Computer Science and Application 计算机科学与应用, 2022, 12(11), 2544-2560
Published Online November 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1211260

Hans汉斯

一种基于摄像头和毫米波雷达的多模态信息融合算法

刘振东，宋春林

同济大学信息与通信工程系，上海

收稿日期：2022年10月17日；录用日期：2022年11月15日；发布日期：2022年11月22日

摘　要

环境感知是高级驾驶辅助系统(Advanced Driver Assistance Systems, ADAS)的关键环节，而摄像头和毫米波雷达是环境感知的核心传感器。利用多源传感器不同模态的互补信息，可以显著提升车辆的自主感知能力，帮助车辆更好地应对复杂场景下的目标检测任务。由于毫米波雷达点云的稀疏性，现有的摄像头和毫米波雷达融合算法，存在对毫米波雷达点云信息利用不充分、缺乏鲁棒性等特点。针对这些问题，本文提出了一种基于点云膨胀的数据级融合算法。算法首先使用提出的最近邻帧同步算法以及空间坐标映射进行多源数据时空对齐。之后，使用提出的基于点云膨胀的增强中心融合网络(Enhanced Center Fusion Net, ECFN)将映射至像素坐标系的毫米波雷达数据进行特征增强，并引入1 × 1的卷积核对输入数据进行降维，实现跨通道的信息交互。此外，ECFN还在损失函数中引入新的速度和深度因子来增强神经网络对雷达点云信息的利用。实验结果表明，增强的中心融合网络ECFN在推理时间稍有增加的情况下，平均精度优于基于单源传感器的算法以及现有的多源融合网络。

关键词

摄像头，毫米波雷达，多模态信息融合，高级驾驶辅助，深度学习

Multi-Modal Information Fusion Algorithm Based on Camera and Millimeter Wave Radar

Zhendong Liu, Chunlin Song

Department of information and Communication Engineering, Tongji University, Shanghai

Received: Oct. 17th, 2022; accepted: Nov. 15th, 2022; published: Nov. 22nd, 2022

Abstract

Environmental awareness is the key link of Advanced Driver Assistance Systems (ADAS), and camera

文章引用: 刘振东, 宋春林. 一种基于摄像头和毫米波雷达的多模态信息融合算法[J]. 计算机科学与应用, 2022, 12(11): 2544-2560. DOI: 10.12677/csa.2022.1211260

and millimeter wave radar are the core sensors of environmental awareness. Using the complementary information of different modes of multi-source sensors can significantly improve the autonomous sensing ability of vehicles, and help vehicles better cope with the target detection task in complex scenes. Due to the sparsity of the millimeter wave radar point cloud, the existing camera and millimeter wave radar fusion algorithms do not make full use of the millimeter wave radar point cloud information and lack of robustness. Aiming at these problems, this paper proposes a data-level fusion algorithm based on point cloud expansion. The algorithm first uses the proposed nearest neighbor frame synchronization algorithm and spatial coordinate mapping to align multi-source data in time and space. Then, the proposed Enhanced Center Fusion Net (ECFN) based on point cloud expansion is used to enhance the features of millimeter wave radar data mapped to the pixel coordinate system, and the 1 × 1 convolution kernel is introduced to reduce the dimension of the input data to realize cross-channel information interaction. In addition, ECFN also introduces new velocity and depth factors into the loss function to enhance the use of radar point cloud information by neural networks. The experimental results show that the average accuracy of the enhanced central fusion network ECFN is better than that of the algorithm based on a single source sensor and the existing multi-source fusion network when the inference time is slightly increased.

Keywords

Camera, Millimeter Wave Radar, Multi-Modal Information Fusion, Advanced Driver Assistance Systems (ADAS), Deep Learning

1. 引言

高级驾驶辅助系统要求车辆具有应对极端情况和全天候环境的能力[1]，由于单传感器在相关场景下的性能存在局限，自动驾驶汽车通常配备不同类型的传感器。多源传感器融合可以利用不同传感器的信息进行互补，提升感知系统的鲁棒性和安全性，但是同时也给设计环境感知系统带来了新的挑战。目前，车辆环境感知常用激光雷达和视觉融合的方案，但是由于激光雷达成本高昂，且摄像头和激光雷达均在恶劣天气下性能衰减明显，也都无法提供目标速度信息，因此使用场景受限。对于毫米波雷达和摄像头的组合，其可以在获得目标丰富的语义信息的同时，感知物体的距离和速度。与摄像头信息相比，毫米波雷达的检测性能受极端天气和光照的影响较小，可以全天候工作，且信息处理所需算力更低，距离和径向速度估计更精确。而摄像头在具有更高的角分辨率和检测精度的同时，可以提供目标轮廓、纹理和颜色分布等语义信息，从而帮助目标分类[2]。

由于摄像头图像数据和毫米波雷达点云数据模态差异很大，因此，如何有效融合多模态信息是一个很大的挑战。现有多模态信息融合方案主要在数据级、特征级和决策级三个方向进行。其中，决策级融合是目前主流信息融合方案。决策级融合首先对多源传感器分别进行信息处理，即先将毫米波雷达信息处理成一个包含物体速度和距离等信息的目标列表[3] [4]，并将视觉信息执行目标检测算法生成包含 2D 位置的目标列表。之后，对多源传感器目标列表信息进行滤波匹配，融合决策信息[5]。当前决策融合主流滤波算法包括贝叶斯理论[6] [7]、卡尔曼滤波算法[8] [9] [10]、模糊子集理论方法和基于证据理论的推理方法[11]等。此外，文献[12] [13]使用雷达信息来验证视觉的检测结果。决策级融合算法可以减小对单传感器感知结果的依赖，增强系统鲁棒性。但是由于其未充分利用多源传感器不同模态的互信息，因此，

并不能显著地提升系统性能。此外，在特征级进行多模态信息融合的方案也被广泛提出。基于卷积神经网络的特征级融合常通过将雷达检测信息转化为图像形式，使用额外的雷达输入分支来辅助目标检测模型对图像特征信息进行学习。特征融合可以使用级联和元素相加等方法[14]，级联将图像特征矩阵和雷达点云特征矩阵进行连接，形成多通道矩阵。而元素相加将多源特征矩阵合并为一个特征矩阵。此外，文献[15]使用空间注意力融合(Space Attention Fusion, SAF)机制生成注意力权重矩阵，融合多源传感器特征，提升检测性能。对于多源传感器的数据级融合，现有算法常利用毫米波雷达生成图像的可分辨单元，再从融合数据中提取信息，用于进一步决策。但由于毫米波雷达点云的稀疏性，现有多源异构融合算法存在融合效率低，雷达信息表达不充分等问题[16]。

针对上述问题，本文提出了一种毫米波雷达点云信息和摄像头视觉信息融合的算法。算法首先使用设计的最近邻帧匹配算法，实现数据帧对齐。其次，对于毫米波雷达点云稀疏的问题，提出了点云膨胀算法，增强毫米波雷达的特征表达。之后，对现有融合网络进行了改进，提出增强中心融合网络，加入多模态信息数据级融合，增强毫米波雷达和摄像头的信息交互。

2. 毫米波雷达与摄像头融合简介

2.1. 毫米波雷达

毫米波雷达是工作在毫米波波段探测的雷达，其频率范围为 30~300 GHz。其主要通过区分发射波与接收回波，通过飞行时间(Time of Flight, ToF)法获取目标距离信息，并利用多普勒原理获取目标的径向速度信息[17]。

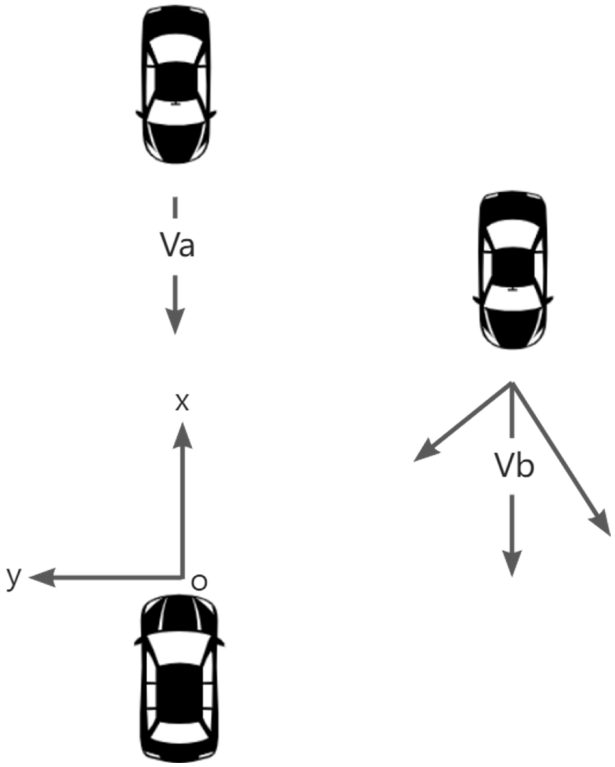

Figure 1. Millimeter wave radar receives objects velocity
图 1. 毫米波雷达获取目标速度

毫米波雷达通常在 BEV 中提供目标的方位角和径向距离、径向速度以及 RCS 等信息。如图 1 所示，

毫米波雷达只能获取车辆坐标系中目标的径向速度，切向速度分量则无法获取。

本文所使用的毫米波雷达传感器的输出维度如表 1 所示。

Table 1. Millimeter wave radar sensor output dimension
表 1. 毫米波雷达传感器输出维度

值	描述
x	目标的 X 维度值，前方为正方向(m)
y	目标的 Y 维度值，左方为正方向(m)
v_x	目标的 X 方向速度值，前方为正方向(m/s)
v_y	目标的 Y 方向速度值，左方为正方向(m/s)
rcs	雷达目标有效截面值

2.2. 摄像头

摄像头通过采集光学图像，为车辆提供视觉信息，是车辆环境感知的核心传感器之一，主要用于目标识别、环境地图构建、车道线检测和目标跟踪等任务。由于其出色的颜色感知能力和较高的角分辨率，为目标分类提供了丰富的语义信息。目前，主要有 CCD 和 CMOS 两种传感器的相机，前者在在噪声，

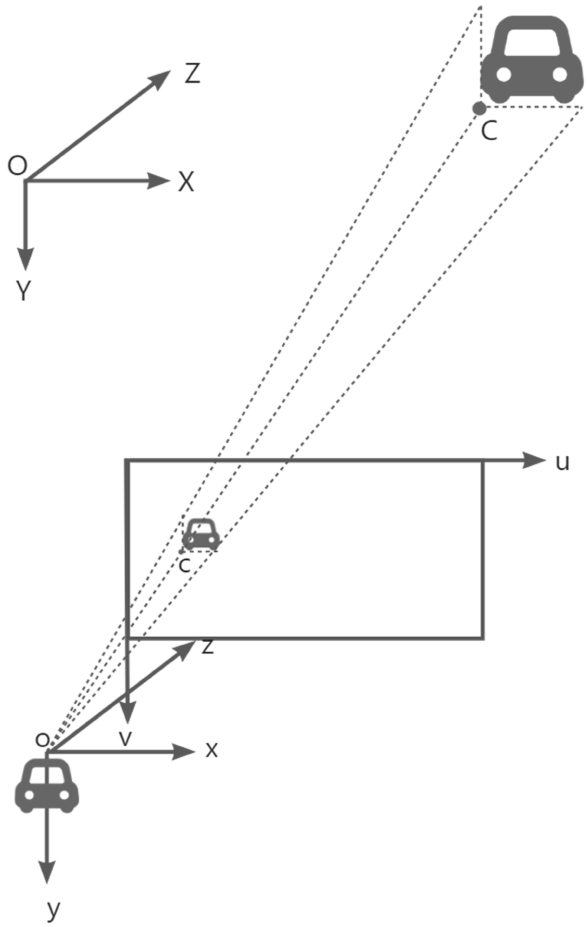

Figure 2. Image information acquisition
图 2. 图像信息获取

动态范围和复杂环境下的可靠性上更具优势，后者则在牺牲一定性能的条件下降低了成本[18]。由于单摄像头无法进行三维感知，其信息仅限于二维图像平面。如图 2 所示，为了从图像信息中获取目标的位置信息，需要进行相机标定，以建立像素坐标系和世界坐标系之间的转换关系。

2.3. 多源信息融合方式

目前常用的毫米波雷达点云信息和摄像头视觉信息数据融合方式包括数据集融合，特征及融合以及决策级融合[2]。

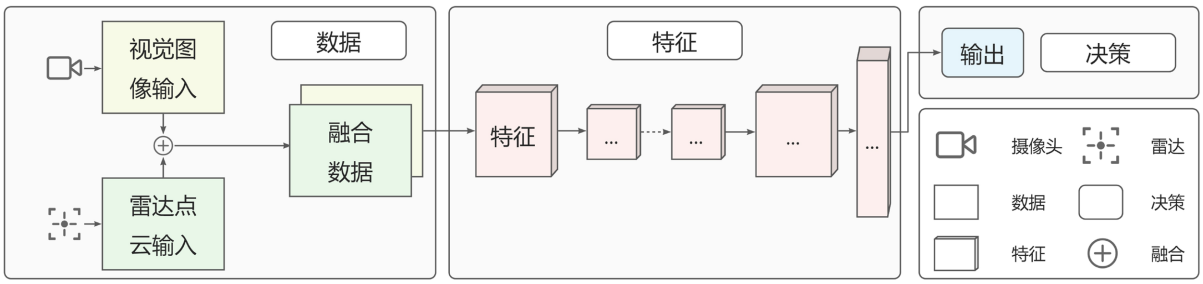

Figure 3. Data-level fusion
图 3. 数据级融合

如图 3 所示，数据级融合是指点云数据与图像数据在数据输入层进行信息融合，之后，再对融合数据进行特征提取和决策输出。由于神经网络输入包含多模态数据完整信息，因此，数据级融合具有最低的信息损耗率和最高的可靠性。

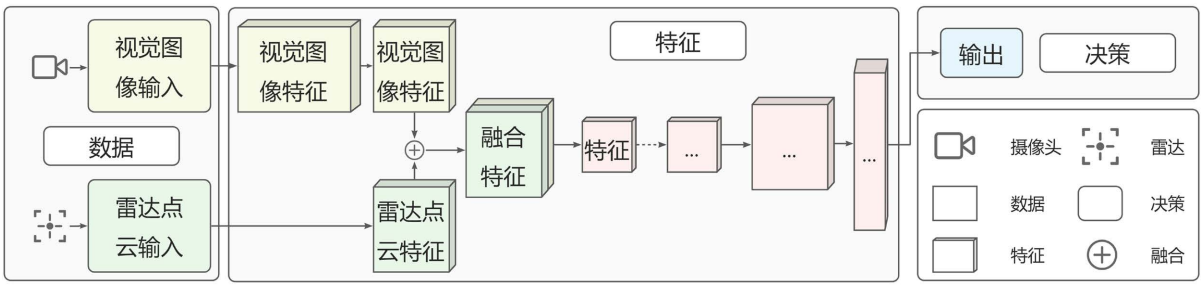

Figure 4. Feature-level fusion
图 4. 特征级融合

如图 4 所示，特征级融合通常分别提取雷达的特征信息和图像的特征信息，并同时输入检测网络，对融合特征进行训练学习，优势在于检测模型可以同时学习多源特征。

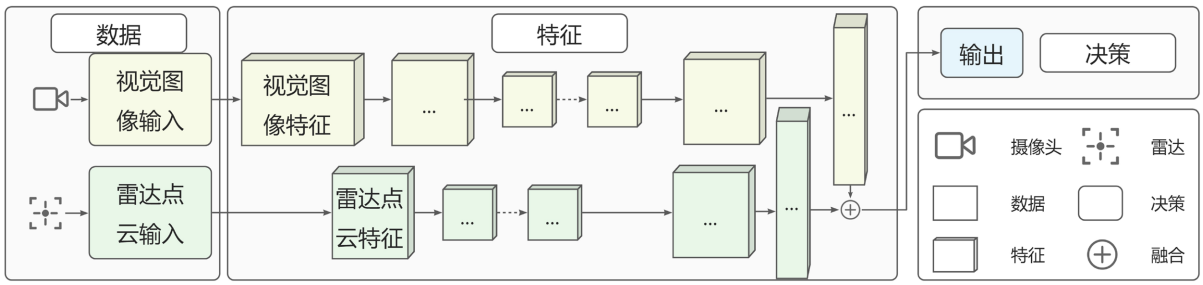

Figure 5. Decision-level fusion
图 5. 决策级融合

如图 5 所示，决策级融合通常对毫米波雷达和摄像机的检测结果信息进行融合决策，可以兼顾雷达的深度信息以及摄像头的视野和分辨率优势。决策级融合的难点在于建模多源传感器不同模态信息的联合概率密度函数。

3. 一种基于点云膨胀的多模态信息融合算法

基于上述问题，本文提出一种基点云膨胀的毫米波雷达和视觉信息融合算法。算法首先对多源异构信息进行时空对齐，之后，使用点云膨胀算法对毫米波雷达点云数据进行数据增强，采用数据级融合制作毫米波雷达和视觉的联合表征，融合多源异构信息，并使用提出的增强中心融合网络进行检测，生成决策信息。

3.1. 多源异构信息时空对齐

由于雷达点云信息和视觉图像信息无法直接融合，因此，算法首先需要对数据进行时空对齐，保证数据的时空一致性。数据的时空对齐包含时间维度的帧同步以及空间维度的坐标变换。其中，对于帧同步，本文设计了基于最近邻帧匹配算法的帧同步机制，并对其帧匹配误差进行了分析。

3.1.1. 基于最近邻帧匹配算法的帧同步

由于不同传感器的时间分辨率差异，多源传感器信息融合首先需要对数据帧进行同步。本文提出基于时间戳的带阈值的最近邻帧匹配算法，对采集的图像帧数据和毫米波雷达点云帧数据进行时间同步。算法流程如图 6 所示，分别提取雷达帧喝图像帧以及其时间戳，并减去对应的帧平均时延，之后，进行最近邻帧匹配。

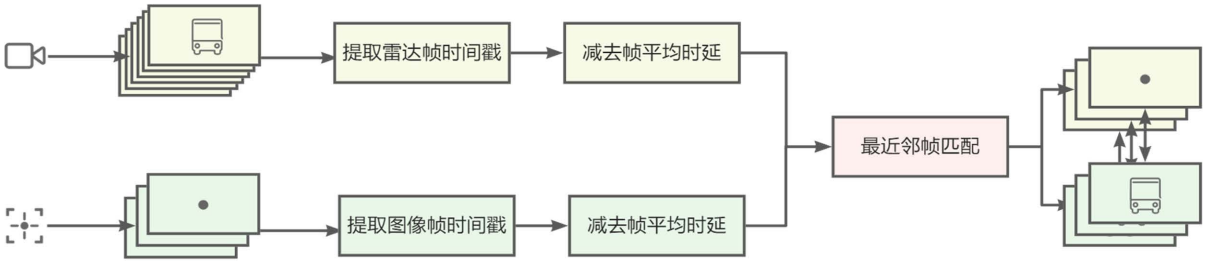

Figure 6. Flow chart of frame synchronization of nearest neighbor frame matching algorithm
图 6. 最近邻帧匹配算法帧同步流程图

设图像数据帧集合定义为：

$$C=\{c_1,c_2,c_3,\cdots,c_m\} \tag{1}$$

雷达数据帧集合定义为：

$$R=\{r_1,r_2,r_3,\cdots,r_n\} \tag{2}$$

其中，图像数据帧和雷达点云真集合时间戳分别为：

$$T_c=\{t_{c1},t_{c2},t_{c3},\cdots,t_{cm}\},T_r=\{t_{r1},t_{r2},t_{r3},\cdots,t_{rn}\} \tag{3}$$

由于摄像机时钟和毫米波雷达时钟与主机时钟并非完全同步，存在延时。因此，设摄像机时钟与主机时钟平均延时为 δ_C，毫米波雷达时钟与主机时钟平均延时为 δ_R。

由于图像数据帧率 f_c 和毫米波雷达帧率 f_r 的差异，选取较低帧率传感器，其帧率记为 $f_{\min}$，设置帧时差阈值 T_{th} 表示帧同步时间精确度，满足

$$\delta \le \frac{1}{2\times f_{\min}} \tag{4}$$

使用滑动窗口法对图像帧数据和毫米波雷达帧数据进行融合，生成结果集。

即对于 $\forall R_j \in R$：

$$O(C,R) = O(C,R) + (C_i, R_j) \ \text{ if } \left|T_{C_i} - T_{R_j}\right| < \delta \tag{5}$$

对于结果集 $O(C,R)$，其中，每组数据 (C_i, R_j)，都是同步时间误差小于帧时差阈值的数据组。

3.1.2. 多模态数据空间对齐

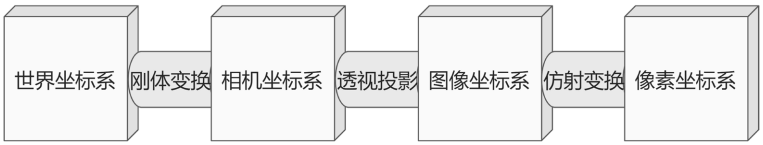

Figure 7. Spatial coordinate system relation diagram
图 7. 空间坐标系关系图

在摄像头的成像系统中，包含世界坐标系、相机坐标系、图像坐标系、像素坐标系这四个坐标系[19]。如图 7 所示，世界坐标系与像素坐标系之间的转化关系为：

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{dX} & -\frac{\cot\theta}{dX} & u_0 \\ 0 & \frac{1}{dY\sin\theta} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} U \\ V \\ W \\ 1 \end{bmatrix} \tag{6}$$

其中，(U,V,W) 是世界坐标系下某一点的物理坐标，(u,v) 是映射至像素坐标系下的像素坐标。s 是尺度因子。

令矩阵：

$$M_1 = \begin{bmatrix} \frac{1}{dX} & -\frac{\cot\theta}{dX} & u_0 \\ 0 & \frac{1}{dY\sin\theta} & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \frac{f}{dX} & -\frac{f\cot\theta}{dX} & u_0 & 0 \\ 0 & \frac{f}{dY\sin\theta} & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{7}$$

则，称矩阵 M_1 为相机的内参矩阵。其中，f 为像距，dX, dY 分别表示 X, Y 方向上的一个像素在相机感光板上的物理长度，(u_0, v_0) 表示光轴与成像平面的交点坐标，θ 表示感光板的横边和纵边之间的角度。

令矩阵：

$$M_2 = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \tag{8}$$

则，称矩阵 M_2 为相机的外参矩阵。其中，R 表示旋转矩阵，T 表示平移矢量。其值取决于世界坐标系和相机坐标系的相对位置。

此外，由于摄像头这种精密光学器件可能由于内部和外部的一些原因存在着畸变。畸变包括径向畸变和切向畸变，其中，三阶径向畸变公式为：

$$\begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix}=\left(1+k_1r^2+k_2r^4+k_3r^6\right)\begin{bmatrix}x\\ y\end{bmatrix} \tag{9}$$

三阶切向畸变公式为：

$$\begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix}=\begin{bmatrix}x\\ y\end{bmatrix}+\begin{bmatrix}2p_1y+p_2\left(r^2+2x^2\right)\\ p_1\left(r^2+2y^2\right)+2p_2x\end{bmatrix} \tag{10}$$

其中，$(\hat{x},\hat{y})$为畸变后归一化图像坐标，(x,y)为无畸变归一化图像坐标。r为像素到图像中心点距离。

因此，为了避免图像数据源的误差，需要对摄像头的参数进行标定，以获取相机的内参矩阵、外参矩阵以及畸变参数。

3.2. 多源异构信息融合

数据融合首先要将毫米波雷达帧数据和摄像头帧数据经过帧同步后筛选出同步帧数据。然后，再将毫米波雷达帧数据经过滤波和坐标变换后映射至像素坐标系。

由于毫米波雷达点云相较于图像像素具有稀疏的特性，给特征提取造成了困难。因此，本文提出了基于毫米波雷达信息的点云膨胀算法，用来在毫米波雷达数据映射到图像上之后，增强毫米波雷达的特征表达。

3.2.1. 雷达点云坐标变换

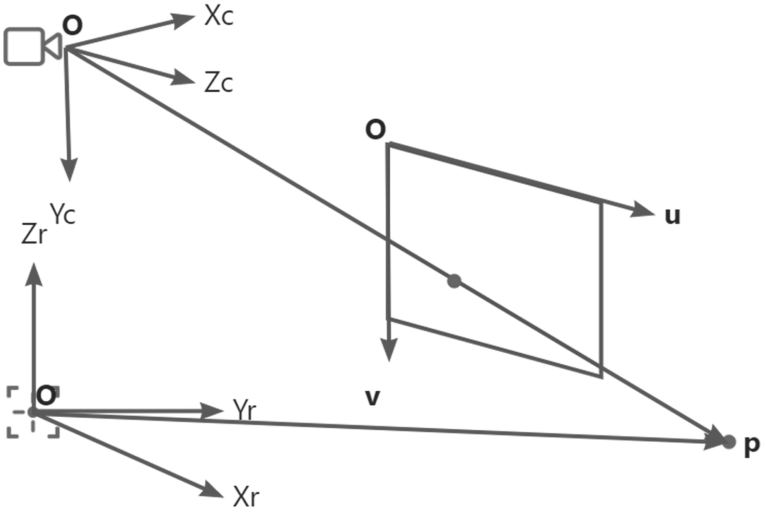

Figure 8. Radar and camera coordinate mapping
图 8. 雷达和摄像头坐标映射

雷达、摄像头与像素坐标系的映射关系[20]如图 8 所示。

(x_c,y_c,z_c)和(x_r,y_r,z_r)分别是毫米波雷达和摄像头的坐标系，(u,v)为像素坐标系。

则毫米波雷达和摄像机与像素之间的变换矩阵为：

$$s\begin{bmatrix}u\\ v\\ 1\end{bmatrix}=\begin{bmatrix}a_x & 0 & u_0 & 0\\ 0 & a_y & v_0 & 0\\ 0 & 0 & 1 & 0\end{bmatrix}\begin{bmatrix}R & t\\ 0^{\mathrm{T}} & 1\end{bmatrix}\begin{bmatrix}x_r\\ y_r\\ z_r\\ 1\end{bmatrix}=M_1M_2\begin{bmatrix}x_r\\ y_r\\ z_r\\ 1\end{bmatrix} \tag{11}$$

其中，M_1和M_2分别代表内参矩阵和外参矩阵。

由于毫米波雷达只提供x_r和y_r两个坐标方向的值，z_r值恒为0，因此，可以简化上式为：

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_r \\ y_r \\ 1 \end{bmatrix} - T \begin{bmatrix} x_r \\ y_r \\ 1 \end{bmatrix} \tag{12}$$

使$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}'$，$V = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}'$，$I = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}'$，$p = \begin{bmatrix} x_r^1 & y_r^1 & 1 \\ \vdots & \vdots & \vdots \\ x_r^n & y_r^n & 1 \end{bmatrix}$，其中，$n$是雷达点云数。转换矩阵$T$可以由最小二乘法得出：

$$T = \begin{bmatrix} \left(\left(P^{\mathrm{T}}P\right)^{-1}P^{\mathrm{T}}U\right)^{\mathrm{T}} \\ \left(\left(P^{\mathrm{T}}P\right)^{-1}P^{\mathrm{T}}V\right)^{\mathrm{T}} \\ \left(\left(P^{\mathrm{T}}P\right)^{-1}P^{\mathrm{T}}I\right)^{\mathrm{T}} \end{bmatrix} \tag{13}$$

3.2.2. 基于点云膨胀算法的数据增强

摄像头所获得的图像信息由于在恶劣天气环境下缺乏鲁棒性，而毫米波雷达信息可以为这种环境为摄像头提供很好的互补信息[21]。为了融合多模态信息，算法需要将毫米波雷达信息映射到图像中，便于后续特征学习。

毫米波雷达信息不包含目标语义信息，无法判断目标形态，为了利用毫米波雷达信息，算法使用雷达点云坐标变换，将毫米波雷达点云变换至图像坐标系，并且增加图像通道，将雷达通道目标RCS信息ρ，和距离信息d以及速度信息的正交分量v_x,v_y投影至像素坐标系中，反应为目标的大小。这样可以让网络在获得雷达信息的同时，不丢失图像的语义信息。

由于毫米波雷达点云具有稀疏的特性，因此使用点云膨胀算法，对雷达点云信息进行膨胀，以增强其在后续目标检测中的特征表达。对毫米波雷达点云信息在图像坐标系的W、H和T三个维度进行膨胀。在T维度的膨胀实现上采用多帧毫米波雷达点云图像进行数据增强，在W和H维度上的膨胀实现上采用在像素平面膨胀目标的宽和高。

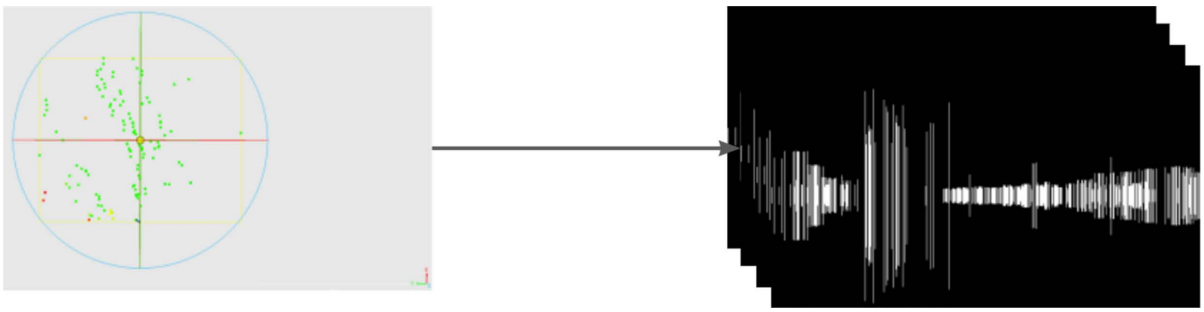

Figure 9. Radar point cloud expansion
图9. 雷达点云膨胀

如图9所示为毫米波雷达点云膨胀前后的效果。由于点云膨胀前后的x维度中心点未发生变化，因此保留的毫米波雷达的角度信息。而点云膨胀后的长宽反映了目标的深度信息。

3.2.3. 增强的中心融合网络 ECFN

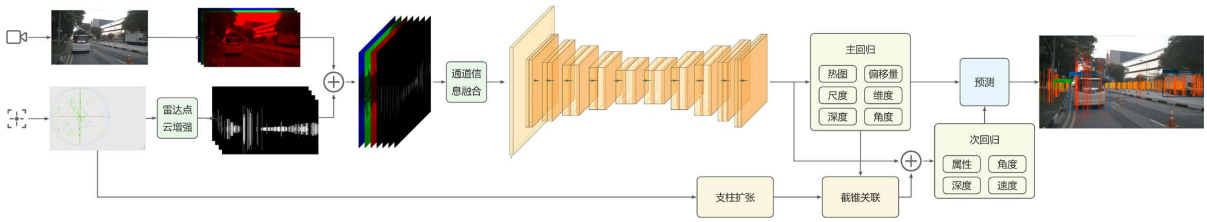

Figure 10. Enhance center fusion net
图 10. 增强中心融合网络

基于深度学习的纯视觉目标检测网络通常以图像 $I \in \mathbb{R}^{W\times H\times 3}$ 作为网络输入。其中，W 和 H 分辨是图像的宽和高的像素值。网络输出则为 $\hat{Y} \in [0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$，其中，$R$ 是网络的下采样率，C 是目标类别数量。其中，预测值 $\hat{Y}_{x,y,c}=p$ 代表对于在 (x,y) 处存在目标 c 的概率是 p。其中，生成特征图的真实值 $Y \in [0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$ 由图像数据的标注生成。网络的代价函数通常描述为 $C=\frac{1}{m}\sum_{i=1}^{m}L\left(\hat{Y}^{(i)},Y^{(i)}\right)$，其中，$L$ 为损失函数。

中心点网络(CenterNet)是一个基于视觉的目标检测器[22]。与 YOLO，SSD，Faster_RCNN 等依靠大量锚点的检测器不同，CenterNet 是一种无锚点目标检测网络。CenterNet 使用目标中心点来表示目标，并使用中心点的偏移量，宽高来得到物体实际的框。CenterNet 使用热图来表示分类信息，每一个类别对应一张热图，使用高斯圆来表示关键点，热图中的高斯圆代表坐标处有目标的中心点。

$$L=L_k+\lambda_s L_s+\lambda_o L_o \tag{14}$$

CenterNet 的损失函数包含三部分，其中，L_k 代表热图的损失，L_s 代表目标长宽预测的损失，L_o 代表中心点偏移值的损失。λ_s，λ_o 分别代表长宽损失和中心点偏移损失的权重。

对于热图损失的计算公式 L_k，表示为：

$$L_k=\frac{-1}{N}\sum_{xyc}\begin{cases}\left(1-\hat{Y}_{xyc}\right)^{\alpha}\log\left(\hat{Y}_{xyc}\right), & \text{if } Y_{xyc}=1\\ \left(1-Y_{xyc}\right)^{\beta}\left(\hat{Y}_{xyc}^{\alpha}\right)\log\left(1-\hat{Y}_{xyc}\right), & \text{otherwise}\end{cases} \tag{15}$$

其中，α 和 β 是用来均衡难易样本和正负样本的超参数，N 是图像关键点数量。求和下标 x,y,c 表示热图上的所有坐标点，其中，c 是目标类别也就是热图数量，x,y 是热图长宽。$\hat{Y}$ 是预测值，Y 是真实值。

对于目标长宽预测损失 L_s，表示为：

$$L_o=\frac{1}{N}\sum_{p}\left\|\hat{O}_{\tilde{p}}-\left(\frac{p}{R}-\tilde{p}\right)\right\|_1 \tag{16}$$

其中，$\hat{O}_{\tilde{p}}$ 表示预测的偏移值，p 为图片中目标中心点坐标，R 为缩放尺度，$\tilde{p}$ 为缩放后中心点的近似整数坐标。

$$L_s=\frac{1}{N}\sum_{k=1}^{N}\left\|\hat{s}_{pk}-s_k\right\|_1 \tag{17}$$

其中，s_{pk} 为预测尺寸，s_k 为真实尺寸。

针对多模态传感器数据源，中心融合网络(CenterFusion) [23]基于 CenterNet 提出了一种特征级的融合算法。算法通过图像创建候选框，并利用相关联的雷达点云信息来调整预测。

CenterFusion 使用 CenterNet 来预测目标的中心点和目标的 3D 边界框。首先，使用改进的深层聚合网络(Deep Layer Aggregation, DLA)作为主干网进行图像特征提取。然后，将提取的特征用于预测目标中心点，以及目标框。

在 CenterNet 基础上，需要将雷达点云信息与图像相关联。CenterFusion 使用目标的 2D 边界框以及其深度来创建 3D 的 ROI，并筛选出目标 ROI 中的雷达点云数据。由于 ROI 中可能存在多个点云，CenterFusion 将最近的点云信息作为输入，并与图像目标相关联，以辅助检测网络的训练。

CenterFusion 相比基于相机的算法显著提升了精度。但是，CenterFusion 存在以下问题：

1、CenterFusion 仅使用雷达模态信息进行辅助训练，以改善相机的检测效果。辅助网络只选取 ROI 区域内最近的点云信息对图像候选框进行辅助训练，对雷达点云信息利用不充分，可能在恶劣天气下存在着性能衰减的问题。

2、CenterFusion 未充分利用雷达点云的 rcs 等通道信息。

针对以上问题，本文对 CenterFusion 进行改进，使用改进后的增强中心融合网络(Enhance CenterFusion Net, ECFN)进行目标检测。

如图 10 所示，为了融合毫米波雷达感知信息，ECFN 首先对毫米波雷达和摄像头帧信息进行时空对齐。毫米波雷达信息被映射到图像上，表示为新增的通道。与摄像机获取的视觉信号相比，毫米波雷达点云数据非常稀疏。为了加强雷达信号在网络中的特征表达，ECFN 首先采用点云膨胀算法对毫米波雷达点云信号进行雷达通道数据增强。

为了融合多通道信息，以便于后续的特征学习，ECFN 以 $I_f \in \mathbb{R}^{W\times H\times(3+n_r)}$ 为输入，其中，n_r 为毫米波雷达信息通道数。受 Inception 网络[24]中 1 × 1 卷积核降维的方法启发，ECFN 在网络检测头处使用 3 个 $(3+n_r)\times1\times1$ 卷积核降维。1 × 1 卷积核最早出现在 Network In Network [25]中，用来加深加宽网络结构。后续在 Inception 中，使用 1 × 1 卷积核在降低大量运算的前提下，降低了特征维度。ECFN 使用 3 个 $(3+n_r)\times1\times1$ 卷积核进行通道压缩，降低特征维度。

ECFN 对 CenterFusion 的损失函数进行了改进，引入毫米波雷达的信息。单依赖视觉图像信息无法对目标速度预测，毫米波雷达可以依赖多普勒效应检测目标径向速度信息。目标速度的预测值 $\hat{v}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times C_v}$，$C_v$ 是雷达点云信息中速度的维度。ECFN 采用速度的 L1 正则损失作为损失函数，表示为：

$$L_v=\frac{1}{N}\sum_{k=1}^{N}\left\|\hat{v}_{\hat{x}_k,\hat{y}_k}-v_k\right\|_1 \tag{18}$$

其中，$v_k\in\mathbb{R}^3$ 是目标的实际速度。

对于目标深度信息，其预测值 $\hat{d}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}}$，ECFN 采用深度的 L2 正则损失作为损失函数，表示为：

$$L_d=\frac{1}{N}\sum_{k=1}^{N}\left\|\hat{d}_{\hat{x}_k,\hat{y}_k}-d_k\right\|_2 \tag{19}$$

其中，$d_k\in\mathbb{R}^3$ 是目标的实际深度。

则最终损失函数为：

$$L_e=L_c+\lambda_v L_v+\lambda_d L_d \tag{20}$$

其中，λ_v，λ_d 分别代表速度损失和深度损失的权重，由于目标深度信息与坐标信息不独立，因此 λ_d 取较小值。

4. 融合结果与分析

本文对所提算法在如表 2 所示硬件设备上进行实验，并分析帧同步算法误差和召回率，以及对摄像

机标定结果的展示，并对本文所提多源异构信息融合网络 ECFN 进行测试分析。

4.1. 多源异构融合实验环境

本文所涉及硬件型号及其制造商如表 2 所示。

Table 2. Hardware equipment and manufacturer
表 2. 硬件设备及制造商

设备	型号(制造商)
摄像头	ARS408-21(Continental AG)
毫米波雷达	LI-USB30-AR023ZWDR(Leopard Imaging)
显卡	GTX2070s(Nvidia)
CPU	Core™ i9-9900K(Intel®)

本文所涉及相关环境版本信息如表 3 所示。

Table 3. Relevant environment information
表 3. 相关环境版本信息

相关环境	版本
Ubuntu	18.04LTS
Python	3.7
Ros	Melodic
Pytorch	1.10
Cuda	11.3
Cudnn	8.2
OpenCV	4.5

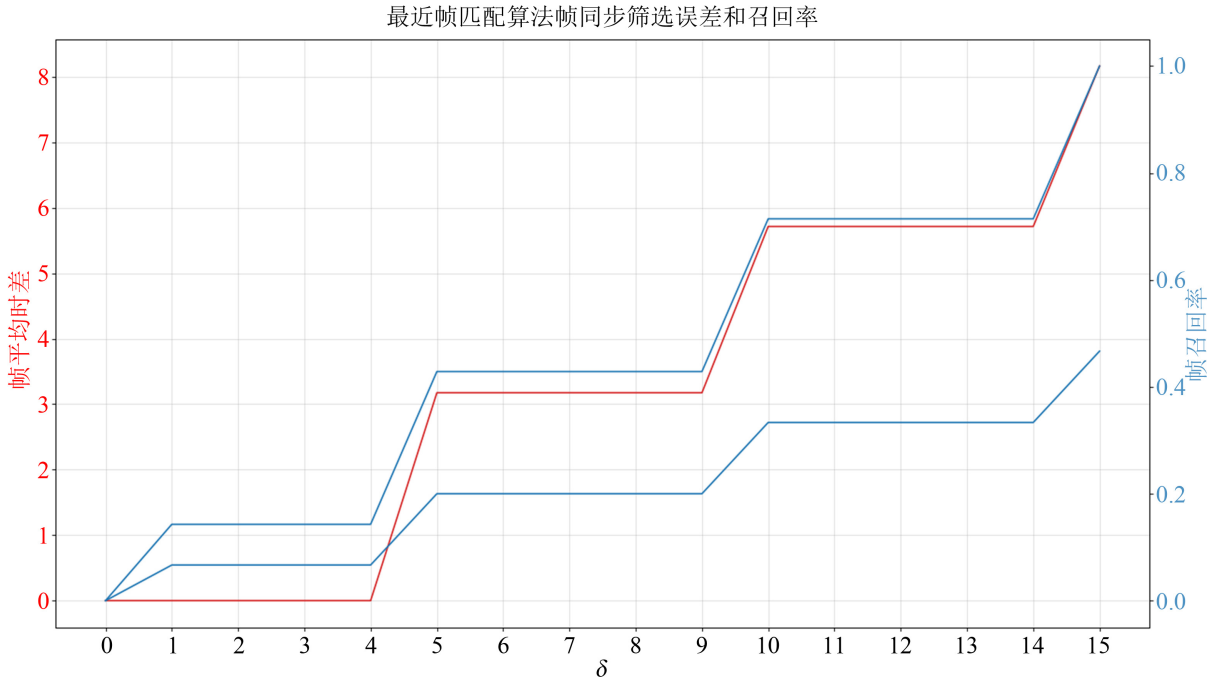

Figure 11. Frame matching algorithm frame synchronization filtering error and recall rate
图 11. 帧匹配算法帧同步筛选误差和召回率

4.2. 最近邻帧匹配算法帧同步分析

采用帧率为 30 fps 的摄像头和帧率为 14 fps 的毫米波雷对本文所提出的最近邻帧匹配算法进行测试分析。同步筛选后的误差和召回率如图 11 所示。

随着最大帧时差 T_{th} 的增加，摄像头数据帧和毫米波雷达数据帧召回率都随之增大，但是与此同时，帧平均时差也随之上升。本实验最大帧时差选择 10 ms，此时，帧平均时差小于 6 ms，且毫米波雷达帧召回率大于 55%。

4.3. 摄像机标定结果分析

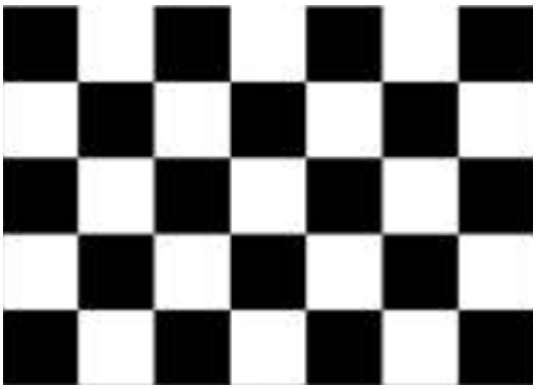

Figure 12. Checkerboard
图 12. 棋盘格

本文采用张正友标定法对相机进行标定。如图 12 所示，使用 6 × 4 大小棋盘格，并使其在摄像头视野范围内的不同位置进行拍摄，得到一组图像。如图 13 所示，对图像中的特征点如标定板角点进行检测，得到标定板角点的像素坐标值，根据已知的棋盘格大小和世界坐标系原点，计算得到标定板角点的物理坐标值。并求解得内参矩阵、外参矩阵和畸变参数矩阵。

如图 13 所示为相机标定校正前(左)和相机标定校正后(右)的拍摄结果。

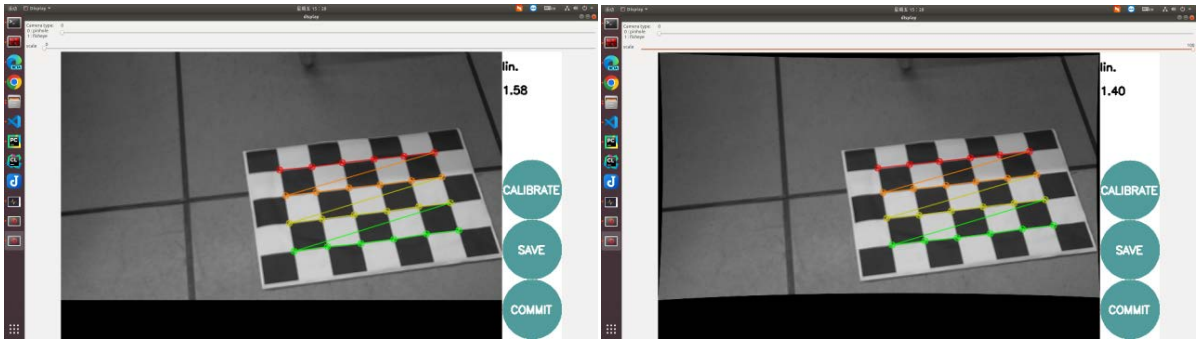

Figure 13. Camera calibration
图 13. 相机标定

4.4. EFCN 检测结果分析

采用公开数据集纽森(nuScenes) [26]对算法现有融合算法 CenterFusion 和本文提出的融合算法 ECFN 进行测试。

测试指标的平均精度(Average Precision, AP)的阈值匹配使用地面上的 2D 中心距离 d。全类平均正确率(mean Average Precision, mAP)类似于 2D 目标检测中的 AP 度量，但匹配策略被从 IoU 替换为 BEV 平面上的 2D 中心距离。如式 21 所示，mAP 是通过平均分类 $\mathbb{C}$ 以及 $\mathbb{D}=\{0.5,1,2,4\}$ 这四种不同距离阈值下的 AP 来计算的。

$$\mathrm{mAP}=\frac{1}{|\mathbb{C}||\mathbb{D}|}\sum_{c\in\mathbb{C}}\sum_{d\in\mathbb{D}}\mathrm{AP}_{c,d} \tag{21}$$

对于所有真阳性(TruePositive，TP)度量，使用 $d=2m$ 的中心距离计算。度量标准如表 4 所示。

Table 4. True positive metrics
表 4. 真阳性度量

度量标准	度量含义
平均平移误差(Average Translation Error, ATE)	二维欧式中心距离
平均尺度误差(Average Scale Error, ASE)	角度对齐后的三维交并比
平均角度误差(Average Orientation Error, AOE)	预测值和真实值之间的最小偏航角差
平均速度误差(Average Velocity Error, AVE)	二维速度差的 L2 范数
平均属性误差(Average Attribute Error, AAE)	1 减去属性分类精度

对于每一个 TP 度量，使用公式 22 计算其全类平均真阳性率(mean True Positive, mTP)。

$$\mathrm{mTP}=\frac{1}{|\mathbb{C}|}\sum_{c\in\mathbb{C}}\mathrm{TP}_c \tag{22}$$

纽森检测分数(nuScenes Detection Score，NDS)是几个指标的组合，其计算方式如式 23 所示：

$$\mathrm{NDS}=\frac{1}{10}\left[5\times\mathrm{mAP}+\sum_{\mathrm{mAP}\in\mathbb{TP}}\left(1-\min\left(1,\mathrm{mTP}\right)\right)\right] \tag{23}$$

其中，$\mathbb{TP}$ 是五个 mTP 度量的集合。通过 mAP 度量和 mTP 度量的加权和来计算 NDS。mAP 的权重为 5，其余 mTP 度量的权重为 1。

使用 CenterFusion 对 nuScenes 数据集的测试集进行测试，结果如表 5 所示。

Table 5. The existing fusion algorithm CenterFusion test results
表 5. 现有融合算法 CenterFusion 测试结果

类别	AP	ATE	ASE	AOE	AVE	AAE
车辆	0.502	0.480	0.160	0.128	0.538	0.870
卡车	0.430	0.641	0.172	0.134	0.114	0.991
公交车	0.576	0.449	0.093	0.123	4.621	0.012
拖车	0.000	1.000	1.000	1.000	1.000	1.000
工程车	0.000	1.000	1.000	1.000	1.000	1.000
行人	0.466	0.600	0.260	0.813	0.896	0.186
摩托车	0.277	0.877	0.330	1.190	0.066	1.000
自行车	0.198	0.542	0.329	0.999	3.512	0.418
交通锥	0.636	0.398	0.344	/	/	/
交通关卡	0.000	1.000	1.000	1.000	/	/

使用本文提出网络 ECFN 对 nuScenes 数据集的测试集进行测试，结果如表 6 所示。

Table 6. The proposed algorithm ECFN test results
表 6. 本文算法 ECFN 测试结果

类别	AP	ATE	ASE	AOE	AVE	AAE
车辆	0.504	0.481	0.156	0.094	0.548	0.872
卡车	0.401	0.624	0.148	0.158	0.119	0.989
公交车	0.529	0.605	0.099	0.120	4.588	0.013
拖车	0.000	1.000	1.000	1.000	1.000	1.000
工程车	0.000	1.000	1.000	1.000	1.000	1.000
行人	0.464	0.588	0.266	0.419	0.887	0.184
摩托车	0.311	0.659	0.340	0.969	0.061	1.000
自行车	0.245	0.659	0.251	0.555	3.456	0.400
交通锥	0.646	0.340	0.276	/	/	/
交通关卡	0.000	1.000	1.000	1.000	/	/

现有算法 CenterFusion 以及本文改进算法 ECFN 的 mAP 及 NDS 分数结果如表 7 所示。

Table 7. Comparison of test results between the proposed algorithm ECFN and the existing fusion algorithm CenterFusion
表 7. 本文算法 ECFN 与现有融合算法 CenterFusion 测试结果对比

算法	mAP	mATE	mASE	mAOE	mAVE	mAAE	NDS
CenterFusion	0.3085	0.6989	0.4687	0.7097	1.4682	0.6846	0.2981
ECFN (本文算法)	0.3101	0.6956	0.4536	0.5907	1.4575	0.6824	0.3128

网络的平均推理时间如表 8 所示。

Table 8. Comparison of running time between the proposed algorithm ECFN and the existing fusion algorithm CenterFusion
表 8. 本文算法 ECFN 与现有融合算法 CenterFusion 运行时间对比

算法	运行时间(ms)	时间增量
CenterFusion	48.1	-
ECFN(本文算法)	49.4	2.7%

测试结果表明，本文提出的基于摄像头和雷达的信息融合算法 ECFN 在公开数据集 nuScenes 上相较于现有融合算法 CenterFusion，在时间耗费增长 2.7%的情况下，mAP 提升了 0.52%，NDS 提升了 4.9%。

基金项目

本文受国家科技重大专项资助(项目编号：2017ZX05005001-005)。

参考文献

[1] 《中国公路学报》编辑部. 中国汽车工程学术研究综述·2017 [J]. 中国公路学报, 2017, 30(6): 1-197. https://doi.org/10.19721/j.cnki.1001-7372.2017.06.001

[2] Wei, Z., Zhang, F., Chang, S., *et al*. (2022) MmWave Radar and Vision Fusion for Object Detection in Autonomous

Driving: A Review. *Sensors*, **22**, Article 2542. https://doi.org/10.3390/s22072542

[3] Langer, D. and Jochem, T. (1996) Fusing Radar and Vision for Detecting, Classifying and Avoiding Roadway Obstacles. *Proceedings of Conference on Intelligent Vehicles IEEE*, Tokyo, 19-20 September 1996, 333-338. https://doi.org/10.1109/IVS.1996.566402

[4] Chavez-Garcia, R.O., Burlet, J., Vu, T.-D. and Aycard, O. (2012) Frontal Object Perception Using Radar and Mono-Vision. 2012 *IEEE Intelligent Vehicles Symposium*, Madrid, 3-7 June 2012, 159-164. https://doi.org/10.1109/IVS.2012.6232307

[5] 张炳力, 詹叶辉, 潘大巍, 等. 基于毫米波雷达和机器视觉融合的车辆检测[J]. 汽车工程, 2021, 43(4): 478-484.

[6] Coué, C., Fraichard, T., Bessiere, P. and Mazer, E. (2002) Multi-Sensor Data Fusion Using Bayesian Programming: An Automotive Application. *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Lausanne, 30 September-4 October 2002, 141-146. https://doi.org/10.1109/IRDS.2002.1041379

[7] Kawasaki, N. and Kiencke, U. (2004) Standard Platform for Sensor Fusion on Advanced Driver Assistance System Using Bayesian Network. *IEEE Intelligent Vehicles Symposium*, Parma, 14-17 June 2004, 250-255. https://doi.org/10.1109/IVS.2004.1336390

[8] Ćesić, J., Marković, I., Cvišić, I. and Petrović, I. (2016) Radar and Stereo Vision Fusion for Multitarget Tracking on the Special Euclidean Group. *Robotics and Autonomous Systems*, **83**, 338-348. https://doi.org/10.1016/j.robot.2016.05.001

[9] Zhong, Z., Liu, S., Mathew, M. and Dubey, A. (2018) Camera Radar Fusion for Increased Reliability in ADAS Applications. *Electronic Imaging*, **30**, art00011. https://doi.org/10.2352/ISSN.2470-1173.2018.17.AVM-258

[10] Kim, D.Y. and Jeon, M. (2014) Data Fusion of Radar and Image Measurements for Multi-Object Tracking via Kalman Filtering. *Information Sciences*, **278**, 641-652. https://doi.org/10.1016/j.ins.2014.03.080

[11] Han, Z.B., Zhang, C.Q., Fu, H.Z. and Zhou, T.Y. (2021) Trusted Multi-View Classification. *ArXiv*, 2102.02051. https://doi.org/10.48550/arXiv.2102.02051

[12] Obrvan, M., Ćesić, J. and Petrović, I. (2016) Appearance Based Vehicle Detection by Radar-Stereo Vision Integration. In: Reis, L.P., *et al*., Eds., *Robot* 2015: *Second Iberian Robotics Conference*, Springer, Cham, 437-449. https://doi.org/10.1007/978-3-319-27146-0_34

[13] Wu, S., Decker, S., Chang, P., Camus, T. and Eledath, J. (2009) Collision Sensing by Stereo Vision and Radar Sensor Fusion. *IEEE Transactions on Intelligent Transportation Systems*, **10**, 606-614. https://doi.org/10.1109/TITS.2009.2032769

[14] Chadwick, S., Maddern, W. and Newman, P. (2019) Distant Vehicle Detection Using Radar and Vision. 2019 *International Conference on Robotics and Automation* (*ICRA*) *IEEE*, Montreal, 20-24 May 2019, 8311-8317. https://doi.org/10.1109/ICRA.2019.8794312

[15] Chang, S., Zhang, Y., Zhang, F., *et al*. (2020) Spatial Attention Fusion for Obstacle Detection Using Mmwave Radar and Vision Sensor. *Sensors*, **20**, Article 956. https://doi.org/10.3390/s20040956

[16] Yadav, R., Vierling, A. and Berns, K. (2020) Radar + RGB Attentive Fusion for Robust Object Detection in Autonomous Vehicles. *IEEE International Conference on Image Processing* (*ICIP*), Abu Dhabi, 25-28 October 2020, 1986-1990. https://doi.org/10.1109/ICIP40778.2020.9191046

[17] 黄昌霸. 车载毫米波雷达目标检测技术研究[D]: [硕士学位论文]. 成都: 电子科技大学, 2020. https://doi.org/10.27005/d.cnki.gdzku.2020.002634

[18] Wang, Z., Wu, Y. and Niu, Q. (2019) Multi-Sensor Fusion in Automated Driving: A Survey. *IEEE Access*, **8**, 2847-2868. https://doi.org/10.1109/ACCESS.2019.2962554

[19] Guo, X.-P., Du, J.-S., Gao, J. and Wang, W. (2018) Pedestrian Detection Based on Fusion of Millimeter Wave Radar and Vision. *Proceedings of the* 2018 *International Conference on Artificial Intelligence and Pattern Recognition*, Beijing, 18-20 August 2018, 38-42. https://doi.org/10.1145/3268866.3268868

[20] 罗逍, 姚远, 张金换. 一种毫米波雷达和摄像头联合标定方法[J]. 清华大学学报(自然科学版), 2014, 54(3): 289-293. https://doi.org/10.16511/j.cnki.qhdxxb.2014.03.005

[21] Nobis, F., Geisslinger, M., Weber, M., Betz, J. and Lienkamp, M. (2019) A Deep Learning-Based Radar and Camera Sensor Fusion Architecture for Object Detection. 2019 *Sensor Data Fusion*: *Trends*, *Solutions*, *Applications* (*SDF*), Bonn, 15-17 October 2019, 1-7. https://doi.org/10.1109/SDF.2019.8916629

[22] Zhou, X., Wang, D. and Krähenbühl, P. (2019) Objects as Points. *ArXiv*, 1904.07850. https://doi.org/10.48550/arXiv.1904.07850

[23] Nabati, R. and Qi, H. (2021) Centerfusion: Center-Based Radar and Camera Fusion for 3d Object Detection. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Waikoloa, 3-8 January 2021, 1527-1536. https://doi.org/10.1109/WACV48630.2021.00157

[24] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016) Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[25] Lin, M., Chen, Q. and Yan, S. (2013) Network in Network. *ArXiv*, 1312.4400. https://doi.org/10.48550/arXiv.1312.4400

[26] Caesar, H., Bankiti, V., Lang, A.H., *et al*. (2020) nuScenes: A Multimodal Dataset for Autonomous Driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 11621-11631. https://doi.org/10.1109/CVPR42600.2020.01164