Operations Research and Fuzziology 运筹与模糊学, 2023, 13(3), 2565-2583
Published Online June 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.133257

Hans汉斯

MS3D-Net：一种端到端的多传感器融合3D检测网络

程家镯，吴训成*，相文彬，吴玉坤

上海工程技术大学，机械与汽车工程学院，上海

收稿日期：2023年4月7日；录用日期：2023年6月23日；发布日期：2023年6月30日

摘　要

随着自动驾驶技术的发展，对车辆环境的3D感知要求越来越高，而多传感器融合可以很好的满足这一要求。针对目前融合技术中存在的网络设计不系统、信息丢失过大和融合策略粗糙问题，本文设计了一种端到端的多传感器融合3D检测网络——MS3D-Net。为秉承系统设计理念找到最优的多模态融合层级，先提出了新的融合层次划分法，再基于Faster-Rcnn源架构的检测模型中通过控制变量法，找到了最适合的特征融合层级；为降低跨模态数据融合过程中的信息损失，设计了新的高维表示，并提出与之对应的融合方法3D-T；为提高融合策略的精细度，提高融合检测精度，受Long Short-Term Memory (LSTM)机制启发拓展设计了中晚期门控递归融合单元，同时为提升图像特征的提取效率，提出了CP卷积。最后在KITTI数据集上进行训练与验证，本文方法在提高检测精度的同时又保证了检测速度。

关键词

传感器融合，高维表示，3D-T，门控递归

MS3D-Net: An End-to-End Multi-Sensor Fusion 3D Detection Network

Jiazhuo Cheng, Xuncheng Wu*, Wenbin Xiang, Yukun Wu

Mechanical and Automotive Engineering, Shanghai University of Engineering and Technology, Shanghai

Received: Apr. 7th, 2023; accepted: Jun. 23rd, 2023; published: Jun. 30th, 2023

Abstract

With the development of autonomous driving technology, 3D perception of the vehicle environ-

*通讯作者。

文章引用: 程家镯, 吴训成, 相文彬, 吴玉坤. MS3D-Net: 一种端到端的多传感器融合 3D 检测网络[J]. 运筹与模糊学, 2023, 13(3): 2565-2583. DOI: 10.12677/orf.2023.133257

ment is becoming more and more demanding, and multi-sensor fusion can meet this requirement very well. To address the problems of unsystematic network design, excessive information loss and rough fusion strategies in current fusion technologies, this paper designs an end-to-end multi-sensor fusion sensing network—MS3D-Net. In order to find the optimal multi-modal fusion hierarchy in adherence to the system design concept, a new fusion hierarchy division method is first proposed, and then the most suitable feature fusion level was found by the control variable method in the detection model based on the Faster-Rcnn architecture. In order to reduce the information loss during cross-modal data fusion, a new high-dimensional representation is designed, and proposes the corresponding fusion method 3D-T. In order to improve the fineness of the fusion strategy and increase the fusion detection accuracy, a medium-late gated recursive fusion unit is extended and designed inspired by the long short-term memory (LSTM) mechanism. At the same time, in order to improve the efficiency of image feature extraction, CP convolution is proposed. Finally, the method is trained and validated on the KITTI dataset, and the detection speed is guaranteed while improving the detection accuracy.

Keywords

Sensor Fusion, Gowey Said, 3D-T, Gated Recursion

1. 引言

近年来，深度学习和自动驾驶的兴起导致了3D检测的快速发展。目前的3D检测方法主要基于LiDAR点云[1] [2] [3] [4] [5]，而点云的稀疏性极大地限制了它们的性能。尤其是在包含遮挡、背景干扰、噪声等的复杂场景中实现对感兴趣目标的检测识别与分割中，大部分融合机制是通过将点云映射到图像中进行融合，然而点云匹配上的像素点较少不能完全利用好图像特征，稀疏的LiDAR点云在远处和遮挡区域提供的信息较差，从而难以生成精确的3D边界框。许多多传感器融合方法提出了解决这个问题的方案。MV3D [6]引入了一种RoI融合策略，在第二阶段融合图像和点云的特征。AVOD [7]提出从图像特征图和BEV特征图中融合全分辨率的特征作物，以实现高召回率。MMF [8]利用2D检测、地面估计和深度完成来辅助3D检测。在MMF中，伪点云用于骨干特征融合，深度完成特征图用于ROI特征融合。尽管他们取得了巨大的成功，但是其融合是统一转换为2D后再进行检测的，缺乏自顶向下的框架搭建，且精度方面仍然处于粗融合层次，以及点云会在3D~2D的时候产生较大的损失。

针对以上问题，本文提出如下解决方案：

- 提出了适应新技术要求的融合层级，分析对比不同融合方法之间的性能，并针对车辆前方环境的特点，自顶向下设计了基于特征级融合方案，使得框架与应用环境更契合。
- 受LSTM机制启发，拓展设计中晚期门控循环融合，有效降低了多传感器之间的对齐偏差，提高融合检测精度。
- 设计一种新的高维度表示，且匹配了专门维度编码器3D-T，专为数据融合设计，有效降低了跨模态数据融合过程中的信息损失。
- 为更好地在三维空间中提取伪点云的特征，提出了CP卷积方法。
- 通过使用数据集进行训练与验证，证明了本文方法的有效性。

2. 相关工作

2.1. 多源异构传感器融合层级

传感器融合是一个十分流行的课题，从其诞生以来的几十年来一直活跃在研究领域。多源异构传感器的数据融合传统包括三个主要步骤：时空标定、数据的对齐、异构数据之间的关联融合，然后就是具体任务步骤，如状态估计、车道线检测和语义分割等[9]。首先传感器融合的划分目前没有明确的定义标准来分类，大部分研究者根据融合的时期不同可以分为：前融合、中间融合和后融合，也可以称为数据级融合、特征级融合以及决策级融合。前融合指的是融合未经过处理的传感器数据；中间融合指的是各类传感器数据先经过特征提取或其他处理后再进行融合，接着再通过别的算法进行运算；后融合指的是各类传感器数据根据不同的处理算法得到处理结果，再通过加权融合得到最终决策的融合。但是，随着现价段模型复杂度的增加，从位置上越来越多的融合被划分为中间融合，这样就不能很好地区分融合层级。因此，本文提出了一个新的划分方法，通过融合后的结果在整个算法网络中充当的作用来划分。这样不仅能更有区分度，还能更好地反应融合对模型算法的作用，具体定义如图 1 所示。

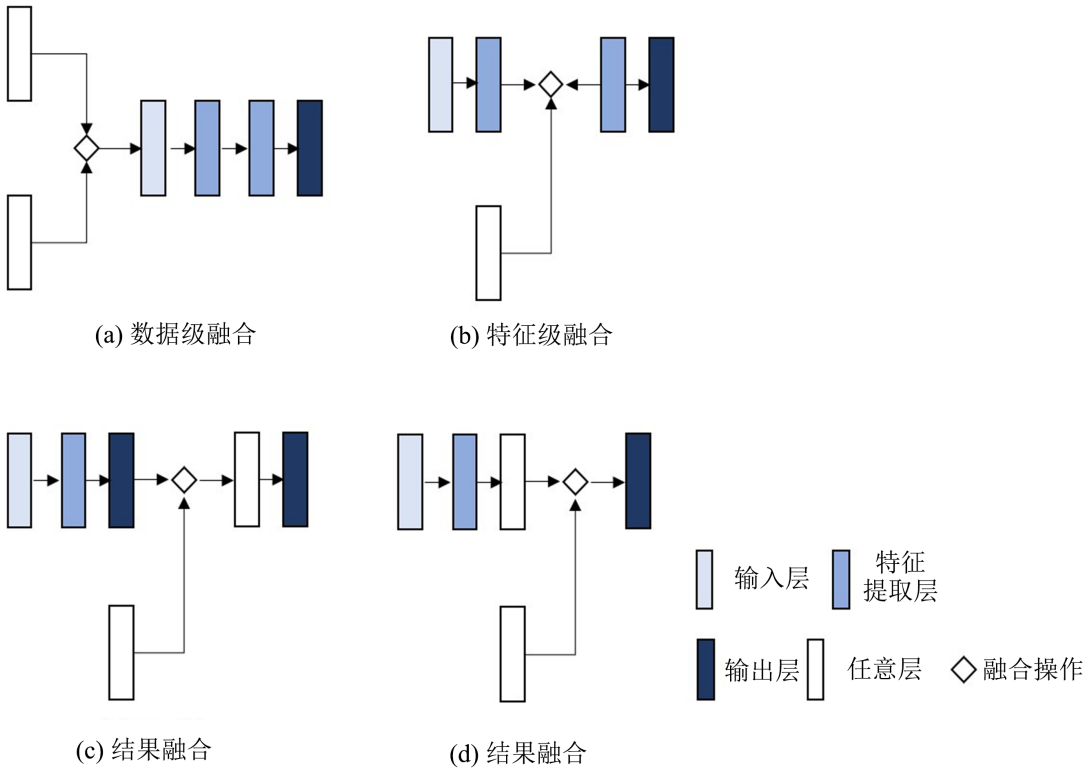


Figure 1. Examples of different fusion levels
图 1. 不同融合层次示例

2.2. 循环神经网络及其变体

循环神经网络(RNN)经常在有时间序列的数据中使用，这样可以更好地处理不同时间序列之间数据的关系。通过简单的循环网络结构如图 2 所示，可以知道网络记忆中上一步的数据信息并且通过该信息影响输出结果。其中 UVW 都是权重矩阵，分别代表输层出到隐藏层、隐藏层到输出层、上一次隐藏层结果到这一次隐藏层的权重；XSO 均是向量值，分别代表输入值、隐藏值和输出值。

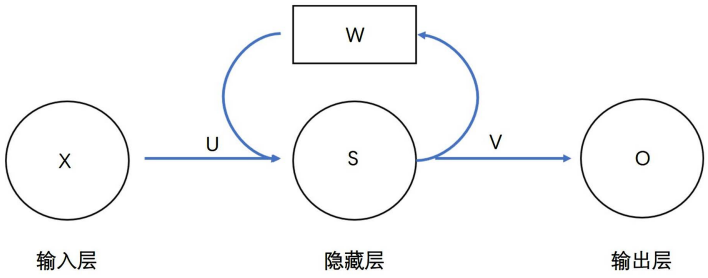


Figure 2. Simple recurrent network structure
图 2. 简单循环网络结构

而在处理时序过长数据时，循环神经网络容易产生梯度爆炸和梯度消失的问题。为了解决这一问题，设计了更复杂的 LSTM 网络。LSTM 首先在 1997 年由 Hochreiter & Schmidhuber 提出，2012 年后随着深度学习在的兴起，LSTM 又在若干学者(Felix Gers, Fred Cummins, Santiago Fernandez, Justin Bayer, Daan Wierstra, Julian Togelius, Faustino Gomez, Matteo Gagliolo, and Alex Gloves)的努力下更新迭代，由此便形成了比较完善的 LSTM 框架，并且在很多领域得到了广泛的应用。其相对循环神经网络更复杂，尤其是针对于记忆的长短、以及应该遗忘何种信息？记住何种信息？的问题通过：遗忘门、输入门、输出门等结构来较为完美地解决。LSTM 单元如图 3 所示，整体来看还是一个循环神经网络架构，只是在隐藏层部分增加了复杂度，多了遗忘门、输入门和输出门三个门控单元，以及隐藏和候选单元状态(ht、ct)。单元状态表示当前 t 时刻的状态内容(记忆内容)，而隐藏状态是模型在时间 t 时的输出。

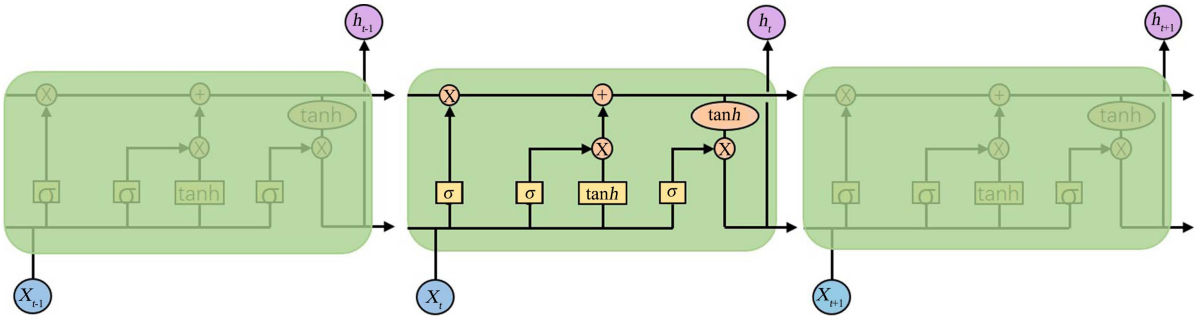


Figure 3. LSTM unit
图 3. LSTM 单元

两个隐藏状态执行重要的功能，即：慢速状态 C_t，用于对抗梯度消失问题；快速状态 h_t，允许 LSTM 在短时间内做出复杂的决定。每个门控状态都执行独立的任务，调节单元状态和隐藏状态的记忆长短与结合程度。

2.3. 数据融合计算方法

在目标检测的融合方法中，出现过多种多样的融合计算方法，主要可以归为连接法、合并法、相加法和子网络法 4 种。

连接法：这种方法是最直接的方法，通过将不同模态的数据直接相连，最普遍的表现形式就是数据在维度上的增加。而直接法中又分为两种，① 把第二个传感器的数据拼接到第一个传感器的数据上，总体的层数是不变的只是每层的尺寸变大，被称为 concatenate；② 在第一个传感器数据通道的基础上将第二个传感器数据平行扩展为新的通道，使得通过拓展通道来融合数据，被称为 extended channel。

合并法：该方法适用于多源异构传感器数据对共同目标输出的融合，首先各个分支数据通过特征提

取后产生各自的候选元素，然后通过该法进行候选框融合[10] [11]，具体的操作通常是加权求和。例如相机和 Lidar 分别获得 RGB 和点云信息，分别通过各自的特征提取网络获得候选框，最后再通过加权求和获得最终的结果。

相加法：通常在特征图融合和 ROI 处理中出现，前者通常是直接将特征图进行合并，后者又分为两种其一是将 ROI 进行合并[6]其二是将 ROI 叠加预设成特征检测的约束条件[12]。

子网络法：这种方法是近年来学者为了提升决策融合的效果提出的，通常是在现有检测网络的结果基础上再设计一个子网络对其进行特征提取，称该子网络为网中网(Network in Network, NiN)。例如对两个分支的结果预测框进行融合，此类融合输出与决策级高层特征融合的方法非常相似。

2.4. 深度补全

深度补全的目的是在彩色图像的引导下，从一个稀疏的深度图中预测产生得到一个密集的深度图。最近，人们提出了许多高效的深度补全方法。文献[13]利用双分支骨干网络实现了一个精确高效的深度补全网络。文献[14]提出了一种多假设的深度表示方法，可以在前景和背景之间锐化深度边界。尽管深度补全任务的主要目的是为下游任务服务，但在三维检测中真正使用深度补全的方法很少。在基于图像的三维物体检测中，有一些工作如文献[15] [16]中使用深度估计来生成伪点云。然而，由于缺乏准确或足够的原始 LiDAR 点云，它们的性能受到很大限制。

3. 跨模态融合方法

本文设计了一个实现智能车辆敏感环境感知的融合网络——MS3D-Net，整体分为三个部分——激光雷达点云流、伪点云流和密集高精融合头，如图 4 所示，其中：① 激光雷达流，仅根据原始点云数据并用 RPN 来生成 3D ROI 特征；② 伪点云流，利用所提出的 CP 卷积提取点特征，并利用稀疏卷积提取体素特征，再经过 RPN 操作得到 3D ROI 特征；③ 密集高精融合头，以网格化的方式融合原始点云和伪点云的 3D ROI 特征并生成粗预测框，且在 LGRF 最小单元之前通过该预测框来修正原始点云和伪点云的 3D ROI 特征，具体是经过裁剪和高维拼接粗预测框，然后经过 LGRF 单元融合，最后将输出结果传输至下一步对应的检测头。具体内容本节将分为六小节来阐述该网络的设计与搭建过程，第一小节通过控制变量方法确定了融合层级，第二小节改良了具体的融合计算方法，第三小节设计了融合框架的主体最小计算单元，第四小节对同步增强操作进行了具体阐述，第五小节具化了 CP 卷积，第六小节匹配了具体的损失函数。

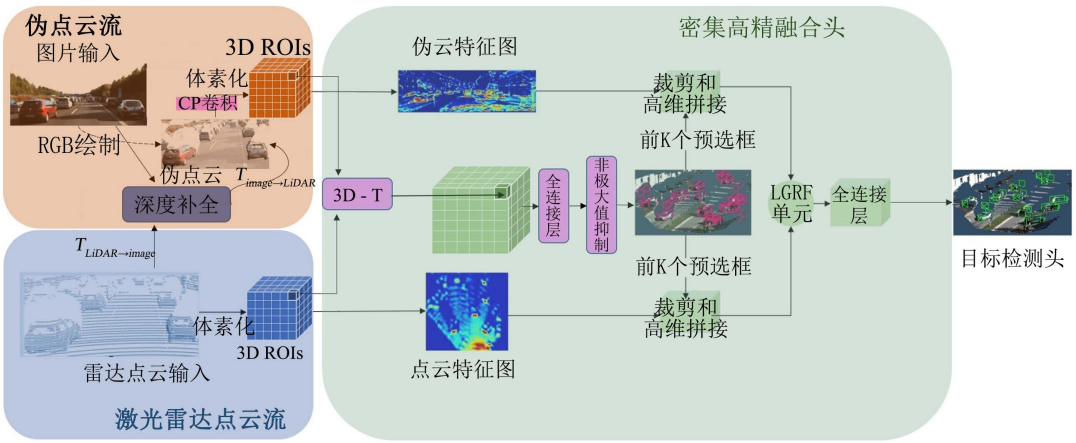


Figure 4. MS3D-Net algorithm diagram
图 4. MS3D-Net 算法图

3.1. 确定融合层级

首先，为了设计出能在各种不同天气中实现良好融合效果的网络模型，本文按照不同融合时期的鲁棒性、实现效果等性能评价指标，分析对比了几种主要的融合层级，来选出最合适的融合层级。本着控制变量准则，先通过开源库 KITTI 收集并制作相同的数据集，接着通过基于统一方法的物体目标检测器，使用相同的训练条件和相同的输入特征，对不同的融合方法进行了定性比较。从融合效果分析、数据冗余性分析两个角度分析融合的合理性，选择最好的融合时期。

1) 检测器源架构

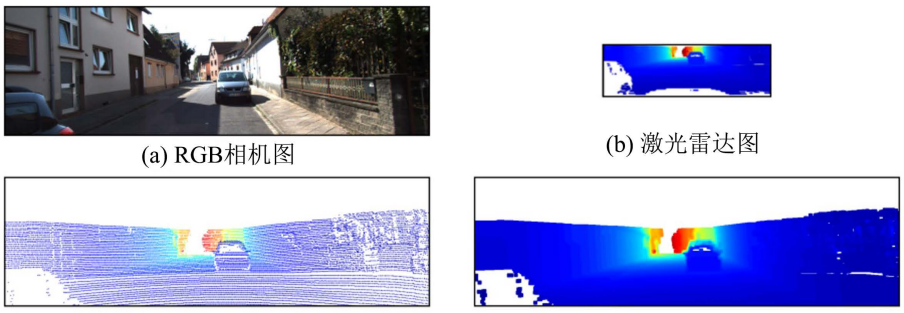
目前，有很多检测器，而这些检测器大多是三种源架构之一的变化或修改得来的，即 Faster-RCNN [17]、R-FCN [18] (基于区域的全卷积网络)和 SSD [19] (单步多目标检测器)。Huang 等人[20]分析了关于这三种源架构的利弊的更多细节，他们在准确性、速度和内存使用方面对它们进行了比较。基于这种比较，本文中选择了 Faster-RCNN，因为它在准确性方面优于其他元架构，而在速度和内存使用方面没有太多的损失。

2) 数据集

收集制作的 KITTI 数据集由 7481 张训练图像和 7518 张测试图像组成，相应的激光雷达数据由 Velodyne 的 64 线激光雷达采集得到。总的来说，标注的数据集包含 6 个不同类别(汽车、卡车、电车、行人、自行车、货车)的大约 8 万个标记的物体。数据集被分成训练集、验证集和测试集，测试集由数据集中前 500 张图像组成，训练集由接下来的 6500 张图像组成，验证集由剩余的图像组成。

3) 预处理

为了通过神经网络融合照相机和激光雷达数据，必须将传感器数据转换为合适的格式，以便将数据送入神经网络。对于 RGB 图片数据，通过 reshape 操作调整为神经网络的预定输入尺寸，同时保持其长宽比不变如图 5 中(a)所示。对于激光雷达数据，为了降低计算损耗先相机视场外的点云数据通过 ROI 操作去除，再将三维激光雷达点投射到图像平面上(如式(1)~(2))，形成一个稀疏的深度图(如式(3)~(8))，最后通过邻域均值法进行插值得到稠密的深度图(如式(9))，如图 5 所示。



(a) RGB相机图　(b) 激光雷达图

(c) 稀疏深度图　(d) 稠密深度图

Figure 5. Data types in preprocessing
图 5. 预处理中的数据类型

$$c = \left[\frac{\arcsin\left(\frac{z}{\sqrt{x^2 + y^2 + z^2}} \right)}{\Delta\phi} \right] \tag{1}$$

$$r=\left[\frac{a\tan 2(y,x)}{\Delta\theta}\right] \tag{2}$$

其中关于激光雷达图像的位置(c, r)，$\Delta\phi$ 和 $\Delta\theta$ 是激光雷达传感器的平均垂直和水平角度分辨率，当 RGB 中存在一个相应的激光雷达映射点是时候，每个激光雷达图像像素(c, r)分配相应的深度值 $d=\sqrt{x^2+y^2}$，不存在时深度值 d 会设置为无穷大。激光雷达图像 B 如图 5 中(b)所示。

$$x_n=-\frac{x}{z} \tag{3}$$

$$y_n=-\frac{y}{z} \tag{4}$$

$$r^2=x_n^2+y_n^2 \tag{5}$$

$$f(r)=1+k_1r^2+k_2r^4+k_3r^6 \tag{6}$$

$$m_d=\begin{pmatrix} f(r)x_n+2k_4x_ny_n+k_5\left(r^2+2x_n^2\right) \\ f(r)y_n+2k_5x_ny_n+k_4\left(r^2+2y_n^2\right) \\ 1 \end{pmatrix} \tag{7}$$

$$\tilde{P}_i=\begin{pmatrix} f_x & \theta & o_x \\ 0 & f_y & o_y \\ 0 & 0 & 1 \end{pmatrix}\cdot m_d \tag{8}$$

式中是使用具有镜头畸变的针孔相机模型来进行矩阵变换的。其中，f_x 和 f_y 是分别是沿 x 轴和 y 轴的焦距，(o_x,o_y) 是相机的光学中心，θ 是其偏斜参数，$k_i, i\in\{1,\cdots,5\}$ 为畸变参数。与激光雷达图像类似，每个投影点 $\tilde{P}_i$ 被赋予深度值 $d=\sqrt{x^2+y^2}$，将其他无映射点的点深度设置为无穷大。该稀疏深度图像 C 的示例如图 5(c)所示。

$$d^*=\sum_{p\in n}\frac{1}{N_n}d_p \tag{9}$$

该式表示以 p 为中心取小邻域 N 内的激光雷达点的平均深度的值，通过插值来填充没有深度信息的像素。该稠密深度图像 D 的示例如图 5(d)所示。

4) 实验分析

最后，基于 Faster-RCNN 目标检测器搭建四种不同层次的融合网络，并通过目标检测任务进行评估各个层级网络的性能。为了模拟各种不同天气中网络的性能，在训练过程种人工加入了恶劣天气的干扰、增加噪声，检测结果如图 6 所示。

如图中，白色区域被随机拟合到相机和深度图像中。这是因为在耀眼的阳光下，相机图像中含有白点，而激光雷达传感器也会因为激光束的反射和吸收而在雨雪中提供较少的点。另外，当将另一个输入流设置为无穷大时，仅使用相机或激光雷达数据模拟传感器故障。

具体各个层级融合网络的检测性能如表 1 所示：

结果表明，传感器数据融合越靠近特征融合，目标检测器的检测率就越高，同时在部分数据失效时(通过增加噪声验证，验证结果如图 6)鲁棒也更好。综上所诉，确定了最符合本文要求的融合层级——特征融合。

Figure 6. Detection results at different noise levels
图 6. 不同噪声下的检测结果

Table 1. Performance test results for converged networks at all levels
表 1. 各层级融合网络性能测试结果

融合方法	输入数据	类别							噪声验证	计算耗时
		小汽车	卡车	电车	行人	自行车	厢式货车	mAP		
无融合	RGB	0.782	0.883	0.878	0.505	0.693	0.780	0.753	0.71	64.54 ms
数据融合	RGB & 稠密图	0.776	0.912	0.871	0.509	0.729	0.778	0.762	0.72	70.98 ms
特征融合	RGB & 稠密图	0.795	0.923	0.864	0.526	0.729	0.789	0.771	0.76	94.57 ms
决策融合	RGB & 稠密图	0.782	0.890	0.866	0.521	0.717	0.787	0.763	0.73	115.43 ms

3.2. 高维表示

针对特征利用率低这一问题，本文提出新的高维表示，使得特征保留尽可能多。高维表示为 $\tilde{S}^i$ 且是一个时间序列，$\tilde{S}^i=\left\{\tilde{S}_1^i,\tilde{S}_2^i,\cdots,\tilde{S}_t^i,\cdots,\tilde{S}_T^i\right\}$，$i\in[1,M]$，其中每个子集包含 (x,y,z,r,g,b,t) 七维数据。本文模型的目标是共同学习最佳传感器融合频率和模式组合，以正确预测期望的分类/回归目标。首先对输入数据进行预处理，使用适当的编码器 3D-T 在时间融合之前将信号数据变换到相同的维度。图像数据经过深度补全，转化到伪云特征图像表示，三维点云经过 3D ROI 操作变换到点云特征图像表示，两个支路的数据均通过 ROI 操作和特征提取器后，先通过裁剪和拼接操作得到新的高维表示后再通过下文的级联单元 LGRF 将 RGB 特征与点云融合得到高维深度图像，如图 7 高维流程图所示。

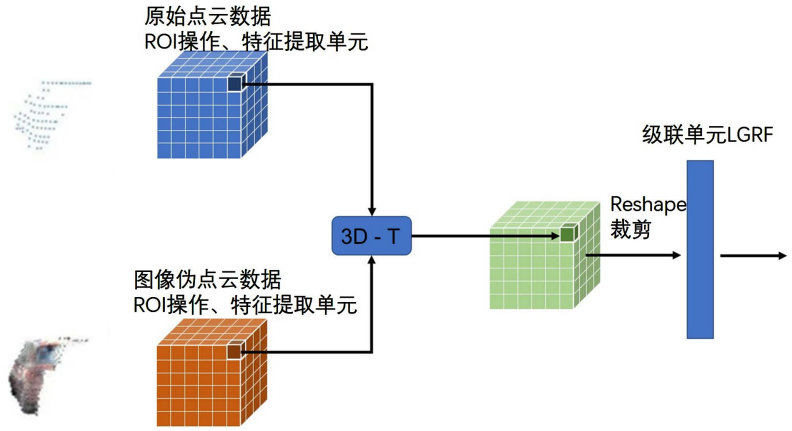


Figure 7. High-dimensional representation algorithm diagram
图 7. 高维表示算法图

1) 图像数据流分支

通过单目摄像头获取得的数据，首先经过高斯处理去除噪声。由上文中得到的稠密深度图像 D，通过空间关系矩阵映射变换如式子(10)~(13)，可以得到一帧伪云 P。映射得到像素点对应的空间坐标后，再通过拼接法(concatenate)得到高维表示(x,y,z,r,g,b,t)。

$$x_n = -x \tag{10}$$

$$f(r) = 1 + k_1 r^2 + k_2 r^4 + k_3 r^6 \tag{11}$$

$$m = \begin{pmatrix} f(x_n) \\ f(y_n) \\ 1 \end{pmatrix} \tag{12}$$

$$\tilde{\boldsymbol{P}}_i = \begin{pmatrix} f_x & \theta & O_x \\ 0 & f_y & O_y \\ 0 & 0 & 1 \end{pmatrix}^{-} \cdot m \tag{13}$$

式中 x 和 y 分别表示 RGB 图像中的像素点坐标，坐标转换矩阵是的空间关系转换的逆矩阵，其他函数定义与上文中一致。

2) 雷达点云数据流分支

通过 64 线激光雷达获取到点云数据后，为了降低计算损耗先相机视场外的点云数据通过 ROI 操作去除。其中，ROI 是根据 VoxelNet 中的 rpn 来进行操作的，如图 8。经过操作后得到点云特征图，且数据维度与前文的伪点云一致。

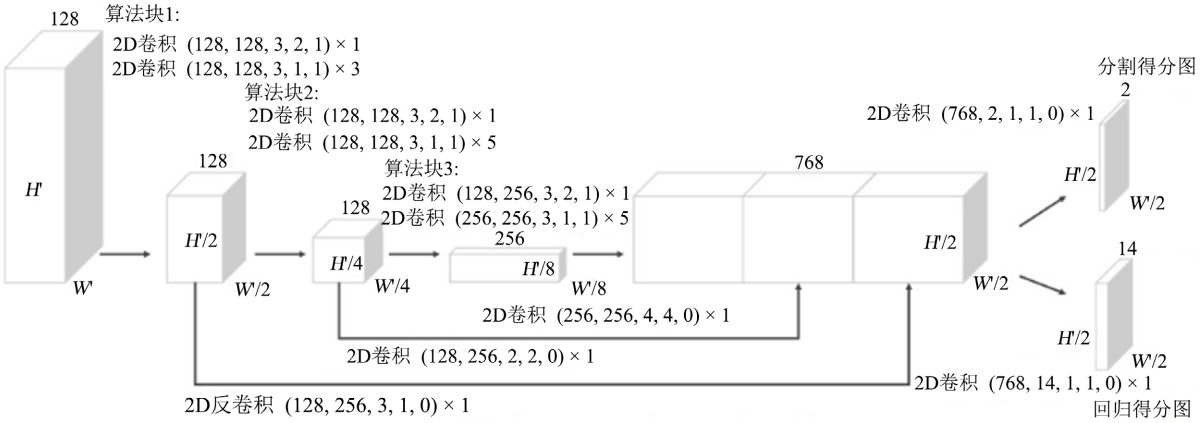
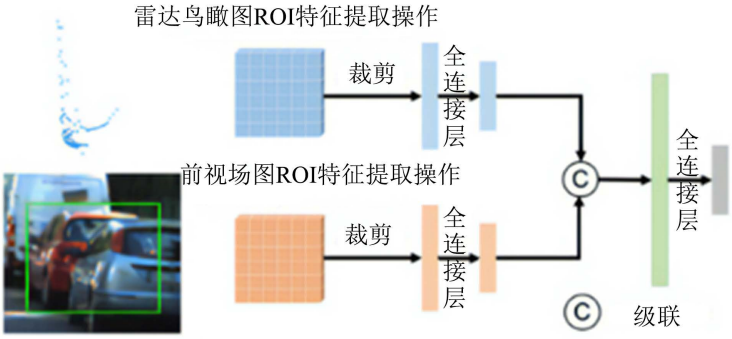


Figure 8. RPN algorithm diagram
图 8. RPN 算法结构图

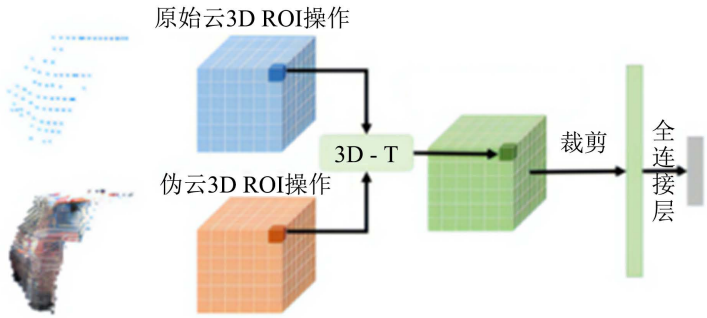
3) 3D-T

由于图像和点云之间的维度差距，先前的工作直接将从 BEV LiDAR 特征图中裁剪的 2D LiDAR ROI 特征和从前视场角图像特征图中剪裁的 2D 图像 ROI 特征连接起来，这是一种粗糙的 ROI 融合策略。在本文的方法中，通过将 2D 图像转换为 3D 伪云，这样可以以更精细的方式融合图像和点云的 ROI 特征，如图 9 所示。本文的 3D-T 由 3D 融合、网格融合和注意力融合组成。

1) 3D 融合：本文使用 3D ROI 来裁剪 3D 原始云和 3D 伪云，其仅包括 3D ROI 中的 LiDAR 特征和图像伪云特征，如图 9(b)所示。以前的方法使用 2D ROI 裁剪图像特征，这将涉及来自其他对象或背景的特征，如图 9(a)所示，这会导致大量干扰，尤其是对于被遮挡的对象。



(a) 先前方法



(b) 本文方法

Figure 9. Principle comparison diagram
图 9. 原理对比图

2) 网格融合：在以前的 ROI 融合方法中，图像 RoI 网格和 LiDAR ROI 网格之间没有对应关系，所以他们直接将图像 ROI 特征和 LiDAR ROI 特征连接起来。而在本文的方法中，由于原始 ROI 特征和伪 ROI 特征的表示方法相同，可以分别融合每一网格的一对特征。这就意味着能够用相应的伪网格特征准确地增强目标物体的每个体素的特征。

3) 注意力融合：为了自适应地融合来自原始 ROI 和伪 ROI 的每对网格特征，本文利用了文献[21]中激励的简单注意力模块。通常，通过预测每对网格的一对权重，并用权重对这对网格特征进行加权，以获得融合后的网格特征。

在此，接下来对本文的 3D-T 进行了详细的描述。令 b 表示一个单一的三维 ROI。接着用 $F^{raw} \in \mathbb{R}^{n\times C}$ 和 $F^{pse} \in \mathbb{R}^{n\times C}$ 分别表示 b 中的原始云 ROI 特征和伪云 ROI 特征。这里(默认为 $6\times6\times6$，遵循的基础模型是 Voxel-RCNN [22]的模型)是三维 ROI 中网格的总数，C 是网格特征通道。F^{raw} 和 F^{pse} 的第 i 个 ROI 网格特征分别表示为 F_i^{raw} 和 F_i^{pse}。随机选定一对 ROI 网格特征(F_i^{raw}，F_i^{pse})，接着再将 F_i^{raw} 和 F_i^{pse} 拼接起来。然后将结果传输到全连接层和激活函数层，产生一对网格特征的权重(w_i^{raw}，w_i^{pse})，其中 w_i^{raw} 和 w_i^{pse} 都是标量。最后，用(w_i^{raw}，w_i^{pse})对(F_i^{raw}，F_i^{pse})加权，得到融合后的网格特征 F_i。其中，F_i 的具体公式表达如下所示：

$$\left(w_i^{raw}, w_i^{pse}\right) = \sigma\left(MLP\left(CONCAT\left(F_i^{raw}, F_i^{pse}\right)\right)\right) \tag{14}$$

$$F_i = MLP\left(CONCAT\left(w_i^{raw}F_i^{raw}, w_i^{pse}F_i^{pse}\right)\right) \tag{15}$$

实际实践中，一个 bath 中的所有 ROI 网格特征对都可以并行处理，因此本文的 3D-T 非常有效。

3.3. 中晚期门控递归融合(Late Gated Recurrent State Fusion, LGRF)最小计算单元

为了解决部分传感器子集被遮挡时融合效果下降和传感器之间时间相关性丢失这两大问题，受LSTM 机制的启发，本文试提出了同时学习融合加权和时间加权的门控递归融合单元(Gated Recurrent Fusion Units, GRFU)，基于此修改设计出了本文的 fusion 中的基本单元——LGRF (Late Gated Recurrent Fusion)最小计算单元。能够做到：1) 延迟融合，并通过 M 个 LSTM 单元并行传递每个传感器数据，允许每个传感器编码器单独决定他们各自的历史与当前传感器输入的利用程度(本文称此为晚期递归融合(Late Recurrent Summation, LRS))，同时能够根据时间维度对齐传感器数据从而减少对齐偏差；2) 为每个传感器定义门，以确定每个传感器编码对融合单元和输出状态的贡献(称之为早期门控递归融合(Early Gated Recurrent Fusion, EGRF))。有了延迟融合，时间相关性就解决了。有了门定义，在部分传感器被遮挡或者失效时就可以实时调整每个传感器的贡献度。该单元整体融合时期偏中后期，鲁棒性能更好。

在下文中，本文首先分别定义了这两个修改模块，最后将两者结合起来定义本文的主模型 LGRF 模型。上文中预处理后的伪点云和点云特征分别接入该模型的C_{t-1}和h_{t-1}接口，再经过LGRF融合单元融合。

1) 晚期门控递归融合LRS

在这个模型中，总共使用了 2 个不同的 LSTM 单元(每个传感器一个)。对于每个模态，分别计算遗忘、输入、输出和单元状态。模型原理图和公式分别显示在图 10 和公式(16)~(18)。

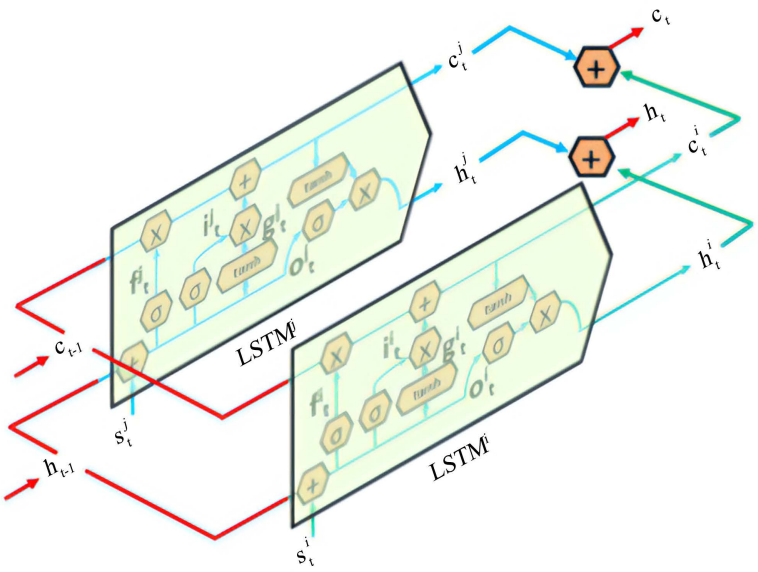


Figure 10. Late recurrent summation unit
图 10. 晚期递归传融合单元

$$
\begin{aligned}
f_t^i &= \sigma\left(W_f^i * s_t^i + U_f^i * h_{t-1} + b_f^i\right),\\
i_t^i &= \sigma\left(W_i^i * s_t^i + U_i^i * h_{t-1} + b_i^i\right),\\
O_t^i &= \sigma\left(W_o^i * s_t^i + U_o^i * h_{t-1} + b_o^i\right),\\
g_t^i &= \tanh\left(W_g^i * s_t^i + U_g^i * h_{t-1} + b_g^i\right)
\end{aligned} \tag{16}
$$

$$
\begin{aligned}
C_t^i &= C_{t-1} \odot f_t^i + i_t^i \odot g_t^i,\\
h_t^i &= O_t^i \odot \tanh\left(c_t^i\right)
\end{aligned} \tag{17}
$$

$$c_t = \sum_{i=1}^{M} c_t^i, h_t = \sum_{i=1}^{M} h_t^i \tag{18}$$

其中，每个门的输入空间转换权重W_*、U_*和偏置B_*都是与每种模态数据匹配不变的，但在不同是时间步长内是共享的。每个 LSTM 单元从过去时间步骤的状态(C_{t-1}^i，h_{t-1}^i)和当前时间步骤的输入S_t^i中接收信息。另外，各个传感器的每个 LSTM 单元没有单独的状态，而是所有的单元都接收来自前一个时间步长的相同的细胞状态(C_{t-1}，h_{t-1})。通过这种建模选择，可以在时间上传播融合后的表征。通过在所有传感器之间共享前一个时间步长的细胞状态(C_{t-1})，该模型可以单独决定是否保留每种模态的信息。最后，所有的隐藏状态(h_t^i)和细胞状态(C_t^i)被添加到一起，产生一个综合的表示h_t和C_t，并将其传输到下一个时间步长。

2) 早期门控递归融合EGRF

晚期融合为模型提供了一些灵活性，可以分别控制各个传感器的记忆信息，但即使在这里，最后的求和也会融合所有的传感器数据(通过假设它们具有相同的权重)。但是，如果能从数据中了解每个传感器对最终融合状态的贡献程度就更有利于融合效果。受 LSTM [23] [24]和 GRU [25]中使用的门控机制的启发，本文在传感器融合模块中也提出了一个类似的曝光控制。对于 M 个传感器，定义了 M − 1 个门(p^*)来控制传感器编码(S_t^i)在最终状态a_t中的曝光。与[25]类似，本文将最后一个传感器的门控定义为$1-\sum_i^{M-1} p^i$。这使得融合结果表示成为单个传感器编码的线性内插。模型示意图和方程式分别如图 11 和公式(19)~(23)所示。首先，使用非线性运算将传感器的嵌入数据转换为相同的维度，如公式(19)。然后，如公式(20)所示，计算 M − 1 门。如公式(21)所示，最后的融合是在每个门与相应的传感器编码相乘并相加后形成联合后进行的。如公式(22)~(23)所示，以 at 为输入进行时间建模。

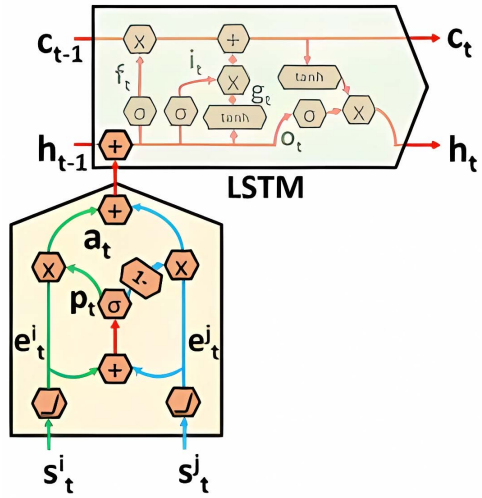


Figure 11. Early gated recurrent fusion unit
图 11. 早期门控循环融合单元

$$e_t^i = relu\left(W_e^i * s_t^i\right) \tag{19}$$

$$p_t^k = \sigma\left(\sum_{i=1}^{M} W_p^i * e_t^i\right), \ \forall k \in [1, M-1] \tag{20}$$

$$a_t = \left(\sum_{k=1}^{M-1} p_t^k \odot e_t^k\right) + \left(1 - \sum_{k=1}^{M-1} p_t^k\right) \odot e_t^k \tag{21}$$

$$\begin{aligned} f_t &= \sigma\left(W_f * a_t + U_f * h_{t-1} + b_f\right), \\ i_t &= \sigma\left(W_i * a_t + U_i * h_{t-1} + b_i\right), \\ o_t &= \sigma\left(W_o * a_t + U_o * h_{t-1} + b_o\right), \\ g_t &= \tanh\left(W_g * a_t + U_g * h_{t-1} + b_g\right) \end{aligned} \tag{22}$$

$$\begin{aligned} c_t &= c_{t-1} \odot f_t + i_t \odot g_t, \\ h_t &= o_t \odot \tanh\left(c_t\right) \end{aligned} \tag{23}$$

门控函数对于得出见解和解释模型内发生的融合的性质很有价值。一旦学会，用户可以通过门控值解释每个传感器的贡献度，并验证它们是否符合人类对数据集中一些任意样本的见解。这种可解释性特征对于涉及安全关键任务的场景至关重要。

3) LGRF

最后，本文使用的中晚期门控递归融合模型，它结合了晚期递归融合(独立控制每个传感器的记忆)和早期门控递归融合(学习如何融合)的最佳方面，以提高时空融合模型的学习性能。

$$e_t^i = relu\left(W_e^i * s_t^i\right) \tag{24}$$

$$p_t^k = \sigma\left(\sum_{i=1}^{M} W_p^i * e_t^i\right), \ \forall k \in [1, M-1] \tag{25}$$

$$a_t^i = \begin{cases} p_t^i \odot e_t^i & \text{if} \quad i \in [1, M-1], \\ \left(1 - \sum_{k=1}^{M-1} p_t^k\right) \odot e_t^i & \text{if} \quad i = M \end{cases} \tag{26}$$

$$\begin{aligned} f_t^i &= \sigma\left(W_f^i * a_t^i + U_f^i * h_{t-1} + b_f^i\right), \\ i_t^i &= \sigma\left(W_i^i * a_t^i + U_i^i * h_{t-1} + b_i^i\right), \\ O_t^i &= \sigma\left(W_o^i * a_t^i + U_o^i * h_{t-1} + b_o^i\right), \\ g_t^i &= \tanh\left(W_g^i * a_t^i + U_g^i * h_{t-1} + b_g^i\right) \end{aligned} \tag{27}$$

$$\begin{aligned} c_t^i &= c_{t-1} \odot f_t^i + i_t^i \odot g_t^i, \\ h_t^i &= o_t^i \odot \tanh\left(c_t^i\right) \end{aligned} \tag{28}$$

$$c_t = \sum_{i=1}^{M} c_t^i, h_t = \sum_{i=1}^{M} h_t^i \tag{29}$$

该模型示意图见图 12。与早期的门控递归融合模型类似，本文把融合门 p_t^* 作为所有传感器编码 e * t 的函数来计算，但并不是对所有传感器输入进行线性插值以得到联合输入状态 at，而是使用门控来控制每个编码传递到传感器特定 LSTM 单元的曝光量。最终的联合单元和隐藏状态是由所有最终单元状态和隐藏状态的输出各自相加计算出来的。

3.4. CP 卷积

每帧伪云定义：对于每一帧伪云 p，将图像中每个像素的 RGB (r,g,b) 和坐标 (u,v) 拼接到其对应的伪点。因此，第 i 个伪点 p_i 可以表示为 $(x_i, y_i, z_i, r_i, g_i, b_i, u_i, v_i)$。

提取伪云特征的一种简单方法是先直接对伪云进行体素化，然后对其执行三维稀疏卷积操作，而实际上它并没有充分发掘伪云中丰富的语义信息和结构信息。PointNet++是提取点的特征的一个很好的例子，但它并不适合伪云。首先，由于伪云的数量庞大，PointNet++中的球查询操作将带来大量的计算。其次，PointNet++不能提取二维特征，因为球查询操作没有考虑二维邻域关系。考虑在内。有鉴于此，需要一个特征提取器可以有效地提取 2D 语义特征和 3D 结构特征。

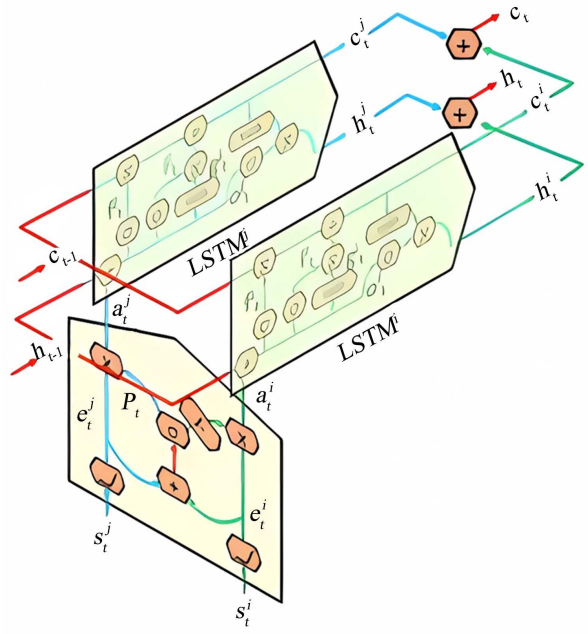


Figure 12. Late gated recurrent fusion unit
图 12. 中晚期门控递归融合单元

1) 图像领域中的ROI感知近邻搜索

基于上述分析，本文提出了 CP 卷积，它在图像域上展开近邻搜索，灵感来自体素查询[22]和网格搜索[26]。通过这种方式，可以克服 PointNet++的不足之处。首先，一个伪点可以在恒定的时间内搜索它的邻近点，这使得它比球查询操作快得多。其次，图像域上的邻域关系使得提取 2D 语义特征成为可能。

然而，不能将所有的伪点投射到当前帧的图像空间进行邻居搜索，因为在 gt 采样的情况下，来自其他帧的伪点可能会导致视场角闭塞遮挡。为此，本文提出了一种 ROI 感知近邻搜索。具体来说，就是根据伪点携带的(u,v)属性，将每个三维 ROI 中的伪点分别投射到其原始图像空间，如图 13 底部所示。这样一来，相互遮挡的伪点就不会成为彼此的邻近点，所以即使它们之间在视场角上存在严重的遮挡，它们的特征也不会相互干扰。

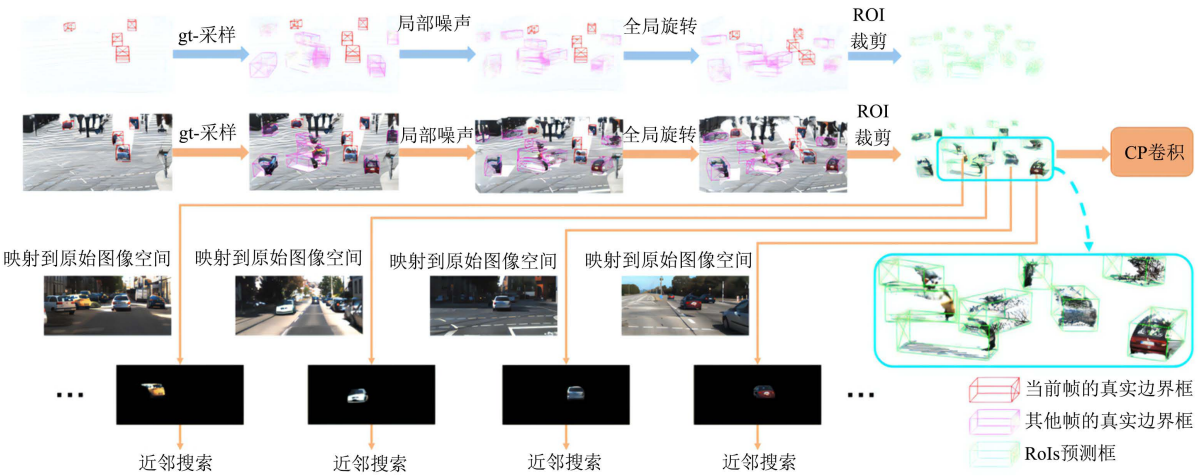


Figure 13. CP convolutional feature extraction schematic
图 13. CP 卷积特征提取示意图

2) 伪点云特征

对于第 1 个伪点 p_i，本文将 p_i 的特征表示为 $f_i=(x_i,y_i,z_i,r_i,g_i,b_i)$，它包括三维几何特征 (x_i,y_i,z_i) 和二维语义特征 (r_i,g_i,b_i)。正如文献[22]中的激励函数，在进行近邻搜索之前，特意在伪点特征上应用了一个全连接层，以降低复杂性。在经过全连接层之后，特征通道被提升至到 C_3，如图 14 所示。

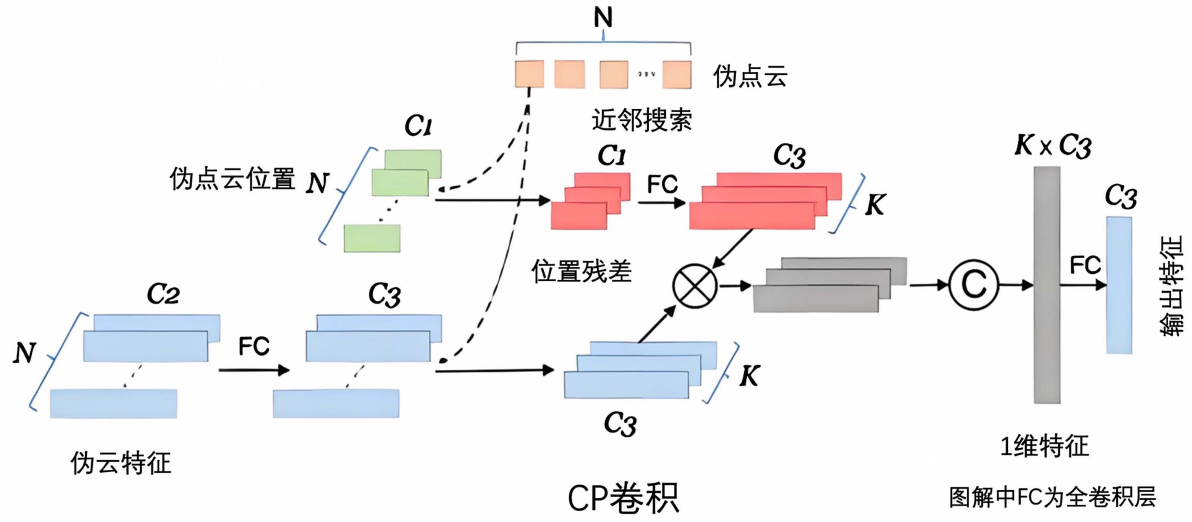


Figure 14. CP convolutional element illustration
图 14. CP 卷积元图解

3) 位置残差

本文利用 p_i 与其邻近点的三维和二维位置残差，使 p_i 的特征能够知道到三维和二维空间的局部关系，这对于提取 p_i 的三维结构特征和二维语义特征尤为重要。对于 p_i 的第 k 个近邻点 p_i^k， p_i 和 p_i^k 之间的位置残差表示见式(30)。

$$h_i^k=\left(x_i-x_i^k,y_i-y_i^k,z_i-z_i^k,u_i-u_i^k,v_i-v_i^k,\left\|p_i-p_i^k\right\|\right) \tag{30}$$

其中 $\left\|p_i-p_i^k\right\|=\sqrt{\left(x_i-x_i^k\right)^2+\left(y_i-y_i^k\right)^2+\left(z_i-z_i^k\right)^2}$。

4) 特征聚合

对于 p_i 的 K 个邻居，首先收集它们的位置并计算位置残差。然后，再在位置残差上应用一个全连接层，将其通道提升到 C_3 以与伪点特征对齐。给定一组相邻特征 $F_i=\left\{f_i^k\in\mathbb{R}^{C_3},k\in1,\cdots,k\right\}$ 和一组相邻位置残差 $H_i=\left\{h_i^k\in\mathbb{R}^{C_2},k\in1,\cdots,k\right\}$，接着用相应的 h_i^k 对每个 f_i^k 进行加权赋值。加权相邻特征被级联而不是最大池化，以获得最大的信息保真度。最后，应用一个全连接层将聚合后的特征通道映射回 C_3。

5) 多层特征融合

本文将三个 CP 卷积堆叠起来以提取伪云的更深层特征。考虑到高层级特征提供了更大的感受野和更丰富的语义信息，而低层级特征可以提供更精细的结构信息，所以将每个 CP 卷积的输出拼接起来，以获得伪云的更全面和更有鉴别力的表示。

4. 实验与分析

本次实验是基于 tensorflow 深度学习框架，通过 python 3.7 编程语言采用逻辑回归方式训练 MS3D-Net 模型，实验硬件设备是英伟达公司的 RTX 3080 显卡，显存 16 G，设置 batch 为 32，学习率为 0.001。

实验的具体流程如图 15 所示，首先基于开源库 KITTI 获取本文所需的车辆环境信息数据。接着制作数据集，总共收集了 14,936 个有效样本数据(每个样本包含图像数据和雷达点云数据)，包含 6 个不同类别(汽车、卡车、电车、行人、自行车、货车)的大约 8 万个标记物体，最后按照 7：2：1 的比例划分训练集、测试集和验证集。

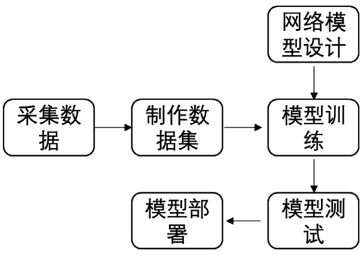


Figure 15. Experiment specific flow chart
图 15. 实验具体流程图

然后进行模型训练，将包含 14,936 个样本的数据集送入本文搭建的模型中进行多次迁移训练，并且与几种模型进行比较得到验证的 P-R 曲线如下图 16 所示。

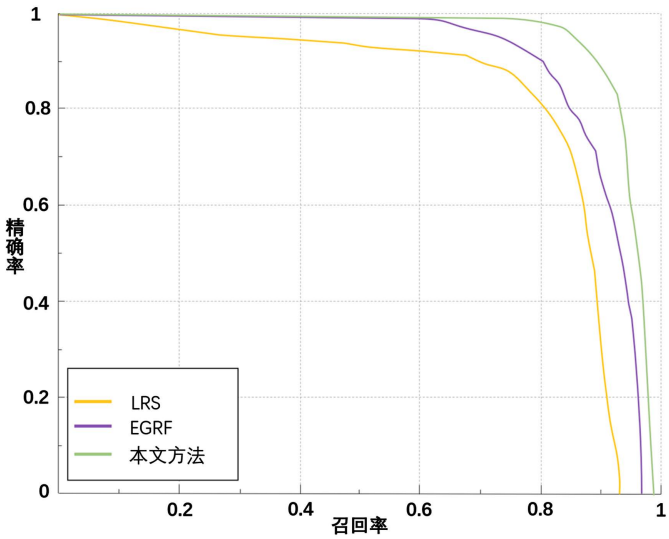


Figure 16. Model comparison PR graph
图 16. 模型对比 PR 曲线图

由上图可知，MS3D-Net 的方法对比其他几种方法无论是在精度(precision)还是召回率(recall)上都更高一些。根据表 2，可以看出，不管是各项的精度还是平均精度(mAP) MS3D-Net 都有着更出色的表现，并且检测速率与 LRS 相近地出色。

Table 2. Table of comparative experimental results by method
表 2. 各方法对比实验结果表

方法	模态组成	3D 检测				平均计算耗时
		mAP	简单	中等	困难	
MV3D	LiDAR + RGB	64.20	74.97	63.63	54.00	115.43 ms

Continued

AVOD	LiDAR + RGB	73.52	83.07	71.76	65.73	94.57 ms
LRS	LiDAR + RGB	76.41	84.37	74.82	70.03	69.97 ms
EGRF	LiDAR + RGB	78.68	88.40	77.43	70.22	73.68 ms
本文方法	LiDAR + RGB	80.79	89.20	80.05	73.11	71.98 ms

最后，基于 tensorflow 框架部署在实验室路端智能设备项目，并且使用了 TensorRT 高性能推理器进行模型加速。主要由服务端模型和客户端软件构成，构建合适的服务端模型来处理客户端的请求；利用 Triton-Client 与 OpenCV 编写客户端脚本，请求服务端模型推理服务，再利用 PyQt5 编写了 GUI，满足相应的功能需求。

安装好 cuda、tensorrt、opencv、docker、NVIDIA 容器、triton 这些所需环境，再将脚本文件打包成可执行文件即完成软件部署。实际部署后，检测结果如图 17 所示。

Figure 17. MS3D-Net deployment detection results graph
图 17. MS3D-Net 部署检测结果图

5. 结论

本文首先对比分析了不同类型融合层对融合性能指标的影响，并且结合车辆敏感环境特点为本文确定了融合框架的层级。然后专门为数据融合设计了新的高维表示，并且提出了与之对应的融合方法 3D-T，有效降低了跨模态数据融合过程中的信息损失。受 LSTM 机制启发，拓展设计了中晚期门控循环状态融合，有效降低了多传感器之间的对齐偏差，减少了部分传感器数据丢失或错误的影响，提升低时间相关性时的融合表现。为了提高图像特征的提取，从图像支流获得更多维度的信息，提出了 CP 卷积方法进行提取。最后通过使用 KITTI 数据集进行训练与验证，分析结果发现本文方法在提高检测精度(融合性能)的同时又保证了检测速度。该方法有效提高复杂场景中的融合效率，为进一步的检测与跟踪提供更准确与鲁棒的信息。

参考文献

[1] Chai, Y., Sun, P., Ngiam, J., *et al*. (2021) To the Point: Efficient 3D Object Detection in the Range Image with Graph Convolution Kernels. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 15995-16004. https://doi.org/10.1109/CVPR46437.2021.01574

[2] Chen, Y., Liu, S., Shen, X. and Jia, J.Y. (2019) Fast Point R-CNN. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October -2 November 2019, 9774-9783. https://doi.org/10.1109/ICCV.2019.00987

[3] Ge, R., Ding, Z., Hu, Y., *et al*. (2017) Real-Time Anchor-Free Single-Stage 3D Detection with IoU-Awareness. arXiv: 2107.14342.

[4] Liu, Z., Zhao, X., Huang, T., *et al*. (2020) Tanet: Robust 3D Object Detection from Point Clouds with Triple Attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 11677-11684. https://doi.org/10.1609/aaai.v34i07.6837

[5] Mao, J., Niu, M., Bai, H., *et al*. (2021) Pyramid R-CNN: Towards Better Performance and Adaptability for 3D Object Detection. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 2703-2712. https://doi.org/10.1109/ICCV48922.2021.00272

[6] Chen, X., Ma, H., Wan, J., Li, B. and Xia, T. (2017) Multi-View 3D Object Detection Network for Autonomous Driving. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6526-6534. https://doi.org/10.1109/CVPR.2017.691

[7] Ku, J., Mozifian, M., Lee, J., *et al*. (2018) Joint 3D Proposal Generation and Object Detection from View Aggregation. 2018 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), Madrid, 1-5 October 2018, 1-8. https://doi.org/10.1109/IROS.2018.8594049

[8] Liang, M., Yang, B., Chen, Y., Hu, R. and Urtasun, R. (2019) Multi-Task Multi-Sensor Fusion for 3D Object Detection. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 7337-7345. https://doi.org/10.1109/CVPR.2019.00752

[9] 钱晓明,黄宇轩,楼佩煌,孙天.基于多传感器融合的跟随 AGV 复合导引技术[J].农业机械学报,2022,53(01):14-22+32.

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

[11] Guan, D.Y., Cao, Y.P., Yang, J.X., Cao, Y.L. and Yang, M.Y. (2019) Fusion of Multispectral Data through Illumination-Aware Deep Neural Networks for Pedestrian Detection. *Information Fusion*, **50**, 148-157. https://doi.org/10.1016/j.inffus.2018.11.017

[12] Matti, D., Ekenel, H.K. and Thiran, J.P. (2017) Combining LiDAR Space Clustering and Convolutional Neural Networks for Pedestrian Detection. 2017 14*th IEEE International Conference on Advanced Video and Signal Based Surveillance* (*AVSS*), Lecce, 29 August -1 September 2017, 1-6. https://doi.org/10.1109/AVSS.2017.8078512

[13] Hu, M., Wang, S., Li, B., *et al*. (2021) Penet: Towards Precise and Efficient Image Guided Depth Completion. 2021 *IEEE International Conference on Robotics and Automation*, Xi'an, 30 May -5 June 2021, 13656-13662. https://doi.org/10.1109/ICRA48506.2021.9561035

[14] Imran, S., Liu, X. and Morris, D. (2021) Depth Completion with Twin Surface Extrapolation at Occlusion Boundaries. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 2583-2592. https://doi.org/10.1109/CVPR46437.2021.00261

[15] Wang, Y., Chao, W.L., Garg, D., *et al*. (2019) Pseudo-Lidar from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 8437-8445. https://doi.org/10.1109/CVPR.2019.00864

[16] You, Y., Wang,Y., Chao, W.L., *et al*. (2019) Pseudo-Lidar + +: Accurate Depth for 3D Object Detection in Autonomous Driving. arXiv: 1906.06310.

[17] Ren, S.Q., He, K.M. and Girshick, R.B. and Sun, J. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv: 1506.01497.

[18] Dai, J., Li, Y., He, K., *et al*. (2016) R-FCN: Object Detection via Region-Based Fully Convolutional Networks. arXiv: 1605.06409.

[19] Liu, W., Anguelov, D., Erhan, D., *et al*. (2016) SSD: Single Shot Multibox Detector. In: Leibe, B., Matas, J., Sebe, N. and Welling, M., Eds., *Computer Vision—ECCV* 2016, Springer, Cham, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[20] Huang, J., Rathod, V., Sun, C., *et al*. (2017) Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors.

2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 3296-3297. https://doi.org/10.1109/CVPR.2017.351

[21] Huang, T., Liu, Z., Chen, X., *et al*. (2020) Epnet: Enhancing Point Features with Image Semantics for 3D Object Detection. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Computer Vision—ECCV* 2020, Springer, Cham, 35-52. https://doi.org/10.1007/978-3-030-58555-6_3

[22] Deng, J., Shi, S., Li, P., *et al*. (2021) Voxel R-CNN: Towards High Performance Voxel-Based 3D Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 1201-1209. https://doi.org/10.1609/aaai.v35i2.16207

[23] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[24] Gers, F.A., Schmidhuber, J. and Cummins, F. (2000) Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, **12**, 2451-2471. https://doi.org/10.1162/089976600300015015

[25] Cho, K., Van Merriënboer, B., Gulcehre, C., *et al*. (2014) Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, 25-29 October 2014, 1724-1734. https://doi.org/10.3115/v1/D14-1179

[26] Fan, L., Xiong, X., Wang, F., Wang, N.Y. and Zhang, Z.X. (2021) RangeDet: In Defense of Range View for Lidar-Based 3D Object Detection. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 2898-2907. https://doi.org/10.1109/ICCV48922.2021.00291