第 59 卷 第 8 期/2022 年 4 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于多模态信息融合的多目标检测算法

刘通，高思洁*，聂为之

天津大学电气自动化与信息工程学院，天津 300072

摘要 随着激光雷达等信息采集设备的发展，目标的智能检测变得越来越重要。为了实现对行人、车辆等目标智能化的检测和识别，提高无人驾驶、城市管理等多方面应用的智能化水平，迫切需要有效的、智能化的目标检测技术。针对激光探测与测距技术(LiDAR)进行三维目标检测时的信息缺失等问题，提出一种基于多模态信息融合的多目标检测算法。网络模型由三个模块组成：LiDAR点云数据处理模块和二维图像数据处理模块分别对点云和RGB图像进行特征提取；信息融合及检测模块根据对应的位置对三维与二维的特征图进行融合，弥补单模态数据的信息缺失，实现特征层面的互补。将融合后的特征图通过三维与二维的区域生成网络分别生成目标检测框，采取后融合的策略，将两种模态的检测框融合得到最终的目标检测结果。通过在点云数据集KITTI和图片数据集VOC2007上的对比实验及相关分析，证明了所提算法的优越性。

关键词 机器视觉；特征融合；目标检测；点云

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202259.0815002

Multitarget Detection Algorithm Based on Multimodal Information Fusion

Liu Tong, Gao Sijie*, Nie Weizhi

School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, *China*

Abstract Developing information acquisition equipment, such as lidar, has made intelligent target detection increasingly important. Recently, there has been an increasing need for an effective and intelligent target detection technology to realize the intelligent detection and recognition of pedestrians, vehicles, and other targets, and improve the intelligence level of unmanned driving, urban management, and other applications. Thus, to solve the lack of information in the use of light detection and ranging (LiDAR) for 3D target detection, this paper proposes a multimodal information fusion-based multitarget detection algorithm. The network model comprises three modules: LiDAR point cloud data processing module, 2D image data processing module, and information fusion and detection module. The first two extracted the point cloud and RGB image features, respectively, whereas the information fusion and detection module merged the three- and two-dimensional feature maps according to the corresponding positions to mitigate the lack of information in the monomodal data and achieve the complementarity of the feature level. The fused feature map generated the target detection frame using the three- and two-dimensional area generation networks and adopted the post-fusion strategy to fuse the detection frames of both modes to obtain the final target detection result. KITTI and VOC2007 datasets were used for evaluation and analysis. Experimental results demonstrated the superiority of the proposed algorithm.

Key words machine vision; feature fusion; object detection; point cloud

收稿日期：2021-03-08；**修回日期**：2021-04-04；**录用日期**：2021-04-21

通信作者：*gaosijie_0112@tju.edu.cn

1 引 言

近年来，随着激光探测与测距（LiDAR）技术的发展，点云数据的获取速度与精确度大大提升。如何实现高效准确的点云目标检测，是智能驾驶[1-2]、遥感、增强现实、虚拟现实[3]等领域的重要问题。与传统的二维目标检测相比，三维目标检测需要更多的输出参数来确定目标的边界框。而由于LiDAR点云的数据特性，在目标检测任务中，常常会面临输入数据分辨率低、纹理和颜色信息缺失、计算开销大等问题，因而更具挑战性。

面对这些问题，多模态信息融合的方法[4]成为了该领域的研究重点。目前，多模态融合方法主要分为三种：早期融合、后期融合、深度融合。早期融合方法在对原始传感器数据进行特征提取之前做特征融合，代表算法为PI-RCNN[5]，该算法直接在三维点上进行逐点连续卷积，并应用点池化和注意集中操作以获得更好的融合性能。后期融合是最为简便的融合方法，仅在决策层进行融合，避免了不同传感器数据差异带来的问题，降低了算法的复杂性。但早期融合和后期融合存在无法充分利用多模态数据间关联性的缺点。深度融合方法在特征层面进行交互，代表方法为MV3D网络[6]，该网络由两个子网络组成，一个用于生成三维目标候选区域，另一个用于多视图特征融合。深度融合对跨模态信息的利用最为充分，但现行方法往往存在对数据对齐敏感、网络结构复杂的缺点。

针对上述问题，本文提出了一种基于多模态融合的多目标检测算法。网络模型由三个模块组成：LiDAR点云数据处理模块和二维图像数据处理模块分别对点云和二维彩色（RGB）图像进行特征提取；信息融合及检测模块根据对应的位置对三维与二维的特征图进行融合，弥补单模态数据的信息缺失，实现特征层面的互补。将融合后的特征图通过三维与二维的区域生成网络分别生成目标检测框，采取后融合的策略，将两种模态的检测框融合得到最终的目标检测结果。在KITTI点云数据集与VOC2007图像数据集上评估了算法性能，并对所提算法与其他单模态与多模态目标检测算法进行了比较分析，相关实验结果证明了所提融合策略的有效性和算法的优越性。

2 相关工作

现行的三维目标检测算法分为三个主要类别：基于图像、基于点云、基于图像和点云的多模态融合。基于二维图像的方法与使用点云的方法在性能上具有很大差距，因此重点关注后两类方法。

2.1 基于二维图像的三维检测

Mousavian等[7]利用二维和三维边界框之间的几何约束来恢复三维信息。Chabot等[8]和Mottaghi等[9]通过计算三维目标和计算机辅助设计（CAD）模型之间的相似度来估计三维目标信息。Wang等[10]和You等[11]探索使用立体图像生成密集点云并使用其进行目标检测。Li等[12]提出的GS3D算法先计算二维检测结果，通过先验知识结合学习算法计算三维检测边界框的尺寸和方位，利用三维表面在二维图像的投影特征进行判别与检测。Li等[13]提出的Stereo R-CNN将Faster R-CNN框架拓展到双目立体视觉，该算法对左右视图进行自动对齐学习，并通过稠密匹配优化三维目标检测结果。相对于三维点云，二维图像缺失了部分空间信息，因此与基于LiDAR的技术相比，这些方法生成的三维边界框精度要低得多。

2.2 基于点云的三维检测

点云技术在三维目标检测中应用广泛，单一传感器避免了多传感器校准和同步问题。在基于点云的三维检测方法中，从原始点云中学习特征的方法及编码方法各不相同。VoxelNet[14]使用体素对原始点云进行编码，并使用三维卷积神经网络来学习体素特征，以进行分类和边界框回归。SECOND算法[15]在VoxelNet的基础上进行了改进，由于原始LiDAR点云的数据结构非常稀疏，因此使用稀疏的三维卷积神经网络，大大减少了推理时间。PointPillars[16]在编码器中使用PointNets[17]，该编码器代表以垂直列组织的点云，后连接二维卷积神经网络（CNN）来执行三维目标检测。与上面讨论的单阶段方法相比，PointRCNN[18]、Fast PointRCNN[19]、STD[20]应用了两阶段架构，该架构首先以自下而上的方式生成三维建议，然后在第二阶段中对这些建议进行完善。PV-RCNN[21]充分利用了三维体素CNN和基于PointNet的集合抽象的优势，以学习更多区分性功能。此外，PartA2-Net[22]在第一阶段预测三维边界框和目标的局部信息，在第二阶段通过局部信息对边界框进行精确定

位。然而，LiDAR点云稀疏的特性导致基于点云的检测方法对远距离物体的检测性能较差，缺乏二维图像的信息多样性，并且由于三维数据的复杂性，计算量往往很高。

2.3　基于多模态融合的三维检测

摄像头-LiDAR融合方法在自动驾驶中被广泛使用。Frustum PointNet[23]、Pointfusion[24]和Frustum ConvNet[25]是二维驱动三维检测器的代表，它们利用成熟的二维检测器生成二维建议并将三维处理域缩小到图像中相应的裁剪区域，但是在只能从三维空间观察的情况下，这种方法可能会失效。MV3D[26]和AVOD[27]将原始点云投影到鸟瞰视图(BEV)中，以形成多通道BEV图像。基于深度融合的二维CNN用于从此BEV图像和前置摄像头图像中提取特征，以进行三维边界框回归。这些方法的整体性能比基于LiDAR的方法差，可能的原因为原始点云转换为BEV图像后会丢失空间信息。为了融合不同模态特征，这些方法往往会对输入数据进行裁剪和调整大小操作，这些操作可能会破坏每个传感器的特征结构。例如池化(pooling)中的下采样操作常常丢弃了位置信息，导致数据各部分空间相对关系被破坏。相机图像是高分辨率的密集数据，而LiDAR点云是低分辨率的稀疏数据，将这两种不同类型的数据结构融合会面临很多困难。强制二维图像和三维LiDAR点云的特征向量具有相同的大小或相等的长度，然后对其进行级联、聚合或平均，可能会导致这些特征向量之间的对应关系不准确。为了在融合中获得更好的对应性，MMF[28]采用连续卷积[29]来构建密集的LiDAR BEV特征图，并与密集图像特征进行逐点特征融合，但存在算法框架复杂的缺点。

3　基于多模态信息融合的多目标检测算法

所提算法的网络结构如图1所示。整个模型分为三部分：1) LiDAR点云数据处理模块，在前视图平面上将LiDAR点云数据均匀划分为若干组块，逐点输入到堆叠体素特征编码(VFE)层得到逐点特征，通过全连接层与逐元素最大池获得逐体素特征，再通过卷积中间层输出三维特征图；2) 二维图像数据处理模块，在二维RGB图像上进行与点云处理模块中相似的分组处理并输入到特征提取网络(该网络由卷积层、池化层和线性整流层组合而成)，输出二维特征图；3) 信息融合及检测模块，对对应位置的点云特征与图像特征进行拼接，将融合后的特征输入区域生成网络，得到目标检测框和预测得分，完成目标检测任务。

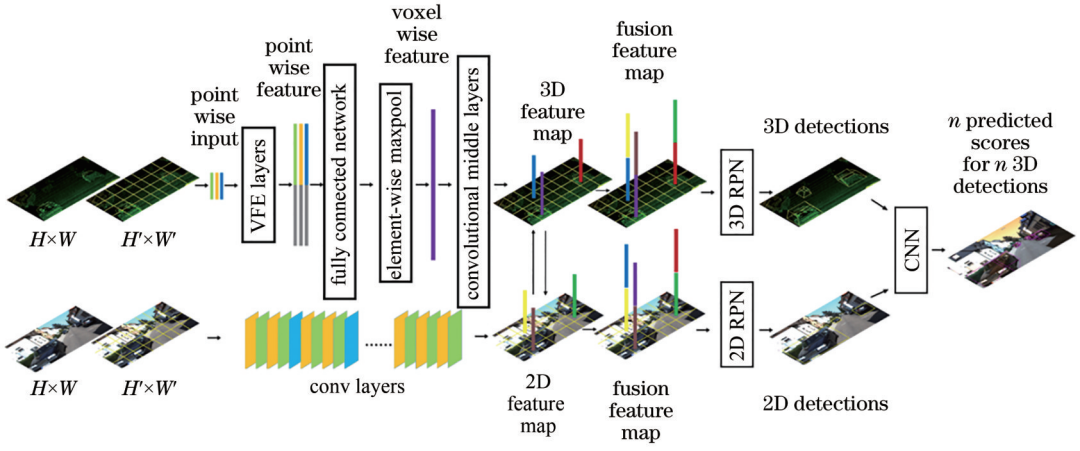

图1　所提算法的网络结构示意图

Fig. 1　Network structure diagram of the proposed algorithm

3.1　LiDAR点云数据处理模块

根据激光雷达与相机之间的位置变换矩阵，将点云转换到相机坐标系中。设点云包含分别沿Z、Y、X轴长度为D、H、W的三维空间，定义高和宽分别为v_H，v_W的体素，则高、宽维度上包含体素的数量为

$$H'=H/v_H,\ W'=W/v_W。\tag{1}$$

简单起见，设H，W是v_H，v_W的倍数。根据所在的体素对点云中的点进行分组。由于LiDAR点云稀疏的特性，每个体素包含的点数目差别较大。为了减少计算量，减小采样偏差，从包含多于T个点的体素中随机采样固定数量的T个点。

将点云输入堆叠体素特征编码层。用相对质心的偏移量扩充表示每个点$\boldsymbol{p}_i$，获得输入特征集：

$$\boldsymbol{V}_{\text{in}}=\left\{\hat{\boldsymbol{p}}_t=\left[x_t,y_t,z_t,r_t,x_t-v_x,y_t-v_y,z_t-v_z\right]^{\mathrm{T}}\in\mathbb{R}^7\right\}_{t=1,2,\ldots,T},\tag{2}$$

式中：r_t是接收的反射率；(v_x,v_y,v_z)为点所在体素的质心坐标。每个$\hat{\boldsymbol{p}}_t$通过全连接网络转换到特征空间，获得点特征$\boldsymbol{f}_t\in\mathbb{R}^{16}$，并通过逐元素的最大池化操作（MaxPooling）获得局部聚合特征$\tilde{\boldsymbol{f}}_t\in\mathbb{R}^{16}$。用$\tilde{\boldsymbol{f}}_t$扩充每个$\boldsymbol{f}_t$，形成点级联特征：

$$\boldsymbol{f}_t^{\text{out}}=\left[\boldsymbol{f}_t^{\mathrm{T}},\tilde{\boldsymbol{f}}_t^{\mathrm{T}}\right]^{\mathrm{T}}\in\mathbb{R}^{32}。\tag{3}$$

输出特征集表示为

$$\boldsymbol{V}_{\text{out}}=\left\{\boldsymbol{f}_t^{\text{out}}\right\}_{t=1,2,\ldots,T}。\tag{4}$$

将第v个VFE层记为$l_{\text{VFE}v}(c_{\text{in}},c_{\text{out}})$，该层将维度为$c_{\text{in}}$的输入特征转换为维度为$c_{\text{out}}$的输出特征。线性层学习维度为$c_{\text{in}}\times(c_{\text{out}}/2)$的矩阵，并且逐点级联产生维度为$c_{\text{out}}$的输出。通过全连接网络将VFE层的输出转换为$R\times C$尺寸并应用MaxPooling操作获得维度为$C$的体素特征。仅对非空体素进行处理，得到一个体素特征列表，每个特征都与特定的非空体素的空间坐标唯一关联。为了减小反向传播期间的内存占用和计算成本，将列表表示为稀疏四维张量，大小为$C\times D'\times H'\times W'$。

在卷积中间层中，使用conv3D$(c_{\text{in}},c_{\text{out}},\boldsymbol{k},\boldsymbol{s},\boldsymbol{p})$表示一个三维卷积算子，其中$c_{\text{in}}$和$c_{\text{out}}$是输入通道和输出通道的数量，$\boldsymbol{k}$，$\boldsymbol{s}$和$\boldsymbol{p}$是分别对应于卷积核大小，步幅大小和填充大小的三维向量。每个卷积中间层依次应用三维卷积、批量归一化（BN）层和线性整流（ReLU）层。卷积中间层在逐渐扩大的感受野中对体素特征进行聚合，从而为形状描述添加更多上下文。

三维特征提取网络的损失函数定义为

$$L=\alpha\frac{1}{N_{\text{pos}}}\sum_i L_{\text{cls}}(p_i^{\text{pos}},1)+\beta\frac{1}{N_{\text{neg}}}\sum_j L_{\text{cls}}(p_j^{\text{neg}},0)+\frac{1}{N_{\text{pos}}}\sum_i L_{\text{reg}}(u_i,u_i^*),\tag{5}$$

式中：p_i^{pos}和p_j^{neg}分别表示正锚点a_i^{pos}和负锚点a_j^{neg}的Softmax输出；$u_i\in\mathbb{R}^7$和$u_i^*\in\mathbb{R}^7$分别是正锚点a_i^{pos}的回归输出和真值；前两项是$\left\{a_i^{\text{pos}}\right\}_{i=1,2,\ldots,N_{\text{pos}}}$和$\left(a_j^{\text{neg}}\right)_{j=1,\ldots,N_{\text{neg}}}$的归一化分类损失；$L_{\text{cls}}$代表二进制交叉熵损失；$\alpha$，$\beta$是平衡相对重要性的正常数；最后一项$L_{\text{reg}}$是回归损失，使用平滑L1（SmoothL1）函数[30]度量。

3.2 二维图像数据处理模块

二维特征提取网络的结构如图1所示。输入图像的宽和高分别为W和H。为了更好地关注局部特征，同时便于与三维数据的对齐和后期融合，对图像进行分组操作，定义每个大小为n_H，n_W的组块，其中n_H，n_W的大小与三维特征提取网络中的v_H，v_W相同。宽、高维度上组块的数目为

$$W'=W/n_W,H'=H/n_H,\tag{6}$$

设H，W是n_H，n_W的整数倍。

二维特征提取网络包括13个卷积（conv）层、4个池化（pooling）层以及13个线性整流（ReLU）层。使用线性整流层的目的在于在网络中引入非线性因素，与其他类型的激活函数相比，线性整流函数能够造成网络的稀疏性，避免过拟合的发生，且计算成本较低。网络生成与原图像对应的特征图（feature map），作为区域候选网络（RPN）的输入特征图。

3.3 信息融合及检测模块

由于对点云与RGB图像均进行了相似的分组操作，各组块的特征具有对应性。在三维特征图与二维特征图上，按所属组块对两个模态的特征进行拼接。RGB图像数据密集且规则，包含了纹理、颜色等信息，但缺失深度信息；点云数据通常稀疏且不规则，但包含了三维几何结构与深度信息。拼接操作在特征层面上实现信息的互补，提高了特征的鲁棒性，避免单一模态数据局限性带来的漏检、误检等问题。

将融合后的三维特征图输入三维区域生成网络。该网络具有三个完全卷积层的块。每个块的第一层通过步幅为2的卷积对特征图进行一半的下采样，其后是步幅为1的卷积序列。在每个卷积层后应用BN和ReLU操作。将每个块的输出上采样为高、宽分别为$H'/2$、$W'/2$的固定尺寸，将每个块的上采样结果拼接为高分辨率特征图。以特征图的每一个点为中心，设置9种不同尺寸的锚框作为初始的检测框。通过两个分支分别对特征图进行映射。第一个分支对每个锚框进行分类，输出概率分数图；第二个分支用于计算锚框相对于目标真值框的回归偏移量，输出平移缩放参数。将前景锚框和边界框回归偏移量共同输入候选层（proposal）中。同样地，将融合后的二维特征图输入二维区域生成网络，输出概率分数图与锚框回归图，最终得到点

云数据与RGB图像中候选区域所属的目标类别及位置。

如果在三维与二维区域生成网络输出的结果中均正确地检测出同一目标，则三维目标边界框在二维图像上的投影与二维目标边界框应具有较高的几何一致性，可以将其作为不同模态检测结果的联系。二维目标检测结果可以表示为

$$\begin{cases} \boldsymbol{P}^{2D}=\left\{\boldsymbol{p}_1^{2D},\boldsymbol{p}_2^{2D},\ldots,\boldsymbol{p}_M^{2D}\right\} \\ \boldsymbol{P}_m^{2D}=\left\{\left[x_{m1},y_{m1},x_{m2},y_{m2}\right],s_m^{2D}\right\} \end{cases}, \tag{7}$$

式中：$\boldsymbol{P}^{2D}$为检测结果的集合；$\boldsymbol{P}_m^{2D}$中的第一项为二维目标检测边界框，第二项为置信度得分。

类似地，三维目标检测结果可以表示为

$$\begin{cases} \boldsymbol{P}^{3D}=\left\{\boldsymbol{p}_1^{3D},\boldsymbol{p}_2^{3D},\ldots,\boldsymbol{p}_N^{3D}\right\} \\ \boldsymbol{P}_n^{3D}=\left\{\left[h_n,w_n,l_n,x_n,y_n,\theta_n\right],s_n^{3D}\right\} \end{cases}, \tag{8}$$

式中：θ_n代表边界框相对Z轴的旋转角度。

采用混合的方式表示两种模态的检测结果：

$$\boldsymbol{T}_{m,n}=\left\{R_{\mathrm{IoU}m,n},s_m^{2D},s_n^{3D},d_n\right\}, \tag{9}$$

式中：第一项表示在图像中的第m个检测结果和点云中的第n个结果的几何一致性，用边界框交并比$R_{\mathrm{IoU}m,n}$表示，如果二维与三维检测网络均正确地检测出同一目标，则三维检测框在二维图像上的投影应与二维检测框具有较大的交并比；第二项内容是二维检测的第m个检测到的物体的置信度分数；第三项内容为在点云场景下的置信度分数；最后一项表示在点云场景下检测到的第n个物体到地面的归一化距离。如果$R_{\mathrm{IoU}m,n}$为0，则将$\boldsymbol{T}_{m,n}$置为空。将非空的$\boldsymbol{T}$输入卷积神经网络，通过最大池化映射到概率得分图。网络的最终输出为三维目标检测框及其对应的概率得分。

3.4 高效实现

参照VoxelNet中的实施细节，将稀疏不均匀的点云转换为密集的张量结构。首先初始化一个$K\times T\times 7$维张量结构以存储体素，并将其输入特征缓冲区，其中K表示非空体素的最大数量，T表示每个体素的最大点数量，每个点的输入编码尺寸为7。这些点在处理之前就被随机化了。对于点云中的每个点，检查对应的体素是否已经存在。使用体素坐标建立哈希表(Hashtable)，可以高效实现体素的查找与初始化。体素输入特征和坐标缓冲区可以通过对点列表进行一次遍历来构造，因此复杂度为$O(n)$。为了进一步提高存储与计算效率，可以仅存储有限数量的体素，而忽略包含点数很少的体素。

4 分析与讨论

4.1 数据集

采用KITTI数据集对算法性能进行评测。KITTI数据集是目前国际上最大的自动驾驶场景下的算法评测数据集，包含7481个用于训练的点云与图像和7518个用于测试的点云与图像，包括汽车、行人和骑自行车的人3种类别。对于每个类别，根据简单、中等、困难三个难度级别评估检测结果，三个难度级别分别根据目标大小、遮挡状态和截断级别确定。对算法进行全面评估，并将训练数据细分为训练集和验证集，得到3712个用于训练的数据样本和3769个用于验证的数据样本。经过分割之后，相同序列的样本不会同时包含在训练和验证集中。

同时，为了验证所提算法中分组操作对二维目标检测性能的影响，在VOC2007数据集上对所提算法中的二维图像目标检测方法进行评估。该数据集包含20个类别，共计9963张图像，其中5011张用于训练，4952张用于测试。

4.2 实施细则

在KITTI验证集上对所提算法及各类对照算法进行评估实验。按照官方评估协议，目标检测需要同时实现目标定位和目标识别两项任务。其中，通过比较预测边框和真值框的交并比(IoU)和阈值的大小判定目标定位的正确性；通过置信度分数和阈值的比较确定目标识别的正确性。以上两步综合判定目标检测是否正确，最终将多类别目标的检测问题转换为“某类物体检测正确、检测错误”的二分类问题，从而可以构造混淆矩阵，使用目标分类的一系列指标评估模型精度。实验中设置汽车类的IoU阈值为0.7，行人和骑自行车者类的IoU阈值为0.5。使用平均精确度(AP)指标，即不同召回率下精确率的均值，对各算法进行比较。对于所提算法，使用KITTI提供的LiDAR数据和RGB图像数据从头开始训练，网络权重参数随机初始化。

4.3 点云随机采样阈值分析实验

为了分析点云随机采样阈值对算法性能的影响，在KITTI验证集汽车类别的三种难度级别上对应用不同采样阈值的所提算法进行对照实验，实验使用平均精度指标对算法精确度进行度量，并记录算法每次检测耗费的平均时间，实验结果如表1所示。随着随机采样阈值的增大，算法的时间开销和

表 1 不同点云采样阈值下所提算法在KITTI验证集上的性能对比

Table 1 Performance comparison of proposed algorithm with different pointcloud sampling thresholds on the KITTI validation set

Pointcloud sampling threshold	Consumed time /ms	AP /%		
		Car Easy	Car Moderate	Car Hard
10	124	60.64	55.13	49.70
20	171	73.25	60.48	56.28
30	225	82.95	67.48	64.22
40	277	81.34	68.61	65.55
50	329	83.20	69.12	66.37

准确率均有提高，分析数据可以发现，在采样阈值大于30之后，算法的平均精度上升幅度很小，而时间开销仍然线性增大。综合考虑检测速度与精度，在后续的实验中将点云的随机采样阈值确定为30。

4.4 分组方法分析实验

为了分析分组方法对算法检测效果的影响，设置了三种不同组块划分方式：$W'=W/n_W=200$，$H'=H/n_H=150$；$W'=W/n_W=400$，$H'=H/n_H=300$；$W'=W/n_W=800$，$H'=H/n_H=600$。

在KITTI数据集三种类别的所有难度级别上对采用三种分组方式的所提算法进行比较实验，实验结果如表2所示。

表 2 不同分组方式下所提算法在KITTI验证集上的平均目标检测精度对比

Table 2 Comparison of average target detection accuracy of the proposed algorithm on KITTI validation set under different grouping methods unit: %

Grouping method		Car			Pedestrian			Cyclist		
W'	H'	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
200	150	78.42	63.14	61.05	55.43	51.56	47.11	63.79	45.23	44.36
400	300	82.95	67.48	64.22	58.90	55.33	50.16	68.42	48.52	46.08
800	600	78.85	64.22	61.50	55.89	51.90	47.68	64.25	46.03	44.88

分析数据可以看出，第二种划分方式取得了最好的效果。实验结果表明，过于稀疏的划分方式会忽略输入数据的局部信息，而过于稠密的划分方式则会由于过于关注局部而忽视特征间的联系。在后续的实验中，所提算法将采用第二种划分方式。

4.5 消融实验

为了分析融合方法在所提算法中的重要性，设置了两个对照算法：第一个对照算法（记为deep fusion）仅对具有两种模态的数据处理模块输出的三维特征图与二维特征图进行特征融合处理，而不对检测框进行后融合，直接将三维区域生成网络输出的检测框与目标类别作为最终的检测结果；第二个对照算法（记为late fusion）对具有两种模态的数据独立进行目标检测，仅在决策阶段对三维与二维的检测结果进行后融合。对照算法与所提算法采用相同的分组方式、网络结构以及参数，在KITTI数据集上进行对比实验。使用AP评估检测性能，实验结果如表3所示，所提算法的检测性能比仅使用特征融合或后融合的对照算法更加优越，证明了特征融合与后融合对目标检测性能具有提升作用。

表 3 不同融合方法在KITTI验证集上的性能对比

Table 3 Performance comparison of different fusion methods on the KITTI validation set unit: %

Fusion strategy	Car			Pedestrian			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Deep fusion	78.85	63.62	61.79	54.31	51.05	47.21	64.38	46.14	43.86
Late fusion	77.48	61.75	59.60	53.08	49.82	45.13	61.79	44.61	40.22
Deep & late fusion	82.95	67.48	64.22	58.90	55.33	50.16	68.42	48.52	46.08

4.6 二维目标检测结果比较

为了分析分组方法在目标检测任务中的作用，在VOC2007数据集上对所提算法中的二维图像目标检测方法与Faster R-CNN进行对比，都采用相同的方法进行训练。为了比较检测性能，加入YOLOv3算法在同一数据集上的表现进行比较。

通过全类平均正确率(mAP)评估检测性能,测试结果如表4所示。测试结果表明,分组处理对二维目标检测性能具有显著的提升作用。

表4 不同二维目标检测算法在VOC2007数据集上的mAP比较

Table 4 mAP comparison of different two-dimensional target detection algorithms on VOC2007 dataset

Method	mAP /%
Faster R-CNN	69.9
YOLOv3	74.4
Proposed method	76.2

4.7 与现行方法及单模态对照方法的对比实验

对于汽车类别,对所提算法与几种性能最佳的算法进行比较。包括基于图像的方法,即Mono3D[31]和3DOP[32];基于LiDAR的方法,即VeloFCN[33]和MV3D。Mono3D、3DOP和MV3D使用预训练的模型进行初始化。以上算法的实验数据来自VoxelNet论文的实验部分,见文献[14]。按照文献中的实验设置,其他现行算法使用预训练的模型进行初始化,再在KITTI数据集上进行训练。对于所提算法,使用KITTI提供的LiDAR数据和RGB图像数据从头开始训练,网络权重参数随机初始化。

为了分析多模态信息融合的重要性,实验中设置了两个单模态对照算法,对照算法网络结构均与本文LiDAR点云数据处理模块相同,仅使用点云数据进行目标检测,第一个对照算法不经分组操作,而第二个对照算法经与所提算法相同的分组操作。采用KITTI数据集提供的LiDAR数据训练对照算法。

与现行方法的对比结果如表5所示。对于汽车类,在所有难度级别上,所提算法的AP指标均明显优于所有其他方法。具体来说,所提算法的性能明显优于基于LiDAR+RGB的代表性方法MV,在简单、中等、困难三个级别上分别超出11.66个百分点、4.80个百分点、7.66个百分点。

与单模态对照算法的对比结果如表6所示。在三维的汽车、行人和骑自行车者检测上对所提算法与两种单模态对照算法进行了比较。由于三维姿势和形状的高度变化较大,对行人和骑自行车者的检测需要更好的三维形状表示。如表6所示,所提算法在所有类别的三种难度的实验中的平均精度均高于两种单模态对照算法,可见融合二维图像中的信息对三维目标检测性能具有提升作用。同时,与不加入分组操作的单模态对照算法相比,加入了分组操作的对照算法的检测精度在所有类别的三种难度上均有提升,证明了分组操作对局部信息的关注在三维目标检测中的有效性。

表5 不同方法在KITTI验证集上的性能对比

Table 5 Performance comparison of different methods on the KITTI validation set unit: %

Method	Modality	Car		
		Easy	Moderate	Hard
Mono3D	Mono	2.53	2.31	2.31
3DOP	Stereo	6.55	5.07	4.10
VeloFCN	LiDAR	15.20	13.66	15.98
MV3D(BV+FV)	LiDAR	71.19	56.60	55.30
MV3D(BV+FV+RGB)	LiDAR+Mono	71.29	62.68	56.56
Proposed algorithm	LiDAR+RGB	82.95	67.48	64.22

表6 所提算法与单模态对照算法在KITTI验证集上的性能对比

Table 6 Performance comparison of proposed algorithm and the monomodal comparison algorithms on the KITTI validation set unit: %

Algorithm	Car			Pedestrian			Cyclist		
	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Mono	72.26	60.41	55.82	44.75	41.37	37.94	56.41	36.93	34.65
Mono+grouping	78.79	64.02	58.65	51.88	48.15	43.29	62.11	41.29	38.90
Proposed algorithm	82.95	67.48	64.22	58.90	55.33	50.16	68.42	48.53	46.08

4.8 速率测试

为了验证所提算法的高效性,设置了一个对照算法,其网络结构与参数与所提算法相同,但在处理点云数据时不使用构建密集张量结构与哈希表

的方法。在 TitanX GPU 和 1.7 GHz CPU 上对所提算法和对照算法进行速率测试。结果显示，所提算法总共耗费的推理时间为 225 ms，其中分组操作花费 5 ms；而对照算法耗费的时间高达 345 ms，其中分组操作花费 120 ms，速度远远低于所提算法。测试结果表明，对点云构建密集张量结构与哈希表后对提升数据处理效率具有显著效果。

5 结 论

提出了一种基于多模态融合的目标检测算法，算法网络结构包含两个子网，分别用于点云特征提取与二维图像特征提取。将 LiDAR 点云与对应的 RGB 图像划分为一一对应的等距组块，分别输入两个子网得到特征图。采用特征融合的策略，对对应组块的特征进行拼接。拼接后的特征图通过区域生成网络输出目标检测框，对三维与二维的检测框进行后期融合，得到最终的检测结果。在 KITTI 点云数据集与 VOC2007 图像数据集上评估了所提算法性能，并与其他单模态与多模态目标检测算法进行了比较分析，相关实验结果证明了所提融合策略的有效性及算法的优越性。

参 考 文 献

[1] 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, June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 3354-3361.

[2] Gomez-Ojeda R, Briales J, Gonzalez-Jimenez J. PL-SVO: semi-direct monocular visual odometry by combining points and line segments[C]//2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 9-14, 2016, Daejeon, Korea (South). New York: IEEE Press, 2016: 4211-4216.

[3] Park Y, Lepetit V, Woo W. Multiple 3D object tracking for augmented reality[C]//2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality, September 15-18, 2008, Cambridge. New York: IEEE Press, 2008: 117-120.

[4] Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: a survey and taxonomy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(2): 423-443.

[5] Xie L, Xiang C, Yu Z X, et al. PI-RCNN: an efficient multi-sensor 3D object detector with point-based attentive cont-conv fusion module[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12460-12467.

[6] Chen X Z, Ma H M, Wan J, et al. Multi-view 3D object detection network for autonomous driving[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6526-6534.

[7] Mousavian A, Anguelov D, Flynn J, et al. 3D bounding box estimation using deep learning and geometry[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5632-5640.

[8] Chabot F, Chaouch M, Rabarisoa J, et al. Deep MANTA: a coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1827-1836.

[9] Mottaghi R, Xiang Y, Savarese S. A coarse-to-fine model for 3D pose estimation and sub-category recognition[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 418-426.

[10] Wang Y, Chao W L, Garg D, et al. Pseudo-LiDAR from visual depth estimation: bridging the gap in 3D object detection for autonomous driving[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 8437-8445.

[11] You Y R, Wang Y, Chao W L, et al. Pseudo-LiDAR++: accurate depth for 3D object detection in autonomous driving[EB/OL]. (2020-02-15)[2021-03-01]. https://arxiv.org/abs/1906.06310.

[12] Li B Y, Ouyang W L, Sheng L, et al. GS3D: an efficient 3D object detection framework for autonomous driving[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 1019-1028.

[13] Li P L, Chen X Z, Shen S J. Stereo R-CNN based 3D object detection for autonomous driving[C]//2019

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 7636-7644.

[14] Zhou Y, Tuzel O. VoxelNet: end-to-end learning for point cloud based 3D object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4490-4499.

[15] Yan Y, Mao Y X, Li B. SECOND: sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): E3337.

[16] Lang A H, Vora S, Caesar H, et al. PointPillars: fast encoders for object detection from point clouds [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 12689-12697.

[17] Charles R Q, Hao S, Mo K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 77-85.

[18] Shi S S, Wang X G, Li H S. PointRCNN: 3D object proposal generation and detection from point cloud [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 770-779.

[19] Chen Y L, Liu S, Shen X Y, et al. Fast point R-CNN[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 9774-9783.

[20] Yang Z T, Sun Y N, Liu S, et al. STD: sparse-to-dense 3D object detector for point cloud[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 1951-1960.

[21] Shi S S, Guo C X, Jiang L, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 10526-10535.

[22] Shi S S, Wang Z, Shi J P, et al. From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(8): 2647-2664.

[23] Qi C R, Liu W, Wu C X, et al. Frustum PointNets for 3D object detection from RGB-D data[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 918-927.

[24] Xu D F, Anguelov D, Jain A. PointFusion: deep sensor fusion for 3D bounding box estimation[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 244-253.

[25] Wang Z X, Jia K. Frustum ConvNet: sliding frustums to aggregate local point-wise features for amodal 3D object detection[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), November 3-8, 2019, Macao, China. New York: IEEE Press, 2019: 1742-1749.

[26] Chen X Z, Ma H M, Wan J, et al. Multi-view 3D object detection network for autonomous driving[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6526-6534.

[27] Ku J, Mozifian M, Lee J, et al. Joint 3D proposal generation and object detection from view aggregation [C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 1-5, 2018, Madrid, Spain. New York: IEEE Press, 2018: 18392975.

[28] 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), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 7337-7345.

[29] Liang M, Yang B, Wang S L, et al. Deep continuous fusion for multi-sensor 3D object detection[M]// Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11220: 663-678.

[30] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-

13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1440-1448.

[31] Chen X Z, Kundu K, Zhang Z Y, et al. Monocular 3D object detection for autonomous driving[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2147-2156.

[32] Chen X Z, Kundu K, Zhu Y K, et al. 3D object proposals using stereo imagery for accurate object class detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(5): 1259-1272.

[33] Li B, Zhang T L, Xia T. Vehicle detection from 3D lidar using fully convolutional network[EB/OL]. (2016-08-29) [2021-03-01]. https://arxiv. org/abs/1608.07916.