同济大学
TONGJI UNIVERSITY

硕士学位论文

(学术学位)

基于信息融合的道路场景解析研究

姓　　名: 李佳航

学　　号: 2230745

学　　院: 电子与信息工程学院

学 科 门 类: 工学

一 级 学 科: 控制科学与工程

二 级 学 科: 模式识别与智能系统

研 究 方 向: 模式识别与智能系统

指 导 教 师: 范睿教授

行 业 导 师:

二〇二五年五月

同濟大學
TONGJI UNIVERSITY

A thesis submitted to
Tongji University in partial fulfillment of the requirements for
the degree of Master of Engineering

Research on Road Scene Parsing Based on Information Fusion

Candidate :	Jiahang Li
Student Number :	2230745
School/Department :	College of Electronic and Information Engineering
Categories :	Engineering
First-level Discipline :	Control Science and Engineering
Second-level Discipline :	Pattern Recognition and Intelligent Systems
Research Fields :	Pattern Recognition and Intelligent Systems
Supervisor :	Prof. Rui Fan

May, 2025

摘要

智能感知技术赋予了自主智能系统在复杂、动态且开放的环境中实现高精度自主导航的能力。而道路场景解析作为智能感知技术的核心任务之一，是实现自主导航的基础性环节。主流道路场景解析网络受限于单模态信息或是原始的异构特征融合策略，难以充分发挥场景信息的潜力，导致网络在低能见度以及复杂背景等挑战性场景下的性能显著下降，实际应用能力不足。为了解决上述问题，本文深入分析了当前道路场景解析方法在异构特征编码、融合以及语义预测能力方面的不足，并针对这些问题提出了一系列改进方案。具体而言，本文的主要工作和贡献概括如下：

1. 提出了一种基于 RGB-法向量图像对输入的信息融合网络，并在可行驶区域与道路破损检测任务上验证了其有效性。该方法能够对可行驶区域和道路破损区域的平面特性进行显式编码，从而获得更强大的性能。此外，相较于现有的特征通道级串联与元素级相加方法，本文引入了基于注意力机制的异构特征融合策略，使得网络能够更加关注与当前任务相关的关键信息，抑制无关信息的干扰，从而有效提高了道路场景解析的性能；
2. 提出了一种适用于 RGB 图像与任意模态视觉信息融合的道路场景解析网络，并通过实验评估了该网络对深度图像、热图像等信息的适用性。该网络有效利用了视觉基础模型，同时结合了 Transformer 的长距离依赖建模能力和卷积神经网络对局部语义的能力，实现了更具模态适配性的异构特征融合，使得网络在复杂环境下表现得更加稳定和准确；
3. 提出了一种以 RGB-深度图像对作为输入的开放词汇场景解析网络。相比于现有仅基于 RGB 图像的开放词汇网络，该网络通过梯度级的任务解耦的思想，能够从源于深度图像空间先验中更好的学习物体轮廓、边缘等特征，从而在复杂的开放场景中实现更精准、鲁棒的物体掩膜提取，同时基于视觉-语言模型实现视觉特征与语言文本特征的对齐，从而赋予网络零样本推理的能力，能够准确对场景中任意新颖语义类别的对象进行解析。

关键词：信息融合，Transformer，卷积神经网络，道路场景解析，基础模型

ABSTRACT

Intelligent perception technology empowers autonomous intelligent systems with the capability to achieve high-precision navigation in complex, dynamic, and open environments. As one of the core tasks of intelligent perception technology, road scene parsing serves as a fundamental component for autonomous navigation.

Mainstream road scene parsing networks are limited by either single-modal information or primitive heterogeneous feature fusion strategies, making it difficult to fully exploit the potential of scene information. This leads to significant performance degradation in challenging scenarios such as low light and cluttered backgrounds, resulting in insufficient practical applicability. To address these issues, this paper conducts an in-depth analysis of the shortcomings of current road scene parsing methods in terms of heterogeneous feature encoding, fusion, and semantic prediction capabilities, and proposes a series of improvements accordingly. Specifically,

The main contributions of this work are summarized as follows:

1. An information fusion network based on RGB-Normal image pairs is proposed, and its effectiveness is validated on freespace detection and road defect detection tasks. This method enables explicit encoding of planar characteristics in freespace and road defect regions, thereby achieving superior performance. Furthermore, compared to existing feature channel-wise concatenation and element-wise addition strategies, an attention-based heterogeneous feature fusion strategy is introduced, allowing the network to focus more on task-relevant key information while suppressing irrelevant interference, effectively improving road scene parsing performance;
2. A road scene parsing network suitable for RGB images fused with arbitrary modality visual information is proposed, and its applicability to depth images and thermal images is evaluated separately through experiments. This network effectively utilizes visual foundation models while combining Transformer's long-range dependency modeling capability with CNN's local semantic aggregation ability, achieving more modality-adaptive heterogeneous feature fusion. This results in more stable and accurate network performance in complex environments;
3. An open-vocabulary scene parsing network using RGB-depth image pairs as input is proposed. Compared to existing RGB-only open-vocabulary networks, this

network employs gradient-level task decoupling to better learn object contours and edge features from spatial priors provided by depth images, enabling more precise and robust mask prediction in complex scenes. At the same time, based on vision-language model, it aligns vision features with text features, thus endowing the network with zero-shot inference capabilities to accurately parse objects of any novel semantic categories in the scene.

Key Words: information fusion, Transformer, convolutional neural network, road scene parsing, foundation model

目录

插图索引

表格索引

第 1 章　引言

1.1　研究背景与意义

在国家《" 十四五" 机器人产业发展规划》的推动下，环境感知技术作为具身智能（Embodied Artificial Intelligence）的核心使能环节，正深刻重构自主移动系统的认知范式。从《新一代人工智能发展规划》到《国家智能制造标准体系建设指南》，政策体系持续强化跨媒体、多模态（信息融合）感知技术的战略价值，特别是在《机器人产业创新发展行动方案》中将复杂环境理解列为共性技术突破方向，明确要求提升机器人在动态开放场景中感知的鲁棒性。这种技术演进趋势不仅推动了感知算法在工业检测、服务机器人等领域的应用，更使其成为行星探测、深海作业等特种机器人系统的关键技术。作为自主系统的“感官延伸”，非结构化环境下的信息融合感知能力已成为上述领域中的核心研究命题。用于行星探测任务的机器人（如图 1.1(a) 所示）通过多源传感器信息的融合实现了对未知地表的地图重建；在远海目标监测（如图 1.1(b) 所示）中，热图像（Thermal Image）与 RGB 图像的信息融合解析显著提升了恶劣海况下的目标追踪精度。面向未来民生需求，智能感知技术正加速向日常生活场景渗透：自动驾驶汽车通过激光雷达与相机传感器信息的前融合机制，实现了复杂城市场景下的实时路况解析（如图 1.1(c) 所示）；智能轮椅则利用 ToF（Time of Flight）相机等多源传感器，鲁棒地完成不同场景下的自主避障决策（如图 1.1(d) 所示）。这些跨领域的应用揭示了一个共性规律：场景要素的精准解析是实现环境认知的基础支撑，

(a) 火星探测任务中的智能机器人　(b) 远海搜救行动中的智能监测系统

(c) 自动驾驶汽车的环境感知系统　(d) 智能轮椅配备的自主避障系统

图 1.1　智能环境感知技术在不同领域中的应用

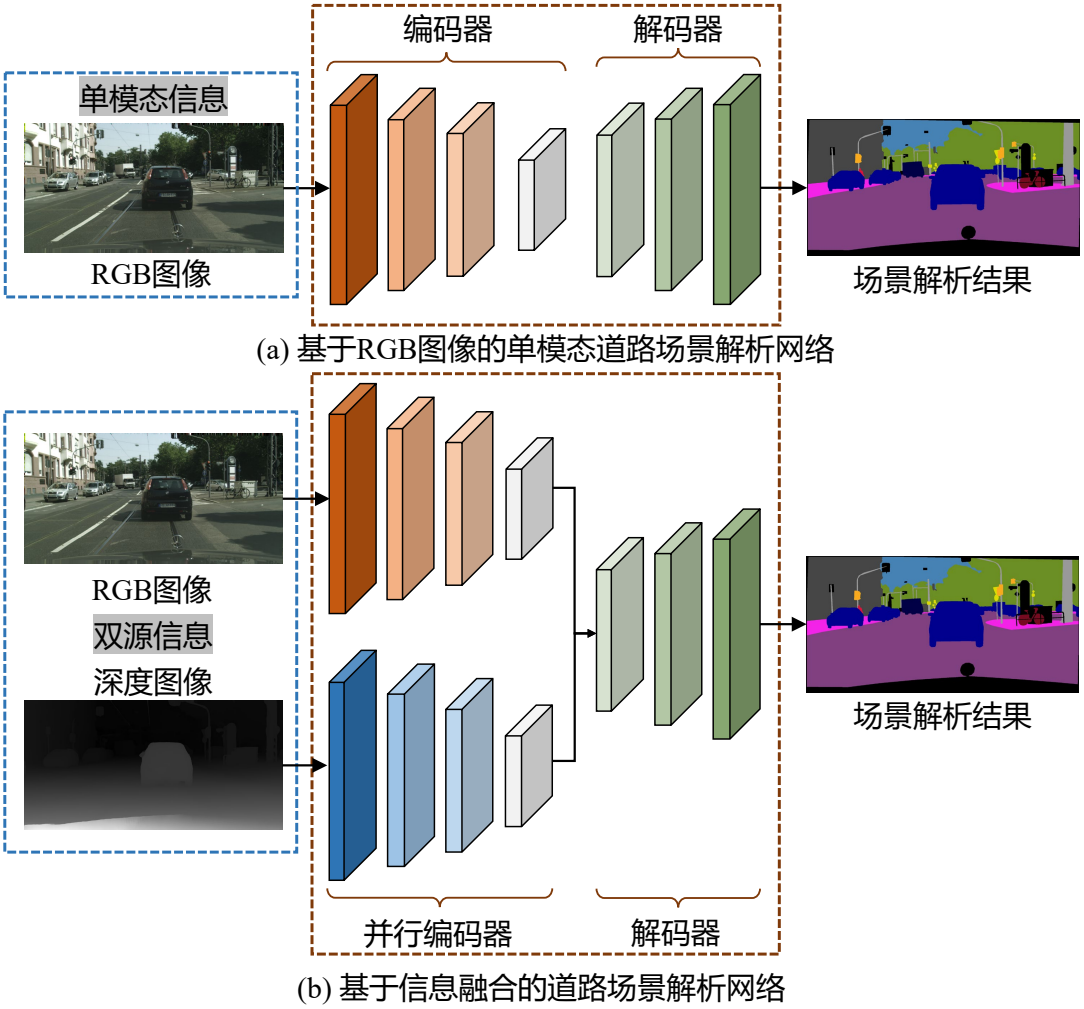


图 1.2　基于RGB 图像的单模态道路场景解析网络架构与基于信息融合的道路场景解析网络架构

而信息融合是突破感知任务瓶颈的一种有效途径。

聚焦自动驾驶与机器人领域，道路场景解析（Road Scene Parsing）任务作为自动驾驶系统的感知中枢，其核心目的是在 RGB 图像上实现对目标对象的像素级语义分类。当前主流方法通过端到端编码器-解码器架构[1] 的场景解析网络完成这一任务，这些网络大都采用单一模态的 RGB 图像作为网络输入（如图 1.2（a）所示）。然而，面对城市道路的动态开放性特征，即昼夜光照突变、雨雾干扰、临时障碍物等复杂场景，单模态网络逐渐暴露出重大缺陷：仅基于 RGB 图像的方法仅能提取场景中的色彩、纹理等特征，在低能见度与复杂背景等条件下会因图像失真而导致网络性能的显著衰减；另一方面，其它多源传感器如激光雷达虽能获取场景的空间几何特征却难以区分物体的色彩纹理特征；毫米波雷达具有稀疏的感知特性，对道路表面微小破损区域的解析能力存在明显不足，难以凭借此类单模态输入实现稳定的场景解析。因此，研究者提出了基于信息融合的方法（如图 1.2（b）所示），通过并行编码器-解码器的架构[2]，同时从 RGB 图像与其他信息：包括深度图像（Depth Image）、视差图像（Disparity Image）、热图

像、法向量图像（Surface Normal Image）等视觉信息中提取异构特征并进行特征融合，实现任务性能的提升。然而，现有异构特征融合方案多采用简单的特征元素相加或特征拼接策略，未能有效应对多源传感器在时空分辨率、特征表示维度等方面的本质差异，导致提取的异构特征难以互补与增强。此外，封闭词汇下的场景解析范式与开放道路场景的动态特性存在根本矛盾，传统基于固定类别集合的监督学习无法对新类别物体进行解析。以上三种问题在极端天气与复杂车流路况下等具有挑战性的道路场景中尤为突出，严重制约了智能感知系统的全天候运行能力。

上述问题的深层次困境在于现有基于信息融合的道路场景解析方法存在一定的局限与空白：在异构特征提取方面，现有网络对于特定数据源信息所编码场景表征的鲁棒性不足，不同信息（如热图像与深度图像）具有明显不同的内在特性，以及传感器噪声、环境干扰等因素都会导致信息对应的异构特征空间出现显著差异，这些因素导致了同一网络无法同时适用于不同的信息输入；在异构特征融合方面，现有方法缺乏对异构特征的自适应建模能力，如深度图像、热图像以及法向量图像等信息在异构特征尺度、语义粒度上的差异使得如特征通道串联或元素级相加等简单的异构特征融合策略易引发特征冲突，从而损害网络性能；在信息融合网络的实际应用层面，基于封闭词汇的预定义语义分类范式与具有开放词汇特性的道路场景特性具有本质冲突。这些系统性缺陷导致现有的道路场景感知系统在具有复杂背景、多变光照等场景中频繁出现误判，严重威胁自动驾驶系统的安全运行。

在上述背景下，本研究致力于构建创新的信息融合道路场景解析架构，重点突破如下三方面：针对异构特征融合的低效性问题，设计基于注意力机制的特征校准框架，实现异构特征自适应的进行择优融合；针对网络复杂场景下泛化能力不足的痛点，建立具有信息自身特性感知能力的非对称特征编码框架，提升网络在光照突变、传感器退化等极端条件下对异构特征潜力的挖掘；针对封闭词汇集的局限性，创新开放词汇解析范式，通过视觉-语言模型的引导扩展网络的语义理解能力。通过上述贡献，本研究旨在为机器人及自动驾驶汽车等智能设备提供全天候、全要素的环境理解能力，助力突破夜间道路、极端天气等挑战性场景中的任务瓶颈，推动现有自主导航系统从“单一感知”向“融合感知”的转型。

1.2　国内外研究现状

深度学习的飞速发展极大地推动了智能感知技术在各大任务领域中的广泛应用[3, 4]。这一趋势在自动驾驶汽车[5]、智能辅助设备[6] 以及服务机器人[7] 等领域体现得尤为突出。当前研究焦点正逐步从单一性能优化转向安全性与用户体

验的协同提升[8]。在此背景下，道路场景解析技术，即对交通环境中各类要素进行像素级语义划分的计算机视觉任务，发挥着关键的支撑作用[3]。

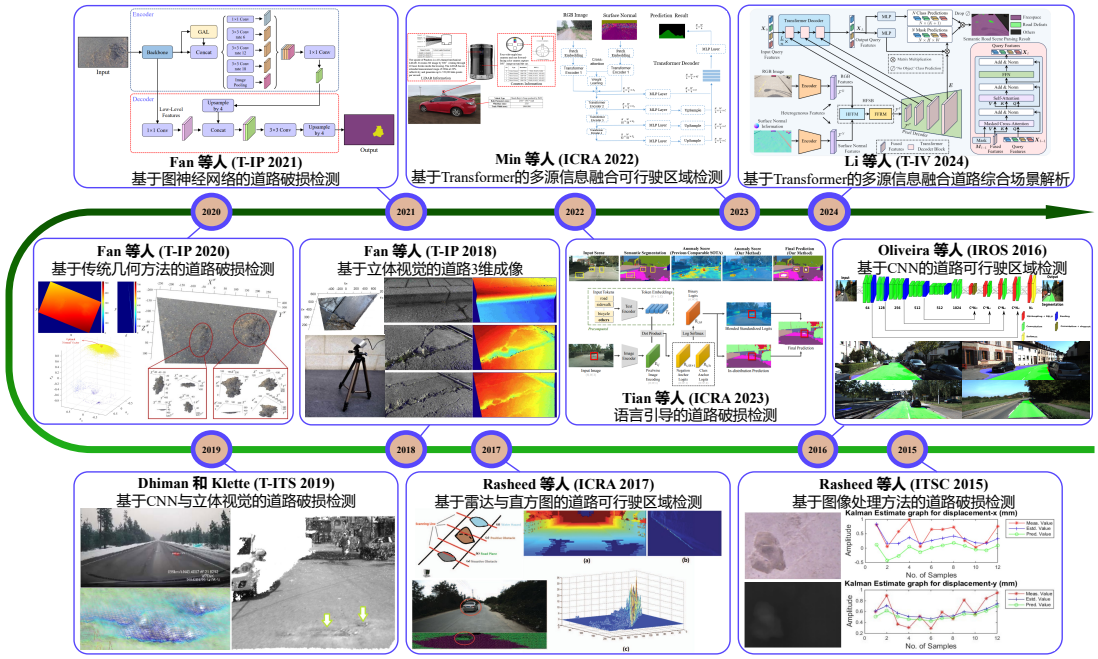


图 1.3　道路场景解析领域近十年代表性研究(2015-2024)[9]

图 1.3展示了过去十年内该领域的代表性研究成果。在深度学习方法兴起之前，基于几何建模的技术路线长期占据主导地位[10]。这类方法通常采用参数化几何模型（如平面、二次曲面等）表征感兴趣区域，通过优化能量函数实现精确提取。典型工作如[11] 将路面建模为二次曲面，利用最小二乘拟合从三维点云中恢复道路几何。后续研究[12] 提出的视差变换算法通过处理密集视差图生成类鸟瞰图投影，使得未损坏区域的视差分布呈现均匀特性，有效区分正负障碍物。此外，基于 B 样条建模的 v-视差分析方法[13, 14] 也为可行驶区域检测提供了新思路。然而，这类方法对道路表面连续性和规则性假设较强，在面对实际道路的复杂几何形变时易受异常值干扰[12]。

卷积神经网络 (Convolutional Neural Network，CNN) 的出现为道路场景解析任务带来了全新的解决方案。基于 CNN 的方法展现出了显著优于传统图像处理和几何建模方法的性能，大大提升了解析精度。例如，在研究[15] 中，研究者使用了编码器-解码器架构的 CNN 网络对投影到鸟瞰图视角的 RGB 图像进行像素级分类，以实现可行驶区域检测任务。然而，该方法在具有挑战性的光照和天气条件下往往表现不佳。为解决此类方法的局限性，后续研究探索了基于并行编码器架构的异特征融合网络，显著提高了道路场景解析的精确性与鲁棒性[16, 17]。研究[2] 从 RGB-深度 (RGB-D) 图像对中提取异构特征，并进行异构特征间的元素级相加以实现特征融合。异构特征中包含了侧重点不同的场景表征，融合带来了对各种场景更深入且全面的理解，相比之前的单模态网络取得了更优的性能。

类似地，SNE-RoadSeg 系列[5, 16, 18] 网络以 RGB-法向量 (RGB-SN) 图像对作为输入，通过逐元素相加或整体注意力机制实现了色彩纹理特征与平面特征的融合。通过将并行编码器网络和稠密连接的类 U-Net[19] 解码器进行组合，SNE-RoadSeg 系列在多个数据集上取得了最先进的性能，包括 KITTI Road[20]、vKITTI2[21] 和 Cityscapes[22]。然而，这些方法仍然具有其局限性：(1) 仅对异构特征进行简单且无差别的融合，(2) 存在过于冗余的网络参数。这两个问题都可能导致特征表示冲突和不准确的场景解析预测，从而在一定程度上制约网络的实际应用价值。

近年来，Transformer 架构[23] 的网络在视觉任务中逐渐展露其优势。相比于 CNN 网络，当有大规模的数据集可用于训练时[24]，Transformer 架构能够达到更高的性能上限，这源于 Transformer 中独特的注意力机制，能够相比 CNN 更有效地建模全局上下文依赖关系[25]。因此，利用注意力机制增强异构特征的融合是一个值得探究的方向。OFF-Net[17] 首次尝试将 Transformer 应用于信息融合道路场景解析任务。通过在大规模的越野道路数据集上进行训练，OFF-Net 在面向越野道路的可行驶区域检测任务中取得了相比之前算法的性能提升。然而，由于 OFF-Net 采用的是轻量级 CNN 解码器，因此其难以被看作一个完全基于 Transformer 架构的网络。此外，该网络在城市道路场景中，特别是在数据规模受限条件下的表现不尽如人意[26]。在后续章节中，我们进一步探索了 Transformer 架构在道路场景解析领域中的潜力，提出了基于 Transformer 的一系列架构，有效提升了道路场景解析性能的上限。

1.2.1　通用场景解析网络

如上所述，场景解析网络具有广泛的可应用场景，可用于显著物体检测，遥感图像分割等领域，许多研究将通用场景解析网络直接或增加专用策略后进行道路场景解析任务。因此，现有的道路场景解析方法可按照图1.4进行分类，我们

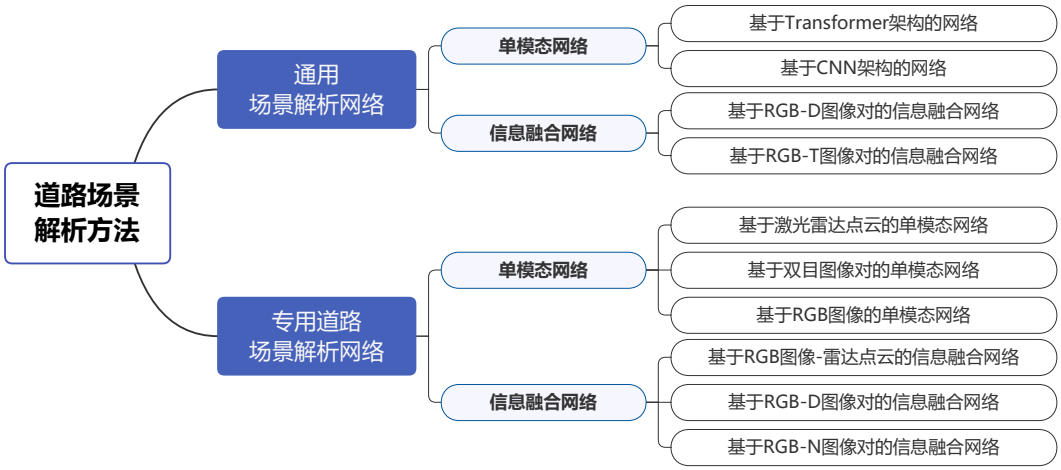


图 1.4　基于深度学习的道路场景解析方法分类

首先对其中的通用场景解析方法进行回顾。

1.2.2　单模态网络

全卷积网络 (Fully Convolutional Network，FCN)[27] 是第一个用于场景解析任务的 CNN 网络，该网络使用单模态 RGB 图像作为输入，FCN 的提出标志着场景解析任务进入深度学习时代。然而，该网络特征下采样过程中会发生严重的特征损失，严重制约了细粒度分割精度。Fast FCN[28] 通过结合空洞卷积与特征金字塔网络，在有效保持感受野的前提下提取图像中的细节特征，同时通过网络模块设计将推理效率提升至 FCN 的 300% 以上。Fast-SCNN[29] 创新性地设计了可学习的特征下采样模块，且能够在算力受限的边缘计算设备上进行实时推理，有效提升了场景解析网络在高分辨率图像上的实时性能。受传统控制领域中的比例-积分-微分 (Proportional-Integral-Derivative，PID) 控制器启发，PIDNet[30] 将 CNN 与 PID 的运行概念相结合，形成了一种新颖的架构，包含三个分支，分别用于提取细节特征、上下文特征以及边界特征，PIDNet 在推理速度和准确性之间达到了最佳平衡。

2017 至 2020 年间，场景解析领域主要探索了金字塔结构的多分辨率特征提取技术[31-33]。例如，DeepLabv3[34] 通过引入并行空洞空间金字塔池化 (Atrous Spatial Pyramid Pooling，ASPP) 模块，有效地捕获了多尺度的上下文特征。然而，DeepLabv3 中的步长操作可能导致目标边界处的精细细节丢失。为克服这一限制，DeepLabv3+[31] 在 DeepLabv3 的基础上引入了一个简洁而有效的解码器，显著改善了场景解析结果，特别是在不同类别边界处的表现。

与上述专注于从低分辨率恢复高分辨率特征图的工作不同[19, 27]，高分辨率网络 (High Resolution Network，HRNet)[35] 在整个特征提取和融合过程中保持高分辨率表示。这种设计通过并行多分辨率子网络的渐进式和重复性多尺度特征融合，实现了更准确的预测，而 PointRend[36] 则在图像分割网络中引入了一种新颖的基于点的渲染技术，通过在迭代细分算法确定的自适应选择位置上进行预测，最终能够实现精确而灵活的分割，适用于实例和场景解析任务。

此外，一些研究工作探索了基于 CNN 架构的注意力机制（不同与 Transformer 中的注意力机制）在场景解析领域的有效性。例如，双重注意力网络 (Dual Attention Network，DANet)[37] 引入了空间和通道注意力模块，有效增强了对上下文信息的理解能力，提高了网络在复杂场景下的性能。在相当一段时间内，许多研究都尝试在场景解析网络中加入注意力机制以求提升性能。然而，注意力机制虽然提高了网络聚焦图像内长距离依赖的能力，但其庞大的计算需求同样为模型部署带来了严重限制。即使相比于使用了大型卷积核的方法，如 ASPP[34]

等重型 CNN 网络，注意力机制模块的训练以及推理成本依然过高。为解决这一问题，一些研究对轻量级的注意力模块进行研究，在一定程度上减轻了网络的计算需求。非局部网络 Non-Local Net[38] 和非对称非局部神经网络 (Asymmetric Non-Local Neural Network，ANN)[39] 仅从特征图中采样少量关键点，大幅降低了计算复杂度。解耦非局部网络 (Disentangled Non-local Neural Network，DNL-Net)[40] 延续了这一路线的研究，针对 Non-Local Net 中存在的优化问题进行改良。DNLNet 提出 Non-Local Net 中的注意力机制可以被分解为成对项和一元项，并通过解耦这两个注意力项间的相互依赖性有效地改善了优化问题，实现了更高效的学习和应用。而全局上下文网络 GCNet（Global Context Network）[41] 针对注意力机制提供了一种简化的与查询无关的公式，在保持非局部网络精度的同时降低了网络的计算开销。交错稀疏自注意力网络 (Interlaced Sparse Self-Attention Network，ISANet)[42] 将密集亲和矩阵分解为两个稀疏矩阵的乘积，从而降低计算负担。与将所有像素视为重建基础的方法[43, 44] 不同，期望最大化注意力网络 (Expectation-Maximization Attention Network，EMANet)[45] 识别更紧凑的基集，显著降低了计算复杂度。这种方法有效简化了大型注意力图的创建，同时大幅减少了内存消耗。此外，上下文引导网络 (Context Guided Network，CGNet)[46] 引入了参数高效的上下文引导模块，该模块整合局部特征与周围上下文，并使用全局上下文进行精炼。借助此模块，CGNet 在 Cityscapes 数据集[22] 上以少于 50 万的网络参数实现了具有竞争力的性能。

2020 年以来，随着 Transformer 架构在计算机视觉领域的普及，基于该架构的场景解析网络，包括 SETR (Segmentation Transformer)[47] 和 SegFormer[24] 等展现出了强大的性能。相比于此前提出的 CNN 注意力机制，这些基于 Transformer 的方法能够更好的捕获长距离依赖关系，为场景解析领域提供了新的网络范式。作为首个基于 Transformer 的通用场景解析网络，Segmentation Transformer (SETR)[47] 借鉴 ViT 的成功经验，将输入图像分割为固定大小的 patch，并通过多层 Transformer 块进行特征编码。随后，SETR 采用 CNN 对编码特征进行逐步上采样，最终实现像素级分类。尽管 SETR 在多个基准数据集上取得了显著成果，但其计算复杂度较高，且在处理多尺度目标时存在局限性为克服这些限制，SegFormer[24] 提出了一种创新的多尺度 Transformer 编码器架构。该架构通过堆叠 Transformer 层并在层间插入卷积操作，有效提取不同尺度的上下文特征。与 SETR 相比，SegFormer 不仅在分割精度上实现了显著提升，特别是在处理尺度变化较大的目标时表现更为优异，同时还大幅降低了计算复杂度，使其更适合实际应用场景。

在层次化 Transformer 架构方面，Swin Transformer[48] 通过引入移位窗口注意力机制，实现了高效的局部-全局特征交互。受 DETR[49] 启发，Segmenter[50] 进一步开发了掩码 Transformer 解码器，在编码和解码阶段均能有效捕获全局上下文

特征。为同时兼顾全局上下文和局部细节，Twins[51] 提出了一种双分支架构：一个分支专注于捕获全局上下文特征，另一个分支则专门处理局部边界细节，从而实现了更精细的分割结果。DPT[52] 则采用了一种独特的特征聚合策略，以 ViT 为骨干网络，从不同阶段收集多分辨率令牌 (token)，并通过与卷积架构的解码器逐步融合，最终生成高分辨率预测。这种设计使得网络能够在每个处理阶段都保持全局感受野，从而产生更细粒度且全局一致性更强的预测结果。在特征表示方面，目标上下文表示 (Object-Contextual Representation, OCR)[53] 基于 Transformer 架构提出了一种新颖的像素表征方法。与传统基于空间位置的多尺度上下文表征策略不同，OCR 通过利用目标类别的语义特征来表征像素，有效区分了同类和异类像素的上下文关系。受此启发，K-Net[54] 引入了一组可学习的核，每个核负责生成特定类别或实例的掩码，从而统一解决了场景解析、实例分割和全景分割等多种图像分割任务。

与上述仅使用 Transformer 作为编码器的逐像素分类场景解析网络不同，MaskFormer[55] 提出了一种全新的场景解析范式，突破了传统逐像素分类方法的局限。该架构通过将查询特征解码为类别特定的掩膜来实现分割，利用多尺度 Transformer 解码器同时为每个类别生成精细的掩膜预测。实验表明，这种范式在多个基准数据集上均优于传统的逐像素分类范式网络。在此基础上，Mask2Former[56] 进一步扩展了 MaskFormer 的能力，通过引入掩码注意力机制，不仅在场景解析任务上取得了新的突破，还将性能优势扩展到了实例分割和全景分割等更广泛的任务中。这些创新不仅推动了场景解析技术的发展，也为计算机视觉领域的其他任务提供了新的思路和方法，本研究在后续章节中基于该掩膜分类范式，对道路场景解析方法进行了一系列的创新。

1.2.3　信息融合网络

通用场景解析领域中的信息融合网络利用从 RGB-D 图像对和 RGB-热（RGB-T）图像对等提取出的异构特征进行特征级融合，以提升场景解析任务的精确性与鲁棒性，相比于专用道路场景解析网络，这些网络的异构特征融合策略往往较为简单，更适用于通用场景中的任务。FuseNet[2] 作为此类方法的先驱，首次将深度图像引入通用场景解析领域。该网络为 RGB 图像和深度图像分别设计了独立的 CNN 编码器，并通过特征元素级相加的方式实现异构特征的融合。MFNet[57] 则利用 RGB 图像和热图像作为输入，构建了一种用于自动驾驶领域场景解析的轻量级网络，实现了网络推理速度与精确性的平衡。类似地，RTFNet[58] 利用 RGB-T 图像对作为输入，开发了一种鲁棒的 CNN 网络，该网络在解码器中加入了稠密连接以生成更清晰的边界，同时保留了细节特征。以上网络作为面向

通用场景解析任务的网络，尽管已在道路场景解析任务中展现出了一定的适用性，但一些针对道路场景设计的专用网络[26] 通常能够在道路场景解析任务中实现更优越的性能。

1.2.4　专用道路场景解析网络

本节将回顾专为道路场景解析任务设计的专用网络。早期的专用网络[59]主要依赖 RGB 图像。该领域的显著进展包括 HA-DeepLabv3+[60] 和 LFD-RoadSeg[61]。前者引入了一种基于立体视觉单应变换的新颖数据增强策略，这种策略能够从目标视角生成合成图像，对参考视角下的图像进行模拟。该方法的性能显著优于其基线网络 DeepLabv3+[62] 和其他最先进的基于立体视觉的方法。研究[61] 提出了 LFD-RoadSeg，这是一种双分支的可行驶区域检测网络。其中第一个分支使用 ResNet-18 骨干网络[63] 提取局部特征，而第二个分支通过同时对图像进行下采样和聚合特征来增强上下文建模。这种特征提取和聚合策略实现了与 ResNet-18 第三编码阶段相当的感受野，同时显著减少了计算时间。随后，该网络使用选择性融合模块计算局部特征和上下文特征之间的像素级注意力，以有效且高效地区分道路和非道路区域。然而，这些仅使用 RGB 图像作为输入的网络仍然对光照和天气条件等环境因素高度敏感[16]。随着激光雷达等 3D 传感器的日益普及，研究者们[64, 65] 在道路场景解析任务中引入这些信息，以提升网络的精确性和鲁棒性，其中最常用的信息包括深度/视差图[66, 67]、激光雷达点云[68, 69] 和法向量图[16]。先前研究[16, 18] 中进行的大量实验一致表明，法向量图和变换后的视差图为道路场景解析提供了最具区分性的空间几何特征，这可以归因于它们表示平面特性的能力。研究[64] 提出了一种仅依赖激光雷达数据的可行驶区域检测网络，其中非结构化点云被转换为俯视图图像。这些图像包含了场景的平均高程和密度等信息，通过这些信息，可将可行驶区域检测任务简化为单尺度问题。随后，研究[65] 提出了一种专为基于激光雷达的场景解析任务设计的 CNN 模型。该研究开发了一种计算效率高的硬件架构并部署在 FPGA（Field-Programmable Gate Array）设备上，每次激光雷达扫描的处理时间仅为 17.59 毫秒。

大部分具有自动驾驶功能的设备，如机器人与自动驾驶汽车普遍配备了相机与激光雷达传感器，因此使用 RGB 图像-雷达点云融合的方法是道路场景解析领域的一个主流研究方向。多个研究团队提出了一系列创新性网络架构[70-72] 来进行场景解析。例如，研究[70] 提出了一种基于双视图融合的 CNN，专门用于可行驶区域检测。该网络创新性地将两组激光雷达点云进行变换表示，以端到端的方式在激光雷达图像和相机透视图中提供像素级可行驶区域检测结果。网络结构中包含了一个专门的映射层，用于将特征从激光雷达图像视图转移到相机

透视视图，从而通过利用这两种表示之间的数据关联来增强网络性能。这种方法优化了对激光雷达数据的利用，能够在复杂的城市环境中产生更准确且鲁棒的可行驶区域检测结果。此外，该团队还提出了另一种可行驶区域检测方法[71]，该方法在条件随机场 (Conditional Random Field，CRF) 框架内整合激光雷达和相机数据，结合空间几何和色彩纹理特征以提高准确性。该网络是一种双分支架构，激光雷达分支中，在 2D 激光雷达距离图像域内应用快速高度差扫描策略，通过几何上采样实现相机图像域中精确的可行驶区域检测，这依赖于精确的激光雷达-相机标定。同时，相机分支使用 FCN 处理 RGB 图像。通过统一的 CRF 框架实现来自激光雷达和相机数据的详细和二进制可行驶区域检测输出的融合，有效地优化了信息的使用，用于稳健的可行驶区域检测。此后，该团队提出了 CLCFNet[72]，设计了一种创新的级联相机-激光雷达融合策略，以两种模式运行：单模态模式，仅使用激光雷达点云；以及多模态模式，结合激光雷达点云和 RGB 图像以适应不同的光照条件。网络架构由三个主要部分组成：激光雷达分割模块，用于从激光雷达点云数据中检测道路点；稀疏到密集模块，用于提高激光雷达特征图的分辨率，以实现更精确的可行驶区域检测；以及相机-激光雷达特征融合模块，将得到的高分辨率雷达特征与 RGB 图像提取出的色彩纹理特征融合，实现了在多种环境下的稳健可行驶区域检测。

其他代表性的相机-激光雷达融合道路场景解析网络包括 LidCamNet[73]、PLARD[68]、PLB-RD[69] 和 USNet[66]。2018 年，研究[73] 提出了 LidCamNet，该网络接受激光雷达点云和 RGB 图像作为输入。首先，非结构化且稀疏的激光雷达点云被投影到相机图像平面上并进行上采样，生成包含空间细节的密集 2D 图像。然后使用多个 FCN 网络进行道路可行驶区域检测，这些 FCN 网络可以处理来自单一传感器的数据，或通过三种融合策略之一：早期融合、晚期融合和交叉融合。在早期和晚期融合策略中，多模态信息在网络内的特定特征级别上进行融合。而交叉融合 FCN 则通过可训练的交叉连接来识别激光雷达和相机分支之间数据整合的最佳特征级别。这提高了提取和利用复杂空间关系的能力，从而提升了可行驶区域检测的准确性。2019 年，研究[68] 提出了一种名为 PLARD 的可行驶区域检测网络，通过融合激光雷达数据增强基于 RGB 图像的道路检测。PLARD 采用两个主要模块：(1) 数据空间适应，将激光雷达数据通过基于高度差的变换与视觉数据空间对齐，以匹配透视视图；(2) 特征空间适应，通过级联融合结构将激光雷达特征与视觉特征整合，以优化检测性能。在[69] 中，研究者提出了 LRDNet+，通过学习转换和融合操作解决激光雷达和视觉特征存在于不同空间的挑战，使用激光雷达数据增强视觉特征。随后，研究[74] 提出了用于可行驶区域检测的网络 PLB-RD，该方法通过深度估计网络模拟激光雷达传感器，对深度信息进行获取，并设计了特征融合模块对 RGB 图像与估计得到的深度图进

行处理，提取其中的异构特征用以增强可行驶区域检测任务的精确性，该方法还开发了一种策略来优化信息流路径以及一种模态蒸馏策略来最小化模型推理阶段的计算需求，在此阶段消除了对深度估计网络的需求。USNet[66] 通过利用 RGB 图像和深度图像以及更高效的特征融合策略，有效地平衡了可行驶区域检测的速度和准确性。具体地说，它采用两个轻量级子网络分别处理 RGB 图像与深度图像，具有不错的实时性能，该方法使用了多尺度池化操作从不同尺度的异构特征中提取不同细粒度的关键特征，以改善像素级分类。此外，设计了一种不确定性感知融合模块利用每种信息源的感知不确定性来指导子网络输出的整合，从而提高场景解析的准确性。

受 FuseNet[2] 的启发，现有最先进的专用网络通常采用并行编码器架构[16]，其中每个编码器从 RGB 图像与其他数据源或模态图像中提取多尺度的异构特征。随后进行异构特征融合，使网络能够对环境有更全面的理解[68]。例如，NIM-RTFNet[75]、SNE-RoadSeg[16]、SNE-RoadSeg+[18] 和 SNE-RoadSegV2[5] 将法向量图像应用到了可行驶区域检测任务中。该系列方法采用密集连接的跳跃连接来增强解码器中的特征保留能力，从而实现了同期最先进的性能。借鉴 Transformer 架构在单模态场景解析领域的成功，OFF-Net[17] 首次尝试将该架构应用于可行驶区域检测任务。该网络利用 SegFormer[24] 的编码器从RGB 图像与法向量图像中分别提取色彩纹理特征与平面几何特征，最终在越野道路数据集中超越了当时最先进 CNN 网络的性能。在上述研究的基础上，研究[26] 采用了一种新颖的 Transformer 解码器架构进行道路场景解析任务。此外，研究[26] 设计了更高效的异构特征融合策略，显著提高了在多个道路场景解析数据集上的任务性能，超越了同期所有的信息融合网络。

1.2.5　现有方法的局限性

尽管现有的道路场景解析研究已经取得了显著进展，但仍存在诸多不足之处，可总结如下：首先，仅依赖 RGB 图像或点云的单模态网络，虽然借鉴了通用场景解析领域的最新网络架构与范式，在色彩纹理特征的挖掘和任务适配性方面表现出色，但由于 RGB 图像在夜间等极端场景下容易失真，而激光雷达点云具有非结构化和稀疏的特点，这些网络在面对复杂、极端场景时性能波动较大，难以保持稳定的表现。

其次，现有的信息融合网络虽然在一定程度上克服了单模态网络的局限性，但仍存在以下不足：

(1) 现有网络大多采用较为原始的异构特征融合策略，未能充分利用信息的潜力。同时，它们往往忽视了通用场景解析领域中的先进网络架构或范式，相

比于单模态网络，性能提升有限。例如，现有的异构特征融合策略通常采用简单的特征通道串联或元素级相加操作，缺乏对特征间复杂关系的建模能力；

(2) 现有网络通常针对 RGB 图像与另一种特定的信息融合设计，将网络迁移至 RGB 图像与其他信息或模态时，网络性能显著下降。此外，这些网络仅采用“无区分性的异构特征编码”与单独设计的异构特征融合模块，忽视了特征提取阶段的重要性。特征提取网络架构与异构特征的适配性不足，限制了网络的性能提升[26]；

(3) 目前的信息融合网络大多针对封闭词汇设计，仅能预测固定类别的对象。在面对实际应用场景中的复杂多变性时，网络无法预测非预定义语义物体，泛化性与适用性有限。例如，现有的道路场景解析网络通常只能识别预定义的道路、车辆等类别，无法应对突发情况或罕见物体的识别。

1.3　主要研究内容

为了弥补现有方法的不足，进一步提升复杂、开放场景下的环境感知能力，尤其是道路场景解析任务的性能，本文使用 RGB 图像与其他信息作为输入，研究基于信息融合的道路场景解析网络。通过更高效地挖掘信息与异构特征的潜力，构建更有效的网络架构与范式，致力于突破自动驾驶汽车、机器人等自主导航系统在复杂道路场景中的瓶颈。具体研究内容包括以下几个方面：

(1) 针对第一个问题，提出了一种基于色彩纹理特征与平面特征融合的信息融合网络。该网络首次将基于 Transformer 架构的掩膜分类范式解码器引入道路场景解析领域，同时设计了基于注意力机制的异构特征自适应融合策略。实验表明，该网络在可行驶区域检测与道路破损检测任务中超越了现有的网络性能；

(2) 针对第二个问题，设计了适用于 RGB-D 图像对、RGB-T 图像对等任意信息作为输入的网络。该网络在道路场景解析领域内首次引入了视觉基础模型，并对其进行了任务适配，实现了更通用、鲁棒的异构特征提取，通过分析 RGB 图像与其他信息的内在特性，设计了非对称的特征编码与特征融合架构，从而实现对信息潜力的深度挖掘。在包括 RGB-D、RGB-T 等多种道路场景解析任务中，该网络均实现了领先的性能；

(3) 针对第三个问题，提出了一种具有零样本推理能力的信息融合道路场景解析网络。该网络基于视觉语言大模型提供的特征对齐能力，能够实现对场景中任意语义类别对象的预测。同时，该网络能够有效利用提取自先验深度图像的空间几何特征，克服了现有仅基于 RGB 图像的开放词汇场景解

析网络在掩膜预测精度不足的问题。在具有杂乱背景等挑战性场景下，显著提升场景解析任务的精确性与鲁棒性与适用性。

1.4　基础理论

该部分将对后续章节中所用到的公共基础理论知识进行介绍，包括基于 Transformer 的注意力机制以及掩膜分类范式的场景解析网络。

1.4.1　Transformer 与其注意力机制

Transformer 架构最初源于自然语言处理领域，由编码器和解码器两部分构成。其中，编码器主要由自注意力机制组成，解码器则同时包含自注意力机制和交叉注意力机制，此外还配备了前馈神经网络 (Feed Forward Network，FFN) 和层归一化 (Layer Normalization，LN) 模块。注意力机制作为 Transformer 的核心组件，能够有效捕获特征间的长距离依赖关系，实现全局特征建模。凭借这一优势，Transformer 架构在 NLP 领域取得了突破性进展。随后，研究者将其引入计算机视觉领域，发展成为与 CNN 并驾齐驱的主流视觉架构。Vision Transformer (ViT)[76] 是首个成功应用于视觉任务的 Transformer 架构，在图像分类任务上展现出超越 CNN 的性能，并被广泛应用于各类下游视觉任务的特征编码。鉴于本研究聚焦于计算机视觉领域，我们将以 ViT 为例说明 Transformer 架构组成。

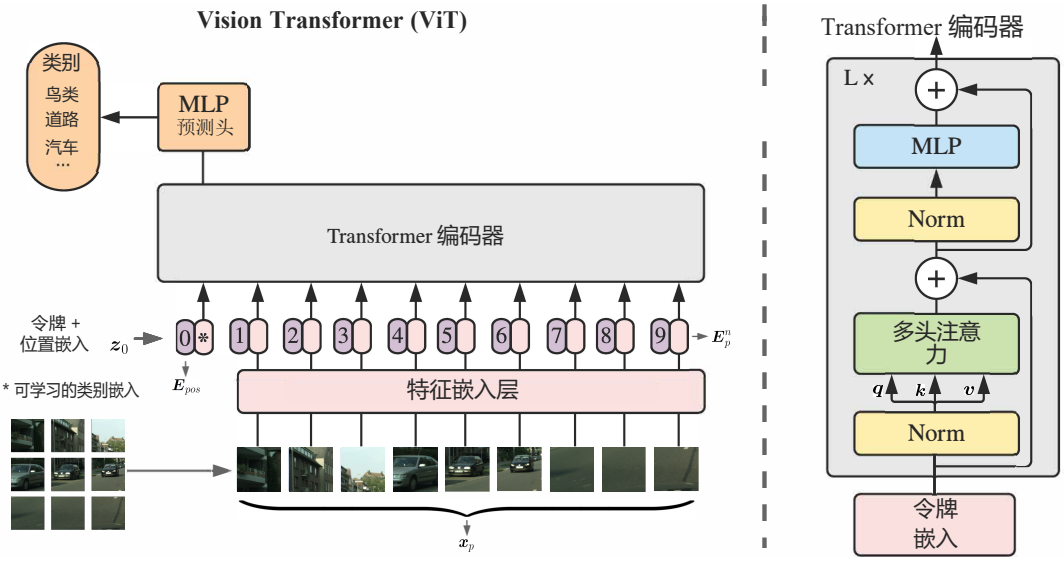


图 1.5　Vision Transformer 网络架构示意图

1.4.1.1　ViT 网络架构

标准 Transformer 架构处理的是一维文本嵌入序列 $\boldsymbol{E}_{text} \in \mathbb{R}^{N\times D}$，其中 N 表示序列长度，D 表示嵌入向量的特征维度。为了将二维图像数据适配到这一架构中，ViT 在进行注意力运算前需要对输入图像进行一系列预处理操作，如图 1.5所示。具体而言，首先将输入图像 $\boldsymbol{I}$ 调整为能被 16 整除的分辨率，随后将图像 $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$ 以 16×16 的步长划分为若干图像块，并将这些图像块展平为序列 $\boldsymbol{X}_p = [\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N] \in \mathbb{R}^{N\times(P^2\cdot 3)}$。其中，$(H,W)$ 代表图像的高度和宽度，$(P^2\cdot 3)$ 表示展平后的特征维度，(P,P) 为图像块的分辨率（此处 $P=16$），最终得到 $N = HW/P^2$ 个图像块，这也是输入 Transformer 的实际序列长度。在现代深度学习框架（如 PyTorch）中，可以通过 Rearrange 函数便捷地实现上述变换，将图像 $\boldsymbol{I} \in \mathbb{R}^{(\frac{H}{P}\cdot P)\times(\frac{W}{P}\cdot P)\times 3}$ 处理为图像块序列 $\boldsymbol{X}_p \in \mathbb{R}^{\frac{HW}{P^2}\times(P^2\cdot 3)}$：

$$\boldsymbol{X}_p = \text{Rearrange}(\boldsymbol{I}) \qquad \in \mathbb{R}^{N\times(P^2\cdot 3)}. \tag{1.1}$$

在 ViT 中，特征维度 D 在不同层间保持不变（但可作为超参数在不同规模网络中调整）。获得图像块序列 $\boldsymbol{X}_p$ 后，ViT 使用可训练的线性映射层 $\boldsymbol{E}$ 将其投影为 D 维的图像嵌入序列 $\boldsymbol{E}_p$，作为注意力机制的输入：$\boldsymbol{E}_p = \boldsymbol{X}_p\boldsymbol{E} \in \mathbb{R}^{N\times D}$，其中 $\boldsymbol{E} \in \mathbb{R}^{(P^2\cdot 3)\times D}$，单个注意力层的整体计算流程可表述如下：

$$\boldsymbol{Z}_0 = [\boldsymbol{x}_{\text{class}};\boldsymbol{E}_p] + \boldsymbol{E}_{pos}, \qquad \boldsymbol{E}_{pos} \in \mathbb{R}^{(N+1)\times D}, \tag{1.2}$$

$$\boldsymbol{Z'}_\ell = \text{MSA}(\text{LN}(\boldsymbol{Z}_{\ell-1})) + \boldsymbol{Z}_{\ell-1}, \qquad \ell = 1\ldots L, \tag{1.3}$$

$$\boldsymbol{Z}_\ell = \text{MLP}(\text{LN}(\boldsymbol{Z'}_\ell)) + \boldsymbol{Z'}_\ell, \qquad \ell = 1\ldots L, \tag{1.4}$$

$$\boldsymbol{Y} = \text{LN}(\boldsymbol{Z}_L^0), \tag{1.5}$$

其中在输入 Transformer 编码器之前，还需要引入一个可学习的" 类别" 嵌入向量 $\boldsymbol{z}_0^0 = \boldsymbol{x}_{\text{class}} \in \mathbb{R}^{1\times D}$，并将其与图像嵌入序列拼接。这里，上标 0 表示该嵌入在拼接后的序列 $\boldsymbol{E}_p$ 中位于首位，下标 0 表示编码器层的索引。在最后一层编码器输出 $\boldsymbol{z}_L^0$ 中，该类别嵌入被视为图像的全局表征，用于下游视觉任务（如公式 1.2所示）。此外，由于 ViT 将二维图像转换为一维序列处理，失去了 CNN 固有的空间结构特性，因此需要为输入序列添加位置编码 $\boldsymbol{E}_{pos}$。该编码可以是基于三角函数的固定编码，也可以是可学习的位置向量，用于保持序列中的空间关系（如公式 1.2所示），将最终得到的嵌入序列作为编码器的输入。

1.4.1.2　标准注意力机制

注意力机制根据查询（Query，$\boldsymbol{Q}$）、键（Key，$\boldsymbol{K}$）与值（Value，$\boldsymbol{V}$）特征矩阵的来源不同可以分为自注意力机制（Self Attention，SA）与交叉注意力机制

（Cross Attention，CA），当 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$、$\boldsymbol{V}$ 的来源相同时被称为自注意力机制，当 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$、$\boldsymbol{V}$ 的来源不同时被称为交叉注意力机制。而标准多头自注意力机制[23] 是 ViT 编码器的核心构建模块。对于输入序列 $\boldsymbol{Z} \in \mathbb{R}^{N\times D}$ 中的每个元素，自注意力机制通过计算其与序列中所有值向量 $\boldsymbol{V}$ 的加权和来获得上下文特征。这里的注意力权重 A_{ij} 反映了序列中任意两个元素之间的关联程度，其计算基于查询向量 $\boldsymbol{Q}^i$ 与键向量 $\boldsymbol{K}^j$ 之间的相似度：

$$[\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}] = \boldsymbol{Z}\boldsymbol{U}_{qkv}, \qquad \boldsymbol{U}_{qkv} \in \mathbb{R}^{D\times 3D_h}, \tag{1.6}$$

$$\boldsymbol{A} = \operatorname{softmax}\left(\boldsymbol{Q}\boldsymbol{K}^{\top}/\sqrt{D_h}\right) \qquad \in \mathbb{R}^{N\times N}, \tag{1.7}$$

$$\mathrm{SA}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{V}. \tag{1.8}$$

1.4.1.3　多头注意力机制

假设 $\boldsymbol{X}_q \in \mathbb{R}^{N_q\times C}$ 为查询矩阵，$\boldsymbol{X}_f \in \mathbb{R}^{N_f\times C}$ 为键和值矩阵，其中 N_q 和 N_f 分别表示查询和特征的序列长度。多头注意力机制（包括多头自注意力 MHSA 和多头交叉注意力 MHCA）可以按以下方式公式化：

$$\boldsymbol{Q}^i,\boldsymbol{K}^i,\boldsymbol{V}^i = \boldsymbol{X}_q\boldsymbol{U}_q^i,\boldsymbol{X}_f\boldsymbol{U}_k^i,\boldsymbol{X}_f\boldsymbol{U}_v^i, \quad i \in [1,n]\cap\mathbb{Z}, \tag{1.9}$$

其中 $\boldsymbol{U}_q^i \in \mathbb{R}^{C\times C_v}$，$\boldsymbol{U}_k^i \in \mathbb{R}^{C\times C_v}$，$\boldsymbol{U}_v^i \in \mathbb{R}^{C\times C_v}$ 表示第 i 个注意力头的查询、键和值的线性投影，C_v 是每个注意力头的投影维度，n 是注意力头的数量。接下来，计算第 i 个注意力头的注意力矩阵 $\boldsymbol{A}^i$，其计算方式为：

$$\boldsymbol{A}^i = \operatorname{softmax}\left(\frac{\boldsymbol{Q}^i(\boldsymbol{K}^i)^{\top}}{\sqrt{C_v}}\right), \tag{1.10}$$

然后，每个头的输出结果可以通过加权求和值向量 $\boldsymbol{V}^i$ 来表示：

$$\mathrm{MHA}(\boldsymbol{X}_q,\boldsymbol{X}_f) = \mathrm{Concat}\left([\boldsymbol{A}^1\boldsymbol{V}^1;\ldots;\boldsymbol{A}^n\boldsymbol{V}^n]\right)\boldsymbol{U}_m, \tag{1.11}$$

其中 $\boldsymbol{U}_m \in \mathbb{R}^{n\cdot C_v\times C}$ 表示所有注意力头的输出结果的线性投影，$\mathrm{Concat}\left([\cdot;\ldots;\cdot]\right)$ 表示将所有 n 个注意力头的输出拼接起来。

1.4.1.4　可变形注意力机制

与标准的多头注意力（MHA）机制相比，可变形注意力（Deformable Attention，DA）仅关注来自特征图的一小部分点作为参考点，而不是整个特征图，从而减少了计算复杂度，同时保持了竞争力的性能。DA 的采样过程可以通过对查询向量 $\boldsymbol{Q}$ 进行线性映射来预测采样偏移量和注意力权重 $\boldsymbol{A}^i$ 来实现。设 $\boldsymbol{p}_q = (h,w)$ 为重新调整后的 $\boldsymbol{Q}$ 上的二维参考点索引，其中 $\boldsymbol{Q} \in \mathbb{R}^{N_q\times C}$。在该参考点处，$\boldsymbol{Q}(\boldsymbol{p}_q) \in \mathbb{R}^{1\times C}$

表示与位置 $\boldsymbol{p}_q$ 对应的特征向量。DA 操作可以表示为：

$$\mathrm{DA}^i(\boldsymbol{Q}(\boldsymbol{p}_q),\boldsymbol{F}) = \sum_{k=1}^{K} \boldsymbol{A}^i_{k,\boldsymbol{p}_q} \cdot \boldsymbol{F}(\boldsymbol{p}_q + \Delta^i_{k,\boldsymbol{p}_q})\boldsymbol{U}^i_v, \tag{1.12}$$

其中 k 表示采样的键的索引，K 是总共采样的键的数量。$\Delta^i_{k,\boldsymbol{p}_q}$ 和 $\boldsymbol{A}^i_{k,\boldsymbol{p}_q}$ 分别表示第 i 个注意力头在参考点 $\boldsymbol{p}_q$ 上第 k 个采样点的采样偏移量和注意力权重。注意力权重 $\boldsymbol{A}^i_{k,\boldsymbol{p}_q}$ 经过归一化，满足 $\sum_{k=1}^{K} \boldsymbol{A}^i_{k,\boldsymbol{p}_q} = 1$。

假设 $\mathcal{X}_f = \{\boldsymbol{F}^l\}_{l=1}^{L}$ 表示多尺度特征图，其中 $\boldsymbol{F}^l \in \mathbb{R}^{H_l \times W_l \times C}$，并且设 $\hat{\boldsymbol{p}}_q \in [0,1]^2$ 为查询向量 $\boldsymbol{Q}$ 参考点的归一化坐标。DA 过程可以扩展为其多尺度可变形注意力（MSDA）的形式，如下所示：

$$\mathrm{MSDA}^i(\boldsymbol{Q}(\hat{\boldsymbol{p}}_q),\mathcal{X}_f) = \sum_{l=1}^{L}\sum_{k=1}^{K} \boldsymbol{A}^{i,l}_{k,\boldsymbol{p}_q} \cdot \boldsymbol{F}^l(\phi_l(\hat{\boldsymbol{p}}_q) + \Delta^{i,l}_{k,\boldsymbol{p}_q})\boldsymbol{U}^i_v, \tag{1.13}$$

其中 l 表示特征图的级别，k 表示采样点的索引。$\Delta^{i,l}_{k,\boldsymbol{p}_q}$ 和 $\boldsymbol{A}^{i,l}_{k,\boldsymbol{p}_q}$ 分别表示在第 i 个注意力头和第 l 个特征级别下，第 k 个采样点的采样偏移量和注意力权重。注意力权重 $\boldsymbol{A}^{i,l}_{k,\boldsymbol{p}_q}$ 经过归一化，满足 $\sum_{l=1}^{L}\sum_{k=1}^{K} \boldsymbol{A}^{i,l}_{k,\boldsymbol{p}_q} = 1$。变换 $\phi_l(\cdot)$ 将 $\hat{\boldsymbol{p}}_q$ 重新缩放到第 l 个特征级别的原始空间大小。

1.4.2　掩膜分类范式场景解析

现有的异构特征融合场景解析网络主要采用基于逐像素分类的解码器范式[2, 58]。在这一范式中，解码器的核心目标是对输入图像的每个像素预测一组 K 维的概率分布：

$$y = \{p_i | p_i \in \Delta^K\}_{i=1}^{H\cdot W}, \tag{1.14}$$

其中，K 代表目标语义类别的总数，$H \cdot W$ 表示图像的像素总数。优化此类网络的方法相对直接：为每个像素分配对应的真值标签 $y^{\mathrm{gt}} = \{y^{\mathrm{gt}}_i | y^{\mathrm{gt}}_i \in 1,\ldots,K\}_{i=1}^{H\cdot W}$，并通过逐像素交叉熵损失函数进行优化：

$$\mathcal{L}_{\text{pixel-cls}}(y, y^{\mathrm{gt}}) = \sum_{i=1}^{H\cdot W} -\log p_i(y^{\mathrm{gt}}_i). \tag{1.15}$$

掩膜分类范式最初源于 Mask R-CNN[77] 在实例级分割任务中的应用。由于传统 CNN 架构难以处理样本中实例数量不固定的情况，该范式在早期并未得到广泛关注。DEtection TRansformer（DETR）[78] 首次基于 Transformer 架构，在物体检测领域引入了基于集群预测的解码方式，通过交叉注意力机制实现实例级特征提取，并使用二部图匹配解决网络优化问题。此后，经过[79] 等工作的改进，逐渐产生了场景解析任务中的掩膜分类范式。相比于逐像素分类的解码器，掩膜

分类范式的网络相对较为复杂，但依赖于更好的实例级特征聚合能力，在场景解析等任务中同样具有优异的性能。

掩膜分类范式将场景解析任务解耦为两个关键部分进行联合优化：(1) 将图像分组为 N 个区域（注意 N 不一定等于 K），每个区域表示为二值化（类别无关）掩膜 $\{m_i | m_i \in [0,1]^{H\times W}\}_{i=1}^{N}$；(2) 将划分的每个区域视为一个整体实例，并且对这个实例预测 K 类别的概率分布。因此，网络最终的输出是 N 个类别-掩膜对，定义为 $z = \{(p_i, m_i)\}_{i=1}^{N}$。与逐像素分类预测不同，掩膜预测的类别概率分布 $p_i \in \Delta^{K+1}$ 除了包含 K 种语义类别外，还引入了一个额外的" 无对象" 类别（$\varnothing$），这个设定是为了满足集群预测的需要。值得注意的是，掩膜分类范式允许预测的多个掩膜具有相同的语义类别，这一机制使其在实例级与语义级场景解析任务中更具通用性。为了优化掩膜分类范式网络，需要对预测的类别-掩膜对 z 与对应的真值标注 $z^{\text{gt}} = \{(c_i^{\text{gt}}, m_i^{\text{gt}}) | c_i^{\text{gt}} \in \{1,\ldots,K\}, m_i^{\text{gt}} \in \{0,1\}^{H\times W}\}_{i=1}^{N^{\text{gt}}}$ 进行匹配，其中 c_i^{gt} 是对应第 i 个真值的语义类别。由于网络预测的集群数量 $|z| = N$ 通常与真值标注数量 $|z^{\text{gt}}| = N^{\text{gt}}$ 不一致，因此需要对 $N \geq N^{\text{gt}}$ 的部分进行填充, 将" 无对象" 类别 $\varnothing$ 分配给这些填充的真值，随后，通过匈牙利算法[80] 对 z 和 z^{gt} 进行二部图匹配，确定每个 z 的优化目标。最终，通过交叉熵类别损失与二值化掩膜损失对每个类别-掩膜对进行监督，实现网络的端到端训练。

1.5　论文组织架构

围绕前文所述的研究内容，本文共分为五个章节展开，其中算法创新部分组织架构如图 1.6所示，整体结构如下：

第 1 章介绍道路场景解析任务的研究背景与意义，总结国内外研究现状，分析现有方法的优势与不足，并提出本文的研究目标与内容，明确需要解决的三个

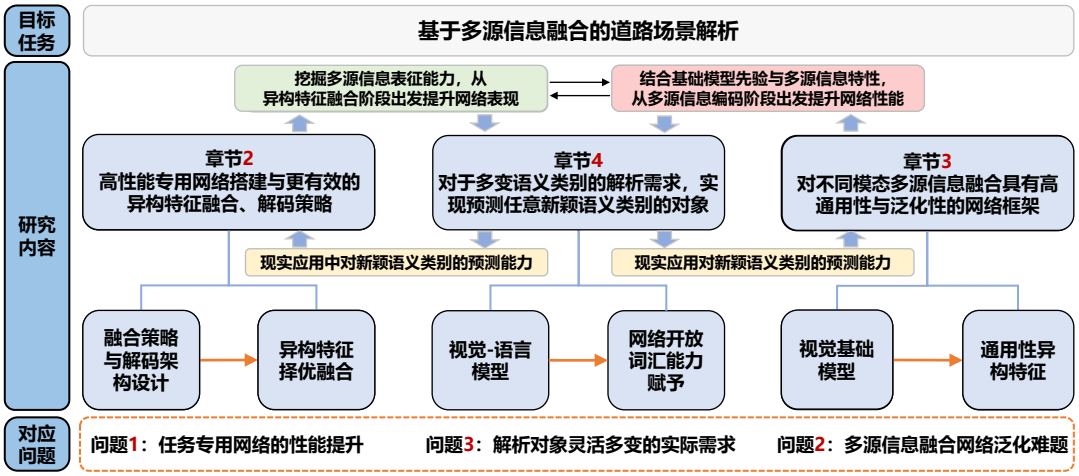


图 1.6　本文算法创新部分组织架构

主要问题；并阐述后续章节所需的基础理论，包括 Transformer 架构、注意力机制以及用于场景解析任务的掩膜分类范式，为后续章节中的研究提供理论支持；

第 2 章到第 4 章介绍本文提出的三种创新的信息融合道路场景解析网络架构，用以突破相关领域内的现有问题，其中第 2 章针对现有信息融合道路场景解析专用网络在 Transformer 架构应用方面的空白以及异构特征融合策略的低效问题，提出了一种以 RGB-SN 图像对作为输入的信息融合网络。该网络首次将基于 Transformer 的掩膜分类解码器引入道路场景解析领域，并设计了基于注意力机制的自适应特征融合策略，在可行驶区域检测与道路破损检测任务中取得了显著性能提升；

第 3 章引入视觉基础模型，提出了一种非对称的特征编码与融合架构，更有效地提取与编码异构特征。通过分析 RGB 图像与其他信息（如深度图像、热图像等）的内在特性，从异构特征编码阶段入手，设计了对不同信息适配性更强的异构特征提取网络，提升了网络的泛化能力，在不同的信息融合道路场景解析任务中实现了领先性能；

第 4 章通过引入视觉-语言模型为信息融合网络赋予零样本推理能力，提出一种基于信息融合的开放词汇场景解析网络。该网络利用视觉语言模型的多模态特征对齐能力，能够预测场景中任意语义类别的对象，克服了现有仅基于 RGB 图像的开放词汇网络的性能局限，同时迎合了信息融合网络在应用中解析对象类别灵活多变的需要，在杂乱背景等挑战性场景下显著提升了解析任务的精确性与鲁棒性。

第 5 章总结了本文的主要工作及研究成果，并对未来的研究方向进行展望。

第 2 章　基于并行编码的双源信息融合道路场景解析网络

针对道路场景中的可行驶区域与破损区域检测任务，本章节提出了一种使用 RGB-法向量图像对作为输入的双源信息融合网络，极大的提升了上述任务的表现。现有的双源信息融合网络普遍以 RGB 图像、深度图像以及视差图像等信息作为输入，提取异构特征并融合以提升网络表现。然而，直接对深度与图像进行编码得到的特征对可行驶区域以及道路破损等对象的表示能力较弱，导致网络性能提升十分有限。另一方面，网络的异构特征融合策略较为原始、单一，导致无法充分发挥异构特征的潜力。针对这些问题，本章将目光转向法向量图像，将其与 RGB 图像共同作为网络输入构建道路场景解析网络，有效提升了在极端天气、光照变化、背景复杂等情况下的任务性能。网络首先通过法向量估计器从深度图像中估计出法向量图像，并使用并行编码器分别提取多尺度的色彩纹理特征与平面特征，设计了基于注意力机制的异构特征同步模块以进一步进行有效的异构特征校准与融合，最后将融合特征送入基于 Transformer 的解码器进行掩膜提取与分类，实现精确的场景解析。在公开与自建的数据基准上的对比实验表明，所提出的网络在包含道路破损等对象的道路场景解析任务种取得了同期最优秀的性能，在道路场景解析任务上表现出了更强的精确性与鲁棒性。

2.1　引言

如前文所述，道路场景解析任务旨在对广泛存在于道路场景中的对象进行像素级语义分类，其中包括了可行驶区域检测以及道路破损检测等任务。在本章中，我们主要关注这两种任务，并构建专用网络以提升任务性能。在上述任务的领域内，基于深度神经网络的方法，由于表现出了相比于传统基于几何的解析方法更强的鲁棒性与泛化性，成为了目前的主流方法。早期的 CNN 时代，如[15] 等研究中提出了基于 CNN 的编码器解码器架构，利用 RGB 图像作为输入来完成可行驶区域检测任务。然而，这种单模态网络的性能十分有限。为了解决这个问题，后续的研究中提出了基于信息融合的方法以进一步提升场景解析的性能。

在早期方法如[2] 中提出的网络是一种双源信息融合网络，该网络从 RGB 图像与深度图像中分别提取异构特征，并通过元素级相加操作进行异构特征融合，相比于同期单模态的网络，异构融合特征提供了更全面的场景理解，从而获得了性能的提升。SNE-RoadSeg[16, 18] 系列网络首次将法向量图像引入可行驶区域检测任务，该系列网络使用双分支的并行 ResNet[63] 编码器分别对 RGB 图像与法

向量图像进行特征编码，并同样通过元素级相加进行异构特征融合，并最终送入基于 U-Net 架构[19] 的稠密连接解码器进行场景解析。结合了 RGB 图像提供的丰富色彩、纹理特征与法向量图像编码的准确物体表面特性，该系列网络在可行驶区域检测任务的权威评测基准 KITTI Road[20] 上取得了同期最精确的结果。然而，该系列网络仅采用不具有“特征选择性”的元素级相加实现异构特征融合，这可能导致异构特征表示之间产生冲突，从而影响最终得预测结果。此外，在多类别场景解析任务中网络的性能十分有限，且在面对道路破损检测这类小任务对象时，预测效果不理想。

另一方面，ViT[76] 的出现让研究者们看到了 Transformer 架构在计算机视觉任务中的巨大潜力，越来越多基于 Transformer 的视觉编码器[24, 48] 被提出，且在下游任务中达到了比肩甚至超越 CNN 架构的性能。Transformer 架构的核心优势在于其中的注意力机制，可根据场景上下文进行自适应的特征交互与建模，实现更为高效的全局特征聚合；相比之下，CNN 架构的感受野有限，且在面对大数量级数据时存在性能饱和等问题[81]。在道路场景解析领域，研究[17] 率先提出了基于 Transformer 架构的可行驶区域检测网络，该网络使用了基于 SegFormer[24] 的权重共享编码器对 RGB 图像与法向量图像进行编码，并采用了基于轻量级 CNN 的解码器，在其原创的 ORFD[17] 数据集中实现了最精确的可行驶区域检测效果。然而该研究仅在特征编码阶段采用了基于 Transformer 的架构，在异构特征融合以及特征解码阶段依旧采用了简单的元素级相加以及 CNN 解码器，对 Transformer 架构的利用程度有限。在道路场景解析领域，利用基于 Transformer 的注意力机制进行异构特征融合以及特征解码的相关研究仍旧存在不足。

在此背景下，本章以 RGB 图像与法向量图像作为网络输入，引入基于 Transformer 的异构特征融合模块与特征解码范式，提出了一种高性能的双源信息融合道路场景解析网络框架。该网络首先通过并行编码器分别从 RGB 图像与估计得到的法向量图像中提取多尺度异构特征；之后利用基于 Transformer 的自注意力机制，在所提取异构特征的空间维度与通道维度分别进行自适应校准与融合，得到异构融合特征；最后引入了基于 Transformer 的掩膜分类解码范式解码融合特征，实现精确的道路场景解析。此外，为弥补领域内道路破损检测任务数据基准短缺的问题，本章基于 Carla[82] 仿真器制作并收集了一个拥有超过 10000 对样本的的大规模合成道路破损检测数据集，包含 RGB 图像、深度图像、视差图像、法向量图像以及像素级的语义标注，并构建了相关评测基准用于领域内的算法对比。

总结而言，本章工作的主要贡献如下：

(1) 本章节设计了一种基于并行编码器的高性能双源信息融合道路场景解析网络 RoadFormer，该网络利用 RGB 图像与对应的法向量图像作为输入，因

而可以学习到道路场景解析任务中对可行驶区域与道路破损等对象类别中更关键的物体表面特性，从而有效提升了道路场景解析任务的精确性与鲁棒性；

(2) 本章节设计了基于 Transformer 注意力机制的异构特征融合策略，对色彩纹理特征与平面特征进行空间以及通道层面的自适应特征建模，更加充分的发挥了所提取异构特征在任务中的潜力；

(3) 本章节通过在多个数据集上进行系统的对比试验，证明了所提出的 RoadFormer 网络与其中异构特征融合策略的有效性，RoadFormer 在公开数据集 KITTI Road[20]，Cityscapes[22]，ORFD[17] 以及自建数据集 SYN-UDTIRI 上都取得了同期最优秀的性能。

2.2　道路场景解析网络 RoadFormer

2.2.1　基于并行编码器的异构特征提取

如图 2.1所示，本章节提出的道路场景解析网络采用 RGB 图像和法向量图像作为双路输入。其中，RGB 图像来自彩色相机，而法向量图像则通过对深度图像的估计获得，深度图像可来源于深度相机或激光雷达等深度传感器。为获取高质量的法向量估计结果，本章采用了一种基于几何方法的高性能法向量估计器[83] 来处理各数据集的深度图像，该估计器无需任何可学习参数。为确保双源信息在时空维度上的对齐性，需要选取同一时刻的 RGB 图像与深度图像，并进行严格的空间配准。这些预处理步骤主要涉及传感器的内外参标定等技术环节，由于相关技术已较为成熟，本章不再详述。经过对齐的 RGB 图像和法向量图像

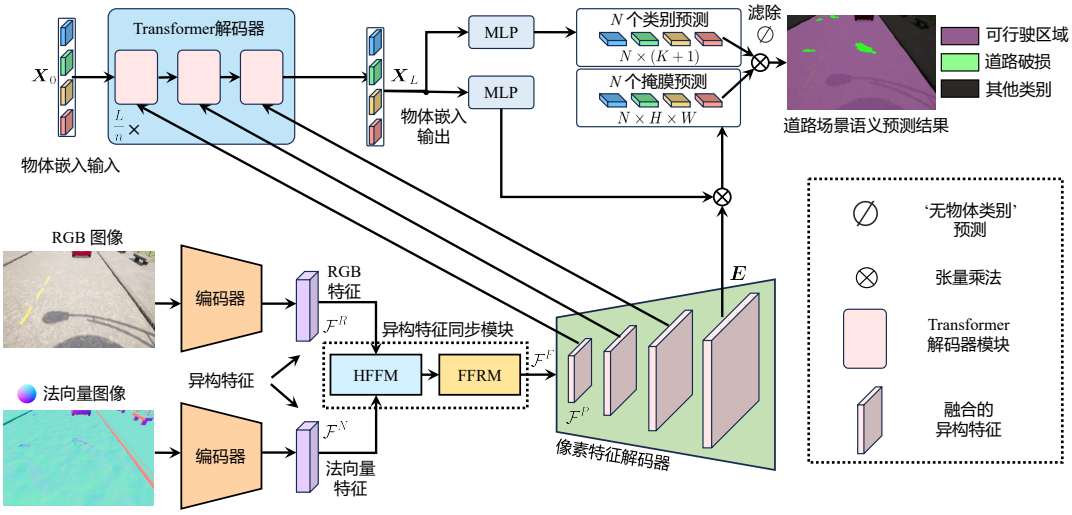


图 2.1　基于 RGB-法向量图像的双源信息融合道路场景解析网络 RoadFormer

分别通过并行编码器，提取多尺度异构特征。这些特征随后经由专门设计的异构特征融合模块进行融合以及像素级解码器增强特征表达，最终采用基于掩膜分类范式的解码器 1.4.2进行不同语义类别的掩膜预测，实现高精度的场景解析。

鉴于多源信息对通常侧重于场景中的不同信息维度（如色彩纹理以及空间几何等），领域内普遍采用并行式编码器架构对异构特征进行特征提取[16, 18]，并行式编码器架构之间不共享权重参数，以确保能够充分捕获异构特征中的特征多样性。本章网络同样采用了该种编码器架构，一个编码器负责从 RGB 图像中提取色彩和纹理特征，另一个编码器则从法向量图像中提取物体表面几何特征。考虑到道路场景解析任务中目标对象具有显著的尺度变化特性，网络需要依托不同细粒度的上下文特征以提取更加完备的场景特征表征，因此本章采用了多尺度的骨干网络架构以提取多尺度异构特征。

在当前领域具有代表性的高性能网络中，包括单模态网络[56] 以及信息融合网络[84] 等研究普遍采用 Swin Transformer[48] 作为视觉骨干网络；另一方面，本章节提出的异构特征融合模块以及后续的解码器都基于 Transformer 架构进行设计，因此本章同样选择采用了 Swin Transformer 构建并行式编码器。然而，不同于目前主流的高性能场景解析网络都依赖于基于 Transformer 架构的骨干网络，本章节进一步证明了 CNN 架构的骨干网络[85] 同样适配场景解析任务，并与基于 Transformer 的特征融合及解码器范式进行有效适配。在后续实验中，我们选择了 ConvNeXt[85] 这一与 Swin Transformer[48] 同期的 CNN 架构骨干网络，并进行了编码器架构选择的消融实验，相关结果将在 2.3.3节中详细讨论。

Swin Transformer　将 ViT 范式的 Transformer 直接应用到视觉领域面临着诸多挑战：首先，视觉图像的尺度变化通常较大且输入尺寸非固定；其次，由于全局注意力机制的计算复杂度与令牌数量呈平方关系，而相较于文本信息，图像通常具有更高的分辨率，这将导致计算开销呈指数级增长。针对上述问题，Swin Transformer[48] 创新性地提出了基于滑动窗口的局部注意力机制和多尺度特征提取策略。其中，滑动窗口注意力机制（如图 2.2所示）包含两个连续的注意力计算阶段：(1) 局部窗口内的注意力计算；(2) 窗口滑动后的跨窗口注意力计算。这种设计将注意力计算限制在恒定大小的窗口范围内进行，一方面引入了类似 CNN 的局部感受野特性从而显著降低计算复杂度，另一方面通过窗口间的特征交互实现了感受野的扩展。

回顾第 1.4.1.1节的内容，对于空间分辨率为 (H,W) 的输入图像，传统 ViT 网络的计算复杂度为 $C\cdot(hw)^2$，Swin Transformer 通过引入窗口注意力机制大幅降低了计算复杂度。具体而言，Swin Transformer 首先将输入划分为 4×4 大小的图像块，随后将这些图像块进一步组织为统一大小的窗口，并交替使用窗口注意力

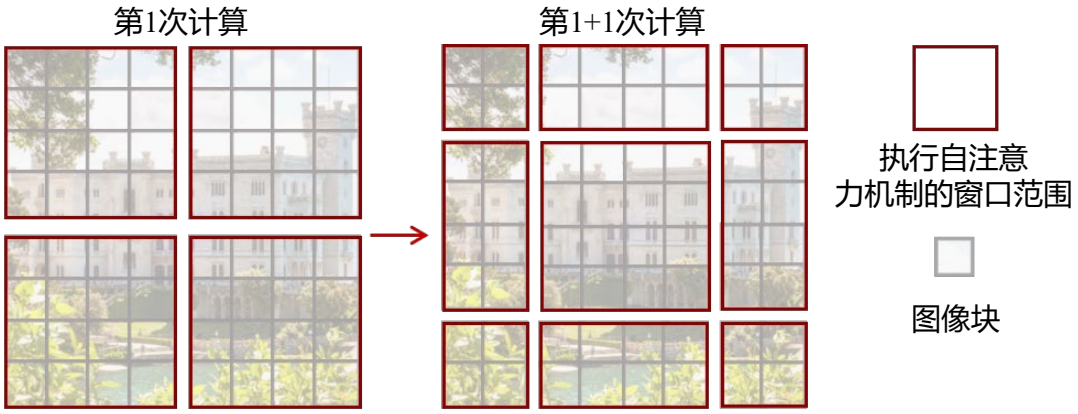
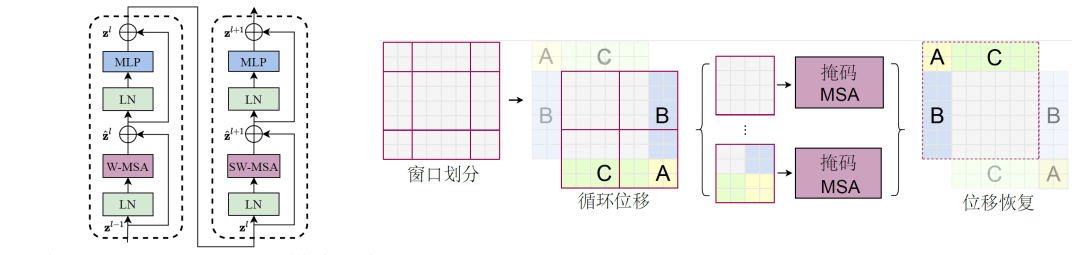


图 2.2　Swin Transformer 中的滑动窗口注意力机制

（W-MSA）与滑动窗口注意力（SW-MSA）在窗口范围内进行掩码自注意力计算（如图 2.3（a）所示）。设每个窗口包含 $M \times M$ 个图像块，则整个特征图共包含 $\frac{hw}{M^2}$ 个窗口。此时单个窗口内的注意力计算复杂度为 $(M^2)^2$，总体计算复杂度为 $\frac{hw}{M^2} \times (M^2)^2 = M^2 \times (hw)$。由于每个窗口内的令牌数量固定且远小于总令牌数，因此相比于 ViT，计算复杂度从令牌数量的平方关系降低至线性关系，实现了显著的效率提升。然而，单纯的窗口注意力机制会限制网络的感受野，因此需要通过窗口滑动策略来促进不同窗口间的特征交互。具体做法是对上一层网络模块的窗口位置进行滑动，并采用如图 2.3（b）所示的循环移动（cyclic shift）方式划定新的窗口范围，随后进行一次新的注意力计算，从而实现窗口间的信息交换。

ConvNeXt　Swin Transformer 凭借其创新的窗口注意力机制在众多下游视觉任务中取得了显著成功。然而，窗口注意力机制本质上仍属于一种局部机制，这在某种程度上印证了 CNN 的局部滑窗机制在视觉任务中的价值。有研究者认为，Swin Transformer 的成功可能并不完全源于注意力机制本身，而是其精心设计的网络架构。基于这一认识，ConvNeXt[85] 探讨了用传统卷积替代局部注意力机制的可能性，并通过架构的优化实现了 CNN 编码器性能的显著提升。在借鉴 Swin Transformer 架构设计的基础上，同期的研究工作 ConvNeXt 对 ResNet[63] 架构进行了以下关键改进：(1) 用深度可分离卷积取代了瓶颈结构（bottleneck）中的 3×3 卷积，同时将网络基础通道数从 64 扩展至 96；(2) 重新设计瓶颈结构中的通道维



(a) 连续的Swin Transformer模块示意图　　(b) 窗口的循环位移示意图

图 2.3　Swin Transformer 网络架构示意图

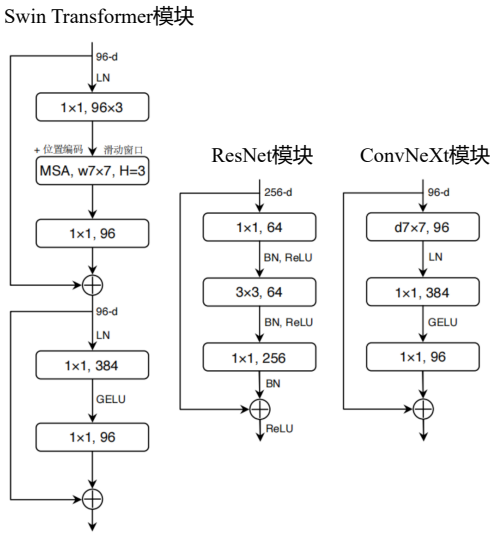


图 2.4　Swin Transformer、ResNet 与 ConvNeXt 模块架构对比图

度变化顺序，从传统的" 大维度-> 小维度-> 大维度" 调整为" 小维度-> 大维度-> 小维度"；(3) 采用更大的 7×7 卷积核扩大感受野。如图 2.4所示，通过这些精心的架构改进，ConvNeXt 网络在多项视觉任务中实现了与 Swin Transformer 相当的性能表现，同时保持了 CNN 架构固有的高效推理特性。

在综合参考了[16, 18] 等代表性研究中的网络规模设计，并权衡性能与计算效率的基础上，本章最终选择了 Swin Transformer-Base 与 ConvNeXt-Base 作为并行式编码器的骨干网络。这两种网络架构具有相近的参数规模，且都能输出步长分别为 481632 的四种尺度的多层次特征表征，能够很好地满足场景解析任务对多尺度特征的需求。具体而言，对于输入的双源图像对，其中一个编码器分支负责从 RGB 图像 $\boldsymbol{I}^R \in \mathbb{R}^{H\times W\times 3}$ 中提取丰富的色彩纹理特征序列 $\mathcal{F}^R = \{\boldsymbol{F}_1^R, \ldots, \boldsymbol{F}_k^R\}$，另一个编码器分支则专门从法向量图像 $\boldsymbol{I}^N \in \mathbb{R}^{H\times W\times 3}$ 中提取物体表面的平面特征序列 $\mathcal{F}^N = \{\boldsymbol{F}_1^N, \ldots, \boldsymbol{F}_k^N\}$。在上述表达式中，$H$ 和 W 分别表示输入图像对的高度和宽度，$\boldsymbol{F}_i^{R,N} \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times C_i}$ 代表第 i 个尺度的特征图，C_i 表示该层特征图的通道数，$S_i = 2^{i+1}$ $(i \in [1,k] \cap \mathbb{Z})$ 表示相应的下采样步长。这种并行式编码使得网络能够同时充分利用 RGB 图像中的色彩纹理特征和法向量图像中的平面几何特征，通过独立的参数学习来适应各自模态的特征分布特性，从而能够更有效地提取各自模态中的判别性特征，为后续的模块提供更加全面和鲁棒的特征表示。

2.2.2　基于注意力机制的异构特征自适应融合

现有的信息融合网络大多采用简单的元素级加法或特征通道级拼接等操作来融合多尺度异构特征。然而，这些简单的融合方式难以充分发挥异构特征的潜力。为此，我们提出了异构特征同步模块（Heterogeneous Feature Synergy Block, HFSB）来实现更有效的特征融合。HFSB 包含异构特征融合模块（Heterogeneous

Feature Fusion Module, HFFM）和融合特征校准模块（Fusion Feature Recalibration Module, FFRM）两个核心组件。并行式编码器提取的多尺度特征依次通过这两个模块完成完整的融合过程。

异构特征融合模块 HFFM　Transformer 凭借其注意力机制强大的特征建模能力，在单模态视觉和多模态视觉-语言等任务中取得了突破性进展[86]。受此启发，我们认为注意力机制同样可以有效应用于基于信息融合的道路场景解析任务中。具体而言，我们基于自注意力机制设计了异构特征同步模块 HFSB（如图 2.5（a) 所示）。该模块通过自注意力机制实现自适应的特征交互，从而增强异构特征的融合效果，提升场景解析性能。HFFM 的具体操作可以用以下公式表示：

$$\boldsymbol{F}_i^H = \mathrm{Reshape}\Big(\mathrm{Norm}\big(\mathrm{Softmax}(\boldsymbol{Q}_i^C \boldsymbol{K}_i^{C\top})\kappa_i \boldsymbol{V}_i^C + \boldsymbol{F}_i^C\big)\Big), \tag{2.1}$$

在这个过程中，首先将色彩纹理特征 $\boldsymbol{F}_i^R$ 和平面特征 $\boldsymbol{F}_i^N$ 在通道维度上拼接，并重塑为融合特征 $\boldsymbol{F}_i^C \in \mathbb{R}^{2C_i \times \frac{H}{S_i}\frac{W}{S_i}}$。该特征随后通过恒等映射分别生成查询嵌入 $\boldsymbol{Q}_i^C$、键嵌入 $\boldsymbol{K}_i^C$ 和值嵌入 $\boldsymbol{V}_i^C$。HFFM 的输出为融合特征 $\boldsymbol{F}_i^H \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times 2C_i}$。此外，参考[37] 的研究，我们引入了可学习的调节因子 κ_i 来动态调控学习到的异构融合特征 $\boldsymbol{F}_i^H$ 的重要程度，以实现更优的特征融合效果。

融合特征校准模块 FFRM　在单模态视觉任务中常用的网络架构中，编码器部分所提取的视觉特征通常包含多个通道，这些特征并非都能对视觉预测产生正

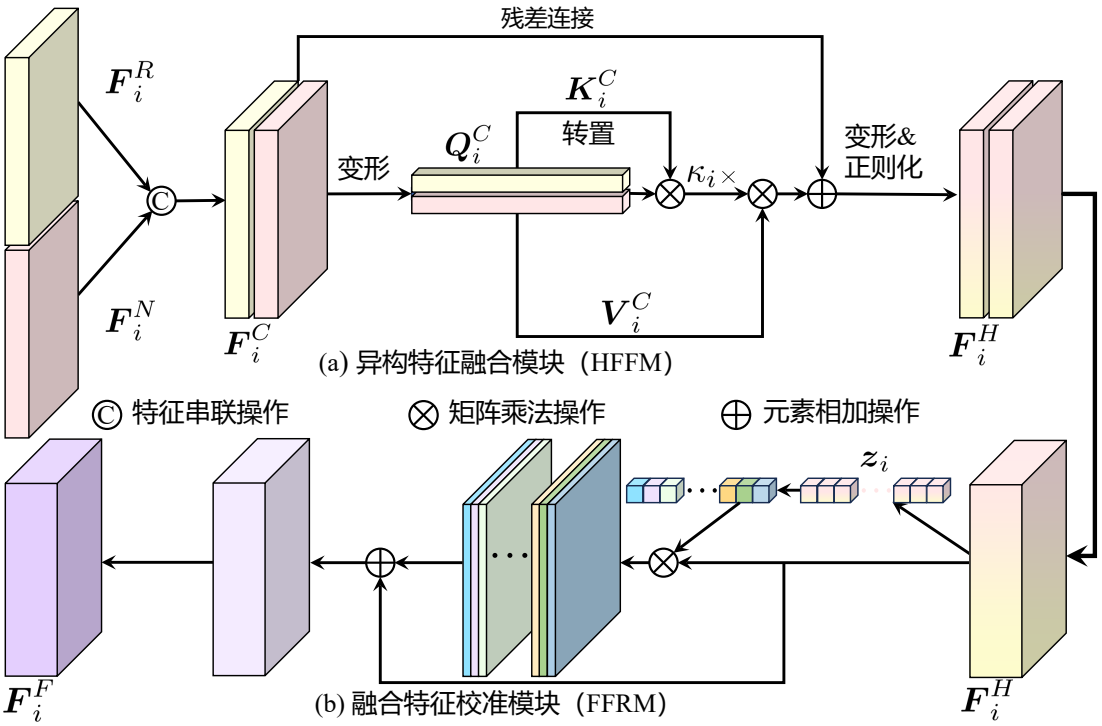


图 2.5　基于注意力机制的异构特征同步模块（HFSB）结构示意图

向作用。事实上，其中部分存在噪声或者与语义无关的通道特征甚至可能严重损害网络的性能。据此，一些工作[87] 提出通过通道注意力机制网络建模不同通道特征之间的内在依赖，以实现自适应的有益特征增强与噪声等无效特征的抑制。在异构特征融合过程中，同样可能产生类似的情况。由于提取的异构特征侧重于场景的不同特性，因此融合得到的特征 $\boldsymbol{F}_i^H$ 可能会劈坏原始特征中关键通道特征的显著性，甚至产生新的不相关特征或噪声特征[88]。针对这个问题，我们基于研究[87] 中提出的压缩激励模块（Squeeze-and-Excitation Block，SEB）提出了 FFRM（如图 2.5(b) 所示）以对得到的异构融合特征进行进一步的校准。我们在 SEB 至上增加了残差链接以增强了模块的训练效率，确保在深度网络能够进行更有效的反向梯度传播[63]。此外，我们引入了额外的深度可分离卷积操作以更好的实现对初步校准特征间的关系建模[89]。所提出的 FFRM 可以表述为如下公式：

$$\boldsymbol{F}_i^F = \underset{1\times1}{\mathrm{Conv}}\left(\boldsymbol{F}_i^H + \left(\boldsymbol{O}\ \mathrm{Sigmoid}\left(\underset{1\times1}{\mathrm{Conv}}(\boldsymbol{z}_i)\right)\right) \odot \boldsymbol{F}_i^H\right), \tag{2.2}$$

其中 $\boldsymbol{O} \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times1\times1}$ 相同形状的全一矩阵，$\odot$ 代表哈达玛积操作，$\boldsymbol{z}_i = [z_{i,1},\ldots,z_{i,2C_i}] \in \mathbb{R}^{1\times1\times2C_i}$ 代表对特征 $\boldsymbol{F}_i^H$ 中的每个通道的特征图进行平均池化操作的结果，平均池化操作过程可表述为：

$$z_{i,j} = \frac{S_i^2}{HW}\sum_{h=1}^{\frac{H}{S_i}}\sum_{w=1}^{\frac{W}{S_i}}\boldsymbol{F}_i^H(h,w,j), \tag{2.3}$$

对于所提出的 HFFM，FFRM 以及之前研究中的 SEB 等特征融合模块的消融实验会在后续章节中进行，可见 2.3.3。

2.2.3　基于掩码注意力的掩膜分类解码器与损失函数设计

基于多尺度可变形注意力的像素特征解码器　在获得多尺度融合特征$\mathcal{F}^F = \{\boldsymbol{F}_1^F,\ldots,\boldsymbol{F}_k^F\}$ 后，本章节引入了多尺度可变形 Transformer 像素解码器以加强跨尺度特征建模能力[56]。该解码器输出 $\mathcal{F}^P = \{\boldsymbol{F}_1^P,\ldots,\boldsymbol{F}_n^P\}$（$n < k$ 通常 $n = 3$），其中 $\boldsymbol{F}_i^P \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times C}$。解码器同时包含一个上采样层，负责将 $\mathcal{F}^F$ 中的低分辨率特征进行上采样，并与最高分辨率特征进行特征相加得到得到高分辨率的像素嵌入 $\boldsymbol{E}$，并通过特征相加生成高分辨率像素嵌入 $\boldsymbol{E}$，该嵌入将在后续 Transformer 解码器中引导掩膜预测鉴于多尺度可变形注意力（MSDA）在先前单模态相关工作[90] 中展现的卓越性能与计算效率，本章在像素特征解码器中采用该操作，以构建像素嵌入特征 $\mathcal{F}^P$。与单模态工作相比，我们的像素解码器创新性地接受来自 RGB 图像与法向量图像的异构融合特征进行特征建模，这使得每个尺度的特征图通道数量较之前的工作增加一倍。为了有效控制网络计算复杂度，在进行多

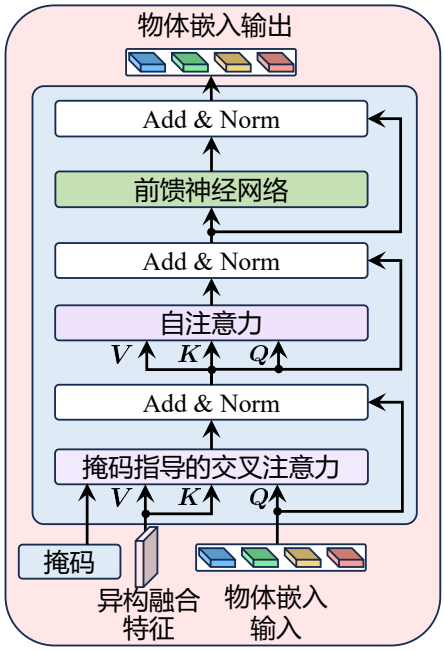


图 2.6　包含掩码交叉注意力的 Transformer 解码器模块示意图

尺度可变形注意力操作前，我们在每个尺度的特征图上应用 1×1 卷积层进行通道压缩。

基于掩码注意力的掩膜分类范式解码器　本章节采用了基于 Transformer 的掩膜分类解码器，该范式的相关基础已在章节 1.4.2中解释过，此处不再赘述。在本章节网络中，通过像素特征解码器输出的多尺度特征图 $\boldsymbol{F}_1^P$ 至 $\boldsymbol{F}_n^P$ 通过解码器对物体查询 $\boldsymbol{X}_0 \in \mathbb{R}^{N \times C}$ 进行迭代式更新，以实现精确的掩膜预测，并借助像素嵌入 $\boldsymbol{E}$ 指导掩膜生成。具体而言，初始化 N 个可学习的 C 维向量作为初始物体查询 $\boldsymbol{X}_0$，并将其输入后续多层 Transformer 解码器。每个解码器层的输出特征均会进行预测与监督。在每个解码器层（如图 2.6所示）内，按序执行以下操作：(1) 通过线性映射操作 $f_Q(\cdot)$ 得到查询特征 $\boldsymbol{Q}_l^D = f_Q(\boldsymbol{X}_{l-1}) \in \mathbb{R}^{N \times C}$，其中 l 代表层索引；(2) 通过线性映射操作分别获取键特征 $\boldsymbol{K}_l^D = f_K(\boldsymbol{F}_i^P) \in \mathbb{R}^{\frac{H}{S_i}\frac{W}{S_i} \times C}$ 和值特征 $\boldsymbol{V}_l^D = f_V(\boldsymbol{F}_i^P) \in \mathbb{R}^{\frac{H}{S_i}\frac{W}{S_i} \times C}$；(3) 将 $\boldsymbol{Q}_l^D$，$\boldsymbol{K}_l^D$ 和 $\boldsymbol{V}_l^D$ 通过掩码交叉注意力机制进行交互，该过程数学表达如下：

$$\boldsymbol{X}_l^C = \mathrm{Softmax}(\boldsymbol{M}_{l-1} + \boldsymbol{Q}_l^D \boldsymbol{K}_l^{D\top}) \boldsymbol{V}_l^D + \boldsymbol{X}_{l-1}, \tag{2.4}$$

其中 $\boldsymbol{M}_{l-1} \in \mathbb{R}^{N \times \frac{H}{S_i}\frac{W}{S_i}}$ 代表注意力掩码，可表示为：

$$\boldsymbol{M}_{l-1}(x,y) = \begin{cases} 0 & \text{当} \boldsymbol{M}_{l-1}(x,y) = 1 \text{时} \\ -\infty & \text{其他情况} \end{cases}, \tag{2.5}$$

其中 $\boldsymbol{M}_{l-1} \in \{0,1\}^{N\times\frac{H}{S_i}\times\frac{W}{S_i}}$ 是通过对第 $l-1$ 层（上一层）解码器模块的掩膜预测结果以 0.5 作为阈值得到的二值化输出；(4) 掩码交叉注意力机制的输出 $\boldsymbol{X}_l^C$ 随后经过标准多头自注意力、前馈神经网络和归一化层，得到该解码器层的物体查询输出 $\boldsymbol{X}_l$。从 $\boldsymbol{F}_1^P$ 至 $\boldsymbol{F}_n^P$ 的多尺度特征图依次输入解码器层并进行交叉注意力（持续 n 层），整个过程重复 $\frac{L}{n}$ 次以充分增强特征更新。

在每层预测阶段，将输出的物体查询 $\boldsymbol{X}_L$ 通过多层感知机（MLP）映射至 $(K+1)$ 维类别预测（包含 K 个语义类别和额外的" 无对象" 类别），并同时将 $\boldsymbol{X}_L$ 映射为掩膜嵌入，通过与逐像素嵌入 $\boldsymbol{E}$ 的点积得到掩膜预测输出。最终，通过对类别预测与掩膜预测进行矩阵乘法，生成语义场景解析输出，每个物体查询对应一个具体的类别-掩膜对。需要注意的是，在训练阶段，对每层解码器的预测进行监督；在测试阶段，仅采用最后一层 $\boldsymbol{X}_L$ 作为网络的最终预测结果。

场景解析损失函数设计　如前文 1.4.2所述，本章节所采用的掩膜分类解码器的损失函数由两个主要部分构成：分类预测损失和掩膜预测损失。其中，分类预测采用标准交叉熵（Cross Entropy，CE）作为损失函数，掩膜预测则结合使用二元交叉熵（Binary CE，BCE）损失与 Dice 损失[91]。分类损失 $\mathcal{L}_{\text{ce}}$ 的数学表达式为：

$$\mathcal{L}_{\text{ce}} = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y}_{k+1}^i \log(\hat{\boldsymbol{y}}_{k+1}^i), \tag{2.6}$$

其中 $\boldsymbol{y}_{k+1}^i$ 表示仅包含 0 或 1 的 $k+1$ 维独热编码类别信息标注，$\hat{\boldsymbol{y}}_{k+1}^i$ 表示网络预测的 $k+1$ 维类别概率分布。用于掩膜预测的损失函数 $\mathcal{L}_{\text{bce}}$ 和 $\mathcal{L}_{\text{dice}}$ 分别定义为：

$$\mathcal{L}_{\text{bce}} = -\frac{1}{N}\sum_{i=1}^{N}(y_b^i \log(\hat{y}_b^i) + (1-y_b^i)\log(1-\hat{y}_b^i)), \tag{2.7}$$

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2\sum_{i=1}^{N} y_b^i \hat{y}_b^i}{\sum_{i=1}^{N} y_b^i + \sum_{i=1}^{N} \hat{y}_b^i}, \tag{2.8}$$

其中 y_b^i 表示包含前背景信息的标注，$\hat{y}_b^i$ 表示相应的二元前背景概率预测。参考[56] 的实验经验，我们为三项损失函数设置如下权重系数：$\lambda_{\text{cls}} = 2.0$，$\lambda_{\text{bce}} = 5.0$，$\lambda_{\text{dice}} = 5.0$。网络的总体损失函数表示为：

$$\mathcal{L} = \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}}. \tag{2.9}$$

2.3　方法验证与实验结果分析

在本节中，首先对网络评估所使用的公开、自建数据集与所使用的评价指标进行简要说明，并介绍网络的实现细节，包括训练策略以及超参数设置，之后将

展示验证所提出网络各个组件部分有效性的消融实验，然后展示与现有代表性的单模态及异构特征融合策略进行性能对比以及结果分析。

2.3.1　数据集与评价指标

SYN-UDTIRI 数据集　针对当前道路场景解析领域缺乏专门的大规模标注数据集这一问题，我们基于 CARLA 仿真器[82] 构建了一个创新性的合成数据集——SYN-UDTIRI。相比于一般的合成数据集，该数据集的独特之处在于融合了道路破损感知领域内的三维重建算法[12, 92] 所获取的真实路面破损数字孪生模型。此外，为了更真实地还原道路表面特征，我们采用随机 Perlin 噪声[93] 模拟了实际道路的复杂纹理材质。考虑到环境因素对道路场景感知的显著影响，我们同样模拟了多样化的光照和天气条件，包括晴天（白天、黄昏、夜晚）和雨天（白天、黄昏、夜晚）等不同场景。数据采集采用了模拟立体相机系统，相机基线距离设置为 0.5 米，并搭载于移动车辆上。最终采集了 10407 对高质量立体图像，分辨率为 $720 \times 1,280$ 像素。除 RGB 图像外，数据集还包含了深度图像、视差图像、法向量图像以及像素级语义标注。语义标注涵盖三个核心类别：可行驶区域、道路破损和其他物体。

KITTI Road 数据集　KITTI Road 数据集[20] 提供了 289 对用于训练和验证的立体图像及其配套的 LiDAR 点云数据，同时包含 289 对用于测试的无标注数据。本章采用了与[16] 一致的数据预处理方法。为了在 KITTI Road 基准测试中提交测试结果，我们在自行划分的训练集上对提出的信息融合场景解析网络进行了针对性微调。为增强模型的泛化能力，我们实施了全面的数据增强策略，包括随机裁剪、水平翻转、光照调整等技术，以应对复杂多变的道路场景。

CityScapes 数据集　CityScapes 数据集[22] 是城市场景语义分割领域的标准数据集之一，包含 2,975 张训练用立体图像和 500 张验证用图像，均具有精细的语义标注。鉴于 KITTI Road 数据集规模相对有限，我们选择在 CityScapes 数据集上开展补充实验，以验证 RoadFormer 在大规模数据集上的性能表现。所有实验评估均基于验证集进行，因测试集不提供真值标注。测试集的性能评估需通过 CityScapes 在线基准测试平台提交结果获取。考虑到本文重点关注道路场景解析，我们对数据集进行了重构，将类别简化为道路和其他两类。值得注意的是，数据集的表面法向量图像是通过两步处理获得的：首先使用在 KITTI 数据集[94] 上训练的 RAFT-Stereo[95] 获取深度图像，然后利用法向量估计器进行推理计算。

ORFD 数据集　ORFD 数据集[17] 是专门面向越野可行驶区域检测设计的综合数据集，包含 12,198 张 RGB 图像及其对应的 LiDAR 点云数据。该数据集的采集场景丰富多样，涵盖了不同地形、天气条件和光照环境，为评估模型在实际应用场景中的鲁棒性提供了可靠基础。本章严格遵循[17] 中规定的数据划分和预处理方法（表面法向量估计除外）开展实验。

2.3.2　实验设置与评估指标

为全面评估所提出网络 RoadFormer 的性能，我们设计了系统的对比实验，将 RoadFormer 与四种单模态网络和五种双源信息融合网络进行对比。其中，单模态网络仅使用 RGB 图像进行训练，而双源信息融合网络同时利用 RGB 图像和法向量图像（采用研究[83] 中提出的法向量估计器利用深度图像进行估计）。为确保实验的公平性，所有网络均训练相同轮次。在训练策略方面，我们采用 AdamW 优化器[96]，并结合多项式学习率衰减策略[97]。具体参数设置如下：初始学习率为 10^{-4}，权重衰减系数为 5×10^{-2}。在对于 ConvNeXt[85] 和 Swin Transformer[48] 作为并行式编码器骨干网络的训练中，我们应用了 10^{-1} 的学习率乘子。此外，为提升模型训练的稳定性和鲁棒性，我们还引入了梯度裁剪机制。

所采用的评价指标　在 SYN-UDTIRI、CityScapes 和 ORFD 数据集上，我们采用了五种在领域内被广泛应用的定量指标来全面衡量模型性能：准确率（Accuracy，Acc）、精确率（Precision，Pre）、召回率（Recall，Rec）、交并比（Intersection over Union，IoU）和 F 分数（F-score，Fsc）。这些评估指标从不同角度刻画了模型的性能表现，其数学定义如下：

$$
\begin{aligned}
\mathrm{Acc} &= \frac{n_{TP} + n_{TN}}{n_{TP} + n_{TN} + n_{FP} + n_{FN}}, \\
\mathrm{IoU} &= \frac{n_{TP}}{n_{TP} + n_{FN} + n_{FP}}, \\
\mathrm{Pre} &= \frac{n_{TP}}{n_{TP} + n_{FP}}, \\
\mathrm{Rec} &= \frac{n_{TP}}{n_{TP} + n_{FN}}, \\
\mathrm{Fsc} &= \frac{2 \times n_{TP}}{2 \times n_{TP} + n_{FN} + n_{FP}},
\end{aligned}
$$

在上述公式中，各统计量的定义如下：n_{TP}（True Positive）：被正确识别为目标类别的像素数量；n_{TN}（True Negative）：被正确识别为非目标类别的像素数量；n_{FP}（False Positive）：被错误识别为目标类别的像素数量；n_{FN}（False Negative）：被错误识别为非目标类别的像素数量。

除了上述性能指标，我们对网络多类别场景解析任务各个语义类别平均交并比（mIoU）以评估网络的场景综合解析能力，同时我们引入了每秒运行帧（Frames Per Second，FPS）这一广泛使用的实时性指标对所提出 RoadFormer 的推理速度进行评估。需要特别说明的是，对于 KITTI Road 基准测试结果，考虑到对比的公正性，本章节采用了其官方规定的评估指标体系，包括最大 F 分数（MaxF）、平均精确度（AP）、精确度（Pre）以及召回率（Rec）, 相关指标的详细计算方法可在官方网站（cvlibs.net/datasets/kitti/eval_road.php）查阅。

2.3.3　网络自身组件的消融实验

尽管当前场景解析领域的主流研究（如[56] 和[98]）普遍采用 Swin Transformer 作为编码器并与 Transformer 解码器相结合，但我们认为这种架构可能并非所有场景解析任务的最优选择。为验证这一假设，我们在 SYN-UDTIRI 和 CityScapes 数据集上开展了系统的消融实验，选取了与 Swin Transformer 同期提出的 CNN 架构 ConvNeXt[85] 作为对照组，综合评估了两种编码器骨干网络与其他网络组件的协同性能。如表 2.1所示的实验结果有力地证实了我们的假设：ConvNeXt 在整体性能上显著优于 Swin Transformer。具体而言，在 IoU 指标上，ConvNeXt 相较于 Swin Transformer 实现了 0.20% 至 1.11% 的提升；在 F-score 指标上也取得了 0.11% 至 0.59% 的优势。这一结果不仅验证了我们提出的 RoadFormer 具有良好

表 2.1　RoadFormer 在 CNN 与 Transformer 架构骨干网络选择上进行的消融实验

数据集	骨干网络	IoU (%)	Fsc (%)	Pre (%)	Rec (%)
SYN-UDTIRI	ConvNeXt	**93.38**	**96.58**	96.58	**96.74**
	Swin	93.18	96.47	**96.59**	96.35
CityScapes	ConvNeXt	**95.80**	**97.86**	**97.74**	**97.97**
	Swin	94.69	97.27	97.25	97.29

表 2.2　对于 RoadFormer 中的异构特征融合模块 HFFM 与 FFRM 有效性与网络推理速度的消融实验

SEB	HFFM	FFRM	IoU (%)	Acc (%)	Fsc (%)	Pre (%)	Rec (%)	FPS
×	×	×	95.11	96.87	97.50	98.12	96.87	21.80
✓	×	×	95.45	97.40	97.67	**97.95**	97.40	21.60
×	×	✓	95.49	97.59	97.69	97.79	97.59	21.60
×	✓	×	95.34	97.67	97.61	97.55	97.67	20.50
×	✓	✓	**95.80**	**97.97**	**97.86**	97.74	**97.97**	20.10

的架构兼容性，能够同时适配基于 CNN 和 Transformer 的骨干网络，更表明在道路场景解析这一特定任务中，ConvNeXt 展现出了更为突出的性能优势。基于这一发现，我们在后续实验中选定 ConvNeXt 作为默认骨干网络。针对网络中关键模块的有效性，我们进一步设计了详细的对比实验，重点考察了 FFRM、HFFM 以及 SEB 三个核心模块的性能贡献。

另一方面，表 2.2所示的异构特征融合模块有效性消融实验数据揭示了以下关键发现：

(1) FFRM 相较于 SEB 表现出更优的性能，在 IoU 指标上实现了 0.04% 的提升；
(2) HFFM 模块相对于基准配置提升了 0.23% 的 IoU 性能；
(3) 当 HFFM 和 FFRM 模块协同工作时，模型性能获得了显著提升，远超过单独使用任一模块的效果。

这些实验结果不仅验证了各个模块的有效性，更证实了它们之间存在显著的协同效应。尤其值得注意的是，尽管 HFSB 模块由于其多尺度重校准和异构特征融合机制的复杂性导致了轻微的计算开销增加，但在实际应用中，我们的网络在配备 NVIDIA RTX 3090 GPU 的环境下处理 352×640 像素分辨率的图像时，仍然保持了约 20 帧/秒的处理速度，完全满足实时应用的性能要求。这一结果表明，我们提出的方法在性能和效率之间取得了很好的平衡。

2.3.4　与代表性方法的对比实验

本节通过在四个具有代表性的数据集（SYN-UDTIRI、CityScapes、ORFD 和 KITTI Road）上进行系统性实验，全面评估了所提出的 RoadFormer 模型的性能。定量实验结果如表 2.5- 2.6所示，定性分析结果如图 2.7- 2.9所示。综合实验结果表明，RoadFormer 在所有测试数据集上均实现了优于现有最先进方法的性能，充分验证了该模型在处理多样化道路场景（包括带有缺陷的合成道路、城市道路和乡村道路）时的卓越性能和鲁棒性。以下将详细分析各数据集上的实验结果：

SYN-UDTIRI　在 SYN-UDTIRI 数据集上，本章节提出的 RoadFormer 网络展现出了显著的性能优势。在验证集上，RoadFormer 的 IoU 达到 93.35%，F-score 达 96.56%，明显优于其他对比方法。其中，SNE-RoadSeg 作为第二好的方法，其 IoU 为 92.00%，F-score 为 95.80%。而性能相对较弱的 FuseNet 虽然在召回率上达到了 97.80% 的高分，但其精确率仅为 68.30%，说明该模型倾向于过度预测道路区域。在测试集上，RoadFormer 同样保持了出色的性能，IoU 和 F-score 分别达到 93.51% 和 96.65%，验证了模型优异的泛化能力。通过对比单模态和信息融合网络的性能差异，我们发现在道路缺陷检测任务中，信息融合网络普遍表现出

表 2.3　在 SYN-UDTIRI 数据集上与代表性场景解析网络的定量对比实验结果，其中信息融合网络都以 RGB-法向量图像对作为输入

	子集	网络名称	IoU (%)	Fsc (%)	Pre (%)	Rec (%)
单模态网络	验证集	Mask2Former	64.29	78.27	83.0	74.05
		SegFormer	52.46	68.82	70.13	67.55
		DeepLabv3+	52.94	69.23	75.23	64.12
		HRNet	52.92	69.21	79.46	61.30
	测试集	Mask2Former	46.91	63.87	73.59	56.41
		SegFormer	36.34	53.31	57.23	49.89
		DeepLabv3+	34.76	51.58	62.54	43.90
		HRNet	35.47	52.37	69.09	42.16
信息融合网络	验证集	FuseNet	67.30	80.40	68.30	**97.80**
		SNE-RoadSeg	92.00	95.80	96.30	95.40
		RTFNet	90.30	94.90	94.10	95.70
		OFF-Net	83.90	91.30	91.70	90.80
		MFNet	89.50	94.50	95.70	93.30
		RoadFormer (本章)	**93.35**	**96.56**	**96.53**	96.59
	测试集	FuseNet	70.70	82.90	72.10	**97.50**
		SNE-RoadSeg	92.10	95.90	**96.70**	95.10
		RTFNet	90.50	95.00	95.50	94.50
		OFF-Net	83.80	91.20	91.90	90.50
		MFNet	87.70	93.50	96.20	90.90
		RoadFormer (本章)	**93.51**	**96.65**	96.61	96.69

更强的鲁棒性。以 Mask2Former 为代表的单模态网络在验证集上取得了 64.29% 的 IoU，但在测试集上性能显著下降至 46.91%。相比之下，使用了表面法向量图像的信息融合网络性能更加稳定，具有较强的精确性与鲁棒性。这种结果是因为法向量图像中包含的平面几何特征能够帮助模型更好地理解道路表面的结构，从而提高对缺陷的检测能力。

CityScapes　在 CityScapes 数据集上的实验结果 2.4展示了一些较为反常的现象。单模态网络表现出较高的稳定性，其中 HRNet 达到了 94.06% 的 IoU，SegFormer、Mask2Former 和 DeepLabv3+ 的性能也非常接近，IoU 指标的波动范围仅为 0.2%。这种稳定性可能得益于这些网络强大的特征提取能力和在大规模数据集上的充分训练。然而，出人意料的是，除 RTFNet 和本章节提出的 RoadFormer 外，其他信息融合网络的性能普遍低于单模态网络。OFF-Net 的性能最弱，IoU 仅为

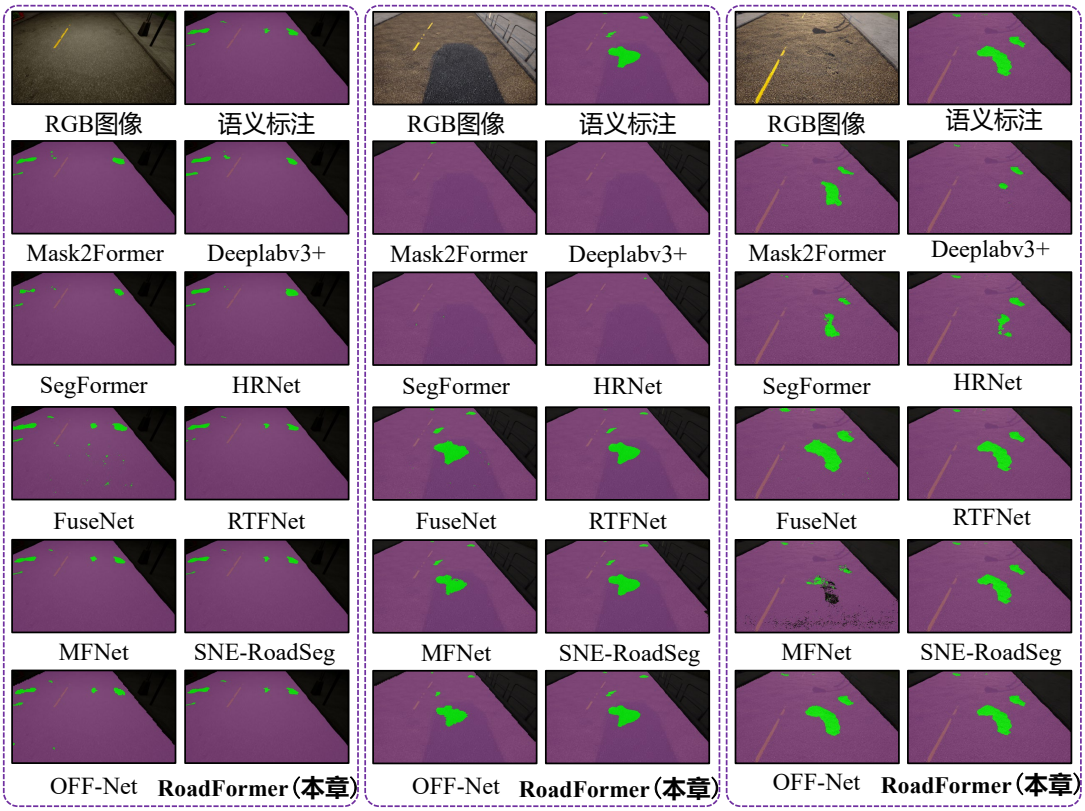


图 2.7　在 SYN-UDTIRI 数据集上与代表性场景解析网络的定性对比实验结果。可行驶区域、道路破损以及其他背景区域分别表示为紫色、绿色以及黑色

表 2.4　在 CityScapes 数据集上与代表性场景解析网络的定量对比实验结果，其中信息融合网络都以 RGB-法向量图像对作为输入

	网络名称	IoU (%)	Fsc (%)	Pre (%)	Rec (%)	mIoU (%)
单模态网络	Mask2Former	93.84	96.82	97.14	96.51	74.80
	SegFormer	93.98	96.90	96.02	97.79	64.51
	DeepLabv3+	93.82	96.81	96.99	96.63	68.66
	HRNet	94.06	96.94	96.29	97.59	70.10
信息融合网络	FuseNet	91.60	95.60	96.00	95.30	52.70
	SNE-RoadSeg	93.80	96.80	96.10	97.50	53.40
	RTFNet	94.10	96.90	96.30	97.60	49.60
	OFF-Net	89.60	94.50	93.40	95.70	39.20
	MFNet	92.10	95.90	94.10	97.70	49.30
	RoadFormer (本章)	**95.80**	**97.86**	**97.74**	**97.97**	**76.20**

89.60%，这可能是由于其简单的特征融合策略难以有效处理来自不同信息的异构特征。MFNet 和 FuseNet 的性能相对较好，但仍未能超越单模态方法。这一现象的主要原因可能在于用于估计表面法向量的视差图的质量问题。由于这些视差图像是直接从预训练的立体匹配网络获得的，其准确性可能受到多种因素的影响。而 RoadFormer 凭借创新的异构特征融合模块，成功克服了这一挑战。能

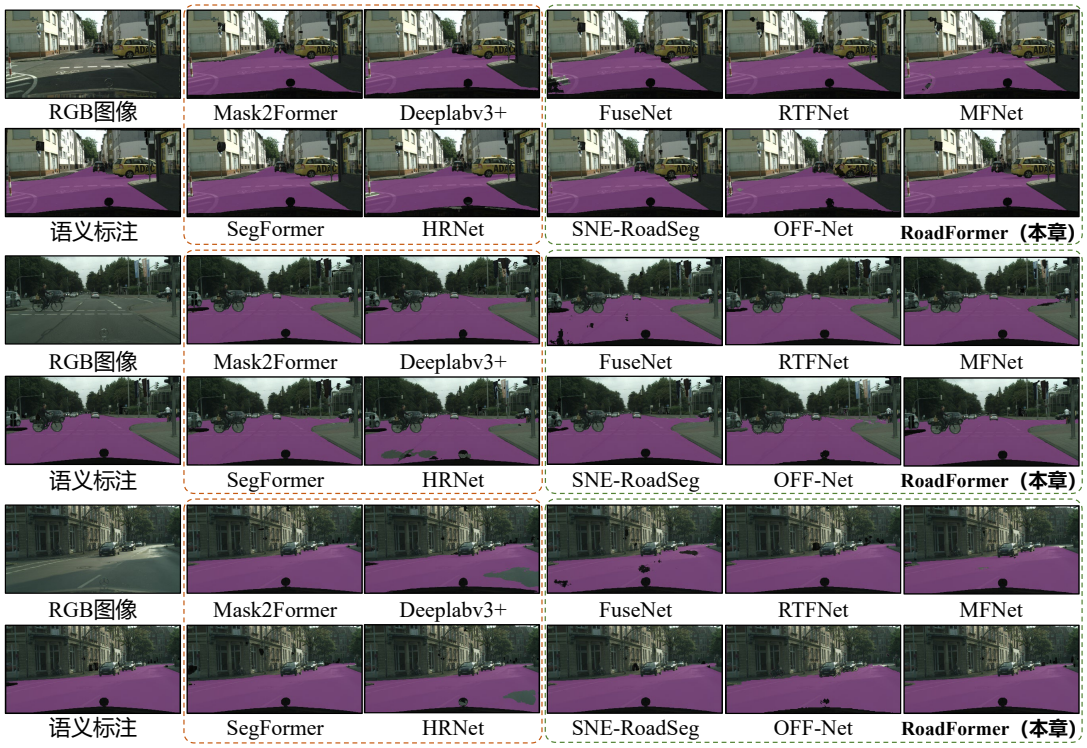


图 2.8　在 CityScapes 数据集上与代表性场景解析网络的定性对比实验结果。可行驶区域与无标签区域在图中分别表示为紫色与黑色

表 2.5　在 ORFD 数据集上与其他代表性信息融合场景解析网络的定量对比实验结果。我们同时汇报了在其原始论文[17] 中的结果与我们重新实验的结果。† 代表在原始论文实验中使用了 RGB-深度图像对作为网络输入，其余网络以 RGB-法向量图像对作为输入

	网络名称	IoU (%)	Fsc (%)	Pre (%)	Rec (%)
原始结果	FuseNet†	66.00	79.50	74.50	85.20
	SNE-RoadSeg	81.20	89.60	86.70	92.70
	OFF-Net	82.30	90.30	86.60	94.30
重新实验结果	FuseNet	59.00	74.20	59.30	**99.10**
	SNE-RoadSeg	79.50	88.60	90.30	86.90
	RTFNet	90.70	95.10	93.80	96.50
	OFF-Net	81.80	90.00	84.20	96.70
	MFNet	81.70	89.90	89.60	90.30
	RoadFormer (本章)	**92.51**	**96.11**	**95.08**	97.17

够自适应地融合和重校准不同模态的特征，使网络在所有评估指标上都取得了最优成绩。特别是在 20 类别的场景解析任务中，RoadFormer 获得了 76.20% 的平均交并比 (mIoU)，显著优于其他方法，充分证明了其在复杂城市场景中的优越性能。

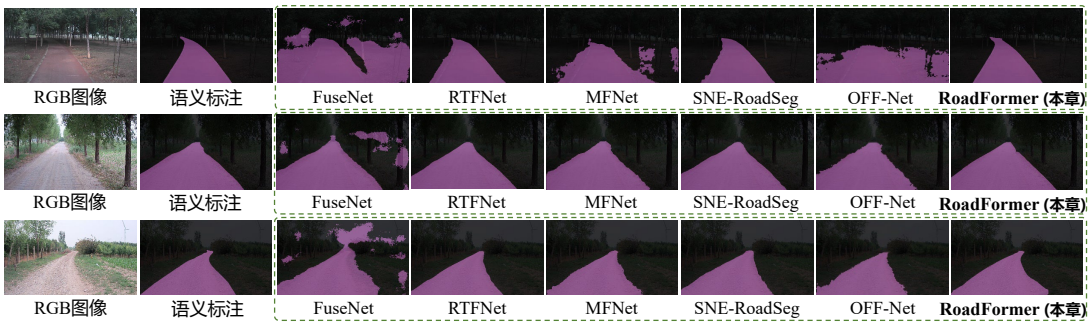


图 2.9　在 ORFD 数据集上与其他代表性信息融合场景解析网络的定性对比实验结果。可行驶区域与背景区域在图中分别表示为紫色与黑色

ORFD　为进行公平的性能比较，我们同时考察了原始论文中报告的结果和重新实现的结果。如表 2.6中所示，在我们重新实现的实验中，RoadFormer 以 92.51% 的 IoU 和 96.11% 的 F-score 领先于所有对比方法。RTFNet 作为第二好的方法，达到了 90.70% 的 IoU 和 95.10% 的 F-score。而 FuseNet 的表现最不理想，尽管其召回率高达 99.10%，但 IoU 仅为 59.00%，这表明该模型在越野场景中存在一定的偏见问题，易出现较为激进的预测。值得注意的是，重新实现的结果与原始论文报告的结果存在一定差异。例如，OFF-Net 在原始论文中报告的 IoU 为 82.30%，而在我们的重新实现中为 81.80%。这种差异可能来源于实现细节的不同，如网络参数的初始化策略、优化器的选择等。此外，越野场景下语义标注的准确性也是一个重要因素。复杂的地形条件和光照变化可能导致标注存在一定的主观性，这种不确定性可能会影响模型的训练效果。

KITTI Road　如表 2.6所示的 KITTI Road 基准测试结果，本章节提出的 RoadFormer 同样展现出了强大的竞争力。模型在 MaxF 指标上达到 97.50%，略高于 SNE-RoadSeg+，但在召回率上表现更优（97.84% 对 97.58%）。PLB-RD 虽然在

表 2.6　在 KITTI Road 数据集上与现有的代表性高精度场景解析网络的定量对比实验结果

网络名称	MaxF (%)	AP (%)	Pre (%)	Rec (%)	排名
NIM-RTFNet[75]	96.02	94.01	96.43	95.62	13
HID-LS[99]	93.11	87.33	92.52	93.71	33
LC-CRF[71]	95.68	88.34	93.62	97.83	15
SNE-RoadSeg[16]	96.75	94.07	96.90	96.61	8
SNE-RoadSeg+[18]	97.50	93.98	**97.41**	97.58	2
PLB-RD[74]	97.42	**94.09**	97.30	97.54	3
LRDNet+[69]	96.95	92.22	96.88	97.02	4
DFM-RTFNet[6]	96.78	94.05	96.62	96.93	7
RoadFormer (本章)	**97.50**	93.85	97.16	**97.84**	**1**

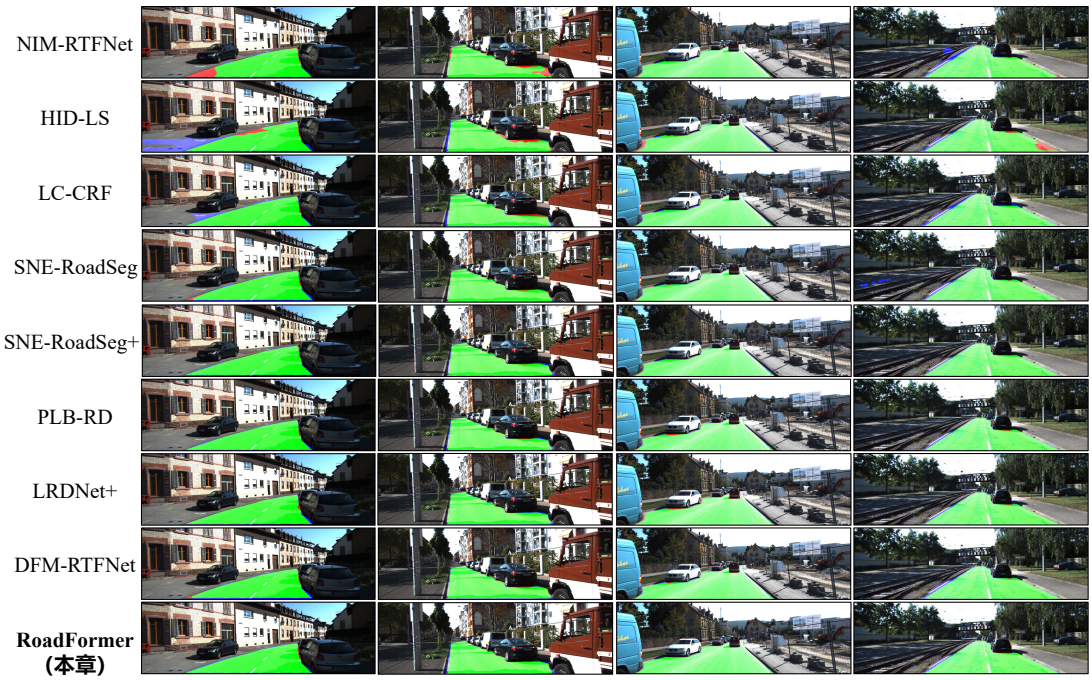


图 2.10　在 KITTI Road 测试集上与现有的代表性高精度场景解析网络的定量对比实验结果。测试结果来自 KITTI 官方评测基准服务器。TP，FP 和 FN 预测结果在图中分别表示为绿色、蓝色和红色

AP 指标上略占优势（94.09% 对 93.85%），但其整体性能略逊于 RoadFormer。而较早提出的 HID-LS 方法表现相对较弱，其 AP 仅为 87.33%，说明近年来深度学习方法在可行驶区域检测任务上取得了显著进展。此外，RoadFormer 优异的召回率指标（97.84%）特别值得关注，这表明网络能够高度完整地识别道路区域，很少出现漏检情况。这一特性对于自动驾驶等实际应用场景极其重要，因为漏检道路区域可能导致严重的安全问题。同时，网络也保持了较高的精确率（97.16%），说明其很好地平衡了预测的精确性与全面性，这种平衡对于自动驾驶等实际应用具有重要意义。

2.4　本章小结

本章提出了一种专为道路场景解析任务设计的高性能双源信息融合网络。该网络由一个并行编码器、一个新颖的异构特征同步模块以及基于 Transformer 架构的掩膜分类解码器组成。相较于领域内的现有研究，本章节提出的网络 RoadFormer 展现出更强大的异构特征融合能力，显著提升了场景解析任务的准确性。在我们自主构建的 SYN-UDTIRI 数据集和三个权威公开数据集上，RoadFormer 全面超越了现有的单模态及信息融合网络，并在业界权威的 KITTI Road 数据基准测试中位居榜首。作为所提出网络的关键组件，基于 Transformer 及其注意力机制的特征融合与解码器架构，相较于传统 CNN，在道路场景解析任务中展现

出显著的性能优势。我们期望在未来将这一方法推广到更广泛的场景理解任务中。本章也揭示了信息选择与任务特性的内在关联：当目标对象具有显著几何特征（如道路可行驶区域检测、路面破损识别等任务）时，以 RGB-法向量图像对的作为双源信息输入能带来最显著的性能提升，而面对通用类别的解析需求时（见本章在 Cityscapes 数据集上的 mIoU 指标 2.4），该组合的优势相对于其他网络则不明显，这提供了一项重要准则：在场景解析领域，信息的输入选择与任务特性保持高度适配。 尽管在准确性方面取得了显著突破，但本章节提出的 RoadFormer 由于采用了并行编码器和多尺度特征融合模块，其实时推理性能仍相对有限，目前仅能达到约 20 帧/秒。提升推理速度将是我们后续研究的重要方向之一，以进一步增强网络在应用场景中的实用性。

第 3 章　基于视觉基础模型的双源信息融合道路场景解析网络

在上一章中，我们提出了一种基于传统并行编码器架构的双源信息（异构特征）融合道路场景解析网络。该网络通过对称的双路传统视觉编码器分别对双源信息进行编码以提取多尺度特征，并深入探究了 Transformer 架构及注意力机制在异构特征融合和场景解析解码过程中的有效性，最终实现了道路场景解析性能的显著提升。然而，仅使用基于 ImageNet 数据集[100] 预训练的传统单模态视觉编码器来提取异构特征，对异构模态本身的特性及优势挖掘仍显不足。若能在特征编码阶段更深入地挖掘双源信息的潜力并增强其特征的表征能力，将能为后续的特征融合及解码阶段提供更丰富的场景先验信息。本章以此为目标，针对不同异构模态的独特优势设计了非对称的编码器架构。其中，我们创新性地引入了先进的视觉基础模型（Vision Foundation Model，VFM），以进一步挖掘双源信息的内在潜力，从而提取蕴含更丰富语义信息的通用性特征。同时，我们提出了一种渐进式的双向多尺度特征融合策略，在特征编码阶段实现了更有效的异构特征提取。基于本章所提出的非对称、渐进式异构特征编码架构，我们构建了双源信息融合道路场景解析网络 HAPNet。该网络在多个包含不同多源信息的道路场景解析数据集上取得了极具竞争力的性能，同时展现出了对不同多源信息的广泛适用性。

3.1　引言

如前一章所述，多源信息（异构特征）的融合能够有效提升道路场景解析任务的性能。其中，RGB 图像提供了场景丰富的色彩与纹理特征，而其他模态的图像（在此统称为"X" 模态，如深度图、热图像等）则反映了场景中的空间几何结构以及显著物体的轮廓特征。这些异构特征从不同维度刻画了场景特征，相互补充、互为佐证。因此，为充分发掘多源信息的潜力，提升特征表示能力，我们需要设计更加适配不同多源信息特点的特征编码与融合架构。然而，在当前道路场景解析领域中，现有的信息融合网络往往采用如[48, 63, 85] 等传统骨干网络构建对称的并行编码器，对 RGB 图像与深度图像、热图像等多源信息的编码流程完全相同。此外，这些编码器仅基于 ImageNet 数据集[100] 进行有监督预训练，之后在特征融合阶段对不同尺度的异构特征使用简单的单尺度融合策略。这种固化的编码范式难以充分挖掘和利用不同多源信息或模态的独特优势，制约了网络

的任务性能。

现有的异构特征编码网络，如并行 ResNet[63]、Swin Transformer[48] 以及 ConvNeXt[85] 等，通常采用有监督的预训练范式。其中，主流的监督预训练方法依赖于具有分类标注的 ImageNet 数据集[100]。尽管 ImageNet 中的数据在一定程度上能够拟合真实世界中的物体语义分布，但其百万级别的数据规模仍然相对有限[101]。传统监督预训练范式对数据标注的依赖性，以及高昂的标注成本，限制了这些网络对互联网上可获取的千万级乃至亿级数据的充分利用。在此背景下，基于自监督学习的视觉基础模型应运而生，如 BEiT 系列[101, 102] 和 DINO 系列[103] 等。这些模型通过利用更大规模的训练数据，在单模态视觉领域的下游任务中取得了卓越性能。然而，这些基础模型在多源信息融合场景解析领域的有效性尚待探索。

此外，现有网络普遍采用对称的同架构骨干网络进行异构特征提取。然而，这种对称架构可能并非异构特征提取的最优选择，这一限制主要源于两个方面：一是 RGB 图像与 X 模态图像之间存在的固有差异[104]，二是 Transformer 与 CNN 等网络架构各自具有的独特优势。具体而言，RGB 图像通常包含丰富的场景色彩与纹理特征，能够提供更完整的全局语义表征[105]，这种全局特征更适合使用基于全局注意力机制的 ViT 架构进行提取；另一方面，深度图像、热图像等（统称为 X 模态）则包含更多场景的局部细节特征[106]（我们称之为局部语义），如物体轮廓、边缘和几何结构等，这些特征通过 CNN 架构能够实现更高效的提取。因此，如何通过合理结合 ViT 与 CNN 架构，设计非对称的双路异构特征提取网络，以更好地发挥多源信息在场景解析任务中的潜力，是一个亟待解决的关键问题。

除了异构特征的编码架构设计，异构特征的融合策略对场景解析任务的性能同样具有重要影响。如前章节 2所述，现有网络普遍采用的异构特征融合范式是对编码器输出的特征在各个尺度上分别进行融合[57, 58]。这种方法仅在编码器输出多尺度特征后对已编码特征进行后处理式融合，且忽略了不同尺度特征之间的语义交互。这种融合策略存在两个主要限制：一方面难以实现深层次的异构特征交互与融合；另一方面，由于不同尺度的异构特征可能存在语义差异，容易产生矛盾的融合特征，从而导致错误的预测结果[84, 107]。以较早期的 MFNet[57] 为例，该网络仅通过对并行式编码器的输出层特征进行简单的元素级相加来实现特征融合。这种" 硬编码" 式的策略完全忽视了 RGB 图像与其他以热图像、深度图像为代表的多源信息之间的内在差异[106]，严重制约了网络的性能上限。因此，如何实现更深层次的异构特征融合，并增强多尺度异构特征之间的语义交互，成为了当前领域亟待解决的另一个关键问题。

针对上述问题，本章首先探究了视觉基础模型在双源信息融合网络中的应

用范式，以更有效地发挥自监督预训练范式与大规模预训练数据的潜力。同时，基于对 RGB 图像与“X”模态图像内在差异的深入分析，我们基于 CNN 架构设计了一种创新性的跨模态空间先验描述子（cross-modal spatial prior descriptor，CSPD），该描述子能够从 RGB 图像与 X 模态图像中鲁棒地提取跨模态的空间先验，该空间先验中蕴含丰富的局部语义。我们将 CSPD 与 VFM 以非对称的方式进行组合，构建了一种新颖的混合架构编码器。在编码过程中，我们提出了渐进式异构特征融合模块（progressive heterogeneous feature integrator，PHFI），实现了全局上下文特征与空间先验的双向交互。这种双向、渐进式的融合策略能够实现异构特征的深层次融合。此外，我们还引入了一个辅助任务来增强融合后特征的局部语义表达，进一步提升了网络在场景解析任务上的表现。总体而言，本章的主要贡献可概括为以下四个方面：

(1) 系统探究了视觉基础模型在双源信息融合道路场景任务中的有效性及其应用范式，通过利用更大规模的预训练数据，实现了更具通用性的场景异构特征提取；
(2) 通过深入分析 RGB 图像与 X 模态图像本质特性的差异，设计了一种更适配的非对称、渐进式异构特征提取与融合网络架构；
(3) 创新性地提出了一种用于增强特征局部语义的辅助任务，并通过系统实验验证了该任务的有效性；
(4) 基于上述编码器架构与辅助任务构建的道路场景解析网络 HAPNet，在多个公开的 RGB-T 与 RGB-D 数据集上都实现了同期领先的性能表现。

3.2　基于 ViT 架构的视觉基础模型

在计算机视觉领域中，Transformer 架构编码器的应用范式主要可分为两类：(1) 原始 ViT 架构[76]；(2) 基于 ViT 模块构建的多尺度变体[48, 108, 109]。普遍而言，后者在视觉相关的密集预测任务（场景解析、深度估计、立体匹配等）中表现更为出色，这主要得益于这些变体引入了更多针对视觉任务优化的架构设计，如 CNN 中的局部空间操作等。这种局部的归纳偏置在视觉任务，尤其是稠密预测任务中具有显著优势[25]。然而，原始 ViT 架构仍具有其独特的优势，这在多模态预训练[110-112] 中体现得尤为明显。与自然语言处理领域中的原始 Transformer 相似，简朴的 ViT 架构对输入数据不施加特定的偏置假设。通过配备不同类型的嵌入层，如图像块嵌入[76]、3D 图像块嵌入[113] 和词元嵌入[23]，模型能够充分利用图像、视频和文本等多模态数据进行预训练，从而习得更为丰富的表示。在此背景下，研究界开始深入探索基于原始 ViT 架构的视觉基础模型 VFM。基础模型是指能够从海量数据中提取知识并迁移至各类下游任务的通用模型。这一概念

最初由斯坦福大学基础模型研究中心（CRFM）于 2021 年提出[114]。这类模型通常采用自监督学习技术，在海量数据上进行训练。通过更大规模的图像与其他模态数据的训练，神经网络能够学习到更具普适性的表征和能力，从而提升下游视觉任务的性能。本章节主要聚焦于可作为视觉编码器的 VFM，暂不讨论针对特定下游任务的基础模型，如用于语义分割任务的 Segment Anything（SAM）系列模型[115, 116] 和用于深度估计任务的 Depth Anything[117] 系列模型。

目前，VFM 的训练范式主要可分为两类：基于掩码图像建模（Masked Image Modelling，MIM）的自监督方法和基于判别式对比学习（Discriminative Contrastive Learning，DCL）的方法。其中，MIM 方法以 BEiT 系列[101, 102] 为代表，而 DCL 方法则以 DINO 系列[103, 118] 为典型代表。

3.2.1　掩码图像建模

BEiT（Bidirectional Encoder representation from Image Transformers）[101] 系列算法提出了一种基于掩码图像建模的自监督训练方法，用于训练 ViT 网络架构。如图 3.1所示，BEiT 的架构主要包含两个核心组件：(1) 基于 ViT 架构的 BEiT 编码器；(2) 离散变分自编码器（discrete variational autoencoder，dVAE）[119]。由于 BEiT 网络结构本身（ViT）已在章节 1.4.1.1中介绍，因此这里不再展开说明，而着重于其自监督预训练策略 dVAE 本质上与 VAE[120] 类似，同样作为图像块编码的关键组件。在 BEiT 预训练之前，需要首先训练一个 dVAE。该过程可概括为：(1) 通过离散变分编码器（Tokenizer）将图像处理为 N 个离散的视觉标记，其中 N 与 ViT 中特征嵌入层得到的序列长度相同；(2) 通过解码器重建原始图像。每个视觉标记是一个取值范围为 $[1, V]$ 的整数，类似于 NLP 领域中的词汇表 $\mathcal{V}$。其

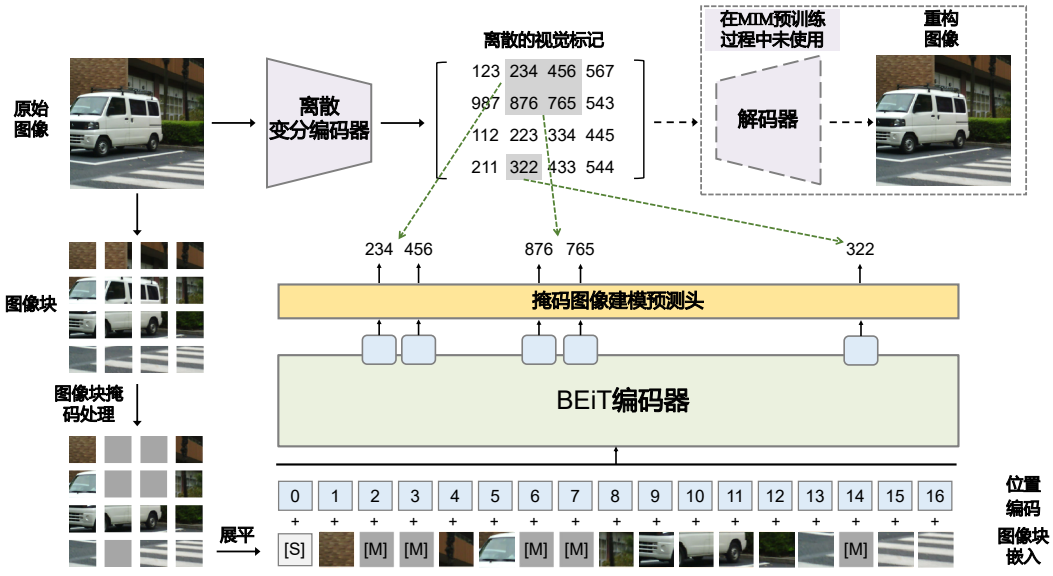


图 3.1　BEiT 视觉基础模型预训练流程示意图

中 $\mathcal{V}=1,\ldots,V$ 包含离散视觉标记的索引，每个 16×16 的图像块经过 Tokenizer 都被映射为词汇表 $\mathcal{V}$ 中的一个词。

在完成 dVAE 的训练后，进入 BEiT 的预训练阶段，具体包含以下步骤：

(1) 给定输入图像 $\boldsymbol{I}$；

(2) 通过 ViT 中的特征嵌入层将 $\boldsymbol{I}$ 分割为图像块嵌入 $\{\boldsymbol{e}_i^p\}_{i=1}^N$；

(3) 通过离散变分编码器将 $\boldsymbol{I}$ 转换为离散视觉标记 $\{\boldsymbol{z}_i\}_{i=1}^N$；

(4) 随机掩码 40% 的图像块嵌入，被掩码的位置集合表示为 $\mathcal{M}\in\{1,\ldots,n\}^{0.4N}$；

(5) 将被掩码的图像块嵌入替换为可学习的嵌入向量 $\boldsymbol{e}_{[m]}\in\mathbb{R}^D$。

此时，输入的图像块嵌入序列可表示为：$\boldsymbol{e}^{\mathcal{M}}=\{\boldsymbol{e}_i^p:i\notin\mathcal{M}\}_{i=1}^N\cup\{\boldsymbol{e}_{[m]}:m\in\mathcal{M}\}_{i=1}^N$。随后，将序列 $\boldsymbol{e}^{\mathcal{M}}$ 输入到 L 层 BEiT 编码器中，得到包含图像全局信息的特征编码表示$\{\boldsymbol{h}_i^L\}_{i=1}^N$。对于掩码位置的编码输出 $\{\boldsymbol{h}_i^L:i\in\mathcal{M}\}_{i=1}^N$，通过预测头进行特征重建，训练目标是最小化这些位置的编码输出与对应的离散视觉标记之间的差异。

3.2.2　判别式对比学习

DINO（DIstillation with NO labels）[118] 提出了一种创新的自监督蒸馏训练方法。如图 3.2所示，该方法的核心在于构建了一个教师网络 $\boldsymbol{g}_{\theta_t}$ 和一个学生网络 $\boldsymbol{g}_{\theta_s}$，两者均采用相同的 ViT 架构。在训练过程中，DINO 首先对输入图像 $\boldsymbol{I}$ 生成多个不同视角（Views）构成的集合 $\mathcal{V}$。该集合包含两种类型的视角：两个不同的全局视角 $\boldsymbol{x}_1^g$ 和 $\boldsymbol{x}_2^g$，以及若干个通过图像裁剪得到的局部视角。训练时，将局部视角输入学生网络，同时将全局视角输入教师网络，这种设计有助于模型学习图像的全局上下文特征。DINO 的训练目标是最小化教师网络和学生网络输出之

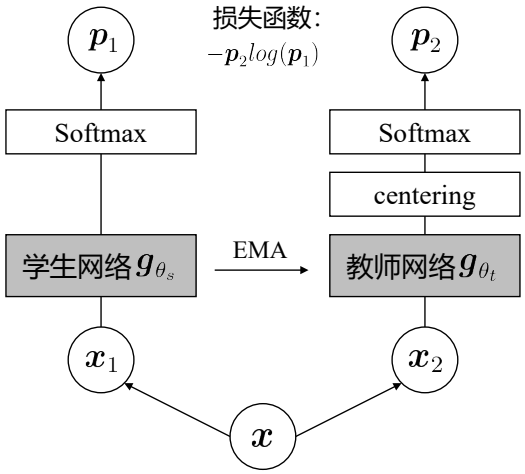


图 3.2　DINO 视觉基础模型预训练流程示意图

间的差异，其优化目标函数可形式化表示为：

$$\min_{\theta_s} \sum_{\boldsymbol{x}\in\{\boldsymbol{x}_1^g,\boldsymbol{x}_2^g\}} \sum_{\substack{\boldsymbol{x}'\in V \\ \boldsymbol{x}'\neq\boldsymbol{x}}} H\left(P_t(\boldsymbol{x}), P_s\left(\boldsymbol{x}'\right)\right), \tag{3.1}$$

其中 $H(a,b)=-a\log b$ 表示交叉熵损失函数。训练过程中，学生模型参数通过随机梯度下降进行更新，而教师模型则采用了两个关键的机制来保证训练的稳定性和效果：

(1) 参数更新机制：教师模型在训练时会进行 θ_t 的梯度计算，并使用指数移动平均（Exponential Moving Average，EMA）更新参数；

(2) 中心化操作：为了减少模型对数据批次大小 B 的依赖，同时防止训练崩溃，DINO 引入了中心化（centering）操作。这一操作本质上是对教师模型输出增加一个动态调整的偏置项 c，表示为 $g_t(x)\leftarrow g_t(x)+c$。该偏置项通过指数移动平均进行更新：

$$c\leftarrow mc+(1-m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}\left(x_i\right) \tag{3.2}$$

种设计不仅提高了训练的稳定性，还有助于模型学习更有判别性的特征表示。通过教师网络和学生网络的双重约束，以及全局视角和局部视角的互补学习，DINO 能够在无监督条件下学习到高质量的视觉特征表示。

3.3　双源信息融合道路场景解析网络 HAPNet

前文 3.2已提到，视觉基础模型能够学习到更丰富且通用的表征。然而，在道路场景解析任务这类密集预测任务的背景下，由于缺乏与密集预测任务相关的细粒度空间先验知识，可能导致其收敛速度较慢且性能偏低（第 3.4.3节的消

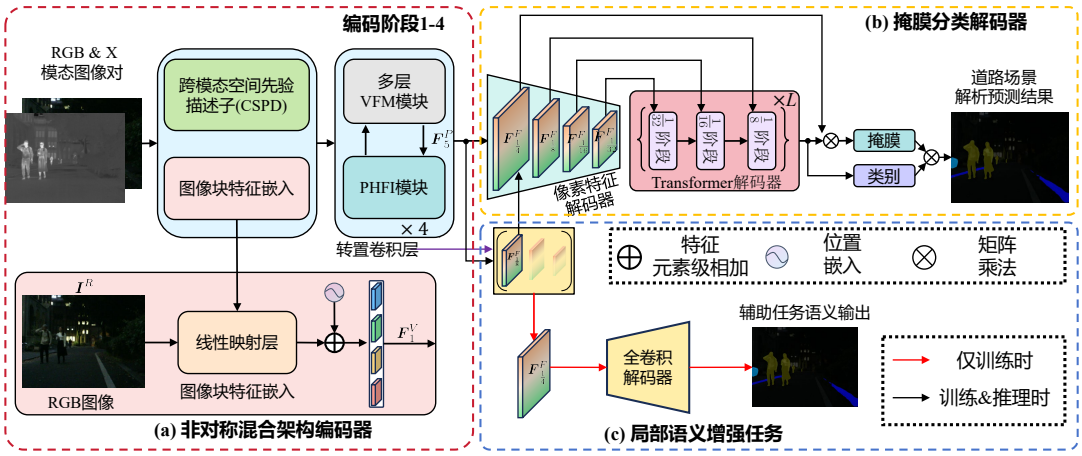


图 3.3　所提出的道路场景解析网络 HAPNet 示意图

融实验验证了该推论）；此外，难以保证基于 RGB 图像或 RGB 图像-文本对训练的 VFM 适用于处理 X 模态图像。为了实现 VFM 在双源信息融合道路场景解析领域中的成功引入，我们需要确保所提出的网络架构具备以下性质：(1) 既能利用 VFM 的通用特征提取能力；(2) 能够有效适配不同多源信息的内在特性进行特征编码；(3) 能够实现有效且深层次的多尺度异构特征融合。本章节所提出的网络 HAPNet 同时具备以上三种特点。该网络接受 RGB 图像与 X 模态图像（不同于上一章仅使用法向量图像的设定）作为输入，并输出高精度的场景解析预测。如图 3.3所示，该网络主体由以下三部分构成：

- 一个基于 VFM 与 CSPD 的非对称异构特征编码器架构，利用 RGB 图像与 RGB-X 图像对分别提取全局上下文特征与多尺度的跨模态空间先验，并且在编码过程中通过 PHFI 渐进式的实现全局上下文特征与空间先验的融合；
- 融合的异构特征经过一个基于 Transformer 的掩膜分类范式解码器输出高精度的语义预测；
- 一个辅助任务，通过深监督技术增强融合特征的局部语义，在推理无需额外参数的情况下进一步增强网络的性能。

3.3.1　基于视觉基础模型的异构特征编解码

相比于领域中现有的对称并行式编码器，所提出的 HAPNet 创新的将基于 ViT 架构的视觉基础模型与基于轻量级 CNN 的 CSPD 模块组合，构成了一个非对称的混合架构双路编码器。相比于对称的并行式编码器中无差别的异构特征提取策略，本章的 HAPNet 有效利用了两种模态的互补优势。它使用 VFM 提取丰富的全局上下文特征，同时通过 CNN 学习双源信息中蕴含丰富局部语义的跨模态空间先验，并在编码过程中通过提出的 PHFI 在不同编码阶段进行有效的全局上下文特征-空间先验融合，产生更具区分度的异构融合特征。因此，我们的解码器能够更有效地利用全局上下文进行准确分类，同时保持对小尺寸目标局部细节的敏感性。

具体来说，首先将 ViT（含 L 个编码器层）均匀划分为四个编码模块，并在每个模块中插入 PHFI 以形成四个编码阶段以促进不同细粒度的异构特征编码。给定 RGB 图像 $\boldsymbol{I}^R \in \mathbb{R}^{H\times W\times 3}$ 及其对应的 X 模态图像 $\boldsymbol{I}^T \in \mathbb{R}^{H\times W\times 3}$ 首先输入 CSPD，提取跨模态空间先验 $\boldsymbol{F}_1^P \in \mathbb{R}^{(\sum_{i=2}^{4}\frac{HW}{S_i^2})\times D}$，作为后续编码阶段的输入之一，其中 $S_i = 2^{i+1}$ $(i \in [2,4]\cap\mathbb{Z})$ 表示对应的特征步长数。随后，$\boldsymbol{I}^R$ 经过线性映射层形成上下文特征 $\boldsymbol{F}_1^V \in \mathbb{R}^{\frac{HW}{16^2}\times D}$，作为后续阶段的另一输入。$\boldsymbol{F}_1^V$ 和 $\boldsymbol{F}_1^P$ 随后通过 PHFI 的两个关键模块：(1) 全局-局部上下文聚合器 (global-local context aggregator，GLCA) 和 (2) 互补上下文生成器 (complementary context generator，

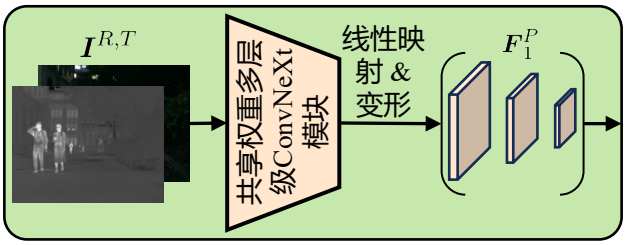


图 3.4　所提出的跨模态空间先验描述子架构

CCG)，在多个编码阶段进行双向的渐进式融合。同时，融合特征在四个 ViT 块中进行全局上下文编码。这种设计使全局上下文特征在通过强大的 VFM 有效捕获全局上下文的同时，捕获精细的空间先验。最终，从四个编码阶段获得融合的多尺度特征 $\mathcal{F}^F = \{\boldsymbol{F}^F_{\frac{1}{S_j}} \in \mathbb{R}^{\frac{H}{S_j}\times\frac{W}{S_j}\times D}\}$，其中 $S_j = 2^{i+1}$ $(j \in [1,4] \cap \mathbb{Z})$，随后输入到掩码分类解码器以生成语义预测 $\boldsymbol{M}^P \in \mathbb{R}^{H\times W}$。

非对称输入模态的 VFM 与 CNN 双路混合编码　为了适配 VFM 架构，在进入 ViT 编码器前首先将 $\boldsymbol{I}^R$ 均匀划分为 16×16 像素大小的图像块，并将其投影为 D 维度的图像块嵌入，提供全局上下文特征 $\boldsymbol{F}^V_1$。在此后的第 i 个编码阶段中，输入的全局上下文特征 $\boldsymbol{F}^V_i$ 和跨模态空间先验 $\boldsymbol{F}^P_i$ 首先在 GLCA 中融合，生成包含更丰富的局部语义的上下文特征 $\hat{\boldsymbol{F}}^V_i \in \mathbb{R}^{\frac{HW}{16^2}\times D}$。随后 $\hat{\boldsymbol{F}}^V_i$ 在 ViT 编码模块中进行全局上下文编码，得到输出 $\boldsymbol{F}^V_{i+1}$，$\boldsymbol{F}^V_{i+1}$ 随后被输入到 CCG 中，与 $\boldsymbol{F}^P_i$ 进行另一次融合。这一步骤通过 $\boldsymbol{F}^V_{i+1}$ 中更新的全局上下文特征对空间先验 $\boldsymbol{F}^P_i$ 中缺少的长距离依赖进行补充，从而生成 $\boldsymbol{F}^P_{i+1}$。通过在四个编码阶段重复上述双向的特征融合，得到最终的空间先验 $\boldsymbol{F}^P_5$ 作为多尺度的异构融合特征，随后将其恢复到其原始的三种空间分辨率，形成多尺度异构特征 $\boldsymbol{F}^F_{\frac{1}{8}}, \boldsymbol{F}^F_{\frac{1}{16}}, \boldsymbol{F}^F_{\frac{1}{32}}$。此外，我们采用 2×2 的转置卷积直接从 $\boldsymbol{F}^F_{\frac{1}{8}}$ 创建 $\frac{1}{4}$ 尺度的特征 $\boldsymbol{F}^F_{\frac{1}{4}}$，该设计是为了避免了得到 $\boldsymbol{F}^F_{\frac{1}{4}}$ 所需注意力操作带来的高计算开销。最终，这些包含四个尺度的异构融合特征构成 $\mathcal{F}^F$，与当前最先进的场景解析网络架构[55] 兼容。仅将 RGB 图像作为 ViT 输入的设计源于我们的假设：使用 VFM 显式编码 X 模态图像可能导致网络收敛问题和表征偏移问题[121]。此外，更侧重于场景几何结构与显著物体轮廓特点的 X 模态图像通常在捕获全局上下文信息方面能力有限，因为具有相似几何形状的物体往往具有相似的空间几何特征或热源特征。例如，在深度图像或热图像中区分足球和篮球、猫和狗等具有类似形状的目标是很困难的。第 3.4.3节中的消融实验证明了所提出的非对称网络设计的有效性。

跨模态空间先验描述子　与 Transformer 架构[25] 相比，CNN 在提取局部细节特征方面展现出了卓越性能，这奠定了其在需要清晰物体边界的密集预测任务

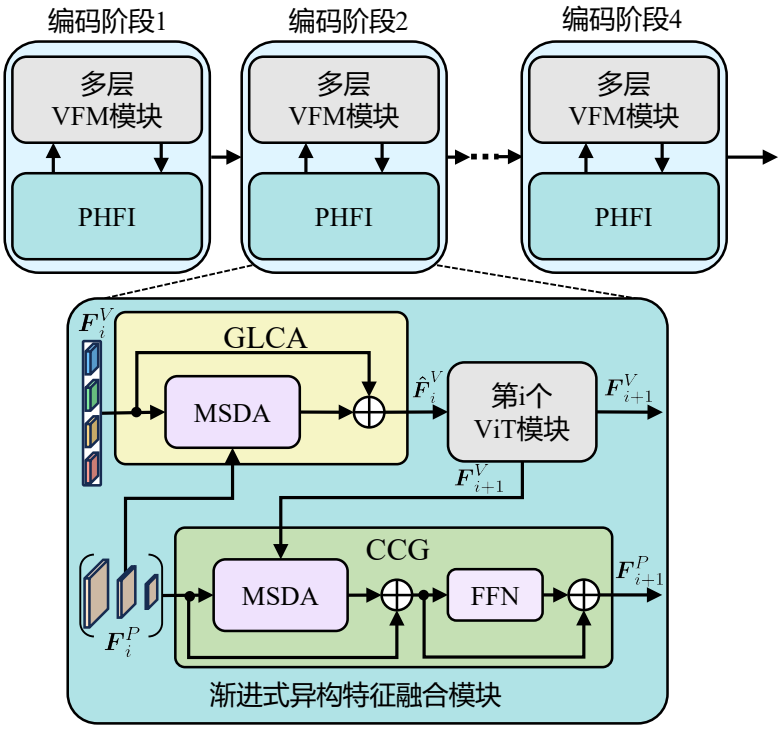


图 3.5　所提出的渐进式异构特征融合模块 PHFI 架构

中的重要性。近期利用 ConvNeXt 进行多模态场景解析的一项研究[122] 证明了 ConvNeXt[85] 在捕获丰富、稳健的视觉特征方面表现出色，优于 ResNet[63] 及其变体[123] 等其他层次化 CNN 架构。鉴于这些优势，我们采用 ConvNeXt 架构构建 CSPD。此外，将 RGB 图像作为互补输入，使 CSPD 能够提取比单独使用 X 模态图像表征能力更佳全面的空间先验。因此，我们使用两个权重共享的 ConvNeXt 网络构建 CSPD（如图 3.4所示），实现了对 RGB-X 图像对中跨模态空间先验的高效提取。鉴于 ConvNeXt 网络在上一章节 2.2.1中已进行过具体介绍，在这里不对其具体架构展开阐述。此外，第 3.4.3节中提供了详细的消融实验，分析了不同 CSPD 构建模块和数据输入策略对整体性能的影响。

为了提取跨模态的空间先验，RGB-X 图像对 $\boldsymbol{I}^R$ 和 $\boldsymbol{I}^T$ 分别输入到一系列权重共享的 ConvNeXt 模块中，分别生成多尺度特征$\mathcal{P}^R = \{\boldsymbol{P}_2^R, \boldsymbol{P}_3^R, \boldsymbol{P}_4^R\}$ 和$\mathcal{P}^T = \{\boldsymbol{P}_2^T, \boldsymbol{P}_3^T, \boldsymbol{P}_4^T\}$，其中 $\boldsymbol{P}_i^{R,T} \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times C_i}$ 表示第 i 阶段的特征，C_i 和 $S_i = 2^{i+1}$ ($i \in [2,4]\cap\mathbb{Z}$) 分别表示对应的特征通道数和步长。随后，我们通过逐元素求和将 $\boldsymbol{P}_i^R$ 和 $\boldsymbol{P}_i^T$ 结合起来，并通过 1×1 卷积将结果降维至 D 通道，得到异构特征$\mathcal{P}^H = \{\boldsymbol{P}_2^H, \boldsymbol{P}_3^H, \boldsymbol{P}_4^H\}$，其中 $\boldsymbol{P}_i^H \in \mathbb{R}^{\frac{H}{S_i}\times\frac{W}{S_i}\times D}$。这三个尺度的特征随后被展平并串联，形成用于后续编码阶段的跨模态空间先验 $\boldsymbol{F}_1^P$。

渐进式异构特征融合模块　如图 3.5所示，PHFI 中的两个基于注意力的组件 GLCA 和 CCG 分别负责在每个 ViT 块前后的双路径特征融合过程。给定第 i 个编码块 ($i \in [1,4]\cap\mathbb{Z}$) 的 $\boldsymbol{F}_i^V$ 和 $\boldsymbol{F}_i^P$，我们首先将它们输入 GLCA 以获得局部语

义增强的输入 $\hat{\boldsymbol{F}}_i^V$。具体而言，$\boldsymbol{F}_i^V$ 作为查询，而 $\boldsymbol{F}_i^P$ 在 GLCA 中作为键和值，进行交叉注意力操作。这一过程可以表述为：

$$\hat{\boldsymbol{F}}_i^V = \boldsymbol{F}_i^V + \kappa_i \mathrm{MHA}(\mathrm{LN}(\boldsymbol{F}_i^V), \mathrm{LN}(\boldsymbol{F}_i^P)), \tag{3.3}$$

其中 LN(·) 表示层归一化（LayerNorm）操作，MHA(·,·) 表示多头注意力操作。参照[124, 125] 等注意力模块中的常见做法，我们引入了一个可学习系数 κ_i，以动态调整 MHA 的权重，实现局部语义与上下文特征的灵活融合。经过第 i 个 ViT 块处理后，得到蕴含丰富全局上下文信息的输出特征 $\boldsymbol{F}_{i+1}^V$。随后，全局上下文特征 $\boldsymbol{F}_{i+1}^V$ 被输入到 CCG 中，进行与空间先验 $\boldsymbol{F}_i^P$ 融合。在这个过程中，$\boldsymbol{F}_i^P$ 作为查询，而 $\boldsymbol{F}_{i+1}^V$ 作为键和值在 CCG 中执行 MHA，如下所示：

$$\hat{\boldsymbol{F}}_i^P = \boldsymbol{F}_i^P + \mathrm{MHA}(\mathrm{LN}(\boldsymbol{F}_i^P), \mathrm{LN}(\boldsymbol{F}_{i+1}^V)). \tag{3.4}$$

MHA 操作后，应用前馈神经网络 (FFN) 进一步处理融合特征，得到更新的空间先验 $\boldsymbol{F}_{i+1}^P$。这些特征已经融合了丰富的局部和全局语义，并将作为下一个编码阶段的输入。在上述两个组件中，MHA 过程基于多尺度可变形注意力 (MSDA)[90] 实现，以减少计算量。

3.3.2　掩膜分类范式解码器

本章节沿用了与章节 2.2.3相同的掩膜分类范式解码器，因此这里不对解码器原理进行重述，仅简要介绍其预测流程。在解码过程中，像素解码器对融合的异构特征 $\{\boldsymbol{F}_{\frac{1}{8}}^F, \boldsymbol{F}_{\frac{1}{16}}^F, \boldsymbol{F}_{\frac{1}{32}}^F\}$ 进行优化，生成 $\hat{\mathcal{F}}^P = \{\hat{\boldsymbol{F}}_{\frac{1}{8}}^F, \hat{\boldsymbol{F}}_{\frac{1}{16}}^F, \hat{\boldsymbol{F}}_{\frac{1}{32}}^F\}$，并从 $\boldsymbol{F}_{\frac{1}{4}}^F$ 生成逐像素嵌入 $\boldsymbol{E}^P \in \mathbb{R}^{C \times \frac{H}{4} \times \frac{W}{4}}$；Transformer 解码器则利用 $\hat{\mathcal{F}}^P$ 来优化固定数量的物体查询，从而生成掩膜嵌入 $\boldsymbol{E}^M \in \mathbb{R}^{Q \times C}$，其中 Q 和 C 分别表示物体查询数量和特征通道数。通过 $\boldsymbol{E}^M$ 与 $\boldsymbol{E}^P$ 的相乘运算，得到掩膜预测 $\boldsymbol{M}^M \in \mathbb{R}^{Q \times \frac{H}{4} \times \frac{W}{4}}$。将 $\boldsymbol{M}^M$ 调整尺寸并与从查询中学习得到的类别预测 $\boldsymbol{E}^C \in \mathbb{R}^{Q \times N}$ 进行点积，得到最终的语义预测 $\boldsymbol{M}^P$，其中 N 表示类别数量（包含一个" 无对象" 类别）。

3.3.3　局部语义增强任务

在我们的掩码分类解码器中，从 $\boldsymbol{F}_{\frac{1}{4}}^F$ 生成的 $\boldsymbol{E}^P$通常提供场景丰富的局部（逐像素）语义信息，而从其他三个尺度特征生成的 $\boldsymbol{E}^M$ 则提供类别相关的聚类中心信息[126]。为了使输入掩码分类解码器的融合特征中蕴含更具易分辨性的局部语义，我们对 $\boldsymbol{F}_{\frac{1}{4}}^F$ 引入深度监督，有效减少其中的无语义区分性的特征以及其中的噪声。具体而言，我们在 $\boldsymbol{F}_{\frac{1}{4}}^F$ 上附加一个轻量级的全卷积（Fully Convolutional Network，FCN）[27] 解码头，并直接监督这个辅助网络产生语义预测。经过两个

卷积层后，生成具有 $(N-1)$ 个通道（不包括" 无目标" 类别）且与 $\boldsymbol{F}_{\frac{1}{4}}^{F}$ 分辨率相同的辅助语义预测。这种设计使相应的编码器层能够产生包含更有区分度的局部语义的融合异构特征 $\boldsymbol{F}_{\frac{1}{4}}^{F}$，进一步确保像素解码器能够生成更具判别性的 $\boldsymbol{E}^{P}$，最终在不增加网络参数和计算复杂度的情况下提升场景解析性能。

3.3.4　损失函数设计

在本章节所提出的 HAPNet 中，除了局部语义增强任务部分以外的损失函数与上一章节相同，包括用于类别预测的交叉熵损失 $\mathcal{L}_{\text{cls}}$）以及用于掩码预测的二元交叉熵损失 $\mathcal{L}_{\text{bce}}$）和 Dice 损失（$\mathcal{L}_{\text{dice}}$）。针对局部语义增强任务，我们使用了标准交叉熵损失（$\mathcal{L}_{\text{ce}}$），总体损失函数可以表示为：

$$\mathcal{L} = \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} \tag{3.5}$$

各个损失函数分量的权重系数分别设定为：$\lambda_{\text{bce}} = 5.0$，$\lambda_{\text{dice}} = 5.0$，$\lambda_{\text{cls}} = 2.0$，以及 $\lambda_{\text{ce}} = 0.4$。对于" 无目标" 类别，我们将 λ_{cls} 的值调整为 0.1，以平衡模型对背景区域的学习。在模型训练过程中，采用匈牙利算法[80] 确定最优的预测-标签匹配方案。

3.4　方法验证与实验结果分析

在本节中，为充分评估 HAPNet 在基于 RGB-X 图像对的道路场景解析任务中的性能，我们选取了热图像以及深度图像两种不同的信息与 RGB 图像组成输入图像对（RGB-T、RGB-D），在相关公开数据集中进行了对比分析。首先将 HAPNet 与其他最先进的双源信息融合网络在以下三个 RGB-T 数据集上进行对比：

3.4.1　数据集与评价指标

MFNet 数据集[57]　这是一个城市驾驶场景解析数据集。其中数据由 InfReC R500 相机进行采集，数据集包含 1569 对精确同步的 RGB-T 图像对，图像分辨率为 640×480 像素。所有图像均在不同时间、不同天气条件下采集，以确保数据的多样性。数据集提供了九个类别的精细语义标注：自行车、行人、汽车、道路车道线、护栏、停车位、减速带、路标锥和背景。数据集收集了在不同光照条件下（包括白天和夜间）获取的数据，以验证模型在各种环境下的鲁棒性。本章节采用与原始论文相同的数据划分策略进行训练和测试。

PST900 数据集[127]　该数据集是为应对极端地下环境挑战而构建的 RGB-T 数据集，包含 894 对高质量配准的 RGB 和热红外图像对。数据采集设备采用 Stereolabs ZED Mini 立体相机（用于 RGB 图像采集）和 FLIR Boson 320 热成像相机（用于热红外图像采集），两种传感器经过严格的时空同步校准，输出分辨率均为 $1,280 \times 720$ 像素。数据集中的场景主要来自隧道、地下停车场等低照度环境，这些场景对传统的单模态网络构成了严峻挑战。数据集提供了五个关键目标的语义标注：背景、灭火器、背包、手持电钻和幸存者，这些类别对地下搜救任务具有重要意义。本章节遵循了研究[127] 的数据划分策略①对数据集进行划分。

KP Day-Night 数据集[128]　该数据集专注于全天候自动驾驶场景理解，包含 950 对经过精确配准的 RGB-T 图像对，图像分辨率为 640×512 像素。数据集包含了大量的昼夜场景对比数据，充分展示了热红外成像在夜间场景感知中的优势。数据集采用与 CityScapes 数据集[22] 相同的 19 类语义标注体系，包括：道路、人行道、建筑物、围墙、围栏、杆子、交通灯、交通标志、植被、地形、天空、行人、骑行者、汽车、卡车、公交车、火车、摩托车和自行车。这种标注体系的一致性便于不同数据集间进行迁移学习和性能对比。本章节按照研究[129] 提供的划分方案，以确保结果的可比性和可靠性。

此外，我们在 NYU-Depth V2 数据集[130] 上进行了实验，以评估我们网络在 RGB-D/HHA 场景解析任务上的泛化与适应能力。

NYU-Depth V2 数据集[130]　该数据集是 RGB-D 场景解析研究中最广泛使用的基准数据集。该数据集由纽约大学使用 Microsoft Kinect 深度相机采集。数据集包含 1,449 对精确配准的 RGB 图像和深度图像，分辨率为 480×640 像素。每张深度图都被处理成对应的 HHA 编码图像，其中 H 通道表示相对高度，第一个 A 通道编码与水平面的角度，第二个 A 通道编码与重力方向的角度。这种编码方式能够有效地将单通道的深度图像转化为适合 CNN 处理的三通道表示。数据集提供了 40 个类别的密集语义标注。本章节采用与[121] 相同的数据集划分策略，以评估 HAPNet 在 RGB-D 场景解析任务上的泛化能力。

3.4.2　网络实现细节和评测指标

HAPNet 的模型训练在 NVIDIA RTX 3090 GPU 上进行，在每个数据集上训练了 200 轮。我们采用 AdamW 优化器[96] 对网络参数进行更新，初始学习率设置为 2×10^{-4}，权重衰减系数为 5×10^{-2}。此外，参照[101] 中的最佳实践，我们对 VFM 编码器应用了 0.9 的层级学习率衰减策略，这有助于稳定模型训练过程并

① 已公开的图像对数量少于文献[127] 中报告的数量

提升性能。在 NYU Depth V2 数据集的实验中，模型输入包括 RGB 图像和深度图像的 HHA 编码表示。HHA 编码通过将深度图像转换为三通道表示（水平视差、相对高度和表面法向量角度），使深度图像中的空间几何特征更易被基于 CNN 架构的网络提取和学习。为了验证模型各个组件的有效性，我们在广泛使用的 MFNet 数据集上进行了一系列消融实验，在基准实验中暂时移除了辅助任务。通过对比表 3.1和表 3.4中的结果可以发现，引入辅助任务后，模型的 mIoU 平均提升了 0.6 个百分点（该结果基于多次实验的平均值，波动范围约为 ±0.2%）。这一实验结果充分证明了辅助任务的有效性，表明其能够帮助模型学习更具判别性的特征表示，从而提升整体性能。

本章节采用了准确率 (Accuracy, Acc) 和交并比 (IoU) 这两个评估指标对网络的场景解析性能进行评估，还计算了所有类别的平均准确率 (mean Accuracy, mAcc) 和平均交并比 (mIoU)，这些指标都已在上一章第 2.3.1节中进行过详细介绍，在此不再进行单独说明。对于在 NYU-Depth V2 数据集上进行的实验，我们额外记录了像素准确率（像素 Acc）指标，该指标的计算方法可见章节 4.4.1。

3.4.3　公开基准数据集对比实验

我们在 MFNet[57]、PST900[127] 和 KP Day-Night[128] 这三个数据集上，分别与 12 个、10 个和 4 个最先进的 RGB-T 场景解析网络进行了定量比较，实验结果可见表 3.1、3.2和 3.3。同时，这三个数据集上的定性对比结果如图 3.6、3.7和 3.8所示。具体而言，本章提出的 HAPNet 在所有数据集上都取得了最高的 mIoU 值：在 MFNet 数据集上超越现有最优方法 0.1-21.8 个百分点，在 PST900 数据集上领先 1.0-32.0 个百分点，在 KP 日夜数据集上提升 2.4-33.6 个百分点。这些实验结果充分证明了 HAPNet 在各种场景下（从复杂的城市驾驶场景到具有挑战性的地下场景）进行 RGB-T 场景解析时的鲁棒性和有效性。值得注意的是，HAPNet 的性能超越了另一个基于掩码分类范式的 Transformer 架构 CRM-RGBTSeg[129]，这验证了我们所提出的混合非对称架构在异构特征融合方面的优越性。与同样采用非对称编码器架构的通用多源信息融合网络 CMNeXt[142] 相比，HAPNet 在 MFNet 数据集上的 mIoU 提升了 1.6 个百分点。这一性能提升主要得益于 HAPNet 采用了更强大的 VFM 和更先进的多尺度交叉注意力机制，该机制能够有效整合来自 VFM 的全局上下文信息以及CNN 提供的多尺度的跨模态空间先验。相比之下，CMNeXt 仅依赖 CNN 和基于池化的特征融合策略。因此，HAPNet 能够生成更具判别性的融合特征用于 RGB-T 场景解析。如图 3.6、3.7和 3.8所示的场景解析结果也直观地展示了 HAPNet 在弱光照条件下进行视觉感知的卓越能力。此外，HAPNet 在 RGB-D/HHA 场景解析任务上也展现出良好的泛化性能。如表 3.5所

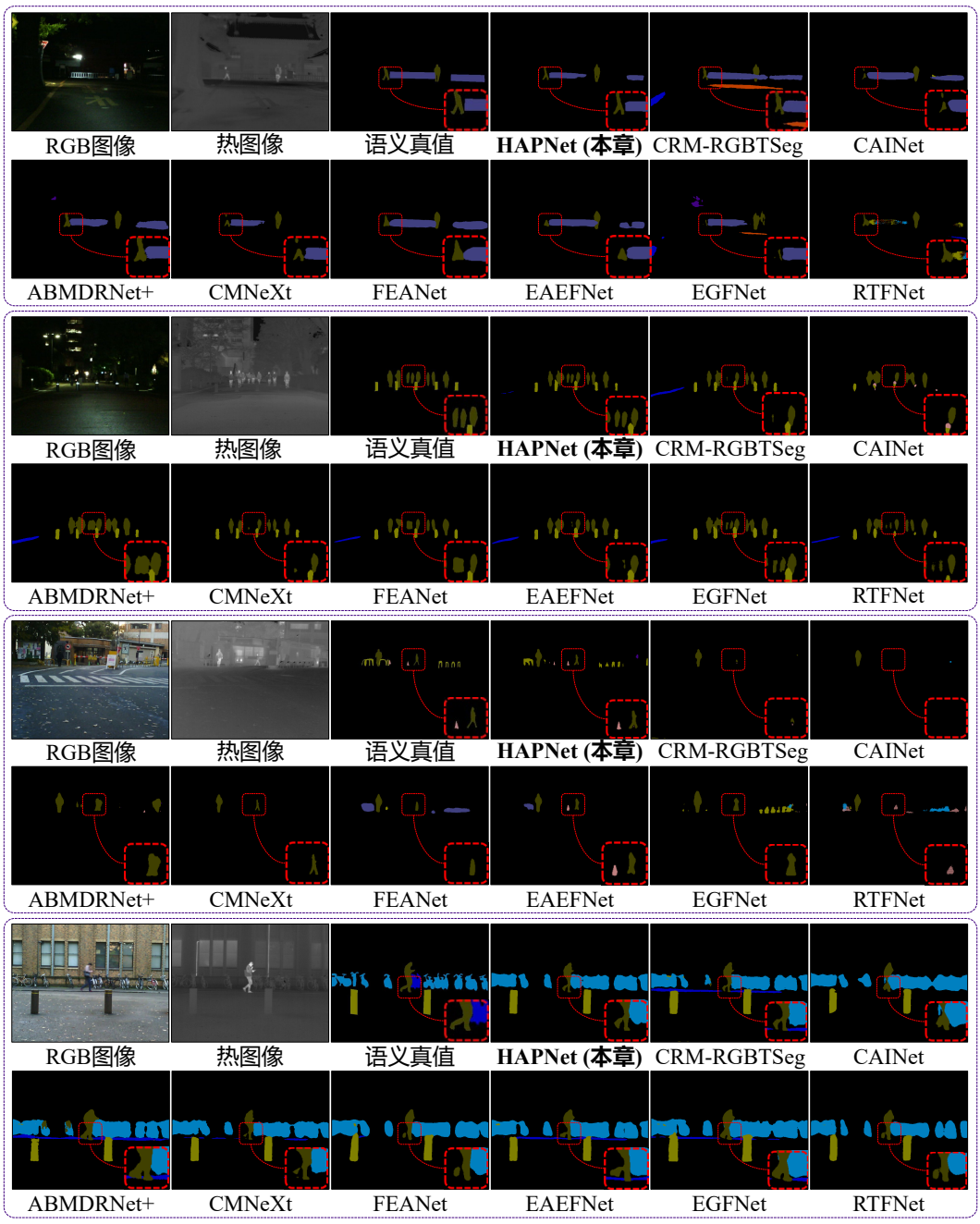


图 3.6　与现有最先进的 RGB-T 场景解析网络在 MFNet 测试集上的定性比较，显著改进的区域在图中已用红色虚线框标出

示，我们的方法在 NYU Depth V2 数据集上优于所有其他 RGB-T 场景解析网络。具体而言，HAPNet 的 mIoU 比 ECGFNet[134] 高 3.5 个百分点，比 RTFNet[58] 高 5.9 个百分点。然而，这些结果也表明 HAPNet 仍落后于专门为 RGB-D/HHA 信息融合开发的最新方法，包括 Omnivore[146]、DFormer-L[121] 和 OmniVec[145]（mIoU 低 2.2-5.8 个百分点）。同时，其性能也低于另外两个最先进的通用信息融合（同时适用于热图像、深度图像、偏振图像等多源信息）网络 CMX 和 CMNeXt[141, 142]（mIoU 低 1.9 个百分点）。我们推测这一性能差距主要源于热红外和深度/HHA 模态之间固有的数据分布差异，这在一定程度上限制了我们的架构在 RGB-D/HHA

表 3.1　与现有最先进的 RGB-T 场景解析方法在 MFNet 测试集上的定量比较（%）。符号"-"表示原始文献中缺失的数据，最佳结果以粗体显示。表中省略了" 背景" 类别的 Acc 和 IoU 指标，但这些数值仍计入相应平均值的计算中

网络名称	车辆		行人		自行车		弯道		停车标志		护栏		彩锥		减速带		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
MFNet[57]	77.2	65.9	67.0	58.9	53.9	42.9	36.2	29.9	19.1	9.9	0.1	8.5	30.3	25.2	30.0	27.7	45.1	39.7
RTFNet[58]	93.0	87.4	79.3	70.3	76.8	62.7	60.7	45.3	38.5	29.8	0.0	0.0	45.5	29.1	74.7	55.7	63.1	53.2
FuseSeg[131]	93.1	87.9	81.4	71.7	78.5	64.6	68.4	44.8	29.1	22.7	63.7	6.4	55.8	46.9	66.4	47.9	70.6	54.5
EGFNet[132]	**95.8**	87.6	89.0	69.8	80.6	58.8	71.5	42.8	48.7	33.8	33.6	7.0	65.3	48.3	71.1	47.1	72.7	54.8
ABMDRNet[133]	94.3	84.8	90.0	69.6	75.7	60.3	64.0	45.1	44.1	33.1	31.0	5.1	61.7	47.4	66.2	50.0	69.5	54.8
ECGFNet[134]	89.4	83.5	85.2	72.1	72.9	61.6	62.8	40.5	44.8	30.8	45.2	11.1	57.2	49.7	65.1	50.9	69.1	55.3
FEANet[135]	93.3	87.8	82.7	71.1	76.7	61.1	65.5	46.5	26.6	22.1	**70.8**	6.6	**66.6**	55.3	77.3	48.9	73.2	55.3
SFAF-MA[136]	94.0	88.1	82.5	73.0	73.9	61.3	63.6	45.6	37.5	29.5	42.2	5.5	57.9	45.7	74.4	53.8	69.6	55.5
ABMDRNet+[137]	95.2	87.1	**92.5**	69.8	76.2	60.9	**72.0**	47.8	42.3	34.2	66.8	8.2	64.8	50.2	63.5	55.0	74.7	56.8
LLE-Seg[138]	91.8	88.6	81.3	73.2	73.7	64.8	62.5	46.8	33.7	30.0	49.1	8.8	55.7	52.5	72.9	**62.4**	71.6	58.4
CAINet[139]	93.0	88.5	74.6	66.3	**85.2**	**68.7**	65.9	**55.4**	34.7	31.5	65.6	9.0	55.6	48.9	**85.0**	60.7	73.2	58.6
EAEFNet[140]	95.4	87.6	85.2	72.6	79.9	63.8	70.6	48.6	47.9	35.0	62.8	14.2	62.7	52.4	71.9	58.3	**75.1**	58.9
CMX[141]	-	90.1	-	75.2	-	64.5	-	50.2	-	35.3	-	8.5	-	54.2	-	60.6	-	59.7
CMNeXt[142]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		59.9
CRM-RGBTSeg[129]	-	90.0	-	75.1	-	67.0	-	45.2	-	**49.7**	-	**18.4**	-	54.2	-	54.4	-	61.4
HAPNet（本章）	95.1	**90.6**	85.5	**75.4**	79.0	67.2	67.9	51.1	**59.5**	48.4	16.0	4.2	65.0	**59.1**	80.3	59.1	70.3	**61.5**

场景解析领域的适用性。尽管取得了最优的性能表现，HAPNet 仍保持着具有竞争力的实时性能。我们的网络包含 169.1M 参数，在 NVIDIA RTX 4090-D GPU 上处理 480Œ480 分辨率的图像时，能够达到约 23 帧/秒（FPS）的推理速度。这一性能表现基本满足自动驾驶系统的实时处理需求。

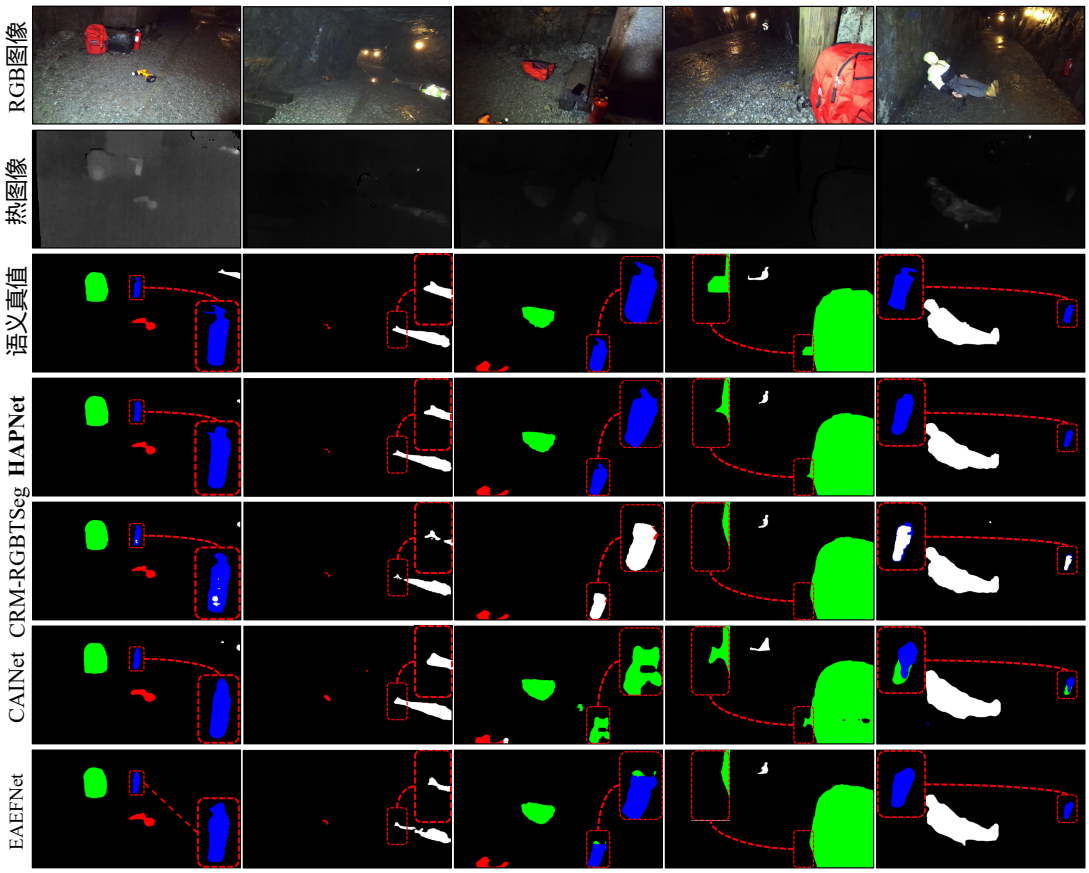


图 3.7　与现有最先进的 RGB-T 场景解析网络在 PST900 测试集上的定性比较，显著改进的区域在图中用红色虚线框标出。

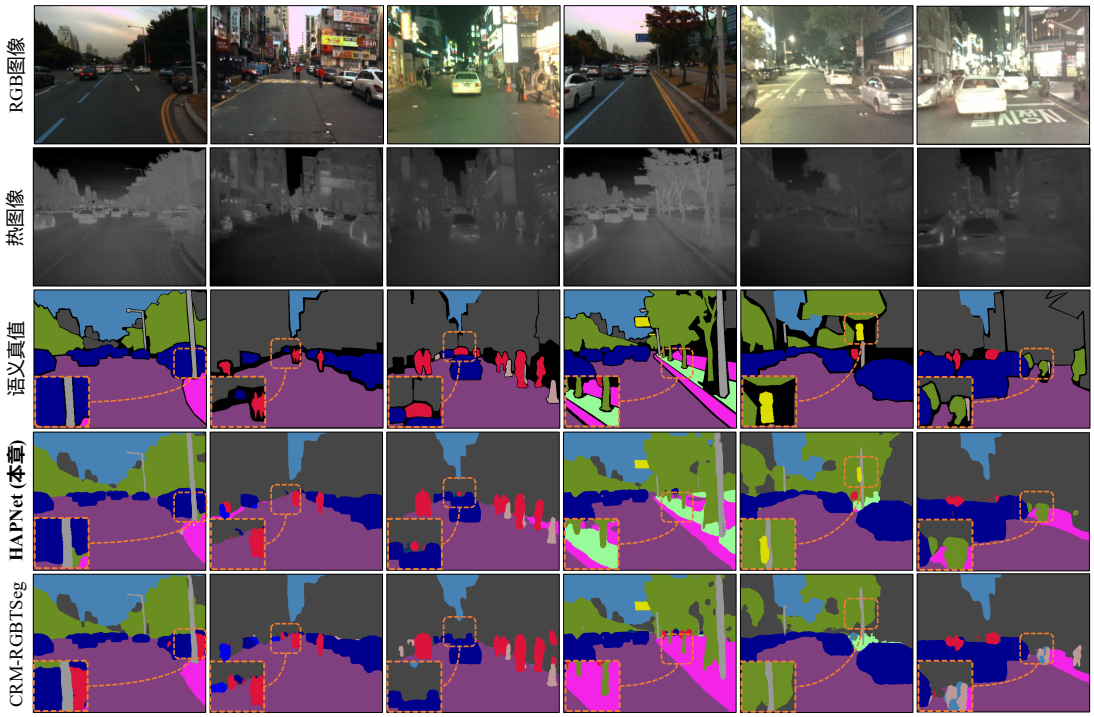


图 3.8　与现有最先进的 RGB-T 场景解析网络在 MFNet 测试集上的定性比较，显著改进的区域在图中已用橙色虚线框标出

表 3.2　与现有最先进的 RGB-T 场景解析方法在 PST900 测试集上的定量比较（%）。符号"-"表示原始文献中缺失的数据，最佳结果以粗体显示

网络名称	背景		灭火器		双肩包		手钻头		幸存者		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
MFNet	-	98.6	-	41.1	-	64.2	-	60.3	-	20.7	-	57.0
RTFNet	-	98.9	-	36.4	-	75.3	-	52.0	-	25.3	-	57.6
EGFNet	-	99.2	-	74.3	-	83.0	-	71.2	-	64.6	-	78.5
ABMDRNet	-	98.7	-	24.1	-	72.9	-	54.9	-	57.6	-	67.3
FEANet	-	-	-	-	-	-	-	-	-	-	91.4	85.5
DBCNet	-	98.9	-	62.3	-	71.1	-	52.4	-	40.6	-	74.5
CAINet	99.6	99.5	**95.8**	80.3	**96.0**	88.0	88.3	77.2	91.3	78.6	**94.2**	84.7
EAEFNet	**99.8**	99.5	92.2	80.4	91.0	87.7	93.0	83.9	79.3	75.6	91.1	85.4
GMNet	**99.8**	99.4	90.2	**85.1**	89.0	83.8	88.2	73.7	80.8	78.3	89.6	84.1
CRM-RGBTSeg	-	**99.6**	-	79.5	-	89.6	-	89.0	-	82.2	-	88.0
HAPNet（本章）	**99.8**	**99.6**	93.9	81.3	95.1	**92.0**	**95.5**	**89.3**	**85.6**	**82.4**	94.0	**89.0**

表 3.5　与现有最先进的信息融合场景解析网络在 NYU-Depth V2 测试集上的定量比较（%）。符号“-”表示原始文献中缺失的数据，最佳结果以粗体显示

输入信息	网络名称	Pixel Acc	mAcc	mIoU	排名
RGB-D	OmniVec[145]	-	-	**60.8**	**1**
	Omnivore[146]			56.8	8
	TokenFusion[147]	79.0	66.9	54.2	16
	DFormer-L[121]	-	-	57.2	5
	AsymFormer[148]	78.5	-	54.1	17
	RTFNet[58]	-	64.8	49.1	59
	ECGFNet[134]	-	65.2	51.5	37
RGB-HHA	CMX[141]	-	-	56.9	6
	CMNeXt[142]	-	-	56.9	6
	SA-Gate[149]	-	-	52.4	31
	HAPNet （本章）	**79.2**	**68.8**	55.0	15

3.4.4　消融实验

首先，我们对 RGB-T 场景解析在不同输入数据组合策略下的性能进行了深入探究。基于非对称双分支架构，我们可以通过九种不同的策略将 RGB-T 数据分别输入到 VFM 和 CSPD 中，详细结果如表 3.6所示。其中，策略 E 和 I 可视

表 3.3　与现有最先进的 RGB-T 场景解析方法在 KP Day-Night 测试集上的定量比较（%）。最佳结果以粗体显示。虽然表中省略了某些类别的准确率（Acc）和交并比（IoU）结果，但这些数值仍计入相应平均值的计算中

网络名称	道路	人行道	建筑物	围栏	电线杆	红绿灯	交通标志	植被	地面	天空	行人	骑行者	汽车	卡车	公共汽车	摩托车	自行车	mIoU
MFNet[57]	93.5	23.6	75.1	0.1	9.1	0.0	0.0	69.3	0.2	90.4	24.0	0.0	69.6	**0.3**	0.3	0.0	0.6	24.0
RTFNet[58]	94.6	39.4	86.6	0.6	0.0	0.0	0.0	81.7	3.7	92.8	58.4	0.0	87.7	0.0	0.0	0.0	0.5	28.7
CMX[141]	97.7	53.8	90.2-	47.1	46.2	10.9	45.1	87.2	34.3	93.5	74.5	0.0	91.6	0.0	59.7	46.1	0.2	46.2
CRM-RGBTSeg[129]	**99.0**	**61.9**	**91.8**	**58.7**	50.6	39.2	55.3	**89.2**	23.2	94.3	**85.2**	2.9	**95.3**	0.0	**80.5**	**66.2**	**54.6**	55.2
HAPNet（本章）	98.6	59.3	91.5	57.0	**58.0**	**41.0**	**56.8**	87.8	**30.4**	**94.8**	81.8	**21.3**	94.2	0.0	69.7	49.9	43.8	**57.6**

表 3.4　在 MFNet 测试集上对不同视觉基础模型（VFMs）的消融实验（%）。"MM" 代表多模态预训练。BEiT 和 BEiTv2 通过掩码图像建模的自监督学习策略进行训练，而 DINOv2 则通过判别式对比学习的自监督学习策略进行训练

VFM	预训练类型	数据集	车辆	行人	自行车	弯道	停车标志	护栏	彩色锥	减速带	mAcc	mIoU
			IoU	IoU	IoU	IoU	IoU	IoU	IoU	IoU		
DeiT-B/16[143]	有监督训练	ImageNet-1K	90.4	75.2	66.6	47.6	36.3	0.7	**59.9**	**60.0**	68.3	59.4
AugReg-B/16[144]	有监督训练	ImageNet-22K	90.3	74.9	66.8	48.2	35.8	6.0	56.6	56.5	74.7	59.3
BEiT-B/16[101]	自监督训练	ImageNet-22K	**90.7**	75.0	**67.6**	48.6	39.1	4.2	55.0	59.0	72.5	59.7
DINOv2-B/14[103]	自监督训练	LVD-142M (MM)[103]	90.0	74.7	66.8	45.1	**42.1**	**9.7**	56.2	59.7	**76.4**	60.3
BEiTv2-B/16[102]	自监督训练	ImageNet-22K (MM)	89.1	**76.0**	67.2	**50.6**	41.3	7.9	58.6	58.7	75.1	**60.9**

表 3.6　在 MFNet 测试集上对不同数据输入策略的消融实验（%）

输入策略	VFM	CSPD	车辆	行人	自行车	弯道	停车标志	护栏	彩色锥	减速带	mAcc	mIoU
			IoU	IoU	IoU	IoU	IoU	IoU	IoU	IoU		
A	RGB + T	RGB + T	89.3	73.9	66.1	48.0	32.4	4.4	55.7	62.6	74.3	58.9
B	RGB + T	RGB	88.3	69.1	64.3	44.3	32.3	4.3	52.9	44.5	67.0	55.3
C	RGB + T	Thermal	85.5	72.7	54.6	39.3	28.5	**8.2**	46.1	55.4	71.5	54.2
D（采用策略）	RGB	RGB + T	89.1	**76.0**	**67.2**	**50.6**	**41.3**	7.9	**58.6**	58.7	**75.1**	**60.9**
E	RGB	RGB	87.9	62.5	63.8	40.7	28.2	5.6	52.1	47.1	66.7	53.9
F	RGB	Thermal	86.7	73.9	60.3	39.1	33.0	7.3	53.6	**59.0**	74.6	56.7
G	Thermal	RGB + T	**89.6**	74.6	65.1	48.9	33.9	1.5	53.9	56.8	68.8	58.1
H	Thermal	RGB	88.4	69.3	65.3	43.2	29.7	4.4	53.4	44.4	67.4	55.1
I	Thermal	Thermal	84.5	71.8	53.8	35.0	24.3	3.8	38.4	54.5	66.6	51.5

表 3.7　在 MFNet 测试集上对不同 CSPD 构建模块选择的消融实验（%）

编码器种类	车辆	行人	自行车	弯道	停车标志	护栏	彩色锥	减速带	mAcc	mIoU
	IoU	IoU	IoU	IoU	IoU	IoU	IoU	IoU		
并行 ResNet-101[63] 共享权重	66.3	48.0	38.4	19.9	9.1	0.0	29.6	23.6	42.3	36.8
并行 Swin-Transformer-S[48] 共享权重	**89.7**	74.1	67.2	47.6	31.0	6.3	52.7	57.0	74.6	58.2
并式 MiT-B4 共享权重[24]	87.8	74.7	64.9	48.4	40.9	**13.1**	51.4	48.1	73.2	58.6
并行 ConvNeXt-S[85] 独立权重	89.6	74.6	66.6	46.3	**45.7**	10.2	56.6	**62.5**	73.6	**61.2**
并行 ConvNeXt-S[85] 共享权重（本章）	89.1	**76.0**	**67.2**	**50.6**	41.3	7.9	**58.6**	58.7	**75.1**	60.9

为单模态版本 (仅使用 RGB 图像或热图像进行编码)，这两种策略分别仅获得了 53.9% 和 51.5% 的平均交并比 (mIoU)。相比之下，同时利用 RGB 图像和热红外图像的策略 A-D 和 F-H 均取得了更优异的性能，这验证了模态互补的重要性。值

表 3.8　在 MFNet 测试集上对 GLCA 和 CCG 有效性的消融实验（%），当两个组件都被移除时，使用 CSPD 提取的跨模态空间先验和使用 VFM 提取的全局上下文在分辨率对齐后通过逐元素相加的方式进行特征融合

GLCA	CCG	Fusion Strategies	车辆	行人	自行车	弯道	停车标志	护栏	彩色锥	减速带	mAcc	mIoU
			IoU	IoU	IoU	IoU	IoU	IoU	IoU	IoU		
		特征元素级相加	89.2	73.5	65.7	50.0	33.0	4.1	52.0	55.4	71.1	57.9
✓		局部语义增强	**89.9**	74.7	67.0	50.0	37.0	5.8	52.9	55.9	72.9	59.1
	✓	全局上下文增强	89.6	74.5	67.0	46.6	36.9	2.9	53.3	53.3	72.4	58.0
✓	✓	**以上两者（本章）**	89.1	**76.0**	**67.2**	**50.6**	**41.3**	**7.9**	**58.6**	**58.7**	**75.1**	**60.9**

表 3.9　在 MFNet 测试集上对不同对称以及非对称编码器架构的消融实验（%）

网络名称 (编码器名称)	车辆	行人	自行车	弯道	停车标志	护栏	彩色锥	减速带	mAcc	mIoU
	IoU	IoU	IoU	IoU	IoU	IoU	IoU	IoU		
并行式 BEiTv2-B/16[102] 特征串联	65.7	57.4	12.7	27.7	8.5	0.0	14.7	20.6	40.9	33.7
并行式 BEiTv2-B/16[102] 元素相加	65.4	57.1	23.3	28.0	8.5	0.0	19.9	16.2	42.6	35.0
并行式 ConvNeXt-S[85] 特征串联	89.7	74.1	66.4	45.3	29.8	4.8	52.8	55.5	69.2	57.4
并行式 ConvNeXt-S[85] 元素相加	**90.7**	74.6	66.9	46.8	32.9	6.6	56.6	52.7	72.0	58.5
HAPNet CMNeXt-B4[142]	88.5	74.0	62.5	44.9	41.1	**12.0**	53.3	**59.0**	71.6	59.3
HAPNet（本章）	89.1	**76.0**	**67.2**	**50.6**	**41.3**	7.9	**58.6**	58.7	**75.1**	**60.9**

得注意的是，策略 D(即将 RGB 图像输入 VFM 分支，同时在 CSPD 中输入 RGB-T 图像对) 取得了最高的 60.9% mIoU。这一结果表明，在 VFM 中编码热红外图像可能会导致特征退化并降低整体性能，这验证了我们在引言部分提出的假设。

我们提出的 CSPD 能够从 RGB 图像和热图像中提取多尺度的空间先验，并

将其与 VFM 的全局上下文特征进行融合，从而隐式地增强特征中的局部语义。为进一步探索 CSPD 的最优结构，我们尝试用其他主流层次化编码器结构替换基础的 ConvNeXt 模块。如表 3.7所示，ConvNeXt 结构展现出最优性能，达到 60.9% 的 mIoU，这一结果优于传统的 ResNet、Swin-Transformer 以及 Mix Vision Transformer (MiT-B4) 等架构，其中 MiT-B4 是此前另一个可使用 RGB-X 图像对作为输入的场景解析网络中采用的复杂结构。此外，我们研究了对两种模态采用独立权重的影响。实验表明，这种策略能带来轻微的性能提升。然而，考虑到精度提升与模型复杂度之间的权衡，我们仍然选择采用权重共享策略，因为权重分离策略仅带来 0.3% 的 mIoU 提升，却显著增加了模型参数量。

我们对 HAPNet 编码器中的重要组件：(1) GLCA 模块和 (2) CCG 模块的有效性进行了验证。如表 3.8所示，移除其中任一模块都会导致 HAPNet 性能显著下降。仅通过空间先验和全局上下文之间的逐元素求和来实现跨模态特征融合的基准设置，其 mIoU 比完整的 HAPNet 低 3.0%。

为验证我们提出的非对称架构的有效性，我们首先构建了基于 BEiTv2 和 ConvNeXt 的对称双编码器基准架构。实验结果显示，这些基准架构的性能均未能达到 HAPNet 的水平。具体而言，对称的 ConvNeXt 和 BEiTv2 架构的 mIoU 分别比 HAPNet 低 3.5% – 2.4% 到 27.2% – 25.9% 个百分点。这些发现进一步支持了我们在引言中提出的假设：对 RGB 图像和热图像采用非对称架构能够更好地发挥各自优势。值得注意的是，对称双编码器 BEiTv2 的性能不佳表明，直接将 VFM 应用于 RGB-T 场景解析任务可能并不合适。我们还将 HAPNet 的编码器与另一个表现优异的非对称编码器进行了比较。为公平起见，我们将 CMNeXt 的编码器与 HAPNet 的解码器结合（简称 HAPNet CMNeXt-B4），结果显示 mIoU 下降了 1.6%，这进一步凸显了我们的编码器设计在 RGB-T 场景解析任务中的优越性。

最后，我们评估了 HAPNet 与多个先进视觉基础模型的兼容性，包括通过传统监督学习训练的模型 (如 DeiT 和 AugReg) 以及基于自监督预训练开发的基础模型 (如 BEiT、BEiTv2 和 DINOv2)。如表 3.4所示，BEiTv2 在 RGB-T 场景解析任务中取得了最高的 60.9% mIoU，优于其他所有模型。基于 BEiTv2 的出色表现，我们将其确定为该任务最适合的视觉基础模型，并将其作为 HAPNet 中的默认 VFM。

3.5　本章小结

在本章中，我们对异构特征提取与融合过程中网络性能提升的关键进行了深入思考，当面对不同的多源信息输入时，应从其本身特性出发进行网络的设

计，才能更充分地发挥多源信息潜力，克服性能瓶颈。具体来说，我们通过分析 RGB 图像和深度图像、热图像等多源信息的固有特性，设计了一种基于 CNN 与 Transformer 混合架构的非对称异构特征提取网络，以充分发挥它们各自的优势，并在领域中率先引入了强大的视觉基础模型。我们首先提出了跨模态空间先验描述子，用于从 RGB-T 图像对中提取蕴含局部语义的空间先验。其次，我们设计了一个渐进式异构特征集成器，该集成器由全局-局部上下文聚合器和互补上下文生成器组成，用于更有效地融合异构特征。此外，我们引入了一个辅助任务来进一步增强融合特征中的局部语义，从而提升融合特征用于解码时的表现。大量实验表明，提出的 HAPNet 在 RGB-T 与 RGB-D 等双源信息融合场景解析任务中实现了同期领先的性能，表现出了网络在不同异构特征融合时的广泛适用性。

然而，尽管 HAPNet 相比现有方法在多个公开数据集中实现了最优秀的性能，但其泛化能力仍有提升空间，如在拥有更复杂多样语义种类对象的开放世界中，如何保证高泛化性的场景解析能力。此外，考虑到场景解析是自动驾驶汽车、移动机器人和无人机等系统中的常见功能，网络的实时性能至关重要，这也是未来领域内工作的一个重要方向。

第 4 章　基于视觉语言模型的开放词汇道路场景解析网络

前两章通过设计不同的异构特征融合策略设计了两种高性能的道路场景解析网络，有效解决了现有方法在特征融合与编码过程中的局限性，所提出的网络架构也在相关人物的权威公开数据集中达到了同期领先的性能。然而，传统数据驱动方法（包括前两章提出的网络）普遍存在网络适用性受限的问题——其性能高度依赖在特定数据集上的有监督微调，当面对实际开放世界中的类别对象分布偏移，即出现了新颖或在训练中未见过的语义类别时，网络性能将显著下降。这一局限性严重制约了智能驾驶系统在真实复杂环境中的实际应用价值。为突破这一限制，本章提出了一种具有开放词汇能力的多源信息融合道路场景解析网络，该网络通过在大规模数据集上进行一次微调即可实现任意场景下对新颖语义类别的准确预测。该突破性能力源自对多模态学习领域中的视觉-语言基础模型（Vision-Language Model，VLM）[86] 的有效利用：此类模型通过基于对比学习[150] 的预训练范式，构建视觉与文本特征对齐的跨模态嵌入空间，通过点积运算即可得到视觉特征与各类别文本描述之间的语义相关性。为实现像素级的细粒度解析，我们引入了适配器网络对基础模型进行任务适配，同时融合 RGB 图像与源自视觉基础模型的先验深度图像，显式地增强网络对目标对象空间几何关系的理解能力。这种多源信息融合机制通过双任务分支架构实现：一方面利用深度线索强化物体边界感知，提升掩膜预测的几何精度；另一方面借助视觉-语言模型的跨模态（跨源）对齐能力，实现开放词汇场景下的语义推理。通过在多个开放词汇基准数据集上的实验，我们验证了本章所提出的基于多源信息融合的开放词汇场景解析网络的有效性。实验结果表明，该网络显著具备开放词汇语义分类能力，同时相较于仅使用 RGB 图像作为输入的开放词汇场景解析网络拥有更强的任务表现，在同期网络中实现了领先的性能。

4.1　引言

近年来，基于深度学习的场景解析方法虽然取得了显著的进步，但其中大多数方法仍停留在“封闭词汇”框架下，即需要在预定义类别的数据集上进行监督微调，以便准确预测特定语义类别的物体；当推理时目标语义类别发生变化，或仅仅是数据集切换时，网络的性能可能发生显著的退化。例如，当网络在仅包含“行人”和“可行驶区域”等类别的训练集上微调后，在推理时若需要预测“人类”或“道路”等语义类别，网络便无法完成相应任务。究其原因，是因为封闭

词汇网络的训练目标仅局限于特定数据集中预定义类别与视觉特征之间的简单对应关系，网络实际上无法理解视觉特征与文本语义之间的内在关联，更不能有效区分诸如“行人”与“人类”或“大巴”与“公交车”等文本描述之间的细微语义差异。与之相反，人类在此类开放词汇分类任务中几乎不受困扰。这是由于人类能够直观地感知不同文本语义之间的相似性，并将之有效地对应到视觉图像上。这种能力源于人类长期大量学习形成的丰富语义知识和整体观念。因此，为使深度网络获得类似于人类的开放词汇视觉理解能力，我们需要构建视觉特征与语言文本中语义特征的对应关系，使网络能够明确地学习到诸如“行人”与“人类”的高度语义相似性，以及“汽车”与“行人”之间显著的语义差异。

近年来，已有一些重要工作[86, 151] 尝试解决上述挑战。其中最具代表性的工作是 Contrastive Language-Image Pre-training（CLIP）[86]，该研究创新性地提出了视觉-文本跨模态对比学习范式，有效地实现了视觉与文本语义的对齐。CLIP 通过构建千万数量级的图像-文本对的大规模数据集，涵盖尽可能广泛的视觉概念，最终成功地学习到了 RGB 视觉图像与语言文本之间的语义对齐关系。

在场景解析领域，近期的研究[152] 利用 CLIP 的开放词汇分类能力，将其应用在了场景解析网络中。如研究[153] 首次尝试了利用特征池化（Feature Pooling）操作使网络学习区域级的视觉特征-文本特征对齐以实现开放词汇场景解析；研究[154] 提出了两段式的开放词汇场景解析架构，首先生成类别无关（Class Agnostic）掩膜，并将该掩膜对应的掩码图像送入 CLIP 网络进行图像分类；除了上述基于掩膜分类范式的方法，还有一些基于逐像素分类范式的研究，如[155, 156] 提出使用代价聚合的方式对像素-文本特征对构建代价体（Cost Volume），并且通过代价聚合得到逐像素的语义分类结果。尽管上述方法在标准场景下表现良好，但其视觉特征的建模仅依赖 RGB 图像，在低光照、复杂背景等挑战性场景中易出现性能波动。

另一方面，正如前文所述，现有的专用道路场景解析网络虽然能够利用来自多源信息作为输入，提取异构视觉特征（对应前章节的异构特征）并融合，在具有挑战性的场景中实现了相比于基于仅 RGB 图像的单模态网络更精确且鲁棒的解析效果；但这些网络[16, 17] 缺乏开放词汇识别能力，需要在具有预定义类别的数据集上进行监督微调，才能实现对特定语义类别物体的准确预测，在一定程度上限制了其应用场景。为了弥补领域内的研究空白，为多源信息融合网络去除这种限制，需要构建具有开放词汇能力的道路场景解析网络，但其中面临着许多困难。(1) 一个比较直观的办法是直接以 CLIP 式的对比学习策略建模异构视觉特征与文本特征的对齐，然而深度图像、热图像等多源信息与 RGB 图像具有明显的域差异，只能通过预训练从头进行训练，而对于这种规模的预训练，若想要模型达到 CLIP 的开放词汇能力，其所需要的计算资源（通常需数千 GPU 小时）与

海量视觉-文本多模态配对数据集是寻常研究无法承受以及拥有的。(2) 另一种方式是利用已经进行过对比学习预训练的 CLIP 模型，并用其视觉-文本特征对齐的约束指导异构视觉特征之间的对齐（如 RGB 图像中的色彩纹理特征与深度图像中的几何特征之间的对齐），这种方法虽然避免了从零开始的预训练，但仍然依赖于精细设计的异构视觉特征融合与特征对齐策略[157, 158]。此外，在 CLIP 模型中引入一种新的异构视觉特征参与原有的特征分布对齐，都会对 CLIP 模型原有的色彩纹理特征与文本特征之间的对齐造成损害，导致其开放词汇性能的灾难性下降，目前尚无开源可复现的相关研究。

面对以上难题，本章节首先深入探究了构建信息融合开放词汇场景解析网络的核心问题。我们通过实验发现，现有基于异构视觉特征融合的范式并不适用于开放词汇任务，并探究了现有方法不奏效的关键所在。具体地说，开放词汇场景解析任务的关键在于开放词汇下语义分类的准确性；如前章节 3所述，深度图像更加侧重对于空间几何特征与物体的轮廓等局部特性的表示，几何特征本身与色彩纹理特征、文本特征存在较大的域差异，简单地沿用现有封闭词汇信息融合网络的做法——即提取几何特征，然后将色彩纹理特征与几何特征进行异构特征融合，再进行开放词汇语义场景解析的范式，不可避免地会在一定程度上破坏 CLIP 原有的视觉-文本特征对齐能力，最终导致 CLIP 模型的开放词汇分类性能下降。针对此问题，本章节巧妙地从掩膜分类的网络范式进行入手，将开放词汇场景解析任务解耦成掩膜预测子任务与掩膜分类子任务，以 RGB 图像与估计得到的先验深度图像作为网络的视觉输入，构建了一个具有开放词汇能力的道路场景解析网络 CLIDA。在该网络中，我们将从深度图像与 RGB 图像提取并融合得到的异构融合特征专门用于掩膜预测子任务，从而获得更加鲁棒、精确的类别无关掩膜。随后，利用这些高精度掩膜对 CLIP 从 RGB 图像中提取的特征（以下称为 CLIP 视觉特征）进行掩膜池化操作，并利用其与 CLIP 从语言文本中提取的 CLIP 文本特征之间的对齐特性完成掩膜分类子任务。两个分支分别使用不同的损失函数进行监督，同时我们构建了梯度解耦策略确保两个子任务的损失不会相互干扰。通过这种精心设计的架构，我们既充分利用异构特征实现了更精确的掩膜预测，又有效保证了 CLIP 模型原有的开放词汇能力不发生灾难性遗忘，从而实现了更加鲁棒、精确的开放词汇场景解析性能。

总结而言，本章工作的主要贡献在以下几个方面：

(1) 本章节通过实验发现了传统封闭词汇网络的异构视觉特征融合范式在开放词汇任务中并不适用，其原因是因为现有的异构特征融合策略会破坏 CLIP 本身通过大数据预训练建立的视觉-文本特征对齐特性，从而造成开放词汇性能下降；

(2) 针对上述发现与结论，本章节基于任务解耦的思想设计了一种开放词汇场

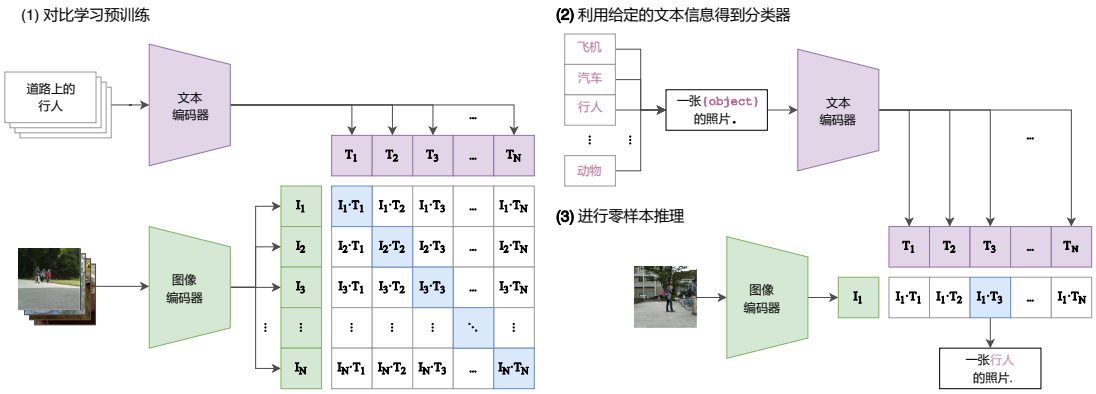


图 4.1　视觉-语言模型 CLIP 的算法原理示意图

景解析网络 CLIDA，该网络以 RGB 图像与先验深度图像提取并融合得到的异构视觉特征用于掩膜预测子任务，并结合该掩膜与仅从 RGB 图像中提取的 CLIP 视觉特征进行开放词汇下的掩膜分类，避免了对 CLIP 特征嵌入空间的破坏；

(3) 本章节通过系统的消融实验证明了所提出网络的有效性，并与现有方法在领域内标准的开放词汇场景解析任务设定下进行了对比实验，实现了同期领先的性能。

本章节提出的开放词汇场景解析网络以 CLIP 模型为基础架构。为便于理解，本章首先在 4.2节对以视觉语言模型 CLIP 为基础的开放词汇分类及场景解析模型中的核心思想进行简要介绍；随后在 4.3节详细说明本章所提出的信息融合开放词汇场景解析网络进行详细介绍； 4.4节将通过消融实验以及对比实验验证所提出网络 CLIDA 的性能，最后，在第4.5节对本章进行总结。

4.2　基于对比学习范式的视觉语言模型 CLIP

传统图像分类网络通常依赖于大规模标注数据集进行监督预训练，如基于 ImageNet[100] 训练的 ResNet[63] 和基于 JFT-300M[159] 训练的 ViT 等。这类方法存在两个显著局限性：首先，高质量标注数据的获取成本高昂；其次，模型的任务迁移能力和泛化性能受限。相比之下，互联网上广泛存在的图像-文本对数据为模型训练提供了更经济高效的选择。OpenAI 团队通过收集 4 亿规模的图像-文本对数据集，创新性地采用对比学习范式训练 CLIP 模型，其核心流程可归纳如下：

(1) 给定批次大小为 N 的图像-文本对，分别通过基于 Transformer[23] 的文本编码器和基于 CNN/Transformer 的图像编码器进行处理。其中，文本编码器将每个文本转换为维度为 D 的特征向量，得到文本嵌入输出 $\boldsymbol{E}_{text} = [\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_N] \in \mathbb{R}^{N \times D}$；图像编码器则输出对应的图像嵌入输出 $\boldsymbol{E}_{image} = [\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_N] \in \mathbb{R}^{N \times D}$；

(2) 在特征嵌入空间中，仅保留索引相同的 $\boldsymbol{I}_i$ 与 $\boldsymbol{T}_i$ 作为正样本对，其余 $N^2 - N$ 个跨模态组合作为负样本对。通过这种对比学习机制，CLIP 模型能够将不同模态的特征映射到统一的特征嵌入空间，实现跨模态特征对齐；

(3) 具体而言，计算 $\boldsymbol{E}_{image}$ 和 $\boldsymbol{E}_{text}$ 的矩阵乘积并进行归一化处理，得到图像-文本对的余弦相似度矩阵。训练目标是最小化以下对比损失函数：

$$min(\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\boldsymbol{I}_i \cdot \boldsymbol{T}_j\right)_{(i\neq j)} - \sum_{i=1}^{N}\left(\boldsymbol{I}_i \cdot \boldsymbol{T}_i\right)) \tag{4.1}$$

如图 4.1所示，该目标函数旨在最大化对角线中蓝色的元素（N 对正样本）的相似度，同时最小化其它非对角线元素（$N^2 - N$ 对负样本）的相似度。

经过预训练后，CLIP 模型具备了强大的跨模态特征对齐能力，可直接应用于零样本图像分类任务，其推理流程如下：

(1) 根据目标任务构建语义类别描述。以 ImageNet 数据集为例，将 1000 个类别词嵌入预定义模板，如下方所示：

```
A photo of {label};
```

以增强语义表达，并通过文本编码器获得文本嵌入矩阵 $\boldsymbol{E}_{text} = [\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_N]$（$N = 1000$）；

(2) 将待分类图像输入图像编码器获得特征向量 $\boldsymbol{I}$，计算其与所有文本特征的相似度得分，选取最大相似度对应的类别作为最终预测结果（例如上图对应的 $\boldsymbol{T}_3$）。

4.2.1　开放词汇场景解析任务

任务定义　开放词汇场景解析任务旨在将输入图像 $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$ 解析为一组语义掩膜及其对应的类别标注：

$$\{y_i\}_{i=1}^{K} = \{(m_i, c_i)\}_{i=1}^{K}, \tag{4.2}$$

其中，K 表示真实掩膜数量，$m_i \in \{0,1\}^{H\times W}$ 为二值化掩膜，c_i 为对应的类别标签。在训练过程中，使用固定的类别集合 C_{train}，而在推理阶段则使用另一个类别集合 C_{test}。在开放词汇场景下，C_{test} 可能包含训练时未见过的新类别，即满足 $C_{\text{train}} \neq C_{\text{test}}$。与现有研究[160] 等一致，我们假定在测试时可以获得类别集合 C_{test} 的自然语言文本表述。

两阶段方法　现有研究[154, 161] 普遍采用两阶段框架解决开放词汇场景解析问题。该框架由以下两个主要组件构成：第一阶段包括一个类别无关的掩膜生成网络

$\mathcal{M}$，其网络参数记作 $\theta_{\mathcal{M}}$，其将输入图像 $\boldsymbol{I}$ 映射为含 N 个掩膜提议 $\{\hat{m}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times H\times W}$：

$$\{\hat{m}_i\}_{i=1}^{N} = \mathcal{M}(\boldsymbol{I};\theta_{\mathcal{M}}), \tag{4.3}$$

第二阶段中，一个 CLIP 适配网络 $\mathcal{P}$ 同时接受图像 $\boldsymbol{I}$ 和掩膜提议输出 $\{\hat{m}_i\}_{i=1}^{N}$ 作为输入，后者用于指导参数冻结的 CLIP 模型（记作 $CLIP^*$，星号表示模型冻结），对掩膜裁剪（crop）后的图像区域进行分类[160, 161]：

$$\{\hat{c}_i\}_{i=1}^{N} = \mathcal{P}(\boldsymbol{I}, \{\hat{m}_i\}_{i=1}^{N}; CLIP^*), \tag{4.4}$$

其中，$\{\hat{c}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times|C|}$ 表示 N 个掩膜对应的类别概率分布，$C \in \{C_{\text{train}}, C_{\text{test}}\}$ 为当前阶段的类别集合，$|C|$ 代表集合的大小。

尽管该框架在开放词汇场景解析任务中取得了显著进展，但仍存在以下两个主要局限性：首先，CLIP 视觉特征需要被重复提取两次，一次用于掩膜预测，另一次用于掩膜分类。导致计算开销加倍，限制了更大规模 CLIP 模型的使用；其次，掩膜生成网络通常需要高分辨率输入（如 1024×1024）以得到高精度的掩膜预测，而 CLIP 模型通常使用低分辨率图像（如 224×224）进行预训练，这种分辨率差异导致特征适配困难，降低了整体效率。

单阶段方法　为克服上述限制，有研究提出将掩膜生成网络与掩膜分类网络合并为一个统一的单阶段框架 $\mathcal{F}$，二者共享同一个 CLIP 预训练的骨干网络（记为 CLIP，此时未冻结）以提取图像 $\boldsymbol{I}$ 的特征：

$$\{\hat{m}_i, \hat{c}_i\}_{i=1}^{N} = \mathcal{F}(\boldsymbol{I}; CLIP, \theta_M), \tag{4.5}$$

然而，研究[162] 表明，直接微调这种简易的单阶段框架，会导致预训练的 CLIP 模型中视觉与文本特征之间特征对齐能力大幅损失，进而导致性能明显下降，尤其是在处理未见过的新类别时表现尤为不佳，称之为灾难性遗忘。此外，训练成本也会显著增加。这一问题的根源在于基于 ViT 架构[76] 的 CLIP 模型对输入尺度的泛化能力有限。尽管部分研究通过引入适配器模块[125, 163] 或代价聚合机制[164] 缓解了这一问题，但仍未完全解决。相比之下，基于 CNN 架构的 CLIP 模型（如 ResNet[63] 和 ConvNeXt[85]）由于全卷积结构[27] 的特性，对不同分辨率输入具有更好的适应性。此外，CNN 骨干网络能够提取多尺度特征，使其能够无缝集成到现有的掩膜分类框架中[56]，为开放词汇场景解析提供了更灵活的解决方案。

4.3　基于掩膜分类范式的开放词汇场景解析网络

如图 4.2所示，本章节所提出的开放词汇场景解析网络 CLIDA 由三个关键部分构成：(1) 基于视觉基础模型的深度估计网络（双源信息输入）、(2) 基于 CLIP 模型的异构特征编码网络以及 (3) 基于掩膜分类范式的开放词汇解码网络。与现有的开放词汇场景解析网络类似，CLIDA 网络首先接受长和宽分别为 H、W 的 RGB 图像 $\boldsymbol{I}^R \in \mathbb{R}^{H\times W\times 3}$ 与语言文本信息输入 CLIP 模型，所得到的 CLIP 视觉特征与 CLIP 文本特征用于后续的开放词汇分类任务。此外，本方法在传统 RGB 图像与语言文本两种模态输入的基础上，创新性地引入深度图像（也可使用其他信息，本章节的后续实验中都使用深度图像作为默认的另一种信息）作为视觉输入 $\boldsymbol{I}^X \in \mathbb{R}^{H\times W\times 3}$，通过异构视觉特征融合增强网络对于局部语义的感知能力，从而提升复杂场景下的解析鲁棒性。$\boldsymbol{I}^X$ 可通过雷达、深度相机、红外相机等传感器获取。然而，考虑到异构传感器的高昂成本与标定难度等问题，在大多数开放词汇任务场景中，这类信息往往难以直接获得。为此，本章采用深度估计方法获取 $\boldsymbol{I}^X$。鉴于任务场景的复杂多变性，我们选择了具有卓越零样本推理能力的视觉基础模型 Depth Anything V2[165] 作为深度估计网络，以确保在开放环境下获得精确的先验深度图像。需要指出的是，由于深度估计网络并非本章节的研究重点，因此后续将不再对其网络结构进行详细阐述。在异构视觉特征融合阶段，通过 CLIP 适配器网络进行色彩纹理特征与几何特征进行高效的融合，进而利用所得的异构融合特征进行掩膜预测子任务，生成掩膜提议预测。随后，结合 CLIP 视觉特征进行掩膜池化操作，并通过计算池化后的 CLIP 视觉特征与 CLIP 文本特征之间的余弦相似度，完成掩膜分类子任务，实现逐掩膜分类。最终，将两个子任务的输出进行整合，得到场景解析预测结果。本节后续将详细介绍所提出的 CLIDA 特征编码网络（见第 4.3.1节）以及开放词汇解码器网络（见第 4.3.3节）。

4.3.1　基于 CLIP 适配器网络的异构特征编码

CLIDA 的特征编码网络主要由异构视觉特征编码和文本特征编码两个分支组成。在视觉特征编码分支中，对于输入的 RGB 图像 $\boldsymbol{I}^R$，首先利用 CLIP 图像编码器提取用于掩膜分类任务的 CLIP 视觉特征。随后，结合其对应的深度图像 $\boldsymbol{I}^X$ 编码得到的几何特征，并通过适配器网络实现 CLIP 视觉特征的增强，以用于掩膜预测子任务。具体而言，$\boldsymbol{I}^R$ 被输入基于 ConvNeXt 架构[85] 的 CLIP 图像编码器，提取多尺度的 CLIP 视觉特征，表示为 $\mathcal{F}^R = \{\boldsymbol{F}_i^R\}, i \in [0,1,2,3]$。对应的特征图 $\{\boldsymbol{F}_0^R, \boldsymbol{F}_1^R, \boldsymbol{F}_2^R, \boldsymbol{F}_3^R\}$ 相对于 $\boldsymbol{I}^R$ 的步长分别为 $\{4,8,16,32\}$；在 $\mathcal{F}^R$ 中，$\boldsymbol{F}_3^R$ 作为 CNN 编码器最后一层的输出，具有最高的语义等级，因此在现有开放词汇场景解析网络框架中[166]，$\boldsymbol{F}_3^R$ 被广泛用于掩码图像的分类任务。本章节沿用了这一

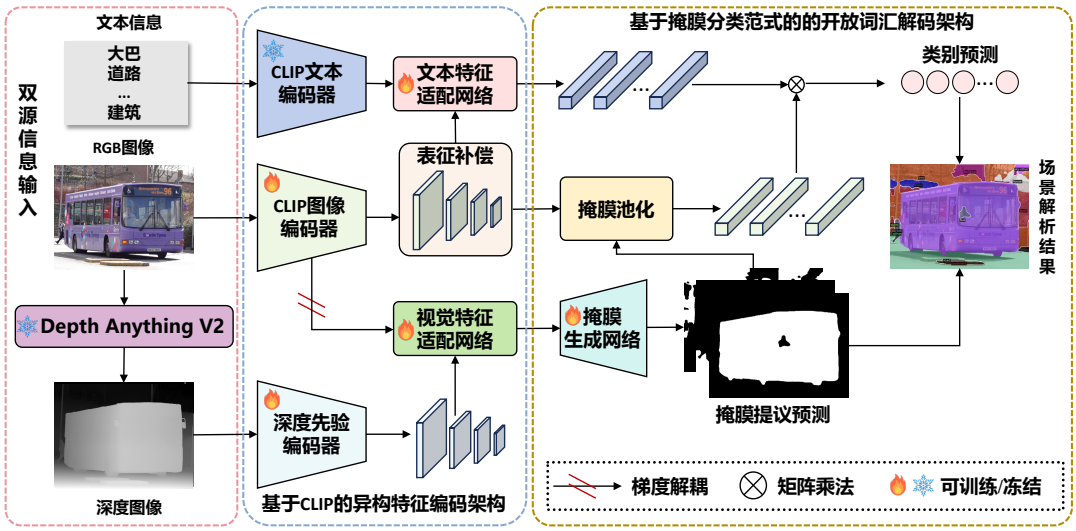


图 4.2　本章所提出的开放词汇场景解析网络 CLIDA 的结构示意图

设定。另一方面，将 $\boldsymbol{I}^X$ 输入同样基于 ConvNeXt 架构的深度先验编码器中，得到多尺度几何特征 $\mathcal{F}^X = \{\boldsymbol{F}_i^X\}$，其中 $i \in [0,1,2,3]$。随后，将来自 RGB 图像与深度图像的两个编码器输出 $\mathcal{F}^R$ 和 $\mathcal{F}^X$ 送入视觉特征适配网络进行异构特征融合，以提升任务表现。

值得注意的是，本章节在 CLIP 图像编码器输出到视觉特征适配网络之间引入了梯度解耦机制。该机制旨在将掩膜预测任务损失到 CLIP 模型之间的梯度从网络计算图中剔除，实现掩膜预测子任务与掩膜分类子任务在梯度层面的任务解耦。这一设计有效避免了不同子任务损失对网络参数的相互影响，确保在提升掩膜预测表现的同时，不会对 CLIP 模型本身视觉-文本特征之间的对齐特性造成灾难性遗忘，从而损害掩膜分类子任务的性能。本章后续将通过详细的消融实验验证所提出机制的有效性及设计原因。

异构视觉特征编码　在此前章节 3，我们已经通过实验验证了结合 VFM 与 CNN 架构的异构特征混合编码架构能够实现更强大的异构特征编码与融合，从而显著提升场景解析任务的表现。基于这一发现，本章旨在通过双源信息进一步增强 CLIP 视觉特征中的局部语义。为此，我们设计了如图 4.3所示的混合编码架

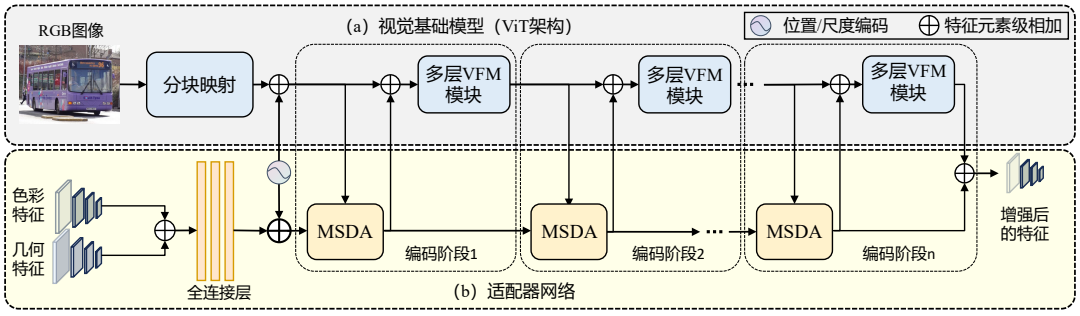


图 4.3　用于增强掩膜预测子任务的视觉特征适配网络架构图

构，作为异构视觉特征 $\mathcal{F}^R$ 和 $\mathcal{F}^X$ 的适配网络。该架构主要由两个分支组成：(1) VFM 分支与 (2) 异构特征适配分支。与上一章节中的架构类似，首先将 RGB 图像 $\boldsymbol{I}^R$ 输入分块映射层，并添加位置编码，得到编码阶段 1 中的全局上下文特征 $\boldsymbol{F}_1^V$。在异构特征适配分支中，输入特征 $\mathcal{F}^R$ 和 $\mathcal{F}^X$ 在对应尺度上分别进行特征通道数量的统一，得到多尺度的异构特征 $\boldsymbol{F}_i^H \in \mathbb{R}^{\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times D}$，其中 $i \in [0,1,2,3]$。这一过程可通过以下公式表示：

$$\underset{1\times1}{\text{Conv}}(\boldsymbol{F}_i^R + \boldsymbol{F}_i^X), \tag{4.6}$$

随后，我们对特征步长为 $\{8,16,32\}$ 的三个尺度的特征进行展平并串联，得到异构空间先验 $\boldsymbol{F}_1^P$。在后续的每个编码阶段中，$\boldsymbol{F}_i^P$ 与 $\boldsymbol{F}_i^V$ 不断进行融合。具体而言，$\boldsymbol{F}_1^P$ 作为查询 Q，$\boldsymbol{F}_i^V$ 作为键和值，通过多尺度可变形交叉注意力机制（MSDA）将 $\boldsymbol{F}_1^P$ 中的丰富局部语义注入 $\boldsymbol{F}_i^V$，从而实现局部语义的增强。该过程可表述为：

$$\hat{\boldsymbol{F}}_i^V = \boldsymbol{F}_i^V + \kappa_i \text{MSDA}(\text{LN}(\boldsymbol{F}_i^V), \text{LN}(\boldsymbol{F}_i^P)), \tag{4.7}$$

其中 $\text{LN}(\cdot)$ 和 $\text{MSDA}(\cdot,\cdot)$ 分别表示层归一化和多尺度可变形交叉注意力，κ_i 为可学习系数，二者的定义与选择原因分别可见章节 1.4.1.4和 3.3.1，此处不再赘述。随后，$\hat{\boldsymbol{F}}_i^V$ 通过多层 VFM 模块进行全局上下文特征编码，以建模特征间的长距离依赖关系，并生成下一个编码阶段的输入 $\boldsymbol{F}_{i+1}^V$。经过四个编码阶段的特征编码后，将最终得到的 $\boldsymbol{F}_5^V$ 进行展平并恢复其原始的空间分辨率。对于步长为 4 的特征 $\boldsymbol{F}_0^H$，我们采用与上一章节相同的处理方式进行上采样，最终得到四个尺度的异构融合特征 $\mathcal{F}^F$。该过程的具体步骤与上一章节一致，此处不再详细展开说明。

文本特征适配网络　CLIDA 的文本特征编码分支由 CLIP 文本编码器以及文本特征适配网络组成，其编码过程如下：给定类别名称集合 $C = C_1, C_2, \ldots, C_n$，首先利用预定义模板[167]（见表4.1）生成与这些类别对应的描述性句子，例如：“a photo of a C_i”，“There is a C_i in the scene...”等。这些句子随后被输入 CLIP 文本编码器，生成每个句子的文本嵌入 $\boldsymbol{e}_{text}$。对于同一类别的不同句式生成的文本嵌入，我们取其平均值，最终得到所有类别的文本嵌入矩阵（$\boldsymbol{E}^T \in \mathbb{R}^{D\times|C|}$），其中

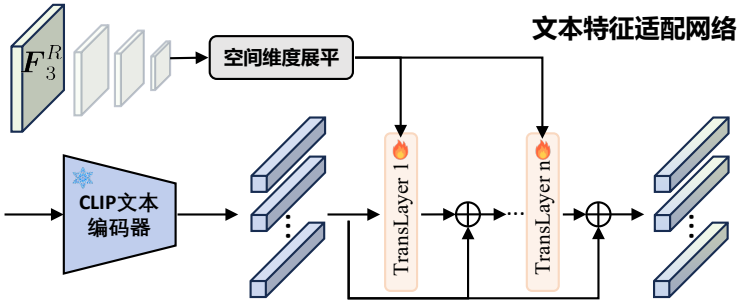


图 4.4　文本特征适配网络架构示意图

表 4.1　本章网络 CLIDA 在实验中所使用的句式模板

模板
“a photo of a { }.”
“This is a photo of a { }”
“There is a { } in the scene”
“There is the { } in the scene”
“a photo of a { } in the scene”
“a photo of a small { }.”
“a photo of a medium { }.”
“a photo of a large { }.”
“This is a photo of a small { }.”
“This is a photo of a medium { }.”
“This is a photo of a large { }.”
“There is a small { } in the scene.”
“There is a medium { } in the scene.”
“There is a large { } in the scene.”

D 为嵌入维度，$|C|$ 为类别数量。为了进一步增强 CLIP 文本编码器输出的表征 $\boldsymbol{E}^T$，我们参考了研究[166] 中的“上下文依赖迁移”方法，构建了文本特征适配网络。该网络通过一系列 Transformer 层对 CLIP 视觉特征与 CLIP 文本特征进行跨模态交叉注意力计算，利用 CLIP 视觉-文本特征之间的对齐特性进行特征融合，从而使最终输出的文本特征对 CLIP 视觉特征具有更强的响应能力，进而提升掩膜分类子任务的表现。如图 4.4所示，该网络以 CLIP 视觉编码器输出的最后阶段特征 $\boldsymbol{F}_3^R$ 和文本嵌入 $\boldsymbol{E}^T$ 作为输入。首先，对 $\boldsymbol{F}_3^R$ 进行空间维度的展平操作，以匹配文本特征的形状，得到 $\boldsymbol{F}_f^R \in \mathbb{R}^{D\times\frac{HW}{32^2}}$；接着，使用 n 个连续的 Transformer 层对 $\boldsymbol{E}^T$ 和 $\boldsymbol{F}_f^R$ 进行处理，并引入了残差连接。具体过程可描述为：

$$\boldsymbol{E}_{i+1}^T = \text{TransLayer}_i(\boldsymbol{E}_i^T, \boldsymbol{F}_f^R) + \boldsymbol{E}_i^T, \quad i = 1, 2, \ldots, l \tag{4.8}$$

在本章节的默认设定中，$l = 2$。文本特征适配网络的最终输出为经过特征适配后的文本嵌入 $\hat{\boldsymbol{E}}^T$。公式 4.8中的 Transformer 层采用了标准注意力机制，其定义如下：

$$\text{TransLayer}(\boldsymbol{X}^q, \boldsymbol{X}^f) = \text{Softmax}\left(\frac{\text{Que}(\boldsymbol{X}^q)\cdot\text{Key}(\boldsymbol{X}^{f\top})}{\sqrt{D}}\right)\cdot\text{Val}(\boldsymbol{X}^f), \tag{4.9}$$

式中的 Que(·)、Key(·) 及 Val(·) 分别代表将特征映射至 $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ 的线性投影函数（关于映射矩阵的定义及注意力机制的详细过程可参阅章节 1.4.1.2），D 为特征维度。此外，公式 4.9中省略了层归一化。值得注意的是，在训练过程中，我们保持 CLIP 文本编码器的参数冻结，仅对适配器网络的参数进行优化。这种设计

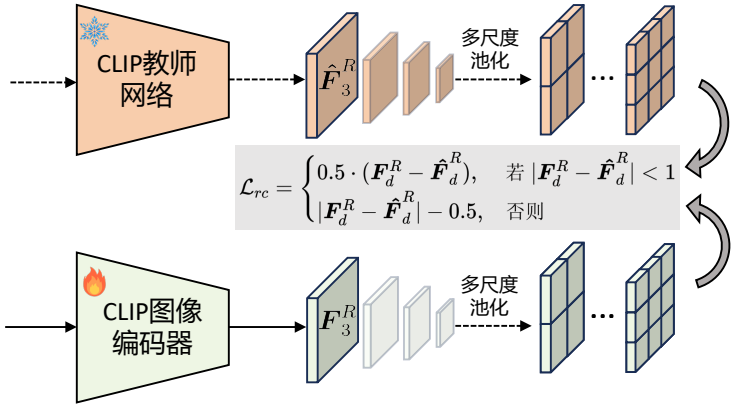


图 4.5　基于特征蒸馏的表征补偿模块流程示意图

不仅确保了文本编码分支优化的高效性，还使得 $\hat{\boldsymbol{E}}^T$ 对于 CLIP 视觉特征 $\mathcal{F}^R$ 的响应更加准确且灵敏，从而提升掩膜分类子任务的表现。

CLIP 网络的表征补偿　正如前文所述，尽管通过梯度解耦的方式在一定程度上避免了 CLIP 网络因掩膜预测子任务的梯度更新而导致其视觉-文本特征对齐特性被破坏，但在开放词汇场景解析网络的训练过程中，掩膜分类子任务的训练目标为像素级精度的掩膜分类，这会导致 CLIP 图像编码器原本与文本特征图像级对齐的输出表征分布发生偏移。此外，基于注意力的掩膜池化操作也会对原有对齐的高维特征产生破坏，从而影响掩膜分类的性能。因此，本章节引入了基于特征蒸馏思想的表征补偿（Representation Compensation, RC）策略。该策略在场景解析领域已有成功的应用案例[166]，其核心思想是在训练过程中利用参数冻结的教师 CLIP 网络对参与训练的 CLIP 图像编码器（学生网络）进行监督，以缓解表征分布偏移的问题。具体实现如图 4.5所示。在 RC 策略中，我们利用冻结的 CLIP 图像编码器（称为 CLIP 教师网络）在训练阶段编码原始的 CLIP 视觉特征。从 CLIP 教师网络和学生网络中分别提取最后阶段的输出特征 $\hat{\boldsymbol{F}}_3^R$ 与 $\boldsymbol{F}_3^R$，并通过特征级约束使两者的输出特征尽可能接近，从而避免灾难性遗忘。然而，现有研究[166] 表明，直接进行逐元素的特征对齐效果有限，因为其无法有效监督区域级特征的差异。因此，我们采用多尺度的平均池化（AvgPooling）方法，生成多尺度特征，并对池化后的特征进行一致性约束。给定任意特征 $f \in \mathbb{R}^{d\times h\times w}$，使用网格尺寸为 $k\times k$ 的平均池化操作可以表示为：

$$f^{pool} = \text{AvgPooling}(f, k), \quad f^{pool} \in \mathbb{R}^{d\times k\times k} \tag{4.10}$$

在默认设置下，我们选择 $K = 1,2,4$ 对 $\hat{\boldsymbol{F}}_3^R$ 和 $\boldsymbol{F}_3^R$ 进行池化操作，分别得到 $1\times 1, 2\times 2, 4\times 4$ 尺度网格池化的特征，记为 $\hat{\boldsymbol{F}}_d^R$ 与 $\boldsymbol{F}_d^R$。具体地：$\hat{\boldsymbol{F}}_d^R = \text{AvgPooling}(\hat{\boldsymbol{F}}_3^R, K)$，$\boldsymbol{F}_d^R = \text{AvgPooling}(\boldsymbol{F}_3^R, K)$。随后，我们使用平滑 L1 损失函

数（Smooth L1）最小化两者之间的差异：

$$\mathcal{L}_{rc} = \text{Smooth L1}(\boldsymbol{F}_d^R, \hat{\boldsymbol{F}}_d^R), \tag{4.11}$$

其中，平滑 L1 损失函数的定义为：

$$\text{Smooth L1}(\boldsymbol{F}_d^R, \hat{\boldsymbol{F}}_d^R) = \begin{cases} 0.5 \cdot (\boldsymbol{F}_d^R - \hat{\boldsymbol{F}}_d^R), & \text{若 } |\boldsymbol{F}_d^R - \hat{\boldsymbol{F}}_d^R| < 1 \\ |\boldsymbol{F}_d^R - \hat{\boldsymbol{F}}_d^R| - 0.5, & \text{否则} \end{cases} \tag{4.12}$$

通过 RC 策略对 $\boldsymbol{F}_3^R$ 的原始 CLIP 视觉特征进行补偿，使得 CLIP 图像编码器在微调过程中能够保持其零样本泛化能力。最后，我们对 $\boldsymbol{F}_3^R$ 采用掩膜池化（Mask Pooling）[162]，为每个掩膜提议预测在空间上对应的 CLIP 视觉特征生成对应的掩膜特征嵌入（$\boldsymbol{F}^M \in \mathbb{R}^{N \times D}$）。

4.3.2　掩膜生成网络

借鉴了现有开放词汇场景解析研究[167]，本章节同样采用 MaskFormer[55, 56] 网络架构作为掩膜生成网络，用于掩膜预测子任务。由于训练过程中采用匈牙利匹配算法[80] 进行标签-预测的最优匹配，因此仅有少量掩膜提议会得到优化。这种匹配策略显著提高了提议生成器的泛化能力，使其能够更有效地生成新类别掩膜。给定异构融合特征 $\mathcal{F}^F$，掩膜生成网络输出数量为 N 的掩膜提议 $\mathcal{M} = \{\boldsymbol{M}_i\}_{i=1}^N \in \mathbb{R}^{N \times \frac{H}{4} \times \frac{W}{4}}$。需要注意的是，在训练阶段，我们对 CLIP 图像编码器到视觉特征适配网络之间的梯度传播进行了截断，以防止掩膜提议预测的损失对 CLIP 模型特征的影响。

4.3.3　基于掩膜分类范式的开放词汇解码器

当获得掩码提议后，即可通过对比学习的方式利用掩膜特征嵌入 $\boldsymbol{F}^M$ 与文本嵌入 $\hat{\boldsymbol{E}}^T$ 进行掩膜分类子任务。具体而言，开放词汇分类器的类别概率预测定义为：$\forall i = 1, \ldots, N$

$$S_{cls} = softmax(\frac{1}{T}\left[cos(\boldsymbol{F}_i^M, \hat{\boldsymbol{E}}_1^T), cos(\boldsymbol{F}_i^M, \hat{\boldsymbol{E}}_2^T), \cdots, cos(\boldsymbol{F}_i^M, \hat{\boldsymbol{E}}_n^T)\right]), \tag{4.13}$$

其中 T 为一个可学习的温度系数（初始值设置为了 0.07），用于调节预测概率分布的敏锐程度；cos 代表余弦相似度函数；$\boldsymbol{F}_i^M$ 为第 i 个掩膜提议对应的特征嵌入；而 $\hat{\boldsymbol{E}}_j^T, j \in [1, n]$ 则为第 j 种语义类别对应的文本嵌入，这些文本嵌入只需生成一次，随后缓存于内存中用于分类无需额外开销。

4.3.4　损失函数设计

对于掩膜分类子任务，本章节采用了掩膜感知损失（mask-aware loss）[167]（$\mathcal{L}_{ma}$，以增强 CLIP 图像编码器对像素级细粒度场景解析任务的敏感性。该损失函数的核心目标是通过利用真实标注的 IoU 分数 S_{IoU} 作为监督信号，将其与掩膜分类预测 CLIP 分类分数对齐，从而引导模型为高质量掩膜提议分配高分类置信度，同时抑制低质量掩膜提议的得分。具体而言，掩膜感知损失通过 Smooth L1 损失函数实现，可表示为：

$$\mathcal{L}_{ma} = \text{Smooth L1}(S_{cls}, S_{IoU}) \tag{4.14}$$

需要注意的是，考虑到 $\mathcal{L}_{ma}$ 可能导致模型在训练类别上过拟合，降低 CLIP 的泛化性能，本章节在训练过程中引入表征补偿损失 $\mathcal{L}_{rc}$（详见第4.3.1节）来补偿 CLIP 的原始表征特性。此外，对于掩膜预测子任务，本章节遵循了 Mask2Former[56] 的方案，采用其原有损失函数（$\mathcal{L}_{mp}$）进行参数优化（具体细节可参考章节2.2.3，此处不再展开阐述）。因此，总的损失函数可表示为：

$$\mathcal{L} = \mathcal{L}_{mp} + \lambda_1 \mathcal{L}_{ma} + \lambda_2 \mathcal{L}_{rc}, \tag{4.15}$$

在实验中我们设定 $\lambda_1 = 1$，$\lambda_2 = 0.1$。需要注意的是，由于在 CLIDA 网络中引入了梯度解耦机制，因此掩膜预测子任务的损失函数不会影响 CLIP 模型本身（详见章节4.3.1）。这一设计确保了掩膜预测子任务和掩膜分类子任务在训练过程中的独立性，从而进一步提升模型的整体性能。

4.4　方法验证与实验结果分析

在本节中，我们首先对开放词汇场景解析领域[166, 168] 内所使用的标准数据基准以及训练设置进行介绍，并阐述网络的实现细节；之后展示了与现有开放词汇场景解析网络的性能比较，最后通过通过系统性的消融实验证明本章节主要贡献的有效性。需要注意的是，为了更充分地评估所提出的多源信息融合道路场景解析网络 CLIDA 在开放词汇任务中的潜力，我们需要与现有的开放词汇场景解析网络进行对比，为保证实验的公平性，本章节严格遵循了这些相关研究中的训练、测试数据集使用、划分以及评价方法，因此本章节的后续实验都在公开的通用语义分割数据集上进行。

4.4.1　数据集与评价方法

现有开放词汇场景解析研究通常在 COCO-Stuff[169] 数据集上进行训练，并在 ADE20K[170] 与 Cityscapes[22] 数据集上进行零样本推理，以评测开放词汇场景解析网络的性能。其中，对于 Cityscapes 数据集的介绍可见章节2.3.1，在此不再赘述。

COCO-Stuff 数据集　COCO-Stuff 是一个大规模语义分割基准数据集，也是开放词汇场景解析领域广泛采用的训练基准。该数据集共包含 171 个经过语义验证的物体与场景类别，涵盖从基础物体到复杂场景的多样化语义信息。本章严格遵循数据集官方划分标准，采用 118,287 张训练图像作为训练集，并选取 5,000 张验证图像作为测试集以客观评估模型泛化能力。

ADE20K 数据集（150 类别）　ADE20K-150（简称 ADE-150）数据集包含 20,000 幅高分辨率训练图像及 2,000 幅验证图像，其标注体系涵盖 150 个经过语义验证的细粒度类别，包含从建筑构件（如屋顶、窗框）、自然要素（如云层、植被）到生活物件（如家具、电子设备）的多层次场景要素。该数据集是评估模型在室内外复杂场景下细粒度分割能力的重要测试基准。

评测指标　为进行定量评估，我们采用了开放词汇场景解析领域的通用做法[161]，使用四项指标衡量网络的性能：

- 平均交并比（mIoU）：所有类别 IoU 的平均值（其相关计算公式可参考章节2.3.2）
- 频数加权交并比（fwIoU）：按类别像素出现频率加权后的 IoU

$$\text{fwIoU} = \sum_{c=1}^{n} \frac{N_c}{N_{\text{total}}} \cdot \frac{TP_c}{TP_c + FP_c + FN_c} \tag{4.16}$$

- 平均准确率（mAcc）：各类别像素级准确率的均值

$$\text{mAcc} = \frac{1}{n} \sum_{c=1}^{n} \frac{TP_c}{TP_c + FN_c} \tag{4.17}$$

- 像素准确率（Pixel Acc）：全图正确分类像素比例

$$\text{Pixel Acc} = \frac{\sum_{c=1}^{n} TP_c}{\sum_{c=1}^{n} (TP_c + FP_c)} \tag{4.18}$$

其中 n 为类别总数，TP_c、FP_c、FN_c 分别表示：被正确识别为类别 c 的像素数量、被错误识别为类别 c 的像素数量、以及被错误识别为非类别 c 的像素数量，N_c 为标注中类别 c 的像素总数，N_{total} 为图像的像素总数。

图 4.6　Depth Anything V2 基础模型预测得到的先验深度图像

4.4.2　网络实现细节

CLIDA 网络中的 CLIP 模型基于 OpenCLIP 开源框架中的 ConvNeXt-Base CLIP 图像编码器构建，并采用 Depth Anything V2[165] 官方仓库中提供的 ViT-L 版本（性能最佳）及其权重进行深度先验的推理，以获得鲁棒的先验深度图像，其示例效果可见图4.6。掩膜生成网络严格遵循 Mask2Former[56] 的默认配置，设置网络产生的的掩膜提议数量为 $N = 100$。训练阶段采用 AdamW 优化器，权重衰减系数设为 0.05，其中 CLIP 图像编码器的学习率设置为 1×10^{-5}，其余模块学习率统一为 1×10^{-4}。由于计算资源的限制，本章节将输入图像与深度先验统一进行 384×384 像素的随机裁剪预处理，在推理时采用滑窗推理机制以保证推理的质量。在硬件配置方面，本章使用了 4 块 NVIDIA RTX 4090 GPU 并对网络进行了 60,000 次迭代训练，整个训练过程仅在 COCO-Stuff 数据集上进行。

4.4.3　消融实验

我们首先对所提出的梯度解耦策略的有效性进行了消融实验。如前文所述，梯度解耦机制的意义在于将掩膜预测与掩膜分类子任务在梯度层面进行分离，使得一个任务的损失导致的梯度不会影响另一个人物的参数更新。根据本章节的

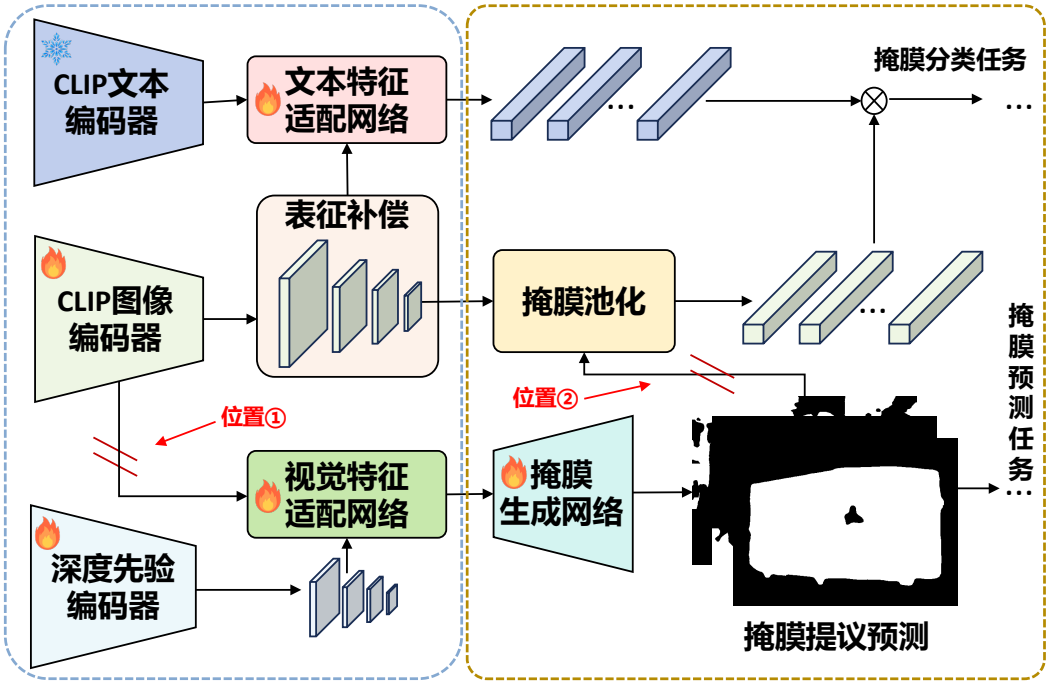


图 4.7　梯度解耦有效性消融实验中解耦节点选择示意图

网络组成，若要实现此目的，则存在两个关键的节点，分别可用于进行梯度解耦，如图4.7中红色箭头标出的位置所示。其中沿着位置 1 反向传播的梯度决定了掩膜预测子任务的掩膜损失是否会影响到 CLIP 图像编码器的参数更新，而沿着位置 2 反向传播的梯度则决定了掩膜分类任务的损失是否会影响到视觉特征适配网络以及深度先验编码器的参数更新。我们对这两个节点分别在 ADE20K 以及 Cityscapes 两种风格的数据集上进行了梯度解耦有效性的对比实验，最终得到了

表 4.2　在 ADE20K 数据集上对不同梯度解耦节点影响进行消融实验（%）得到的定量结果

梯度解耦节点	**mIoU** (%)	fwIoU (%)	mAcc (%)	Pixel Acc (%)
无	31.29	**60.36**	46.64	**72.33**
位置 1（本章）	**33.96**	59.54	**52.76**	71.51
位置 2	30.72	58.82	46.36	71.06
位置 1+2	33.68	59.70	52.04	71.64

表 4.3　在 Cityscapes 数据集上对不同梯度解耦节点影响进行消融实验（%）得到的定量结果

梯度解耦节点	**mIoU** (%)	fwIoU (%)	mAcc (%)	Pixel Acc (%)
无	42.70	78.76	54.60	87.02
位置 1（本章）	44.86	78.73	56.13	87.21
位置 2	43.20	78.55	54.58	86.94
位置 1+2	**45.41**	**79.63**	**56.77**	**87.81**

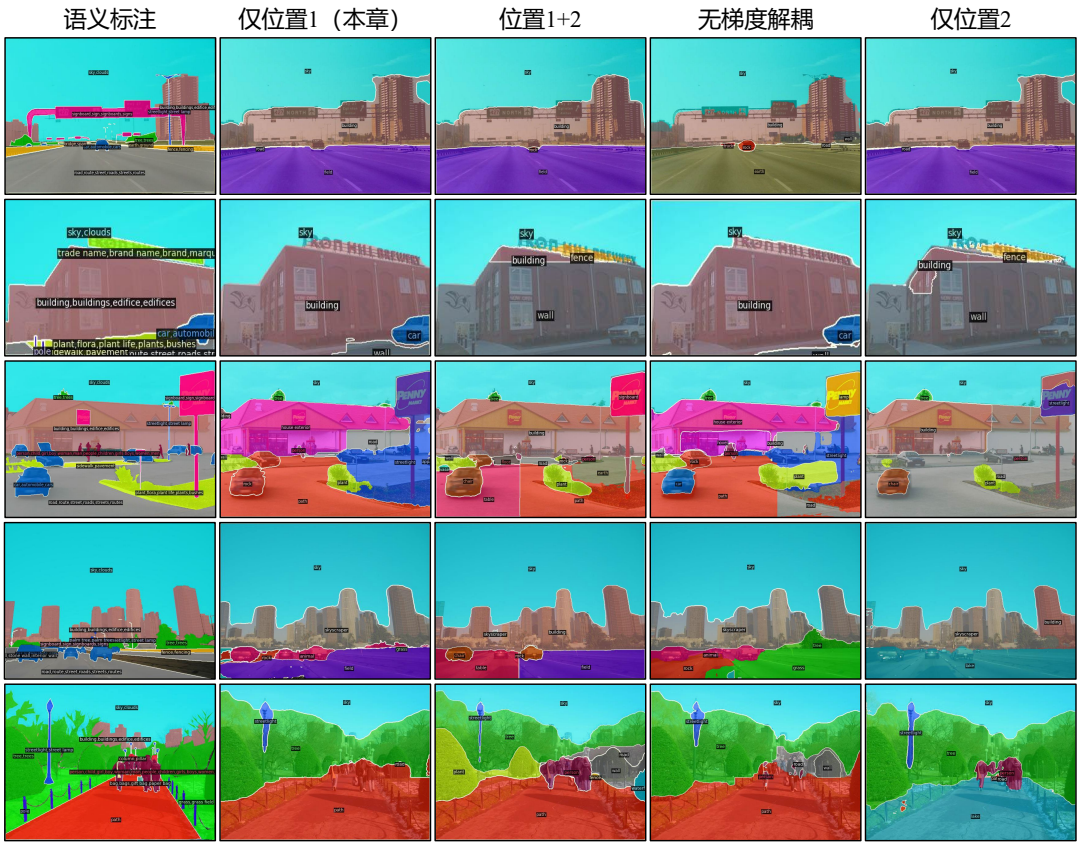


图 4.8　在 ADE20K 数据集上对不同梯度解耦节点影响进行消融实验（%）得到的定性结果

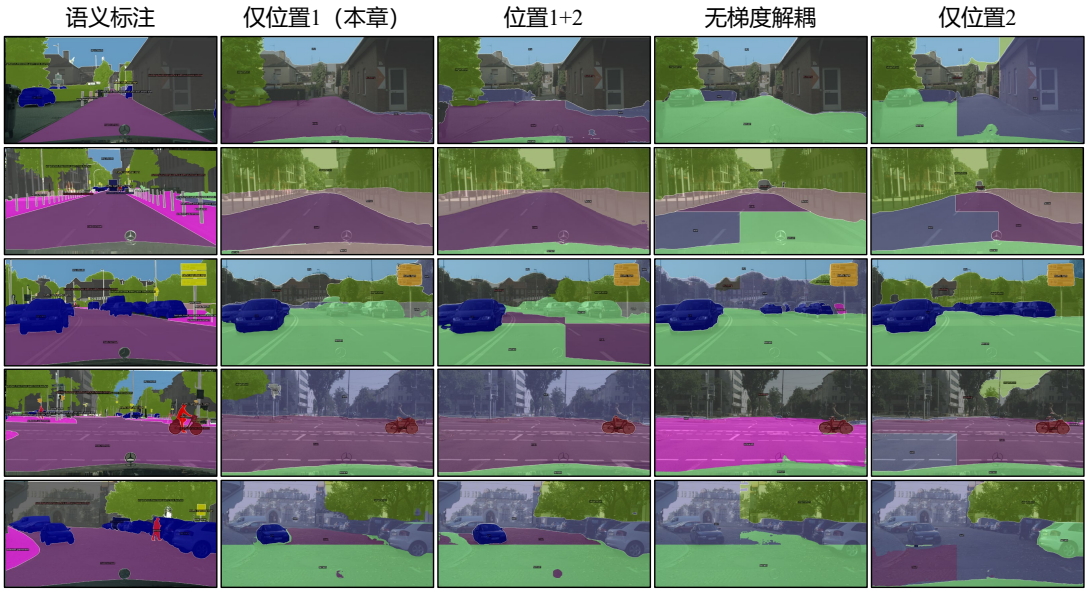


图 4.9　在 Cityscapes 数据集上对不同梯度解耦节点影响进行消融实验（%）得到的定性结果

如表 4.2- 4.3以及图4.8-4.9所示的定量及定性结果。可以明显看出，仅在位置 1 进行梯度解耦取得了最佳的 mIoU 指标（33.96%），远远超出了不进行梯度解耦以及仅在位置 2 进行解耦的变体。这进一步体现了本章节所提出的梯度解耦策略的有效性，也论证了之前提到的掩膜预测任务损失会影响 CLIP 图像编码器中的

参数，破坏其视觉-文本特征对齐的特性；而对于同时在位置 1 和 2 进行梯度解耦的变体来说，其性能在 ADE20K 数据集上略低于（0.28%）本章最终的选择，但在 Cityscapes 数据集上略高，这充分说明对于开放词汇场景解析任务，引入掩膜分类子任务对掩膜预测部分网络的影响，能够在一定程度上提升整体任务的表现，反之则可能会损害网络的性能。

本章节同样对所引入先验深度图像的有效性进行了消融实验，通过将网络输入的先验深度图像替换为对应的 RGB 图像，从而取消了空间信息的引入，实验的定量结果如表 4.4所示。可以看到，当网络失去显式编码的空间几何特征以后，网络的 mIoU 指标下降了 1.21%，说明了引入先验深度图像能够有效提升开放词汇场景解析任务的精确性和鲁棒性。

另一方面，基于梯度解耦实现的掩膜分类与掩膜预测任务解耦，本章节因而可以借助章节3.3中提出的异构特征编码架构为基础，在不影响网络开放词汇分类能力的前提下构建更强大的掩膜预测网络。在本章节提出的视觉特征适配网络中，我们首先对其中 VFM 的有效性进行了消融实验。考虑到开放词汇网络训练的效率，我们挑选了在上一章节封闭词汇场景解析任务中中表现最佳的两种 VFM，DINOv2 与 BEiTv2 进行了实验，结果如表 4.5所示，使用 DINOv2 作为视觉特征适配网络中的 VFM 在本章节的任务中实现了更为强大的性能，这一结果与上一章第3.4.4节中的实验结论略有不同。

此外，如前文所述，开放词汇场景解析任务需要通过引入先验深度图像提升掩膜预测子任务的性能，这一任务通常需要实现更加精确的局部语义特征，而并非全局特征。因此，我们仅选择了上一章节中提出的 GLCA 模块进行先验深度图像中的局部语义特征注入，实现异构特征融合，这一点与上一章节中的双向融合不同；我们通过实验结果论证了该思路的正确性，如表 4.6所示，我们分别引入 GLCA 模块与 CCG 模块，以及同时引入上述两个网络组件进行了实验（上述组件的作用请读者见章节3.3.1），最终证明了仅使用 GLCA 将深度图像中局部语义特征融入异构融合特征 $\mathcal{F}^F$ 能够实现最佳的场景解析性能，也同时能够避免同时引入 CCG 模块而导致过重的网络结构。

表 4.4　在 ADE20K 数据集上对引入双源信息有效性的消融实验（%）

双源信息	**mIoU** (%)	fwIoU (%)	mAcc (%)	Pixel Acc (%)
RGB 图像 +RGB 图像	32.75	59.02	**53.07**	70.83
RGB 图像 + 先验深度图像（本章）	**33.96**	**59.54**	52.76	**71.51**

4.4.4　公开数据基准对比实验

如表 4.7所示是与现有最先进的开放词汇场景解析网络在 ADE20K-150 数据集上的零样本推理性能对比，相比于现有工作，本章节所提出基于信息融合的网络能够在其中取得领先的性能。从表中我们可以发现，开放词汇场景解析任务的性能与所使用视觉-语言模型的规模成正比，通常在网络架构相同或相似的前提

表 4.5　对视觉特征适配网络中不同 VFM 有效性的消融实验（%）

VFM	**mIoU** (%)	fwIoU (%)	mAcc (%)	Pixel Acc (%)
BEiTv2[102]	33.75	60.02	51.98	71.94
DINOv2[103]	**33.96**	59.54	**52.76**	71.51

表 4.6　对视觉特征适配网络中模块 GLCA 与 CCG 有效性的消融实验（%）

模块名称	**mIoU** (%)	fwIoU (%)	mAcc (%)	Pixel Acc (%)
仅 CCG	32.70	58.50	51.52	70.72
仅 GLCA（本章）	**33.96**	59.54	**52.76**	71.51
以上二者	33.59	59.48	51.91	**71.53**

网络名称	视觉-语言模型	预训练数据集	mIoU	会议（刊物）
LSeg+[153]	ALIGN RN101	COCO-Stuff	13.00	ECCV 2022
OpenSeg[153]	ALIGN RN101	COCO-Stuff	15.30	ECCV 2022
LSeg+[153]	ALIGN EN-B7	COCO-Stuff	18.00	ECCV 2022
OpenSeg[153]	ALIGN EN-B7	COCO-Stuff	21.10	ECCV 2022
OpenSeg[153]	ALIGN EN-B7	COCO-Stuff+Loc. Narr.†	28.60	ECCV 2022
SimSeg[154]	CLIP ViT-B/16	COCO-Stuff	21.10	ECCV 2022
OvSeg[152]	CLIP ViT-B/16	COCO-Stuff	24.80	CVPR 2023
SCAN[171]	CLIP ViT-B/16	COCO-Stuff	30.80	CVPR 2024
MaskCLIP[160]	CLIP ViT-L/14	COCO-Stuff	23.70	ICML 2022
SimSeg[154]	CLIP ViT-L/14	COCO-Stuff	21.70	ECCV 2022
OvSeg[153]	CLIP ViT-L/14	COCO-Stuff	29.60	CVPR 2023
SCAN[171]	CLIP ViT-L/14	COCO-Stuff	33.50	CVPR 2024
R-SAN[172]	CLIP ViT-L/14	COCO-Stuff	32.10	T-PAMI 2023
FC-CLIP[162]	CLIP ConvNeXt-B	COCO-Stuff	31.1	NeurIPS 2023
SED[156]	CLIP ConvNeXt-B	COCO-Stuff	31.60	CVPR 2024
MAFT+[166]	CLIP ConvNeXt-B	COCO-Stuff	33.60	ECCV 2024
CLIDA（本章）	CLIP ConvNeXt-B	COCO-Stuff	**33.96**	-

表 4.7　与现有最先进的开放词汇场景解析方法在 ADE20K 验证集上进行的零样本推理定量比较（%），† 代表该网络在预训练时使用了研究[173] 中开源的视频数据集（Localized Narratives）进行了额外的训练

下，使用更大规模的视觉语言模型能够有效提升网络的性能。此外，由于场景解析任务本身是一种像素级细粒度的分类任务，基于 CNN 架构的视觉语言模型本身就能够提取出多尺度且与文本特征对齐的特征，因此通常能够取得更好的性能，近年来受到更多研究的使用。

4.5　本章小结

本章节探究了现有基于信息融合的网络在开放环境以及实际应用中，仅能实现对于预定义类别物体的预测，从而导致其泛化性以及适用性受限的问题。相比于具有二维空间结构的 RGB 图像、深度图像、热图像等视觉信息，语言文本信息在结构上具有非空间特性，因此难以与视觉特征直接进行融合以提升任务性能。对此，本章利用视觉-语言模型提供的视觉-文本特征对齐能力，实现对于语言文本信息的间接利用。同时本章节引入了视觉基础模型提供的先验深度图像，最终构建了一个以 RGB 图像与先验深度图像作为输入的双源信息融合开放词汇场景解析网络。在此过程中，我们首先发现了现有异构特征融合策略会破坏视觉语言模型本身的视觉-文本特征对齐的特性，从而极大的损害网络开放词汇能力，并引入了梯度解耦的思想，实现任务解耦，设计了视觉特征适配网络，在保证网路开放词汇能力的前提下，实现更加精确、鲁棒的掩膜预测，构建了性能领先的开放词汇场景解析网络，在弥补现有信息融合网络适用性不足的同时，为开放词汇场景解析任务在具有复杂背景等挑战性场景中的性能提升提供了新的范式。

第 5 章　结论与展望

5.1　研究总结

本文围绕基于信息融合的道路场景解析展开系统性研究，针对复杂环境下自动驾驶感知系统的性能瓶颈问题，提出了一系列创新性解决方案。道路场景解析作为感知系统的重要组成，在无人驾驶、智能交通等领域中具有具有基础而重要的作用。然而主流方法或仅依赖于单模态信息（如 RGB 图像），这些方法在光照变化、天气恶劣以及场景复杂情况下表现出精度和鲁棒性不足；或依赖于较为原始的多源信息融合框架，使得任务精度无法获得有效提升。为此，本文通过深度整合 RGB 图像、激光雷达点云及文本语义等多源信息，在异构特征融合机制、模型架构设计和任务范式创新三个维度实现了重要突破，显著提升了道路场景解析的精确性、鲁棒性与泛化能力。主要研究成果可归纳如下：

(1) 本文提出了以 RGB-法向量图像对作为输入的高性能信息融合网络与更加高效的异构特征融合策略，大大提升了可行驶区域检测与道路破损检测等任务的性能。具体来说，我们基于 Transformer 中注意力机制设计了动态自适应的异构特征融合框架，区别于传统特征串联或元素相加的方法，能够根据场景动态突出关键特征、抑制无关特征，在道路破损检测等任务中展现出显著优势。此外，该方法是道路场景解析领域中对掩膜分类范式任务解码的首次尝试，证明了基于 Transformer 的多层迭代解码范式能够更有效的利用异构融合特征，弥补了领域内道路破损检测任务训练数据匮乏的问题。

(2) 建立了基于基础模型的通用异构特征提取与融合体系。针对现有信息融合道路场景解析网络对于不同应用场景或任务的鲁棒性与泛化性不足，以及异构特征表征能力低下的问题，本文提出了一种基于视觉基础模型的双源信息融合网络。通过对基础模型在特征编码阶段进行任务适配，实现了更具通用性的鲁棒特征提取，并针对 RGB 与异构视觉模态的内在特性，构建了 Transformer 与 CNN 混合的异构特征提取框架：采用 Transformer 网络捕获 RGB 图像的全局语义关联，配合 CNN 网络提取深度/热红外等异构数据的局部细节特征。通过设计多尺度特征对齐模块与跨模态交互单元，实现了全局上下文与局部特征的协同优化，在主流多模态场景解析任务中展现出卓越的泛化性能，显著提升了模型在光照突变、恶劣天气等复杂场景下的鲁棒性。

(3) 开创开放词汇场景解析新范式。针对现有信息融合网络对预定义类别的依赖，无法解决实际开放场景中的对于多变语义类别对象的预测问题，本文基于视觉-语言大模型提出了一种具有开放词汇能力的信息融合网络，有效增强了信息融合网络在面对不同语义类别对象时的泛化性与适用性。具体而言，设计双分支解耦架构：掩膜分类分支利用预训练的视觉语言模型支撑开放词汇的语义推理；掩膜预测分支则融合色彩特征与几何特征，通过几何线索增强物体边界表征。这种设计充分考虑了不同信息的特性差异：深度图像虽未与文本空间对齐，但其蕴含的几何特征能有效提升掩膜预测的轮廓精度。通过梯度解耦机制，网络能够充分发挥双源信息潜力，使 RGB 图像专注于高级语义信息理解，而深度图像增强网络对于局部细节的理解。所提出的网络在主流开放词汇基准中展现出了卓越的性能，为信息融合感知系统赋予了理解未知对象类别的能力。

最后，在应用层面，本研究构建的创新方法展现出广泛的应用潜力。信息融合算法可显著提升自动驾驶系统在低光照、极端天气等复杂场景下的感知可靠性，为决策控制模块提供更精准的环境认知。在自动驾驶与智能交通等领域，本研究提出的方法可支持复杂路况下的精确环境感知、道路智能巡检等关键任务，推动基础设施维护向智能化转型。总而言之，道路场景解析作为智能交通和自动驾驶的关键环节，仍将是一个不断演进的研究领域。未来的工作将围绕更高效的异构特征融合、更强大的模型泛化能力以及更广泛的跨领域集成展开。相信通过学术界和产业界的共同努力，基于信息融合的智能感知技术将持续完善，为构建安全、高效的自主导航系统奠定更加坚实的基础。

5.2　研究展望

随着近年来深度学习领域基础模型、大模型等概念的兴起，包括环境感知等任务的性能都已达到了一个新的高度。尽管本文部分章节加入了大模型的相关设计，但受限于算力等因素的影响，本文中的工作其本质是对大模型的一种微调策略，而未涉及原生的大模型预训练。这些策略归根到底是无法与具有大数据驱动力与可扩展性的原生大模型相媲美。因此，我认为对于以道路场景解析任务为代表的环境感知技术，在如今的时代有以下内容值得探索：

(1) 包含深度图像、法向量图像以及热图像等双源信息的多模态预训练理论体系构建。如今自动驾驶等领域中拥有超过百万数量级的开源多模态数据集，设计合理的预训练策略，将 RGB 图像与前述多源信息在预训练阶段进行融合，在大数据驱动下学习更全面的场景上下文信息，这很可能比针对特定任务设计的微调范式更加有效，从大规模数据中学习，大模型也将具有

更强的鲁棒性。

(2) 当前提出的异构特征融合模块虽然提高了精度，但其计算复杂度和实时性能存在缺陷。为了有效编码异构特征，网络常常需要多个骨干网络进行异构特征编码，这通常带来了成倍的计算复杂度与模型参数。可以探索更高效的特征融合算法和模型压缩技术，例如设计轻量级注意力机制或蒸馏已有模型，以降低计算开销并满足实际应用对实时性的要求，并保证相当的网络性能。

(3) 自动驾驶领域的端到端模型已成为一种流行的范式，受到了各大科技公司以及研究团队的关注。在这种网络架构中，感知、决策以及控制等阶段被模块化，同时在一个网络种进行端到端的训练，有效避免了当前传统自动驾驶系统中不同阶段间的“代沟”问题，能否在这种端到端范式的网络中的感知模块加入多源传感器信息，提升端到端系统对于环境先验感知的精度以及鲁棒性，从而进一步提升整体系统的稳定性上限。

参考文献

[1] BADRINARAYANAN V, et al. SEGNET: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[2] HAZIRBAS C, et al. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture[C]//13th Asian Conference on Computer Vision (ACCV). 2017: 213-228.

[3] MACAULAY M O, SHAFIEE M. Machine learning techniques for robotic and autonomous inspection of mechanical systems and civil infrastructure[J]. Autonomous Intelligent Systems, 2022, 2(1): 8.

[4] GOODALE M A. Lessons from human vision for robotic design[J]. Autonomous Intelligent Systems, 2021, 1(1): 2.

[5] FENG Y, et al. SNE-RoadSegV2: Advancing Heterogeneous Feature Fusion and Fallibility Awareness for Freespace Detection[J]. IEEE Transactions on Instrumentation and Measurement, 2024.

[6] WANG H, et al. Dynamic Fusion Module Evolves Drivable Area and Road Anomaly Detection: A Benchmark and Algorithms[J]. IEEE Transactions on Cybernetics, 2021, 52(10): 10750-10760.

[7] RUBAGOTTI M, et al. Perceived safety in physical human–robot interaction—A survey[J]. Robotics and Autonomous Systems, 2022, 151: 104047.

[8] KC S. Enhanced pothole detection system using YOLOX algorithm[J]. Autonomous Intelligent Systems, 2022, 2(1): 22.

[9] FAN R, et al. A Glance Over the Past Decade: Road Scene Parsing towards Safe and Comfortable Autonomous Driving[J]. Autonomous Intelligent Systems, 2025, 5(1): 1-15.

[10] MA N, et al. Computer vision for road imaging and pothole detection: a state-of-the-art review of systems and algorithms[J]. Transportation safety and Environment, 2022, 4(4): tdac026.

[11] FAN R, et al. Pothole Detection Based on Disparity Transformation and Road Surface Modeling[J]. IEEE Transactions on Image Processing, 2020, 29: 897-908.

[12] FAN R, et al. Rethinking Road Surface 3-D Reconstruction and Pothole Detection: From Perspective Transformation to Disparity Map Segmentation[J]. IEEE Transactions on Cybernetics, 2021, 52(7): 5799-5808.

[13] WEDEL A, et al. B-spline modeling of road surfaces for freespace estimation[C]//2008 IEEE Intelligent Vehicles Symposium (IV). 2008: 828-833.

[14] WEDEL A, et al. B-Spline Modeling of Road Surfaces With an Application to Free-Space Estimation[J]. IEEE Transactions on Intelligent Transportation Systems, 2009, 10(4): 572-583.

[15] LU C, et al. Monocular Semantic Occupancy Grid Mapping With Convolutional Variational Encoder–Decoder Networks[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 445-452.

[16] FAN R, et al. SNE-RoadSeg: Incorporating Surface Normal Information into Semantic Segmentation for Accurate Freespace Detection[C]//European Conference on Computer Vision (ECCV). 2020: 340-356.

[17] MIN C, et al. ORFD: A Dataset and Benchmark for Off-Road Freespace Detection[C]//2022 International Conference on Robotics and Automation (ICRA). 2022: 2532-2538.

[18] WANG H, et al. SNE-RoadSeg+: Rethinking Depth-Normal Translation and Deep Supervision for Freespace Detection[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2021: 1140-1145.

[19] RONNEBERGER O, et al. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//Medical Image Computing and Computer-Assisted Intervention (MICCAI). 2015: 234-241.

[20] FRITSCH J, et al. A New Performance Measure and Evaluation Benchmark for Road Detection Algorithms[C]//16th International IEEE Conference on Intelligent Transportation Systems (ITSC) 2013. 2013: 1693-1700.

[21] CABON Y, et al. Virtual KITTI 2[J]. Computing Research Repository (CoRR), 2020, abs2001.10773.

[22] CORDTS M, et al. The Cityscapes Dataset for Semantic Urban Scene Understanding[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 3213-3223.

[23] VASWANI A, et al. Attention is All you Need[J]. Advances in Neural Information Processing Systems (NeurIPS), 2017, 30.

[24] XIE E, et al. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers[J]. Advances in Neural Information Processing Systems (NeurIPS), 2021, 34: 12077-12090.

[25] LI K, et al. UniFormer: Unifying Convolution and Self-Attention for Visual Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[26] LI J, et al. RoadFormer: Duplex Transformer for RGB-normal Semantic Road Scene Parsing[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(7): 5163-5172.

[27] LONG J, et al. Fully Convolutional Networks for Semantic Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 3431-3440.

[28] WU H, et al. FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation[J]. Computing Research Repository (CoRR), 2019, abs/1903.11816.

[29] POUDEL R P, et al. Fast-SCNN: Fast Semantic Segmentation Network[C]//The British Machine Vision Conference (BMVC). 2019.

[30] XU J, et al. PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 19529-19539.

[31] CHEN L C, et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 801-818.

[32] HE J, et al. Dynamic Multi-Scale Filters for Semantic Segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 3562-3572.

[33] ZHAO H, et al. Pyramid Scene Parsing Network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 2881-2890.

[34] FLORIAN L C, et al. Rethinking Atrous Convolution for Semantic Image Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

[35] SUN K, et al. Deep High-Resolution Representation Learning for Human Pose Estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 5693-5703.

[36] KIRILLOV A, et al. PointRend: Image Segmentation as Rendering[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 9799-9808.

[37] FU J, et al. Dual Attention Network for Scene Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 3146-3154.

[38] WANG X, et al. Non-Local Neural Networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 7794-7803.

[39] ZHU Z, et al. Asymmetric Non-Local Neural Networks for Semantic Segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2019: 593-602.

[40] YIN M, et al. Disentangled Non-local Neural Networks[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2020: 191-207.

[41] CAO Y, et al. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). 2019: 0–0.

[42] HUANG L, et al. Interlaced Sparse Self-Attention for Semantic Segmentation[J]. Computing Research Repository (CoRR), 2019, abs1907.12273.

[43] WANG X, et al. Non-Local Neural Networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 7794-7803.

[44] ZHAO H, et al. PSANet: Point-Wise Spatial Attention Network for Scene Sarsing[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 267-283.

[45] LI X, et al. Expectation-Maximization Attention Networks for Semantic Segmentation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 9167-9176.

[46] WU T, et al. CGNet: A Light-weight Context Guided Network for Semantic Segmentation[J]. IEEE Transactions on Image Processing, 2020, 30: 1169-1179.

[47] ZHENG S, et al. Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 6881-6890.

[48] LIU Z, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 10012-10022.

[49] CARION N, et al. End-to-End Object Detection with Transformers[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2020: 213-229.

[50] STRUDEL R, et al. Segmenter: Transformer for Semantic Segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 7262-7272.

[51] CHU X, et al. Twins: Revisiting the Design of Spatial Attention in Vision Transformers[J]. Advances in Neural Information Processing Systems (NeurIPS), 2021, 34: 9355-9366.

[52] RANFTL R, et al. Vision Transformers for Dense Prediction[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2021: 12179-12188.

[53] YUAN Y, et al. Object-Contextual Representations for Semantic Segmentation[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2020: 173-190.

[54] ZHANG W, et al. K-Net: Towards Unified Image Segmentation[J]. Advances in Neural Information Processing Systems (NeurIPS), 2021, 34: 10326-10338.

[55] CHENG B, et al. Per-Pixel Classification is Not All You Need for Semantic Segmentation[J]. Advances in Neural Information Processing Systems (NeurIPS), 2021, 34: 17864-17875.

[56] CHENG B, et al. Masked-attention Mask Transformer for Universal Image Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 1290-1299.

[57] HA Q, et al. MFNet: Towards Real-Time Semantic Segmentation for Autonomous Vehicles with Multi-Spectral Scenes[C]//2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2017: 5108-5115.

[58] SUN Y, et al. RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes[J]. IEEE Robotics and Automation Letters, 2019, 4(3): 2576-2583.

[59] LEVI D, et al. StixelNet: A Deep Convolutional Network for Obstacle Detection and Road Segmentation[C]//Proceedings of the British Machine Vision Conference (BMVC):vol. 12. 2015: 4.

[60] FAN R, et al. Learning Collision-Free Space Detection From Stereo Images: Homography Matrix Brings Better Data Augmentation[J]. IEEE/ASME Transactions on Mechatronics, 2021, 27(1): 225-233.

[61] ZHOU H, et al. Exploiting Low-level Representations for Ultra-Fast Road Segmentation[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(8): 9909-9919.

[62] CHEN L C, et al. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[63] HE K, et al. Deep Residual Learning for Image Recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770-778.

[64] CALTAGIRONE L, et al. Fast LIDAR-based Road Detection Using Fully Convolutional Neural Networks[C]//IEEE Intelligent Vehicles Symposium (IV). 2017: 1019-1024.

[65] LYU Y, et al. ChipNet: Real-Time LiDAR Processing for Drivable Region Segmentation on an FPGA[J]. IEEE Transactions on Circuits and Systems, 2018, 66(5): 1769-1779.

[66] CHANG Y, et al. Fast Road Segmentation via Uncertainty-aware Symmetric Network[C]// IEEE International Conference on Robotics and Automation (ICRA). 2022: 11124-11130.

[67] SUN J Y, et al. Reverse and Boundary Attention Network for Road Segmentation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). 2019: 0–0.

[68] CHEN Z, et al. Progressive LiDAR adaptation for road detection[J]. IEEE/CAA Journal of Automatica Sinica, 2019, 6(3): 693-702.

[69] KHAN A A, et al. LRDNet: Lightweight LiDAR Aided Cascaded Feature Pools for Free Road Space Detection[J]. IEEE Transactions on Multimedia, 2022: 1-13.

[70] GU S, et al. Two-View Fusion based Convolutional Neural Network for Urban Road Detection[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2019: 6144-6149.

[71] GU S, et al. Road Detection through CRF based LiDAR-Camera Fusion[C]//2019 International Conference on Robotics and Automation (ICRA). 2019: 3832-3838.

[72] GU S, et al. A Cascaded LiDAR-Camera Fusion Network for Road Detection[C]//IEEE International Conference on Robotics and Automation (ICRA). 2021: 13308-13314.

[73] CALTAGIRONE L, et al. LiDAR–camera fusion for road detection using fully convolutional neural networks[J]. Robotics and Autonomous Systems, 2019, 111: 125-131.

[74] SUN L, et al. Pseudo-LiDAR-Based Road Detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(8): 5386-5398.

[75] WANG H, et al. Applying Surface Normal Information in Drivable Area and Road Anomaly Detection for Ground Mobile Robots[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2020: 2706-2711.

[76] DOSOVITSKIY A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. International Conference on Learning Representations (ICLR), 2020: 1-22.

[77] HE K, et al. Mask R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017: 2961-2969.

[78] CARION N, et al. End-to-End Object Detection with Transformers[C]//European Conference on Computer Vision (ECCV). 2020: 213-229.

[79] WANG H, et al. Max-Deeplab: End-to-End Panoptic Segmentation with Mask Transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 5463-5474.

[80] KUHN H W. The Hungarian Method for the Assignment Problem[J]. Naval Research Logistics Quarterly, 1955, 2(1-2): 83-97.

[81] HAN K, et al. A Survey on Vision Transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 87-110.

[82] DOSOVITSKIY A, et al. CARLA: An Open Urban Driving Simulator[C]//Conference on Robot Learning (CoRL). 2017: 1-16.

[83] FENG Y, et al. D2NT: A High-Performing Depth-to-Normal Translator[C]//2023 IEEE International Conference on Robotics and Automation (ICRA). 2023: 12360-12366.

[84] FENG Y, et al. SNE-RoadSegV2: Advancing Heterogeneous Feature Fusion and Fallibility Awareness for Freespace Detection[J]. ArXiv preprint arXiv:2402.18918, 2024.

[85] LIU Z, et al. A ConvNet for the 2020s[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 11976-11986.

[86] RADFORD A, et al. Learning Transferable Visual Models From Natural Language Supervision[C]//International Conference on Machine Learning (ICML). 2021: 8748-8763.

[87] HU J, et al. Squeeze-and-Excitation Networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 7132-7141.

[88] MAI S, et al. Analyzing Multimodal Sentiment Via Acoustic- and Visual-LSTM With Channel-Aware Temporal Convolution Network[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1424-1437.

[89] CHOLLET F. Xception: Deep Learning With Depthwise Separable Convolutions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 1251-1258.

[90] ZHU X, et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection[C]//International Conference on Learning Representations (ICLR). 2020: 1-16.

[91] MILLETARI F, et al. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation[C]//2016 fourth International Conference on 3D Vision (3DV). 2016: 565-571.

[92] FAN R, et al. Road Surface 3D Reconstruction Based on Dense Subpixel Disparity Map Estimation[J]. IEEE Transactions on Image Processing, 2018, 27(6): 3025-3035.

[93] PERLIN K. An Image Synthesizer[J]. ACM Siggraph Computer Graphics, 1985, 19(3): 287-296.

[94] MENZE M, GEIGER A. Object Scene Flow for Autonomous Vehicles[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 3061-3070.

[95] LIPSON L, et al. RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching[C]//2021 International Conference on 3D Vision (3DV). 2021: 218-227.

[96] LOSHCHILOV I, HUTTER F. Decoupled Weight Decay Regularization[C]//International Conference on Learning Representations (ICLR). 2018: 1-18.

[97] CHEN L C, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4): 834-848.

[98] JAIN J, et al. OneFormer: One Transformer to Rule Universal Image Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 2989-2998.

[99] GU S, et al. Histograms of the Normalized Inverse Depth and Line Scanning for Urban Road Detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 20(8): 3070-3080.

[100] KRIZHEVSKY A, et al. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in Neural Information Processing Systems (NeurIPS), 2012, 25.

[101] BAO H, et al. BEiT: BERT Pre-Training of Image Transformers[J]. ArXiv preprint arXiv:2106.08254, 2021.

[102] PENG Z, et al. BEiT v2: Masked Image Modeling With Vector-Quantized Visual Tokenizers[J]. ArXiv preprint arXiv:2208.06366, 2022.

[103] OQUAB M, et al. DINOv2: Learning Robust Visual Features Without Supervision[J]. Transactions on Machine Learning Research, 2023.

[104] LIU J, et al. Revisiting Modality-Specific Feature Compensation for Visible-Infrared Person Re-Identification[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 7226-7240.

[105] SEICHTER D, et al. Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). 2021: 13525-13531.

[106] ZHOU W, et al. GMNet: Graded-Feature Multilabel-Learning Network for RGB-Thermal Urban Scene Semantic Segmentation[J]. IEEE Transactions on Image Processing, 2021, 30: 7790-7802.

[107] HUANG J, et al. RoadFormer+: Delivering RGB-X Scene Parsing through Scale-Aware Information Decoupling and Advanced Heterogeneous Feature Fusion[J]. IEEE Transactions on Intelligent Vehicles, 2024.

[108] DONG X, et al. CSWin Transformer: A General Vision Transformer Backbone With Cross-Shaped Windows[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 12124-12134.

[109] WANG W, et al. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2021: 568-578.

[110] ZHU X, et al. Uni-Perceiver: Pre-Training Unified Architecture for Generic Perception for Zero-Shot and Few-Shot Tasks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 16804-16815.

[111] ZHU J, et al. Uni-Perceiver-MoE: Learning Sparse Generalist Models with Conditional MoEs[J]. Advances in Neural Information Processing Systems (NeurIPS), 2022, 35: 2664-2678.

[112] WANG W, et al. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 19175-19186.

[113] LIU Z, et al. Video Swin Transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 3202-3211.

[114] BOMMASANI R, et al. On the Opportunities and Risks of Foundation Models[J]. ArXiv preprint arXiv:2108.07258, 2021.

[115] KIRILLOV A, et al. Segment Anything[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023: 4015-4026.

[116] RAVI N, et al. SAM 2: Segment Anything in Images and Videos[J]. ArXiv preprint arXiv:2408.00714, 2024.

[117] YANG L, et al. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 10371-10381.

[118] CARON M, et al. Emerging Properties in Self-Supervised Vision Transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 9650-9660.

[119] RAMESH A, et al. Zero-Shot Text-to-Image Generation[C]//International Conference on Machine Learning. 2021: 8821-8831.

[120] KINGMA D P. Auto-Encoding Variational Bayes[J]. ArXiv preprint arXiv:1312.6114, 2013.

[121] YIN B, et al. DFormer: Rethinking RGBD Representation Learning for Semantic Segmentation[J]. International Conference on Learning Representations (ICLR), 2024.

[122] LI J, et al. RoadFormer: Duplex Transformer for RGB-Normal Semantic Road Scene Parsing[J]. ArXiv preprint arXiv:2309.10356, 2024.

[123] XIE S, et al. Aggregated Residual Transformations for Deep Neural Networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 1492-1500.

[124] HUANG S, et al. FaPN: Feature-aligned Pyramid Network for Dense Image Prediction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 864-873.

[125] CHEN Z, et al. Vision Transformer Adapter for Dense Predictions[J]. International Conference on Learning Representations (ICLR), 2023.

[126] YANG X, et al. PolyMaX: General Dense Prediction with Mask Transformer[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2024: 1050-1061.

[127] SHIVAKUMAR S S, et al. PST900: RGB-Thermal Calibration, Dataset and Segmentation Network[C]//2020 IEEE International Conference on Robotics and Automation (ICRA). 2020: 9441-9447.

[128] KIM Y H, et al. MS-UDA: Multi-Spectral Unsupervised Domain Adaptation for Thermal Image Semantic Segmentation[J]. IEEE Robotics and Automation Letters, 2021, 6(4): 6497-6504.

[129] SHIN U, et al. Complementary Random Masking for RGB-Thermal Semantic Segmentation[J]. ArXiv preprint arXiv:2303.17386, 2023.

[130] SILBERMAN N, et al. Indoor Segmentation and Support Inference from RGBD Images[C]//European Conference on Computer Vision (ECCV). 2012: 746-760.

[131] SUN Y, et al. FuseSeg: Semantic Segmentation of Urban Scenes Based on RGB and Thermal Data Fusion[J]. IEEE Transactions on Automation Science and Engineering, 2020, 18(3): 1000-1011.

[132] ZHOU W, et al. Edge-Aware Guidance Fusion Network for RGB–Thermal Scene Parsing[C]//Proceedings of the AAAI Conference on Artificial Intelligence:vol. 363. 2022: 3571-3579.

[133] ZHANG Q, et al. ABMDRNet: Adaptive-weighted Bi-directional Modality Difference Reduction Network for RGB-T Semantic Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 2633-2642.

[134] ZHOU W, et al. Embedded Control Gate Fusion and Attention Residual Learning for RGB–Thermal Urban Scene Parsing[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(5): 4794-4803.

[135] DENG F, et al. FEANet: Feature-Enhanced Attention Network for RGB-Thermal Real-time Semantic Segmentation[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2021: 4467-4473.

[136] HE X, et al. SFAF-MA: Spatial Feature Aggregation and Fusion With Modality Adaptation for RGB-Thermal Semantic Segmentation[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-10.

[137] ZHAO S, et al. Mitigating Modality Discrepancies for RGB-T Semantic Segmentation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023.

[138] GUO X, et al. Low-Light Enhancement and Global-Local Feature Interaction for RGB-T Semantic Segmentation[J]. IEEE Transactions on Instrumentation and Measurement, 2025.

[139] LV Y, et al. Context-Aware Interaction Network for RGB-T Semantic Segmentation[J]. IEEE Transactions on Multimedia, 2024.

[140] LIANG M, et al. Explicit Attention-Enhanced Fusion for RGB-Thermal Perception Tasks[J]. IEEE Robotics and Automation Letters, 2023, 8(7): 4060-4067.

[141] ZHANG J, et al. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(12): 14679-14694.

[142] ZHANG J, et al. Delivering Arbitrary-Modal Semantic Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 1136-1147.

[143] TOUVRON H, et al. Training data-efficient image transformers & distillation through attention[C]//International Conference on Machine Learning (ICML). 2021: 10347-10357.

[144] STEINER A P, et al. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers[J]. Transactions on Machine Learning Research, 2022.

[145] SRIVASTAVA S, SHARMA G. OmniVec: Learning Robust Representations With Cross Modal Sharing[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2024: 1236-1248.

[146] GIRDHAR R, et al. OMNIVORE: A Single Model for Many Visual Modalities[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 16102-16112.

[147] WANG Y, et al. Multimodal Token Fusion for Vision Transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 12186-12195.

[148] DU S, et al. AsymFormer: Asymmetrical Cross-Modal Representation Learning for Mobile Platform Real-Time RGB-D Semantic Segmentation[J]. ArXiv preprint arXiv:2309.14065, 2023.

[149] CHEN X, et al. Bi-directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation[C]//European Conference on Computer Vision (ECCV). 2020: 561-577.

[150] CHEN T, et al. A Simple Framework for Contrastive Learning of Visual Representations[C]//International Conference on Machine Learning (ICML). 2020: 1597-1607.

[151] JIA C, et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision[C]//International Conference on Machine Learning (ICML). 2021: 4904-4916.

[152] LIANG F, et al. Open-Vocabulary Semantic Segmentation With Mask-Adapted CLIP[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 7061-7070.

[153] GHIASI G, et al. Scaling Open-Vocabulary Image Segmentation with Image-Level Labels[C]//European Conference on Computer Vision (ECCV). 2022: 540-557.

[154] XU M, et al. A Simple Baseline for Open-Vocabulary Semantic Segmentation with Pre-trained Vision-Language Model[C]//European Conference on Computer Vision (ECCV). 2022: 736-753.

[155] CHO S, et al. CAT-Seg: Cost Aggregation for Open-Vocabulary Semantic Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 4113-4123.

[156] XIE B, et al. SED: A Simple Encoder-Decoder for Open-Vocabulary Semantic Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 3426-3436.

[157] ZHAO G, et al. Open-Vocabulary RGB-Thermal Semantic Segmentation[C]//European Conference on Computer Vision (ECCV). 2024: 304-320.

[158] YU M, et al. Open-RGBT: Open-vocabulary RGB-T Zero-shot Semantic Segmentation in Open-world Environments[J]. ArXiv preprint arXiv:2410.06626, 2024.

[159] SUN C, et al. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era[C]// Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017: 843-852.

[160] DING Z, et al. Open-Vocabulary Universal Image Segmentation with MaskCLIP[J]. International Conference on Machine Learning (ICML), 2022.

[161] XU J, et al. Open-Vocabulary Panoptic Segmentation With Text-to-Image Diffusion Models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 2955-2966.

[162] YU Q, et al. Convolutions Die Hard: Open-Vocabulary Segmentation with Single Frozen Convolutional CLIP[J]. Advances in Neural Information Processing Systems (NeurIPS), 2023, 36: 32215-32234.

[163] XU M, et al. Side Adapter Network for Open-Vocabulary Semantic Segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 2945-2954.

[164] ZHOU C, et al. Extract Free Dense Labels from CLIP[C]//European Conference on Computer Vision (ECCV). 2022: 696-712.

[165] YANG L, et al. Depth Anything V2[J]. Advances in Neural Information Processing Systems (NeurIPS), 2024, 37: 21875-21911.

[166] JIAO S, et al. Collaborative Vision-Text Representation Optimizing for Open-Vocabulary Segmentation[C]//European Conference on Computer Vision (ECCV). 2024: 399-416.

[167] JIAO S, et al. Learning Mask-aware CLIP Representations for Zero-Shot Segmentation[J]. Advances in Neural Information Processing Systems (NeurIPS), 2023, 36: 35631-35653.

[168] DING J, et al. Decoupling Zero-Shot Semantic Segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022: 11583-11592.

[169] CAESAR H, et al. COCO-Stuff: Thing and Stuff Classes in Context[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 1209-1218.

[170] ZHOU B, et al. Scene Parsing through ADE20K Dataset[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 633-641.

[171] LIU Y, et al. Open-Vocabulary Segmentation with Semantic-Assisted Calibration[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 3491-3500.

[172] XU M, et al. SAN: Side Adapter Network for Open-Vocabulary Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15546-15561.

[173] VOIGTLAENDER P, et al. Connecting Vision and Language With Video Localized Narratives[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 2461-2471.

致谢

转眼间，我在同济的三年就要画上句号了。以前看着学长学姐们毕业离校，听他们说着不舍的话，总觉得毕业离我还很遥远。直到现在轮到自己，才真正体会到这种心情——原来告别是这么复杂的一件事。在这里，有太多人值得我用心感谢。

首先要衷心感谢我的导师范睿教授。有人说，选择导师就像二次投胎，那我一定是非常幸运，才遇到如此年轻且优秀的启蒙恩师。三年前，当我以考研专业最后一名的成绩进入同济，几乎无人问津时，是范老师给了我继续深造的机会。这份知遇之恩，改变了我的人生轨迹。如果没有范老师的接纳，现在的我可能正在某个不感兴趣的岗位上庸庸碌碌，而不会发现自己对科研的热爱。

我永远记得写第一篇论文时，范老师不厌其烦地带着我推导公式、修改语法，更是亲自执笔润色。在办公室的无数个日夜，恩师以其渊博的学识和严谨的治学态度，为我指明了科研之路。从项目申请到答辩实践，范老师始终以“精益求精”的学术标准要求我们，这种精神将成为我毕生的追求。范老师渊博的学识令人敬佩，但更让我感动的是他对学生的宽容和理解。无论我们犯了什么错，他总是先给予鼓励，再指出改进的方向。“高山仰止，景行行止”，虽然我可能永远达不到范老师的高度，但我会一直以他为榜样。

特别感谢任倩老师在我求学路上的鼎力相助。无论是录取过程中的波折，还是就读期间的困惑，任老师总是及时给予我支持与鼓励。

感谢这三年来遇到的师兄、同门、和朋友。

感谢所有给予我帮助的师兄师姐们：周光亮师兄：虽然相处只有短短一年，但你对待科研的严谨态度、刨根问底的精神，始终鞭策着我不断自省；马纳川师兄：从你身上我学会了如何更好地与人相处，如何从容应对生活中的困难；苏帅师兄：在我研一基础最薄弱的阶段，是你不厌其烦地帮助我进步;钟献有师兄:感谢你在代码方面的悉心指导，让我对这个科研必备工具有了更深理解；林啸、彭云、杨静葳、方琴、何宗涛、刘创伟、张屹康等师兄师姐：感谢你们在这段旅程中给予的关心和帮助，特别要感谢你们包容我这个“话痨”。

感谢冯翊、郭思岑、马羽、郭子瞻、张孟谈、赵鸿博、黄知为、唐冠峰、樊嘉禾等同门在学术上给我的启发；感谢魏星、陈春春、田浩、杨佳怡等 627 实验室的伙伴们，是你们营造的融洽氛围让我度过了美好的三年时光；感谢电信楼 625-629 的所有同学，能与如此优秀的你们共同进步，是我莫大的荣幸；感谢遇到的其他老师、师兄以及朋友，在我的学术研究进入低谷，前行之路遇到困难的

时候，给予我的无私帮助。

最后，要深深感谢我的父母和家人。你们从不苛求我的成就，只关心我的健康与快乐。这份不求回报的爱，是我今生最宝贵的财富。

临别在即，千言万语难诉衷肠。惟愿以诸葛武侯之言寄怀：“今当远离，临表涕零，不知所言。”

李佳航

2025 年 5 月

同济 · 嘉定

个人简历、在读期间取得的学术成果

已发表论文:

[1] **Li J.**, Zhang Y., Yun P., Zhou G., Chen Q., Fan R.: RoadFormer: Duplex Transformer for RGB-Normal Semantic Road Scene Parsing[J]. 第一作者. IEEE Transactions on Intelligent Vehicles, Vol.: 9, No.: 7, pp.: 5163 - 5172, 2024. DOI: 10.1109/TIV.2024.3388726. (中科院 TOP 一区，IF: 14；对应论文第二章).

[2] Huang J*, **Li J.***, Jia, N., Sun, Y., Liu, C., Chen, Q., Fan, R.: RoadFormer+: Delivering RGB-X Urban Scene Parsing through Scale-Aware Information Decoupling and Advanced Heterogeneous Feature Fusion[J]. 共同第一作者. IEEE Transactions on Intelligent Vehicles, 2024. DOI: 10.1109/TIV.2024.3448251. (中科院 TOP 一区，IF: 14).

[3] Guo S.*, **Li J.***, Feng Y., Zhou D., Zhang D., Su S., Zhu X., Chen Q., Fan R.: UDTIRI: An Online Open-Source Intelligent Road Inspection Benchmark Suite[J]. 共同第一作者. IEEE Transactions on Intelligent Transportation Systems, Vol.: 25, No.: 8, pp.: 9920 - 9931, 2024. DOI: 10.1109/TITS.2024.3351209. (中科院 TOP 一区，IF: 7.9).

[4] Fan R., **Li J.**, Li J., Wang J., Long Z., Jia N., Liu Y., Wang W., Bocus J. M., Vityazev S., Chen X., Xiao J., Andreev S., Lu H., Dvorkovich A.: A Glance over the Past Decade: Road Scene Parsing towards Safe and Comfortable Autonomous Driving[J]. 导师一作，学生二作. Autonomous Intelligent Systems, Vol.: 5, No.: 8, pp.: 1 - 15, 2025. DOI: 10.1007/s43684-025-00096-y. (对应论文第一章)

[5] Huang J., **Li J.**, Vityazev S., Dvorkovich A., Fan R.: DepthMatch: Semi-Supervised RGB-D Scene Parsing through Depth-Guided Regularization[J]. 第二作者. IEEE Signal Processing Letters (JCR 二区，IF: 3.2).

[6] Fan R., Zhang Y., Guo S., **Li J.**, Feng Y., Su S., Zhang Y., Wang W., Jiang Y., Bocus J. M., Zhu X., Chen Q.: Urban Digital Twins for Intelligent Road Inspection[C]. 第四作者. 2022 IEEE International Conference on Big Data (IEEE Big Data), pp.: 5110 - 5114, 2022. DOI: 10.1109/BigData55660.2022.10021042. (CCF-C 会议)

[7] Yun P., Liu Y., Yan X., **Li J.**, Wang J., Tai L., Jin N., Fan R., Liu M.: 3D Object Detection in Autonomous Driving[J]. Autonomous Driving Perception: Fundamentals and Applications. Springer Nature Singapore, pp.: 139-173, 2023. DOI: 10.1007/978-981-99-4287-9_5. (章节书合著者)

[8] Guo S., Jiang Y., **Li J.**, Zhou D., Su S., Bocus J. M., Zhu X., Chen Q., Fan R.: Road Environment Perception for Safe and Comfortable Driving[J]. Autonomous Driving Perception: Fundamentals and Applications. Springer Nature Singapore, pp.: 357-387, 2023. DOI: 10.1007/978-981-99-4287-9_11. (章节书合著者)

投稿论文:

[1] **Li J.**, Yun P., Chen Q., Fan R.: HAPNet: Toward Superior RGB-Thermal Scene Parsing via Hybrid, Asymmetric, and Progressive Heterogeneous Feature Fusion[J]. 第一作者. Science China Information Sciences (中科院 TOP 一区，IF: 7.3, 已投在审，ID: SCIS-2025-0566, 对

应论文第三章).

[2] **Li J.**, Fan R.: OVDepth: Incorporating Depth Prior into Open-Vocabulary Scene Parsing via Task Decoupling[J]. **第一作者**. IEEE Transactions on Circuits and Systems for Video Technology (中科院 TOP 一区，IF: 8.4, 待投，**对应论文第四章**).

[3] Tang G., Wu Z., **Li J.**, Fan R.: TiCoSS: Tightening the Coupling between Semantic Segmentation and Stereo Matching within A Joint Learning Framework[J]. **第三作者**. IEEE Transactions on Automation Science and Engineering (中科院二区，IF: 5.9, 已投在审，ID: T-ASE-2025-420.R1).

[4] Zhang Y., Liu C.W., **Li J**, Chen Y., Cheng J., Fan R.: Establishing Reality-Virtuality Interconnections in Urban Digital Twins for Superior Intelligent Road Inspection[J]. **第三作者**. IEEE Robotics and Automation Letters (中科院二区，IF: 4.6, 已投在审，ID: 25-1726).

专利与软著:

[1] 发明专利：一种基于 RGB 和热红外信息的多模态融合语义感知方法，专利申请号：202410541391.1. [导师第一发明人，学生第二发明人] (实质审查)

[2] 发明专利：一种数据融合道路场景语义感知方法、装置及介质，专利申请号：202311379862.5. [导师第一发明人，学生第二发明人] (实质审查)

[3] 发明专利：一种基于数据融合的城市场景语义分割方法和电子设备，专利申请号：2024107844207. [导师第一发明人，学生第二发明人] (实质审查)

[4] 软件著作权：基于色彩和法向量信息的数据融合道路可行驶区域与破损检测系统 V1.0，登记号：2024SR0841764.

荣誉获奖:

[1] 2023-2024 年度同济大学国奖奖学金（电子与信息工程学院 2024 年度硕士总分第一名）

[2] 2025 年同济大学校级优秀毕业生

[3] 2023-2024 年度同济大学优秀学生

[4] 2023-2024 年度同济大学 MIAS Group 学术之星