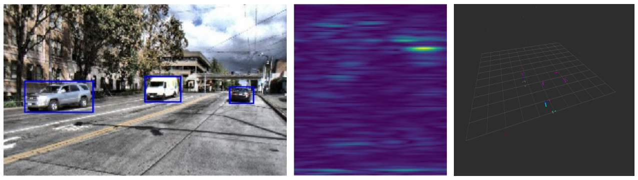
Rethinking of Radar's Role: A Camera-Radar Dataset and Systematic Annotator via Coordinate Alignment (Supplementary Document)

Yizhou Wang[1], Gaoang Wang[2], Hung-Min Hsu[1], Hui Liu[1,3], Jenq-Neng Hwang[1]
[1]University of Washington, Seattle, WA, USA
[2]Zhejiang University, Hangzhou, China
[3]Silkwave Holdings Limited, Hong Kong, China
{ywang26,hmhsu,huiliu,hwang}@uw.edu, gaoangwang@intl.zju.edu.cn

1. Radar Data Representations

In this section, we would like to introduce the two radar data representations, i.e., radio frequency (RF) image and radar points, in detail. The examples of an RGB image, an RF image, and a frame of radar points are shown in Fig. 1.



(a) RGB Image (b) RF Image (c) Radar Points

Figure 1. Examples for RGB image, RF image, and radar points. (a) Three cars in the RGB image; (b) Annotation is extremely difficult on the corresponding RF image; (c) Radar points are sparse but have relative precise locations.

RF images are in radar range-azimuth coordinates and can be described as a bird's-eye view (BEV) representation, where the x-axis denotes azimuth (angle) and the y-axis denotes range (distance). For an FMCW radar, it transmits continuous chirps and receives the reflected echoes from the obstacles. After the echoes are received and pre-processed, we implement the fast Fourier transform (FFT) on the samples to estimate the range of the reflections. A low-pass filter (LPF) is then utilized to remove the high-frequency noise across all chirps in each frame at the rate of 30 FPS. After the LPF, we conduct a second FFT on the samples along different receiver antennas to estimate the azimuth angle of the reflections and obtain the final RF images. The radar points can then be derived from these frequency images through the peak detection algorithm. This pre-processing pipeline is shown in Fig. 2. After being transformed into RF images, the radar data become a similar format as image sequences, which can thus be directly processed by an image-based CNN.

2. Derivation for Camera-Radar BCP

We start from projecting a point $(r, \theta) \in \mathcal{R}$ to $\mathbf{x}^c = (x^c, y^c, z^c)$. Before that, we first do a polar to Cartesian coordinate transformation,

$$\begin{aligned} x^r &= r\sin(\theta), \\ z^r &= r\cos(\theta). \end{aligned} \tag{1}$$

This point can be transformed to $\mathcal{W}^c$ by sensor calibration of the translation from camera to radar $\mathbf{t}_{cr} = [t_{cr,x}, t_{cr,y}, t_{cr,z}]^\top$,

$$\begin{aligned} x^c &= x^r + t_{cr,x}, \\ z^c &= z^r + t_{cr,z}. \end{aligned} \tag{2}$$

Here, we ignore the rotation transformation between $\mathcal{W}^r$ and $\mathcal{W}^c$ since both sensors are well-calibrated with the same orientation.

To calculate y^c, we first consider the ground plane with only pitch rotation angle φ. Assuming a small φ, which is a valid assumption in driving scenarios, y^c_φ can be approximated by the similar triangles rule as

$$\begin{aligned} y^c_\varphi &= \left(\frac{h}{\sin(\varphi)} - r\right) \cdot \tan(\varphi) \approx \left(\frac{h}{\sin(\varphi)} - r\right) \cdot \sin(\varphi) \\ &= h - r\sin(\varphi). \end{aligned} \tag{3}$$

By taking the second rotation angle γ into consideration, we then have

$$y^c_\gamma = x^c \tan(\gamma). \tag{4}$$

Therefore, the overall y^c can be represented as

$$\begin{aligned} y^c &= y^c_\varphi - y^c_\gamma = h - r\sin(\varphi) - x^c\tan(\gamma) \\ &= h - \sqrt{(x^c)^2 + (z^c)^2} \cdot \sin(\varphi) - x^c \tan(\gamma). \end{aligned} \tag{5}$$

Finally, we project $\mathbf{x}^c \in \mathcal{W}^c$ to $\mathcal{C}$ by the camera intrinsic

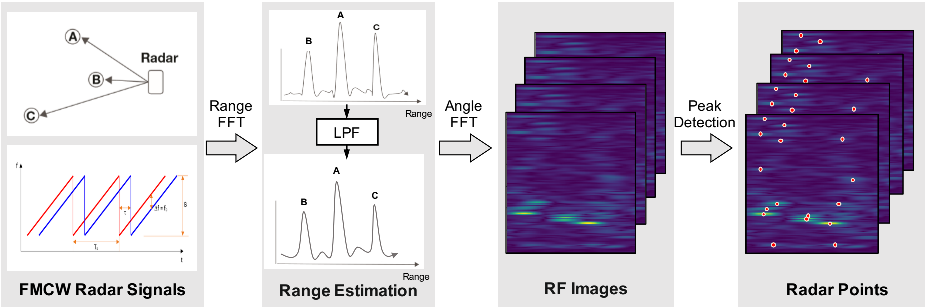


Figure 2. The workflow of the RF image generation from the raw radar signals.

matrix $\mathbf{K}$,

$$s\begin{bmatrix}u\\v\\1\end{bmatrix}=\mathbf{Kx^c}=\begin{bmatrix}f_x&0&c_x\\0&f_y&c_y\\0&0&1\end{bmatrix}\begin{bmatrix}x^c\\y^c\\z^c\end{bmatrix}\tag{6}$$

Thus, we have

$$\begin{aligned}u&=f_x\frac{x^c}{z^c}+c_x,\\v&=f_y\frac{y^c}{z^c}+c_y.\end{aligned}\tag{7}$$

Put it together, we can project a point $(r,\theta)\in\mathcal{R}$ to $(u,v)\in\mathcal{C}$ by the following $\mathcal{R}\to\mathcal{C}$ projection, denoted as $P(\cdot)$,

$$[u,v]^\top=P(r,\theta),\tag{8}$$

where

$$\begin{aligned}u&=f_x\frac{r\sin(\theta)+t_{cr,x}}{r\cos(\theta)+t_{cr,z}}+c_x,\\v&=f_y\frac{h-r\sin(\varphi)-(r\sin(\theta)+t_{cr,x})\tan(\gamma)}{r\cos(\theta)+t_{cr,z}}+c_y.\end{aligned}\tag{9}$$

After the $\mathcal{R}\to\mathcal{C}$ projection is properly derived, we would like to consider the other direction, i.e., $\mathcal{C}\to\mathcal{R}$ projection. Since the projection in Eq. 8 is a 2D-to-2D projection without losing any degrees of freedom (DoF), it can be reversed. Therefore, we define the $\mathcal{C}\to\mathcal{R}$ projection as

$$[r,\theta]^\top=P^{-1}(u,v).\tag{10}$$

Here, the $\mathcal{C}\to\mathcal{W}^c$ projection can be derived as

$$\begin{aligned}x^c&=\frac{h\hat{x}}{\sqrt{1+\hat{x}^2}\sin(\varphi)+\hat{x}\tan(\gamma)+\hat{y}},\\z^c&=\frac{h}{\sqrt{1+\hat{x}^2}\sin(\varphi)+\hat{x}\tan(\gamma)+\hat{y}},\end{aligned}\tag{11}$$

where

$$\begin{aligned}\hat{x}&=\frac{x^c}{z^c}=\frac{u-c_x}{f_x},\\\hat{y}&=\frac{y^c}{z^c}=\frac{v-c_y}{f_y}.\end{aligned}\tag{12}$$

Note that $\mathcal{W}^c\to\mathcal{R}$ projection just contains the camera to radar translation and a Cartesian to polar coordinate transformation, which is trivial and will not be elaborated in detail.

3. Object Location Similarity Details

We propose the object location similarity (OLS) to represent the similarity between a detection i and a ground truth j to be

$$\mathrm{OLS}(i,j)=\exp\left\{\frac{-d_{ij}^2}{2(s_j\kappa_{cls})^2}\right\},\tag{13}$$

where d is the distance (in meters) between the two points in an RF image; s is the object distance from the sensors; and κ_{cls} is a per-class constant that represents the error tolerance for class cls, which can be determined by the object average size of the corresponding class. Since OLS is reasonably distributed between 0 and 1, we treat it as a good representation of the localization error, and use it as the matching threshold for the following evaluation metrics, i.e., i and j are matched if $\mathrm{OLS}(i,j)>0.5$.

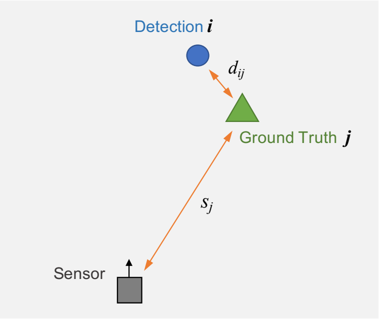


Figure 3. Illustration of the proposed point-based similarity OLS between a detection i and a ground truth j.

Fig. 3 is an illustration of the above OLS metric, where the blue dot represents a detection point and green triangle represents a ground truth. The key parameters are shown in Fig. 3.