuanbin Zhao, Feifei Gao, *Fellow, IEEE*, and Geoffrey Ye Li, *Fellow, IEEE*

***Abstract*—Integrated sensing and communications (ISAC) has been deemed as a key technology for the sixth generation (6G) wireless communications systems. In this paper, we explore the inherent clustered nature of wireless users and design a multi-user based environment reconstruction scheme. Specifically, we first select users based on the estimation precision of channel's multipath, including the line-of-sight (LOS) and the non-line-of-sight (NLOS) paths, to enhance the accuracy of environment reconstruction. Then, we develop a fusion strategy that merges communications signalling with camera image to increase the accuracy and robustness of environment reconstruction. The simulation results demonstrate that the proposed algorithm can achieve a remarkable sensing accuracy of centimeter level, which is about 17 times better than the scheme without user selection. Meanwhile, the fusion of communications data and vision data leads to a threefold accuracy improvement over the image only method, especially under challenging weather conditions like raining and snowing.**

***Index Terms*—ISAC, environment reconstruction, multi-user selection, multi-modal fusion**

I. Introduction

The swift progress in artificial intelligence (AI) has notably propelled the development of sensing-assisted communications technology [1]–[6]. Various studies have employed sensing data, such as vision and radar, to improve the efficiency and quality of wireless communications. Reversely, wireless communications systems can fulfill the dual role of sensing the surrounding environment and data transmission [7]. The transmitted signals complexly interact with the environment during their journey to the receivers. Thus, the ultimately received signals carry a rich range of critical environment information or data. In fact, the expansion of communication frequency and bandwidth in the coming 6G is opening new opportunities for communication-assisted sensing.

Recently, integrated sensing and communications [8] (ISAC) has been proposed to facilitate concurrent high-speed data transmission and high-precision sensing. The development of ISAC gradually converges into two categories: moving target sensing [9]–[12] and static environment reconstruction [13]–[16]. In [10], a simultaneous beam training and target sensing scheme has been proposed. In [11], a root-MUSIC-based algorithm has been developed to estimate the kinematic parameters of identified moving targets in a cluttered environment. The ESPRIT-based moving target sensing method in [12] can achieve super-resolution and low-complexity estimation of the targets' parameters. On the aspect of environment reconstruction, the concept of simultaneous localization and mapping (SLAM) can construct a comprehensive radio map based on the multipath channel state information (CSI). The belief propagation (BP) based SLAM algorithm in [13] utilizes the association of specular multipath components (MPCs) with geometric features to reconstruct the environment. In [14], a Bayesian approach has been designed for communication-driven SLAM by extracting the soft information of channel parameters. The angle-based SLAM algorithm in [15] extends the classic BP SLAM algorithm. The multiple-model probability hypothesis density filter and map fusion routine in [16] can effectively map the radio environment.

However, existing works on reconstructing the environment are merely based on one single user, resulting in limited sensing information, sparse reconstruction results, and low reliability. Actually, the advantage of communications-sensing over radar-sensing lies in the fact that the former can utilize information from the massive users in communications systems. Therefore, it has been demonstrated in [17] that multi-user sensing can enhance the sensing performance. A centralized multi-user collaborative mapping and positioning approach has been proposed in [18]. The robust SLAM algorithm in [19] extends the classic BP-based SLAM algorithm to multi-user scenarios. However, due to variations in user positions and user orientations, the degree of alignment between the beam scanning direction and the angle of reflection paths varies among different users, resulting in different accuracies in environment reconstruction when only one user is used for sensing.

Nevertheless, the constructed environment point clouds are sparse due to the sparsity of the mmWave propagation paths. Traditional method utilizes visual sensors, which can capture rich information to reconstruct the environment. However, visual sensors have a limited detection range and therefore are highly influenced by weather conditions.

Some studies have focused on the fusion of radar data and vision data for environment reconstruction and depth estimation. The fusion of radar and vision has been proposed

B. Lin and F. Gao are with the Department of Automation, Tsinghua University, State Key Lab of Intelligent Technologies and Systems, Tsinghua University, State Key for Information Science and Technology (TNList), Beijing 100084, P. R. China (e-mail: feifeigao@ieee.org; linb20@mails.tsinghua.edu.cn).

C. Zhao is with the State Key Laboratory of Intelligent Technologies and Systems, Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Institute for Artificial Intelligence Tsinghua University (THUAI), Tsinghua University, Beijing 100084, China, and also with the Senior Engineer China Telecom Corporation Sichuan Branch, Chengdu 610000, China (e-mail: zcb23@mails.tsinghua.edu.cn).

Geoffrey Ye Li is with the Department of Electrical and Electronic Engineering, Imperial College London, SW7 2BU London, U.K. (e-mail:Geoffrey.Li@imperial.ac.uk).

in [20] for 3D object detection. The modified encoder-decoder deep convolutional neural network (CNN) in [21] can fuse the camera's and radar's measurements for depth reconstruction. The geometric method in [22] performs 3D reconstruction using a panoramic microwave radar and a camera. The point cloud reconstruction approach in [23] fuses millimeter wave radar data and vision data. In [24], the authors explored the possibility of achieving a more accurate depth estimation by fusing monocular images and radar points using a deep neural network (DNN).

Due to the high similarity between radar systems and communications systems [25], it is possible to fuse the communications signals in ISAC with vision data for environment reconstruction. However, there are still several challenges when fusing ISAC and vision:

(i) The sensing information obtained from communications is sparser compared to radar.
(ii) Radar sensing operates in a self-transmit and self-receive manner, capturing only reflection information while communication sensing involves both direction and reflection information, which requires extra effort to identify the reflection paths.
(iii) The sensing angle range of radar sensing is well-defined while the sensing angle range of communication sensing is random.
(iv) The fusion mechanism of communications data with vision data is not yet clearly defined.

In this paper, we reconstruct environment based on multi-user selection and multi-sensor fusion. The main contributions of this paper are as follows:

- We propose a criterion to select users to enhance the accuracy of environment reconstruction.
- We leverages the clustered nature of wireless users to enable rich sensing information.
- We fuse the information from ISAC and vision, and design a multi-modal fusion network (MMFN) to obtain accurate and robust environment reconstruction.
- We adopt the meta-learning strategy to train the MMFN to guarantee the effectiveness of the MMFN across different user quantities.

The rest of this paper is organized as follows. Section II introduces the multi-user selection and multi-sensor fusion based system model. Section III presents the evaluation criterion for assessing the sensing capabilities of users and proposes the user selection algorithms. Section IV designs the multi-modal fusion network for environment reconstruction. Section V provides the dataset generation and simulation results and Section VI draws the conclusion.

II. System Model

We consider an orthogonal frequency-division multiplexing (OFDM) mmWave communications system with one BS and N users. The BS is equipped with a uniform planar array (UPA) of N_t antennas, and the user is equipped with a UPA of N_r antennas. Denote the set of users as $\mathcal{U} = \{u_1, u_2, \cdots, u_N\}$. Considering the cost of hardware deployment, both the BS and the users adopt a fully analog architecture with only one radio frequency (RF) chain. Assume that

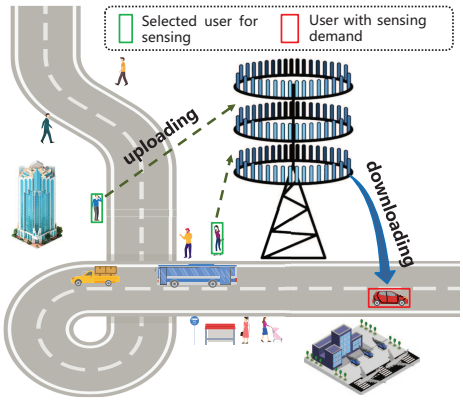


Fig. 1: The multi-user, multi-sensor fusion system for environment reconstruction.

there are L_c multipath components (MPC) between the user and the BS. The wireless parameters of the l-th MPC provided by the wideband mmWave geometric channel model [26] are: complex path gain α_l, time delay τ_l, azimuth angle of departure (AoD) $\phi_{t,l}$, elevation AoD $\theta_{t,l}$, azimuth angle of arrival (AoA) $\phi_{r,l}$, and elevation AoA $\theta_{r,l}$. The channel for the n-th OFDM symbol is

$$\mathbf{H}[n] = \sqrt{N_r N_t} \sum_{l=0}^{L_c-1} \alpha_l g(nT - \tau_l)\mathbf{a}(\phi_{r,l}, \theta_{r,l})\mathbf{a}^*(\phi_{t,l}, \theta_{t,l}), \tag{1}$$

where $g(\cdot)$ is the shaping pulse, $T = \frac{1}{B}$ is the symbol period, B is the bandwidth, $\mathbf{a}(\phi_{r,l}, \theta_{r,l})$ is the array steering vector at the receiver, and $\mathbf{a}(\phi_{t,l}, \theta_{t,l})$ is the array steering vector at the transmitter. Denote K as the total number of OFDM subcarriers. The frequency domain channel at subcarrier k is

$$\mathbf{H}[k] = \sum_{n=0}^{L-1} \mathbf{H}[n] e^{-j\frac{2\pi k}{K}n}, \tag{2}$$

where L is the maximum discrete-time delay of the channel.

Analog beamforming is employed at both the transmitter and the receiver based on fixed beam codebooks. Denote the transmit beam codebook as $\mathcal{F}_t = \{\mathbf{a}(\phi^b_{t,i}, \theta^b_{t,j}) | \phi^b_{t,i} \in \Phi^b_t, \theta^b_{t,j} \in \Theta^b_t\}$ and the receive beam codebook as $\mathcal{F}_r = \{\mathbf{a}(\phi^b_{r,p}, \theta^b_{r,q}) | \phi^b_{r,p} \in \Phi^b_r, \theta^b_{r,q} \in \Theta^b_r\}$, where Φ^b_t is the candidate set of transmit azimuth angle, Θ^b_t is the candidate set of transmit elevation angle, Φ^b_r is the candidate set of receive azimuth angle, and Θ^b_r is the candidate set of receive elevation angle[1]. At the beam management phase [27] of each user, the BS and the user perform exhaustive beam sweeping among all directions in the transmit and the receive codebooks. To simplify notation, we use quad (i, j, p, q) to represent performing beamforming with beam pair $\{\mathbf{a}(\phi^b_{t,i}, \theta^b_{t,j}), \mathbf{a}(\phi^b_{r,p}, \theta^b_{r,q})\}$. Given a quad (i, j, p, q), the received signal power is

$$y_{(i,j,p,q)} = \sum_{k=0}^{K-1} |\mathbf{a}^H(\phi^b_{r,p}, \theta^b_{r,q})\mathbf{H}[k]\mathbf{a}(\phi^b_{t,i}, \theta^b_{t,j})|^2. \tag{3}$$

[1]The angles in Φ^b_t and Θ^b_t are uniformly distributed.

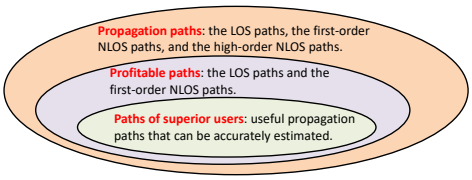


Fig. 2: The inclusion relationship between paths.

After exhaustive beam sweeping, the user recodes the signal powers in a four dimension power map (PM) tensor $\mathbf{P}_u \in \mathbb{R}^{|\Phi_t^b|\cdot|\Theta_t^b|\cdot|\Phi_r^b|\cdot|\Theta_r^b|}$, where $\mathbf{P}_u[i,j,p,q] = y_{(i,j,p,q)}$, and then feeds back the PM tensor to the BS.

III. User Selection Based Environment Reconstruction

The environment reconstruction is generally realized by collecting the reflection points of the first-order non-line-of-sight (NLOS) paths to a *point set* $\mathcal{P}$ [28]. Specifically, we use the line-of-sight (LOS) path to calculate the location of the user, and then utilize the user's location as well as the first-order NLOS paths to calculate the reflection points. Hence, in order to obtain accurate points, we have to precisely estimate the angles of the LOS and the first-order NLOS paths from the PM tensor, which includes the powers of the LOS path, the powers of the first-order NLOS paths, the powers of the high-order NLOS paths, and the powers of the noise. We define the *profitable paths* as the LOS paths and the NLOS paths that are beneficial for environment reconstruction as shown in Fig. 2. Subsequently, we define the superior users as the ones whose LOS and first-order NLOS paths can be precisely estimated from the PM tensor. We divide all user set $\mathcal{U}$ into two sets, $\mathcal{U}_s$ and $\mathcal{U}_n$, according to whether they are superior users or not. The steps of inferring whether a user is a superior user are:

(i) to extract the profitable paths that include the LOS paths and the first-order NLOS paths;
(ii) to distinguish whether the profitable paths are LOS paths or first-order NLOS paths;
(iii) to judge whether the LOS paths and the first-order NLOS paths can be accurately estimated.

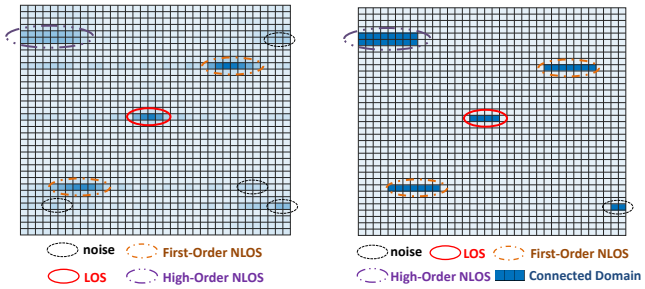
The above steps will be elaborated subsequently.

A. Identify The Profitable Paths

Among the PM tensors, the powers in the directions of the propagation paths are large while the noise powers in other directions are small. Then, we can use power as a threshold T_u to distinguish the propagation paths and the noise. Specifically, element $\mathbf{P}_u[i,j,p,q] \geq T_u$ is regarding as the propagation path while $\mathbf{P}_u[i,j,p,q] < T_u$ is regarding as the noise. A good threshold value T_u should make the inter-class variance between the propagation path and the noise large [29]. The steps to calculate the inter-class variance are as follows:

- Discretizing the values in the $\mathbf{P}_u$ into K values[2] to reduce computational complexity.

[2] K is a customizable parameter. When K is large, the discrete value is close to the continuous value. Then the calculated threshold will be accurate but complexity will be high.



(a) Original power map. (b) Binarized power map.

Fig. 3: An example of power map.

- Counting the number of occurrences for each discrete power, represented as n_k.
- Calculating the average power of the propagation paths as

$$m_1 = \sum_{k=1}^{T_u} n_k \cdot \sum_{k=1}^{T_u} \frac{k n_k}{\sum_{j=1}^{K} n_j}. \tag{4}$$

- Calculating the average power of noise as

$$m_2 = \sum_{k=T_u+1}^{K} n_k \cdot \sum_{k=T_u+1}^{K} \frac{k n_k}{\sum_{j=1}^{K} n_j}. \tag{5}$$

- Calculating the average power of fully $\boldsymbol{P}_u$ as

$$m_G = \sum_{k=1}^{K} \frac{k n_k}{\sum_{j=1}^{K} n_j}. \tag{6}$$

- Calculating the inter-class variance as

$$\sigma^2(T_u) = \frac{(m_1 - m_G)^2}{\sum_{k=1}^{T_u} n_k} + \frac{(m_2 - m_G)^2}{\sum_{k=T_u+1}^{K} n_k}. \tag{7}$$

Then, the optimal T_u should maximize the inter-class variance as

$$T_u^* = \arg\max_{1 \leq T_u \leq K} \sigma^2(T_u). \tag{8}$$

Next, we mark the propagation paths by setting the elements in $\mathbf{P}_u$ that are greater than or equal to T_u^* as "1"; otherwise, setting the elements in $\mathbf{P}_u$ as "0", that is,

$$\begin{aligned} \mathbf{P}_u^b[i,j,p,q] &= \text{binarize}(\mathbf{P}_u[i,j,p,q], T_u^*) \\ &= \begin{cases} 0, & P_u[i,j,p,q] < T_u^*, \\ 1, & P_u[i,j,p,q] \geq T_u^*. \end{cases} \end{aligned} \tag{9}$$

Define the area among $\mathbf{P}_u^b$ whose elements are all "1" and are adjacent to each other as a *connected domain* as shown in Fig. 3(b). Denote n_u connected domains of $\mathbf{P}_u^b$ as $\mathcal{C}_1^u, \mathcal{C}_2^u, \cdots, \mathcal{C}_{n_u}^u$. Each $\mathcal{C}_k^u$ encompasses the angular coordinates of the k-th connected domain. The presence of a connected domain suggests the possible existence of a propagation path. Denote the angles (AoA and AoD) of a propagation path as $(\hat{i}, \hat{j}, \hat{p}, \hat{q})$. Since the angles close to that of the propagation path also yield high powers, a valid propagation path will yield high values for multiple adjacent elements in $\mathbf{P}_u$ as shown in Fig. 3(a). Then one propagation path would yield multiple elements among the connected domain as shown in

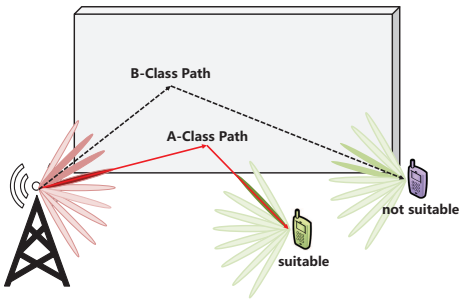


Fig. 4: The AoA and AoD of the NLOS path may not align precisely with the angles in the codebook.

Fig. 3(b). Conversely, a connected domain consisting of only one or a few elements may indicate a false alarm without a propagation path. Moreover, among the propagation paths, the angle spreads of the high-order NLOS paths are generally larger than that of the profitable paths [30]. Hence, as shown in Fig. 3(b), a connected domain consisting of relatively more elements may indicate a higher-order NLOS path. Then, we define a connectivity factor to indicate the profitable path as

$$c_u^k = \text{sign}\left(|\mathcal{C}_u^k| - \text{thr}_c\right) \cdot \text{sign}\left(\text{thr}_h - |\mathcal{C}_u^k|\right) + 1, \tag{10}$$

where $|\mathcal{C}_u^k|$ represents the size of $\mathcal{C}_u^k$, $\text{sign}(\cdot)$ represents the signum function, thr_c and thr_h represent the minimum size required for a connected domain to be considered as having a propagation path and having a high-order NLOS path, respectively. If size $|\mathcal{C}_u^k|$ of a connected domain is greater than thr_c but less than thr_h, then we set $c_u^k = 2$, which implies the connected domain $\mathcal{C}_u^k$ encompassing a useful propagation path; otherwise, we set $c_u^k = 0$, which implies the connected domain $\mathcal{C}_u^k$ not encompassing a profitable path. Moreover, if $\mathcal{C}_u^k$ encompasses a profitable path, then the angles of the highest received signal power in $\mathcal{C}_u^k$ are the elevation AoD, azimuth AoD, elevation AoA, and azimuth AoA of this path.

B. Distinguish The LOS Path and The First-Order NLOS path

After obtaining the connected domain with a profitable path, we need to recognize whether this path is a LOS path or a first-order NLOS path. Denote $\theta_t^{u,k}$, $\phi_t^{u,k}$, $\theta_r^{u,k}$, and $\phi_r^{u,k}$ as the elevation AoD, azimuth AoD, elevation AoA, and azimuth AoA of the path in $\mathcal{C}_u^k$, respectively. Note that if the AoA and AoD of a path are complementary angles satisfying $\tan(\theta_t^{u,k}) = \tan(\pi - \theta_r^{u,k})$ and $\tan(\phi_t^{u,k}) = \tan(\pi - \phi_r^{u,k})$, then this path is deemed as a LOS path; otherwise, it is deemed as a first-order NLOS path. Hence we design the reflection factor as

$$\begin{aligned} r_u^k = \text{sign}\Big[&\text{thr}_{tan} - |\tan(\theta_t^{u,k}) - \tan(\pi - \theta_r^{u,k})| \\ &- |\tan(\phi_t^{u,k}) - \tan(\pi - \phi_r^{u,k})|\Big] + 1, \end{aligned} \tag{11}$$

where thr_{tan} is the maximum LOS tolerance of the tangent values' difference between the AoA and AoD. In other words, if $|\tan(\theta_t^{u,k}) - \tan(\pi - \theta_r^{u,k})| + |\tan(\phi_t^{u,k}) - \tan(\pi - \phi_r^{u,k})| \leq \text{thr}_{tan}$, then we deem the path among $\mathcal{C}_u^k$ as a LOS path; otherwise, we deem this path as a first-order NLOS path.[3]

C. Determine Whether A Path Can Be Accurately Estimated

We determine whether the angles of a path can be accurately estimated based on whether the angles of the path matches the angles in the codebook. As shown in Fig. 4, the AoA and AoD of a path may not align precisely with the angles in the codebook, resulting in angle estimation error. We then define paths with AoA and AoD aligning precisely with the angles in the codebook as A-Class paths, and paths with either AoA or AoD not aligning precisely with the angles in the codebook as B-Class paths. From Fig. 4, the AoA and AoD estimations of A-Class paths are more accurate compared to those of B-Class paths. Hence, we believe that a path can be accurately estimated only if it is an A-Class path.

Assume $qd_k^u = (i_k^u, j_k^u, p_k^u, q_k^u)$ is the coordinate of the angles with the highest power among $\mathcal{C}_k^u$, and denote the adjacent set of qd_k^u as $\mathcal{D}_k^u$, i.e.

$$\begin{aligned} \mathcal{D}_k^u = \{&(i_k^u - 1, j_k^u, p_k^u, q_k^u), (i_k^u + 1, j_k^u, p_k^u, q_k^u), \\ &(i_k^u, j_k^u - 1, p_k^u, q_k^u), (i_k^u, j_k^u + 1, p_k^u, q_k^u), \\ &(i_k^u, j_k^u, p_k^u - 1, q_k^u), (i_k^u, j_k^u, p_k^u + 1, q_k^u), \\ &(i_k^u, j_k^u, p_k^u, q_k^u - 1), (i_k^u, j_k^u, p_k^u, q_k^u + 1)\}. \end{aligned} \tag{12}$$

Denote $\widehat{qd_k^u}$ and $\widetilde{qd_k^u}$ as the coordinates of the angles with the highest and the lowest powers among $\mathcal{D}_k^u$. In the connected domain of an A-Class path, the highest power is obviously higher than the powers of adjacent angles. Consequently, the power difference among the angles in the adjacent set is minimal, leading to a small value of the ratio $\frac{\mathbf{P}_u[\widehat{qd_k^u}]}{\mathbf{P}_u[\widetilde{qd_k^u}]}$. Conversely, in the connected domain of a B-Class path, there may be one or more angles in the adjacent set with considerably higher powers, resulting in a large value of the ratio $\frac{\mathbf{P}_u[\widehat{qd_k^u}]}{\mathbf{P}_u[\widetilde{qd_k^u}]}$. Hence, we design a power factor to indicate the existence of the A-Class path as

$$p_u^k = \text{sign}\left[\text{thr}_{pow} - \frac{\mathbf{P}_u[\widehat{qd_k^u}]}{\mathbf{P}_u[\widetilde{qd_k^u}]}\right] + 1, \tag{13}$$

where thr_{pow} refers to the maximum value of $\frac{\mathbf{P}_u[\widehat{qd_k^u}]}{\mathbf{P}_u[\widetilde{qd_k^u}]}$ for a connected domain to be considered as having an A-Class path. For multi-user environment reconstruction, excluding the B-Class paths may make point set $\mathcal{P}$ sparser but more accurate. Conversely, incorporating the B-Class paths will bring numerous unacceptable error points. Therefore, we only utilize the A-Class paths to calculate the reflection points.

Based on the connectivity factor, the reflection factor, and

[3]The introduction of thr_{tan} aims to enhance the robustness of NLOS path detection. For instance, when the AoA and AoD of a LOS path are not centered on the grid, their estimations become imperfect, resulting in non-zero values of $|\tan(\theta_t^{u,k}) - \tan(\pi - \theta_r^{u,k})|$ and $|\tan(\phi_t^{u,k}) - \tan(\pi - \phi_r^{u,k})|$. Then the path will be mistakenly identified as an NLOS path, leading to an erroneous point calculated from this path.

the power factor, we can calculate the user selection factor as

$$\begin{aligned} s_u &= (\sum_{k=1}^{n_u} s_{u,los}^k) \cdot (\sum_{k=1}^{n_u} s_{u,nlos}^k) \\ &= \left[\sum_{k=1}^{n_u} \frac{1}{8} c_u^k \cdot (2 - r_u^k) \cdot p_u^k\right] \cdot \left[\sum_{k=1}^{n_u} \frac{1}{8} c_u^k \cdot r_u^k \cdot p_u^k\right], \end{aligned} \tag{14}$$

where $\sum_{k=1}^{n_u} \frac{1}{8} c_u^k \cdot (2 - r_u^k) \cdot p_u^k$ and $\sum_{k=1}^{n_u} \frac{1}{8} c_u^k \cdot r_u^k \cdot p_u^k$ represent the number of LOS paths and NLOS paths that can be accurately calculated from $\mathbf{P}_u$. Moreover, s_u represents the number of reflection points that can be accurately calculated from $\mathbf{P}_u$. Hence, $s_u \geq 1$ indicates that the user u is a superior user; otherwise $s_u = 0$.

D. Calculate The Reflection Points

For a superior user, we first use OFDM ranging to calculate the length of the LOS path [31], [32].

In an OFDM system, signals are modulated onto subcarriers with different frequencies to facilitate transmission. For the same distance, the received signals of different subcarriers exhibit different phase shifts. During the beam sweeping stage, the received phase of the m-th subcarrier is $\frac{2\pi f_m d}{c}$ when beamforming is performed on the LOS direction, where d is the length of the LOS path, f_m is the frequency of the m-th subcarrier, and c is the speed of light. Hence, the length of the LOS path can be estimated by

$$d^\star = \arg\max_d \left| \sum_{m=1}^{M} \exp\left(\mathrm{j}\left(\varphi_m - \frac{2\pi f_m d}{c}\right)\right)\right|, \tag{15}$$

where φ_m is the phase of the m-th subcarrier measured at the receiver. Denote the location of the BS as (x_b, y_b, z_b), the elevation AoA of the LOS path as θ_{los}, and the azimuth AoA of the LOS path as ϕ_{los}. Then, location (x_u, y_u, z_u) of the user can be calculated by

$$\begin{cases} \dfrac{y_u - y_b}{x_u - x_b} = \tan(\phi_{los}), \\ \dfrac{z_u - z_b}{\sqrt{(x_u - x_b)^2 + (y_u - y_b)^2}} = \tan\left(\dfrac{\pi}{2} - \theta_{los}\right), \\ \left\| [x_u, y_u, z_u]^T - [x_b, y_b, z_b]^T \right\|_2 + \\ \left\| [x_u, y_u, z_u]^T - [x_b, y_b, z_b]^T \right\|_2 = d^\star. \end{cases} \tag{16}$$

After obtaining the location of the user, we utilize azimuth AoD $\phi_{t,nlos}$, elevation AoD $\theta_{t,nlos}$, azimuth AoA $\phi_{r,nlos}$, and elevation AoA $\theta_{r,nlos}$ of the first-order NLOS path to calculate the reflection point (x_u^p, y_u^p, z_u^p) as

$$\begin{cases} \dfrac{y_u^p - y_b}{x_u^p - x_b} = \tan(\phi_{t,nlos}), \\ \dfrac{z_u^p - z_b}{\sqrt{(x_u^p - x_b)^2 + (y_u^p - y_b)^2}} = \tan\left(\dfrac{\pi}{2} - \theta_{t,nlos}\right), \\ \dfrac{y_u^p - y_u}{x_u^p - x_u} = \tan(\phi_{r,nlos}), \\ \dfrac{z_u^p - z_u}{\sqrt{(x_u^p - x_u)^2 + (y_u^p - y_u)^2}} = \tan\left(\dfrac{\pi}{2} - \theta_{r,nlos}\right). \end{cases} \tag{17}$$

By recording the reflection points of the superior users, we obtain the point set $\mathcal{P}$.

E. Surface Fitting from The Points

However, the point set is a discrete approximation of environment reconstruction. We then smoothen the environment reconstruction result by generating statistical surfaces that approximate the points in $\mathcal{P}$. Specifically, we use the K-Means algorithm in [33] to cluster the points in $\mathcal{P}$ and then fit a surface to the points in each cluster. We propose to represent each surface by a high-order polynomial,

$$\begin{aligned} z =& c_0 + c_1 x + c_2 y + c_3 x^2 + c_4 xy + c_5 y^2 + \\ & c_6 x^3 + c_7 x^2 y + c_8 x y^2 + c_9 y^3, \end{aligned} \tag{18}$$

where $c_0, c_1, \cdots, c_9$ are the coefficients of the polynomial and (x, y, z) is the coordinate of the point in the surface. Next, we calculate $c_0, c_1, \cdots, c_9$ by the following steps.

- Denote $z_1(x, y) = c_0 + c_1 x + c_2 y$. We employ multivariate linear regression (MLR) [34] to fit the plane z_1 and obtain proper c_0, c_1, and c_3 that ensures a high degree of proximity between the points in $\mathcal{P}$ and the plane $z_1(x, y)$.
- Denote $z_2(x, y) = c_3 w_1 + c_4 w_2 + c_5 w_3 + z_1(x, y)$, where $w_1 = x^2$, $w_2 = xy$, and $w_3 = y^2$. We utilize MLR to fit $z_2(x, y)$ and obtain proper c_3, c_4, and c_5.
- Denote $z(x, y) = c_6 w_4 + c_7 w_5 + c_8 w_6 + c_9 w_7 + z_1(x, y)$, where $w_4 = x^3$, $w_5 = x^2 y$, $w_6 = xy^2$, and $w_7 = y^3$. We utilize MLR to fit $z(x, y)$ and obtain proper c_6, c_7, c_8, and c_9.

IV. Multi-Modal Fusion Based Environment Reconstruction

Although the environment reconstruction based on multi-user communications has denser points than that based on single-user communications, it still cannot support complex sensing tasks, such as autonomous driving. Assume that the users are equipped with both communications devices and cameras. We will leverage both the vision sensing and ISAC to obtain more comprehensive environment reconstruction. As shown in Fig. 5, the user first downloads the point set $\mathcal{P}$ from the BS, and then resorts to a multi-modal fusion network (MMFN) to fuse $\mathcal{P}$ and the image for depth estimation. The MMFN consists of a sensing-with-vision (SWV) module to extract the features from the image, a sensing-with-communications (SWC) module to extract the features from point set $\mathcal{P}$, and a fusion and prediction (FP) module to fuse the features and predict the depth map.

A. Sensing-with-Vision Module

We adopt the state-of-the-art Mixing Datasets for Zero-shot Cross-dataset Transfer (MiDaS) [35] as the SWV module, which has been pretrained on 10 distinct datasets to ensure high quality and great generalization.

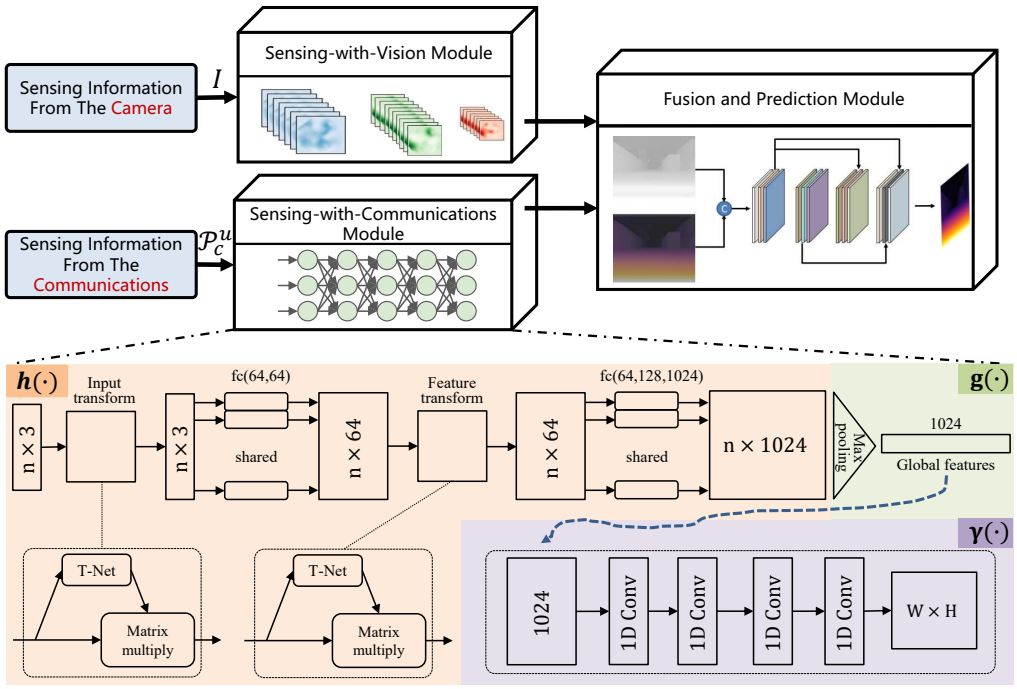


Fig. 5: The structure of the multi-modal fusion network.

B. Sensing-with-Communications Module

Note that, point set $\mathcal{P}$ provided by the BS is based on the world coordinate system while the image is based on the camera coordinate system [36]. In order to achieve better fusion results, the point set $\mathcal{P}$ and the image should be in the same coordinate system. Hence, we first convert $\mathcal{P}$ to the camera coordinate system.

Denote the rotation angles of the camera along each axis as (ϕ_x, ϕ_y, ϕ_z). Then the rotation matrices along each axis are

$$\begin{aligned}\mathbf{R}_x &= \begin{bmatrix} \cos\phi_x & -\sin\phi_x & 0 \\ \sin\phi_x & \cos\phi_x & 0 \\ 0 & 0 & 1 \end{bmatrix}, \\ \mathbf{R}_y &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi_y & \sin\phi_y \\ 0 & -\sin\phi_y & \cos\phi_y \end{bmatrix}, \\ \mathbf{R}_z &= \begin{bmatrix} \cos\phi_z & 0 & -\sin\phi_z \\ 0 & 1 & 0 \\ \sin\phi_z & 0 & \cos\phi_z \end{bmatrix}.\end{aligned} \tag{19}$$

The rotation matrix of the camera can be calculated by $\mathbf{R} = \mathbf{R}_x \cdot \mathbf{R}_y \cdot \mathbf{R}_z$.

Denote the homogeneous coordinate[4] of the reflection point in the world coordinate system as $(x_u^p, y_u^p, z_u^p, 1)$, and denote the relative displacement of the camera and the user as $\mathbf{T} = [t_x, t_y, t_z]^T$. Then the coordinate of the point in the camera coordinate system will be

$$\begin{bmatrix} x_{u,c}^p \\ y_{u,c}^p \\ z_{u,c}^p \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_u^p \\ y_u^p \\ z_u^p \\ 1 \end{bmatrix}. \tag{20}$$

After converting all points in $\mathcal{P}$ to the camera coordinate system, we record them in a new set $\mathcal{P}_c^u = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and input $\mathcal{P}_c^u$ into the SWV module. However, the point set $\mathcal{P}_c^u$ is unordered, which means that changing the order of points in $\mathcal{P}_c^u$ does not alter the information it contains. Hence, we introduce specific symmetrizations to ensure that the output features of the SWV module remain consistent regardless of the input points' order. The symmetrical function of the SWC module is designed as

$$f\left(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}\right) = \gamma(g\left(h\left(\boldsymbol{x}_1\right), \ldots, h\left(\boldsymbol{x}_n\right)\right)), \tag{21}$$

where $h(\cdot)$ represents the function of the multi-layer perceptron (MLP) shared by all points, $g(\cdot)$ represents the max pooling function, which is a symmetric function, and $\gamma(\cdot)$ represents the function of the MLP for feature extraction.

The SWC module is shown in Fig. 5. In order to enhance the network's adaptability to different input point sets, we let the points in $\mathcal{P}_c^u$ undergo an input transform process [38], which mainly relies on a T-net. The T-net generates an affine transformation matrix, which is then applied directly to the input points. Next, a shared fully-connected (FC) layer is employed to extract features for each individual point. Following the FC, a "feature transform" module is utilized to transform the extracted features into a suitable domain using another T-net. The transformed features are subsequently passed through an FC to obtain deep features. The deep

[4]Homogeneous coordinates use N+1 dimensions to represent N-dimensional coordinates to deal with geometric problems in perspective space [37]. In perspective space, two parallel lines can meet at infinity. Using homogeneous coordinates, the translation of an object can be conveniently represented by a linear transformation.

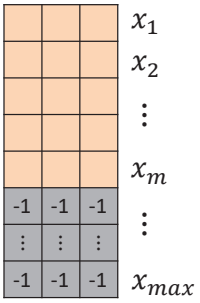


Fig. 6: The mask method for the input of the communications sensing module.

features are passed into a max pooling function to generate the global features. Then the global features are fed into four 1D CNN layers. The output of the final CNN is reshaped to $\boldsymbol{F}_{cs} \in \mathbb{R}^{W \times H}$.

C. Fusion and Prediction Module

The structure of the FP module is shown in Fig. 5. Denote the output the SWV module as $\boldsymbol{D}_p \in \mathbb{R}^{W \times H}$. We combine $\boldsymbol{D}_p$ and $\boldsymbol{F}_{cs}$ as $\boldsymbol{F} = \text{cat}(\boldsymbol{F}_{cs}, \boldsymbol{D}_p)$. Then $\boldsymbol{F}$ is input into several CNNs with skip connections. The final output of the fusion module is the predicted depth map $\hat{\boldsymbol{D}}$.

We utilize the root mean-squared error (RMSE) of the predicted depth for all pixels as the loss function

$$\mathcal{L} = \frac{1}{N}\sqrt{\frac{1}{W \times H}\sum_{d_i \in \boldsymbol{D}, \hat{d}_i \in \hat{\boldsymbol{D}}} \left| d_i - \hat{d}_i \right|^2}, \tag{22}$$

where $\boldsymbol{D}$ is the ground truth of the depth map and N is the size of the dataset.

D. Meta-Learning based Training Strategy for Any User Quantity

For MMFN, if the quantity of input points changes, then we have to train a new network; otherwise, the performance of MMFN may suffer significant degradation. To generalize the network across different user quantities, we design a meta-learning-based[5] training strategy. Specifically, we denote the depth estimation through the fusion of images and N_m^u users' points as task-m, where $N_m^u \in \{N_1^u, N_2^u, \cdots, N_M^u\}$ is the possible user quantity. Then we train MMFN for these tasks sequentially, where the initial parameters of task-$(m+1)$ is the trained parameters of task-m.

However, when the quantity of input points varies, the input dimension of the communications-sensing module also changes. We then design an input masking method illustrated in Fig. 6 to ensure the immutability of the input dimension. Assume that the maximum number of users is N_{max}. Then we generate a matrix $\boldsymbol{P}_m \in R^{N_{max} \times 3}$ as the input for task-m whose first m rows is the m points' coordinates in the camera coordinate system and the last $(N_{max} - m)$ rows are masked as -1. The meta-learning based training strategy is illustrated in Algorithm 1.

[5]Meta-learning [39] is a machine learning approach that focuses on developing algorithms or models capable of learning and adapting to new tasks or environments quickly. The training methodology of meta-learning involves training a model for a specific number of epochs on a given task and utilizing the parameters of that model as the initial values for training another task.

Algorithm 1 Meta Learning Based Training Strategy for Depth Estimation

Require: Training dataset $\mathcal{D}$, number of iterations n_{iter}, number of iterations of each task n_m, initialized trainable parameters of the SWC module $\boldsymbol{\Theta}_c$, initialized trainable parameters of the FP module $\boldsymbol{\Theta}_f$, trained parameters of MiDaS $\boldsymbol{\Theta}_M$, and the maximum number of users N_{max}.
Ensure: Trained parameters of SWC module $\boldsymbol{\Theta}_c$ and trained parameters of the FP module $\boldsymbol{\Theta}_f$.
for k = 1 to n_{iter} **do**
  for n = 1 to n_m **do**
    - Draw mini-batch $\mathcal{D}_k$: a random subset of $\mathcal{D}_m$
    - Prepare the input of MiDaS $\boldsymbol{I}$
    - Generate the input of the SWC module $\boldsymbol{P}_c^m$ by masking $(N_m ax - m)$ rows
    - Estimate the dense depth $\boldsymbol{D}_i$ by MiDaS and the image $\boldsymbol{I}$
    - Estimate the domain map $\boldsymbol{F}_{dm}$ by the SWC module and the point cloud $\boldsymbol{P}_c^m$
    - Calculate the output depth by fusing $\boldsymbol{D}_i$ and $\boldsymbol{F}_{dm}$ based on fusion module
    - Compute the loss: $\mathcal{L}$
    Back-propagation Phase:
    - Use Adam optimizer to update $\boldsymbol{\Theta_c}$ and $\boldsymbol{\Theta_f}$
  end for
end for

E. Evaluation Metrics of Depth Estimation

The evaluation metrics of depth estimation includes the following.

- Root Mean-Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{d_i \in \boldsymbol{D}, \hat{d}_i \in \hat{\boldsymbol{D}}} \left| d_i - \hat{d}_i \right|^2}.$$

- Root Mean-Squared Error logscale (RMSElog):

$$\text{RMSElog} = \sqrt{\frac{1}{N}\sum_{d_i \in \boldsymbol{D}, \hat{d}_i \in \hat{\boldsymbol{D}}} \left| \log d_i - \log \hat{d}_i \right|^2}.$$

- Mean-Absolute Error (MAE):

$$\text{MAE} = \sqrt{\frac{1}{N}\sum_{d_i \in \boldsymbol{D}, \hat{d}_i \in \hat{\boldsymbol{D}}} \left| d_i - \hat{d}_i \right|}.$$

- Root Mean-Squared Error logscale (MAElog):

$$\text{MAElog} = \sqrt{\frac{1}{N}\sum_{d_i \in \boldsymbol{D}, \hat{d}_i \in \hat{\boldsymbol{D}}} \left| \log d_i - \log \hat{d}_i \right|}.$$

TABLE I: Environment Reconstruction under Different User Selection Factors

Factor	Without Selection	Connectivity	Reflection	Power	Connectivity + Reflection	Connectivity + Power	Reflection + Power	ALL
RMSE	0.9233	0.8535	0.5141	0.4351	0.2266	0.1584	0.4026	**0.0556**

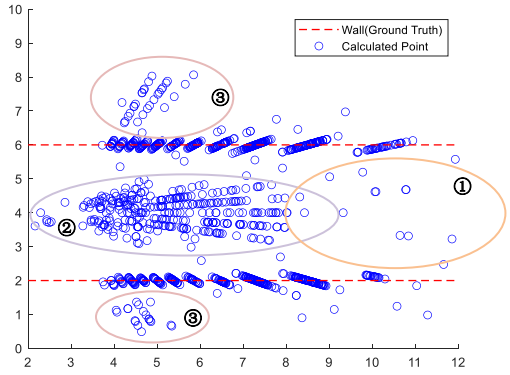


Fig. 7: The calculated points without sensing user selection

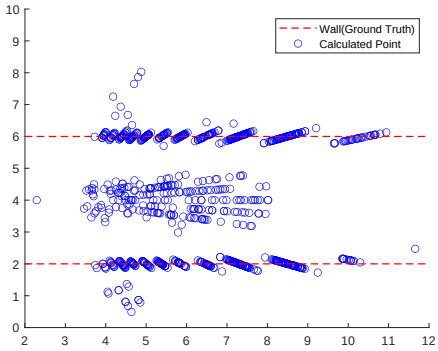


Fig. 8: The environment reconstruction based on the connectivity factor.

- Absolute Relative Error (AbsRel):

$$\text{AbsRel} = \frac{1}{N} \sum_{d_i \in \boldsymbol{D}, \hat{d}_i \in \hat{\boldsymbol{D}}} \frac{\left| d_i - \hat{d}_i \right|}{d_i}.$$

- Square-Relative Error (SqRel):

$$\text{SqRel} = \frac{1}{N} \sum_{d_i \in \boldsymbol{D}, \hat{d}_i \in \hat{\boldsymbol{D}}} \frac{\left| d_i - \hat{d}_i \right|^2}{d_i}.$$

- δ_n threshold:

$$\delta_n = \left| \left\{ \hat{d}_i : \max_{d_i \in \boldsymbol{D}, \hat{d}_i \in \hat{\boldsymbol{D}}} \left(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i} \right) < 1.25^n \right\} \right| / |\boldsymbol{D}|.$$

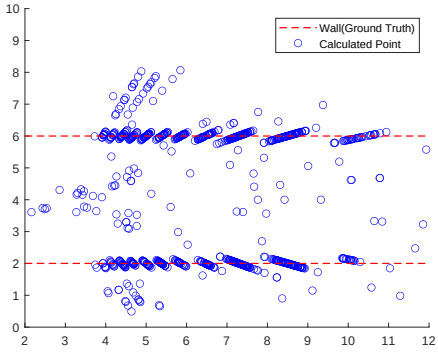


Fig. 9: The environment reconstruction based on the reflection factor.

V. Simulation Results

In this section, we generate the dataset and evaluate the performance of the proposed multi-user selection and multi-modal fusion based environment reconstruction.

A. Dataset Generation

We consider a communications scenario involving numerous pedestrians equipped with communications devices as well as cars equipped with both communications devices and cameras. We use CARLA [40] to build the simulation scenario and utilize the cameras in the scenario to capture the images.[6] Next, we use the 3D ray-tracing package "propagationModel" of MATLAB [41] to calculate the wireless parameters. The communication frequency is set to 28 GHz and the bandwidth is set to 40 MHz. The BS is equipped with a uniform planar array (UPA) of 8×8 antennas and the user is equipped with a UPA of 4×4 antennas.

Without user selection, the points are calculated and presented in Fig. 7, which are plagued by numerous errors. Then we test the impact of the connectivity factor, the reflection factor, and the power factor on environment reconstruction. By separately incorporating these three factors, the environment reconstruction results are shown in Fig. 8, Fig. 9, and Fig. 10 respectively. Note that the goal of the connectivity factor is to eliminate noise and high-order NLOS paths. From Fig. 8, many irregular noise points in circle ① of Fig. 7 can be eliminated after introducing the connectivity factor. The purpose of the reflection factor is to avoid misidentifying the LOS path as a first-order NLOS path. From Fig. 9, a large number of erroneous points between the BS and the user as shown in circle ② of Fig. 7 can be eliminated after introducing the reflection factor. The objective of the power factor is to ensure that the selected user has an A-

[6]CARLA has been meticulously designed to streamline the development, training, and validation of autonomous driving systems. In addition to providing open-source code and protocols, CARLA offers a wealth of freely accessible open digital assets, including urban layouts, buildings, and vehicles, all tailored to this domain. CARLA empowers users with the ability to customize sensor suites and environmental conditions to suit their specific needs, while also enabling full control over both static and dynamic actors. Furthermore, CARLA boasts functionalities such as map generation, facilitating extensive experimentation and thorough testing of autonomous driving systems.

TABLE II: Depth estimation metrics under different weather conditions

Weather	Method	RMSE ↓	RMSE_{log} ↓	MAE ↓	MAE_{log} ↓	AbsRel ↓	SqRel ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Sunny	image	5.281	0.226	1.289	0.413	0.173	12.828	0.692	0.960	0.996
	fusion	**1.874**	0.091	0.895	0.243	0.062	0.187	0.976	0.996	0.998
Rainy	image	10.315	0.437	2.054	0.538	0.281	15.557	0.554	0.815	0.906
	fusion	**3.564**	0.176	1.355	0.362	0.126	0.524	0.812	0.977	0.994
Snowy	image	18.339	0.981	3.160	0.892	0.675	26.220	0.150	0.337	0.497
	fusion	**4.802**	0.230	1.508	0.403	0.178	1.244	0.751	0.937	0.978

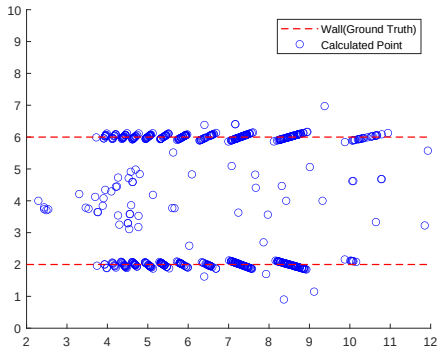


Fig. 10: The environment reconstruction based on the power factor.

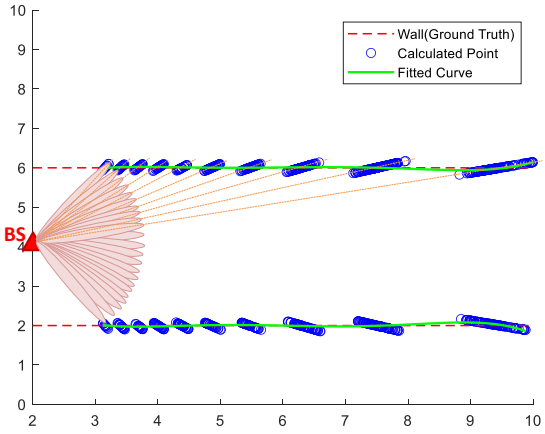


Fig. 11: The environment reconstruction based on the proposed user selection algorithm.

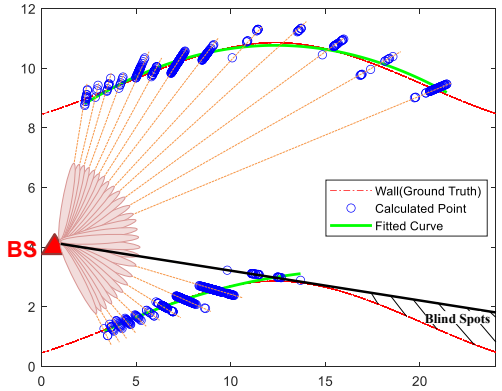


Fig. 12: The environment reconstruction of irregular walls based on the proposed user selection algorithm.

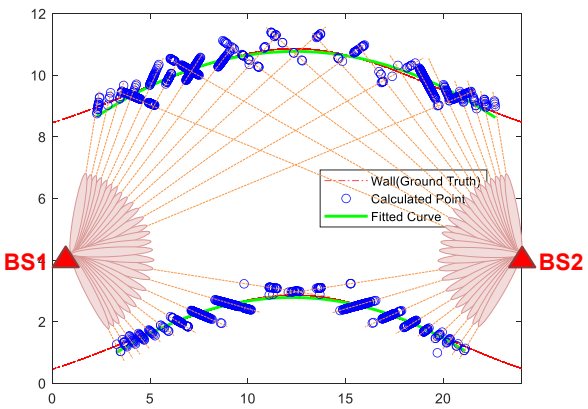


Fig. 13: The environment reconstruction of irregular walls based on two BSs.

class path, which can potentially result in highly accurate environment reconstruction. As depicted in Fig. 10, the power factor effectively eliminates numerous error points near the ground truth points in circle ③ of Fig. 7.

The environment reconstruction result based on all three factors is shown in Fig. 11. An interesting phenomenon is that the calculated points all fall in the directions of the transmit beams and are clustered in distribution. This is because that the points lie between these two clusters belong to the first-order NLOS paths whose angles do not align with the angles of the transmit beams, and the user selection algorithm excludes the first-order NLOS paths. The RMSE under different user selection factors is shown in Tabel I. It can be seen that the RMSE of environment reconstruction without user selection is 0.9233. However, by employing the proposed user selection method, the RMSE decreases to 0.0556, resulting in a significant improvement in sensing accuracy.

In addition to the reconstruction of regular walls, as proposed in [14], [42], we next attempt to reconstruct irregular walls. However, in the environment with irregular walls, there may be blind spots when using one single BS for sensing. As shown in Fig. 12, the wall in the bottom right corner cannot be sensed because there is no first-order NLOS paths with reflection points located in the blind spots. To eliminate the blind spots and obtain a complete environment reconstruction result, we utilize two BSs that are placed on the left and right sides respectively for the sensing of the irregular walls. Each BS first utilizes the proposed multi-user selection algorithm to select the superior users and generate the point set. Since

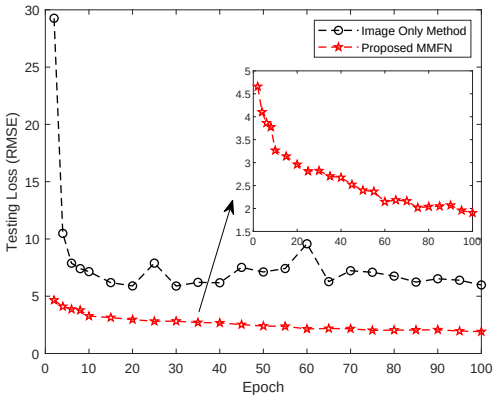


Fig. 14: Testing loss of the proposed MMFN and the image only method.

Fig. 15: The ground truth of the depth map.

the point set obtained from each BS is relatively accurate, we then directly take the union of the two point sets from the two BSs and fit the curves as shown in Fig. 13. It can be seen that the multi-user selection based environment reconstruction can effectively reconstruct irregular walls as well. Moreover, the calculated points from each BS all fall in the directions of the transmit beams from the BS, which is similar to the phenomenon in Fig. 11.

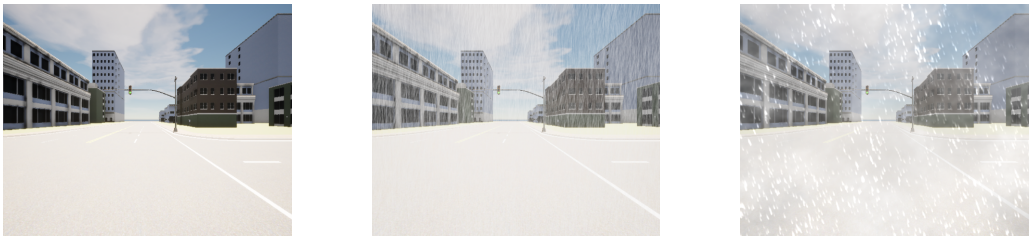
Next, we evaluate the performance of the proposed multi-modal fusion based environment reconstruction under different weather conditions. Specifically, we consider fusing the image with the BS's point set by the proposed MMFN to accurately estimate the depth of each pixel within the image. We train the MMFN on the sunny dataset. The testing RMSE during the training process is shown in Fig. 14. The image-only method yields an RMSE of 5.281, while the integration of communications sensing information and visual sensing information achieves an RMSE of 1.874. We further examine the robustness of the MMFN under rainy and snowy weather conditions. The ground truth of the depth map is presented in Fig. 15. In the depth map, the varying shades of color represent the proximity of objects, where darker colors indicate closer distances and lighter colors represent farther distances. For the image-only method, the predicted depth map in a sunny environment is displayed in Fig. 16(d), the predicted depth map in a rainy environment is displayed in Fig. 16(e), and the predicted depth map in a snowy environment is displayed in Fig. 16(f). With the proposed MMFN, the predicted depth map in a sunny environment is displayed in Fig. 16(g), the predicted depth map in a rainy environment is displayed in Fig. 16(h), and the predicted depth map in a snowy environment is displayed in Fig. 16(i). It can be seen that the image-only method is comparable to the MMFN in sunny conditions. However, in rainy and snowy conditions, the depth estimation performance of the image-only method is relatively poor, whereas the proposed MMFN still maintains a high accuracy. Therefore, the performance of the image-only method and the proposed MMFN across different evaluation metrics is presented in TABLE II, where a downward arrow indicates smaller values are better while an upward arrow indicates larger values are better. In rainy conditions, the proposed MMFN achieves a depth estimation RMSE of 3.564, which is approximately 2.9 times better than the image-only method. Similarly, in snowy conditions, the proposed MMFN achieves an RMSE of 4.802, which is approximately 3.8 times better than the image only method. The proposed MMFN demonstrates strong robustness to different weather conditions.

To answer how many communications users are required for environment reconstruction, we next test the accuracy of depth estimation with different numbers of users. In Fig. 17, the dashed line represents the MMFN trained using the traditional training strategy while the solid line represents the MMFN trained using the proposed meta-learning based strategy. For the traditional training approach, the maximum user number is set to $N_{max} = 80$. During the training process, the input of the MMFN consists of the sensing data from N_{max} communications users. However, during testing, when u_t users' sensing data is used, the remaining $(N_{max} - ut)$ rows of the input are filled with $(0, 0, 1)$[7]. It can be seen from Fig. 17 that the depth estimation error gradually decreases with the increasing of user number, which also demonstrates that the clustering characteristics of the communication system can bring significant benefits to ISAC. Moreover, the traditional training approach requires an increase in the number of users to 50 before the performance reaches a satisfactory level. However, with the proposed meta-learning based training strategy, the performance converges to a satisfactory level after the number of users reaches 10. The proposed meta-learning based training strategy facilitates the activation of performance leaps, making it easier to trigger significant improvements as the number of users increases.

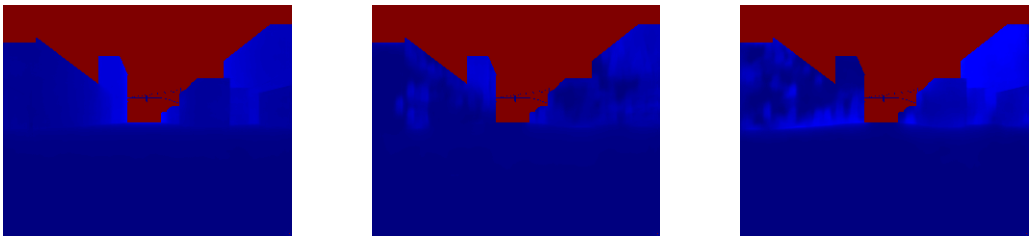
VI. Conclusions

In this paper, we propose a multi-user based environment reconstruction scheme, where the BS collects the beam scanning information of the ubiquitous users to compute the environment point set. Moreover, we propose an evaluation criterion for sensing users and use this criterion to select users who can yield accurate reflection points. The point set is then

[7]The filled row is $(0, 0, 1)$, where $(0, 0)$ represent the coordinate of the input image and the value 1 represents the depth of the coordinate $(0, 0)$ is the farthest. In other words, the filled $(0, 0, 1)$ represents the depth information of the top-left corner of the image, indicating the farthest distance (sky) information.



(a) Image of a sunny day (b) Image of a rainy day (c) Image of a snowy day



(d) Depth estimation by the image-only at a sunny day (e) Depth estimation by the image-only at a rainy day (f) Depth estimation by the image-only at a snowy day

(g) Depth estimation by the proposed MMFN at a sunny day (h) Depth estimation by the proposed MMFN at a rainy day (i) Depth estimation by the proposed MMFN at a snowy day

Fig. 16: The depth estimation under different weather conditions.

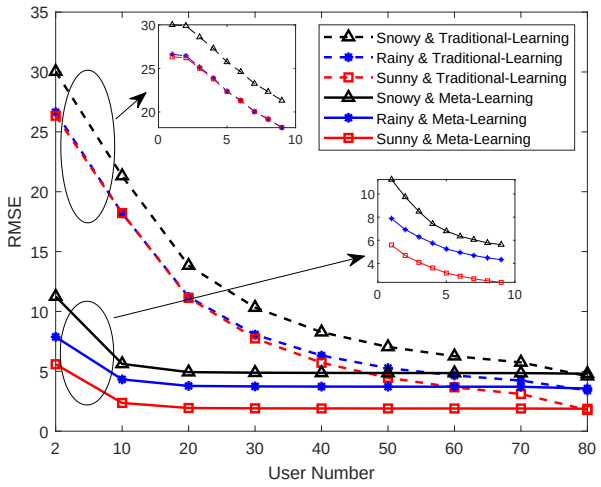


Fig. 17: The RMSE when different users participate in the sensing task.

distributed to users with sensing requests. Each user fuses the received point set with their own images by the proposed MMFN to achieve a more dense and comprehensive depth map. Based on the user selection, we attained an impressive 16-fold enhancement in the precision of point set estimation. Additionally, The proposed MMFN significantly enhances the accuracy of environment reconstruction under challenging weather conditions like raining and snowing. Moreover, we propose a meta-learning-based training approach that enables the network to be effective under any number of users, which greatly improves the deployability and feasibility of MMFN.

REFERENCES

[1] W. Xu, F. Gao, X. Tao, J. Zhang, and A. Alkhateeb, "Computer vision aided mmWave beam alignment in V2X communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2699–2714, Apr. 2023.

[2] B. Lin, F. Gao, Y. Zhang, C. Pan, and G. Liu, "Multi-camera views based beam searching and bs selection with reduced training overhead," *IEEE Trans. Commun.*, 2024.

[3] J. Chen, F. Gao, X. Tao, G. Liu, C. Pan, and A. Alkhateeb, "Computer vision aided codebook design for MIMO communications systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 5, pp. 1558–2248, May. 2023.

[4] G. Charan, M. Alrabeiah, and A. Alkhateeb, "Vision-aided dynamic blockage prediction for 6g wireless communication networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*. IEEE, Jun. 2021, pp. 1–6.

[5] Z. Qin, F. Gao, B. Lin, X. Tao, G. Liu, and C. Pan, "A generalized semantic communication system: from sources to channels," *IEEE Wireless Commun.*, vol. 30, no. 3, pp. 18-26, Jun. 2023.

[6] Z. Zhang, Y. Zhang, J. Zhang, and F. Gao, "Adversarial training-aided time-varying channel prediction for TDD/FDD systems," *China Commun.*, vol. 20, no. 6, pp. 100-115, Jun. 2023.

[7] Z. Wei, F. Liu, C. Masouros, N. Su, and A. P. Petropulu, "Toward multi-functional 6G wireless networks: Integrating sensing, communication, and security," *IEEE Commun. Mag.*, vol. 60, no. 4, pp. 65–71, Apr. 2022.

[8] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 6, pp. 1728–1767, Jun. 2022.

[9] H. Luo, Y. Wang, J. Zhao, H. Wu, S. Ma, and F. Gao, "Integrated sensing and communications in clutter environment," *arXiv preprint arXiv:2311.01674*, 2023.

[10] K. Chen, C. Qi, O. A. Dobre, and G. Y. Li, "Simultaneous beam training and target sensing in isac systems with RIS," *IEEE Trans. Wireless Commun.*, 2023.

[11] D. Luo, H. Wu, H. Luo, B. Lin, and F. Gao, "Moving target sensing for ISAC systems in clutter environment," *arXiv preprint arXiv:2311.01700*, 2023.

[12] Y. Xiang, Y. Gao, X. Yang, S. Kang, and M. Shao, "An esprit-based moving target sensing method for MIMO-OFDM ISAC systems," *IEEE Commun. Lett.*, 2023.

[13] E. Leitinger, F. Meyer, F. Hlawatsch, K. Witrisal, F. Tufvesson, and M. Z. Win, "A belief propagation algorithm for multipath-based SLAM," *IEEE Trans. Wireless Commun.*, vol. 18, no. 12, pp. 5613-5629, Dec. 2019.

[14] Q. Wang, X. Yuan, C. Xu, and X. Wang, "A bayesian approach to communication-driven SLAM based on diffuse reflection model," *IEEE Wireless Commun. Lett.*, vol. 12, no. 7, pp. 1279-1283, July. 2023.

[15] J. Yang, C.-K. Wen, J. Xu, H. Que, H. Wei, and S. Jin, "Angle-based SLAM on 5G mmwave systems: Design, implementation, and measurement," *IEEE Internet Things J.*, vol. 10, no. 20, pp. 17755-17771, Oct. 2023.

[16] H. Kim, K. Granström, L. Gao, G. Battistelli, S. Kim, and H. Wymeersch, "5G mmwave cooperative positioning and mapping using multi-model phd filter and map fusion," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3782–3795, Jun. 2020.

[17] Z. Sun, Z. Yu, B. Guo, B. Yang, Y. Zhang, and D. W. K. Ng, "Integrated sensing and communication for effective multi-agent cooperation systems," *IEEE Commun. Mag.*, 2024.

[18] X. Chu, Z. Lu, D. Gesbert, L. Wang, and X. Wen, "Vehicle localization via cooperative channel mapping," *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 5719–5733, Jun. 2021.

[19] J. Yang, C.-K. Wen, S. Jin, and X. Li, "Enabling plug-and-play and crowdsourcing slam in wireless communication systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1453–1468, Mar. 2021.

[20] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3D object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pp. 1527–1536, 2021.

[21] U. Niesen and J. Unnikrishnan, "Camera-radar fusion for 3-D depth reconstruction," in *Proc., 2020 IEEE Intelligent Vehicles Symp. (IV)*, Las Vegas, NV, USA, 2020, pp. 265-271.

[22] G. El Natour, O. A. Aider, R. Rouveure, F. Berry, and P. Faure, "Radar and vision sensors calibration for outdoor 3D reconstruction," in *Proc. IEEE Int. Conf. Robot. Autom.*, Seattle, WA, USA, 2015, pp. 2084–2089.

[23] F. Yang, L. Zhan, H. Wei, L. Chengyu, and D. Shi, "Mmwave radar and vision fusion for semantic 3D reconstruction," in *2023 IEEE 7th ISEMC*, Hangzhou, China, 2023, pp. 1-4.

[24] J.-T. Lin, D. Dai, and L. Van Gool, "Depth estimation from monocular images and sparse radar data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Las Vegas, NV, USA, 2020, pp. 10233-10240.

[25] K. Zhang, Z. Li, W. Yuan, Y. Cai, and F. Gao, "Radar sensing via OTFS signaling," *China Commun.*, vol. 20, no. 9, pp. 34–45, Sep. 2023.

[26] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

[27] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "NR: The new 5G radio access technology," *IEEE Commun. Standards Mag.*, vol. 1, no. 4, pp. 24–30, Dec. 2017.

[28] Z. Mou and F. Gao, "Millimeter wave wireless communication assisted three-dimensional simultaneous localization and mapping," *arXiv preprint arXiv:2303.02617*, 2023.

[29] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[30] B. H. Fleury, "First-and second-order characterization of direction dispersion and space selectivity in the radio channel," *IEEE Trans. Inf. Theory*, vol. 46, no. 6, pp. 2027–2044, Sep. 2000.

[31] Y. Sun, J. Li, T. Zhang, R. Wang, X. Peng, X. Han, and H. Tan, "An indoor environment sensing and localization system via mmwave phased array," *J. Commun. Inf. Netw*, vol. 7, no. 4, pp. 383–393, Dec. 2022.

[32] H. Luo, F. Gao, W. Yuan, and S. Zhang, "Beam squint assisted user localization in near-field integrated sensing and communications systems," *IEEE Trans. on Wireless Commun.*, 2023.

[33] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, p. 1295, 2020.

[34] N. Draper and H. Smith, "Wiley series in probability and statistics. 1998," *Applied Regression Analysis*.

[35] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2020.

[36] R. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. Robot. Autom.*, vol. 3, no. 4, pp. 323–344, Aug. 1987.

[37] E.A. Maxwell, "The methods of plane projective geometry based on the use of general homogeneous coordinates," *Cambridge Univ. Press*, 1946.

[38] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3D classification and segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, July 2017.

[39] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[40] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Proc. 1st Annu. Conf. Robot Learn.*, 2017, pp. 1–16.

[41] L. Model, "MATLAB: Create rf propagation model - matlab propagationmodel," https://ww2.mathworks.cn/help/antenna/ref/propagationmodel.html, 2023.

[42] J. Yang, C.-K. Wen, and S. Jin, "Hybrid active and passive sensing for slam in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 7, pp. 2146–2163, July. 2022.