MULTIMODAL SENSING-AIDED BEAMFORMING OPTIMIZATION FOR OFDM SYSTEMS

Yinghan Li and Wei Yu

Department of Electrical and Computer Engineering
University of Toronto, Toronto, ON, Canada, M5S 3G4

ABSTRACT

This paper proposes a multimodal learning method that leverages camera images, LiDAR points, and RF pilots to optimize beamforming in orthogonal frequency division multiplexing (OFDM) systems under non-line-of-sight (NLoS) and dynamically changing environments. Existing image-based and LiDAR-based methods typically rely on estimating the user location for beamforming optimization, and are limited to line-of-sight and static environments. In contrast, dynamic NLoS scenarios require modeling the influence of obstacles that serve as reflecting surfaces. This paper focuses on these dynamic NLoS scenarios. We propose integrating LiDAR data and camera images with pilots to capture spatial and material information not only about the users but also about surrounding obstacles. The implicit NLoS channel characteristics embedded in spatial and material information can then be utilized to optimize beamforming in dynamic NLoS scenarios. Specifically, we propose a learning-based framework that utilizes a keypoint detection network to extract two-dimensional spatial and material information from images. The information from images then guides the selection of LiDAR points to precise three-dimensional geometric information. Finally, a graph neural network (GNN) integrates the processed image features, selected LiDAR data, and pilots to optimize beamforming vectors. Simulation shows that the proposed method achieves higher spectral efficiency than pilot-only methods.

1. INTRODUCTION

Modern wireless communication systems widely use orthogonal frequency division multiplexing (OFDM) to maintain robust communication performance [1–3]. In OFDM system, the frequency band is divided into multiple orthogonal subcarriers, which mitigates inter-symbol interference and improves spectrum efficiency. However, designing beamforming in OFDM system is challenging, as optimizing beamforming across the subcarriers requires accurate channel state information (CSI), which in turn requires a large number of pilots.

Pilot-based OFDM beamforming methods have been extensively studied in the literature [4, 5]. For the subcarriers with pilots, their CSI is typically estimated using least-squares (LS) or minimum mean-square error (MMSE) methods [3]. The CSI of the remaining subcarriers is then reconstructed using interpolation or filtering techniques, such as linear, spline, or Wiener interpolation [3, 6]. While effective with dense pilot placement, these methods suffer performance degradation in rich multipath channels, where simple interpolation often fails to capture the variations of the frequency response accurately.

Recent advances in sensing technologies have enabled the acquisition of spatial information about the propagation environment, which can implicitly provide CSI beyond what is available from the pilots. For exmaple, camera images have been used to improve beamforming optimization [7–9], and LiDAR sensors have similarly been used for beam selection in mmWave systems [10]. These approaches demonstrate the potential of incorporating sensing modalities beyond pilots into wireless optimization. However, most existing methods assume static line-of-sight (LoS) conditions, where user equipment (UE) location is considered the dominant factor in determining the channel and corresponding beamforming vector. In more practical non-line-of-sight (NLoS) scenarios with dynamic obstacles, UE location alone is insufficient to characterize the channel, and methods built on this assumption have limited applicability. This motivates the joint use of pilots, images, and LiDAR data for beamforming optimization, but to the best of authors' knowledge, this has not been explored in prior works.

This paper proposes a multimodal beamforming framework for OFDM systems under NLoS and dynamically changing environments. Unlike prior works that optimize beamforming based on user location under static LoS conditions, our method leverages both cameras and LiDAR to capture 3D spatial and material information about UEs and obstacles that act as reflecting surfaces. The points of reflection offer reliable information about the propagation paths, which are particularly valuable in OFDM systems for characterizing CSI of subcarriers. Therefore, the proposed multimodal method has the potential to significantly improve beamforming performance in dynamic NLoS environments.

Specifically, in the proposed method, a detection network first extracts two-dimensional (2D) spatial and material information of UEs and obstacles from camera images. The information from the camera then guides the selection of LiDAR points to construct a more concise three-dimensional (3D) description of the environment. Finally, a graph neural network (GNN) processes the selected LiDAR data, image-derived spatial and material information, and pilots to optimize beamforming. The proposed method is able to achieve higher performance as compared to pilot-based methods.

2. SYSTEM MODEL

We consider a downlink OFDM system where the base station (BS) employs a planar antenna array with M elements and the UE has a single antenna. A camera and a LiDAR sensor are deployed to capture visual and spatial information about the environment. Let N denote the number of subcarriers, and define the channel matrix as $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N] \in \mathbb{C}^{M\times N}$, where $\mathbf{h}_n \in \mathbb{C}^M$ denotes the channel vector between the M BS antennas and the UE on the n-th subcarrier. The channel on subcarrier n is modeled as:

$$\mathbf{h}_n = \sum\nolimits_{\ell=1}^{L} \mathbf{a}_\ell e^{-j2\pi \tilde{f}_n \tau_\ell} \tag{1}$$

where $\mathbf{a}_\ell \in \mathbb{C}^M$ and τ_ℓ denote the per-antenna complex gain vector and propagation delay of path ℓ, respectively, and $\tilde{f}_n$ is the baseband frequency of subcarrier n. We assume perfect synchronization and

a sufficiently long cyclic prefix (CP), and neglect inter-symbol and inter-carrier interference in this paper. Let s_n denote the symbol transmitted on subcarrier n, and let $\mathbf{w}_n \in \mathbb{C}^M$ be the beamforming vector on subcarrier n. The vector $\mathbf{w}_n$ is subject to the per-subcarrier power constraint $\|\mathbf{w}_n\|_2^2 \leq P$, where P is the maximum transmit power. The received signal at the UE on subcarrier n is

$$r_n = \mathbf{h}_n^H \mathbf{w}_n \, s_n + z_n, \tag{2}$$

where $z_n \sim \mathcal{CN}(0, \sigma^2)$ denotes the additive white Gaussian noise (AWGN). The downlink aggregate spectral efficiency across all N subcarriers is

$$R_{\text{sum}} = \sum\nolimits_{n=1}^{N} \log_2 \left(1 + \frac{|\mathbf{h}_n^H \mathbf{w}_n|^2}{\sigma^2}\right). \tag{3}$$

From (3), we observe that the spectral efficiency depends on the $\mathbf{h}_n^H \mathbf{w}_n$ on each subcarrier. Therefore, it is essential to obtain CSI before optimizing the beamforming. In this paper, camera images, LiDAR points, and pilots are utilized as the data from which we extract channel-related information.

2.1. Camera Image Processing

We first utilize camera images to extract material and 2D spatial information of the environment. Let $\mathbf{I} \in \mathbb{R}^{S_x \times S_y \times 3}$ denote the image, where S_x and S_y are the width and height in pixels, and the third dimension corresponds to the RGB channels. From camera image $\mathbf{I}$, a detection function $\mathcal{G}^{\text{img}}(\cdot)$ extracts bounding boxes $\mathbf{B}$, keypoints $\mathbf{K}$, and material labels $\mathbf{M}$ as

$$(\mathbf{B},\, \mathbf{K},\, \mathbf{M}) = \mathcal{G}^{\text{img}}(\mathbf{I}), \tag{4}$$

where $\mathbf{B} \in \mathbb{R}^{\hat{N} \times 4}$ denotes the bounding boxes of the detected objects, $\mathbf{K} \in \mathbb{R}^{\hat{N} \times 8}$ represents the associated keypoints, and $\mathbf{M} \in \mathbb{R}^{\hat{N}}$ represents the material labels. Here, $\hat{N}$ is the maximum number of detectable objects in the image. When the number of detected objects is smaller than $\hat{N}$, the missing entries are padded with -1.

2.2. LiDAR Point Processing

Next, we utilize LiDAR points to extract 3D spatial information of the UE and obstacles. Let $\mathbf{L} \in \mathbb{R}^{N^{\text{L}} \times 3}$ represent the set of LiDAR points collected at the sensor, where N^{L} denotes the total number of raw points. Since raw LiDAR data contain redundant points, bounding boxes $\mathbf{B}$ and material labels $\mathbf{M}$ obtained from camera images are used to select relevant points from $\mathbf{L}$ and assign each point with a material label. The selected point set $\tilde{\mathcal{L}}$ is obtained through the LiDAR selection function $\mathcal{S}^{\text{lidar}}(\cdot)$ as

$$\tilde{\mathcal{L}} = \mathcal{S}^{\text{lidar}}(\mathbf{L}, \mathbf{B}, \mathbf{M}), \tag{5}$$

where $\tilde{\mathcal{L}}$ represents the selected LiDAR points with their assigned material label. The i-th element of $\tilde{\mathcal{L}}$ can be expressed as $\tilde{\mathcal{L}}_i = (\mathbf{L}_i, m_i)$, where $\mathbf{L}_i \in \mathbb{R}^3$ denotes the 3D coordinates of a selected LiDAR point and $m_i \in \mathbf{M}$ is the associated material label. The selected set $\tilde{\mathcal{L}}$ provides geometric and material information of the 3D environment, which contains channel characteristics because the positions and orientations of obstacles strongly influence wireless signal reflections.

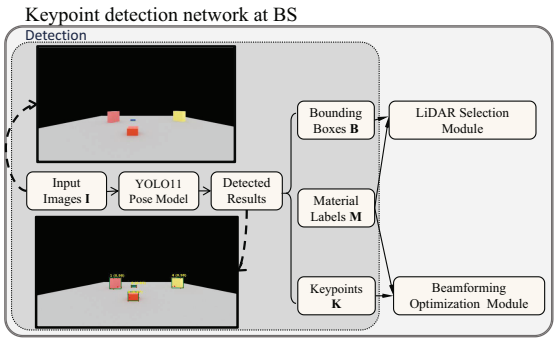


Fig. 1. Architecture of the keypoint detection network.

2.3. Pilot-Based RF Information

In addition to visual and LiDAR sensing, we also utilize uplink pilots to provide direct RF-domain information about the channel. We consider a time-division duplex (TDD) system in which the downlink CSI can be inferred from uplink pilots. We adopt a comb-type pilot arrangement, where N_p pilot symbols are uniformly inserted into the N subcarriers with spacing $L = N/N_p$, because the channel across the subcarriers are highly correlated. Let $\mathbf{x} = [x_0, x_1, \ldots, x_{N-1}]^\top \in \mathbb{C}^N$ denote the uplink frequency-domain transmit vector. At subcarrier indices $n = mL$, $m = 0, 1, \ldots, N_p - 1$, the elements x_{mL} correspond to the known pilot symbols. The non-pilot subcarriers are assumed to be zero. The received pilot on subcarrier n is given by

$$\mathbf{y}_n = \mathbf{h}_n \, x_n + \mathbf{z}_n, \tag{6}$$

where $\mathbf{z}_n \sim \mathcal{CN}(\mathbf{0}, \hat{\sigma}^2 \mathbf{I}_M)$ is additive white Gaussian noise.

2.4. Multimodal Beamforming Optimization

Traditional methods first estimate the channel from the pilots and then optimize beamforming based on the estimated CSI. In contrast, we bypass explicit channel estimation and directly determine the beamformers from the uplink pilots and multimodal features as $\{\mathbf{w}_n\}_{n=1}^N = \mathcal{F}(\tilde{\mathcal{L}}, \mathbf{K}, \mathbf{M}, \{\mathbf{y}_n\}_{n=1}^N)$. Note that the beamformers are not restricted to some codebook. In this case, the multimodal optimization problem is formulated as

$$\underset{\{\mathbf{w}_n\}_{n=1}^N}{\text{maximize}} \quad R_{\text{sum}}, \tag{7a}$$

$$\text{subject to} \quad \|\mathbf{w}_n\|_2^2 \leq P \tag{7b}$$

$$\{\mathbf{w}_n\}_{n=1}^N = \mathcal{F}(\tilde{\mathcal{L}}, \mathbf{K}, \mathbf{M}, \{\mathbf{y}_n\}_{n=1}^N). \tag{7c}$$

Solving (7) requires extracting and fusing information from three heterogeneous data sources, which is a highly challenging task. In this paper, we adopt a data-driven approach that learns the function $\mathcal{F}(\cdot)$ directly from data.

3. DETECTION AND LIDAR SELECTION

In this section, we describe how raw camera images and LiDAR point clouds are transformed into structured inputs suitable for multimodal learning. The processing consists of two main functions: the detection block $\mathcal{G}^{\text{img}}(\cdot)$, which extracts bounding boxes, keypoints, and material labels from images, and the LiDAR selection block $\mathcal{S}^{\text{lidar}}(\cdot)$, which filters and labels LiDAR points using the image-derived bounding boxes as regions of interest.

3.1. Image-Based Detection Block

The first component is the detection block, which corresponds to the function $\mathcal{G}^{\text{img}}(\cdot)$ in (4). As illustrated in Fig. 1, the camera image $\mathbf{I}$ is processed by a fine-tuned YOLOv11-pose model [11]. The model detects the target objects in the scene and outputs 2D bounding boxes $\mathbf{B}$, material labels $\mathbf{M}$, and 2D keypoints $\mathbf{K}$.

In this work, we further augment 2D image with LiDAR information. The motivation for doing so is as follows. The 3D spatial information is essential for beamforming in NLoS scenarios, because the beamforming vector usually aligns with the dominant propagation paths, which are determined by the relative 3D positions and orientations of the UE and surrounding reflecting obstacles. Although the 2D image can already provide material information and projected geometry, it lacks depth information and suffers from scale and occlusion ambiguities, which limit its ability to accurately infer these propagation paths. In contrast, a LiDAR sensor can supply accurate 3D coordinates of the environment, which resolve the ambiguities of 2D images and provide the geometric cues needed to identify dominant reflecting paths for the beamforming. For this reason, we incorporate LiDAR sensing to complement the camera by providing accurate 3D spatial information of the environment.

3.2. LiDAR Selection Block

The second component is the LiDAR selection block, which refines the LiDAR input before feature extraction. Instead of inputting all raw LiDAR points into the neural network, which would lead to high computational complexity, we use information from the camera to focus on the most relevant regions of interest. Let $\mathbf{R}^{\text{CL}} \in \mathbb{R}^{3\times3}$ and $\mathbf{t}^{\text{CL}} \in \mathbb{R}^3$ denote the relative rotation and translation between the LiDAR and camera coordinate systems. For a LiDAR point $\mathbf{L}_i \in \mathbb{R}^3$ expressed in the LiDAR coordinate system, its coordinates in the camera coordinate system are obtained by $\mathbf{L}_i^{\text{C}} = \mathbf{R}^{\text{CL}}\mathbf{L}_i + \mathbf{t}^{\text{CL}}$.

Then, we project $\mathbf{L}_i^{\text{C}}$ onto the image plane, where the corresponding pixel coordinates are $\mathbf{L}_i^{\text{I}} = (u_i, v_i)$. Using the camera intrinsic matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$, the projection is given by $\lambda[u_i, v_i, 1]^\top = \mathbf{K}\mathbf{L}_i^{\text{C}}$, where λ is a homogeneous scaling factor of the camera.

Let $\mathcal{L}^{\text{I}} = \{\mathbf{L}_i^{\text{I}}\}_{i=1}^{N^{\text{L}}}$ denote the set of all LiDAR points projected onto the 2D image plane. Let $\mathbf{b}_j$ denote the bounding box of the j-th detected object in the image, extracted from the j-th row of the matrix $\mathbf{B}$. It is defined as $\mathbf{b}_j = [\underline{x}_j, \underline{y}_j, \bar{x}_j, \bar{y}_j]$, where $\underline{x}_j, \underline{y}_j$ are the lower pixel coordinates and $\bar{x}_j, \bar{y}_j$ are the upper pixel coordinates of the bounding box. From the detection network, we obtain the material label of the object within the bounding box $\mathbf{b}_j$, denoted by m_j. We define the region in the image covered by the j-th bounding box as

$$\mathcal{B}_j \triangleq \{(u, v) \mid \underline{x}_j \le u \le \bar{x}_j, \underline{y}_j \le v \le \bar{y}_j\}. \tag{8}$$

For the j-th bounding box, the set of indices of projected LiDAR points lying inside $\mathcal{B}_j$ is defined as

$$\mathcal{I}_j \triangleq \{\, i \mid \mathbf{L}_i^{\text{I}} \in \mathcal{L}^{\text{I}},\ \mathbf{L}_i^{\text{I}} \in \mathcal{B}_j\}. \tag{9}$$

To reduce complexity while preserving details of small objects, we randomly select N^{S} points from $\mathcal{I}_j$ to form the subset $\mathcal{S}_j \subseteq \mathcal{I}_j$ with $|\mathcal{S}_j| = N^{\text{S}}$. If $|\mathcal{I}_j| < N^{\text{S}}$, zero-padding is applied so that $\mathcal{S}_j$ always contains N^{S} elements. The filtered set of projected LiDAR points associated with the j-th bounding box is defined as

$$\tilde{\mathcal{L}}_j^{\text{I}} = \{(\mathbf{L}_i^{\text{I}},\, m_j) \mid i \in \mathcal{S}_j\}. \tag{10}$$

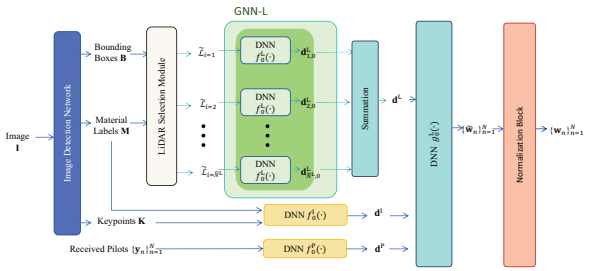


Fig. 2. Architecture of the multimodal GNN framework.

By mapping the selected projected points back to their original 3D LiDAR coordinates, we obtain the final filtered and labeled LiDAR set $\tilde{\mathcal{L}}$ as

$$\tilde{\mathcal{L}} = \bigcup\nolimits_j \{(\mathbf{L}_i, m_j) \mid i \in \mathcal{S}_j\}, \tag{11}$$

where the cardinality is $|\tilde{\mathcal{L}}| = \tilde{N}^{\text{L}} = N^{\text{S}}\hat{N}$.

The set $\tilde{\mathcal{L}}$ from LiDAR, together with material labels $\mathbf{M}$ and keypoints $\mathbf{K}$ from the camera, serve as inputs to the subsequent multimodal learning framework.

In this way, the camera detection module and LiDAR selection module convert raw camera images and point clouds into compact, structured inputs that capture both geometry and material information. This is particularly valuable for MIMO-OFDM systems, because the obtained 3D spatial information implicitly provides reflecting path information and thereby captures the channel characteristics on each subcarrier, which can be utilized to improve the efficiency of beamforming optimization.

4. MULTIMODAL GNN FRAMEWORK

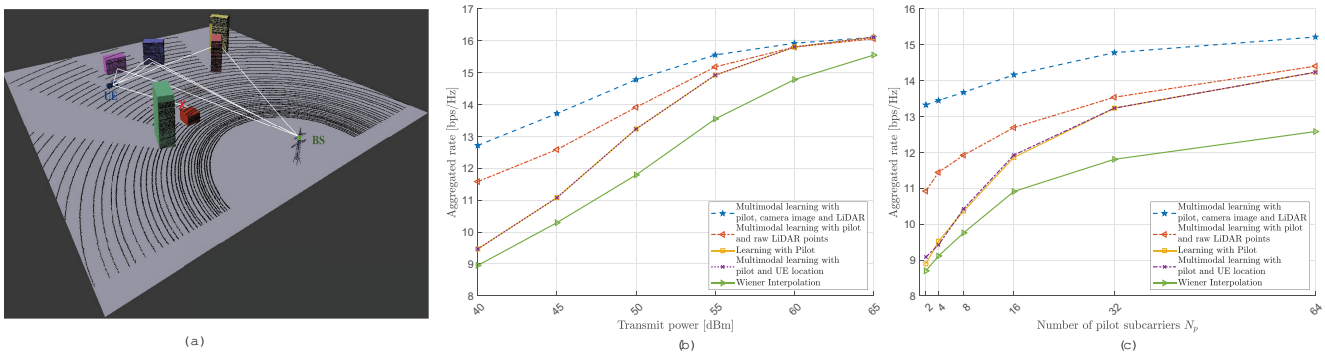
As illustrated in Fig. 2, the proposed framework integrates camera images, LiDAR points, and pilots for beamforming optimization. The processed data consist of: (i) the filtered LiDAR set $\tilde{\mathcal{L}}$, (ii) material labels $\mathbf{M}$ and keypoints $\mathbf{K}$ from camera images, and (iii) the received pilots $\mathbf{y}_n$.

The proposed multimodal optimization framework utilizes a GNN to extract features from LiDAR points and deep neural networks (DNNs) to process pilots as well as material and keypoint information from camera images. Data from the three modalities are first processed separately and then fused for joint optimization.

4.1. Multimodal Neural Network

In beamforming optimization, the order in which LiDAR points are listed is irrelevant, since the beamforming decision depends only on the overall set of spatial points. This permutation invariance property motivates the use of GNNs, which are designed to process unordered data. The GNN-L in Fig. 2 represents the LiDAR feature extraction neural network. Given the selected LiDAR subsets $\{\tilde{\mathcal{L}}_i\}_{i=1}^{\tilde{N}^{\text{L}}}$ obtained from the LiDAR selection module, each subset $\tilde{\mathcal{L}}_i$ is first mapped to an initial node feature through a DNN $f_0^{\text{L}}(\cdot)$, i.e., $\mathbf{d}_{i,0}^{\text{L}} = f_0^{\text{L}}(\tilde{\mathcal{L}}_i)$.

The features $\mathbf{d}_{i,0}^{\text{L}}$ are then combined using a permutation-invariant aggregation function $\rho(\cdot)$, such as element-wise summation or averaging, i.e., $\mathbf{d}^{\text{L}} = \rho(\{\mathbf{d}_{i,0}^{\text{L}}\}_{i=1}^{\tilde{N}^{\text{L}}})$. The aggregated feature $\mathbf{d}^{\text{L}}$ captures the essential LiDAR information for beamforming optimization, which is subsequently fused with the features extracted from camera images and pilots for multimodal optimization.



(a) (b) (c)

Fig. 3. Simulation scenario in NLoS environment and simulation results: (a) simulation scenario in Blensor; (b) aggregate spectral efficiency versus transmit power; (c) aggregate spectral efficiency versus number of pilot subcarriers.

The design ensures that the final LiDAR representation $\mathbf{d}^{\mathrm{L}}$ captures the essential information from all subsets while remaining invariant to the ordering of points.

In parallel, the detected keypoints $\mathbf{K}$ and material information $\mathbf{M}$ extracted from the image detection network are jointly processed by a fully connected network $f_0^I(\cdot)$ to generate the image-derived feature vector, i.e., $\mathbf{d}^{\mathrm{I}} = f_0^{\mathrm{I}}(\mathbf{K}, \mathbf{M})$. This feature vector encodes the 2D spatial information and material types obtained from visual observations and provides complementary information to the LiDAR-derived features for subsequent multimodal fusion.

Similarly, the received uplink pilots $\{\mathbf{y}_n\}_{n=1}^N$ are processed into the feature vector through another DNN $f_0^{\mathrm{P}}(\cdot)$, i.e., $\mathbf{d}^{\mathrm{P}} = f_0^{\mathrm{P}}(\{\mathbf{y}_n\}_{n=1}^N)$. The resulting feature vector encodes the underlying channel characteristics from the pilots.

4.2. Normalization and Loss Function

The extracted features from LiDAR, camera images, and pilots are concatenated to form the multimodal representation $\mathbf{d}^M = [\mathbf{d}^L,\ \mathbf{d}^I,\ \mathbf{d}^P]$. The vector $\mathbf{d}^M$ is then passed to a subsequent DNN block, which further updates the combined features for beamforming optimization. Unlike single-carrier systems, we consider an OFDM system, where the channel varies across subcarriers, and design subcarrier-specific beamforming vectors for effective optimization. The preliminary beamformer for the n-th subcarrier is obtained as $\{\tilde{\mathbf{w}}_n\}_{n=1}^N = g_0^{\mathrm{L}}(\mathbf{d}^M)$.

Finally, the output beamformers are normalized to satisfy the power constraint as $\mathbf{w}_n = \sqrt{P}\frac{\tilde{\mathbf{w}}_n}{\|\tilde{\mathbf{w}}_n\|_2}$, where $\mathbf{w}_n$ denotes the normalized beamforming vector for the i-th subcarrier.

Given the beamformers $\{\mathbf{w}_n\}_{n=1}^N$ produced by the multimodal network, the achievable sum-rate for each sample is computed using (3). The training loss is defined as the negative spectral efficiency.

5. DATASET GENERATION AND PERFORMANCE EVALUATION

5.1. Dataset Generation and Benchmarks

In this section, we describe the dataset generation process and evaluate the proposed multimodal beamforming framework.

The simulated scenario is shown in Fig. 3(a). The CSI and camera images are obtained from the Sionna ray-tracing library [12]. LiDAR data are generated using the Blensor toolbox [13], which simulates realistic LiDAR point clouds by modeling the scanning patterns of the LiDAR sensor. The BS is equipped with a 3×3 uniform planar array. The OFDM system consists of 64 subcarriers with a spacing of 240 kHz, operating at a carrier frequency of 3.5 GHz. The cameras have a horizontal field of view of 136° and a resolution of 1280×420 pixels. The noise power is set to -20 dBm.

We compare the proposed method with several benchmarks. *Learning with Pilot* uses only the received pilots as input to the neural network for beamforming optimization. *Multimodal with Pilot and UE Location* combines pilots with the UE location to highlight the importance of obstacle information. *Multimodal with Pilot and LiDAR without Selection* uses pilots and randomly selected LiDAR points as input. *Wiener Interpolation* estimates CSI of the subcarriers with pilots using LMMSE and reconstructs the full channel based on frequency-domain correlation.

5.2. Simulation Results

We examine the performance in terms of the aggregated rate over all subcarriers in Figs. 3(b) and 3(c). Fig. 3(b) shows the aggregate spectral efficiency performance versus transmit power, where the proposed method consistently achieves the highest rate. In contrast, adding only the UE location to pilots provides almost no benefit compared with the pilot-only method, because in NLoS scenarios the location alone does not capture the channel characteristics of reflected paths. *Multimodal learning without LiDAR selection* offers some gains but suffers from redundancy in the raw point clouds. *Wiener Interpolation* performs worse than the learning-based methods. Fig. 3(c) shows the spectral efficiency performance versus the number of pilot subcarriers, where the proposed algorithm again outperforms all baselines, especially when the number of pilots is small. Together, Figs. 3(b) and 3(c) demonstrate that the proposed algorithm can achieve high spectral efficiency even when the number of pilots is limited.

6. CONCLUSION

This paper proposes a multimodal learning method that combines pilots, camera images, and selected LiDAR for beamforming in OFDM systems in NLOS conditions. By leveraging both sensing data and pilots, the method reduces pilot overhead while maintaining high spectral efficiency. Simulations show that the proposed method outperforms pilot-only or pilot-plus-UE-location baselines, which demonstrates the usefulness of fusing LiDAR and camera image information for beamforming optimization in OFDM systems.

7. REFERENCES

[1] S. Weinstein and P. Ebert, "Data transmission by frequency-division multiplexing using the discrete fourier transform," *IEEE Trans. Commun. Technol.*, vol. 19, no. 5, pp. 628–634, Oct. 1971.

[2] J. A. C. Bingham, "Multicarrier modulation for data transmission: an idea whose time has come," *IEEE Commun. Mag.*, vol. 28, no. 5, pp. 5–14, May. 1990.

[3] S. Colieri, M. Ergen, A. Puri, and A. Bahai, "A study of channel estimation in OFDM systems," in *Proc. IEEE 56th Veh. Technol. Conf. (VTC Fall)*, Sep. 2002, vol. 2, pp. 894–898.

[4] Yinsheng Liu, Zhenhui Tan, Hongjie Hu, Leonard J. Cimini, and Geoffrey Ye Li, "Channel estimation for OFDM," *IEEE Commun. Surv. Tutor.*, vol. 16, no. 4, pp. 1891–1908, Fourth Quarter 2014.

[5] C. Qi, G. Yue, L. Wu, Y. Huang, and A. Nallanathan, "Pilot design schemes for sparse channel estimation in ofdm systems," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 4, pp. 1493–1505, Apr. 2015.

[6] H. Minn and V. K. Bhargava, "An investigation into time-domain approach for OFDM channel estimation," *IEEE Trans. Broadcast.*, vol. 46, no. 4, pp. 240–248, Dec. 2000.

[7] Shuaifeng Jiang and Ahmed Alkhateeb, "Computer vision aided beam tracking in a real-world millimeter wave deployment," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2022, pp. 142–147.

[8] Weihua Xu, Feifei Gao, Xiaoming Tao, Jianhua Zhang, and Ahmed Alkhateeb, "Computer vision aided mmWave beam alignment in V2X communications," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2699–2714, 2023.

[9] Weihua Xu, Feifei Gao, Yong Zhang, Chengkang Pan, and Guangyi Liu, "Multi-user matching and resource allocation in vision aided communications," *IEEE Trans. Commun.*, vol. 71, no. 8, pp. 4528–4543, Aug. 2023.

[10] Aldebaro Klautau, Nuria González-Prelcic, and Robert W. Heath, "LiDAR data for deep learning-based mmWave beam-selection," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 909–912, Jun. 2019.

[11] Glenn Jocher and Jing Qiu, "Ultralytics YOLOv11," `https://github.com/ultralytics/ultralytics`, 2024, Version 11.0.0, License: AGPL-3.0.

[12] Jakob Hoydis, Sebastian Cammerer, Fayçal Ait Aoudia, Avinash Vem, Nikolaus Binder, Guillermo Marcus, and Alexander Keller, "Sionna: An open-source library for next-generation physical layer research," *arXiv preprint*, Mar. 2022.

[13] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, "Blensor: Blender sensor simulation toolbox," in *Proc. Int. Symp. Advances in Visual Computing*, 2011, pp. 199–208.