arXiv:2302.01665v2 [cs.CV] 6 Oct 2023

CVTNet：面向自动驾驶环境中基于激光雷达的地点识别的跨视图Transformer网络

Junyi Ma，Guangming Xiong，Jingyi Xu，Xieyuanli Chen*

摘要—基于激光雷达的地点识别（LPR）是自动驾驶车辆在无GPS环境中识别曾到访地点的关键组成部分之一。现有大多数LPR方法使用输入点云的常规表示，未考虑不同视图，因而可能无法充分利用激光雷达传感器中的信息。本文提出了一种基于跨视图Transformer的网络，称为CVTNet，用于融合由激光雷达数据生成的距离图像视图（RIVs）和俯视图（BEVs）。该网络利用视图内Transformer提取视图内部的相关性，并利用视图间Transformer提取两个不同视图之间的相关性。在此基础上，我们提出的CVTNet能够端到端在线地为每次激光扫描生成偏航角不变的全局描述符，并通过当前查询扫描与预建数据库之间的描述符匹配，检索之前见过的地点。我们在三个采用不同传感器配置和环境条件采集的数据集上评估了该方法。实验结果表明，与最先进的LPR方法相比，我们的方法对视点变化和长时间间隔具有更强的鲁棒性。此外，我们的方法具有更好的实时性能，运行速度快于典型的激光雷达帧率。我们方法的实现已作为开源项目发布于：https://github.com/BIT-MJY/CVTNet。

索引词—激光雷达地点识别；自动驾驶；Transformer网络；多视图融合

I. 引言

地点识别能够确定车辆在先前见过的环境中的当前全局位置，是机器人同步定位与建图（SLAM）和全局定位的重要组成部分。在线运行时，它通过直接回归相似度 [1], [2] 或进行描述符匹配 [3], [4], [20]，从数据库中检索与当前查询最相似的参考扫描。基于激光雷达的地点识别方法 [5]–[7] 能够应用于大规模户外环境，因为它们对光照和天气变化具有鲁棒性。人们已将激光雷达数据的不同表示形式用作 LPR 方法的输入，例如

三维点云 [4], [20],体素 [8]–[10], 正态分布变换单元 [11], [12], 俯视图（鸟瞰图） [6], [13],[14], 和距离图像视图（RIV） [1], [5]。然而，很少有方法将它们融合起来，以利用不同视图的优势。Yin 等人 [15]提出了 FusionVLAD，将不同的输入形式结合起来，但他们只是使用全连接

本研究得到国家自然科学基金委员会项目（编号 52372404）的资助。J. Ma、G. Xiong 和 J. Xu 就职于北京理工大学。X. Chen 就职于国防科技大学。

*通讯作者邮箱：xieyuanli.chen@nudt.edu.cn

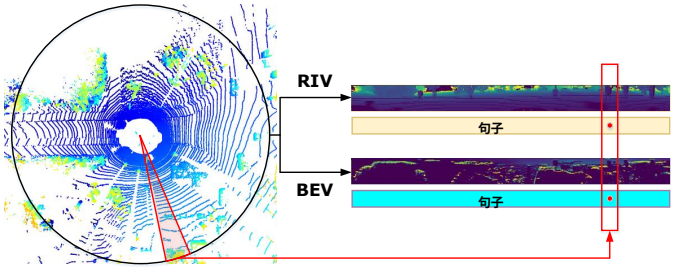

图 1：RIV 与鸟瞰图之间的对应关系产生了对齐的特征句子，供我们提出的 CVTNet 使用。

层来融合特征，因此尚未深入挖掘不同表示之间的内在关系。

为弥补这一空白，本文提出了一种名为 CVTNet 的新型跨视图Transformer网络，用于基于激光雷达的地点识别。该网络利用安装在自动驾驶车辆上的三维激光雷达生成的多层RIV和鸟瞰图。利用同一激光雷达扫描的RIV与鸟瞰图之间的对应关系（如图1所示），首先将这两个输入视图压缩为对齐的类句子特征。随后，我们提出的 CVTNet 使用视图内和视图间Transformer，从不同视图的对齐句子中提取视图内及跨视图相关性。基于这一新颖的Transformer网络，我们的方法以端到端的方式为每个地点生成偏航角不变全局描述符，最终通过描述符匹配实现鲁棒的地点识别。

本文的主要贡献是提出一种端到端 Transformer 网络，通过融合激光雷达数据的距离图像视图和鸟瞰图，实现可靠的长期地点识别。与现有采用常规输入形式的 LPR 方法不同，我们的方法同时利用激光雷达数据的距离图像视图和鸟瞰图。此外，我们将原本的单通道距离图像视图和鸟瞰图扩展为多层格式，使 Transformer 网络能够重新调整不同距离和高度处激光雷达数据的权重，从而生成更可靠、更具区分性的特征。据我们所知，这是首项从激光雷达点云中挖掘距离图像视图与鸟瞰图特征之间潜在对齐关系，并利用这一特性为地点生成可区分表征的工作。我们提出的 CVTNet 也是首个利用交叉注意力融合多视图激光雷达数据，并为激光雷达地点识别生成偏航角不变全局描述符的网络。得益于所设计的偏航角不变架构，CVTNet 能够抵抗视点变化，即使车辆朝相反方向行驶，也能实现可靠的地点识别；这一点已在 NCLT、KITTI 和自录数据集上得到了全面评估。

实验结果表明，无论是在回环检测还是地点识别方面，我们提出的方法都优于最先进方法。此外，本文还提供了针对内部和视图间 Transformer 模块的消融研究，以及对偏航角不变性的验证实验。

总而言之，我们提出三项主张：(i) 使用多视角激光雷达数据，在室外大规模环境中实现地点识别和回环检测的最先进性能；(ii) 基于所提出的航向角不变架构，即使视角不同，也能准确识别地点；(iii) 实现在线运行，运行时间不到 50 毫秒，快于典型的激光雷达帧率。我们在多个数据集上进行的实验评估支持上述所有主张。

本文其余部分安排如下。第二节将探讨相关工作。第三节将详细介绍我们提出的 CVTNet 架构及相关数学推导。第四节将展示实验设置、结果及相应分析。最后，第五节总结全文。

II. 相关工作

地点识别已在机器人学和计算机视觉领域得到广泛研究 [16], [17]。本文仅关注基于激光雷达的地点识别（LPR）。

LPR 是一个经典的研究课题，已有许多传统的手工设计方法被提出 [6], [13], [14], [18],[19]。近年来，随着神经网络技术的发展，越来越多的基于学习的方法取得了显著进展。现有研究中，有些方法将原始点云或其体素化和分割形式作为模型输入。例如，Uy 等人 [20]首次将常用的 PointNet [21]与可微分的 VLAD 架构 [17], 相结合，为后续从点云中提取全局描述子的方法提供了基础框架。Liu 等人 [3] 提出了 LPD-Net，该方法利用图神经网络聚合从点云中提取的多个局部特征。Dub´e 等人 [22]提出的 SegMap 首先从点云中提取分段，并将其作为所设计卷积神经网络的输入。Vidana-pathirana 等人 [4] 提出的 LoGG3D-Net 直接对原始点云进行稀疏卷积，同时考虑局部一致性损失和场景级损失。相比之下，Komorowski [9]提出的 MinkLoc3D 将稀疏体素化表示与卷积相结合，以解决无序集合问题。最近，Cattaneo 等人 [7] 提出的 LCDNet 使用经过改造的 PV-RCNN架构提取特征，并据此生成点云的全局描述子，以及估计查询与参考扫描之间的相对位姿。

尽管基于点云的激光雷达位置识别方法已取得不错的效果，但其中大多数无法在线运行，因此阻碍了其在现实应用中的部署。为解决这一问题，许多其他研究采用激光雷达数据的类图像表示形式，即距离图像视图或鸟瞰图，以提高识别效率并探索更具代表性的属性。例如，OverlapNet [1]利用距离图像，直接回归查询扫描与参考扫描之间的重叠值和相对偏航角。基于重叠这一概念，OverlapTransformer [5] 利用天然具备的偏航角等变性，构建了偏航角不变架构。Zhao 等人 [23] 提出的 SphereVLAD++ 也使用距离图像视图提取视点不变全局描述子。与距离图像视图不同，鸟瞰图会损失部分精细的深度信息，但能保留稳定的高度信息，这对长期定位很有帮助。例如，Scan Context [6] 和 Intensity Scan Context [13]均利用俯视图将每个点云编码为特征矩阵。LiDAR-Iris [19]进一步将基于上下文的方法拓展到频域，并使用傅里叶变换。受 LiDAR-Iris 启发，BV-Match [14] 也借鉴了频域的优势，提取关键点和 BVFT 描述子，用于基于词袋的检索。CCL [24] 提出使用对比学习提取更具区分性的地点描述符，以取得更好的 LPR 效果。

上述所有 LPR 方法都只使用单一格式的点云作为输入，未结合不同格式来进一步提升所提取特征和描述符的表征能力。Yin 等人的 FusionVLAD [15] 是首个考虑原始点云不同视图输入的工作。然而，该方法仅使用全连接网络来融合提取的特征，无法充分利用多视图中的不同信息。Transformer [25]近年来已被用于更好地提取地点识别所需的特征 [5], [8], [12]。然而，目前尚无工作成功使用 Transformer 融合来自不同输入形式或视图的特征。在本文中，我们提出了一种新颖的 Transformer 网络，用于提取 RIV 与鸟瞰图之间及各自内部的相关性。我们明确对齐来自同一激光雷达数据的 RIV 和鸟瞰图特征，从而利用交叉Transformer更好地发现双视图中句子式特征之间的关系，并提取更具区分性的全局描述子，以实现更准确的地点识别。

III. 所提方法

我们提出的方法概览如图2所示。CVTNet由三个模块组成：多视图多层生成器（MMG）、对齐特征编码器（AFE）和多视图融合模块（MVF）。MMG首先按距离和高度区间划分区域，分别进行球面投影和俯视投影，以生成多层输入数据（见第III-A节）。随后，AFE以多层RIV和鸟瞰图为输入，并使用全卷积编码器生成句子式特征。接着，它应用视图内Transformer提取各视图压缩特征的内部相关性（见第III-B节）。最后，将增强后的特征体输入MVF模块。该模块首先使用视图间Transformer融合来自不同视图的对齐特征，然后在单视图特征和融合特征上应用NetVLAD及多层感知机（MLP），生成最终的一维全局描述符（见第III-C节）。最终描述符对偏航角旋转具有不变性（见第III-D节）。我们使用带有重叠标签的三元组损失训练网络，以更好地区分正例和负例（见第III-E节）。

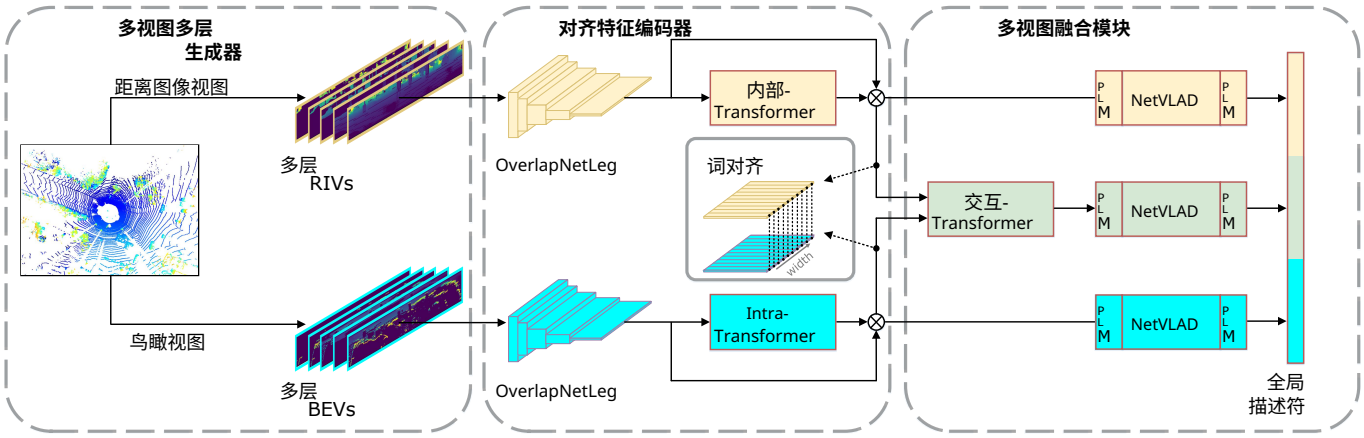

图 2：我们提出的 CVTNet 的流程概览。首先，它为两个独立分支生成多层 RIV 和 BEV图像，并结合 OverlapNetLeg 与内部 Transformer 模块，分别对对齐的类句子特征进行编码。然后，利用交互 Transformer 模块提取跨视图的内部相关性。最后，分别在三个分支上使用带有多层感知机的 NetVLAD，生成用于地点识别快速检索的最终偏航角不变全局描述符。

A. 多视图多层生成器

对于LPR方法而言，从不同扫描中提取判别性特征至关重要。现有方法通过单次球面投影或俯视投影获取RIV或鸟瞰图，这会导致数据丢失，从而降低检索性能。为了充分利用激光雷达数据中的所有信息，同时保持较快的处理速度，我们提出在不同距离和高度下进行多次投影，并基于空间划分构建多层输入。通过这种方式，我们提出的网络能够学习对不同距离和高度的信息进行差异化加权，从而从扫描中提取更具代表性的特征。

更具体地说，我们提出了多视图多层生成器。对于球坐标系中的距离图像视图，它首先根据预设的距离区间 $\{s_0, s_1, s_2, s_3, ..., s_q\}$ 对区域进行离散化，得到划分后的空间 $E^r = \{E_1^r, E_2^r, E_3^r, ..., E_q^r\}$。然后，它对输入点云 $\mathcal{P}$ 中不同距离区间内的激光点应用球面投影，生成多层距离图像。类似地，对于欧几里得坐标系中的鸟瞰图，它根据高度区间 $\{t_0, t_1, t_2, t_3, ..., t_q\}$ 对空间进行离散化，得到 $E^b = \{E_1^b, E_2^b, E_3^b, ..., E_q^b\}$，并为所有区间生成多层鸟瞰图图像。

激光雷达点 $\boldsymbol{p}_k \in \mathcal{P}$，$\boldsymbol{p}_k = (x_k,\ y_k,\ z_k)$ 与对应距离图像视图 $\mathcal{R}_i$ 中像素坐标 $(u_k^r,\ v_k^r)$ 之间的对应关系可表示为：

$$\begin{pmatrix} u_k^r \\ v_k^r \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left[1 - \arctan(y_k, x_k)/\pi\right] w \\ \left[1 - \left(\arcsin(z_k/r_k) + \mathrm{f}_{\mathrm{up}}\right)/\mathrm{f}\right] h \end{pmatrix}, \quad (1)$$

其中，$r_k = ||\boldsymbol{p}_k||_2 \in [s_{i-1}, s_i]$ 是对应点的距离测量值，$\mathrm{f} = \mathrm{f}_{\mathrm{up}} + \mathrm{f}_{\mathrm{down}}$ 是传感器的垂直视场，w，h 分别是生成的距离图像的宽度和高度。

我们通过俯视投影将相同的激光雷达点投影到鸟瞰图中。同一激光雷达点 $\boldsymbol{p}_k \in \mathcal{P}$ 与相应鸟瞰图像中的像素坐标 $(u_k^b,\ v_k^b)$ 之间的对应关系 $\mathcal{B}_j$ 可表示为：

$$\begin{pmatrix} u_k^b \\ v_k^b \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left[1 - \arctan(y_k, x_k)/\pi\right] w \\ \left[r_k'/\mathrm{f}'\right] h \end{pmatrix}, \quad (2)$$

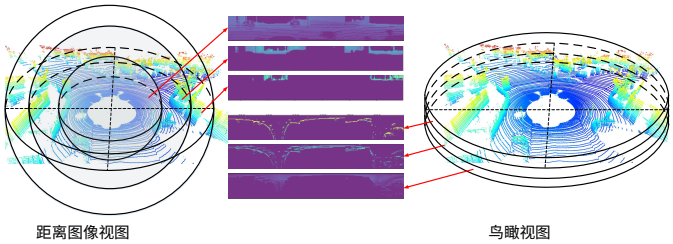

图 3：RIV（左）和BEV（右）的多层生成。

其中 $r_k' = ||(x_k, y_k)||_2 \in [t_{j-1}, t_j]$，$\mathrm{f}'$ 是最大感知范围。

在本研究中，为了更好地对齐，我们令RIV和BEV图像具有相同的宽度和高度，从而得到 $u_k^r = u_k^b$。因此，RIV和BEV图像中索引相同的列均来自同一空间扇区，如图1所示。最后，MMG生成多层RIV $\mathbb{R} = \{\mathcal{R}_i\}$ 和多层BEV $\mathbb{B} = \{\mathcal{B}_j\}$，如图 3所示。

B. 对齐特征编码器

为了更好地融合激光雷达数据的不同表示，CVTNet 使用对齐特征编码器提取每种表示的粗粒度特征，并将它们对齐以便后续融合。首先，沿通道维度拼接所有距离图像视图 $\mathbb{R}$ 和鸟瞰图 $\mathbb{B}$，得到大小为 $q \times h \times w$的 $\mathcal{R}$ 和 $\mathcal{B}$。然后，使用两个架构相同的 OverlapNetLeg [1], [5]模块，将输入压缩为类似句子的特征。OverlapNetLeg 是一个全卷积编码器，仅沿垂直维度压缩距离图像视图和鸟瞰图，而不改变宽度维度，以避免离散化误差影响偏航等变性。随后，使用视图内Transformer增强各分支输出的粗粒度特征。对于 RIV 分支，我们将增强后的特征记为 $\mathcal{A}_0^r = \mathrm{AFE}_r(\mathcal{R})$；对于 BEV 分支，则记为 $\mathcal{A}_0^b = \mathrm{AFE}_b(\mathcal{B})$。两者的大小均为 $c \times 1 \times w$，其中 c 表示特征通道数。

如 [5]，所证明，RIV 分支的特征具有偏航角等变性：

$$\mathcal{A}_0^r C = \mathrm{AFE}_r(\Pi_r(R_\theta \mathcal{P})), \quad (3)$$

其中 $\Pi_r(\cdot)$ 是式（1）中详细说明的球面投影，C 表示通过矩阵右乘对特征进行列移位，θ 是偏航角变化量，而 R_θ 是点云 $\mathcal{P}$的偏航旋转矩阵。

本文进一步表明，我们提出的鸟瞰图分支也具有偏航角等变性。根据式（ Π_b（·）中详述的2）所示的自上而下投影，点的偏航旋转对应于鸟瞰图上的特定水平移位：

$$\begin{pmatrix} u_k^{b\prime} \\ v_k^{b\prime} \end{pmatrix} = \begin{pmatrix} u_k^b + s \\ v_k^b \end{pmatrix}, \quad s = \frac{1}{2}\left[1 - \theta\pi^{-1}\right] w, \tag{4}$$

$$\mathcal{B}C = \Pi_b(R_\theta \mathcal{P}), \tag{5}$$

其中，C 和 R_θ 分别表示相同的列移位和偏航旋转。

由于 OverlapNetLeg 和视图内Transformer已证明鸟瞰图分支具有偏航角等变性 [5]，，因此我们也获得了具有偏航角等变性的特征：

$$\mathcal{A}_0^b C = \mathrm{AFE}_b(\Pi_b(R_\theta \mathcal{P})), \tag{6}$$

在使用交叉Transformer融合不同视图并提取跨模态相关性 [26], [27]之前，必须先对齐来自不同模态的特征。为此，我们首先确保来自两个不同视图的视图间Transformer输入在空间上对齐。我们令输入的 RIV 和鸟瞰图图像具有相同的宽度和高度，因此 $u_k^r = u_k^b$ 如第III-A节所述成立。同一个激光点投影到两幅图像的同一列，因此一个空间扇区对应于两幅图像中相同的列索引。为保持对齐，我们进一步将网络架构设计为仅压缩高度维度，并保留来自同一空间扇区的信息，使其聚合到两个输出特征的同一列中。我们沿通道维度拼接 OverlapNetLeg 的输出特征与经视图内Transformer增强的特征。因此，AFE 的输出特征 $\mathcal{A}_0^r$ 和 $\mathcal{A}_0^b$ 在宽度维度上设计为词对齐。

C. 多视图融合模块

我们的 CVTNet 在设计的多视图融合模块中融合多视图特征。该模块由一个交互 Transformer 和三个 NetVLAD-MLP 组合构成 [17]，如图所示。2。首先，它使用所提出的交互 Transformer 模块融合来自不同视图的两个分支的特征，详见图 4。交互 Transformer 模块中也包含两个分支，每个分支由 l 个堆叠的 Transformer 块组成，通过使用一个视图的查询特征，以及另一个视图的键特征和值特征，提取不同视图之间的相关性。

对于 RIV 分支，我们将提取的特征体表示为通过 l_{th} Transformer 块的多头自注意力（MHSA）得到 $\mathcal{A}_l^r$，随后 Transformer 的交叉注意力机制可表示为：

$$\bar{\mathcal{A}}_l^r = \mathrm{Attention}(\mathcal{Q}_{l-1}^r, \mathcal{K}_0^b, \mathcal{V}_0^b) = \mathrm{softmax}\left(\frac{\mathcal{Q}_{l-1}^r {\mathcal{K}_0^b}^T}{\sqrt{d_k}}\right)\mathcal{V}_0^b, \tag{7}$$

其中，$\mathcal{Q}_{l-1}^r$ 基本上是经过层归一化的 $\mathcal{A}_{l-1}^r$ 在 RIV 中的查询分块，而 $\mathcal{A}_0^r$ 是 RIV 分支中 AFE 的输出特征。$\mathcal{K}_0^b$，$\mathcal{V}_0^b$ 是键和值

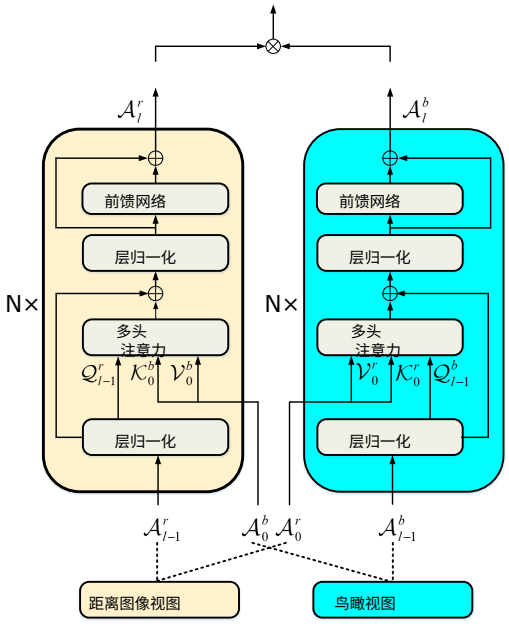

图 4：视图间Transformer 模块。

分块，来自经过层归一化的 $\mathcal{A}_0^b$，即 BEV 分支中 AFE 的输出特征。d_k 表示分块的维度。随后将 $\mathcal{A}_l^r$ 输入前馈网络（FFN）和层归一化（LN），生成 l_{th} Transformer 块的注意力增强特征 $\mathcal{A}_l^r$，其公式如下：

$$\mathcal{A}_l^r = \mathrm{FFN}(\mathrm{LN}(\bar{\mathcal{A}}_l^r + \mathrm{LN}(\mathcal{A}_{l-1}^r))) + \mathrm{LN}(\bar{\mathcal{A}}_l^r + \mathrm{LN}(\mathcal{A}_{l-1}^r)). \tag{8}$$

对于鸟瞰图分支，我们对称地执行相同的操作：

$$\bar{\mathcal{A}}_l^b = \mathrm{Attention}(\mathcal{Q}_{l-1}^b, \mathcal{K}_0^r, \mathcal{V}_0^r) = \mathrm{softmax}\left(\frac{\mathcal{Q}_{l-1}^b {\mathcal{K}_0^r}^T}{\sqrt{d_k}}\right)\mathcal{V}_0^r, \tag{9}$$

$$\mathcal{A}_l^b = \mathrm{FFN}(\mathrm{LN}(\bar{\mathcal{A}}_l^b + \mathrm{LN}(\mathcal{A}_{l-1}^b))) + \mathrm{LN}(\bar{\mathcal{A}}_l^b + \mathrm{LN}(\mathcal{A}_{l-1}^b)). \tag{10}$$

然后，我们将最后一个 Transformer 块中的 $\mathcal{A}_l^r$ 与 $\mathcal{A}_l^b$ 拼接，得到中间融合特征 $\mathcal{A}_l^f$。

最后，我们的 MVF 在 RIV 和鸟瞰图特征 $\mathcal{A}_0^r$ 和 $\mathcal{A}_0^b$以及交叉融合特征 $\mathcal{A}_l^f$ 上使用 NetVLAD-MLP 组合。NetVLAD-MLP 已广泛用于 LPR [3], [7], [12]，以生成全局描述符。我们还利用三个 NetVLAD-MLP 组合，将两个视图内Transformer和一个跨视图变换器间模块的特征压缩为全局描述子 $[g^r, g^b, g^f]$。

D. 偏航角不变性

我们进一步从数学上证明，CVTNet 生成的多视图融合描述符具有偏航角不变性，因此我们的方法对视点变化非常鲁棒。我们已在第III-B 节证明，对齐特征编码器的输出在式（3）和式（6）中具有偏航角等变性。正如 [20], NetVLAD 具有置换不变性，而原始点云偏航旋转导致其输入发生的列移位，可视为通道维度上一组一维向量的重新排序 [5]。因此，RIV 分支和鸟瞰图分支中 NetVLAD 的输出不受偏航旋转影响，全局描述子 g^r 和

g^b 具有偏航角不变性。接下来，我们将证明融合后的跨视图特征 g^f 也具有偏航角不变性。

注意，$\mathcal{A}_0^r$ 和 $\mathcal{A}_0^b$ 共享相同的位移，C 该位移由单个激光雷达数据的一次偏航旋转引起，已在第III-B节中介绍。因此，对于旋转后的视图间Transformer中的第一个 Transformer 块，公式（7）变为：

$$\begin{aligned}\widetilde{\mathcal{A}}_1^r &= \text{Attention}(C\mathcal{Q}_0^r, C\mathcal{K}_0^b, C\mathcal{V}_0^b),\\ &= \text{softmax}\Big(\frac{C\mathcal{Q}_0^r(C\mathcal{K}_0^b)^T}{\sqrt{d_k}}\Big)C\mathcal{V}_0^b,\\ &= C\,\text{softmax}\Big(\frac{\mathcal{Q}_0^r\mathcal{K}_0^{bT}}{\sqrt{d_k}}\Big)\mathcal{V}_0^b = C\bar{\mathcal{A}}_1^r. \end{aligned} \tag{11}$$

类似地，对于后续堆叠的交叉 Transformer 模块，也有：

$$\begin{aligned}\widetilde{\mathcal{A}}_l^r &= \text{Attention}(C\mathcal{Q}_{l-1}^r, C\mathcal{K}_0^b, C\mathcal{V}_0^b),\\ &= \text{softmax}\Big(\frac{C\mathcal{Q}_{l-1}^r(C\mathcal{K}_0^b)^T}{\sqrt{d_k}}\Big)C\mathcal{V}_0^b,\\ &= C\,\text{softmax}\Big(\frac{\mathcal{Q}_{l-1}^r\mathcal{K}_0^{bT}}{\sqrt{d_k}}\Big)\mathcal{V}_0^b = C\bar{\mathcal{A}}_l^r, \end{aligned} \tag{12}$$

这意味着跨视图 Transformer 块的多头自注意力输出具有偏航角等变性。注意，在公式（11）和公式（12）中，位移 C 采用左乘，这是因为输入特征的第一维是宽度维度，第二维是嵌入维度。由于前馈网络和层归一化对于特征 [5]，也具有偏航角等变性，并且最后的拼接在通道维度上进行，因此视图间 Transformer 的输出是偏航角等变的跨视图特征体。随后，使用带有多层感知机的 NetVLAD 生成偏航角不变全局描述符 g^f。由于这三个部分均具有偏航角不变性，我们最终的全局描述符也具有偏航角不变性。

E. 网络训练

接下来 [1], [5], ，我们使用扫描对之间更合适的重叠度来监督网络，而不是使用真实距离。对于每个训练元组，我们使用一个查询描述符 $\mathcal{G}_a$、k_p 个正样本描述符 $\{\mathcal{G}_p\}$ 和 k_n 个负样本描述符 $\{\mathcal{G}_n\}$ 来计算惰性三元组损失：

$$\begin{aligned}&\mathcal{L}_T(\mathcal{G}_q, \{\mathcal{G}_p\}, \{\mathcal{G}_n\}) =\\ &k_p(\alpha + \max_p(d(\mathcal{G}_q, \mathcal{G}_p))) - \sum_{k_n}(d(\mathcal{G}_q, \mathcal{G}_n)),\end{aligned} \tag{13}$$

其中，α 是用于避免损失为负的间隔，$d(\cdot)$是欧几里得距离平方。在我们的工作中，重叠率大于 0.3 的一对扫描被视为正样本，否则视为负样本。此外，我们为三元组损失设置了 $k_p = 6$、k_n =6 和 $\alpha = 0.5$ 。我们使用三元组损失，使查询与所有采样负样本全局描述符之间的距离远大于查询与最难正样本全局描述符之间的距离。



图 5：自动驾驶车辆在城市环境中采集的自录数据集。

IV. 实验评估

实验评估旨在展示我们方法的性能，并支持以下论断：我们的方法能够：(i) 使用多视角激光雷达数据，在大规模户外环境中实现地点识别和回环检测的最先进性能；(ii) 基于所提出的偏航角不变架构，尽管视角不同，仍能良好地识别地点；(iii) 实现快速运行，运行时间低于 50 毫秒。

A. 实现与实验设置

为了评估我们提出方法的定位性能，我们使用了三个环境条件和传感器配置各不相同的数据集，包括公开可用的 NCLT 数据集 [28] 和 KITTI 里程计序列 [29]，以及我们在城市驾驶环境中采集的自录数据集。我们的自录数据集由一辆自动导引车采集，该车配备了激光雷达传感器（禾赛PandarXT，32 线）、广角相机（SENSING SG2，HFOV 106°，VFOV56°）、RTK全球导航卫星系统（Asensing INS570D，0.1° 横滚/俯仰误差 0.2°，偏航误差，2 厘米 + 1 ppm，RTK定位）、微型计算机（英特尔至强E，3.5 GHz，80 W），以及一块Nivida Tesla T4（16 GB 显存，70 W），如图5所示。

1）RIV输入：对于采用 32 线激光雷达数据的 NCLT 和自录数据集的 RIV输入，我们将距离区间设为{[0 米，15 米], [15 米，30 米], [30 米，45 米], [45 米，60 米]}，以生成 4 层距离图像。我们还使用 60 米范围内的所有激光点生成距离图像，并将其与 4 层距离图像拼接，得到尺寸为5×32×900的最终 RIV输入。对于 KITTI序列，我们将距离区间设为{[0 米，15 米], [15 米，30 米], [30 米，45 米], [45 米，80 米]}，以保持与现有研究相同的最大探测距离 [1], [5]，从而进行公平比较。

2）BEV输入：对于NCLT数据集和自录数据集的BEV输入，我们将高度区间设为 {[−4 m、0 m], [0 m、4 m], [4 m、8 m], [8 m、12 m]}，以生成4层BEV图像，并将其与由60 m范围内所有激光点投影得到的BEV图像拼接。对于KITTI序列，我们将高度区间设为 {[−3 m −1.5 m]、[−1.5 m、0 m], [0 m 1.5 m]、[1.5 m、5 m]}，以适应其数据分布。多层BEV的尺寸与多层RIV相同。

3）CVTNet配置：我们的CVTNet使用与OverlapTransformer相同的OverlapNetLeg [5]。对于所有Transformer模块，我们将嵌入维度设为 $d_{model} = 256$，注意力头数量设为 $n_{head} = 4$，前馈层的中间维度设为 $d_{ffn} = 1024$。按照 [5], 的设置，我们在每个视图内Transformer模块中使用一个自注意力Transformer块，并在交互Transformer模块中使用两个交叉Transformer模块；随后的消融研究验证了这是最佳配置。对于所有NetVLAD模块，我们将中间特征维度设为 d_{inter} =1024，输出特征维度设为 $d_{output} = 256$，聚类中心数量设为 $d_K = 64$。CVTNet的输出描述符是一个大小为 1×768的向量。针对每个数据集进行评估时，我们提出的CVTNet训练30轮，使用ADAM优化器更新网络权重，初始学习率为5e-6，并每5步应用0.9的权重衰减。其他相关训练配置已在第III-E节中说明。测试设置遵循先前工作 [30], ，并将在以下章节中介绍。我们将开源CVTNet的实现，以提供更多网络细节和评估配置。

4) 基线配置：在以下实验中，我们将所提出的 CVTNet 与多种最先进的方法进行比较，包括用于实时检索且不进行移位匹配的 PointNetVLAD [20],Scan Context [6]、MinkLoc3D [9], Over- lapTransformer [5], ，以及另一种多视角激光雷达地点识别方法 FusionVLAD [15]。除 FusionVLAD 生成融合后的1×2048描述符外，上述所有基于学习的基线方法生成的描述符大小均为 1×256，且均采用各自原始的配置。Scan Context 生成的特征矩阵设为 20×60。请注意，由于 FusionVLAD 未开源，我们根据其原始论文重新实现了该方法；实验中所有其他基线方法的参数配置也均遵循其原始论文。

B. 地点识别评估

第一项实验验证了我们的观点：在室外大规模环境中使用多视角激光雷达数据时，我们的方法在地点识别和回环检测方面达到了最先进的水平。

接下来，我们使用 NCLT数据集评估长距离地点识别任务中，使用序列 2012-01-08 训练模型，并使用序列 2012-02-05、2012-06-15、2013-02-23 和 2013-04-05 评估识别性能。除了 NCLT 数据集，我们还录制了自己的数据集，其中包含在同一环境中多次重复的反向驾驶情况。这类情况在自动驾驶中很常见，但对地点识别方法而言颇具挑战。我们使用序列-1作为训练数据库，并将另外两次行驶记录（序列-2和序列-3）作为查询序列进行评估。它们行驶路线相同，但方向相反。相比之下，KITTI 里程计序列用于评估回环检测，其中使用序列 03–10进行训练、序列 02进行验证，并使用序列 00进行评估。

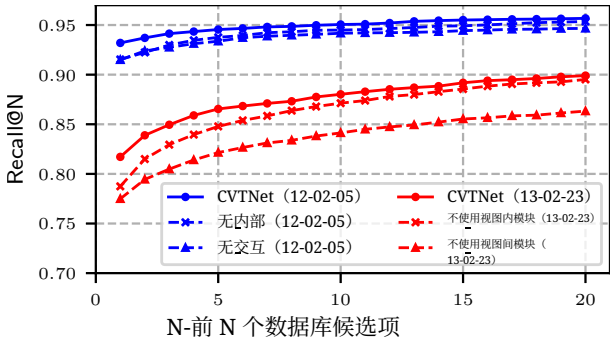

图 6：视图内与视图间 Transformer 模块的消融研究。

我们使用平均前1召回率（AR@1）、前5召回率（AR@5）和前20召回率（AR@20）作为 NCLT 和自录数据集评估的指标。对于 KITTI 数据集，我们还使用曲线下面积和F1最大分数作为评估指标，具体介绍见 [5]。评估结果如表 I 和表 II 所示。可以看出，在地点识别指标方面，我们提出的 CVTNet 性能优于所有最先进的基线方法。与 OverlapTransformer 相比，CVTNet 使用距离图像视图和鸟瞰图的组合，而非仅使用距离图像作为输入，使其在识别时间间隔较大的地点时性能有所提升，在 NCLT 数据集上的 AR@1 提高了 7.1% ～ 28.4% 。与同样考虑不同视图的 FusionVLAD 相比，我们提出的 CVTNet 在 NCLT 数据集上的 AR@1 高出14.6% ～35.1% ，在自录数据集上的 AR@1 高出 2.1% ～ 4.2%。在 KITTI 数据集上进行回环检测时，我们的方法也显著优于所有基线方法。

*C. Transformer*模块的消融研究

本消融研究探讨视图内和视图间Transformer模块的作用。我们将CVTNet与不使用Transformer模块的变体进行比较。我们使用NCLT数据集中序列2012- 02-05和2013-02-23的平均召回率AR@N 评估所有变体。结果如图6所示。可以看出，同时使用视图内和视图间Transformer模块的CVTNet始终优于所有变体。此外，仅使用视图间Transformer的CVTNet优于仅使用视图内Transformer的版本，这表明，与视图内Transformer相比，视图间Transformer对增强最终全局描述符的判别能力贡献更大。当地点识别的时间间隔较大时，使用这两类Transformer的优势更加明显；比较在不同时间记录的两个序列的结果即可得出这一结论。

D. 交叉 *Transformer* 模块消融研究

这项消融研究旨在展示 CVTNet 视图间Transformer模块中交叉 Transformer 模块数量的影响。请注意，为了降低内存消耗，我们仅将本方法与 $1 \sim 3$ 交叉 Transformer 模块进行比较。NCLT数据集上的实验结果如表III所示。可以看出，包含两个交叉 Transformer 模块的视图间 Transformer优于包含

表 I：NCLT数据集上的识别性能比较

方法	2012-02-05			2012-06-15			2013-02-23			2013-04-05		
	AR@1	AR@5	AR@20	AR@1	AR@5	AR@20	AR@1	AR@5	AR@20	AR@1	AR@5	AR@20
Scan Context [6]	0.767	0.836	0.909	0.581	0.637	0.724	0.481	0.564	0.726	0.418	0.496	0.649
PointNetVLAD [20]	0.746	0.823	0.875	0.612	0.720	0.782	0.469	0.604	0.719	0.449	0.576	0.683
MinkLoc3D [9]	0.802	0.864	0.926	0.630	0.685	0.774	0.507	0.616	0.751	0.482	0.587	0.685
OverlapTransformer [5]	0.861	0.899	0.930	0.639	0.697	0.780	0.536	0.645	0.764	0.496	0.603	0.715
FusionVLAD [15]	0.786	0.870	0.922	0.462	0.579	0.702	0.510	0.643	0.754	0.429	0.553	0.667
CVTNet（本文方法）	**0.932**	**0.946**	**0.957**	**0.793**	**0.835**	**0.871**	**0.817**	**0.866**	**0.899**	**0.780**	**0.826**	**0.869**

表 II：KITTI 和自录数据集上的识别性能比较

方法	序列-2			序列-3（反向）			KITTI	
	AR@1	AR@5	AR@20	Recall@1	AR@5	AR@20	AUC	F1最大值
Scan Context [6]	0.877	0.958	0.986	0.564	0.710	0.810	0.836	0.835
PointNetVLAD [20]	0.913	0.936	0.957	0.400	0.521	0.623	0.856	0.846
MinkLoc3D [9]	0.948	0.971	0.993	0.450	0.653	0.808	0.894	0.869
OverlapTransformer [5]	0.974	0.987	0.995	0.776	0.878	0.931	0.907	0.877
FusionVLAD [15]	0.971	0.990	0.997	0.865	0.923	0.960	0.882	0.860
CVTNet（我们的方法）	**0.992**	**0.996**	**0.998**	**0.907**	**0.951**	**0.993**	**0.911**	**0.880**

TABLEIII：CVTNet 在 NCLT 数据集上使用不同数量交叉 Transformer 模块的比较 et

数量	运行时间 [毫秒]	2012-02-05			2012-06-15			2013-02-23			2013-04-05		
		AR@1	AR@5	AR@20	AR@1	AR@5	AR@20	AR@1	AR@5	AR@20	AR@1	AR@5	AR@20
1	5.46	0.924	0.943	0.955	0.781	0.826	0.860	0.798	0.843	0.893	0.749	0.797	0.840
2	6.49	**0.932**	**0.946**	**0.957**	**0.793**	**0.835**	**0.871**	**0.817**	**0.866**	**0.899**	**0.780**	**0.826**	**0.869**
3	7.32	0.923	0.940	0.955	0.783	0.824	0.861	0.787	0.840	0.889	0.757	0.807	0.856

一个和三个交叉 Transformer 模块的视图间Transformer。原因可能是，更多的 Transformer 块需要更多训练数据和时间才能取得良好性能。我们还在表III中展示了在 NCLT数据集的序列 2012-01-08 中生成一个全局描述符的平均运行时间。可以看出，交叉 Transformer 模块越多，提取特征所需的时间就越长。因此，我们在 CVTNet 中使用2个 Transformer 块，以兼顾效果与效率。

E. 多层输入研究

在本实验中，我们使用自录数据集验证为 CVTNet 采用多层输入的优势。我们修改了 Overlap-NetLeg 第一层的输入通道数，使其能够接收单通道输入。实验结果如表 IV 所示。可以看出，采用多层输入的 CVTNet 优于仅使用单通道输入的基线。我们提出的多层输入明确针对背景中被遮挡的激光点问题，提供了更多点云信息，从而使网络能够对来自不同子空间的数据重新加权，增强提取特征的判别能力。

为进一步验证所提出的多层输入形式的有效性，我们还将其应用于另一种网络 FusionVLAD，如表 IV 所示。多层输入形式也提升了 FusionVLAD 的性能，这表明我们提出的多层输入形式能够很好地推广到不同方法中。尽管如此，由于设计了新颖的内部与跨层Transformer网络，我们的方法仍显著优于采用多层输入的 FusionVLAD。

表 IV：多层输入研究

方法	AR@1	AR@5	AR@20
FusionVLAD（单通道）	0.865	0.923	0.960
FusionVLAD（多层）	0.868	0.930	0.970
CVTNet（单通道）	0.896	0.948	0.987
CVTNet（多层）	**0.907**	**0.951**	**0.993**

F. 偏航角不变性研究

我们已在第IV-B节中使用自录数据集展示了所提出的 CVTNet 对视点变化的鲁棒性。在本实验中，我们进一步设计了一种更直接的方法，以验证所设计的网络架构对于同一地点采集的激光雷达扫描的偏航旋转具有严格的不变性。我们将序列 2012-02-05 中的每个查询扫描沿偏航轴每隔 30 度旋转一次，并在由序列 2012-01-08 构建的数据库中搜索相同地点。我们将本方法与最先进的基线方法进行比较，并以 Recall@1 作为评判标准，展示偏航角变化对地点识别性能的影响。实验结果如图7所示。可以看出，在所有旋转条件下，CVTNet 的 Recall@1 始终优于其他方法。由于所设计的架构具有偏航旋转不变性，CVTNet 和 OverlapTransformer 均不受沿偏航轴旋转的影响。FusionVLAD 受到偏航角旋转的影响并不明显，但并不具备严格的偏航旋转不变性。相比之下，随着偏航角增大，MinkLoc3D 和 PointNetVLAD 的性能迅速下降，这表明它们的网络架构对视点变化的鲁棒性不足。

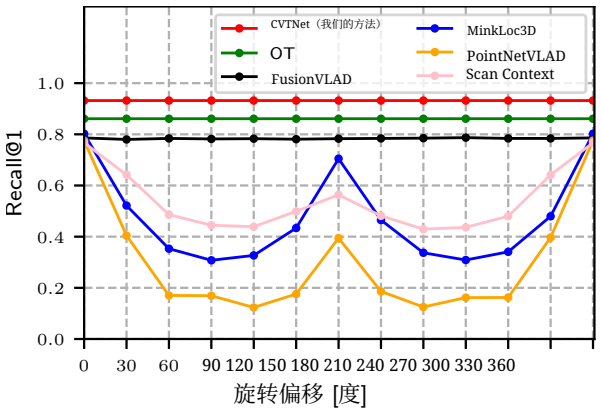

图 7：偏航角不变性研究。

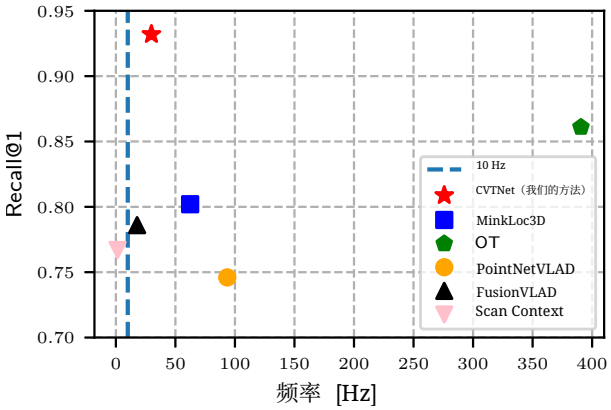

图 8：与最先进方法的运行时间比较。

G. 运行时间和内存消耗

我们进一步比较 CVTNet 与所有基线方法的实时性能，并评估其内存消耗。实验在配备 Intel i7-11700K CPU 和 Nvidia RTX 3070 GPU 的系统上进行。我们使用 NCLT 序列 2012-01-08 作为数据库，序列 2012-02-05 作为查询，计算全局描述符生成以及在20 个候选结果中检索所需的总运行时间；对于每个查询激光扫描，检索范围为数据库中的28127 个激光扫描。除 Scan Context 外，均使用 FAISS库 [31] 来加速搜索过程；Scan Context 只能使用专门用于基于平移的匹配的检索函数 [6]。图8 展示了运行频率与前1召回率的关系。CVTNet 的运行频率为29。85 Hz（总计33。50ms，其中描述符生成耗时15。03ms，检索耗时18。47ms），约为典型激光雷达传感器帧率10Hz 的 3 倍，同时实现了最先进的地点识别性能。此外，CVTNet 的参数数量为21。38M，表明我们的方法内存消耗较低。

V. 结论

本文提出了一种新颖的跨视图Transformer网络，用于基于激光雷达的地点识别。点云被投影为多层RIV和鸟瞰图，作为本方法的输入，而非传统的单通道形式。所提出的网络采用视图内Transformer来捕获各个视图内部的相关性，同时使用视图间Transformer提取不同视图之间的跨视图相关性。将本方法与最先进方法进行实验对比，结果验证了其在地点识别和回环检测方面的优越性能。此外，实验结果还表明，所提出的CVTNet能够适应视点变化，具有偏航角不变性，并可在线运行，因此能够用于实际的自动驾驶应用。

参考文献

[1] X. Chen, T. L¨abe, A. Milioto, T. R¨ohling, J. Behley, and C. Stachniss, “OverlapNet：一种用于计算激光雷达扫描相似度、并应用于闭环和定位的孪生网络”，《*Autonomous Robots*》，第 46 卷，第 61–81 页，2021 年。[2] J. Ma, S. Wang, K. Zhang, Z. He, J. Huang, and X. Mei, “通过卷积自编码器和运动一致性实现快速、鲁棒的回环检测”，*IEEE*工业信息学汇刊，第 18 卷，第 6 期，第 3681–3691 页，2022 年。[3] Z. Liu, S. Zhou, C. Suo, P. Yin, W. Chen, H. Wang, H. Li, and Y.-H. Liu, “Lpd-net：面向大规模地点识别和环境分析的三维点云学习”，载于*IEEE/CVF*国际计算机视觉会议（*ICCV*）论文集，2019 年。[4] K. Vidanapathirana, M. Ramezani, P. Moghadam, S. Sridharan, and C. Fookes, “Logg3d-net：用于三维地点识别的局部引导全局描述符学习”，载于*IEEE*国际机器人与自动化会议（*ICRA*）论文集，2022 年。[5] J. Ma, J. Zhang, J. Xu, R. Ai, W. Gu, and X. Chen, “Overlaptransformer：一种高效且偏航角不变的激光雷达地点识别Transformer网络”，*IEEE*机器人与自动化快报（*RA-L*），第 7 卷，第 3 期，第 6958–6965 页，2022 年。[6] G. Kim and A. Kim, “Scan context：用于三维点云地图中地点识别的自我中心空间描述符”，载于*IEEE/RSJ*国际智能机器人与系统会议（*IROS*）论文集，2018 年。[7] D. Cattaneo, M. Vaghi, and A. Valada, “Lcdnet：用于激光雷达SLAM的深度回环检测与点云配准”，《*IEEE*机器人学汇刊》，第 38 卷，第 4 期，第 2074–2093 页，2022 年。[8] Z. Fan, Z. Song, H. Liu, Z. Lu, J. He, and X. Du, “Svt-net：用于大规模地点识别的超轻量稀疏体素Transformer。” AAAI，2022 年。[9] J. Komorowski, “Minkloc3d：基于点云的大规模地点识别”，载于*IEEE*冬季计算机视觉应用会议（*WACV*）论文集，2021 年。[10] M. Y. Chang, S. Yeon, S. Ryu, and D. Lee, “Spoxelnet：针对拥挤室内空间三维点云的基于球形体素的深度地点识别”，载于*2020 IEEE/RSJ*智能机器人与系统国际会议（*IROS*）。IEEE，2020 年，第 8564–8570 页。[11] A. Zaganidis, A. Zerntev, T. Duckett, and G. Cielniak, “在SLAM中使用NDT直方图进行语义辅助闭环”，载于*2019 IEEE/RSJ*智能机器人与系统国际会议（*IROS*）。IEEE，2019 年，第 4562–4568 页。[12] Z. Zhou, C. Zhao, D. Adolfsson, S. Su, Y. Gao, T. Duckett, and L. Sun, “Ndt-transformer：使用正态分布变换表示进行大规模三维点云定位”，载于*IEEE*国际机器人与自动化会议（*ICRA*）论文集，2021 年。[13] H. Wang, C. Wang, and L. Xie, “Intensity scan context：对强度与几何关系进行编码以实现回环检测”，载于*IEEE*国际机器人与自动化会议（*ICRA*）论文集，2020 年。[14] L. Luo, S.-Y. Cao, B. Han, H.-L. Shen, and J. Li, “Bvmatch：使用鸟瞰图像进行激光雷达地点识别”，*IEEE*机器人与自动化快报，第 6 卷，第 3 期，第 6076–6083 页，2021 年。[15] P. Yin, L. Xu, J. Zhang, and H. Choset, “Fusionvlad：用于无视点限制的三维地点识别的多视图深度融合网络”，*IEEE*机器人与自动化快报，第 6 卷，第 2 期，第 2304–2310 页，2021 年。[16] X. Ye and J. Ma, “用于视觉地点识别的邻域流形保持匹配”，*IEEE*工业信息学汇刊，第 1–10 页，2022 年。[17] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “Netvlad：用于弱监督地点识别的CNN架构”，载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，2016 年。

[18] L. Li, M. Yang, C. Wang, and B. Wang, “面向工业车辆的混合滤波框架鲁棒定位”，*IEEE*工业信息学汇刊，第 14 卷，第 3 期，第 941–950 页，2018 年。[19] Y. Wang, Z. Sun, C.-Z. Xu, S. E. Sarma, J. Yang, and H. Kong, “用于回环检测的激光雷达虹膜”，载于*IEEE/RSJ*国际智能机器人与系统会议（*IROS*）论文集，2020 年。[20] M. A. Uy and G. H. Lee, “Pointnetvlad：基于深度点云的大规模地点识别检索”，载于*IEEE*计算机视觉与模式识别会议论文集，2018 年。[21] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet：用于三维分类和分割的点集深度学习”，载于*IEEE*计算机视觉与模式识别会议论文集，2017 年。[22] R. Dub´e, A. Cramariuc, D. Dugas, H. Sommer, M. Dymczyk, J. Nieto, R. Siegwart, and C. Cadena, “Segmap：使用数据驱动描述符进行基于分段的建图与定位”，《国际机器人研究杂志》，第 39 卷，第 2–3 期，第 339–355 页，2020 年。[23] S. Zhao, P. Yin, G. Yi, and S. Scherer, “Spherevlad++：基于注意力且信号增强的视点不变描述符”，*arXiv*预印本 *arXiv:2207.02958*，2022 年。[24] J. Cui and X. Chen, “Ccl：用于激光雷达地点识别的持续对比学习”，*IEEE*机器人与自动化快报，第 8 卷，第 8 期，第 4433–4440 页，2023 年。[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need”，载于神经信息处理系统进展会议（*NIPS*）论文集，2017 年。[26] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, “用于多视图序列学习的记忆融合网络”，载于人工智能进展会议（*AAAI*）论文集，2018 年。[27] Pham, Hai and Liang, Paul Pu and Manzini, Thomas and Morency, Louis-Philippe and P´oczos, Barnab´as, “Found in Translation：通过模态间的循环平移学习鲁棒联合表征”，载于人工智能进展会议（*AAAI*）论文集，2019 年。[28] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, “密歇根大学北校区长期视觉与激光雷达数据集”，《国际机器人研究杂志》（*IJRR*），第 35 卷，第 9 期，第 1023–1035 页，2016 年。[29] A. Geiger, P. Lenz, and R. Urtasun, “我们准备好迎接自动驾驶了吗？KITTI视觉基准套件”，载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，2012 年。[30] J. Ma, X. Chen, J. Xu, and G. Xiong, “Seqot：一种利用序列激光雷达数据进行地点识别的时空变换器网络”，*IEEE*工业电子汇刊，2022 年。[31] J. Johnson, M. Douze, and H. J´egou, “使用图形处理器进行十亿级相似性搜索”，*IEEE*大数据汇刊，第 7 卷，第 3 期，第 535–547 页，2019 年。