# Integrated Multimodal Sensing and Communications in 6G Network

Fanyi Meng
fanyimeng@link.cuhk.edu.cn

June 14, 2024

## 1 Introduction

Driven by emerging bandwidth-intensive applications such as Ultra-High Definition (UHD) 3D video, Virtual Reality (VR), and Augmented Reality (AR), the demand for the capacity of wireless communication networks is increasing [1]. Multiple Input Multiple Output (MIMO) is considered an important technique since it can centralize the signal on the user's location, thereby improving transmission efficiency.

However, this approach still faces significant challenges in practical scenarios. On one hand, as the number of antennas increase, the dimension of the Channel State Information (CSI) also increases, making CSI estimation more complex and inevitably increasing the overhead of pilot signals, which in turn occupies communication resources. On the other hand, when users are in high-speed motion, their positions and corresponding CSI change rapidly over time, leading to a mismatch between real and estimated CSI. As a result, the direction of the beams would not align with the user's location, leading to a decrease in signal strength and reduced bandwidth [2].

Recently, referencing extensive work on Integrated Sensing and Communication (ISAC), many researchers have explored using the radar sensing capabilities of base stations to address this issue [3]. By receiving the echo communication signals reflected from the users, base stations can accurately estimate the user's position, speed, and angle of arrival (AOA). This allows the base station to perform predictive beamforming, which improves communication performance. However, these theories rely on mathematical assumptions about sensing channels. In complex RF environments, ISAC signals inevitably experience significant interference, which can adversely affect sensing results, further compromising beam tracking effectiveness and communication throughput.

The aforementioned work motivates us to utilize sensing results to enhance communication. This research proposal is dedicated to enhancing the integration of multimodal sensing and communication. As mentioned in [4, 5], multimodal data can also provide detailed environmental information related to communication performance. This proposal is organized as follows. In Section 2, we will introduce the current work on using multimodal sensors to assist communication. In Section 3, we will discuss some of the existing shortcomings and deficiencies in this area. Finally, in Section 4, we will propose potential solutions to address these issues.

## 2 Literature Review

It has been widely studied that sensors can be used to help with many problems in communication such as beam prediction, blockage prediction, and channel prediction. As sensors provide

rich multipath channel-related information, the transmitter and receiver can avoid complex channel estimation and pilot overhead, showcasing the potential of using sensors to enhance communication performance.

Numerous studies have leveraged various sensors to enhance communication systems. This includes the use of cameras, radar, GPS, IMU, and other sensors. Here we list cameras and radar/LiDAR as representative examples, primarily because of their extensive application, especially within the realm of autonomous driving. Additionally, these two modalities offer rich environmental data that can substantially augment communication capabilities. Furthermore, we will examine the representation of currently available datasets to provide a comprehensive understanding of the field.

**Vision-aided communication** As one of the most common sensors, the camera has been widely deployed in various mobile devices due to its universality, low cost, and high resolution. Cameras provide rich visual information, capturing detailed images and videos of the environment. This visual data is invaluable for a wide range of applications, from object detection and recognition to scene understanding. In [6], considering that segmenting different vehicles in pictures is relatively easy, the authors designed a vision-aided communication scheme to help the base station efficiently achieve user matching and the corresponding resource allocation. In [7], the authors used PSPnet to extract multiple semantic features in the image, such as "building", "location", etc., and used these features to assist in beam prediction and blockage prediction. In [8], the authors collected a dataset in the real world and validated the camera's effectiveness. A millimeter-wave TX was placed on the vehicle, while a stationary millimeter-wave RX and a camera were placed at the roadside. They trained a machine learning model where the input is the photo taken and the output is the strongest or top 5 beam indexes. In [9], the camera was used to estimate the states of the UAVs, which would help the BS to quickly build initial access to multiple UAVs.

**Radar/LiDAR-aided communication** Compared to cameras, LiDAR and radar are also widely used because they can provide distance information that vision alone cannot offer. Their sensing results, point clouds, reflect the signals of electromagnetic waves scattered back by environmental objects. These point clouds contain information about the positions, velocities, and Radar Cross Section (RCS) of environmental objects, making them valuable references for communication. In [10], a distributed architecture that utilized LiDAR on vehicles for line-of-sight detection and beam selection was proposed, confirming the potential of LiDAR to reduce beam training overhead. In [11], LiDAR information was analyzed to classify dynamic and static scatterers. They examined the distributions of parameters such as distance, angle, and power related to these scatterers. Based on this analysis, they achieved channel modeling informed by these parameters. In [12], they used vehicle-mounted millimeter-wave radar to estimate the vehicle's motion state and yaw and employed an Extended Kalman Filter (EKF) algorithm to fuse pilot information, achieving robust parameter estimation with just one pilot symbol, thereby reducing beam tracking overhead. In [13], similar to [8], they collected a dataset of radar and communication and then trained a machine learning model to predict the beam index.

**Related Datasets** As mentioned before, different types of sensing data significantly improve communication performance. Many of the methods leveraging this data are based on machine learning, necessitating the use of widely recognized datasets. Similar to autonomous driving, current datasets are primarily categorized into real-world data and simulated data. For instance, Viwi [14] established a data-generating framework that simulates wireless data using REMCOM and combines it with visual data from the same scenes using Blender. Similarly, $M^3SC$ [15]

simulates camera and LiDAR data based on AirSim and wireless data using REMCOM under different weather conditions. FLASH [16] collected multimodal sensor data and mmWave radio data via an autonomous vehicle. Additionally, Deepsense [17] provides a large-scale, real-world dataset comprising co-existing and synchronized multimodal sensing and communication data from over 40 deployment scenarios, including vehicle-to-infrastructure, vehicle-to-vehicle, reconfigurable intelligent surfaces, pedestrians, and drone communication.

# 3 Research Problem

Although extensive work has already been discussed in this direction, there are still two aspects that are insufficient.

1. **Lack of datasets and simulators:** The current datasets include a variety of scenarios and sensor data. However, in real-world measurements, the current dataset primarily considers single-base station and single-user experimental scenarios, without accounting for potential communication interference. In a communication network, signals from the same frequency originating from different base stations can interference the transmission signal. This may lead to other beams also receiving a certain power gain, further impacting the algorithm's performance. While simulated data can provide channel information in this scenario, the enormous computational load of ray tracing in communication makes it difficult to perform large-scale simulations across multiple base station.

2. **Insufficient Algorithm Interpretability:** Currently, many works use sensor data as input and communication parameters as labels for network training. This approach is based on the assumptions that machine learning can deeply capture the underlying relationships between sensing results and communication. In some scenarios, this approach is highly effective. For instance, in beam prediction tasks, the strongest beam index typically corresponds to the direct path between the transmitter and receiver in LOS scenarios. However, for NLOS paths and conditions like low light, the interpretability of machine learning results remains unclear, making it challenging to ensure that trained models are applicable in different scenarios.

# 4 Methodology

Based on the two issues mentioned in the previous section, we have proposed corresponding solutions.

### Communication Aided Multimodal Sesning Simulator

Sionna [18], a link-level simulator produced by Nvidia, uses TensorFlow as its backend to support machine learning tasks for communication. Based on this, we plan to build a fast and effective simulator One of its toolboxes, Sionna RT [19], enables ray tracing on GPUs, significantly reducing the computation time and resources required for CSI simulation. Additionally, the scenes in Sionna RT are imported through Blender, making it compatible with other perception simulation software such as Airsim and Blensor. This integration makes it possible to build a multi-modal simulator and quickly validate machine learning algorithm performance, providing a valuable opportunity for advancement in this field. The workflow of our proposed simulator is summarized as follows:

1. **Scene Import:** We will use OpenStreetMap to import existing 3D maps from the internet, bringing real-world physical environments into the simulation. Blender models will be used to simulate the movement of vehicles, drones, and other objects.

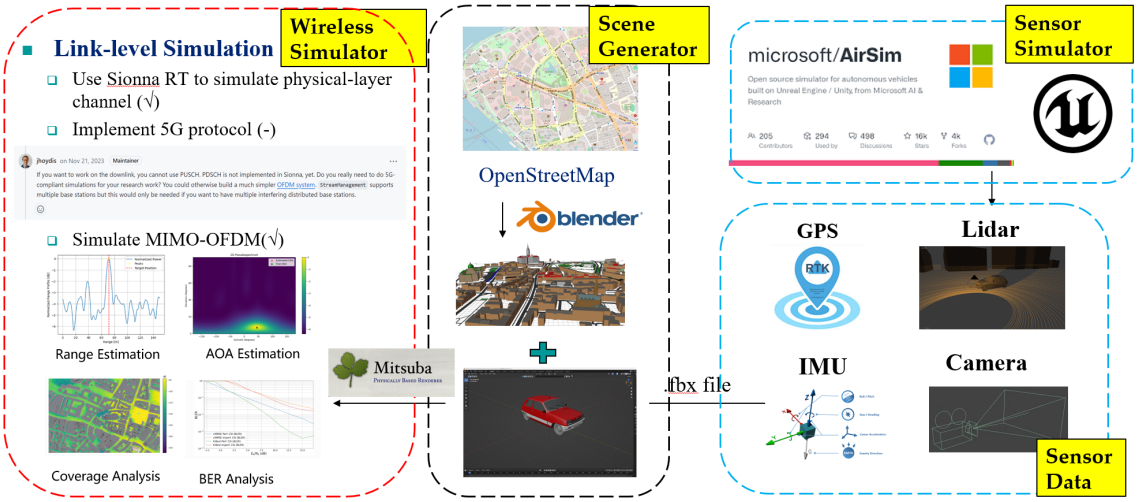


Figure 1: Workflow of the Simulator

2. **Communication Simulation:** Using Sionna RT, we will simulate communication in the previously edited environments. With ray-tracing obtained CSI, we can extract any required parameters, such as beam index and occlusions. We have successfully deployed this part of the code and extended it to OFDM RF-ISAC.

3. **Multi-Modal Sensing Simulation:** The same environment will be imported into Airsim, where we will use Airsim's toolbox to simulate cameras, LiDAR, and other sensors. This part has been initially deployed, but further development is needed for actual alignment and validation.

## Environment-aware Algorithm Design

The fundamental reason sensors can assist communication lies in their ability to provide rich environmental information, which can be utilized for channel modeling. Current end-to-end methods lack in-depth analysis of the environment, resulting in a certain deficiency in interpretability. We aim to dissect the concept of "environment" in a more precise manner. Instead of obtaining comprehensive environmental information through methods like SLAM, we propose leveraging machine learning to extract implicit features of the environment.

For example, NeRF[20] introduces a method for environmental representation based on neural radiance fields. NeRF implicitly represents the physical environment using particles characterized by their color and intensity, and learns these features using images captured from different camera poses. During the inference phase, the corresponding images from different viewpoints can be generated simply by inputting the camera's pose. NeRF2[21] builds upon this idea by modeling the electromagnetic wave propagation environment, based on the fact that light is a type of electromagnetic wave. Specifically, it simulates the attenuation and radiance of received signals to represent the signal strength at the receiver. By training with actual data, it derives the corresponding values in the environment, enabling the inference of signal reception strength at different transmitter positions during the inference phase.

These two methods inspire us to use implicit features to represent the environment, and further, we hope to combine both approaches. We aim to explore how the environmental features provided by sensors can be linked with the environmental features of electromagnetic waves. Although this method is challenging because RF signals operating at communication frequencies are more prone to reflection, diffraction, and scattering, given that their frequencies

are much lower than visible light, it still makes sense for the development of more robust sensing-assisted communication algorithms.

## 5 Conclusion

In this research proposal, we review the current work on using multimodal sensors to assist communication. We summarize the existing achievements and potential issues of the current work and conduct two preliminary exploration of potential solutions to these related problems. Future work will focus on refining these methods and validating their effectiveness in different scenarios.

## References


[1] Wang Yi, Wei Zhiqing, and Feng Zhiyong. Beam training and tracking in mmwave communication: A survey. *China Communications*, 2024.

[2] Chenhao Qi, Peihao Dong, Wenyan Ma, Hua Zhang, Zaichen Zhang, and Geoffrey Ye Li. Acquisition of channel state information for mmwave massive mimo: Traditional and machine learning-based approaches. *Science China Information Sciences*, 64:1–16, 2021.

[3] Fan Liu, Weijie Yuan, Christos Masouros, and Jinhong Yuan. Radar-assisted predictive beamforming for vehicular links: Communication served by sensing. *IEEE Transactions on Wireless Communications*, 19(11):7704–7719, 2020.

[4] Xiang Cheng, Haotian Zhang, Jianan Zhang, Shijian Gao, Sijiang Li, Ziwei Huang, Lu Bai, Zonghui Yang, Xinhu Zheng, and Liuqing Yang. Intelligent multi-modal sensing-communication integration: Synesthesia of machines. *IEEE Communications Surveys & Tutorials*, 26(1):258–301, 2024.

[5] Umut Demirhan and Ahmed Alkhateeb. Integrated sensing and communication for 6g: Ten key machine learning roles. *IEEE Communications Magazine*, 61(5):113–119, 2023.

[6] Weihua Xu, Feifei Gao, Yong Zhang, Chengkang Pan, and Guangyi Liu. Multi-user matching and resource allocation in vision aided communications. *IEEE Transactions on Communications*, 71(8):4528–4543, 2023.

[7] Yuwen Yang, Feifei Gao, Xiaoming Tao, Guangyi Liu, and Chengkang Pan. Environment semantics aided wireless communications: A case study of mmwave beam prediction and blockage prediction. *IEEE Journal on Selected Areas in Communications*, 41(7):2025–2040, 2023.

[8] Gouranga Charan, Muhammad Alrabeiah, Tawfik Osman, and Ahmed Alkhateeb. Camera based mmwave beam prediction: Towards multi-candidate real-world scenarios. *arXiv preprint arXiv:2308.06868*, 2023.

[9] Cui Yanpeng, Zhang Qixun, Feng Zhiyong, Qin Wen, Zhou Ying, Wei Zhiqing, and Zhang Ping. Sensing-assisted accurate and fast beam management for cellular-connected mmwave uav network. *China Communications*, pages 1–19, 2024.

[10] Marcus Dias, Aldebaro Klautau, Nuria González-Prelcic, and Robert W. Heath. Position and lidar-aided mmwave beam selection using deep learning. In *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5, 2019.

[11] Ziwei Huang, Lu Bai, Mingran Sun, and Xiang Cheng. A lidar-aided channel model for vehicular intelligent sensing-communication integration. *arXiv preprint arXiv:2403.14185*, 2024.

[12] Cen Liu, Guangxu Zhu, Fan Liu, Yuanwei Liu, and Kaibin Huang. Successive pose estimation and beam tracking for mmwave vehicular communication systems. In *2023 IEEE Globecom Workshops (GC Wkshps)*, pages 13–19. IEEE, 2023.

[13] Umut Demirhan and Ahmed Alkhateeb. Radar aided 6g beam prediction: Deep learning algorithms and real-world demonstration. In *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 2655–2660, 2022.

[14] Muhammad Alrabeiah, Andrew Hredzak, Zhenhao Liu, and Ahmed Alkhateeb. Viwi: A deep learning dataset framework for vision-aided wireless communications. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, pages 1–5, 2020.

[15] Xiang Cheng, Ziwei Huang, Lu Bai, Haotian Zhang, Mingran Sun, Boxun Liu, Sijiang Li, Jianan Zhang, and Minson Lee. M3sc: A generic dataset for mixed multi-modal (mmm) sensing and communication integration. *China Communications*, 20(11):13–29, 2023.

[16] Batool Salehi, Jerry Gu, Debashri Roy, and Kaushik Chowdhury. Flash: Federated learning for automated selection of high-band mmwave sectors. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, pages 1719–1728, 2022.

[17] Ahmed Alkhateeb, Gouranga Charan, Tawfik Osman, Andrew Hredzak, Joao Morais, Umut Demirhan, and Nikhil Srinivas. Deepsense 6g: A large-scale real-world multi-modal sensing and communication dataset. *IEEE Communications Magazine*, 61(9):122–128, 2023.

[18] 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.

[19] Jakob Hoydis, Faycal Ait Aoudia, Sebastian Cammerer, Merlin Nimier-David, Nikolaus Binder, Guillermo Marcus, and Alexander Keller. Sionna rt: Differentiable ray tracing for radio propagation modeling. In *2023 IEEE Globecom Workshops (GC Wkshps)*, pages 317–321, 2023.

[20] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[21] Xiaopeng Zhao, Zhenlin An, Qingrui Pan, and Lei Yang. Nerf2: Neural radio-frequency radiance fields. In *Proc. of ACM MobiCom '23*, pages 1–15, 2023.