智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

形状补全引导的Transformer点云目标检测方法

周静, 胡怡宇, 黄心汉

引用本文:
周静,胡怡宇,黄心汉. 形状补全引导的Transformer点云目标检测方法[J]. 智能系统学报, 2023, 18(4): 731–742.
ZHOU Jing,HU Yiyu,HUANG Xinhan. Shape completion–guided Transformer point cloud object detection method[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(4): 731–742.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202210038

DOI: 10.11992/tis.202210038
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20230323.0854.002.html

形状补全引导的 Transformer 点云目标检测方法

周静[1], 胡怡宇[1], 黄心汉[2]
（1. 江汉大学 人工智能学院, 湖北 武汉 430056; 2. 华中科技大学 人工智能与自动化学院, 湖北 武汉 430074）

摘　要: 针对雷达传感器采集到的场景点云中存在大量远距离或位于遮挡视角的形状缺失的低质量目标, 其几何信息不足难以被识别, 影响检测精度的问题, 本文提出一种基于形状补全引导的 Transformer 点云目标检测方法 (shape completion-guided transformer point cloud object detection method, STDet), 通过增强低质量目标形状特征来有效提升目标检测精度, 利用 Pointformer 主干网络提取场景点云特征以生成初始候选框, 基于特征分离预测的形状补全模块重构候选框中残缺目标的完整形状点云; 构建 Transformer 几何特征增强模型, 融合目标完整形状信息及空间位置信息至各目标点特征中, 并感知各目标点不同邻域掩码范围内的局部结构信息与全局几何特征的注意力相关性, 以获取关键几何信息增强的目标全局几何特征; 基于该特征引导生成精细化的目标检测框。在 KITTI 数据集上的实验结果表明, 该方法在存在大量形状残缺低质量目标的困难场景中检测精度较基准算法提升了 4.96%, 大量消融实验证明了该方法所构建的形状补全算法和 Transformer 几何特征增强模型的有效性。

关键词: 3D 目标检测; 低质量目标; 特征分离; 形状补全; Transformer; 多尺度; 邻域掩码; 特征增强
中图分类号: TP391.41　　**文献标志码:** A　　**文章编号:** 1673−4785(2023)04−0731−12

中文引用格式: 周静, 胡怡宇, 黄心汉. 形状补全引导的 Transformer 点云目标检测方法 [J]. 智能系统学报, 2023, 18(4): 731–742.
英文引用格式: ZHOU Jing, HU Yiyu, HUANG Xinhan. Shape completion-guided Transformer point cloud object detection method[J]. CAAI transactions on intelligent systems, 2023, 18(4): 731–742.

Shape completion-guided Transformer point cloud object detection method

ZHOU Jing[1], HU Yiyu[1], HUANG Xinhan[2]
(1. School of Artificial Intelligence, Jianghan University, Wuhan 430056, China; 2. School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: Aiming at the problem that in the point cloud of scenes collected by the LIDAR sensor, there are lots of low-quality objects with missing shapes due to long distance or occlusion, whose geometric information are too insufficient to be recognized, so that the detection accuracy is affected. Hence, a shape completion-guided Transformer point cloud object detection method (STDet) is proposed to improve the object detection precision by enhancing shape features of the low-quality objects. The features of the point clouds are acquired by the Pointformer backbone network to generate the initial candidate box. Then, the shape completion module predicted based on feature separation is designed to reconstruct a complete shape of point clouds of the incomplete objects within the candidate box. A Transformer geometric feature enhancement module is established, which integrates the complete shape information and spatial location knowledge of the object into its point-wise feature to perceive the attention correlation between the local structure information and the global geometric features within different neighborhood masks, so as to acquire the global geometric feature with enhanced critical geometric knowledge of the objects. Finally, the refined object detection boxes are generated under the guidance of global geometric features. Experimental results on KITTI data set show that compared with the benchmark algorithm, the proposed method improves detection accuracy by 4.96% in scenes with abundant low-quality objects of incomplete shapes. Meanwhile, the effectiveness of the proposed shape completion algorithm and Transformer geometric feature encoding module is proved by extensive ablation experiments.

Keywords: 3D object detection; low-quality object; feature separation; shape completion; Transformer; multi-scale; neighboring mask; feature enhancement

近年来, 随着人工智能和深度学习技术的兴起, 带动自动驾驶、智能机器人等领域的飞速发展。对于以感知三维空间环境信息为基础的自动

收稿日期: 2022−10−29.　**网络出版日期:** 2023−03−23.
基金项目: 国家自然科学基金项目(62106086); 湖北省自然科学基金项目(2021CFB564).
通信作者: 周静. E-mail: zhj131@jhun.edu.cn.

驾驶系统，能够有效感知和检测环境物体的 3D 目标检测技术对其至关重要。但由于三维环境的复杂性及数据采集设备固有的不平衡性，3D 物体检测技术的准确性和实时性受到限制，因此，研究 3D 目标检测技术非常具有挑战性。

目前，由于激光雷达设备扫描获得的 3D 点云数据受环境影响较小且保留了三维环境空间中原始的几何信息，以此更容易学到三维物体的空间位置和姿态信息[1]。因此，基于 3D 点云的目标检测方法获得了业界广泛关注。

考虑到点云数据的不规则性，很多研究工作如文献 [2-4] 通过多视图投影对点云数据进行预处理，然后采用成熟的 2D 目标检测技术预测出目标的二维空间位置信息，最后通过反投影估计目标的高度及深度信息以生成 3D 边界框。然而，不恰当的视图融合方式不利于目标检测任务，且将点云投影至二维视图会损失三维深度信息，因此，另一类研究工作[5-7] 采用基于体素的方法直接将原始点云空间划分为规则三维体素网格并用规则的卷积操作提取体素特征，以更好地保留三维深度信息；然而单一尺度的体素特征蕴含的目标空间结构信息较为粗糙，特征表达能力有限，影响检测性能，因此，文献 [8-9] 构建多尺度体素特征以学习目标丰富的空间信息，文献 [10-11] 提出将粗粒度的体素特征表示转化为细粒度的点特征，以充分利用点特征中细粒度的 3D 结构信息来增强粗粒度体素特征表示，提高检测精度。

上述基于体素的方法能充分学习体素表示下物体的三维几何信息，但是量化点云数据为三维体素网格不可避免地会模糊目标的边界信息。因此，直接学习点云数据而无需对点云进行体素化的基于点的方法受到了广泛关注。典型的基于点的方法如点区域卷积网络 (point regions convolutional neutral network, PointRCNN)[12] 采用 PointNet++提取丰富的 3D 逐点特征以生成鲁棒的 3D 候选框，然后基于框内局部点云细化生成 3D 目标检测框。然而，目标的前景点在场景点云中占比较少，而 PointNet++平等地采样前景点和背景点，使得仅有少量的前景点被采样以用于聚合特征，从而限制目标特征表达，影响候选框的生成。为此，文献 [13-14] 改进了 PointNet++中的点采样策略以提升采样中心点中前景目标点的占比；文献 [15] 通过引入图神经网络来增强目标前景点的特征表达。这些方法通过增强前景点的占比和特征能捕获更丰富的目标局部几何信息，但忽略了目标点云中不同的点对于检测任务的贡献不同，导致目标的关键信息未能获得有效关注，影响了生成的检测框的准确性。由于 Transformer 结构擅于捕捉目标全局上下文关联并能增强目标关键信息注意力，且具有置换不变性，适用于直接编码原生的点云数据，因此，许多最新基于点的方法结合 Transformer 结构进一步增强目标的关键特征的表达能力。如文献 [16] 结合 Transformer 编码各点间的全局信息，以捕捉目标的关键几何信息迭代地细化候选框参数；文献 [17] 结合 PointNet++的归纳偏置属性与 Transformer 结构构建鲁棒的点特征提取主干网络，以捕获目标局部邻域内的结构关联，增强目标的局部关键结构特征表达；文献 [18] 提出 Pointformer 方法，基于 Transformer 探索场景及目标点云间的全局上下文相关性，以加权目标重要结构细节，增强目标的关键几何特征的注意力，生成准确的目标检测框。虽然上述方法在大量点云场景中取得了不错的检测性能，但是随着传感器的距离及采集角度的变化，各目标的点云密度分布差异较大。距离传感器较远或是位于遮挡视角下的目标通常具有稀少的点云数量和残缺的几何形状，难以提供充足的几何信息以被准确感知和定位，而前述方法忽略了这一数据层面上点云分布不均的问题，导致其在存在大量遮挡或远距离的低质量目标的三维复杂场景中检测精度受限。

针对这一问题，同时考虑到基于点的方法能够有效保留原生点云的几何信息，本文采用基于点的方法作为主干网络，结合补全算法与擅于捕捉目标点云全局上下文信息的 Transformer 模型，构建形状补全引导的 Transformer 点云目标检测方法 STDet。首先，以 Pointformer 作为特征提取主干网络生成初始候选框；然后，基于形状补全（shape completion, SC）模块重构候选框中低质量目标点云的形状点集，以恢复有益于检测任务的目标几何形状信息；并构建 Transformer 几何特征增强模型（transformer geometric feature enhancement, TGE）以结合 Transformer 模型的全局相关性学习机制与卷积操作的归纳偏置属性，融合候选框内补全形状信息和原始空间位置信息至各目标点特征中，并建模各目标点的局部结构特征和全局几何特征间的相关性，增强目标的关键几何特征以提高检测精度。在 KITTI 数据集上的实验结果表明，本方法能显著提升在具有大量远距离或遮挡目标的复杂场景中的检测精度。

1　总体检测框架

本文提出的点云目标检测方法 STDet 的整体结构框架如图 1 所示，其主要包括 3 个部分：1）候选框生成网络（proposal generation network, PGN）；

2)形状补全模块 SC; 3)Transformer 几何特征增强模块 TGE。对于给定的场景点云, STDet 在第 1 阶段基于 PGN 网络提取场景点云特征并以自下而上的方式生成初始目标候选框; 然后在第 2 阶段针对初始候选框中形状缺失的低质量目标, 通过形状补全模块 SC 重构目标的完整表面形状点集, 并基于多尺度邻域掩码 Transformer(multi-scale neighboring mask transformer, MMT)模型构建 Transformer 几何特征增强模块 TGE, 以基于目标的原始空间位置信息和恢复的形状点集对初始候选框进行细化。TGE 模块首先融合目标重构的完整几何形状信息与原始空间位置信息, 以得到具有丰富几何信息的区域形状特征, 然后将其送入至多尺度邻域掩码 Transformer 模型中进行编码细化以增强有益于框位置回归的关键几何信息, 得到精细化几何特征。最后基于该特征进行置信度分类预测和框位置回归计算, 以生成准确的目标检测框。

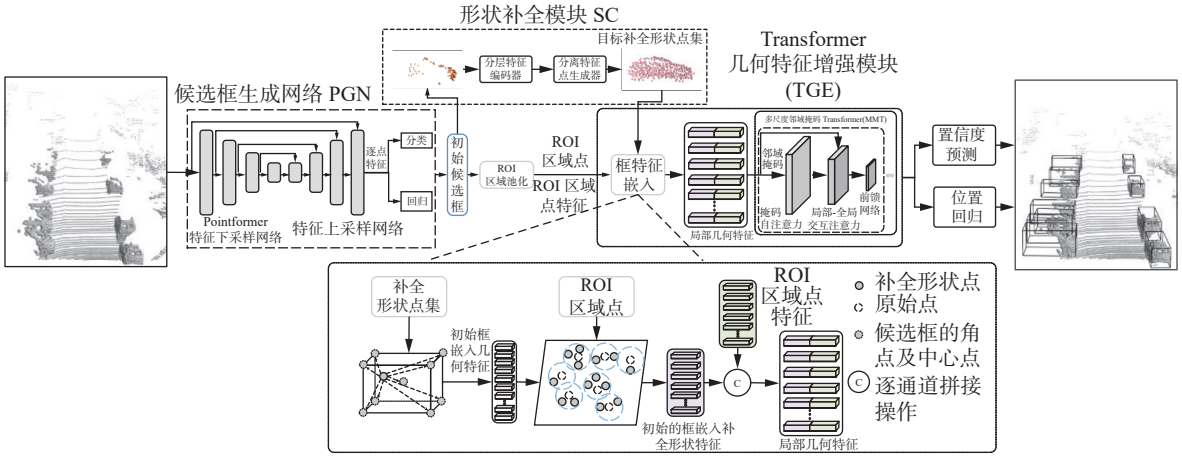


图 1　STDet 检测方法整体结构

Fig. 1　Overall structure of STDet object detection method

2　形状补全引导 Transformer 目标检测

2.1　PGN 网络

Pointformer 是性能优异的基于 Transformer 结构的点云目标检测主干网络, 由局部 Transformer、全局 Transformer 及局部−全局 Transformer 等模块组成。对于输入的场景点云, 首先由局部 Transformer 对其进行下采样获得各局部区域中心点, 建模目标局部区域中各点间的依赖关系; 然后, 由局部−全局 Transformer 模块聚合采样得到的低分辨率中心点集的局部特征和未采样的高分辨率点集的全局特征, 以捕捉不同尺度目标点集间的局部与全局特征语义相关性, 增强场景中目标关键区域的注意力; 再由全局 Transformer 学习场景点云中全局上下文表示, 捕获各目标点集间的几何相似性, 以准确感知目标的空间位置信息。经过堆叠的 Pointformer 模块学习目标高层级空间语义特征, 再由反向插值操作将各中心点特征传播给局部区域中的非中心点, 恢复目标点云分辨率, 以逐点生成目标检测框。

Pointformer 能有效建模点云数据间的长程依赖关系, 并捕捉目标局部区域和全局场景点云之间的空间上下文相关性, 增强目标关键特征的注意力, 从而提升检测网络的特征学习能力, 有利于 3D 检测框的生成。因此, 本文基于 Pointformer 主干网络在第 1 阶段构建候选框生成模块, 以感知目标的空间位置信息, 生成目标初始候选框。

2.2　基于分离特征的形状补全模块

对于第 1 阶段生成的候选框, 通过非极大值抑制方法保留高置信度候选框以进行第 2 阶段细化。由于场景中存在形状残缺、点云稀少的低质量目标, 其缺乏充足的几何信息以细化对应的候选框的参数, 导致难以生成准确的检测框。因此, 本文提出在第 2 阶段细化网络中融合点云形状补全任务来恢复候选框中低质量目标的完整形状点集, 以增强目标的几何信息, 优化检测框的生成。

通常的补全方法基于局部特征恢复目标细节, 并借助全局特征恢复目标整体形状, 然而其所提取的全局和局部特征中缺乏明确的缺失部分特征线索, 使得预测生成的缺失部分点云易与输入的已知部分点集发生混淆, 影响重构点云的形状合理性。因此, 本文构建一种基于点云特征分离的形状补全模块, 通过分离不同层次的局部特征与全局特征得到残差特征, 以此预测包含目标缺失区域点集的形状点云, 从而生成合理的目标完整形状点集。

对于场景点云中形状残缺的目标, 首先基于其对应的候选框参数在场景点云中划分出相应区

域，然后在区域中采集固定大小的点集$\boldsymbol{X}_{\text{input}} \in \mathbf{R}^{N_0\times 3}$输入至形状补全模块 SC 中进行形状恢复。如图 2 所示，SC 模块由分层特征编码器和基于分离特征的点生成器构成。

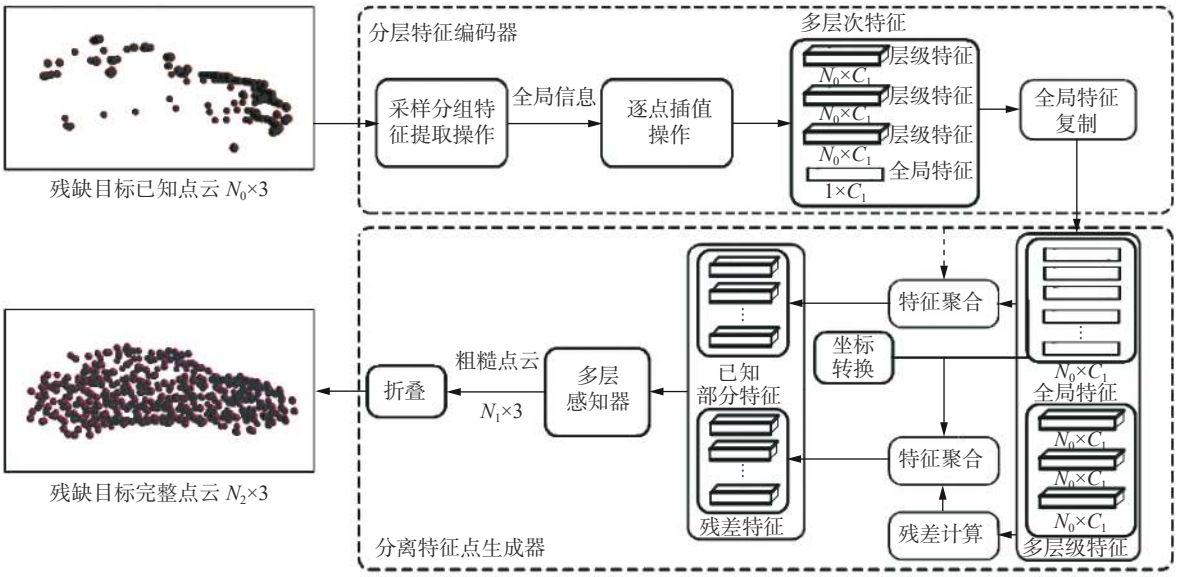


图 2　基于分离特征的形状补全模块

Fig. 2　Shape completion module based on the separated features

2.2.1　分层特征编码器

对于候选框中形状残缺的目标点集$\boldsymbol{X}_{\text{input}}$，分层特征编码器采用改进的 PointNet++网络对其进行编码，以获取多层级局部特征，并迭代地融合不同层次的局部信息以获得蕴含多重语义信息的全局特征。

如图 3 所示，在分层特征编码器中，首先基于最远点采样 (farthest point sampling, FPS) 和特征分组操作逐点学习目标点集$\boldsymbol{X}_{\text{input}}$的特征，得到不同分辨率的中心点集及不同层级的局部特征，再对局部特征进行最大池化操作以提取出不同层级的全局信息，并将其聚合得到全局特征。然后融合该全局特征与各层级的局部特征，并进行反向插值得到多层级局部特征$\{\boldsymbol{F}^1_{\text{level}}, \boldsymbol{F}^2_{\text{level}}, \boldsymbol{F}^3_{\text{level}}, \boldsymbol{F}^i_{\text{level}} \in \mathbf{R}^{N_0\times C_1}, i=1,2,3\}$，其中 N_0 为点云的点数，C_1 表示多层级局部特征的特征维度，并由最后一层的特征提取得到全局特征$\boldsymbol{F}_{\text{global}} \in \mathbf{R}^{1\times C_1}$，其融合了多层级局部结构细节与整体形状信息。

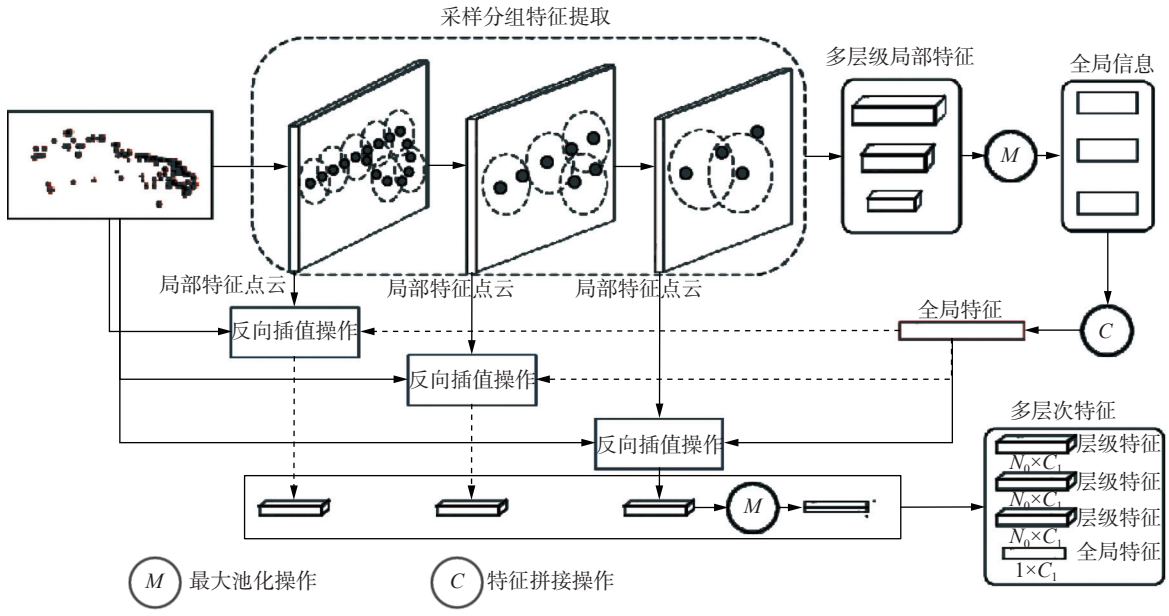


图 3　分层特征编码器结构示意

Fig. 3　Structure diagram of the hierarchical feature encoder

2.2.2 基于分离特征的点生成器

上述多层次局部特征和全局特征保留了目标已知部分点云的整体结构及局部细节, 但并未捕获明确的目标缺失部分点云的几何信息。因此, 在基于分离特征的点生成器中, 通过分离已知部分的局部特征与表征整体形状信息的全局特征来获得残差特征, 并基于残差信息重构生成包含缺失区域的目标粗糙形状点云。首先, 由分层特征编码器获取的多层次局部特征和全局特征构建目标已知部分特征 $\boldsymbol{F}_{\text{know}} \in \mathbf{R}^{N_0 \times C_2}$; 其次, 从全局特征中分离出局部特征得到残差特征 $\{\boldsymbol{F}^1_{\text{res}}, \boldsymbol{F}^2_{\text{res}}, \boldsymbol{F}^3_{\text{res}} | \boldsymbol{F}^i_{\text{res}} \in \mathbf{R}^{N_0 \times C_1}\}$, 再由残差特征构成包含目标缺失部分信息的特征 $\boldsymbol{F}_{\text{missing}} \in \mathbf{R}^{N_0 \times C_2}$。

$$\boldsymbol{F}_{\text{know}} = \text{cat}\left(\boldsymbol{F}^1_{\text{level}}, \boldsymbol{F}^2_{\text{level}}, \boldsymbol{F}^3_{\text{liwal}}, \boldsymbol{X}_{\text{input}}, \rho(\boldsymbol{F}_{\text{global}})\right) \tag{1}$$

$$\boldsymbol{F}^i_{\text{res}} = \varphi\left(\rho(\boldsymbol{F}_{\text{global}}) - \boldsymbol{F}^i_{\text{level}}\right), i = 1,2,3 \tag{2}$$

$$\boldsymbol{F}_{\text{mising}} = \text{cat}\left(\boldsymbol{F}^1_{\text{res},}, \boldsymbol{F}^2_{\text{res}}, \boldsymbol{F}^3_{\text{ies}}, \varphi(\boldsymbol{X}_{\text{input}})\rho(\boldsymbol{F}_{\text{global}})\right) \tag{3}$$

式中: $\varphi(\cdot)$为多层感知器, $\rho(\cdot)$为扩展特征维度的复制操作, cat(·)为在特征通道维度上的拼接操作。

为基于残差信息和多层级局部特征生成包含缺失部分点云的目标完整形状点集, 将残缺目标点云的已知部分特征 $\boldsymbol{F}_{\text{know}}$ 和包含目标缺失部分信息的特征 $\boldsymbol{F}_{\text{missing}}$ 分别复制 B_1 和 B_2 次, 并通过多层感知器将其投影扩展至高维空间, 以重构形成具有 $N_1(N_1=(B_1+B_2)\times N_0)$ 个点的粗略的目标形状补全点集 $\boldsymbol{X}_{\text{coarse}} \in \mathbf{R}^{N_1 \times 3}$。为进一步细化生成密集的完整形状点集, 以 $\boldsymbol{X}_{\text{coarse}}$ 中每个点为中心, 通过折叠操作, 在各中心点的局部邻域内生成 B_3 个点, 得到包含 $N_2(N_2=N_1\times(B_3+1))$ 个点的密集的完整形状点集 $\boldsymbol{X}_{\text{all}} \in \mathbf{R}^{N_2 \times 3}$。

2.3 Transformer 几何特征增强模型

为利用由 SC 模块生成的目标完整形状点集来增强目标的几何特征, 本文提出一种 Transformer 几何特征增强模型 TGE, 其首先融合 SC 模块重构的完整形状信息与候选框内的原始空间上下文信息, 然后构建多尺度邻域掩码 Transformer 以捕获不同掩码范围内的点集的局部结构特征与全局几何特征的相关性, 细化各点的空间几何信息, 得到候选框中关键信息增强的目标几何特征, 用以细化候选框。

2.3.1 几何信息嵌入

由于三维候选框中各点的相对位置信息有利于确定目标边界, 细化候选框的位置参数, 因此, 本文将初始候选框与框内补全重构的完整形状点集之间的相对位置信息嵌入至该点集的各点特征中, 以得到目标的补全形状特征。

对于候选框及框内的完整形状点集 $\boldsymbol{X}_{\text{all}}$, 首先计算 $\boldsymbol{X}_{\text{all}}$ 中各点与对应候选框的 8 个角点之间的相对距离, 对于候选框中任一点 $\boldsymbol{p}_i \in \boldsymbol{X}_{\text{all}}$, 其与候选框的第 j 个角点 $\bar{p}_j$ 之间的距离为 $\boldsymbol{\mu}^j_i = \bar{\boldsymbol{p}}_j - \boldsymbol{p}_i$; $\boldsymbol{p}_i$ 与候选框的中心点 $\bar{\boldsymbol{p}}_0$ 之间的距离为 $\boldsymbol{\mu}^0_i = \bar{\boldsymbol{p}}_0 - \boldsymbol{p}_i$。随后将相对距离 $\boldsymbol{\mu}^j_i$ 和 $\boldsymbol{\mu}^0_i$ 通过线性层 ϕ 嵌入至形状补全点集以获得补全形状特征 $\boldsymbol{G}' = \left\{\boldsymbol{g}'_1, \boldsymbol{g}'_2, \cdots, \boldsymbol{g}'_{N_2}\right\} \in \mathbf{R}^{N_2 \times D}$:

$$\boldsymbol{g}'_i = \phi\left(\boldsymbol{\mu}^0_i, \boldsymbol{\mu}^1_i, \cdots, \boldsymbol{\mu}^8_i\right), i = 1,2,\cdots,N_2 \tag{4}$$

同时, 根据候选框参数对原始点云和由 PGN 模块生成的逐点特征进行感兴趣区域池化操作, 得到框内目标的区域点集 $\boldsymbol{X}_{\text{RoI}} \in \mathbf{R}^{N\times 3}$ 和蕴含原始空间位置信息的区域点特征, 以 $\boldsymbol{X}_{\text{RoI}}$ 为中心在给定半径范围内聚集补全形状特征 $\boldsymbol{G}'$, 得到局部形状特征, 将其与区域点特征拼接得到区域几何特征 $\boldsymbol{G}^{(l)} \in \mathbf{R}^{N\times D}$, 其融合了候选框内目标的完整形状信息及原始空间位置信息。将区域几何特征 $\boldsymbol{G}^{(1)}$ 及对应的区域点集 $\boldsymbol{X}_{\text{RoI}}$ 送入至多尺度邻域掩码 Transformer 模型中, 以进一步学习框内各目标点的局部结构特征和全局几何特征间的相关性, 细化区域几何特征, 增强目标关键几何信息。

2.3.2 多尺度邻域掩码 Transformer

多尺度邻域掩码 Transformer 模型 MMT 的结构如图 4 所示, 其由邻域掩码自注意力模块和几何感知交叉注意力模块构成。其中, 邻域掩码自注意力模块结合局部特征学习机制中的归纳偏置属性和自注意力机制, 以建模不同邻域点集范围内几何特征间的结构关联, 提取目标局部区域内重要的结构特征。几何感知交叉注意力模块通过编码不同尺度全局几何特征和局部结构特征间的上下文相关性, 探索框内区域点集中各点的注意力权重, 进一步增强目标的关键几何特征。

对于第 s 层的区域点集 $\boldsymbol{X}_{\text{RoI}}$ 和输入特征 $\boldsymbol{G}^{(s)} \in \mathbf{R}^{N\times D}$, 多尺度邻域掩码 Transformer 首先采用邻域掩码自注意力模块进行掩码自注意力计算。即, 首先通过邻域分组操作在区域点集 $\boldsymbol{X}_{\text{RoI}}$ 中划分出各点邻域半径 $r^{(s)}$ 内的点, 若第 i 个点和第 j 个点的坐标均在彼此半径 $r^{(s)}$ 的范围内, 则将二进制邻域掩码值 $\boldsymbol{M}^{(1)}_{i,j}$ 赋为 1, 反之为 0, 由此构成大小为 $N\times N$ 的二进制邻域掩码矩阵 $\boldsymbol{M}^{(s)}$:

$$\boldsymbol{M}^{(s)}_{i,j} = \begin{cases} 1, & d(ij) \leqslant r^{(s)} \\ 0, & d(ij) > r^{(s)} \end{cases} \tag{5}$$

式中 $d(ij)$ 为 $\boldsymbol{X}_{\text{RoI}}$ 中第 i 个点和第 j 个点的距离。

然后, 基于邻域掩码矩阵对输入特征 $\boldsymbol{G}^{(s)}$ 进行掩码自注意力计算, 以学习其邻域内的局部结构信息:

$$\boldsymbol{G}^{(s+1)}_{\text{local}} = \sum_{h=1}^{H} \delta\left(\eta\left(\frac{\left(\boldsymbol{G}^{(s)}\tilde{\boldsymbol{Q}}^{(s)}_h\right)\left(\boldsymbol{G}^{(s)}\tilde{\boldsymbol{K}}^{(s)}_h\right)^{\mathrm{T}}}{D/H}, \boldsymbol{M}^{(s)}\right)\right)\left(\boldsymbol{G}^{(s)}\tilde{\boldsymbol{V}}^{(s)}_h\right) \tag{6}$$

式中: $\delta(\cdot)$ 为 softmax 归一化函数。即首先通过线

性神经网络实现 3 个大小均为 $D\times(D/H)$ 的线性投影矩阵$\tilde{\boldsymbol{Q}}_h^{(s)}$、$\tilde{\boldsymbol{K}}_h^{(s)}$和$\tilde{\boldsymbol{V}}_h^{(s)}$，并将其分别与输入特征$\boldsymbol{G}^{(s)}$相乘，以将$\boldsymbol{G}^{(s)}$映射至不同的注意力子空间中，得到相应的查询向量、关键向量和值向量，以此在不同的注意力子空间中查询出目标的关键几何信息。随后，将查询向量和关键向量相乘，以计算各点间的特征相关性，获取自注意力权重。同时为保留目标各点邻域范围内的局部结构关联，基于邻域掩码矩阵$\boldsymbol{M}^{(s)}$对自注意力权重执行掩码填充操作$\eta(\cdot)$，使得在各点邻域范围内的权重值不变，而在邻域范围以外的权重值被赋为负无穷，然后，将掩码填充后的自注意力权重矩阵与对应的值向量相乘，以对各注意力子空间中的关键邻域几何信息进行注意力加权，得到局部几何注意力特征，随后聚合 H 个注意力子空间中的局部几何注意力特征，得到关键结构信息增强的局部区域特征$\boldsymbol{G}_{\text{local}}^{(s+1)}\in\mathbf{R}^{N\times D}$，其蕴含了目标重要的局部结构和几何位置信息。

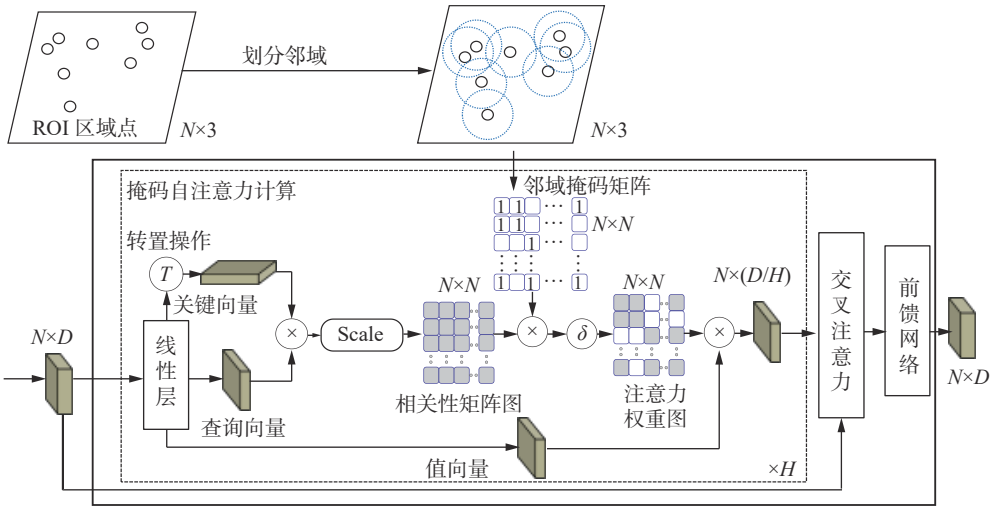


图 4 多尺度邻域掩码 Transformer 模块

Fig. 4 Multi-scale neighboring mask Transformer module

为进一步建模目标点云的局部结构特征与全局几何特征间的交互关联，准确感知目标的几何边界，构建几何感知交叉注意力模块将上一层 MMT 模型输出的全局几何特征$\boldsymbol{G}^{(s)}$聚合至本层多尺度邻域掩码对应的局部结构特征$\boldsymbol{G}_{\text{local}}^{(s+1)}$中，使得网络能够有效学习感知目标点云的多层级空间几何信息并加权目标关键几何特征。具体地，分别使用线性神经网络实现 3 个线性投影矩阵$\hat{\boldsymbol{Q}}_h^{(s)}$、$\hat{\boldsymbol{K}}_h^{(s)}$、$\hat{\boldsymbol{V}}_h^{(s)}\in\mathbf{R}^{D\times(D/H)}$，并将线性矩阵$\hat{\boldsymbol{Q}}_h^{(s)}$与本层局部结构特征$\boldsymbol{G}_{\text{local}}^{(s+1)}$相乘，以将$\boldsymbol{G}_{\text{local}}^{(s+1)}$投影至对应的注意力空间中得到查询向量，同时分别将线性矩阵$\hat{\boldsymbol{K}}_h^{(s)}$和$\hat{\boldsymbol{V}}_i^{(s)}$与上一层的全局几何特征$\boldsymbol{G}^{(s)}$相乘，以投影得到相应的关键向量和值向量。然后，基于查询向量、关键向量和值向量进行几何感知交叉注意力计算，以学习目标的多尺度邻域掩码对应的局部结构特征与全局几何特征间的相关性权重，加权目标关键的全局几何信息，感知目标的整体几何边界，并使用前馈神经网络 (feed forward network, FFN) 进行更新，最终得到蕴含全局几何关联的注意力加权几何特征$\boldsymbol{G}^{(s+1)}\in\mathbf{R}^{N\times D}$：

$$\boldsymbol{G}^{(s+1)}=\text{FFN}\left(\sum_{h=1}^{H}\left(\delta\left(\left(\boldsymbol{G}_{\text{local}}^{(s+1)}\hat{\boldsymbol{Q}}_h^{(s)}\right)\left(\boldsymbol{G}^{(s)}\hat{\boldsymbol{K}}_h^{(s)}\right)^{\mathrm{T}}/(D/H)\right)\right)\left(\boldsymbol{G}^{(s)}\hat{\boldsymbol{V}}_h^{(s)}\right)\right) \tag{7}$$

将$\boldsymbol{G}^{(s+1)}$与区域点集$\boldsymbol{X}_{\text{RoI}}$输入至下一层的 MMT 模型进一步进行细化编码。通过 s 个 MMT 模型逐层细化各目标点的几何形状信息，重新加权关键几何特征，编码得到关键几何信息增强的精细化目标全局几何特征。将该特征送入至检测头进行置信度预测和位置回归，生成微调后的目标检测框及其类别信息。

3 损失函数

STDet 以端到端的方式进行训练，损失函数由第 1 阶段候选框生成损失L_{rpn}和第 2 阶段候选框细化损失L_{box}组成：

$$L=L_{\text{rpn}}+L_{\text{box}} \tag{8}$$

候选框生成损失由基于点的分类损失和候选框回归损失组成：

$$L_{\text{rpn}}=L_{\text{seg}}+\frac{1}{\beta}\sum L_{\text{reg}}^{a} \tag{9}$$

式中β为场景点云中前景点的数目。

由于复杂场景中的前景点与背景点的分布不均衡，前景点过少，为了平衡 2 类样本点，同时增强网络对困难样本点的关注以优化训练，本文采用 focal-loss 作为分类损失：

$$L_{\text{seg}}(\lambda) = -\alpha(1-\lambda)^{\gamma}\log(\lambda)$$
$$\lambda = \begin{cases} \bar{\lambda}, & \text{前景点的置信度} \\ 1-\bar{\lambda}, & \text{背景点的置信度} \end{cases} \tag{10}$$

式中：γ 因子减少网络对易分样本的损失；平衡因子α调节正负样本的重要性，训练时设置α=0.25，γ=2；$\bar{\lambda}$为样本点的置信度。

候选框回归损失为

$$L_{\text{reg}}^{j} = \sum_{u\in\{x,y,z,l,w,h,\theta\}} L_{\text{smooth}-L_1}\left(\hat{e}_j^u, e_j^u\right) \tag{11}$$

式中：回归损失函数采用 smooth-L_1 实现；$\hat{e}_j^u$为网络预测的第 j 个前景点与对应的候选框中心点之间的残差；e_j^u为回归目标。

候选框细化阶段的总损失由置信度损失和回归损失组成：

$$L_{\text{refine}} = L_{\text{cls}} + \frac{1}{N_{\text{box}}}\sum_{i=1}^{N_{\text{box}}}\sum_{u\in\{x,y,z,l,w,h,\theta\}} L_{\text{smooth}-L_1}\left(\hat{t}_j^u, t_j^u\right) \tag{12}$$

式中：置信度损失L_{cls}为交叉熵损失；回归损失采用 smooth-L_1 实现；N_{box}为前景框个数；t_i^u为一阶段第 i 个前景候选框与真实目标框之间的残差；作为回归目标，$\hat{t}_i^u$为第 i 个前景候选框和预测框间的残差。

4　实验分析

4.1　数据集及参数

KITTI 数据集作为 3D 目标检测基准数据集，提供了 7 481 个训练样本，其中 3 712 个样本被划分为训练集，3 769 个样本被划分为验证集，并提供了 7 518 个测试样本构成测试集；同时，根据场景中目标的遮挡、截断程度，KITTI 数据集中的目标样本被划分成“容易”、“中等”和“困难”3 种难度级别。本文采用 KITTI 数据集提供的训练集对 STDet 进行训练，并基于测试集和验证集分别对该模型进行评估与验证。

训练过程在 GeForce RTX 3 090 Ti 型号的 GPU 上进行，共训练 100 轮，学习率设为 0.002，训练批次大小为 6。同时，形状补全模型 SC 在 ShapeNet Car 数据集上的预训练轮数为 80，训练批次大小设为 16。同时，采用不同交并比（intersection over union, IoU）阈值下的平均检测精度 (average precision, AP) 值，以及 3 个难度级别的平均检测精度均值 (mean value of average precision, mAP) 作为网络检测性能的评估指标。

4.2　评估与分析

为验证 STDet 检测方法的有效性，分别在 KITTI 的验证集和测试集上对其最优训练模型进行了验证和测试，得到在 3 种难度级别样本中的检测精度，与其他主流的三维目标检测方法进行对比，结果如表 1、2 所示。对比方法包括基于体素的方法，如稀疏嵌入卷积检测 (sparsely embedded CONvolutional detection, SECOND) 网络[6]、点柱网络 PointPillars[7]、基于稀疏 Transformer 的点柱网络 (sparse Transformer based pointpillars, STPointpillars)[19]、三维交并比损失 (three-dimensional intersection over union loss, 3D IoU Loss) 网络[20]、基于体素化图卷积网络 (voxel-based graph convolution network, VGCN)[21] 的检测方法、Part-A^2 方法[11]、混合的点−体素表示 (hybrid voxel-point representation, HVPR)[22] 检测方法、点−体素区域卷积网络 (point-voxel region convolution neural network, PVRCNN)[23]、掩码引导注意力的无锚 3DSSD (mask-guided attention for anchor-free 3DSSD, MGAF-3DSSD) 方法[24] 等，以及基于点的方法 PointRCNN、Pointformer、点−图卷积神经网络 (point-graph neural network, PointGNN)[15] 和点−区域图卷积网络 (point region graph convolution network, PointRGCN)[25] 等。

表 1 给出了本文方法 STDet 在 KITTI 验证集上与其他方法在车辆目标类别上 3D 的 AP 值的对比结果。

表 1　KITTI 验证集上不同目标检测算法的 AP 值对比
Table 1　Comparison of AP values of different object detection algorithms on KITTI val split　%

检测方法	mAP	AP		
		简单	中等	困难
SECOND	81.48	88.61	78.2	77.22
PointPillars	79.76	87.50	77.01	74.77
3D IOU Loss	81.58	89.16	78.33	77.25
VGCN	82.34	89.25	79.21	78.58
Part-A^2	82.49	89.47	79.47	78.54
PVRCNN	83.91	89.35	83.69	78.70
STPointPillars	82.09	88.53	79.58	78.15
PointRGCN	81.50	88.37	78.54	77.60
PointRCNN	81.63	88.88	78.63	77.38
PointGNN	81.20	87.89	78.34	77.38
Pointformer	82.86	90.05	79.65	78.89
STDet (本文方法)	83.12	89.90	80.27	79.19

由表 1 可知，对比其他方法，STDet 取得了优异的检测性能。对于中等级别和困难级别目标，对比基准方法 Pointformer，STDet 的检测精度分别提升了 0.62% 和 0.3%。对比性能优异的基于体素的检测方法 Part-A^2 和 PVRCNN，STDet 在困难级别目标上分别取得了 0.65% 和 0.49% 的精度提升。

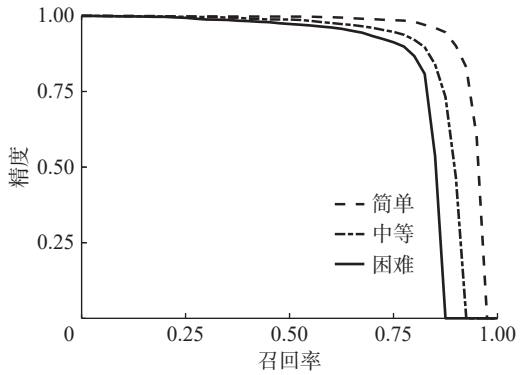
表 2 给出了本文提出的 STDet 在 KITTI 测试集上与其他检测方法在车辆目标类别上 3D 的 AP 值及推理时间的对比结果。

表 2　不同目标检测算法在 KITTI 测试集上 AP 值对比
Table 2　Comparison of AP values of different object detection algorithms on KITTI test split

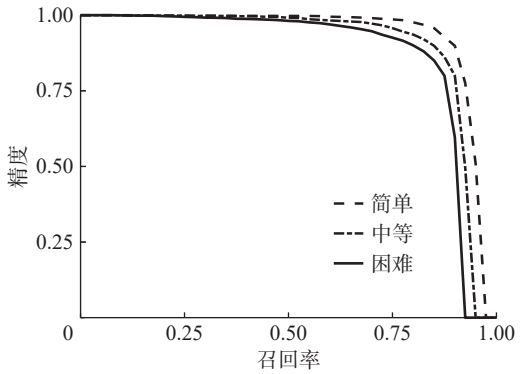
检测方法	AP/%			时间/s
	简单	中等	困难	
PointPillars	82.580	74.310	68.990	0.016
3D IOU Loss	86.160	76.500	71.390	0.080
MGAF-3DSSD	88.160	79.680	72.390	0.065
Part-A^2	87.810	78.490	73.510	0.083
HVPR	86.380	77.920	73.040	0.028
PointRGCN	85.970	75.730	70.600	0.260
PointGNN	88.330	79.470	72.290	0.600
PointRCNN	86.960	75.640	70.700	0.100
Pointformer	87.130	77.060	69.250	0.120
STDet (本文方法)	87.110	79.810	74.210	0.180

由表 2 可知，与基准方法 Pointformer 对比，本文方法在中等、困难难度级别目标样本上的检测精度分别提高了 2.75% 和 4.96%，可知本文方法在困难级别样本中精度提升显著。同时，对比其他方法，STDet 在中等和困难级别目标样本上也取得了最高的检测精度，对于困难级别的目标样本，与基于体素的方法 Part-A^2、3D IOU Loss、HVPR 和 MGAF-3DSSD 相比，STDet 取得了明显的精度提升；与基于点的检测方法 PointRCNN、PointGNN、Pointformer 和 PointRGCN 相比，STDet 至少能获得 1.92% 的提升。另外，由表 2 中给出的本文方法与其他检测方法在推理时间上的对比结果可知，增加的 SC 和 TGE 模块使得 STDet 相较基准方法的单帧推理时间略有增长。

图 5 给出 Pointformer 与 STDet 在 KITTI 验证集中鸟瞰视角下的精度−召回率曲线。



(a) 基准方法 Pointformer



(b) 本文方法 STDet

图 5　KITTI 验证集中鸟瞰视角下的 PR 曲线
Fig. 5　Bird-view PR curves on KITTI val split

由图 5 可知，STDet 在 3 种难度级别上的检测性能相比 Pointformer 均取得了较为明显的提升。

以上实验结果证明了本文方法 STDet 在 3 种难度级别样本中均有良好的表现，尤其是对于低质量困难级别目标样本，STDet 的检测精度 AP 相对基准方法获得了 4.96% 的显著提升，表现出优异的检测性能，充分证明了本文方法能有效提升具有大量低质量目标的复杂场景中的目标检测精度。

4.3　消融实验

为验证本文提出的网络中各模块的有效性，分别对形状补全模块 SC 和 Transformer 几何特征增强模块 TGE 进行消融实验。消融实验均在 KITTI 的训练集上进行训练，在验证集上进行验证。

4.3.1　SC 及 TGE 模块消融实验

在相同的第 1 阶段主干网络下，分别使用基于 SC 和 TGE 模块构建不同的第 2 阶段细化网络，所得到的检测精度如表 3 所示。

表 3　SC 模块和 TGE 模块的消融实验结果
Table 3　Ablation experimental results of SC and TGE modules　%

SC	TGE	PointNet++	AP		
			简单	中等	困难
×	×	√	89.76	79.46	78.59
√	×	√	89.80	79.85	78.86
×	√	×	89.85	79.94	78.91
√	√	×	89.90	80.27	79.19

表 3 中第 1 行数据为未使用 SC 和 TGE 模块的基准网络所得到的 AP 值，其仅采用 PointNet++ 网络细化第 1 阶段的特征；第 2 行数据是采用 PointNet++而非 TGE 模块来细化 SC 重构的形状点集得到的 AP 值；第 3 行数据是仅采用 TGE 模块直接细化第 1 阶段特征所得的 AP 值，最后一

行数据是本文提出的融合了 SC 模块和 TGE 模块的 STDet 网络的检测精度值。

由表 3 可知, 对比基准方法, 本文分别使用 TGE 模块或 SC 模块增强框中低质量目标特征时, 由于普通目标形状完整且包含点数较多, 在第 2 阶段采样过少的输入点进行特征细化时, 所得的采样点中能保留一定的目标形状信息但原始关键点占比明显下降, 因此仅恢复目标的完整形状使得检测精度在简单级别(普通目标)上无明显变化; 但中等或困难目标自身形状残缺, 点集分布稀疏, 在第 2 阶段采样过少的输入点时, 采样点数相对于其自身点数无较大差异, 即关键点占比基本不变但形状信息缺失严重, 因此, 使用 SC 模块或 TGE 模块对其进行形状补全或特征增强, 可使其检测精度有明显提升。

为进一步探讨 SC 模块和 TGE 模块对候选框中目标特征细化的有效性, 将 STDet 的主干网络替换为具有代表性的基于体素的方法 SECOND 和 Pointpillars, 即使用 SC 模块和 TGE 模块在第 2 阶段细化由 SECOND 或 Pointpillars 网络生成的候选框, 所得的检测精度如表 4 所示。

表 4　主干网络的消融实验

Table 4　Ablation experiment of the backbone network　%

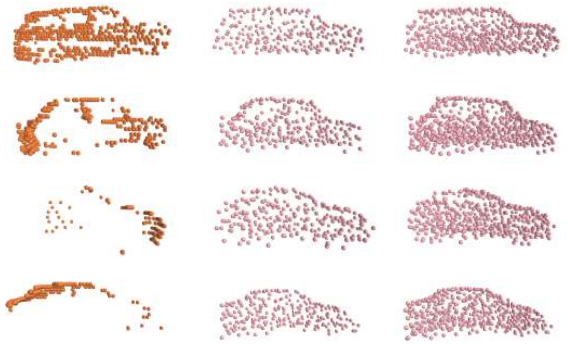
方法	AP		
	简单	中等	困难
SECOND	88.61	78.62	77.22
SECOND+SC+TGE	88.98	79.67	78.68
Pointpillars	87.50	77.01	74.77
Pointpillars +SC+TGE	88.26	79.10	77.18
本文方法	89.90	80.27	79.19

对比可知, 将 SC 和 TGE 模块增加至 SECOND 或 Pointpillars 网络中可分别在困难难度级别上获得 1.46% 和 2.41% 的检测精度提升, 以此证明对于基于体素的检测方法, 本文所设计的 SC 和 TGE 模块也能有效提升其检测精度。同时, 由表 4 中第 2、4、5 行的检测结果可知, 本文方法的检测精度仍高于增加了 SC 和 TGE 模块的 SECOND 及 Pointpillars 网络, 因此, 进一步证明了本文方法的有效性。

由上述分析可知, 使用 SC 和 TGE 模块确实能有效提升检测精度。为进一步深入分析 SC 模块对于检测性能的作用以及 TGE 模块中各成分的有效性, 本文分别针对 SC 和 TGE 模块进行了消融实验。

4.3.2　SC 模块有效性分析

图 6 给出了不同目标点数经过 SC 模块所得到的补全结果, 共包括 4 组可视化效果图, 图 6(a) 前 2 组表示包含 512 个点的残缺目标车辆, 图 6(a) 后 2 组表示包含 256 个点的残缺目标车辆, 图 6(b)、(e) 分别表示补全生成的形状点集经过下采样后得到的包含 256、512 个点的完整目标点云。对比可知, 图 6(a) 前 2 组图中的点数较多, 普通目标自身形状较为完整, 经过补全后得到的形状和原始轮廓无明显差异, 而图 6(a) 后 2 组图中点数较少为明显缺失了完整轮廓的残缺困难目标, 补全后生成的密集点集具有清晰完整的形状, 能为定位框提供充分的几何形状信息, 因此补全操作能有效增强困难目标的形状特征。



(a) 原始点　(b) 256 个补全点　(c) 512 个补全点

图 6　SC 模块补全结果

Fig. 6　Completion results of SC module

为进一步验证补全输出的形状点集蕴含的完整形状特征对检测性能的影响, 分别采样得到不同点数的原始点集, 然后基于原始点集聚合补全形状特征或原始空间上下文特征, 以得到不同点数的原始点集对应的空间几何特征, 并将其输入至 MMT 模块进行细化增强, 所得到的检测精度如表 5 所示。

表 5　SC 模块的消融实验

Table 5　Ablation experiment of SC module　%

点数	原始特征	补全特征	AP		
			简单	中等	困难
256	√	×	89.85	79.94	78.91
512	√	×	90.03	80.06	78.99
256	√	√	89.90	80.27	79.19
512	√	√	90.05	80.32	79.17

表 5 中第 1、2 行数据为不加补全模型时, 采用 256 或 512 个原始点作为 MMT 的输入时的检测精度, 第 3、4 行数据为增加补全模型后, 分别

使用 256 或 512 个原始点来聚集补全特征并输入至 MMT 进行特征增强后得到的检测精度。由表 5 中第 1、2 行可知，对于简单级别目标，对其采样 512 个点所得到的检测精度明显高于仅采样 256 个点。而使用补全模型后，表 5 中第 3、4 行所呈现的简单级别的目标检测精度没有明显提升且仍有一定差异，分析认为，简单级别目标大多为无明显遮挡且点数较多的普通目标，对于这类普通目标，当仅从中采样出较少的 256 个点进行特征细化时，目标关键点信息会被丢失，造成精度下降。同时，在对普通目标使用补全模型时，由于其采样点仍能较为均衡地表示一定的目标形状，所以补全操作的均衡性对简单级别目标作用不显著，并未带来明显的精度提升。另外，对比第 1、2 行中不同点数的中等和困难级别目标的检测精度差异，表 5 中第 3、4 行分别采样 256 或 512 个原始点来聚合补全特征作为 MMT 输入时，所得到的检测精度明显提升且不同点数下的精度差异明显减小。这是由于中等和困难级别目标点数稀少，形状缺失严重，因此补全操作可以通过恢复其完整形状点集来均衡其点数分布，使其在补全后对于不同的输入点数均能捕获丰富的几何信息，取得优异的检测精度。

同时，由于采用 512 个原始点时，MMT 运行时间过长，因此，本文选择 256 个原始采样点来聚合补全特征和框的位置信息以进行候选框细化，来有效提升目标检测精度。

4.3.3　TGE 模块有效性分析

为选定 TGE 模块中多尺度邻域掩码 Transformer(MMT) 最优的堆叠层数，针对不同层数进行对比实验，结果如图 7 所示，可知堆叠 3 层 MMT 进行几何特征细化可取得最优的检测精度。

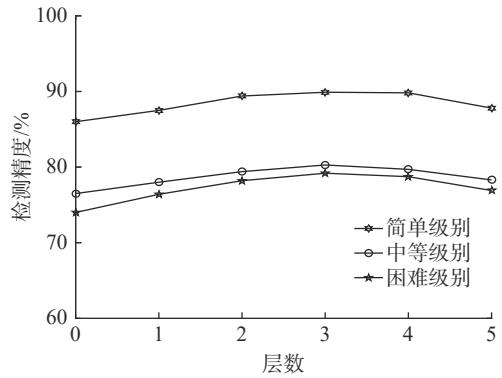


图 7　MMT 模块的层数对比实验结果

Fig. 7　Comparison experiment results for layer number of MMT module

为进一步验证本文设计的多尺度邻域掩码的有效性，对比分别采用单一尺度邻域掩码、多尺度邻域掩码以及无邻域掩码的 Transformer 模型在低质量困难级别样本上所获得的检测精度，如图 8 所示，可知无邻域掩码的 Transformer 模型在训练后期检测精度低于具有邻域掩码的模型，证明了邻域掩码机制的有效性。同时，相对于单一尺度邻域掩码模型，多尺度邻域掩码 Transformer 可充分建模不同尺度感受野的局部与全局信息，有效利用归纳偏置特性，取得了更高的检测精度。

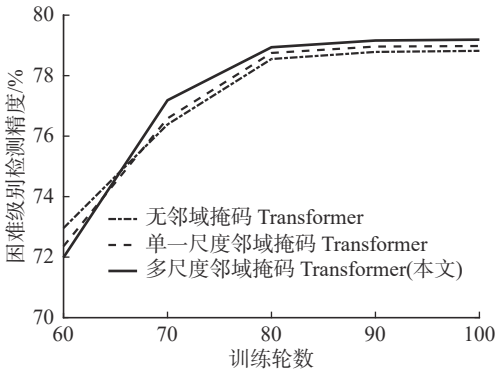


图 8　不同尺度邻域掩码的 Transformer 模型精度比较

Fig. 8　Comparison of detection accuracy of the Transformer models with different scale neighborhood masks

为验证 TGE 模块中候选框的几何信息嵌入对框细化的影响，在保留原始特征的前提下，分别聚合补全特征或嵌入了候选框信息的补全特征，并使用多尺度邻域掩码 Transformer 进行细化，所得的检测精度如表 6 所示。

表 6　候选框几何信息嵌入效果

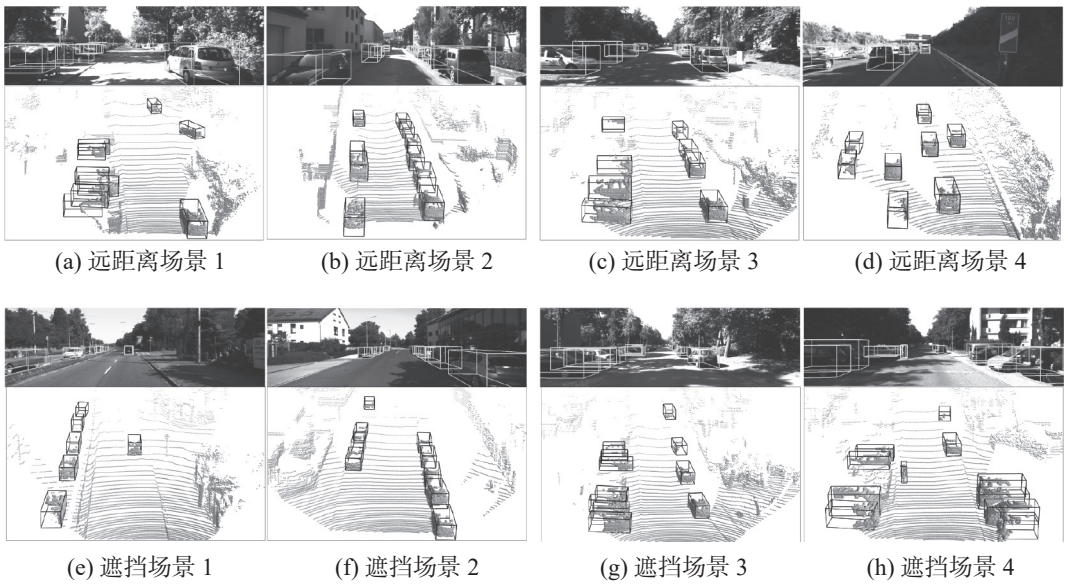
Table 6　Effect of embedding the box geometric information　%

框几何信息	AP(3D)		
	简单	中等	困难
×	89.84	80.09	79.01
√	89.90	80.27	79.19

由表 6 可知，将框的几何信息嵌入至补全特征后对检测精度提升较为明显，证明了候选框与框内点集的相对位置信息能有效辅助网络回归生成准确的目标检测框。

4.4　可视化定性分析

为进一步直观地验证本文方法的有效性，基于 STDet 最优的训练模型对 KITTI 验证集数据进行了测试并给出了 8 组可视化效果图，如图 9 所示，每一组可视化结果的上图是当前真实的道路场景图，其中的三维线框表示目标的真实检测框；下图表示在当前点云场景中的检测结果图，其中的三维框表示网络预测的目标框。



(a) 远距离场景 1　(b) 远距离场景 2　(c) 远距离场景 3　(d) 远距离场景 4

(e) 遮挡场景 1　(f) 遮挡场景 2　(g) 遮挡场景 3　(h) 遮挡场景 4

图 9　STDet 在复杂场景中检测结果可视化

Fig. 9　Visualizes detection results in complex scenes of STDet

从图 9 中可以看出, 在存在各种点云分布不均匀且稀疏的低质量目标的困难场景中, STDet 方法对位于不同距离及位置相互遮挡的目标, 均能生成准确的目标检测框, 具有一定的鲁棒性, 以此进一步证明了本文提出的 STDet 方法在复杂场景下有较好的检测性能。

5　结束语

本文提出一种形状补全引导的 Transformer 点云目标检测方法 STDet, 在含有大量低质量困难目标的复杂场景中能够取得较高的检测精度。对于 PGN 网络生成的初始候选框, 通过构建基于特征分离的形状补全算法预测生成框内目标合理的完整形状点云, 并设计一种 Transformer 几何特征增强模块, 其首先聚合目标补全点集与目标原始空间位置特征, 然后通过多尺度邻域掩码 Transformer 捕获不同邻域掩码范围内点集的局部结构特征与全局几何信息的相关性权重, 加权增强目标关键特征以生成精细化的目标检测框。在 KITTI 数据集上充分的实验结果证明了本文方法的有效性。

参考文献:

[1] 张新钰, 邹镇洪, 李志伟, 等. 面向自动驾驶目标检测的深度多模态融合技术 [J]. 智能系统学报, 2020, 15(4): 758–771.
ZHANG Xinyu, ZOU Zhenhong, LI Zhiwei, et al. Deep multi-modal fusion in object detection for autonomous driving[J]. CAAI transactions on intelligent systems, 2020, 15(4): 758–771.

[2] LIANG Ming, YANG Bin, CHEN Yun, et al. Multi-task multi-sensor fusion for 3D object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition . Piscataway: IEEE, 2019: 7337−7345.

[3] ZHOU Yin, SUN Pei, ZHANG Yu, et al. End-to-end multi-view fusion for 3D object detection in LiDAR point clouds[EB/OL]. (2019−10−15)[2022−10−29]. https://arxiv.org/abs/1910.06528.

[4] DENG Jiajun, ZHOU Wengang, ZHANG Yanyong, et al. From multi-view to hollow-3D: hallucinated hollow-3D R-CNN for 3D object detection[J]. IEEE transactions on circuits and systems for video technology, 2021, 31(12): 4722–4734.

[5] ZHOU Yin, 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. Piscataway: IEEE, 2018: 4490−4499.

[6] YAN Yan, MAO Yuxing, LI Bo. SECOND: sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[7] 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. Piscataway: IEEE, 2019: 12689−12697.

[8] YE Maosheng, XU Shuangjie, CAO Tongyi. HVNet: hybrid voxel network for LiDAR based 3D object detection[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 1628−1637.

[9] DENG Jiajun, SHI Shaoshuai, LI Peiwei, et al. Voxel R-CNN: towards high performance voxel-based 3D object detection[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI, 2021:

1201−1209.

[10] HE Chenhang, ZENG Hui, HUANG Jianqiang, et al. Structure aware single-stage 3D object detection from point cloud[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition . Piscataway: IEEE, 2020: 11870−11879.

[11] SHI Shaoshuai, WANG Zhe, SHI Jianping, 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.

[12] SHI Shaoshuai, WANG Xiaogang, LI Hongsheng. PointRCNN: 3D object proposal generation and detection from point cloud[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 770−779.

[13] YANG Zetong, SUN Yanan, LIU Shu, et al. 3DSSD: point-based 3D single stage object detector[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition . Piscataway: IEEE, 2020: 11037−11045.

[14] ZHANG Yifan, HU Qingyong, XU Guoquan, et al. Not all points are equal: learning highly efficient point-based detectors for 3D LiDAR point clouds[EB/OL]. (2022−03−21)[2022−10−29]. https://arxiv.org/abs/2203.11139.

[15] SHI Weijing, RAJKUMAR R. Point-GNN: graph neural network for 3D object detection in a point cloud[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 1708−1716.

[16] LIU Ze, ZHANG Zheng, CAO Yue, et al. Group-free 3D object detection via transformers[C]//2021 IEEE/CVF International Conference on Computer Vision . Piscataway: IEEE, 2021: 2929−2938.

[17] MISRA I, GIRDHAR R, JOULIN A. An end-to-end transformer model for 3D object detection[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 2886−2897.

[18] PAN Xuran, XIA Zhuofan, SONG Shiji, et al. 3D object detection with pointformer[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 7459−7468.

[19] 韩磊, 高永彬, 史志才. 基于稀疏 Transformer 的雷达点云三维目标检测 [J]. 计算机工程, 2022, 48(11): 104−110, 144.
HAN Lei, GAO Yongbin, SHI Zhicai. Three-dimensional object detection of LiDAR point cloud based on sparse transformer[J]. Computer engineering, 2022, 48(11): 104−110, 144.

[20] ZHOU Dingfu, FANG Jin, SONG Xibin, et al. IoU loss for 2D/3D object detection[C]//2019 International Conference on 3D Vision . Piscataway: IEEE, 2019: 85−94.

[21] 赵毅强, 艾西丁·艾克白尔, 陈瑞, 等. 基于体素化图卷积网络的三维点云目标检测方法 [J]. 红外与激光工程, 2021, 50(10): 281–289.
ZHAO Yiqiang, ARXIDIN A, CHEN Rui, et al. 3D point cloud object detection method in view of voxel based on graph convolution network[J]. Infrared and laser engineering, 2021, 50(10): 281–289.

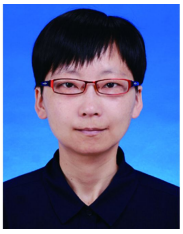
[22] SHI Shaoshuai, GUO Chaoxu, JIANG Li, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 10526−10535.

[23] LI Jiale, DAI Hang, SHAO Ling, et al. Anchor-free 3D single stage detector with mask-guided attention for point cloud[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York: ACM, 2021: 553−562.

[24] NOH J, LEE S, HAM B. HVPR: hybrid voxel-point representation for single-stage 3D object detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 14600−14609.

[25] ZARZAR J, GIANCOLA S, GHANEM B. PointRGCN: graph convolution networks for 3D vehicles detection refinement[EB/OL]. (2019−11−27)[2022−10−29]. https://arxiv.org/abs/1911.12236.

作者简介：



周静，教授，主要研究方向为深度学习与智能算法、智能机器视觉、点云目标检测、图像处理与模式识别。主持国家自然科学基金项目、湖北省自然科学基金项目和横向科研项目 20 余项，发明专利 10 余项，武汉市优秀青年教师。发表学术论文 30 余篇。

胡怡宇，硕士研究生，主要研究方向为智能机器视觉、深度学习与智能算法、三维目标检测。

黄心汉，教授，博士生导师，中国人工智能学会会士，智能机器人专业委员会名誉主任，享受国务院政府特殊津贴，湖北省有突出贡献的中青年专家。主要研究方向为智能控制、智能机器人、信息融合、图像处理与模式识别。主持国家自然科学基金项目、国家 863 计划、国家科技支撑计划及省部级和横向科研项目 60 余项，授权发明专利 11 项。发表学术论文 300 余篇，出版专著 4 部、译著 2 本。