分类号：TP391.4　　单位代码：10422
密　级：公开　　学　　号：

山东大学
SHANDONG UNIVERSITY

硕士学位论文

Thesis for Master Degree

（专业学位）

论文题目：基于语义引导的场景识别算法研究

Study on Semantic-guided Scene Recognition

作　者　姓　名　宋传鑫
培　养　单　位　控制科学与工程学院
专　业　名　称　电子信息
指　导　教　师　马昕　教授
合　作　导　师

2024 年 5 月 30 日

原 创 性 声 明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律责任由本人承担。

论文作者签名：_____ 日 期：2024 年 5 月 30 日

关于学位论文使用授权的声明

本人同意学校保留或向国家有关部门或机构送交论文的印刷件和电子版，允许论文被查阅和借阅；本人授权山东大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

(保密论文在解密后应遵守此规定)

论文作者签名：_____导师签名：_____日 期：2024 年 5 月 30 日

摘　要

场景识别是计算机视觉领域的一个研究热点，旨在识别图像中展现的场景类别，如“卧室”、“客厅”和“厨房”等。这一技术是机器人在未知环境下执行任务的前提，同时在自动驾驶和人机交互等众多领域发挥着关键作用。虽然深度学习技术的兴起使计算机视觉领域取得了显著进展，但场景识别仍然面临重大挑战。这主要源自场景中复杂多变的空间布局导致的类内巨大差异，以及不同场景中出现的相同物体引起的类间相似现象。本文以深度学习为核心技术路径，基于场景内语义信息的引导，对场景识别算法进行了深入研究，开展了以下创新性工作：

（1）为充分探索场景内复杂多变的空间布局，提出了基于语义空间关系建模的场景识别方法。首先，借助语义分割技术解耦场景中的空间信息，避免空间无关信息对建模过程的干扰；其次，设计了一个自适应置信度滤波器，有效缓解语义分割性能有限所引起的语义歧义问题；然后，根据语义分割结果的特性构建了语义空间关系模块，探索场景内空间信息的高级表示；最后，在探索出的空间特征的基础上，构建了组合语义空间关系模块，全面利用场景内信息提高场景识别精度。在多个公开数据集上的实验结果表明，所提方法对空间关系的充分建模能够显著提高场景识别性能。

（2）针对不同场景中相同语义的物体引起的类间相似问题，进一步提出了融合语义空间关系和物体共现性建模的场景识别框架。首先，利用语义空间关系模块和图像特征提取模块提取场景特征；然后，构建了语义节点特征聚合模块将场景特征分配至各物体语义节点，为物体赋予场景相关特征，藉此区分不同场景中相同语义的物体，并生成语义节点特征序列；最后，设计了全局-局部依赖模块深入挖掘语义节点特征与全局特征之间的远程相关性，进而在建模物体共现性的过程中精细调整全局特征，提升场景识别性能。实验结果表明，融合场景特征建模物体共现性，可以有效克服跨场景共存物体带来的类间相似问题，获得更具区分度的场景表示。

（3）针对上述方法对物体信息的启发式利用可解释性不足的问题，从知识推理角度提出了基于物体间鉴别性图建模的场景识别方法。首先，基于概率统计捕捉物体与场景之间的鉴别性关系，并运用贝叶斯推理将该关系转化为物体间鉴别性原型；然后，提出了一个鉴别性图网络，将像素级场景特征设定为图的节点，并借助物体间先验知识将节点之间的鉴别相关性编码为边；最后，通过图卷积操作，将物体间鉴别相关性与场景特征显式融合，实现更具可解释性地利用物体信息辅助场景识别。在公开数据集上的实验结果表明，从知识推理角度入手融合物体知识和场景特征，使网络在单分支架构下即可获得极具竞争力的场景识别性能。

（4）针对物体信息提取对算力的高需求限制了上述语义引导方法的实际部署这一问题，提出了基于类间语义相似度原型的场景识别方法。首先，基于语义分割和统计分析捕获场景中物体信息分布情况，并据此构建类间语义相似度原型；然后，设计了渐进标签软化和批量级对比损失两种训练策略，前者将相似度原型嵌入到标签中，利用置信度渐进策略指导网络训练，后者则基于相似度原型衡量训练批次中样本间的相似度要求；最后，在训练过程中基于上述两种策略增强场景识别网络的特征辨识能力，从而提升场景识别性能。在公开数据集上针对多个基准网络的比较实验表明，相似度原型及相关策略可实现在不增加额外计算负担的前提下，利用物体信息提升场景识别性能。

关键词：场景识别；语义知识；注意力机制；图神经网络；数据推理

ABSTRACT

Scene recognition is a research hotspot in the field of computer vision, aiming to identify the scene categories presented in images, such as "bedroom," "living room," and "kitchen." Scene recognition is a prerequisite for robots to perform tasks in unknown environments and plays a key role in various fields such as autonomous driving and human-computer interaction. Although the emergence of deep learning techniques has led to dramatic advances in computer vision, scene recognition still faces significant challenges. This mainly stems from the large intra-class differences caused by the complex and variable spatial layout, as well as the inter-class similarity caused by the coexisting objects across scenes. Centering on deep learning as the core technological pathway and guided by the semantic information within scenes, this thesis conducts an in-depth study of scene recognition, resulting in the following innovative works:

(1) To fully explore the complex and variable spatial layout within scenes, a semantic spatial relationship modeling-based scene recognition method is proposed. First, the spatial information in the scene is decoupled with the help of semantic segmentation, avoiding the interference of space-irrelevant information in the modeling process. Second, an Adaptive Confidence Filter is designed to mitigate the semantic ambiguity caused by the limited precision of semantic segmentation. Then, based on the semantic segmentation result characteristics, a Semantic Spatial Relation Module is designed to explore the advanced representation of spatial information within scenes. Finally, a Combined Semantic Spatial Relation Module is constructed based on the explored spatial features to comprehensively utilize the in-scene information to improve the scene recognition accuracy. Experimental results on several publicly available datasets show that adequate modeling of spatial relations significantly enhances scene recognition performance.

(2) To address the inter-class similarity caused by objects with the same semantics across different scenes, this thesis proposes a scene recognition framework that integrates semantic spatial relation and object co-occurrence modeling. First, the Semantic Spatial Relation Module and the Image Feature Extraction Module are used to extract scene features. Then, the Semantic Node Feature Aggregation Module is constructed to allocate scene-related features to each object semantic node, so as to differentiate the objects with the same semantics in different scenes and generate the semantic node feature sequence. Finally, the

Global-Local Dependency Module is designed to deeply explore the long-range correlation between semantic node features and global features, and then fine-tune the global features when modeling object co-occurrence, improving the scene recognition performance. Experimental results show that integrating scene features to model object co-occurrence can effectively overcome the inter-class similarity caused by coexisting objects across scenes, obtaining a more distinctive scene representation.

(3) To address the insufficient interpretability in the heuristic use of object information by existing methods, a scene recognition method based on inter-object discriminative graph modeling is proposed. First, capturing the object-scene discriminative relationships based on probabilistic statistics, which are transformed into an Inter-Object Discriminative Prototype using Bayesian inference. Then, a Discriminative Graph Network is proposed, setting pixel-level scene features as graph nodes, and encoding the discriminative correlations between nodes into edges with the help of prior knowledge among objects. Finally, through graph convolution operation, we explicitly fuse inter-object discriminative correlations with scene features, and realize a more interpretable use of object information to assist scene recognition. Experimental results show that integrating object knowledge and scene features from a knowledge inference perspective allows the network to achieve competitive scene recognition performance under a single-branch architecture.

(4) Considering that the high computational demand for object information extraction limits the practical deployment of semantic-guided scene recognition methods, this thesis proposes an inter-class semantic similarity prototype-based scene recognition method. First, capturing the object information distribution within scenes based on semantic segmentation and statistical analysis, and the inter-class semantic Similarity Prototype is constructed accordingly. Then, we design two training strategies: Gradient Label Softening and Batch-level Contrastive Loss. The former embeds the similarity prototype in soft labels and uses a confidence gradient strategy to guide network training. The latter uses the similarity prototype to measure the similarity requirements between samples in each mini-batch. Finally, improving the feature discrimination ability of the scene recognition network based on the above two strategies during training, so as to enhance scene recognition performance. Comparative experiments on public datasets for multiple benchmark networks show that the similarity prototype and related strategies can utilize object information to improve scene recognition performance without adding extra computational burden in practice.

Key words: Scene recognition; Semantic knowledge; Attention mechanism; Graph neural network; Data reasoning

目　　录

摘　　要 I
ABSTRACT III
1 绪论 1
1.1 研究背景与意义 1
1.2 国内外研究现状 3
1.2.1 非语义引导的场景识别方法 3
1.2.2 基于语义引导的场景识别方法 5
1.3 研究问题与挑战 7
1.4 本文主要研究内容及章节安排 8
2 基于语义空间关系建模的室内场景识别方法 11
2.1 引言 11
2.2 语义空间关系特征提取 12
2.2.1 自适应置信度滤波器 13
2.2.2 空间关系建模过程 13
2.3 面向场景识别的全局特征提取 14
2.3.1 PlacesCNN 模块 15
2.3.2 基于条带深度卷积的全局聚合模块 15
2.4 场景识别常用数据集介绍 16
2.4.1 MIT-67 数据集 16
2.4.2 SUN397 数据集 17
2.4.3 Places 数据集 17
2.5 实验结果与分析 19
2.5.1 实验设置 19
2.5.2 语义空间关系模块架构的消融实验 20
2.5.3 物体信息数量的影响 21
2.5.4 不同的特征聚合策略的对比 22
2.5.5 与现有算法的性能比较 23
2.6 本章小结 25

3 融合语义空间关系和物体共现性建模的室内场景识别方法 27
3.1 引言 27
3.2 融合语义空间关系和物体共现性的场景识别框架 28
3.2.1 总体架构 28
3.2.2 语义空间关系模块 29
3.2.3 图像特征提取模块 29
3.2.4 语义节点特征聚合模块 29
3.2.5 全局-局部依赖模块 31
3.2.6 场景分类器 32
3.3 实验结果与分析 33
3.3.1 实验设置 33
3.3.2 不同的语义节点特征聚合策略的比较 33
3.3.3 不同的编码器组合策略的对比 35
3.3.4 网络组成成分分析 35
3.3.5 特征可视化 36
3.3.6 与现有算法的性能比较 39
3.4 本章小结 40
4 基于物体间鉴别性图建模的室内场景识别方法 41
4.1 引言 41
4.2 基于物体间鉴别性图建模的室内场景识别方法 42
4.2.1 总体架构 42
4.2.2 物体间鉴别性原型 43
4.2.3 鉴别性图网络 46
4.3 实验结果与分析 49
4.3.1 实验设置 49
4.3.2 物体间鉴别性原型的设计细节选择 50
4.3.3 超参数选择 52
4.3.4 网络组成成分分析 54
4.3.5 特征可视化 54
4.3.6 不同 CNN 架构的对比 56
4.3.7 与现有算法的性能比较 56

4.4 本章小结 58
5 基于语义嵌入的类间相似度原型的场景识别方法 59
5.1 引言 59
5.2 基于语义嵌入的类间相似度原型的场景识别方法 60
5.2.1 总体概述 60
5.2.2 基于语义嵌入的相似度原型 61
5.2.3 基于相似度原型的渐进标签软化 64
5.2.4 基于相似度原型的批量级对比损失 67
5.3 实验结果与分析 69
5.3.1 实验设置 69
5.3.2 算法框架中不同组件配置的对比 70
5.3.3 网络组成成分分析 71
5.3.4 特征可视化 75
5.3.5 计算成本评估 76
5.3.6 与现有算法的性能比较 77
5.4 本章小结 79
6 总结与展望 81
6.1 工作总结 81
6.2 研究展望 82
参考文献 85
致　　谢 93
攻读硕士学位期间的科研成果与所获奖励 95

CONTENTS

Chinese abstract.....I
English abstract III
Chapter 1 Introduction 1
1.1 Research background and significance 1
1.2 Research status at home and abroad.....3
1.2.1 Non-semantic guided scene recognition methods.....3
1.2.2 Semantic guided scene recognition methods 5
1.3 Issues and challenges 7
1.4 Main research content and chapter arrangement of this thesis.....8
Chapter 2 Semantic spatial relation modeling based indoor scene recognition.....11
2.1 Introduction 11
2.2 Semantic spatial relation feature extraction 12
2.2.1 Adaptive confidence filter 13
2.2.2 Process of spatial relationship modeling.....13
2.3 Global feature extraction for scene recognition 14
2.3.1 PlacesCNN module 15
2.3.2 Global aggregation module based on strip depth-wise convolution 15
2.4 Introduction to common scene recognition datasets 16
2.4.1 MIT-67 dataset.....16
2.4.2 SUN397 dataset.....17
2.4.3 Places dataset.....17
2.5 Experimental results and analysis 19
2.5.1 Experimental setup.....19
2.5.2 Ablation experiment of semantic spatial relation module architecture 20
2.5.3 Effect of the number of object information.....21
2.5.4 Comparison of different feature aggregation strategies 22
2.5.5 Comparison with the state-of-the-arts.....23
2.6 Summary of this chapter.....25
Chapter 3 Semantic-guided spatial relation and object co-occurrence modeling for indoor

scene recognition 27
3.1 Introduction 27
3.2 Scene recognition framework fusing semantic spatial relationships and object co-occurrence 28
3.2.1 Overall architecture 28
3.2.2 Semantic spatial relation module 29
3.2.3 Image feature extraction module 29
3.2.4 Semantic node feature aggregation module 29
3.2.5 Global-local dependency module 31
3.2.6 Scene classifier 32
3.3 Experimental results and analysis 33
3.3.1 Experimental setup 33
3.3.2 Comparison of different semantic node feature aggregation methods 33
3.3.3 Comparison of different encoder combination strategies 35
3.3.4 Network component analysis 35
3.3.6 Comparison with the state-of-the-arts 39
3.4 Summary of this chapter 40
Chapter 4 Inter-object discriminative graph modeling for indoor scene recognition 41
4.1 Introduction 41
4.2 Inter-object discriminative graph modeling for indoor scene recognition 42
4.2.1 Overall architecture 42
4.2.2 Inter-object discriminative prototype 43
4.2.3 Discriminative graph network 46
4.3 Experimental results and analysis 49
4.3.1 Experimental setup 49
4.3.2 Design detail options for inter-object discriminative prototype 50
4.3.3 Hyperparameter options 52
4.3.4 Network component analysis 54
4.3.5 Feature visualization 54
4.3.6 Comparison of different CNN architectures 56
4.3.7 Comparison with the state-of-the-arts 56

4.4 Summary of this chapter 58
Chapter 5 Semantic-embedded similarity prototype for scene recognition 59
5.1 Introduction 59
5.2 Semantic-embedded similarity prototype for scene recognition 60
5.2.1 Overall overview 60
5.2.1 Semantic-embedded similarity prototype 61
5.2.3 Similarity prototype-based gradient label softening 64
5.2.4 Similarity prototype-based batch-level contrastive loss 67
5.3 Experimental results and analysis 69
5.3.1 Experimental setup 69
5.3.2 Comparison of different component configurations in algorithmic framework 70
5.3.3 Network component analysis 71
5.3.4 Feature visualization 75
5.3.5 Computational cost evaluation 76
5.3.6 Comparison with the state-of-the-arts 77
5.4 Summary of this chapter 79
Chapter 6 Summary and outlook 81
6.1 Work summary 81
6.2 Research outlook 82
Reference 85
Acknowledgments 93
Academic achievements and awards during the study for master's degree 95

1 绪论

1.1 研究背景与意义

随着科技的飞速发展，人工智能已经成为推动现代社会发展的核心力量。2017 年，国务院发布我国《新一代人工智能发展规划》[1]，标志着人工智能被正式纳入国家的重点发展战略。2022 年，科技部等六个部门又印发了《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》[2]，强调我国对于推动人工智能技术市场化应用的坚定支持。在人工智能的众多研究领域中，计算机视觉技术一直是研究的热点。通过模仿人类视觉系统的工作原理，计算机视觉技术让机器具备“看见”和“理解”世界的能力，为机器代替人类执行智能化任务奠定了基础。

作为计算机视觉领域中的一个重要研究主题，场景识别（Scene recognition）因其广阔的应用前景而备受关注。场景识别的目标是为给定的图像或视频帧分配预定义的场景类别标签，如图 1-1 所示，这些类别标签是对场景内部的物体、布局以及整体风格的综合性概括[3]。随着人工智能技术的不断进步，场景识别已被广泛应用于智能机器人[4,5]、自动驾驶汽车[6,7]、图像和视频检索[8]以及人机交互[9]等众多领域，成为实现智能化应用的关键基础。



图 1-1 几个场景类别和物体类别的对比

Fig. 1-1 Comparison of several scene categories and object categories

具体来说，在智能机器人领域[4,5]，场景识别赋予了机器人理解和适应周遭环境的能力，使其能够更有效地执行复杂任务，从而大幅提升了工作效率和安全性。在自动驾驶领域[6,7]，实时场景识别使自动驾驶车辆能够在复杂多变的交通环境中迅速做出准确的决

策与响应，极大地提高了行驶安全性。在图像和视频检索领域[8]，通过场景识别技术理解图像和视频中的内容，检索技术突破了传统的基于元数据或关键词的方法，实现了基于实际视觉信息的高精度检索，显著提升了检索效率。在人机交互领域[9]，通过识别用户当前所处的场景，智能设备能够为其提供更加情境化的服务，使人机交互技术更加贴合用户的需求。

综上所述，场景识别技术广泛应用于多个智能领域，具有极高的研究意义和应用前景。近年来，得益于海量数据的积累和计算能力的提升，深度学习技术在计算机视觉领域取得了突破性进展，这一进步也为场景识别任务的发展提供了强大动力。然而，场景识别的挑战性远超一般的物体分类（Object classification）任务。以图 1-1 为例，同一场景类别内多变的空间布局（如“图书馆”、“机房”和“餐厅”）导致了显著的类内差异问题；而不同场景中共存的物体（如“椅子”同时出现在“机房”和“餐厅”场景中、“架子”同时出现在“图书馆”和“鞋店”场景中）也带来了类间相似问题。这些挑战限制了神经网络的泛化能力，使其在场景识别任务上的表现不如在处理物体分类等基础任务时那样出色。总体而言，场景图像的复杂性主要源于图像各个局部区域中多样的物体分布，因此，这些区域的物体语义信息（如图 1-2 所示）之间的关系可以为场景识别提供重要线索。最近对场景本质的研究[10]也指出，场景识别与场景中的语义信息有着密切的关联。因此，基于场景内语义信息引导的场景识别算法值得深入研究。



图 1-2 场景内语义信息的示例

Fig. 1-2 Examples of semantic information within scenes

本文在国家自然科学基金-区域创新发展联合基金项目“可视度受限环境下跨光谱多传感信息融合的机器人语义感知与交互协作”（U21A20486）和山东省自然科学基金重大基础研究项目“面向康养服务机器人的人-机-环境共融基础理论与关键技术”（ZR2019ZD07）的资助下，针对现有的难点问题展开深入研究，提出了一系列基于语义引导的场景识别算法，为机器人环境感知提供技术支撑。

1.2 国内外研究现状

目前，场景识别已成为计算机视觉领域的热门研究方向，近年来涌现出了各种各样的场景识别算法。本节按照是否利用场景内的语义信息将这些方法分为了两大类：非语义引导的场景识别方法和基于语义引导的场景识别方法。如图 1-2 所示，本文所指的语义信息是场景内各个局部区域内的具体物体，如床、窗户和屏幕等。利用这些场景内的语义信息进行辅助的方法，被称为语义引导的建模方法。下面首先介绍非语义引导的场景识别方法，然后介绍与本文紧密相关的基于语义引导的场景识别方法。

1.2.1 非语义引导的场景识别方法

（1）基于手工特征表示的方法

21 世纪初期，早期的场景识别方法主要依赖于手工设计的局部视觉描述符，如尺度不变特征转换（Scale Invariant Feature Transform，SIFT）[11]、方向梯度直方图（Histogram of Oriented Gradient，HOG）[12]和局部二值模式（Local Binary Pattern，LBP）[13]等。这些描述符能够提取图像中的低级特征，专注于捕捉局部细节，但由于缺乏语义和全局信息，其应用范围主要限于简单的自然场景（如沙滩等）。然而，面对复杂场景时，这些描述符的性能表现通常较为有限。

为了改善场景识别效果，研究者们开发了多种策略，将上述低级手工描述符汇聚成中级表示形式[14]。一个典型的例子是视觉词袋模型（Bag of Visual Words，BoVW）[15]，即首先基于局部描述符提取低级视觉特征，然后通过 BoVW 技术将这些特征聚类成词袋表示。BoVW 在场景识别任务中效果显著，但忽略了场景中的空间结构信息。为了解决这一问题，Lazebnik 等[16]提出了空间金字塔匹配（Spatial Pyramid Matching，SPM）技术，即先将图像划分为细小的子区域，并分别提取各个区域的词袋表示，然后利用 SPM 技术整合这些区域表示，从而弥补了缺失的空间结构信息，在自然场景数据集上取得了明显的性能提升。2009 年，麻省理工学院的研究人员[17]进一步构建了室内场景数据集 MIT-67，并提出了一个基于原型（Prototype）的模型，结合局部和全局信息进行室内场景识别，为室内场景识别取得了标志性的进展。

总体而言，通过多种编码方法增强的手工图像表示法在场景识别领域取得了显著成就。然而，由于手工描述符本身的低级和易受干扰特性，其在提升场景图像表示的区分能力方面存在局限。随着场景识别需求的不断发展，手工描述符难以有效表征日益复杂的场景图像。

（2）基于深度神经网络的方法

自 2009 年以来，得益于计算能力的显著增强和大型数据集（如 ImageNet[18]和 Places365[19]）的推出，深度学习技术在计算机视觉领域取得了突破性进展。众多深度神

经网络架构，包括 AlexNet[20]、VGGNet[21]、ResNet[22]以及 Vision Transformer[23]等，已在此领域取得了卓越成就。相较于传统的手工特征，这些网络在数据丰富的环境中能够自动学习并提取出更具鉴别能力的特征，极大地推动了图像分类技术的发展。随着深度学习技术的不断革新，越来越多的研究开始尝试利用深度神经网络提取用于场景识别的特征表示。

基于深度特征与传统特征编码相结合的方法：在深度学习技术涌现初期，研究者们倾向于将神经网络特征与传统特征编码方法结合使用。2015 年，加利福尼亚大学的研究者们[24]提出，将语义费雪向量（Fisher Vector，FV）计算为神经网络的 *Softmax* 层所输入的高斯混合 FV，他们不仅利用这种语义 FV 进行场景识别，还在后续工作[25]中进一步将其实现为了端到端的方式。此外，文献[26]也创新性地提出了一个字典学习层，用以替换深度神经网络架构中的全连接层和激活层，基于字典结构，他们设计了一种新型的标签鉴别性回归器，并通过优化字典选择进一步增强特征的稀疏表示，获得了更高的识别精度。

上述方法通过传统编码方法有效整合了深度神经网络所提取的特征，显著提升了场景识别性能。尽管如此，随着数据量的持续增长，传统特征编码方法在整合能力上逐渐显得不足，研究者开始减少对这些编码方法的依赖。

基于显著区域的方法：得益于在大型数据集上预训练的卷积神经网络（Convolutional Neural Network，CNN）权重[18,19]的开源，基于显著区域的场景识别算法[27-30]逐渐崭露头角。例如，2018 年，Adi-Red[27]提出利用预训练的 ResNet 网络定位图像中不同尺度的显著区域，并且在调整这些区域至适当尺寸后，利用三个独立的网络分支提取多尺度的特征表示，显著提升了场景识别的准确性。类似地，中科院计算所的蒋树强团队[28]于 2020 年提出了与原型无关的场景布局（Prototype-agnostic Scene Layout，PaSL）选取场景内的代表性区域。具体来讲，PaSL 利用预训练的 CNN 生成的卷积激活图定位图像中的鉴别性区域，并进一步将这些区域特征作为图的节点，将节点特征之间的空间及相似度关系编码为图的边，利用图卷积网络（Graph Convolutional Network，GCN）[31]探索区域间的上下文关系，从而进一步提升了场景识别的性能。这些研究依托于深度神经网络强大的特征鉴别能力，通过建模显著区域特征之间的上下文关系显著提高了场景识别性能。

综上所述，深度学习技术已经使非语义引导的场景识别领域取得了卓越的进展。这些方法主要依托于预训练神经网络的特征提取和鉴别能力，选取具有代表性的特征进行场景识别。然而，不同的场景类别可能共享易于产生相似场景表示的物体信息，导致网络面临难以处理的类间相似问题[32]，即使在同一场景类别内，也可能出现多样化的空间布局，这进一步增加了类内差异问题[28]。这些问题严重限制了深度神经网络在场景识别领域的泛化能力[33]，导致预训练网络的高响应区域可能错误地聚焦在非鉴别区域。为了

应对这些挑战，越来越多的研究者开始专注于探索场景中的物体上下文关系，以期基于场景内语义信息的引导提升场景识别性能。

1.2.2 基于语义引导的场景识别方法

（1）基于手工特征表示的方法

早在手工描述符时期，研究者们就认识到了场景内物体语义信息对于理解场景的重要性。2007 年，加拿大的研究人员[34]探讨了通过分析九种常见语义概念（包括天空、水、草、树干、叶子、田地、岩石、花朵和沙子）的出现频率描述自然场景图像的方法，尝试利用不同场景之间的共性辅助场景识别。然而，九种语义概念的表达能力有限，仅适用于描述简单的自然场景，难以全面准确地表示复杂场景[14]。进一步地，斯坦福大学的李飞飞等研究者[35]借助一种“广义物体卷积”作为物体检测器对图像进行多尺度卷积处理，并依托于卷积响应，构建了一个“物体库”表示框架，他们以物体库为基础，通过空间金字塔匹配技术整合物体特征，获得了更具描述性的场景表示，有效提高了场景识别的性能。

上述研究基于手工设计的语义描述符获取场景内的物体语义信息，通过探索这些信息之间的相互关系辅助场景识别，取得了值得注意的效果。尽管如此，手工描述符因其固有的低泛化能力，在处理较为复杂的场景时，提取语义信息的能力有限，这限制了其对场景识别的辅助效果。

（2）基于深度神经网络的方法

随着大规模数据集的出现和深度学习技术的发展，物体分类、目标检测和语义分割等技术日渐成熟。这促使研究者们越来越多地探索将这些技术与场景识别结合的可能性，催生了一系列基于深度学习和语义引导的先进场景识别方法。

基于区块级特征上下文的方法：基于在大规模数据集[18]预训练的物体分类网络，一些研究[36-38]首先利用稠密网格将场景图像细分为若干区块（Patch），然后通过分类网络提取这些区块的特征以辅助场景识别。例如，南京大学的王利民教授团队[36]提出了 PatchNet，为场景内的各个区块生成区域级语义特征。利用语义编码技术将这些特征聚合为全局表示，从而实现场景识别。进一步地，Xie 等[37]在区块级特征的基础上，提出了一种层次化卷积特征编码方法，具体来讲，利用非负稀疏分解模型（Non-Negative Sparse Decomposition，NNSD）学习不同尺度区块间的共享内容，以生成全局特征，在此基础上，采用类间线性编码（Inter-Class Linear Coding，ICLC）捕获鉴别性内容，通过层次化编码方式获得更精确的场景表示，显著提高了场景识别性能。这些方法通过建模不同区块特征之间的上下文关系辅助场景识别，取得了显著效果。

上述研究依赖于随机裁剪或者固定尺寸的网格划分场景区块，然而，这种做法可能导致单个物体被分散到多个区块中，或多个物被集中到一个区块内，随之而来的歧义问

题限制了模型性能。

基于物体间空间关系建模的方法：为了克服上述挑战，研究者们[33,39,40]转向利用目标检测或语义分割这种物体信息提取技术获取场景内物体的具体分布情况。考虑到场景内复杂多变的空间布局引起的巨大类内差异，一些研究[41-44]通过探索物体间的空间关系进行场景识别。例如，Song 等[41]在获取场景内物体的位置信息后，提出利用“物体-关系-物体”三元组探索物体之间的空间顺序关系，他们基于循环神经网络（Recurrent Neural Network，RNN）将该三元组编码为特征表示，以辅助场景识别。相比之下，于 2022 年提出的 ARG-Net[42]则通过语义分割技术获取场景中各个区域的语义信息，该方法从图（Graph）的角度入手，将不同语义区域的特征作为图的节点，而给个区域之间的空间度量关系作为图的边，并基于图卷积网络获得极具辨识性的图像表示，从而实现场景识别。这些方法通过探索场景内部的空间布局，显著提高了场景表示的辨识度，取得了令人鼓舞的场景识别效果。

上述研究通常只关注一种空间关系，然而实际上，场景中的空间关系不仅限于顺序和度量关系，还涉及到拓扑关系（如枕头被床所包围）。现有研究未能实现对场景中多样化空间关系的充分探索。

基于物体共现性建模的方法：除了复杂多变的空间布局，跨场景共存的物体也带来了不同场景类别之间的高相似问题，一些研究[32,45,46]通过建模不同场景中的物体共现性辅助场景识别。例如，清华大学的鲁继文教授团队提出了 SDO-Net[32]，该方法首先分析不同场景中物体之间的共现模式，选取具有代表性的物体，随后基于这些代表性物体知识丢弃非鉴别性物体相关的特征，以缓解跨场景常见物体带来的类间相似问题，从而提升场景间的可辨别性。类似地，香港中文大学的研究团队于 2021 年提出的 BORM[45]也采用了概率统计手段建模场景中的物体共现关系，该方法利用代表性物体提升物体特征的鉴别性，并将物体特征与场景特征相结合，以实现场景识别。这些方法首先利用神经网络对场景内的物体信息进行编码，在代表性物体知识的辅助下生成物体特征，然后将物体特征与输入场景的全局特征相结合，从而提升场景识别性能。

需要注意的是，某些场景中的代表性物体可能是一致的，例如，床既是“卧室”场景的代表性物体，也是“病房”场景的代表性物体，这种现象给上述基于物体共现性建模的方法带来了重大挑战。此外，上述方法大多采用独立分支分别处理输入场景以及场景内物体信息，在这种多分支架构中，场景特征和物体特征通过级联或者相加等方式融合，然而这些融合方式是完全启发式的。该融合方式可能会导致物体知识与场景特征的次优融合[46]。

总体而言，现有的基于语义引导的场景识别方法通过对场景内物体信息的充分利用，显著增强了场景特征的鉴别性能，获得了卓越的识别效果。然而，这些方法在实际部署时，需依赖目标检测或语义分割等技术获取场景中的具体物体信息，这一过程往往

伴随着远高于常规分类网络的计算成本。这不仅给计算资源有限的边缘设备带来了沉重负担，也严重影响了这些方法在实际应用中的运行效率。

1.3 研究问题与挑战

通过上述对场景识别技术的研究现状进行分析，可以看到迄今为止国内外科研人员已经开展了大量的研究工作，先后提出了很多先进的场景识别方法，也取得了令人瞩目的成果。尽管如此，目前基于语义引导的场景识别研究仍处于技术攻坚阶段，还存在很多需要解决的关键问题与挑战。

（1）场景内复杂多变的空间布局难以充分建模

现有工作多依赖于骨干网络提取场景特征，并通过手工定义的方式探索特征之间的某一类空间关系。然而，场景内的空间关系包括顺序、度量和拓扑关系[47]，仅仅依靠手工定义显然难以全面捕捉场景中的复杂空间布局。此外，骨干特征中的颜色等空间无关信息带来的损失下降不利于优化网络表征空间信息的能力。因此，如何有效规避空间无关信息的干扰，并实现对场景内语义空间关系的充分建模，是场景识别研究中的一个关键问题。

（2）不同场景中存在相同语义的物体引起的类间相似问题

除了场景内的空间上下文信息，不同场景中的物体共现信息对于场景识别同样意义重大。现有的基于物体共现的研究通过分析不同场景中物体的概率分布，推理出与场景类别紧密相关的代表性物体，取得了显著成效。然而，一些场景类中的代表性物体可能是一致的（如病房和卧室场景中的床），随之而来的类间相似问题对基于物体共现性的方法提出了巨大挑战。

（3）对场景内物体信息的利用可解释性有限

现有的基于物体信息辅助的场景识别研究通常采用独立的分支分别处理输入场景和物体信息，并试图从端到端的角度整合场景特征和物体特征。然而，这种多分枝架构对物体信息的完全启发式的利用可解释性有限，容易导致物体知识和场景特征的次优融合。因此，从更具可解释性的角度利用物体信息中的鉴别性知识辅助场景识别，是一个值得深入研究的方向。

（4）物体信息提取过程对算力的巨大需求限制了算法在边缘设备的实际部署

基于语义引导的场景识别方法在实际部署时，需要通过目标检测或者语义分割等技术从输入图像中提取出物体语义信息。物体信息提取过程所需的计算资源远超常规分类网络的需求。这不仅给算力有限的边缘设备带来了沉重负担，也严重影响了场景识别技术在实际应用中的运行速度。因此，有必要探讨如何在不增加额外计算负担的前提下，利用语义信息辅助提升场景识别网络的性能。

1.4 本文主要研究内容及章节安排

围绕 1.3 小节中对基于语义引导的场景识别方法现存问题的分析，本文旨在探讨如何充分且高效地利用语义信息提升场景识别性能，并据此开展了一系列深入的研究。本文的结构框架如图 1-3 所示。

全文共分为六个章节，各章节内容的安排如下：

第一章，绪论。本章首先阐述了场景识别任务的研究背景与意义；其次，简要回顾了场景识别领域的研究现状；然后，归纳总结了当前基于语义引导的场景识别技术所面临的问题与挑战；最后，介绍了本文的主要研究内容及章节安排。

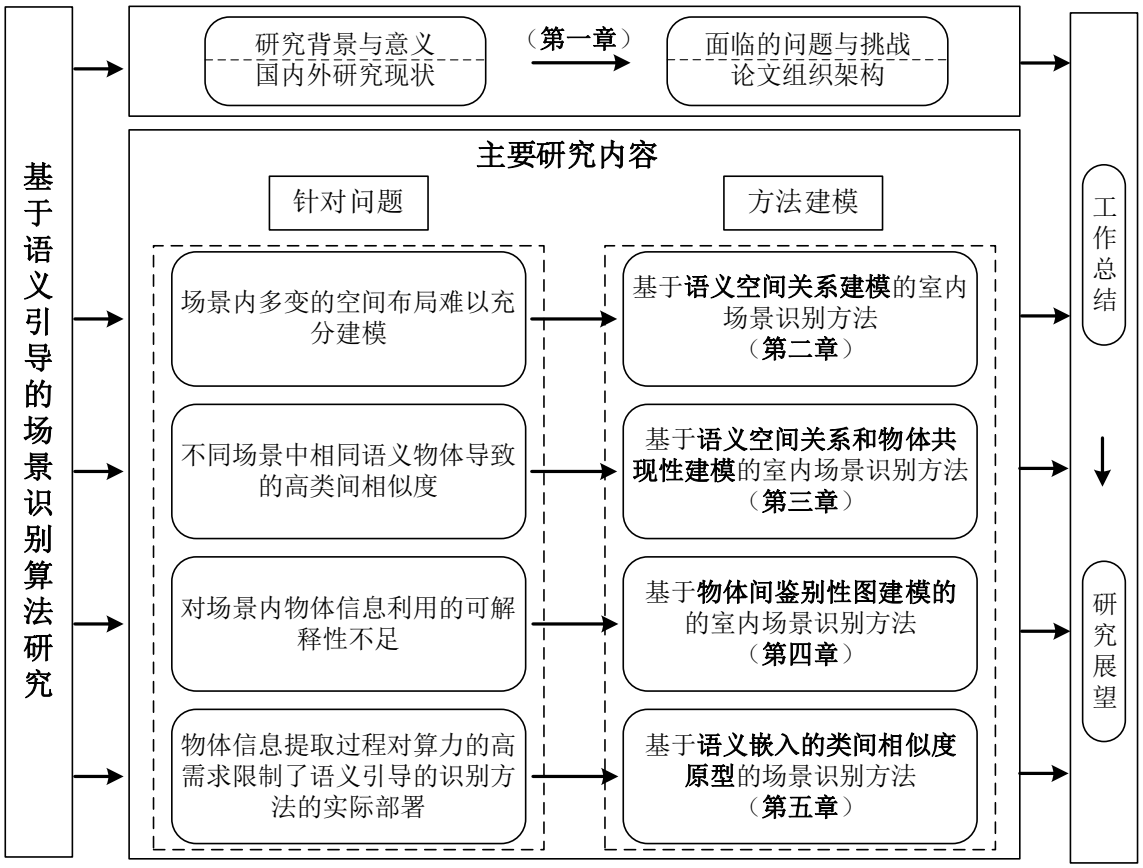


图 1-3 本文的组织架构

Fig. 1-3 Organizational structure of this thsis

第二章，基于语义空间关系建模的室内场景识别方法。本章针对场景内多变的空间布局难以充分建模的问题，提出了一种语义空间关系建模方法。首先，借助语义分割技术解耦输入图像中物体的空间信息，并设计自适应置信度滤波器以缓解语义分割精度有限所引起的语义歧义问题；然后，根据语义分割结果的特性，结合现有网络构建了一个语义空间关系模块，以探索场景内的语义空间关系；最后，在探索出的空间信息的基础

上，构建了组合语义空间关系模块，确保对场景内信息的全面利用。

第三章，融合语义空间关系和物体共现性建模的室内场景识别方法。本章针对不同场景中相同语义物体导致的类间相似度问题，提出先将场景相关特征分配给物体，再建模物体共现性的解决思路。首先，基于图像特征提取模块和第二章设计的语义空间关系模块提取场景特征；其次，构建了语义节点特征聚合模块，将场景特征分配至各物体语义节点，为物体赋予场景相关特征，藉此区分跨场景相同语义的物体，并生成语义节点特征序列；然后，构建了全局-局部依赖模块，通过注意力机制深入挖掘语义节点特征间的长期共现关系，并以此为基础精细调整全局特征，提升场景识别性能。

第四章，基于物体间鉴别性图建模的室内场景识别方法。针对场景内物体信息利用的可解释性不足的问题，本章提出了一种基于数据推理的方法，旨在以更具可解释的方式利用物体信息中隐藏的鉴别性知识辅助场景识别。首先，基于概率统计方法处理所有训练图像的语义分割结果，构建物体间鉴别性原型以建模物体间相关性；然后，将像素级场景特征作为图节点，并借助物体间鉴别性原型中的先验知识编码节点之间的鉴别性相关性；最后，通过图卷积和映射操作，将物体间鉴别性知识和场景特征显示融合，提高场景识别性能。

第五章，基于语义嵌入的类间相似度原型的场景识别方法。针对物体信息提取过程对算力的高需求限制了语义引导的场景识别方法的实际部署这一问题，本章提出了一种高效利用物体信息的方法。首先，基于语义分割和统计分析捕获各场景类别中物体信息的分布情况，并据此量化类别间相似度，构建类间语义相似度原型；然后，设计了渐进标签软化和批量级对比损失两种策略，利用相似度原型中的先验知识指导网络训练过程；最终实现在不增加额外计算负担或网络参数的前提下，利用物体信息提升场景识别性能。

第六章，总结与展望。对本文的主要研究成果进行了总结与归纳，并对尚未解决的问题进行了深入探讨，继而对未来可继续深入研究的方向做出了展望。

2 基于语义空间关系建模的室内场景识别方法

2.1 引言

如图 2-1 所示，相比于室外场景，室内场景识别由于其多样化的空间布局，面临更加严峻的挑战。因此，开发一种有效的室内场景表示方法十分关键。最近关于场景本质的研究[10]表明，场景识别与其所包含的物体间上下文关系紧密相关。基于这一点，本章旨在深入探索物体间的空间关系，利用场景内的语义空间信息辅助场景识别。



图 2-1 一些室内场景和室外场景的示例

Fig. 2-1 Some examples of indoor scenes and outdoor scenes

最近已有研究通过建模物体间的空间上下文关系辅助场景识别，取得了不错的成效。例如，文献[41]利用物体-关系-物体三联体探索物体间的空间顺序关系；文献[28,42]分析了场景中物体区域之间的空间度量关系。然而，空间关系的范畴不仅限于顺序和度量，还包括拓扑关系（如枕头被床包围的概念）[47]。显然，仅依靠人工定义的方式对一种空间关系进行建模，无法全面分析场景中复杂多变的空间布局。启发于 DNN 强大的拟合能力[22]，本章利用 DNN 以端到端的方式自适应建模场景内的空间上下文信息。需要注意的是，文献[28,42]也提出为骨干网络分配语义标签以分析物体区域特征之间的空间度量关系，但这些方法的有效性主要依赖于骨干网络提取特征的鉴别性能力。然而，骨干特征中的与空间关系无关的信息（如颜色），可能会在训练过程中阻碍对网络表示空间上下文能力的优化。相比之下，本章构建了一个语义空间关系模块（Semantic Spatial Relation Module，SSRM），旨在全面探索场景中物体间的所有空间关系。首先，借助语义分割技术解耦出输入图像中的空间信息，避免不相关信息的干扰，确保网络输入的

纯净性。然后，利用神经网络探索语义分割结果内部物体之间的空间关系。在 SSRM 建模出的空间特征的基础上，本章进一步构建了组合语义空间关系模块（Combined Semantic Spatial Relation Module，CSSRM），确保对场景内所有信息的充分利用。

此外，鉴于语义分割技术的精度有限，所有基于语义分割的方法均面临语义歧义的消极影响，这也是很少有研究直接从语义分割结果入手进行场景识别的原因之一。现有方法通常使用固定的置信度阈值策略[32,42]缓解歧义问题，但这种策略的泛化能力不足，随着数据量的增加，效果有限。为应对这一挑战，本章设计了一个简单高效的自适应置信度滤波器（Adaptive Confidence Filter，ACF），ACF 可以根据每个图像的具体情况，自适应地选择需要过滤的语义歧义点，并取得了显著的效果。

2.2 语义空间关系特征提取

考虑到场景内的空间关系包括拓扑（如枕头被床包围）、顺序（如椅子在桌子的后面）和度量关系（如画板和人之间的距离）[47]，仅仅以人工定义的方式不足以分析场景中如此多样化的空间布局。因此，基于 DNN 强大的拟合能力[22]，本章构建了语义空间关系模块（SSRM），以端到端的方式自适应建模场景内的空间上下文关系。图 2-2 展示了 SSRM 的整体框架。

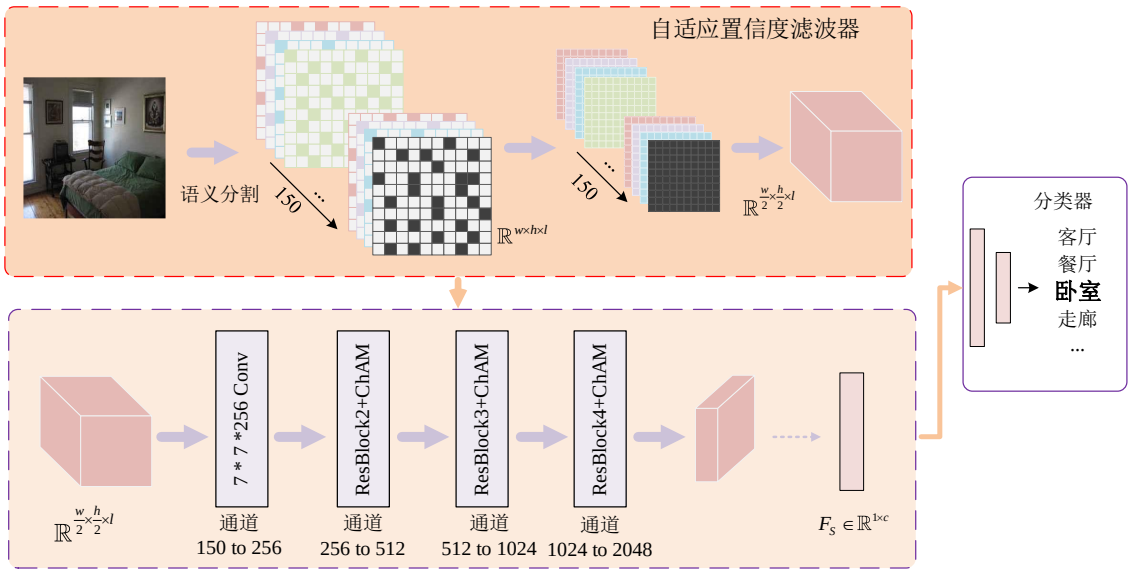


图 2-2 语义空间关系模块（SSRM）

Fig. 2-2 Semantic Spatial Relation Module (SSRM)

首先，为了在训练过程中避免空间无关信息的消极影响，将给定图像 $I \in \mathbb{R}^{w\times h\times 3}$ 输入至一个语义分割网络，以获得语义分割得分张量（Semantic segmentation score tensor）$M \in \mathbb{R}^{w\times h\times l}$（其中 $l=150$）[48,49]。张量 M 仅包含纯净的空间信息，排除了非必要的信息

干扰。然后，将 M 作为网络的输入，确保此模块仅处理与空间相关的信息，从而深入挖掘场景中的空间上下文关系。SSRM 将生成一个空间上下文特征 $F_S \in \mathbb{R}^{1\times c}$（$c=2048$），它是给定图像 I 中空间上下文信息的集中体现。

2.2.1 自适应置信度滤波器

由于语义分割网络的精度限制，分割结果出现误差在所难免。现有的方法[32,42]试图通过固定的置信度阈值过滤歧义点。然而，该策略泛化性不足，尤其是在处理大型数据集时效果有限。为了缓解语义歧义带来的影响，本节设计了一个自适应置信度过滤器（ACF），基于每个图像的具体状态，自适应地过滤其语义分割得分张量。

具体来讲，ACF 利用一个 2×2 的滤波器对 M 的各个通道（Channel）进行平滑处理。在应用滤波器的每个区域，ACF 只选取该区域内置信度最高的像素。对 M 的所有通道进行处理后，生成了一个新的张量 $M'\in\mathbb{R}^{\frac{w}{2}\times\frac{h}{2}\times l}$。图 2-3 是 ACF 应用于 M 的一个示例。M 对应的语义分割标签图，其每个内部点代表了 M 中在 $1\times1\times l$ 的范围内置信度最高的通道位置。而在 M' 对应的语义分割标签图中，每个内部点则反应了 M 中 $2\times2\times l$ 范围内置信度最高的通道位置。通过利用范围内的自适应筛选而非固定的置信度阈值来过滤语义分割得分张量，ACF 能够根据每张图像的具体特性提高语义分割结果的精度。此外，ACF 还有助于减少后续网络处理的输入尺寸，从而降低计算成本。本章将在第 2.5 小节中详细比较和分析计算成本。

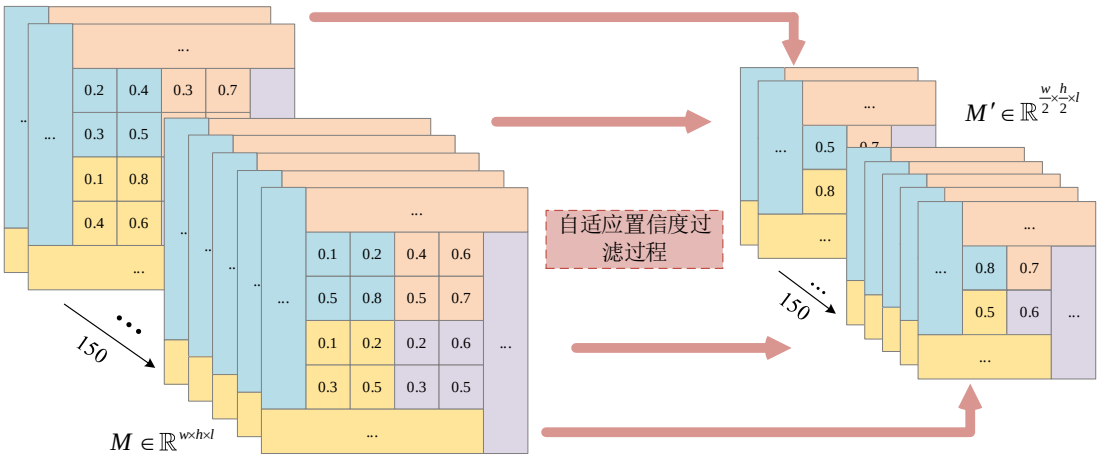


图 2-3 自适应置信度滤波器的实施示例

Fig. 2-3 An illustration of Adaptive Confidence Filter

2.2.2 空间关系建模过程

接下来，利用残差块（ResBlocks）处理过滤后的语义分割得分张量 M'，以探索物体区域之间的空间关系。ResBlocks 源自原始的 ResNet50 架构[22]，包括几个关键的残差基本块（即 *ResBlock* 2、*ResBlock* 3 和 *ResBlock* 4）。考虑到 $M'_{(i,j)}$ 中的各通道值表示图

像 I 中对应位置的预测语义概率[48,49]，本章追随 SAS-Net[33]的做法，在几个残差块内部引入了通道注意力模块（Channel Attention Module，ChAM)[50]，以帮助网络更好地聚焦于关键物体。图 2-4 展示了 ChAM 作用于 ResBlock 的示意图。更具体地说，给定特征图 $F \in \mathbb{R}^{w^0 \times h^0 \times l^0}$，ChAM 首先采用平均池化和最大池化操作聚合 F 的空间信息，得到特征向量 F_{avg} 和 F_{max}。F_{avg} 和 F_{max} 经过一个共享权重的多层感知机（Multi-Layer Perceptron，MLP）处理后并逐元素相加。最后，经过 *Sigmoid* 激活函数处理，生成通道注意力图 $M_c(F) \in \mathbb{R}^{1\times 1\times l^0}$：

$$M_c(F) = \sigma\left(MLP\left(F_{avg}\right) + MLP\left(F_{max}\right)\right) = \sigma\left(W_1\left(W_0 Avg\left(F\right)\right) + W_1\left(W_0 Max\left(F\right)\right)\right) \tag{2-1}$$

其中，σ 表示 *Sigmoid* 激活函数，$W_0 \in \mathbb{R}^{(l^0/r)\times l^0}$ 和 $W_1 \in \mathbb{R}^{l^0\times(l^0/r)}$ 是 MLP 的权重，每个 W_0 之后都有一个 *ReLU* 激活函数，$r = 16$ 是缩减率[50]。随后，$M_c(F)$ 被用于加权特征图 F：

$$F' = F \oplus M_c(F) \odot F \tag{2-2}$$

其中 $\odot$ 表示元素级乘积（Hadamard product），$\oplus$ 表示元素级相加。

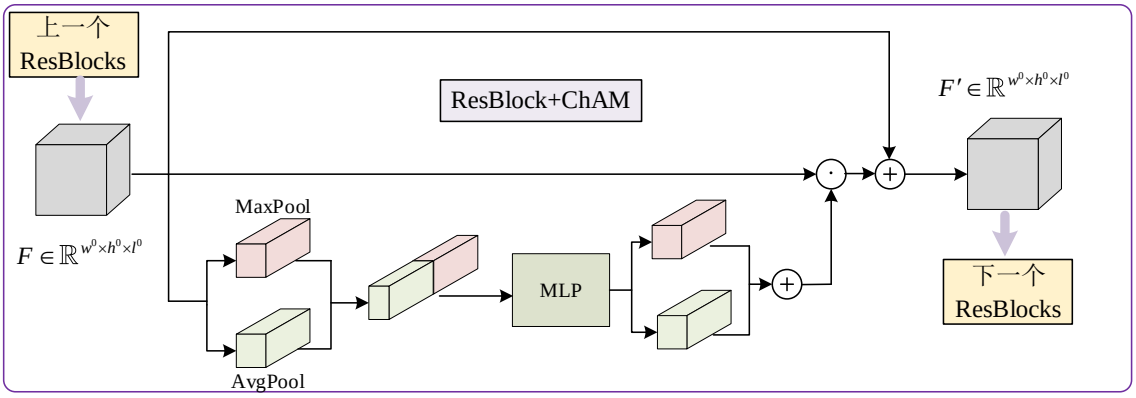


图 2-4 ResBlock 与通道注意力模块结合的示意图

Fig. 2-4 Illustration of ResBlock in combination with the Channel Attention Module

通过上述对 M' 中物体区域之间的空间关系的探索，可以获得基于语义的空间特征 F_S，它包含丰富的空间上下文信息。

2.3 面向场景识别的全局特征提取

鉴于输入到 SSRM 的语义分割得分张量未包含输入图像的颜色和纹理等信息，为了充分利用图像蕴含的信息，本章将 SSRM 和 PlacesCNN[19]相结合，构建了组合语义空间关系模块（Combined Semantic Spatial Relation Module，CSSRM），图 2-5 展示了 CSSRM 的整体框图。

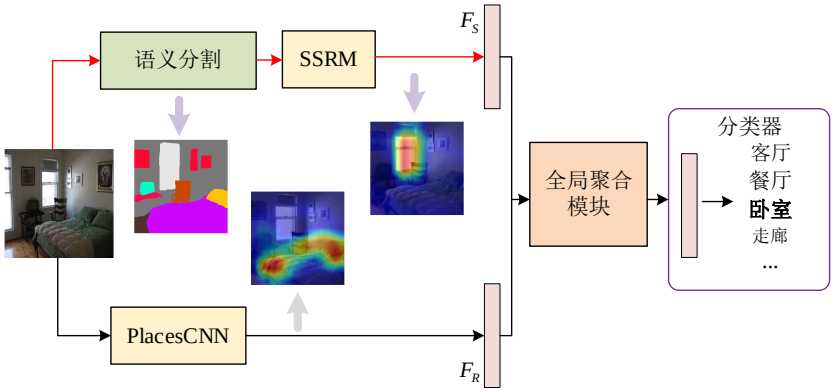


图 2-5 组合语义空间关系模块

Fig. 2-5 Combined Semantic Spatial Relation Module

2.3.1 PlacesCNN 模块

本模块的骨干网络为以 ResNet50（去除了池化层）[22]为基础的 PlacesCNN[19]模型。本模块的输入是给定的原始红绿蓝（Red Green Blue，RGB）图像 $I \in \mathbb{R}^{w \times h \times 3}$，输出是一个高级深度特征 $F_R \in \mathbb{R}^{1 \times c}$。为确保与其他现有算法的公平比较，当在 MIT-67 数据集[17]上对本章算法进行评估时，本模块在 Places365 数据集[19]上进行了权重预训练；而在 Places_7 和 Places_14 数据集[19]上对本章算法进行评估时，本模块在 ImageNet 数据集[18]上进行了权重预训练。

2.3.2 基于条带深度卷积的全局聚合模块

前面的处理过程成功提取了空间特征 $F_S \in \mathbb{R}^{1 \times c}$ 和深度特征 $F_R \in \mathbb{R}^{1 \times c}$。接下来，如图 2-6 所示，构建了一个全局聚合模块，深入挖掘 F_S 和 F_R 之间的互补信息。挖掘 F_S 和 F_R 之间的互补信息是进行场景识别的关键，受 MobileNet[51]和 OTS-Net[52]的启发，本节使用了一种从通道（Channel）维度聚合两个特征的策略，即条带深度卷积（Strip Depth-Wise Convolution）。2.5 小节将把此架构与其他聚合架构进行比较，以彰显其优越性。下面对本模块进行详细介绍。

首先，将 F_R 和 F_S 的各个通道进行级联，得到 $F' \in \mathbb{R}^{2 \times c}$。随后，通过一层 MLP 对其进行处理（使用了残差结构以防止过拟合），输出 $F'' \in \mathbb{R}^{2 \times c}$：

$$F'' = F' + \phi\left(W_n\left(Drop\left(\phi\left(W_m F' + b_m\right)\right)\right) + b_n\right) \tag{2-3}$$

其中 ϕ 表示 *GeLu* 激活函数，$W_m \in \mathbb{R}^{l \times c}$、$b_m \in \mathbb{R}^{l}$、$W_n \in \mathbb{R}^{l \times c}$ 和 $b_n \in \mathbb{R}^{c}$ 是 MLP 内部两个全连接层的权重（Weight）和偏置（Bias），$l = 8192$。*Drop* 表示比例为 0.1 的随机失活函数（*Dropout*）。

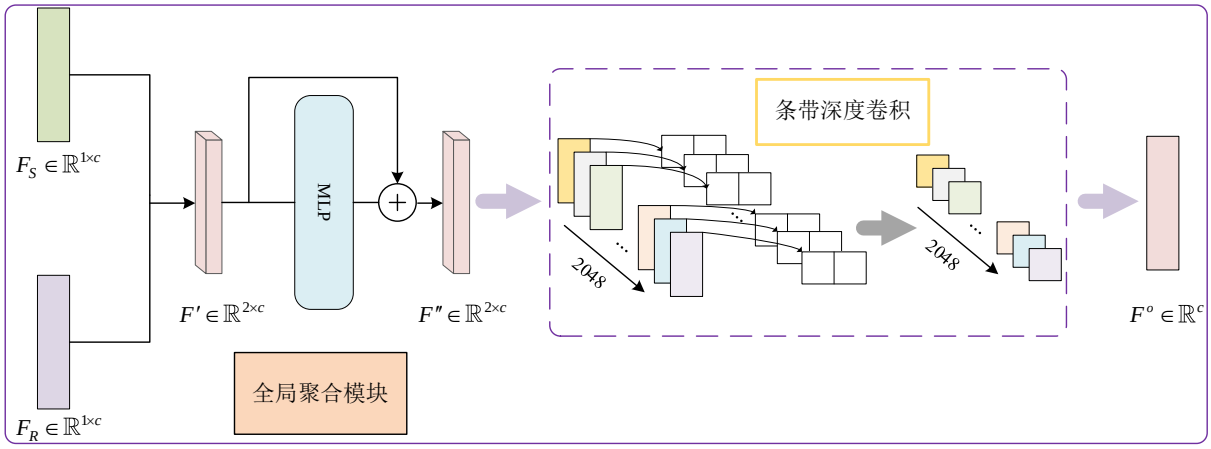


图 2-6 全局聚合模块

Fig. 2-6 Global Integration Module

随后，通过深度（Depth-Wise，DW）卷积对 F'' 的各个通道进行卷积处理，对 F_R 和 F_S 实现通道维度的精细结合，以深入挖掘二者之间的互补信息。最终得到的全局聚合特征 $F^o \in \mathbb{R}^c$ 的每个通道值可以被表示为：

$$F_m^o = \sum_{i=1}^{2} K_{i,m} \cdot F''_{i,m} \tag{2-4}$$

其中 K 表示大小为 $2\times1\times M$ 的深度卷积核。K 中的第 m_{th} 个卷积核被应用于 F'' 的第 m_{th} 个通道，以生成 F^o 的第 m_{th} 个通道值，$M=2048$。

最后，全局聚合特征 F^o 将被送入一个全连接层以进行最终的场景预测。

2.4 场景识别常用数据集介绍

为了促进场景识别领域的发展，业内学者们陆续发布了多个大规模的场景识别数据集，本节对三个广泛使用的数据集进行简要介绍。

2.4.1 MIT-67 数据集

MIT-67 数据集[17]由来自 67 个不同室内场景类别的 15620 张图片组成，其中每张图片在最短边上的分辨率不低于 200 像素。这些场景类别可被分为居家环境、商店、办公区域、休闲场所和公开场合 5 大部分。由于室内场景多变的空间布局引发的类内差异，以及跨场景共存的常见物体导致的类间相似性，MIT-67 数据集颇具挑战性。该数据集适用于以场景理解和环境感知为核心的视觉认知任务，目前由麻省理工学院持续维护。每个场景类别包含至少 100 张图片，官方的评估方案[17]指定 80 张图片用于训练，20 张图片用于测试。数据集示例如图 2-7 所示。

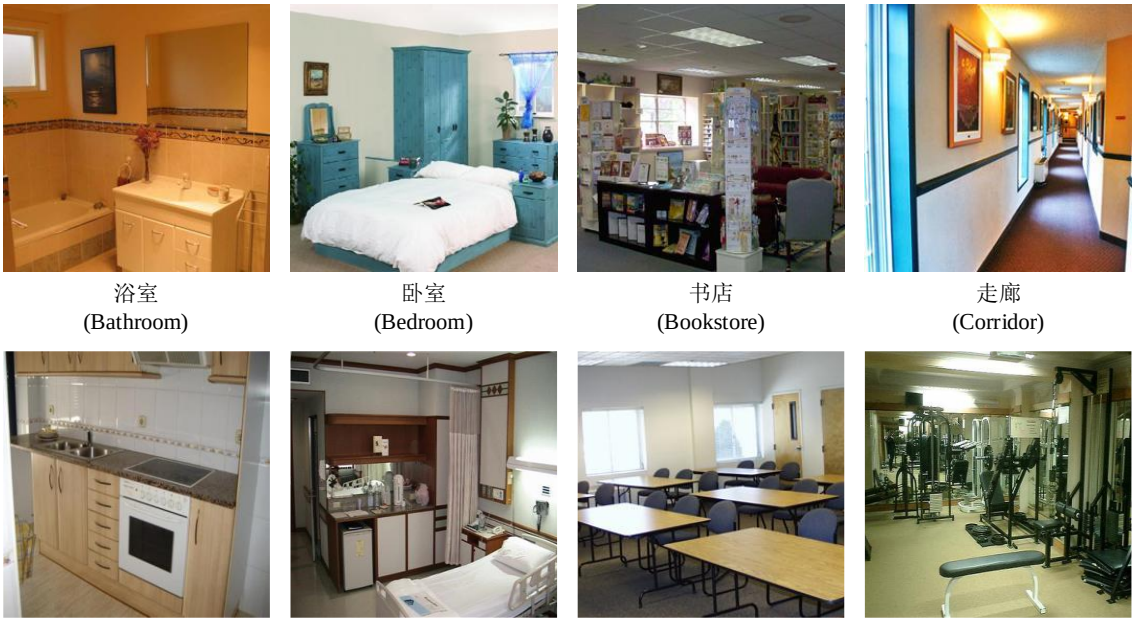


图 2-7 MIT-67 数据集的部分场景类别

Fig. 2-7 Subset of scene categories of MIT-67 dataset

2.4.2 SUN397 数据集

SUN397 数据集[53]是一个覆盖室内外场景的庞大数据集，汇集了来自 397 个不同场景类别的共计 1108754 张图片，每张图片的分辨率大约为 500×300 像素。该数据集包括 175 个室内场景类别、220 个室外场景类别，以及 2 个融合了室内外元素的场景类别。遵循官方的评估体系[53]，每个类别分配 50 张图片进行训练，另外 50 张图片用于测试。数据集示例如图 2-8 所示。

2.4.3 Places 数据集

Places 数据集[19]是目前最大的以场景为中心的数据集之一，涵盖了大约 180 万张训练图像和 365 种不同的室内外场景类别，是迄今为止面向场景分类领域中最具挑战性的数据集。每个场景类别包含了 5000 张训练图像及 100 张评估图像。为了与其他室内场景识别方法[45,52]公平比较，本文采用了该数据集的两个简化版本：Places_7 与 Places_14。其中，Places_7 聚焦于 7 种室内场景类别，包括浴室、卧室、走廊、餐厅、厨房、客厅和办公室，数据集示例如图 2-9 所示。而 Places_14 则包括 14 种室内场景类别，涵盖阳台、卧室、餐厅、家庭办公室、厨房、客厅、楼梯、浴室、壁橱、车库、家庭影院、自助洗衣店、游戏室以及酒吧，数据集示例如图 2-10 所示。



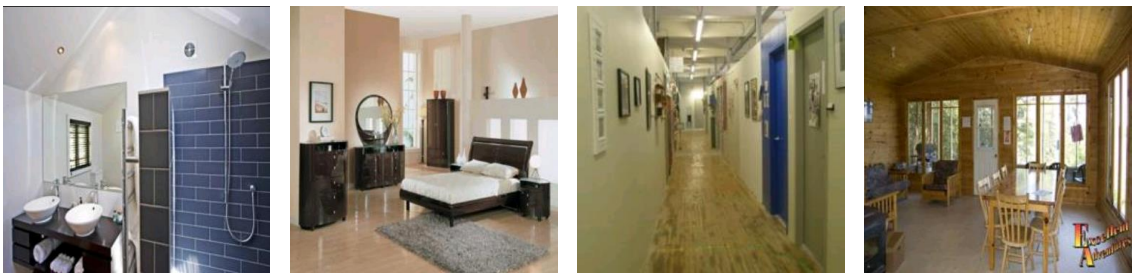
飞机机舱 (Airplane cabin)　面包店 (Bakery shop)　浴室 (Bathroom)　沙滩 (Beach)



图书馆 (Library)　办公室 (Office)　玩具店 (Toyshop)　山丘 (Hill)

图 2-8 SUN397 数据集的部分场景类别

Fig. 2-8 Subset of scene categories of SUN397 dataset



浴室 (Bathroom)　卧室 (Bedroom)　走廊 (Corridor)　餐厅 (Dining room)



厨房 (Kitchen)　客厅 (Living room)　办公室 (Office)

图 2-9 Places_7 数据集的场景类别

Fig. 2-9 Scene categories of Places_7 dataset



图 2-10 Places_14 数据集的部分场景类别

Fig. 2-10 Subset of scene categories of Places_14 dataset

2.5 实验结果与分析

本节对提出的基于语义区域关系建模的室内场景识别框架在 MIT-67、SUN397（取其 175 个室内类别）、Places_7 和 Places_14 这 4 个公开数据集上进行了评估。下面首先介绍实验的基本设置，然后通过多项消融实验验证各个组件的有效性，最后将所提算法与当前先进的算法进行比较。

2.5.1 实验设置

本章针对室内场景识别，使用 MIT-67、SUN397（取其 175 个室内类别）、Places_7 和 Places_14 数据集验证所提算法。

（1）语义分割

在本文所有的实验中，语义分割（Semantic segmentation）过程均由在 ADE20K 数据集[49]上预训练的 Vision Tranformer Adapter[48]进行。对于给定的图像 $I \in \mathbb{R}^{w \times h \times 3}$，语义分割的结果是一个得分张量（Score tensor）$M \in \mathbb{R}^{w \times h \times l}$，该张量可以用来生成图像 I 的语义分割标签图（Label map）$L \in \mathbb{R}^{w \times h}$。图像 I 中位置 (i, j) 的语义预测概率被表示为 $M_{i,j} \in \mathbb{R}^{1 \times 1 \times l}$，其中 $l = 150$ 是语义标签的类别总数。标签图 L 中位置 (i, j) 处的值 L_{ij} 表示 I 中对应位置的语义标签。

（2）训练细节和超参数

为深入探索输入场景中的空间信息，在训练语义空间关系模块（SSRM）时，需要消除空间无关信息的影响。因此，本节实施了一个两阶段（Two-stage）的训练过程。在第一阶段，SSRM 和 PlacesCNN 模块被分别独立训练。在第二阶段，则固定了 SSRM 和 PlacesCNN 模块的权重，仅对后续模块进行训练。为简化训练流程，采用了 ALI-G 优化器[54]优化网络参数。该优化器仅需设定一个初始学习率超参数，在本节的所有实验中，将初始学习率统一设置为 0.1。在训练的第二阶段，为了避免过拟合，在最终的分类器中引入了随机失活正则化技术（*Dropout*），设定失活概率为 0.8。

本文的方法是在 PyTorch[55]和 SAS-Net[33]的开源框架下实现。在最终评估所提方法的性能时，采用了标准的 10-crop 评估策略[33,56]，以保证与其他方法的公平比较。

2.5.2 语义空间关系模块架构的消融实验

本小节在 MIT-67 和 SUN397 数据集上对不同的语义空间关系模块（SSRM）的架构进行了对比评估，研究了自适应置信度滤波器（ACF）和通道注意力模块（ChAM）对识别精度的影响。表 2-1 展示了实验结果。

如表 2-1 所示，利用 ACF 对语义分割得分张量进行过滤处理，不仅显著提升了识别性能，还大幅降低了计算成本。与直接使用原始得分张量相比，将 ACF 整合至 SSRM 后，在 MIT-67 数据集上的识别准确率提高了 3.43%至 5.75%，在 SUN397 数据集上也实现了 2.365%至 3.705%的准确率增长，同时，浮点运算数（Floating Point Operations，FLOPs）减少了 17.89G。此外，ChAM 的引入进一步提高了 SSRM 的识别性能，而这仅增加了微小的计算成本。

表 2-1 SSRM 的不同架构的消融结果

Tab. 2-1 Ablation results for different architectures for the SSRM

模型架构	预训练	MIT-67 (%)	SUN397 (%)	FLOPs (G)
ResNet50	无	64.403	50.529	27.2
4×4 ACF + ResNet50	无	70.149	52.894	9.31
2×2 ACF + ResNet50	无	69.627	53.035	9.31
ResNet50_ChAM	无	69.851	54.624	27.3
4×4 ACF + ResNet50_ChAM	无	73.284	57.4	9.32
2×2 ACF + ResNet50_ChAM	无	**74.403**	**58.329**	**9.32**
2×2 ACF +ResNet50_ChAM	Places365	81.642	66.953	9.32

接下来，本节探讨了使用不同过滤尺寸的 ACF 对 SSRM 识别性能的影响。结果显示，相较于应用 4×4 ACF，采用 2×2 ACF 可令 SSRM 在 MIT-67 和 SUN397 数据集上的

识别准确率分别增加 1.12%和 0.929%。这一现象表明，尽管 ACF 能有效缓解由于语义歧义引发的负面影响，但也可能导致输入张量中部分物体信息的丢失。因此，在选取 ACF 的过滤域尺寸时，应兼顾物体信息保留与语义歧义缓解之间的平衡。值得注意的是，尽管 4×4 ACF 可能会导致更多物体信息的丢失，其仍旧在 MIT-67 和 SUN397 数据集上分别使 SSRM 的识别准确率显著增长了 3.43%和 2.776%，进一步凸显了 ACF 在 SSRM 架构中的关键作用。

根据上述实验结果，在最终选定的 SSRM 架构中，首先使用 2×2 大小的 ACF 处理语义分割得分张量，然后利用集成了 ChAM 的 ResNet50 架构进行后续处理。同时，如表 2-1 最后一行所示，本节还在 Places365 数据集上选定的模型架构进行了预训练，从而在下游数据集上达到了更高的识别精度。

2.5.3 物体信息数量的影响

为了探索物体信息对 SSRM 的影响，本节进行了一系列消融实验，在 MIT-67 和 Places_14 两个数据集上探讨了物体信息数量变化对场景识别性能的影响。图 2-11 展示了实验结果，其中横轴代表 SSRM 所利用的场景内物体信息类别的数量，范围区间设定为 70 至 150，间隔为 10。

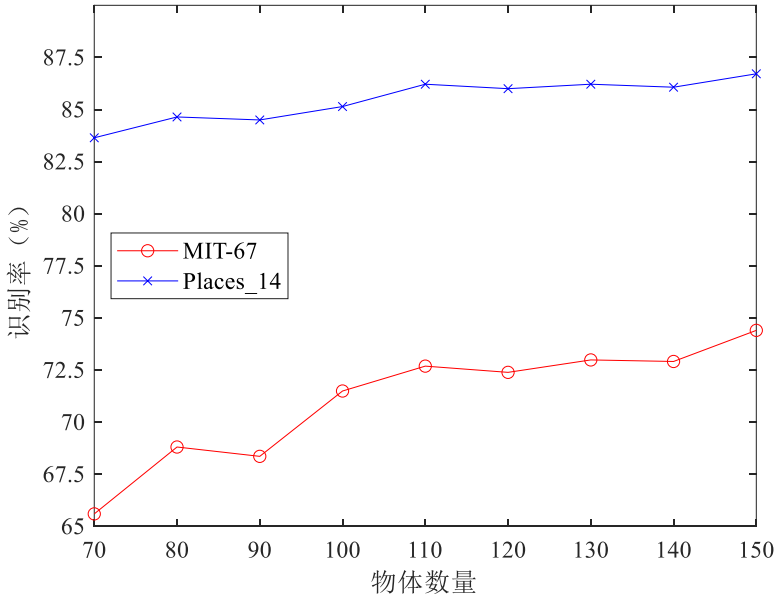


图 2-11 物体信息数量的消融研究

Fig. 2-11 Albation study on the number of object information

注：按照 ADE20K 数据集中的物体类别列表为顺序，选取了对应数量的物体信息。

从图 2-11 可以明显看出，场景识别的准确率与所使用的物体信息数量之间存在显著的正相关性。随着物体信息类别数量的增加，识别精度大大提高。例如，在 MIT-67 数

据集上，使用 150 个语义类别相较于使用 70 个语义类别，识别准确率提高了 8.81%；同样地，在 Places_14 数据集上，准确率也提高了 3.073%。此外，值得注意的是，物体信息数量的增加并不总是导致识别精度的线性提升。在某些情况下，识别精度的提升幅度较小，甚至出现轻微下降。例如，当物体类别数量从 80 增至 90 时，在 Places_14 数据集上的识别精度反而降低了 0.143%，而在 MIT-67 数据集上，准确率也低于前者 0.448%。这一现象是因为此时新增的物体类别为汽车、卡车、路灯、塔等与室内场景关联度较低的物体类别，因此，物体语义类别与场景类别的相关性对 SSRM 的有效性也十分关键。

2.5.4 不同的特征聚合策略的对比

为验证所采用的全局特征聚合模块架构的优越性，本小节在 MIT-67 数据集上进行了实验，对比了多个聚合 RGB 深度特征 F_R 和语义空间特征 F_S 的方法，并将其与两个基准模型（PlacesCNN 和 SSRM）进行了比较，实验结果如表 2-2 所示。为便于理解，图 2-12 简要展示了上述两种聚合策略。其中，特征级联（Concatenation）策略曾在 BORM[45] 中被采用以达到类似目的，而语义门控（Semantic gating）策略则在 SAS-Net[33]中有类似应用。

表 2-2 对不同的全局聚合模块架构的消融实验

Tab. 2-2 Ablation results for different architectures of Global Aggregation Module

结合方式	精度（%）	FLOPs (k)
PlacesCNN	84.776	-
SSRM	81.642	-
条带卷积	88.731	143.4
特征级联	88.582	274.4
语义门控	85.0	137.2

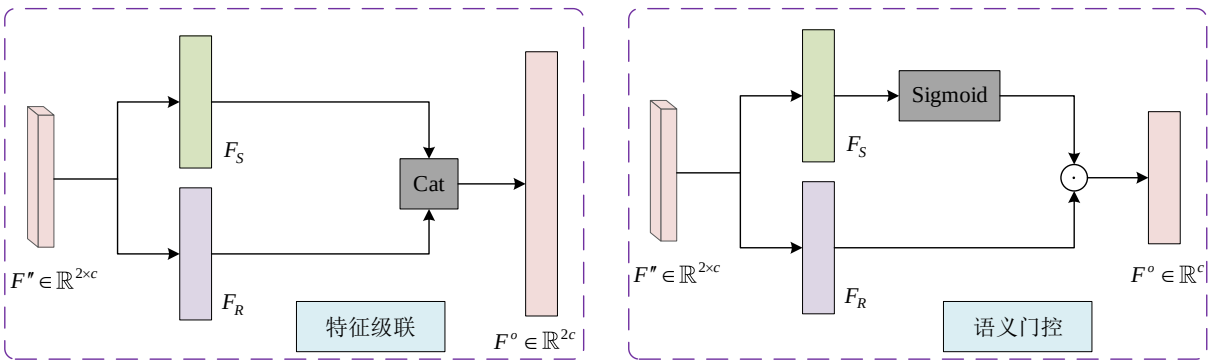


图 2-12 特征级联和语义门控的图例

Fig. 2-12 Illustration of feature concatenation and semantic gating

表 2-2 的结果显示，无论采用哪种聚合策略，都能获得比两个基准模型更好的性能。这强调了在场景识别任务中深入挖掘语义空间特征与 RGB 深度特征之间互补信息的关键意义。进一步地，相较于特征级联策略，条带卷积不仅减少了计算资源（FLOPs）消耗，还在准确性上有所提高，这凸显了条带卷积在挖掘特征互补信息方面的独特优势。意外的是，先前在 SAS-Net[33]中表现优异的语义门控机制，在本实验中的性能却是最低的，仅比 PlacesCNN 高出 0.224%。这可能是因为 *Sigmoid* 激活函数改变了语义空间特征的尺度，从而对其与 RGB 深度特征的融合产生了负面影响，这说明语义门控可能不适用于本章的网络架构。经过综合比较，条带卷积因其最佳的性能表现被选为特征 F_R 和 F_S 的聚合策略。

2.5.5 与现有算法的性能比较

本节在 MIT-67、SUN397、Places_7 和 Places_14 四个室内场景数据集上将所提出的方法与现有的先进方法进行了比较。除非另有说明，所有比较方法的结果均直接引用自它们各自的原始论文。

（1）基于 SSRM 的比较

在这一部分，本节将 SSRM 与其他仅依赖语义分割或物体检测结果获取场景表示的方法进行了对比，结果如表 2-3 所示。为了确保公平比较，在 MIT-67 上进行验证时，先在 Places365 数据集上对 SSRM 进行了预训练；而在 Places_7 和 Places_14 数据集上，则从头开始训练 SSRM。由于其他相关研究并未在 SUN397 数据集上进行验证，故本节未展示 SSRM 在 SUN397 数据集上的表现。从实验结果可以明显看出，SSRM 展现出了显著的性能优势。这进一步验证了在室内场景识别任务中，对语义空间关系进行建模的重要性和有效性。

表 2-3 SSRM 与现有先进算法的比较

Tab. 2-3 Comparison of SSRM with the state-of-the-art methods

算法	Places_14 (%)	Places_7 (%)	MIT-67 (%)
Deduce (Φ_{obj})[39]	47.00	62.60	-
BORM[45]	74.90	83.10	-
OTS-Net[52]	85.90	90.10	-
SAS-Net (SemBranch)[33]	-	-	73.43
SSRM	**86.71**	**93.14**	**81.64**

（2）基于 CSSRM 的比较

如表 2-4 和表 2-5 所示，本小节将本章所提出的完整模型（CSSRM）与现有的先进场景识别方法进行了对比。可以明显看出，CSSRM 在有效性和泛化性能方面均优于大

多数现有的室内场景识别算法。相比于依靠场景中物体的语义信息辅助场景识别的其他方法[32,33,39,42,45,52]，CSSRM 展现出了更优越的识别性能，这不仅验证了本章对语义空间信息的探索极为有效，也强调了深入研究 RGB 深度特征表示与语义空间特征表示之间互补信息的重要性。此外，CSSRM 还优于当前通过多分支策略获取场景多尺度信息的方法[32,33,38,45,58,59]，进一步表明了将场景中的物体语义信息作为一个额外信息源并提取其高级表示的有效性。总的来讲，这些实验结果验证了所提出的方法在室内场景识别领域的优越性和广泛适用性。

表 2-4 在 MIT-67 和 SUN397 数据集上 CSSRM 与现有先进算法的比较

Tab. 2-4 Comparison of CSSRM with the state-of-the-art methods on MIT-67 and SUN397

算法	MIT-67 (%)	SUN397 (%)
Dual CNN-DL[26]	76.56	70.13
NNSD+ICLC[37]	84.30	64.78
SDO-Net[32]	86.76	73.41
SAS-Net[33]	87.10	74.04
CCF-Net[38]	87.30	-
DeepScene-Net[57]	71.00	-
ARG-Net[42]	88.13	75.02
PL + AP + AI + IM[58]	88.06	74.12
LGN[28]	88.06	73.85
MR-Net[59]	88.08	73.98
MASR[60]	88.50	75.01
CSSRM	**88.73**	**75.48**

表 2-5 在 Places_14 和 Places_7 数据集上 CSSRM 与现有先进算法的比较

Tab. 2-5 Comparison of CSSRM with the state-of-the-art methods on Places_14 and Places_7

算法	Places_14 (%)	Places_7 (%)
Word2Vec[61]	83.70	-
Deduce[39]	-	88.10
BORM[45]	85.80	90.10
OTS-Net[52]	85.90	90.10
AGCN[62]	86.00	91.70
CSSRM	**88.71**	**93.43**

2.6 本章小结

本章构建了一个语义空间关系模块（SSRM），旨在深入挖掘室内场景中复杂多变的空间上下文信息。首先，在语义分割技术的帮助下将场景中的关键空间信息解耦出来，避免空间无关信息的影响，确保 SSRM 能够专注于对空间关系的探索。然后，设计了一个自适应置信度滤波器（ACF），有效缓解了语义分割结果中的歧义所带来的负面影响，为处理语义分割的不确定性提供了一种切实可行的解决方案。在上述基础上，构建了一个专门针对语义分割结果（即语义得分张量）的网络架构，该架构整合了 ResNet 和通道注意力机制，以确保 SSRM 能够全面理解和表达场景中的空间布局。此外，考虑到 SSRM 的输入过滤了颜色和纹理等空间无关的信息，将 SSRM 与以 RGB 图像为输入的 PlacesCNN 进行整合，并基于条带深度卷积设计了一个全局特征聚合模块，深入挖掘和利用两个模块间的互补特征，进而优化最终的场景表示。在多个公开场景数据集上进行的实验验证了所提方法的有效性。与现有的场景识别方法相比，本章方法展现出了更优异的性能。

尽管本章方法通过建模物体间的空间上下文信息提升了场景识别精度，但未充分考虑跨场景共存的鉴别性物体可能导致的类间相似问题。因此，下一章将致力于研究解决此问题的方法，以期更有效地利用物体信息辅助场景识别。

3 融合语义空间关系和物体共现性建模的室内场景识别方法

3.1 引言

本文第二章探讨了通过神经网络深入挖掘场景中的空间上下文信息以辅助场景识别的可行性。除了空间信息，场景中的物体共现性对于场景识别同样具有重要意义。已有一些研究[32,41,45]致力于建模物体共现性以实现场景识别。这些方法通过计算不同场景中物体的概率分布，推断出与特定场景类别紧密相关的鉴别性物体，取得了显著的效果。然而，如图 3-1 所示，一些场景中的鉴别性物体可能是相同的（例如，“病房”和“卧室”场景中的“床”，“教室”和“餐厅”场景中的“桌子”），这对基于物体共现性建模的方法提出了巨大挑战。



图 3-1 一些室内场景的示例

Fig. 3-1 Some examples of indoor scenes

本章提出了一个新颖的思路应对上述挑战，灵感来源于人类视角：如图 3-1 所示，尽管“卧室”和“病房”中的鉴别性物体相同，人们仍能轻松区分这两个场景。这是因为同一物体在不同场景类别中往往呈现出不同的特性（例如，“卧室”和“病房”中“床”的明显差异）。基于这一观察，本章提出在探索不同场景中物体的共现性之前，先将给定场景的相关特征分配到这些物体。这样一来，就能使网络像人类一样区分具有相同语

义的物体。然而，随之而来的问题是，由于同一物体在不同的场景中具有了不同的特征，在本研究中手工统计方法[32,45]不再适用于探索物体共现性。幸运的是，Transformer[63]的出现提供了一种灵活的注意力机制架构，能够捕获序列特征之间的长期依赖关系。一些基于 Transformer 的方法[64-67]通过注意力机制探索局部特征之间的全局交互作用以获得鉴别性特征。受此启发，本章构建了一个全局-局部依赖编码器，以端到端的方式建模物体特征之间的长期共现性。

基于上述思路，本章在第二章构建的语义空间关系模块（SSRM）的基础上，进一步提出了 SpaCoNet，旨在通过同时建模场景中的空间关系（**Spa**tial relation）和物体共现性（**Co**-occurrence of objects）进行室内场景识别。

3.2 融合语义空间关系和物体共现性的场景识别框架

3.2.1 总体架构

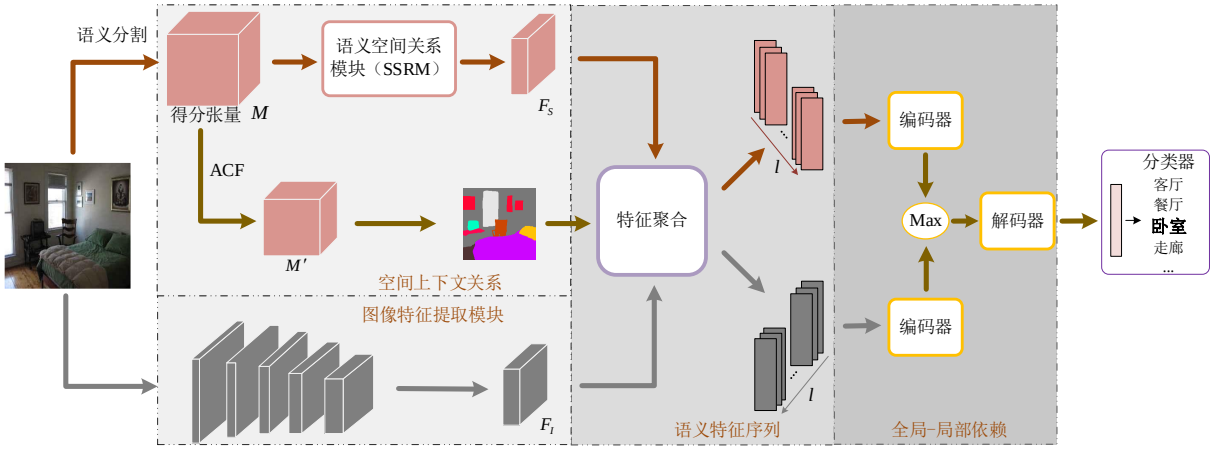


图 3-2 SpaCoNet 的整体架构

Fig. 3-2 Pipeline of the proposed SpaCoNet

图 3-2 展示了所提出的 SpaCoNet 的整体流程。该框架包括五个主要模块：（1）语义空间关系模块，（2）图像特征提取模块，（3）语义节点特征聚合模块，（4）全局-局部依赖模块，以及（5）场景分类器。语义空间关系模块负责生成描述场景中各物体区域间空间关系的特征图；图像特征提取模块则负责提供来自 PlacesCNN 的特征图，PlacesCNN 在大型数据集 Places365[19]上进行了预训练，以获得 RGB 图像的高级表示；这两个模块生成的特征图，连同输入图像的语义分割标签图，被送入语义节点特征聚合模块。该模块依据语义标签图，分别对两个特征图进行特征聚合，生成两组语义特征序列，并将这些序列传递给全局-局部依赖模块。然后，全局-局部依赖模块通过注意力机制挖掘这些语义特征序列内部的长期共现关系，并利用这些共现信息优化全局特征。最

后，经过精细优化的全局特征被送入场景分类器，进行场景类别预测。

3.2.2 语义空间关系模块

鉴于语义空间关系模块（SSRM）已经在第二章得到了详细介绍，本节不再对其进行冗余描述。与第二章的操作不同的是，本章移除了 SSRM 中的全局平均池化层，旨在保留像素级的细节信息。这样一来，通过 SSRM 得到了基于语义的空间特征 $F_S \in \mathbb{R}^{\frac{w}{32}\times\frac{h}{32}\times c}$，为后续语义节点特征聚合模块的处理提供了便利。

3.2.3 图像特征提取模块

除空间关系外，场景中的物体共现也是识别场景的重要因素。虽然空间关系本身也隐含了一定程度的多物体共现信息，但其主要关注点在于探究物体间的位置关系，而不涉及物体自身的属性（如颜色）。这也是本章使用解耦的语义分割得分张量 M 作为 SSRM 的输入探索空间关系的原因。然而，由于 M 不包含物体本身的特征信息，SSRM 缺乏区分可能出现在不同场景中的鉴别性物体的能力，导致对物体共现性的探索不够充分。因此，构建了一个针对图像 I 的图像特征提取模块（Image Feature Extraction Module，IFEM），以提取完整的图像深度特征 F_I。通过结合 F_I 和 F_S，可以将场景相关特征与后续模块中的物体联系起来，进而深入挖掘场景中物体的共现性。

具体来说，与第二章的操作类似，为充分利用图像中包含的信息，将 SSRM 与作为图像特征提取模块的 PlacesCNN 相结合。PlacesCNN 由在 Places 数据集[19]上预训练的几乎完整的 ResNet50 架构组成。需要注意的是，PlacesCNN 排除了 ResNet50 中的池化层，原因是池化层会引入平移不变性（Translation invariance），这种不变性会导致在顶层特征层中模糊不同节点特征之间的差异[68]。这种模糊阻碍了后续模块对全局-局部依赖关系的探索。

给定图像 $I \in \mathbb{R}^{w\times h\times 3}$，IFEM 为其生成一个高级深度特征 $F_I \in \mathbb{R}^{\frac{w}{16}\times\frac{h}{16}\times c}$。然后，特征 F_S 和 F_I 被送入语义节点特征聚合模块中，实现将场景相关特征与物体语义的结合。注意，为了便于后续处理，采用双线性插值算法对 F_S 进行了插值处理，使其分辨率与 F_I 保持一致。

3.2.4 语义节点特征聚合模块

为了区分不同场景中的相同物体，本节设计了语义节点特征聚合模块，该模块将输入场景的相关特征分配到场景内的各个物体上。通过依次将空间特征 F_S 和深度特征 F_I 输入到本模块，可以得到这两个特征各自对应的语义特征序列。

对于每幅图像，首先从 SSRM 中获取其语义分割得分张量 M，进而根据 M 中的概率关系生成语义分割标签图 L。标签图 L 能够确保本模块从 F_S 和 F_I 中提取出的每个像素级特征匹配到对应的物体上。然而，这个过程面临两个问题：

1）语义歧义：由于语义分割技术的精度有限，标签图 L 内可能存在歧义点，这可能对后续工作产生不利影响。

2）分辨率不匹配：标签图 L 的分辨率与输入图像的分辨率一致，但与特征图的分辨率不一致。

本节采用了下面介绍的特征聚合方法解决上述问题，本模块的结构框架如图 3-3 所示。

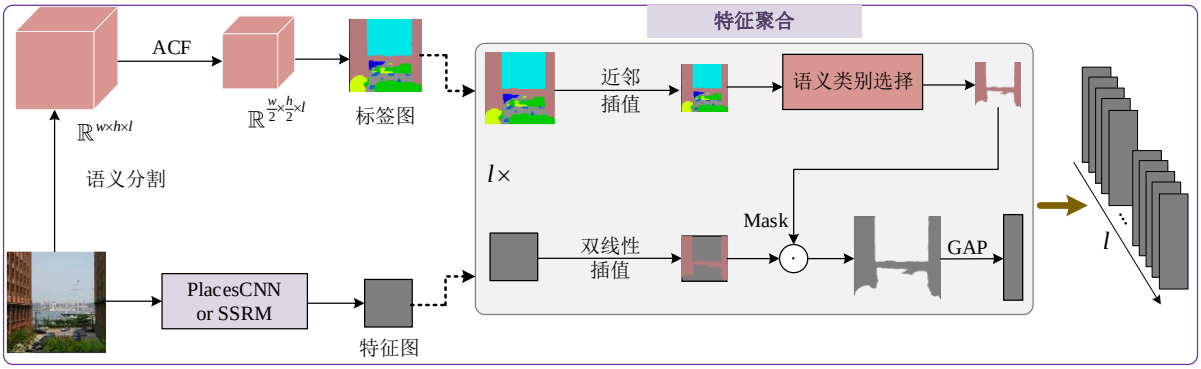


图 3-3 语义节点特征聚合模块

Fig. 3-3 Semantic Node Feature Aggregation Module

对于语义歧义问题，首先在得分张量 M 上应用第二章提出的自适应置信度滤波器（ACF），得到 M'。然后，利用精炼后的张量 M' 生成标签图 $L'\in\mathbb{R}^{\frac{w}{2}\times\frac{h}{2}}$。这样一来，分割结果的可靠程度就得到了提高。

对于分辨率不匹配问题，在得到 L' 后，利用最近邻插值算法降低其分辨率，使其与特征图 F_I 和 F_S 一致。随后，为 l 个物体类别中的每个物体语义 o 生成相应的二值掩膜 L^o：

$$L^o_{i,j}=\begin{cases}1 & L'_{i,j}=o\\ 0 & L'_{i,j}\neq o\end{cases} \tag{3-1}$$

对特征图 F_I 和 F_S 分别应用二值掩膜 L^o 进行操作，以提取与物体 o 相关的图像特征和空间特征。然后，通过平均池化处理，即可获得物体 o 的深度特征向量 $x^i_{rgb}\in\mathbb{R}^c$ 和空间特征向量 $x^i_{spa}\in\mathbb{R}^c$。该过程可表述为：

$$\begin{aligned}x^i_{rgb}&=AveragePooling\left(F_I\odot L^o\right)\\ x^i_{spa}&=AveragePooling\left(F_S\odot L^o\right)\end{aligned} \tag{3-2}$$

其中 $\odot$ 代表元素乘积（Hadamard product）

然后，将各个物体的特征向量堆叠在一起，形成深度聚合特征 $X_{rgb}\in\mathbb{R}^{l\times c}$ 和空间聚合特征 $X_{spa}\in\mathbb{R}^{l\times c}$：

$$X_{rgb} = concat\left(x_{rgb}^{0}, x_{rgb}^{1}, x_{rgb}^{2}, ..., x_{rgb}^{149}\right)$$
$$X_{spa} = concat\left(x_{spa}^{0}, x_{spa}^{1}, x_{spa}^{2}, ..., x_{spa}^{149}\right) \tag{3-3}$$

最后，将 X_{rgb} 和 X_{spa} 输送给全局-局部依赖模块，以探索语义特征序列中的共现信息。

3.2.5 全局-局部依赖模块

鉴于同一物体在不同场景中展现出不同的特征，传统的概率统计方法已不再适用于探索语义特征序列内部的共现性。因此，本文设计了全局-局部依赖模块，该模块基于注意力机制探索语义特征序列内部的长距离依赖关系，进而优化和调整全局特征。整个过程如图 3-4 所示。

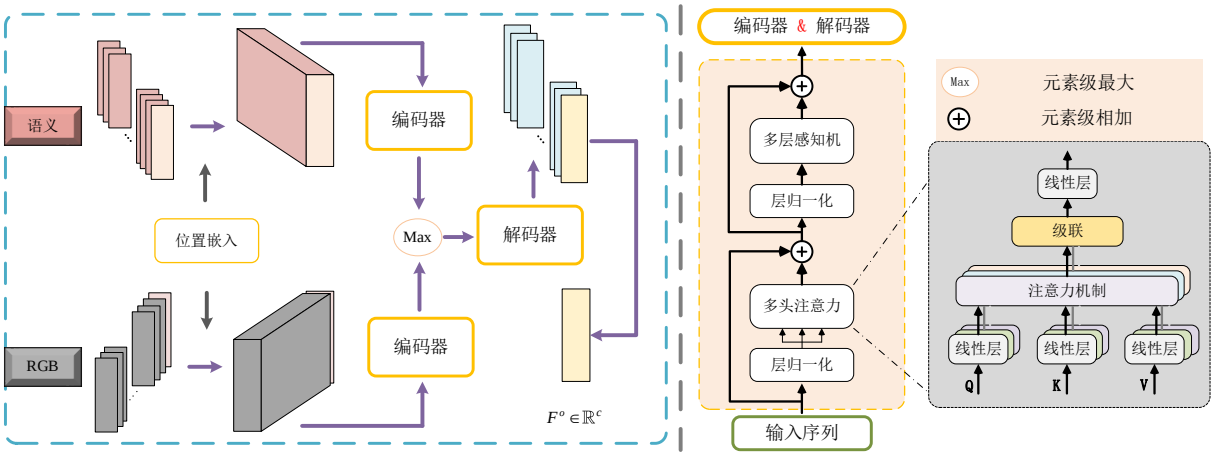


图 3-4 全局-局部依赖模块的详细结构

Fig. 3-4 Detailed structure of Global-Local Dependency Module

具体来说，通过语义节点特征聚合模块，本文构建了深度特征序列 $X_{rgb} \in \mathbb{R}^{l \times c}$ 和空间特征序列 $X_{spa} \in \mathbb{R}^{l \times c}$。考虑到最终目标是利用语义特征序列内部的长距离依赖优化全局表示。因此，首先分别对 F_I 和 F_S 进行全局平均池化处理以提取各自的全局特征，并将这两个全局特征分别融合到 X_{rgb} 和 X_{spa} 中，生成特征序列 $X_{rgb}^{1} \in \mathbb{R}^{(l+1) \times c}$ 和 $X_{spa}^{1} \in \mathbb{R}^{(l+1) \times c}$：

$$X_{rgb}^{1} = concat\left(X_{rgb}, GlobalAvgPooling\left(F_I\right)\right)$$
$$X_{spa}^{1} = concat\left(X_{spa}, GlobalAvgPooling\left(F_S\right)\right) \tag{3-4}$$

同时，考虑到记录与每个节点特征相对应的物体语义类别的必要性，进一步添加了位置嵌入 $P_{emb}^{rgb} \in \mathbb{R}^{(l+1) \times c}$ 和 $P_{emb}^{spa} \in \mathbb{R}^{(l+1) \times c}$，以生成编码后的语义特征序列 $X_{spa}^{2} \in \mathbb{R}^{(l+1) \times c}$ 和 RGB 特征序列 $X_{rgb}^{2} \in \mathbb{R}^{(l+1) \times c}$。这样一来，两个编码器的输入为：

$$\begin{aligned} X_{rgb}^{2} &= X_{rgb}^{1} + P_{emb}^{rgb} \\ X_{spa}^{2} &= X_{spa}^{1} + P_{emb}^{spa} \end{aligned} \tag{3-5}$$

考虑到这两个特征序列之间的模态差异，首先分别探索它们各自的内部相关性，然后将二者结合起来，探索整体信息的内在关系。如图 3-4 所示，本模块采用注意力机制，深入挖掘了局部语义特征节点之间以及这些节点与全局特征节点之间的长距离依赖。启发于 Vision Transformer[23]，一个编码器块包含三个核心部分：层归一化（Layer normalization）、多头自注意力（Multi-head Self-Attention，MSA）和多层感知机（Multi-Layer Perceptron，MLP）。层归一化用于规范化数据分布，增强模型的泛化能力。接着是 MSA。之后，采用残差连接传递信息以缓解过拟合。然后，再次应用层归一化，并通过 MLP 输出编码器块的最终结果。编码器处理图像聚合特征 $X_{rgb}^{2} \in \mathbb{R}^{(l+1)\times c}$ 和空间聚合特征 $X_{spa}^{2} \in \mathbb{R}^{(l+1)\times c}$ 的处理流程可表述为：

$$\begin{aligned} X_{rgb}^{3} &= X_{rgb}^{2} + MSA\left(LN\left(X_{rgb}^{2}\right)\right) \\ X_{rgb}^{4} &= X_{rgb}^{3} + MLP\left(LN\left(X_{rgb}^{3}\right)\right) \\ X_{spa}^{3} &= X_{spa}^{2} + MSA\left(LN\left(X_{spa}^{2}\right)\right) \\ X_{spa}^{4} &= X_{spa}^{3} + MLP\left(LN\left(X_{spa}^{3}\right)\right) \end{aligned} \tag{3-6}$$

通过两个编码器分别探索 X_{rgb}^{2} 和 X_{spa}^{2} 内部节点间的长期依赖关系之后，采用元素级最大值的方式将二者融合：

$$X_{o} = Max\left(X_{spa}^{4}, X_{rgb}^{4}\right) \tag{3-7}$$

为有效利用节点间的长期依赖关系优化场景表示，X_{o} 被送入一个解码器，旨在磨合图像特征与空间特征之间的差异，并深入挖掘所有信息的内在联系。解码器的结构与编码器类似，可表述为：

$$\begin{aligned} X_{o}^{1} &= X_{o} + MSA\left(LN\left(X_{o}\right)\right) \\ X_{o}^{2} &= X_{o}^{1} + MLP\left(LN\left(X_{o}^{1}\right)\right) \end{aligned} \tag{3-8}$$

通过上述过程，在所有特征节点之间建立了稳固的长期共现性，并输出了优化后的特征表示 X_{o}^{2}。

3.2.6 场景分类器

为了优化场景表示，本章采用注意力机制（编码器-解码器架构）建模全局节点和局部节点之间的长期依赖。为防止过拟合，只采用了一层编码器和一层解码器。全局-局部依赖建模过程产生了优化后的特征表示 X_{o}^{2}。如图 3-4 所示，在分类阶段，从 X_{o}^{2} 中提取出经过精细打磨的全局特征节点 $F_{o} \in \mathbb{R}^{c}$，作为最终的全局表示。F_{o} 被送至一个全

连接层以获得最终的场景预测，并使用交叉熵函数作为损失函数：

$$L=-\sum_{n=1}^{N}y_n\log\frac{exp\left(F_o\left(n\right)\right)}{\sum_{m=1}^{N}exp\left(F_o\left(m\right)\right)} \tag{3-9}$$

其中，y 代表真实标签，F_o 是本模块的输出结果。

3.3 实验结果与分析

本节首先介绍实验设置，并执行消融实验以评估各个模块对所提方法的具体影响。然后，将所提方法与当前最先进的方法进行了对比分析。

3.3.1 实验设置

本章针对室内场景识别，使用 MIT-67、SUN397（取其 175 个室内类别）、Places_7 和 Places_14 数据集验证所提算法。

与第二章一致，语义分割过程由在 ADE20K 数据集[49]上预训练的 Vision Tranformer Adapter[48]进行。对于给定的图像 $I\in\mathbb{R}^{w\times h\times 3}$，语义分割的结果是一个得分张量 $M\in\mathbb{R}^{w\times h\times l}$，该张量可以用来生成 I 的语义分割标签图 $L\in\mathbb{R}^{w\times h}$。图像 I 中位置 (i,j) 的语义预测概率被表示为 $M_{i,j}\in\mathbb{R}^{1\times 1\times l}$，其中 $l=150$ 是语义类别的总数。在标签图 L 中，位置 (i,j) 处的值 L_{ij} 表示 I 中相应位置的语义标签。

与第二章的原因一致，本章也采用了两阶段的训练方式。首先分别训练 SSRM 和 IFEM。在第二阶段，固定 SSRM 和 IFEM 的权重，仅对后续模块进行训练。使用 ALI-G 优化器[54]优化网络参数。ALI-G 仅需要一个初始学习率超参数，在所有的实验中这个参数均设为 0.1。在训练的第二阶段，为了防止过拟合，在场景分类器中应用了随机失活正则化（*Dropout*），设定失活概率为 0.8。

本文的方法是在 PyTorch[55]和 SAS-Net[33]的开源框架下实现。在评估最终性能时，采用了标准的 10-crop 评估策略[33,56]，以保证与其他方法的公平比较。

3.3.2 不同的语义节点特征聚合策略的比较

将语义分割产生的标签图与场景特征相结合，是语义节点特征聚合模块的一个关键环节。该模块输出的语义特征序列，对于后续模块探索语义节点间的长期依赖至关重要。因此，有必要通过对比实验选取最适合的特征聚合方式。如表 3-1 所示，本节尝试了多种聚合策略整合标签图和特征图。在表 3-1 中，“过滤”代表使用指定尺寸的自适应置信度滤波器（ACF，详见第 2 章）过滤语义分割得分张量；*Nearest* 指使用最近邻插值法将语义标签图插值至特定大小；*Bilinear* 指使用双线性插值法将特征图插值至特定大小。为便于后续分析表述，表 3-1 中为所有聚合策略分配了编号。此外，首先通过元素

级最大值直接结合 SSRM 和 IFEM 的输出特征，然后经过全局平均池化并将其送入分类器，以此结果作为基线性能，编号为实验 0。

表 3-1 不同的特征聚合方法的消融结果（%）

Tab. 3-1 Ablation results for different feature aggregation methods (%)

实验	得分张量	标签图	特征图	MIT-67 (%)
0	-	-	-	88.731
1	-	*Nearest* to 14×14	-	89.851
2	2×2 过滤	*Nearest* to 14×14	-	**90.746**
3	4×4 过滤	*Nearest* to 14×14	-	90.149
4	-	-	*Bilinear* to 224×224	89.776
5	2×2 过滤	-	*Bilinear* to 112×112	90.224
6	4×4 过滤	-	*Bilinear* to 56×56	90.299
7	-	*Nearest* to 56×56	*Bilinear* to 56×56	90.075
8	2×2 过滤	*Nearest* to 56×56	*Bilinear* to 56×56	90.672

如表 3-1 所示，无论采用哪种策略聚合标签图和特征图，本章所提出的方法均优于基线实验 0。这一现象验证了全局-局部依赖建模的有效性和泛化性。此外，对比实验 1、2、3；实验 4、5、6；以及实验 7、8，可以明显观察到，在得分张量上使用 ACF 比直接使用原始得分张量产生了更佳的性能，这进一步凸显了 ACF 在利用语义分割的方法中的优势。同时，比较实验 2 与 3 可以发现，使用 4×4 过滤域的 ACF 处理后的得分张量作为网络输入产生的识别精度略逊于使用 2×2 过滤域的 ACF，这与表 2-1 中展现的趋势相吻合，再次验证了在应用 ACF 处理得分张量时，需同时考虑语义歧义的过滤和图像中的物体信息的保留。

进一步地，对比实验 2、5、8 可以发现，将特征图插值到更大尺寸以对其执行更精细的语义标签分配，反而不利于提升场景识别性能。这一现象的出现可能是因为插值算法所引起的像素级特征位置的微小偏移，增加了将语义标签错误分配给特征的风险。对实验 5 和实验 6 的比较也可以支持上述结论，对特征图执行较大尺寸的插值导致输入经过 2×2 ACF 过滤的得分张量（实验 5）在识别率上反而不如 4×4 ACF 过滤的张量（实验 6）。这一现象与前文对 ACF 效果的分析相悖，进一步凸显了对场景特征进行插值产生的负面影响。幸运的是，使用未经插值的特征图作为输入不但避免了上述负面影响，还能提高语义节点特征聚合模块的处理效率。因此，本节最终选用实验 2 的配置生成语义特征序列，首先使用 2×2 ACF 处理得分张量，生成合适的语义标签图；然后，利用最近邻插值调整标签图的尺寸，使其与场景特征图尺寸一致；最后，将调整后的标签图与特

征图结合，生成最终的语义特征序列。

3.3.3 不同的编码器组合策略的对比

为充分探索由图像特征提取模块（IFEM）和语义空间关系模块（SSRM）产生的特征之间的互补信息，本节比较了四种不同的编码器组合策略：元素级乘积、特征级联、元素级相加和元素级最大值。表 3-2 展示了上述策略在 MIT-67 和 SUN397 数据集上产生的消融结果。结果表明，元素级最大值策略的性能超过了其他三种方法。相较于特征级联策略，元素级最大值策略在实现更高性能的同时，消耗的计算资源更少。与元素级乘积、相加策略相比，最大值策略在消耗相同的计算资源的情况下，更有效地保留了关键信息，从而使得解码器能够更深入地解析语义特征节点之间的依赖关系。因此，最终选用元素级最大值策略作为两个编码特征的组合方式。

表 3-2 不同的编码器组合方法的消融结果

Tab. 3-2 Ablation results for different Encoder combination methods

策略	MIT-67 (%)	SUN397 (%)
元素级乘积	89.925	75.965
特征级联	90.299	75.859
元素级相加	90.373	75.965
元素级最大值	**90.746**	**76.153**

3.3.4 网络组成成分分析

本节在 MIT-67 数据集上，对所提出的 SpaCoNet 的各个模块进行了消融研究，以评估其四个核心组成部分对模型性能的影响：PlacesCNN（作为基线）、语义空间关系模块（SSRM）、编码器和解码器（即全局-局部依赖模块）。首先，对 PlacesCNN 的输出特征执行全局平均池化操作，并将其输入到分类器中，以此作为基线。相关的实验结果汇总在表 3-3 中。

如表 3-3 所示，与基线 PlacesCNN 相比，SpaCoNet 取得了 5.776%的精度提升，验证了本章对空间上下文和物体共现性建模的巨大价值。此外，相比于第二章介绍的 CSSRM 架构，可以发现通过全局-局部依赖模块探索场景中物体之间的长距离依赖关系，能够进一步将准确率提高 2.015%，这强调了研究场景内物体共现性对于提高识别准确率的重要性。此外，本节还探索了仅使用编码器的输出进行场景识别的效果，可以看到其性能略低于完整的编码器-解码器架构。这一结果表明，解码器能够有效缓解 PlacesCNN 和 SSRM 之间的模态差异，进而更有效地挖掘和利用二者之间的互补信息。

表 3-3 网络组成部分的消融分析

Tab. 3-3 Ablation study of all network components

PlacesCNN	SSRM	编码器	解码器	精度（%）
✓	-	-	-	84.970
-	✓	-	-	81.642
✓	✓	-	-	88.731（CSSRM）
✓	✓	✓	-	90.075
✓	✓	✓	✓	**90.746**
较基线（PlacesCNN）的精度提升				5.776

3.3.5 特征可视化

本小节进行了一系列可视化实验，以定性地展示和分析所提方法的有效性。首先，基于梯度的类激活映射（Gradient-based Class Activation Map，Grad-CAM）[69]进行热力图可视化，从实例级角度探索所提出的模块如何调整网络的关注区域。随后，使用 t-分布随机邻域嵌入（t-distributed Stochastic Neighbor Embedding，t-SNE）[70]对整个数据集样本进行可视化，从数据集级角度评估所提方法对网络鉴别性能力的影响。

（1）实例级可视化

本部分使用 Grad-CAM 来可视化所提出的模块对网关注区域的影响。如图 3-5 所示，从左到右依次为：原始彩色图像、语义分割结果、基于基线模型（PlacesCNN）的热力图、结合 PlacesCNN 与 SSRM（探索空间关系）产生的热力图，以及基于所提出的 SpaCoNet（进一步探索物体共现性）的热力图。

如图 3-5 所示，所提出的模块有效调整了网络的关注区域，使网络能够更有效地利用图像中的物体进行场景预测。例如，在卧室场景中，基线模型主要关注“床”（第三列），而在引入空间关系探索后（第四列），网络进一步关注了“窗户”；在 SpaCoNet 中（第五列），注意力还扩展到了“枕头”。同时，在某些情况下，仅仅探索空间关系就足以表示场景。例如，在书店场景中，探索空间关系使网络感知到更远处楼上的“书架”（第四列）。这表明空间关系为网络提供了额外的补充信息，使其能够关注更细致的场景细节。然而，由于不存在更多显著的鉴别性区域，进一步探索物体共现性（第五列）只是扩展了网络对“书架”的关注。

此外，在某些情况下，探索空间关系对网络的帮助是有限的，此时整合场景特征以探索物体共现性则更为有益。例如，在机房场景中，仅仅探索空间关系（第四列）未能有效调整基线模型的关注区域（第三列）。这可能是因为“桌子”在多个场景中（如餐厅）都很常见。相比之下，SpaCoNet（第五列）将场景相关特征分配给物体并建模物体

共现性，使网络能够识别机房中“桌子”的独特特征，从而进一步将“桌子”识别为场景预测的关键区域。类似地，在病房场景中，由于“电脑”在其他场景（如机房）中也很常见，第三和第四列未将“电脑”视为场景预测的贡献因素。相比之下，SpaCoNet（第五列）通过整合场景特征来探索物体共现性，进一步将“电脑”视为场景预测的重要元素。这些例子充分展示了结合场景特征建模物体共现性在场景识别中的重大意义。

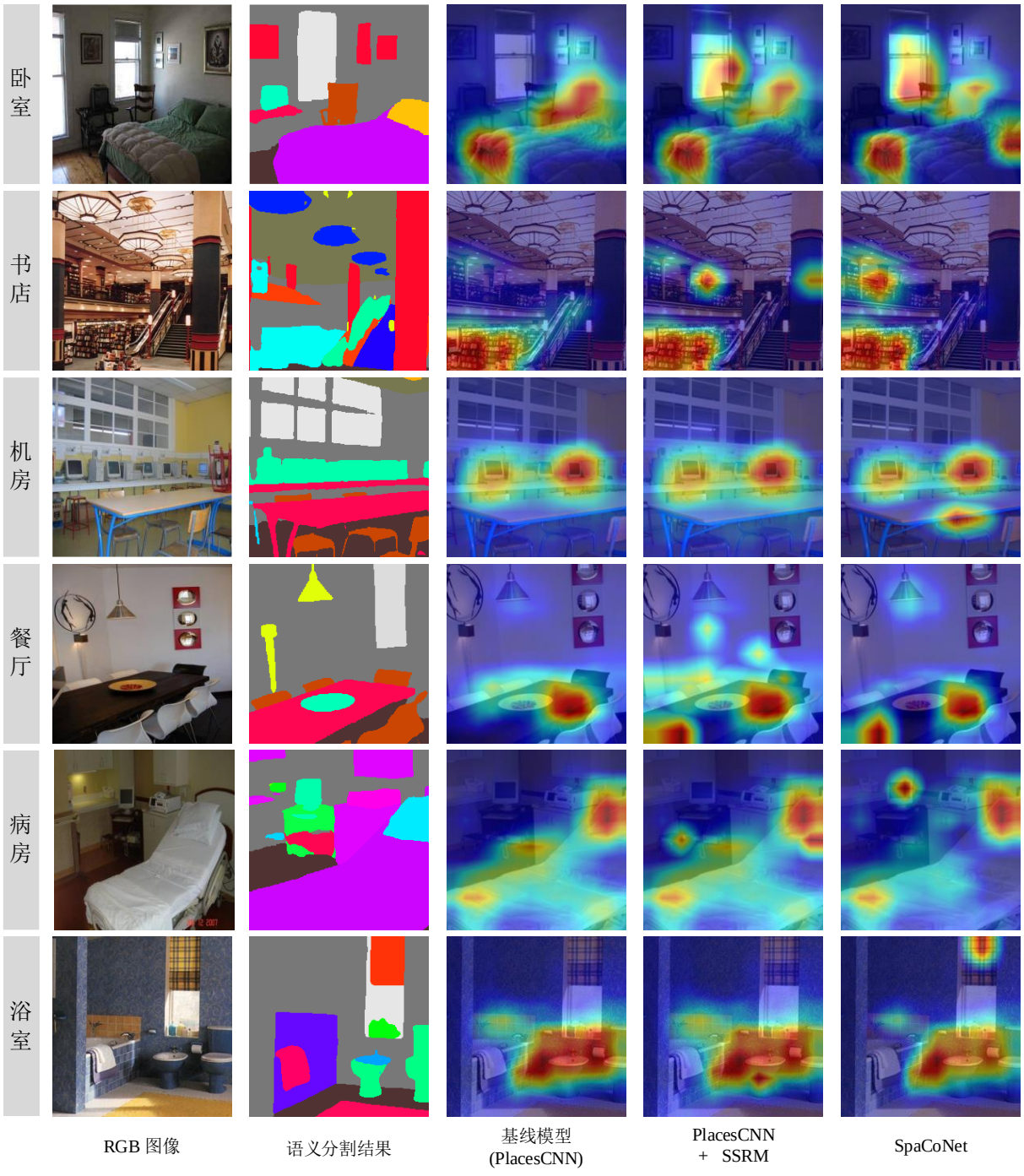


图 3-5 实例级特征可视化

Fig. 3-5 Instance-level feature visualization

（2）数据集级可视化

本部分在 MIT-67、Places_7、Places_14 和 SUN397 数据集的测试集上对提出的 SpaCoNet 和基线模型（PlacesCNN）进行了评估，精度对比如表 3-4 所示。此外，为直观比较两个模型即将输入场景分类器的特征，采用 t-SNE[70]技术，通过在图 3-6 中绘制这些特征的二维分布图进行可视化分析。图 3-6 的上半部分展示了 PlacesCNN 提取的特征的可视化效果，下半部分则展示了 SpaCoNet 的输出特征的可视化。

表 3-4 特征可视化的消融研究

Tab. 3-4 Ablation study of feature visualization

方案	MIT-67 (%)	Places_7 (%)	Places_14 (%)	SUN397 (%)
基线（PlacesCNN）	84.970	93.0	87.643	73.129
SpaCoNet	90.746	94.286	89.714	76.153

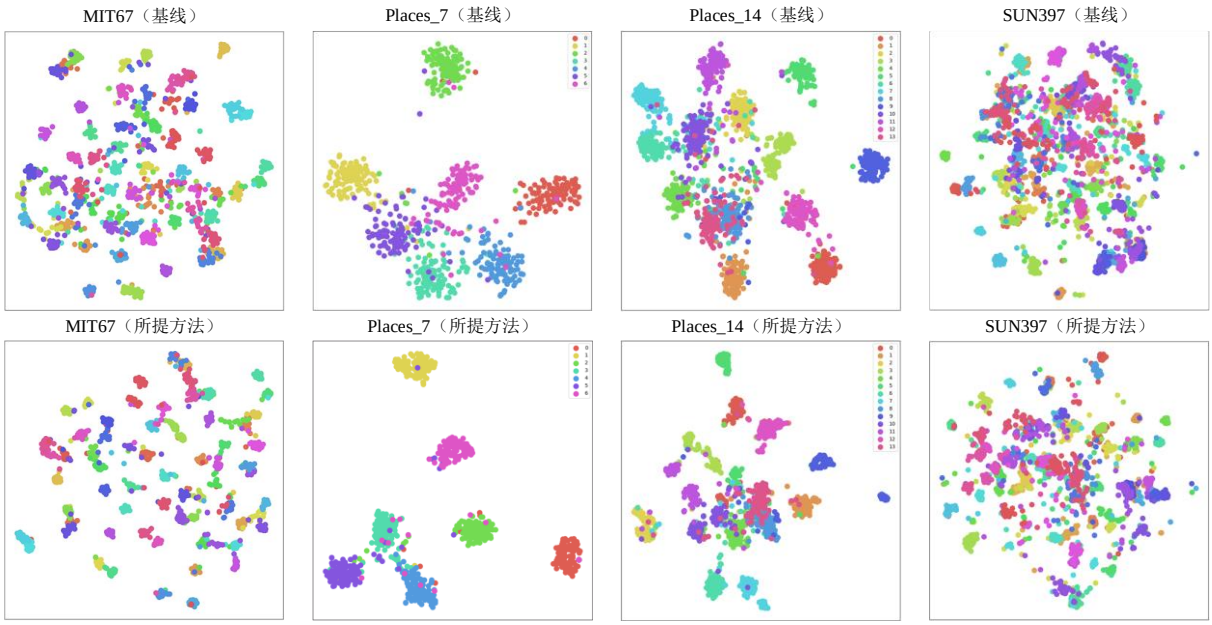


图 3-6 测试集的特征分布可视化

Fig. 3-6 Feature distributions of test sets

注：每个点代表一张图片，相同颜色的点代表属于同一类别的图片。

如图 3-6 所示，相比于基线特征分布，SpaCoNet 使同一类别的场景特征分布更加集中，不同类别的场景特征分布更加分散。这种现象在类别较少的 Places_7 和 Places_14 数据集中尤为明显，混合在一起的不同类别点更少。此外，即使在包含大量类别的 MIT-67 和 SUN397 数据集中，SpaCoNet 仍显著增强了不同场景类之间的边界，并减少了相同类别之间的差异。这些现象表明，同时建模物体间的空间关系和长期共现性显著提高了场景特征的判别能力，从而带来了更优越的场景识别性能（参见表 3-4）。

3.3.6 与现有算法的性能比较

如表 3-5 和表 3-6 所示，本节将 SpaCoNet 与多个现有的先进算法在 MIT-67、SUN397、Places-7 和 Places-14 数据集的表现上进行了对比。

表 3-5 在 MIT-67 和 SUN397 数据集上 SpaCoNet 与现有先进算法的比较

Tab. 3-5 Comparison of SpaCoNet with the state-of-the-art methods on MIT-67 and SUN397

算法	输入尺寸	MIT-67 (%)	SUN397 (%)
Dual CNN-DL[26]	224×224	76.56	70.13
NNSD+ICLC[37]	224×224	84.30	64.78
SDO-Net[32]	224×224	86.76	73.41
SAS-Net[33]	224×224	87.10	74.04
CCF-Net[38]	512×512	87.30	-
DeepScene-Net[57]	224×224	71.00	-
ARG-Net[42]	448×448	88.13	75.02
PL + AP + AI + IM[58]	448×448	88.06	74.12
LGN[28]	448×448	88.06	73.85
MR-Net[59]	448×448	88.08	73.98
FCT-Net[64]	224×224	89.17	76.06
AdaNFF[71]	256×256	-	74.18
CSSRM	224×224	88.73	75.48
SpaCoNet	224×224	**90.75**	**76.15**

注：CSSRM 为本文第二章所提出的算法

可以看到，SpaCoNet 超越了大多数现有方法。与那些同样利用物体信息辅助场景识别的算法[32,33,39,42,45,52]相比，本章方法展现了更优异的性能。这验证了通过深入挖掘物体间的空间关系和长期共现依赖性进行场景识别的有效性。此外，相较于那些采用多分支策略捕获场景的多尺度信息的方法[33,38,45,58,59]，SpaCoNet 同样表现出色，这进一步凸显了利用物体信息作为额外信息源的重要意义。值得注意的是，尽管有些研究通过增大网络的输入尺寸提高识别准确率[28,38,42,58,59,71]，本章所提出的方法仅使用 224×224 的输入尺寸就实现了卓越的性能，有效降低了对计算资源的需求。上述性能比较结果进一步验证了所提方法在室内场景识别领域的优越性和良好的泛化能力。

此外，本文第二章介绍的算法 CSSRM 通过建模场景内的空间上下文关系进行场景识别。与之相比，本章提出的 SpaCoNet 在充分建模场景内空间上下文关系的基础上，进一步对物体共现性进行了探索，有效缓解了不同场景中具有相同语义的物体所导致的类间相似性问题，从而获得了更优越的场景识别性能。

表 3-6 在 Places_14 和 Places_7 数据集上 SpaCoNet 与现有先进算法的比较

Tab. 3-6 Comparison of SpaCoNet with the state-of-the-art methods on Places_14 and Places_7

算法	输入尺寸	Places_14 (%)	Places_7 (%)
Word2Vec[61]	224×224	83.70	-
Deduce[39]	224×224	-	88.10
BORM[45]	224×224	85.80	90.10
OTS-Net[52]	224×224	85.90	90.10
AGCN[62]	224×224	86.00	91.70
CSSRM	224×224	88.71	93.43
SpaCoNet	224×224	**89.71**	**94.29**

注：CSSRM 为本文第二章所提出的算法

3.4 本章小结

本章提出了一个新颖的室内场景识别框架 SpaCoNet，通过同时建模场景内的空间信息和物体间共现性来构建场景表示。首先，利用在第二章中构建的语义空间关系模块（SSRM）提取场景内的空间上下文特征，利用图像特征提取模块提取图像整体特征；然后，为深入探索不同场景中物体间的共现性，并有效区分不同场景中具有相同语义的物体，提出了一种新颖的方法，将场景相关特征与物体语义紧密结合，进而将语义分割技术可分割的所有物体类别建模为特征节点；然后，利用注意力机制挖掘物体语义特征序列中的长期共现性，从而生成了更具鉴别意义的场景表示。在一系列的定性和定量实验中，本章算法展现出了卓越的性能和竞争力。

尽管如此，本章算法主要从启发式的角度出发利用场景内的物体信息辅助场景识别，其可解释性相对较弱。下一章将采用推理式方法深入挖掘隐藏在物体信息中的鉴别性知识，从更具可解释性的角度利用物体信息增强场景特征的鉴别力。

4 基于物体间鉴别性图建模的室内场景识别方法

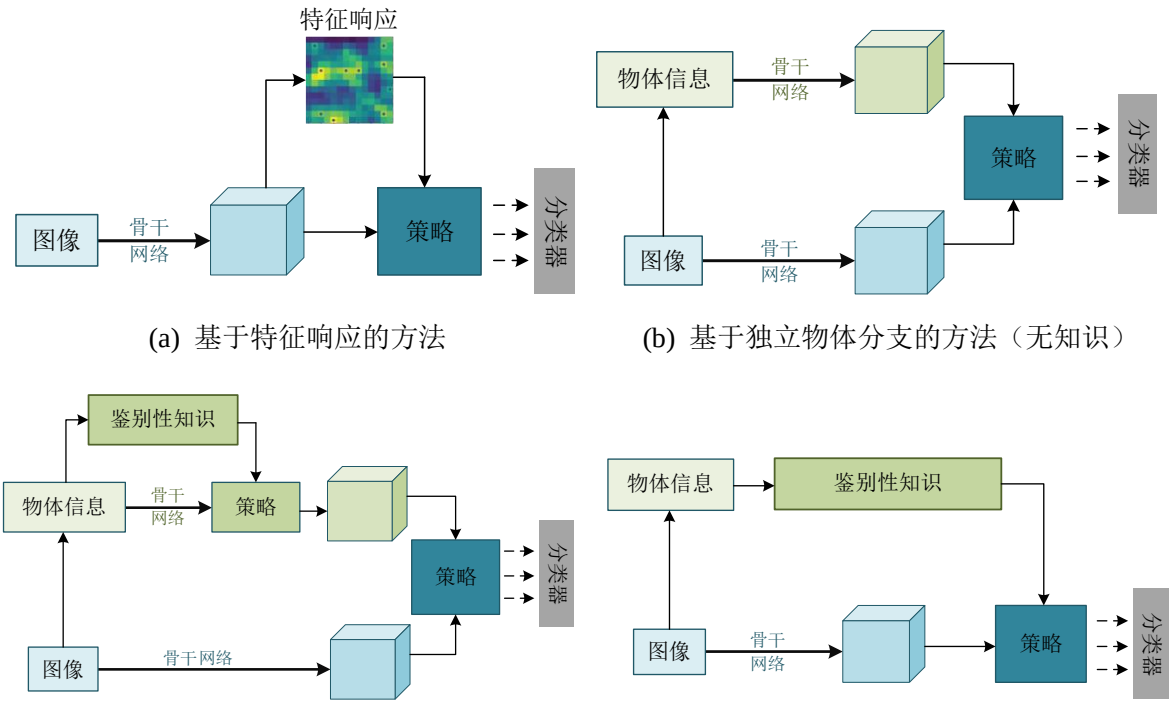
4.1 引言

室内场景识别不仅需要探索同一场景类别内的共同特征，还需要探索不同场景类别间的独特特征。然而，多变的场景布局和跨场景共存的物体使得这些特征难以捉摸。因此，如何提取具有高鉴别性的图像表示一直是室内场景识别领域的核心问题。

为提高场景特征的鉴别能力，一些研究使用在大型数据集上预训练的卷积神经网络（CNN）进行特征选择。如图 4-1（a）所示，这些基于特征响应的方法，通过分析卷积通道响应[27,28,72]或类激活图[59]提取输入图像的代表性特征，并进一步通过分析这些特征之间的关系实现场景识别。显然，这种方法高度依赖预训练网络的特征鉴别能力。然而，神经网络固有的不确定性意味着 CNN 的高响应区域[59]可能聚焦在非鉴别性区域。神经网络的这种“黑盒”特性促使研究者转向利用场景内具体的物体信息获得更具鉴别性的图像表示。

为利用物体信息辅助提高场景特征的鉴别性，如图 4-1（b）所示，一些研究[10,33,38]，包括本文第二章和第三章提出的算法，设计了专门用于提取与物体信息相关的特征的网络分支。通过将这些物体相关特征与场景特征结合，形成更具区分力的表示。此外，如图 4-1（c）所示，也有一些方法[32,43-46]尝试从概率角度提取鉴别性物体知识，并基于这些知识优化物体特征的提取过程。这两种方法都采用了多分支架构，独立处理场景图像和物体信息。在这种架构中，场景特征和物体特征通过级联或者相加等方式融合，但这些融合方式往往是启发式的，可能导致物体知识与场景特征的次优融合。

为了克服上述限制，近期一些方法[42,52]提出将物体语义赋予给场景特征，通过探索物体区域间的上下文关系实现场景识别。这些方法提供了对物体信息更具可解释性的利用，并表现出了卓越的性能。但是，这些方法倾向于对每个区域特征一视同仁，忽视了对鉴别性物体知识的利用。此外，这种构建区域级特征的方式可能会弱化存储在像素中的宝贵信息的充分利用[73]。研究表明，某些鉴别性物体只在特定场景中频繁出现，而普通物体则在不同场景中的出现概率类似[32]。受此启发，本文认为与鉴别性物体相关的特征应该受到更多的关注。据此，如图 4-1（d）所示，提出将鉴别性物体知识映射至场景特征之间的相关性上，旨在从推理角度对像素级场景特征进行微调，从而生成更具鉴别力的图像表示。



(a) 基于特征响应的方法　(b) 基于独立物体分支的方法（无知识）

(c) 基于独立物体分支的方法（有知识）　(d) 本章方法

图 4-1 现有的方法和本章方法的图例

Fig. 4-1 Illustrations of existing and proposed methods

4.2 基于物体间鉴别性图建模的室内场景识别方法

4.2.1 总体架构

最近的研究[10,32]强调了鉴别性物体在场景识别中的重要性，指出这些物体往往只出现在特定场景中，与在各种场景下普遍存在的普通物体不同。这一见解促使本章更加关注与鉴别性物体相关的场景特征，以获得更具鉴别性的图像表示。

随之而来的挑战是如何量化对与物体相关的场景特征分配的关注度。启发于 SDO-Net[32]，本研究采用贝叶斯概率统计方法，构建了一个物体间鉴别性原型（Inter-Object Discriminative Prototype，IODP），以建模物体之间的相关性。在此基础上，基于图（Graph）论将 IODP 中的先验知识与场景特征结合起来。更具体地讲，提出了一个鉴别性图网络（Discriminative Graph Network，DGN），在 DGN 中，像素级的场景特征被定义为图的节点（Node）。通过语义分割技术获得这些节点位置的物体语义后，利用 IODP 确定这些节点之间的鉴别性关系，然后将这种关系编码为图的边（Edge）。DGN 通过图卷积和映射操作（GCN）优化每个像素级场景特征，最终产生更具鉴别能力的图像表示。图 4-2 展示了本章方法的概念性工作流程，接下来的章节将对该流程进行详细描述。

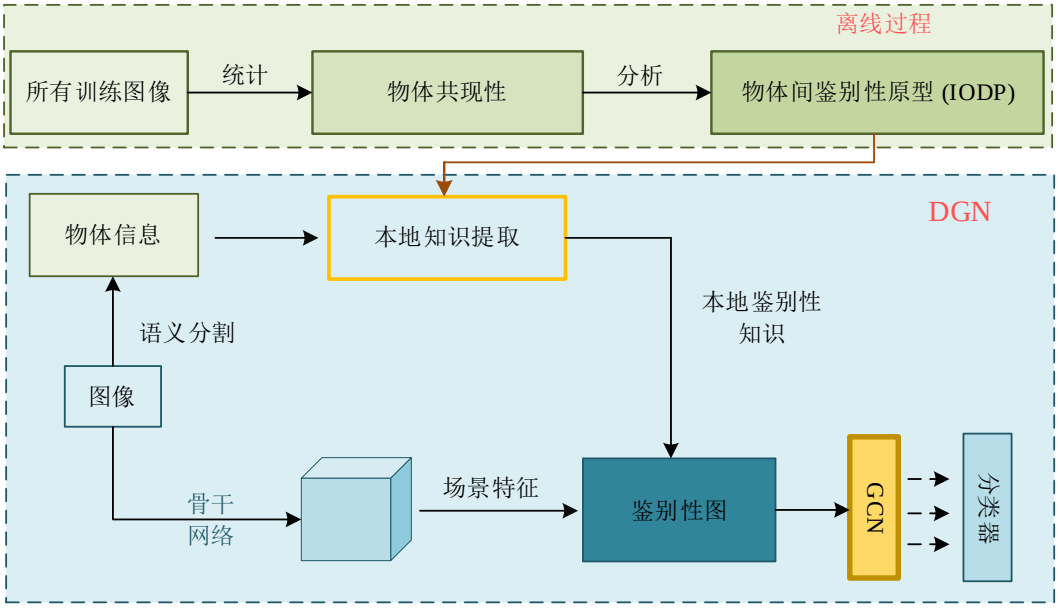


图 4-2 物体间鉴别性原型和图建模的整体工作流程

Fig. 4-2 Overall workflow of Inter-Object Discriminative Prototype and Graph modeling

4.2.2 物体间鉴别性原型

受SDO-Net[32]的启发，本研究采用贝叶斯概率统计方法深入挖掘物体知识。SDO-Net主要关注单个物体的鉴别性。然而，考虑到本研究的最终目标是挖掘场景特征之间的鉴别相关性，本章将焦点转移到了探索物体间的鉴别性知识，提出了一个物体间鉴别性原型（IODP），如图 4-3 所示。

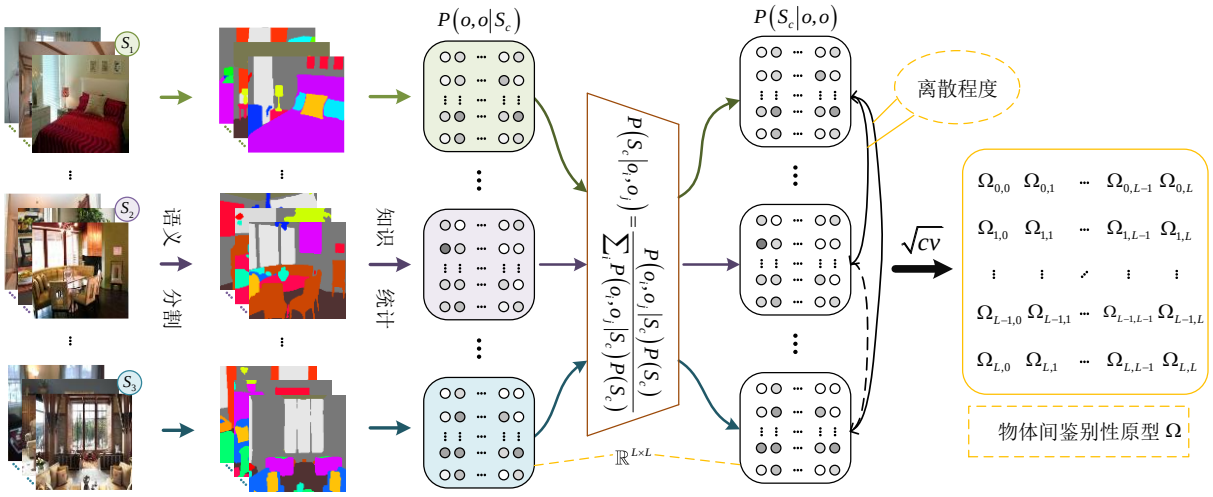


图 4-3 物体间鉴别性原型的构建过程描述

Fig. 4-3 Process description of constructing Inter-Object Discriminative Prototype

注：整个过程基于场景数据集的训练数据实现，不同颜色的块代表不同的场景类别。

如图 4-3 所示，首先对场景数据集中的训练数据进行语义分割处理，并通过统计方

法确定物体对在各个场景中的分布情况。给定一个有C个场景类别的数据集，用S_c表示其第c_{th}个场景类，$c\in\{1,2,...,C\}$。通过处理S_c中的每一个实例，可以计算物体o_i和o_j同时出现在S_c中的概率$P(o_i,o_j|S_c)$。Places_7 和 Places_14 数据集的数据量相对充足，共现概率可以直接计算。然而，MIT-67 和 SUN397 数据集相对有限的数据量则对数据代表能力提出了挑战。因此，通过假设不同物体独立同分布以估算后两个数据集中的共现概率，即首先分别计算o_i和o_j的分布，再将二者结合起来近似计算联合分布$P(o_i,o_j|S_c)$，从而缓解潜在的误差。上述过程可表示为：

$$P(o_i,o_j|S_c)=\begin{cases}\dfrac{N_{o_i,o_j}}{N_{S_c}}, & Places_7\ or\ Places_14\\ P(o_i|S_c)P(o_j|S_c)=\dfrac{N_{o_i}\times N_{o_j}}{N_{S_c}^2}, & MIT67\ or\ SUN397\end{cases} \tag{4-1}$$

其中，N_{S_c}表示场景类S_c中的实例总数。N_{o_i,o_j}表示同时出现物体o_i和o_j的实例数量。N_{o_i}和N_{o_j}分别指出现o_i和o_j的实例数量，其中i和j的取值范围均为$[1,L]$，而L表示语义分割技术可分割的物体类别总数。本文将在章节 4.3 中将对上述两种计算方式产生的性能进行比较。

在计算出所有的物体共现分布之后，即可通过贝叶斯后验概率公式推导出$P(S_c|o_i,o_j)$：

$$P(S_c|o_i,o_j)=\frac{P(o_i,o_j|S_c)P(S_c)}{\sum_c P(o_i,o_j|S_c)P(S_c)} \tag{4-2}$$

其中，$P(S_c)$表示场景类S_c的概率，$P(S_c)=1/C$。

给定场景中存在物体o_i和o_j，可以为所有场景类别构建后验概率矩阵$P(S|o_i,o_j)$：

$$P(S|\,o_i,o_j)=\left[P(S_1|\,o_i,o_j),\ldots,P(S_c|\,o_i,o_j),\ldots,P(S_C|\,o_i,o_j)\right] \tag{4-3}$$

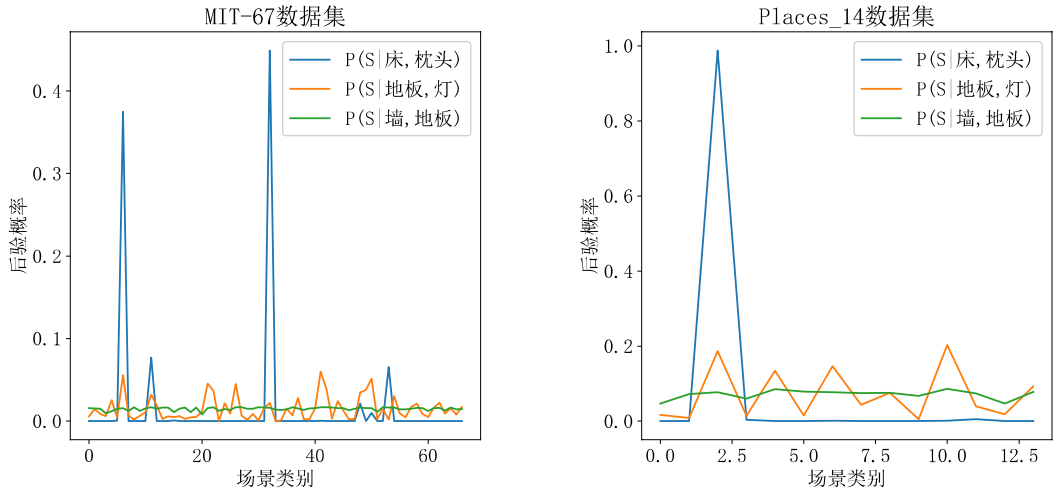
图 4-4 可视化了针对不同场景类别，物体对{床，枕头}，{地板，灯}和{墙，地板}生成的后验概率矩阵。可以看到，物体对{床，枕头}只对特定场景呈现高概率，在其他场景中显示出较低的概率，这说明{床，枕头}具有很强的鉴别性意义。相比之下，{墙，地板}对所有场景都呈现出相似的概率，显然对场景识别的帮助微乎其微。{地板，灯}的鉴别性意义则介于{床，枕头}和{墙，地板}之间。显然，后验概率矩阵分布的波动越大，物体对的鉴别性能力越强。基于此发现，本章提出采用$P(S|o_i,o_j)$的离散程度衡量o_i和o_j之间的鉴别性相关度，并尝试了几种常用的指标衡量离散程度，包括极差R、标准差σ和离散系数cv：

$$\theta_{i,j}=\begin{cases}\max\left(P\left(S|o_i,o_j\right)\right)-\min\left(P\left(S|o_i,o_j\right)\right), & R\\ \sigma\left(P\left(S|o_i,o_j\right)\right), & \sigma\\ cv\left(P\left(S|o_i,o_j\right)\right)=\dfrac{\sigma\left(P\left(S|o_i,o_j\right)\right)}{\mu\left(P\left(S|o_i,o_j\right)\right)}, & cv\end{cases} \tag{4-4}$$

其中 $\theta_{i,j}$ 代表 $P\left(S|o_i,o_j\right)$ 的离散程度。$\sigma(\cdot)$ 代表 $P\left(S|o_i,o_j\right)$ 的标准差，$\mu(\cdot)$ 代表 $P\left(S|o_i,o_j\right)$ 的算数平均数。

此外，由于上述推理过程忽略了图像中的一些细节（如颜色等），可能会导致最终的物体间鉴别性原型 Ω 过于锐利而脱离实际情况。为了避免这种情况的发生，通过钝化 $\theta_{i,j}$ 得到鉴别性相关值 $\Omega_{i,j}$：

$$\Omega_{i,j}=\sqrt{\theta_{i,j}} \tag{4-5}$$



(a) MIT-67 数据集　　(b) Places_14 数据集

图 4-4 后验概率可视化

Fig. 4-4 Posterior probabilities visualization

本文将在章节 4.3 进行消融实验，选择衡量物体间鉴别性相关性的最佳方式。计算完每个物体对之间的鉴别性相关性后，即可构建最终的物体间鉴别性原型 $\Omega\in\mathbb{R}^{L\times L}$（IODP）。生成 IODP 的伪代码见算法 4-1。IODP 可以为神经网络提供丰富的物体间鉴别性知识。

算法 4-1：物体间鉴别性原型构建过程

Data： L 代表语义分割可分割的物体类别的数量

Input： C 个场景类别的训练实例：$S_c,\ c=1,...C$

Output： 物体间鉴别性原型 $\Omega\in\mathbb{R}^{L\times L}$

for $c=1$ to C **do**
  提取 S_c 内所有实例的语义分割标签图；
  for $i=1$ to L **do**
    for $j=1$ to L **do**
      处理所有标签图，获得物体共现分布 $P\left(o_i,o_j\middle|S_c\right)$；
      $$P\left(S_c\middle|o_i,o_j\right)=\frac{P\left(o_i,o_j\middle|S_c\right)P\left(S_c\right)}{\sum_c P\left(o_i,o_j\middle|S_c\right)P\left(S_c\right)}$$
    end for
  end for
end for
for $i=1$ to L **do**
  for $j=1$ to L **do**
    $$P\left(S\middle|\ o_i,o_j\right)=\left[P\left(S_1\middle|\ o_i,o_j\right),\ldots,P\left(S_c\middle|\ o_i,o_j\right),\ldots,P\left(S_C\middle|\ o_i,o_j\right)\right]$$
    $$\theta_{i,j}=cv\left(P\left(S\middle|o_i,o_j\right)\right)=\frac{\sigma\left(P\left(S\middle|o_i,o_j\right)\right)}{\mu\left(P\left(S\middle|o_i,o_j\right)\right)}$$
    $$\Omega_{i,j}=\sqrt{\theta_{i,j}}$$
  end for
end for
return Ω；

4.2.3 鉴别性图网络

为利用构建出的 IODP 增强场景特征的鉴别性能力，本节提出了鉴别性图网络（DGN）。图 4-5 展示了 DGN 的整体流程。

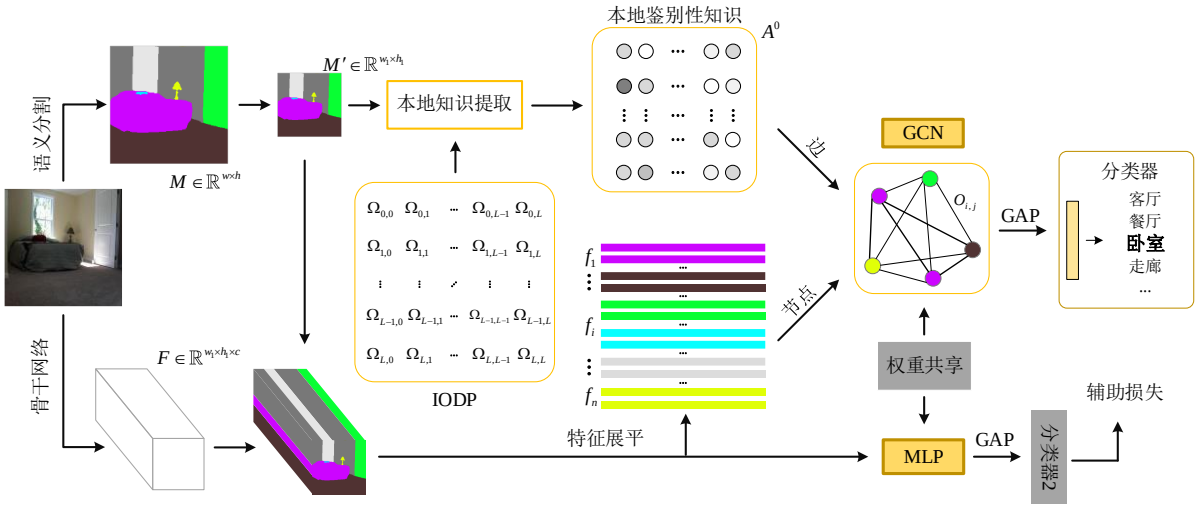


图 4-5 鉴别性图网络（DGN）的整体流程

Fig. 4-5 Overall process of Discriminative Graph Network (DGN)

注：本地知识提取过程见算法 4-2。

算法 4-2：从 IODP 中提取本地鉴别性知识的过程

Input： $V=[f_1, f_2, ..., f_i, ... f_n]$，IODP Ω，$M''=[m_1, m_2, ..., m_i, ..., m_n]$

Output： 本地鉴别性知识 $A^0 \in \mathbb{R}^{n \times n}$

for $i=1$ to n **do**
  从 M'' 中获得 f_i 对应的物体语义 m_i；
  for $j=1$ to n **do**
    从 M'' 中获得 f_j 对应的物体语义 m_j；
    $A^0_{i,j} = \Omega_{m_i, m_j}$
  end for
end for
return A^0；

（1）构建鉴别性图

给定图像 $I \in \mathbb{R}^{w \times h \times 3}$，其语义分割标签图可表示为 $M \in \mathbb{R}^{w \times h}$。首先，将 I 输入到骨干网络中，以提取对应的场景特征 $F \in \mathbb{R}^{w_1 \times h_1 \times c}$；然后，通过对 M 应用近邻插值算法，获得与 F 分辨率一致的 $M' \in \mathbb{R}^{w_1 \times h_1}$，$M'$ 可提供 F 中各个像素级特征对应的物体语义信息。基于这些表示和 IODP，本节构建了一个鉴别性图 $G=(V,E,A)$，其中 V 表示图中的节点，E 表示图的边，A 表示邻接矩阵。以往的研究[28,42]将区域级特征聚合为节点，这可能会破坏场景特征的原始拓扑性质[73]。与这些方法不同，得益于 IODP 同时具有不同物体之间和相同物体之间的鉴别性知识，本研究利用所有像素级特征构建节点 V，确保场景特

征信息保持完整。

具体来讲，场景特征 F 中所有的像素级特征都被用于构建图节点 $V=[f_1,f_2,...,f_i,...,f_n]$，其中 $f_i\in\mathbb{R}^c$ 表示 F 的第 i_{th} 个像素级特征，$n=w_1\times h_1$。节点之间的边 E 被编码为节点对应物体语义之间的鉴别性关系（由 IODP 提供）。为便于处理，首先将 M' 转换为一维形式的 $M''=[m_1,m_2,...,m_i,...,m_n]$，以与 V 中的节点一一对应。n 个节点之间的本地鉴别性知识 $A^0\in\mathbb{R}^{n\times n}$ 可以通过算法 4-2 从 IODP 中提取出来。以图 4-5 中的输入图像为例，图 4-6 展示了相应的 A^0 的定性表示。

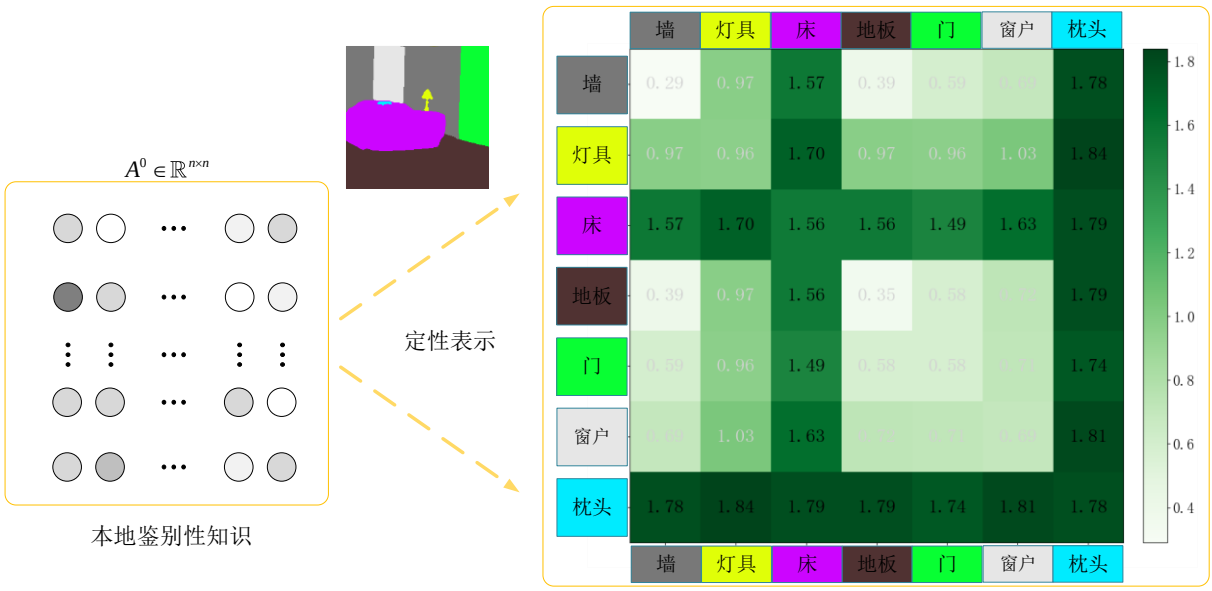


图 4-6 本地鉴别性知识的定性表示

Fig. 4-6 Qualitative representation of the local discriminative knowledge

注 1：所展现的定性表示并非 A^0，而是 A^0 中所有的两两节点之间的鉴别性相关程度。

注 2：颜色越深，对应的物体之间的鉴别性相关程度越强。

如图 4-6 所示，A^0 中节点之间的鉴别值超出了区间 $[0,1]$，这不利于后续处理。因此，对 A^0 进行行归一化处理，生成鉴别性图的邻接矩阵 A：

$$A_{i,j}=A^0_{i,j}/sum\left(A_{i,:}\right) \tag{4-6}$$

按照上述步骤，即可成功构建本章的鉴别性图 G。

（2）图卷积特征映射

前面的步骤构建了鉴别性图 $G=(V,E,A)$，其中 $V\in\mathbb{R}^{n\times c}$ 代表节点集，$A\in\mathbb{R}^{n\times n}$ 是相应的邻接矩阵。考虑到图卷积网络（Graph Convolution Network，GCN）在图优化方面的优越性，将其应用于提出的鉴别性图，以获得最优的图像表示。应用 GCN 的过程可以表示为：

$$V^* = \eta(D^{-1}\tilde{A}VW)$$
$$\tilde{A} = A + I_N \tag{4-7}$$

其中V^*表示优化后的节点表示，$W \in \mathbb{R}^{c\times d}$是可训练的权重矩阵，$d$是 GCN 的隐藏维度。$D_{i,j} = \sum_j A_{i,j}$是$A$的度矩阵，$\eta$表示激活函数*Sigmoid*。

为避免过拟合，只采用一层图卷积层获得优化后的节点表示$V^* \in \mathbb{R}^{n\times d}$。

（3）场景分类器和辅助损失

对V^*应用全局平均池化（Global Average Pooling，GAP），即可获得优化后的图像表示f^o。接下来，将f^o送入一个全连接层进行场景类别预测。在训练阶段，*Softmax*函数和交叉熵损失函数被用于计算分类损失l_o。

本文通过建模鉴别性图，将 IODP 中的物体间先验知识融入到场景特征中，以增强图像表示的鉴别性能力。然而，一个潜在的问题是：在训练阶段，对骨干网络参数的调整可能是次优的。更具体地说，通过 DGN 利用 IODP 中的物体间先验知识提高场景特征的鉴别性，虽然减少了模型损失，但这种损失下降可能会引起骨干网络参数的次优调整。为避免这一问题，本研究在 DGN 中引入了一个辅助分类器。该分类器生成的辅助损失专门用于优化骨干网络参数，从而促进其更精细的调整。

具体来讲，节点特征表示V被送入 GCN 层的同时，也被送入一个多层感知机（MLP）。MLP 层的权重与 GCN 层的隐藏层权重W共享，以确保参数的一致性。随后，在 MLP 层后添加一个辅助场景分类器，以生成辅助损失l_a。既利用 IODP 提供的先验知识增强特征鉴别性能力，又确保了对骨干网络参数的充分训练。需要注意的是，辅助分类器的权重是独立于主分类器的。

通过结合两个分类器的损失，即可得到最终的损失l：

$$l = l_o + \lambda l_a \tag{4-8}$$

其中，λ是调整辅助损失贡献度的超参数。注意，辅助损失只被用于训练过程。

4.3 实验结果与分析

本节通过广泛的对比实验验证所提出的室内场景识别方法的有效性。首先介绍实验的设置细节；随后，分析关键参数及其最优值，并对所提出的方法进行了消融研究；最后，展示并分析了定性和定量实验结果。

4.3.1 实验设置

本章针对室内场景识别，使用 MIT-67、SUN397（取其 175 个室内类别）、Places_7 和 Places_14 数据集验证所提算法。

（1）语义分割

与前面两章的方法保持一致，语义分割过程由在 ADE20K 数据集[49]上预训练的 Vision Tranformer Adapter[48]进行。考虑到语义分割精度的固有限制，在生成语义分割标签图之前，使用了第二章提出的自适应置信度过滤器（ACF）对语义分割得分张量进行了预处理。为提高计算效率，直接对语义分割标签图进行近邻插值，确保其分辨率与骨干网络输出的特征图一致。需要指出的是，提高分辨率可能会增强识别性能，但这并非本研究的工作重点。

（2）训练细节和超参数

骨干网络的输入图像分辨率固定为224×224。在训练阶段，通过 Adam 优化器[74]对模型进行了 30 个周期（Epoch）的训练，且所有参数均采用 Xavier 初始化进行了随机初始化。在训练的初始阶段，图模块的学习率为 0.001，并在第 10、15 和 20 个训练周期时下降至原来的十分之一。为了保持预训练骨干网络的特征提取能力，在训练初始阶段其学习率被设置为 0.00001，并在第 10 个周期进一步减少至原来的十分之一。此外，使用了权重衰减（值为 0.00001）对模型进行了正则化处理。对于 MIT-67 和 SUN397 数据集，批处理大小（Batch size）设为 32；对于 Places_7 和 Places_14 数据集，由于其更大的数据量和更快的收敛速度，批处理大小被调整为 64。所有实验均在单个 NVIDIA 3090 GPU 上基于 Pytorch[55]和 SAS-Net[33]框架完成。为确保实验的可重复性，所有训练过程的随机种子均被设置为 304。

本章采用 ConvNext[75]作为骨干网络进行特征提取。为节约计算资源，首先利用在 ImageNet 1k[18]上预训练的 ConvNext_Tiny 进行消融实验，随后采用在 ImageNet 22k[18]上预训练的 ConvNext_Base 评估所提方法的性能。在评估最终性能时，采用了标准的 10-crop 评估策略[33,56]，以保证与其他方法的公平比较。

（3）基线

为便于后续表述，本章构建了一个基线模型，该模型将骨干网络的输出特征作为输入，经过平均池化处理，并通过全连接层进行场景预测。该模型的性能将作为后续所有实验的基线。

4.3.2 物体间鉴别性原型的设计细节选择

本节将进行对比实验以确定章节 4.2.2 中描述的生成 IODP 的设计细节。为便于比较，在本节的所有实验中，暂时将辅助损失中的λ设置为 1，GCN 中的隐藏维度d配置为 768。

（1）不同数据集的物体共现计算方式选择

本小节对比了不同的物体共现计算方式对最终的场景识别性能的影响。为了便于比较，这里暂时采用离散系数cv衡量物体之间的鉴别性相关度。实验结果如表 4-1 所示。

表 4-1 中，“非独立分布”指通过常规统计方法计算两个物体的共现分布，即式（4-1）中所定义的$\frac{N_{o_i,o_j}}{N_{S_c}}$；而“独立分布”则是指首先分别计算两个物体的分布，再基于独立同分布假设估算二者的共现分布，即式（4-1）中的$\frac{N_{o_i} \times N_{o_j}}{N_{S_c}^2}$。更多细节可参见式（4-1）。

表 4-1 不同的物体共现计算方式产生的分类精度（%）

Tab. 4-1 Effect of using different ways to compute the object co-occurrence on accuracy (%)

数据集	非独立分布	独立分布
MIT-67	81.791	**82.015**
SUN397	66.094	**66.471**
Places_7	**92.714**	92.229
Places_14	**88.571**	88.5

根据表 4-1 所示结果，可以观察到在 Places_7 和 Places_14 数据集上，以常规方式计算两个物体间的共现概率分布（即“非独立分布”）产生了相对较好的性能；而在 MIT-67 和 SUN397 数据集上，基于两个物体各自的独立分布估算二者共现性（即“独立分布”）则展现出更优的性能。这一发现与章节 4.2.2 中的讨论保持一致，可能源于不同数据集的数据量差异。

基于上述实验结果，在本章后续所有实验中，将根据数据集的规模选择不同的策略计算物体共现性。对于数据量有限的 MIT-67 和 SUN397 数据集，采用“独立分布”方式计算物体共现性；而对于包含充足数据的 Places_7 和 Places_14 数据集，则采用“非独立分布”方式计算物体共现性。

（2）物体间鉴别性相关程度的衡量方式选择

为确定量化物体间鉴别性相关程度的最优方式，本小节深入比较分析了式（4-4）和（4-5）中的几种衡量指标及其变体对性能的影响，结果如图 4-7 所示。在图 4-7 中，用空心圆圈出的点是产生最高精度的衡量方式。R 代表极差 $Range$，σ 代表标准差，cv 代表离散系数。

如图 4-7 所示，显而易见，无论采用何种衡量方式评估物体间的鉴别性相关程度，所提出的方法均在所有数据集上显著提升了网络性能，这进一步验证了使用 IODP 为场景识别过程提供先验知识的鲁棒性和泛化性能。此外，不论采用哪种衡量方式，对结果进行平方根转换均能够进一步提高识别精度。这主要是因为，在统计物体分布的过程中，忽略了场景中的一些细节信息（如颜色），可能导致计算出的物体间鉴别性关系比实际情况更尖锐，平方根转换可以有效缓和这一问题。值得注意的是，在大多数数据集上，

$\sqrt{cv}$ 取得了最佳的识别性能。因此，在本章后续的所有实验中，均使用 $\sqrt{cv}$ 衡量物体间鉴别性相关程度。

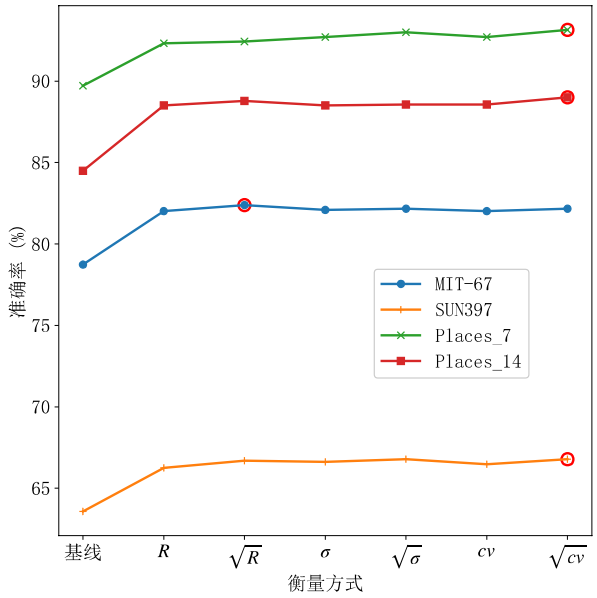


图 4-7 不同的物体间鉴别性相关程度的衡量方式的效果

Fig. 4-7 Effect of using different ways to measure inter-object discriminative correlation

注：用空心粗体圆圈圈出的精度最高的节点。

4.3.3 超参数选择

GCN 中的两个超参数，即隐藏维度 d 和辅助损失中的 λ，对本章方法的性能有着重要影响。本小节以 ConvNext_Tiny 作为骨干网络，通过一系列实验确定这些超参数的最优值。

首先探究不同的隐藏维度 d 对模型识别精度的影响。为便于比较，此时 λ 被固定为 1。图 4-8 展示了不同 d 值对识别精度产生的影响，其中最高准确率点被一个空心粗圈标出。在图 4-8 中，In_Dim 指由骨干网络提取的特征图的通道维度（对于 ConvNext_Tiny，In_Dim=768；对于 ConvNext_Base，In_Dim=1024）。

如图 4-8 所示，一个普遍趋势是：随着隐藏维度的增加，识别准确率先是逐渐提高，当隐藏维度增长到 In_Dim 后，准确率便开始下降。这一现象可能是由于过高的隐藏维度导致了模型过拟合。基于上述实验结果，将隐藏维度 d 设置为 In_Dim，保持参数调整灵活性的同时，避免过拟合的风险。

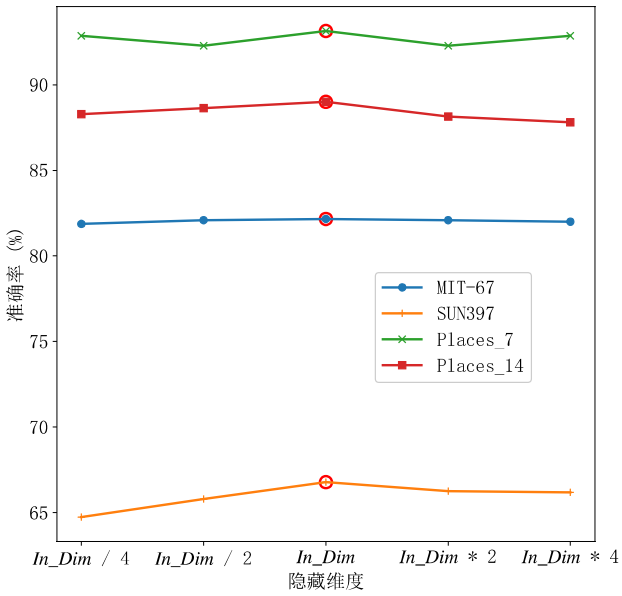


图 4-8 GCN 中不同的隐藏维度 d 的影响

Fig. 4-8 Effect of hidden dimension d in GCN

注：用空心粗体圆圈圈出的精度最高的节点。

除了隐藏维度 d，辅助损失中的超参数 λ 同样十分关键。如章节 4.2.3 所述，虽然引入 IODP 中的先验知识能够提升网络的整体性能，但也可能对骨干网络的鉴别性能力的优化产生消极影响。为缓解这一问题，在训练过程中引入了辅助损失，以维持骨干网络固有的鉴别性能力。表 4-2 展示了不同的 λ 对识别精度的影响。结果显示，在数据集 MIT-67、Places_7 和 Places_14 上，当 λ 设置为 0.25 时，模型达到最佳性能；而在数据集 SUN397 上，当 λ 设置为 0.5 时，模型性能最优。

表 4-2 结合辅助损失时 λ 对分类精度的影响（%）

Tab. 4-2 Influence of λ on classification accuracy (%) when combined with auxiliary loss

λ	0	0.25	0.5	1	2
MIT-67	82.090	**82.239**	81.791	82.164	82.09
SUN397	66.741	66.753	**66.929**	66.776	66.506
Places_7	92.286	**93.143**	92.143	93.143	93.086
Places_14	88.357	**89.071**	88.5	89	88.714

上述实验确定了最优超参数配置。具体如下：对于 MIT-67、Places_7 和 Places_14 数据集，$\{d = In_Dim, \lambda = 0.25\}$；对于 SUN397 数据集，$\{d = In_Dim, \lambda = 0.5\}$。这些参数设置将被作为本章节后续实验的默认配置。

4.3.4 网络组成成分分析

本节基于 MIT-67 和 Places_14 数据集，对所提方法进行了全面的消融研究。首先，分析了使用 IODP 向已经训练好的基线模型提供先验知识的影响（仅在测试时使用 IODP）；然后，探讨了辅助损失对模型识别性能的贡献。实验结果如表 4-3 所示。注意，当仅在测试阶段使用 IODP 时，移除了 GCN 层中的隐藏层和激活函数，即将式（4-7）中的$V^* = \eta(D^{-1}\tilde{A}VW)$简化为$V^* = D^{-1}\tilde{A}V$。这样便可将 IODP 中的先验知识即插即用到训练好的基线模型。

表 4-3 在 MIT-67 和 Places_14 数据集上的消融研究（%）

Tab. 4-3 Ablation studies on MIT-67 and Places_14 datasets (%)

基线	训练时使用 IODP	测试时使用 IODP	辅助损失	MIT-67	Places_14
✓	-	-	-	78.731	84.5
-	-	✓	-	79.478	87.357
-	✓	✓	-	82.090	88.357
-	✓	✓	✓	**82.239**	**89.071**

在表 4-3 中，最佳结果以粗体突出显示，相较于基线模型，分别在 MIT-67 和 Places_14 数据集上实现了 3.508%和 4.571%的性能提升。此外，使用 IODP 向训练好的基线网络提供先验知识（仅在测试时使用 IODP），也能显著提升模型的识别性能，特别是在 Places_14 数据集上的提升幅度达到了 2.857%。值得注意的是，这一成果是在不增加额外训练负担或新增任何网络参数的前提下实现的，充分展示了直接将 IODP 应用于已训练模型的有效性，并有力验证了 IODP 中的鉴别性先验知识的重要性，以及从图（Graph）角度利用这些先验知识的可行性。当在训练和测试阶段同时使用 IODP 时，模型性能获得了进一步的提升。这是因为在训练过程中，隐藏层等可以加强 IODP 中的先验知识与场景特征的磨合，确保了 IODP 先验知识的更充分利用。同时，引入辅助损失进一步改善了网络性能。这一结果与第 4.2.3 节中的分析相符，表明在训练过程中整合 IODP 以增强特征鉴别性确实可能会削弱对骨干网络参数的训练，而所提出的辅助损失确保了对骨干网络参数的有效训练。

4.3.5 特征可视化

为定性分析本章所提出的方法，本小节采用了基于梯度的类激活映射（Gradient-based Class Activation Map，Grad-CAM）[69]技术，以可视化基线模型和集成所提方法的模型的关注区域。可视化结果如图 4-9 所示，从左到右依次为：原始 RGB 图像、对应语义分割结果、基于基线模型的关注区域可视化、仅在测试阶段应用 IODP

的可视化，以及基于完整 DGN 方法的关注区域可视化。

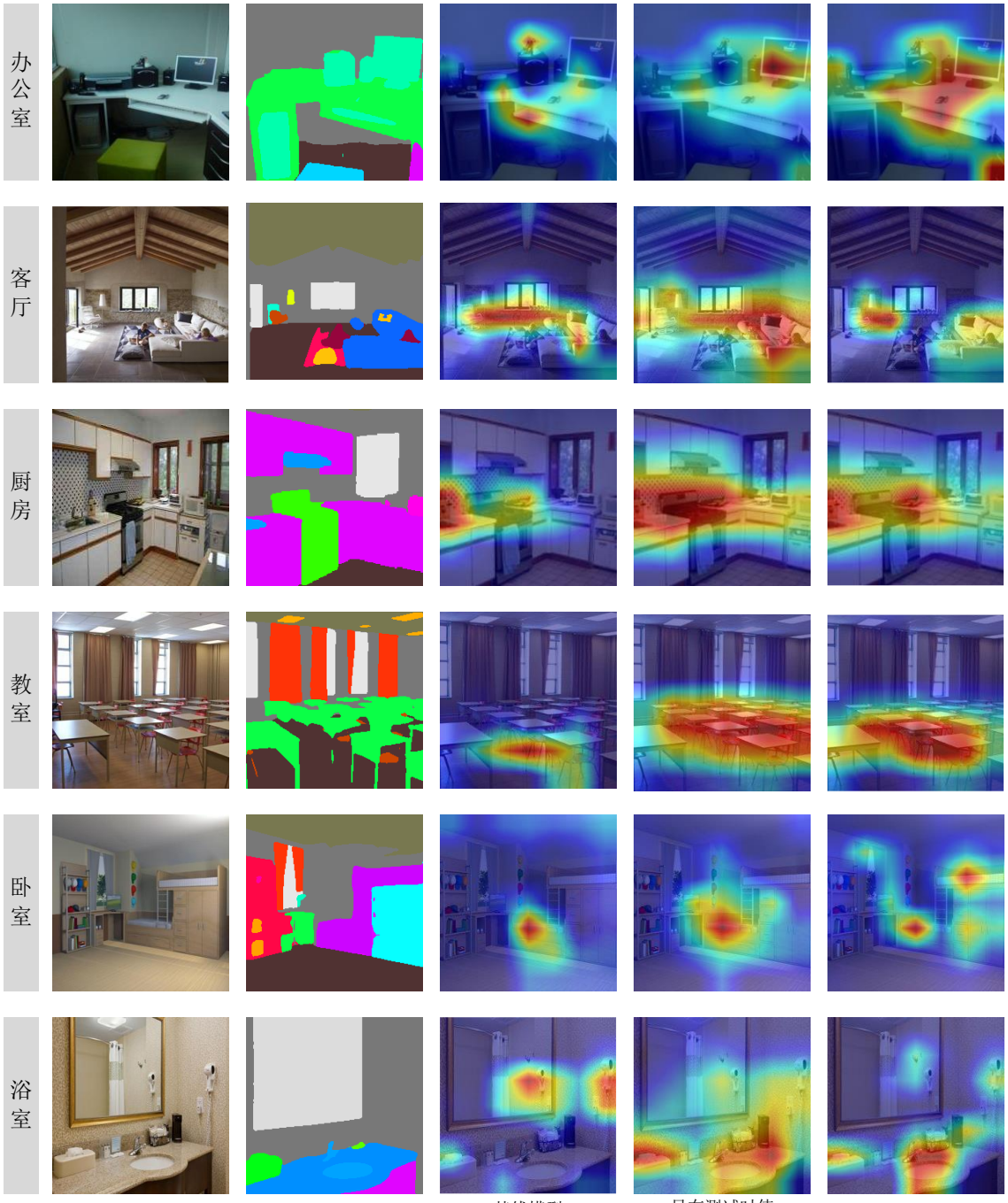


图 4-9 定性可视化结果

Fig. 4-9 Qualitative visualization result

注：网络更强调的区域在可视化结果中呈现出更深的红色。

如图 4-9 所示，1）相比于基线模型的可视化结果，引入 IODP（图中的第 4 列和第 5 列）有助于模型将注意力集中于鉴别性物体相关区域，从而提升了网络的识别性能。2）仅在测试阶段使用 IODP 时（图中第 4 列），基于 IODP 提供的先验知识对基线网络的

关注区域进行了优化调整，获得了更佳的性能表现。但是，这种做法也存在一定的局限性：可能会导致关注区域的过度泛化，使得网络的焦点区域过大，引入了额外噪声，这可能会降低特征的鉴别性能力。这一问题的原因在于，IODP 提供的先验知识仅关注了物体间的鉴别性关系，无法基于颜色等细节因素调整场景特征。相比之下，同时在训练和测试阶段使用 IODP 能够综合考虑先验知识与其他细节信息，通过隐藏层等对网络参数和鉴别性图进行细致调整，进一步优化关注区域的合理性与有效性。

4.3.6 不同 CNN 架构的对比

前述实验以 ConvNext_Tiny 为骨干网络确定了本章算法的设计细节和超参数配置，消融研究和可视化分析也验证了所提方法在场景识别任务中的显著效果。为进一步验证所提算法的有效性，本节采用具有更强大的特征提取能力的 ConvNext_Base 作为骨干网络，并在 MIT-67、SUN397、Places_7 和 Place_14 数据集上进行了性能评估。实验结果如表 4-4 所示。可以看到，得益于 ConvNext_Base 卓越的特征提取能力，其在各个场景数据集上均实现了更高的识别精度。更进一步地，在加入本节算法之后，精度得到了进一步的提升。这进一步验证了 IODP 和 DGN 的有效性和泛化性。最后，在表 4-4 的最后一行，使用标准的 10-crop 测试方法[33,56]评估了所提出的 DGN-Net。需要注意的是，由于本文所使用的 Places 数据集是已经预处理至网络输入的尺寸的版本，10-crop 对网络在 Places 数据集上的评估没有影响。

表 4-4 基于不同的骨干网络在多个数据集上的验证（%）

Tab. 4-4 Validation of our method on several datasets based on different models (%)

骨干网络	方法	MIT-67	SUN397	Places_7	Places_14
ConvNext_Tiny	基线	78.731	63.576	89.714	84.5
	DGN-Net	82.239	66.929	93.143	89.071
ConvNext_Base	基线	87.388	75.424	93.857	88.957
	DGN-Net	89.254	78.671	94.286	89.914
	DGN-Net (10-crop)	90.373	79.765	94.286	89.914

4.3.7 与现有算法的性能比较

在 MIT-67、SUN397、Places_7 和 Places_14 这四个数据集上，本节将 DGN-Net 与现有的先进算法进行了比较。这些对比的算法既包括基于通用深度学习架构的方法，也包括利用物体信息辅助场景识别的方法。表 4-5 和表 4-6 展示了比较结果。

表 4-5 在 MIT-67 和 SUN397 数据集上 DGN-Net 与现有先进算法的比较（%）

Tab. 4-5 Comparison of DGN-Net with the state-of-the-art methods on MIT-67 and SUN397 (%)

算法	输入尺寸	MIT-67	SUN397
Adi-Red [27]	224×224	-	73.59
LGN [28]	448×448	88.06	74.06
Scene Essence [10]	224×224	83.92	68.31
MR-Net[59]	448×448	88.08	73.98
SDO-Net[32]	224×224	86.76	73.41
SAS-Net [33]	224×224	87.10	74.04
CCF-Net [38]	512×512	87.30	-
ARG-Net [42]	448×448	88.13	75.02
PL + AP + AI + IM[58]	448×448	88.06	74.12
FCT-Net[64]	224×224	89.17	76.06
CSSRM	224×224	88.73	75.48
SpaCoNet	224×224	90.75	76.15
DGN-Net	224×224	**90.37**	**79.77**

注：CSSRM、SpaCoNet 分别为本文第二章、第三章所提出的算法。

表 4-6 在 Places_14 和 Places_7 数据集上 DGN-Net 与现有先进算法的比较（%）

Tab. 4-6 Comparison of DGN-Net with the state-of-the-art methods on Places_14 and Places_7 (%)

算法	输入尺寸	Places_14	Places_7
Word2Vec[61]	224×224	83.70	-
Deduce[39]	224×224	-	88.10
BORM[45]	224×224	85.80	90.10
OTS-Net[52]	224×224	85.90	90.10
AGCN[62]	224×224	86.00	91.70
CSSRM	224×224	88.71	93.43
SpaCoNet	224×224	89.71	94.29
DGN-Net	224×224	**89.91**	**94.29**

注：CSSRM、SpaCoNet 分别为本文第二章、第三章所提出的算法。

如表 4-5 和表 4-6 所示，所提出的 DGN-Net 优于大多数现有的算法。首先，相比于基于卷积响应进行特征选择的方法[27,28,38,59,62]，DGN 展现了更高的识别准确率，这验证了使用具体的物体信息提高特征鉴别性能力的实质可行性。接下来，将焦点转移向基于

物体辅助的方法。方法[10,32,33,39,45]通过多分支架构分别处理场景图像和物体信息，并通过级联或求和等启发式手段将场景特征和物体特征结合，以获得场景表示。与这些方法相比，DGN-Net 采用单分支架构，从推理角度利用物体知识优化场景特征，不仅提高了算法的可解释性，还避免了多分支架构可能导致的计算负担。进一步地，相比于其他同样采用图卷积技术辅助场景识别的算法[10,42,62]，DGN-Net 通过将鉴别性先验知识集成到图架构，实现了更优越的性能。这进一步凸显了挖掘物体间鉴别性先验知识的重要性。

此外，本文第二章算法 CSSRM 和第三章算法 SpaCoNet 从启发式角度出发，利用多分支架构处理物体信息。与它们相比，本章算法从知识推理的角度入手，利用物体信息中的隐藏知识提高场景特征的鉴别性能力，不但提高了可解释性，而且在单分支架构下就取得了与 CSSRM 和 SpaCoNet 相当甚至更优的场景识别性能。尽管在 MIT-67 数据集上，SpaCoNet 的准确率比 DGN-Net 高出 0.373%，但考虑到 SpaCoNet 依赖于多分支架构，DGN-Net 在轻量化方面表现更佳。此外，DGN-Net 在 SUN397 和 Places_14 数据集上的准确率分别要比 SpaCoNet 高出 3.612%和 0.2%，展示出了更强大的泛化能力。

值得一提的是，尽管方法[42,52]也采用单分支架构利用物体信息，但它们主要关注于物体信息的像素级位置，忽略了隐藏在物体信息中的鉴别性知识，导致性能提升有限。此外，方法[32,45]也采取了类似的统计策略挖掘物体之间的鉴别性关联，但在这些方法中，物体知识仅用于优化处理物体信息的分支特征，对物体知识的利用相对有限。这也验证了物体间鉴别性知识在场景识别的重要作用，而 DGN 能够有效利用这一知识，将场景特征转化为更具鉴别性的图像表示。

4.4 本章小结

本章提出从概率角度捕捉物体与场景之间的鉴别性关系，并将其转化为物体间鉴别性原型（IODP）。IODP 中蕴含的丰富的物体与物体之间的鉴别性知识，可被用于辅助场景识别网络提高性能。具体来讲，本研究通过构建一个鉴别性图网络（DGN）实现鉴别性辅助过程。DGN 以像素级场景特征作为图节点，并将 IODP 中的物体间鉴别性知识转化为像素级特征之间的相关性。通过图卷积和映射操作，将物体间鉴别性知识和场景特征转化为鉴别性图像表示。在 MIT-67、SUN397、Places_7 和 Places_14 数据集上进行的定性和定量实验，验证了物体间鉴别性知识在场景识别中的关键作用，并展示了从图建模角度利用这些知识改善场景特征的有效性和泛化能力。

然而，本章提出的算法，以及几乎所有基于物体信息引导的场景识别算法，都面临一个共同的挑战：物体信息提取技术固有的高计算负担。这严重限制了基于语义引导的场景识别算法在边缘设备上的实际部署能力。针对这一问题，下一章将进一步探讨如何在不增加额外计算负担的前提下，利用语义信息辅助提升场景识别网络的性能。

5 基于语义嵌入的类间相似度原型的场景识别方法

5.1 引言

如图 5-1 所示，由于场景内多变的空间布局带来的类内差异和场景间共存的物体带来的类间相似性，大量研究通过探索场景内的语义知识辅助场景识别。尽管取得了优越的性能，但当在边缘设备实际部署训练好的模型时，这些基于语义辅助的方法需要使用目标检测或语义分割技术获得场景内具体的物体信息（如窗户等）。这些物体信息提取技术往往需要比一般的分类网络高得多的计算成本，不仅对计算资源有限的边缘设备构成了沉重负担，也严重阻碍了语义引导的方法在实际应用中的运行效率。因此，有必要探索对物体语义信息更高效的利用方式，以辅助场景识别而不增加额外的计算需求。

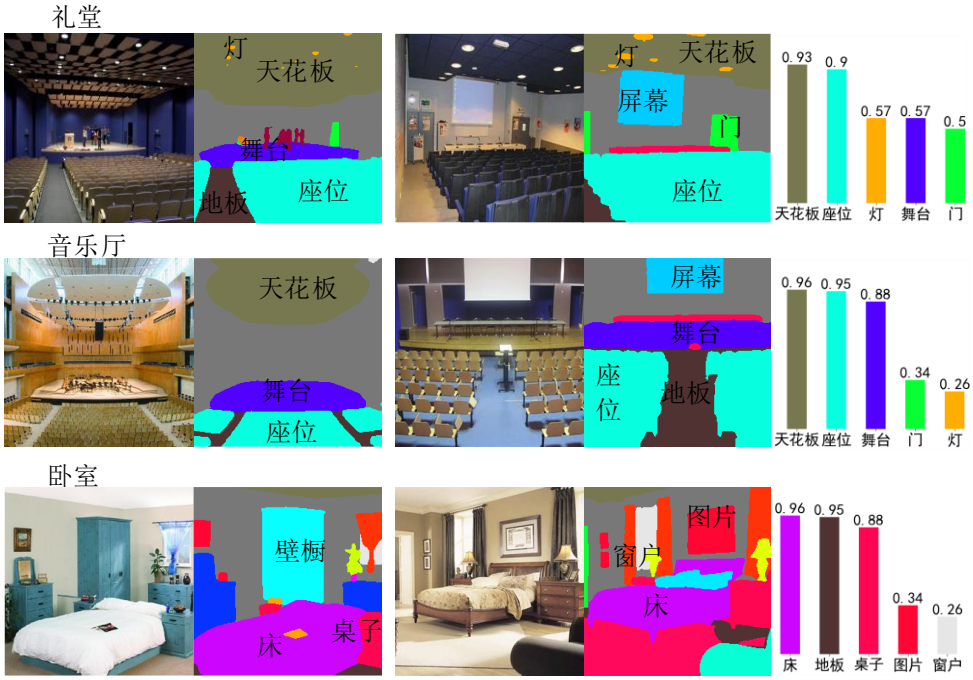


图 5-1 共现物体导致的场景类间相似性

Fig. 5-1 Similarity due to object co-occurrence in scene recognition

注：右侧的柱状图展现的是该场景类中出现频率最高的五个物体。

基于物体信息的引导探索场景内的特征上下文关系这一传统做法[27,28,33,38,42]显然不再适用于上述目的。因此，本研究将关注点从场景内的物体区域间关系转移到了场景间的宏观相关性上。这一转变的动机启发于一个人类视角：不同类别之间的高相似性是场景识别任务的一个挑战，但并非所有的场景类别之间都高度相似。以图 5-1 为例，虽然

场景类“礼堂”与“音乐厅”的相似度很高，但这它们与“卧室”的差异却非常明显。因此，人类会有意识地花费更多精力区分那些更相似的场景，从而学习场景类别。受此启发，本研究认为，如果在训练过程中网络能够了解到不同类别间的相似度，就能够更加专注于区分那些相似的场景类别，从而找到更合适的拟合方向。此外，研究表明骨干网络由于其参数的冗余性[76]，拥有巨大的性能提升潜力，但这需要精细的训练指导来激发。有关场景本质的研究[10]指出，不同场景之间的相似性往往是其内部共存的物体引起的。因此，本章提出将物体语义知识嵌入到一个相似度原型中，该原型可以在训练阶段为网络提供类间相似度的先验知识，有效提高拟合后网络的识别精度。

具体来讲，本章基于语义分割和统计分析，捕获训练数据中物体语义信息的分布情况，生成类级语义表示。这些表示被用于量化场景类别间的相似度，最终构建出一个类间语义相似度原型。随后，本章从两个角度出发，利用此原型辅助模型训练：标签平滑（Label smoothing）和对比损失（Contrastive loss）。

标签平滑[77]是一种正则化策略，通过使用软化的标签替代硬标签，减轻模型对于预测的过度自信。不过，该策略对所有非目标类别一视同仁，忽略了它们与目标类别的相关性差异。对比损失[78]是一种度量学习策略，通过建立度量函数来比较样本对以训练网络，旨在使特征嵌入空间中相同类别的样本更接近，不同类别的样本更远离。然而，类似的一个主要局限是，在该策略中所有正对和负对之间的距离都是相等的，这种僵硬的边界设置无法反映出相似类别和不相似类别样本之间的真正差异。

所提出的相似度原型为上述问题提出了解决办法。从标签平滑角度，本章提出了一种渐进标签软化（Gradient Label Softening，GLS）方法，将相似度原型中的知识嵌入到软化的标签中，并通过置信度渐变策略在训练过程中逐步引导模型的拟合方向。从对比损失角度，受工作[78-80]的启发，本章提出了一种即插即用的批量级对比损失（Batch-level Contrastive Loss，BCL）策略，利用相似度原型调整类间相似度的阈值，并进一步将此策略扩展到类内对比损失，确保在训练阶段对语义先验知识的充分利用。本章在 MIT-67、SUN397、Places_7 和 Places_14 数据集上评估了所提方法，在多个基准上进行的综合评估表明，本章方法在不增加额外训练负担或网络参数的前提下，可以使现有的先进方法更进一步，取得更优越的性能。

5.2 基于语义嵌入的类间相似度原型的场景识别方法

5.2.1 总体概述

本章节首先借助语义分割技术、训练数据和统计学方法为每个场景类别生成语义表示，随后利用这些类级语义表示构建了一个相似度原型，其中包含了丰富的类间相似度知识。在此基础上，提出了两种利用相似度原型辅助网络训练的方法：渐进标签软化

（GLS）和批量级对比损失（BCL）。通过上述训练过程，使网络在不增加任何额外参数或计算负担的情况下即可获得更优越的场景识别性能。

这种即插即用的技术可以轻易融入到大部分的场景表示学习框架中，只需要简短几行代码和很小的计算开销即可实现。

5.2.2 基于语义嵌入的相似度原型

相似度原型的推理过程可以分为两步：语义表示和标签相关性。如图 5-2 所示，首先使用语义知识为每个场景类生成语义表示，随后以这些类级语义表示为基础探索类间标签相关性，这些相关性被嵌入到所提出的相似度原型中。

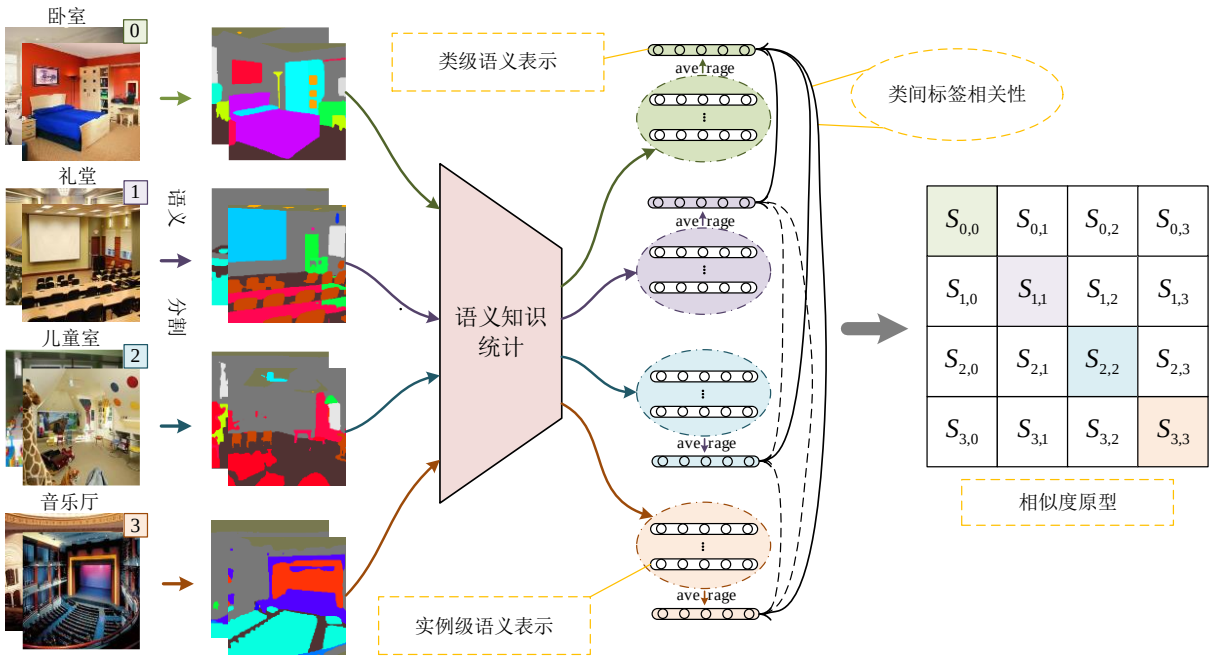


图 5-2 为场景数据集制作相似度原型的整体过程概述

Fig. 5-2 Process illustration of making a similarity prototype for a scene dataset

注 1：不同的颜色块表示不同的场景类别。

注 2：推导出的相似度原型呈现为一个维度等于场景类别数的矩阵，该矩阵的对角线元素均为 1。

（1）类级语义表示

给定一个场景数据集，将其训练数据表示为 $D_T=\{(X_c,y_c)\}_{c=1}^{C}$，其中 X_c 代表第 c_{th} 个场景类的所有训练实例，$y_c\in\{1,2,...,C\}$ 是 X_c 的场景标签。对每一个场景类 X_c，可进一步将之表示为 $X_c=\{x_c^n,y_c\}_{n=1}^{N}$，其中 $x_c^n\in\mathbb{R}^{W\times H\times 3}$ 表示第 c_{th} 个场景类的第 n_{th} 个训练实例。为获得类级语义表示，本节整合了 X_c 的所有实例级表示。

算法 5-1 展示了生成类级语义表示的伪代码。针对每一个场景实例 x_c^n，首先将其输入到语义分割网络，获得相应的语义分割标签图 $M_c^n \in \mathbb{R}^{W\times H}$。$M_c^n$ 表示 x_c^n 中每个像素点所对应的物体语义标签，可以表示为 $M_c^n = \left\{M_c^n(w,h)\right\}_{\substack{1\le w\le W\\ 1\le h\le H}}$。$M_c^n$ 随后被用于推理 x_c^n 的实例级语义表示 $s_c^n \in \mathbb{R}^L$。L 代表语义分割技术可分割的物体类别的数量（$L=150$）。更具体地说，对每个 $l \in [1, L]$，如果存在 (w,h) 使 $M_c^n(w,h)=l$，那么 $s_c^n(l)=1$；相反，如果对任何 (w,h) 都 $M_c^n(w,h)\neq l$，那么 $s_c^n(l)=0$。执行上述过程 L 次，即可得到 x_c^n 的实例级语义表示 s_c^n。

算法 5-1：类级语义表示生成过程

Data： L 代表语义分割可分割的物体类别的数量
Input： 训练实例 $x_c^n, c=1,...C; n=1,...,N$
Output： 类级语义表示 $S_c \in \mathbb{R}^L$

```
for c = 1 to C do
    for n = 1 to N do
        实例级语义表示 s_c^n ∈ R^L 初始化为一个零向量;
        M_c^n = Semantic Segmentation(x_c^n)
        for l = 1 to L do
            if l in M_c^n then
                s_c^n(l) = 1
            end if
        end for
    end for
    S_c = (1/N) Σ_{n=1}^{N} s_c^n
end for
return S_c;
```

通过重复上述过程 N 次，即可获得第 c_{th} 个场景类 X_c 中所有的实例级语义表示，以此为基础，即可计算 X_c 的类级语义表示 $S_c \in \mathbb{R}^L$：

$$S_c = \frac{1}{N}\sum_{n=1}^{N} s_c^n \tag{5-1}$$

其中 N 表示第 c_{th} 个场景类中训练实例的数量。

循环上述过程C次，即可获得所有的类级语义表示。

接下来，本节将从余弦相似度和欧式距离两个角度，度量不同场景类别之间的标签相关性。使用$S_{i,j}$表示第i_{th}和第j_{th}个场景类别之间的标签相关性，下面对其生成过程进行详细描述。

（2）基于余弦相似度的类间标签相关性

余弦相似度（Cosine similarity），通过计算两个向量之间夹角的余弦值捕捉二者之间的相似程度。在本文中，它被用来量化两个类级语义表示之间的标签相关性，其计算方式如下：

$$S_{i,j} = CosineSimilarity\left(S_i, S_j\right) = \frac{S_i \bullet S_j}{\left\|S_i\right\|\left\|S_j\right\|} \tag{5-2}$$

其中$S_i \bullet S_j$表示S_i和S_j的内积，$\left\|S_i\right\|$和$\left\|S_j\right\|$表示它们各自的欧式范数。

$S_{i,j}$反映了第i_{th}和第j_{th}个场景类别之间的标签相关性。$S_{i,j}$的值越大，表示相关性越强，意味着这两个类之间的相似度越高，越难区分；反之，数值越小，相关性越弱，表明二者更容易区分。

（3）基于欧式距离的类间标签相关性

除了余弦相似度之外，本节还尝试了欧氏距离（Euclidean distance）衡量类间标签相关性。欧氏距离通过计算两个语义表示之间的几何距离量化它们之间的差异程度，计算方式为：

$$EuclideanDistance(S_i, S_j) = \sqrt{\sum_{l=1}^{L}\left(S_{il} - S_{jl}\right)^2} \tag{5-3}$$

其中$L=150$表示语义表示S_i和S_j的维度，S_{il}和S_{jl}分别表示S_i和S_j在第l_{th}维的值。

式（5-3）衡量的是两个类级语义表示之间的差异程度。为确保度量出的标签相关性与类间相似性正相关，本节使用指数函数对欧氏距离进行转换，从而得到标签相关性$S_{i,j}$：

$$S_{i,j} = \exp\left(-EuclideanDistance\left(S_i, S_j\right)\right) \tag{5-4}$$

经过上述过程，$S_{i,j}$的值越高，表示相关性越强；而值越低，表示相关性越弱。

根据上述两个衡量结果，可以构建两个相似度原型$S \in \mathbb{R}^{C \times C}$，其中$S_{i,j}$表示第$i_{th}$和第$j_{th}$个场景类之间的标签相关性。图 5-3 展示了针对 Places_14 数据集的基于余弦的相似度原型。可以看到，“卧室”与“客厅”的相关性比与“厨房”的相关性更紧密，而“厨房”与“餐厅”的相关性比与“卧室”的相关性更紧密。这些现象与人类常识相契合，足够作为网络的先验知识。

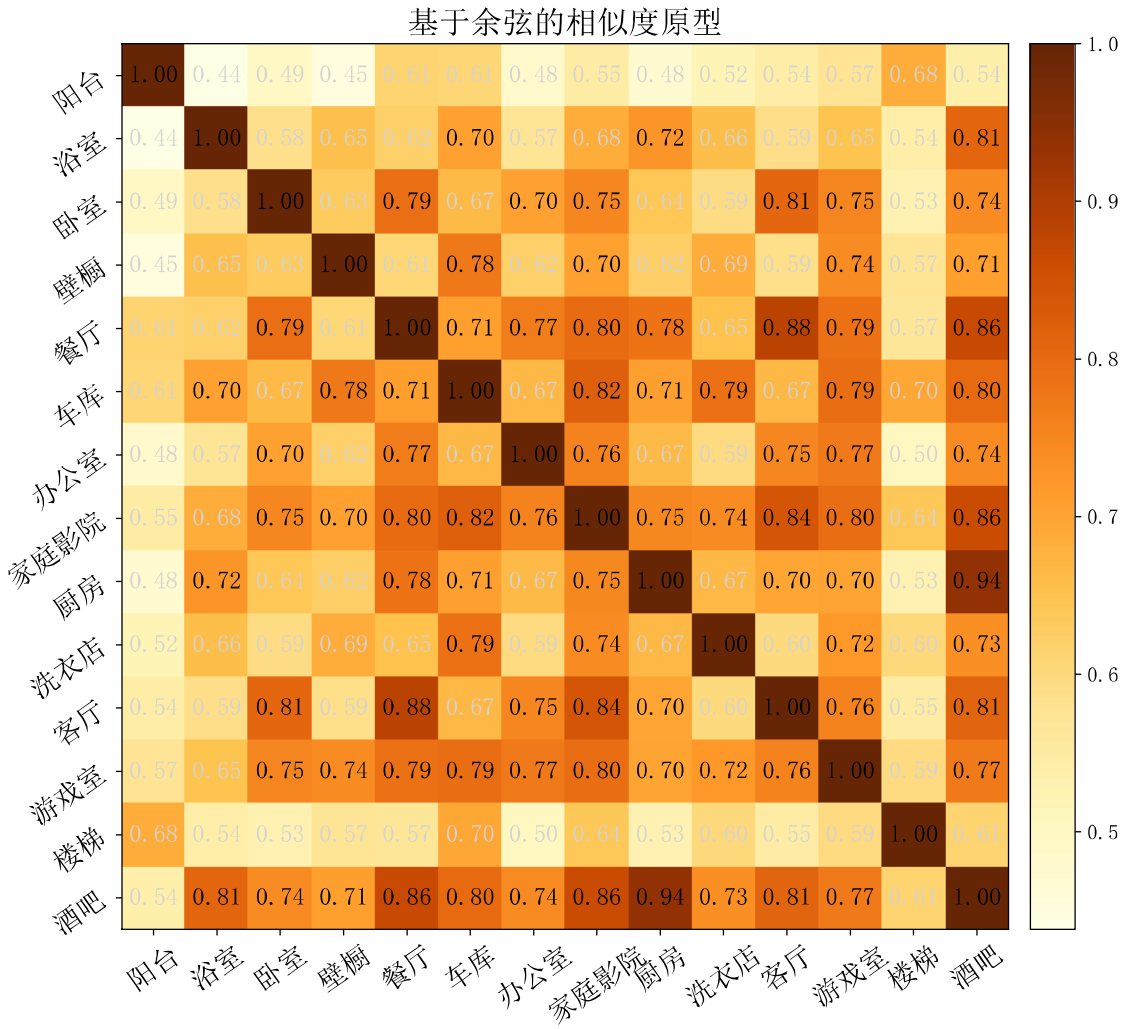


图 5-3 针对 Places_14 数据集的基于余弦的相似度原型

Fig. 5-3 Cosine-based similarity prototype for Places_14 dataset

注：越深的颜色表示相应的场景类别之间的相似度越强。

上述过程成功将场景类之间的相似度相关性嵌入到了相似度原型中。下面将从两个角度介绍如何在不增加任何参数的前提下利用相似度原型辅助网络训练。

5.2.3 基于相似度原型的渐进标签软化

本小节从标签软化角度利用相似度原型辅助模型训练。与标签平滑正则化（Label Smoothing Regularization，LSR）[77]使用人为定义的平滑值软化标签不同，有了相似度原型提供先验知识，本节根据不同场景类别之间的相似度量化标签软化的程度。

具体来讲，给定一个有C个场景类的数据集，通过章节 5.2.2 可以得到其相似度原型$S \in \mathbb{R}^{C\times C}$。首先对$S$的进行行归一化得到可用的软标签形式$S_{norm}$：

$$S_{norm} = S \oslash \left(S \bullet E_{C \times C} \right) \tag{5-5}$$

其中$\oslash$表示元素级除法，$E_{C \times C}$表示维度为$C \times C$的全一矩阵。

然而，S_{norm}中不同场景类的软标签的目标类别置信度不同，这可能导致模型训练过程中的类不平衡问题。因此，本研究进一步精细调整了S_{norm}，确保每个类别对应软标签的目标类别置信度一致。通过修改S的对角线元素实现这一点。

具体来讲，给定$S_{c,c}$，其软标签的目标类别置信度σ的计算过程为：

$$\sigma = \frac{S_{c,c}}{\sum_{i=1}^{C} S_{i,c}} = \frac{S_{c,c}}{S_{c,c} + \sum_{i=1, i \neq c}^{C} S_{i,c}} \tag{5-6}$$

对式（5-6）进行转换，则给定目标类别置信度σ'，其相应的$S_{c,c}$应赋值为：

$$S_{c,c} = \frac{\sigma'}{1 - \sigma'} \sum_{i=1, i \neq c}^{C} S_{i,c} \tag{5-7}$$

接下来，采用S_{norm}中的最大的目标类别置信度统一作为所有类别的置信度，更新S得到S^1：

$$\begin{aligned} value_start &= \max\left(S_{norm} \right) \\ S_{i,j}^{1} &= \begin{cases} \dfrac{value_start}{1 - value_start} \sum_{i=1, i \neq c}^{C} S_{i,j} & i = j \\ S_{i,j} & i \neq j \end{cases} \end{aligned} \tag{5-8}$$

随后，再次利用式（5-5），即可得到可用的归一化矩阵S_{norm}^{1}。在S_{norm}^{1}中，所有的类别标签拥有一致的目标类别置信度，同时目标类别和非目标类别之间也保持着合理的置信度差异。

鉴于相似度原型的推理过程忽略了场景中的颜色等细节信息，直观上讲，训练好的网络模型区分不同场景类别的能力应优于纯粹的知识统计（相似度原型）。因此，在训练过程中逐渐提高目标类别的置信度，驱使网络朝着合适的方向逐步提高其鉴别性能力。在这个过程中，仍然基于相似度原型保持场景类别之间的置信度差异，确保网络对语义先验知识的充分利用。在*STEP*次迭代（Epoch）后，所有的类标签又变为了硬标签。上述过程的伪代码如算法 5-2 所示。同时，图 5-4 以 Places_7 数据集为例，展示了其S_{norm}^{1}随着迭代次数的变化。

随着迭代次数的变化，S_{norm}^{1}所提供的软标签将与交叉熵损失函数一起用于模型训练：

$$\mathcal{L}_{CE} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{C} \left(S_{norm}^{1}\left[j, target[i] \right] \cdot \log\left(p\left[i, j \right] \right) \right) \tag{5-9}$$

其中B表示一个训练批次中的图像数量，p表示网络输出的预测概率，*target*表示这个训练批次的标签。

算法 5-2：渐进标签软化算法

Input： 相似度原型 S

Output： 训练标签

for $epoch = 1$ to End_Epoch **do**

  $value = value_start + \left(0.99 - value_start\right) \cdot \dfrac{epoch - 1}{STEP}$

  $S^1 = S.copy()$

  for $c = 1$ to C **do**

    $S^1_{c,c} = \dfrac{value}{1 - value} \sum_{i=1, i \neq c}^{C} S_{i,c}$

  end for

  $S^1_{norm} = S^1 \oslash \left(S^1 \cdot E_{C \times C} \right)$ # 软标签

  if $value > 0.99$ **then**

    $S^1_{norm} = I_{C \times C}$ # 硬标签

  end if

  return S^1_{norm} # 训练标签

end for

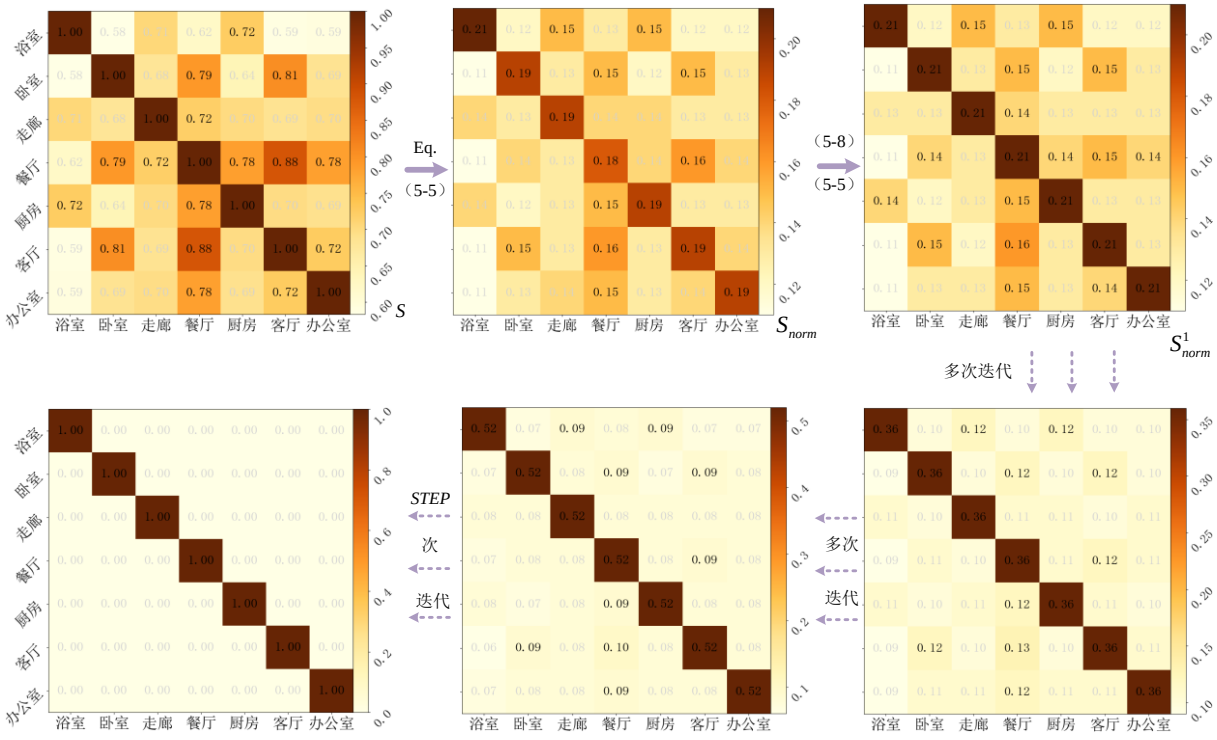


图 5-4 以 Places_7 数据集的基于余弦的相似度原型为例，S^1_{norm} 随迭代次数而变化

Fig. 5-4 Taking cosine-based similarity prototype for Places_7 as an example, S^1_{norm} varies with epoch

5.2.4 基于相似度原型的批量级对比损失

本小节从度量学习（对比损失）角度利用相似度原型辅助模型训练。对比损失致力于通过最大化同一类别内的相似度和最小化不同类别间的相似度来优化特征空间。通过将相似度原型整合到对比损失中，可摒弃手动定义阈值的需求，转而依据统计得到的相似度原型设定更合适的边界。图 5-5 直观可视化了所提出的批量级对比损失（BCL）的操作。给定一对输入 x_i^m 和 x_j^n，对比损失 $\mathcal{L}_{contrastive}$ 可以计算为：

$$\mathcal{L}_{\text{contrastive}} = (1-\gamma)\max\left\{0, p_{x_i^m,x_j^n} - S_{i,j}\right\} + \gamma\max\left\{0, \max\left\{S_{i,j}, j \neq i\right\} - p_{x_i^m,x_j^n}\right\} \tag{5-10}$$

其中 $p_{x_i^m,x_j^n}$ 表示特征 x_i^m 和 x_j^n 之间的相似度。当一对输入属于场景类别时，$\gamma=1$；否则，$\gamma=0$。

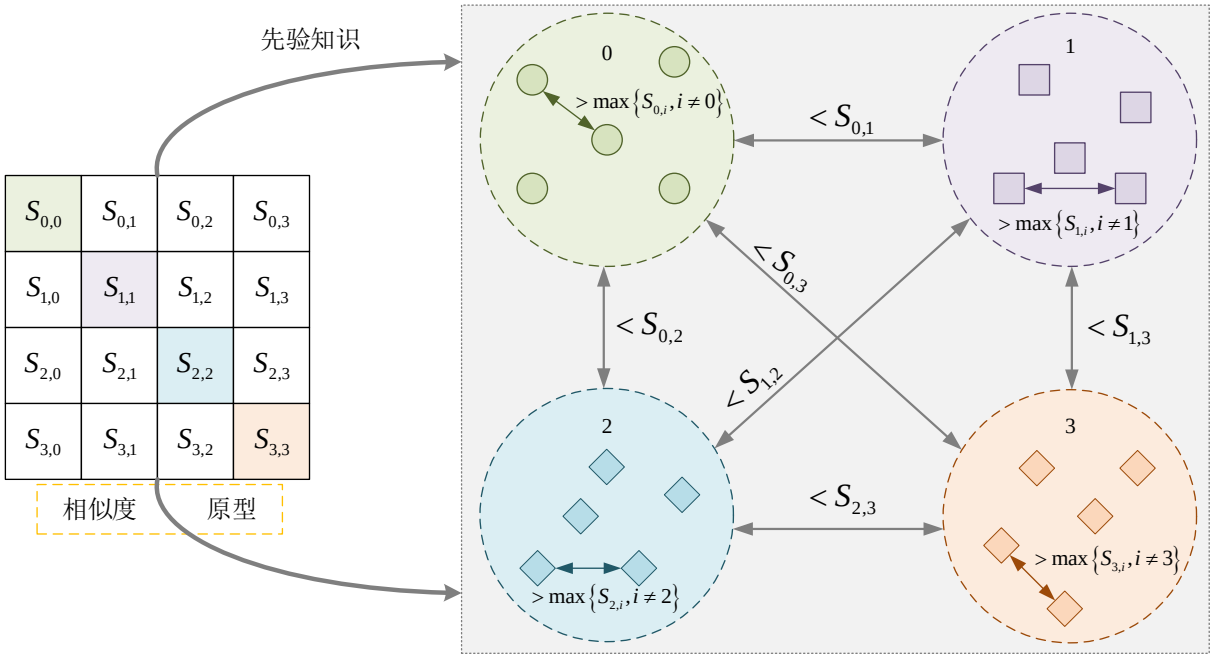


图 5-5 基于相似度原型引导下的对比损失函数的操作图例

Fig. 5-5 Illustration of Contrastive Loss guided by Similarity Prototype

注：不同的颜色块代表不同的场景类别。

为确保本策略的高度即插即用性，启发于 N-pair loss[80]和 Structured loss[81]，本章基于迷你批次（Mini-batch）计算对比损失。更具体地说，假定网络输入的迷你批次包含 B 个随机采样的场景图像，将网络输出预测记为 $p \in \mathbb{R}^{B\times C}$，实际标签记为 $target \in \mathbb{R}^{B}$。使用 $target$ 作为相似度原型的行索引，可以得到 $target$ 中 B 个样本之间的相似度原型变体 $S_{batch} \in \mathbb{R}^{B\times B}$：

$$S_{batch} = S[target] \bullet S[target]^T \tag{5-11}$$

随后，使用余弦相似度或者欧式距离计算网络的输出 p 中 B 个样本之间的相似度矩阵 $p_{matrix} \in \mathbb{R}^{B \times B}$。

鉴于相似度原型的推理过程忽略了场景中的颜色等细节，训练好的网络的鉴别性能力应强于统计推理出的相似度原型。因此，网络预测出的场景类之间的相似度应低于推导出的类间相似度。相应地，可以设计出一个类间对比损失矩阵 $\mathcal{L}_{inter_matrix}$：

$$\mathcal{L}_{inter_matrix} = p_{matrix} - S_{batch} \tag{5-12}$$

鉴于只需要考虑异常情况，因此类间对比损失 $\mathcal{L}_{inter}$ 可以计算为：

$$\mathcal{L}_{inter} = Mean\left(\max\left(\mathcal{L}_{inter_matrix}, 0\right)\right) \tag{5-13}$$

然而，上述过程只考虑了不同场景类别的样本，未能充分利用迷你批中同一类别样本的特征。这是因为同一类样本之间的相似度无法通过相似度原型进行有效衡量，而同一类别的不同样本之间的相似度要达到“1”也是不现实的。然而，在场景识别任务中，只需要确保模型输出的同一类别样本之间的相似度高于其与其他类别样本之间的相似度（如图 5-5 所示）。因此，本节在对比损失中加入了一层类内对比机制，以充分利用来自同一类样本的特征信息。

为计算类内对比损失，首先生成了一个相似度原型 S 的变体：自相似度矩阵 S''。该矩阵只关注类内相似度，它将类内相似度阈值设置为相似度原型中目标场景类与所有其他场景类之间相似度的最大值：

$$S''_{i,j} = \begin{cases} \max\left\{S_{i,k}, i \neq k\right\} & i = j \\ 0 & i \neq j \end{cases} \tag{5-14}$$

类似地，通过式（5-11）可以得到 $target$ 中 B 个样本之间的类内相似度原型变体 $S''_{batch} \in \mathbb{R}^{B \times B}$。随后，可以计算类内对比损失 $\mathcal{L}_{intra}$：

$$\begin{aligned} \mathcal{L}_{intra_matrix} &= S''_{batch} - p_{matrix} \\ \mathcal{L}_{intra} &= Mean\left(\max\left(\mathcal{L}_{intra_matrix}, 0\right)\right) \end{aligned} \tag{5-15}$$

至于度量类内对比损失是否有效，本章将在第 5.3.2 节中进行对比实验验证这一点。

当然，对比损失起到的毕竟是辅助作用，最后的损失由交叉熵损失和对比损失结合实现：

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B} \log\left(\frac{\exp\left(p_{i,target_i}\right)}{\sum_{j=1}^{C}\exp\left(p_{i,j}\right)}\right) + \left(\mathcal{L}_{inter} + \mathcal{L}_{intra}\right) \tag{5-16}$$

5.3 实验结果与分析

本节对所提出的相似度原型的有效性进行了实验验证。首先，在 MIT-67 和 SUN397 数据集上分别评估了基于相似度原型的渐进标签软化（GLS）和批量级对比损失（BCL）训练策略；然后，在 Places_7 和 Places_14 数据集上进行了可视化分析；最后，将所提出的方法与现有的先进算法进行了比较。

5.3.1 实验设置

本章针对场景识别，使用 MIT-67、SUN397、Places_7 和 Places_14 数据集验证所提算法。

（1）语义分割

与前面的章节一致，语义分割过程由在 ADE20K 数据集[49]上预训练的 Vision Tranformer Adapter[48]进行。语义分割网络输出一个标签图 M，其分辨率与输入图像一致。M 中的每个像素 (w,h) 都被赋予一个值 M_{wh}，代表输入图像中对应位置的物体语义标签。

（2）训练细节和超参数

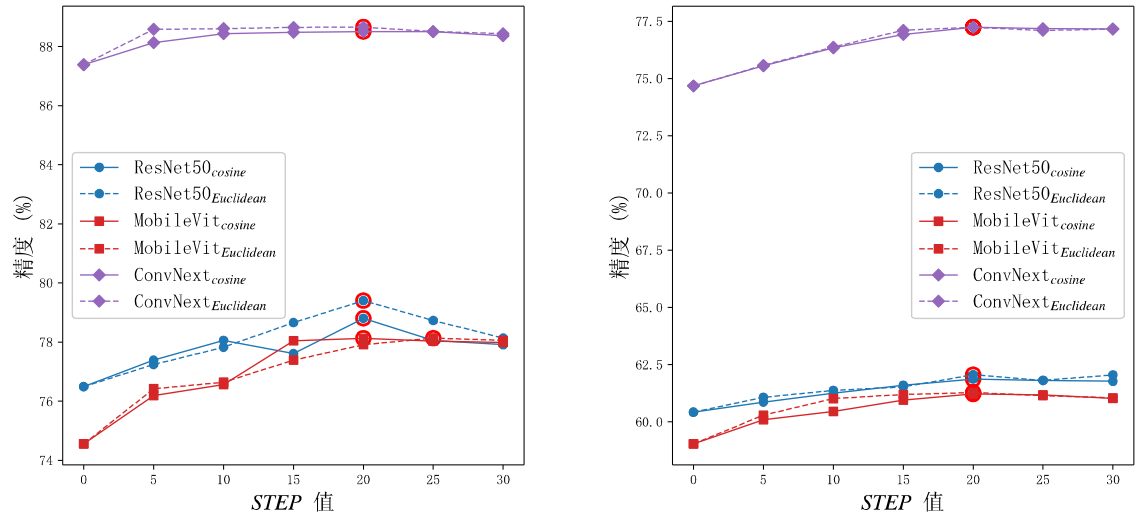
所提出的相似度原型及其衍生策略可轻松集成到大量现有模型中。本章节使用了 8 个广泛使用的预训练模型验证所提算法的有效性和泛化性能：ResNet50-IN1k[22]、VGG16-IN1k[21]、MobileNetV3-IN1k[82]、ShuffleNetV2-IN1k[83]、MobileVIT_S-IN1k[84]、ConvNextBase-IN22k[75]、VITBase-IN22k[23]和 SwinTransformerBase-IN22k[85]。这里，后缀“IN1k”指该模型在 ImageNet 1k 数据集[18]上进行了预训练，后缀“IN22k”则表示该模型在 ImageNet 22k 数据集[18]上进行了预训练。

本节对所有模型的训练均使用 Adam 优化器[74]。在 MIT-67 和 SUN397 数据集上进行实验时，模型的最后一个全连接层的初始学习率被设为 0.001，而其他预训练层的学习率则被设为 0.00001，并在第 10、15 和 20 个迭代轮次时将学习率衰减至原来的 0.1 倍。训练的批次大小设置为 32，权重衰减系数设置为 0.00001，对模型进行了 100 个训练周期的迭代。针对 Places 数据集，鉴于其更大的数据规模和更快的收敛速度，批次大小被调整为 64，并将训练周期减少至 30 个。对于在 ImageNet 22k 上预训练的模型，考虑到其较大的参数规模和快速的收敛性，训练周期同样被设置为 30 个。特别地，对于 MobileVIT 模型，由于其独特的架构特性，初始学习率被提高了 10 倍，以增强模型拟合效率。所有实验均在单个 NVIDIA 3090 GPU 上基于 Pytorch[55]和 SAS-Net[33]开源框架完成。

5.3.2 算法框架中不同组件配置的对比

（1）渐进标签软化策略中超参数 *STEP* 的选择

鉴于超参数 *STEP* 在渐进标签软化（GLS）训练策略中的关键作用，本小节基于模型 ResNet、MobileVIT 和 ConvNext，在 MIT-67 和 SUN397 数据集上进行了一系列实验，以比较不同的 *STEP* 配置下模型性能的差异。如图 5-6 所示，基于余弦（Cosine）相似度和基于欧式（Euclidean）距离的相似度原型展现了类似的辅助效果。进一步分析可以发现：（1）无论 *STEP* 取何值，采用相似度原型中的先验知识辅助训练所产生的模型性能始终优于使用硬标签训练的基准模型，这一结果有力验证了将相似度原型嵌入训练标签的可行性。（2）通过系统地调整 *STEP* 的值，可以看到模型精度随 *STEP* 值从 0 增加到 20 时逐渐提升，并在 *STEP*=20时基本达到峰值，之后进一步增加 *STEP* 的值对模型性能的影响微乎其微。因此，为确保一致性，在后续所有模型训练过程中， *STEP* 的值均被设定为 20。



(a) 在 MIT-67 数据集上的评估　　(b) 在 SUN397 数据集上的评估

图 5-6 不同的 *STEP* 值对拟合后模型性能的影响

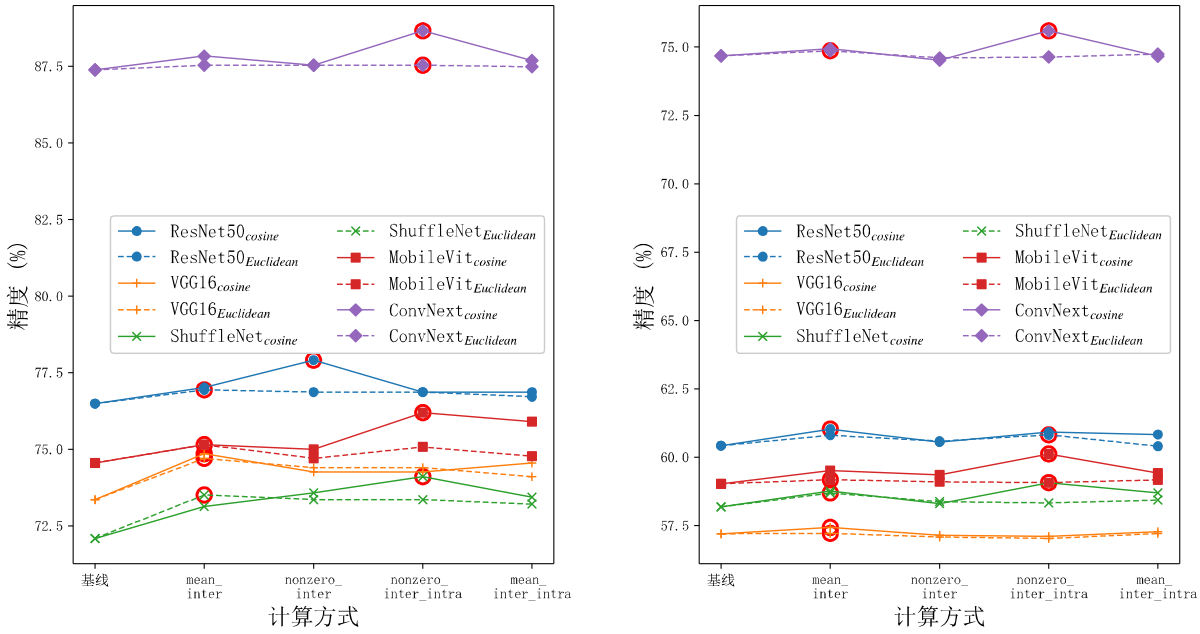
Fig. 5-6 Effect of different *STEP* on the performance of trained model

注：用空心粗体圆圈圈出的节点是最高的精度点。

（2）批量级对比损失策略的计算方式

给定一个对比损失矩阵（详见 5.2.4 节），有两个方式计算最终的对比损失：全局平均（记作“mean”）和非零值平均（即平均过程只考虑矩阵中的非零值的数量，记作“nonzero”）。本小节对比了这两种计算方式对模型性能的影响，并讨论了对类内对比损失（详见 5.2.4 节）的考虑所带来的影响。如图 5-7 所示，使用多个网络模型在 MIT-67 和 SUN397 数据集上进行了对比实验。为便于表示，给定对比损失矩阵 $\mathcal{L}_{inter_matrix}$ 和 $\mathcal{L}_{intra_matrix}$，在图 5-7 中，使用“mean_inter”表示对 $\mathcal{L}_{inter_matrix}$ 执行全局平均以计算对比

损失，“mean_intra_intra”则表示对 $\mathcal{L}_{inter_matrix}+\mathcal{L}_{intra_matrix}$ 执行全局平均；“nonzero_inter”表示对 $\mathcal{L}_{inter_matrix}$ 执行非零值平均，“mean_intra_intra”则表示对 $\mathcal{L}_{inter_matrix}+\mathcal{L}_{intra_matrix}$ 执行非零值平均。



(a) 在 MIT-67 数据集上的评估　　(b) 在 SUN397 数据集上的评估

图 5-7 不同的对比损失计算方式对模型性能的影响

Fig. 5-7 Effect of different computation of contrastive loss on model performance

注：用空心粗体圆圈圈出的节点是最高的精度点。

从图 5-7 可以明显看出，与基线（即传统训练策略）相比，采用相似度原型从对比损失的角度辅助训练可以有效提高模型性能。这一现象再次有力支持了将相似度原型作为先验知识用于辅助训练的有效性。进一步地，无论是基于欧式距离的相似度原型还是基于余弦相似度的原型，都能有效促进模型性能的提升。综合所有实验结果，可以发现一个普遍现象：在只考虑类间对比损失时，“mean”（全局平均）的表现更好；在同时考虑类间和类内对比损失时，“nonzero”（非零值平均）的表现更好。因此，在接下来的实验中将采用这两种对比损失计算方式，即“mean_inter”和“nonzero_inter_intra”，辅助模型训练。

5.3.3 网络组成成分分析

（1）渐进标签软化训练策略的性能评估

本节用八种常见模型对基于相似度原型的渐进标签软化（GLS）策略进行了评估。在 MIT-67 和 SUN397 数据集上应用 GLS 策略训练了这些模型，并将其性能与使用硬标签训练产生的性能（基线）以及采用传统标签平滑正则化（LSR）策略[77]训练产生的性

能进行了比较，结果汇总在表 5-1。

如表 5-1 所示，无论在轻量级模型还是在复杂模型上，GLS 策略均能显著提升场景识别的性能，验证了 GLS 在不同网络架构下出色的泛化能力。具体来说，在 MIT-67 数据集上，将 GLS 应用于 ResNet50 产生的准确率达到了 79.403%，超出了基线性能 2.91%。相比之下，传统的 LSR 策略所带来的性能提升则较为有限，在某些情况下甚至会对一些模型产生负面影响。例如，在 SUN397 数据集上，应用了 GLS 的 ResNet50 可以达到 62.06%的准确率，不仅比使用硬标签训练的基线性能提高了 1.637%，也比采用 LSR 训练的性能提高了 3.032%。值得注意的是，SUN397 数据集包含 397 个不同的场景类别，这进一步凸显了所提方法在处理大规模数据集时的稳定性和有效性。

表 5-1 在 MIT-67 和 SUN397 数据集上对渐进标签软化（GLS）策略的评估（%）

Tab. 5-1 Evaluation of Gradient Label Softening (GLS) on MIT-67 and SUN397 datasets (%)

数据集	模型	参数量	基线	LSR[77]	GLS 余弦相似度	GLS 欧式距离
MIT-67	ResNet50	25M	76.493	76.417-0.076	78.805+2.312	**79.403**+2.910
MIT-67	VGG16	138M	73.358	74.078+0.72	74.104+0.746	**75.0**+1.642
MIT-67	MobileNet	5.4M	69.179	68.507-0.672	71.045+1.866	**71.119**+1.94
MIT-67	ShuffleNet	7M	72.090	73.209+1.119	**74.925**+2.835	74.701+2.611
MIT-67	MobileVit	6M	74.552	75.522+0.97	**78.134**+3.582	77.91+3.353
MIT-67	SwinT	87M	88.681	88.657-0.024	**89.03**+0.349	88.906+0.225
MIT-67	VIT	86M	86.194	86.642+0.448	86.940+0.746	**87.015**+0.821
MIT-67	ConvNext	89M	87.388	87.239-0.149	88.507+1.119	**88.657**+1.269
SUN397	ResNet50	25M	60.423	59.028-1.395	61.863+1.44	**62.060**+1.637
SUN397	VGG16	138M	57.204	57.950+0.746	58.503+1.299	**58.715**+1.511
SUN397	MobileNet	5.4M	53.375	53.043-0.332	**55.657**+2.282	54.625+1.25
SUN397	ShuffleNet	7M	58.181	57.950-0.231	59.385+1.204	**60.08**+1.899
SUN397	MobileVit	6M	59.033	59.283+0.25	61.219+2.186	**61.28**+2.247
SUN397	SwinT	87M	75.063	74.826-0.237	77.013+1.95	**77.194**+2.131
SUN397	VIT	86M	75.078	75.043-0.035	**75.758**+0.68	75.677+0.6
SUN397	ConvNext	89M	74.675	75.017+0.342	**77.274**+2.599	77.239+2.564

（2）批量级对比损失训练策略的性能评估

本节应用基于相似度原型的批量级对比损失（BCL）策略辅助模型训练，并将其性能与基线（不使用对比损失训练）以及传统对比损失（CL）策略[78]进行了对比，在 CL 策略中，相同类别的样本对的相似度阈值被设定为“1”，不同类别的样本对的相似度

阈值则被设置为“0”。在 MIT-67 和 SUN397 数据集上，通过对多个模型应用上述策略进行实验，结果汇总在表 5-2。

表 5-2 中的结果显示，在大部分情况下，应用 BCL 策略辅助训练的模型性能要优于基线，也优于传统的 CL 策略。具体来说，在 MIT-67 数据集上，采用 BCL 训练的 ShuffleNet 相较于基线，性能提升了 2.014%，而采用传统的 CL 策略仅比基线提升了 1.044%。在 SUN397 数据集上，采用 BCL 训练的 ConvNext 性能比基线提升了 0.917%，而传统的 CL 产生的性能反而比基线下降了 0.043%。这些现象表明，传统 CL 策略由于过度强调同类样本之间的相似性以及非同类样本之间的差异性，可能会对模型的识别性能产生负面影响。这进一步凸显了在训练过程中引入相似度原型的先验知识，对于提高模型性能的可行性。

表 5-2 在 MIT-67 和 SUN397 数据集上对批量级对比损失（BCL）策略的评估（%）

Tab. 5-2 Evaluation of Batch-level Contrastive Loss (BCL) on MIT-67 and SUN397 datasets (%)

数据集	模型	参量	基线	CL[78]	BCL 余弦相似度 mean_inter	BCL 欧式距离 mean_inter	BCL 余弦相似度 nonzero_inter_intra	BCL 欧式距离 nonzero_inter_intra
MIT-67	ResNet50	25M	76.493	76.642	**77.014**+0.521	76.94	76.866	76.866
MIT-67	VGG16	138M	73.358	74.327	**74.851**+1.493	74.701	74.254	74.403
MIT-67	MobileNet	5.4M	69.179	68.358	**69.478**+0.299	68.433	68.582	68.433
MIT-67	ShuffleNet	7M	72.090	72.564	73.134	73.507	**74.104**+2.014	73.358
MIT-67	MobileVit	6M	74.552	74.776	75.149	75.149	**76.194**+1.642	75.075
MIT-67	SwinT	87M	88.681	89.328	89.179	**89.701**+1.02	89.403	89.328
MIT-67	VIT	86M	86.194	86.342	86.269	86.418	**86.493**+0.299	**86.493**+0.299
MIT-67	ConvNext	89M	87.388	87.462	87.835	87.537	**88.657**+1.269	87.537
SUN397	ResNet50	25M	60.423	60.635	**61.033**+0.61	60.806	60.916	60.821
SUN397	VGG16	138M	57.204	57.09	**57.436**+0.232	57.219	57.108	57.033
SUN397	MobileNet	5.4M	53.375	53.169	53.270	53.713	**54.055**+0.68	**54.055**+0.68
SUN397	ShuffleNet	7M	58.181	58.176	58.761	58.695	**59.073**+0.892	58.332
SUN397	MobileVit	6M	59.033	59.264	59.511	59.179	**60.121**+1.088	59.073
SUN397	SwinT	87M	75.063	74.7	75.037	75.088	**75.95**+0.887	74.877
SUN397	VIT	86M	75.078	75.264	75.416	75.314	**75.521**+0.443	75.275
SUN397	ConvNext	89M	74.675	74.632	74.942	74.866	**75.592**+0.917	74.635

值得指出的是，在某些特定情况下，BCL 策略可能会对模型性能产生不利影响。例如，在 MIT-67 数据集上评估 MobileNet 时，除了基于余弦相似度的“mean_inter”策略外，其他三种 BCL 策略都对模型性能有所损害。不过，传统 CL 策略所导致的性能下降

更为严重。因此，这一情况可能是由于在 MIT-67 数据集上对比损失策略不适用于 MobileNet 网络架构。尽管如此，在发生这种冲突的情况下，BCL 基于余弦相似度的“mean_inter”策略的应用仍能进一步提升模型性能。这再次验证了在模型训练过程中利用相似性原型提供的先验知识的价值。

（3）结合双训练策略的性能评估

本节尝试了同时利用渐进标签软化（GLS）和批量级对比损失（BCL）策略辅助模型训练，使用多个模型在 MIT-67 和 SUN397 数据集上进行了性能评估，实验结果如表 5-3 所示。

表 5-3 的结果表明，与基线策略（使用硬标签训练）相比，GLS 和 BCL 的结合显著提高了模型的识别准确率。具体而言，在 MIT-67 数据集上，同时使用 GLS 和 BCL 策略使 MobileVIT 的准确率提高了 4.552%；类似地，在 SUN397 数据集上，ConvNext 的准确率提高了 2.731%。需要强调的是，这一性能提升并未伴随网络参数或计算成本的增加，这一点对实际应用来说尤为重要。这些结果强有力地验证了所提出的相似性原型在提高模型性能方面的重要性及其广泛的适用性。

表 5-3 在 MIT-67 和 SUN397 数据集上结合 BCL 和 GLS 策略的评估（%）

Tab. 5-3 Evaluation of combining BCL and GLS on MIT-67 and SUN397 datasets (%)

数据集	模型	参量	基线	GLS+BCL 余弦 mean_inter	GLS+BCL 欧式 mean_inter	GLS+BCL 余弦 nonzero_inter_intra	GLS+BCL 欧式 nonzero_inter_intra
MIT-67	ResNet50	25M	76.493	**79.104**+2.611	78.881	78.433	78.806
MIT-67	VGG16	138M	73.358	75.0	75.522	73.881	**75.746**+2.388
MIT-67	MobileNet	5.4M	69.179	**71.045**+1.866	70.224	70.149	70.223
MIT-67	ShuffleNet	7M	72.090	**75.298**+3.208	74.402	74.179	74.328
MIT-67	MobileVit	6M	74.552	**79.104**+4.552	77.836	74.851	77.910
MIT-67	SwinT	87M	88.681	89.03	89.104	89.03	**89.179**+0.498
MIT-67	VIT	86M	86.194	**87.388**+1.194	86.866	86.567	86.791
MIT-67	ConvNext	89M	87.388	**88.805**+1.417	88.433	88.358	88.433
SUN397	ResNet50	25M	60.423	62.086	62.161	**62.358**+1.935	62.045
SUN397	VGG16	138M	57.204	58.519	**58.982**+1.778	57.516	58.872
SUN397	MobileNet	5.4M	53.375	55.869	56.166	54.232	**56.317**+2.942
SUN397	ShuffleNet	7M	58.181	59.788	**60.171**+1.99	59.954	**60.171**+1.99
SUN397	MobileVit	6M	59.033	61.274	61.365	60.71	**61.476**+2.433
SUN397	SwinT	87M	75.063	77.118	77.083	**77.294**+2.231	77.108
SUN397	VIT	86M	75.078	75.682	75.768	**75.768**+0.69	75.597
SUN397	ConvNext	89M	74.675	77.259	77.259	**77.406**+2.731	77.229

进一步地，本节对比了使用 GLS 策略获得的最佳性能、使用 BCL 策略获得的最佳性能以及同时使用这两种策略获得的最佳性能。表 5-4 汇总了这些结果。

如表 5-4 所示，与 GLS 策略相比，单独使用 BCL 策略对模型性能的提升相对有限。这一现象的原因在于，为了保持所提出的方法即插即用特性，本研究直接将 BCL 应用于随机提取的小批量（Mini-batch）图像，而没有对图像选择施加特定限制。这一点不同于其他常用的对比损失算法[78-81,86]，后者强调精心挑选用于训练的图像对以达到更优的性能表现。对表 5-4 的进一步分析揭示了 GLS 和 BCL 策略之间的互补效应。在大多数情况下，结合这两种训练策略可以实现高于单独使用任一策略的识别准确率。这一发现再次验证了所提出的相似性原型的潜力，并表明在利用相似性原型进行模型训练方面，仍有进一步的探索空间。

表 5-4 在 MIT-67 和 SUN397 数据集上三种训练策略的比较（%）

Tab. 5-4 Comparation of three training strategies on MIT-67 and SUN397 datasets (%)

数据集	模型	参数量	基线	GLS	BCL	GLS+BCL
MIT-67	ResNet50	25M	76.493	**79.403**	77.014	79.104
MIT-67	VGG16	138M	73.358	75.0	74.851	**75.746**
MIT-67	MobileNet	5.4M	69.179	**71.119**	69.478	71.045
MIT-67	ShuffleNet	7M	72.090	74.925	74.104	**75.298**
MIT-67	MobileVit	6M	74.552	78.134	76.194	**79.104**
MIT-67	SwinT	87M	88.681	89.03	**89.701**	89.179
MIT-67	VIT	86M	86.194	87.015	86.493	**87.388**
MIT-67	ConvNext	89M	87.388	88.657	88.657	**88.805**
SUN397	ResNet50	25M	60.423	62.060	61.033	**62.358**
SUN397	VGG16	138M	57.204	58.715	57.436	**58.982**
SUN397	MobileNet	5.4M	53.375	55.657	54.055	**56.317**
SUN397	ShuffleNet	7M	58.181	60.08	59.073	**60.171**
SUN397	MobileVit	6M	59.033	61.28	60.121	**61.476**
SUN397	SwinT	87M	75.063	77.194	75.95	**77.294**
SUN397	VIT	86M	75.078	75.758	75.521	**75.768**
SUN397	ConvNext	89M	74.675	77.274	75.592	**77.406**

5.3.4 特征可视化

为全面评估所提出的模型训练方法的有效性，本小节在 Places_7 和 Places_14 数据集上对模型进行了可视化分析。首先，使用传统的硬标签训练方法训练了一个 ResNet50

模型，作为性能基准。然后，采用所提出的训练策略对另一个 ResNet50 模型进行了训练。上述训练完成后，从训练好的模型中提取特征图（位于网络倒数第二层的特征），并应用 t-分布随机邻域嵌入（t-SNE）[70]技术对这些特征进行了降维处理，以便于直观展示和比较。通过在图 5-8 中绘制这些特征的二维表示，可以直观地观察到验证集中图像的分布情况。

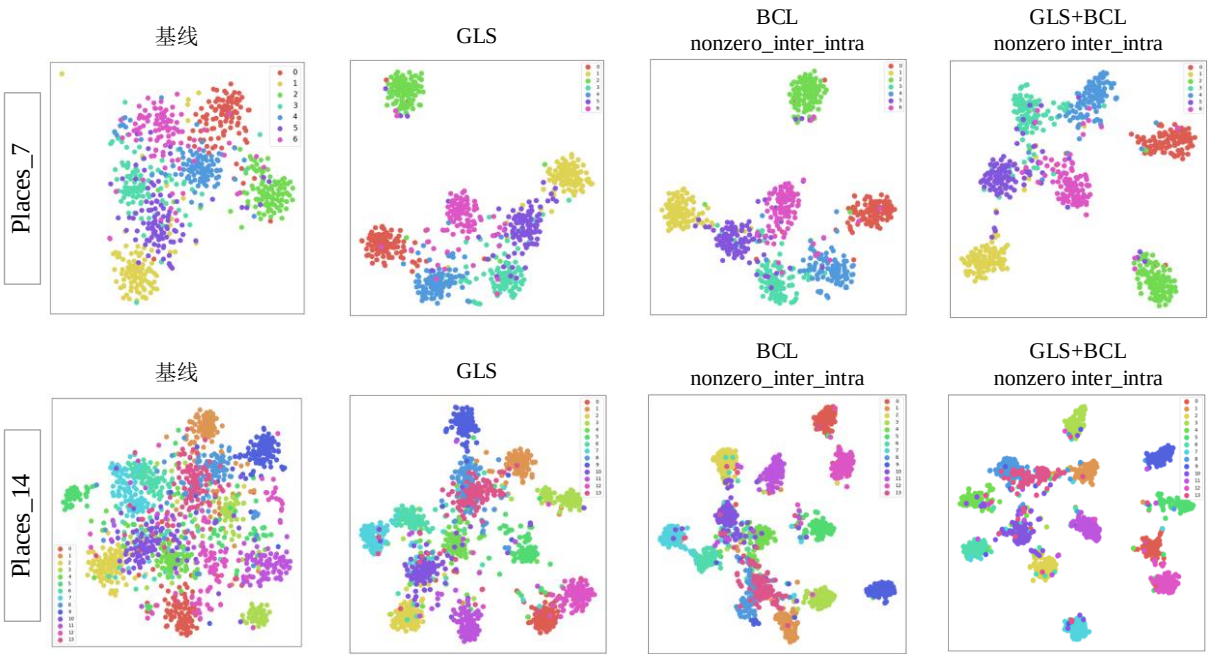


图 5-8 使用 t-SNE 在 Places_7 和 Places_14 验证集上的可视化

Fig. 5-8 Visualization on Places_7 and Places_14 validation sets using t-SNE

注：每个点代表一张图片，相同颜色的点代表属于同一类别的图片。

通过分析图 5-8，可以观察到相比于基线，采用本章所提出的任一模型训练策略均显著提升了网络的鉴别性能力。通过所提出的算法训练的模型区分不同场景类别的实例的能力得到了显著增强，凝聚同一场景类别下的实例的能力也实现了进一步的提升。直观来看，所提出的 GLS 和 BCL 训练策略产生了类似的可视化结果，而将这两种辅助策略结合应用则产生了一个更具区分能力的网络，使不同场景类别之间的界限更加清晰。这一结果与表 5-4 中的实验数据相吻合，进一步验证了所提出的相似度原型的潜力。综合本节的定性可视化分析，可以肯定的是，所提出的方法对模型训练起到了非常显著的促进作用。

5.3.5 计算成本评估

为验证所提方法在实际部署中卓越的计算效率，本节通过一个比较分析，展示本地训练和实际部署之间的计算成本差异。具体来讲，选用 ConvNext_Base[75]作为骨干网络，VIT Adapter[48]作为语义分割网络，将它们各自的浮点运算数（FLOPs）和推理吞吐量记

录在表 5-5。表 5-5 中，FLOPs 数据来源于各自的原始文献。至于推理吞吐量，考虑到这两篇文献使用了不同的测试设备，为确保比较的公平性，这里在同一块 NVIDIA 3090 GPU 上运行了这两个模块来测试推理吞吐量。鉴于 GLS 和 BCL 在训练过程中仅占用了微量的 FLOPs，因此表 5-5 中并未展示它们的数据。注意，由于 GLS 和 BCL 在实际部署阶段不再使用，这种遗漏不会影响本节的结论。

表 5-5 训练和实际部署中计算成本的比较

Tab. 5-5 Comparison of computational cost during training and practical deployment

网络架构	训练阶段 FLOPs (G)	实际运行阶段 FLOPs (G)	训练阶段 吞吐量 (image/s)	实际运行阶段 吞吐量 (image/s)
骨干网络	15.4	15.4	115	115
语义分割	473	0	0.09	-
汇总	488.7	**15.4**	<0.09	**115**

从表 5-5 的数据可以明显看出，本章所提出的方法仅依赖于训练好的骨干网络实现实时的场景识别。通过有效地消除对语义分割技术的依赖，大幅降低了模型所需的 FLOPs，从而在实际部署中显著提升了模型的推理吞吐量。这一结果与本章最初的预期相符，有力地验证了所提方法在实际部署中显著的计算效率优势。此外，通过在实际应用中省去语义分割过程，也减轻了对所选用的语义分割技术的推理速度的要求。例如，尽管 VIT Adapter 在推理速度上的局限性使其可能不适用于边缘设备的实际部署，但由于本章方法只在模型训练阶段使用该技术，能够在不牺牲实际部署可行性的同时，充分利用其高精度优势。

5.3.6 与现有算法的性能比较

如表 5-6 所示，本小节首先以 ConvNext 作为骨干网络，将本章所提出的方法与现有的先进算法在 MIT-67 和 SUN397 数据集上进行了比较。然后，对第四章所提出的 DGN-Net 应用本章所提出的训练策略进行了训练，以便与本文其他章节提出的算法做进一步对比分析。根据表 5-6 中的结果，与大多数现有方法相比，本章方法展现了卓越的场景识别性能。

现有的基于物体信息辅助的方法[32,33,42,59]（包括本文其他章节提出的方法）在实际应用中需要大量的计算资源提取场景中的物体信息。不同于这些方法，本章方法仅在训练阶段使用物体知识增强网络的特征提取能力，从而避免了在实际应用中提取物体信息的计算负担。值得注意的是，尽管在实际应用中不再直接利用物体信息，本章方法仍然展现了比大多数基于物体辅助的方法更优越的识别性能，这进一步凸显了所提出的相似性原型的优势和实用价值。此外，本节还在表 5-6 中展示了采用各种组合技术提升性能

的最新算法。虽然这些算法结合了多尺度信息[37,58,87-89]或多模型组合[32,58,59,89]等因素，本章所提出的方法始终保持着领先地位，基本上达到了最先进的场景识别性能水平。

表 5-6 在 MIT-67 和 SUN397 数据集上不同算法的比较（%）

Tab. 5-6 Comparison of different methods on MIT-67 and SUN397 (%)

算法	输入尺寸	MIT-67	SUN397
MLR+CFV+FCR[87]	256×256	82.24	64.53
NNSD+ICLC[37]	224×224	84.30	64.78
Multi-Scale CNN[88]	889×889	80.97	70.17
MP[89]	640×640	86.90	72.60
SOSF+CFA+GAF[40]	608×608	**89.51**	**78.93**
Multi-Resolution CNN[56]	336×336	86.70	72.00
SDO-Net[32]	224×224	86.76	73.41
SAS-Net [33]	224×224	87.10	74.04
ARG-Net [42]	448×448	88.13	75.02
PL + AP + AI + IM[58]	448×448	88.06	74.12
MR-Net[59]	448×448	88.08	73.98
CSSRM	224×224	88.73	75.48
SpaCoNet	224×224	90.75	76.15
DGN-Net	224×224	90.37	79.77
ConvNext + GLS + BCL	224×224	**88.81**	**77.41**
DGN-Net + GLS + BCL	224×224	91.42	80.37

注 1：CSSRM、SpaCoNet、DGN-Net 分别为本文第二章、第三章、第四章所提出的算法。

注 2：DGN-Net + GLS + BCL 指使用本章所提出的训练策略训练 DGN-Net（迭代次数等超参数配置与第四章完全一致）。

此外，与本文第三章提出的 SpaCoNet 和第四章提出的 DGN-Net 相比，尽管本章所提出的算法在识别性能上不具备优势，但它在实际部署中具有非常显著的优势。具体而言，该算法无需执行耗费大量计算资源的物体信息提取过程，从而大幅提高了运行效率。进一步地，为与其他章节的算法公平比较，本节采用本章介绍的训练方法（基于欧式距离的相似度原型结合 GLS+$BCL_{nonzero_inter_intra}$）对 DGN-Net 进行了重新训练，如表 5-6 最后一行所示。可以看到，本章的方法使 DGN-Net 的识别性能更进一步，超越了本文中所有其他算法。这一发现再次验证了本章提出的相似度原型的即插即用特性和强大的泛化能力，甚至能够提升已经经过精细调整的场景识别算法的性能。

综上所述，表 5-6 提供了充分的证据，彰显了本章所提出的相似性原型在提升场景

识别性能方面的显著优势和实际应用价值。

5.4 本章小结

本章提出将物体语义知识嵌入到一个类间语义相似度原型中，实现在不增加任何网络参数的前提下辅助模型训练。在相似度原型的基础上，提出了渐进标签软化（GLS）和批量级对比损失（BCL）两种训练策略，以充分挖掘和利用该原型在辅助场景识别方面的潜力。具体来说，GLS 通过将相似度原型嵌入到软化后的标签中，利用置信度渐进增强策略引导网络训练；BCL 则基于相似度原型衡量每个迷你批次中类内样本对和类间样本对的相似度要求。

通过实施上述策略，本研究在场景识别任务上取得了显著进展，并首次验证了在不增加计算负担的情况下，物体信息可以有效提高场景识别的性能。在几个公开场景数据集（MIT-67、SUN397、Places_7 和 Places_14）上进行的综合评估表明，所提出的相似度原型及相关策略能够有效提升网络识别性能，并在实际部署中避免了增加额外的网络参数或计算成本。

6 总结与展望

6.1 工作总结

本文以深度学习方法为技术基础，围绕基于语义引导的场景识别这一领域展开研究。研究伊始，针对场景内复杂多变的空间布局，设计了一个充分考虑空间信息特性的语义空间关系建模网络，实现全面理解并表达场景内的语义空间布局；在此基础上，面对跨场景共存物体导致的高类间相似度挑战，提出将场景相关特征信息与物体语义紧密结合，并通过注意力机制自适应探索物体共现性，生成更具区分度的场景表示；进一步地，考虑到对物体信息的完全启发式利用可解释性有限，从提高可解释性的推理式角度入手，提出了一种基于物体间先验知识和鉴别性图网络的场景识别方法；最后，考虑到物体信息提取对计算资源的高需求可能限制语义引导方法的实际部署，提出了一个高效利用物体信息的语义相似度原型，借助标签平滑和对比损失训练策略，实现在不增加计算负担的前提下利用物体信息提升场景识别的性能。本文的主要贡献总结如下：

（1）为深入探究场景内物体间的复杂空间布局，第二章设计了一个语义空间关系模块（SSRM）。首先，借助语义分割技术解耦出场景内纯净的空间信息，以避免无关信息对空间上下文建模的干扰；其次，设计了自适应置信度滤波器（ACF），基于图像的具体状态自适应地优化其语义分割得分张量，有效缓解了语义分割技术精度有限带来的语义歧义问题；然后，结合通道注意力机制构建了一个充分考虑语义分割得分张量特性的网络架构，确保 SSRM 对场景内空间信息的全面利用；最后，又结合 PlacesCNN 和条带深度卷积构建了组合语义空间关系模块（CSSRM），实现对场景内所有信息的全面利用。在多个公开数据集上的综合实验表明，本章算法达到了目前先进性能。

（2）针对第二章未考虑的不同场景中相同语义物体所引发的高类间相似问题，第三章提出了先为物体赋予场景相关特征，再建模物体共现性的解决思路。在 SSRM 的基础上，提出了融合语义空间关系和物体共现性建模的室内场景识别网络。首先，基于 SSRM 和图像特征提取模块提取场景特征；然后，设计了一个语义节点特征聚合模块，将输入场景的相关特征分配给各个物体语义节点，并生成语义节点特征序列，藉此区分不同场景中相同语义的物体；最后，基于注意力机制设计了全局-局部依赖模块，利用编码器-解码器架构深入挖掘语义节点特征序列中的共现信息，以此构建更具区分度的场景表示。实验结果表明，通过融合场景特征建模物体共现性，使本章算法实现了比第二章算法更优越的场景识别性能。

（3）以上两章算法从完全启发式角度利用场景内的物体信息辅助场景识别，可解

释性有限。围绕该问题，第四章提出从推理角度以更具可解释性的方式利用物体知识辅助场景识别。首先，基于概率统计方法捕捉物体与场景之间的鉴别性关系，并将这种关系转化为物体间鉴别性原型（IODP），IODP 中包含了丰富的物体间鉴别性知识；然后，设计了一个鉴别性图网络（DGN），将像素级场景特征定义为图节点，并借助 IODP 中的物体间鉴别性先验知识编码节点间的关联，形成图的边。最后，通过图卷积和映射操作，实现物体间鉴别性知识与场景特征的显式融合，形成极具鉴别性的图像表示。实验结果表明，从知识推理角度融合物体知识和场景特征，使本章算法在单分支架构下即可获得与前两章相当甚至更优的场景识别性能。

（4）以上三章所提出的算法在部署到边缘设备时，均需消耗大量计算资源提取场景内的物体信息，这极大加重了边缘设备的运行负担。针对物体信息提取对算力的高需求限制了语义引导方法的实际部署这一问题，第五章提出了一种高效利用物体信息的方法。首先，基于语义分割技术和统计分析方法捕获场景中物体信息的分布情况，并据此构建了一个类间语义相似度原型；然后，基于相似度原型设计了渐进标签软化（GLS）和批量级对比损失（BCL）两种策略，实现在不增加任何网络参数的情况下辅助模型训练。GLS 将相似度原型嵌入到软化标签中，并利用置信度渐进策略指导网络训练；BCL 则基于相似度原型衡量训练批次中类内和类间样本之间的相似度要求。在公开数据集上的综合评估表明，上述策略可以实现在不增加额外计算负担或网络参数的前提下，利用物体信息提升场景识别性能。

6.2 研究展望

本文围绕基于语义引导的场景识别方法的现存问题，开展了一系列相关的研究工作，并取得了一些研究成果。尽管如此，以下研究方向仍需进一步探索：

（1）现有的基于语义引导的方法受限于其所使用的语义分割技术能够识别的物体类别数量。举例来说，本文中采用的语义分割技术仅能识别 150 种物体，在一些场景中该技术可能无法分割所有出现的物体。当这些未被识别的物体对该场景具有鉴别性意义时，这一局限性可能会对最终的识别结果产生较大的负面影响。

有两种可能的策略解决这个问题。首先，可以考虑训练能够识别更多物体的语义分割技术，但随之而来的挑战是需要对大量的数据进行精确标注，这无疑是一项资源密集型的工作；或者，也可以探索半监督或无监督学习方法，使网络能够自动识别场景中的物体，进而克服对语义分割技术的依赖，这个方向值得深入研究。

（2）本文第五章提出的相似度原型首次验证了在不增加额外计算开销的情况下，通过物体信息显著提升场景识别性能的可能性，为后续研究开辟了新的探索路径。期待这项工作可以启发后续研究开发出更高效的基于语义辅助的场景识别方法。同时，第

5.3.3 节的实验分析表明，所提出的相似度原型在增强场景识别性能方面具有未被完全挖掘的潜力。鉴于此，未来的研究可以继续从模型架构、损失函数设计或特征提取方法等方面进行探索，充分挖掘该原型的应用潜力。

参考文献

[1] 中华人民共和国国务院办公厅. 关于印发新一代人工智能发展规划的通知[R/OL]. (2017-07-20) [2024-05-24]. https://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.

[2] 科技部等六部门. 关于印发《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》的通知 [R/OL]. (2022-07-29) [2024-05-24]. https://www.gov.cn/zhengce/zhengceku/2022-08/12/content_5705154.htm.

[3] Xie L, Lee F, Liu L, et al. Scene recognition: a comprehensive survey[J]. Pattern Recognition, 2020, 102: 107205.

[4] 余慧瑾, 方勇纯. 基于改进型自主发育网络的机器人场景识别方法[J]. 自动化学报, 2021, 47(7): 1530-1538.

[5] Wang Y, Qiu Y, Cheng P, et al. Hybrid CNN-transformer features for visual place recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(3): 1109-1122.

[6] Smith W H B, Milford M, McDonald-Maier K D, et al. OpenSceneVLAD: Appearance invariant, open set scene classification[C]. International Conference on Robotics and Automation, Philadelphia, USA, 2022: 4578-4584.

[7] 姚芷馨, 张太红, 赵昀杰. 基于卷积神经网络的多模型交通场景识别研究[J]. 计算机技术与发展, 2022, 32(7): 93-98.

[8] Boyd K, McAllister P, Mulvenna M, et al. Designing multimodal video search by examples (MVSE) user interfaces: UX requirements elicitation and insights from semi-structured interviews[C]. Proceedings of the European Conference on Cognitive Ergonomics, New York, USA, 2023: 1-8.

[9] Zeng Q. Construction of intelligent art guidance system based on intelligent robot-assisted painting equipment[C]. International Conference on Electronics and Sustainable Communication Systems, Coimbatore, India, 2022: 1345-1348.

[10] Qiu J, Yang Y, Wang X, et al. Scene essence[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 8322-8333.

[11] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60: 91-110.

[12]Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, 2005, 1: 886-893.

[13]Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.

[14]宋新航. 局部到全局的场景识别技术研究[D]. 中国科学院大学, 2017.

[15]Sivic, Zisserman. Video Google: A text retrieval approach to object matching in videos[C]. Proceedings of the IEEE International Conference on Computer Vision, Nice, France, 2003: 1470-1477.

[16]Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, 2006: 2169-2178.

[17]Quattoni A, Torralba A. Recognizing indoor scenes[C]. IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 413-420.

[18]Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]. IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009: 248-255.

[19]Zhou B, Lapedriza A, Khosla A, et al. Places: a 10 million image database for scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(6): 1452-1464.

[20]Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25.

[21]Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[22]He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770-778.

[23]Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[24]Dixit M, Chen S, Gao D, et al. Scene classification with semantic fisher vectors[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, 2015: 2974-2983.

[25]Dixit M, Li Y, Vasconcelos N. Semantic fisher scores for task transfer: using objects to

classify scenes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(12): 3102-3118.

[26] Liu Y, Chen Q, Chen W, et al. Dictionary learning inspired deep network for scene recognition[C]. Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018, 32(1).

[27] Zhao Z, Larson M. From volcano to toyshop: Adaptive discriminative region discovery for scene recognition[C]. Proceedings of the ACM International Conference on Multimedia, Seoul, Republic of Korea, 2018: 1760-1768.

[28] Chen G, Song X, Zeng H, et al. Scene recognition with prototype-agnostic scene layout[J]. IEEE Transactions on Image Processing, 2020, 29: 5877-5888.

[29] López-Cifuentes A, Escudero-Viñolo M, Bescós J, et al. Attention-based knowledge distillation in scene recognition: the impact of a dct-driven loss[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023.

[30] Khan A, Chefranov A, Demirel H. Building discriminative features of scene recognition using multi-stages of inception-ResNet-v2[J]. Applied Intelligence, 2023, 53(15): 18431-18449.

[31] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. arXiv preprint arXiv:1609.02907, 2016.

[32] Cheng X, Lu J, Feng J, et al. Scene recognition with objectness[J]. Pattern Recognition, 2018, 74: 474-487.

[33] López-Cifuentes A, Escudero-Vinolo M, Bescós J, et al. Semantic-aware scene recognition[J]. Pattern Recognition, 2020, 102: 107256.

[34] Vogel J, Schiele B. Semantic modeling of natural scenes for content-based image retrieval[J]. International Journal of Computer Vision, 2007, 72: 133-157.

[35] Li L J, Su H, Fei-Fei L, et al. Object bank: a high-level image representation for scene classification & semantic feature sparsification[J]. Advances in Neural Information Processing Systems, 2010, 23.

[36] Wang Z, Wang L, Wang Y, et al. Weakly supervised patchnets: describing and aggregating local patches for scene recognition[J]. IEEE Transactions on Image Processing, 2017, 26(4): 2028-2041.

[37] Xie L, Lee F, Liu L, et al. Hierarchical coding of convolutional features for scene recognition[J]. IEEE Transactions on Multimedia, 2019, 22(5): 1182-1192.

[38] Sitaula C, Aryal S, Xiang Y, et al. Content and context features for scene image representation[J]. Knowledge-Based Systems, 2021, 232: 107470.

[39] Pal A, Nieto-Granda C, Christensen H I. Deduce: diverse scene detection methods in unseen challenging environments[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, Macau, China, 2019: 4198-4204.

[40] Sun N, Li W, Liu J, et al. Fusing object semantics and deep appearance features for scene recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29(6): 1715-1728.

[41] Song X, Jiang S, Wang B, et al. Image representations with spatial object-to-object relations for RGB-D scene recognition[J]. IEEE Transactions on Image Processing, 2019, 29: 525-537.

[42] Zeng H, Song X, Chen G, et al. Amorphous region context modeling for scene recognition[J]. IEEE Transactions on Multimedia, 2022, 24: 141-151.

[43] Pereira R, Garrote L, Barros T, et al. A deep learning-based indoor scene classification approach enhanced with inter-object distance semantic features[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, Prague, Czech Republic, 2021: 32-38.

[44] Song X, Liu C, Zeng H, et al. Composite object relation modeling for few-shot scene recognition[J]. IEEE Transactions on Image Processing, 2023.

[45] Zhou L, Cen J, Wang X, et al. BORM: bayesian object relation model for indoor scene recognition[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, Prague, Czech Republic, 2021: 39-46.

[46] Choe S, Seong H, Kim E. Indoor place category recognition for a cleaning robot by fusing a probabilistic approach and deep learning[J]. IEEE Transactions on Cybernetics, 2021, 52(8): 7265-7276.

[47] Egenhofer M. A mathematical framework for the definition of topological relations[C]. Proceedings of the International Symposium on Spatial Data Handing, Zurich, Switzerland, 1990: 803-813.

[48] Chen Z, Duan Y, Wang W, et al. Vision transformer adapter for dense predictions[J]. arXiv preprint arXiv:2205.08534, 2022.

[49] Zhou B, Zhao H, Puig X, et al. Scene parsing through ade20k dataset[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 633-641.

[50] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]. Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 3-19.

[51] Howard A G, Zhu M, Chen B, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[52] Miao B, Zhou L, Mian A S, et al. Object-to-scene: learning to transfer object knowledge to indoor scene recognition[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, Prague, Czech Republic, 2021: 2069-2075.

[53] Xiao J, Hays J, Ehinger K A, et al. Sun database: large-scale scene recognition from abbey to zoo[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, USA, 2010: 3485-3492.

[54] Berrada L, Zisserman A, Kumar M P. Training neural networks for and by interpolation[C]. Proceedings of the International Conference on Machine Learning, PMLR, 2020: 799-809.

[55] Paszke A, Gross S, Massa F, et al. Pytorch: an imperative style, high-performance deep learning library[J]. Advances in Neural Information Processing Systems, 2019, 32.

[56] Wang L, Guo S, Huang W, et al. Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs[J]. IEEE Transactions on Image Processing, 2017, 26(4): 2055-2068.

[57] Yee P S, Lim K M, Lee C P. DeepScene: scene classification via convolutional neural network with spatial pyramid pooling[J]. Expert Systems with Applications, 2022, 193: 116382.

[58] Zeng H, Song X, Chen G, et al. Learning scene attribute for scene recognition[J]. IEEE Transactions on Multimedia, 2019, 22(6): 1519-1530.

[59] Lin C, Lee F, Xie L, et al. Scene recognition using multiple representation network[J]. Applied Soft Computing, 2022, 118: 108530.

[60] Zhang J, Ainam J P, Zhao L, et al. Scene recognition with objectness, Attribute and Category Learning[J]. arXiv preprint arXiv:2207.10174, 2022.

[61] Chen B X, Sahdev R, Wu D, et al. Scene classification in indoor environments for robots using context based word embeddings[J]. arXiv preprint arXiv:1908.06422, 2019.

[62] Zhou L, Zhou Y, Qi X, et al. Attentional graph convolutional network for structure-aware audiovisual scene classification[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-15.

[63] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.

[64] Xie Y, Yan J, Kang L, et al. FCT: fusing CNN and transformer for scene classification[J]. International Journal of Multimedia Information Retrieval, 2022, 11(4): 611-618.

[65] Wang Y, Qiu Y, Cheng P, et al. Hybrid CNN-transformer features for visual place recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(3): 1109-1122.

[66] Lv P, Wu W, Zhong Y, et al. SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-12.

[67] Zhu B, Xie J, Gao X, et al. A fusiform network of indoor scene classification with the stylized semantic description for service-robot applications[J]. Expert Systems with Applications, 2024, 245: 122979.

[68] Springenberg J T, Dosovitskiy A, Brox T, et al. Striving for simplicity: the all convolutional net[J]. arXiv preprint arXiv:1412.6806, 2014.

[69] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization[C]. Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 2017: 618-626.

[70] Van der Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11).

[71] Zou Z, Liu W, Xing W. AdaNFF: a new method for adaptive nonnegative multi-feature fusion to scene classification[J]. Pattern Recognition, 2022, 123: 108402.

[72] Yuan Y, Xiong Z, Wang Q. Acm: adaptive cross-modal graph convolutional neural networks for rgb-d scene recognition[C]. Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, USA, 2019, 33(01): 9176-9184.

[73] Yang Y, Tang X, Cheung Y M, et al. SAGN: semantic-aware graph network for remote sensing scene classification[J]. IEEE Transactions on Image Processing, 2023, 32: 1011-1025.

[74] Kingma D P, Ba J. Adam: a method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[75] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 11976-11986.

[76] Hou Y, Ma Z, Liu C, et al. Network pruning via resource reallocation[J]. Pattern Recognition, 2024, 145: 109886.

[77] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2818-2826.

[78] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with

application to face verification[C]. IEEE Computer Society Conference on Computer Vision and pattern Recognition, San Diego, USA, 2005, 1: 539-546.

[79] Ni J, Liu J, Zhang C, et al. Fine-grained patient similarity measuring using deep metric learning[C]. Proceedings of the ACM on Conference on Information and Knowledge Management, Singapore, Singapore, 2017: 1189-1198.

[80] Sohn K. Improved deep metric learning with multi-class n-pair loss objective[J]. Advances in Neural Information Processing Systems, 2016, 29.

[81] Oh Song H, Xiang Y, Jegelka S, et al. Deep metric learning via lifted structured feature embedding[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 4004-4012.

[82] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019: 1314-1324.

[83] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: practical guidelines for efficient cnn architecture design[C]. Proceedings of the European conference on Computer Vision, Munich, Germany, 2018: 116-131.

[84] Mehta S, Rastegari M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer[J]. arXiv preprint arXiv:2110.02178, 2021.

[85] Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 10012-10022.

[86] Hoffer E, Ailon N. Deep metric learning using triplet network[J]. arXiv preprint arXiv:1409.1556, 2014.

[87] Xie G S, Zhang X Y, Yan S, et al. Hybrid CNN and dictionary-based models for scene recognition and domain adaptation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 27(6): 1263-1274.

[88] Herranz L, Jiang S, Li X. Scene recognition with cnns: objects, scales and dataset bias[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 571-579.

[89] Song X, Jiang S, Herranz L. Multi-scale multi-feature context modeling for scene recognition in the semantic manifold[J]. IEEE Transactions on Image Processing, 2017, 26(6): 2721-2735.

致　　谢

行文至此，百感交集。时光荏苒，三年的硕士生活即将结束。

由衷感谢我的导师马昕教授对我的悉心培养和指导。马老师是一位出色的科研引路人。犹记得科研初期自己的进度缓慢，是马老师耐心地指导我，教会我如何高效地阅读论文并深入总结。马老师严谨细致的科研态度、和煦真诚的待人方式，对我产生了深远的影响，值得我终身学习。当我在科研道路上遇到挫折与迷茫时，马老师让我明白，科研之路本就充满曲折与挑战。同时，非常感谢马老师为我提供出国参加学术会议的机会，这极大地开阔了我的视野。在未来的科研旅程中，我将以马老师为榜样，坚定信念，砥砺前行。

在申博过程中，感谢王超群老师和济南大学的赵钦君老师对我提供的帮助与指点。

感谢 326 这个大家庭中的每一位成员。感谢武寒波师姐在我小论文撰写期间提供的帮助，感谢李翔师兄对实验室服务器的维护与贡献，感谢王凯师兄对实验室事务的管理与安排，感谢李刚师兄、田皓宇师兄、刘涵池师姐、于江磊师兄、程金龙师兄、单益飞师兄、孙甲浩、张昊、高兆志、王煦等实验室的同门；同样感谢我的室友蒋泽翠、李祎。祝愿大家都能得偿所愿，前程似锦。

最后，我要感谢求学路上一直在背后默默支持我的父母和姐姐，你们的支持是我前行的动力。

这三年的点滴历历在目，有过意气风发，也有过失意沮丧；有过自信满满，也有过忧虑不安。我将永远珍藏这段经历，始终提醒自己不忘初心，砥砺前行。

我们无法改变过去，但可以努力争取自己的未来。共勉！

攻读硕士学位期间的科研成果与所获奖励

1 学术论文

[1] **Chuanxin Song**, Xin Ma. SRRM: Semantic Region Relation Model for Indoor Scene Recognition[C]. 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 2023. (EI, Oral)

[2] **Chuanxin Song**, Hanbo Wu, Xin Ma. Inter-object Discriminative Graph Modeling for Indoor Scene Recognition. Knowledge-Based Systems. (SCI, Q1, Revise)

[3] **Chuanxin Song**, Hanbo Wu, Xin Ma, Yibin Li. Semantic-embedded Similarity Prototype for Scene Recognition. Pattern Recognition. (SCI, Q1, Major Revision)

[4] **Chuanxin Song**, Hanbo Wu, Xin Ma. Semantic-guided modeling of spatial relation and object co-occurrence for indoor scene recognition. Expert Systems with Applications. (SCI, Q1, Under Review)

2 发明专利

[1] 马昕，**宋传鑫**，宋锐，荣学文，周乐来，王超群，李贻斌，基于组合语义区域关系模型的室内场景识别方法及系统，申请号：2023103406980，申请日期：2023-03-29，状态：公开实审（导师一作，学生二作）

3 参与项目

[1] 国家自然科学基金-区域创新发展联合基金项目：可视度受限环境下跨光谱多传感信息融合的机器人语义感知与交互协作（U21A20486）

[2] 山东省自然科学基金重大基础研究项目：面向康养服务机器人的人-机-环境共融基础理论与关键技术（ZR2019ZD07）

[3] 国家重点研发计划课题：基于双模三维视觉传感器的服务机器人人机交互研究及应用（2018YFB1305803）

4 所获奖励与荣誉

[1] 中国研究生电子设计大赛华北赛区一等奖（第一位次）

[2] 山东大学创新创业活动先进个人

[3] 山东大学硕士生一等奖学金

[4] 山东大学优秀研究生

学位论文评阅及答辩情况表

<table>
<tr><td rowspan="6">论文评阅人</td><td colspan="2">姓　名</td><td>专业技术职　务</td><td>是否博导（硕导）</td><td colspan="2">所 在 单 位</td><td>总体评价</td></tr>
<tr><td colspan="2">匿名评阅</td><td></td><td></td><td colspan="2"></td><td>优秀</td></tr>
<tr><td colspan="2">匿名评阅</td><td></td><td></td><td colspan="2"></td><td>优秀</td></tr>
<tr><td colspan="2">匿名评阅</td><td></td><td></td><td colspan="2"></td><td>优秀</td></tr>
<tr><td colspan="2"></td><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td colspan="2"></td><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td rowspan="9">答辩委员会成员</td><td colspan="2">姓　名</td><td>专业技术职　务</td><td>是否博导（硕导）</td><td colspan="3">所 在 单 位</td></tr>
<tr><td>主席</td><td>于宁波</td><td>教授</td><td>博导</td><td colspan="3">南开大学</td></tr>
<tr><td rowspan="7">委

员</td><td>田新诚</td><td>教授</td><td>博导</td><td colspan="3">山东大学控制科学与工程学院</td></tr>
<tr><td>宋锐</td><td>教授</td><td>博导</td><td colspan="3">山东大学控制科学与工程学院</td></tr>
<tr><td>范永</td><td>高级工程师</td><td>硕导</td><td colspan="3">山东优宝特智能机器人有限公司</td></tr>
<tr><td></td><td></td><td></td><td colspan="3"></td></tr>
<tr><td></td><td></td><td></td><td colspan="3"></td></tr>
<tr><td></td><td></td><td></td><td colspan="3"></td></tr>
<tr><td colspan="2">答辩委员会对论文的总体评价</td><td>优秀</td><td>答辩秘书</td><td></td><td>答辩日期</td><td>2024 年 5 月 24 日</td></tr>
<tr><td colspan="2">备注</td><td colspan="6"></td></tr>
</table>