ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11464129

VISION-FIRST MULTIMODAL FRAMEWORK FOR SHORT-TERM MMWAVE BLOCKAGE PREDICTION

Pratik Raj[1], *Madhav Gopal Rayate*[1], *Praval Gupta*[1], *Kaushik Bharatiya*[1], *Rimalapudi Sarvendranath*[1]

[1]Indian Institute of Technology Tirupati, India

ABSTRACT

Predicting user equipment (UE) blockage helps to maintain reliable millimeter-wave communication links. For Task 1 of the CONVERGE Signal Processing Challenge, we present a vision-first multimodal framework for short-term blockage (no, partial, full) prediction using RGB images and radio data. A YOLO detector-based vision pipeline localizes the complete UE and the blocker (curtain). It computes a normalized geometric occlusion ratio between the bounding boxes of the UE and the curtain, and maps it to blockage-class probabilities. Independently, a radio-based pipeline predicts blockage using radio measurements. A vision-first fusion strategy prioritizes the vision-based prediction when the UE and blocker are present in the frame, and falls back to radio-based inference otherwise. The proposed system predicts no, partial, and full blockage states at a prediction horizon of 142 ms and is evaluated using a leave-one-experiment-out protocol, achieving strong weighted F1 performance above 0.93 across multiple experiments.

***Index Terms*—** Blockage prediction, multimodal fusion, mmWave, RGB, radio features.

1. INTRODUCTION

The CONVERGE challenge, part of the ICASSP 2026 Signal Processing grand challenge, provides a multimodal dataset with synchronized visual and radio measurements collected in an indoor mmWave testbed [1, 2]. The challenge focuses on predicting blockages between gNodeB and user equipment (UE). In recent studies, vision-aided approaches have been shown to improve mmWave beam selection and human-induced blockage prediction [3, 4]. In [3], a convolutional neural network (CNN) is trained to perform mmWave beam selection and to predict blockages. Similarly, [5] shows that vision-only sensing can predict future mmWave link blockages by learning environmental changes. A deep learning-based method is employed in [6] to predict dynamic blockages in 6G mmWave systems using a sequence of RGB images and wireless beamforming data. Different from the above works, we propose a vision-first multimodal architecture for blockage prediction.

The work is supported by the DST INSPIRE research grant.

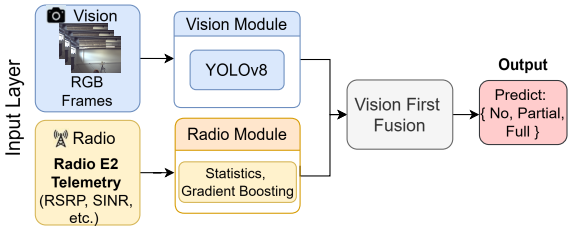


Fig. 1. Overview of the proposed multimodal architecture.

2. EXPERIMENTAL SETUP AND OBJECTIVE

Data for each experiment is collected in an indoor environment, with a fixed UE placed in front of a mmWave radio unit (RU) mounted on a robotic arm. A movable RF-shield curtain placed between the RU and UE creates the blockage. Movement of the curtain and the robotic arm creates different blockage scenarios. Each scenario provides synchronized RGB images, radio measurements, and frame-level ground-truth labeled as *no*, *partial*, or *full* blockage [1].

Objective: Given synchronized multimodal observations at time t, the objective is to predict the blockage state at time $t + 142$ ms, corresponding to 7 fps camera operation. The main challenge here is to perform three-class classification on a highly imbalanced dataset with significantly fewer samples in the partial blockage class.

3. PROPOSED MULTIMODAL ARCHITECTURE

We propose a modular multimodal architecture shown in Fig. 1. It has independent vision and radio pipelines combined via lightweight decision-level fusion logic.

Vision Pipeline estimates blockage by explicitly computing the geometric occlusion from RGB images.[1] From the provided experiments, we have collected 350 images with 640×640 dimensions, covering all possible blockage states. To mitigate the limited size of the CONVERGE dataset, data augmentation is employed. We augmented horizontally and vertically, applied brightness ($\pm 20\%$) and exposure ($\pm 15\%$) adjustments, and Gaussian-blurred images with kernel sizes

[1]Depth is not used as the focus is on blockages and not on the position.

Authorized licensed use limited to: Indian Institute of Technology. Downloaded on May 02,2026 at 02:15:51 UTC from IEEE Xplore. Restrictions apply.

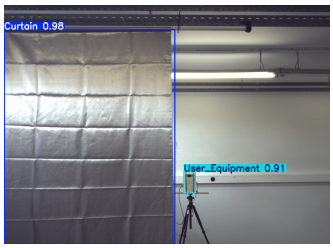


Fig. 2. YOLOv8 detected bounding boxes for UE and curtain.

up to 1 pixel to form an 840-image dataset. A YOLOv8 detector is used, as it balances detection accuracy and latency, and is further trained on an augmented dataset to detect a bounding box for the entire UE rather than selecting a part of it. The updated model weights are used for inference [7, 8].

Trained YOLOv8 outputs the bounding boxes of the complete UE and curtain as shown in Fig. 2. For each frame, we select the UE instance with the highest confidence score. If the UE is absent but the curtain is present in the frame, the vision pipeline predicts full blockage. If both are present, we compute the area of the intersection between the UE and the curtain bounding boxes and normalize it by the UE bounding box area to obtain the *occlusion ratio* O_r. Then, we compare this ratio with empirically determined thresholds T_{no}, T_{full} such that $T_{\text{full}} > T_{\text{no}}$ to output probabilities

$$(P^v_{\text{no}}, P^v_{\text{par}}, P^v_{\text{full}}) = \begin{cases} (1,0,0), & \text{If } O_r < T_{\text{no}}, \\ (0,0,1), & \text{If } O_r > T_{\text{full}}, \\ (0.2(1-\gamma), 0.75, 0.2\gamma), & \text{Otherwise}, \end{cases}$$

where $\gamma = (O_r - T_{\text{no}})/(T_{\text{full}} - T_{\text{no}})$ is the occlusion ratio normalized with the width of the partial blockage case. The class with the maximum probability is predicted.

Radio Pipeline: Let $x_t = (RSRP, RSRQ, SINR, CQI, ULMCS, ULBLER)$, be the radio measurements at time t. We compute their statistics, such as mean, standard deviation, minimum, and maximum, over a short window of length W, i.e., $\{x_{t-W+1}, \ldots, x_t\}$. In addition, we also compute their temporal derivatives such as $\Delta x_t = x_t - x_{t-1}$, and acceleration, $\Delta x_t - \Delta x_{t-1}$, rolling averages, and rolling standard deviation (volatility) over a window size K. Therefore, for each base radio feature in x_t, we compute four statistics and six temporal/rolling features. As x_t has six radio features, we get a 60 dimensional feature vector, i.e., $\mathbf{f}_t \in \mathcal{R}^{60}$. It is passed as input to the Gradient Boosting classifier, which performs well on small and tabular datasets [9], to predict the probabilities $(P^r_{\text{no}}, P^r_{\text{par}}, P^r_{\text{full}})$. Then the class with the maximum probability is predicted.

Vision-First Fusion: Here, we combine the outputs of the vision and radio pipelines to output the final prediction. If UE or a curtain is present, the vision pipeline's output is taken as the final prediction. Otherwise, the radio pipeline's prediction is used. The vision-first approach is used because the radio pipeline performs worse under partial blockage.

Table 1. Performance metrics and experimental scenarios

Test	Macro F1	Wt. F1	Scenario
Performance comparison			
Vision only	0.88	0.99	Fixed arm, curtain moving up
Radio only	0.19	0.24	Fixed arm, curtain moving up
Proposed [2]	0.84	–	–
Baseline [2]	0.63	–	–
Leave-one-experiment-out evaluation			
exp1	0.91	0.95	Moving arm, aligned RU
exp2	0.64	0.93	Moving arm, misaligned RU
exp3	1.00	1.00	Fixed arm, curtain moving up
exp4	0.81	0.98	Fixed arm, curtain moving down
exp5	0.88	0.99	Fixed arm, curtain moving up

4. EXPERIMENTAL RESULTS AND DISCUSSION

We have set $T_{\text{no}} = 0.05$, $T_{\text{full}} = 0.45$, $W = 0.1$ second and K = 3, 5. These thresholds were selected empirically based on validation experiments and reflect the observed transition between partial and full blockage. For the gradient boosting classifier, we used 300 estimators with a max depth of 7 and a learning rate of 0.05.

Table 1 compares the weighted F1-score and the macro F1-scores of our vision-first multimodal pipeline with vision only, radio only, and baseline [2]. In the first part of Table 1, we see that when tested on exp5, the vision-only and radio-only architectures that do not perform fusion achieve macro F1-scores of 0.88 and 0.19, respectively. This significant improvement in performance by the vision-only model motivated our vision-first approach. On the test dataset of the CONVERGE challenge, our fusion-based model achieved a macro F1-score of 0.84, which is significantly higher than the 0.63 achieved by the baseline solution [2]

In the second part of the Table 1, for *evaluation*, a leave-one-experiment-out strategy is adopted. Out of the five experiments in the CONVERGE dataset, we trained on four and tested on the remaining one. We observe that the weighted F1-score is above 0.93 across all experiments, indicating the robustness of our proposed pipeline. We see that *exp3* has a perfect weighted F1-score of 1 as all samples belong to full blockage. We observe the lowest macro F1 and weighted F1-scores for the *exp2*. This is because, in most samples, both the UE and the curtain are out of the camera's view, and the final prediction depends on the radio pipeline. Experiments *exp1*, *exp4*, and *exp5* have different macro F1-scores with *exp4* having lowest score. This happens because the number of partial-blockage samples is lower in *exp1* and *exp5* than in *exp4*. However, the *exp4* and *exp5* have similar weighted F1-score, indicating similar overall classification behavior. For *exp5*, all the partial and full blockage cases are predicted correctly, and 3 out of 83, no-blockage samples are predicted as partial blockage due to the small overlap between the curtain and the UE. On the NVIDIA RTX 3060 Mobile GPU, the reported inference time is around 64 ms per prediction across all test cases in the leave-one-experiment-out evaluation strategy.

Authorized licensed use limited to: Indian Institute of Technology. Downloaded on May 02,2026 at 02:15:51 UTC from IEEE Xplore. Restrictions apply.

5. REFERENCES

[1] Filipe B. Teixeira, Manuel Ricardo, André Coelho, Hélder P. Oliveira, Paula Viana, Nuno Paulino, Helder Fontes, Paulo Marques, Rui Campos, and Luis M. Pessoa, "CONVERGE: A vision-radio research infrastructure towards 6g and beyond," in *2024 Joint European Conference on Networks and Communications 6G Summit (EuCNC/6G Summit)*, June 2024, pp. 1015–1020.

[2] Jichao Chen, Filipe B. Teixeira, Francisco M. Ribeiro, Ahmed Alkhateeb, Manuel Ricardo, Luis M. Pessoa, and Dirk Slock, "CONVERGE Challenge: Multimodal learning for 6G wireless communications," in *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2026.

[3] Muhammad Alrabeiah, Andrew Hredzak, and Ahmed Alkhateeb, "Millimeter wave base stations with cameras: Vision-aided beam and blockage prediction," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, May 2020, pp. 1–5.

[4] A. Alkhateeb, I. Beltagy, and S. Alex, "Machine learning for reliable mmwave systems: Blockage prediction and proactive handoff," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov. 2018, pp. 1055–1059.

[5] Gouranga Charan and Ahmed Alkhateeb, "Computer vision aided blockage prediction in real-world millimeter wave deployments," in *2022 IEEE Global Communications Conference Workshops (GLOBECOM Workshops)*, Dec 2022, pp. 1711–1716.

[6] Gouranga Charan, Muhammad Alrabeiah, and Ahmed Alkhateeb, "Vision-aided dynamic blockage prediction for 6G wireless communication networks," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, June 2021, pp. 1–6.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 779–788.

[8] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in *Computer Vision – ECCV 2024*, 2024, pp. 1–21.

[9] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?," in *Advances in Neural Information Processing Systems*, 2022, vol. 35, pp. 507–520.

Authorized licensed use limited to: Indian Institute of Technology. Downloaded on May 02,2026 at 02:15:51 UTC from IEEE Xplore. Restrictions apply.