基于多级特征融合的体素三维目标检测网络

张吴冉[a]，胡春燕[a]，陈泽来[a]，李菲菲[b]
（上海理工大学 a.光电信息与计算机工程学院 b.医疗器械与食品学院，上海 200093）

摘要：目的 为精确分析点云场景中待测目标的位置和类别信息，提出一种基于多级特征融合的体素三维目标检测网络。**方法** 以 2 阶段检测算法 Voxel−RCNN 作为基线模型，在检测一阶段，增加稀疏特征残差密集融合模块，由浅入深地对逐级特征进行传播和复用，实现三维特征充分的交互融合。在二维主干模块中增加残差轻量化高效通道注意力机制，显式增强通道特征。提出多级特征及多尺度核自适应融合模块，自适应地提取各级特征的关系权重，以加权方式实现特征的强融合。在检测二阶段，设计三重特征融合策略，基于曼哈顿距离搜索算法聚合邻域特征，并嵌入深度融合模块和 CTFFM 融合模块提升格点特征质量。**结果** 实验于自动驾驶数据集 KITTI 中进行模拟测试，相较于基线网络，在 3 种难度等级下，一阶段检测模型的行人 3D 平均精度提升了 3.97%，二阶段检测模型的骑行者 3D 平均精度提升了 3.37%。**结论** 结果证明文中方法能够显著提升目标检测性能，且各模块具有较好的移植性，可灵活嵌入到体素类三维检测模型中，带来相应的效果提升。

关键词：三维目标检测；残差融合；自适应融合；特征增强；三重特征融合
中图分类号：TP311 **文献标识码：**A **文章编号：**1001-3563(2022)15-0042-12
DOI：10.19554/j.cnki.1001-3563.2022.15.005

Voxel-based 3D Object Detection Network Based on Multi-level Feature Fusion

ZHANG Wu-ran[a], *HU Chun-yan*[a], *CHEN Ze-lai*[a], *LI Fei-fei*[b]
(a. School of Optical-electrical and Computer Engineering b. School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

ABSTRACT: The work aims to accurately analyze the location and classification information of the object to be tested in the point cloud scene, and propose a voxel-based 3D object detection network based on multi-level feature fusion. The two-stage Voxel-RCNN was used as the baseline network. In the first stage, the Sparse Feature Residual Dense Fusion Module (SFRDFM) was added to propagate and reuse the level-by-level features from shallow to deep, to achieve full interactive fusion of 3D features. The Residual Light-weight and Efficient Channel Attention (RL-ECA) mechanism was added to the 2D backbone network to explicitly enhance channel feature representation. A multi-level feature and multi-scale kernel adaptive fusion module was proposed to adaptively extract the weight information of the multi-level features, to achieve a strong fusion with a weighted manner. In the second stage, a Triple Feature Fusion Strategy (TFFS) was designed to aggregate neighborhood features based on the Manhattan distance search algorithm, and a Deep Fusion Module (DFM) and a Coarse to Fine Fusion Module (CTFFM) were embedded to improve the quality of grid features. The algorithm in this paper was tested in the autonomous driving data set KITTI. Compared with the baseline network at three difficulty levels, the average 3D accuracy of pedestrians in the first stage detection model was improved by 3.97%, and

收稿日期：2022–05–16
基金项目：上海市高校特聘教授（东方学者）岗位计划（ES2015XX）
作者简介：张吴冉（1995—），男，上海理工大学硕士生，主攻计算机视觉与目标检测。
通信作者：胡春燕（1976—），女，硕士，上海理工大学讲师，主要研究方向为图像处理与模式识别、计算机视觉等。

the average 3D accuracy of cyclists in the second stage detection model was improved by 3.37%. The experimental results prove that the proposed method can effectively improve the performance of object detection, each module has superior portability, and can be flexibly embedded into the voxel-based 3D detection model to bring corresponding improvements.
KEY WORDS: 3D object detection; residual fusion; adaptive fusion; feature enhancement; triple feature fusion

随着无人驾驶、室内移动机器人等技术的发展，大量研究人员开始关注三维目标检测领域。基于三维目标检测可获取目标物体类别、位置、三维尺寸及姿态等更加详细的信息，借助检测结果可实现对周围环境的精确感知，保证设备安全运行。

目前，三维目标检测算法主要分为 2 类：基于点云表示[1-4]的方法和基于点云和图像多模态融合[5-7]的方法。其中基于点云表示的方法又可分为体素[8-11]方式（Voxel-Based）和原始点[12-14]方式（Point-Based）。其中体素网络以较快的推理速度广受欢迎。此类算法在点云采样阶段采用网格化处理，将离散的点云均匀分割成立体体素。但此种方法在采样过程中会导致信息丢失，影响目标检测效果。原始点的方式直接从初始点云数据中提取特征，相较于网格化的方式保留更多目标细粒度信息，但是逐点特征提取带来高昂的计算代价。基于多模态融合的方法则通过增加图像数据处理分支，对点云分支进行信息补充，缓解小目标物体漏检问题，但异构数据融合困难，计算复杂度较高，网络推理速度较低。文中为平衡网络精度和实时化性能，选取体素检测算法 Voxel–RCNN[15]作为基线网络，并在此网络上进行分析和改进，实现对中小目标检测效果的提升。Voxel–RCNN 网络第 1 阶段和 SECOND[16]结构雷同，主要由 3 个部分组成：体素特征编码模块、三维稀疏特征提取模块、二维主干网络。其中，体素特征编码模块对输入点云进行均匀采样和特征处理，得到体素级特征表示。三维稀疏特征提取模块对输入的体素特征进行稀疏化及卷积运算等相关操作，实现对浅层特征的深层抽象。二维主干网络于二维鸟瞰图上进行最终检测，生成三维检测框。然而，SECOND 网络在稀疏卷积特征提取模块仅使用简单的卷积塔结构对特征进行下采样抽象，忽略了多层特征之间的信息补充。二维主干网络由下采样层、上采样层组成，虽然在上采样层进行特征堆叠实现特征粗略的融合，但是忽略了多级特征之间的相关性。Voxel–RCNN 第 2 阶段的精化模块仅对最高级的体素稀疏特征进行小范围特征搜索，忽视了低级特征和多范围邻域特征的重要性。为了解决以上不足之处，文中对 Voxel–RCNN 网络进行改进，设计基于多级特征融合的体素三维目标检测网络。

1　网络设计

提出的基于多级特征融合的体素三维目标检测网络结构见图 1，主要包含 4 个部分：稀疏特征残差密集融合模块、残差轻量化高效通道注意力机制、多级特征及多尺度核自适应融合模块和三重特征融合策略。文中主要改进点如下。

1）在三维稀疏特征提取部分设计稀疏特征残差密集融合模块（Sparse Feature Residual Dense Fusion Module，SFRDFM）。为了高效地处理三维特征，使用三维稀疏卷积[17]和子流形卷积[18]算法，设计稀疏特征残差半密集融合层，混合叠加此卷积层形成主干，缓解特征冗余的同时加强逐层特征之间的信息交流补充。

2）在二维主干网络模块中，通过降低特征通道数量以降低计算量，同时增加残差轻量化高效通道注意力机制（Residual Light-Weight Efficient Channel Attention Mechanism，RL-ECA）对损失的通道信息进行补充增强，减少计算量的同时提升了检测器性能。

3）在二维主干网络上采样阶段增加了多级特征及多尺度核自适应融合模块（Module of Multi-Level Feature And Multi-Scale Kernel Adaptive Fusion，MFMKAF），通过编码多级特征依赖关系，自适应地融合低层空间特征，中层复合特征和高层语义特征，实现多级特征之间的交流融合，进一步提升特征表达能力。

4）在第二阶段精化模块部分，设计三重特征融合策略（Triple Feature Fusion Strategy，TFFS），包含多级特征融合、多范围分组聚合和多尺度格点采样策略，组合以上 3 种策略用于二次搜索聚合体素稀疏特征。并设计 2 种不同的格点特征融合模块：深度融合模块（Deep Fusion Module，DFM）对输入特征进行多重提取压缩融合；由细粒度到粗粒度的融合模块（Coarse to Fine Fusion Module，CTFFM）自适应地融合输入特征，生成更具区分性的格点特征，进一步精化三维建议框。

相较于二维图像数据，点云数据由于维度增加导致计算复杂度指数提升，数量庞大的点云数据使得逐点处理耗费巨大算力。为能够高效的处理三维数据，文中使用均匀体素采样算法对输入点云进行信息降维。首先，将大范围点云数据进行场景截取，得到 $L\times W\times H$ 范围内的数据，数据中的每个点由点 P 表示，包含的初始属性有三维空间坐标 (x,y,z) 和点反射强度 r。然后分别于 X,Y,Z 3 个维度对点云进行均匀分割，得到大小相同的立体体素，每个体素的大小表示为 $v_L\times v_H\times v_w$，体素分辨率为 $L'\times H'\times W'$，其中 $L'=L/v_L$，$W'=W/v_W$，$H'=H/v_H$。最后对输入

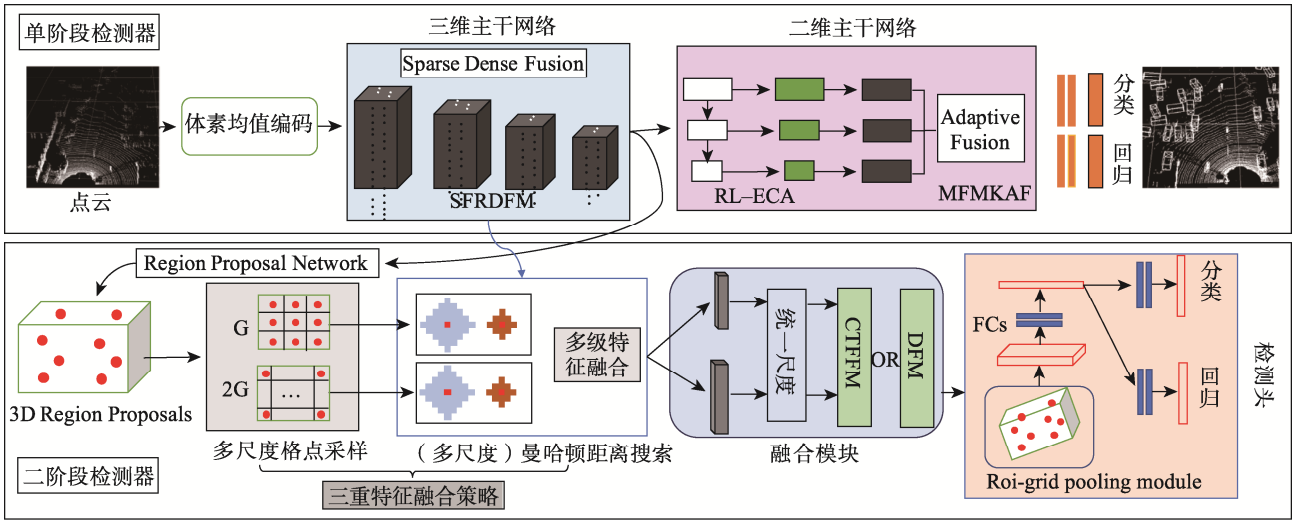

图 1　文中提出的三维目标检测网络框架图

Fig.1 Structure of 3D object detection network proposed in the work

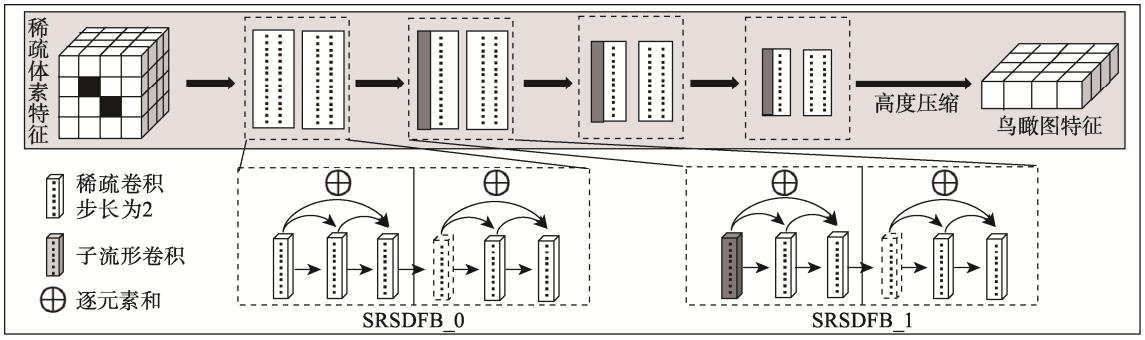

图 2　稀疏特征残差密集融合模块

Fig.2 Sparse feature residual dense fusion module (SFRDFM)

的点云信息 $P_i=\left\{\left[x_i,y_i,z_i,r_i\right]\in\mathcal{R}^4\right\}_{i=1,\ldots,i\leqslant T}$（$T$ 表示点云最大点数）进行逐体素的点分组，得到分组后的体素 $V_{k=1,\ldots,k\leqslant K}=\left\{P_I^k\right\}$（$K$ 表示体素最大数量）。文中沿用基线网络 SECOND 中体素编码模式，直接对体素内的所有点进行均值池化，获得均匀体素级特征：

$$\hat{V}_{x,y,z}^k=\left\{\left[v_x^k,v_y^k,v_z^k,v_r^k\right]\in\mathcal{R}^4\right\}_{k=1,\ldots,k\leqslant K} \tag{1}$$

接下来使用卷积算法对特征进行深层处理。

1.1　稀疏特征残差密集融合模块

在点云体素化过程中有超过 90%体素为空值体素，传统三维稠密[19]卷积会遍历所有区域，加大计算代价和内存负担的同时，还会导致稠密数据失真。为了进一步的提升体素特征提取算法的实时性，Graham 等[17]提出稀疏卷积（SC）和子流形卷积算法[18]（SSC）替换稠密卷积，保证稀疏算法仅在稀疏化数据上运行，核心思想是通过输入数据的稀疏性限制输出数据的稀疏性，降低三维卷积操作的计算量和内存占用。为缓解稀疏卷积（SC）随着网络深度的扩展出现稀疏性弱化的问题，增加子流形卷积（SSC）算法处理数据，此算法仅对输入的非空值区域进行相应运算，且只对非空值区域赋值，最大程度保持数据稀疏性。对于深层特征提取网络而言，多层特征图包含多尺度详细信息，这些信息对于场景中目标的检测是非常有用的。Voxel-RCNN 的三维骨架是经典的由浅入深式稀疏卷积下采样结构，考虑到此结构忽视了各层特征之间的信息交流，损失大量的细粒度信息。为缓解以上问题，文中在此基础上设计了稀疏特征残差密集融合模块（Sparse Feature Residual Dense Fusion Module，SFRDFM），模块结构见图 2。由于子流形卷积（SSC）对有值位置作强制限制导致一定程度的信息丢失，稀疏卷积（SC）带来位置信息失真的缺点，文中采用稀疏卷积和子流形卷积混用的结构平衡两种算法带来的问题。首先叠加 5 层子流形卷积对输入的稀疏体素数据进行特征处理，再叠加一层稀疏卷积（SC）和 4 层子流形卷积（SSC）继续提取特征。密集融合前 3 层和后 3 层稀疏特征，这里称为稀疏特征残差半密集融块（Sparse feature Residual Semi-Dense fusion Block，SRSDFB），以半数融合 5

层特征方式，防止过多特征叠加冗余，影响检测效果。区别于文献[20]通道堆叠（Concatenation）方式，模块使用逐元素相加进行融合，达到稳定网络训练、降低计算代价、复用浅层特征的作用。通过使用步长为 2 的稀疏卷积和 ReLU 激活函数进行特征下采样，得到 3 组不同尺度的稀疏特征，起到特征由低到高的抽象、增大感受野和降低特征维度的作用。理论上可以对 SRSDFB 叠加更多子流形卷积层设计更深的特征提取模块，但考虑到推理时间的消耗和参数复杂度提升，仅使用 5 层叠加形式。

1.2　二维特征自适应融合模块

如果在三维特征图上生成锚框（Anchor）将出现数量过多的空三维框，导致计算资源的负担和正负锚框不平衡的问题。而在自动驾驶场景中，目标物体基本处于地面上，目标空间位置相对固定，位于 Z 轴的高度信息变化较小，为了进一步降低计算复杂度，将三维特征图沿着 Z 轴方向向下投影得到二维鸟瞰特征图表示，再基于鸟瞰图进行三维框的估计。文中二维主干网络模块见图 3，此模块在基线网络的基础上进行改进，在初始的 2 层卷积塔结构上增加第 3 层卷积块，并增加残差轻量化高效通道注意力机制和多级特征及多尺度核自适应融合模块。

1.2.1　二维卷积塔模块

本模块由常规的特征下采样和上采样结构组成，模块架构图如图 3 中左框图所示。文中在基线网络 Voxel-RCNN 的第 1 阶段卷积下采样模块增加一组卷积块得到 3 层卷积塔结构，加深网络特征提取能力和尺度变化，此 3 组卷积块均是常规的 5 层二维卷积堆叠组成，用作提取鸟瞰图的语义信息。文中将自上而下的 3 组卷积块命名为卷积块_0，卷积块_1，卷积块_2。将卷积块_0 的步长设置为 1，输出通道数为 64，卷积块_1 和卷积块_2 的步长设置为 2，输出通道递增为 128 和 256，此操作对特征进行提取抽象的同时起到缩减特征尺度和增大感受野的功能。

其中卷积块_0 可以保留更细节的目标位置信息，卷积块_1 可提取到相对细节的位置信息和语义信息，卷积块_2 提取得到更加抽象的语义信息。上采样结构则使用转置卷积操作对下采样模块中输出的不同尺度的特征图进行尺度恢复，并且固定 3 组特征通道数为 128，相比于原网络，此操作加深卷积层的同时降低了通道数量，虽然损失了一些有效信息，但是增加了小尺度的特征计算，能够提升大目标的检测性能，而且通过压缩特征通道数量去降低计算代价，维持计算量的平衡。

1.2.2　残差轻量化高效通道注意力机制

由于在上采样阶段减少了特征通道数量，虽然降低了计算复杂度，但是会损失一些有效信息导致特征质量降低，为缓解此问题，在该模块中增加残差轻量化高效通道注意力机制（Residual Light-weight Efficient Channel Attention Mechanism，RL-ECA）对以上 3 组特征的通道信息进行特征增强。该模块的网络架构见图 4。沿用文献[21]的网络框架，此文献中注意力模块首重轻量化及高效性，通过使用一维卷积实现跨通道的信息交互来降低计算复杂度。区别于其他注意力模块在特征提取阶段的维度压缩操作，此模块通过保持通道数量恒定的方式，保留更多通道信息。文中在此基础上进行简单修改，移除自适应卷积核提取函数，固定一维卷积提取核的尺寸为 3，增加残差融合操作对输出特征进行有效补偿，详细过程见式（2）。

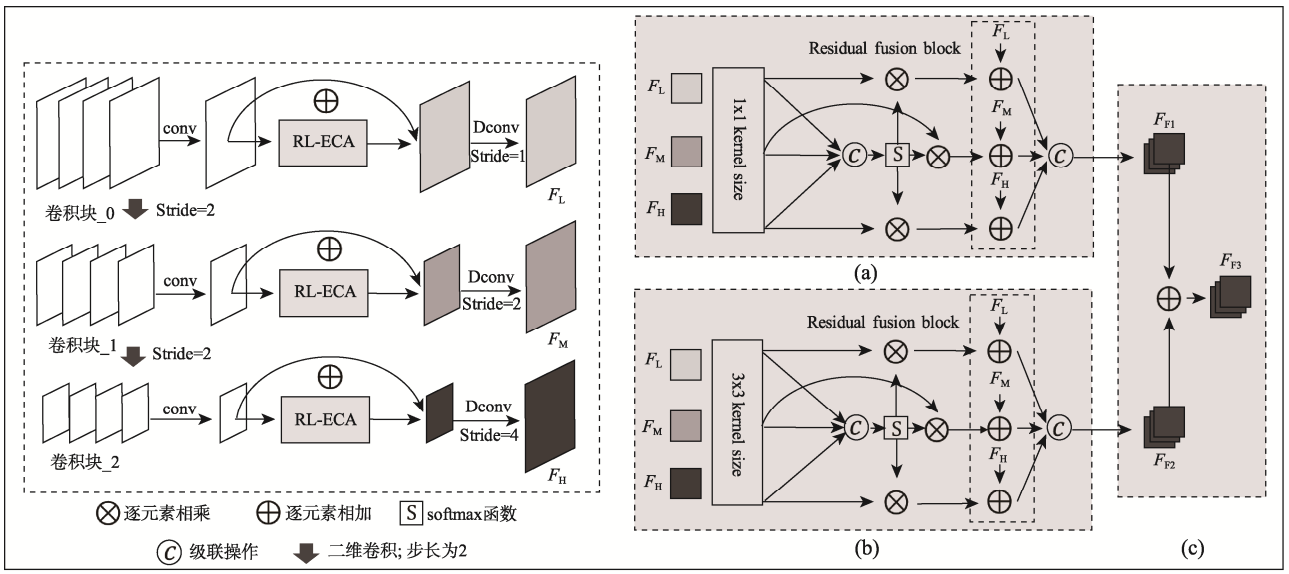

图 3　多级特征及多尺度核自适应融合模块

Fig.3 Module of multi-level feature and multi-scale kernel adaptive fusion (mfmkaf)

$$\begin{cases} W=\sigma\left[W_2 Relu\left(W_1\left(g\left(X\right)\right)\right)\right] \\ F_{\text{new}}=W\times X+X_{\text{res}} \end{cases} \tag{2}$$

式中：X 为输入特征；g 为全局平均池化；σ 为 sigmoid 函数。

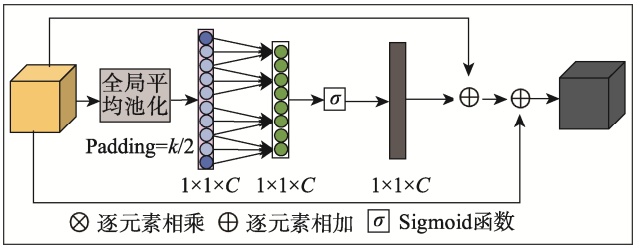

图 4　残差轻量化高效通道注意力机制
Fig.4 Residual light-weight efficient channel attention mechanism

1.2.3　多级特征及多尺度核自适应融合模块

在卷积塔结构中获取了 3 种不同级别的特征，分别是低层空间特征、中层复合特征和高层语义特征。基线网络中作者仅对多级特征进行简单的堆叠融合（这里称为弱融合操作），没有充分利用不同级别特征的依赖关系。考虑到多级特征对于目标精确定位和分类的重要性[22-23]，文中设计了多级特征及多尺度核自适应融合模块（Module of Multi-Level Feature And Multi-Scale Kernel Adaptive Fusion，MFMKAF）对 3 种级别的特征进行深层的融合。此模块的网络框架如图 3 右半部分 a、b、c 3 个框图所示。首先使用多尺度的卷积核将 3 组特征图压缩成 1 维通道，对其空间信息进行自适应的特征提取。如模块 a 所示，使用 1×1 尺寸卷积核对多级特征分别处理，然后在通道维度上对 3 组 1 维的权重图进行堆叠拼接（Concatenation），并使用 Softmax 函数归一化建立三组特征之间的关联性得到空间权重，详细过程见式（3）。

$$W_{\text{l}}^{i};W_{\text{m}}^{i};W_{\text{h}}^{i}=S\left[F\left(C_{k=i,i=\{1,2\}}^{\text{share}}\left(F_{\text{L}},F_{\text{M}},F_{\text{H}}\right)\right)\right] \tag{3}$$

式中：F 表示堆叠融合（Concatenation）；S 表示 Softmax 函数；C 表示卷积算子。

将 3 组权重和相应的输入特征逐元素相乘后逐通道堆叠融合（Concatenation），再增加残差融合块 R（Residual Fusion Block，RFB）将输入特征以加和的方式融合到新特征上，从而实现多级特征自适应的强融合，详细过程见式（4）。

$$F_{F1}=F[R(F_{\text{L}}\times W_{\text{l}}^{1}+F_{\text{M}}\times W_{\text{m}}^{1}+F_{\text{H}}\times W_{\text{h}}^{1})] \tag{4}$$

模块 b 整体流程似模块 a，详细过程见式（5）。

$$F_{F2}=F[R(F_{\text{L}}\times W_{\text{l}}^{2}+F_{\text{M}}\times W_{\text{m}}^{2}+F_{\text{H}}\times W_{\text{h}}^{2})] \tag{5}$$

二者区别在于不同尺度的核操作，模块 a 采用 1×1 核，能够提取更详细的小目标位置信息，模块 b 采用 3×3 核，能够提取较大目标位置信息，交替使用模块 a、b 能够让网络拟合不同的任务要求。模块 c 则是将模块 a、b 输出的特征进一步的相加融合，从而得到更具表达能力的新特征，详细过程见式（6）。

$$F_{\text{F3}}=\left(F_{\text{F1}}-R\right)+\left(F_{\text{F2}}-R\right)+R \tag{6}$$

1.3　三重特征融合策略

体素化三维检测网络分为单阶段和两阶段三维检测器[24]，两者主要区别在于两阶段算法增加了区域建议模块（RPN）和精化模块。其中，精化模块的主要作用是对区域建议模块得到的三维建议框进一步的细化处理，一定程度上增加了计算量，但对精度提升较大。

一阶段检测器将特征处理成二维鸟瞰特征表达，降低了计算代价，但忽略了三维空间结构信息。Voxel–RCNN 通过增加二阶段精化模块，对具有完整三维结构的体素稀疏特征进行相关操作，恢复特征的三维结构上下文信息。首先基于 RPN 网络对鸟瞰特征进行处理，生成大量三维建议框（3D Region Proposals）。然后将三维框进行网格分割，将分割格点作为关键点保存并映射回稀疏体素特征空间，基于关键点对邻域内的体素特征进行二次采集，获取的格点特征用于进一步精化三维框。

1.3.1　多级特征融合

对于稀疏的点云场景而言，低级特征具备更多的细粒度信息，为进一步获取信息量丰富的格点特征，Voxel–RCNN 采用多级特征融合策略。具体结构见图 5 中的模块 a。通过对各个级别的稀疏体素特征进行曼哈顿距离搜索，将采集的 L2、L3 和 L4 级体素特征进行堆叠（Concatenate）融合，然后进行三维候选框的进一步精化。

1.3.2　多范围分组聚合

对于场景检测任务而言，目标局部邻域的范围大小选取尤为重要，搜索范围越小，能采集到的有效信息越少，相反，搜索范围越大，能采集的信息越多，但过大范围会引入更多的背景噪声，影响检测性能。文中对曼哈顿距离搜索算法设置 2 种大小不同的度量距离，分别为 R 和 $2R$，同时作用于特征空间进行信息采集，获取基于格点的多范围邻域特征，最后将两种范围内的邻域特征进行堆叠（Concatenate）融合。详细结构见图 5 中的模块 b。

1.3.3　多尺度格点采样

文中通过格点采样的方式对邻域范围信息进行聚合，其中格点参数设置极为重要。首先将三维建议框分割成 2 种尺寸不同的表示，分别为（3×3×3）的粗粒度和（6×6×6）的细粒度表示，其中 G=3，$2G$=6。然后基于 2 组格点进行特征采样和融合。详细结构见图 5 中的模块 c。对于模块 c 生成的 2 组格点特征，先使用上采样算法统一尺度，后为进一步融合以上 2 组特征，设计了 2 种格点特征融合模块（Grid Feature

Fusion Module；GFFM），分别为深度融合模块（Deep Fusion Module，DFM）和由粗粒度到细粒度的融合模块（Coarse to Fine Fusion Module，CTFFM），详细结构见图 6。其中深度融合模块使用叠加方式获取融

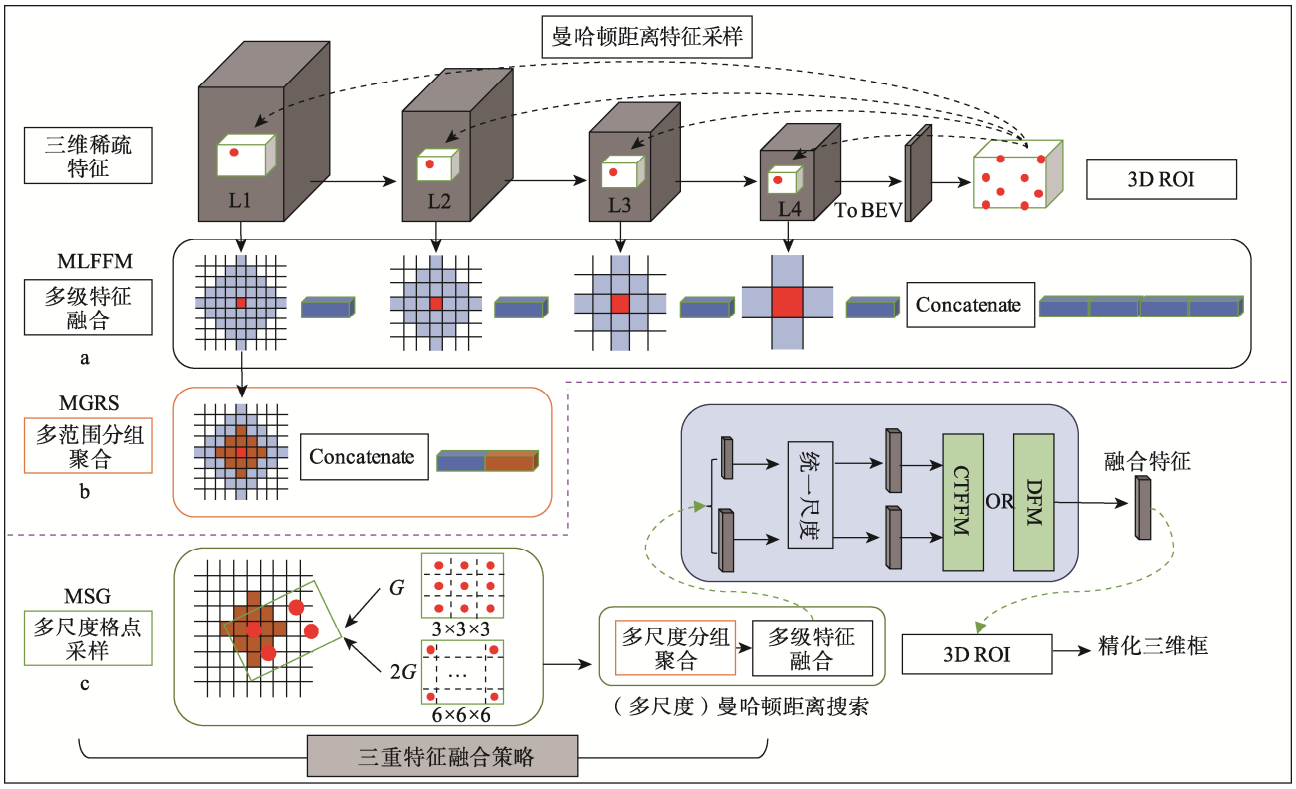

图 5　三重特征融合策略

Fig.5 Triple feature fusion strategy

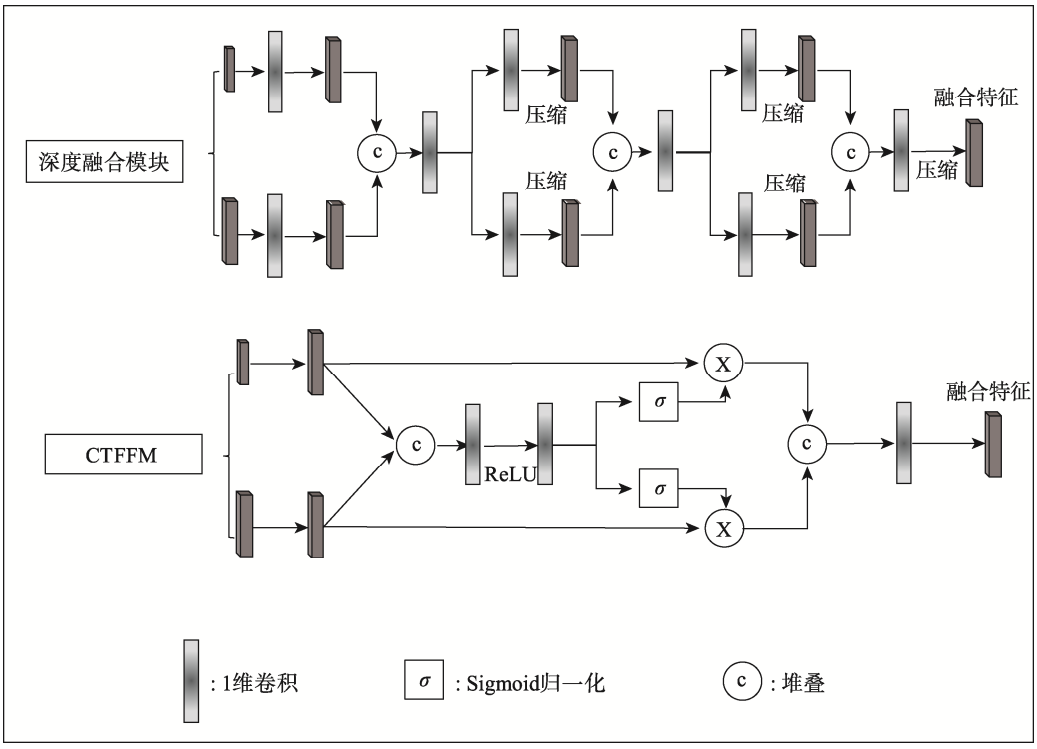

图 6　格点特征融合模块

Fig.6 Grid Feature Fusion Module

合特征，后接 3 组双层卷积层以重复叠加和压缩方式深层次的融合 2 组特征。为了更好的编码两组不同尺度格点特征的依赖关系，CTFFM 模块将两组特征堆叠融合后，接入卷积模块和 Sigmoid 进行归一化，生成粗粒度格点特征和细粒度格点特征的权值关系，有辨别地融合 2 组特征，生成具有更高语义和更多细节的输出特征。

1.4 损失函数

为优化网络，文中使用和文献[4]相同的锚框设置和损失函数，对于每个锚框（Anchors），使用 7 维向量表示框的位置，1 维向量表示类别信息。本网络需要预测汽车、行人、骑行者三种类别，不同类别需要匹配不同的 IOU 阈值来筛选正负锚框，分别计算出 3 种类别的锚框和真实框的交并比。对于汽车而言，如果交并比大于 0.6 则被认为是正锚框，小于 0.45 则被认为是负锚框，其他锚框不做训练使用，行人和骑行者的设定阈值为[0.35, 0.5]。

文中损失函数设置主要分为 2 个部分。

第 1 部分为 RPN 损失函数，详细见式（7）。

$$L_{\text{rpn}}=\frac{1}{N_p}[\sum_i l_{\text{cls}}\left(p_i^a,g_i^*\right)+\mathbb{I}\left(g_i^*\geqslant 1\right)\sum_i l_{\text{reg}}\left(d_i^a,t_i^*\right)] \tag{7}$$

式中：N_p 为正锚框数量；p_i^a 和 d_i^a 为分类和回归分支的预测结果；g_i^* 和 t_i^* 为分类标签和回归基准目标。$\mathbb{I}\left(g_i^*\geqslant 1\right)$ 表示回归损失仅基于正锚框计算。l_{cls} 和 l_{reg} 为分类损失和回归损失。分类损失使用 Focal loss[28]计算，回归损失使用 Huber Loss 计算。

第 2 部分为检测头损失，详细见式（8）。

$$L_{\text{head}}=\frac{1}{N_{\text{s}}}[\sum_i l_{\text{cls}}\left(p_i,l_i^*\right)+\mathbb{I}\left(I_i\geqslant \vartheta_{\text{reg}}\right)\sum_i l_{\text{reg}}\left(d_i^a,t_i^*\right)] \tag{8}$$

式中：N_{s} 为于训练阶段采样的建议框数量；$\mathbb{I}\left(I_i\geqslant \vartheta_{reg}\right)$ 表示仅有 IOU 大于等于 ϑ_{reg} 的建议框才介于回归损失计算。分类置信度使用 Entropy Loss 计算，回归分支使用 Huber Loss 计算。

2 实验结果与分析

2.1 实验配置

文中网络使用的服务器硬件配置为：Linux64 位操作系统：Ubuntu 18.04，英伟达 RTX 3080 10 GB 显卡。

环境配置为：Pytorch1.8.0、python3.7.2、CUDA11.3、CUDNN11.3。

网络参数设置：将点云数据进行范围切割，范围为[0, 70.4]，[−40, 40]，[−3, 1]（单位：米），对切割好的点云数据进行体素化，其中单个体素的分辨率为[0.05, 0.05, 0.1]。设置 3 种类别的锚框（Anchors）尺寸：汽车为[1.9, 3.6, 1.56]、行人为[0.6, 0.8, 1.73]、骑行者为[0.6, 1.76, 1.73]，计算锚框和真实框的交并比（IOU）并根据设定阈值筛选正负锚框。在训练时使用初始学习率为 0.003 的 Adam[29]优化器，优化动量参数为 0.9，该实验在单个 GPU 上训练，batch_size 设置为 2，一共训练 80 个 epochs。

2.2 数据集和数据评估

实验使用 KITTI[30]数据集，使用不同的传感器对市区、乡村、高速公路等主要场景进行数据采集，其中三维点云数据由 64 线激光雷达扫描获得，根据数据场景中目标遮挡程度，目标尺寸，截断程度等因素，将目标难易度划分为 3 个等级：简单、中等、困难。根据训练和测试要求划分，获取 7 481 个样本的训练集和 7518 个样本的测试集，再将训练集被进一步的划分为 3 712 个训练样本和 3 769 个验证样本。参考文献[11,16]的测试和验证标准，文中对目标中汽车（Car）、行人（Pedestrians）、骑行者（Cyclist）3 种主要类别进行评估。

为验证文中算法性能，实验结果将和当前的主流网络进行对比。使用平均精度（Average Precision）作为评估指标，设置汽车（Car）交并比的阈值为 0.7，行人（Pedestrians）和骑行者（Cyclist）阈值为 0.5，并对简单、中等、困难等级的目标分别进行验证。

2.3 主流网络对比实验

该小结将网络验证结果和当前的主流三维目标检测网络结果进行比较，表 1、表 2 分别展示了汽车（Car）、骑行者（Cyclist）、行人（Pedestrian）在 3D 和鸟瞰图指标下的检测精度。并且对简单（Easy）、中等（Moderate）、困难（Hard）3 个等级的目标分别进行评估。

实验结果如上图表 1 和表 2 所示，其中*表示基线网络（baseline），由表 1 和表 2 结果可知，增加了 SFRDFM、RL–ECA 和 MKMKAF 的一阶段网络和 SECOND 网络比较，在 3D 指标下行人类别的 3 种难度等级检测精度分别提升了 6.37%、3.26%和 2.28%，在鸟瞰图指标下分别提升了 5.02%、2.84%和 2.22%，并且在汽车和骑行者两种中大型目标类别均有小幅提升。

在一阶段结构基础上增加 TFFS 和 CTFFM 的二阶段网络和 Voxel-RCNN 比较，在 3D 指标下骑行者类别的 3 种难度等级检测精度分别提升了 3.2%、3.92%和 3%，在鸟瞰图指标下提升了 1.81%、3.07%和 2.99%，并且在汽车和骑行者均有不同程度的提升和下降。由此可证明提出的方法能够有效提升检测器性能。

表 1　KITTI 数据集中不同类别在 3D 指标下和主流网络结果对比

Tab.1 Comparison of the results of pedestrians, cars and cyclists in the KITTI data set with the mainstream network under 3D indicators

方法	模态	阶段	行人			汽车			骑行者		
			简单	中等	困难	简单	中等	困难	简单	中等	困难
MV3D[25]	L+C	Two	—	—	—	71.29	62.68	56.56	—	—	—
F–PointNet[26]	L+C	Two	70.00	61.32	53.59	83.76	70.92	63.65	77.15	56.49	53.37
ContFuse[27]	L+C	Two	—	—	—	86.32	73.25	67.81	—	—	—
AVOD–FPN[5]	L+C	Two	—	—	—	84.41	74.44	68.65	—	—	—
VoxelNet[9]	L	Two	57.86	53.42	48.87	81.97	65.46	62.85	67.17	47.65	45.11
TANet[11]	L	Two	—	—	—	87.52	76.64	73	84.53	61.64	57.44
PointPillars[10]	L	One	57.75	52.29	47.90	86.46	77.28	74.65	80.04	62.60	59.52
Point-GNN[8]	L	One	—	—	—	87.89	78.34	77.38	—	—	—
SECOND*[16]	L	One	56.23	52.52	48.53	88.02	78.19	77.03	81.03	67.74	63.88
Ours	L	**One**	**One**	**62.60**	**55.78**	**50.81**	**88.31**	**78.35**	**77.19**	**82.01**	67.61
对比实验结果 1	—	—	+6.37	+3.26	+2.28	+0.29	+0.16	+0.16	+0.98	−0.13	+0.16
Voxel–RCNN*[15]	L	Two	65.63	59.90	55.08	89.30	83.75	78.72	89.20	70.91	68.07
Ours	L	**Two**	65.04	59.74	54.57	**89.66**	**84.30**	**79.08**	**92.40**	**74.83**	**71.07**
对比实验结果 2	—	—	−0.59	−0.16	−0.51	+0.36	+0.55	+0.36	+3.2	+3.92	+3

注：L+C 表示激光雷达和相机多模态融合方法；*表示基线网络；加粗数字表示最优效果。

表 2　KITTI 数据集中不同类别在鸟瞰图指标下和主流网络结果对比

Tab.2 Comparison of the results of pedestrian, car and cyclist in KITTI data set with the mainstream network under aerial view indicators

方法	模态	阶段	行人			汽车			骑行者		
			简单	中等	困难	简单	中等	困难	简单	中等	困难
MV3D[25]	L+C	Two	—	—	—	86.55	78.1	76.67	—	—	—
F–PointNet[26]	L+C	Two	72.38	66.39	59.57	84.02	88.16	76.44	81.82	60.03	56.32
ContFuse[27]	L+C	Two	—	—	—	95.44	87.34	82.43	—	—	—
AVOD–FPN[5]	L+C	Two	—	—	—	88.53	83.79	77.73	—	—	—
VoxelNet[9]	L	Two	65.95	61.05	56.98	89.6	84.81	78.57	74.41	52.18	50.49
TANet[11]	L	Two	—	—	—	—	—	—	—	—	—
PointPillars[10]	L	One	61.63	56.27	52.60	89.65	87.17	84.37	82.25	66.11	62.55
Point-GNN[8]	L	One	—	—	—	89.82	88.31	87.16	—	—	—
SECOND*[16]	L	One	60.48	56.27	53.46	89.49	87.71	86.42	83.29	70.77	68.28
Ours	L	**One**	**65.50**	**59.11**	**55.68**	**89.93**	**88.11**	**87.12**	**84.15**	70.40	67.67
对比实验结果 1	—	—	+5.02	+2.84	+2.22	+0.44	+0.4	+0.7	+0.86	−0.37	−0.61
Voxel–RCNN*[15]	L	Two	66.84	62.64	56.88	90.39	88.32	87.81	91.40	72.96	69.68
Ours	L	**Two**	**68.18**	62.01	56.04	**90.41**	**88.53**	**88.09**	**93.21**	**76.03**	**72.67**
对比实验结果 2	—	—	+1.34	−0.63	−0.84	+0.02	+0.21	+0.28	+1.81	+3.07	+2.99

注：L+C 表示激光雷达和相机多模态融合方法；*表示基线网络；加粗数字表示最优效果。

2.4 消融实验

此章节对上文一阶段的三组模块和二阶段的 2 组模块分别进行组合实验。

一阶段：表 3 中，组合 1、2 和 3 可知，SFRDFM 模块、RL–ECA 模块和 MFMKAF 单模块均提升了 3 种类别目标的检测效果，证明了 3 种模块的有效性。组合 4 使用 2 种模块，中等难度下，行人类别目标检测精度有较高上升，其他类别有所下降。组合 5、6 和 7 是 3 种模块的组合实验，从组合 7 的实验结果来看，小尺度核的模块 a 能够显著提升行人类的小目标效果，由组合 6 的实验结果可知，中尺度核的模块 b 能够提升骑行者此类中型目标检测效果，组合 5 实验结果可知融合模块 c 能够提升目标检测综合性能，但单一类别检测性能方面有所降低。综合以上实验证明改进网络能够更好地学习小目标的特征信息，并且灵活的模块搭配能够应对更多的任务要求。

二阶段：表 4 中，组合 1 和组合 2 表示三重融合策略中第 3 个多尺度格点采样策略的分解实验，格点尺度分别为 3 和 6 的二阶段网络检测结果，尺度为 3 时，行人效果较好，尺度为 6 时，汽车和骑行者效果提升显著。组合 3 是融合 2 种尺度格点特征的检测结果，相较组合 1 和组合 2，3 种目标类别精度均有提升。组合 4 表示在三重融合策略基础上增加了深度融合模块（DFM），结果表明此模块能较高提升骑行者指标。组合 5 在三重特征融合策略基础上增加了 CTFFM 模块，相较于深度融合模块，此模块能够进一步提升骑行者指标，且在汽车类别也有小幅提升。

表 3 SFRDFM、RL–ECA 和 MFMKAF 3 组模块组合对比实验（一阶段）

Tab.3 Comparison experiment of three sets of module combination of SFRDFM, RL-ECA and MFMKAF (the first stage)

方法	SFRDFM	RL–ECA	MFMKAF			行人			汽车			骑行者		
			a	b	c	简单	中等	困难	简单	中等	困难	简单	中等	困难
组合 1	√					59.06	53.81	49.62	88.49	78.64	77.52	80.74	66.61	61.60
组合 2		√				57.48	53.25	48.58	87.17	78.03	76.86	81.01	64.86	60.25
组合 3					√	57.09	52.53	47.25	88.48	78.41	77.14	80.19	65.83	61.41
组合 4	√	√				59.01	54.54	48.99	87.54	78.00	76.66	79.14	67.35	63.30
组合 5	√	√			√	58.02	52.79	47.97	88.22	78.30	77.17	82.14	67.69	63.12
组合 6	√	√		√		56.55	52.91	47.88	88.19	78.29	76.94	83.40	68.18	64.54
组合 7	√	√	√			62.60	55.78	50.81	88.31	78.35	77.19	82.01	67.61	64.04

表 4 TFFS 和 GFFM 2 组模块组合对比实验（二阶段）

Tab.4 Comparison experiment of two sets of module combination of TFFS and GFFM (the second stage)

方法	TFFS		GFFM		行人			汽车			骑行者		
	G=3	2G=6	DFM	CTFFM	简单	中等	困难	简单	中等	困难	简单	中等	困难
组合 1	√				65.91	60.17	55.21	89.11	79.07	78.46	85.74	71.92	68.67
组合 2		√			64.84	59.20	54.66	89.37	84.60	78.84	86.94	73.39	70.97
组合 3	√	√			65.37	60.07	54.92	89.45	84.47	78.91	87.97	74.26	71.40
组合 4	√	√	√		65.50	60.34	55.40	89.18	83.76	78.78	91.74	73.46	70.22
组合 5	√	√		√	65.04	59.74	54.57	89.66	84.30	79.08	92.40	74.83	71.07

2.5　点云结果可视化

对改进网络的检测结果进行可视化，一共处理了 6 组场景，每组场景分别由原始图像、基线网络和文中网络（一阶段）和（二阶段）可视化结果 4 张图像组成。

可视化图见图 7。从图 7a 的点云可视化实例分析可知，文中检测网络均可很好的学习到汽车类别信息，并有效提升了汽车精度。图 7b—e 场景中，基线网络出现大量的误检结果，如图 7 中矩形框所示，将杂物识别成汽车行人等类别，而文中检测网络误检结果逐渐变少。图 7f 场景下，文中网络能够很好的检测行人目标，且遮挡问题情况下，依然能够正确分类汽车目标，而基线网络错把汽车识别成行人。以上可视化结果可直观表明文中算法的有效性。

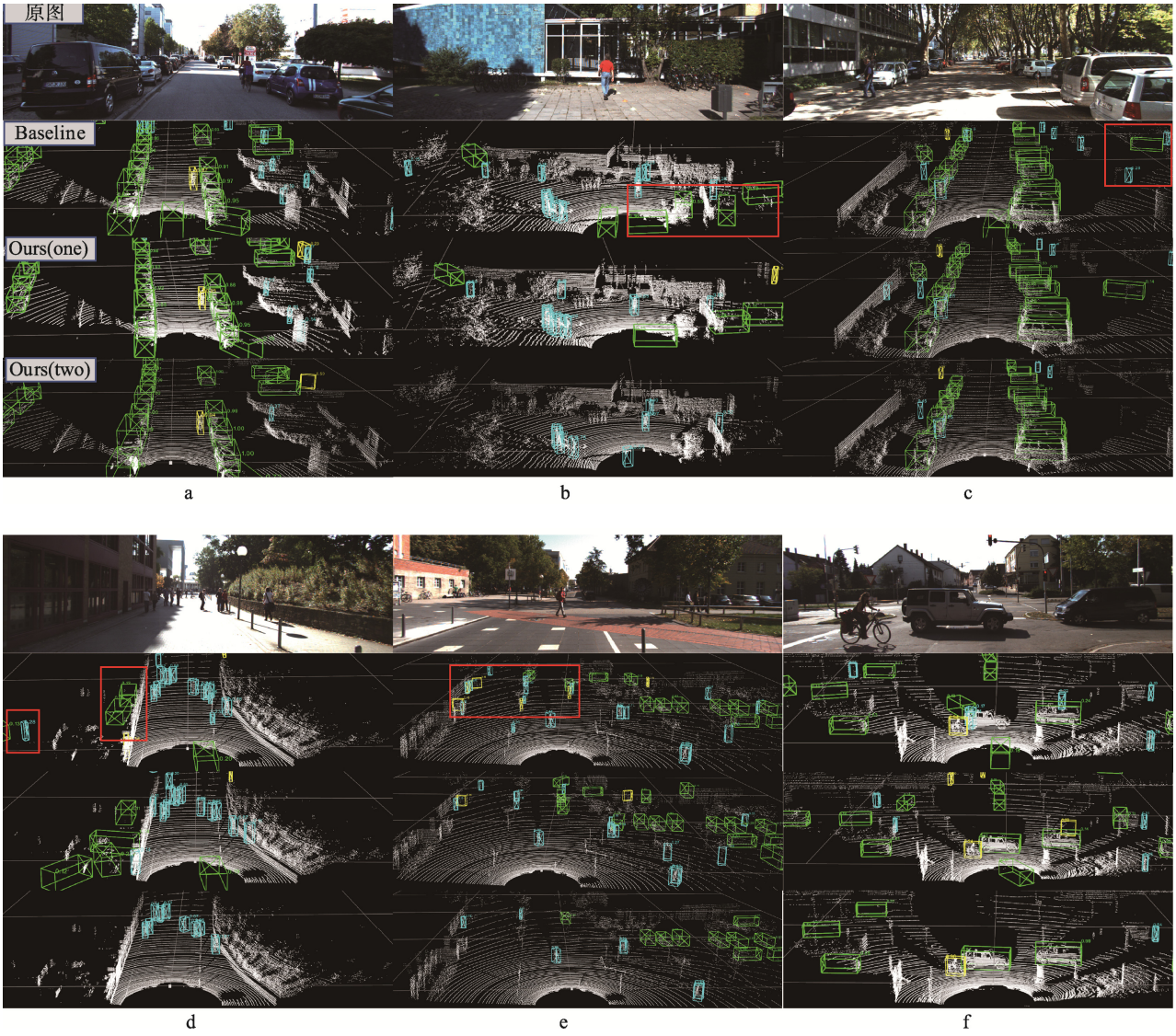

图 7　点云检测结果可视化

Fig.7 Visualization of point cloud detection results

3　结语

文中以体素两阶段网络为基础，于一阶段框架中增加稀疏特征密集融合模块，对稀疏特征逐层进行半密集融合，加强浅层小目标特征复用的同时减少特征冗余。使用轻量化残差高效通道注意力机制稳定计算量的同时对通道特征进行增强。提出多级特征及多尺度核自适应融合模块，以不同尺度的卷积核自适应编码多级特征之间的依赖关系，设置 3 种不同的融合模块以适应不同的任务要求。在 2 阶段，设计了三重特征融合策略，对三维体素稀疏特征空间进行密集的邻域信息搜索聚合，并提出深度融合模块（DFM），使用 3 组双层卷积块对格点特征进行多层次的深度特征抽象融合。另外，还设计 CTFFM 模块分析 2

组格点特征的依赖关系，有区分性地融合 2 组特征以提升特征表达能力，从而进一步提升了检测框的输出质量。

参考文献：

[1] MEYER G P, LADDHA A, KEE E, et al. LaserNet: An Efficient Probabilistic 3d Object Detector for Autonomous Driving[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 12677-12686.

[2] QI C R, SU H, MO K, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 652-660.

[3] QI C R, YI L, SU H, et al. PointNet++: Deep Hierarchical Feature Learning on Point Sets in A Metric Space[J]. Advances in neural information processing systems, 2017: 30-39.

[4] BELTRÁN J, GUINDEL C, MORENO F M, et al. BirdNet: A 3d Object Detection Framework from Lidar Information[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3517-3523.

[5] KU J, MOZIFIAN M, LEE J, et al. Joint 3D Proposal Generation and Object Detection from View Aggregation[C]// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018: 1-8.

[6] LIANG M, YANG B, CHEN Y, et al. Multi-task and Multi-sensor Fusion for 3D Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 7345-7353.

[7] PAIGWAR A, ERKENT O, WOLF C, et al. Attentional PointNet for 3D Object Detection in Point Clouds[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019: 1357-1369.

[8] YAN Y, MAO Y, LI B. SECOND: Sparsely Embedded Convolutional Detection[J]. Sensors, 2018, 1: 3337-3344.

[9] SHI W, RAJKUMAR R. Point-GNN: Graph Neural Network for 3D Object Detection in A Point Cloud[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1711-1719.

[10] ZHOU Y, TUZEL O. Voxelnet: End-to-end Learning for Point Cloud Based 3D Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4490-4499.

[11] LANG A H, VORA S, CAESAR H, et al. PointPillars: Fast Encoders for Object Detection from Point Clouds[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12697-12705.

[12] LIU Z, ZHAO X, HUANG T, et al. TANet: Robust 3D Object Detection from Point Clouds with Triple Attention[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11677-11684.

[13] DENG J, SHI S, LI P, et al. Voxel-RCNN: Towards High Performance Voxel-Based 3D Object Detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 1201-1209.

[14] SHI S, WANG X, LI H. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 770-779.

[15] 李文举，储王慧，崔柳，等. 结合图采样和图注意力的 3D 目标检测方法[J/OL]. 计算机工程与应用, 2022: 1-9. http://kns.cnki.net/kcms/detail/11.2127.TP.20220422.1214.006.html
LI Wen-ju, CHU Wang-hui, CUI Liu, et al. 3D Object Detection Method Combining on Graph Sampling and Graph Attention[J/OL]. Computer Engineering and Applications, 2022: 1-9. http://kns.cnki.net/ kcms/detail/11.2127.TP.20220422.1214.006.html.

[16] DING Z, HAN X, NIETHAMMER M. Votenet: A Deep Learning Label Fusion Method for Multi-Atlas Segmentation[C]// Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019: 202-210.

[17] GRAHAM B. Sparse 3D Convolutional Neural Networks[C]// Proceedings of the British Machine Vision Conference, 2015: 356-368.

[18] GRAHAM B, ENGELCKE M, VAN DER MAATEN L. 3D Semantic Segmentation with Submanifold Sparse Convolutional Networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 9224-9232.

[19] YAN C, SALMAN E. Mono3D: Open Source Cell Library For Monolithic 3D Integrated Circuits[J]. IEEE Transactions on Circuits and Systems I, 2017, 65(3): 1075-1085.

[20] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely Connected Convolutional Networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 4700-4708.

[21] WANG Q, WU B, ZHU P, et al. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11534-11542.

[22] YOO J H, KIM Y, KIM J, et al. 3D-CVF: Generating Joint Camera and Lidar Features Using Cross-View Spatial Feature Fusion for 3D Object Detection[C]// Proceedings of 16th European Conference on Computer Vision (ECCV), 2020: 720-736.

[23] ZHENG W, TANG W, CHEN S, et al. CIA-SSD: Confident IoU-aware Single-Stage Object Detector from Point Cloud [C]//Proceedings of the AAAI conference on artificial intelligence. 2021, 35(4): 3555-3562.

[24] SHI S, GUO C, JIANG L, et al. PV-RCNN: Point-voxel Feature Set Abstraction for 3D Object Detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10529-10538.

[25] CHEN X, MA H, WAN J, et al. Multi-View 3D Object Detection Network for Autonomous Driving[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1907-1915.

[26] QI C R, LIU W, WU C, et al. Frustum Pointnets for 3D Object Detection from RGB-D Data[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 918-927.

[27] LIANG M, YANG B, WANG S, et al. Deep Continuous Fusion for Multi-sensor 3D Object Detection[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 641-656.

[28] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal Loss for Dense Object Detection[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 2980-2988.

[29] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[J]. International Conference for Learning Representations, 2014, 21(12): 6980-6995.

[30] GEIGER A, LENZ P, STILLER C, et al. Vision Meets Robotics: The KITTI Dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

责任编辑：曾钰婵