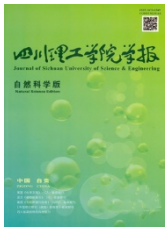


四川轻化工大学学报(自然科学版)
Journal of Sichuan University of Science & Engineering(Natural Science Edition)
ISSN 2096-7543,CN 51-1792/N

《四川轻化工大学学报(自然科学版)》网络首发论文

题目：基于多模态融合的 3D 目标检测技术研究
作者：曾恒，姚娅川
收稿日期：2024-06-11
网络首发日期：2025-04-21
引用格式：曾恒，姚娅川．基于多模态融合的 3D 目标检测技术研究[J/OL]．四川轻化工大学学报(自然科学版). https://link.cnki.net/urlid/51.1792.N.20250421.1308.002

基于多模态融合的 3D 目标检测技术研究

曾　恒[1a,2]，姚娅川[1b,2]

(1. 四川轻化工大学 a.自动化与信息工程学院，b.物理与电子工程学院，四川 宜宾 644000；
2. 人工智能四川省重点实验室，四川 宜宾 644000）

摘　要：针对自动驾驶领域远距离目标的漏检问题，提出一种融合相机与毫米波雷达数据的改进 CenterFusion 的 3D 目标检测模型。首先，引入早期融合策略将雷达数据映射到图像平面上，并将其与图像数据结合形成多通道输入，以增强网络模型的抗干扰能力。其次，在特征融合网络后引入注意力机制，使模型聚焦于融合特征图关键信息提取，有效提高了 3D 目标检测的准确度。然后，进一步改进损失函数解决正负样本不均衡问题。最终，模型在 NuScenes 数据集上进行对比实验和消融实验，结果表明，改进模型相较于传统的 CenterFusion 模型平均检测精度均值提高了 1.5%，NuScenes 检测分数提高了 2.1%，有效提高了远距离目标的检测能力。

关键词：自动驾驶；传感器融合；3D 目标检测；早期融合；注意力机制

中图分类号：TP391　　**文献标志码**：A

引　言

随着自动驾驶技术的快速发展，对于高精度、高可靠性的 3D 目标检测算法的需求日益增长，并逐步发展成为一种趋势[1]。3D 目标检测作为自动驾驶系统的核心技术之一[2]，它可以精确识别出车辆周围的目标，通过估计它们的位置、尺寸和运动状态，为自动驾驶车辆的路径规划和决策控制提供准确的空间语义信息[3]。

3D 目标检测技术利用多种传感器数据，包括相机、激光雷达和毫米波雷达，实现对环境中的静态和动态物体进行精确识别。目前，3D 目标检测方法主要分为 3 大类：基于相机图像的检测方法、基于雷达点云的检测方法和基于多传感器融合的检测方法[4]。

基于相机图像的方法中，Shi 等[5]提出了一种双目立体视觉模型 StereoCenterNet，该模型通过融合双目图像特征，同时预测 2D 和 3D 检测框，并采用不同的策略来优化投影误差，使模型能够有效适应不同遮挡程度的检测物体，从而增强了模型在复杂环境下的鲁棒性，能够提供更为精确的 3D 目标检测结果。但是在重复纹理或无纹理区域对准确估计视差具有挑战性，导致深度估计不准确。Zhang 等[6]提出了 Mono Flex 模型，通过在 3D 边界框关键点估计过程中引入深度信息，并以多检测头策略对物体边界框关键点和深度信息进行估计，以此提升模型的 3D 目标检测性能。但是，该模型没能针对被遮挡的物体进行有效处理，且容易受到光照条件的影响。

基于雷达点云的方法中，Zheng 等[7]提出了一种场景感知雷达学习框架，采用感知序列混合增强技术，进一步提高毫米波雷达在各种环境下的目标检

收稿日期：2024-06-11
基金项目：四川省科技厅重大专题项目(2018GZDZX0045)
通信作者：姚娅川(1968-),女,教授,硕士,研究方向为图像处理、智能控制,(E-mail)610851229@qq.com

测性能；并基于 3D 自动编码器用于雷达目标检测，针对不同场景下的雷达数据特征，动态调整数据增强策略，从而在复杂的实际应用场景中，实现了更高的准确率和鲁棒性。但是，该模型在雷达信号受到其他物体的遮挡或截断时，会影响目标检测的完整性和准确性，且对于动态场景下的检测准确性有待提高。Yang 等[8]提出了一种采用 Transformer 架构的雷达点云三维检测模型（PVT-SSD 模型），该模型通过体素化将不规则的点云数据转换为规则的体素网格，进而从每个体素网格中提取局部特征并利用其自注意力机制捕捉体素间的全局信息。但是在将连续的点云数据转换为离散的体素网格时会引入量化误差，从而影响目标物体的检测精度。

基于传感器融合的方法中，Yin 等[9]提出一种新颖的多视角投影（MVP）模型，该模型利用摄像机的内参矩阵，将 RGB 图像中的二维检测结果转换为伪点云，以此增强稀疏的雷达点云数据，进而提升雷达的 3D 目标检测模型的性能。但是，该模型在两阶段细化模块仅使用俯视图的特征，没有充分利用由算法生成的高分辨率虚拟点。Kim 等[10]提出了一种新的相机雷达网融合网络模型（CRN），该模型使用精确的雷达点将相机图像特征转换为 BEV 来解决图像中空间信息的缺乏，并使用多模态可变形注意力机制在 BEV 中聚合图像和雷达特征图，以解决多传感器数据输入之间的空间对齐问题。但是，该模型在将图像特征点由 2D 转化为 3D 时，对目标深度估计存在一定的误差，且缺乏雷达的高度信息。车俐等[11]提出一种基于交叉注意力机制的方法融合毫米波雷达和相机数据。首先，将毫米波雷达和相机进行空间对齐，并将对齐后的点云投影成点云图像，最后将点云图像和相机图像输入到包含 AF 结构的 CenterNet 网络中进行训练并生成空间注意力权重，以增强相机中关键特征，从而提高原网络检测各种大小目标的性能。但是，该模型的点云图像质量容易受到点云稀疏性的影响，且在点云转换过程中可能会损失一些原始的三维空间信息，从而导致特征提取的质量和准确性降低。Nabati 等[12]提出了一种融合毫米波雷达点云信息和图像数据的 3D 目标检测方法，以改进的 CenterNet 作为图像检测的骨干网络对图像进行特征提取，再对雷达点云进行柱状扩展，并提出一种视锥关联机制将毫米波雷达检测点与图像检测准确 2D 边框信息进行关联，最后将图像特征图与毫米波雷达特征图进行融合得到最终的 3D 边框信息。但是，该模型在 3D 检测过程中过于依赖相机进行初步检测，在雨雾等天气条件下，会降低相机检测的准确性和可靠性，缺乏对外部干扰的鲁棒性。

综上所述，基于图像的 3D 目标检测容易受到光照变化和恶劣天气条件的影响，并且无法提供精确的深度信息，从而导致目标的检测性能下降。基于雷达点云的 3D 目标检测缺乏目标的颜色和纹理信息，以及点云分辨率过低和易受到其他雷达源的干扰，从而导致误检或漏检。基于传感器融合的 3D 目标检测受限于不同类型数据的空间和时间同步对齐，以及在融合过程传感器故障，导致检测的鲁棒性和精度降低。因此，对于上述研究存在的问题，改进的模型在原模型的基础上引入早期融合策略[13]，以增强目标检测网络模型的鲁棒性来解决目标漏检问题，并在图像与毫米波雷达特征融合后引入改进的金字塔挤压注意力（Pyramid Squeeze Attention，PSA）[14]模块，提高 3D 目标检测的精度。

1 CenterFusion 模型

CenterFusion 通过融合相机图像与雷达点云数据进行 3D 目标检测，其网络架构如图 1 所示。

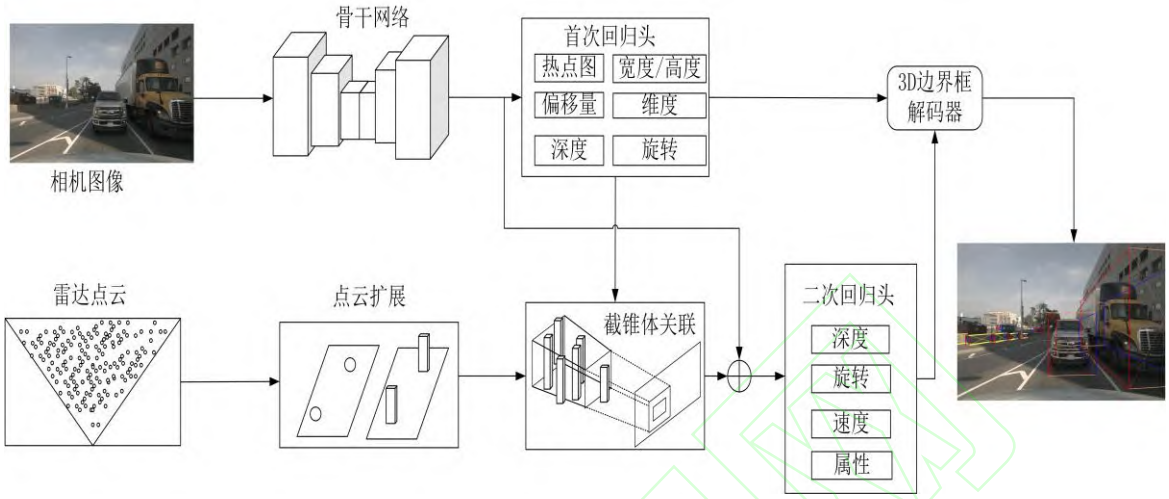


图 1 CenterFusion 网络模型

在该网络模型中，主要分为基于图像处理的目标检测分支、基于雷达处理的截锥体关联网络分支以及二次回归特征融合网络 3 个部分。在基于图像处理的目标检测分支中，改进的深度聚合（Deep Layer Aggregation，DLA）网络作为 CenterNet[15]网络结构的骨干网络对输入图像进行特征提取，再由 1 个 256 通道的3×3卷积和 1 个1×1卷积构成的首次回归头预测每个目标的2D边界框以及初步3D边界框。

在基于雷达处理的截锥体关联网络分支中，由于毫米波雷达不能提供高度信息，将毫米波雷达的目标检测点进行柱状扩展，并采用一种截锥体关联方法，使用目标对象准确的 2D 边界框以及其估计的深度和大小来为目标对象创建3D感兴趣区域（RoI）截锥体，以此过滤掉不在该区域的雷达点。如果该区域内存在多个雷达点云柱体，以最近的雷达点云柱体与目标相进行关联。

在二次回归特征融合网络中，将由截锥体关联网络的毫米波雷达点云特征信息与目标检测网络的图像特征信息按通道进行拼接，构成二次回归头新的输入特征，用于预测目标的 3D 边界框。二次回归头由 3 个3×3的卷积和 1 个1×1的卷积构成。最后，将首次回归头和二次回归头的检测结果输入到3D边界框解码器中生成更准确的 3D 边界框。

2 模型改进

在改进的 CenterFusion 模型中，采用了一种早期融合策略，实现雷达特征与图像数据融合，并构成一个多模态数据输入。进一步地，在特征融合网络后引入了注意力模块，该模块能够使网络模型专注于关键特征的提取。此外，为了解决类别不平衡问题，对损失函数进行了调整，以优化正负样本的分布。改进模型网络结构如图 2 所示。

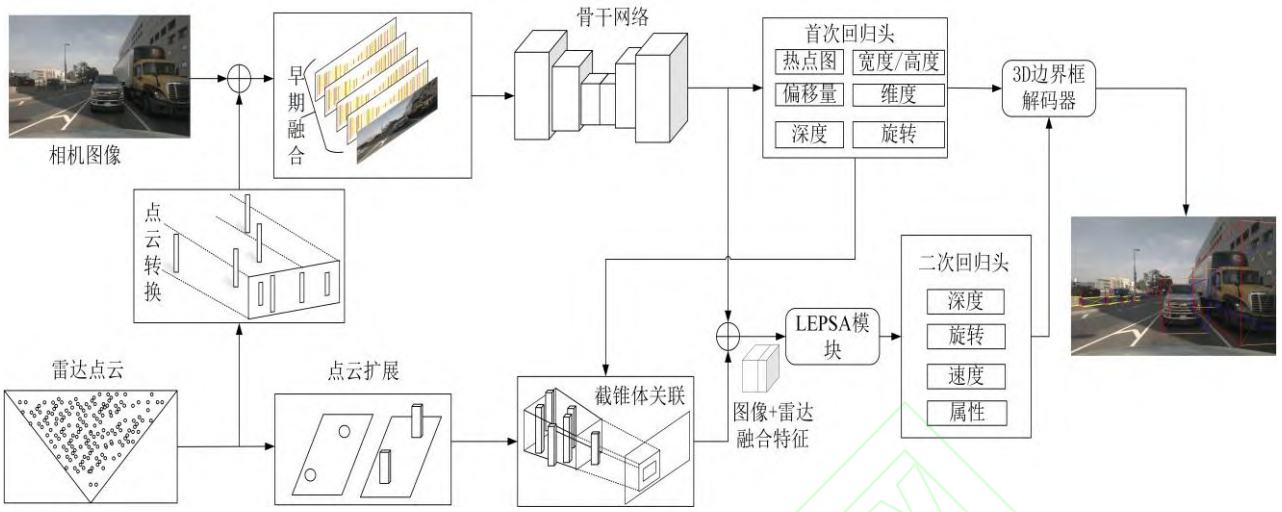


图 2 改进 CenterFusion 模型网络结构

2.1 早期融合

在 CenterFusion 的图像目标检测网络分支中，只依赖相机进行初始目标检测，对于远距离以及在复杂环境中的目标存在漏检的问题。因此，本文在 CenterFusion 模型的基础上引入早期融合策略。在早期融合阶段，将毫米波雷达点云包含的特征信息通过坐标转换投影到图像平面中，作为附加通道并入到目标检测网络的输入中，同时改变目标检测网络的输入通道数，修改网络的卷积层，用一个新的二维卷积层替换原始层，这个新的卷积层具有更多的输入通道，以容纳额外的雷达数据。再将原始相机图像的权重复制到新卷积层的对应通道上，而雷达数据的权重则依赖于后续训练过程的学习。使目标检测网络能够处理多通道输入数据，以此来提升网络模型的性能。

在早期融合中，毫米波雷达点云的数据非常稀疏。为了弥补这个问题，首先，将毫米波雷达扫描帧进行聚合，然后将毫米波雷达点通过坐标转换投影到二维图像平面，并将毫米波雷达点在尺寸、高度以及宽度上进一步放大，从而让雷达图像特征变得更密集。坐标转化关系如图 3 所示。

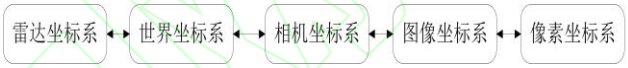


图 3 坐标系转换关系图

首先，通过坐标系之间的转换关系可以得到像素坐标系与世界坐标系之间的转换关系如式(1)所示：

$$Z_c\begin{bmatrix} x_v \\ y_v \\ 1 \end{bmatrix}=\begin{bmatrix} f/d_x & 0 & u_o & 0 \\ 0 & f/d_y & v_o & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} R_{\mathrm{W}} & T_{\mathrm{W}} \\ O^{\mathrm{T}} & 1 \end{bmatrix}\begin{bmatrix} X_{\mathrm{W}} \\ Y_{\mathrm{W}} \\ Z_{\mathrm{W}} \\ 1 \end{bmatrix} \tag{1}$$

其中，(x_v,y_v) 为目标在像素坐标系下的坐标；$(X_{\mathrm{W}},Y_{\mathrm{W}},Z_{\mathrm{W}})$ 为目标在世界坐标系下的坐标；f 为相机的焦距；d_x 和 d_y 分别表示像素坐标系 x 轴和 y 轴方向的物理尺寸。(u_o,v_o) 表示成像平面中心点在像素坐标系下的坐标；R_{W} 和 T_{W} 分别为世界坐标系到相机坐标系的旋转矩阵与平移矩阵；$\begin{bmatrix} f/d_x & 0 & u_o & 0 \\ 0 & f/d_y & v_o & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$ 为相机的内参矩阵；$\begin{bmatrix} R_{\mathrm{W}} & T_{\mathrm{W}} \\ O^{\mathrm{T}} & 1 \end{bmatrix}$ 为相机的外参矩阵。

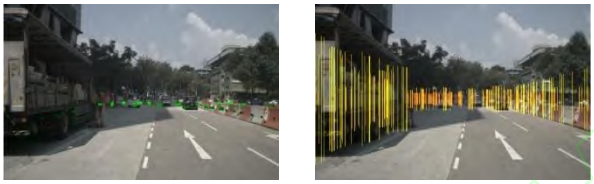
其次将雷达坐标系转换到世界坐标系中。假设目标被雷达探测到的目标点 Q 的位置信息为 (r,θ)，那么该目标在世界坐标系中对应的坐标如下所示：

$$(x_r,y_r,z_r)=(r\sin\theta,0,r\cos\theta) \tag{2}$$

然后，通过测量雷达与世界坐标系的偏差，可以得到偏移向量$e=[e_x,e_y,e_z]$，则目标在世界坐标系中的坐标$(x_{\mathrm{w}},y_{\mathrm{w}},z_{\mathrm{w}})$为：

$$\begin{cases} X_{\mathrm{w}}=x_r+\boldsymbol{e}_x \\ Y_{\mathrm{w}}=y_r+\boldsymbol{e}_y \\ Z_{\mathrm{w}}=z_r+\boldsymbol{e}_z \end{cases} \tag{3}$$

最后，经过上述变换使毫米波雷达点云投影到图像中，垂直投影线的起点从 3D 空间的地面开始，高度设定为$3\ \mathrm{m}$，毫米波雷达点垂直投影线的高度随着与摄像机原点距离的增加而减小，如图 4 所示。



(a) 投影至图像的雷达点　　(b) 雷达点扩展为投影线

图 4 雷达点投影到图像平面

利用雷达点云中提供的深度、径向速度和反射截面（Radar Cross Section，RCS）信息，将这些雷达回波特征信息以像素值的方式存储在特征图像中，并作为附加的图像通道连接到 RGB 输入图像，构成目标检测网络的输入。早期融合示意图如图 5 所示，显示的附加图像通道对应于毫米波雷达深度d、x轴与z轴径向速度以及 RCS 的值，白色在输入通道中显示为 0。

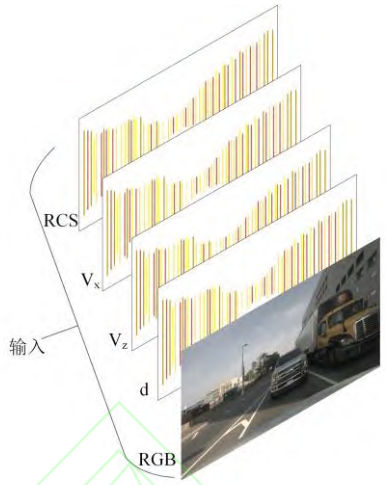


图 5 早期融合示意图

2.2 LEPSA 模块

为了更加准确地提取融合后特征图中的多尺度信息，引入 PSA 模块，并在此基础上设计了一种轻量高效金字塔挤压注意力机制（Lightweight and Efficient Pyramid Squeeze Attention，LEPSA）。首先，使用多尺度金字塔卷积结构来整合输入特征图像的多维度信息；其次，通过 SPC（Spatial-Channel Compression）模块，实现对输入张量通道的高效压缩与合并，这使得能够在每个通道上捕捉到尺度各异的空间特征，进而构建出具有多尺度特性的通道特征图；进一步地，部署 SEWeight（Selective Enhancement Weighting）机制，该机制负责从多尺度特征中提取关键的注意力权重，形成针对通道的注意力向量；随后，Softmax 函数被应用于这些注意力向量，以获得校准后多尺度注意力权重；最后，将这些权重与相应的特征图进行元素级乘法操作，从而得到多尺度特征信息的特征图。LEPSA 模型结构如图 6 所示。

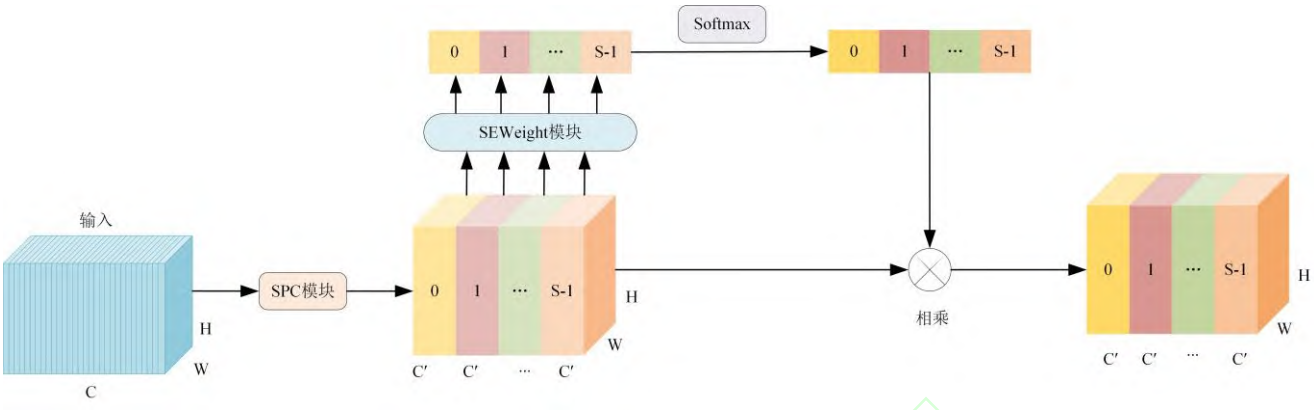


图 6 LEPSA 模型结构图

在 LEPSA 模型结构中，SPC 模块采用多分支的结构提取输入特征图的空间信息，其工作机制如图 7 所示。

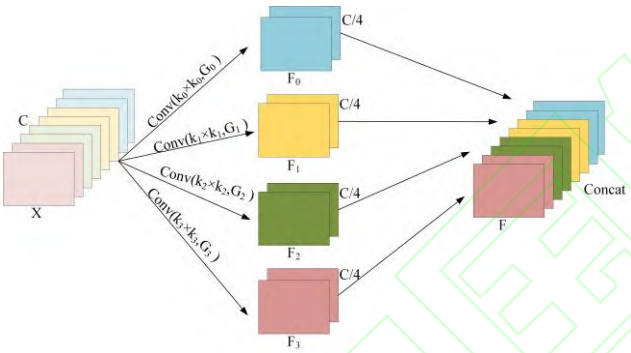


图 7 SPC 模块工作机制图

LEPSA 模型结构中，C 表示为输入的通道维度数，S 表示被等分的份数，具有不同尺度的特征图每个特征映射具有公共通道维数 $C'=C/S$，对于输入向量，采用分组卷积应用于卷积核。其中，多尺度内核大小和组大小的关系如式(4)所示：

$$G=2^{\frac{K-1}{2}} \tag{4}$$

其中，K 为内核大小；G 为组大小。多尺度特征图生成函数和预处理特征图如式(5)所示：

$$\begin{cases} F_i=Conv(k_i\times k_i,G_i)(X) & i=0,1,2,\cdots,S-1 \\ F=Concat([F_0,F_1,\cdots,F_{S-1}]) \end{cases} \tag{5}$$

其中，X 表示输入的特征图，H 表示所输入特征图的高度，W 表示所输入特征图的宽度，第 i 个核大小 $k_i=2\times(i+1)+1$，第 i 个组大小 $G_i=2^{(k_i-1)/2}$，$F_i\in R^{C'\times H\times W}$ 表示不同比例的特征图，$F\in R^{C\times H\times W}$ 为整个多尺度预处理特征图，*Concat* 为将不同比例的特征图进行拼接操作。

SEWeight 模块从多尺度预处理特征图中提取通道注意力权重信息，得到不同尺度的注意力权重向量，如式(6)所示：

$$\begin{cases} Q_i=SEWeight(F_i), & i=0,1,2,\cdots,S-1 \\ Q=Q_0\oplus Q_1\oplus\cdots\oplus Q_{S-1} \end{cases} \tag{6}$$

其中，$Q_i\in R^{C'\times1\times1}$ 为注意力权重，$\oplus$ 为 *Concat* 算子，Q 为多尺度注意力权重向量。

然后使用 Softmax 对获得的多尺度注意力权重重新校准，以实现局部和全局通道注意力之间的交互，并以级联的方式获得整个通道注意力权重，如式(7)所示：

$$\begin{cases} att_i=\mathrm{Softmax}(Q_i)=\exp(Q_i)\Big/\sum_{i=0}^{S-1}\exp(Q_i) \\ att=att_0\oplus att_1\oplus\cdots\oplus att_{S-1} \end{cases} \tag{7}$$

其中，att_i 为重新校准后的权重；att 为重新校准后的多尺度通道注意力权重。

最后，将重新校准的注意力权重与相应的特征图进行点乘操作，如式(8)所示：

$$\begin{cases} Y_i=F_i*att_i & i=1,2,3,\cdots,S-1 \\ Out=Concat([Y_0,Y_1,\cdots,Y_{S-1}]) \end{cases} \tag{8}$$

其中，$*$ 表示通道乘法；Y_i 表示多尺度通道注意力权重的特征图；*Out* 为最后输出。

2.3 损失函数

CenterNet 的损失函数由热力图损失、中心点偏移损失和宽高大小损失构成，其损失函数如式(9)所

示：

$$L_{\text{det}} = L_{\text{heatmap}} + \lambda_{\text{off}} L_{\text{off}} + \lambda_{\text{wh}} L_{\text{wh}} \tag{9}$$

其中权重系数 $\lambda_{\text{wh}} = 0.1$，$\lambda_{\text{off}} = 1$。

采用 Focal Loss 作为热力图损失 L_{heatmap}，如式(10)所示：

$$L_{\text{heatmap}} = -\frac{1}{N}\sum_{\tilde{x}\tilde{y}c}\begin{cases}\left(1-\mathrm{Y}_{\tilde{x},\tilde{y},c}\right)^{\alpha_1}\log\left(\mathrm{Y}_{\tilde{x},\tilde{y},c}\right) & \text{if } \mathrm{Y}_{\tilde{x},\tilde{y},c}=1 \\ \left(1-\mathrm{Y}_{\tilde{x},\tilde{y},c}\right)^{\beta}\mathrm{Y}_{\tilde{x},\tilde{y},c}^{\alpha_2}\log\left(1-\mathrm{Y}_{\tilde{x},\tilde{y},c}\right) & \text{if } \mathrm{Y}_{\tilde{x},\tilde{y},c}\neq 1\end{cases} \tag{10}$$

其中，N 为输入图像中关键点的数量；α_1、α_2 和 β 均为超参数；$\mathrm{Y}_{\tilde{x},\tilde{y},c}$ 为图像中关键点的真实位置；$\mathrm{Y}_{\tilde{x},\tilde{y},c}$ 为图像中关键点位置的预测值。在 CenterNet 原始网络中将 α_1、α_2 设置为 2，β 设置为 4，但在实验过程中发现，图像中存在正样本的数量太少从而导致目标检测框的置信度不高。当 α_1 设置为 2 时，对于正样本的惩罚权重是非常小的。因此，将式中的 α_1 的值设置为 1，增大正样本的惩罚力度，使所训练的模型在预测的结果上具有更高的置信度。

网络的目标中心点偏移损失函数 L_{off} 如式(11)所示：

$$L_{\text{off}} = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{o}_{\tilde{x}_k,\tilde{y}_k} - \left(\frac{p_{k,c_k}}{K} - \tilde{p}_{k,c_k}\right)\right| \tag{11}$$

其中，$\hat{o}_{\tilde{x}_k,\tilde{y}_k}$ 为预测后偏移量；p_{k,c_k} 为图像中心坐标值；$\tilde{p}_{k,c_k}$ 为缩放后的中心点近似坐标值；K 为缩放因子。

网络的目标框宽高损失函数 L_{size} 如式(12)所示：

$$L_{\text{wh}} = \frac{1}{N}\sum_{k=1}^{N}\left|\hat{s}_{\tilde{x}_k,\tilde{y}_k} - s_k\right| \tag{12}$$

其中，$\hat{s}_{\tilde{x}_k,\tilde{y}_k}$ 为目标预测的宽高大小；s_k 为目标真实的宽高大小。

3 实验与分析

3.1 实验环境

模型训练过程均在服务器上进行。服务器配置的 CPU 是主频为 2.3 GHz 的英特尔至强 Platinum 8336C，GPU 为 NVIDIA GeForce RTX 4090，显存大小为 24 GB。实验软件环境平台为 Ubuntu 22.04.3 LTS，CUDA 版本为 12.1.0，开发语言为 Python3.8，深度学习框架为 Pytorch2.1.0。

3.2 数据集

目前，在自动驾驶领域的 3D 目标检测研究中 KITTI[16]、waymo[17]和 nuScenes[18]等数据集因其全面性和多样性而广受青睐。本研究选择使用 NuScenes 数据集，该数据集结合毫米波雷达、激光雷达和相机等传感器，收集约 15 h 的真实街道驾驶数据。NuScenes 数据集包含 1000 个独立场景，每个场景持续 20 s，并为 23 个不同的类别和 8 个属性提供了详尽的 3D 边界框注释。本研究采用 NuScenes 数据集中的相机和毫米波雷达数据，以验证所提方法的有效性。

3.3 评估指标

为了评估模型的性能，本文使用的评估指标包括平均精度（AP）和全类平均精度（mAP）以及 Nuscenes 检测分数（NDS）。

mAP 是基于精确度（P）和召回率（R）来计算的，其计算为：

$$P = \frac{TP}{TP + FP} \tag{13}$$

$$R = \frac{TP}{TP + FN} \tag{14}$$

其中，*TP* 为真阳性数；*FP* 为假阳性数；*FN* 为假阴性数。*AP* 的计算公式为：

$$AP = \int_0^1 P(R)\,\mathrm{d}R \tag{15}$$

mAP 是由所有类别的 *AP* 计算得来，表示为：

$$mAP = \frac{1}{C}\sum_{c\in C} AP_c \tag{16}$$

其中，C 为 10 种数据类型。

Nuscenes 检测分数 *NDS* 是将 *mAP* 指标与 5 个 TP 指标结合起来计算的，表示为：

$$NDS = \frac{1}{10}\left[5mAP + \sum_{mTP \in TP}(1 - \min(1, mTP))\right] \quad (17)$$

其中，5 个 TP metrics 指标分别为平均平移误差（*mATE*）、平均尺度误差（*mASE*）、平均方向误差（*mAOE*）、平均速度误差（*mAVE*）和平均属性误差（*mAAE*）。*mTP*（mean TP metric）是对于每个类别的 TP 计算所有类别的平均度量，如式(18)所示：

$$mTP = \frac{1}{C}\sum_{c \in C} TP_c \quad (18)$$

3.4 实验结果

对改进模型针在远距离以及复杂环境下检测目标是否存在漏检的情况进行测试，在相同的环境下，通过对比 CenterNet、CenterFusion 模型中的各项指标数据见表 1。

表 1 算法性能对比

模型	*NDS*↑	*mAP*↑	*mATE* ↓	*mASE* ↓	*mAOE* ↓	*mAVE* ↓	*mAAE* ↓
CenterNet	0.331	0.304	0.713	0.258	0.574	1.387	0.661
CenterFusion	0.452	0.331	0.649	0.263	0.534	0.543	0.143
Ours	0.473	0.346	0.593	0.249	0.523	0.498	0.128

注：“↑”表示指标越大越好；“↓”表示指标越小越好。

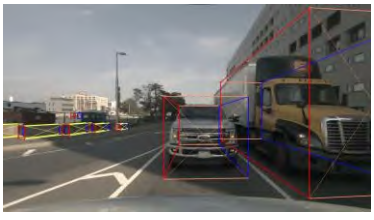
由表 1 可知，本文所提出改进模型的 *NDS* 相较于 CenterFusion 模型提升了 2.1%，其 *mAP* 提升了 1.5%。并且，本文模型的各项误差指标相较于原模型得到下降，尤其是对于 *mAVE* 和 *mAAE*，误差下降的效果尤为显著。此外，改进模型与 CenterNet 和 CenterFusion 模型 10 个类别的目标检测精度见表 2。

表 2 算法对不同类别目标的检测精度对比

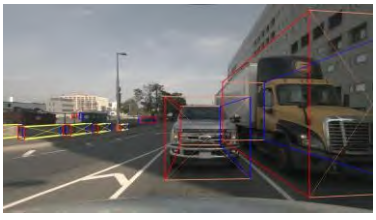
模型	不同类别目标的检测精度									
	Car	Truck	Bus	Trailer	Const.	Pedest.	Motor.	Bicycle	Traff.	Barrier
CenterNet	0.461	0.237	0.327	0.135	0.035	0.364	0.249	0.233	0.551	0.452
CenterFusion	0.525	0.265	0.368	0.148	0.054	0.388	0.303	0.227	0.563	0.471
Ours	0.534	0.269	0.371	0.156	0.065	0.421	0.345	0.242	0.576	0.479

由表 2 可知，本文模型在不同检测目标类别中的精度指标都得到了提升，相较于 CenterFusion 模型，Car、Truck、Bus、Trailer、Const.、Pedest.、Motor.、Bicycle、Traff.、Barrier 分别提升了 0.9%、0.4%、0.3%、0.8%、1.1%、3.3%、4.2%、1.5%、1.3%、0.8%。其检测效果如图 8 所示。

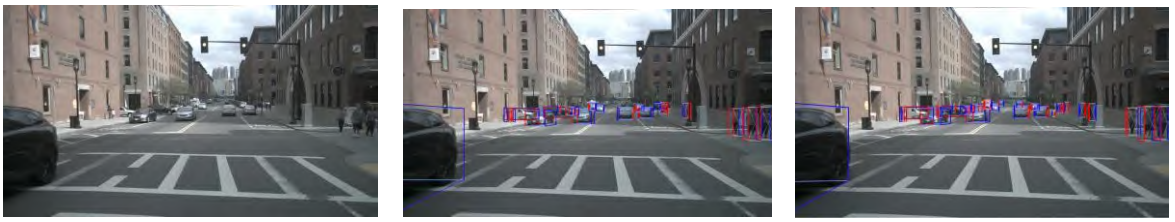
(a1) 原图一



(a2) CenterFusion 模型检测一



(a3) 改进模型检测一



(b1) 原图二　　(b2) CenterFusion 模型检测二　　(b3) 改进模型检测二

图 8 检测结果图

图 8 中为原图、CenterFusion 模型和本文改进模型的检测结果。从图 8(a2)可以看出，CenterFusion 模型在该组可视化结果图中未能检测到白色小车左后方远距离的黑色小车；在下面一组可视化结果图（图 8(b2)）中未能检测出复杂道路场景中位于远距离的两辆黑色小汽车，而在改进的模型中都能精确检测到远距离的目标，以此验证本文改进模型对远距离目标有更好的检测精度。

综上所述，本文模型的各项指标相较于其他模型得到了提升。首先，模型采用了早期融合策略，该策略将毫米波雷达数据与相机数据相结合，以实现更为精确的检测；其次，引入了 LEPSA 注意力机制，用于处理融合后的特征图，能够从多个尺度提取空间信息，从而提高特征的表达能力，并进一步增强目标检测性能。这种多尺度的空间信息提取，为模型提供了更为丰富的上下文信息，进一步加强融合后的特征提取和利用，从而增强目标检测的准确性。

3.5 消融实验

为了验证本文所提出的改进方案对提升网络性能的有效性，将提出的早期融合策略、注意力机制模块与改进的损失函数进行消融实验。在相同实验环境下，实验结果见表 3。

表 3 不同改进方案下消融实验

改进方案	早期融合	LEPSA	损失函数	*NDS* ↑	*mAP* ↑	*mATE*↓	*mASE*↓	*mAOE*↓	*mAVE*↓	*mAAE*↓
1	×	×	×	0.452	0.331	0.649	0.263	0.534	0.543	0.143
2	√	×	×	0.459	0.338	0.633	0.261	0.528	0.531	0.135
3	√	×	√	0.463	0.339	0.617	0.258	0.527	0.529	0.133
4	√	√	×	0.470	0.345	0.609	0.251	0.524	0.505	0.131
5	√	√	√	0.473	0.346	0.593	0.249	0.523	0.498	0.128

注：“↑”表示指标越大越好；“↓”表示指标越小越好。

由表 3 可知，方案 1 是基模型的实验结果；方案 2 通过加入早期融合策略后，模型的 *NDS* 和 *mAP* 都得到了提升。这是因为在早期融合中，将毫米波雷达特征信息与 RGB 图像融合进行检测，解决了单一相机特征不足导致的漏检问题。

方案 4 相较于方案 2，在引入早期融合策略的同时加入注意力机制模块，模型的 *NDS* 和 *mAP* 得到了进一步提升，证明了加入注意力机制模块的有效性。

方案 5 相较于方案 4 改进了损失函数，模型的参数指标得到了提升。综上所述，本文模型通过引入早期融合策略、注意力机制模块和改进的损失函数，使模型的性能得到了极大的提升。

4 结束语

针对自动驾驶领域中 3D 目标检测在远距离以及

复杂环境下存在漏检的问题，提出了一种多模态融合的 3D 目标检测模型。该模型在 CenterFusion 的基础上采用早期融合策略，以增强网络模型的抗干扰能力；引入轻量高效的金字塔挤压注意力模块，进一步提升模型的精度；此外，改进的损失函数有效缓解正负样本不均衡的问题。实验结果表明，本文提出的方法在 nuScenes 数据集上相较于原模型的 *mAP* 提升了 1.5%， *NDS* 提高了 2.1%，增强了远距离目标的检测能力。未来的研究方向将聚焦于优化传感器融合策略，通过整合更多类型的传感器，进一步增强模型在极端天气条件下的鲁棒性。同时，鉴于现有的高精度3D目标检测模型存在计算资源消耗较大，难以满足自动驾驶对实时性的要求，未来的研究还需致力于开发轻量级的网络架构以及高效的推理算法，在模型精度与计算效率之间寻求更优的平衡。

参考文献：

[1] ARNOLD E,AL-JARRAH O Y,DIANATI M,et al.A survey on 3D object detection methods for autonomous driving applications[J].IEEE Transactions on Intelligent Transportation Systems,2019,20(10):3782-3795.

[2] 窦允冲,侯进,曾雷鸣,等.基于反馈机制与空洞卷积的道路小目标检测网络[J].计算机工程,2023,49(1):287-294.

[3] LIANG W,XU P F,GUO L,et al.A survey of 3D object detection[J].Multimedia Tools and Applications,2021,80(19):29617-29641.

[4] 梁振明,黄影平,宋卓恒,等.自动驾驶中基于深度学习的 3D 目标检测方法综述[J].上海理工大学学报,2024,46(2):103-119.

[5] SHI Y,GUO Y,MI Z,et al.Stereo CenterNet-based 3D object detection for autonomous driving[J]. Neurocomputing,2022,471:219-229.

[6] ZHANG Y P,LU J W,ZHOU J.Objects are different:flexible monocular 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition,Nashville,USA,June 20-25,2021: 3288-3297.

[7] ZHENG Z,YUE X,KEUTZER K,et al.Scene-aware learning network for radar object detection[C]//Proceedings of the 2021 International Conference on Multimedia Retrieval,Taipei,China,August 21-24,2021:573-579.

[8] YANG H,WANG W,CHEN M,et al.PVT-SSD:single-stage 3D object detector with point-voxel transformer[C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition,Vancouver, Canada,June 18-22,2023:13476-13487.

[9] YIN T W,ZHOU X Y,KRÄHENBÜHL P.Multimodal virtual point 3D detection[J].Advances in Neural Information Processing Systems,2021,34:16494-16507.

[10] KIM Y,SHIN J,KIM S,et al.CRN:camera radar net for accurate,robust,efficient 3D perception[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV),Paris,France,October 1-6,2023:17569-17580.

[11] 车俐,吕连辉,蒋留兵.AF-CenterNet:基于交叉注意力机制的毫米波雷达和相机融合的目标检测[J].计算机应用研究,2024,41(4):1258-1263.

[12] NABATI R,QI H R.CenterFusion:center-based radar and camera fusion for 3D object detection[C]//2021 IEEE/CVF Winter Conference on Applications of Computer Vision,Waikoloa,USA,January 3-8,2021: 1526-1535.

[13] NOBIS F,GEISSLINGER M,WEBER M,et al.A deep learning-based radar and camera sensor fusion architecture for object detection[C]//2019 Sensor Data Fusion:Trends,Solutions,Applications,Bonn,Germany,October 15-17, 2019:1-7.

[14] ZHANG H,ZU K,LU J,et al.EPSANet:an efficient pyramid squeeze attention block on convolutional neural network[C]//Proceedings of the Asian Conference on Computer Vision,Macao,China,December 4-8,2022:1161-1177.

[15] 邢雪,王彬,王馨田.基于 CenterNet 编码优化的车辆目标检测模型[J].长江信息通信,2024,37(5):73-77.

[16] GEIGER A,LENZ P,URTASUN R.Are we ready for autonomous driving?The KITTI vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition,Providence, USA,June 16-21,2012:3354-3361.

[17] SUN P,KRETZSCHMAR H,DOTIWALLA X,et al.Scalability in perception for autonomous driving:waymo open dataset[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle,USA,June 13-19,2020:2443–2451.

[18] CASEAR H,BANKITI V,LANG A H,et al.NuScenes:a multimodal dataset for autonomous driving[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),Seattle,USA,June 13-19,2020: 11618-11628.

引用格式：

中　文：曾恒,姚娅川.基于多模态融合的 3D 目标检测技术研究[J].四川轻化工大学学报(自然科学版),2025,38(4)

英　文：ZENG H,YAO Y C.Research on 3D object detection technology based on multimodal fusion[J].Journal of Sichuan University of Science & Engineering (Natural Science Edition),2025,38(4)

Research on 3D Object Detection Technology Based on Multimodal Fusion

ZENG Heng[1a,2] *, YAO Yachuan* [1b,2]

(1a. School of Automation and Information Engineering, 1b. School of Physics and Electronic Engineering, Sichuan University of Science & Engineering, Yibin 644000, China;
2. Artificial Intelligence Key Laboratory of Sichuan Province, Yibin 644000, China)

Abstract: Aiming at the problem of missed detection of long-distance targets in the field of autonomous driving, an improved 3D target detection model based on CenterFusion is proposed, which combines camera and millimeter wave radar data. First of all, the early fusion strategy is introduced to map the radar data to the image plane and combine it with the image data to form multi-channel input to enhance the anti-jamming ability of the network model. Secondly, after the feature fusion network, the attention mechanism is introduced to make the model focus on the key information extraction of the fusion feature map, which effectively improves the accuracy of 3D target detection. Then, the loss function is further improved to solve the problem of imbalance between positive and negative samples. Finally, the proposed model is used to carry out comparative experiments and ablation experiments on nuScenes data sets, and the results show that the average detection accuracy of the improved model is 1.5% higher than that of the traditional CenterFusion model, and the nuScenes detection score of the improved model is 2.1% higher, effectively improving the detection ability of long-distance targets.

Key words: autonomous driving; sensor fusion; 3D target detection; early fusion; attention mechanism