2021 IEEE Intelligent Transportation Systems Conference (ITSC)
Indianapolis, USA. September 19-21, 2021

2021 IEEE International Intelligent Transportation Systems Conference (ITSC) | 978-1-7281-9142-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/ITSC48978.2021.9564904

3D Detection and Tracking for On-road Vehicles with a Monovision Camera and Dual Low-cost 4D mmWave Radars

Hang Cui*†, Junzhe Wu†, Jiaming Zhang, Girish Chowdhary and William R. Norris

Abstract— **High resolution 4D millimeter wave radar has been increasingly used for robust 3D detection and tracking of on-road vehicles. Rich point clouds generated by 4D radars can not only provide more reliable detection in harsh weather environments, but also offers 3D tracking capabilities for on-road objects. In this paper, a convolutional neural network (CNN) with cross fusion strategy is proposed for 3D on-road vehicle detection. The trained CNN model was also tested with dual low-cost 4D millimeter wave radars and a single monovision camera. An extended version of radar-camera calibration in three dimensions and 3D tracking with an extended Kalman filter (EKF) were also presented. The detection results showed that the proposed convolutional neural network model outperformed the one used on the Astyx dataset which provided up to 1500 radar detection points, on average, per frame.**

Keywords - 4D radar, 3D detection and tracking, EKF, CNN

I. Introduction

Millimeter wave (mmWave) radar has been used in advanced driver assistance systems (ADAS) for years due to advantages from immunity to adverse weather conditions, such as fog, rain, snow, dust, and glare, as well as the capability to estimate target velocity with Doppler measurements. Successful applications, both in stand-alone or sensor fusion scenarios, have shown its robustness and reliability in accurate measurements of distances and relative velocities (using Doppler measurements) of detected objects in harsh environments [1]. Recently, TI released their low-cost high resolution 77 GHz frequency modulated continuous wave (FMCW) radar chips which offer 4 GHz bandwidth and range resolution up to 3.75 cm [2]. Different from the CW and 24 GHz radar sensor, a FMCW radar increases its reliability by modulating the transmitted signal with a predefined sequence. The reduced wavelength enables smaller multiple input multiple output (MIMO) antenna arrays while providing better performance for environmental scanning. The state-of-the-art TI AWR1843 chipset [3] has three transmitters and four receivers for object detection in 4 dimensions, namely the x, y, z positions in space and the Doppler velocity of the detected object. The principle of FMCW radar is to transmit electromagnetic waves whose frequencies are linearly increased with time, also known as chirps, and to calculate objects' range, velocity, and arrival of angle (AoA). The calculations are performed with range-FFT (1D), Doppler-FFT (2D), and angle-FFT across multiple receiver antennas, respectively. Unlike other commercial 77 GHz automotive mmWave radars [4]–[6], users have absolute source code level control of the TI AWR1843 radars' programming and configurations except for the locked front-end radio subsystem. This allows for developers to customize their radar chirp profiles, frame profiles or apply advanced radar techniques, such as beam steering, according to their applications. With the real-time DCA1000EVM data capture card [7], developers can also record the raw ADC data for post-processing via the LVDS interface of the TI radars.

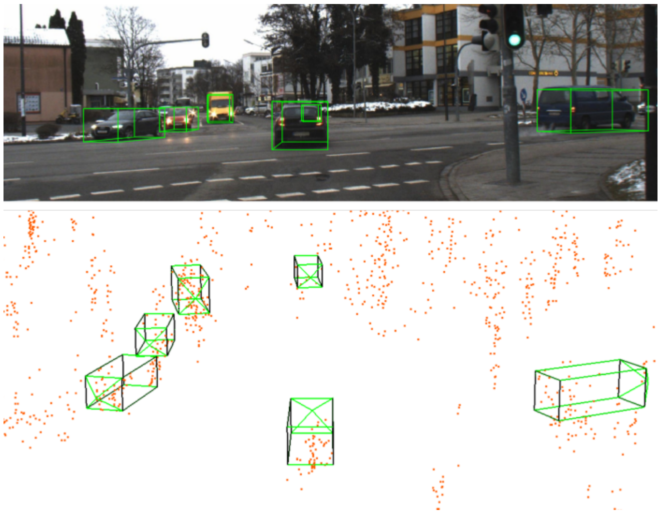

Fig. 1. An example of 3D detection results from the proposed convolutional neural network model of this paper on the Astyx dataset [8] [9].

mmWave radars have been successfully used in ADAS, such as adaptive cruise control (ACC), automatic emergency braking (AEB) and lane change assist (LCA). However, they usually can not generate enough dense point clouds for object recognition and classification compared with the LiDAR sensors. Classification tasks like identifying humans, motorcycles, or vehicles could not be done by only inspecting the radar cross section (RCS) of the detected objects. Instead, radar data fused with the data from other sensors, such as cameras and LiDAR, can lead to improvements in both accuracy and efficiency for object detection and tracking. However, LiDAR is not widely used in cars currently due to its high cost. Furthermore, the cost presents a significant barrier for adoption beyond current advanced autonomous vehicles where it is employed. Whereas, mmWave radars are commonly installed on modern cars and work in adverse

H. Cui and J. Zhang are with the Mechanical Science & Engineering Department, J. Wu and G. Chowdhary are with the Agricultural & Biological Engineering Department, W. Norris is with the Industrial & Enterprise Systems Engineering Department, University of Illinois Urbana-Champaign, Champaign, IL 61820, USA. †These authors contributed equally. *Corresponding author

lighting and weather conditions. Given these advantages, there has been a large amount of research focused on the methods for fusing monocular vision data with mmWave radar to improve object detection and tracking performance. [10], [11] shows a radar-camera sensor fusion method inspired by a human vision system with a vehicle running up to 30 km/h. [12] applied deep learning techniques, such as YOLO v2 for vehicle detection, fused the camera results with radar data. [13], [14] used a radar region proposal network which mapped radar detection points to an image coordinate system to generate proposals based on an objects' distances. [9], [15] presented how bird's eye view (BEV) images and radar point clouds can be used as inputs to neural networks to predict 3D bounding boxes. Instead of using peak detection from radar point clouds, [16] used image-like tensors for vehicle detection. [17] demonstrated the early fusion of radar and camera data on a deep learning architecture to enhance the accuracy and robustness for vehicle detection. In this paper, the rich point clouds generated by 4D imaging radars were further explored for on-road vehicle 3D detection and tracking.

The main contributions of this paper include: (1) the presentation of a customized low-cost automotive mmWave radar configurations based on application requirements. (2) the implementation of an extended version for radar-camera calibration in three dimensions and 3D tracking with an extended Kalman filter (EKF). (3) a proposed convolutional neural network model with cross fusion strategy to integrate multi-modal features for on-road vehicle 3D detection. The rest of the paper is organized as follows. In Section II, the perception system framework is introduced, including the radar subsystem, camera subsystem and radar-camera sensor calibration. In Section III, the vehicle detection and tracking framework and associated techniques are discussed. In Section IV, the experimental results are discussed. In Section V, a conclusion and future work are provided.

II. Perception System Framework

The perception system consisted of a single monovision camera mounted on the front roof of the vehicle and dual TI AWR1843 radars mounted on each side of the camera and separated by 36 cm as shown in Fig. 2. The camera was connected to a PC with a USB3.1 cable running at 30 fps with a resolution of 2304 × 1536. Each radar was connected to the PC with a UART-to-USB cable at a baud rate of 921600 for radar data transmission.

A. Radar and Camera Subsystems

The TI AWR1843Boost EVM radar had on board etched antennas with three transmitters and four receivers and a built-in phase lock loop (PLL) to enable detecting and tracking multiple objects with their distances and angles [18]. The AWR1843 radars were configured as the ultra-short-range radars (USRR) with detection range up to 30 m and a field of view of 120 degrees in the azimuth plane. The USRR alternates chirps in a frame transition with three transmitters, namely TX1, TX2, and TX3, to improve the angle resolution by a factor of three. The MIMO configuration synthesizes an array of twelve virtual RX antennas in the corresponding planes (8 virtual RX antennas in the azimuth plane and 4 virtual RX antennas in the elevation plane) [19].

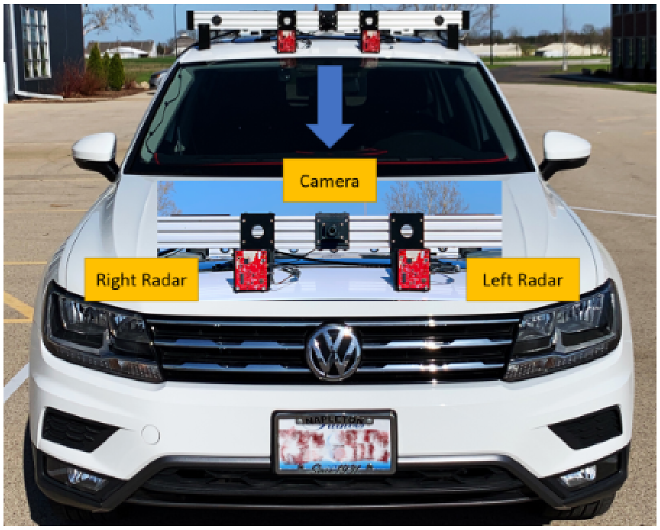

Fig. 2. An overview of sensor configurations

As shown in Fig. 3, the integrated AWR1843 FMCW radar had built-in analog to digital converters (ADC) with a max intermediate frequency (IF) of 10 MHz. It is capable of operation the radar transceiver in the 76-81 GHz operating frequency. Developers can choose either a narrow bandwidth of 1 GHz (76-77 GHz) or a wide bandwidth of 4 GHz (77-81GHz) according to their application scenarios. The AWR1843 chipset also integrates the master subsystem (MSS, Cortex-R4F), digital signal processing (DSS, C674x DSP) subsystem and radar hardware accelerator. These components are responsible for radar configuration and control, radar detection algorithm implementation, and radar signal processing, such as FFT, CFAR-CA. A second dedicated Cortex-R4F MCU was used as the radio processor subsystem for continuous monitoring, RF calibration, and self-test. The CAN and CAN-FD interfaces were provided on this single chip radar sensor for automotive applications.

With the help of the TI mmWave sensing estimator [20], developers can easily calculate the chirp parameters based on their radar application requirements. This includes max range, max velocity, range resolution, velocity resolution, and then fine tune them during the tests. A set of chirps in sequence forms a frame which was used for periodic radar object detection. Different types of chirps can co-exist in a frame for advanced sub-frame configurations. For this application, the maximum detectable range was 30 m, the range resolution was 0.12 m, the max velocity was 30 km/h, the velocity resolution was 2 km/h, and the measurement rate was 30 Hz. The typical detectable object chosen was a car which usually had an RCS of 5 m^2. The typical FMCW chirp and its configuration parameters are shown in Fig. 4. The configuration parameters shown in Table I. were calculated based on the application requirements,

TABLE I
CHIRP CONFIGURATION PARAMETERS

Parameter	USRR
Start frequency (GHz)	77
Frequency slope (MHz/μs)	46.83
ADC sample frequency (ksps)	10407
Number of ADC samples	277
Idle time (μs)	5
ADC valid start time (μs)	3.9
Ramp end time (μs)	31.52
Frame periodicity (ms)	33.33
Number of transmitters, receivers	3,4

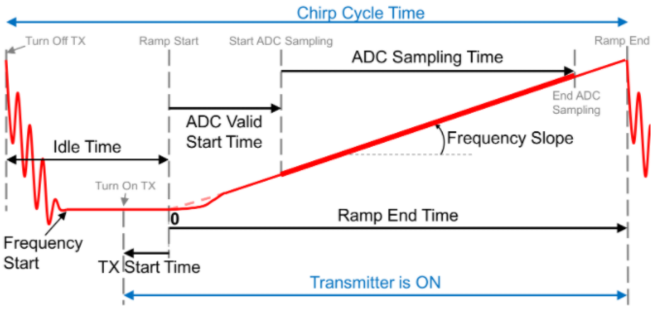

Fig. 4. The chirp schematic of the TI FMCW radar [20]

and how each of them influences the performance of the radar system. According to the mmWave sensing estimator user's guide [20], the chirp configuration parameters were calculated under the assumptions that the maximum idle time was 7 μs, the maximum ADC valid start time was 12.2 μs, and the sampling frequency was set to the minimum possible within the device limits.

The single monovision camera used for image capturing was the NileCAM30_USB 3.4MP GMSL camera. It had a wide field of view of 118 degrees with an image format in UYVY or MJPEG. The resolution was configured as 2304 × 1536 running at 30 fps.

B. Sensor Calibration

Sensor calibration was an essential step before sensor fusion. Accurate and careful radar-camera calibration can ensure good object detection and tracking performance. For this application, the radar and camera were individually calibrated first. Then the radar and camera radar-image data pairs were jointly calibrated using a direct linear transformation (DLT) [21].

1) Radar and Camera Individual Calibration: The TI AWR1843 radar sensor has an internal processor for the calibration routines and self-monitoring to stabilize the radar front-end performance across temperature ranges [22]. The mechanisms of calibration and monitoring were implemented using a combination of TI hardware and firmware. As for the object range detection calibration, a strong corner reflector was used with an edge length of 196 mm and effective RCS of 26.21 dBsm at 5 m to 30 m away from the AWR1843 radar. As a result, the detected ranges of the reflector were accurate enough, so range bias was not used. A 9×7 chessboard with 108 mm squares was used to calculate the distortion coefficients and the camera matrix of the monovision camera.

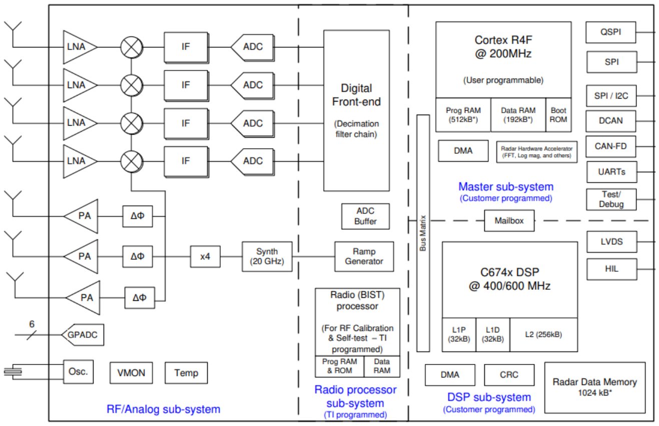

Fig. 3. The functional block diagram of the AWR1843 chipset [3]

$$\sigma = \frac{4\pi a^4}{3\lambda^2} \tag{1}$$

Where a is the edge length of the corner reflector, λ is the wave length and σ is the cross section area.

2) Radar and Camera Joint Calibration: According to the results from [21], DLT was used with pre-conditioning normalization as the radar-camera joint calibration method. Different patterns of the five strong reflectors were set at different ranges and heights for the calibration, shown in Fig. 5. For better performance, the left radar and right radar with the monovision camera were calibrated individually and 100 radar-camera data pairs were collected at different ranges and heights of the strong corner reflector for both radars.

Since the mmWave radars used in this application had x, y, z dimensions in the 3D space, the DLT method described in [21] was extended for the radar-camera calibration from 2 dimensions to 3 dimensions. First, denote $\{\mathbf{p}_i = [x_i, y_i, z_i]^T\}_{i=1}^N$ and $\{\mathbf{q}_i = [u_i, v_i]^T\}_{i=1}^N$ as N data pairs from radar and camera, respectively. The goal of this calibration was to estimate a three-dimensional projection transformation which maps the radar data to a two dimensional image.

$$\bar{\mathbf{q}} = \mathbf{H}\bar{\mathbf{p}} \tag{2}$$

Where $\bar{\mathbf{p}} = [x, y, z, 1]^T$ and $\bar{\mathbf{q}} = [u, v, 1]^T$ are the homogeneous coordinates of the radar and image points, respectively. H is the transformation matrix, denoted as $\mathbf{H} = [h_{ij}]_{3\times4}$.

From (2), a radar point, $p_i = [x_i, y_i, z_i]^T$ can be mapped to an image point $q_i = [u_i, v_i]^T$, shown in formula (3)

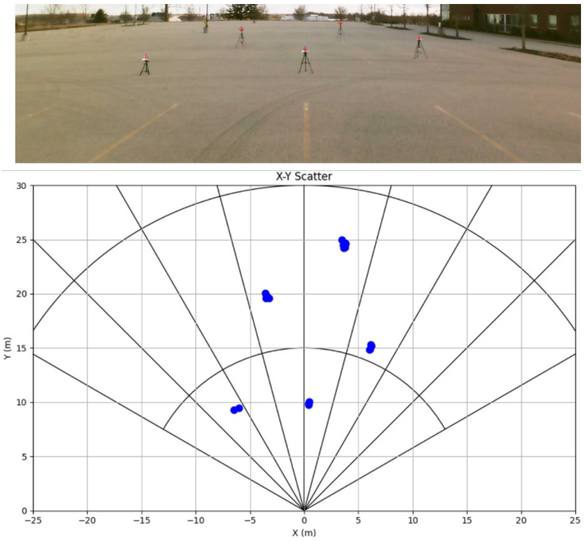

Fig. 5. The calibration of a single monovision camera and dual radars

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} = \begin{bmatrix} \frac{h_{11}x_i+h_{12}y_i+h_{13}z_i+h_{14}}{h_{31}x_i+h_{32}y_i+h_{33}z_i+h_{34}} \\ \frac{h_{21}x_i+h_{22}y_i+h_{23}z_i+h_{24}}{h_{31}x_i+h_{32}y_i+h_{33}z_i+h_{34}} \end{bmatrix} \tag{3}$$

If $\mathbf{h} = [h_{11}, h_{12}, h_{13}, h_{14}, h_{21}, h_{22}, h_{23}, h_{24}, h_{31}, h_{32}, h_{33}, h_{34}]^T$, the formula (3) can be modified to formula (4)

$$\mathbf{A}_i\mathbf{h} = 0 \tag{4}$$

Where

$$\mathbf{A}_i = \begin{bmatrix} -x_i & -y_i & -z_i & -1 & 0 & 0 & 0 & 0 & u_ix_i & u_iy_i & u_iz_i & u_i \\ 0 & 0 & 0 & 0 & -x_i & -y_i & -z_i & -1 & v_ix_i & v_iy_i & v_iz_i & v_i \end{bmatrix}$$

From the collected radar-camera data pairs, the optimization problem becomes (5)

$$\underset{\|\mathbf{h}\|_2=1}{argmin} \|\mathbf{Ah}\|_2^2 \tag{5}$$

Where $\mathbf{A} = [\mathbf{A}_1^T, \ldots, \mathbf{A}_N^T]^T$, in our case $N = 100$, and $\|\cdot\|$ is the Euclidean norm of a vector.

Pre-conditioning normalization was recommended to be used together with DLT. $\mathbf{T}_p$ and $\mathbf{T}_q$ were denoted as the normalization matrix for $\bar{\mathbf{p}}_i$ and $\bar{\mathbf{q}}_i$, where

$$\widetilde{\mathbf{p}}_i = \mathbf{T}_p\bar{\mathbf{p}}_i = \begin{bmatrix} \sqrt{2}/d_1 & 0 & 0 & -\sqrt{2}m_1/d_1 \\ 0 & \sqrt{2}/d_1 & 0 & -\sqrt{2}m_2/d_1 \\ 0 & 0 & \sqrt{2}/d_1 & -\sqrt{2}m_3/d_1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \bar{\mathbf{p}}_i \tag{6}$$

$$\widetilde{\mathbf{q}}_i = \mathbf{T}_q\bar{\mathbf{q}}_i = \begin{bmatrix} \sqrt{2}/d_2 & 0 & -\sqrt{2}n_1/d_2 \\ 0 & \sqrt{2}/d_2 & -\sqrt{2}n_2/d_2 \\ 0 & 0 & 1 \end{bmatrix} \bar{\mathbf{q}}_i \tag{7}$$

Where m_1, m_2, and m_3 are the elements of the mean vector $\mathbf{m} = \frac{\sum_i \mathbf{p}_i}{N}$, n_1, and n_2 are the elements of the mean vector $\mathbf{n} = \frac{\sum_i \mathbf{q}_i}{N}$, d_1 and d_2 are the average of the norms of $\{\mathbf{p}_i - \mathbf{m}\}_{i=1}^N$ and $\{\mathbf{q}_i - \mathbf{n}\}_{i=1}^N$, respectively. After performing the normalization using $\mathbf{T}_p$ and $\mathbf{T}_q$, DLT was applied to the normalized radar-camera data pairs $\{\widetilde{\mathbf{p}}_i\}_{i=1}^N$ and $\{\widetilde{\mathbf{q}}_i\}_{i=1}^N$ to obtain $\widetilde{\mathbf{H}}$.

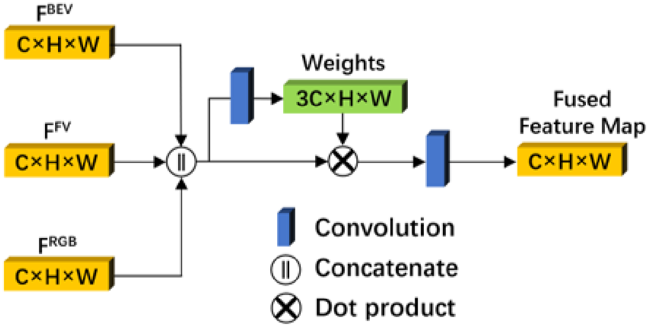

Fig. 6. SSMA block: modality-specific feature maps were concatenated and fed into convolutional layer to generate the weight matrix. The dot product of the weight matrix and the input concatenated matrix was fed into another convolutional layer to compute the fused feature map. [23]

Then, an arbitrary detected radar point $[x_i, y_i, z_i]^T$ was mapped to a 2D image with formula (8).

$$\begin{bmatrix} \widehat{u} \\ \widehat{v} \end{bmatrix} = \begin{bmatrix} \frac{\widehat{h}_{11}x_i+\widehat{h}_{12}y_i+\widehat{h}_{13}z_i+\widehat{h}_{14}}{\widehat{h}_{31}x_i+\widehat{h}_{32}y_i+\widehat{h}_{33}z_i+\widehat{h}_{34}} \\ \frac{\widehat{h}_{21}x_i+\widehat{h}_{22}y_i+\widehat{h}_{23}z_i+\widehat{h}_{24}}{\widehat{h}_{31}x_i+\widehat{h}_{32}y_i+\widehat{h}_{33}z_i+\widehat{h}_{34}} \end{bmatrix} \tag{8}$$

where $\widehat{\mathbf{H}} = \mathbf{T}_q^{-1}\widetilde{\mathbf{H}}\mathbf{T}_p$

III. Methodology

In this application, dual radars were used to generate radar point clouds in 3D and Doppler velocities. A convolutional neural network was trained to detect cars in 3D space based on a multi-view representation of detected radar points and RGB images. An extended Kalman filter (EKF) was used for tracking the car in three dimensions running on the DSP subsystem of the radars.

A. Convolutional Neural Network for Detection

A similar neural network architecture as described in [9] was used in this research. However, the self-supervised model adaptation (SSMA) block [23] was used to fuse different feature maps at the pixel level and turned the deep fusion scheme in [9] into a cross fusion scheme. The radar point cloud from each frame was used to generate a front view (FV) image and a bird's eye view (BEV) image. A 3D region proposal network was then used to generate proposals based on RGB images and BEV images of the radar point clouds. These proposals were projected into three feature maps. In order to combine information from these three feature maps with the projected proposals, a cross fusion scheme was employed. The SSMA block shown in Fig. 6 was used to fuse multi-view feature maps at the pixel level. The architecture of the convolutional neural network is shown in Fig. 7. The output of the last SSMA block was used

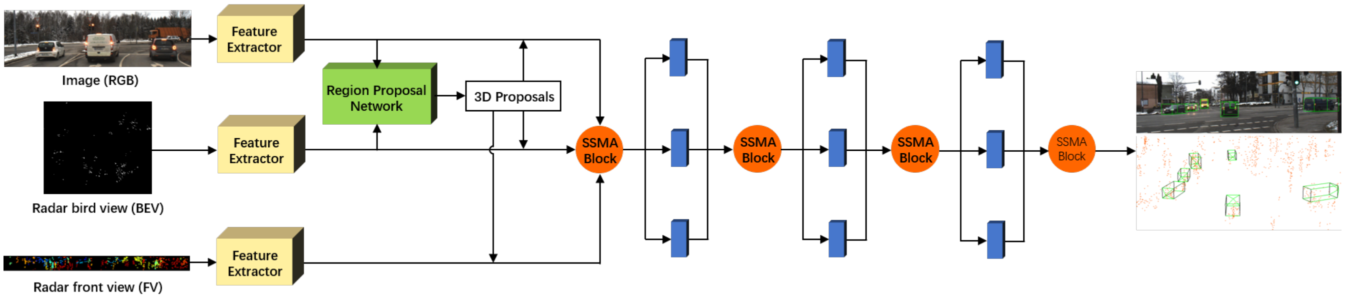

Fig. 7. Convolutional neural network structure: The feature extractors are shown in yellow, the region proposal network is shown in green, the SSMA blocks are shown in orange, and the convolutional layers are shown in blue.

to predict 3D bounding boxes encoded with four corners and two heights which represented the distance from the top plane and bottom plane of the bounding box to the ground plane [24]. The object orientations were computed from the predicted 3D box corners.

B. Dataset and Training

The dataset from [8] was used and split into a training set and a test set with a ratio of 4:1. The same data augmentation methods in [9] were used during training due to the small size of the dataset. These included horizontally flipping images, point clouds, ground truth boxes and adding noise to the camera images [25]. The trained weights using the camera and LiDAR data of the KITTI dataset were set as the initial weights for training the proposed convolutional neural network on the Astyx dataset. The network was trained for 30000 iterations with a learning rate of 0.0001 and a mini batch size of 1. The training took 5 hours and 25 minutes on one Nvidia GTX 1070 GPU.

C. Vehicle Tracking

The 3D tracking for on-road vehicles was implemented in the DSP subsystem of the TI AWR1843 radars. The radar detection points were clustered using DBSCAN [26] to get the center coordinates and radial velocities. The measurement vector $\mathbf{z}_t$ is given below.

$$\mathbf{z}_t = \begin{bmatrix} x_t & y_t & z_t & \dot{r}_t \end{bmatrix}^T \tag{9}$$

Given the estimated positions and velocities of multiple detected cars, the extended Kalman filter for 3D object tracking [27] was used.

The state vector μ_t is defined in (10).

$$\mu_t = \begin{bmatrix} x_t & y_t & z_t & \dot{x}_t & \dot{y}_t & \dot{z}_t \end{bmatrix}^T \tag{10}$$

The measurement vector was related to the state vector via

$$\mathbf{z}_t = \mathbf{H}(\mu_t) + \mathbf{v}_t \tag{11}$$

Where H is a non-linear transformation

$$\mathbf{H}(\mu_t) = \begin{bmatrix} x_t \\ y_t \\ z_t \\ \frac{x_t\dot{x}_t + y_t\dot{y}_t + z_t\dot{z}_t}{\sqrt{x_t^2 + y_t^2 + z_t^2}} \end{bmatrix} \tag{12}$$

And $\mathbf{v}_t = \begin{bmatrix} v_t^x & v_t^y & v_t^z & v_t^{\dot{r}} \end{bmatrix}$ is a vector of measurement noise with covariance matrix $\mathbf{R}$ given by $\mathbf{R} = \text{diag}\begin{bmatrix} \sigma_x^2 & \sigma_y^2 & \sigma_z^2 & \sigma_{\dot{r}}^2 \end{bmatrix}$.

The object movement with time is given in (13).

$$\mu_t = \mathbf{F}\mu_{t-1} + \mathbf{w}_t \tag{13}$$

Where

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & 0 & T & 0 & 0 \\ 0 & 1 & 0 & 0 & T & 0 \\ 0 & 0 & 1 & 0 & 0 & T \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{14}$$

And T is the sampling interval. The vector $\mathbf{w}_t$ represents process noise associated with covariance matrix Q and is given by

$$\mathbf{w}_t = \begin{bmatrix} w_t^x & w_t^y & w_t^z & w_t^{\dot{x}} & w_t^{\dot{y}} & w_t^{\dot{z}} \end{bmatrix}^T \tag{15}$$

The nonlinear function $\mathbf{H}(\cdot)$ maps μ_t to $\mathbf{z}_t$ of (11) is simplified by using the first term in the Taylor series expansion of $\mathbf{H}(\mu_t)$; i.e.

$$\mathbf{z}_t = \mathbf{H}(\bar{\mu}_t) + \mathbf{J_H}(\bar{\mu}_t)(\mu_t - \bar{\mu}_t) + \mathbf{v}_t \tag{16}$$

Where $\mathbf{J_H}(\cdot)$ is the Jacobian matrix given in (17). The prediction and measurement update are shown from (18) to (22)

$$\mathbf{J_H}(\mu) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{\partial \dot{r}}{\partial x} & \frac{\partial \dot{r}}{\partial y} & \frac{\partial \dot{r}}{\partial z} & \frac{\partial \dot{r}}{\partial \dot{x}} & \frac{\partial \dot{r}}{\partial \dot{y}} & \frac{\partial \dot{r}}{\partial \dot{z}} \end{bmatrix} \tag{17}$$

Prediction:

$$\bar{\mu}_t = \mathbf{F}\mu_{t-1} \tag{18}$$

$$\bar{\mathbf{P}}_t = \mathbf{F}\mathbf{P}_{t-1}\mathbf{F}^T + \mathbf{Q} \tag{19}$$

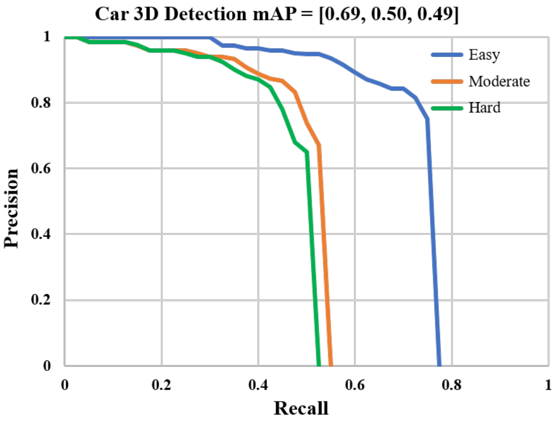

Fig. 8. Precision-recall curve for car 3D detection using radar and camera

TABLE II

COMPARISON OF PERFORMANCE

	Easy		Moderate		Hard	
	AP	AHS	AP	AHS	AP	AHS
3DRC	61.00	N/A	48.00	N/A	45.00	N/A
Ours	**69.50**	68.16	**50.05**	48.59	**49.13**	47.61

Measurement update:

$$\mathbf{K}_t = \frac{\bar{\mathbf{P}}_t \mathbf{J}_{\mathbf{H}}^T(\bar{\mu}_t)}{\mathbf{J}_{\mathbf{H}}(\bar{\mu}_t)\bar{\mathbf{P}}_t \mathbf{J}_{\mathbf{H}}^T(\bar{\mu}_t) + \mathbf{R}} \tag{20}$$

$$\mu_t = \bar{\mu}_t + \mathbf{K}_t[\mathbf{z}_t - \mathbf{H}(\bar{\mu}_t)] \tag{21}$$

$$\mathbf{P} = [\mathbf{I} - \mathbf{K}_t \mathbf{J}_{\mathbf{H}}(\bar{\mu}_t)]\bar{\mathbf{P}}_t \tag{22}$$

IV. EXPERIMENTS AND RESULTS

The Astyx test set was used with the trained model for the evaluation of 3D on-road vehicle detection, shown in Fig. 1. Rich radar point clouds collected per frame were converted into FV and BEV images together with the RGB images as the inputs for the proposed convolutional neural network. The pyramid attached to the detected 3D bounding box indicated the forward direction of vehicle. For a better evaluation, ground truth was split into three categories, namely easy, moderate and hard [9]. The easy category implies that only fully visible cars were evaluated, whereas in the hard category all of the cars were evaluated. For the moderate category, fully occluded cars were excluded from evaluation. The test set was evaluated using 3D average precision (AP) and Average Heading Similarity (AHS) [15] at a 0.5 intersection over union (IoU) threshold. The comparison of the proposed convolutional neural network and 3DRC [9] is shown in Table II. The 3D APs of the proposed method were 69.50%, 50.05%, and 49.13% for easy, moderate, hard categories, respectively, as shown in Fig. 8. From the results, the presented method outperformed the 3DRC [9] by 9.50% in easy, 2.05% in moderate, and 4.13% in hard categories, respectively. This could be due to better initial training weights, additional SSMA blocks and radar FV images.

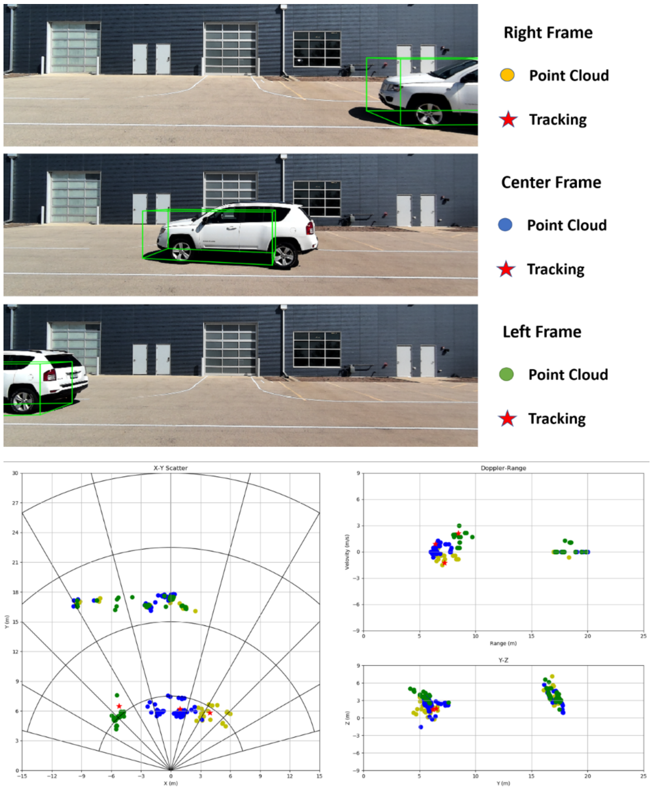

Fig. 9. The real scene tests of trained convolutional neural network model with dual low-cost radars and a single monovision camera. The red stars indicate the tracking results in the corresponding frame. Different colors were used for radar detection points in different frames.

The trained convolutional neural network model was used for testing with the dual low-cost radars and the single monovision camera. Three different frames are shown in Fig. 9 and around 100 valid radar detection points were generated per frame. The static objects between 15 and 18 meters were detected in all frames. In the right frame, the detected car was approaching the radars and in the left frame the detected car was moving away from the radars. The red stars in Fig. 9 indicate the 3D tracking results from the radars. However, there was no ground truth for tracking result analysis during the tests and it will be considered for future work. In Fig. 9, all the 3D bounding boxes have more than 60% overlap with the detected car. This indicates that the trained model from the high-resolution radar dataset, namely the Astyx dataset, can be applied to the low-cost radar-camera settings shown in Fig. 2. The overall 3D bounding results were not as good as those on the test set and this could be caused for several reasons. First, even though dual 4D radars were used to generate rich point clouds, the number of the detected points were still far less than the ones given in the Astyx dataset per frame. Second, the detection performance of each radar was customized by setting different parameter requirements on max range, range resolution, etc, which may not be consistent with the configurations of the radar used for collecting the Astyx dataset. Third, the angular resolutions in the azimuth and elevation plane of the TI AWR1843 radars were not as good as the HiRes 6455 radar due to hardware features.

V. Conclusion and Future Work

In this paper, a proposed convolutional neural network model with cross fusion strategy was used for 3D detection of on-road vehicles. Then, the trained model was tested with dual low-cost 4D mmWave radars and a single monovision camera. An extended version of radar-camera calibration in three dimensions and 3D tracking with EKF were also presented. The results of the proposed neural network on the test dataset showed promising detection accuracy using only inputs from 4D radars and a single monovision camera. The results of the trained model on real scene tests were not as good as those on the Astyx test set, which has been discussed in Section IV. In the future, a higher resolution 4D mmWave radar, for example, the TI AWR2243 cascaded imaging radar [28], will be used to collect more data for training and testing. The performance of 3D tracking using EKF will be analyzed with ground truth. The early fusion of radar-camera data, such as the raw ADC data or heat map data, will be explored on different neural network architectures.

ACKNOWLEDGMENT

The authors want to thank for Texas Instruments Inc, and Dan Block for providing the AWR1843BOOST EVMs for this research.

References

[1] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pp. 1–7, IEEE, 2019.

[2] S. Rao, "Intro to mmwave sensing: Fmcw radars." https://training.ti.com/intro-mmwave-sensing-fmcw-radars-module-1-range-estimation?context=1128486-1139153-1128542, 2017. Accessed: 2017-04-28.

[3] TI, *AWR1843 Single-Chip 77- to 79-GHz FMCW Radar Sensor*. Texas Instrument, 12500 TI Blvd. Dallas, TX 75266, 2020.

[4] APTIV, *Aptiv Electronically Scanning RADAR*. AutonomouStuff, 306 Erie Avenue Morton, IL 61550, 10 2020.

[5] C. E. Services, *ARS 408-21 Long Range Radar Sensor 77 GHz*. AutonomouStuff, 306 Erie Avenue Morton, IL 61550, 9 2018.

[6] Smarmicro, *AUTOMOTIVE SENSOR UMRR-11 TYPE 132*. AutonomouStuff, 306 Erie Avenue Morton, IL 61550, 1 2021.

[7] TI, *DCA1000EVM Data Capture Card User's Guide*. Texas Instrument, 12500 TI Blvd. Dallas, TX 75266, 5 2019.

[8] M. Meyer and G. Kuschk, "Automotive radar dataset for deep learning based 3d object detection," in *2019 16th European Radar Conference (EuRAD)*, pp. 129–132, IEEE, 2019.

[9] M. Meyer and G. Kuschk, "Deep learning based 3d object detection for automotive radar and camera," in *2019 16th European Radar Conference (EuRAD)*, pp. 133–136, IEEE, 2019.

[10] X. Wang, L. Xu, H. Sun, J. Xin, and N. Zheng, "Bionic vision inspired on-road obstacle detection and tracking using radar and visual information," in *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 39–44, IEEE, 2014.

[11] X. Wang, L. Xu, H. Sun, J. Xin, and N. Zheng, "On-road vehicle detection and tracking using mmw radar and monovision fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 7, pp. 2075–2084, 2016.

[12] J. Kim, Ž. Emeršič, and D. S. Han, "Vehicle path prediction based on radar and vision sensor fusion for safe lane changing," in *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 267–271, IEEE, 2019.

[13] R. Nabati and H. Qi, "Rrpn: Radar region proposal network for object detection in autonomous vehicles," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3093–3097, IEEE, 2019.

[14] R. Nabati and H. Qi, "Radar-camera sensor fusion for joint object detection and distance estimation in autonomous vehicles," *arXiv preprint arXiv:2009.08428*, 2020.

[15] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–8, IEEE, 2018.

[16] B. Major, D. Fontijne, A. Ansari, R. Teja Sukhavasi, R. Gowaikar, M. Hamilton, S. Lee, S. Grzechnik, and S. Subramanian, "Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[17] T.-Y. Lim, A. Ansari, B. Major, D. Fontijne, M. Hamilton, R. Gowaikar, and S. Subramanian, "Radar and camera early fusion for vehicle detection in advanced driver assistance systems," in *Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems*, 2019.

[18] TI, *xWR1843 Evaluation Module (xWR1843BOOST) Single-Chip mmWave Sensing Solution*. Texas Instrument, 12500 TI Blvd. Dallas, TX 75266, 5 2020.

[19] TI, *Getting Started Guide mmWave LAB Medium Range Radar*. Texas Instrument, 12500 TI Blvd. Dallas, TX 75266.

[20] TI, *MMWave Sensing Estimator*. Texas Instrument, 12500 TI Blvd. Dallas, TX 75266.

[21] J. Oh, K.-S. Kim, M. Park, and S. Kim, "A comparative study on camera-radar calibration methods," in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 1057–1062, IEEE, 2018.

[22] TI, *Self-Calibration in TI's mmWave Radar Devices*. Texas Instrument, 12500 TI Blvd. Dallas, TX 75266, 6 2018.

[23] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *International Journal of Computer Vision*, pp. 1–47, 2019.

[24] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1907–1915, 2017.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[26] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, pp. 226–231, 1996.

[27] M. Z. Ikram and M. Ali, "3-d object tracking in millimeter-wave radar for advanced driver assistance systems," in *2013 IEEE Global Conference on Signal and Information Processing*, pp. 723–726, IEEE, 2013.

[28] TI, *AWR2243 Single-Chip 76- to 81-GHz FMCW Transceiver*. Texas Instrument, 12500 TI Blvd. Dallas, TX 75266, 2 2020.