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/ICASSP55912.2026.11463713

# WEIGHTED VISION-RADIO FUSION FOR SHORT-TERM BLOCKAGE PREDICTION IN MMWAVE SYSTEMS

*Aadhya Agrawal*[1], *Puppala Jyothi Srinidhi*[1], *R.V.S.R Rupesh*[1],*Tubati Vangmayi*[1], Rimalapudi Sarvendranath [1]

[1]Indian Institute of Technology Tirupati, India

## ABSTRACT

Blockage prediction improves the reliability of millimeter-wave communication systems. We propose a multimodal framework for Task 1 of the CONVERGE Signal Processing Challenge that combines synchronized RGB images and radio measurements. A YOLOv8-based vision module detects the UE and blocker curtain and computes a bounding-box occlusion ratio, which is mapped to the probabilities of no, partial, and full blockage states. In parallel, a gradient boosting classifier estimates blockage state probabilities from structured radio features. The final prediction is obtained through weighted late fusion of the vision and radio module outputs. Evaluated on the CONVERGE dataset using a leave-one-experiment-out protocol, the proposed approach achieves macro and weighted F1-scores of $0.83$ and $0.99$, respectively, with strong performance for no and full blockage cases.

***Index Terms*—** Blockage prediction, mmWave communications, multimodal, late fusion, 6G networks

## 1. INTRODUCTION

Millimeter-wave (mmWave) communication offers high data rates due to the large available bandwidths. However, it's performance is highly sensitive to the blockages caused by users and surrounding objects [1]. Accurate and proactive blockage prediction is therefore critical for reliable beam management. For blockage prediction, deep learning models using radio data have been proposed in [2] while vision-assisted framework is developed in [3, 4]. A cross-band neural network approach for mmWave beam and blockage prediction is introduced in [5], and YOLOv8 based object detection has recently been applied to mmWave radar data in [6]. Different from earlier works, we propose a weighted multimodal late fusion approach that performs blockage prediction by jointly leveraging the image and radio data of a mmWave system provided by the CONVERGE infrastructure [7]. In it, a YOLO based vision module and a gradient boosting based radio module predict the blockage state in parallel. The final prediction is based on the weighted fusion of the outputs from the vision and radio modules.

The work is supported by the DST INSPIRE research grant.

**Table 1**. CONVERGE chamber experimental setup.

| Exp | BS Motion | Curtain Motion | Beam Alignment | Blockage Transition |
|---|---|---|---|---|
| 1 | Left–right | Static | Central | Blockage ↔ No-blockage |
| 2 | Left–right | Static | Non-aligned | Blockage ↔ No-blockage |
| 3 | Static | Up | Fixed | Always blocked |
| 4 | Static | Down | Fixed | No-blockage → Blockage |
| 5 | Static | Up | Fixed | Blockage → No-blockage |

## 2. PROBLEM SETUP AND DATASET DESCRIPTION

The experimental setup consists of a base station (BS) mounted on a robotic arm, equipped with a synchronized RGB-D camera and radio measurement infrastructure. The robotic arm controls BS movement. The camera mounted on the BS captures images at 7 fps, while radio measurements are simultaneously collected through the E2 interface, ensuring time-aligned multimodal data streams. The dataset includes RGB-D frames and radio E2 measurements with frame-level labels for no, partial, or full blockage states collected across five separate experiments conducted in the CONVERGE chamber [7] as shown in Table 1.

*Problem Statement:* For the given multimodal data, the objective is to predict the blockage label at $t + dt$ ($dt = 0.142$ s). The main challenge in this task is to handle the imbalanced dataset, with significantly fewer samples for partial blockage compared to the no and full blockage scenarios. The main goal is to improve the overall F1 score.

## 3. PROPOSED SOLUTION

We propose a multimodal fusion framework for blockage prediction that combines RGB images and radio data, as shown in Fig. 1. It consists of three modules: (i) Vision-based occlusion estimation, (ii) Radio-based blockage inference, and (iii) Weighted late fusion of probabilistic outputs of vision and radio modules.

*YOLO based Vision Module*: We observe that blockage-relevant occlusion is localized over the antenna array on the user equipment (UE). Therefore, we employ a YOLOv8 to output bounding boxes of the curtain that acts as a blockage and the antenna array on the UE as shown in Fig. 2. Here, YOLOv8 is chosen due to its accurate bounding-box



Authorized licensed use limited to: Indian Institute of Technology. Downloaded on May 02,2026 at 02:14:08 UTC from IEEE Xplore. Restrictions apply.

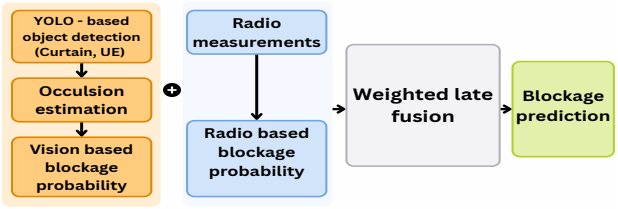



**Fig. 1**. Block diagram of the proposed framework

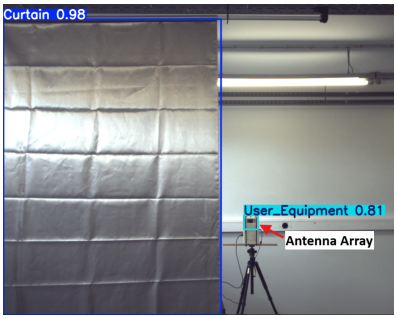



**Fig. 2**. YOLO Detection of Curtain and Antenna Array

detection on small objects and moderate computational overhead [8]. We manually created antenna array labeled data and fine-tuned the YOLOv8 parameters based on the labeled RGB data for accurate detection. We did not consider the disparity information, as blockage does not depend on the depth. To mitigate dataset imbalance and improve generalization, data augmentation techniques, such as random brightness adjustment, orientation changes, and contrast variations, are performed generating a larger training dataset.

Let $BB_{\text{UE}}$ and $BB_{\text{cur}}$ denote the bounding boxes of the UE and curtain, respectively. We compute the occlusion ratio as the area under the intersection of these boxes divided by the area of the UE box. It is given by $\frac{\text{Area}(BB_{\text{UE}} \cap BB_{\text{cur}})}{\text{Area}(BB_{\text{UE}})}$. This bounding-box-based calculation is simpler and more robust than segmentation. We estimate vision based class probabilities $p_{\text{vis}}(c)$ from a single frame at a time for $c \in \{\text{no}, \text{partial}, \text{full}\}$ instead of hard labels to perform uncertainty-aware fusion.

*Radio Based Blockage Estimation and Fusion*: We employ a radio based blockage estimator to address visual occlusion misdetection caused by lighting variations and limited field of view (FOV). For each time step, radio channel quality indicators such as reference signal received power (RSRP), received signal strength indicator (RSSI), signal-to-interference-plus-noise ratio (SINR) and their first- and second-order temporal differences, 3 and 5 frame rolling averages, and 3 and 5 frame variability measures are used to capture instantaneous signal quality and short-term dynamics.

We then train a gradient boosting classifier (GBC) [9, 10] to map the radio feature vector to blockage classes. We chose it due to its strong performance in small data regimes and suitability for tabular data compared to deep neural networks. The classifier outputs probabilities $p_{\text{rad}}(c)$.

*Fusion:* We employ weighted probability averaging with,

$$p_{\text{fused}}(c) = \alpha\, p_{\text{vis}}(c) + (1-\alpha)\, p_{\text{rad}}(c), \qquad (1)$$

where fusion weight $\alpha$ is chosen based on empirical validation, prioritizing visual information.

**Table 2**. Evaluation Results

| Test Experiment | Alpha ($\alpha$) | Macro F1 | Wt. F1 |
|---|---|---|---|
| Exp 5 (Vision only) | 1.0 | 0.83 | 0.98 |
| Exp5 (Radio only) | 0.0 | 0.56 | 0.31 |
| Baseline [7] | - | 0.63 | - |
| Proposed Model [7] | 0.9 | 0.76 | - |
| Exp 1 | 0.9 | 0.88 | 0.95 |
| Exp 2 | 0.9 | 0.73 | 0.95 |
| Exp 3 | 0.9 | 1.00 | 1.00 |
| Exp 4 | 0.9 | 0.79 | 0.99 |
| Exp 5 | 0.9 | 0.83 | 0.99 |

## 4. RESULTS

The YOLO detector is trained on $640 \times 640$ resolution images, selecting the highest-confidence UE instance per frame using default parameters of the ultralytics module. The GBC uses 200 estimators (depth 7, learning rate 0.05, subsample 0.8).

*Performance Benchmarking*: Table 2 compares the performance of the baseline and our proposed approach with vision only and radio only approaches that do not perform fusion. The vision only approach achieves a macro F1-score of 0.83 much higher than the radio only approach, which achieves a score of 0.56. This motivates our vision prioritized weighted fusion with $\alpha = 0.9$. When tested on the CONVERGE challenge dataset, our proposed multimodal fusion approach achieves a macro F1 score of 0.76, outperforming the baseline score of 0.63 [7]. This performance boost over baseline is due to accurate detection of the antenna array and curtain by the YOLOv8 detector and the fusion strategy.

Table 2 also reports results of our model under the leave-one-experiment-out protocol on the experiments listed in Table 1. Exp. 1 and 3 exhibit stable performance as both the UE and curtain remain within the FOV. In Exp. 2, the UE is out of the FOV, and the predictions rely on radio data. Exp. 4 and 5 show similar weighted F1-scores but different macro F1-scores due to class distribution differences. Hence, we see that the dataset composition and fine-grained occlusion patterns influence per-experiment performance. For Exp. 5, the macro F1-score is 0.83, and the weighted F1-score is 0.99. F1-scores of 1.0, 0.5, and 0.99 are observed for no, partial, and full blockage classes, respectively. Misclassifications occur primarily between partial and full blockage cases, with two partial blockage samples predicted as full blockage. A per sample latency of 76 ms is observed for the final testing.



Authorized licensed use limited to: Indian Institute of Technology. Downloaded on May 02,2026 at 02:14:08 UTC from IEEE Xplore. Restrictions apply.

## 5. REFERENCES


[1] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[2] Y. Wu, M. Alrabeiah, and A. Alkhateeb, "Deep learning for moving blockage prediction in millimeter wave wireless networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 3935–3947, 2019.

[3] S. Kumar, M. Alrabeiah, and A. Alkhateeb, "Vision-aided indoor blockage prediction for millimeter wave communications," in *IEEE Global Communications Conference (GLOBECOM)*, 2020, pp. 1–6.

[4] X. Zhang, Y. Wang, M. Alrabeiah, and A. Alkhateeb, "Computer vision aided blockage prediction for millimeter wave communications," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4841–4853, 2020.

[5] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmwave beam and blockage prediction using sub-6 GHz channels," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 550–563, 2021.

[6] M. Raimondi, G. Ciattaglia, A. Nocera, L. Senigagliesi, S. Spinsante, and E. Gambi, "mmDetect: Yolo-based processing of mmwave radar data for detecting moving people," *IEEE Access*, 2023.

[7] J. Chen, F. B. Teixeira, F. M. Ribeiro, A. Alkhateeb, M. Ricardo, L. M. Pessoa, and D. 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.

[8] 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.

[9] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[10] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022.




Authorized licensed use limited to: Indian Institute of Technology. Downloaded on May 02,2026 at 02:14:08 UTC from IEEE Xplore. Restrictions apply.