文章编号:1671-4598(2025)07-0263-09 DOI:10.16526/j.cnki.11-4762/tp.2025.07.032 中图分类号:TP391 文献标识码:A

面向复杂交通环境的多源融合实时语义场景补全算法

林伟烜[1]，叶仕通[2]
(1. 广州华商职业学院 人工智能与数据学院，广州 511300；
2. 广州华商学院 人工智能学院，广州 511300)

摘要：面向自动驾驶与智慧交通需求，实时且精准地重建车辆周围的三维语义场景对保障交通安全与优化出行效率具有重要意义；针对复杂交通环境中存在多源异质数据以及遮挡严重、光照多变等难题，文章提出一种多源融合实时语义场景补全算法；通过跨模态自注意力策略融合摄像机、激光雷达和毫米波雷达信息，实现对遮挡区域的精确感知与语义推断；利用时空上下文建模在序列数据中捕捉目标动态变化，显著提升场景一致性与补全完整度；实验结果表明，与采用四维稀疏卷积的主流基线方法相比，所提算法在遮挡处理与推理速度上均取得显著优势。

关键词：多源融合；语义场景补全；时空上下文；实时感知；自动驾驶

Real-time Semantic Scene Completion Algorithm with Multi-source Fusion in Complex Traffic Environments

LIN Weixuan[1], YE Shitong[2]
(1. College of Artificial Intelligence and Data, Guangzhou Huashang Vocational College, Guangzhou 511300, China;
2. School of Artificial Intelligence, Guangzhou Huashang College, Guangzhou 511300, China)

Abstract: Facing the needs of autonomous driving and intelligent transportation, it is of great significance for the real-time and accurate reconstruction of 3D semantic scenes around vehicles to ensure traffic safety and optimize travel efficiency. Aiming at complex traffic environments where there are multiple sources of heterogeneous data, as well as serious occlusion and variable illumination, this paper proposes a multi-source fusion real-time semantic scene complementation algorithm. By using a cross-modal self-attention strategy and integrating information from a camera, LiDAR and millimeter-wave radar, the perception and semantic inference of occluded regions are accurately achieved. Spatio-temporal contextual modeling is used to capture the dynamic changes of the target in the sequence data, which significantly improves the consistency and completeness of scene. Experimental results show that the proposed algorithm has significant advantages over mainstream baseline methods with four-dimensional sparse convolution in both occlusion processing and inference speed.

Keywords: multi-source fusion; semantic scene complementation; spatio-temporal context; real-time sensing; autonomous driving

0 引言

随着智能交通系统和自动驾驶技术的迅猛发展，环境感知在实现自主驾驶、安全驾驶和高效交通管理等方面扮演着至关重要的角色。准确、实时地构建车辆周围环境的三维场景，不仅有助于提高目标检测、路径规划和决策制定的精度，而且对复杂交通环境中的障碍物避让、行人安全监控等问题也具有显著的应用价值[1-2]。

收稿日期:2025-01-25； **修回日期**:2025-03-10。
基金项目:2024年广东省普通高校自然科学类平台和项目(2024ZDZX3035);2024年广州华商职业学院校级科研课题(GZHSKY2024019)。
作者简介:林伟烜(1983-),女,博士,副研究员。
通讯作者:叶仕通(1981-),男,博士,教授。
引用格式:林伟烜,叶仕通.面向复杂交通环境的多源融合实时语义场景补全算法[J].计算机测量与控制,2025,33(7):263-271,279.

为了满足这一需求，语义场景补全作为一种重要的三维重建技术，越来越受到学术界和工业界的关注。语义场景补全的目标是基于传感器获取的部分环境信息，推断和补全缺失的区域，并为每个补全体素赋予准确的语义标签，以完成对场景的完整描述[3-5]。

在自动驾驶与智能交通场景中，实时语义感知面临着严峻的挑战。一方面，由于传感器自身的视野与探测原理限制，单一模态往往难以准确、完整地捕捉复杂交通场景的全局信息[6-7]。例如摄像机在夜晚光照不足或强逆光条件下极易丢失关键细节。激光雷达在雨雾天气下易出现噪声干扰和目标点云稀疏化，导致目标边界不清晰。毫米波雷达尽管能提供稳定的目标运动信息，但分辨率低，无法直接捕捉目标的精细结构与语义细节。这些模态的单一使用在遇到复杂遮挡、动态变化或恶劣天气等实际场景时，往往表现出明显的感知不足。另一方面，真实交通环境中，车辆、行人、骑行者与固定设施等目标频繁发生动态交互，导致大量目标间存在局部甚至整体遮挡，进而形成感知盲区或局部信息缺失。现有研究多基于简单的多模态特征拼接或独立模态建模，难以高效融合不同传感器之间的冗余与互补信息，从而导致在严重遮挡区域或传感器盲区内的语义补全精度明显下降[8-9]，无法满足自动驾驶与智能交通对场景理解准确性与实时性的双重要求。因此研究如何有效利用多模态数据的跨模态优势，建立更具鲁棒性和实时性的场景补全方法，已成为当前领域亟待解决的重要问题。

为了解决这一问题，近年来，研究人员开始探索多模态数据融合的方案，通过结合来自不同传感器的信息，如摄像机、激光雷达、毫米波雷达等，来提高场景补全的精度和鲁棒性[10-11]。多模态传感器的融合不仅能弥补单一传感器在信息获取上的局限，还能提高算法在动态环境中的适应能力。例如激光雷达能够提供精确的几何结构信息，毫米波雷达则能在恶劣天气或光照条件下保持较好的检测效果，而摄像机则能提供高分辨率的纹理和颜色信息。因此如何有效地将多模态传感器的数据进行融合，成为提升语义场景补全性能的关键。

尽管多模态融合在一定程度上能够提升场景补全的效果，现有研究仍面临着几个问题：如何在保证计算效率的前提下进行实时的多模态数据融合，以适应自动驾驶等高时效性的应用需求；如何有效地处理时空信息，特别是在动态场景中补全被遮挡或无法观测的部分；如何保证补全结果的语义一致性和全局一致性，避免局部错误的放大影响全局性能[12-14]。针对这些问题，本文提出了一种面向复杂交通环境的多源融合实时语义场景补全算法（简称 MFR-SSC），该算法通过结合多模态数据与时空上下文信息，能够在保证实时性的基础上，精确补全遮挡区域，并提供更为精细的语义标注。

1 总体设计

本文所提出的面向复杂交通环境的多源融合实时语义场景补全算法（MFR-SSC），主要包括多模态语义特征抽取、跨模态特征融合、时空上下文建模和场景补全推理 4 个核心模块。

在多模态语义特征抽取模块中，通过独立的深度特征提取网络，分别对摄像机图像、激光雷达点云和毫米波雷达信号进行初步特征编码。针对不同模态的传感器数据特性，设计了适配的特征提取策略：对图像数据采用深度卷积与自注意力机制，捕捉丰富的纹理和语义细节。对激光雷达数据采用三维稀疏卷积以提取高精度几何结构特征。对毫米波雷达数据采用二维卷积与通道注意力机制强化目标运动信息与抑制杂波干扰，从而生成适合跨模态融合的统一深度语义特征表示。

在获得各模态单独编码的特征之后，算法进一步通过跨模态特征融合模块对多源数据进行高效整合。为克服各模态在空间分辨率和信息完整性上的差异，本研究引入了一种基于跨模态自注意力机制的融合方案，以每一模态作为查询，主动从其他模态获取互补特征。通过引入可学习的门控机制对不同模态的贡献权重进行动态调节，从而提升对遮挡区域和信息不确定区域的感知准确度。

为了更好地捕捉交通环境中场景目标的动态变化，算法在融合特征的基础上引入了时空上下文建模模块。通过时空自注意力网络捕捉相邻帧之间特征令牌的长短期依赖关系，进一步挖掘连续帧之间目标运动信息和场景上下文变化趋势，缓解因瞬时遮挡或感知盲区而导致的目标位置丢失或语义歧义等问题。引入全局记忆机制，进一步提高了长期场景一致性，优化了遮挡区域的语义补全效果。

基于前面获得的融合特征与时空特征表示，算法在场景补全推理模块中，通过构建三维占据网格与图卷积推断，实现对遮挡区域的精细修正与语义预测。该模块利用多任务学习网络对体素进行占据状态与语义类别的初步估计，再通过构建空间图结构并施加图卷积推断，对局部不确定区域进行语义与几何特征传播，以实现全局语义一致性与精确场景补全。

图 1 为 MFR-SSC 方法的详细算法流程图。

2 算法详细设计

2.1 多模态语义特征抽取

为应对复杂交通环境下多种传感器数据的实时融合需求，本研究在图像、激光雷达以及毫米波雷达 3 种典型模态上分别构建深度神经网络，用以提炼高保真且可融合的语义特征表示。通过多尺度卷积网络[15-16]、三维稀疏卷积[17-18]以及自适应注意力[19-20]结构的综合运用，

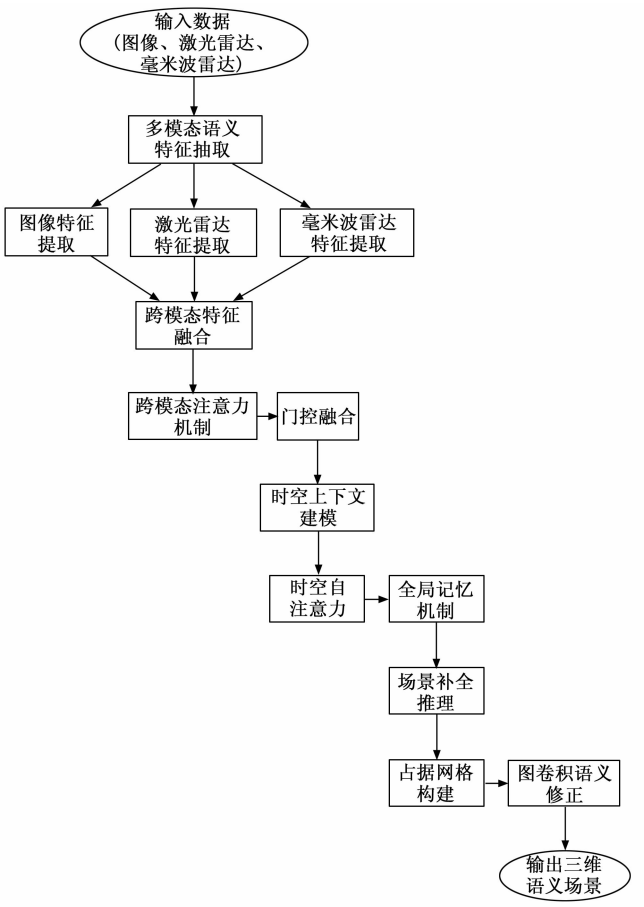

图 1　MFR-SSC 算法流程图

可以在不同传感器的优势维度中捕捉关键目标与背景上下文信息，并跨模态特征融合、时空上下文建模和场景补全推理提供高精度输入。

记 $I\in\mathbb{R}^{H\times W\times 3}$ 为摄像机拍摄的 RGB 图像，$P=[p_i=(x_i,\ y_i,\ z_i,\ r_i)]$ 表示激光雷达输出的点云数据，$R\in\mathbb{R}^{U\times V}$ 为毫米波雷达在距离－速度域或距离－方位域的回波反射映射。针对每一模态，定义其对应的特征提取函数为：$\Phi_{\text{img}}(I),\Phi_{\text{lidar}}(P),\Phi_{\text{radar}}(R)$，分别输出多维度的深度语义表示 $\boldsymbol{F}_{\text{img}}$，$\boldsymbol{F}_{\text{lidar}}$，$\boldsymbol{F}_{\text{radar}}$。针对图像模态，通过多尺度卷积网络获取局部纹理与语义分割级别的信息，令第 l 层二维卷积输出为：

$$X_{\text{img}}^{(l)}=\sigma(W_{\text{img}}^{(l)}*X_{\text{img}}^{(l-1)}+b_{\text{img}}{}^{(l)}),l=1,\cdots,L \tag{1}$$

其中：σ 是非线性激活函数，$*$ 表示卷积运算，$X_{\text{img}}^{(0)}=I$。为了在最深层卷积输出 $X_{\text{img}}^{(L)}$ 上进一步引入自注意力结构以获得全局上下文，令 Z_{img} 表示展平后的像素序列，定义多头自注意力过程为：

$$Q=Z_{\text{img}}W_Q,K=Z_{\text{img}}W_K,V=Z_{\text{img}}W_V \tag{2}$$

$$A=\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \tag{3}$$

$$Z'_{\text{img}}=AV \tag{4}$$

$$F_{\text{img}}=FFN(Z'_{\text{img}}) \tag{5}$$

其中：d 为隐空间的维度，W_Q，W_K，W_V 为可学习参数，$FFN(\cdot)$ 表示前馈网络，实现对 Z'_{img} 的进一步线性变换与激活。为准确表征三维空间结构，我们采用稀疏体素化与三维稀疏卷积对激光雷达点云进行深度编码。将 P 按体素大小 Δx，Δy，Δz 离散为稀疏占据表示 V_{lidar}，对体素非空区域施加三维稀疏卷积：

$$X_{\text{lidar}}^{(m)}=\sigma(W_{\text{lidar}}^{(m)}*_3X_{\text{lidar}}^{(m-1)}+b_{\text{lidar}}^{(m)}) \tag{6}$$

其中：$*_3$ 表示仅在有效体素激活点上计算的三维卷积算子。在经过多层堆叠后得到 $X_{\text{lidar}}^{(M)}$，并经全局池化或局部特征聚合输出 F_{lidar}。对于毫米波雷达模态，我们令 R 表示其在距离－速度域或距离－方位域上的原始回波矩阵。引入二维卷积网络与自适应通道注意力，以区分目标反射回波与背景杂波。卷积网络的第 k 层输出为：

$$X_{\text{radar}}^{(k)}=\sigma(W_{\text{radar}}^{(k)}*X_{\text{radar}}^{(k-1)}+b_{\text{radar}}^{(k)}) \tag{7}$$

并在末端施加自适应通道注意力 ACA（$\cdot$），令：

$$F_{\text{radar}}[c,h,w]=\alpha_c\cdot X_{\text{radar}}^{(K)}[c,h,w] \tag{8}$$

其中：α_c 为通道 c 的注意力系数，通过全局平均池化与两层全连接网络获取，从而强化对速度差异明显的运动目标通道，并抑制非相关杂波通道。在完成对图像、激光雷达及毫米波雷达的各自深度编码后，分别获得：$F_{\text{img}},F_{\text{lidar}},F_{\text{radar}}$，为跨模态融合与时空补全操作提供可对齐或可拼接的中间特征表示。

2.2　跨模态特征融合

在完成对图像、激光雷达与毫米波雷达的单独深度编码之后，分别获得了 F_{img}、F_{lidar} 和 F_{radar} 3 种特征表示。由于不同模态在空间分辨率、尺度层次以及信息不确定性方面存在差异，若直接拼接或简单加权，容易导致特征冗余或关键信息损失。因此本研究提出一套基于跨模态注意力与可学习门控机制的融合方法，旨在充分挖掘多源感知在遮挡补全、语义细粒度和目标运动特性上的互补优势。

令 $F_m\in\mathbb{R}^{N_m\times d}$ 表示模态 m 的深度特征序列，其中 $m\in\{\text{img},\ \text{lidar},\ \text{radar}\}$ 为激活单元的数量，d 为隐空间维度。为建立跨模态交互，我们在特征层面设计了多头交互注意力，令模态 m 在融合时主动查询模态 n 的特征表征，以赋予网络在局部噪声或缺失信息条件下的动态调整能力。对每一对模态（m，n）定义可学习参数 W_{Q_m}，W_{K_n}，$W_{V_n}\in\mathbb{R}^{d\times d}$，将 F_m 投影为查询 Q_m，Fn 投影为键 K_n 与值 V_n，即：

$$Q_m=F_mW_{Q_m},K_n=F_nW_{K_n},V_n=F_nW_{V_n} \tag{9}$$

接着我们计算跨模态注意力分布 $A_{m,n}$：

$$A_{m,n}=\text{softmax}\left(\frac{Q_mK_n^T}{\sqrt{d}}\right) \tag{10}$$

其中：softmax 在行方向上进行归一化处理。得到的融合向量 $Z_{m,n}$ 为：

$$Z_{m,n} = A_{m,n} V_n \tag{11}$$

模态 m 的特征可从模态 n 的角度获得补充信息，如图像模态可利用点云的几何分布来增强遮挡区域的识别能力，而雷达模态则可借助图像的高分辨率纹理提升目标检测的准确度。为在多模态间实现全面交互，我们定义了一个三模态交互循环：对（m，n）∈ {img，lidar，radar} {img，lidar，radar}，$m \neq n$，分别计算 $Z_{m,n}$ 并累加形成初步融合结果 H_m：

$$H_m = \sum_{n \in \{\text{img},\text{lidar},\text{radar}\}, n \neq m} Z_{m,n} \tag{12}$$

在获得 $\{H_{\text{img}}, H_{\text{lidar}}, H_{\text{radar}}\}$ 之后，还需要进一步考虑各模态的信噪比、时空置信度等因素，我们引入可学习门控向量 $g_m \in \mathbb{R}^d$ 对融合结果进行加权：

$$F'_m = H_m \odot \sigma_{\text{sigmoid}}(g_m) \tag{13}$$

其中：⊙为逐通道点乘，σ_{sigmoid}（·）对门控向量进行 [0，1] [0，1] [0，1] 区间的映射。该步可使模型在融合过程中自动学习信源选择：当某模态受到恶劣天气或遮挡干扰时，相应门控值会变小，降低其干扰度，从而强化信息可靠且置信度高的模态对场景理解的贡献。最终我们将所有模态的加权结果拼接或叠加形成统一的融合特征 F_{fusion}：

$$F_{\text{fusion}} = \Psi(F'_{\text{img}}, F'_{\text{lidar}}, F'_{\text{radar}}) \tag{14}$$

其中：Ψ（·）可简单设为沿特征维度的串接操作，或包含额外的前馈网络以进一步压缩和增强跨模态表达。该方法的伪代码如下：

Algorithm 1 跨模态特征融合过程

Require：多模态特征序列 F_{img}，F_{lidar}，F_{radar}：隐空间维度 d

Ensure：统一融合特征 F_{fusion}

1：定义可学习参数：对于每个模态 m，n，准备 W_{Q_m}，W_{K_n}，$W_{K_n} \in \mathbb{R}^{d \times d}$，

以及门控向量 $g_m \in \mathbb{R}^d$。

2：跨模态注意力交互：

3：for 模态对（m，n），其中 $m \neq n$ 且 m，$n \in$ {img，lidar，radar} do

4：计算查询、键、值：

$Q_m \leftarrow F_m W_{Q_m}$，$K_n \leftarrow F_n W_{K_n}$，$V_n \leftarrow F_n W_{V_n}$。

5：计算注意力权重：

$$A_{m,n} \leftarrow \text{softmax}\left(\frac{Q_m K_n^T}{\sqrt{d}}\right).$$

6：获得融合向量：

$$Z_{m,n} \leftarrow A_{m,n} V_n.$$

7：end for

8：多模态聚合：

9：for 每个模态 $m \in$ {img，lidar，radar} do

10：将来自其他 模态的跨模态特征累加：

$$H_m \leftarrow \sum_{n \neq m} Z_{m,n}.$$

11：利用可学习门控向量加权：

$$F'_m \leftarrow H_m \odot \sigma\text{Sigmoid}(gm)$$

其中：⊙为逐通道点乘，σSigmoid（·）是 Sigmoid 激活函数。

12：end for

13：融合输出：

$$F_{\text{fusion}}\ \psi(F'_{\text{img}};\ F'_{\text{lidor}}\ F'_{\text{radar}}$$

其中：ψ（·）表示在特征维度的拼接或额外前馈网络映射。

14：return F_{fusion}

2.3 时空上下文建模

在多源特征融合得到 F_{fusion} 之后，若仅对单帧数据进行补全推理，往往难以克服交通场景中存在的大量动态扰动与遮挡问题。车辆、行人或其他目标可能在局部时刻被遮挡，又会在后续帧出现可观测的特征，且道路环境的语义结构具有时间连续性和运动规律。为此我们引入时空上下文建模模块，通过对多帧融合特征施加时空自注意力与全局记忆机制，在更长时间尺度上追踪与补全未观测区域。

1）序列输入与帧级特征表示：

令 $\{I_t, P_t, R_t\}_{t=1}^T$ 分别表示在离散时间步 $t \in$ {1，2，…，T} 时采集到的图像、激光雷达点云和毫米波雷达数据。经过前两节的多模态特征抽取与跨模态融合，得到对应的融合特征 $F_t \in \mathbb{R}^{N \times d}$，其中 N 表示融合后特征令牌的数量，d 为隐空间维度。将这些时序特征构成序列（F_1，F_2，…，F_T）。作为时空建模的输入。此时，在相邻帧 F_t 与 $F_(t+1)$ 之间，交通场景中的目标和背景语义分布会随着车辆运动或目标本身运动而发生变化，需要通过深度网络在时间维度上捕捉动态关联。

2）时空自注意力与全局记忆机制：

为兼顾帧间的细粒度目标跟踪与全局环境演化，我们采用一种 Spatio-Temporal Transformer 结构，如图 2 所示。将时间序列中的每一帧特征看作一组可交互的令牌，通过自注意力机制建立跨帧与跨空间位置的依赖。记 F_t 的展平表示为 $Z_t \in \mathbb{R}^{N \times d}$。在时空自注意力中，定义查询、键与值矩阵为：

$$\boldsymbol{Q}_t = Z_t W_Q, K_s = Z_s W_K, V_s = Z_s W_V \tag{15}$$

对于任意两帧 $t \neq s$，可计算注意力分布 $A_{t,s}$：

$$A_{t,s} = \text{softmax}\left(\frac{Q_t K_s^T}{\sqrt{d}}\right) \tag{16}$$

并对值矩阵 $\boldsymbol{V}_s$ 加权求和得到跨帧融合表示：

$$Z_{t,s} = A_{t,s} \boldsymbol{V}_s \tag{17}$$

最终对 $s \in$ {1，…，T} 的所有帧求和累加，形成

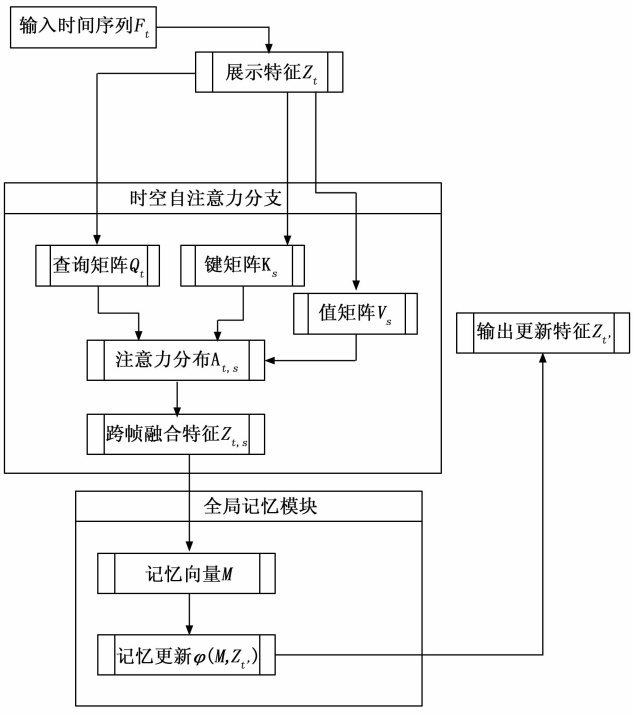

图 2　Spatio-Temporal Transformer 结构

对帧 t 的更新：

$$Z_t = \sum_{s=1}^{T} Z_{t,s} \tag{18}$$

由于 Z_t 整合了来自所有其他帧的信息，相比单帧特征能够显著减少短时遮挡或局部缺失对补全精度的影响。为进一步保留长期的环境先验信息，引入全局记忆机制，在模型中设定全局记忆向量 $M\in\mathbb{R}^{M\times d}$（$M$ 为记忆槽个数），随时间不断更新并存储道路布局、静态障碍物或重复出现的运动模式。记 M 的更新过程为：

$$M \leftarrow \varphi(M, Z'_t) \tag{19}$$

其中：$\varphi(\cdot)$ 为可学习的网络函数，如自注意力或门控循环单元。在推理时，通过引入对 M 的查询，可从全局记忆中检索相关先验进行补全，对大范围的静态场景（如建筑物、路面标志）与周期性运动行为（如红绿灯变化、人流高峰）具有较好的刻画能力。

3）时空自注意力与全局记忆机制：

为确保在多帧融合时维持语义一致性，特别针对同一目标在相邻帧的标签保持连续且平滑，引入时空一致性约束。令 Y_t 表示第 t 帧场景的语义预测结果（如分割图或点云分类标签），通过以下损失项约束相邻帧的预测相似度：

$$L_{\text{consistency}} = \sum_{t=1}^{T-1}\sum_{i=1}^{N} \| Y_t^{(i)} - Y_{t+1}^{(i)} \|_2 \cdot \omega(i) \tag{20}$$

其中：$Y_t^{(i)}$ 表示第 i 个像素（或体素）的预测标签向量，$\omega(i)$ 为权重函数，可根据置信度或运动速度进行加权，使快速移动或低置信度区域得到更多关注。

2.4　场景补全推理

在完成多模态融合与时空上下文建模后，得到了对复杂交通场景的统一深度特征序列 $\{Z_t\}_{t=1}^{T}$。由于交通环境中常存在遮挡、采样盲区或噪声导致的缺失信息，如何利用前面提取的多源语义表征对这些区域进行有效补全，进而生成完整的三维语义场景至关重要。本研究提出将占据网格表示与图卷积推断相结合，在空间域和图结构域同时施加约束，从而更准确地恢复缺失区域的几何与语义属性。

1）三维占据网格构建：

令 $V_t=\{v_{t,i}\}_{i=1}^{N_v}$ 表示在时间步 t 对场景进行离散化后得到的体素集合，每个体素 $v_{t,i}$ 与一个固定的三维坐标（$X_{t,i}$，$Y_{t,i}$，$Z_{t,i}$）以及占据状态 $\omega_{t,i}\in\{0,1\}$ 关联，其中 $\omega_{t,i}=1$ 表示体素已被探测到有物体占据，$\omega_{t,i}=0$ 则为未知或未观测状态。为了同时融合时空特征 Z_t 与局部几何信息，将 Z_t 中的特征令牌与 V_t 进行对应对齐或插值映射，得到体素级别的特征 $H_{t,i}\in\mathbb{R}^d$。若某体素在传感器视野之外，则初始特征记为零向量或使用最近邻体素的特征进行填充。

2）占据状态与语义类别估计：

$$y_{t,i} = [y_{t,i}^{(occ)}, y_{t,i}(\text{"}sem\text{"}))] \tag{21}$$

其中：$y_{t,i}^{(occ)}\in\{0,1\}$ 表征体素是否被物体占据，$y_{t,i}^{(sem)}\in\{1,\cdots,C\}$ 为体素所属语义类别（例如路面、车辆、行人、建筑等）。通过一个多任务学习的网络 $\Omega(\cdot)$ 对体素特征 $H_{t,i}$ 进行分类，定义：

$$y_{t,i} = \Omega(H_{t,i}) \tag{22}$$

其损失函数可综合二值交叉熵用于占据状态，以及多类别交叉熵用于语义分类：

$$L_{\text{voxel}} = \sum_{t=1}^{T}\sum_{i=1}^{N_v} L_{\text{BCE}}[y_{t,i}^{(occ)}, \hat{y}_{t,i}^{(occ)}] + L_{CE}[y_{t,i}^{(sem)}, \hat{y}_{t,i}^{(sem)}] \tag{23}$$

其中：$\hat{y}_{t,i}^{(occ)}$ 和 $\hat{y}_{t,i}^{(sem)}$ 分别为网络输出的预测值。然而，仅依赖局部特征与单体素分类难以解决由于遮挡或感知盲区导致的大面积信息缺失。为此，本研究在体素间引入图结构推断机制，对相邻体素施加约束，使得空间上邻近的体素在补全时具备更高的语义一致性。

3）图卷积推断与遮挡区域修正：

我们构建体素级别的有向图 $G_t=(V_t, E_t)$，其中 V_t 为节点集合，对应时间步 t 的体素，$E_t\subseteq V_t\times V_t$ 为边集，表征空间邻域关系或相似度度量。令 $\boldsymbol{A}_t\in\mathbb{R}^{N_v\times N_v}$ 表示邻接矩阵，对于节点 i 与 j 之间的边权 $A_(t, i, j)$，可基于三维坐标距离或特征相似度定义，例如：

$$\boldsymbol{A}_{t,i,j} = \begin{cases} \exp(-\beta\| H_{t,i} - H_{t,j} \|_2^2, & \text{若 } \left\| \begin{matrix}(X_{t,i},Y_{t,i},Z_{t,i}) - \\ (X_{t,j},Y_{t,j},Z_{t,j})\end{matrix} \right\| < \delta) \\ 0, & \text{否则} \end{cases} \tag{24}$$

其中：δ 为空间邻域阈值，β 控制相似度衰减速率。

在图结构 G_t 上施加图卷积推断，利用邻域节点的信息来修正未观测区域或低置信度体素的占据和语义预测。定义图卷积更新规则为：

$$H_{t,i}^{(l+1)} = \sigma\left[\sum_{j\in N(i)} A_{t,i,j}\left(W^{(l)}\ H_{t,j}^{(l)} + b^{(l)}\right)\right] \quad (25)$$

其中：$N(i)$ 表示节点 i 的邻域集合，$W^{(l)}$，$b^{(l)}$ 为第 l 层图卷积的可学习参数，σ 是激活函数。在经过 L 层迭代后，得到修正后的体素特征 $H_{t,i}^{(L)}$。该过程能在存在显著遮挡时，通过邻居体 aa 素的几何与语义关联将信息传播到缺失区域，进而弥补单体素预测的局限。最后通过整合图卷积输出 $H_{t,i}^{(L)}$ 到多任务网络 $\Omega(\cdot)$ 中重新估计 $\hat{y}_{t,i}^{(occ)}$ 与 $\hat{y}_{t,i}^{(sem)}$，实现对障碍物、自车道及其他交通要素的全局一致性补全。该方法的伪代码如下：

Algorithm 2 场景补全推理

时空融合特征序列 $\{Z_t'\}_{t=1}^{T}$，

Require：体素集合 $\{V_t\}_{t=1}^{T}$，多任务网络 $\Omega(\cdot)$，

邻域阈值 δ，衰减因子 β，图卷积层数 L。

Ensure：完整三维语义场景（含占据状态与类别）

1：网格构建与初步分类：

2：for $t=1$ to T do

3：for $i=1$ to N_v do4：将 Z_t' 与体素坐标（$X_{t,i}$，$Y_{t,i}$，$Z_{t,i}$）对齐，得到 $H_{t,i}\in\mathbb{R}^d$；

5：若体素超出视野，则令 $H_{t,i}=0$ 或插值近邻。

6：通过网络 $\Omega(\cdot)$ 得到占据与类别预测：

$$Y_{t,i} = \Omega H_{t,i} = \left[Y_{t,i}^{(com)}, Y_{t,i}^{(sem)}\right].$$

7：end for

8：end for

9：图卷积推断修正：

10：for $t=1$ to T do

11：构建体素级别图 $G_t=(V_t, E_r)$ 并计算邻接矩阵 $\boldsymbol{A}t$：

$$\boldsymbol{A}_{t,i,j} = \begin{cases} \exp(-\beta\| H_{t,i} - H_{t,j}\|^2), & \text{若距离} < \delta, \\ 0, & \text{否则} \end{cases}.$$

12：for $l=1$ to L do

13：for$i=1$ to N_v do

14：

$$H_{t,i}^{(l+1)} \leftarrow \sigma\left\{\sum_{j\in N(i)} \boldsymbol{A}_{t,i,j}\left[W^{(l)}\ H_{t,j}^{(l)} + b^{(l)}\right]\right\}.$$

15：end for

16：end for

17：用修正后 $H_{t,i}^{(L)}$ 再次输入 $\Omega(\cdot)$ 更新预测，消除遮挡与盲区误差。

18：end for

19：return 全局一致的体素占据与语义结果

3 实验仿真及分析

3.1 实验条件

为了确保对本文提出的多源融合实时语义场景补全算法的评估具有充分的代表性和可比性，我们的实验条件设置如下：硬件方面，主要依托一台配备双路 Intel Xeon E5-2698 v4 CPU（合计 40 核、2.2 GHz 主频）、256 GB DDR4 内存以及两块 NVIDIA Tesla V100 GPU（32 GB 显存）的大型服务器进行离线训练与验证，在车规级 NVIDIA DRIVE Orin 平台上进行推理测试，以考察算法在真实道路场景下的实时性表现。软件方面，操作系统我们选择 Ubuntu 20.04 LTS，深度学习框架采用 PyTorch 1.10.0，并结合 MinkowskiEngine 实现 3D/4D 稀疏卷积相关操作。

3.2 实验数据集及评估指标

本文的实验数据主要来自 nuScenes、KITTI 公开多模态自动驾驶数据集。数据集包含多种复杂场景与环境条件，如城市中心区域、住宅街区、高速公路及复杂路口等，场景类型涵盖日间、夜间、晴朗、雨雾等不同天气状况，且包含大量典型的遮挡场景，如车辆交叉遮挡、行人与车辆相互遮挡等情形，能够充分考察算法在真实交通环境下的泛化能力与鲁棒性。为进一步验证本文算法在真实复杂交通条件下的有效性与实时性，我们还选取了本研究团队在城市道路、隧道入口、高架桥底部及繁忙十字路口等典型城市复杂路段自主采集的数据。这些数据采集平台统一配备了 3 台 RGB 相机（分辨率 1 280×720，帧率 30 Hz）、Velodyne HDL-64 线激光雷达、毫米波雷达多种传感器。实验总数据规模超过 20 万帧，涵盖多种交通场景与天气环境，可有效代表真实自动驾驶场景中的多样性与复杂性。

为全面评估本文提出的 MFR-SSC 算法在复杂交通环境下的场景补全性能，本研究选取了交并比（IoU，intersection over union）、补全度（Completeness）、精确度（Precision）和帧率（FPS，frames per second）多种评价指标。这些指标分别从语义补全的空间准确性、几何完整性、预测可靠性及实时性 4 个维度，能客观、严谨地衡量了算法在实际复杂场景下的表现。

1）IoU 用于衡量算法对各类目标语义补全结果与真实标注之间的空间重合程度，定义为：

$$\mathrm{IoU} = \frac{|V_{\mathrm{pred}} \cap V_{gt}|}{|V_{\mathrm{pred}} \cup V_{gt}|} \quad (26)$$

其中：V_{pred} 为算法预测的体素集合，V_{gt} 为真实标注的体素集合。IoU 值越高表明算法预测结果与真实场景吻合度越好。

2）Completeness 用于评价被算法成功预测补全的体素占真实语义场景中应当补全的体素的比例，定义为：

$$\text{Completeness} = \frac{|V_{\text{pred}} \cap V_{gt}|}{|V_{gt}|} \tag{27}$$

该指标能够反映算法在遮挡或盲区区域补全缺失信息的能力，值越高表示算法对真实缺失区域的补全越完整。3）Precision 用于评估算法预测补全的体素中正确的体素所占比例，定义为：

$$\text{Precision} = \frac{|V_{\text{pred}} \cap V_{gt}|}{|V_{pred}|} \tag{28}$$

4）为进一步量化算法在多类别体素语义预测上的综合性能，本文还采用了平均交并比指标 mIoU，定义为所有类别 IoU 的算术平均值：

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \text{IoU}_c \tag{29}$$

其中：C 为语义类别的总数，IoU_c 表示第 c 类目标的 IoU 值。mIoU 越高表示算法对不同类别目标语义预测的整体表现越好。

（5）为验证算法在实际交通环境中的实时性表现，我们计算每秒可处理的帧数 FPS：

$$\text{FPS} = \frac{N_{\text{frames}}}{T_{\text{process}}} \tag{30}$$

其中：N_{frames} 为处理的帧数，T_{process} 为处理这些帧所花费的时间。FPS 值越高，代表算法的实时性越好，越适合实际应用于自动驾驶与智能交通场景。

3.3　实验具体处理

1）数据预处理：

对图像数据，我们统一将原始 RGB 图像调整为尺寸为 448×448，并归一化到区间 [0，1]。激光雷达数据采用体素化处理，体素尺寸设定为 0.1 m×0.1 m×0.1 m，删除反射强度过低或离群点。毫米波雷达数据使用二维傅里叶变换转换为距离－速度谱图，尺寸统一调整为 128×128，并标准化处理。关键核心代码片段如下：

```
#图像数据预处理
from torchvision import transforms
image_transform = transforms.Compose([
transforms.Resize((448, 448)),
transforms.ToTensor(),
transforms.Normalize(mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225])
])


#激光雷达数据体素化处理
def voxelize_point_cloud(points, voxel_size=0.1):
coords = np.floor(points[:, :3] / voxel_size).astype(np.int32)
coords, unique_indices = np.unique(coords, axis=0, return_index=True)
voxel_features = points[unique_indices, 3:]
return coords, voxel_features
```

2）特征提取方法及参数：

图像特征提取模块采用基于 ResNet-50 的多尺度卷积网络，初始化权重来自 ImageNet 预训练模型，特征维度为 512。激光雷达数据采用基于 3D 稀疏卷积网络（Minkowski Engine）结构，稀疏卷积核尺寸设置为 3×3×3，特征维度为 128。毫米波雷达数据特征提取采用二维卷积网络，卷积核尺寸为 3×3，配合通道注意力机制，输出维度为 64。

3）模型训练与参数设置：

本文的训练过程采用了端到端方式，模型训练具体参数设置如下：优化器采用 AdamW，初始学习率为 1e-4，权重衰减设定为 1e-4。使用余弦退火学习率调整策略（Cosine Annealing LR），训练总 epoch 为 100，批量大小为 8。损失函数组合采用语义分类交叉熵与占据预测的二分类交叉熵进行加权求和。关键核心代码片段如下：

```
# 损失函数定义
criterion_semantic = nn.CrossEntropyLoss()
criterion_occupancy = nn.BCEWithLogitsLoss()
for epoch in range(100):
for images, lidar_points, radar_data, labels in dataloader:
optimizer.zero_grad()
outputs = model(images, lidar_points, radar_data)
loss_sem = criterion_semantic(outputs['semantics'], labels['semantics'])
loss_occ = criterion_occupancy(outputs['occupancy'], labels['occupancy'])
loss = loss_sem + loss_occ
loss.backward()
optimizer.step()
scheduler.step()
```

3.4　实验结果分析

本研究进行的实验主要对算法的遮挡区域补全能力、体素级别的语义精度以及推理帧率进行验证，与采用四维稀疏卷积的 4D Minkowski Convolution [21] 进行深入比较，通过多种评价指标综合呈现算法在复杂交通环境中的优劣势。

为检验算法在遮挡场景下的补全能力，特别挑选了 nuScenes 数据集中多车并排行驶、行人与大型货车等目标相互遮挡的复杂路段，并从 KITTI 数据集中额外筛选出包含临时障碍物遮蔽或狭窄车道并行车辆的片段。表 1 汇总了本文提出的 MFR-SSC 与采用四维稀疏卷积的 4D Minkowski Convolution（以下简称“4D Mink”）在车辆、行人与骑行者等主要目标遮挡区域上的补全表现。

表 1 4D Mink 与 MFR－SSC 在遮挡区域场景补全对比

类别/指标	4D Mink	MFR-SSC
Vehicle-IoU/%	68.9	72.5
Vehicle-Completeness/%	75.2	78.7
Vehicle-Precision/%	78.6	82.1
Pedestrian-IoU/%	61.3	64.8
Pedestrian-Completeness/%	74.9	78.1
Pedestrian-Precision/%	72.7	77.3
Cyclist-IoU/%	59.4	63.2
Cyclist-Completeness/%	70.1	74.6
Cyclist-Precision/%	71.8	76.3
Overall-IoU/%	62.8	65.4
Overall-Completeness/%	78.5	82.2
Overall-Precision/%	80.1	84.6

从表 1 数据可见，在多类目标的遮挡情形下，MFR-SSC 均有不同程度的优势，特别是在“Vehicle”和“Pedestrian”这两类与交通安全最为相关的目标上，IoU、Completeness 及 Precision 相比 4D Mink 均有约 3～5% 的提升。当摄像机视角出现大面积遮挡时，4D Mink 主要依赖连续帧点云的几何变化，遇到多车并列或大型车辆遮挡时，点云信息极度稀疏，往往难以准确补全被遮挡目标的体素形状。在车辆急加速、行人快速穿越马路等动态场景中，MFR-SSC 借助跨模态自注意力和全局记忆单元对不同时刻的特征进行关联，能有效捕捉到目标运动轨迹并保持连续帧预测的一致性。4D Mink 也考虑了时间维度，但缺乏对多种传感器数据的深度融合，无法在完全遮挡的短时刻填补结构缺失。

为了进一步量化算法在不同置信度下对场景中各体素的语义判别能力，本文在 nuScenes 与 KITTI 数据集的混合测试集上进行了多次实验。对每个体素预测的语义标签输出一个置信度，并以多种阈值将体素分为“保留”或“舍弃”；对保留体素计算在语义标签上的平均交并比（mIoU），即可得到不同阈值下的精度。图 3 显示了 MFR-SSC 与 4D Mink 在阈值从 0.3 递增到 0.9 的实验结果。

MFR-SSC 在各置信度阈值下的 mIoU 均超过 4D Mink，并且随着阈值升高，虽然整体精度都有不同程度的下降，但 MFR-SSC 保持的降幅更为平缓，对高置信度体素的判别依旧具备明显优势。MFR-SSC 在各置信度阈值下的 mIoU 均超过 4D Mink，并且随着阈值升高，虽然整体精度都有不同程度的下降，但 MFR-SSC 保持的降幅更为平缓，对高置信度体素的判别依旧具备明显优势。无论是低阈值还是高阈值阶段，MFR-SSC 的 mIoU 均高出 4D Mink 约 3～4 个百分点，说明在光照不佳或存在较大遮挡时，多模态融合与全局记忆机制能够填补感知盲区，从而输出更精细化的语义预测结果。

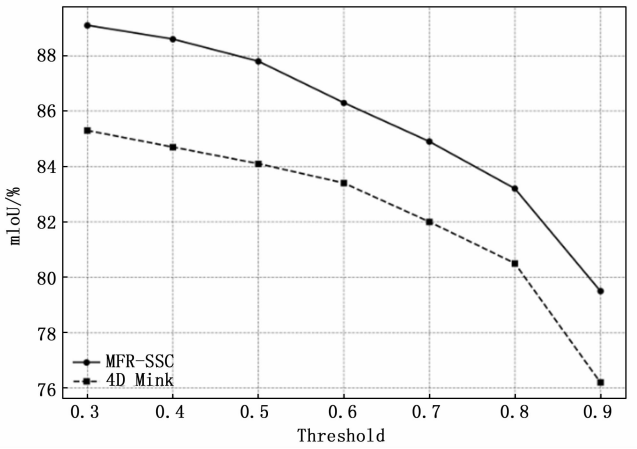

图 3 不同置信度阈值下的体素级别语义精度

在自动驾驶或城市感知场景中，传感器（如激光雷达、摄像机、毫米波雷达）的有效探测范围和分辨率往往随着距离增加而显著下降。为评估算法在不同距离区间的识别精度，本文以 0～100 m 范围为例，将测试数据按 10 m 步长划分为 10 个区段，在每个区段内分别统计体素级 Precision，比较本文提出的 MFR-SSC 和基线方法 4D Mink 的差异。图 4 显示了相应结果。

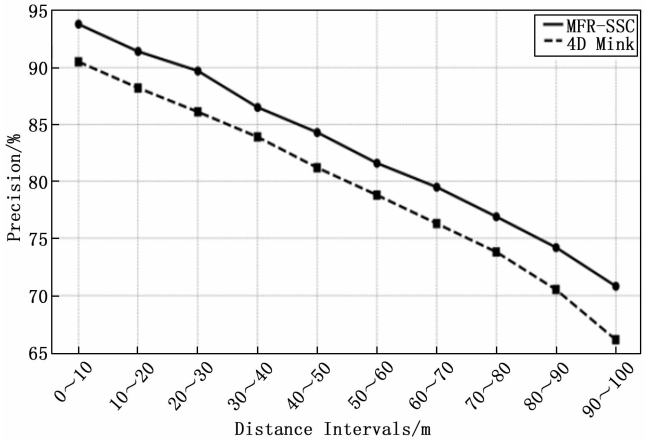

图 4 不同距离区间的体素级 Precision（%）对比

在靠近传感器的范围内，点云密度最高且摄像机纹理细节更丰富，MFR-SSC 和 4D Mink 均能获得较高的精度。随着距离增大，点云开始变得稀疏，摄像机分辨率与激光回波信号强度也随之下降。如果有遮挡出现，缺失信息往往更明显。在此范围内，MFR-SSC 利用跨模态注意力和时空一致性仍可保持高于 4D Mink 约 2～3 个百分点的 Precision，展现出对部分遮挡和噪声散射的更好抑制能力。当目标出现在 50 m 以外，激光点云趋于稀疏，摄像机也难以分辨运动物体的细节，毫米波雷达数据受杂波影响增大，此时整体精度水平显著下

滑，但 MFR-SSC 依然比 4D Mink 的表现更好。

为了评估算法在不同时间窗口设置下的实时性能，本研究在固定的硬件环境中，使得时间序列长度从 1 帧逐步增加至 10 帧的条件下，对所提出的 MFR-SSC 与对比算法 4D Mink 的推理帧率（FPS）进行测量。图 5 显示了两种算法随时间窗口变化的平均帧率，用于分析在需要更长时序信息时各自性能的下降幅度及可用性。

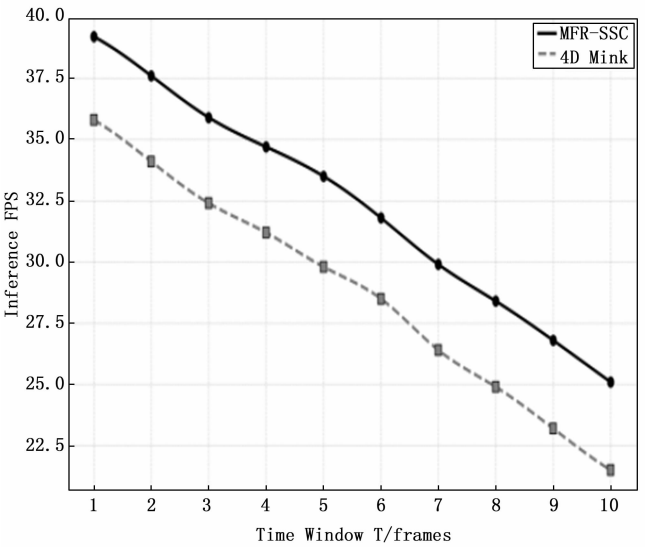

图 5　不同时间窗口下的推理帧率（FPS）对比

在时间窗口较小的情况下（$T=1$ 至 3），MFR-SSC 和 4D Mink 均能维持每秒 30 帧以上的处理速度，基本满足城市路况下对时序感知的要求。随着窗口长度增大，计算量同步提升，两种算法帧率均有一定程度下滑，但 MFR-SSC 在 T＝10 时依然可保持约 25 FPS，而 4D Mink 已降至 21.5 FPS，显示前者在时空融合的并行化和稀疏激活优化上更具优势。

4　结束语

本文针对复杂交通环境下的实时语义场景补全需求，提出了一种融合多模态感知信息与时空上下文的多源融合实时语义场景补全算法（MFR-SSC）。通过跨模态自注意力策略整合摄像机、激光雷达与毫米波雷达的冗余与互补信息，以及基于时空 Transformer 或图卷积的序列建模方法，算法在遮挡处理、动态目标跟踪以及全局场景一致性等方面均展现出良好的鲁棒性与高精度。实验结果表明，与 4D Minkowski Convolution 等代表性基线算法相比，MFR-SSC 在遮挡区域的场景补全精度和推理速度上均取得了显著提升，尤其在多车并线或复杂路口场景下，能维持更好的语义完整度和实时性。

但 MFR-SSC 仍有一些问题尚待进一步研究，例如当场景中存在极端气象条件（如暴雨、浓雾）或多辆大型车叠加遮挡时，部分传感器数据可能同时失效，导致补全精度和时序稳定性下降。在超大规模动态场景中，如何进一步提升模型的时间效率以及对稀疏异常数据的鲁棒处理能力仍需深入探讨。在未来的工作中，我们将考虑引入更多类型的环境先验知识或自监督机制，并与车路协同、联邦学习等新兴范式结合，进一步扩展多源融合实时语义场景补全的应用范围与实际落地价值。

参考文献：

[1] 张　康，安泊舟，李　捷，等. 基于 RGB-D 图像的语义场景补全研究综述 [J]. 软件学报，2023，34（1）：444－462.

[2] XIA Z，LIU Y，LI X，et al. Scpnet：Semantic scene completion on point cloud [C] //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023：17642－17651.

[3] 肖海鸿，吴秋遐，李玉琼，等. 三维补全关键技术研究综述 [J]. 光学精密工程，2023，31（5）：667－696.

[4] ROLDAO L，DE CHARETTE R，VERROUST-BLONDET A. 3D semantic scene completion：A survey [J]. International Journal of Computer Vision，2022，130（8）：1978－2005.

[5] 耿信财，施祥鹏，黄书发，等. 基于语义信息和彩色图像引导的道路场景深度补全框架 [J]. 测绘通报，2024，（s2）：191－196.

[6] 赵文红，王　巍，万子璐. 基于时空 Transformer 特征融合的车辆轨迹预测 [J]. 通信学报，2024，45（11）：267－276.

[7] FENG S，YAN X，SUN H，et al. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment [J]. Nature communications，2021，12（1）：748－756.

[8] LI Y，LI S，LIU X，et al. Sscbench：A large-scale 3d semantic scene completion benchmark for autonomous driving [C] //2024 IEEE/RSJ International Conference on Intelligent Robots and Systems（IROS）. IEEE，2024：13333－13340.

[9] RIST C B，EMMERICHS D，ENZWEILER M，et al. Semantic scene completion using local deep implicit functions on lidar data [J]. IEEE transactions on pattern analysis and machine intelligence，2021，44（10）：7205－7218.

[10] 毕　欣，翁才恩，王　瑜，等. 4D 毫米波雷达点云与视觉图像自动外参配准方法 [J]. 华南理工大学学报（自然科学版），2025，53（1）：74－83.

[11] 史鹏涛，危康乐，吴　昊，等. 自动驾驶环境下基于语义分割的激光雷达与相机外参标定方法 [J]. 激光与光电子学进展，2024，61（24）：306－311.

（下转第 279 页）