中国科学: 信息科学 2024 年 第 54 卷 第 6 期: 1419–1440
SCIENTIA SINICA Informationis
论文

基于信息论的智能驾驶可解释多模态感知

张新钰[1,2,3*], 国纪龙[1,3], 李骏[1,2,3], 李德毅[4], 张世焱[1,3], 沈思甜[1,3], 吴凡[1,3], 刘华平[4]

1. 清华大学智能绿色车辆与交通全国重点实验室, 北京 100084
2. 北京航空航天大学交通学院, 北京 100191
3. 清华大学车辆与运载学院, 北京 100084
4. 清华大学计算机科学与技术系, 北京 100084
* 通信作者. E-mail: xyzhang@tsinghua.edu.cn

收稿日期: 2023–06–14; 修回日期: 2023–10–26; 接受日期: 2023–11–13; 网络出版日期: 2024–06–13

国家重点研发计划 (批准号: 2018YFE0204300) 和国家自然科学基金 (批准号: 62273198, U1964203) 资助项目

摘要 智能驾驶汽车已成为人们关注的热点话题之一. 然而, 现有的智能驾驶技术仍面临一系列挑战, 如交通障碍物的遮挡所引起的模型漏检, 以及当汽车驶入隧道等光线骤变的场景时所引起的传感器感知精度下降导致的误检问题等. 为保证复杂场景下车辆的感知安全, 智能驾驶多模态感知技术应运而生. 然而, 现有的多模态融合方法仍局限于对检测精度的提升, 缺乏感知过程的可解释性, 并缺少对模型感知过程的评价指标. 本文从信息论角度出发, 按照通信模型的构建方法对感知模型进行设计, 提出了一种基于信源信道联合编码理论的多模态融合感知模型, 从理论上对模型的感知过程进行解释. 同时, 提出了一种新的评价指标 —— 平均信息熵变 (average entropy variation, AEV), 用 AEV 来实时反映模型与外界感知交互过程中的稳定性. 进而, 对多模态模型的感知过程进行量化分析, 增加模型的可解释性. 最后, 与其他的感知模型在 KITTI 数据集的评估结果进行比较, 我们的模型在经过相似的网络结构时平均信息熵变下降到 0.5904, 更好地保证了检测任务的感知安全.

关键词 可解释性, 信息论, 信源信道联合编码, 多模态融合, 智能驾驶

1 引言

智能驾驶是未来城市交通发展的必然趋势 [1], 而驾驶安全是人们关注的首要问题. 随着自动驾驶车辆造成的人员伤亡事故发生, 自动驾驶汽车的安全问题不仅受到社会舆论的关注, 还面临政府和媒体的质疑 [2]. 同时, 现实场景中复杂的道路行车环境和天气因素给智能驾驶安全带来了挑战. 面对复

引用格式: 张新钰, 国纪龙, 李骏, 等. 基于信息论的智能驾驶可解释多模态感知. 中国科学: 信息科学, 2024, 54: 1419–1440, doi: 10.1360/SSI-2023-0086
Zhang X Y, Guo J L, Li J, et al. Information-theoretic-based interpretable multimodal perception for intelligent vehicles (in Chinese). Sci Sin Inform, 2024, 54: 1419–1440, doi: 10.1360/SSI-2023-0086

杂多变的交通环境, 智能驾驶车辆的安全成为了当今行业最关键的挑战之一. 其中, 感知安全是提升智能汽车安全性面临的首要问题 [3].

智能驾驶感知安全问题主要集中于感知任务 [4, 5]. 为满足智能驾驶对感知任务的较高需求, 许多先进的算法都给出了解决方案, 以实现更高的精度、更快的速度 [6]. 然而, 基于单模态的感知算法 [7] 往往受限于传感器的性能. 当大型车辆将目标部分遮挡时, 仅采用单模态传感器容易造成目标的漏检与误检. 此外, 当车辆驶入隧道等光线骤变的场景时, 由于光线的突然改变, 部分受光照影响较大的传感器容易出现感知错误, 无法满足智能驾驶车辆的感知安全需求. 而日益成熟的多传感器融合技术因能突破单一传感器的性能局限, 利用不同模态的互补特征补充单模态在不良光照等场景下的特征损失, 成为感知系统发展的必然趋势, 从而有效提高车辆的感知能力与驾驶安全性 [8]. 但传统的融合方法 [9~11] 通常是采用结果级融合的方式, 难以达到检测结果中目标数量或类别差异下的有效匹配, 从而导致目标的漏检和误检, 为感知安全带来风险 [12]. 此外, 现有的多模态感知模型通常基于深度学习算法, 在部分感知任务中表现优异 [13]. 但这类感知模型往往通过实验结果驱动进行设计, 通过拟合大量数据来优化参数, 存在可解释性差和感知功能底层机理难以阐释的问题, 无法保证车辆遮挡以及光线骤变等特殊场景下的正确感知. 同时, 传统深度学习网络难以评估检测结果的可信度, 在适应复杂动态环境下的感知安全方面存在严重不足. 并且, 绝大多数多模态融合模型将感知结果的准确率作为其主要的评估指标, 无法保证模型在同外界环境实时感知交互过程中的可靠性, 难以评估模型的泛化能力与感知过程的可信度 [14]. 为此, 无人驾驶领域迫切需要相关理论对融合网络进行合理解释, 建立能够描述网络底层机理的理论模型, 指导融合网络设计, 并对模型进行系统的评估, 从而保证智能汽车的感知安全.

本文着力于提高感知模型的可解释性, 基于信息论中通信模型 [15] 的设计方法提出了一种多传感器的特征深度融合方法. 首先, 提出了一种基于特征融合的多模态融合方法, 利用多头注意力融合模块融合不同模态之间的特征信息, 避免了结果级融合存在的目标漏检问题. 同时, 由于多模态特征之间的相互补充与矫正, 模型即使面临车辆遮挡以及光线骤变等特殊场景, 依然能够保证感知安全与准确. 其次, 应用信息论中基于信源信道联合编码的理论对感知模型中的特征提取和特征融合理论进行解释, 在增强复杂场景下感知能力的同时保证模型的可解释性. 最后, 构建了一种新的评价指标: 平均信息熵变 (average entropy variation, AEV), 对模型与外界的感知交互过程进行实时评估, 在模型感知交互过程中, 通过输出模型结构的特征信息的熵变程度对模型的感知交互稳定性进行了定量评价, 丰富了感知模型的评估方法, 增强了模型评估检测的可信度.

本文后续结构组织如下: 第 2 节介绍相关工作, 第 3 节基于信息论对可解释性的多模态感知模型设计思路以及可解释性的感知过程实时评估计方法进行设计与分析, 第 4 节介绍多模态感知模型的具体结构以及感知过程的底层机理, 第 5 节依据所提思路进行实验验证, 第 6 节指出智能驾驶多模态特征融合未来发展趋势和面临的挑战, 第 7 节对全文进行总结.

2 相关研究

本文主要研究基于信源信道联合编码的智能驾驶多模态感知, 研究包括智能驾驶感知安全、智能驾驶多模态深度融合感知和信源信道联合编码理论 3 个方面.

2.1 智能驾驶感知安全

智能驾驶感知安全是指智能汽车可以自动识别附近的安全障碍物, 以便避免发生事故. 它可以通

过雷达、视觉系统、超声波传感器等技术来实现, 以帮助司机更快、更准确地检测到危险, 从而防止发生事故 [16]. 面对复杂多变的交通环境, 如何确保智能驾驶安全是智能驾驶车辆走向工业生产的关键问题和难题.

Zhang 等 [17] 提出了一种基于双目摄像机的场景安全评价方法和一种综合安全评价模型, 可以对不同级别的自动驾驶采用不同的实时监控方法, 进一步提高自动驾驶的安全性. Rodionova 等 [18] 提出了利用度量时态逻辑 (metric temporal logic, MTL) 规范的鲁棒性作为连续安全度量, 自动探索自动驾驶汽车安全模型的性能. Wu 等 [19] 提出了一种基于博弈论组合 TOPSIS 模型的驾驶安全评价算法, 该方法能有效评价驾驶员驾驶行为的安全性. Luo 等 [20] 建立了城市道路交通安全评价的综合体系, 然后基于此评价体系利用模糊算法建立了城市交通状况的评价模型, 该模型可用于评价城市道路交通安全. Cai 等 [21] 基于车辆 OBD 驾驶行为数据及信息熵理论, 提出了城市道路交通安全风险预估方法, 能够精准、有效地进行交通事故预防预警. Sun 等 [22] 重点分析了真实驾驶情景下智能互联汽车 (intelligent connected vehicles, ICV) 的车辆避障主动安全控制, 并基于 ICV 技术进行了驾驶风险感知.

2.2 智能驾驶多模态深度融合感知

在智能驾驶领域, 感知周围环境是实现驾驶安全的前提. 目前, 普遍使用的单一模态传感器经常面临视野遮挡、数据稀疏等问题, 尤其在极端恶劣或复杂多变的环境中, 存在鲁棒性差以及准确率较低的情况 [23]. 为此, 研究者考虑利用相机图像和激光雷达等不同模态传感器之间的互补性实现多模态融合 [24], 从而提高感知准确性和精度, 提升智能驾驶车辆的鲁棒性和安全性.

Liang 等 [25] 提出利用多个相关任务来实现精确的多传感器三维目标检测, 通过融合不同层次的信息可以帮助网络更好地学习表征. Gao 等 [26] 提出了一种用于自动驾驶汽车视觉和光检测与测距雷达融合的目标分类方法, 保证了目标分类策略的有效性和效率. Wang 等 [27] 提出了一种在深度卷积神经网络 (convolutional neural networks, CNN) 融合激光雷达点云和相机捕捉图像的新方法. Prakash 等 [28] 提出了一种新颖的多模态融合 Transformer, 使用注意力来集成图像和 LiDAR 表示. Piergiovanni 等 [29] 提出了一种用于对齐三维点云和 RGB 传感器信息的 4D-Net, 4D-Net 能够更好地使用运动线索和密集的图像信息. Deng 等 [30] 提出了一种深度卷积神经网络来解决一般的多模态图像恢复和多模态图像融合问题. Sun 等 [31] 设计了一种新型的深度 MFNet 网络, 该网络可以利用多模态 VHR 航拍图像和激光雷达数据及其模内特征, 完成对周围任务的感知. 在机器人领域, 单一模态信息会限制机械手对物体的识别、抓取能力. Xiong 等 [32] 提出变分贝叶斯高斯 (Bayesian Gaussian) 混合条件生成对抗网络 (Bayesian Gaussian mixture-conditional generative adversarial network, BGM-CGAN) 的跨模态多样性噪声数据生成式方法.

2.3 可解释的信源信道联合编码理论

深度学习网络在智能驾驶领域已经取得了充分的发展, 但大多数精确的决策支持系统仍然是复杂的 “黑匣子”, 无法直接了解其内部逻辑和工作方式, 存在可信度和可解释性的问题 [33]. 为了促进各领域对机器学习系统的理解和信任, 系统的可解释性是必不可少的. 近年来, 学术界一直致力于可解释模型和解释方法的开发, 并从信息论的视角进行深度学习可解释性研究. 其中, 信源信道联合编码理论解决了可解释性的问题, 即从多个信号源中同时提取特征, 有效结合多个信号源的信息, 从而提高特征的准确性和鲁棒性 [34].

Liu 等 [35] 在结构化数据传输方向领域设计了一种基于深度学习架构的联合信源信道编码方案, 该方案明显优于独立的信源信道编码. Dommel 等 [36] 考虑了一种基于信源信道联合 (joint source-

channel, JSC) 编码的更有效的解决方案, 该方案通过基于类型的多址 (type-based multiple access, TBMA) 的非正交推广. Li 等 [37] 从深度学习模型的迁移机制和认知机理两个方面进行分析, 得到了关于迁移学习、归因方法应用和模型鲁棒性评估方面的相关结论, 填补了现有研究的空白. Abdallah 等 [38] 研究了一种信源信道联合编码方案, 以提高图像传输性能. Jankowski 等 [39] 提出了一种基于深度神经网络 (deep neutral network, DNN) 的针对检索任务的压缩方案, 该方法不仅提高了端到端精度, 而且简化和加快了编码操作, 这对功率和延迟受限的物联网应用非常有利. Wang 等 [40] 设计了一种基于特征融合的 MTL 网络 (feature fusion based MTL network, FFMNet), 用于联合目标检测和语义分割, 采用 JSCC 编码的 FFMNet 对各种信道条件具有较强的鲁棒性, 优于单独信源信道编码方案. Wang 等 [41] 提出了一种潜在空间强化学习方法, 可以获得提供语义信息和环境理解的可解释状态, 并用于高速公路入口匝道处自动驾驶车辆的可解释决策.

目前的驾驶感知安全研究主要集中在评判车辆面临危险场景时的决策举措, 从感知模型的感知过程分析驾驶感知安全的研究相对较少. 现有的基于多模态的感知模型虽然在感知精度上取得了较好的结果, 但通常不具有可解释性, 为驾驶感知安全带来了风险. 同时, 现有的研究对感知模型的评价主要集中在感知结果上, 对感知过程的评估相对较少. 首先, 本文提出了一种基于信源信道联合编码理论的多模态融合感知模型, 从提高模型感知过程稳定性的角度提升驾驶感知安全. 其次, 在模型设计过程中运用信息论的相关理论对模型进行设计, 以此提高模型的可解释性. 最后, 提出一种新的评价指标平均信息熵变 (AEV), 从感知过程角度对模型进行评价, 克服了传统方法仅从感知结果角度对模型进行评价导致缺乏可解释性的问题, 丰富了对模型的评价方式, 同时增强了模型的可解释性.

3 基于联合编码的可解释性多模态感知模型设计

分离定理证明了在信源信道模块长度趋近无穷时, 将信源编码与信道编码分开设计在理论上是最优的 [42]. 但由于实际应用中无法满足信源信道模块无限长的条件, 采用信源信道联合编码的方式效果往往优于分开编码 [43]. 因此本文选用联合编码的方式对模型进行搭建, 将信源编码定义为数据压缩的过程, 将信道编码定义为添加冗余信息的过程. 基于信息论中通信模型的设计方法, 对感知模型进行设计, 使感知过程可以运用信息论中信源信道编码的相关理论进行解释, 增强了模型的可解释性. 此外, 本文为了更好地用熵值变化表示模型稳定性, 对信息熵进行建模, 并基于所构建的熵变指标提出了一种针对模型感知过程的评价指标 —— 平均信息熵变, 对模型的感知过程的稳定性进行了评估, 使模型的感知性能可以更好地观测和量化, 增强了模型的可解释性.

3.1 信源信道编码的感知模型

传统无线通信系统中, 通常将信源编码与信道编码分开进行设计, 具有良好的简单性和通用性, 但没有达到最佳效率. 信源符号间具有一定的相关性和分布不均匀性, 使得信源数据存在冗余度, 信源编码通过数据压缩减少冗余增加系统有效性. 相反地, 信道编码通过在信息序列中适当添加冗余度纠正传输过程中的错误, 增加系统可靠性 [44]. 一定意义上, 二者是相互矛盾的. 因此, 合理匹配信源编码和信道编码是设计信源信道联合编码的重点. 根据香农 (Shannon) 第一定理:

$$\lim_{N\to\infty}\frac{\bar{L}_N}{N}=H_r(S), \tag{1}$$

其中, 信源的信息熵 $H_r(S)$ 是无失真信源压缩的极限值, 根据香农第一定理, 可知无失真的信源编码, 数据压缩的极限值为平均每个信源符号所需最少的 r 元码元数为信源的熵 $H_r(S)$. 此外根据香农公式

以及香农第二定理:

$$C = B \cdot \log_2 \left(1 + \frac{S}{N}\right), \tag{2}$$

其中, B 为信道带宽, $\frac{S}{N}$ 为信噪比, 单位为 dB, 信息传输率表示为

$$R = (1/T) \times \log_2 N, \tag{3}$$

其中, T 为一个数字脉冲信号的宽度 (全宽码) 或重复周期 (归零码), 单位为 s; 一个数字脉冲也称为一个码元, N 为一个码元所取的有效离散值个数. 当信道的信息传输率不超过信道容量时 ($R < C$), 采用合适的信道编码方法可以实现任意高的传输可靠性, 但若信息传输率超过了信道容量 ($R < C$), 就不可能实现可靠的传输. 因此, 为实现减少编码后信源数据的冗余度, 通过将通信系统设计为信源信道联合编码可以实现高效的信息传输, 使用最小化的信道符号来表达信源, 并对信源编码后的数据进行独立信道编码, 增加冗余度以降低错误率和干扰性 [45]. 给定具有独立同分布 $p(x)$ 的信源和有界的失真度量 $d(\hat{x}|x)$, 操作意义下的率失真函数等于信息论意义下的率失真函数, 即

$$R(D) = R^{(I)}(D) = \min_{\sum_{x,\hat{x}} p(x)q(\hat{x}|x)d(\hat{x}|x) \leqslant D} I(X;\hat{X}). \tag{4}$$

由香农第三定理可知, 只要码长足够长, 总可以找到一种信源编码, 使编码后的信息传输率 R' 略大于率失真函数 $R(D)$, 而码的平均失真度 $d(C)$ 不大于给定的允许失真度 D, 即 $R' \geqslant R(D)$, $d(C) \leqslant D$. 香农第三定理证明存在最佳编码方法, 但在实际应用中需要考虑符合实际信源的率失真函数 $R(D)$ 计算比较困难的问题以及寻找合适的达到极限值 $R(D)$ 的编码方法的问题 [46].

我们使用卷积神经网络对信源编码和信道编码进行联合设计, 以实现端到端的最优性能. 其中, 信源编码使用合适的网络得到结构化数据, 经过对准后作为信道编码的输入, 在信道编码中将两种模态的融合特征作为冗余信息补充到主模态中, 从而纠正信息在网络传输过程中的错误. 对于多模态信源信道联合编码, 我们通过协同调节信源编码与信道编码中的注意力机制的权重矩阵来调节网络参数, 实现不同模态间信源信道编码之间的相互作用. 感知模型中加入了信源编码以及信道编码模块, 使得模型在进行特征提取时具有明确的方向, 即信息熵减少的方向, 对模型特征提取的方向进行了理论解释. 在特征融合过程中, 运用信道编码添加冗余的理论解释了特征融合的过程, 即融合过程为向主模态添加冗余信息的过程. 系统地解释了模型在完成感知任务过程中的感知机理, 增强了模型的可解释性.

3.2 多模态融合网络的信息熵建模

由于网络传输信息的过程存在干扰, 信息传输的稳定性降低, 从而影响模型的感知安全能力 [47]. 熵值变化的稳定程度作为网络稳定性的量化度量, 其变化情况直接与网络性能相关联. 编码程度优化的本质在于联合编码模块中深度神经网络 (DNN) 模型中实际输出值 $f_\theta = \Pr[\hat{Z}|X]$ 与真实值 $f_\theta^* = \Pr[Z|X]$ 间的贝叶斯 (Bayes) 误差优化问题, 即优化使输出值与真实值之间信息熵趋向于零的过程. 在传统方法中, 若只有预估的概率分布 Q, 使用估计得到的概率分布, 计算估计的信息熵为

$$\text{Entropy } = \mathbb{E}_{x\sim Q}[-\log Q(x)]. \tag{5}$$

但由于概率分布与真实分布存在差异, 因此对于有真实预测分布 P 的网络模型训练, 通常采用交叉熵对估计的信息熵进行替代, 交叉熵的计算为

$$H(P,Q) = \mathbb{E}_{x\sim P}[-\log Q(x)], \tag{6}$$

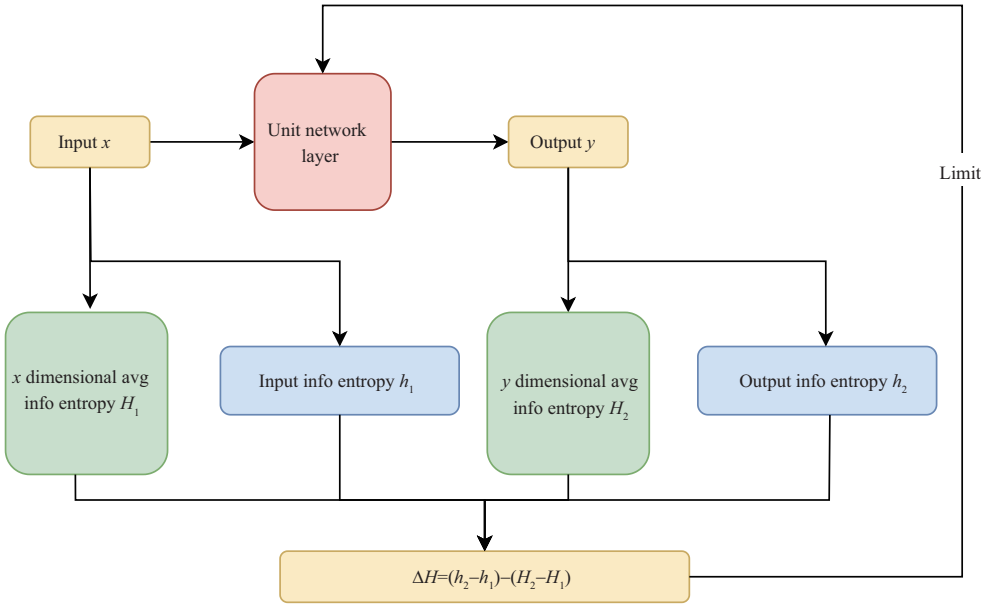


图 1　(网络版彩图) 熵变指标

Figure 1　(Color online) Entropy change indicator

并在模型优化过程中采用交叉熵损失函数对概率分布的相似程度进行度量 [48]: 假设真实分布为 y, 网络输出分布为 $\hat{y}$, 总的类别数为 n, 当前 batch 的样本数为 m. 在单类别任务中, 一个图像对应一个标签, 其交叉熵损失函数为

$$\text{Loss} = -\sum_{i=1}^{n} y_i \log \hat{y}_i, \tag{7}$$

对应一个 batch 的交叉熵损失函数:

$$\text{Loss} = -\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} y_{ji} \log \hat{y_{ji}}. \tag{8}$$

在多类别任务中, 一个图像样本可以有多个标签, 把网络最后一层的每个神经元视为任务中的一个类别, 交叉熵损失值的计算属于二项分布:

$$\text{Loss} = -y \log \hat{y} - (1-y)\log(1-\hat{y}), \tag{9}$$

对应一个 batch 的交叉熵损失函数:

$$\text{Loss} = \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{n} -y_{ji} \log \hat{y_{ji}} - (1-y_{ji})\log(1-\hat{y_{ji}}). \tag{10}$$

对于没有真实预测分布的情况, 对输出数据进行还原后将输入数据作为真实分布进行比较, 计算获得交叉熵损失值.

由于贝叶斯误差优化方法的评估方式只能对感知结果的准确程度进行表征, 但是无法有效反映感知过程中信息熵的变化程度, 本文对网络的信息熵进行建模, 如图 1 所示. 通过计算单位网络层输入和输出对应维度的平均分布信息熵 H, 建立对应的基准信息熵 H_1, H_2; 计算输入、输出概率分布信息

熵 h_1, h_2 在平均分布上增加的信息量, 得到与基准之间的相对熵 $h_1 - H_1, h_2 - H_2$; 其中单位网络层是指一层神经网络, 本文通过对单层神经网络的信息熵进行计算, 可以更好地反映出模型特征提取过程数据的变化情况, 从而有效地评估模型的稳定程度. 在面对由多层网络构成的网络结构的情况下, 在计算网络整体的熵变指标时, 首先对各层网络的熵变指标分别进行计算, 而后将各层网络的熵变指标的期望作为最终模型网络层的熵变指标. 整理相对熵的变化式, 得到类似于交叉熵的熵变指标:

$$\Delta H = (h_2 - h_1) - (H_2 - H_1). \tag{11}$$

同时, 该熵变指标同样适用于多模态网络模型, 通过计算不同模态网络层输出的熵变指标来反映各层特征提取过程信息熵的变化程度. 通过对熵变指标的快速计算, 实现对各网络层信息熵的变化程度进行有效表征, 从而更好地反映数据通过网络结构时模型感知过程的稳定程度.

3.3 基于可解释性的感知过程实时评估

深度学习被认为应该在和外界实时的交互和迭代中完成 [49], 因此, 我们认为对模型感知过程的评估也应该在和外界的交互过程中实时进行, 根据外界的现实场景实时评价感知过程的稳定程度, 保证感知过程的准确与稳定. 为准确分析深度融合网络的感知过程提供量化度量, 本文设计深度融合网络的信息熵计算方法, 通过引入一种熵的计算方法来描述失真受限编码过程中实时采集的真实场景数据信息量的变化, 从而量化模型中特征提取与特征融合后数据信息熵的变化趋势与规律, 对模型感知任务的过程进行定量解释.

在通信模型和信息论中, 熵常被用作信息量的度量. 熵值越低, 信息量越大 [50]. 而在一个可信度较高的通信模型中, 信息压缩网络各层信息变化是稳定的, 保证了信息压缩的平滑性, 在信息穿过网络各层时, 保持熵变不变, 可以防止信息的突然失真 [51]. 在感知模型中, 对于一个具有特定功能的特征提取网络, 原始输入数据经过每个特征提取层后形成一个新的特征图, 而在这个过程中会出现信息量的变化, 即信息熵的变化. 但是, 由于在特征提取时没有引入新的信息, 过滤掉了大量与目标任务无关的信息, 只保留了与目标任务相关的信息, 该过程同样为信息熵值减小的过程, 因此可以用信息熵来反映模型中数据特征的压缩情况. 通过对经过网络层数据特征信息熵的计算可以对网络层的特征提取过程进行量化解释, 即保留有效信息去除冗余, 减小信息熵的过程. 同时, 在感知模型同一特征提取网络中, 网络各层的特征提取能力与该层参数的规模呈正相关. 因此, 当特征图经过参数数量相同、结构相似的若干个连续的特征提取层时, 每次特征提取后都应保持相似的信息减少量, 以保证网络的平滑信息压缩. 而对于一个网络稳定性低的模型, 在处理某些存在噪声信息的数据时, 特征图信息量在经过特征提取器每一层时的变化会变得不规则, 而不是平滑减少, 熵值甚至可能增加. 因此, 我们可以通过计算经过感知模型相似网络结构的熵值变化反映出模型进行特征提取过程的稳定程度, 为感知模型的可解释性提供了量化度量. 最终, 我们选用熵值的变化程度来反映检测网络的可信度.

基于熵的期望来反映各层输出数据的压缩程度需要一种方法来估计信息压缩网络中各层输出的熵. 此外, 估计熵需要对输出数据的分布进行概率建模. 在信息压缩中, 通常将卷积网络作为选择. 为了将卷积神经网络建模为概率模型, 我们将其设置如下. 卷积核产生的特征通道 $\{x_1, x_2, \ldots, x_i\}$ 被认为是多维连续型随机变量 X 的样本, 其中, i 是通道数, 每个通道中的值的个数是 X 的维数 d. 因为任何网络的每一层的输出都可以被认为是连续型随机变量 X, 而任何层的输出都可以被选为 X 的一组样本 x_i, 概率建模也可以应用于除卷积神经网络之外的其他神经网络结构, 计算数据分布的熵. 因此, 所提出的卷积神经网络概率建模方法可以推广到其他神经网络结构.

本文通过计算连续随机变量的微分熵来根据未知的连续随机变量 X 的概率分布来估计熵, 其中,

$f(x)$ 为 X 的概率密度函数. 微分熵 $h(X)$ 定义如下:

$$h(X) = -\int f(x)\log f(x)\mathrm{d}x. \tag{12}$$

随机变量的概率分布是事先不知道的, 因此概率密度函数是未知的, 并且在该概率分布中只有有限数量的样本值可用. 因此, 本文使用 KNN 方法来计算信息熵. 因为连续变量是离散的采样, 并使用 n 个样本来近似整个样本空间, 每个样本点被扩展成一个 d 维超球体, 球体的半径是样本点和最近的样本点之间的距离. 当变量在样本空间中均匀分布时, 每个样本点的概率可近似为 $1/n$. 由于随机变量在样本空间中的分布是未知的, 与均匀分布可能存在较大差异, 利用样本在空间中的分布对随机变量在空间中的分布进行修正. 样本空间中样本的密度和稀疏性直接影响每个样本点附近的概率密度. 每个样本点的离散概率估计为

$$p(x_i) = [(n-1)\cdot r_d(x_i)^d]\cdot V_d^{-1}, \tag{13}$$

其中, n 为样本个数, $r_d(x_i)$ 为样本 x_i 与其最近样本点之间的 d 维欧氏距离, V_d 为 d 维空间中单位球面的体积. 随机变量 X 的熵的估计值 [52] 为

$$H(X) = \frac{1}{n}\sum_{i=1}^{n}[-\log p(x_i)] + \gamma, \tag{14}$$

其中, γ 是欧拉 – 马瑟罗尼 (Euler-Mascheroni) 常数, 约等于 0.5772. K 近邻熵估计方法将每个样本点与其最近的样本点之间的距离扩展到与其最近的第 k 个样本点之间的距离, 随机变量 X 的熵估计变为

$$H(X,k) = -\psi(k) + \psi(n) + \log V_d + \frac{d}{n}\sum_{i=1}^{n}\log r_{d,k}(x_i), \tag{15}$$

其中, ψ 是 D_i-gamma 函数, $\psi(1) = -\gamma$, $\psi(n) \sim \log(n-1)$, $r_{d,k}(x_i)$ 是样本 x_i 与其最近的第 k 个样本点之间的 d 维欧氏距离. 最后, 我们用上述计算出的熵值反映网络层输出信息量, 将平均信息熵变 AEV 定义为

$$\mathrm{AEV}(t) = \frac{\sum_{n=1}^{N}(\Delta H_n(t) - \widehat{\Delta H}(t))^2}{N}. \tag{16}$$

我们使用 $H(t)$ 作为网络各层实时输出的熵值 (如图 2 所示), 然后得到网络各层的熵变量 $\Delta H_n(t)$, $\Delta H_n(t) = H_{n+1}(t) - H_n(t)$, 其中, n 是网络各层的索引. AEV(t) 可以有效地反映单模态网络以及多模态网络的稳定程度. 在单模态网络中, 通过对数据经过相似网络结构信息熵的变化程度计算 AEV(t), 同时, 在如 U-net 类型具有的特殊结构的网络中, 我们依然可以分别对编码器和解码器的部分进行 AEV(t) 数值的计算, 并计算均值以反映模型的编码和解码过程. 具体地, 我们可以对编码器中的特征提取模块中的 block 计算 AEV(t), 再对解码器中特征恢复模块 block 计算 AEV(t), 计算均值作为 AEV(t) 的最终取值, 从而通过 AEV(t) 的数值来反映单模态模型感知过程的稳定程度. 在多模态网络中, 先分别计算数据经过不同模态网络结构数据压缩部分时 AEV(t) 的数值, 再将模型不同模态数据压缩与特征融合过程中的 AEV(t) 数值进行叠加来求取整体多模态网络的 AEV(t) 数值, 从而判断模型整体感知过程的稳定程度. 通过 AEV(t) 的数值来反映模型在进行感知的过程中相似网络层在信息压缩过程中熵变的稳定程度, 当 AEV(t) 的数值越低时, 模型的感知过程就越稳定. 通过评价指标 AEV(t) 的引入, 补充了模型质量的评估方法, 从感知过程的角度对模型进行评判, 更能实时反映模型感知的状态. 此外, 通过对模型感知过程稳定程度的评估, 也增加了人们对模型感知结果的信心, 体现了模型可解释性的意义与价值.

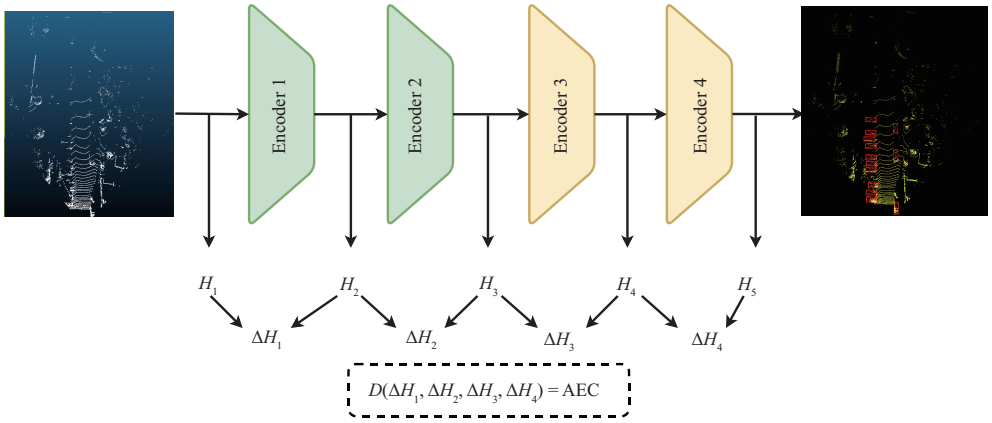


图 2　(网络版彩图) 提出的评价指标 AEV 的详细说明

Figure 2　(Color online) An detailed illustration of our proposed AEV method

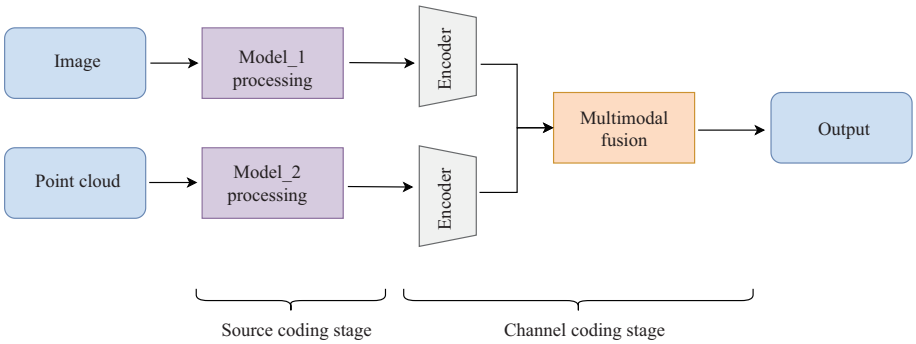


图 3　(网络版彩图) 信源信道联合编码

Figure 3　(Color online) Joint source-channel coding

4　基于联合编码的多模态特征深度融合网络

在感知模型的设计过程中, 通常存在一种在准确性和简单性或可解释性之间的权衡 [53]. 经典的基于规则或专家系统 [54] 是高度可解释的, 但不是非常准确 (或健壮) 的. 本节针对智能驾驶感知安全的需求, 突破主流多模态深度融合网络构建方法, 从信息论的思想出发, 构建基于信源信道联合编码的多模态融合模型, 提出基于联合编码的多模态深度融合方法, 如图 3 所示. 其中, 信源编码模块与信道编码模块均按照通信模型的设计思路进行设计与解释, 与传统的基于深度卷积的网络结构相比, 模型具有更强的可解释性, 可以更好地对模型的特征提取以及特征融合过程进行解释与评估.

4.1　基于信源编码的特征提取网络

由于信源符号之间存在分布不均匀等特点, 信源存在一定的冗余度. 通过信源编码对信源符号进行变换, 能够减少冗余, 从而提高通信有效性 [55]. 具体而言, 信源编码过程需要针对信源输出符号序列的统计特性, 寻找一种方法以确保最可能的信源符号由最短的码字表示 [56]. 该方法需要保证后者的各码元所载荷的平均信息量最大, 且能保证无失真地恢复原来的符号序列. 在通信模型中, 信源数据的压缩编码方案一般采用基于卷积神经网络、循环神经网络的自编码器结构来进行 [57]. 自编码器信源编码 (source coding with autoencoders) 是一种用于压缩数据的技术, 它可以通过将原始数据转换

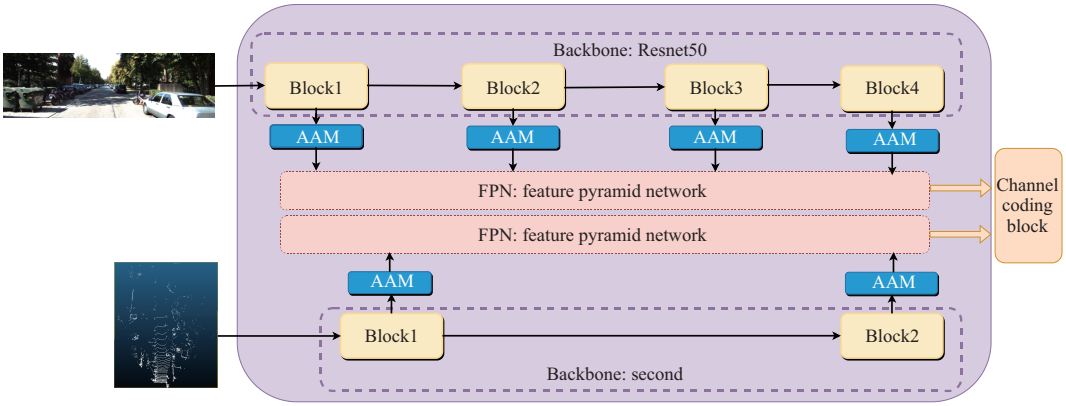


图 4　(网络版彩图) 信源编码

Figure 4　(Color online) Source coding

为更小的表示来减少数据量进行传输. 自编码器信源编码可以通过学习训练将原始数据映射到更小的维度 [58]. 同时, 自编码器信源编码可以在减少数据量的同时保持原始数据的重要特征, 从而使其能够被重新恢复. 而在本文提出的感知模型中, 我们借鉴自编码器的思想, 设计一种能够将信源数据映射到更小维度的神经网络, 在减少冗余信息 (数据压缩) 的同时, 保证特征信息的稳定以及全面. 为了保证压缩后的信源数据能够无失真地恢复原来的符号序列, 我们选用一种轴向注意力机制 (axial attention model, AAM) 对数据压缩方向进行指导, 通过改变注意力权重调整数据压缩方向, 保证了压缩过程的稳定性以及准确性. 在网络中引入注意力机制, 通过利用 Query 和 Key 计算权重系数, 再用该系数对 Value 加权求和, 使感知模型能够更有效地抽取有效特征, 从而保证了数据压缩的准确性. 同时, 感知模型的数据压缩方向通过调整注意力权重的方法得以确定, 保证了数据压缩朝着特征信息增加、冗余信息减少的方向进行. 本文通过在主干网络的模块间加入自注意力机制, 强化模型捕捉数据内部相关性的能力, 实现基于深度学习方法的信源编码过程.

基于信源编码的特征提取网络结构如图 4 所示. 首先使用稀疏嵌入的目标检测卷积 (sparsely embedded convolutional detection, SECOND) 网络作为骨干结构, 处理激光雷达输入的 3D 点云图像. 其整体结构依次包括: 点云体素化、体素特征提取、3D 稀疏卷积块、RPN 层 (下采样块 + 上采样块), 以及分类分支、回归分支和车头方向分支. SECOND 网络在 VoxelNet 的 Convolutional Middle Layer 基础上引入稀疏卷积操作 [59], 如图 5 所示. 本模型的信源编码部分, SECOND 网络每个块的输出首先经过轴向注意力机制模块, 而后使用特征金字塔 (feature pyramid network, FPN) 网络对两个注意力模块的输出进行融合.

同时, 我们使用残差神经网络 (Resnet50) 作为骨干网络, 处理输入的 RGB 图像. 4 个块逐层改变信息提取的语义层次. 和 SECOND 网络相同, 本模型的信源编码部分, 在 Resnet50 网络的每两个块之间加入空洞卷积模块. 每个块的输出经过轴向注意力模块后, 再由 FPN 网络进行金字塔式的特征融合. 由此实现不同语义层次下, 自注意力机制指导的特征提取. 两个模态提取出的特征图将被输入信道编码进行模态融合.

4.1.1　轴向注意力机制模块

综合考量信源编码模型复杂度和模型效果, 我们选用轴向注意力机制模块进一步提取特征. 在详细介绍轴向注意力之前, 首先回顾普通的基于自回归建模的注意力机制计算方法. 一个自注意力层将

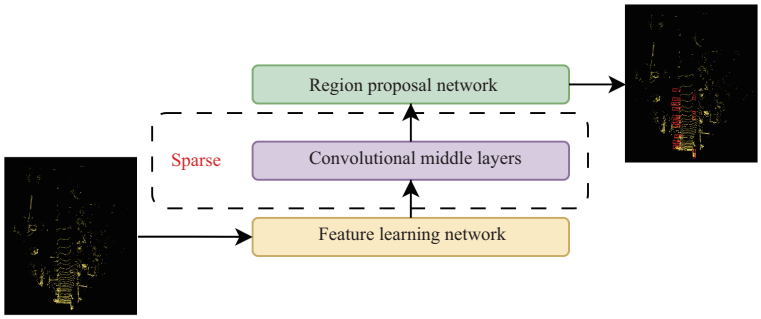


图 5 (网络版彩图) 基于 VoxelNet 改进的 SECOND 网络框架

Figure 5 (Color online) Based on VoxelNet's improved SECOND network framework

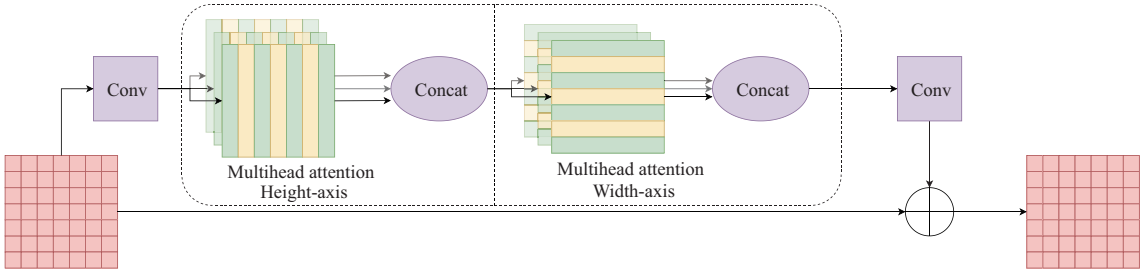


图 6 (网络版彩图) 轴向注意力机制

Figure 6 (Color online) Axial attention mechanism

一个 $N \times D$ 的矩阵 X 作为输入, 通过以下公式生成输出序列 Y, Y 也是一个 $N \times D$ 的矩阵.

$$\begin{aligned} &Q = XW_Q,\ K = XW_K,\ V = XW_V, \\ &A = \mathrm{softmax}(QK^{\mathrm{T}}/\sqrt{D}),\ Y = AV, \end{aligned} \tag{17}$$

其中, W_Q, W_K 和 W_V 是 $D \times D$ 的参数矩阵, 分别负责将矩阵 X 的每个条目投影为键、查询和值. 输出序列 Y 的每个条目是由注意力矩阵 A 加权的 V 中值的线性组合, 注意力矩阵 A 本身由所有查询和键向量对之间的相似性计算得到.

本文信源编码部分使用的轴向注意力机制的做法为, 对于 k 个轴, $\mathrm{Attention}_k(x)$ 构件在输入序列 x 的 k 轴上执行自注意力计算操作, 沿着 k 轴混合信息, 同时保持沿着其他轴的信息独立. 同一轴向内的采样顺序为逐个通道、逐个列、逐个位置. 轴向注意力的感受野是目标像素的同一行 (或者同一列) 的 W (或 H) 个像素. 和传统的自注意力机制相比, 轴向注意力机制大大降低了计算复杂度. 它的效率使我们能够关注大的区域, 并建立模型来学习全局信息的交互. 图 6 为轴向注意力机制在二维图像上的应用示意.

4.1.2 金字塔式特征融合模块

卷积神经网络由浅到深, 语义信息越来越丰富, 但特征图越来越小, 分辨率越来越低. FPN 融合模块同时利用低层特征高分辨率和高层特征的高语义信息, 将浅层的信息传递到深层, 以解决深层特征图容易忽略小目标的问题, 提升特征提取效果 [60].

整个信源编码网络通过对不同模态输入数据的规格分解, 实现对输入信息不同语义层次的解读;

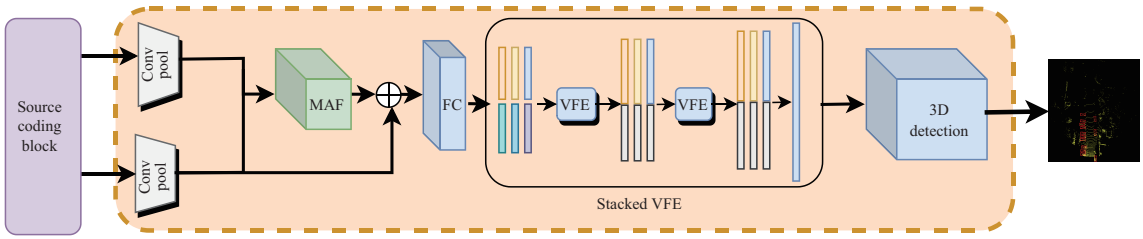


图 7 (网络版彩图) 信道编码

Figure 7 (Color online) Channel coding

通过对每个语义下特征图的自注意力计算, 突出了该层次中的重点信息, 优化了特征提取, 尽可能去除了冗余信息和无关信息; 最后通过金字塔式特征模块将不同语义层次下的信息整合.

4.2 基于信道编码的跨模态融合网络

由于信息在网络传输过程中, 不可避免地会出现扰动以及衰弱等现象, 因此需要通过信道编码的方式对信号进行相应的纠错, 来增强信息传输时信道的抗干扰能力. 为此, 我们构建了如图 7 所示的信道编码网络. 信道编码通过添加噪声来克服传输过程带来的干扰, 添加的噪声会抵消外界的干扰, 使得传输的信号和外界的干扰信号不会发生重叠, 从而抵消外界的干扰, 使传输的信号可以正常传输 [61]. 本文的感知模型借鉴信道编码的方式, 在点云特征中引入 “噪声”, 用来抵消特征提取过程中造成的损失. 为了保证引入的 “噪声” 能够抵消数据传输过程中的损失干扰, “噪声” 的选取十分重要. 由于单一模态下的点云数据存在数据稀疏, 以及在大雨、大雾天气中激光雷达存在检测效果下降等问题 [62], 点云数据在感知模型中经过卷积运算时容易造成特征提取错误或者特征提取不足等问题, 而通过合适的 “噪声” 引入, 可以对点云数据进行补充矫正, 从而能够完成高效的特征提取与数据传输. 同时, 信道编码过程之所以能够检出和校正接收比特流中的差错, 是因为加入了一些冗余比特, 把几个比特上携带的信息扩散到更多的比特上 [63]. 因此, 我们可以将 “噪声” 定义为一种冗余信息, 用于抵消外界的干扰, 并能够对数据进行矫正与检出. 在感知模型中, 我们选择激光雷达数据信息的冗余特征与互补特征作为冗余信息, 也就是通信模型中的噪声, 通过将冗余信息补充到点云数据中, 保证点云数据在感知模型中的稳定传输. 其中, 激光雷达的冗余信息主要是辅模态中包含的相似信息以及互补信息, 例如大雨或大雾天气下雷达失真但视觉传感器感知正确的语义特征等. 通过将冗余信息添加到主模态的数据信息中, 感知正确的辅模态的语义特征对主模态的语义特征进行矫正, 从而保证了感知结果的准确性以及感知过程的稳定性.

4.2.1 多头注意力机制融合模块

在冗余特征的选取中, 为了使特征提取的过程更加准确稳定, 我们选取一组与原始的点云特征具有较高的相关性以及互补性的特征作为冗余特征. 具体地, 我们设计了一个如图 8 所示的多头注意力机制融合 (multihead attention fusion, MAF) 模块来获取冗余特征, 该模块利用模块中的注意力机制来纳入图像信息以及激光雷达信息之间的互补信息, 从而起到对主模态信息进行检测和补偿的作用. 该模块的输入为两种不同模态的特征图, RGB 图像模态的特征图记为 $F_r^{\text{in}} \in \mathbb{R}^{N_r \times D_r}$, 3D 点云模态的特征图记为 $F_l^{\text{in}} \in \mathbb{R}^{N_l \times D_l}$, 其中, N_r 代表二维图像模态的通道数, N_l 代表三维点云模态的通道数, D_r 代表二维图像特征矩阵的维度, D_l 代表三维点云特征矩阵的维度.

为了获取具有互补性的融合两种不同模态的冗余特征, 我们分别用线性变换求取了 MAF 模块

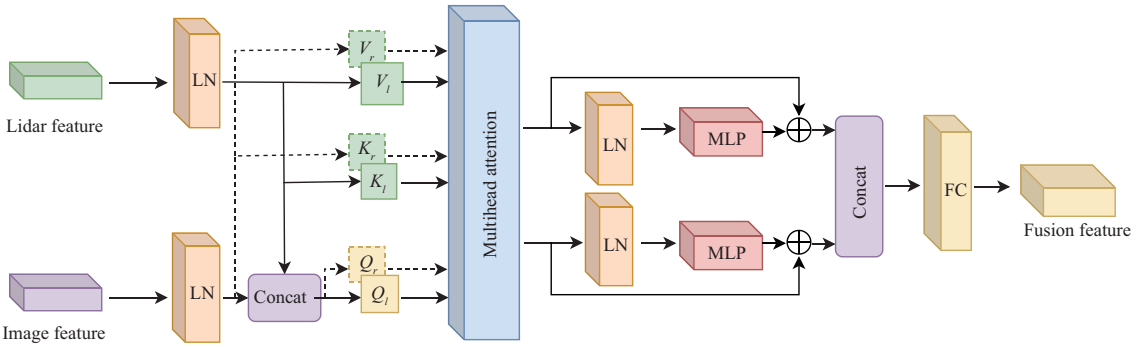


图 8 (网络版彩图) 多头注意力机制融合模块

Figure 8 (Color online) Multihead attention fusion

RGB 图像模态所对应的权重矩阵 Q_r, K_r, V_r, 以及 3D 点云模态所对应的权重矩阵 Q_l, K_l, V_l:

$$Q_r = \text{concat}(F_r^{\text{in}}, F_l^{\text{in}})M_r{}^q,\ K_r = F_r{}^{\text{in}}M_r{}^k,\ V_r = F_r{}^{\text{in}}M_r{}^v, \tag{18}$$

$$Q_l = \text{concat}(F_l^{\text{in}}, F_r{}^{\text{in}})M_l{}^q,\ K_l = F_l{}^{\text{in}}M_l{}^k,\ V_l = F_l{}^{\text{in}}M_l{}^v, \tag{19}$$

其中, $M_r{}^q \in \mathbb{R}^{D_r \times q_r}, M_r{}^k \in \mathbb{R}^{D_r \times k_r}, M_r{}^v \in \mathbb{R}^{D_r \times v_r}$ 分别为 RGB 图像模态下的初始权重矩阵; $M_l{}^q \in \mathbb{R}^{D_l \times q_l}, M_l{}^k \in \mathbb{R}^{D_l \times k_l}, M_l{}^v \in \mathbb{R}^{D_l \times v_l}$ 分别为 3D 点云模态下的初始权重矩阵. 之后我们将各模态的注意力计算为

$$A_r = \text{softmax}\left(\frac{Q_r {K_r}^{\mathrm{T}}}{\sqrt{k_r}}\right)V_r, \tag{20}$$

$$A_l = \text{softmax}\left(\frac{Q_l {K_l}^{\mathrm{T}}}{\sqrt{k_l}}\right)V_l. \tag{21}$$

之后通过一个非线性变换分别将不同模态的输出转换成与输入特征尺寸相同的注意力特征矩阵并将其与变换前的特征相加获得注意力特征 A_r' 与 A_l':

$$A_r' = \text{MLP}(A_r) + A_r, \tag{22}$$

$$A_l' = \text{MLP}(A_l) + A_l. \tag{23}$$

最后将不同模态的注意力特征进行拼接, 形成一个新的特征矩阵 Z, 再将拼接后的特征矩阵 Z 与权重向量 W_o 相乘从而得到最终与输入特征尺寸相同的特征矩阵:

$$Z = \text{concat}(A_r', A_l'), \tag{24}$$

$$F^{\text{add}} = ZW_o.$$

通过两种特征图的拼接, 保证了注意力图语义特征的全面和准确. 其中, F^{add} 中包含了主模态以及辅模态的特征信息, 其与主模态的特征信息相比, 既包含了主模态的部分特征信息, 又包含了辅模态的补充信息. 因此将其与原始输出特征 F_{in} 进行拼接形成冗余特征, 将 F^{add} 作为冗余信息引入到主模态的特征中, 使得特征信息扩展到更高维度的矩阵上. 之后通过非线性变换将冗余特征变换成与主模态特征相同尺寸的特征矩阵:

$$F^{\text{out}} = \text{concat}(F^{\text{in}}, F^{\text{add}})W. \tag{25}$$

与 LiDAR 点云相比, RGB 图像包含了更丰富的颜色和纹理信息 [64]. 在信道编码过程中, 为了提高模型感知过程的检测精度以及稳定性, 我们从 RGB 图像中提取高级语义特征作为辅模态, 补充到 LiDAR 点云的主模态中. 通过模块中的注意力机制将不同模态的特征进行融合, 获得包含互补信息的融合特征, 从而将获取信道编码中的冗余信息作为 "噪声" 加入到主模态的点云信息中, 对主模态中添加冗余信息, 使得主模态的点云特征拓展到更高的维度上, 从而避免了信源编码部分数据的过度压缩所可能带来的网络的不稳定, 保证模型中的特征信息在传输过程不受外界干扰造成损失 [65]. 同时, 本文提出的感知模型, 从信息论的角度出发, 解释了多模态融合的过程是一个通过添加冗余特征抵抗外界干扰的过程, 利用冗余特征中相关性以及互补性的特征与单一模态的语义特征融合, 使得模型中的特征更加完整全面, 抵消了卷积运算为数据带来的部分特征损失, 保证了模型的稳定性, 同时增强了模型的可解释性.

4.2.2 多层体素特征编码模块

多层的体素特征编码 (stacked voxel feature encoding, VFE) 为特征学习网络, 通过对融合图像信息的点云进行编码, 将 3D 空间划分为等间距的体素, 同时将点云分组为体素 [66]. 使用体素特征编码层对体素进行编码, 将体素内的每个点连接一个全连接网络提取逐点的特征, 并对这些特征进行元素级的池化形成局部聚合特征, 最后将局部聚合特征串接到每个点的特征上, 从而完成一次特征提取, 最后依次堆叠 VFE 层可以获得更高层次的特征 [67]. 堆叠的 VFE 层的输出通过一组卷积中间层, 来聚集逐渐扩展的感受野内的体素特征, 同时附加上下文信息来提高检测性能. 在卷积中间层之后, 区域提议网络执行 3D 目标检测, 该网络由 3 个完全卷积层组成. 在每个卷积层之后, 应用 BN 和 ReLU 操作. 然后将每个块的输出上采样到固定大小, 并连接起来以构建高分辨率特征图. 从而完成最后的检测任务.

该网络改进了传统的基于点融合的方法, 通过注意力机制使图像特征与雷达点云信息建立关联并融合. 首先该网络使用预训练的检测网络从图像中提取基于图像的语义的高级特征图, 以及基于点云的特征信息. 而后通过 MAF 模块将两种特征信息融合获取融合特征, 将融合特征作为 "噪声" 与主模态特征拼接, 通过权重矩阵对最后特征进行矫正. 最后将通过 MAF 模块与卷积层的堆叠输出特征经过一层 VFE 处理, 获取最后用于检测阶段任务的特征. 这种方法的优点在于, 该网络可以通过 MAF 模块从两种模态中总结有用的融合信息, 帮助信道编码获取冗余信息, 从而完成信道编码过程. 此外, 该方法利用 LiDAR 点云并将相应的图像特征补充到 3D 点云的坐标特征中. 所获取的最终特征中包含更多的语义信息, 从而使该网络结构最终能够获取更好的检测结果.

信道编码在整体框架中, 运用信息论的编码理论, 对感知模型完成不同模态的融合过程进行了重新设计. 借鉴通信模型中向信道中引入噪声的方式 [68], 制定了新的融合方法, 将融合特征作为 "噪声" 加入到主模态中, 以保证主模态的数据信息能够抵消网络所带来的干扰. 同时, 其余模态特征中的冗余信息以及互补信息的补充, 使各个模态数据传输过程中能够抵消掉网络结构以及外界噪声带来的干扰, 保证了数据传输的准确与稳定. 从底层机理解释了多模态感知模型的融合过程以及不同模态融合后所带来的性能提升的原因, 提升了模型的可解释性.

4.3 基于信源信道联合编码的多模态特征融合

基于信源信道联合编码的设计可以根据信道的状态和噪声水平来自适应地调整编码策略, 从而提高通信系统的性能. 这种自适应性能在深度学习模型中可以通过模型训练来实现. 本文使用卷积神经网络对信源编码和信道编码进行联合设计保证编码过程的稳定准确, 从而使得模型端到端的性能达到

表 1 AEV 有效性

Table 1 Effectiveness of AEV

Setting		Value	AEV			Confidence		
Model	Dataset		Clean	Noise1	Noise2	Clean	Noise1	Noise2
VoxelNet	KITTI	Mean	0.015	0.008	0.009	0.495	0.248	0.248
		Change (%)	0.0	−48.5	−39.1	0.0	−49.9	−49.9
PointPillars	KITTI	Mean	0.012	2.086	0.008	0.487	0.344	0.344
		Change (%)	0.0	17475.6	−36.1	0.0	−29.3	−29.3
PointPillars	nuSenes	Mean	0.034	1.918	0.016	0.168	0.128	0.128
		Change (%)	0.0	5494.7	−54.5	0.0	−23.7	−23.7

最优. 信源编码使用合适的网络得到的结构化数据作为信道编码的输入, 信道编码也选用合适的网络结构来获取冗余特征, 从而将原始特征拓展到更高的维度上来提高信息传输的可靠性. 信道编码包含多层卷积网络, 因此增加了特征提取网络的深度, 信道编码则应用 MAF 模块获取冗余信息后与主模态特征融合. 本文提出了基于多模态信源信道联合编码的深度融合网络方案, 利用信道编码中的 MAF 模块的注意力机制以可变的权重矩阵更新调整注意力参数来使得信道编码过程能够更好地进行主模态与辅模态融合. 将融合特征作为冗余信息, 利用包含两种模态特征的融合特征, 将原始特征拓展到更高的维度, 补充了原始特征的同时, 又能起到一定的检错和矫正作用, 保证了模型感知过程中的稳定与准确.

同时由于信源编码与信道编码中均包含随训练过程变化的权重矩阵, 因此本文将信源编码与信道编码中的权重矩阵, 作为深度学习过程中的训练参数, 通过同时对信源信道编码中的权重矩阵进行动态调整, 实现信源信道编码之间的相互作用, 完成信源信道联合编码的训练过程.

5 实验

本文的实验在 NVIDIA A100-SXM4-80GB 设备上进行, 基于 MMDetection3D 框架, 使用 PyTorch 建立模型. MMDetection3D 是基于 PyTorch 的开源对象检测工具箱, 面向通用 3D 检测的新一代平台, 是 MMLab 开发的 OpenMMLab 项目的一部分. 首先对 AEV 评价指标的可靠性进行验证实验, 证明 AEV 对模型感知过程稳定性评估的可靠性. 而后, 对本文提出的基于信源信道联合编码的可解释多模态感知模型进行了消融实验验证, 证明了模型结构的可靠性. 最后, 将本文提出的感知模型与现有的感知模型进行比较, 证明了模型在保证较高的感知精度的同时有着极强的感知稳定性.

5.1 AEV 评价指标的可靠性分析

AEV 数值的变化可以对模型的感知过程的稳定性进行判别, 当数据中包含大量无用噪声时, 模型感知过程容易出现异常, 各层输出数据信息熵变化剧烈程度, 可以通过 AEV 的数值进行判断. 因此, 我们比较了 VoxelNet 模型和 PointPillars 模型在 KITTI 以及 nuSenes 数据集上, AEV 对异常数据造成的感知过程的不稳定程度, 然后计算处理模型中数据集的结果的置信度与 AEV. 将实验结果的 AEV 与实验结果的置信度进行比较, 证明 AEV 对感知过程稳定性评估的可靠性, 结果如表 1 所示.

通过将模型获得的置信度与处理不同测试数据集后的 AEV 值进行比较. 研究发现, 在处理不同模型下添加不同噪声的数据集时, 添加噪声后的置信度低于不添加噪声时的置信度. 可以识别出此时的感知过程是异常的, 但是, 当噪声添加率不同时, 结果的置信度变化很小, 对数据异常的敏感性很低,

表 2 消融实验

Table 2 Ablation study

Model	AAM	FPN	MAF	mAP		AEV		
				bbox	BEV	AEV-channel	AEV-source	AEV-joint
Without-AAM	×	√	√	93.3073	90.0549	0.628824	0.131481	0.760305
Without-FPN	√	×	√	92.8058	87.9710	0.710557	0.007787	0.718344
Without-MAF	√	√	×	93.0922	89.1565	0.630544	0.003330	0.633874
Ours	√	√	√	95.0187	90.7390	0.587259	0.003189	0.590448

并且没有区别, 不能判断噪声添加率. 因此, 置信度存在对添加噪声所引起感知过程不稳定的过度解释问题, 但在这种情况下的 AEV 值可以识别数据异常. 并且, 它对一些具有特定比例噪声的数据特别敏感, 并且可以检测不同大小噪声的添加. 也就是说, 当使用相同的模型来获得相同的检测结果时, 使用 AEV 作为指标可以更好地判断数据异常所引起的感知过程的稳定程度, 并且可以在一定程度上辨别检测过程中出现的异常数据, 从而能够对感知过程的稳定程度进行可靠评估.

5.2 感知模型的结构分析

在模型设计过程中, 为了增加信源编码过程中数据压缩以及信道编码中信息矫正的稳定程度, 分别引入 AAM, FPN, MAF 模块. 为了验证各个模块对模型感知过程以及感知结果的影响和各个模块起到的作用, 我们对各个模块进行了消融实验, 结果如表 2 所示.

表 2 比较了不同网络结构下模型的稳定性, 通过分别去除整个模型中的 AAM, FPN, MAF 模块, 改变模型结构, 避免了不同模块之间相互作用的结果对测试的影响. 实验结果表明, AAM 模块加入后, 模型在 bbox 下的 mAP 上升了 1.83%, 在 BEV 下的 mAP 上升了 0.76%, 同时 AEV 下降了 22.3%; FPN 模块加入后, 模型在 bbox 下的 mAP 上升了 2.38%, 在 BEV 下的 mAP 上升了 3.15%, 同时 AEV 下降了 17.8%. 证明了 AAM 模块以及 FPN 模块能够有效地提升模型的感知精度以及感知过程的稳定程度. 模型性能提升的原因在于, 随着 AAM 模块的引入, 信源编码模块在数据压缩过程能够有效地朝着冗余信息减少的方向进行, 保证了数据压缩过程数据信息的准确全面. 同时, FPN 模块通过对不同层次数据信息的拼接, 避免了模型在数据压缩过程中的失真, 保证了模型的感知精度以及感知过程的稳定性. 加入 MAF 模块后, 模型在 bbox 下的 mAP 上升了 2.33%, 在 BEV 下的 mAP 上升了 3.05%, 同时 AEV 下降了 6.54%, 证明了 MAF 模块能够很好地提升模型的检测精度以及模型感知过程的稳定性. 模型性能提升的原因在于, MAF 模块对主模态的语义信息进行了检错与矫正, 通过冗余信息与互补信息的加入, 保证了主模态语义信息在模型感知过程中的准确与稳定.

5.3 感知模型的性能比较

本文提出的方法在 KITTI 三维目标检测数据集上进行了验证, 该数据集包含 7481 个训练样本和 7518 个测试样本 [69], 有 3 个不同的层次: 简单、中等和困难, 它们是根据对象的大小、可见性 (遮挡) 和截断来确定的 [70]. 进一步将训练集分割为训练与验证集, 避免来自相同序列的样本包含在两个集中. 本文的模型便是基于 MMDetection3D 框架, 通过构建信源编码模块以及信道编码模块等完成搭建任务的, 并将模型在 KITTI 数据集上进行训练, 训练完成后遵循标准的 KITTI 评估方案 (IoU = 0.7) 来对模型进行性能测试, 探索模型检测目标的能力. 同时, 用测试集中的数据对模型感知过程的稳定性 AEV 进行评估, 以数据集中各组数据的 AEV 平均值作为模型的整体评价指标对模型进行评估.

表 3 不同模型性能比较

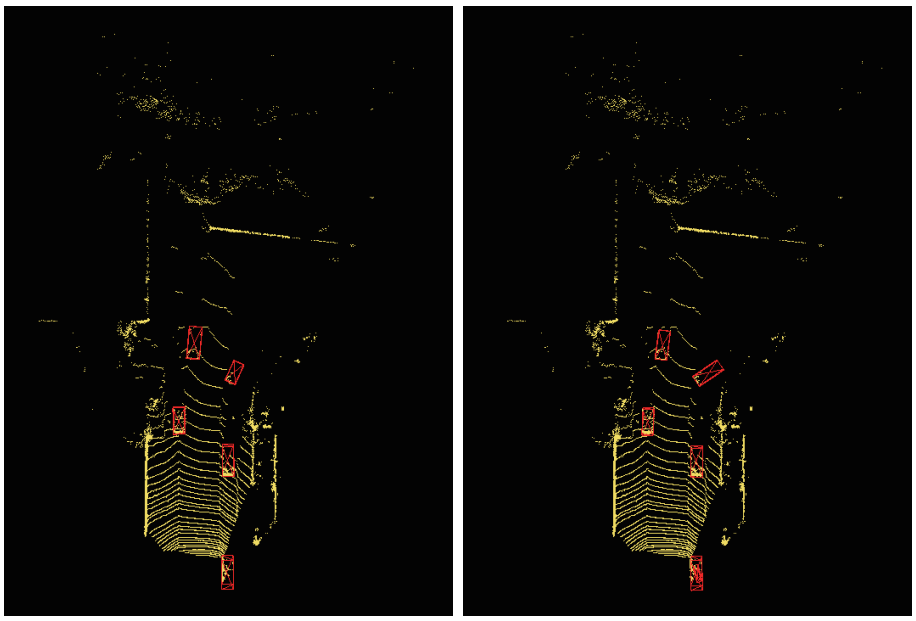
Table 3 Comparison of performance of different models

Model	mAP		AEV		
	bbox	BEV	AEV-channel	AEV-lidar	AEV-joint
Smoke	90.8558	17.6157	–	0.34768	0.3476
Second	95.2876	90.0579	–	5.515044	5.5150
Mvxnet	96.2173	92.293	2.203327	1.184249	3.3875
EPNet	97.7698	94.364	1.061605	0.00275	1.0643
Sourse coding	93.0922	89.1565	0.63054	0.00333	0.6338
Channel coding	93.4052	91.6867	0.75995	0.09278	0.8527
Ours	95.0187	90.739	0.587259	0.003189	0.5904

表 3 展示了本文提出的联合编码的方法与现有的应用与 KITTI 的评估方法在 bbox 以及鸟瞰图 (BEV) 上的全类平均精度 (mAP) 得分, 以及 AEV 得分比较. 与其他的传统感知算法相比, 本文提出的方法在感知精度上有着相对较好的评分结果, 在 bbox 以及 BEV 视角下感知结果的 mAP 值均在 90 以上, 具备良好的感知能力. 原因在于, 雷达信息与图像信息的融合, 使得不同模态的语义信息之间能够进行互补以及矫正, 从而保证了感知结果的准确性. 然而, 本文模型在 BEV 视角下感知结果的 mAP 指标低于一些对比算法, 其原因在于本文所提出的基于信源信道联合编码的网络结构采用了通信模型的编码方式来处理任务. 这种方法提高了多模态模型特征提取的效果, 同时也增强了特征提取与特征融合过程的可解释性, 使得模型可以更好地理解和解释其感知任务中的特征. 在模型的结构设计中, 信源编码模块引入了 AAM 等模块. 这些模块的引入有助于确保模型能够有效地提取和编码特定的信号或特征. 然而, 此类模块也使得模型的结构复杂程度有所增加, 导致模型可能会在某些情况下过度拟合部分数据, 从而导致精度略有衰减. 尽管 mAP 指标略低于一些对比算法, 但仍然保持在 90 以上. 这意味着该模型仍然具有很高的性能, 能够满足感知任务的需求. 此外, 本文提出的模型基于信息论中通信模型的信源信道联合编码的设计思路进行设计, 模型具备较强的可解释性, 而这类经典的基于规则的模型具备高度可解释的同时准确性相对较差, 本文通过模型结构优化以及参数优化等方法, 使得模型在保证可解释性的同时, 也达到了优异的检测精度. 同时, 我们的模型与其他的传统感知模型相比, 在 AEV 评价指标中表现出来极佳的性能, AEV 指标下降到 0.5904. 由此可见, 我们的模型结构在感知过程中的稳定性更强, 其中的信源编码模块有效地去除了输入数据中的冗余信息, 同时信道编码模块也高效地完成了对主模态语义信息的检错与矫正, 保证模型能够稳定地提取和融合不同模态的语义特征, 从而获得更好的感知稳定性. 模型在保持稳定的同时仍可以完成感知任务的需求, 感知结果如图 9 所示. 本文提出的模型由于轴向注意力的加入, 使得模型对轴向特征更为关注, 从而可以更加稳定地提取特征. 并且, 信道编码模块中通过 MAF 模块的作用使得主模态的特征得到了更好的补充和修正, 从而获得更利于网络提取的融合特征图, 使得网络在信源编码的特征提取部分以及信道编码的融合后的特征提取部分均能够更为稳定地进行提取, 从而获得更高的稳定性. 因此, 网络中相似结构的熵变更为平稳, 从而有着较低的 AEV 得分. 证明了我们的模型在保持较好的感知能力的同时, 有着较高的稳定性.

6 智能驾驶多模态特征融合未来发展趋势和面临的挑战

近年来, 用于自动驾驶感知任务的多模态融合方法取得了快速进展, 不同种模态逐渐应用到检测网络中, 能够提取更高级的特征, 从而更好地完成感知任务. 然而, 在自动驾驶感知任务中仍然存在一



(a) (b)

图 9 (网络版彩图) 感知结果

Figure 9 (Color online) Perceptual outcome. (a) Ground truth; (b) predicted result

定的问题与挑战.

由于相机与 LiDAR 的内在与外在参数存在着很大的差异, 因此两种数据无法直接应用在多模态融合的任务中, 需要在新的坐标系下重新组织. 一些传统的解决方法是利用外部的校准矩阵将 LiDAR 点直接投影到像素上, 从而完成数据之间的融合. 但是这种方法通常无法准确地进行逐像素对齐, 无法获得较好的融合结果. 此外, 除了这种坐标变换的方法, 还可以通过一些周围信息作为补充, 通过图卷积的方式, 来提高对其的效果以及准确性. 同时, 在输入和特征空间的转换过程中, 还存在一些其他的信息丢失. 通常, 降维操作的投影不可避免地会导致大量信息丢失, 例如将 3D LiDAR 点云映射到 BEV 图像中. 因此, 通过将两种模态数据映射为另一种专为融合而设计的高维表示, 未来的工作可以有效地利用原始数据, 同时减少信息丢失.

也要增强对来自多个维度和来源的信息的有效利用. 大多数研究都专注于前视图中的单帧多模态数据. 结果, 其他有意义的信息没有得到充分利用, 例如语义、空间和场景上下文信息. 在自动驾驶场景中, 许多具有显式语义信息的下游任务可能会极大地提高目标检测任务的性能. 例如, 车道检测可以直观地为检测车道之间的车辆提供额外帮助, 语义分割结果可以提高目标检测性能. 但是语义分割如今还面临着实时性和精确性之间有矛盾的问题. 通常情况下, 精确性越高, 则实时性越低, 所以算法模型很难同时满足高精确性和高实时性的要求. 传统的语义分割还面临着受到环境干扰、边缘分割不够准确等问题, 因此具有一定的局限性. 未来的研究可以通过检测车道、红绿灯和标志等各种下游任务, 共同构建完整的城市景观场景语义理解框架, 以辅助感知任务的执行.

7 总结

本文提出了一种基于信源信道联合编码的多模态融合网络, 用以提高模型在复杂的交通环境下感知过程的准确与稳定, 从而提升了模型的感知安全性. 根据信息论中的编码理论设计信源编码以及信

道编码结构, 按照通信模型中自编码的思想对信源编码模块进行设计, 运用引入轴向注意力机制, 使信源数据压缩过程更加稳定, 并向着冗余减少的方向进行, 解释模型的压缩方向, 增强了模型的可解释性. 在信道编码模块中, 将包含主模态与辅模态的融合特征作为噪声, 加入到主模态中来抵消网络结构以及外界的噪声干扰, 解释了多模态融合的底层机理, 增强了模型的可解释性. 最后, 提出了一种基于信息熵的评价指标 AEV 对模型与外界的感知交互过程进行实时评估. AEV 根据信息熵变程度衡量模型感知交互的稳定性, 从量化角度对模型感知过程稳定性进行评估. 这种新的评价指标增强了模型的可解释性, 达到了提高模型的感知安全的效果. 实验结果表明, 本文提出的感知模型 AEV 指标在 KITTI 数据集上降低到 0.5904, 相较于其他感知模型的信息熵变化程度明显下降, 增强了模型的感知安全性.

参考文献

1 Freudendal-Pedersen M, Kesselring S, Servou E. What is smart for the future city? Mobilities and automation. Sustainability, 2019, 11: 221

2 Othman K. Public acceptance and perception of autonomous vehicles: a comprehensive review. AI Ethics, 2021, 1: 355–387

3 Guan T, Han Y, Kang N, et al. An overview of vehicular cybersecurity for intelligent connected vehicles. Sustainability, 2022, 14: 5211

4 Chen B, Pei X F, Chen Z F. Research on target detection based on distributed track fusion for intelligent vehicles. Sensors, 2020, 20: 56

5 Liu T, Zhao Y, Wei Y C, et al. Concealed object detection for activate millimeter wave image. IEEE Trans Ind Electron, 2019, 66: 9909–9917

6 Meiring G A M, Myburgh H C. A review of intelligent driving style analysis systems and related artificial intelligence algorithms. Sensors, 2015, 15: 30653–30682

7 Chen B H, Huang S C, Kuo S Y. Error-optimized sparse representation for single image rain removal. IEEE Trans Ind Electron, 2017, 64: 6573–6581

8 Suk H I, Lee S W, Shen D G. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. Neuroimage, 2014, 101: 569–582

9 Du X X, Ang M H, Rus D. Car detection for autonomous vehicle: LIDAR and vision fusion approach through deep learning framework. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017

10 Pang S, Morris D, Radha H. CLOCs: camera-LiDAR object candidates fusion for 3D object detection. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020. 10386–10393

11 Zhao X M, Sun P P, Xu Z G, et al. Fusion of 3D LIDAR and camera data for object detection in autonomous vehicle applications. IEEE Sens J, 2020, 20: 4901–4913

12 Xu J L, Luo C B, Chen X Y, et al. Remote sensing change detection based on multidirectional adaptive feature fusion and perceptual similarity. Remote Sens, 2021, 13: 3053

13 Salazar-Gomez G A, Saavedra-Ruiz M A, Romero-Cano V A. High-level camera-LiDAR fusion for 3D object detection with machine learning. 2021. ArXiv:2105.11060

14 Zhao K, Ma L F, Meng Y, et al. 3D vehicle detection using multi-level fusion from point clouds and images. IEEE Trans Intell Transp Syst, 2022, 23: 15146–15154

15 Zou Z H, Zhang X Y, Liu H P, et al. A novel multimodal fusion network based on a joint coding model for lane line segmentation. Inf Fusion, 2022, 80: 167–178

16 Muhammad K, Ullah A, Lloret J, et al. Deep learning for safe autonomous driving: current challenges and future directions. IEEE Trans Intell Transp Syst, 2021, 22: 4316–4336

17 Zhang X Y, Shao W B, Zhou M, et al. A scene comprehensive safety evaluation method based on binocular camera. Robotics Autonomous Syst, 2020, 128: 103503

18 Rodionova A, Alvarez I, Elli M S, et al. How safe is safe enough? Automatic safety constraints boundary estimation

for decision-making in automated vehicles. In: Proceedings of IEEE Intelligent Vehicles Symposium (IV), 2020
19 Wu Z Y, Chen G D, Yao J J. A driving safety evaluation algorithm based on topsis model of game theory combination. In: Proceedings of the 7th Asia International Symposium on Mechatronics: Volume II, 2020
20 Luo Q, Hu S-G, Gong H-W, et al. Model building and research of urban road traffic safety evaluation system. J Guangxi Univ (Nat Sci Ed), 2017, 42: 587–592 [罗强, 胡三根, 龚华炜, 等. 城市道路交通安全评价体系研究与模型构建. 广西大学学报 (自然科学版), 2017, 42: 587–592]
21 Cai X Y, Lei C L, Peng B, et al. Road traffc safety risk estimation based on driving behavior and information entropy. China J Highw Transp, 2020, 33: 190–201
22 Sun C, Zheng S F, Ma Y L, et al. An active safety control method of collision avoidance for intelligent connected vehicle based on driving risk perception. J Intell Manuf, 2021, 32: 1249–1269
23 Javed A R, Usman M, Rehman S U, et al. Anomaly detection in automated vehicles using multistage attention-based convolutional neural network. IEEE Trans Intell Transp Syst, 2021, 22: 4291–4300
24 Li X L. Multi-modal cognitive computing. Sci Sin Inform, 2023, 53: 1–32 [李学龙. 多模态认知计算. 中国科学: 信息科学, 2023, 53: 1–32]
25 Liang M, Yang B, Chen Y, et al. Multi-task multi-sensor fusion for 3D object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019
26 Gao H B, Cheng B, Wang J Q, et al. Object classification using CNN-based fusion of vision and LIDAR in autonomous vehicle environment. IEEE Trans Ind Inf, 2018, 14: 4224–4231
27 Wang Z N, Zhan W, Tomizuka M. Fusing bird's eye view LIDAR point cloud and front view camera image for 3D object detection. In: Proceedings of IEEE Intelligent Vehicles Symposium (IV), 2018
28 Prakash A, Chitta K, Geiger A. Multi-modal fusion transformer for end-to-end autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021
29 Piergiovanni A, Casser V, Ryoo M S, et al. 4D-Net for learned multi-modal alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 15435–15445
30 Deng X, Dragotti P L. Deep convolutional neural network for multi-modal image restoration and fusion. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 3333–3348
31 Sun Y J, Fu Z L, Sun C X, et al. Deep multimodal fusion network for semantic segmentation using remote sensing image and LiDAR data. IEEE Trans Geosci Remote Sens, 2022, 60: 1–18
32 Xiong P W, Tong X B, Song A G, et al. Robotic cross-modal generative adversarial network based on variational Bayesian Gaussian mixture noise model. Sci Sin Inform, 2021, 51: 104–121 [熊鹏文, 童小宝, 宋爱国, 等. 基于变分贝叶斯高斯混合噪声模型的机器人跨模态生成对抗网络. 中国科学: 信息科学, 2021, 51: 104–121]
33 Ma Y F, Wang Z Y, Yang H, et al. Artificial intelligence applications in the development of autonomous vehicles: a survey. IEEE/CAA J Autom Sin, 2020, 7: 315–329
34 Zamanipour M. A novelty in Blahut-Arimoto type algorithms: optimal control over noisy communication channels. IEEE Trans Veh Technol, 2020, 69: 6348–6358
35 Liu T, Chen X C. Attention-based neural joint source-channel coding of text for point to point and broadcast channel. Artif Intell Rev, 2022, 55: 2379–2407
36 Dommel J, Utkovski Z, Stańczak S, et al. Joint source-channel coding and Bayesian message passing detection for grant-free radio access in IOT. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020
37 Li W J, Yang W, Liu Y X, et al. Research and exploration on the interpretability of deep learning model in radar image. Sci Sin Inform, 2022, 52: 1114–1134 [李玮杰, 杨威, 刘永祥, 等. 雷达图像深度学习模型的可解释性研究与探索. 中国科学: 信息科学, 2022, 52: 1114–1134]
38 Abdallah A B, Zribi A, Dziri A, et al. Adaptive joint source-channel coding using multilevel codes for unequal error protection and Hierarchical Modulation for SPIHT image transmission. In: Proceedings of IEEE 19th Mediterranean Microwave Symposium (MMS), 2019
39 Jankowski M, Gündüz D, Mikolajczyk K. Deep joint source-channel coding for wireless image retrieval. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020
40 Wang M Y, Zhang Z C, Li J H, et al. Deep joint source-channel coding for multi-task network. IEEE Signal Process Lett, 2021, 28: 1973–1977

41 Wang H J, Gao H B, Yuan S H, et al. Interpretable decision-making for autonomous vehicles at highway on-ramps with latent space reinforcement learning. IEEE Trans Veh Technol, 2021, 70: 8707–8719

42 Seo J, Kim S, Kang J. Neural joint source-channel coding via Bernoulli latent straight-through estimator. J Commun Netw, 2022, 24: 679–685

43 Hong S H, Xu Z P, Liu S Y, et al. Optimization design for joint source-channel coding and decoding system based on double protograph low-density parity-check codes. J Xiamen Univ Nat Sci, 2021, 60: 586–597

44 Ren Y B, Zhang Y, Liu Y W, et al. DNA-based concatenated encoding system for high-reliability and high-density data storage. Small Methods, 2022, 6: e2101335

45 Jose S T, Kulkarni A A. Shannon meets von Neumann: a minimax theorem for channel coding in the presence of a jammer. IEEE Trans Inform Theor, 2020, 66: 2842–2859

46 van Wyk M A, Ping L, Chen G R. Multivaluedness in networks: Shannon's noisy-channel coding theorem. IEEE Trans Circuits Syst II, 2021, 68: 3234–3235

47 Sun P, Boukerche A. Security enhancing method in vehicular networks by exploiting the accurate traffic flow prediction. In: Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), 2021

48 Yeung M, Sala E, Schönlieb C B, et al. Unified focal loss: generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Computized Med Imag Graph, 2022, 95: 102026

49 Li D Y. Creating a warm agricultural machinery with AI. J Intell Syst, 2022, 17: 859–859

50 Gray R M. Entropy and Information Theory. New York: Springer, 2011

51 Xin G T, Fan P Y. EXK-SC: a semantic communication model based on information framework expansion and knowledge collision. Entropy, 2022, 24: 1842

52 Yang H B, Zhang S Y, Yang Z Y, et al. Eloss in the way: a sensitive input quality metrics for intelligent driving. 2023. ArXiv:2302.00986

53 Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of IEEE International Conference on Computer Vision, 2017

54 Saibene A, Assale M, Giltri M. Expert systems: definitions, advantages and issues in medical field applications. Expert Syst Appl, 2021, 177: 114900

55 Wang H, Zhang W, Jing Y Z, et al. Controversial variable node selection-based adaptive belief propagation decoding algorithm using bit flipping check for JSCC systems. Entropy, 2022, 24: 427

56 Babu S A, Raj R J S, VM A X, et al. DCT based enhanced tchebichef moment using Huffman encoding algorithm (ETMH). In: Proceedings of the 3rd International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021

57 LeClair A, Haque S, Wu L F, et al. Improved code summarization via a graph neural network. In: Proceedings of the 28th International Conference on Program Comprehension, 2020

58 Saidutta Y M, Abdi A, Fekri F. Joint source-channel coding over additive noise analog channels using mixture of Variational Autoencoders. IEEE J Sel Areas Commun, 2021, 39: 2000–2013

59 Yan Y, Mao Y X, Li B. SECOND: sparsely embedded convolutional detection. Sensors, 2018, 18: 3337

60 Yang X Q, Duan L L. MPTC-FPN: a multilayer progressive FPN with transformer-CNN based encoder for salient object detection. IEEE Access, 2022, 10: 98816–98827

61 Yu Q Y, Lin H C, Chen H H. Intelligent radio for next generation wireless communications: an overview. IEEE Wireless Commun, 2019, 26: 94–101

62 Bai L, Li Y G, Cen M, et al. 3D instance segmentation and object detection framework based on the fusion of Lidar remote sensing and optical image sensing. Remote Sens, 2021, 13: 3288

63 El-Bakary E M, El-Shafai W, El-Rabaie S, et al. Efficient secure optical DWT-based watermarked 3D video transmission over MC-CDMA wireless channel. J Opt, 2023, 52: 2068–2089

64 Zheng W Q, Xie H, Chen Y F, et al. PIFNet: 3D object detection using joint image and point cloud features for autonomous driving. Appl Sci, 2022, 12: 3686

65 Yang J, Li B. Semantic segmentation of 3D point cloud based on self-attention feature fusion group convolutional neural network. Optics Precision Eng, 2022, 30: 840–853

66 Guo Y L, Wang H Y, Hu Q Y, et al. Deep learning for 3D point clouds: a survey. IEEE Trans Pattern Anal Mach Intell, 2021, 43: 4338–4364

67 Sindagi V A, Zhou Y, Tuzel O. MVX-Net: multimodal voxelnet for 3D object detection. In: Proceedings of International Conference on Robotics and Automation (ICRA), 2019
68 Jiang Y, Kim H, Asnani H, et al. Turbo autoencoder: deep learning based channel codes for point-to-point communication channels. In: Proceedings of Advances in Neural Information Processing Systems, 2019
69 Al-refai G, Al-refai M. Road object detection using Yolov3 and Kitti dataset. Int J Adv Comput Sc, 2020, 11: 48–53
70 Trivedi M M. Attention monitoring and Hazard assessment with bio-sensing and vision: empirical analysis utilizing CNNs on the KITTI dataset. In: Proceedings of IEEE Intelligent Vehicles Symposium (IV), 2019

Information-theoretic-based interpretable multimodal perception for intelligent vehicles

Xinyu ZHANG[1,2,3*], Jilong GUO[1,3], Jun LI[1,2,3], Deyi LI[4], Shiyan ZHANG[1,3], Sitian SHEN[1,3], Fan WU[1,3] & Huaping LIU[4]

1. *National Key Laboratory of Intelligent Green Vehicles and Transportation, Tsinghua University, Beijing* 100084*, China*;
2. *College of Transportation, Beijing University of Aeronautics and Astronautics, Beijing* 100191*, China*;
3. *School of Vehicle and Mobility, Tsinghua University, Beijing* 100084*, China*;
4. *Department of Computer Science and Technology, Tsinghua University, Beijing* 100084*, China*
* Corresponding author. E-mail: xyzhang@tsinghua.edu.cn

Abstract Intelligent driving has become one of the most compelling topics of interest. Nevertheless, current intelligent driving technologies still face challenges, such as missed detection due to large vehicle occlusion and false detection caused by sensor accuracy degradation in sudden light changes. Multimodal perception technology for intelligent vehicles has emerged to ensure the safety of vehicle perception in complex scenarios. However, the existing multimodal fusion methods are still limited to the improvement of detection accuracy, lack of interpretability of the perception process and lack of evaluation indexes for the model perception process. In this paper, from the information theory perspective, we design the perception model according to the communication model. We propose a multimodal fusion perception model based on joint source-channel coding theory to explain the perception process of the model theoretically. At the same time, we propose a new evaluation index, average entropy variation (AEV), which is used to reflect the stability of the model during its perceptual interaction with the outside world in real time. Further, the perceptual process is quantified and analyzed to increase the interpretability of the model. Finally, we compare the evaluation results with other advanced perceptual models in the KITTI dataset, and our model decreases the average entropy variation to 0.5904, which better ensures the perceptual safety of the detection task.

Keywords interpretability, theory of information, joint source-channel coding, multimodal fusion, intelligent driving