

基于时空投影卷积的激光雷达点云语义分割系统

丁奔, 焦继超

(北京邮电大学电子工程学院, 北京 100876)

摘要: 激光雷达点云语义分割可以提供丰富的感知信息, 这对自动驾驶和机器人系统至关重要。传统的点云语义分割系统为了实现高效实时的点云分割, 通常会将三维的点云数据转换为二维的距离图像, 再通过卷积神经网络处理。然而, 距离图像的表示忽视了连续点云间的时序信息, 并且针对真实图像设计的卷积操作也无法充分利用距离图像中点的空间坐标信息。为了解决这一问题, 本文通过点云间的深度信息变化生成残差图像, 为系统输入增加了时间维度上的信息。同时, 本文提出了一个针对距离图像和残差图像的时空投影卷积。它不仅能利用距离图像中的空间坐标信息和残差图像中的时序信息, 还能动态地结合二者, 构成强大的时空特征。在 SemanticKITTI 数据集上的实验结果表明, 本文提出的时空投影卷积可以提高现有系统的精度。

关键词: 计算机视觉; 语义分割; 激光雷达; 点云

中图分类号: TP391.4

LiDAR Point Cloud Semantic Segmentation System based on Spatio-Temporal Projection Convolution

DING Ben, JIAO Jichao

(School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876)

Abstract: Semantic segmentation of LiDAR point clouds can provide rich perception information, which is crucial for autonomous driving and robotic systems. Traditional point cloud semantic segmentation systems, for efficient and real-time segmentation, typically convert 3D point cloud data into 2D distance images, which are then processed by convolutional neural networks. However, the representation of distance images overlooks the temporal information between continuous point clouds, and convolution operations designed for real images fail to fully utilize the spatial coordinate information in distance images. To solve this problem, this paper generates residual images through depth information changes between point clouds, adding temporal dimension information to system inputs. Additionally, this paper proposes a spatio-temporal projection convolution for distance images and residual images. This not only makes use of the spatial coordinate information in distance images and temporal information in residual images, but it also dynamically combines both, forming powerful spatio-temporal features. Experiment results on SemanticKITTI datasets indicate that the spatio-temporal projection convolution proposed in this paper can enhance the accuracy of existing systems.

Key words: Computer vision; Semantic Segmentation; LiDAR; Point cloud

0 引言

激光雷达点云语义分割可以从激光雷达点云中获取对周围环境的理解, 是当前自动驾驶领域的研究热点之一。然而, 由于激光雷达点云的无序性、不规则性, 传统的卷积神经

作者简介: 丁奔 (1999-), 男, 硕士, 主要研究方向: 三维感知

通信联系人: 焦继超 (1982-), 男, 副教授、博导, 主要研究方向: 多传感器融合的广域室内无缝定位。

E-mail: jiaojichao@bupt.edu.cn

网络无法直接处理点云数据。SqueezeSeg^[1]开创了基于投影的方案。即先通过球形投影将三维点云转换为二维投影图像，再输入给卷积神经网络。这实现对点云的间接处理。在此基础上，研究人员们又提出了许多优化算法。RangeNet++^[2]提出了一种在点云中运行的、支持 GPU 的 KNN 算法，缓解了分割结果的边界模糊问题。SqueezeSegV3^[3]提出了一种空间自适应卷积，能根据图像中的特征分布动态分配权重。FIDNet^[4]提出了一种无参数的解码器，大大降低了模型的参数量和计算复杂度。不过，以上基于投影的算法往往忽视了对连续点云间时序信息的利用。此外，它们使用的网络也往往是对现有图像分割模型的沿用。这未能考虑到点云投影图像与真实图像的不同。

通过对现有方法的分析，本文提出了一个基于时空投影卷积的激光雷达点云语义分割系统。首先，本文在现有投影方案的基础上增加了残差图像，与距离图像一起输入神经网络，为系统引入了时间维度上的信息。其次，本文针对距离图像和残差图像的数据特点，设计了一种时空投影卷积。它可以很好地提取距离图像和残差图像中的空间特征和时间特征，并通过注意力机制动态地结合二者，构成强大的时空特征。在 SemanticKITTI 数据集上的实验结果证明了本文提出的时空投影卷积的有效性。

1 球形投影

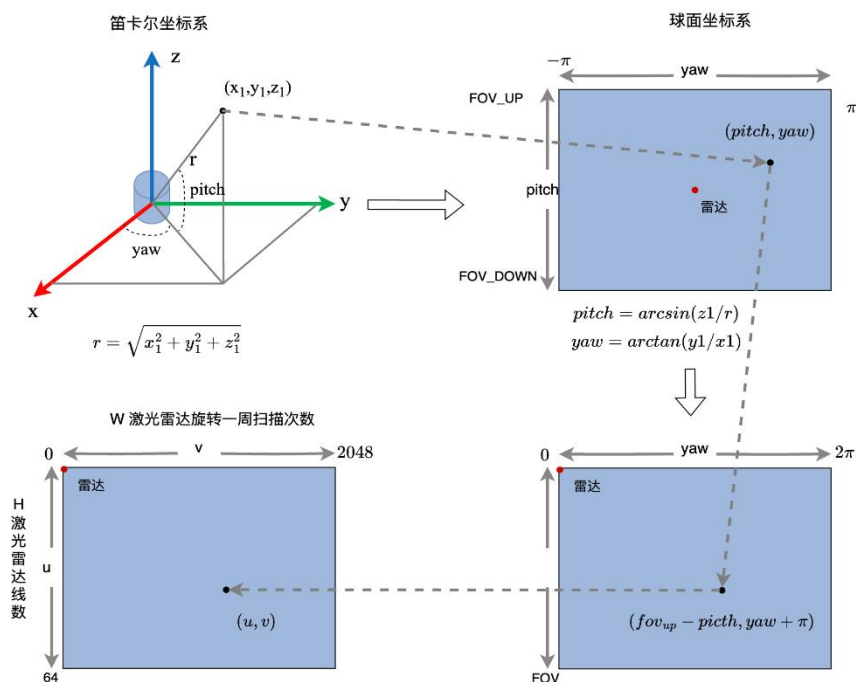


图 1 球形投影

Fig. 1 Spherical Projection

图 1 展示了球形投影的过程。将激光雷达作为坐标系原点。那么点 P 与原点的连线便于 XY 平面（即地面）形成一个角度，我们称之为俯仰角（pitch）。点 P 与原点的连线在 XY 平面的投影与 X 轴形成一个角度，我们称之为偏转角（yaw）。这样我们便获得了点云中每个点的极坐标。最后再通过归一化处理，我们就可以得到每个点在投影图像中的二

65 维坐标 (u, v) 。球形投影的数学定义如下：其中， fov 是激光雷达的垂直视场角。 H 和 W 则是投影图像的宽和高。

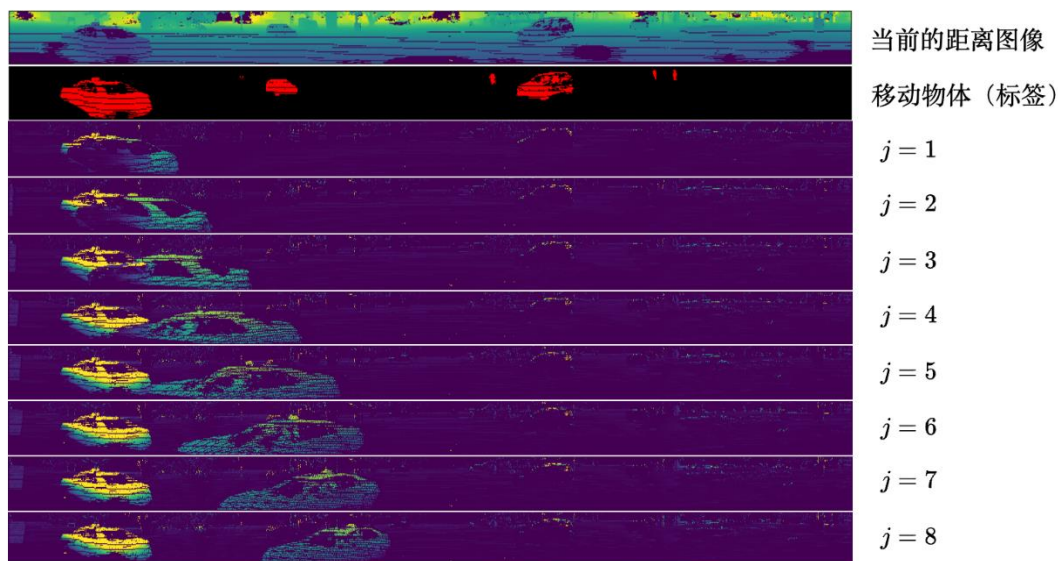
$$u = [1 - (pitch + |fov_{down}|) \times fov^{-1}] \times H \quad (1)$$

$$v = 0.5 \times [1 + yaw \times \pi^{-1}] \times W \quad (2)$$

2 方法设计

70 2.1 距离残差图像

距离残差图像是由一张距离图像和多张残差图像组合而成的伪图像。其中，距离图像是对当前帧点云球形投影得到的；残差图像是对不同帧点云的信息差球形投影得到的。图 2 展示了距离图像和多张残差图像，其中 j 指的是当前点云与前 j 帧点云之间的残差图像。



75 图 2 距离图像和残差图像

Fig. 2 Range Image and Residual Image

80 距离图像中每个像素包含以下信息：点在笛卡尔坐标系下的空间位置 (x, y, z) 、点离激光雷达数据采集中心的距离 (r) 以及反射强度 (i) 。这与真实图像的 (r, g, b) 通道信息完全不同。然而，传统卷积是为处理常规图像而设计的，它的效果完全取决于每个像素的所在位置，这使得传统卷积可能无法充分利用距离图像每个点的笛卡尔坐标所携带的丰富的几何信息。

85 残差图像中每个像素的值是两帧点云中点的深度信息差。残差图像的概念最早出现于视频识别领域。在视频识别中，残差图像是一种用于识别两帧或多帧间的差异的图像。它是通过计算一幅图像与另一幅图像的像素级别的差异得到的。这种差异可以反映出视频中的动态变化，如物体的移动或场景的变更。同样地，通过计算一帧点云与另一帧中的深度信息变化，我们也可以生成针对点云数据的残差图像。深度信息变化的计算公式如下：其中 r 是点离激光雷达数据采集中心的距离， i, k 是点云的序号。

$$d_{k,i}^l = \frac{|r^i - r_i^{k \rightarrow l}|}{r^i} \quad (3)$$

2.2 时空投影卷积

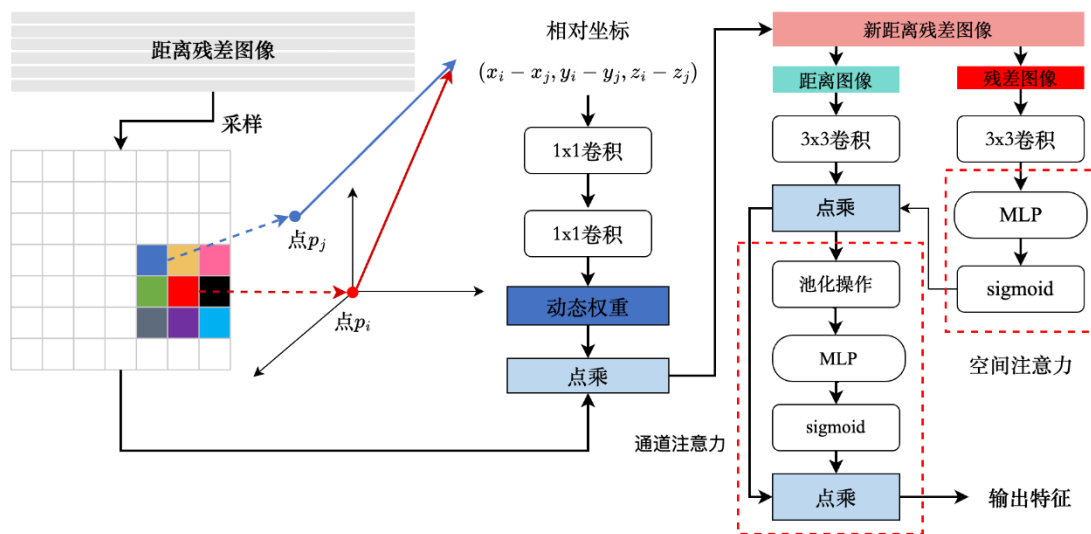


图3 时空投影卷积

Fig. 3 Spatio-Temporal Projection Convolution

图3展示了本文提出的一种时空投影卷积（Spatio-Temporal Projection Convolution，简称 STPC）。它的运行流程如下：第一步是采样操作，STPC 会从距离图像中收集每个点周围点的信息。通过计算中心点（红色方块）与采集点（其他颜色方块）之间的坐标差，就能还原出中心点所属物体的几何结构，获得中心点的空间几何信息。第二步是从每个点的空间几何信息中获取点的动态空间权重。首先通过两个点卷积构成的全连接层对上一步获取的相对坐标矩阵进行特征提取，最后将提取到的空间动态权重与原始的距离残差图像相乘，完成对原始距离残差图像的加权。第三步是时空信息的聚合，我们首先通过两个标准的卷积层分别对距离图像和残差图像进行特征提取。然后，通过本文设计的空间注意力，从残差图像中找出值得关注的移动物体，并以此作为权重与距离图像相乘完成时间信息与空间信息的融合。最后，通过使用通道注意力，为时空特征合理地分配不同特征通道的权重。

STPC 中的空间注意力由一个全连接层和 sigmoid 函数组成。我们首先通过全连接层从残差图像中生成单通道的特征图，这代表了残差图像中每个点的注意力权重。接着，将这个注意力权重输入一个 sigmoid 激活函数，使得每个元素的值落在 0 和 1 之间。这样就可以将这些值作为权重，通过元素乘法应用到距离图像上，从而对不同的空间位置赋予不同的重要性，进而可以突出距离图像中重要的移动物体。

STPC 中的通道注意力由一个平均池化函数、一个最大池化函数、一个全连接层和一个 sigmoid 函数组成。首先，我们通过平均池化操作学习目标物体的平均程度信息，通过最大池化学习到目标物体的最关键性信息。接下来，我们将平均池化和最大池化的输出分别输入到全连接层，并将二者的输出相加，从而生成特征图。在完成这些步骤之后，我们对特征图应用 sigmoid 函数以获取通道注意力的权重，并通过元素乘法将这些权重应用到

115 空间注意力机制提取到的特征上。

3 实验分析

在本节中，我们首先介绍了实验的实现细节，包括实验设置、数据集和评价指标。然后，我们在 SemanticKITTI 数据集上对比了不同方法加入时空投影卷积后的性能。最后，我们进行了消融实验，以验证本文提出的时空投影卷积的有效性。

120 3.1 数据集介绍

SemanticKITTI 数据集在 2019 年被提出，这是一个针对自动驾驶场景中点云语义分割任务设计的开源数据集。SemanticKITTI 数据集使用 Velodyne 64 线激光雷达进行数据采集，每帧点云有近 12 万个点。它采集于德国卡尔斯鲁厄市，覆盖的场景包括市区、郊区和高速公路。SemanticKITTI 数据集有 22 个序列，一共包括 43,551 个扫描。其中，序列 125 00 到 10（19,130 个扫描）被用于训练，而序列 11 到 21（20,351 个扫描）被用于测试。

3.2 评估策略

为了确保公平比较，本文的评估策略选择采用 mIoU，也就是平均交并比，来检测不同方法的性能。mIoU 可被定义为对每个类别的 IoU（Intersection over Union）进行的平均操作。IoU 衡量的是预测的矩形框或像素区域（在分割任务中）与真实矩形框或像素区域 130 之间交集和并集的比率。这是一个非常适合评估预测准确性和覆盖度的直观度量：IoU 值越大，预测的既准确又完全。计算 mIoU 时，首先单独计算每个类别的 IoU，然后对所有的 IoU 值进行简单平均。这样的计算方法可以确保各个类别的重要性得到平衡，避免过于注重大类别的精度，而忽视小类别的影响。该指标的数学定义如下：

$$mIoU = \frac{1}{n} \sum_{c=1}^n \frac{TP_c}{TP_c + FP_c + FN_c} \quad (4)$$

135 其中 c 是预测的类别，TP 是预测值与真实值都为 c 的像素点的数量，FP 是预测值不为 c 而真实值为 c 的像素点数量，FN 是预测值为 c 为真实值不为 c 的像素点数量。

3.3 对比实验

本文在 SemanticKITTI 的测试集上验证了 FIDNet 加入 STPC 卷积后的性能。FIDNet 是一个轻量级的编解码网络，其编码器部分采用了 ResNet34 的结构，其解码器无参数的， 140 全部由双线性插值组成。

在训练阶段，网络使用的距离残差图像由一张距离图像和一张残差图像组成。网络使用随机梯度下降优化器，初始动量值设置为 0.9，学习率为 0.01，学习衰减率设置为 0.03。此外，其余设置均与 FIDNet 中保持一致。

表 1 SemanticKITTI 单帧性能对比

145 Tab. 1 SemanticKITTI Single Scan Benchmark

Approach	Size	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	vegetation	trunk	mean-IoU
PointNet ^[5]	50k pts	46.3	1.3	0.3	0.1	0.8	0.2	0.2	0.0	61.6	15.8	35.7	31.0	4.6	14.6
PointNet++ ^[6]		53.7	1.9	0.2	0.9	0.2	0.9	1.0	0.0	72.0	18.7	41.8	46.5	13.8	20.1
SPLATNet ^[7]		66.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	70.4	0.8	41.5	72.3	35.9	22.8
LatticeNet ^[8]		92.9	16.6	22.2	26.6	21.4	35.6	43.0	46.0	90.0	59.4	74.1	81.7	63.6	52.9
RandLANet ^[9]		94.2	26.0	25.8	40.1	38.9	49.2	48.2	7.2	90.7	60.3	73.7	81.4	61.3	53.9
KPConv ^[10]		96.0	30.2	42.5	33.4	44.3	61.5	61.6	11.8	88.8	61.3	72.7	84.8	69.2	58.9
BAAF-Net ^[11]		95.4	31.8	35.5	48.7	46.7	49.5	55.7	33.0	90.9	62.2	74.4	82.7	63.4	59.9
RangeNet++	2048x64	91.4	25.7	34.4	25.7	23.0	38.3	38.8	4.8	91.8	65.0	75.2	80.5	55.1	52.2
3D-MiniNet ^[12]		90.5	42.3	42.1	28.5	29.4	47.8	44.1	14.5	91.6	64.2	74.5	82.8	60.8	55.8
SqueezeSegV3		92.5	38.7	36.5	29.6	33.0	45.6	46.2	20.1	91.7	63.4	74.8	82.0	58.7	55.9
SalsaNext ^[13]		91.9	48.3	38.6	38.9	31.9	60.2	59.0	19.4	91.7	63.7	75.8	81.8	63.6	59.5
FIDNet	2048x64	93.9	54.7	48.9	27.6	23.9	62.3	59.8	23.7	90.6	59.1	75.8	84.5	64.4	59.5
FIDNet+STPC		91.8	40.7	38.0	35.1	36.9	59.8	57.3	40.9	91.8	66.1	76.4	83.3	63.5	60.3

表 1 是在 SemanticKITTI 数据集上不同算法的精度对比。它展示了不同算法的平均精度，以及在移动物体较多的分类上的精度。可以看出在加入 STPC 卷积后，FIDNet 的总体精度上升了 0.8%。并且对动态物体较多的分类识别精度提升很大，例如对卡车（truck）的识别精度提升了 7.5%、对其他车辆（other-vehicle）的识别精度提升了 13%，而对摩托车手（motorcyclist）的识别精度提升 17.2%。

除此之外，本文还在 SemanticKITTI 的验证集上测试了加入 STPC 后，不同点云分割模型的提升。SemanticKITTI 验证集即训练集中的 08 序列。在验证集上测试性能时，不会将其加入训练序列。

表 2 STPC 性能对比

Tab. 2 Performance Comparison For STPC

	RangeNet++	SqueezeSegV3	SalsaNext	FIDNet
without	51.8	55.2	58.5	58.3
with	53.3	56.6	59.7	59.7

如表 2 所示，在加入了 STPC 卷积后，RangeNet53、SqueezeSegV3、SalsaNext 和 SalsaNext 模型的分割精度均出现了不同程度的上升。其中，RangNet53 的精度上升了 1.5%，SqueezeSegV3 的精度上升了 1.4%，SalsaNext 的精度上升了 1.2%，FIDNet 的精度上升了 1.4%。这证明了 STPC 卷积的有效性。

160

3.4 消融实验

残差图像的数量可能会影响到 STPC 卷积的性能以及网络的运行速度。本文进行了消融实验，对比了不同数量的残差图像对网络性能的影响，实验结果如图 4、图 5 所示：

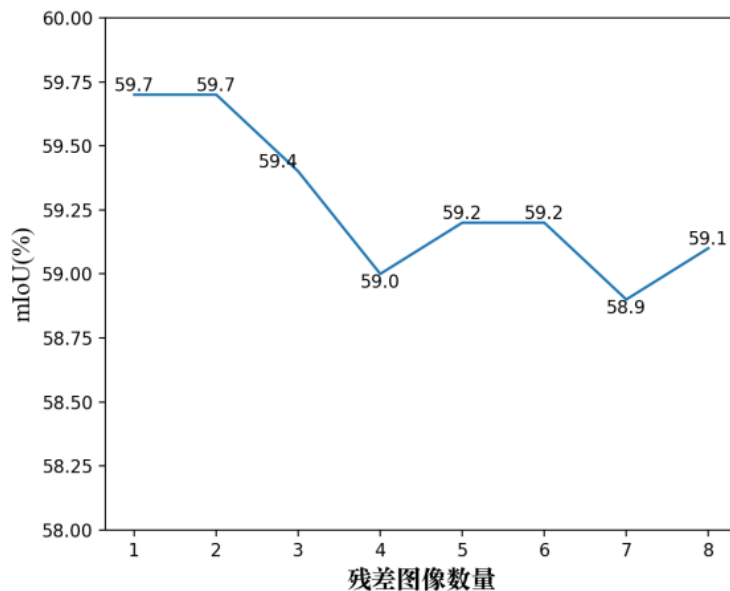


图 4 残差图像数量对精度的影响

165

Fig. 4 Accuracy Change on Residual Image Number

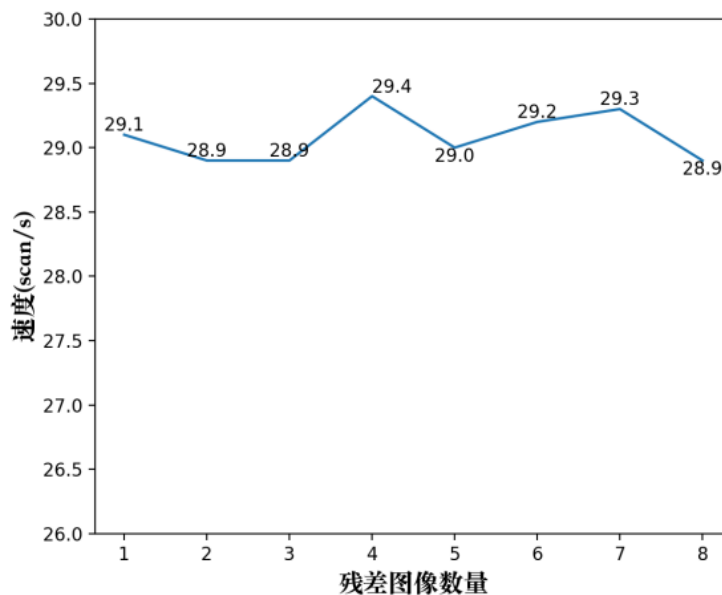


图 5 残差图像数量对速度的影响

Fig. 5 Speed Change on Residual Image Number

170

从图中可以看出，残差图像数量对运行速度影响并不大。网络的运行速度基本保持在 29 帧/秒（1 帧内的差别属于机器运行时的正常波动）。其主要原因是，与 FIDNet 的运行时间相比，STPC 所需的运行时间太少。因此残差图像数量不会影响整体运行速度。同时，从实验中可以观察到，残差图像的数量在一定程度上会影响 STPC 卷积的性能。当残

差图像数量较少时，如 1 和 2 时，网络的精度最高。而当残差图像数量过多时，如 7 和 8 时，网络的精度反而较低。经过分析，本文认为这是由残差图像的数据特性引导的。残差图像是当前点云与前 i 帧点云之间的深度距离差值投影到当前的距离图像视角生成的。然而，正如图 2 所示，引入更多的残差图像并不会使网络获取更多的信息，因为不同残差图像之间提供的信息差别并不大。此外，当 i 值过大时，那些移动速度快的物体可能会在残差图像中重复出现，这会对 STPC 的性能产生负面影响。

最后，我们通过点云可视化图直观地展示 STPC 对 FIDNet 的提升效果。从图 6 和图 7 的对比可以看出，加入 STPC 后，FIDNet 中错误点数量明显减少。

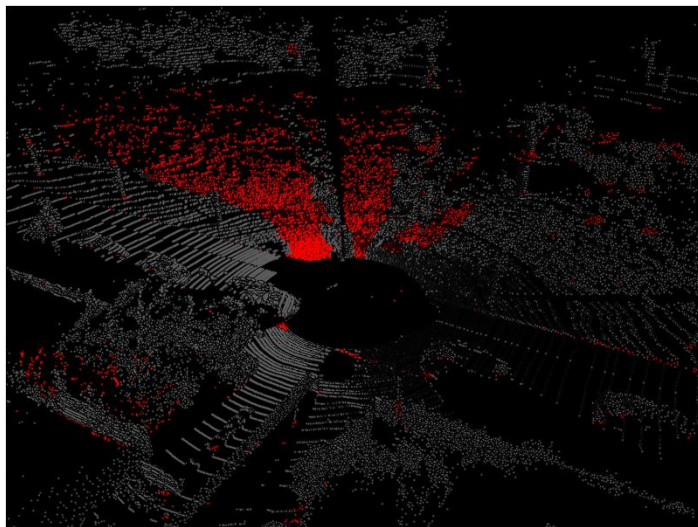


图 6 FIDNet 预测结果坏点

Fig. 6 Error Points On FIDNet

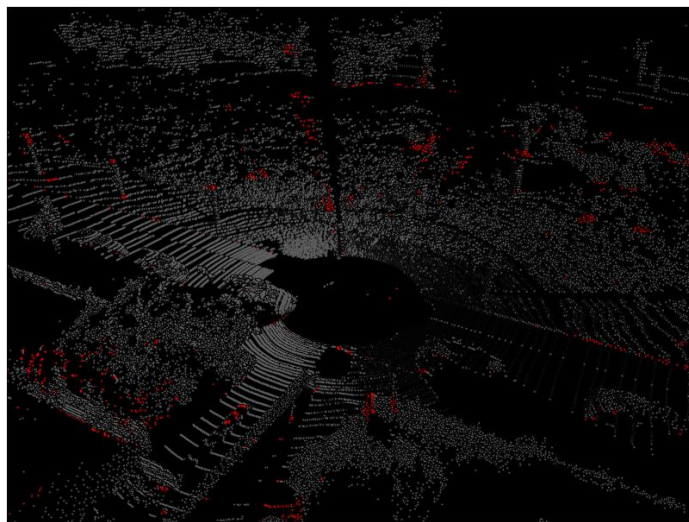


图 7 本文方法的预测结果坏点

Fig. 7 Error Points On Ours

4 结论

本文提出了一个基于距离残差图像的时空投影卷积。不同与传统的卷积操作只是关注像素间的相对位置，它可以从距离残差图像中每个点的笛卡尔坐标中提取三维空间权重。

190 此外, 它还能将这个空间权重与时序信息动态结合, 这能有效地提高现有语义分割系统的精度。本文在 SemanticKITTI 数据集上进行了实验验证, 结果表明本文提出的时空投影卷积有效的提升现有点云语义分割系统的精度, 证明了本文算法的有效性。

[参考文献] (References)

- 195 [1] WU B, WAN A, YUE X, et al. SqueezeSeg: Convolutional Neural Nets with Recurrent CRF for Real-Time Road-Object Segmentation from 3D LiDAR Point Cloud[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). 2018: 1887-1893.
- [2] MILIOTO A, VIZZO I, BEHLEY J, et al. RangeNet ++: Fast and Accurate LiDAR Semantic Segmentation[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2019: 4213-4220.
- 200 [3] XU C, WU B, WANG Z, et al. SqueezeSegV3: Spatially-Adaptive Convolution for Efficient Point-Cloud Segmentation[C]//VEDALDI A, BISCHOF H, BROX T, et al. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 1-19.
- [4] ZHAO Y, BAI L, HUANG X. FIDNet: LiDAR Point Cloud Semantic Segmentation with Fully Interpolation Decoding[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2021: 4453-4458.
- 205 [5] 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.
- [6] QI C R, YI L, SU H, et al. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space[C]//Advances in Neural Information Processing Systems: Vol. 30. Curran Associates, Inc., 2017.
- 210 [7] SU H, JAMPANI V, SUN D, et al. SPLATNet: Sparse Lattice Networks for Point Cloud Processing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2530-2539.
- [8] LUO X, XIE Y, ZHANG Y, et al. LatticeNet: Towards Lightweight Image Super-Resolution with Lattice Block[C]//VEDALDI A, BISCHOF H, BROX T, et al. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 272-289.
- 215 [9] HU Q, YANG B, XIE L, et al. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11108-11117.
- [10] THOMAS H, QI C R, DESCHAUD J E, et al. KPConv: Flexible and Deformable Convolution for Point Clouds[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6411-6420.
- 220 [11] QIU S, ANWAR S, BARNES N. Semantic Segmentation for Real Point Cloud Scenes via Bilateral Augmentation and Adaptive Fusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 1757-1767.
- [12] ALONSO I, RIAZUELO L, MONTESANO L, et al. 3D-MiniNet: Learning a 2D Representation From Point Clouds for Fast and Efficient 3D LIDAR Semantic Segmentation[J]. IEEE Robotics and Automation Letters, 2020, 5(4): 5432-5439.
- 225 [13] CORTINHAL T, TZELEPIS G, ERDAL AKSOY E. SalsaNext: Fast, Uncertainty-Aware Semantic Segmentation of LiDAR Point Clouds[C]//BEBIS G, YIN Z, KIM E, et al. Advances in Visual Computing. Cham: Springer International Publishing, 2020: 207-222.