·激光雷达感知与自动驾驶多模态3D目标检测技术专题·

基于改进Transformer-无迹卡尔曼滤波器的智能车辆多模态3D目标检测方法*

张哲宁[1,2]　刘祯[1,2]　王化强[1,2]

（1.湖北文理学院纯电动汽车动力系统设计与测试湖北省重点实验室，襄阳 441053；2.湖北文理学院汽车与交通工程学院，襄阳 441053）

【摘要】为提升智能车辆环境感知目标检测的准确性，提出一种特征融合3D目标检测方法。首先，对Transformer模型进行改进，借助多头自注意力机制，充分考虑数据空间的局部特征信息，并保留不同的特征权重，以提取点云与图像特征；然后，借助无迹卡尔曼滤波器（UKF）设计图像与点云多模态融合系统，最终实现3D目标检测；最后，采用KITTI数据集和实车数据集对模型进行训练和推理，并与多种主流算法进行对比。验证结果显示，与应用广泛的截锥卷积网络（F-ConvNet）、视锥点云网络（F-PointNet）等主流多模态融合算法相比，所提出的目标检测模型多类别平均精度分别提升了0.34百分点和3.03百分点，车辆和骑行者对象的检测平均精度分别提升了2.52百分点和9.32百分点，且该模型在实车数据推理中的表现与训练评测结果基本一致。

主题词：无人驾驶　Transformer　多传感器信息融合　多模态感知　无迹卡尔曼滤波器

中图分类号：TK411　　**文献标志码：**A　　**DOI:** 10.19620/j.cnki.1000-3703.20241037

Design and Experimental Validation of an Enhanced Transformer-UKF Multi-Modal 3D Object Detection System for Intelligent Vehicles

Zhang Zhening[1,2], Liu Zhen[1,2], Wang Huaqiang[1,2]

（1. Pure Electric Vehicle Power System Design and Testing Hubei Provincial Key Laboratory, Hubei University of Arts and Sciences, Xiangyang 441053; 2. School of Automotive and Transportation Engineering, Hubei University of Arts and Sciences, Xiangyang 441053）

【Abstract】To enhance the accuracy of environmental perception and object detection for intelligent vehicles, this paper proposes a novel feature fusion-based 3D object detection method. First, the Transformer model is improved by leveraging a multi-head self-attention mechanism to fully capture local spatial feature information while preserving diverse feature weights, enabling the extraction of point cloud and image features. An Unscented Kalman Filter (UKF) is then utilized to design a multimodal fusion system that combines image and point cloud data, ultimately achieving 3D object detection. The proposed model is trained and tested on both public and private datasets and is compared against several state-of-the-art algorithms. Experimental results demonstrate that the proposed feature fusion-based 3D object detection model achieves an improvement in mean Average Precision (mAP) of 0.34 percentage points and 3.03 percentage points compared to widely used algorithms such as F-ConvNet and F-PointNet, respectively. For vehicle and cyclist detection tasks, the mAP increases by 2.52 percentage points and 9.32 percentage points, respectively. Moreover, the inference results on real-world vehicle data align closely with the training evaluation outcomes. These findings indicate that the proposed algorithm improves the accuracy of object detection in perception systems to a certain extent and provides a valuable reference for the design of object detection systems in autonomous driving.

Key words: Autonomous driving, Transformer, Multi-sensor information fusion, Multimodal perception, Unscented Kalman Filter (UKF)

基金项目：湖北省自然科学基金创新发展联合基金项目（2022CFD083）；“新能源汽车与智慧交通”湖北省优势特色学科群项目（XKTD022023）；襄阳市科技计划项目（高新领域）重点项目（2021ABH004233）。

通信作者：刘祯（1984—），博士，副教授，主要研究方向为智能驾驶技术等，liuzhen@hbuas.edu.cn。

【引用格式】张哲宁，刘祯，王化强．基于改进Transformer-无迹卡尔曼滤波器的智能车辆多模态3D目标检测方法[J]．汽车技术，2026(2)：39-47.

ZHANG Z N, LIU Z, WANG H Q. Design and Experimental Validation of an Enhanced Transformer-UKF Multi-Modal 3D Object Detection System for Intelligent Vehicles[J]. Automobile Technology, 2026(2): 39-47.

1 前言

近年来，无人驾驶汽车环境感知技术发展迅速，目标检测作为环境感知的基础性关键环节备受关注。在各类环境感知方案中，视觉图像与激光雷达（Light Detection And Ranging，LiDAR）点云数据的融合感知因可显著提升系统冗余度和容错性而独具优势[1]。

在智能汽车目标检测领域，主流方式为深度融合，如Qi等提出的视锥点云网络（Frustum PointNet，F-PointNet）[2]就是在其PointNet[3]与PointNet++[4]的基础上结合视觉检测器实现分类与回归。Yoo等提出的三维跨域融合（3D Cross View Fusion，3D-CVF）算法[5]通过跨域特征映射（Cross-View Feature Mapping）模块将图像特征转换至鸟瞰图视角，并利用自适应门控融合网络（Adaptive Gated Fusion Network）对图像与点云特征进行加权融合。Wang等提出的截锥卷积网络（Frustum ConvNet，F-ConvNet）[6]采用滑动锥体卷积、局部特征聚合等技术，专注于非遮挡目标检测。Liang等提出深度连续融合（Deep Continuous Fusion）算法[7]实现多传感器数据的无缝融合，提升了检测准确性与鲁棒性，并进一步提出多任务多传感器融合（Multi-task Multi-sensor Fusion，MMF）算法[8]，通过多任务学习融合多传感器数据提高三维目标检测的准确性和鲁棒性。Chen等[9]提出了全稀疏体素网络（VoxelNeXt），直接基于稀疏体素特征进行3D物体检测。Yang[10]等提出了多尺度多深度融合（Multiple Scales with Multi-Depth Fusion，MSMDFusion）方法，利用跨模态交互增强检测与跟踪性能。Li等[11]提出了局部-全局融合网络（Local-to-Global fusion Network，LoGoNet），在不同层次实现激光雷达与相机数据的有效对齐。

Transformer模型拥有自注意力机制（Self-Attention），可并行训练并保留全局信息，适用于目标检测任务。因此，Dosovitskiy等将Transformer模型引入视觉任务，设计出用于处理图像深度学习的视觉Transformer（Vision Transformer，ViT）[12]，增强图像任务的整体关联性；点云Transformer（Point Cloud Transformer）[13]通过引入多层次的自注意力机制与可变形卷积（Deformable Convolution），提升了对点云局部形变的鲁棒性。

综上，多模态感知融合能有效结合图像和激光雷达的优势。为此，本文提出一种新的三维目标检测算法，使用改进的视觉Transformer模型提取图像特征，使用改进的点云Transformer提取点云特征，充分保留其空间局部信息与特征权重，并在模型中间层使用无迹卡尔曼滤波器（Unscented Kalman Filter，UKF）对两类特征向量进行估计与融合，从而提高环境感知的准确性。

2 环境特征提取网络设计

2.1 改进的Transformer架构

在图像数据和点云数据融合前，需要对其特征向量进行提取[14]。由于本文在数据融合的中间层次进行特征数据融合，故需对Transformer进行改进，使其能够直接在原始Transformer的中间层输出编码（Encode）后特征向量矩阵。

原始图像与点云本身的数据性质不同，故在该改进架构中需经过不同的处理模块。输入的图像信息将划分为一系列区块，经位置编码处理，通过线性嵌入映射为一个二维特征向量矩阵；点云信息经过级联的线性层-批归一化-激活函数组合模块（Linear+BatchNorm+ReLU，LBR）进入更高层特征空间，点云信息本身即具备位置特征，因此可以在提取特征后插入位置特征而不需要额外的编码。点云与图像经过各自的输入模块处理后，通过自注意力机制和前馈神经网络的处理计算，获得神经网络生成的特征向量矩阵，如图1所示。

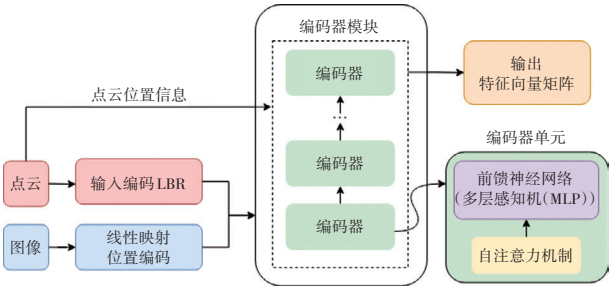


图1 改进的Transformer模型

2.2 视觉图像特征提取

本文通过改进ViT输出层的架构输出有序的特征向量矩阵，该图像特征提取网络的整体框架如图2所示。

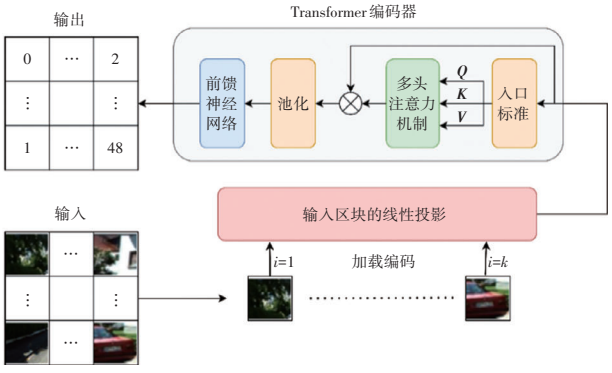


图2　图像特征提取网络的整体框架示意

将输入图像划分为一系列图像块，对其一一对应编码并进行线性投影，设图像的RGB通道数C=3，图像块的分辨率为$P\times P$，将维度映射为D维，同时保持图像区块数一致：

$$D=P^2C \tag{1}$$

随后，编码器（Encoder）利用自注意力机制使模型在全局范围内学习每个图像块间的相关性。这种全局关注的能力使得ViT能够捕捉到图像中不同部分之间的长距离依赖关系，有助于更好地理解图像的整体语境：

$$A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=S\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

式中：A为注意力函数（Attention函数），$\boldsymbol{Q}$为查询（Query）矩阵，$\boldsymbol{K}$为键值（Key）矩阵，$\boldsymbol{V}$为价值（Value）矩阵，S为归一化指数函数（Softmax函数），d_k为$\boldsymbol{Q}$、$\boldsymbol{K}$的维度。

相较于添加了注意力机制的神经网络模型，Transformer在d_k维的缩放点积注意力基础上提出多头注意力机制，将d_k维分为多个头，形成多个子空间，使模型关注不同方面的信息，进而学习更丰富的特征，设$\boldsymbol{W}_i^Q$、$\boldsymbol{W}_i^K$、$\boldsymbol{W}_i^V$为各自的权重维度矩阵，则有：

$$M(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=C(h_1,\cdots,h_h)\boldsymbol{W}^o \tag{3}$$

$$h_i=A(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V) \tag{4}$$

式中：M为多头注意力机制的输出函数，$\boldsymbol{W}^o$为多头注意力中用于整合各子空间结果的权重矩阵，h_i第i个子空间对应的自注意力计算结果。

在本文中，图像数据首先通过位置编码引入关于不同区域的空间信息。然后，通过自注意力机制和前馈神经网络的协同作用，模型深入挖掘图像内部的关联性，提升了特征的表征能力，同时通过池化过程对得到的特征进行整合和提炼。

与传统ViT模型相比，图像特征数据并没有经过解码器（Decoder）的处理，而是直接输出。这一简化改进是为了保留模型对图像特征的更丰富表示，并为下一步的融合过程提供更具信息量和表征力的输入，从而为后续融合步骤提供更为强大和灵活的特征表示。

2.3　激光点云特征提取

点云特征计算对描述点云的形状、结构和其他属性至关重要。本文在点云Transformer模型的基础上进行改进，实现包含位置信息且语义丰富的特征矩阵输出，设计的点云特征提取网络整体框架如图3所示。

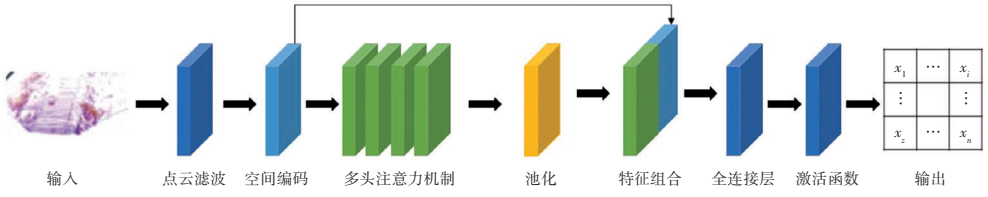


图3　点云特征提取网络的整体框架示意

首先，在点云数据的预处理阶段进行点云分割和地面滤波，以提高点云输入数据的质量。在点云位置编码阶段，采用一种自适应的位置编码策略，以更好地适应不同点云结构的特点。具体而言：

$$P_{\mathrm{E}}(p,2i)=\sin(p/10\,000^{2i/d_{\mathrm{m}}}) \tag{5}$$

$$P_{\mathrm{E}}(p,2i+1)=\cos(p/10\,000^{2i/d_{\mathrm{m}}}) \tag{6}$$

式中：P_{E}为位置编码策略函数；p为输入序列中点云的位置；i为点云向量中的维度索引；d_{m}为点云向量的嵌入维度，用于归一化。

进而，针对输入序列中不同位置的点云和点云向量内部的不同位置可以产生唯一的位置编码P_{E}。

通过调整多头注意力的权重分配方式，进一步优化了网络，使其更专注于点云中具有重要结构的局部区域，从而增强了点与点之间的关联性。随后，分别通过池化增强特征表示并通过特征组合对不同特征进行匹配编码。接下来，通过前馈神经网络对包含全连接层L_{i}和激活函数R_{E}的子网络结构使用2个残差连接：

$$X_h=L_{\mathrm{i}}(R_{\mathrm{E}}(L_{\mathrm{i}}(X_{\mathrm{a}}))) \tag{7}$$

式中：X_{a}为输入，X_h为最终输出。

经编码器处理，直接输出点云的特征向量数据，为下一步的融合提供输入。

3　基于UKF的多模态特征融合模型设计

特征级融合能够充分挖掘并利用不同数据源的

信息，进而有效增强对目标任务的预测性能[15]。本文采用特征级融合的特征向量拟合方式。如图4所示，根据获得的图像与点云原始数据，使用相应的特征提取器Transformer模型提取各自的特征向量，并据此设计一种非线性卡尔曼滤波器进行点云和图像连续特征融合。

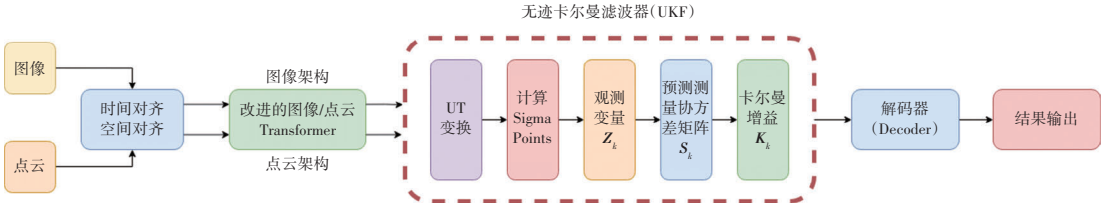


图4　多模态特征融合流程

在融合点云和图像特征向量前，确保二者的时间和空间对齐是非常重要的。时间对齐确保采集或生成的点云和图像的时间戳同步，空间对齐则确保点云和图像采集或生成的坐标系一致。为保证多模态融合中的空间一致性，本文使用KITTI数据集附带的calib_xxx.txt文件，将LiDAR点云投影至图像平面。依据相机的内参矩阵以及相机-激光雷达联合外参矩阵等参数，将LiDAR点云映射到图像坐标系，确保图像与点云特征在空间上对齐。且从KITTI数据集获得的数据能够保证时间差没有累计漂移，再根据时间戳保证输入点云与图像数据在时空上一一对应。

相较于通用滤波器，UKF[16]更适合处理图像与点云之间的非线性关系。UKF通过使用无迹变换(Unscented Transform，UT)来处理非线性关系，无迹变换通过在非线性函数周围选择一组代表性的点，更准确地估计非线性变换[17]，如图5所示。

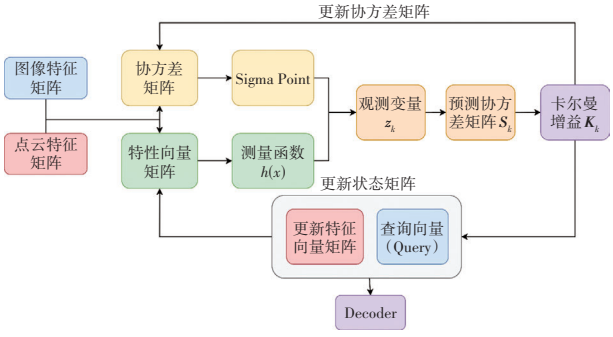


图5　特征融合UKF滤波器

为此，本文设计UKF用于图像和点云特征的多模态特征融合。首先，将图像特征矩阵和点云特征矩阵连接成更大的特征向量$\boldsymbol{x}_k$，表示k时刻的状态向量，并设置初始协方差矩阵$\boldsymbol{P}_0$，用于表示对初始状态的不确定性：

$$\boldsymbol{x}_k=\begin{bmatrix} I_{\mathrm{F}} \\ P_{\mathrm{CF}} \end{bmatrix} \tag{8}$$

$$\boldsymbol{P}_0=\begin{bmatrix} \sigma_{\mathrm{pc}}^2 & 0 \\ 0 & \sigma_{\mathrm{img}}^2 \end{bmatrix} \tag{9}$$

式中：I_{F}为图像特征(Image Features)，P_{CF}为点云特征(Point Cloud Features)，σ_{pc}^2为图像特征向量的方差，σ_{img}^2为点云特征向量可信度的方差。

无迹卡尔曼滤波器中的关键步骤之一是生成西格玛点(Sigma Points)，这些点在捕捉状态变量的不确定性和非线性特性方面发挥着重要作用，通过对状态向量进行采样生成，在非线性系统中能够更好地反映潜在的系统变化。

对式(7)协方差矩阵$\boldsymbol{P}_0$进行楚列斯基(Cholesky)分解，得到一个下三角矩阵$\boldsymbol{L}$，使得$\boldsymbol{P}_0=\boldsymbol{L}\boldsymbol{L}^{\mathrm{T}}$。在无迹卡尔曼滤波器中，通常选择$(2n+1)$个西格玛点，其中$n$为状态向量$\boldsymbol{x}_k$的维度，$\boldsymbol{X}_k^{(i)}$为包含$(2n+1)$个列向量的矩阵，每个列向量都是一个西格玛点：

$$\begin{aligned} \boldsymbol{X}_k^{(0)} &= \boldsymbol{x}_k \\ \boldsymbol{X}_k^{(i)} &= \boldsymbol{x}_k+\sqrt{(n+\lambda)\boldsymbol{P}_k}, \quad i=1,\cdots,n \\ \boldsymbol{X}_k^{(i+n)} &= \boldsymbol{x}_k-\sqrt{(n+\lambda)\boldsymbol{P}_k}, \quad i=1,\cdots,n \end{aligned} \tag{10}$$

式中：$\boldsymbol{x}_k$为k时刻系统的状态向量，λ为控制可调参数，$\boldsymbol{P}_k$为系统状态与测量的交叉协方差。

然后，计算预测测量均值和协方差矩阵，依据各特征可信度的权重将测量函数$h(x)$设置为：

$$h(\boldsymbol{x}_k)=\alpha\cdot(\boldsymbol{x}_k)+(1-\alpha)\cdot(\boldsymbol{x}_k) \tag{11}$$

式中：$a\in(0,1)$为特征融合权重系数。

随后将第i个西格玛点根据测量函数$h(x)$映射到测量空间得到预测测量值，即得到第i个观测变量：

$$z_k^i=h(\boldsymbol{X}_k^{(i)}) \tag{12}$$

式中：z_k^i为第k时刻的第i个观测变量。

依据计算预测测量均值$\bar{z}_k$、权重ω_i，可以获得预测测量协方差矩阵$\boldsymbol{S}_k$：

$$\bar{z}_k=\sum_{i=0}^{2n}\boldsymbol{W}_i^{\mathrm{m}}z_k^i \tag{13}$$

式中：$\boldsymbol{W}_i^{\mathrm{m}}$为第$i$个分量对应的均值权重矩阵。

$$\boldsymbol{S}_k=\frac{1}{2n+1}\cdot\sum_{i=0}^{2n}\omega_i(z_k^i-\bar{z}_k)(z_k^i-\bar{z}_k)^{\mathrm{T}} \tag{14}$$

计算卡尔曼增益$\boldsymbol{K}_k$：

$$\boldsymbol{K}_k=\boldsymbol{P}_k\boldsymbol{S}_k^{-1} \tag{15}$$

其中，$\boldsymbol{P}_k$可表示为：

$$\boldsymbol{P}_k = \sum_{i=0}^{2n} \boldsymbol{W}_i^c (\boldsymbol{X}_k^{(i)} - \boldsymbol{x}_k)(\boldsymbol{z}_k^i - \bar{\boldsymbol{z}}_k)^{\mathrm{T}} \tag{16}$$

式中：$\boldsymbol{P}_k$为k时刻状态向量$\boldsymbol{x}_k$对应的协方差矩阵，$\boldsymbol{W}_i^c$为第i个分量对应的协方差权重。

最后，更新状态$\hat{\boldsymbol{x}}_k$和协方差矩阵$\boldsymbol{P}_k$，并将更新后的状态作为UKF的输出结果：

$$\hat{\boldsymbol{x}}_k = \bar{\boldsymbol{x}}_k + \boldsymbol{K}_k \boldsymbol{y}_k \tag{17}$$

$$\boldsymbol{P}_k = \bar{\boldsymbol{P}}_k - \boldsymbol{K}_k \boldsymbol{S}_k \boldsymbol{K}_k^{\mathrm{T}} \tag{18}$$

式中：$\hat{\boldsymbol{x}}_k$为k时刻系统状态的最优估计值，$\bar{\boldsymbol{x}}_k$为k时刻系统状态的先验估计值，$\boldsymbol{y}_k$为系统k时刻的测量残差，$\bar{\boldsymbol{P}}_k$为k时刻状态的先验协方差矩阵。

本文使用无迹卡尔曼滤波器融合点云特征向量和图像特征向量，可以提高感知系统的准确性和稳定性。在输出阶段，无迹卡尔曼滤波器输出系统状态的估计，包括融合特征向量矩阵$\hat{\boldsymbol{x}}_k$和一组可学习的查询向量。二者输入解码器，通过多头注意力机制生成最终的检测结果。

4 试验验证与结果分析

本文采用KITTI数据集[18]作为模型训练数据集，使用实车采集的数据集进行模型训练结果的评测。本文试验所采集的数据源自湖北文理学院北区工科楼前方的封闭道路。数据采集全程使用python脚本驱动在机器人操作系统（Robot Operating System，ROS）环境下预先设置的激光雷达与相机节点，从而实现多传感器的同步采集。

根据本文设计的Transformer与连续特征融合模型，首先分别利用ViT模型与点云Transformer对RGB图像和激光点云信息进行特征提取。随后，通过这两个模型生成两组特征向量，进而利用无迹卡尔曼滤波器进行高效融合。融合后的特征向量组通过特征提取器进行进一步处理，以进行检测与输出。

4.1 试验环境

为验证本文提出的Transformer模型与卡尔曼滤波器特征融合在车辆智能驾驶感知系统中的有效性，在Python3.8、touch1.11.0、CUDA10.1环境下使用公开数据集进行仿真训练和评测，所使用计算机和配置环境的硬件参数包括：中央处理器为Inter® Core i9-13900KF；图形处理单元（Graphics Processing Unit，GPU）为NVIDIA GeForce RTX 4090，GPU内存容量为24 GB；设置训练轮次（Epochs）为8轮，批大小（Batch-Size）为256。

4.2 评价标准

常见的环境感知评价标准包括交并比（Intersection-over-Union，IoU）阈值、查准率（Precision）、查全率（Recall）、平均精度（Average Precision，AP）、平均精度均值（mean Average Precision，mAP）等[19]。本文经由无迹卡尔曼滤波器解码输出，因此选用AP和mAP在多个维度上对Transformer模型与卡尔曼滤波器特征融合算法进行评价，并且输出可视化结果加以分析：

$$P_{\mathrm{A}} = \int_0^1 H_p(r)\,\mathrm{d}r \tag{19}$$

$$P_{\mathrm{mA}} = \frac{1}{N}\sum_{i=1}^{N} P_{\mathrm{A}i} \tag{20}$$

式中：P_{A}为平均精度，P_{mA}为平均精度均值，$H_p(r)$为平均精度函数，r为查全率（Recall），N为待检测的目标类别数量。

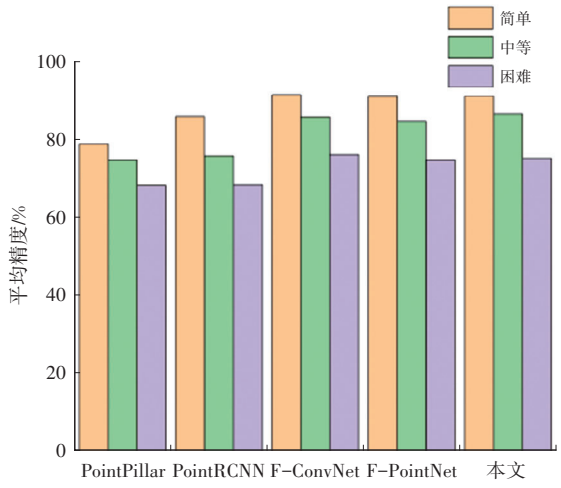
4.3 试验结果分析

选用点云区域卷积神经网络（Point Region-based Convolutional Neural Network，PointRCNN）[20]、PointPilla等激光点云算法与F-ConvNet、F-PointNet等主流多模态融合算法对本文提出的Transformer模型与特征融合算法进行对比验证，选用KITTI数据集分别在简单（Easy）、中等（Moder）和困难（Hard）场景下进行试验。

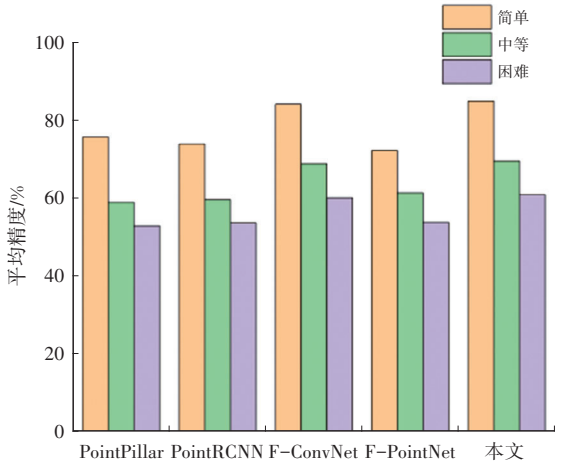
表1展示了本文方法与各类神经网络算法在车辆（Car）、行人（Pedestrian）、骑行者（Cyclists）类别上的平均精度和平均精度均值对比结果。车辆和骑行者的平均精度结果如图6所示，可以看出，由于目标的点云尺寸较大且包含丰富的语义信息，本文提出的方法通过多头自注意力机制并经由UKF融合获得的特征向量更为明显，特别值得注意的是，车辆类别和骑行者类别的平均精度均呈现出显著提升的趋势。

表1 本文方法与各类方法在KITTI数据集上的试验结果 %

方法	mAP	车辆三维检测AP			行人三维检测AP			骑行者三维检测AP		
		简单	中等	困难	简单	中等	困难	简单	中等	困难
PointPillar	60.58	78.85	74.69	68.13	51.93	43.39	40.90	75.65	58.87	52.83
PointRCNN	60.78	85.94	75.76	68.32	49.43	41.78	38.63	73.93	59.60	53.59
F-ConvNet	67.89	91.51	85.84	76.11	57.04	48.96	44.33	84.16	68.88	60.05
F-PointNet	65.20	91.17	84.67	74.77	57.13	49.57	45.48	72.26	61.37	53.78
本文	68.23	91.23	87.19	74.18	56.09	48.28	44.39	83.03	70.69	59.02



（a）车辆



（b）骑行者

图6　各类算法在车辆和骑行者类别上的平均精度

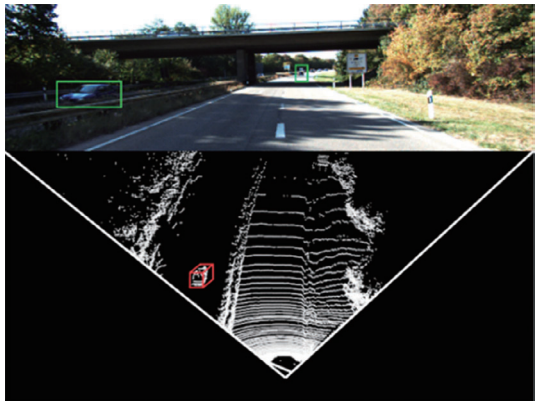
由表1和图6可知，本文方法在处理大尺寸目标点云时表现优异，能有效利用语义信息，显著提升了车辆与骑行者类别的检测性能。基于各单类别表现的全面提升，本文方法在多类别检测任务中实现了平均精度均值的显著提高。

4.4　结果可视化分析

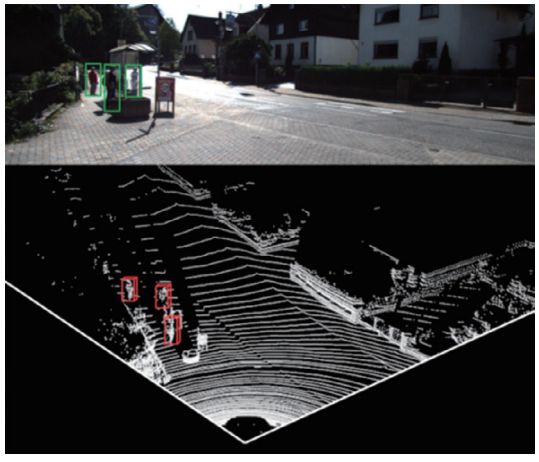
图7所示为本文特征融合模型的部分3D边界框预测结果，呈现了二维图像与点云的检测可视化效果。



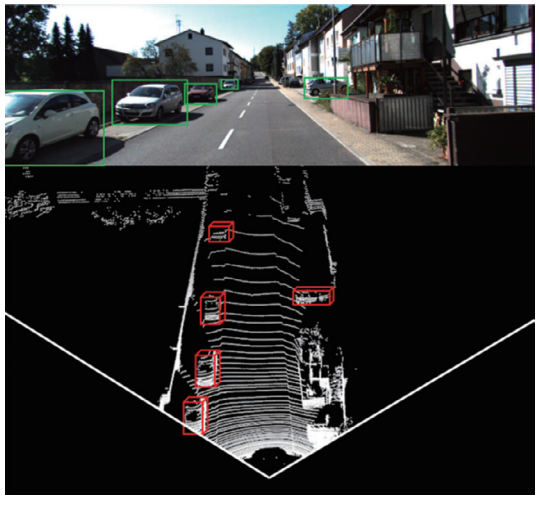
（a）示例1



（b）示例2



（c）示例3



（d）示例4

图7　KITTI数据集上的可视化检测结果

由图7可知，对于场景中不同距离范围的各类目标，本文构建的网络可以输出较为精确的边界框。从图7a、图7b、图7d中可以看出，即使在图片中只出现了车辆的小部分，本文特征融合模型仍然能够在点云中精确地检测出目标物体；从图7d中可以看出，本文特征融合模型在存在大量复杂检测对象的环境下仍可输出较为稳定的检测结果。图7b中出现了漏检的情况，是由于

距离过远而被检测目标周围几乎没有点云，导致目标边界框预测不准。

4.5 消融试验

为了分析融合方法在本文算法中的重要性，设置了2个对照算法：rmUKF，移除UKF融合模块，仅对图像和点云特征进行简单拼接，验证UKF在特征融合中的作用；rmTRS，将改进的Transformer替换为PointRCNN，以验证Transformer在特征融合中的作用。

表2对比了不同模块消融后的性能。模型在移除UKF模块后，在评测集中mAP下降3.4百分点左右，表明其通过动态权重优化显著提升了融合效果；将改进的Transformer模块替换为PointRCNN提取特征时，mAP下降约1.7百分点，证明本文的改进Transformer机制设计更适配图像与点云的多模态特征提取。消融试验结果证明了本文提出的改进Transformer-UKF能够提升目标检测模型的性能。

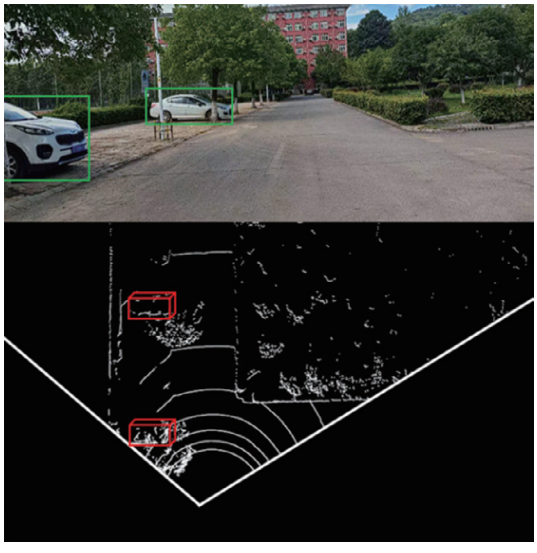
表2 消融试验结果 %

方法	mAP	车辆三维检测AP			行人三维检测AP			骑行者三维检测AP		
		简单	中等	困难	简单	中等	困难	简单	中等	困难
rmUKF	64.79	89.11	82.19	70.37	52.18	43.69	39.34	81.48	66.46	58.43
rmTRS	66.54	90.38	84.64	72.35	54.46	45.32	42.92	82.47	67.94	60.65
本文方法	68.23	91.23	87.19	74.18	56.09	48.28	44.39	83.03	70.69	59.02

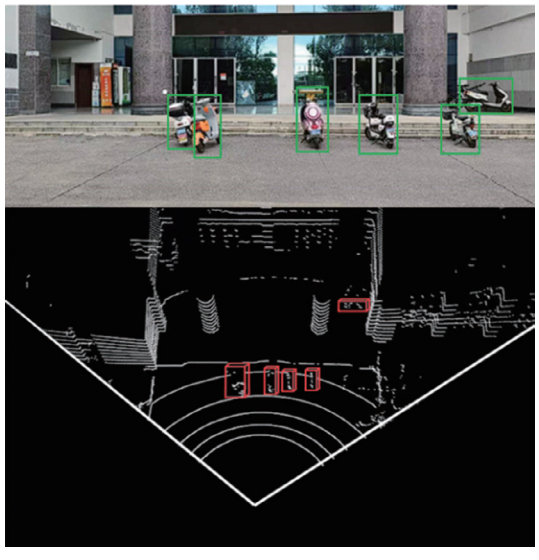
4.6 实车数据集推理验证

为验证训练后模型在真实场景下的泛化性能，本文基于湖北文理学院封闭测试场地采集的实车数据集开展推理验证试验。数据采集过程采用ROS框架下的多传感器同步系统，确保时空对齐精度，在良好光照条件、稳定天气及低密度干扰场景下采集并构建高质量数据集。

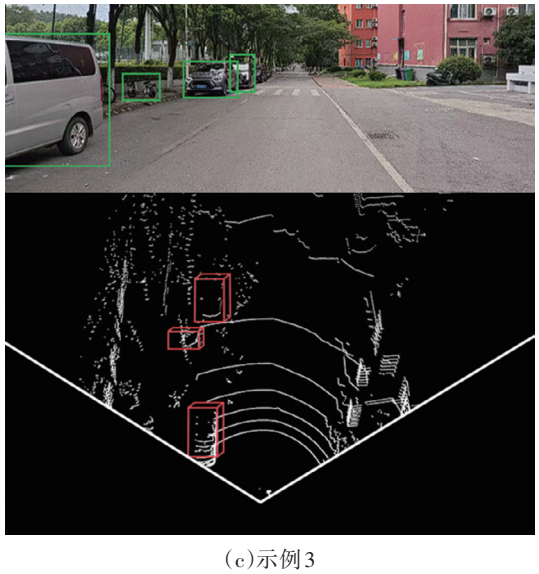
推理验证试验结果如图8所示：由图8a和图8c可知，即使处于较复杂的环境下，模型也能够输出较为稳定的检测结果；由图8b可知，虽然出现了点云的稀疏与遮挡情况，但由于图像补充了语义信息，模型仍然能够正确识别出目标。上述结果表明，该模型在实际数据推理阶段仍可准确识别目标，尤其是在光照稳定、目标分布清晰的封闭场景中，检测边界框定位准确。



(a)示例1



(b)示例2



(c)示例3

图8 实车数据集上的可视化检测结果

5 结束语

本文将深度学习Transformer架构与无迹卡尔曼滤波器相结合，提出了一种多模态融合感知算法，通过改进视觉Transformer模型提取图像特征，通过改进点云Transformer提取点云特征，并在两者的中间层次借助无迹卡尔曼滤波器对两组特征向量进行状态估计，从而实现了图像与点云的高效特征融合。

试验结果表明，相较于传统的激光点云算法，本文算法在各类场景下均取得了一定的提升，mAP获得了8百分点左右的大幅提升。相较于应用广泛的多模态融合算法F-PointNet和F-ConvNet，mAP分别提升了0.34百分点和3.03百分点，特别是相较于F-PointNet，在车辆和骑行者对象的中等难度检测任务中，平均精度分别提升了2.52百分点和9.32百分点，充分证明了通过多模态数据检测3D目标的有效性，并且在实车数据集中也能推理出较为合理的结果。通过对目标点云尺寸和语义信息的充分考虑，本文提出的方法在车辆、行人和骑行者等多类别目标检测任务中表现出色。

参 考 文 献

[1] 刘永刚, 于丰宁, 章新杰, 等. 基于激光点云与图像融合的3D目标检测研究[J]. 机械工程学报, 2022, 58(24): 289-299.
LIU Y G, YU F N, ZHANG X J, et al. Research on 3D Object Detection Based on Laser Point Cloud and Image Fusion[J]. Journal of Mechanical Engineering, 2022, 58(24): 289-299.

[2] CHARLES R Q, LIU W, WU C X, et al. Frustum PointNets for 3D Object Detection from RGB-D Data[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 918-927.

[3] CHARLES R Q, SU H, MO K C, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 77-85.

[4] QI C R, YI L, SU H, et al. Pointnet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space[C]// Proceedings of the Advances in Neural Information Processing Systems. Long Beach, CA, USA: NeurIPS, 2017: 5105-5114.

[5] YOO J H, KIM Y, KIM J S, et al. 3D-CVF: Generating Joint Camera and LiDAR Features Using Cross-View Spatial Feature Fusion for 3D Object Detection[C]// Computer Vision-ECCV, Lecture Notes in Computer Science. Switzerland: Springer, Cham, 2020.

[6] WANG Z X, JIA K. Frustum ConvNet: Sliding Frustums to Aggregate Local Point-Wise Features for Amodal 3D Object Detection[C]// IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macau, China: IEEE, 2019: 1742-1749.

[7] LIANG M, YANG B, WANG S L, et al. Deep Continuous Fusion for Multi-Sensor 3D Object Detection[C]// 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer-Verlag Berlin, 2018: 663-678.

[8] LIANG M, YANG B, CHEN Y, et al. Multi-Task Multi-Sensor Fusion for 3D Object Detection[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 7337-7345.

[9] CHEN Y K, LIU J H, ZHANG X Y, et al. VoxelNeXt: Fully Sparse VoxelNet for 3D Object Detection and Tracking[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 21674-21683.

[10] JIAO Y, JIE Z Q, CHEN S X, et al. MSMDFusion: Fusing LiDAR and Camera at Multiple Scales with Multi-Depth Seeds for 3D Object Detection[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 77-85.

[11] LI X, MA T, HOU Y N, et al. LoGoNet: Towards Accurate 3D Object Detection with Local-to-Global Cross-Modal Fusion[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC, Canada: IEEE, 2023: 178-185.

[12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C]// 9th International Conference on Learning Representations. Online, United States: OpenReview.net, 2021.

[13] GUO M H, CAI J X, LIU Z N, et al. PCT: Point Cloud Transformer[J]. Computational Visual Media, 2021, 7(2): 187-199.

[14] 吴庆, 彭育辉, 黄炜, 等. 基于点云鸟瞰图的实时车辆目标检测[J]. 汽车技术, 2023(9): 35-42.
WU Q, PENG Y H, HUANG W, et al. Real-Time Vehicle Target Detection Based on Point Cloud Bird's Eye View[J]. Automobile Technology, 2023(9): 35-42.

[15] 任柯燕, 谷美颖, 袁正谦, 等. 自动驾驶3D目标检测研究综述[J]. 控制与决策, 2023, 38(4): 865-889.
REN K Y, GU M Y, YUAN Z Q, et al. 3D Object Detection Algorithms in Autonomous Driving: A Review[J]. Control and Decision, 2023, 38(4): 865-889.

[16] OM K, BOUKOROS S, NUGALIYADDE A, et al. Modelling Email Traffic Workloads with RNN and LSTM Models[J]. Human-Centric Computing and Information

Sciences, 2020, 10(1): 39-44.

[17] 陈华志. 基于雷达/红外数据融合多目标跟踪滤波算法的研究[D]. 成都: 电子科技大学, 2011.
CHEN Z H. Research on Multi-Target Tracking Filtering Algorithm Based on Radar/Infrared Data Fusion[D]. Chengdu: University of Electronic Science and Technology, 2011.

[18] 彭育辉, 郑玮鸿, 张剑锋. 基于深度学习的三维目标检测方法综述[J]. 汽车技术, 2020(9): 1-7.
PENG Y H, ZHENG W H, ZHANG J F, et al. A Review of Deep Learning Based 3D Target Detection Methods[J]. Automobile Technology, 2020(9): 1-7.

[19] EVERINGH M, VAN GOOL L, WILLIAMS C K I, et al. The Pascal Visual Object Classes (VOC) Challenge[J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[20] SHI S S, WANG X G, Li H S, et al. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 770-779.

（责任编辑　斛　畔）

修改稿收到日期为2025年4月7日。

《汽车文摘》征文

《汽车文摘》（月刊）于1963年7月3日创刊，由国务院国有资产监督管理委员会主管、中国第一汽车集团有限公司主办，为中国汽车工程学会会刊。《汽车文摘》以“览全球汽车技术文献，指中国汽车技术之道”为使命，以打造“中国汽车前沿与创新技术传播与交流的重要平台”为愿景，致力于成为汽车领域最具影响力的综述类期刊。

《汽车文摘》深耕电动化、智能化、网联化、共享化和智能制造5大方向和10大领域，聚焦新能源、燃料电池与混合动力汽车、智能网联汽车、氨氢融合零碳和碳中性燃料、汽车安全、健康与舒适、碳达峰与碳中和、生命周期评价（LCA）与技术经济分析、智能制造、材料轻量化与一体化压铸、飞行汽车前沿与创新技术综述论文，揭示相关领域的新动态、新趋势、新技术和新进展，为广大科研和工程技术人员进一步发展这一领域提供新突破口、新出发点和新基准。

欢迎高等院校师生、研发工程技术人员、技术管理人员，充分发挥专业领域优势，深度挖掘国内外高影响力学术期刊与其它文献，形成某个技术领域前沿综述。

《汽车文摘》选题范围：

电动化：混合动力关键技术；动力电池关键技术；先进充电技术；电驱动系统及电力电子技术；智能底盘及子系统线控关键技术；燃料电池动力系统设计与优化。

智能化：新型电子电气架构；自动驾驶感知、决策与运动控制；智能新能源汽车测试评价方法与工具链；车辆智能安全技术。

网联化：智能网联云控技术；车用通信及网络技术；车路协同技术；汽车人因、人机交互与智慧座舱；信息安全与预期功能安全；车网融合（V2G）及应用。

低碳化：汽车节能与排放技术；清洁能源动力系统技术；碳达峰、碳中和；氢燃料制、储、运、加及安全管控技术；生命周期评价（LCA）、标准法规与技术经济分析；氨氢融合零碳和碳中性燃料。

轻量化：新能源汽车新材料技术；混合材料轻量化设计；一体化压铸。

共享化：区块链技术与移动出行；车辆大数据挖掘方法与应用案例。

燃料电池：电池堆、电池系统与基础设施。

智能制造：机器人与自动化控制、四大工艺、物流技术、设计-制造-服务。

颠覆式出行：飞行汽车；未来低空智能交通体系及其关键技术。

汽车安全：主被动安全与融合；智能安全；健康与舒适。

《汽车文摘》发表论文的独特优势：

《汽车文摘》是国家级刊物、中国汽车工程学会会刊、汽车领域唯一的综述期刊。《汽车文摘》不收版面费、4个月左右可发稿。

投稿要求：

1. 综述篇幅在10 000～15 000字（6~10页），图文并茂，图、表和公式非原创要求标注引用文献；

2. 请按科技论文要求撰写文章摘要，摘要中文字数在200±10字；

3. 文章必须附有公开发表、体现本领域最新研究成果和高影响力出版物作为参考文献，一般要求参考文献在20篇以上，一半左右为外文参考文献，且在文中标注所引用文献；

4. 来稿保密审查工作由作者单位负责，确保署名无争议，文责自负；

5. 切勿一稿多投。

《汽车文摘》投稿网址：http://www.qcwz.cbpt.cnki.net

邮箱：autodigest@faw.com.cn

《汽车文摘》编辑部