专题综述

自动驾驶场景多模态融合感知

上海交通大学　马超

近年来，自动驾驶相关技术逐渐成为了当前的研究热点，其目的是辅助或者代替人类进行交通工具的操控。相比于传统的人工驾驶，自动驾驶在安全性、便利性、高效性等诸多方面具有显著的优势，这对交通运输和生产安全有着重大意义。自动驾驶的技术栈包括软硬件的结合，其中软件部分包括： 1）环境定位模块； 2）环境感知模块； 3）决策规划模块。作为自动驾驶系统中至关重要的一环，环境感知算法对后续的决策和路径规划至关重要，吸引了大量研究人员的关注。在自动驾驶场景中，激光雷达和相机传感器是两种常用的环境感知传感器。一方面，激光雷达通过获取 3D 空间中的激光点云，提供高精度的定位，但是稀疏无序的点云缺少表观纹理特征。另一方面，相机提供的 RGB 图像能为目标检测提供丰富的语义信息，但是由于缺乏深度估计，很难对物体的 3D 位置进行准确的预测。

基于上述背景，在业界，不同公司针对自己的业务特点和技术理解，面向不同的自动驾驶级别选择了不同的传感器配置和算法方案，其主要分为两派：纯视觉感知方案以及多模态融合方案。以 Tesla 公司为代表的一派主张采用纯视觉感知方案，仅依靠来自多个摄像头的图像进行环境感知与决策。纯视觉方案更接近人类驾驶直觉，成本相对较低但容易受到环境中光照等因素的影响；另一派以谷歌 Waymo 为代表采用多传感器融合的方案，通过对摄像头、激光雷达等传感器数据融合进行环境感知，这使得车辆对物体的位置、距离和大小的感知更加准确，但由于多传感器融合方案要求配备更多的传感器设备并对计算芯片有更高的算力要求，因此成本较高。然而，随着技术的发展，激光雷达成本不断降低，十年间车用激光雷达价格从 10 万美元价格区间已经下降到 100 美元区间，这使得多模态融合感知算法逐渐显示出其优势。相比于纯视觉方案，使用多模态融合方案的优势主要体现在三点：

- 安全优势：人类司机驾驶仅仅依靠双眼而无需使用雷达进行辅助是各公司主张纯视觉感知方案的主要依据。然而，纯视觉方案虽然更接近人眼感知直觉，但在当前，人工智能与人类感知仍相差巨大，且正如人眼会存在误判所导致的每年有大量交通事故的发生，纯视觉算法也在许多场景下存在误判，而安全性对自动驾驶车辆至关重要。因此，自动驾驶技术首要任务是实现更安全的驾驶而非仅仅是模拟人类司机驾驶习惯，多模态方案中激光雷达的加入可以大幅度提高感知算法对障碍物的检出能力，保证更为安全可靠的自动驾驶技术的实现。
- 数据需求优势：自动驾驶场景是一高度复杂的开放环境，纯视觉感知算法通常需要使用大规模的不同种类和场景的真实驾驶数据进行训练，以使模型应对多变的驾驶场景。而激光雷达通过发射多种激光并接收反射光的方式检测障碍物，这减轻了对数据的依赖，更具可靠性与可解释性。同时，采用多模态融合方案一定程度上规避了大型公司的数据垄断带来的技术垄断，能有效促进自动驾驶技术相关研究更为快速的发展。
- 鲁棒性优势：不同传感器具有不同的优点和局限性。例如，激光雷达可以提供准确的距离信息，但在雨雪等恶劣天气下可能会受到影响。而摄像头则可以提供更丰富的视觉信息，但在低光等条件下可能会

受到影响。通过将不同传感器的信息融合起来，可以弥补单一传感器的局限性，不同传感器采集的信息可以互相补充，从而形成一个更全面、更准确的场景理解。同时，激光雷达对目标距离和速度的准确感知能力有助于自动驾驶车辆做出更为准确的驾驶决策，提高场景感知的鲁棒性。

总之，多模态融合方案可以利用不同传感器间的互补性，弥补单一传感器的局限性，提高场景感知的准确性和可靠性，并且更好地适应复杂场景，因此在自动驾驶等场景感知应用中具有重要的优势。 3D 目标检测和跟踪算法能够对环境中的物体进行分类定位和关联，是环境感知的核心环节。目前自动驾驶场景下的多模态融合感知算法主要集中在研究相机与激光雷达两种模态下的环境感知。有效地融合这两种互补的模态，不仅能提升检测任务的精度，也能够丰富跟踪任务中的匹配线索。

一、多模态融合3D检测

1.1. 多模态数据融合

如图 1 所示，目前已有的基于多模态 3D 检测方法，根据两种传感器的融合阶段可以分为：1）结果层级的融合，2）候选框层级的融合和 3）点云层级的融合：

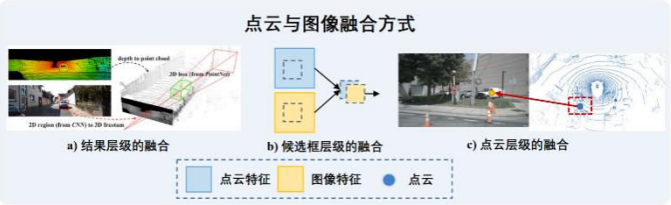


图 1　三种点云与图像融合方式

结果层级的融合方法直接利用 2D 图像上所获取的二维候选框区域与对应的点云特征进行融合。2018 年，Qi 等人提出了 Frustum PointNets[1]。该方法首先直接利用已有的 2D 目标检测器在图像上获取 2D 候选框，并将对应到 2D 候选框内的点云视锥（frustum）裁剪出来。接着将视锥的点云经过一个类似 PointNet 的模块获取分割信息并过滤前景点以进一步用于 3D 候选框的回归。F-ConvNet[2]在此基础上进行了改进，实现了更细粒度和多层级的点云特征提取，从而获得更优的检测性能。Frustum PointNets 系列方法借助成熟的 2D 检测算法提供一定程度上的先验知识，从而减小了 3D 搜索空间。然而，级联方法的缺点是它严重依赖于 2D 检测器提供候选框从而导致漏检情况。

基于候选框的方法在候选框层面上以 ROI pooling 的形式融合点云与图像特征，从而进一步的优化候选框。MV3D[3]和 AVOD[4]是基于候选框融合的典型方法。为了更好的对特征进行融合，如图 2 所示，工作[5]提出一两阶段融合框架，在第一阶段通过分别提取图像与点云特征并利用一种联合 anchor 机制生成区域候选框，在第二阶段进一步融合生成的 2D 与 3D 候选框中的密集特征。尽管基于候选框的融合方法更好的利用了不同模态的信息，但这类方法往往存在速度慢且计算量大的问题。

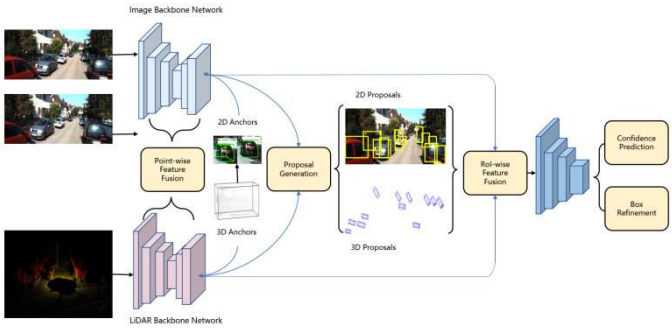


图 2　相机与激光点云两阶段融合框架[5]

点云层级的融合则是直接建立点与图像的关联，将图像信息增强到点云上，作为点云的额外输入特征。其中一部分方法[13-16]的目标是构建一个 BEV 的图像特征，在和点云特征结合后，在鸟瞰图上做 3D 检测，这种直接构建 BEV 特征的思路常常会引入特征模糊。目前在多模态融合方法上，采用更多的是点到点的思路，即直接将点云投影到图像上并抓取相应的图像特征后以点对点的方式直接增强到点云上。这种方式直接建立了点云与图像的联系，减少了信息的损失。PointPainting[6]利用来自图像的语义分割信息来增强点云的输入。如图 3 所示，该方法首先利用 2D 分割网络对每个图像像素进行分类，然后将点云直接投影到分割掩码中，将对应的分割分数作为图像特征附加到每个点云上。最后，将“绘制”后的点云部署到任意纯点云 3D 检测器中用于定位和分类。

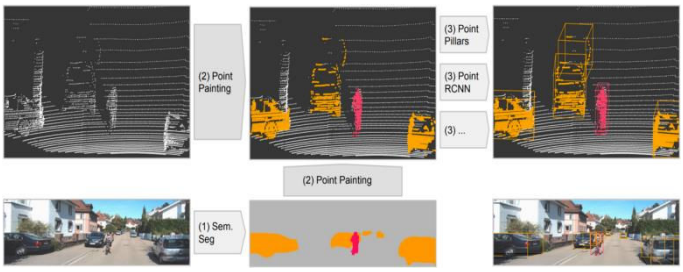


图 3　PointPainting 算法框架图[6]

尽管 PointPainting 取得了明显的改进，但是分割信息只为点云提供了类别，并没有充分利用图像信息，可以认为是一种紧致但次优的图像表征。直观上，图像的高维 CNN 特征包含了更丰富的外观纹理信息和更大的感受野，应该更适合与点云融合。因此，PointAugmenting[7]方法选择图像特征作为增强点云的输入，模型的网络结构如图 4 所示。该方法在点云检测器 Centerpoint[17]的基础上，主要做了两点改进用于融合图像和点云信息：首先，基于先前的结论，将点云投影到图像上，抓取点云相应的图像特征并增强到点云上；此外，考虑到点云与图像特征之间存在数据特性与数据分布的差异，在 3D backbone 中额外添加了一个图像分支，用于处理图像特征，最后两个模态在 BEV 特征上进行融合。

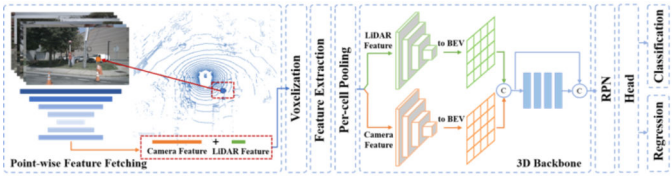


图 4　PointAugmenting 算法框架图[7]

1.2. 多模态数据增广

另一方面，数据增强是视觉感知领域提升感知精度的最为重要的手段，而由于模态间数据的差异，大部分数据增强方法无法直接迁移到多模态场景。例如，在点云检测器的训练中，常常采用 GT-Paste[8]增强方法将其他场景中的物体粘贴到当前训练场景。GT-Paste 可以缓解数据集的类别不平衡问题，并加速模型的收敛。但是这种有效的数据增强方式并不能直接迁移到多模态的场景中，因为这种粘贴方式会破坏点云和图像之间的一致性关联。因此 PointAugmenting 同时提出了一种数据增强方式，能够使得 GT-Paste 适用于多模态的检测器。如图 5(a)和(d)所示，汽车存在于原始场景中，它的点云由黄色表示；骑行者是当前想要粘贴到当前场景的虚拟物体，它的点云由绿色表示。从观察者的角度来看，由于粘贴的自行车在原始 3D 场景中被汽车部分遮挡，导致在图像视角上，两个物体的图像块存在部分重叠。如果直接将虚拟物体的图像块粘贴到图像上，如图 5(b)和(e)所示，则在图像块的重叠区域中，投影的物体点云可能会获取不匹配的图像信息。为解决此问题，PointAugmenting 保持物体之间的遮挡关系，并从观察者的角度过滤那些被遮挡的点云。对于图像数据，所有虚拟物体和原始物体将按照由远到近的顺序，将其对应的图像块粘贴到原始图像上，从而与点云的遮挡关系保持一致。

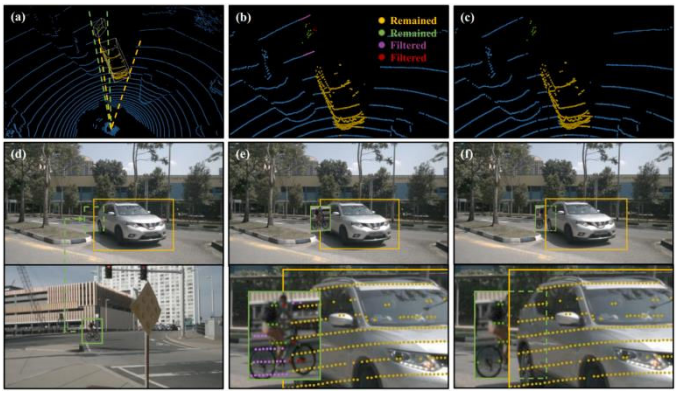


图 5　多模态数据增强示例

基于上述信息融合与增强方法，PointAugmenting 在 nuScenes 和 waymo 数据上都取得了明显的精度提升，在 nuScenes 上达到了同期最好的检测结果，比纯点云的基线算法 CenterPoint 精度提升了 6.5 个点。

1.3. 时序多模态融合

时序多帧数据为多模态融合感知提供多视角、运动特征等额外信息，能够进一步提升物体检测精度。目前大多数方法对时序的利用是将来自不同帧的输入投影到同一关键帧中，以起到增强点云稠密度的作用。但是简单的投影合并在运动场景中存在明显的拖尾效应，因此只适合融合时间跨度小的非关键帧。为了获取更长时间序列中更丰富的信息，设计合理的时序数据的融合方

案是非常有意义的。尽管最近的工作[9]对学习时序多模态模型进行了早期尝试，但实际上，它是使用点连接的预处理方案进行时间融合，也就是把时序信息和多模态信息融合的建模视为独立的两个部分。相比之下，对时序多模态信息进行显式的融合建模的方法更有利于充分利用未对齐的互补信息。

为了解决上述问题，工作[10]提出了一种时序多模态融合模型(LiDAR Image Fusion Transformer，LIFT)，可以直接学习四维时序多模态信息的相互对齐。具体来说，如图 6 所示，所提出的模型包含一个格状特征编码器和一个跨模态垮时间注意力模块。首先，在格状特征编码器中，LIFT 获取对应点的相机特征并进行柱状特征提取以将激光雷达点和逐点相机特征投影到鸟瞰图表示上。通过保持相对较少数量的网格，LIFT 能够有效地计算网格间的相互交互和网格内的细粒度注意力。同时，该项工作通过设计一种 4D 位置编码来进行时序多模态数据的四维位置定位。进一步地，为了减少自注意力模块的计算开销，LIFT 还设计了稀疏的窗口分区机制进一步舍弃不含激光点的空窗口区域来降低计算量，并构建了金字塔上下文结构以扩大特征感受野。

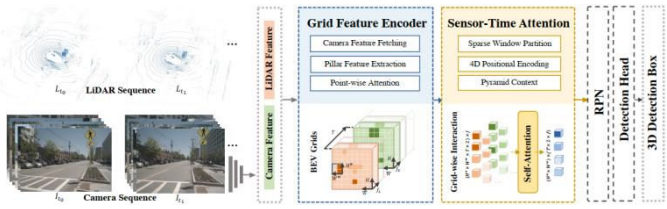


图 6　时序多模态融合 3D 目标检测 LIFT 框架[10]

在之前单帧的点云检测模型中，一种基于随机物体粘贴的数据增强方案因为能够增强训练数据的多样性，被验证对模型训练很有帮助。但是在时序多模态数据中，普通的粘贴方案无法保持跨模态和跨时序的一致性，因此如图 7 所示，LIFT 扩展了传统数据增强方案，把单个点云物体的粘贴扩展为序列点云和对应的序列图片块的粘贴，从而保证了粘贴在时序多模态训练数据中的增强物体保持了时空一致性。在 nuScenes 数据集上进行的大量定性与定量实验表明时序信息对多模态检测的促进作用。

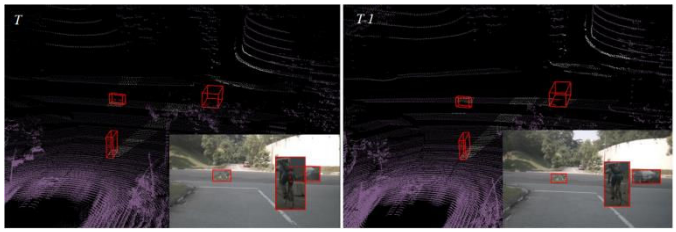


图 7　跨模态垮时序数据增强可视化

二、多模态融合3D跟踪

通常在检测跟踪框架中，检测结果与轨迹被表示成和位置有关的点，并假设连续帧之间的位置变化在局部区域内，进而用点位置距离进行检测结果和轨迹的关联计算。尽管取得了不错的效果，由于缺乏丰富的表观信息，这种仅依赖位置进行匹配的方法在较大运动变化和存在噪声检测结果的情况下往往会失败。多模态 3D 目标检测与跟踪算法 AlphaTrack[11]同时考虑 3D 物体位置和表观变化。如图 8 所示，该方法用独立的 3D 卷积网络分别提取两个模态的特征，并在点级别结合激光雷达点云和对应的图像特征；其次，为了获取 3D 对象的表观信息，以在跟踪算法中能够显式地利用表观线索，该方法为模型附加了一个表观分支以学习实例级的表观特征。最后，该工作进一步提出了一个三阶段跟踪算法以在算法中同时利用位置线索和表观线索进行匹配关联。第一阶段是基于位置的匹配，以中心点距离作为位置相似度初步匹配检测和跟踪轨迹；然后 AlphaTrack 以表观特征的余弦距离作为表观相似度对第一阶段的结果进行合理性筛选，即将相似度排序靠后的匹配对予以删除，最终在第三阶段中用表观特征对剩余未匹配对进行重匹配。通过以上三个步骤，实现了在跟踪关联中显式地利用位置和表观两种信息。AlphaTrack 在 nuScenes 数据集 3D 跟踪上取得同期最佳结果（在排行榜上占据榜首位置长达一年）。

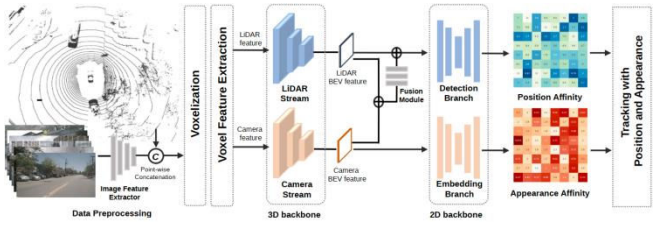


图 8　AlphaTrack 整体架构[11]

三、通用跨模态知识蒸馏

基于上述相关研究工作背景可以发现，在 3D 目标检测器中，单模态检测器模型简单但检测精度较低，而多模态检测器检测性能好但系统复杂度高。跨模态知识蒸馏方法可以根据已有模型进行知识迁移以提高目标模型检测精度，然而现有的跨模态知识蒸馏框架限定了教师与学生检测器的模态，即限定了教师和学生的模态为 LiDAR 和 camera 以及 Fusion 和 LiDAR，而现实中不同检测器都有广泛应用场景，因此应用范围受到了限制。从以上问题出发，构建统一的知识蒸馏框架，对教师和学生的任意模态组合均适用变得越来越重要。通过对不同模态检测器的结构进行分解寻找一致的中间特征表示，可以发现现有的表现较好的检测器通常会在 BEV 视角下进行检测，并且具有统一的流程。如图 9 所示，检测器首先对输入数据提取特征，并将特征投影到 BEV 视角下，得到 low-level 的 BEV 特征，然后对该特征进一步处理，得到 high-level 的 BEV 特征，最后由一个检测头进行预测，输出 response 特征用于生成预测结果。基于以上的统一分解形式，对于不同模态的教师和学生检测器，就可以在这些一致的中间特征表示上进行知识蒸馏。

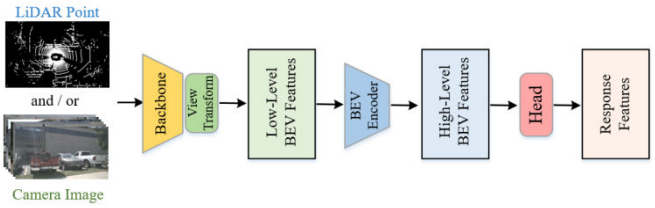


图 9　3D 目标检测器通常遵循统一的流程

针对上述背景，工作[12]提出一通用跨模态知识蒸馏框架 UniDistill。如图 10 所示，该方法提出三种蒸馏损失，分别为特征蒸馏(feature distillation)、关系蒸馏(relation distillition) 以及响应蒸馏(response distillation) 模块。首先，为消除背景影响以及平衡不同尺度的检测框对蒸馏损失的占比，在特征蒸馏与关系蒸馏中，UniDistill 对每一个 GT 检测框只选取了关键的 9 个点以分别对 low-level 的 BEV 特征以及余弦相似度进行对齐。同时，由于在检测框中心的预测值较为准确，因此对于响应蒸馏部分，UniDistill 针对每一个 GT 检测框中间的一个高斯掩膜区域的特征进行了对齐。

在 nuScenes 数据集对四种教师与学生的模态组合方式 (Fusion→LiDAR/Camera, LiDAR→Camera, Camera→LiDAR)进行验证，实验结果表明 UniDistill 对提高学生检测器性能的优异表现。

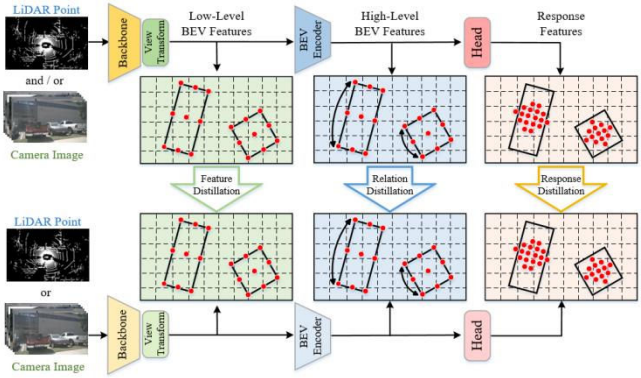


图 10　UniDistil 算法整体框架[12]

四、总结与展望

本文首先从当前自动驾驶的感知模块选用纯视觉感知方案还是多模态融合方案这一热点问题出发，对视觉感知中多模态融合方案的优势进行了探讨。围绕自动驾驶场景的多模态融合感知，分别介绍了多模态融合 3D 检测与跟踪领域的当前研究进展，并介绍了跨模态知识蒸馏方法帮助检测模型提高精度的进展。基于目前已取得的实验结果和结论，笔者认为多模态融合感知领域还存在较大的深入研究空间，比如以下四个方向：

- 数据融合算法的优化：现有的跨模态融合方案大都是基于投影关系进行的，这样的融合方案虽然在实际应用中是计算简单、并验证有效的，但考虑到不同传感器数据之间存在一定的标定误差，投影方案不可避免地也会引入一定的误差，此外不同模态之间的信息可靠度在不同的应用环境下也是不稳定的。因此如何进一步优化跨模态数据的融合是非常有价值的，具体来说，可以考虑以下两种优化方向：1) 优化投影关系的准确度。传感器间的投影误差是有可能通过深度学习方案缓解的；2) 优化模态间的信息衡量标准。不同模态的特征可靠性是可以通过一定的衡量标准进行评估的，以此提供在不同环境下更合理融合的参考。

- 基于主动学习、迁移学习、自监督学习等方案的数据标注优化研究：3D 数据的标注对于感知模型的训练是至关重要的，然而精确的 3D 数据标注需要高昂的人力成本。因此存在两种降低标注成本的方向，一是降低标注的数目，即在无标注或部分标注的情况下也能取得性能优秀的感知模型；二是改善粗糙标注的质量，即通过深度学习方案自主提升低质量标注的精度，以降低精确标注的成本。其中主动学习作为一种自动挖掘错误标注、难标注的方案，在辅助进行高质量标注的过程中能够提供一种低成本的技术支持。此外迁移学习能够帮助模型提升在无标注数据域上的泛化性能，更轻便简洁的迁移方案对实际应用具有重要研究价值。
- 感知与决策的耦合：在自动驾驶系统中，感知和决策是紧密耦合的。多模态感知提供了丰富的环境信息，但同时也增加了感知与决策之间的耦合程度。为了实现高效的自动驾驶，需要在感知和决策之间建立有效的信息传递机制。使得感知与决策模块相互配合，这方面的研究可分为两个方向：1）研究如何将多模态感知的结果进行有效的融合和分析，以支持决策模块的实时决策，这需要形成对道路拓扑结构、高精度地图以及动态障碍物的联合感知与理解；2）研究统一的端到端感知与决策联合预测模型。对上述方向的推进有助于实现简洁高效的自动驾驶统一框架。
- 多模态感知的安全性和可靠性：自动驾驶技术是一个非常复杂的系统，需要确保其在各种情况下都能够安全、可靠地运行。因此，在设计和优化多模态感知算法时，必须考虑到各种异常情况和边界条件，并采取相应的措施以确保系统的安全性和可靠性。在自动驾驶中，不确定性来源于多个方面，包括传感器的噪声、遮挡、动态环境等，这可能会导致自动驾驶系统出现错误的决策，甚至造成事故。为了实现稳定可靠的自动驾驶，需要研究如何在多模态感知中有效地管理不确定性。因此，需要采用多种冗余机制来确保系统的可靠性和安全性。这方面的研究可以从两方面进行：1）研究如何在数据融合过程中考虑不确定性；2）研究如何在感知和决策之间传递不确定性信息，以支持决策模块在不确定性条件下进行决策。

随着自动驾驶技术的不断发展和应用，多模态感知技术作为其中最核心的一部分，也将不断得到完善和优化。目前，该领域仍然存在的亟待解决关键问题具有重要研究价值。通过不断的创新和发展，自动驾驶系统将实现更高的环境感知精度和决策能力，进而实现高效、精准、可靠和安全的环境感知技术。

责任编委 王金甲

参考文献

[1] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, Leonidas J. Guibas. Frustum pointnets for 3d object detection from rgb-d data, CVPR 2018.

[2] Wang, Zhixin, and Kui Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection, arXiv:1903.01864, 2019.

[3] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, Tian Xia. Multi-view 3d object detection network for autonomous driving, CVPR 2017.

[4] Ku J, Mozifian M, Lee J, et al. Joint 3d proposal generation and object detection from view aggregation, IROS 2018.

[5] Ming Zhu, Chao Ma, Pan Ji, Xiaokang Yang. Cross-Modality 3D Object Detection, WACV 2021.

[6] Vora S, Lang A H, Helou B, et al. Pointpainting: Sequential fusion for 3d object detection, CVPR 2020.

[7] Chunwei Wang, Chao Ma, Ming Zhu, Xiaokang Yang. PointAugmenting: Cross-Modal Augmentation for 3D Object Detection, CVPR 2021.

[8] Yan, Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. Sensors, 18(10):3337, 2018.

[9] Piergiovanni A J, Casser V, Ryoo M S, et al. 4D-Net for Learned Multi-Modal Alignment, ICCV 2021.

[10] Yihan Zeng, Da Zhang, Chunwei Wang, Chao Ma, et al. LIFT: Learning 4D LiDAR Image Fusion Transformer for 3D Object Detection, CVPR 2022.

[11] Yihan Zeng, Chao Ma, Ming Zhu, Zhiming Fan, and Xiaokang Yang, Cross-Modal 3D Object Detection and Tracking for Auto-Driving, in IROS 2011

[12] Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, Chao Ma, UniDistill: A Universal Cross-Modality Knowledge Distillation Framework for 3D Object Detection in Bird's-Eye View, CVPR 2023.

[13] Ming Liang, Bin Yang, Shenlong Wang, Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection, ECCV 2018.

[14] Ming Liang, Bin Yang, Yun Chen, Rui Hu, Raquel Urtasun. Multi-task multi-sensor fusion for 3d objectdetection, CVPR 2019.

[15] 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, arXiv:2004.12636, 2020.

[16] Tianwei Yin, Xingyi Zhou, Philipp Krahenbuhl. Center-based 3d object detection and tracking, CVPR 2021.

马超

上海交通大学人工智能研究院长聘轨副教授，博士生导师。上海交通大学与加州大学默塞德分校联合培养博士。2016 至 2018 年澳大利亚机器人视觉研究中心(阿德莱德大学)博士后研究员。中国图象图形学学会优博，上海市浦江人才。主要研究方向是多模态物件检测与跟踪。担任中国图象图形学学会青年工作委员会副秘书长、中国图象图形学学会优博俱乐部轮值主席。谷歌学术总引用近 1 万次，自 2020 年起连续入选爱思唯尔中国高被引学者。研究成果应用于华为达芬奇芯片及其无人驾驶 MDC 平台，获华为技术合作领域 2021 年度优秀技术成果奖。

Email: chaoma@sjtu.edu.cn

专题综述

重新思考点云配准中的生成和选择过程

陈志[1]，孙琨[2]，杨帆[1]，郭琳[1]，陶文兵[1]
[1] 华中科技大学 [2] 中国地质大学（武汉）

一、摘要

误匹配消除是基于特征的点云配准方法的重要步骤。本论文重新思考了基于传统的 RANSAC 方法中的模型生成和模型选择过程。具体来说，对于模型生成，本文提出了一种新的二阶兼容性度量 (SC^2) 来计算匹配对之间的相似性。本文从概率的角度证明了该度量能够极大地降低模型生成中采样一致性集合时采样到误匹配的概率，从而提升模型生成过程中的采样稳定性。对于模型选择，本文提出一种特征和空间一致性引导的截断倒角距离 (FS-TCD) 来评估生成模型的质量。它综合地考虑了全局对齐质量、几何信息和特征信息，缓解了现有度量过度依赖特征匹配准确性的问题。所提出的方法在室内以及室外数据集上取得了当前最好的配准性能。论文已被 TPAMI 2023 收录，代码已开源：https://github.com/ZhiChen902/SC2-PCR-plusplus。

二、引言

三维刚体点云配准旨在恢复两个具有重叠区域的点云之间的刚体变换。常用的基于特征的方法先为点云中的每个点提取局部特征描述子并建立粗略的点云对应关系，然后通过稳健的模型估计算法，从含有误匹配（外点）的粗匹配关系中寻找正确匹配（内点）并估计刚体变换。本研究主要关注如何在粗匹配中含有大量错误匹配的情况下进行点云配准。

RANSAC[1]最早采用迭代的策略来进行模型估计。然而，它需要大量的采样来保证算法的收敛，并且在内点率过低的情况下并不能保证一定能找到正确解。一些方法通过空间兼容性来解决 RANSAC 的问题，它们利用变换的刚体属性，即：空间中任意两个点经过刚体变换之后长度不变。因此，一阶空间兼容性度量认为如果两个匹配对之间的空间距离差越小，他们的相似度越大。由于正确匹配对（内点）之间的空间距离差理论上应为 0，这样两个内点的相似性就会比较大，从而在内点之间形成聚类效应来方便采样。

然而一阶空间兼容性存在模糊性，即外点也有可能与内点有很高的相似度，如图 1(b) 中黄色底纹的方格。当前许多研究利用传统方法[2,3]或深度学习框架[4,5]来减轻模糊性带来的问题，虽然一定程度上缓解了这一问题，但是在内点率很低的情况下有时也会失效。为了解决这一问题，我们提出一个二阶兼容性测度来度量两个匹配对之间的相似性。具体来说，我们首先二值化一阶空间

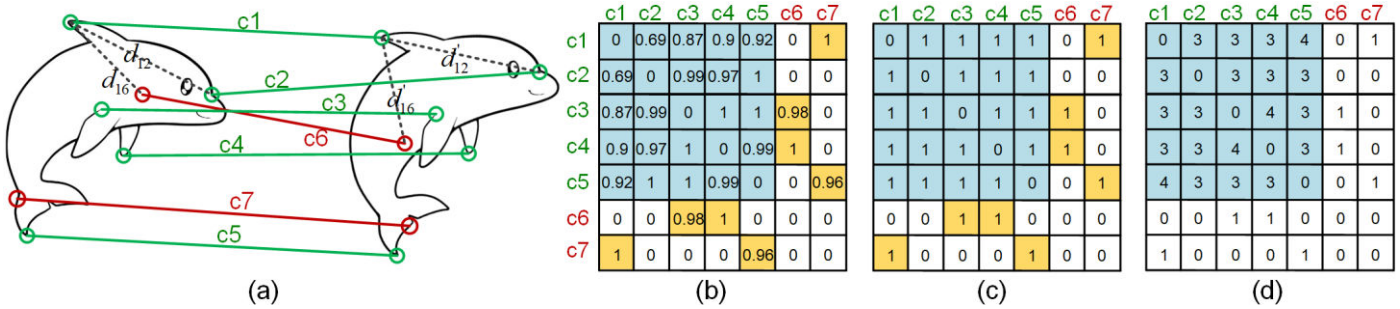


图 1　二阶空间兼容性动机及示意图

兼容性测度（1 代表两个匹配对是兼容的，0 代表不兼容），如图 1（c）。然后对于任意两个兼容的匹配对，我们计算与他们共同兼容的匹配对的个数（可以表示为兼容性的二阶形式）作为他们的相似性。由于内点之间都是相互兼容的，因此任意两个内点之间的相似度至少为所有匹配对中内点的个数（除去这两个匹配对自身），而内点和外点之间并没有这样的性质。如图 1(d)所示，在所提出的二阶兼容性度量中，内点和外点之间的高相似性被抑制了。我们从概率的角度证明了二阶兼容性度量发生模糊性事件的概率远小于一阶度量，证明了它能够更稳定地用于内点采样，从而提升模型生成的效率和鲁棒性。

除了模型生成，另一个问题是如何从生成的多个模型中选择出最好的模型。当前的方法通常采用内点计数（IC）的方式，即利用生成的模型对齐两个点云，然后统计对齐误差小于某一阈值的匹配点对个数作为 IC 值，并选择 IC 值最大的模型作为最终结果。然而，这种选择方式依赖初始匹配对的准确性。当初始匹配中准确匹配对个数较小时，即使是估计出了正确的模型，它对应的 IC 值可能也比较小，使得它无法被选择为最终的结果。为了解决这个问题，本文提出了一种特征和空间一致性引导的截断倒角距离 (FS-TCD) 作为模型选择的度量。它从倒角距离这种全局度量出发，通过引入特征信息和空间一致性信息的引导，解决了直接将倒角距离作为模型选择度量的低效和不稳定因素，同时保留了倒角距离的全局性。由于全局信息的引入，该度量也缓解了 IC 度量对于初始匹配对的依赖性。

基于所提出的二阶兼容性 (SC^2) 和特征和空间一致性引导的截断倒角距离（FS-TCD），本文重新构建了一个点云稳健姿态估计方法。

三、二阶兼容性分析

为了分析用于采样的度量的有效性，我们定义了一个模糊性事件的概率：

$$\mathrm{P}_{am}(M) = \mathrm{P}\left(M_{in,out} > M_{in,in}\right)$$

其中M是一个具体的度量。$M_{in,out}$是内点和外点之间的相似度，而$M_{in,in}$是两个内点之间的相似度。P()表示一个事件发生的概率。当$M_{in,out} > M_{in,in}$ 时，外点就会变

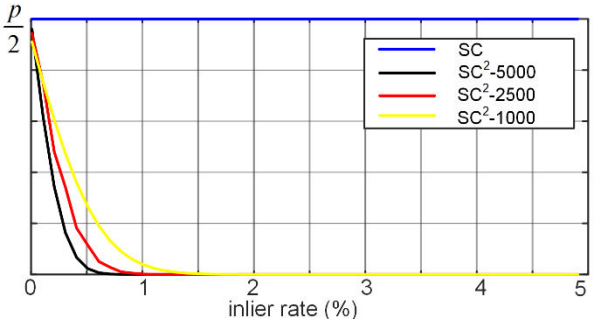


图 2　一阶和二阶度量模糊性概率对比（SC^2-N,N=5000,2500,1000 为匹配对个数）

成内点在该度量下的近邻，从而基于度量的采样就越不稳定。因此，概率值越小，基于该度量的采样越稳定。

传统的一阶兼容性度量 SC 通常被定义为如下形式：

$$SC_{ij} = \phi\left(d_{ij}\right), d_{ij} = \left| d\left(x_i, x_j\right) - d\left(y_i, y_j\right)\right.$$

其中ø为一个单调递减函数，d_{ij}为两个匹配对之间的距离差。本文提出的二阶兼容性 (SC^2) 具有如下形式：

$$C_{ij} = \begin{cases} 1, d_{ij} \leq d_{thr} \\ 0, d_{ij} > d_{thr} \end{cases},$$

$$SC_{ij}^2 = C_{ij} \cdot \sum_{k=1}^{N} C_{ik} \cdot C_{kj}.$$

其中C_{ij}是将d_{ij}经过阈值化过后的{0,1}值。二阶空间兼容性统计了任意两个兼容的匹配的共同兼容匹配对个数。

为了对比传统的一阶兼容性和提出的二阶兼容性用于模型生成时的鲁棒性，我们分别推导了它们的模糊性概率公式，并做出了他们的分布图，如图 2 所示。从图中可以看出，所提出的方法可以大大降低模糊性概率值，从而提升采样的稳定性。

四、特征和空间一致性引导的截断倒角距离

在 RANSAC 中，在生成大量模型后，通常通过内点计数（IC）的方式选择最好的模型。IC 度量的计算方式为：通过估计的模型对齐初始匹配对，并统计成功对齐的匹配对个数。然而，这种方法受限于初始匹配对的准确率。也就是说，即使是正确的模型，它的 IC 值最大为初始匹配对中正确匹配对的个数。图 3（a）展示了两对待匹配点云以及正确匹配对的个数，图 (b) 是一个错误估计的模型，而图 (c) 是一个正确估计的模型。然而，由于初始匹配对中正确匹配的数量过少，正确模型的 IC

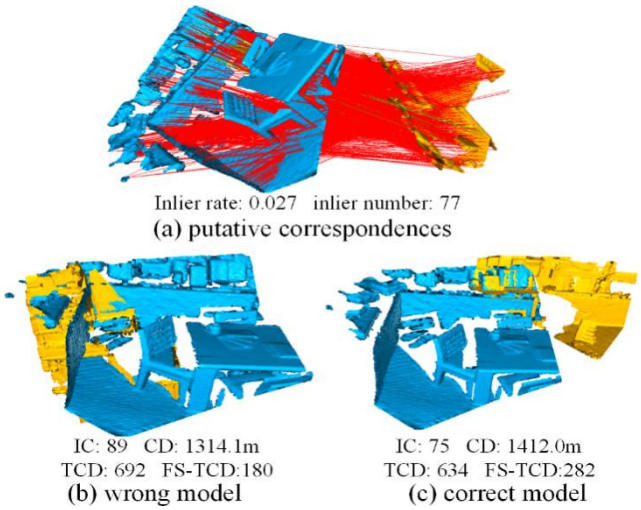


图 3　不同的模型选择度量对比（IC、TCD 和 FS-TCD 越高越好，CD 越低越好）

值只有 75，而错误模型的 IC 值反而偶然地大于正确模型的 IC 值，使得正确模型无法被选择出来。

因此，我们考虑利用倒角距离（CD）这种全局度量作为模型选择依据。然而，直接应用 CD 是不行的，因为它需要在全局范围内搜索近邻，使得它用于大量的模型选择时比较低效。同时，它没有考虑点云的部分重叠问题。为了解决这一问题，我们首先将 CD 改写成截断的形式（TCD），使得它不考虑非重叠区域的对齐质量。然后，我们用特征和空间一致性来引导倒角距离的搜索空间，这种方式既能缩短 CD 的搜索时间从而提升效率，又能通过特征和空间一致性的引导来减少 CD 的偶然误差。从图 3 中可以看出，所提出的 FS-TCD 对正确模型和错误模型有更好的区分度。

五、总体方法

基于所提出的二阶兼容性（SC^2）和特征和空间一致性引导的截断倒角距离（FS-TCD）度量，我们设计了一个高效的配准算法，名为 SC^2-PCR++。方法流程图如图 4 所示。具体来说，方法的输入是待匹配的点云对以及为它们分别提取的描述子。在模型生成步骤中，我们首先通过一对一匹配建立初始匹配集合，然后为它们构建逐匹配的 SC^2 矩阵。然后，我们通过谱分解的方式结合非极大值抑制选择一些可靠的匹配作为种子点，目的是减少采样次数来提升算法效率。在这之后，我们通过一个两阶段的采样方式为每个种子点构造一个一致性集合。在第一阶段，我们选择与每个种子点相似度得分最高的 K1 个匹配构建一个局部相似度矩阵。在第二个阶段中在局部集合中做进一步筛选来剔除潜在的误匹配并保留 K2 个匹配对。最后，我们通过局部谱匹配的方式结合可微的奇异值分解来为每一个种子点的一致性集合求一个刚体变换。

在模型选择阶段，我们首先构建一对多的匹配关系矩阵用作后续计算 FS-TCD 的引导。尽管 FS-TCD 相比于 CD 十分高效，但是如果为每个种子点生成的刚体变换都计算一个 FS-TCD 度量仍然十分耗时。因此，我们采用 IC 度量首先对生成的刚体变化做一个过滤。我们首先为每个变换计算 IC 值，并选择出 IC 值最高的N_S个作为候选模型，最后通过计算 FS-TCD 并选择得分最高的最终的结果。

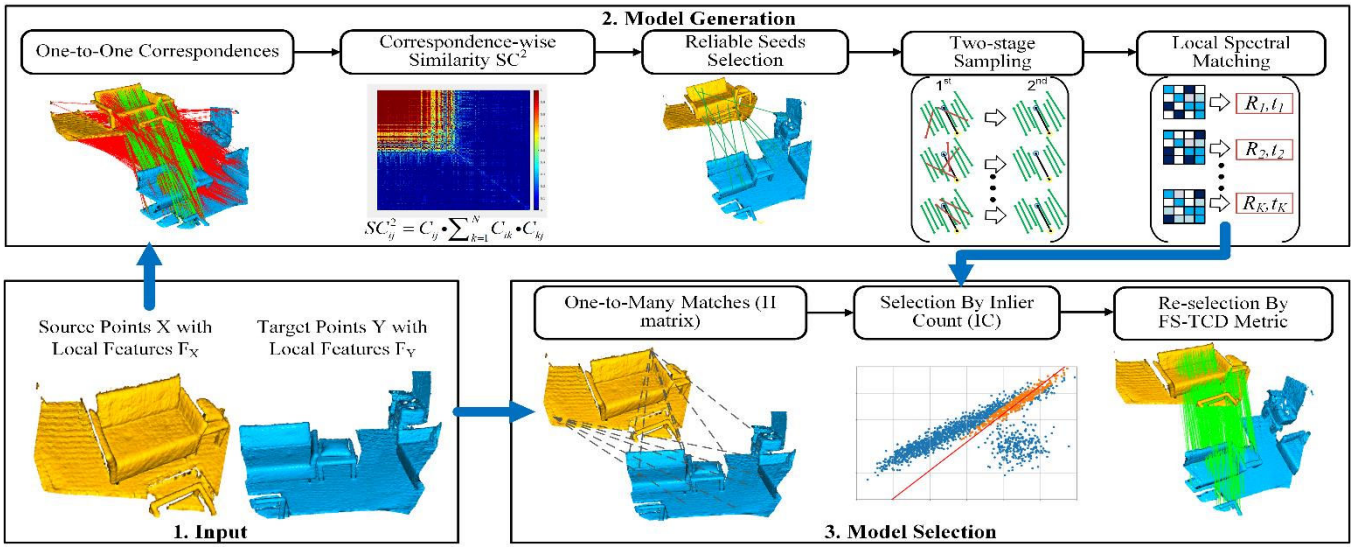


图 4　方法总体流程

表 1　在室内配准数据集 3DMatch 的结果

	FPFH [7] (traditional descriptor)						FCGF [8] (learning-based descriptor)						
	RR(%) ↑	RE(deg) ↓	TE(cm) ↓	IP(%)↑	IR(%)↑	F1(%) ↑	RR(%) ↑	RE(deg) ↓	TE(cm) ↓	IP(%)↑	IR(%)↑	F1(%) ↑	Time (s)
OM-Net* [12]	-	-	-	-	-	-	35.90	4.16	10.50	-	-	-	**0.08**
RegTR* [13]	-	-	-	-	-	-	92.00	**1.57**	**4.90**	-	-	-	0.18
DGR [14]	32.84	2.45	7.53	29.51	16.78	21.35	88.85	2.28	7.02	68.51	79.92	73.15	1.53
DHVR [5]	67.10	2.78	7.84	60.19	64.90	62.11	91.93	2.25	7.08	80.20	78.15	78.98	3.92
PointDSC [4]	77.57	**2.03**	**6.38**	68.45	71.56	69.75	92.85	2.08	6.51	78.91	86.23	82.12	0.10
FGR [2]	40.91	4.96	10.25	6.84	38.90	11.23	78.93	2.90	8.41	25.63	53.90	33.58	0.89
TEASER [15]	75.48	2.48	7.31	73.01	62.63	66.93	85.77	2.73	8.66	**82.43**	68.08	73.96	0.07
GC-RANSAC [16]	67.65	2.33	6.87	48.55	69.38	56.78	92.05	2.33	7.11	64.46	**93.39**	75.69	0.55
RANSAC-4M [1]	66.10	3.95	11.03	64.27	59.10	61.02	91.44	2.69	8.38	78.88	83.88	81.04	2.86
CG-SAC [3]	78.00	2.40	6.89	68.07	67.32	67.52	87.52	2.42	7.66	75.32	84.61	79.90	0.27
SC^2-PCR [9]	83.98	2.18	6.56	72.48	78.33	75.10	93.28	2.08	6.55	78.94	86.39	82.20	0.11
SC^2-PCR++	**87.18**	2.10	6.64	**76.49**	**81.72**	**78.82**	**94.15**	2.04	6.50	80.57	87.69	**83.71**	0.28

六、实验结果

表 1 展示了所提出的方法与其他传统方法和深度学习方法在室内数据集 3DMatch[6]上的对比结果。我们报告了配准召回率（RR）、平均旋转误差（RE）、平均平移误差（TE）、匹配准确率（IP）、匹配召回率（IR）和匹配 F1 分数等评价指标。其中配准召回率为最重要的指标，因为它直观地反应了姿态估计正确的点云对的比例。为了更加全面地对比误匹配消除性能，我们分别将所有的误匹配消除方法与传统描述子 FPFH[7]和深度学习描述子 FCGF[8]结合。除了对比了一些误匹配消除方法，我们也对比了一些端到端的点云配准方法（带*号的方法）。由于这些方法不需要匹配，因此对于这些方法我们没有报告与匹配准确性相关的指标（IP，IR，F1）。为了更清晰地展示所提出方法的性能，我们还对比了我们方法之前的版本（SC^2-PCR[9]），该方法对应的工作被 CVPR2022 录用。从表中的结果可以看出，SC^2-PCR 和 SC^2-PCR++取得了当前最好的效果，无论是与传统的点云描述子 FPFH 结合还是与深度学习描述子 FCGF 结合。SC^2-PCR++相比于 SC^2-PCR 也取得了明显的性能提升。效率方面，SC^2-PCR 取得了和深度学习方法相当的效率。SC^2-PCR++效率低于 SC^2-PCR，这是由于 SC^2-PCR++引入了更加精细的模型选择策略，增加了一定的时间消耗。考虑到性能的显著提升，增加的时间消耗也是可以接受的。

表 2 展示了所提出的方法与其他传统方法和深度学习方法在 3DLoMatch[10]的对比结果。3DLoMatch 由重叠率较低的待匹配点云对组成，比较具有挑战性。我们分别选取了当前三种比较有代表性的基于深度学习的描述子与误匹配消除方法结合，包括 FCGF[8]，

表 2　在低重叠率配准数据集 3DLoMatch 上的结果

	FCGF [8]						
	RR↑	RE↓	TE↓	IP↑	IR↑	F1↑	Time(s)
DHVR [5]	54.41	4.14	12.56	41.96	38.60	39.22	3.55
DGR [14]	43.80	4.17	10.82	42.22	38.96	39.05	1.48
PointDSC [4]	56.09	3.87	**10.39**	44.51	52.38	47.57	**0.10**
FGR [2]	19.99	5.28	12.98	27.63	19.16	19.98	1.32
RANSAC [1]	46.38	5.00	13.11	40.70	44.61	42.02	2.86
CG-SAC [3]	52.31	3.84	10.55	42.16	47.02	44.61	0.25
SC^2-PCR [9]	57.83	3.77	10.46	44.87	53.69	48.38	0.11
SC^2-PCR++	**61.15**	**3.72**	10.56	**47.12**	**56.52**	**50.85**	0.26
	Predator [10]						
	RR↑	RE↓	TE↓	IP↑	IR↑	F1↑	Time(s)
DHVR [5]	65.41	4.97	12.33	54.75	54.66	53.70	3.55
DGR [14]	59.46	**3.19**	10.01	51.38	54.24	51.62	1.48
PointDSC [4]	68.89	3.43	9.60	56.55	67.52	60.82	**0.10**
FGR [2]	35.99	4.77	11.64	47.18	38.76	39.10	1.32
RANSAC [1]	64.85	4.28	11.04	56.44	65.68	60.01	2.86
CG-SAC [3]	64.01	3.86	10.94	56.88	64.12	59.25	0.25
SC^2-PCR [9]	69.46	3.46	**9.58**	56.98	67.47	61.08	0.11
SC^2-PCR++	**71.59**	3.45	9.61	**59.61**	**70.17**	**63.73**	0.26
	GeoTransformer [11]						
	RR↑	RE↓	TE↓	IP↑	IR↑	F1↑	Time(s)
DHVR [5]	73.83	4.49	10.21	61.06	71.85	64.21	2.71
PointDSC [2]	77.82	3.00	8.71	63.65	76.87	68.39	0.09
RANSAC [1]	77.48	3.37	9.69	**64.91**	73.98	68.68	2.03
CG-SAC [3]	76.92	3.34	9.81	62.10	75.27	67.05	0.22
LGR [11]	77.20	2.99	8.58	64.47	76.04	68.86	**0.05**
SC^2-PCR [9]	78.33	3.04	8.81	64.63	76.67	69.19	0.08
SC^2-PCR++	**78.72**	**2.96**	**8.56**	64.80	**77.02**	**69.55**	0.24

Predator[10]和 GeoTransformer[11]。从表 2 中展示的结果可以看出，无论与哪种描述子结合，SC^2-PCR++都取得了当前最好的性能。尤其是在特征描述子的表达能力相对较弱时（如 FCGF），SC^2-PCR++取得的性能提升尤为明显。这是由于 SC^2-PCR++对低内点率的情况有较好的鲁棒性。

为了更全面地评价方法的性能，我们在表 3 中展示了我们的方法在室外点云配准数据集 KITTI 上和其他方法的对比结果。

为了展示配准在下游任务中的应用，我们将配准方法用于室内地图重建并在 ICL_NUIM 数据集上对比了一些经典的 SLAM 方法。从表 4 的结果中可以看出，SC^2-PCR++在其中两个场景上取得了最低的轨迹误差，并且平均轨迹误差为所有方法中最小的。

表 3　在室外数据集 KITTI 上的结果

	FPFH [7]						
	RR↑	RE↓	TE↓	IP↑	IR↑	F1↑	Time(s)
DHVR [5]	-	-	-	-	-	-	-
DGR [14]	77.12	1.64	33.10	78.39	54.12	62.15	2.29
PointDSC [4]	98.20	0.35	8.13	92.85	93.87	93.11	0.45
FGR [2]	5.23	0.86	43.84	4.93	0.05	0.10	3.88
RANSAC [1]	74.41	1.55	30.20	78.50	52.66	60.72	5.43
CG-SAC [3]	74.23	0.73	14.02	78.64	60.82	67.11	0.73
SC^2-PCR [9]	**99.64**	**0.32**	7.23	93.63	95.89	94.63	**0.31**
SC^2-PCR++	**99.64**	**0.32**	**7.19**	**94.07**	**96.19**	**95.00**	0.86
	FCGF [8]						
	RR↑	RE↓	TE↓	IP↑	IR↑	F1↑	Time(s)
DHVR [5]	**99.10**	**0.29**	**19.80**	-	-	-	0.83
DGR [14]	98.20	0.34	21.70	72.19	78.06	75.13	2.29
PointDSC [4]	98.02	0.33	21.03	82.00	90.84	85.83	0.45
FGR [2]	89.54	0.46	25.72	**95.13**	4.25	8.18	3.88
RANSAC [1]	98.02	0.39	23.17	81.89	90.36	85.52	5.43
CG-SAC [3]	97.84	0.37	22.91	81.85	90.84	85.74	0.73
SC^2-PCR [9]	98.20	0.33	20.95	82.01	91.03	85.90	**0.31**
SC^2-PCR++	98.56	0.32	20.61	82.17	**91.23**	**86.09**	0.86

表 4　多路配准数据集 ICL_NUIM 上的轨迹误差

	Living1	Living2	Office1	Office2	AVG
ElasticFusion	66.61	24.33	13.04	35.02	34.75
InfiniTAM	46.07	73.64	113.8	105.2	85.68
BAD-SLAM	fail	40.41	18.53	26.34	-
Multiway + DGR	21.06	21.88	15.76	11.56	17.57
Multiway + PointDSC	20.25	15.58	13.56	11.30	15.18
Multiway + DHVR	22.91	16.37	**12.58**	10.90	15.69
Multiway+ FGR	78.97	24.91	14.96	21.05	34.98
Multiway + RANSAC	110.9	19.33	14.42	17.31	40.49
Multiway + SC^2-PCR	18.68	**14.31**	14.63	11.95	14.90
Multiway + SC^2-PCR++	**17.56**	14.37	13.24	**9.49**	**13.67**

责任编委　崔海楠

参考文献

[1] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM.

[2] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In European Conference on Computer Vision, pages 766–782. Springer, 2016.

[3] Siwen Quan and Jiaqi Yang. Compatibility-guided sampling consensus for 3-d point cloud registration. IEEE Transactions on Geoscience and Remote Sensing, 58(10):7380–7392, 2020.

[4] Bai X, Luo Z, Zhou L, et al. Pointdsc: Robust point cloud registration using deep spatial consistency[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15859-15869.

[5] Junha Lee, Seungwook Kim, Minsu Cho, and Jaesik Park. Deep hough voting for robust global registration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15994–16003, 2021.

[6] Zeng A, Song S, et al. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1802-1811.

[7] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In 2009 IEEE international conference on robotics and automation, pages 3212–3217. IEEE, 2009.

[8] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In Proceedings of the IEEE International Conference on Computer Vision, pages 8958–8966, 2019.

[9] Chen Z, Sun K, Yang F, et al. Sc2-pcr: A second order spatial compatibility for efficient and robust point cloud registration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13221-13231.

[10] Huang S, Gojcic Z, Usvyatsov M, et al. Predator: Registration of 3d point clouds with low overlap[C]//Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2021: 4267-4276.

[11] Qin Z, Yu H, Wang C, et al. Geometric transformer for fast and robust point cloud registration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11143-11152.

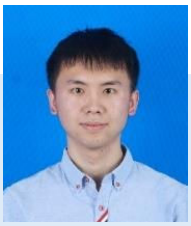
[12] Xu H, Liu S, Wang G, et al. Omnet: Learning overlapping mask for partial-to-partial point cloud registration[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 3132-3141.

[13] Yew Z J, Lee G H. Regtr: End-to-end point cloud correspondences with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 6677-6686.

[14] Choy C, Dong W, Koltun V. Deep global registration[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2514-2523.

[15] Yang H, Shi J, Carlone L. Teaser: Fast and certifiable point cloud registration[J]. IEEE Transactions on Robotics, 2020, 37(2): 314-333.

[16] Barath D, Matas J. Graph-cut RANSAC[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6733-6741.



陈志

华中科技大学在读博士生。主要研究方向为图像匹配、三维点云配准等。
Email: z_chen@hust.edu.cn

孙琨

中国地质大学（武汉）副教授。主要研究方向为多视图图像匹配、三维重建和点云处理。
Email: sunkun@cug.edu.cn

杨帆

华中科技大学在读博士生。主要研究方向为三维点云配准和深度学习几何学。
Email: fanyang@hust.edu.cn

郭琳

华中科技大学在读硕士生。主要研究方向为点云配准和图像匹配等。
Email: linguo@hust.edu.cn

陶文兵

华中科技大学教授。主要研究方向为三维配准、多视图立体几何、表面重建、图像分割等。
Email: wenbingtao@hust.edu.cn

热点追踪

基于布朗桥扩散模型的图像翻译

南昌航空大学　李波　薛凯韬　刘彬
Cardiff University　Yu-Kun Lai

一、引言

图像到图像的翻译问题是计算机视觉（Computer Vision，CV）领域的一个重要问题，它是指在两个不同的图像域之间建立映射关系。许多计算机视觉领域的问题都可以看作是图像翻译问题，比如图像去雾去噪，灰度图像着色，图像补全等。

现有的图像翻译方法通常基于生成对抗网络（Generative Adversarial Network，GAN）。但是基于 GAN 的方法存在训练不稳定和模式坍塌的问题。其他的方法诸如基于变分自编码器（Variational Autoencoder，VAE）的方法，基于自回归模型（Autoregressive Models，AM）的方法并不能达到和基于 GAN 的方法相同的质量和泛化能力。近年来扩散概率模型（Diffusion Probabilistic Models，DPM）在图像生成上已经可以达到与 GAN 方法相当的质量。并且有一些条件扩散模型被用于图像翻译任务。但是这些条件扩散模型将条件信息直接加入到 U-Net 中，这种加入条件信息的机制缺乏清晰的理论基础保证。

我们提出了一种基于布朗桥扩散模型（Brownian Bridge Diffusion Models，BBDM）的图像翻译方法。如图 1 所示，传统的条件扩散模型从高斯噪声出发，通过直接将条件信息与每一步的噪声图像拼接之后输入神经网络来达到条件生成的目的。布朗桥扩散模型采用了截然不同的条件机制。本文首次提出利用布朗桥随机过程直接建模两个不同图像域之间的映射，可以从理论上很好的保证扩散过程可以稳定的从条件域出发，最终收敛到目标图像域的分布。在不同的图像翻译，诸如语义图像生成真实图像，边界图像生成真实图像，风格迁移，图像补全，灰度图像着色任务上的实验证明我们的布朗桥扩散模型可以生成高质量和多样的结果。

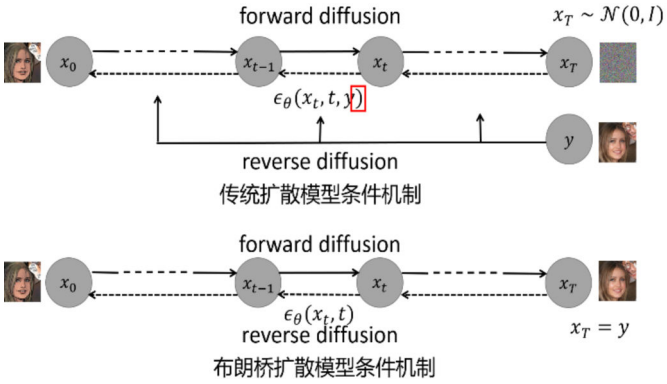


图 1　传统扩散模型和布朗桥扩散模型

二、基于布朗桥扩散模型的图像翻译总览

布朗桥（Brownian Bridge）是一种连续时间上的随机过程，它的分布为在条件$B_0 = a, B_1 = b$下的维纳过程。在基于布朗桥的图像翻译任务中假设条件图像的域表示为$\boldsymbol{A}$，目标图像域表示为$\boldsymbol{B}$，布朗桥随机过程在初始时刻的分布视为条件图像域$\boldsymbol{A}$，在最终时刻的分布视为目标图像域$\boldsymbol{B}$，这样就可以利用布朗桥随机过程实现图像域之间的转换。

基于布朗桥的图像翻译的整体过程如图 2 所示。

为了降低计算复杂度，提升模型的泛化性能，我们采用了隐式扩散模型的思想。对于一张条件图像$\boldsymbol{I_A} \in \boldsymbol{A}$，用预训练好的 VQGAN 的编码器提取出条件图像的隐式特征$\boldsymbol{L_A}$，以压缩图像的冗余信息，然后在隐空间将条件图像的隐式特征通过布朗桥扩散模型转换到目标图像的隐式特征$\boldsymbol{L_B}$，最后利用预训练好 VQGAN 的解码器从目标图像的隐式特征中解码出目标域的图像 $\mathbf{I_B} \in \mathbf{B}$。

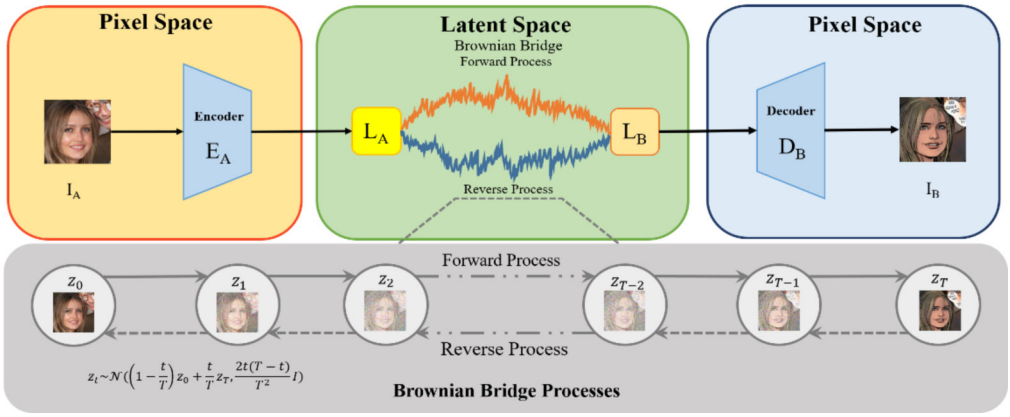


图 2　基于布朗桥扩散模型的图像翻译框架图

三、布朗桥扩散模型原理

我们利用布朗桥扩散模型实现域之间的转换。为了更一般化的表示布朗桥扩散模型原理，首先定义纯净的数据分布为$x_0 \in q(x_0)$，条件信息的分布为$y \in q(y)$，布朗桥随机过程则需要在给定y的情况下得到其对应的x_0。为了达到这个目的，与其他扩散模型类似，布朗桥扩散模型也分为前向过程（Froward Process）和反向过程（Reverse Process）。

前向过程中每一个时刻的边缘分布可以表示为：

$$q_{BB}(x_t|x_0,y) = \mathcal{N}(x_t;(1-m_t)x_0+m_t y,\delta_t I)$$

其中$m_t=\frac{t}{T}$，$\delta_t = 2(m_t-m_t^2)$.

为了得到前向中间过程的每一步的分布，我们将这个过程建模为马尔科夫链（Markov Chain）：

$$q_{BB}(x_{1:T}|x_0,y) = \prod_{t=1}^{T} q_{BB}(x_t|x_{t-1},y)$$

利用马尔科夫链的性质可以得到前向中间过程的分布的表示：

$$q_{BB}(x_t|x_{t-1},y) = \mathcal{N}(x_t;\frac{1-m_t}{1-m_{t-1}}x_{t-1} + \frac{m_t-m_{t-1}}{1-m_{t-1}}y,\delta_{t|t-1}I)$$

其中$\delta_{t|t-1} = \delta_t - \delta_{t-1}\left(\frac{1-m_t}{1-m_{t-1}}\right)^2$.

之后利用马尔科夫链的无后效性，可以得到后验分布：

$$q_{BB}(x_{t-1}|x_t,x_0,y) = \mathcal{N}(x_{t-1};\frac{\delta_{t-1}(1-m_t)}{\delta_t(1-m_{t-1})}x_t + (1-m_{t-1})\frac{\delta_{t|t-1}}{\delta_t}x_0 + \left(m_{t-1}-m_t\frac{\delta_{t-1}(1-m_t)}{\delta_t(1-m_{t-1})}\right)y,\tilde{\delta}_t = \frac{\delta_{t|t-1}\delta_{t-1}}{\delta_t})$$

反向的过程也定义为马尔科夫链，但是反向中间过程的均值需要通过神经网络预测：

$$p_\theta(x_{t-1}|x_t,y) = \mathcal{N}(x_{t-1}|\mu_\theta(x_t,t),\tilde{\delta}_t I)$$

训练的过程通过最小化变分下界来对齐前向中间过程的后验分布和反向的中间过程，也即约束均值相同，经过一定的重参数化技巧可以得到最终的训练目标：

$$\mathbb{E}_{x_0,y,\epsilon}\left[\left\|m_t(y-x_0)+\sqrt{\delta_t}\epsilon-\epsilon_\theta(x_t,t)\right\|_2^2\right]$$

但是把布朗桥随机过程建模为马尔科夫链的一个弊端是反向采样的过程和前向过程的步数要完全一样，时间代价线性正比于步数T。为了提高采样效率，我们借鉴了去噪扩散隐式模型的方法，将前向过程建模为非马尔科夫链的形式，这样在保证训练目标完全不变的情况下，只需要对采样过程进行一定的改动就能实现快速采样，具体的对于采样序列$\{1,2,\dots,T\}$的任意一个子序列$\{\tau_1,\tau_2,\dots,\tau_S\}$，将后验分布定义为如下形式就可以实现跳步的采样从而大大节省时间代价：

$$q_{BB}(x_{\tau_{s-1}}|x_{\tau_s},x_0,y) = \mathcal{N}(x_{\tau_{s-1}};(1-m_{\tau_{s-1}})x_0+m_{\tau_{s-1}}y + \sqrt{\delta_{\tau_{s-1}}-\sigma_{\tau_s}^2}\frac{1}{\sqrt{\delta_{\tau_s}}}(x_{\tau_s}-(1-m_{\tau_s})x_0-m_{\tau_s}y),\sigma_{\tau_s}^2 I)$$

当$\sigma_t^2 = \tilde{\delta}_t$时，这个过程与建模为马尔科夫链过程是等价的。

四、实验内容

为了验证布朗桥扩散模型在图像翻译应用上的有效性，我们在语义图像生成真实图像，边界图像生成真实图像，风格迁移，图像补全，灰度图像着色任务上进

行了实验。实验结果证明，我们提出的基于布朗桥扩散模型的图像翻译方法不仅可以生成高质量和多样的结果，而且在不同的任务上有着较好的泛化性能。部分实验结果如图 3 所示。

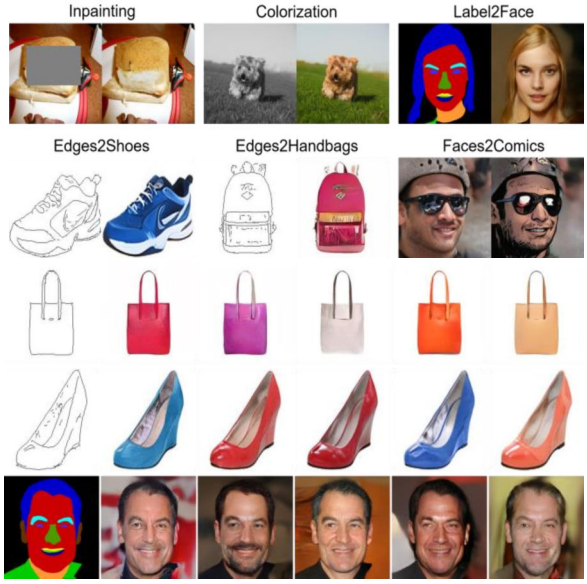


图 3　基于布朗桥扩散模型的图像翻译实验结果

五、总结和展望

本文介绍了基于布朗桥扩散模型的图像翻译方法的基本原理以及相关的实验内容。相较于传统的条件扩散模型，布朗桥扩散模型拥有更清晰的理论基础。实验的结果也证明，基于布朗桥扩散模型的图像翻译可以生成高质量和多样的结果。但是目前的布朗桥扩散模型应用于图像翻译需要成对的训练数据，而且需要较大的数据集才能有比较好的效果。因此，将布朗桥扩散模型应用于不成对的图像翻译并降低其对数据集大小的依赖是未来工作的重要方向。

该成果被国际学术会议 The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR2023) 接收。

责任编委　储珺

李波

南昌航空大学，数学与信息科学学院教授，主要研究方向为图深度学习、计算机视觉。
Email: libo@nchu.edu.cn

薛凯韬

南昌航空大学，在读硕士，CCF 学生会员。主要研究方向为深度学习、计算机视觉等。
Email: xuekt98@gmail.com

刘彬

南昌航空大学，数学与信息科学学院讲师，主要研究方向为数字几何处理、计算机视觉。
Email: nyliubin@nchu.edu.cn

Yukun Lai

Cardiff University, Professor of School of Computer Science and Informatics, research interests include computer graphics, geometric modelling and processing, computer vision, image processing.
Email: LaiY4@cardiff.ac.uk