# Supplementary Materials for "Echoes Beyond Points: Unleashing the Power of Raw Radar Data in Multi-modality Fusion"

## Supplementary Material

## A Proof for column and pillar correspondence

In this section, we will provide a detailed proof for the correspondence between pillar in radar coordinate and column in camera coordinate as described in Section 4.2 of our main paper.

Suppose that the rotation matrix and translation vector between radar and camera are:

$$R = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix}, T = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}. \tag{S1}$$

Then for a point located at $(r_{bev}, \phi_{bev}, z)$ under radar's polar coordinate, it has cartesian coordinate $(x_R, y_R, z_R)$:

$$x_R = r_{bev} \cos(\phi_{bev}), y_R = r_{bev} \sin(\phi_{bev}), z_R = z, \tag{S2}$$

and it can be projected to the camera frame using extrinsic parameters:

$$\begin{pmatrix} x_C \\ y_C \\ z_C \end{pmatrix} = R \begin{pmatrix} x_R \\ y_R \\ z_R \end{pmatrix} + T = \begin{pmatrix} a_1 r_{bev} \cos\phi_{bev} + b_1 r_{bev} \sin\phi_{bev} + c_1 z + d_1 \\ a_2 r_{bev} \cos\phi_{bev} + b_2 r_{bev} \sin\phi_{bev} + c_2 z + d_2 \\ a_3 r_{bev} \cos\phi_{bev} + b_3 r_{bev} \sin\phi_{bev} + c_3 z + d_3 \end{pmatrix}. \tag{S3}$$

Using the intrinsic parameters, the camera coordinate $(x_C, y_C, z_C)$ can be correlated with image pixel position $(x_I, y_I)$ using:

$$\frac{x_I - u_0}{f_x} = \frac{x_C}{z_C}, \frac{y_I - v_0}{f_y} = \frac{y_C}{z_C}, \tag{S4}$$

where $f_x$, $f_y$, $u_0$, and $v_0$ are intrinsic parameters. Our goal is to find a situation that the pillar is projected as a column on the image plane. Under this condition, $x_I$ should be irrelevant with $z$, i.e. in the equation below:

$$\frac{x_I - u_0}{f_x} = \frac{x_C}{z_C} = \frac{a_1 r_{bev} \cos\phi_{bev} + b_1 r_{bev} \sin\phi_{bev} + c_1 z + d_1}{a_3 r_{bev} \cos\phi_{bev} + b_3 r_{bev} \sin\phi_{bev} + c_3 z + d_3}. \tag{S5}$$

The coefficients of z should be 0, which means by combining the relationship between rotation matrix $R$ with roll $\alpha$, pitch $\beta$, and yaw $\gamma$ [1], we can derive:

$$\begin{cases} c_1 = \sin\gamma \sin\alpha + \cos\gamma \sin\beta \cos\alpha = 0 \\ c_3 = \cos\beta \cos\alpha = 0. \end{cases} \tag{S6}$$

Considering second row of Eq. (S6), there are two solutions $\beta = \pm\frac{\pi}{2}$ or $\alpha = \pm\frac{\pi}{2}$. Considering the first solution $\beta = \pm\frac{\pi}{2}$, the first row of Eq. (S6) is transformed to:

$$\sin\gamma \sin\alpha \pm \cos\gamma \cos\alpha = 0, \tag{S7}$$

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

which means:

$$\cos(\gamma \mp \alpha) = 0. \tag{S8}$$

Thus we have:

$$\begin{aligned} &\beta = \frac{\pi}{2}, \gamma = \alpha \pm \frac{\pi}{2}, \alpha \in [-\pi, \pi] \ , \\ or,\ &\beta = -\frac{\pi}{2}, \gamma = -\alpha \pm \frac{\pi}{2}, \alpha \in [-\pi, \pi] \, . \end{aligned} \tag{S9}$$

For another solution $\alpha = \pm\frac{\pi}{2}$, the first row of Eq. (S6) is transformed to:

$$\sin\gamma \sin\alpha + \cos\gamma \sin\beta \cos\alpha = \pm \sin\gamma = 0. \tag{S10}$$

Thus we have:

$$\alpha = \pm\frac{\pi}{2}, \gamma = 0 \ or \ \pi, \beta \in [-\pi, \pi] \, . \tag{S11}$$

As a result, the final solution is:

$$\begin{aligned} &\beta = \frac{\pi}{2}, \gamma = \alpha \pm \frac{\pi}{2}, \alpha \in [-\pi, \pi] \, , \\ or,\ &\beta = -\frac{\pi}{2}, \gamma = -\alpha \pm \frac{\pi}{2}, \alpha \in [-\pi, \pi] \, , \\ or,\ &\alpha = \pm\frac{\pi}{2}, \gamma = 0 \ or \ \pi, \beta \in [-\pi, \pi] \, . \end{aligned} \tag{S12}$$

In standard ADS, the LiDAR or radar coordinate system usually has a z-axis pointing up, while the camera coordinate system has a y-axis pointing up or down. So a widely adopted practice first exchanges the y and z axes, which means a roll $\alpha = \pm\frac{\pi}{2}$. After that, a pitch $\beta$ around the vertical y-axis is executed, while further rotation around the z-axis is not required, as shown in Figure S1. In other words, the roll $\beta$ can be any value between $-\pi$ and $\pi$, while yaw $\gamma$ is set to 0. This is exactly the situation of the third solution of Eq. (S12). And under this condition, we have expression of $R$ according to [1]:

$$R = \begin{pmatrix} \cos\beta & \pm\sin\beta & 0 \\ 0 & 0 & \mp 1 \\ -\sin\beta & \pm\cos\beta & 0 \end{pmatrix}, T = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}. \tag{S13}$$

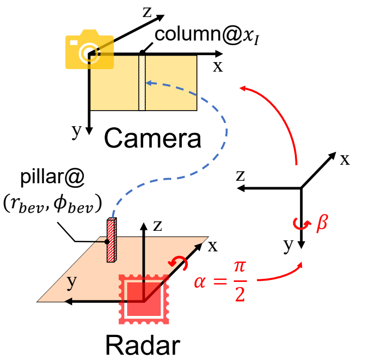


Figure S1: Illustration of coordinate system transformation.

Thus, using Eq. (S3) and Eq. (S5), we can get the expression of column index $x_I$. If $\alpha = \frac{\pi}{2}$, we have:

$$x_I = u_0 + f_x \frac{r_{bev}\cos(\beta - \phi_{bev}) + d_1}{-r_{bev}\sin(\beta - \phi_{bev}) + d_3}. \tag{S14}$$

If $\alpha = -\frac{\pi}{2}$, we have:

$$x_I = u_0 + f_x \frac{r_{bev}\cos(\beta + \phi_{bev}) + d_1}{-r_{bev}\sin(\beta + \phi_{bev}) + d_3}. \tag{S15}$$

For RADIal, the rotation matrix from the radar coordinate system to the camera coordinate system is:

$$R = \begin{pmatrix} 0.0465 & -0.9989 & -0.0051 \\ -0.0476 & 0.0029 & -0.9989 \\ 0.9978 & 0.0467 & -0.0474 \end{pmatrix}, \tag{S16}$$

which is the approximation of $(\alpha, \beta, \gamma) = (\pi/2, -\pi/2, 0)$. And thus we approximately have:

$$x_I \approx u_0 + f_x \frac{-r_{bev}\sin\phi_{bev} + d_1}{r_{bev}\cos\phi_{bev} + d_3}. \tag{S17}$$

It can be inferred that $x_I$ is almost irrelevant to $z$, which means one pillar in the polar coordinate of radar corresponds to a column in the image.

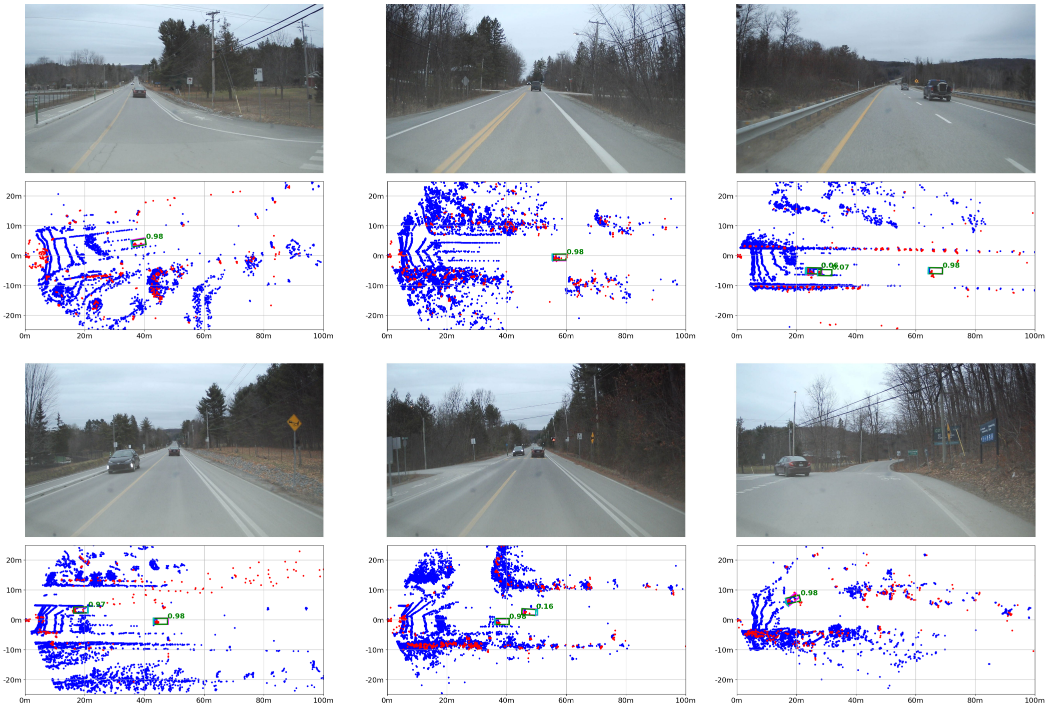
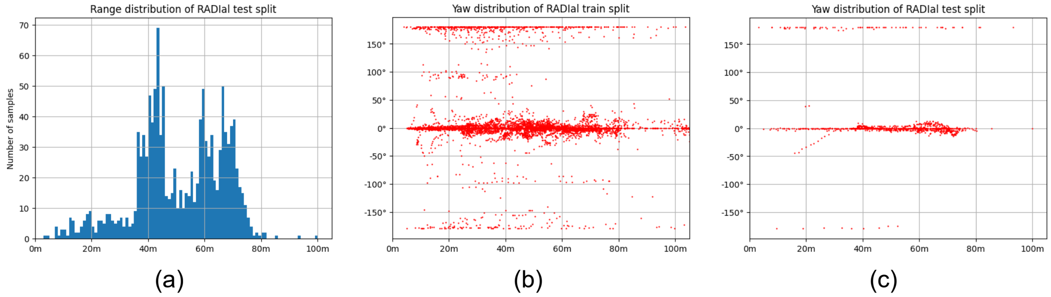


Figure S2: Visualization of ground truths and predictions from RT and image fusion. The ground truth bounding boxes are in pink, while the predicted bounding boxes are in green with the confidence score on its upper right. The LiDAR points and the radar points are respectively in blue and red. The rear edge of each car is in cyan. We recommend readers to zoom in for better detail.



(a)

(b)

(c)

Figure S3: Statistics of RADIal dataset. Image (a) corresponds to sample distribution over range, while images (b) and (c) correspond to yaw distribution over the range of train and test data splits.

# B Visualization and Dataset Statistics

To better illustrate the performance of our method, we visualize the predictions and ground truths in Figure S2. As can be seen from the cases in the first and the second column, our method can accurately predict the position of vehicles. But there are several cases in which the accurate prediction does not get high confidence, as shown in the upper case in the third column. On the other hand, the orientation estimation error may be significant when the yaw of the vehicle is large, shown in the inferior case in the third column. The main reason is that the samples with large yaw are relatively rare in the dataset, as shown in Figure S3 (b) and (c).

As shown in Figure S3 (a), the samples of test data are rather rare within 30 meters and out of 80 meters. To comprehensively explore the detection performance of objects, we divide the range into 0-50m and 50-100m, which correspond to short-range and long-range detection performance respectively.

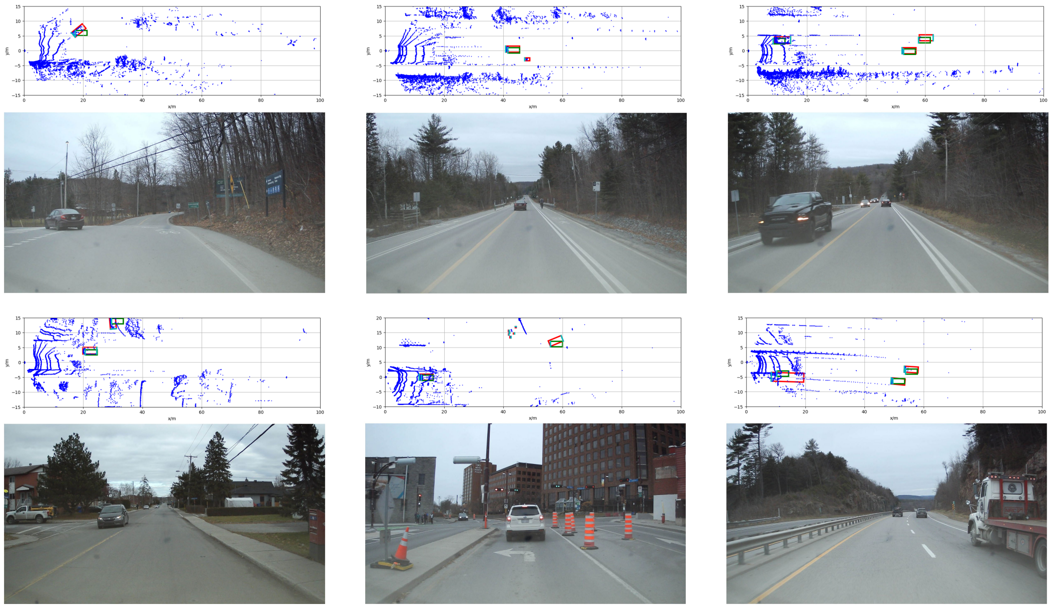


Figure S4: The visualization of the old labels and new labels under Bird's Eye View. The old labels include 3D center points and fixed scale and orientation, denoted as green boxes. The new labels are denoted as red boxes, which we offered real scale and orientation, as well as objects' accurate center position. The rear edge of each bounding box is colored in cyan. The left, middle, and right column respectively denotes the situation of large heading angle, missed traffic participants, and vehicles of large scale. We recommend readers to zoom in for better detail.

Table S1: LET-BEV-AP results with different longitudinal tolerance $T_l$. The performances of different modality combinations of our algorithm on RADIal Dataset test split are presented, using refined 3D bounding box annotations. C, L, and RT respectively refer to the camera, LiDAR, and range-time data. The results with LiDAR are on gray background.

| Modality | | | $T_l = 0.1$, LET-BEV-AP@0.7(%)↑ | | | $T_l = 0.3$, LET-BEV-AP@0.7(%)↑ | | |
|---|---|---|---|---|---|---|---|---|
| C | RT | L | Overall | 0 - 50m | 50 - 100m | Overall | 0 - 50m | 50 - 100m |
| ✓ | | | 56.92±3.91 | 78.98±1.41 | 36.20±5.81 | 81.36±1.15 | 83.30±2.01 | 77.00±0.43 |
| | ✓ | | 55.04±1.25 | 68.56±4.11 | 52.30±1.82 | 55.34±1.91 | 69.79±4.59 | 51.02±0.07 |
| | | ✓ | 86.07±0.44 | 88.55±0.79 | 93.17±0.10 | 86.41±0.47 | 89.72±0.03 | 93.22±0.04 |
| ✓ | ✓ | | 88.86±0.62 | 92.81±1.37 | 94.81±0.52 | 88.56±0.53 | 92.58±0.70 | 94.57±0.29 |

We further visualize scenarios that are corrected by our new labels. It is worth noting that the original annotation and prediction of previous methods are centers of objects' visible faces, and they assign a fixed scale and zero orientation when calculating Average Precision (AP). But vehicles of large angles and scales are normal in daily scenarios, as shown in Figure S4. Our new annotations based on dense LiDAR points correct this, helping a more comprehensive evaluation of existing methods. We also annotate other traffic participants such as pedestrians and cyclists. But due to their limited numbers (4 cyclists, 23 pedestrians in training split, and 17 cyclists, 0 pedestrians in test split), we don't add more fine-grained class labels.

## C Complementary property of radar and camera

To highlight the complementary nature of radar and camera properties, we investigated the performance of LET-BEV-AP with varying longitudinal tolerance ($T_l$), as presented in Table S1. Notably, when using a larger $T_l$ of 0.3, the camera's performance experiences a significant improvement. This suggests that the relatively lower performance of long-range LET-BEV-AP under $T_l$ of 0.1 primarily stems from depth errors. By reducing the impact of longitudinal estimation, the camera's exceptional

Table S2: Recall with different IoU thresholds. The performances of different modality combinations of our algorithm on RADIal Dataset test split are presented, using refined 3D bounding box annotations. C, L, and RT respectively refer to the camera, LiDAR, and range-time data. The results with LiDAR are on gray background.

| Modality | | | BEV recall@0.7(%)↑ | | | BEV recall@0.3(%)↑ | | |
|---|---|---|---|---|---|---|---|---|
| C | RT | L | Overall | 0 - 50m | 50 - 100m | Overall | 0 - 50m | 50 - 100m |
| | ✓ | | 62.73±1.24 | 73.89±2.11 | 61.21±0.82 | 86.27±0.43 | 92.13±0.71 | 93.40±0.54 |
| | | ✓ | 86.38±0.00 | 89.11±0.08 | 93.35±0.08 | 91.40±0.08 | 96.13±0.17 | 96.95±0.16 |
| ✓ | ✓ | | 86.18±0.13 | 89.13±0.6 | 92.53±0.56 | 91.78±0.85 | 95.82±0.50 | 97.67±1.05 |

Table S3: BEV detection performances of different modality combinations of our algorithm on RADIal Dataset test split. Refined 3D bounding box annotations are applied. C, L, ADC, RD, RT, RA, RP respectively refer to the camera, LiDAR, ADC data, range-Doppler spectrum, range-time data, range-azimuth map, and radar point cloud. The best result without LiDAR of each metric are in **bold**, and the results with LiDAR are on gray background.

| Modality | | | | | | | BEV AP@0.7(%)↑ | | | LET-BEV-AP@0.7(%)↑ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | ADC | RT | RD | RA | RP | L | Overall | 0 - 50m | 50 - 100m | Overall | 0 - 50m | 50 - 100m |
| ✓ | | | | | | | 10.07±1.10 | 21.68±3.11 | 2.21±0.59 | 56.92±3.91 | **78.98±1.41** | 36.20±5.81 |
| | ✓ | | | | | | 49.64±0.43 | 61.37±1.57 | 47.92±1.48 | 56.52±2.36 | 68.93±1.87 | 54.68±1.76 |
| | | ✓ | | | | | 48.26±0.92 | 62.33±4.00 | 45.05±0.98 | 55.04±1.25 | 68.56±4.11 | 52.30±1.82 |
| | | | ✓ | | | | 47.34±0.73 | 59.31±1.64 | 42.53±1.98 | 54.96±0.67 | 67.02±1.67 | 50.59±0.26 |
| | | | | ✓ | | | 50.67±1.03 | 65.05±1.59 | 42.98±0.51 | 58.53±0.13 | 72.95±0.78 | 51.41±0.54 |
| | | | | | ✓ | | **53.55±0.70** | **66.19±1.72** | **47.41±1.06** | **62.59±1.14** | 76.45±2.06 | **56.73±1.33** |
| | | | | | | ✓ | 84.47±0.15 | 86.92±0.61 | 91.91±0.13 | 86.07±0.44 | 88.55±0.79 | 93.17±0.10 |
| ✓ | ✓ | | | | | | 84.85±0.22 | 87.68±1.54 | 91.46±1.08 | 88.66±0.63 | 92.43±1.15 | 94.99±0.66 |
| ✓ | | ✓ | | | | | **84.92±0.98** | 87.56±1.58 | 91.06±1.18 | 88.86±0.62 | 92.81±1.37 | 94.81±0.52 |
| ✓ | | | ✓ | | | | 83.31±0.39 | 87.26±0.56 | 89.16±0.44 | 88.24±0.06 | 92.18±0.43 | 93.26±0.73 |
| ✓ | | | | ✓ | | | 84.77±0.65 | **87.93±0.60** | **91.48±0.28** | **89.54±0.54** | **93.83±0.56** | **94.87±0.58** |
| ✓ | | | | | ✓ | | 82.35±0.93 | 86.97±1.01 | 86.39±0.74 | 88.35±0.78 | 93.07±0.80 | 92.77±0.69 |
| ✓ | | | | | | ✓ | 86.35±1.15 | 88.81±1.40 | 94.25±1.68 | 88.44±0.79 | 91.63±1.34 | 95.31±1.28 |

angular perception accuracy becomes more pronounced. Conversely, the LET-BEV-AP performance of the radar-only variant remains relatively unchanged. This finding indicates that the radar's depth estimation is quite accurate and minimally affected by the value of $T_l$. Similarly, the highly accurate LiDAR-only variant and camera and RT fusion variant don't suffer from performance fluctuation as well, which supports our conclusion.

To further investigate the primary factor affecting radar performance, we conducted a comparison of BEV recall at IoU thresholds of 0.7 and 0.3, as outlined in Table S2. It can be deduced that by using a less stringent IoU threshold, the recall of the RT-only variant experiences a significant increase. This result suggests that the RT-only variant's limitation mainly lies in its inferior angular localization ability. But by combining the strengths of camera and radar modalities, the model can leverage the advantages of both modalities, leading to a significant performance boost.

## D Doppler domain multiplexing and coordinate system

As mentioned in [2], Doppler Domain Multiplexing(DDM) is used to distinguish received radar signals from different transmitters. The mechanism can be best illustrated in Range-Doppler (RD)

Table S4: Ablation on the coordinate system. Experiments are carried on RADIal dataset, with RT data and image as input.

| | BEV AP@0.7(%)↑ | | | LET-BEV-AP@0.7(%)↑ | | |
|---|---|---|---|---|---|---|
| Coordinate | Overall | 0 - 50m | 50 - 100m | Overall | 0 - 50m | 50 - 100m |
| Polar | 84.92±0.98 | 87.56±1.58 | 91.06±1.18 | 88.86±0.62 | 92.81±1.37 | 94.81±0.52 |
| Sphere | 85.02±0.60 | 88.02±0.59 | 91.76±0.98 | 89.97±0.61 | 93.35±0.32 | 96.41±1.21 |

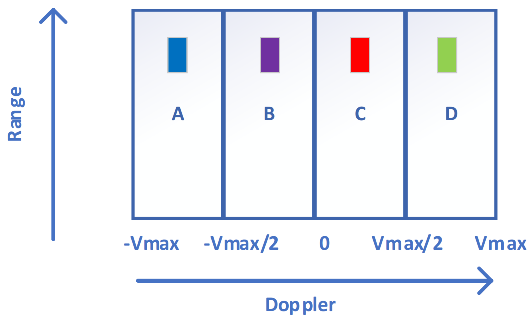


Figure S5: The illustration of Doppler Domain Multiplexing(DDM). This figure shows a toy example of how to use DDM to distinguish four transmitters.

Table S5: BEV detection performances of different modality combination on KRadar Dataset test split with KITTI protocol of 40 recall positions. C, RA, RP respectively refer to the camera, range-azimuth map, and radar point cloud. The best result of each metric is in **bold**.

| | | BEV AP(%)↑ | | | 3D AP(%)↑ | | |
|---|---|---|---|---|---|---|---|
| Training set | Method | AP@0.3 | AP@0.5 | AP@0.7 | AP@0.3 | AP@0.5 | AP@0.7 |
| KRadar-20-train | EchoFusion(RP) | 55.74 | 43.94 | 20.56 | 53.40 | 29.21 | 2.90 |
| KRadar-20-train | EchoFusion(RA) | 54.45 | 42.28 | 17.97 | 51.75 | 24.65 | 3.85 |
| KRadar-20-train | EchoFusion(RP+C) | 66.46 | 54.17 | 26.39 | 58.62 | **34.67** | 3.96 |
| KRadar-20-train | EchoFusion(RA+C) | **68.70** | **55.90** | **29.65** | **66.68** | 34.33 | **5.34** |

spectrum format of data, as shown in Figure S5. Without DDM, the valid measurable speed interval is $[-V_{max}, V_{max}]$. Suppose we have four transmitters, Tx 1/2/3/4, and one object with speed in $[0, V_{max}/2]$. The echo signals from Tx 1/2/3/4 will respectively fall into sections C, D, A, B by using DDM, showing a periodical pattern. But now we can only measure speed in $[0, V_{max}/2]$. Similar phenomenon can be observed in the figure of Range Doppler data, i.e. the third image in **??** in the paper.

We also carried out experiments on coordinate system of BEV space. Though the range dimension is under sphere system, we found that a polar system can achieve on par performance, as shown in Table S4. Thus we apply the polar system in the paper.

## E More experimental results

Referring to Table S3, we have included the detection performance when our model consumes the ADC data or range-Doppler spectrum as input. It is worth noticing that we add an extra complex linear layer initialized with FFT coefficients for ADC data to replace the range FFT and the side-lobe suppression. In the case of the single modality version, the performance discrepancy between RT and these two modalities remains within the margin of error. When considering the fusion of image data, the performance of RD slightly lags behind that of RT, while that of ADC is almost the same as RT. These findings suggest that the integration of image data significantly mitigates the importance of conventional radar signal processing in accurate radar-based detection.

Besides, we offer ablation of different modality combinations on KRadar. The results are reported in Table S5. Here, we train all models with training split. The RA map shows considerable superiority over radar point cloud when fusing with the monocular image, which aligns with our research motivation and results on RADIal dataset.

## References

[1] Richard M Murray, Zexiang Li, S Shankar Sastry, and S Shankara Sastry. *A mathematical introduction to robotic manipulation*. CRC press, 1994.

[2] Julien Rebut, Arthur Ouaknine, Waqas Malik, and Patrick Pérez. Raw high-definition radar for multi-task learning. In *CVPR*, 2022.