中国科学院大学

University of Chinese Academy of Sciences

博士学位论文

有限监督条件下的视觉目标定位及关系推理

作者姓名：朱艺

指导教师：焦建彬 教授

中国科学院大学电子电气与通信工程学院

学位类别：工学博士

学科专业：信号与信息处理

培养单位：中国科学院大学电子电气与通信工程学院

2020 年 11 月

Visual Object Recognition and Relationship Reasoning Under Limited Supervision

A dissertation submitted to the

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in Signal and Information Processing

By

Yi Zhu

Supervisor: Professor Jianbin Jiao

School of Electronic, Electrical and Communication Engineering

University of Chinese Academy of Sciences

November, 2020

中国科学院大学
学位论文原创性声明

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工作所取得的成果。尽我所知，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研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关保存和使用学位论文的规定，即中国科学院大学有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

场景理解是计算机视觉领域中的一个重要研究方向，在自动驾驶、智能监控、机器人导航等工业应用领域得到了广泛的应用。场景理解涉及视觉目标定位和对目标间语义关系的推理，为众多下游视觉任务如图像检索、视觉问答、视觉语言导航等提供了研究基础。基于深度学习的目标定位和关系推理模型往往依赖大数据监督训练，人工标记大量训练数据十分耗时耗力，且容易因为标注人员的主观性产生噪声标签，因此可用于训练模型的样本质量和数量往往十分有限。

在有限监督条件下，目标定位及关系推理方法面临着待识别目标外观多变、目标间关系种类繁多，且训练样本不充足或标注不精确的问题。有效解决这些问题才能进一步推动场景理解的研究模型和应用系统的性能提升。本文针对有限监督条件下的目标定位及关系推理问题进行了深入研究，内容与成果包括：

（1）针对弱监督目标定位任务中无法定位完整目标区域的问题，提出了一种目标概率预选模块。该模块基于目标内特征响应的语义和空间关联的先验，引导网络通过学习来激活更加完整的目标范围，从而得到更加准确的锚点定位和框定位结果。进一步地，为了得到精细的分割定位结果，提出了一种实例激活图技术。该技术从含噪声的目标预选掩膜中学习类无关的填充权值，并基于该权值生成目标的实例分割结果。

（2）针对视觉关系检测任务中由于关系类别的长尾分布导致的不常见类别识别困难的问题，提出了一种渐近知识驱动变换器。该变换器使用外部知识监督以约束目标区域间的连接关系，使得特征学习聚焦在关联性更强的目标区域。为了使模型能够兼容来自真实世界中不同域的常识知识，并使特征学习不受冗余和噪声知识的干扰，提出了可配置图推理技术。该技术将传统关系推理路径分解为多个子路径，通过学习动态地配置基于知识的视觉关系推理路径，自适应地对不同的关系特征进行知识增强，从而提升视觉关系模型的预测精度。当关系类别的标注示例极度稀少时，模型训练样本偏差较大，使得模型难以充分学习区分不同的关系类别，这导致模型鲁棒性大大降低。针对此我们提出了场景图攻击技术，该技术通过攻击节点语义促使模型学习来保证全局语义一致性，从而提升模型精确度和泛化性。

（3）针对视觉场景理解与自然语言结合时的跨模态信息难以匹配的问题，提出一种跨模态记忆网络。该网络在视觉对话导航任务中将视觉场景与对话语言在时序上进行关联，通过学习编码关于历史动作决策的丰富的跨模态记忆信息，辅助当前的导航动作决策。

本文的研究成果，为弱监督目标定位问题提出了一种一体化的粗定位框架和一种实例级别分割定位框架，为目标间的视觉关系检测问题提出了一种可渐近地集成不同类型知识的特征变换器、一种能够从海量知识图中挑选知识的图推理技术以及一种针对场景语义图的对抗攻击技术，为解决深度学习框架中训练样本标注稀少或不够精确的问题提供了一种新的方法，为实际应用中的场景理解问题提供了新的研究思路和方向。

关键词：有限监督，场景理解，视觉目标检测，视觉关系检测，图模型

Abstract

Scene understanding is an important research topic in computer vision, and has been widely used in industrial applications such as autonomous driving, intelligent monitoring, and robot navigation. Scene understanding involves the recognition of visual objects and the reasoning of semantic relations between object pairs, providing a research foundation for many downstream visual tasks such as image retrieval, visual question and answer, and visual language navigation. Deep learning based models for object localization and relationship reasoning models usually rely on big data training. Manual labeling of these data will be very time-consuming and labor-intensive, and it is easy to generate noise labels due to the subjectivity of annotators. Therefore, the quality and quantity of the training samples are extremely limited.

Under limited supervision, object localization and relationship reasoning methods are faced with the problem of recognizing objects and relations with a variety of appearances and various types, and insufficient training samples or inaccurate labeling. Solving these problems would further promote the performance of research models and application systems for scene understanding. This paper conducts research on object localization and relationship reasoning under limited supervision. The contributions include:

(1) Proposing a Soft Proposal Network (SPN) model, which addresses the problem of weakly supervised object localization that fail to identify the whole object extent. SPN learns to guide network learning based on the priors about the semantic and spatial correlations between the feature responses within each object, leading the model to activate more complete object region; thus obtaining more accurate predictions on point localization and bounding box localization. To further obtain fine-detailed segmentation masks, an Instance Activation Mapping (IAM) technique is proposed to learn class-agnostic filling weights from noisy proposal masks. Based on these weights instance segmentation results are generated for objects.

(2) To address the problem in identifying infrequent relationships from the long-

tailed relationship categories in visual relationship recognition, a Progressive Knowledge-driven Transformer (PKT) is proposed to utilize knowledge priors as external supervision to restrain the connection between object regions and force the feature learning focus on more relevant regions. To make the model compatible with common sense knowledge from different domains in real world, a Configurable Graph Reasoning (CGR) technique is proposed to prevent feature learning from the interference of redundant and noisy knowledge connections. The technique decomposes the traditional reasoning path into multiple sub paths, and learns to dynamically compose knowledge-enhanced reasoning paths, adaptively performs knowledge enhancement for each relation features; thus improving the accuracy of relationship prediction. When the labeled examples of relationship classes are extremely rare, the model training is insufficient and the sample deviation is large, which greatly reduces the robustness of the model. We further propose a Scene Graph Attack technique, which promotes the model to learn to defend the global semantics from the semantic attack on nodes, and improves the model precision and generalization.

(3) To solve the cross-modal information matching problem when combining visual scene understanding and natural language, a Cross-modal Memory Network (CMN) is proposed to sequentially associate the visual scenes with the dialogue sentences in the visual dialogue navigation task. The network encodes historical information about previous actions for the navigator. The rich cross-modal memory information of decision-making helps the current action decision.

This dissertation proposes an integrated coarse localization framework and an instance-level segmentation localization framework for the weakly-supervised object localization problem, proposes a feature transformer that can progressively incorporate different types of knowledge, a graph reasoning technique that can select knowledge from large knowledge bases, and an adversarial attack technique that attacks scene semantic graphs for the visual relationship detection problem, proposes a new idea for the problem of sparse and inaccurate labeling of training samples in the deep learning framework, provides new research ideas and directions for real-world scene understanding.

Keywords: Limited Supervision, Scene Understanding, Visual Object Detection, Visual Relationship Detection, Graph Model

目 录

图形列表

表格列表

第 1 章　绪论

1.1　研究背景和意义

计算机视觉中的场景理解问题致力于让机器能够像人一样理解场景图像的语义信息。场景理解在自动驾驶、视觉问答系统、视觉语言导航等多个应用场景中有着广泛的应用，是构成智能视觉平台和系统的关键性技术。要从全局角度理解图像中所表达的视觉语义信息，除了识别和定位场景中的视觉目标以外，还需进一步识别和推理目标之间的语义关联。

目标定位是场景理解中的一个基础问题，是众多视觉研究和应用的重要前提。目标定位是目标识别问题中的一类具体任务，是指识别出图片中所包含的目标类别标签并给出目标在图片中的坐标位置。目标定位结果的信息精细程度不同，定位形式也不同，包括锚点定位、框定位和分割定位，如图1.1所示。视觉目标的定位结果越精细，其定位难度越大。处于自然场景中的视觉目标其外观非常多样化，且因为受不同环境中光照、遮挡和嘈杂背景的影响而变得难以识别。

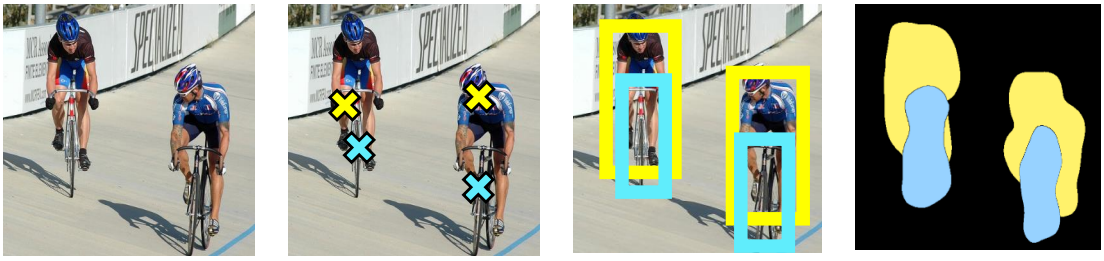


图 1.1 视觉目标定位结果示例

Figure 1.1 Illustration of visual object localization results

近年来，深度学习技术在计算机视觉领域取得了巨大的成功，在经典视觉任务如图片分类、目标识别和定位等任务上的性能大幅提升。基础视觉识别任务上的突破性进展促使研究者们开始思考如何让机器拥有更智能、更抽象的视觉认知，能够像人类一样理解当前所处的场景，与环境或人类互动并辅助人类做出决策。基于目标识别和定位的结果进一步推理目标间的关系，这些视觉关系能够将整个场景中彼此独立的视觉目标关联起来，构建一种结构化的场景表示，即场景图，其中节点代表目标，边代表目标间的语义关系，如图1.2所示。视觉关系检

测任务要求预测目标标签、目标框坐标以及目标间的语义关系类别。场景图描述了视觉目标间的互动关系，为场景理解的下游任务如图像检索、图像描述、视觉问答和视觉语言导航等提供了更加紧致和准确的场景信息表示。

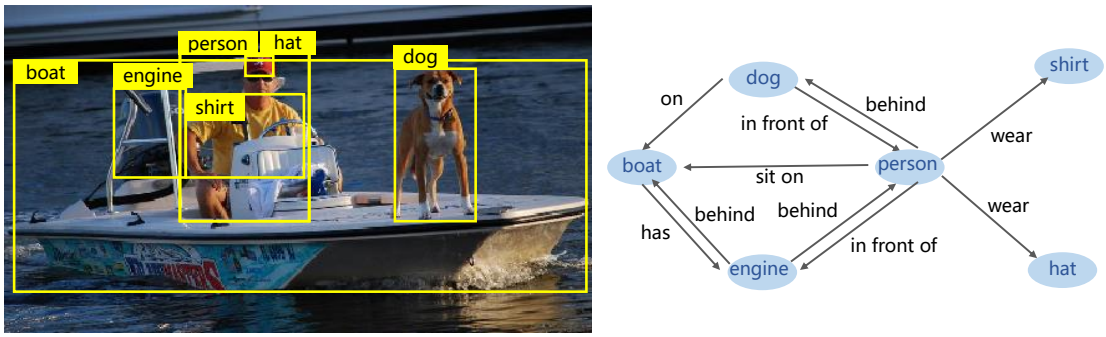


图 1.2 视觉关系检测结果示例

Figure 1.2 Illustration of visual relationship detection results

基于深度学习技术的方法和模型对训练数据有较高的要求，需要大量的标注数据来训练网络参数以保证模型性能。随着互联网的发展，图像视频数据急剧增长，基于这些数据以全监督方式训练深度学习模型需要消耗大量的人工标注成本。使用精细标注的数据训练的模型可能由于过拟合的原因导致泛化性较差，难以在真实场景中被广泛地应用。

人类学习认识世界中的事物只需要少量的样本和粗略的指导，并结合已有知识或先验信息增强认知。使用大量精确标注的数据训练深度神经网络模型并不符合人类视觉认知的规律。大量的科研人员致力于研究视觉认知规律和原理，针对实际应用场景设计新的神经网络模型，以实现以下目标：（1）在监督信息有限的条件下进行特征与模型学习，（2）从海量未标注数据中挖掘可辅助模型学习的信息，（3）提升模型的计算效率和泛化性。

本文涉及的有限监督存在两层含义：（1）监督信息的精细程度有限，如无监督学习、弱监督学习。弱监督目标定位任务要求模型在仅有图片级别的标注的情况下学习识别和定位图片中的目标；（2）可用的监督信息的数量有限，如半监督学习、少样本学习。视觉关系检测任务中，关系类别在真实世界中呈长尾分布，不常见类往往只有极少量的训练样本可用于模型学习。

有限监督条件下的目标定位及关系推理模型需要在标注数据有限或不够准确的情况下学习定位视觉目标，捕捉目标间的语义关系，构建更抽象、更全面的

场景语义表达。弱监督目标定位算法可以以更少的标注代价训练模型以定位视觉目标，其所需的数据和图片级标签可从互联网中轻易获取。但是由于图片级标注信息不够精确，使得弱监督模型的训练更具挑战性。已有许多工作致力于解决该问题，但其方法的性能距离全监督模型依然有着明显的差距。在当前网络数据极速增长的时代，智能平台对弱监督算法的需求将会更加紧迫。视觉关系检测基于目标识别和定位的结果，预测每对目标之间的语义关系类别。视觉关系类别在真实场景中的分布极不均匀，导致模型无法基于较少的训练样本充分学习关系类别。现有方法致力于引入先验知识来辅助训练，但同时也容易引入数据集偏差，使得模型性能和泛化性受限。随着智能设备在生活中的普遍应用，如何在复杂多变的实际场景中准确地推理视觉目标间的关系，对视觉场景进行更抽象、更全面的理解变得尤为重要。

1.2　国内外研究现状

1.2.1　研究现状

弱监督目标定位方法通常使用分步策略，即首先提取目标候选框[1–3]，然后对候选框进行分类和打分[4–7]。最早的目标候选框提取算法通过在不同尺度的图像上逐像素地滑动窗口来提取目标的定位候选框，这种方法产生百万量级的候选框，可以保证较高的查全率，但是由于候选框数量过多，导致候选框特征提取和分类模型的计算效率大大降低。为了提升模型效率，Selective Search[1]、EdgeBoxes[2] 和 MCG (Multiscale Combinatorial Grouping)[3] 等方法基于图像的颜色、纹理、边缘等底层视觉信息生成目标候选框，在保证查全率的同时减少候选框数目至两千个左右。该类目标候选框提取方法不需要使用任何标注信息，在目标定位和检测任务中得到了广泛的应用。基于目标候选框的弱监督目标定位，一种代表性的方法是 WSDDN (Weakly Supervised Deep Detection Network)[6]，该方法通过联合学习预选框和目标分类来显著提高性能，在此基础上，ContextLoc[7] 试图在学习过程中扩展或收缩固定的预选框以利用目标周围的上下文信息，从而提升弱监督目标定位的准确度。ProNet[8] 使用多个并行卷积神经网络进行多尺度缩放，以预测可能的目标区域，然后通过级联网络对这些预选区域进行分类。另外一些方法基于深度神经网络设计端到端的弱监督目标定位框架。其主要思想是基于图像分类网络进行改进，去除全连接层以保留深度特征的空间结构

性，使用特征图的最高响应[9]或均值响应[10]预测图像标签，不同的类别响应图上被激活的图像区域即为该类别的目标定位结果。此外，通过对深度卷积神经网络的滤波器进行分析，可以根据层间激活过程在网络输出的二维特征图上定位目标的判别性区域，从而获取目标锚点定位和框定位结果。对于更具挑战性的分割定位任务，其相关研究依然在初期探索阶段。BoxSup[11]使用目标框生成的二值化掩膜作为分割的弱监督信息，构造的伪监督掩膜用来训练全监督实例分割模型。PRM (Peak Response Map)[12]方法尝试只使用图像级标签来学习预测目标实例分割结果，提出利用分类网络获取类峰值响应来提取实例感知的视觉线索，并利用该线索从预选掩膜中检索目标实例分割掩膜。

视觉关系检测作为计算机视觉中最具挑战性的问题之一，经典的思路是首先基于图像构建全连接图，将视觉目标（节点）根据它们之间的视觉关系（边）连接起来，然后在图结构上迭代地进行消息传播以推断目标和关系的标签。为了降低传播过程在密集连接图上的运算成本，现有工作通过计算关系预选[13]，将图划分为子图以进行层次化推理[14]，或使用诸如树结构[15]之类的先验结构来初始化目标节点之间的连接。CMAT (Counterfactual Critic Multi-Agent Training)[16]将场景图生成问题建模为对于全连接场景图上的节点和边的一个序列的选择过程。VG (Visual Genome) 数据集上的统计分析[17]显示，视觉场景中的目标具有很强的结构规律性，并且在已知目标对的情况下，关系类别是高度可预测的，例如 “person” 和 “bike” 的关系通常更可能是 “ride” 或 “next to”。基于以上观测，Motifs[17]和 KERN (Knowledge-Embedded Routing Network)[18]研究如何利用目标对和视觉关系类别之间的统计关联来辅助模型做出正确的预测。这些工作预先假设了一个关系推理路径，即{主体（subject），客体（object）→ 谓词（predicate）}，并使用统计先验来辅助推理。该统计信息通常来源于模型的训练集，在容量和通用性上和来自真实场景中的视觉关系知识存在差距。为了提升先验知识的有效性、真实性和通用性，一些研究人员探索从维基百科等大量数据中收集外部知识，并通过知识蒸馏的方式将这些知识引入到关系检测模型的学习过程中[19]。KB-GAN[20]使用目标的预测标签从大规模语言数据 ConceptNet[21]中检索与当前目标可能关联的视觉关系的类别分布，并通过注意力机制对检索结果进行筛选，与关系特征进行融合，从而促进视觉关系的推理。

1.2.2　存在的问题

真实世界中的场景理解任务其监督信息往往十分有限，模糊的标注信息容易导致模型学习陷入局部最优，稀少的标注信息容易导致模型学习不充分。基于图像分类网络的弱监督目标定位框架只有图片级标签可用于监督模型训练，其网络生成的目标响应往往只包含对分类有判别性的目标局部区域，而非目标的完整范围。实际场景中目标间的视觉关系往往十分多样化，且类别分布极度不均匀，常见的关系类有充足的样本示例供模型训练，不常见类的样本示例则非常稀少。虽然使用数据集先验统计信息能够一定程度上缓解该问题，但是数据集类别标签定义以及分布的偏差常常导致所学关系识别模型泛化性较差。

1.3　本文主要研究内容与贡献

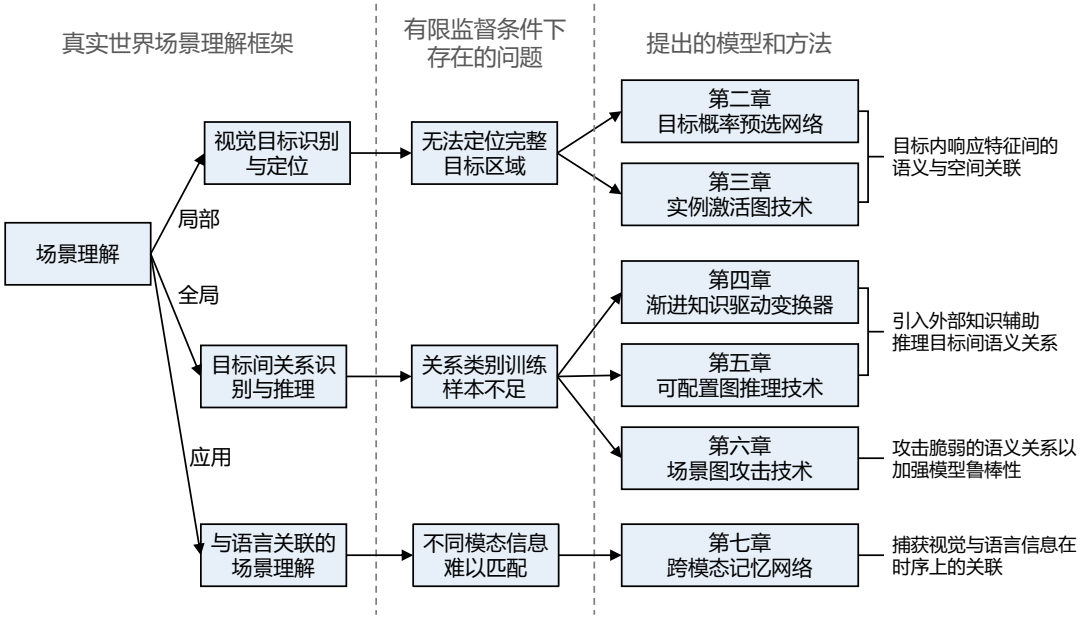


图 1.3 本文研究内容关系图

Figure 1.3 The relationship between the research content of this thesis

本文的研究内容如图1.3所示。真实世界的场景理解往往面临训练样本不足、标注信息不准确等问题，本文从局部识别、全局理解和实际应用三个方面，从目标区域特征关联与信息传播学习的角度，研究有限监督条件下场景理解面临的问题。对于场景理解中的视觉目标定位任务，本文致力于解决其在仅有图像级监督信息时无法定位完整目标区域的问题，针对性地提出目标概率预选网络和实

例激活图技术；对于目标间视觉关系推理任务，本文致力于解决其在关系类别训练样本不足的情况下无法进行准确预测的问题，针对性地提出渐近知识驱动变换器和可配置图推理技术以引入外部知识辅助模型的学习，提出场景图攻击技术以增强视觉关系检测模型的鲁棒性。在实际应用中，对于视觉场景理解与语言信息难以匹配的问题，提出跨模态记忆网络，从时序角度捕获跨模态信息之间的关联。本文的主要创新点包括以下几个方面：

（1）提出了一种目标概率预选模块，用于弱监督目标定位任务。基于随机游走算法，根据目标在深度特征图上的响应之间的语义和空间相似性在目标区域内进行置信度传播，从而激活更完整的目标区域。该模块可以插入在深度卷积神经网络中的任意层，其运算的时间代价极小，几乎可忽略不计。该模块与原始网络联合优化，引导网络学习聚焦在目标置信度更高的区域，从而提升目标分类、锚点定位和框定位结果。

（2）提出了一种实例激活图技术，用于弱监督目标实例分割任务。基于带噪声的预提取候选掩膜，学习填充稀疏且不完整的实例响应，从而得到精确的实例分割掩膜。该方法致力于总结和捕获目标内部像素之间类无关的关联和一致性，所学填充权值可直接以无监督方式应用于其他数据集和任务。

（3）提出了一种渐近知识驱动变换器，用于视觉关系检测任务。基于堆叠的多头注意力模块，引入一系列多样化的外部知识引导图，使得模型可以自适应地学习目标区域间的语义关联。外部知识引导目标特征编码更多地关注与其相关的其他区域，从而促进模型对不频繁的关系类别的预测。

（4）提出了一种可配置图推理技术，用于视觉关系检测任务。将传统的目标关系推理路径进行分解，得到多个二元子推理路径。该技术对每个子推理路径基于常识知识图进行增强，通过可学习的配置模块对每张图片选取并组合得到更具决定性的推理路径，从而更加准确地预测视觉关系。

（5）提出了一种场景图攻击技术，用于少样本视觉关系检测任务。该技术攻击场景图中的目标节点使其被模型预测为错误的类别，同时模型通过全局语义约束来抵御场景图攻击。该攻击促使模型学习在保持原场景图语义不变的情况下，产生新的关系变种。新的关系可用于辅助更新未标记数据的伪标签信息以促进模型训练，并且增强模型的鲁棒性。

（6）提出了一种跨模态记忆网络，用于视觉语言导航任务。基于注意力机制，

将真实场景的视觉信息和交互产生的语言信息进行关联，编码时序信息，捕获历史动作决策的状态，从而辅助当前场景中的导航动作预测。

1.4　本文的组织结构

第一章，绪论。论述有限监督条件下视觉目标定位和关系推理的研究背景和意义，分析当前有限监督框架下视觉目标识别和关系推理方法的不足之处，明确本文的主要研究内容和贡献。

第二章，目标概率预选模块。首先分析弱监督目标定位任务中难以定位完整目标区域的问题，并指出其原因在于分类网络只能激活具有判别力的目标部件。根据同个目标内的特征响应存在语义与空间相似性的特性，提出基于深度特征图相似性传播的目标概率预选模块，可与图像分类网络联合优化以激活目标区域，并给出实验分析和验证。

第三章，实例激活图技术。首先分析现有弱监督实例分割任务中无法预测精确的目标实例掩膜的问题，以及过度依赖预选掩膜带来的精度和效率的问题，提出基于类无关目标置信度传播的实例激活图技术，从含有噪声的预选掩膜中总结学习将局部视觉线索填充为一个完整的实例分割掩膜，设计了更加高效的弱监督实例分割框架，并在最后对该框架进行了充分的实验分析。

第四章，渐近知识驱动变换器。首先论述现有目标关系检测方法面临的关系类别的长尾分布带来的难以充分学习不常见关系的问题，随后指出现有方法在引入外部知识时的局限性；然后基于目标区域类别间的共现关联，提出渐近知识驱动变换器，以灵活且高效的形式向特征编码过程中注入视觉常识知识，并对模型的性能和耗时进行了全面的评估和分析。

第五章，可配置图推理技术。首先论述现有方法遵循的固定的视觉关系推理路径带来的学习偏差的问题，提出将推理路径分解为多个单依赖子路径，基于常识知识连接通过图传播更新特征表达，再动态挑选有效的推理路径，最终对模型进行实验分析和验证。

第六章，场景图攻击技术。首先分析在少样本标记下视觉关系检测面临的模型学习不充分而导致的泛化性差的问题，从未标记数据中挖掘训练样本，迭代地攻击变换已有关系示例并更新未标记样本的伪标签，学习保卫全局场景图语义不受局部节点攻击而损害。最终在多个不同的少样本学习设定下进行实验验证。

第七章，跨模态记忆网络。提出挖掘和编码历史视觉特征和语言特征之间的跨模态关联来促进视觉对话导航任务，这种记忆感知的特征编码包含了先前导航过程中的动作决策信息，有助于当前步的动作预测。最后，在 Matterport3D 模拟器上提供的真实环境中进行评估测试。

第八章，总结与展望。总结本文的主要研究内容，提出对未来工作方向的展望，包括开集关系识别问题、基于对比学习的少样本关系检测等。

第 2 章　基于概率目标预选的弱监督目标定位

基于图像分类网络的弱监督目标定位框架，其网络中的滤波器可以被看作是一种目标检测器[22]，网络中输出的深度特征图可被融合在一起用于生成类别响应图 CAM（Class Activation Map)[10]。该类别响应图指示出不同目标类别的判别性模式的空间分布，展示出其在弱监督条件下定位目标的能力。如图2.1所示，在训练过程中如果没有先验的目标区域知识，传统的卷积神经网络在目标区域的激活和定位任务中存在以下三个问题：1）易受背景噪声干扰，例如，目标“cow” 的定位结果易受背景 “grass” 干扰；2）易受共现模式的干扰，例如，“rail” 与目标 “train” 常常共同出现，被误分为 “train”；3）难以定位完整的目标区域，例如，目标 “person” 的定位结果往往只包含头部区域。

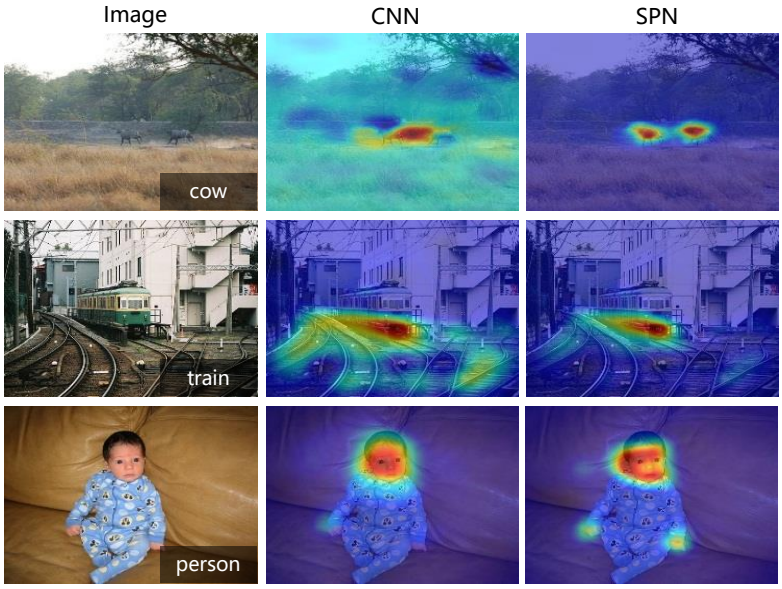


图 2.1 CNN 和 SPN 生成的目标类别响应图可视化

Figure 2.1 Visualization of Class Activation Maps (CAM) for generic CNN and SPN

2.1　模型概述

本章提出目标概率预选模块 SP（Soft Proposal），该模块通过计算目标预选图来引导网络的学习聚焦在目标区域，并抑制共现噪声的干扰，使得模型能够激活更精细、更完整的目标区域，如 “hand” 对于类别 “person”。SP 模块提取的目

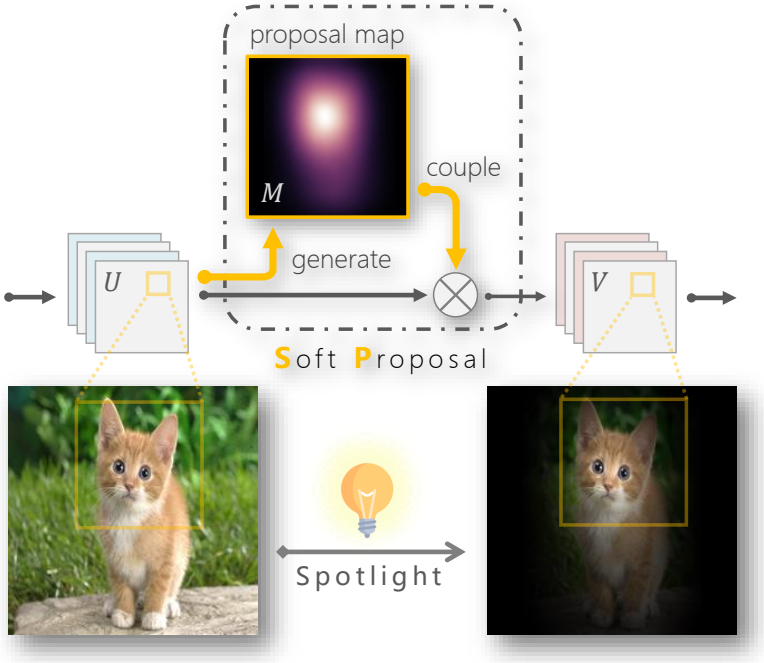


图 2.2 SP 模块示意图

Figure 2.2 Illustration of the proposed Soft Proposal (SP) module

标预选是一种二维的响应图，该图中的每个像素值代表对应感受野区域的目标置信度得分。SP 模块生成的目标预选以概率方式与深度特征图相结合，这不仅避免了阈值调整，而且可以汇总所有精细的响应值信息以提高精度。SP 模块可以被插入在经典卷积神经网络结构的任意位置，将 SP 模块应用于经典的 CNN（包括 CNN-S，VGG 和 GoogLeNet），可得到 SPN（Soft Proposal Network）。SPN 只需使用图片级别的标注信息来训练，在训练过程中，目标预选图迭代地更新并被映射回深度特征图上，进一步地和网络参数进行联合优化。SPN 可以学习更好的以目标为中心的卷积滤波器，并激活更完整的视觉目标区域，从而提升弱监督的目标定位性能。

2.2 基于概率目标预选的特征学习

2.2.1 生成概率目标预选

目标概率预选 $M \in \mathbb{R}^{N\times N}$ 是一种反映目标置信度的响应图，由 SP 模块基于深度特征图生成，生成过程见图 2.3。SP 模块可以被插入在神经网络任意层之后，不失一般性，此处我们假设插入在第 l 个卷积层之后。定义 $U^l \in \mathbb{R}^{K\times N\times N}$ 为第 l 个卷积层输出的深度特征图，其中 K 为特征图的通道数，$N\times N$ 代表特征图的

空间尺寸。特征图 U^l 上每个位置 (i,j) 对应一个深度特征向量 $\mathbf{u}_{ij}^l = U_{\cdot,i,j}^l \in \mathbb{R}^K$，由该位置在 U^l 中的所有 K 个通道的响应值组成。为了生成目标概率预选图 M，我们首先定义一个全连接有向图 G，用于连接 U^l 上的每个位置节点，权值矩阵为 $D \in \mathbb{R}^{N^2 \times N^2}$，其中 $D_{iN+j,pN+q}$ 代表节点 (i,j) 到节点 (p,q) 的边的权值。对于权值矩阵 D 的计算，我们考虑从两种角度去衡量目标预选：（1）属于同个目标类别的图片区域有着更相似的深度特征；（2）空间上的近邻区域更可能在语义上存在相关性。我们通过将特征差异和空间距离相结合的相似度度量来反映目标置信度：

$$\begin{aligned} D'_{iN+j,pN+q} &\triangleq \|\mathbf{u}_{ij}^l - \mathbf{u}_{pq}^l\| \cdot L(i-p, j-q), \\ L(a,b) &\triangleq exp(-\frac{a^2+b^2}{2\epsilon^2}), \end{aligned} \tag{2.1}$$

其中 ϵ 被设为 $0.15N$。进一步地，再对每个节点发出的所有边的权值进行归一化：

$$D_{a,b} = \frac{D'_{a,b}}{\sum_{a=1}^{N} D'_{a,b}}. \tag{2.2}$$

基于权值矩阵 D，SP 模块通过随机游走算法来生成目标预选图 M。随机游走算法会迭代地在与周围环境差异较大的节点上累积目标置信度。一个节点从进入的有向边接收置信度，沿着出去的边散播置信度，因此一个节点的目标置信度可以沿着边流动到其他节点，如图2.3所示。

为了便于使用随机游走算法进行计算，我们首先将二维的目标预选概率图 M 重组为包含 N^2 个元素的向量，并将向量值初始化为 $\frac{1}{N^2}$。M 通过权重矩阵 D 的迭代乘积进行更新：

$$M \leftarrow D \times M. \tag{2.3}$$

公式 2.3中的迭代更新过程是特征向量集中度度量（eigenvector centrality measure）[23] 的变体，它输出一个预选图表示深层特征图上每个位置的目标概率置信度。在网络学习过程中，权重矩阵 D 是以深度特征图 U^l 为基础，而 U^l 则以第 l 层的卷积层参数 W^l 为基础。为了显示这种相互依赖性，可将公式2.3更新为：

$$M \leftarrow D\big(U^l(W^l)\big) \times M. \tag{2.4}$$

给定深度特征图 U，公式2.4通常在大约十次迭代后达到其稳定状态，其输出 M 向量被重组为二维预选图 $M \in \mathbb{R}^{N \times N}$。

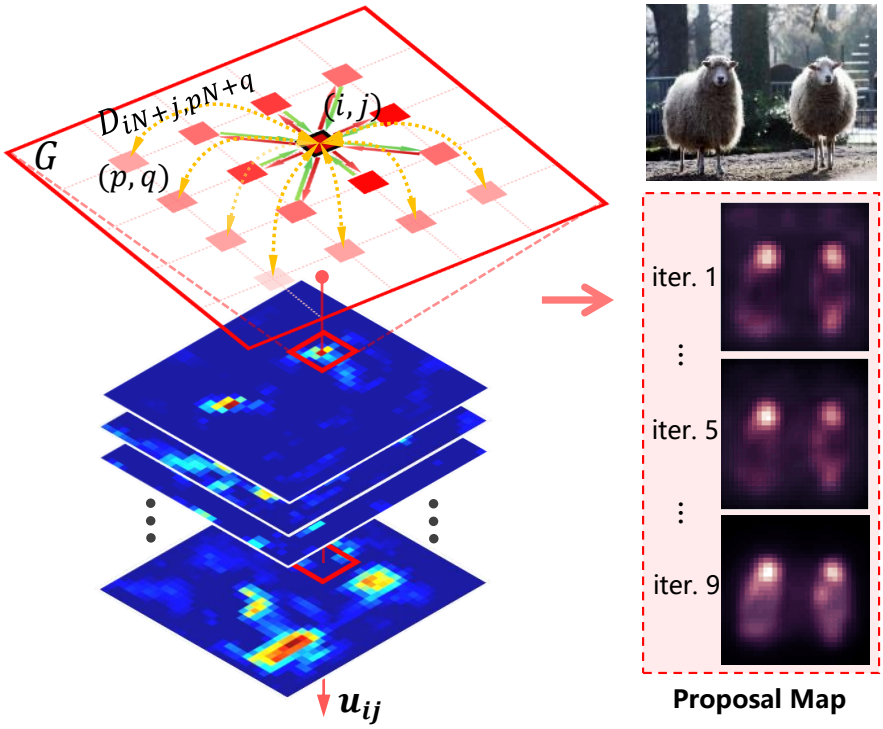


图 2.3 目标预选生成过程示意图

Figure 2.3 Soft Proposal Generation in a single SPN feedforward pass

2.2.2 概率目标预选与深度特征融合

用深度特征图以弱监督的方式生成的预选图可以看作是一种目标概率图，它表示可能的目标区域。从图像表示的角度来看，该预选图着重于“感兴趣的目标区域”，这些区域有助于图像分类。M 可以通过 SP 模块集成到深度卷积神经网络的端到端学习中，如图2.2，以汇总来自深度响应的特定于图像的判别信息。

在 SPN 的前向传播过程中，深度特征图 U 和 SP 模块计算所得的目标预选图 M 通过 Hadamard 积运算进行融合，得到 $V \in \mathbb{R}^{N\times N}$：

$$V_k = U_k^l(W^l) \circ M, k=1,2,...,K, \tag{2.5}$$

其中，下标 k 为通道索引，“$\circ$” 代表逐元素相乘。融合所得的深度特征图 V 将继续向前传播，以预测目标分类结果 $y \in \mathbb{R}^C$，C 为目标类别数。接着根据数据集标注的图片标签 t 为每个样本计算分类预测误差 $E = \ell(y,t)$，$\ell(\cdot)$ 为交叉熵损失函数。在 SPN 的反向传播过程中，梯度通过目标概率预选图来进行重分配：

$$\begin{aligned} W^l &= W^l + \Delta W(M) \\ \Delta W(M) &= -\eta \frac{\partial E}{\partial W^l}(M) \end{aligned} \tag{2.6}$$

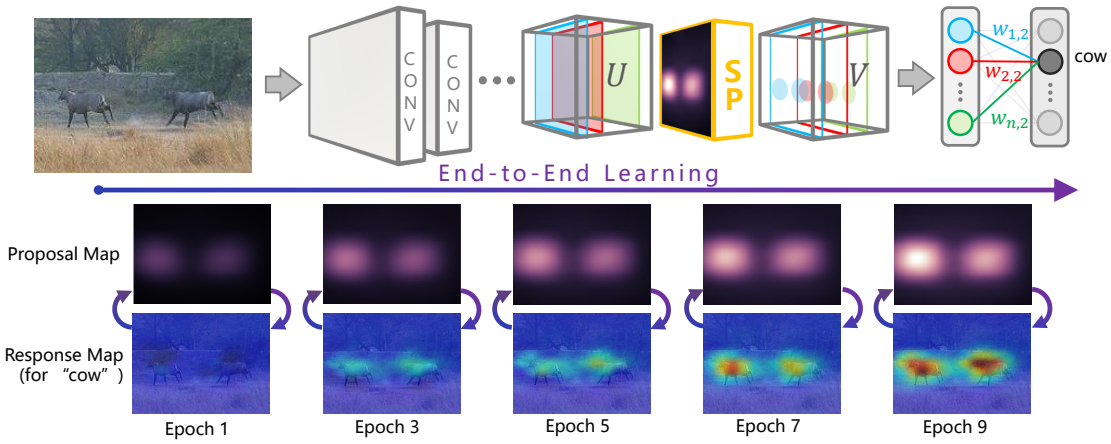


图 2.4 SPN 学习过程示意图

Figure 2.4 Illustration of the learning process of Soft Proposal Network (SPN)

其中，η 为神经网络的学习率，$\Delta W(M)$ 表示参数 W^l 依赖 M，因为卷积层滤波器的梯度 $\frac{\partial E}{\partial W^l}$ 依赖 M，如公式2.9所示。由于 W^l 依赖于 M，SPN 可以学习到含有丰富信息的图片区域，并且抑制背景噪声。结合公式2.4定义的目标概率预选生成，公式2.5定义的目标概率预选融合，和公式2.6定义的反向传播过程，可以清楚地看到，U^l，W^l，和 M 彼此互相依赖。在网络训练期间，卷积滤波器参数 W^l 基于公式2.6进行更新，U^l 也会随之更新。接着，通过随机游走，预选图 M 会被更新。目标预选图 M 引导 SPN 的学习逐渐地聚焦在 U^l 中更可能是目标的区域，忽略背景噪声区域，从而有效地学习更具判别性的滤波器 W^l。目标预选图和卷积神经网络的联合学习过程如图2.4所示。

2.2.3　弱监督深度响应激活

SPN 通过首先使用空间池化层将深层特征图聚合到特征向量，然后使用完全连接层将这种特征向量连接到图像类别，来执行弱监督学习任务，如图 2.4所示。该网络架构使用来自网络末端目标分类任务的弱监督信息来激活潜在的目标区域。在 SPN 的前向传播过程中，预选图 M 由插入在第 l 个卷积层之后的 SP 模块生成，SP 模块的输入为第 l 个卷积层输出的深度特征图：

$$U_j^l = (\sum_{i \in S_j} U_i^{l-1} * W_{ij}^l + b_j^l) \circ M, \tag{2.7}$$

其中 S_j 是输入特征图的集合，b_j^l 是加性偏置，W_{ij}^l 是连接 U^{l-1} 中第 i 个输入图和 U^l 中第 j 个输出图的卷积滤波器。在 SPN 的反向传播过程中，误差通过 δ 从

算法 1 Learning Soft Proposal Network

Input: 带有分类标签的训练图片

Output: 网络参数，每幅图片的目标预选图

repeat
  初始化 M 中的每个元素值为 $\frac{1}{N^2}$
  repeat
    $M \leftarrow D\left(U^l(W^l)\right) \times M$
  until 达到稳定状态
  $V = U^l(W^l) \circ M$, 前向传播
  $W^l = W^l + \Delta W(M)$, 反向传播
  for 所有的卷积层 l **do**
    $U^l = W^l * U^{l-1}$
  end for
until 学习收敛

第 $l+1$ 层向第 l 层传播：

$$\begin{aligned}\delta^l = \frac{\partial E}{\partial U^l} &= \frac{\partial E}{\partial U^{l+1}} \frac{\partial U^{l+1}}{\partial U^l} \\ &= \delta^{l+1} \frac{\partial [(U^l * W^{l+1} + b^l) \circ M]}{\partial U^l} \\ &= \delta^{l+1} * W^{l+1} \circ M.\end{aligned} \tag{2.8}$$

预选图 M 使得网络学习聚焦在富含信息量的区域以及更值得学习和关注的位置。因为 M 会随着梯度 δ 进行流动，所以在 CNN 最顶层插入一个 SP 模块即可通过网络学习辅助优化整个网络的滤波器参数。在计算得到 δ^l 后，可计算得卷积滤波器的梯度：

$$\begin{aligned}\frac{\partial E}{\partial W_{ij}^l} &= \sum_{p,q} (\delta_j^l)_{pq} (\mathbf{x}_i^{l-1})_{pq} \\ &= \sum_{p,q} (\delta_j^{l+1} * W_{j\cdot}^{l+1})_{pq} M_{pq} (\mathbf{x}_i^{l-1})_{pq},\end{aligned} \tag{2.9}$$

计算对偏置的梯度：

$$\begin{aligned}\frac{\partial E}{\partial b_{ij}^l} &= \sum_{p,q} (\delta_j^l)_{pq} \\ &= \sum_{p,q} (\delta_j^{l+1} * W_{j\cdot}^{l+1})_{pq} M_{pq},\end{aligned} \tag{2.10}$$

其中 $W_{j\cdot}^{l+1}$ 表示第 $l+1$ 层参与了计算 U_j^l 的卷积滤波器参数，$(\mathbf{x}_i^{l-1})_{pq}$ 表示在 U_i^{l-1} 上以 (p,q) 为中心的窗口区域。根据公式 2.9和公式 2.10，表示目标置信度的预选图 M 和梯度图在弱监督激活过程中进行融合，驱动 SPN 来学习更加有用的视觉信息和模式。对于弱监督目标定位，我们为第 c 个类别计算响应图[10]：

$$R_c = \sum_k w_{k,c} \cdot \hat{U}_k \circ M \tag{2.11}$$

其中，$\hat{U}_k$ 是最后一个卷积层输出的第 k 个特征图，$w_{k,c}$ 是全连接层连接第 c 个输出节点和第 k 个特征向量的权值，如图 2.4所示。

2.3　实验结果及分析

我们将 SP 模块插入到现有的 CNN 网络（如 VGG，GoogLeNet）中构成 SPN，并在一系列基准上对模型进行评估。首先，我们将 SP 模块生成的目标概率预选和传统目标预选生成方法进行比较，以评估 SPN 所生成的目标预选的质量以及 SP 模块引入的时间开销。接着，通过弱监督锚点定位任务评估 SPN 学习以目标为中心的滤波器并生成精确的目标响应图的能力。进一步地，我们通过弱监督框定位任务来评估模型发现更精细更全面的视觉线索的能力。最后，通过大规模图像分类任务，评估模型学习更具判别力的目标分类视觉证据的能力。我们使用随机梯度下降优化器和交叉熵损失函数来训练 SPN 模型。训练过程中，权重衰减值设置为 0.0005，动量设置为 0.9，初始学习率设置为 0.01。

2.3.1　目标预选的质量

在 VOC2007 数据集上，我们定义一个目标能量度量值来评估生成的预选质量。主要和传统的目标预选生成方法进行对比，如 Selective Search[1]，Edge-Boxes[2] 和 RPN（Region Proposal Network）[24]。对于这些传统方法，我们需要将一系列目标预选框的坐标转换为预选图的形式。转换时，预选图的每个像素的能量值为包含该像素的所有预选框的置信度得分之和。对于 SPN，可将生成的预选图上采样到原图大小以直接获得预选图，如图2.5，第一行为输入图片，第二行为目标预选图叠加在图片上。第三行显示预选图上得分较高的前 100 个位置对应在原图上的感受野区域。将预选图进行标准化后，计算其落在真值标注框内的能量得分。由目标能量度量值的定义可知，该能量值的范围为 $[0.0, 1.0]$。如表2.1所示，RPN 和 SPN 的测试是在 NVIDIA Tesla K80 GPU 上进行测试。由于

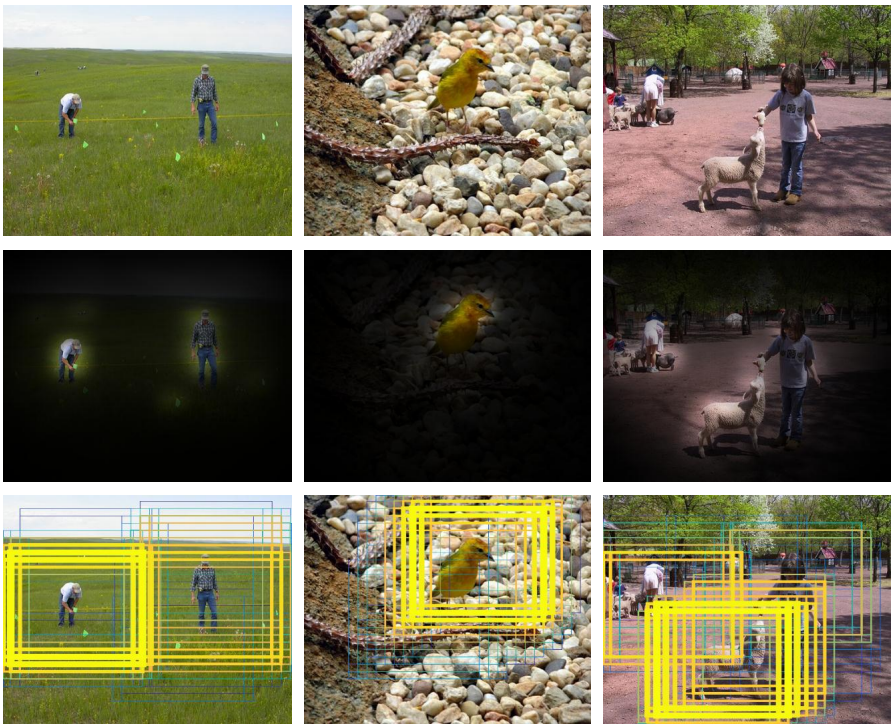


图 2.5 SP 模块生成的目标预选示例

Figure 2.5 Proposal examples generated by Soft Proposal module

算法复杂性和并行性的限制，Selective Search 和 EdgeBoxes 暂时只能在 CPU 上进行测试。SPN 生成的预选图质量高于传统手工设计方法如 Selective Search 和 EdgeBoxes，并与全监督的学习方法 RPN 性能差距较小。然而，SPN 的时间优势非常明显，远远小于所比较的方法。SPN 的实现是基于 GPU 并行化计算，简单且高效。从表中可以看到 SPN 生成预选图的代价几乎可以忽略不计。如图2.6(a) 所示，相比于 Selective Search 和 EdgeBoxes，SPN 能够使网络学习更好地聚焦在小目标区域，尽管预选图的分辨率较低。如图2.6(b) 所示，SPN 生成的预选图能够迭代地进化，并且在训练过程中和原网络参数进行联合优化。

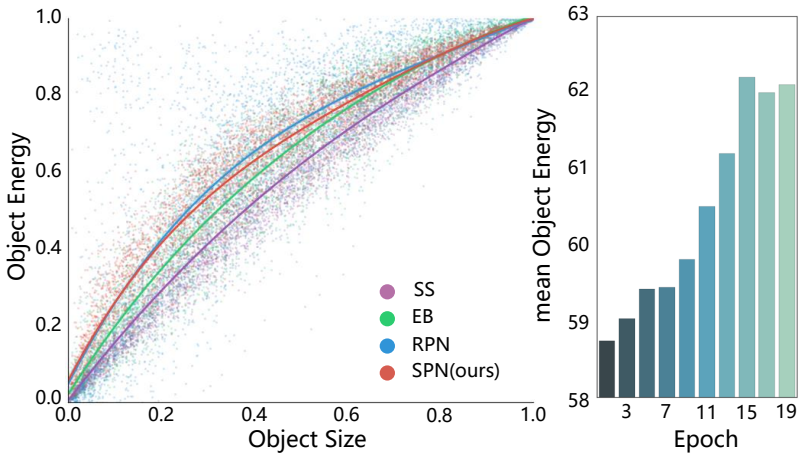
2.3.2 弱监督锚点定位

不预测锚点标签：为了评估 SPN 是否可以学习到更加具有判别力的卷积滤波器进而生成更加精确的响应图，我们在弱监督锚点定位任务上进一步测试训练的模型。我们选择三种常用的 CNN 结构，包括 CNN-S[25]，VGG16[26]，和 GoogLeNet[27]，并在其卷积层的最后一层之后插入 SP 模块，从而将它们更新为 SPN。所有的 SPN 都在 VOC2007 训练集上使用相同的超参进行微调，然后我

Method	ObjectEnergy(%)	Time(ms)
Selective Search [1]	53.7	2000
EdgeBoxes [2]	58.8	200
RPN (supervised) [24]	**63.3**	10.5
SPN (weakly supervised)	62.2	**0.9**

表 2.1 基于 VOC2007 测试集的目标预选图质量评估

Table 2.1 Proposal quality evaluation on VOC2007 test set



(a) Object Energy distribution　(b) Object Energy evolution

图 2.6 目标置信度能量值统计

Figure 2.6 Statistical analysis of Object Energy

们使用真值标注标签来计算每个类的响应图。我们参考 c-MWP[28] 计算锚点定位结果：如果响应最大值落在真值目标标注框内（允许 15 像素的误差），则记为一次命中（Hits），否则的话，记一次失误（Misses）。我们通过计算命中率 $Acc = \frac{Hits}{Hits+Misses}$ 来度量每个类的定位准确度，最终的锚点定位结果为每个类定位准确度的平均值。对于 VOC2007 数据集，我们使用两种测试集，**All** 代表使用整个测试集，和 **Diff.** 表示使用一个子集，其中的图片包含多个目标类别并且包含小目标（目标区域小于图片面积的四分之一）。如表2.2所示，我们将常见的 CNN 更新为 SPN 可以带来显著的性能提升。具体地，SP-VGGNet 在 **All** 集合上超过 c-MWP 的性能达 7.5% (87.5 % vs 80.0 %)，在 **Diff.** 上高出 11.3% (78.1% vs 66.8%)。SP-GoogLeNet 在 **All** 和 **Diff.** 上分别比 c-MWP 高出 3.1% and 6.8%。SPN

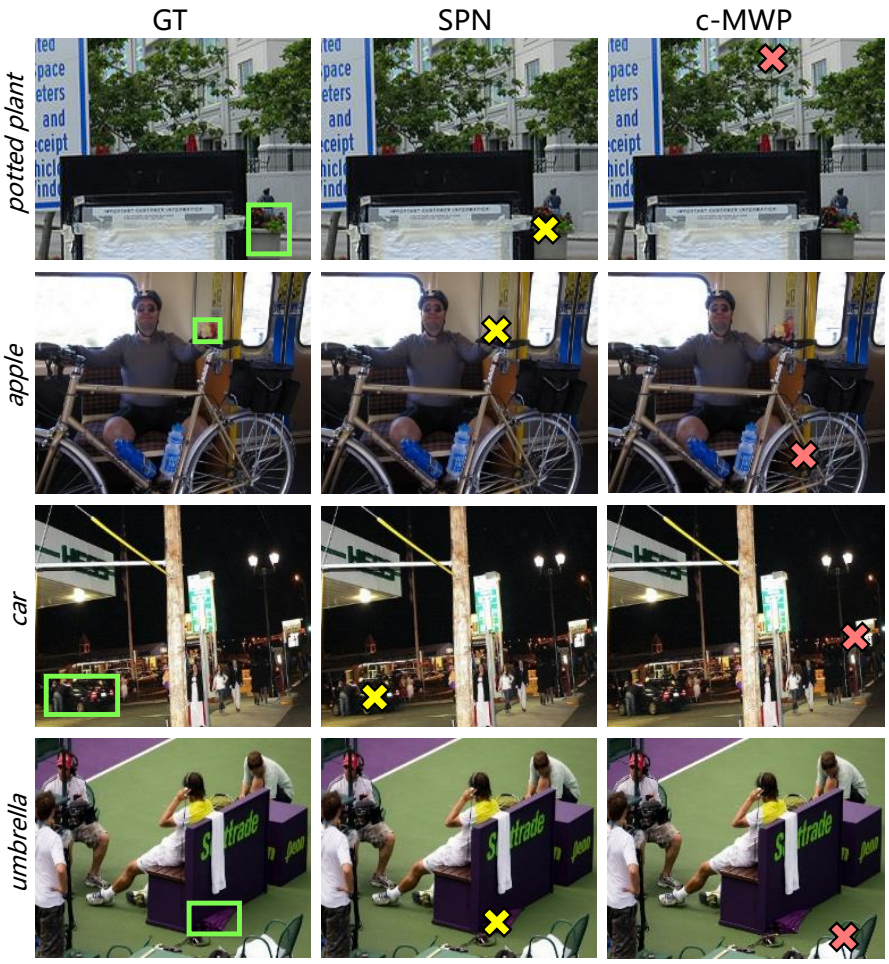


图 2.7 锚点定位结果示例

Figure 2.7 Examples of pointing localization

在锚点定位任务上的显著提升证实了 SP 模块的有效性，引导网络学习更加以目标为中心的滤波器，从而激活更加准确的目标区域。根据表2.2我们观测到：（1）SP-VGGNet 比 SP-GoogLeNet 有着更高的性能，这是因为 SP-VGGNet 的感受野区域小于 SP-GoogLeNet。当感受野区域间的交叠较少时，SP 模块中的目标置信度传播才更加能够发挥优势；（2）模型在 **Diff.** 集合上的性能提升大于在 **All** 集合上的，证实了 SPN 在混杂的场景中提取预选图的能力。

预测锚点标签：进一步地，我们在更具挑战的预测锚点标签任务上测试 SPN。该任务要求网络不仅正确预测目标类别是否存在，而且正确预测目标的锚点定位结果，即最大响应值点落在目标框内（容忍度为 18 像素）[9]。我们将预训练的 VGG16 升级为 SPN，并在 VOC2012 和 COCO2014 数据集上微调 20 步，结果见表2.3。在未使用多尺度设定的情况下，SPN 的性能显著优于目前最好方法[32]，

Method	CNN-S	VGG16	GoogLeNet
Center	69.5/42.6	69.5/42.6	69.5/42.6
Grad[29]	78.6/59.8	76.0/56.8	79.3/61.4
Deconv[30]	73.1/45.9	75.5/52.8	74.3/49.4
LRP[31]	68.1/41.3	-	72.8/50.2
CAM[10]	-	-	80.8/61.9
MWP[28]	73.7/52.9	76.9/55.1	79.3/60.4
c-MWP[28]	78.7/61.7	80.0/66.8	85.1/72.3
SPN	**81.8/66.7**	**87.5/78.1**	**88.2/79.1**

表 2.2 VOC2007 测试集 (All/Diff.) 上弱监督锚点定位结果

Table 2.2 Pointing localization accuracy (%) on VOC2007 test set (All/Diff.)

Method	mAP (%)	
Dataset	VOC	COCO
Oquab *et al.*[9]	74.5	41.2
Sun *et al.*[8]	74.8	43.5
Bency[32]	77.1	49.2
SPN	**82.9**	**55.3**

表 2.3 VOC2012 和 COCO2014 校验集上的锚点定位结果

Table 2.3 Pointing localization mAP (%) results on VOC2012 and COCO 2014 val. set

在 VOC2012 高出 5.8% mAP，在 COCO2014 上高出 6% mAP。该评测结果证明了 SP 模块为原始 CNN 带来的更加准确的定位能力并且同时保持了原有的分类能力。在后面的实验中，我们将进一步验证 SPN 的分类能力。

2.3.3　弱监督目标框定位

尽管在学习期间没有目标使用任何目标级别的标注，SPN 模型依然能够在类别响应图的帮助下估计目标框。我们根据目标类别的真值标注计算类别响应图，并根据均值阈值将响应图转化为二值图。将二值图上采样回原始图像大小后，

图 2.8 VOC2007 测试集弱监督目标框定位结果

Figure 2.8 Bounding box localization results on the VOC2007 test set

Method	CAM	c-MWP	MWP	Fb	SPN
LocErr (%)	48.1	57.0	38.7	38.8	**36.3**

表 2.4 ILSVRC2014 校验集弱监督目标框定位结果

Table 2.4 Bounding box localization errors on ILSVRC2014 val. set

对前景像素取最小包围框作为目标框的预测结果。我们使用定位错误率 LocErr（Localization Error）来评估弱监督目标框定位结果。如表2.4所示，SPN 模型的性能优于其他比较方法。从图2.8可以看出，已有方法倾向于激活每个目标类别最具判别力的区域，如人脸，而 SPN 通过学习目标置信度区域，引导网络发现更多精细的目标信息，如手和腿。对于目标类别沙发和桌子，SPN 的能够通过置信度传播激活更多的目标碎片区域。

2.3.4 图像分类

预测目标类别在图像中是否存在并不要求模型去精确地定位完整的目标区域，SP 模块生成目标预选图引导网络学习更加聚焦于目标区域可以帮助抑制

Method	ImageNet	COCO	VOC
GoogLeNetGAP[10]	35.0/13.2	54.4	83.4
SP-GoogLeNetGAP	**33.5/12.7**	**56.0**	**84.2**

表 2.5 图像分类结果对比

Table 2.5 Comparison of image classification results

背景噪声对学习的干扰，从而提升图像分类性能。我们使用 GoogLeNet 的一个简化网络，GoogLeNetGAP[10]，并将其更新为 SPN。我们使用随机梯度下降优化器训练模型，在 ILSVRC2014 数据集上训练 90 轮。如表2.5所示，第二列为在 ILSVRC2014 校验集上的分类错误率 top-1/top-5 error(%)，第三、四列为在 VOC2007 测试集和 COCO 校验集上的分类正确率 mAP(%)。插入 SP 模块后，图像分类性能提升了 1.5%，说明 SPN 能够学习更加富含信息量的特征表示。我们接着在 COCO2014 和 VOC2007 上分别训练 50 轮和 20 轮，并进行测试，结果显示，SP-GoogLeNetGAP 在 COCO2014 上提升 1.6%，在 VOC2007 提升更加显著，高达 4.5%。实验结果进一步说明了 SP 模块在弱监督框架下生成的目标概率预选在目标定位和分类任务上的有效性。

2.4　本章小结

本章我们提出了一种目标概率预选模块，该模块引导网络的学习聚焦在目标区域，抑制背景噪声和共现噪声的干扰，使得模型能够激活更精细、更完整的目标区域。我们将 SP 模块插入到传统的 CNN（例如 VGG 和 GoogLeNet）中，得到目标概率预选网络 SPN。在 SPN 中，模型基于深度特征图生成迭代进化的目标预选图，然后将其投影回去，引导卷积滤波器通过统一的学习过程发现更富含信息量的目标定位的视觉线索。实验结果显示，SPN 在弱监督的定位和分类任务上的性能明显优于已有方法，证明了 SPN 模型能够引导网络学习富含信息量的目标区域，激活更加完整的目标范围。

第 3 章　基于实例响应图学习的弱监督目标分割

弱监督目标锚点定位和框定位的结果可以为实际应用提供一定的辅助和指导，如 X 光安检，算法只需预警违禁品出现，并表明大致位置以辅助人类进一步检查确认。另一些应用则需要精确定位目标的边界，如自动驾驶。本章进一步研究弱监督实例分割任务，在使用图像级标注的情况下学习预测目标实例分割掩膜。PRM[12]（Peak Response Map）方法首先为每个类别生成响应图 CAM[10]，从而指示可用于鉴别目标类别的感受野区域。响应图上的峰值，即局部极值，被收集起来并反向解析，从而得到峰值响应图 PRM，用于高亮属于目标实例的富含信息的区域。PRM 虽然可以定位每个目标的判别性部件，但并不足以定位整个目标范围。如图3.1所示，PRM 高亮狗的头部而忽略了狗的身体区域，因为狗的头部相比于身体对分类任务来说是更具判别力的区域。

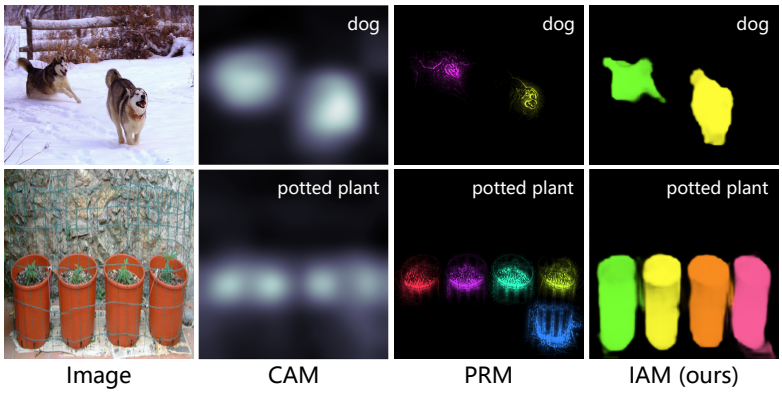


图 3.1 CAM，PRM，IAM 深度激活图对比

Figure 3.1 Comparison of the deep activation maps from CAM, PRM and IAM

3.1　模型概述

在本章，我们提出一种实例范围填充方法。该方法基于 PRM 生成实例激活图，致力于在弱监督框架下基于局部判别性响应学习填充实例完整范围。我们首先利用通过已有的 PRM 方法[12] 获得的不完整的区域响应，并基于 PRM 从嘈杂的预选掩膜中收集像素级别的伪监督信息。然后使用该伪监督信息来训练填充模型，该模型以 PRM 为输入，学习填充图像中对应的目标范围，从而得到实

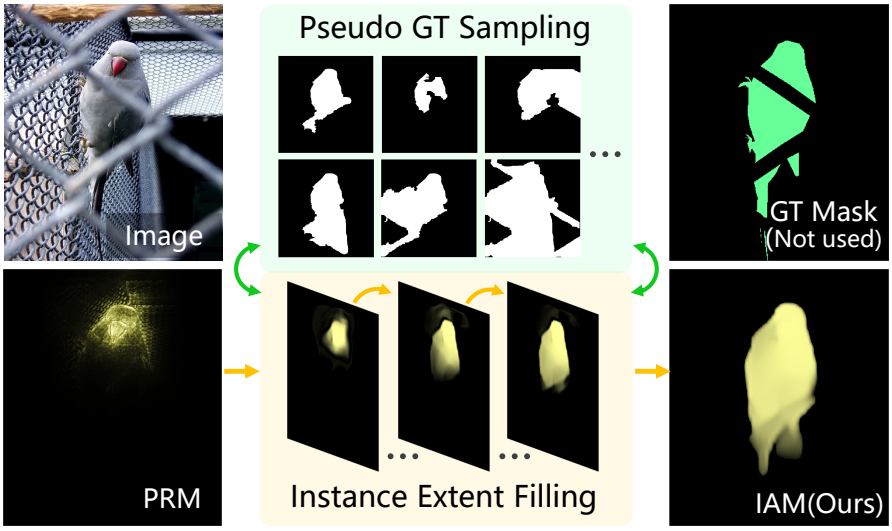


图 3.2 IAM 方法示意图

Figure 3.2 Illustration of Instance Activation Maps

例级别的激活图。实例激活图 IAM（Instance Activation Map）指定了空间布局和详细的实例边界，如图3.2所示。使用轻量级且对 GPU 友好的稠密条件随机场（DenseCRF）进行后处理修饰边缘后可直接得到实例掩膜结果。实验结果显示，IAM 显著提高了最新的弱监督实例分割性能，并将测试速度提升了一个数量级。

我们的方法从图像级标签和不精确的预提取目标预选掩膜中学习实例范围知识。实验表明，模型所学知识在其他域和未见过的目标类别中都有很好的泛化性。这使得本章所提出的方法能够扩展到许多其他与定位目标范围相关的视觉任务上。

3.2 实例响应图特征学习

在本节中，我们首先回顾已有工作 PRM[12]，并基于该工作从分类网络中提取局部的实例响应区域。接着，我们介绍本章所提的实例范围填充方法，包括伪监督信息的收集和填充模块的设计。最后，我们探讨所提方法的实质并说明模型实现细节。基于所提实例响应图的弱监督实例分割框架如图3.3所示。

3.2.1 尖峰响应图简介

我们使用 PRM 方法从分类网络中提取实例级别的局部响应。分类网络首先被转换成全卷积网络，通过移除全局池化层并且将全连接层的权值转化为 1×1 的卷积滤波器。经过一次前向传播，该全卷积网络输出类别响应图 CAM $M \in$

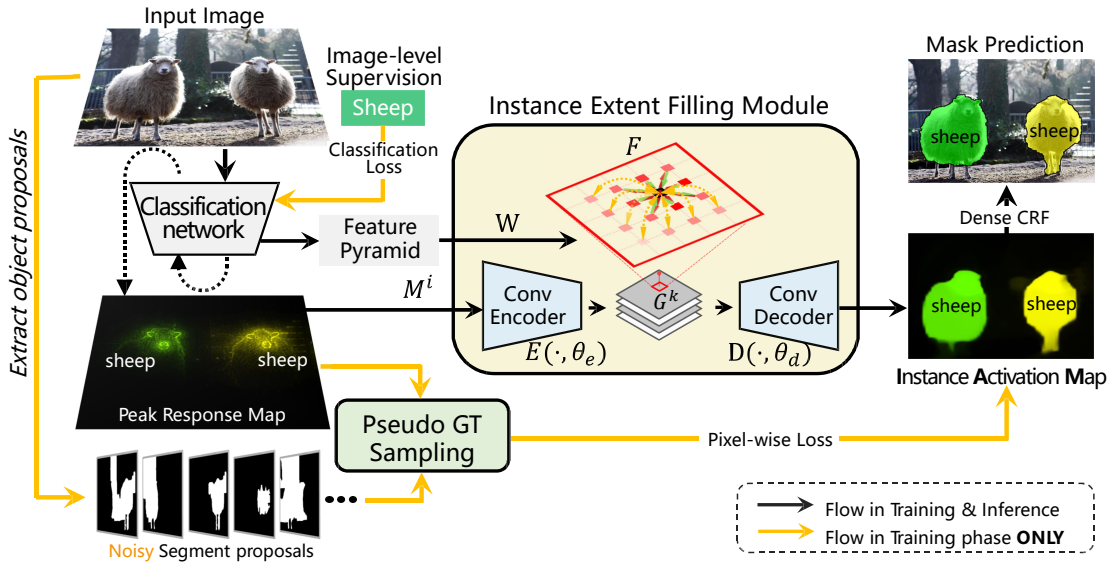


图 3.3 基于实例响应图的弱监督实例分割框架

Figure 3.3 Learning Instance Activation Maps for weakly supervised instance segmentation

$\mathbb{R}^{C\times N\times N}$，其中 C 代表类别数，$N\times N$ 是响应图的空间尺寸。对于第 c 个类别的响应图 M^c，我们检测其上的类别尖峰响应（局部极值），将尖峰响应取均值用来预测第 c 个图像类别的置信度。

在训练阶段，分类损失函数驱动网络学习多个具有判别力的类别响应尖峰，这些尖峰响应基于网络相邻层之间自顶向下的空间位置关联进行反向传播，从而生成尖峰响应图：

$$P(U_{pq}^k)=\sum_{k\in\mathcal{I}_c}\sum_{(p,q)\in\mathcal{H}_{ij}^k}P(U_{pq}^k|V_{ij}^c)P(V_{ij}^c), \tag{3.1}$$

其中 $P(U_{pq}^k|V_{ij}^c)=\sum_{(i,j)\in\mathcal{N}_{pq}^c}Z_{pq}\times U_{ij}^k\hat{W}_{pq}$，$U,V$ 是两个相邻层的输出。$\mathcal{I}_c$ 是连接到 V^c 的特征图的集合，$\mathcal{H}_{ij}^k$ 是 U 中通过非负权值连接 V_{ij}^c 的位置的集合，$\mathcal{N}_{pq}^c$ 是 V 中通过非负权值 $\hat{W}$ 连接 U_{pq}^c 的位置的集合。为了防止负数权值对最终的响应图产生影响，它们在 ReLU 层即被丢弃。Z_{pq} 是一个标准化因子用于保证转移概率相加和为一。

我们构建一个初始的概率图，其中峰值响应的位置值设为 1.0，通过公式3.1中所定义的概率传播，可以逐层确定浅层的哪些响应位置对深层的哪些响应有贡献，从而最终生成尖峰响应图 $\mathcal{M}$，用以高亮判别力的目标实例区域，如图3.3所示。尖峰响应图 $M^i\in\mathcal{M}$ 的空间尺寸和原始输入图像相同，并且它所预测的类

别标签是对应的类别尖峰响应所在的通道索引值。在进一步对尖峰响应图进行处理之前，我们对响应图根据通道维度进行标准化。

3.2.2 生成实例激活图

本章所提出的弱监督实例分割框架，通过学习一个填充模块对不完整的尖峰响应图复原完整的目标范围，可生成实例激活图。

收集伪监督信息：基于底层视觉的目标建议方法通常使用以下假设：同个目标内部的像素具有一致的颜色，纹理和或封闭边界，从而估计与类无关的目标预选掩膜。尽管伪监督信息不完整且嘈杂，但是大量精细的目标预选掩膜可以在统计意义上覆盖该目标，并且足以供网络学习如何填充局部区域中的目标范围。给定一幅图片，我们首先对每幅图片提取预选掩膜集合 S。接着，我们计算每个 PRM $M^i \in \mathcal{M}$ 和每个掩膜 $S^j \in S$ 的匹配得分：

$$f_{ij} = \alpha \cdot M^i * S^j + M^i * \hat{S}^j \tag{3.2}$$

其中，$\hat{S}_j$ 是根据形态梯度算子计算得到的预选边界掩膜，α 是一个类无关的平衡因子。匹配得分综合考虑了 M^i 和 S^j 的范围以及边缘匹配程度。对每个 PRM 的匹配得分进行排序后，我们保留得分最高的 k 个预选掩膜来提供局部正确的目标范围。当 k 增大时，会引入更多不准确的预选掩膜，导致 k 个预选掩膜互相之间差异增大，这将影响模型的学习。因此，我们接着计算 $\mathcal{M}$ 和 k 个预选掩膜之间的交叠度 $\max_i M^i * S^j$，保留大于 0.2 的掩膜。

在训练期间，对于每个 PRM，我们都会从前 k 个预选掩膜中随机抽取一个掩膜，用于在每个前向传播中构建伪监督掩膜。我们的方法与 PRM[12] 中的预选掩膜检索过程之间的区别有两个方面：1）我们方法中的掩膜用于学习实例填充模块，而 PRM 的掩膜用于直接生成最终的预测；2）我们使用随机策略对多个预选掩膜进行采样，而 PRM 仅检索得分最高的单个预选掩膜。因此，实例响应图方法的优点可总结为：1）避免了在推理阶段进行提议预选掩膜检索，因而大大提高了推理速度；2）可以从多个局部正确的预选掩膜中进行统计学习，总结出完整精确的目标范围。

实例范围填充：从上面产生的每个 {PRM-预选掩膜} 组合对中，我们从预选掩膜和图像特征中学习目标范围的常识知识，将 PRM 恢复为完整的目标范围。为了达到这个目的，我们开发了可微的目标范围填充模块，遵循编码-填充-解码

的框架。编码过程 $E(\cdot,\theta_e)$ 表示一个小型的卷积神经网络的前向传播过程，该网络包含 Conv-BatchNorm-ReLU-MaxPooling 层。θ_e 代表该网络中所有可学习的参数。通过 E 对输入 PRM M^i 进行空间尺度压缩，从而使填充模块能够更有效地捕获不同空间位置之间的长距离依赖。编码后的实例级别视觉线索 $E(M^i,\theta_e)$ 被嵌入到了一个特征空间，以减缓 PRM 中的噪声响应对填充模块的学习的影响。解码过程 $D(\cdot,\theta_d)$ 也是一个网络的前向传播过程，它的网络结构与编码器对称。经过编码-填充-解码后，可得到实例激活图 IAM。在这个过程中，空间位置和语义信息都被充分考虑，从而得到更加准确的目标范围。填充过程 F 的具体实现为一个包含 N 个填充步的迭代的概率传播过程：

$$\begin{aligned} G^k &= F(G^{k-1}, W) \\ G^0 &= E(M^i,\theta_e), \end{aligned} \tag{3.3}$$

其中 $G^k \in \mathbb{R}^{C\times N\times N}$ 是经过 k 次迭代后的特征，$0<k\le N$, $W\in\mathbb{R}^{C\times(N\times N)\times(r\times r)}$ 是填充权值，由分类及网络的中间输出特征图 M 构建而成，即 $W=R(M,\theta_r)$，R 代表特征金字塔结构[33]，该结构使用一系列 1×1 卷积对不同层级的特征图进行上采样，并从网络的深层到浅层进行融合，如图3.3所示。θ_r 是特征金字塔中的可学习网络参数。如图3.4所示，在填充过程的每一步，G^k 的第 c 个通道上的空间位置根据他的邻居像素进行更新：

$$G^k_{ij} = \sum_{u,v\in\mathcal{N}_{ij}} Y_{ij} W_{c;i,j;u,v} E^{k-1}_{uv}(M^i,\theta_e), \tag{3.4}$$

其中 Y_{ij} 是一个标准化因子以保证 $\sum_{u,v\in\mathcal{N}_{ij}} W_{c;i,j;u,v}=1$，$\mathcal{N}_{ij}$ 代表坐标 (i,j) 的 r^2 个邻居，填充过程的最大迭代次数与编码后的特征图一致，设为 N，以保证填充迭代能够到达特征图上的任意位置。学习的填充权重 W 的过程如图3.5所示，我们可视化每个像素的八个相邻邻居，其中角度表示相应方向的邻居，长度表示权值。我们计算通道上 W 的平均值，然后从每张图中减去平均值以抑制与所有邻居连接的平坦区域。在缩放区域中可以看到，填充权重清楚地标识出了实例边界。从第三行的示例可以看出，即使填充模块在数据集中不是有效的目标类别，我们的填充模块也可以成功识别 “bed” 的边界。这表明我们的方法可以学习目标范围的常识，该常识还可以推广到未见过的目标类别。

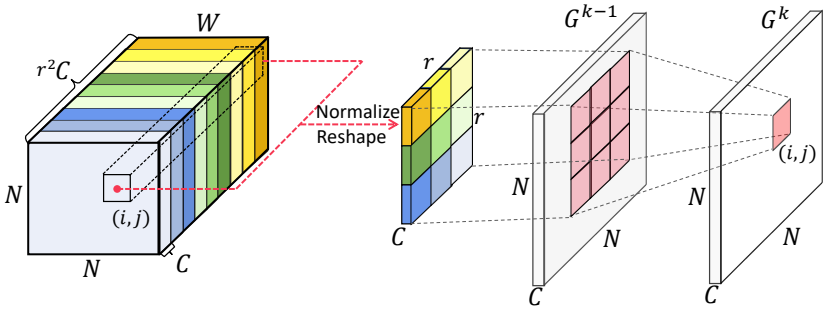


图 3.4 填充过程示意图

Figure 3.4 Illustration of the filling process

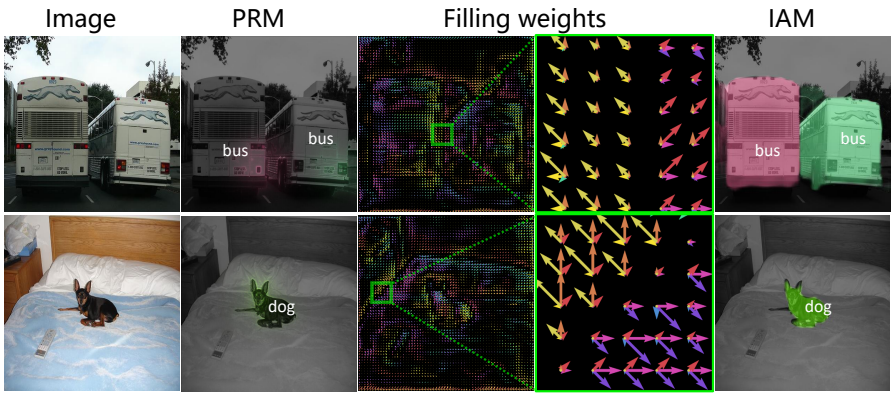


图 3.5 填充权值可视化

Figure 3.5 Visualization of the filling weights

3.2.3 模型实现

训练细节：我们提出的模型已通过图像级标签和与类无关的预选掩膜进行了训练。我们基于标准的 ResNet50[34] 体系结构来实现我们的方法。首先，我们使用多标签软边际损失函数和随机梯度下降优化器对装配有峰值刺激的骨干网络进行图像分类[12] 的训练，学习率为 0.01。然后，我们使用二进制交叉熵损失优化填充模块。随机梯度下降优化器的初始学习率设置为 0.1。我们使用基网络 ResNet50 的第 2、3、4 个残差模块的输出来构建特征金字塔。根据已有方法[12]，我们使用 MCG[3]（Multi-scale Combinatorial Grouping）框架联合高质量的区域架构和卷积方向边界[35] 来提取预选掩膜。我们的方法对于预选掩膜算法的选择没有限制。

后处理：我们的模型生成的实例级别激活图 IAM 能够覆盖目标实例范围，我们选择卷积条件随机场 ConvCRF[36] 对实例激活图进行进一步的边界优化。这

与以前的工作[12] 不同，因为其生成的 PRM 只能定位目标局部，因此不得不采用计算繁复的预选掩膜检索策略来获取更加完整的目标范围。实验显示，我们的方法不仅可以保持最佳性能，而且还可以将推理时间缩短一个数量级（每个图像从 3 秒缩短到 0.3 秒）。

3.2.4　方法原理探讨

我们的方法利用了来自分类网络的实例感知视觉线索和预选提供的目标先验，通过实例范围填充操作来学习完整的对象范围。在训练阶段，它实际上实现了一种特殊的 “语义拼接”，它收集目标碎片，将破碎的语义信息吸收到卷积神经网络中，然后学习拟合完整的目标范围。在测试阶段，尖峰响应图被看作是对应于目标判别性部件的语义锚点，范围填充模块基于这些锚点学习生成实例激活图（IAM），此过程类似于经典的 “泛洪” 过程[37]。不同之处在于，泛洪是基于灰度稳定图像区域定义，而 IAM 是为语义稳定区域定义的。实验结果表明，我们的方法可以从冗余且嘈杂的提议分段中学习实例级别的目标分割掩膜，并为弱监督实例分割提供了新的更加高效的解决方案。

3.3　实验结果及分析

我们在几种流行的基准上评估了所提的实例激活图技术。首先我们将目标实例范围填充网络与最新的弱监督实例分割方法进行了比较，证明了我们方法的有效性。接着，我们统计分析以测量所生成的 IAM 的质量，实验表明我们的方法可以生成覆盖目标范围的实例感知激活图。最后，我们将训练好的填充模块直接应用于细粒度的目标框定位任务和显著性检测任务，进一步验证了我们方法的泛化能力。

3.3.1　弱监督实例分割

我们在 PASCAL VOC 2012[38] 数据集上评估实例激活图的弱监督实例分割性能，并与现有基准进行比较。遵循常规做法，我们采用已有工作[39] 中介绍的增强数据集将原始训练集从 1464 张图像扩展到 10582 张图像，并在 1449 张图像上进行验证。

数值结果：表 3.1中展示了实例分割评估结果，我们报告平均准确率 mAP（mean Average Precision）在 IoU 阈值分别为 0.25，0.5 和 0.75 的结果。参与评

估的 IAM 模型分别为 IAM-S1，IAM-S5 和 IAM-S9，IAM-Sk 对应表示伪监督掩膜的采样个数 k。实验结果表明，IAM-S5 模型的性能高出现有最佳方法 1.6%，2.0%，和 2.9%。该模型在较高的 IoU 阈值 0.75 上取得了更加显著的提升（相比于 0.25 和 0.5），说明我们的方法在生成高质量实例激活图以及捕获精细目标边缘的有效性。此外，我们还使用平均最佳交叠 ABO（Average Best Overlap）[40] 作为评价指标。实验结果显示，IAM-S5 的 ABO 得分提升了 4.3%，证明了 IAM 恢复完整目标范围的能力。我们根据不同的从目标框生成掩膜的策略[11] 构建一些基准方法，包括：1）Rect.，直接使用目标框作为实例掩膜；2）Ellipse，在目标框内取一个最大椭圆；3）MCG，用目标框在 MCG 预选掩膜中选择一个最佳匹配的掩膜。

预选采样数 k 的影响：在填充过程中，我们的模型从伪监督掩膜中学习生成实例激活图 IAM，这些伪监督掩膜通过随机预选掩膜采样获得。填充模块从匹配得分较高的前 k 个预选掩膜中学习总结关于目标范围的常识知识。表3.1中评估了采样数 k 的影响。我们首先将 k 设为 1 来验证我们的模型是否能够通过一个不准确的预选掩膜探索目标范围的共性。结果显示，IAM-S1 相比于已有方法在 $mAP^r_{0.25}$，$mAP^r_{0.5}$，$mAP^r_{0.75}$ 和 ABO 上取得了一致的性能提升。当我们增大 k 到 5 时，IAM-S5 的性能表现优于 IAM-S1，说明我们的方法能够从带噪声的掩膜中总结出目标范围的常识知识。尽管如此，我们的模型学习会在 k 增大至 9 时受到影响，因为更多的错误的预选掩膜会在训练阶段被采样到，这使得我们的模型难以总结学习正确的目标范围。

推断时间：表3.2中展示了模型推断阶段的耗时，PRM 依赖预选掩膜检索过程（每张图检索耗时 3.0 秒，生成预选掩膜 8.1 秒），该过程十分耗时。相对的，IAM 能够学习去填充目标范围并通过条件随机场来微调掩膜边缘（每张图耗时 0.3 秒），并且能够提升实例分割性能。

可视化结果：图3.6中展示了一些 IAM 的实例分割结果，包含成功的和失败的案例。在第一列中，我们的方法可以区分具有复杂纹理的目标实例。第二列的例子中，我们的方法在杂乱场景中表现良好。第三列中，当待分割目标和其他物体十分靠近时，其分割掩膜也能被完整地定位。在第四和第五列中，展示我们方法对来自不同比例和不同类别的目标能够有效地区分，这表明 IAM 可以从分类网络中提取类感知和实例感知激活。最后一列显示了 IAM 预测错误的情况。首

Method		$mAP^r_{0.25}$	$mAP^r_{0.5}$	$mAP^r_{0.75}$	ABO
Ground Truth Box	Rect.	78.3	30.2	4.5	47.4
	Ellipse	81.6	41.1	6.6	51.9
	MCG	69.7	38.0	12.3	53.3
Baselines constructed from Weakly Supervised Object Localization					
CAM[10]	Rect.	18.7	2.5	0.1	18.9
	Ellipse	22.8	3.9	0.1	20.8
	MCG	20.4	7.8	2.5	23.0
SPN[41]	Rect.	29.2	5.2	0.3	23.0
	Ellipse	32.0	6.1	0.3	24.0
	MCG	26.4	12.7	4.4	27.1
MELM[42]	Rect.	36.0	14.6	1.9	26.4
	Ellipse	36.8	19.3	2.4	27.5
	MCG	36.9	22.9	8.4	32.9
Weakly Supervised Instance Segmentation					
PRM[12]		44.3	26.8	9.0	37.6
IAM-S1		45.6	28.3	10.4	41.5
IAM-S5		**45.9**	**28.8**	**11.9**	**41.9**
IAM-S9		45.7	27.8	10.5	41.7

表 3.1 弱监督实例分割结果对比

Table 3.1 Comparison of weakly supervised instance segmentation results

	Feed-Forward	Proposal Retrieval	CRF	Total
PRM[12]	0.05	3.0 (+8.1)	N/A	11.15
IAM (Ours)	0.07	N/A	0.3	0.37

表 3.2 推断时间（秒）对比

Table 3.2 Per-image inference time (seconds).

图 3.6 弱监督实例分割结果示例

Figure 3.6 Weakly supervised instance segmentation examples

先，如果没有准确的实例感知线索，我们的方法会漏检一个实例。通常，IAM 可能会因大面积区域的颜色或纹理差异而被误导，有时会在连接模糊或空心物体的各个部分时出现问题，IAM 有时也无法识别彼此相似的拥挤对象的边界。

3.3.2 统计分析

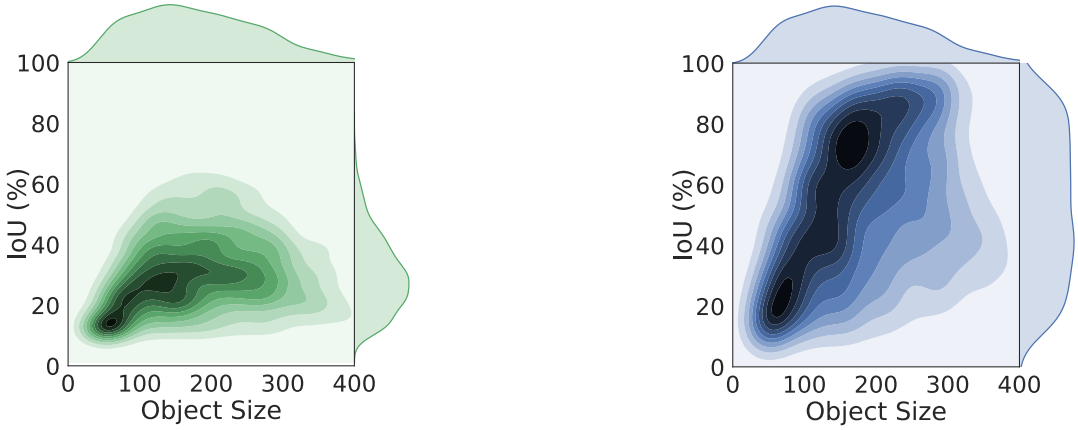
我们根据目标尺寸和目标类别对 IAM 进行一系列统计分析，并和已有工作 PRM 进行对比。使用真实标注的分割掩膜去匹配 IAM，并通过测量最匹配的 IoU 来判断是否重叠。与标注掩膜完全重合的 IAM，预测的实例激活图 M 和标注掩膜 $\mathcal{T}$ 之间的 IoU 必须接近 100%，使用度量值：

$$m = \max_{\theta, T_i \in \mathcal{T}} \frac{area(f_b(M,\theta) \cap T_i)}{area(f_b(M,\theta) \cup T_i)}, \tag{3.5}$$

其中函数 $f_b(M,\theta) = M \geq \theta$ 基于概率化的 IAM，在阈值 $\theta \in (0,1)$ 范围内生成一系列掩膜并得到其中最匹配的一个作为输出。

基于目标尺寸的分布采样：我们首先可视化 PRM 和 IAM 的 IoU 值的密度，以观测 IAM 是否能够覆盖不同尺寸的目标，如图3.7所示。在图3.7a中展示了来自 PRM 的采样样本，大部分样本都聚集在 IoU 值小于 50% 的区域，并且很难覆盖到较大尺寸的目标。相对地，图3.7b中更多的 IAM 样本分布在更高的 IoU 值的区域，并且在较大尺寸的目标上表现更好。

基于目标类别的分布采样：我们进一步计算每个类别的平均 IoU，以分析不



(a) PRM[12]　　**(b) IAM (Ours)**

图 3.7 PRM 和 IAM 的样本密度统计图

Figure 3.7 The density map of samples from PRMs and IAMs

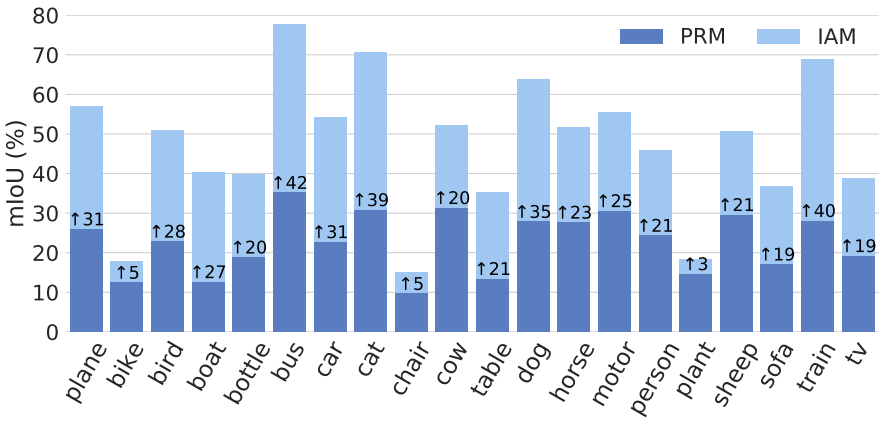


图 3.8 每个类别的平均 IoU(%)

Figure 3.8 Per-class mean IoU (%) of PRMs and IAMs

同对象类别的影响，如图3.8所示。我们的方法在所有类别上实现了一致性的提升。在 “bus” 和 “cat” 上，IAM 大大超过 PRM（约 40%）。因为 PRM 可以突出显示判别性的部分，例如 “bus” 的轮胎和 “cat” 的头，而 IAM 可以覆盖整个对象区域。

3.3.3　在未知类别上的泛化性

经过弱监督实例分割训练所得的填充模块，可直接应用于定位细粒度目标类别和未知类别的目标范围。在不对实例范围填充模块进行任何微调或重新训练的情况下，此过程可以视为无监督域自适应的一种尝试。

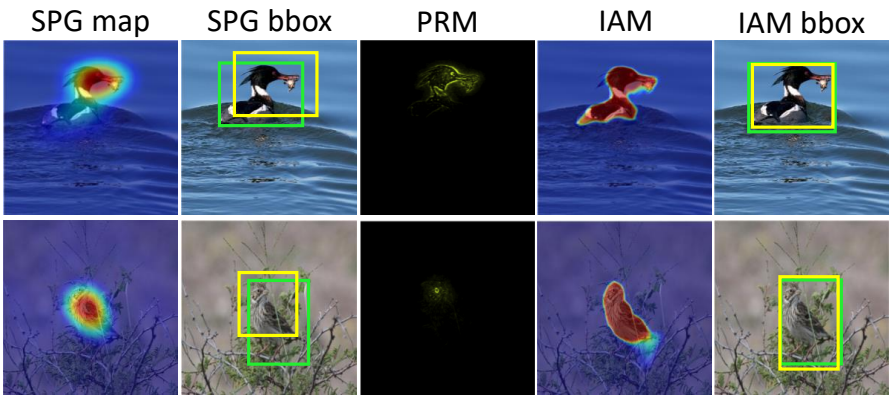


图 3.9 细粒度鸟类的目标定位结果可视化结果

Figure 3.9 Visualization of object localization from fine-grained bird species

Methods	GoogLeNet-GAP[10]	ACol[44]	SPG[45]	IAM (Ours)
Loc. Err	59.00	54.08	53.36	**52.21**

表 3.3 在 CUB-200-2011 测试集上的定位错误率对比

Table 3.3 Localization error (%) on CUB-200-2011 test set

定位来自精细类别的目标：我们使用预先训练的模型在 CUB-200-2011 数据集[43] 中对鸟类进行定位，该数据集包含 200 种鸟类的 11788 张图像，其中 5994 张图像用于训练，5794 张图像用于测试。我们选择该数据集来验证从 PASCAL VOC 2012 数据集中学到的目标范围共性知识是否可以迁移到包含许多不常见的鸟类目标的精细类别上。VOC 2012 训练集中鸟类类别仅包含 705 张图像，且不区分更精细的子类别。我们首先使用预先训练的 IAM-S5 模型计算图像的 IAM，然后使用平均值阈值提取边界框。如果预测的类别标签正确，且预测框与真实框之间的重叠度大于 0.5，则认为边界框具有正确的定位预测。为了进行公平的比较，我们使用 GoogLeNet-GAP 预测的类别响应图来辅助提取 PRM，而我们模型预测的目标范围是类无关的。在表3.3中，我们将定位结果与弱监督的定位方法进行了比较。我们发现，IAM 在细粒度鸟类上定位效果表现良好，平均错误率为 52.21%。这表明我们的模型虽然在其他域的数据和任务上进行训练，但仍可以推广到来自各式各样的子类的目标。图3.9显示 IAM 可以覆盖整个目标范围，并且能够进一步提升弱监督目标框定位的精度。

对未知类别的目标进行显著性检测：对于人类来说，即使我们不知道一个目

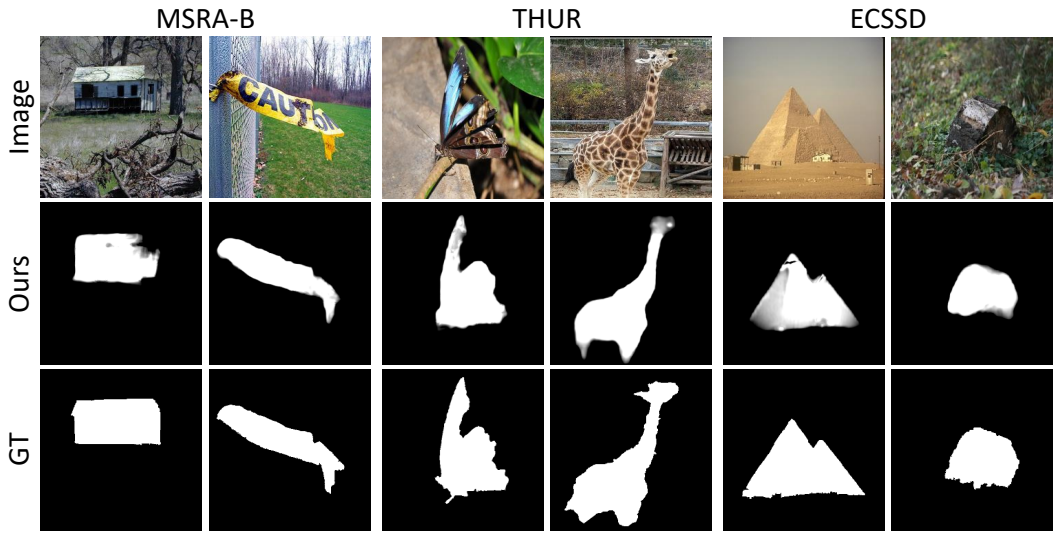


图 3.10 显著性目标检测结果示例

Figure 3.10 Salient object detection examples

标是什么类别，我们也可以确认该目标的范围，这激发我们的模型探索如何学习与类别无关的一个目标内在区域的共性。为了证实我们的模型是否可以对未知类别的目标进行定位，我们将其应用于目标显著性检测任务。我们使用在 ImageNet 上预先训练的 Resnet50 分类网络来提取实例感知的视觉线索，这些视觉线索提供了显著性目标区域的粗略位置。然后将这些线索和图像输入在 VOC12 上训练的填充网络。在 IAM 上执行 ReLU 之后，我们获得了显著性检测结果。我们使用 F 值度量（F_β）来评估三个流行的显著性检测数据集的性能，包括 THUR[46]，MSRA-B[47] 和 ECSSD[48]。在表3.4中我们与最新方法进行比较，包括三种基于深度学习框架的监督方法，三种基于手工特征的无监督方法以及两种基于深度学习的无监督模型。结果表明，尽管该我们的模型没有针对特定任务进行训练，但其性能仍然可与常用的目标范围定位方法不相上下。图3.10展示了一些显著性检测结果，这些结果表明，即使目标类别的外观与 VOC12 目标类别的外观之间存在较大差距，IAM 仍可以起到填充对象范围的作用，从而进一步验证了我们方法的泛化能力。

Finetune	Use GT	Methods	THUR	MSRA-B	ECSSD
✓	✓	DSS[49]	0.7081	0.8941	0.8796
✓	✓	NLDF[50]	-	0.8970	**0.8908**
✓	✓	DC[51]	0.6940	**0.8973**	0.8315
✓	✗	SBF[52]	-	-	0.7870
✓	✗	Multi-Noise[53]	0.7322	0.8770	0.8783
✗	✗	DRFI[54]	0.5613	0.7282	0.6440
✗	✗	RBD[55]	0.5221	0.7508	0.6518
✗	✗	DSR[56]	0.5498	0.7227	0.6387
✗	✗	IAM (Ours)	**0.7364**	0.8643	0.8613

表 3.4 显著性目标检测的平均 F 值结果

Table 3.4 Mean F-measure on salient object detection

3.4 本章小结

我们提出了一种实例激活图技术，针对弱监督框架下难以定位精确且完整的目标范围的问题，基于目标局部响应和对应的目标分割预选掩膜来学习如何填充目标的实例分割掩膜。该技术从分类网络中提取实例感知视觉线索，根据预测的范围填充权值迭代地拟合随机采样的伪监督预选掩膜，从大量含有噪声的监督信息中总结学习目标分割掩膜内像素的一致性。我们的方法实现了一种特殊的“语义拼接”，它收集了大量的包含噪声的目标分割掩膜碎片，将破碎的语义信息总结到卷积神经网络模型中，以学习与类无关的目标范围知识。实验表明，在常用数据集上，它极大地改善了现有技术，提升了弱监督实例分割性能。在已有方法 PRM 的基础上大大地提升了模型测试速度。IAM 在实例级弱监督学习问题上显示出巨大潜力，且可以被直接应用到未知类别的目标定位和分割任务上。

第 4 章　基于渐近知识驱动变换器的视觉关系检测

本文中弱监督定位的研究其核心是基于深度特征图上的图传播模型，图的结点是深度特征图中的像素，节点间的边表示像素特征之间的语义相似性和空间相似性。在本章中，我们基于图传播模型，进一步探索目标间的视觉语义关系，图的节点是定位到的目标区域，节点间的边描述了更加抽象且富含高阶语义的视觉关系。视觉关系检测任务致力于预测成对的目标区域之间的高级语义关系，这些关系可以帮助构建视觉场景的结构化抽象，从而促进场景理解的下游任务，如视觉问答[57]，图像描述[58] 和图像检索[59] 等。视觉关系检测（即，识别并定位目标，并预测目标间的语义关系）也是当前一个研究热点问题，其难点在于，关系类别往往呈长尾分布，频繁出现的关系类有充足的训练样本，而不频繁的关系类别常因为样本量稀少而难以学习。

图 4.1 基于目标类别标签猜测关系类别

Figure 4.1 Top-k guesses of predicate labels based on object labels

尽管在计算机视觉中识别视觉关系的任务很困难，人类却可以通过将目标关系的常识应用和适配到每个特定场景来执行准确而快速的视觉推理，这启发研究人员们探索如何借助关于视觉关系的常识来有效地推断不同场景中的视觉关系。研究发现[17]，当目标的类别确定时，关系类别会更容易预测。并且 top-5 盲猜的准确率高达 95%，如图4.1所示。我们基于 Transformer 设计渐进知识驱动的上下文信息编码机制，通过引入先验知识来辅助不频繁关系类别的学习。不同类型的先验知识构成一系列知识图，为目标区域之间的注意力的学习提供外部监督信息，从而使得区域的特征编码更富含相关区域的上下文信息，得到更加紧

凑且更具判别力的关系特征编码。

如图4.2所示，从常识知识中提取的先验关系图可以为每个特定场景中的实际关系图的预测提供适当的指导。知识图的常见用法是用来初始化区域之间相关性的权重值，而我们的工作旨在开发知识驱动的关系编码模块，该模块不仅可以显式地合并任何种类的常识知识以实现更好的语义推理，而且还可以链接目标区域，并以适应性的方式考虑外部知识。

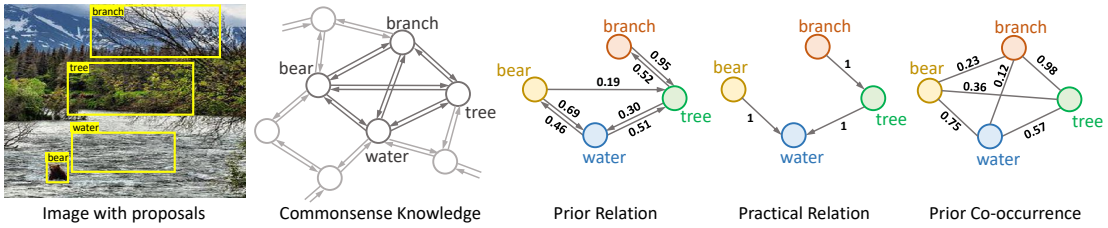


图 4.2 关于视觉关系的常识知识示意图

Figure 4.2 Illustration of the commonsense knowledge about visual relationships

4.1 模型概述

为了更好地探索有关视觉关系的结构化知识提供的丰富线索，我们提出了渐进知识驱动的变换器 PKT（Progressive Knowledge-driven Transformer），以渐进和自适应的方式引入从常识中提取的指导图，从而限制了目标区域之间的联系并强制区域特征聚焦在和其最相关的区域上下文。具体而言，PKT 通过堆叠的多头注意模块在目标区域之间传播和组合关系感知上下文信息，该模块通过将引导图视为外部监督来学习每对视觉目标的自适应上下文连接，而不是固定连接。指导图可以是具有定向或间接边缘，边缘权重的概率或整数数值，对于特定图像或特定数据集分布的任何形式的视觉常识知识图。图4.2展示了可用于构建指导图的三种不同类型的知识图。PKT 通过学习使现实世界场景中共享的视觉常识知识适配每幅图像中的视觉目标区域，为目标区域对生成通用的上下文连接，并生成具有更好的全局语义一致性和合理性的关系特征表示，从而促进视觉关系的识别。来自视觉常识的外部指导图可以帮助模型有效地规范目标区域对之间关系的不平衡分布，并减少不频繁关系预测的模糊性。

大量实验表明，PKT 在视觉关系检测的多个子任务上的性能均优于基线方法。在平均召回率指标上，PKT 相比于现有方法的性能提升更加显著，因为该指

标倾向于鼓励模型对更多来自不频繁关系类的样本进行更准确的预测。PKT 在训练速度和模型参数方面拥有巨大优势，因为它继承了高度可并行化的多头注意机制的优点。知识迁移实验的结果进一步验证了 PKT 对于来自跨数据集和跨任务知识的良好的泛化性。

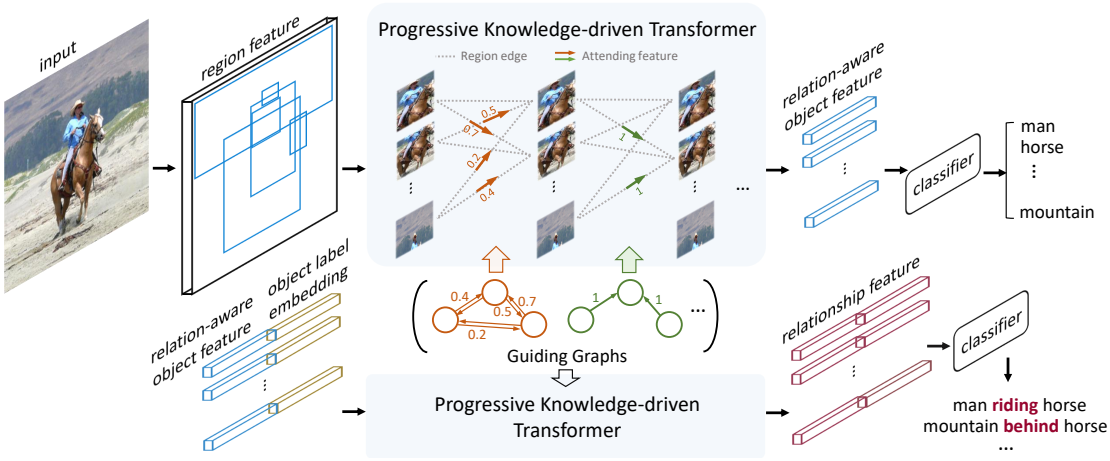


图 4.3 基于知识渐近驱动变换器的视觉关系检测框架

Figure 4.3 Progressive Knowledge-driven Transformer (PKT) for visual relationship detection

4.2　知识渐近驱动变换器

在本节中，我们首先建模视觉关系检测任务，并将框架分解为包括渐进式知识驱动的变换器（PKT）和常规分类网络模块。接着，我们详细介绍了 PKT 的结构，以及如何由视觉关系的常识中提取的引导图逐步地驱动学习。然后，我们实例化 PKT，以生成编码了输入特征之间一系列语义关联的关系感知的新特征。视觉关系检测模型架构如图4.3所示。

4.2.1　模型定义

对于一幅图片 I 中的视觉关系定义一个结构化的有向语义图 G，包含（1）一组目标检测框 $B=\{b_1,b_2,...,b_n\}$，其中 $b_i\in\mathbb{R}^4$；（2）对应于 B 中目标框的目标类别 $O=\{o_1,o_2,...,o_n\}$，其中 $o_i\in\mathbb{R}^{C_o}$，且 C_o 为目标类别个数；（3）目标区域两两之间的视觉关系类别 $R=\{r_1,r_2,...,r_m\}$。我们首先建立视觉关系检测模型，将其分解为三个模型：

$$\mathrm{P}(G|I)=\mathrm{P}(B|I)\mathrm{P}(O|B,I,S)\mathrm{P}(R|O,B,I,S), \tag{4.1}$$

其中 S 代表一系列的引导图 ($S_1 \rightarrow S_2 \rightarrow ... \rightarrow S_l$)，用于将关于视觉关系的常识知识引入到模型的学习中。

首先，目标检测模块 P($B|I$) 输出一组目标框和对应的目标类别，对于目标分类模块 P($O|B, I, S$)，我们进一步将其分解为一个 PKT 作为特征提取器 **PKT**($V|B, I, S$) 和一个特征分类器 P($O|V$)：

$$\mathrm{P}(O|B, I, S) = \mathbf{PKT}(V|B, I, S)\mathrm{P}(O|V), \tag{4.2}$$

其中 V 表示关系感知的目标特征，该特征是基于引导图 S 进行编码而来。引导图 $S_i \in S$ 可以是多种形式的能够反映目标间不同角度的关联的关系图，如图4.2所示。类似地，关系分类模块 P($R|O, B, I, S$) 也被分解为一个 PKT 用以向关系特征注入常识知识，以及一个关系特征分类器。这里 “视觉关系” 指三元组 $\langle$主体 (subject), 谓词 (predicate), 客体 (object)$\rangle$，这里 “谓词” 表示两个目标之间的语义关系标签。

4.2.2 模型方法细节

本章所述视觉关系检测框架的核心，是知识渐近驱动变换器 PKT（Progressive Knowledge-driven Transformer），PKT 是一个一般化的模块，以一种灵活和可转移的形式将视觉常识知识渐近地集成到特征学习过程中。具体地，PKT 使用一系列堆叠的多头注意力模块用来更新特征，其所学的注意力由外部引导图信息进行监督。引导图可以是任意形式的知识图，可以是有向图和无向图、软性或硬性的数值权值、来自特定图片或特定数据集的知识。

在 PKT 的每一层，引导图作为外部监督和多头注意力权值计算损失，促使特征表达更多地关注和它更可能存在关联的区域，从而生成更加紧致的特征表达。PKT 中的注意力模块基于多样化的视觉常识知识增强区域特征，具体地，如图4.4所示。

输入特征 $V = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n\}, \mathbf{v}_i \in \mathbb{R}^d$，PKT 首先对每个 $\mathbf{v}_i$ 学习转换参数 W^K, W^Q and $W^V \in \mathbb{R}^{d \times d}$。给定一个引导图 S_i，每个特征表示 $\mathbf{v}_i$ 学习基于注意力权值 A_h 去关注其他的目标区域特征 $\mathbf{v}_j$：

$$A_h(\mathbf{v}_i, \mathbf{v}_j) = \mathrm{softmax}((W_h^Q \mathbf{v}_j)(W_h^K \mathbf{v}_i)^T / \sqrt{d}), \tag{4.3}$$

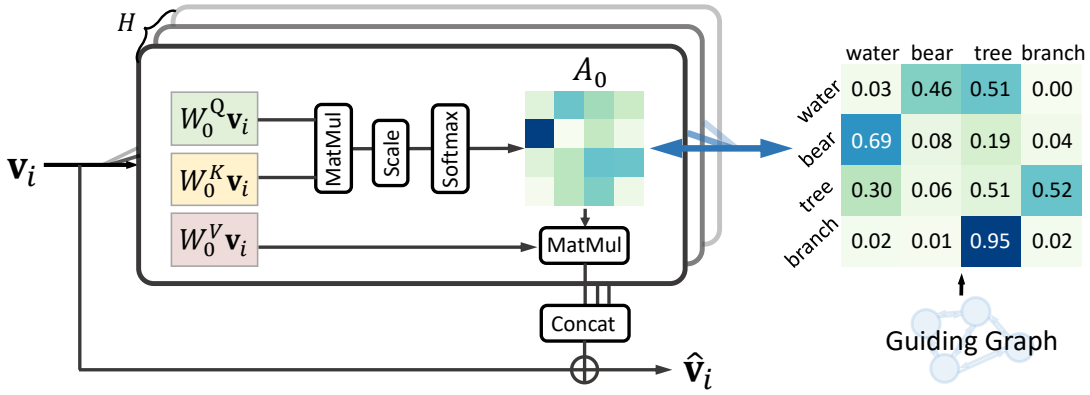


图 4.4 PKT 中的多头注意力模块示意图

Figure 4.4 Illustration of the H-head attention mechanism in PKT

$$\hat{\mathbf{v}}_i = \mathbf{v}_i + \text{concat}_{h=1}^{H}\left\{\sum_{j=1}^{N} A_h(\mathbf{v}_i, \mathbf{v}_j) W_h^V \mathbf{v}_j\right\}, \tag{4.4}$$

$$\mathcal{L}_{attn} = \sum_{S_i \in S}\sum_{h=1}^{H} f(A_h, S_i), \tag{4.5}$$

其中 PKT 在连接 H 个注意头之后应用了残差连接，并且通过二进制交叉熵损失函数 $f(\cdot)$ 来计算 A_h 与指导图 S 之间的损失 L_{attn}。在多头注意力层之后是一个特征变换模块，包括两层前馈网络和 LayerNormalization 层。

PKT 的优点主要在于三个方面：首先，由于多头注意模块的堆叠架构，PKT 模块的实现轻量化且高度可并行化；其次，PKT 具有灵活性，可以轻松地合并任何形式的外部知识图来构建紧凑的关系特征表示；最后，PKT 以自适应方式将视觉常识与每个图像的视觉概念联系起来，因而该模型能够兼容跨数据集和跨任务场景中的常识知识。

4.2.3　视觉关系检测

视觉关系检测需要在复杂且混乱的场景中检测目标区域并识别它们之间的关系。PKT 通过逐步合并从关于视觉关系的常识中提取的引导图来学习目标区域之间的语义关联，从而为目标区域生成更加紧凑的关系感知的特征表示。具有更好的全局语义连贯性和合理性的特征表达对于视觉关系的识别至关重要，特别是对于不常见类别的视觉关系。

目标分类：首先将 PKT 实例化为一个 PKT^{obj}，用于对目标特征进行特征编

码，其输入为多种特征的联合形式，包括目标区域的特征 $\mathbf{v}_i^o \in \mathbb{R}^{4096}$，图像的全局上下文特征 $\mathbf{c} \in \mathbb{R}^{4096}$ 以及目标区域坐标 b_i 和标签 l_i^o 的嵌入特征经过可学习权值 $W_1 \in \mathbb{R}^{4\times200}$ 和 $W_2 \in \mathbb{R}^{C\times200}$ 的变换特征。对该联合特征 $[\mathbf{v}_i^o, W_1 b_i, W_2 l_i^o, \mathbf{c}]$ 使用全连接层进行降维，可得新的特征 $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^d$，该特征被输入到 PKT^{obj} 中去生成关系感知的特征表达。与涉及复杂的循环或顺序模块的已有的目标分类模块相比，仅基于注意力机制的 PKT 不仅提供了集成外部指导图的可能，还具有更高的可并行性，所需的训练时间和模型参数明显更少。

关系分类：由于 PKT 可以灵活地引入关于目标区域关系的多种形式的指导图以进行渐进式关系编码，因此我们将 PKT 实例化为 PKT^{rel}，以进一步构造用于关系分类的特征表示，如图4.3。对于每个区域，输入特征向量是通过将由 PKT^{obj} 编码的特征与预测目标标签的嵌入特征进行串联来构造的，见图4.3。然后通过一个全连接层以获得 PKT^{rel} 的输出特征向量，从而得到 8192-d 编码的特征，对于主体（subject）特征和客体（object）特征，该特征可以分为两个 4096-d 特征向量。对于每个可能的主体-客体对，我们将主体特征和客体特征以及两个区域的并集边界框的特征进行组合，以构建他们之间的关系的特征。最后采用全连接层来预测每个目标对之间的关系类别。

引导图：我们提出的 PKT 不仅可以显式地参考任何形式的常识知识的指导图以进行特定的视觉推理，而且还可以以自适应方式将外部知识与每个图像的视觉观测联系起来。在这里，我们从关于视觉关系的三种知识形式（图4.2）中介绍引导图。先验关系图 S^r 是有向图，其边缘表示目标类之间成对语义连接的出现频率。边缘权重是浮点数，并且分布是特定于数据集的。实际关系图 S^t 表示每个图像中的语义关系的真实情况，为定向边，整数权重，并且是特定于图像的分布。先前的共现图 S^c 使用具有浮点权重的边来描述出现在同一图像中的任何两个视觉元素的频率。

鉴于 PKT 使用外部知识图作为指导的高度灵活性，通过多头注意力模块的堆叠，逐步演化的注意力能够逐步扩展到更多其他的相关目标区域，而不仅仅是相邻区域。我们的视觉关系检测模型所使用的指导图以及对应的学习到的注意力如图4.5中所示。在第一行中，我们可以看到，根据学习到的注意力，“ear”，“eye”，“head” 和 “trunk” 的表示倾向于关注在 “elephant” 上。在第二行中，“pizza” 和 “table” 之间的注意力变得比指导图中的更强。在知识图的指导下以自适应方

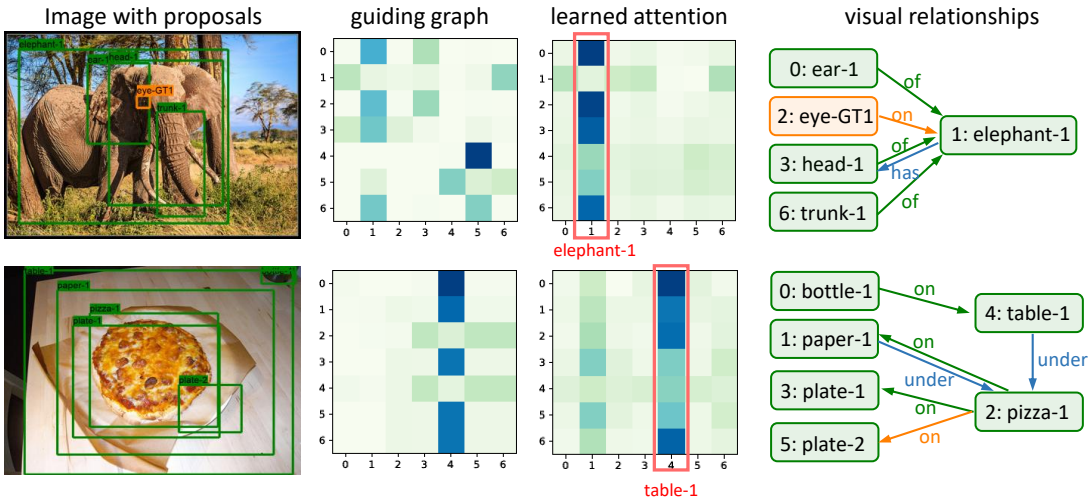


图 4.5 引导图和所学注意力权值的可视化示例

Figure 4.5 Visualization examples of the guiding graphs and the learned attention weights

式学习注意力权重，并基于每个图像中具体的视觉观测来计算注意力权重，从而约束和重分配对象之间的连接。

4.2.4　方法原理探讨

PKT 旨在通过从包含视觉常识的指导图中学习视觉概念之间的自适应联系来丰富特征表示。指导图可以视为“课程[60]”，它可以教会 PKT 以渐进方式利用每一层区域之间的关系。指导图可以是重复知识图的序列，从而形成一个简单的课程来关注特定类型的知识。另一方面，包括混合类型的知识图的指导图可以形成包含了各种视觉常识知识的课程。

4.3　实验结果及分析

我们基于几种流行的基准评估 PKT，首先介绍实验设置，例如训练详细信息，数据集和度量。其次将 PKT 与现有的视觉关系检测方法进行了比较，以证明我们方法的有效性和效率。接着我们对用不同种类的引导图实例化的 PKT 进行了消融研究，显示了 PKT 在合并知识图的不同组合中的灵活性。最后，可转移的实验验证了我们的 PKT 在跨数据集和跨任务场景中都与知识图兼容的优良特性。

Dataset	Training Set			Testing Set			#ObjCls	#RelCls
	#Img	#Rel	Ratio	#Img	#Rel	Ratio		
VRD[59]	4,734	30,355	1:6	954	7,638	1:8	100	70
VG[61]	62,723	439,063	1:7	26,446	183,642	1:7	150	50
VG-MSDN[61,62]	46,164	507,296	1:11	10,000	111,396	1:11	150	50
VG-DR-Net[61,63]	67,086	798,906	1:12	8,995	26,499	1:3	399	24

表 4.1 实验所用数据集的统计信息

Table 4.1 Dataset statistics of VG and three cleansed version of the raw VG

4.3.1 实验设定介绍

实现细节：我们采用带有 VGGNet 的 Faster RCNN 作为我们的基网检测器，检测器的训练使用随机梯度下降算法，输入图像的大小为 592×592。随后，我们冻结检测器的权重以训练 PKT^{obj} 和 PKT^{rel}。我们堆叠了四个 4 头部注意模块以构造 PKT。我们的模型首先使用真实边界框标注进行训练，然后使用检测器预测的区域预选进行完善。目标分类和关系分类模块都使用交叉熵损失函数。

数据集介绍：我们在四个流行的数据集上评估我们的方法，包括 VRD[59]（Visual Relationship Detection）数据集，VG[61]（Visual Genome）数据集，以及对 VG 数据集的两个清除了标注噪声的版本：VG-MSDN[62] 和 VG-DR-Net[63]。数据集的详细统计信息见表4.1，#Img 代表图片数，#Rel 代表关系数，Ratio 代表每幅图中平均的关系数目。#Object 和 #Pred 分别代表目标和关系的类别数。

任务介绍：为了全面地和已有方法进行对比，我们在四个子任务上评估模型：(1) Predicate Classification (**PredCls**)，给定图像的目标真实标注框和标签，预测目标间的关系类别。(2) Scene Graph Classification (**SGCls**)，给定图像的目标真实标注框，预测目标标签以及目标间的关系类别。(3) Scene Graph Generation (**SGGen**)，给定一幅图像，预测目标标注框、目标标签、目标间的关系类别。(4) Visual Phrase Detection (**PhrDet**)，给定一幅图像，预测存在关系的目标的联合框，目标分别的标签，对应的关系类别。

评价指标：Recall@K（R@K）[59] 指标通常用于早期方法中，以评估视觉关系检测方法，其中，Recall@K（K=20、50、100）得分表示真实关系三元组出现在图像中预测的前 K 个三元组中的比例。由于该指标很容易被最频繁关系的表现所

Method	SGCls			PredCls			SGGen		
	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
IMP [64]	31.7	34.6	35.4	52.7	59.3	61.3	14.6	20.7	24.5
MOTIFS [17]	32.9	35.8	36.5	58.5	65.2	67.1	21.4	27.2	30.3
KERN [18]	-	36.7	37.4	-	65.8	67.6	-	27.1	29.8
VCTree [15]	35.2	38.1	38.8	60.1	66.4	68.1	22.0	27.9	31.3
PKT (Ours)	33.5	36.1	36.8	60.4	66.3	68.0	21.7	27.0	30.2

表 4.2 VG 数据集上单关系预测结果 Recall@K 的对比

Table 4.2 Comparison of the single prediction results (Recall@K) on VG dataset

主导，因此一些最近的著作[15,18] 建议使用新指标，即 mean Recall@K（mR@K），该指标首先计算 R@K 分别来自每个关系类别的样本，然后计算每个关系类别的 R@K 分数的平均值。由于我们提出的 PKT 旨在更好地利用常识来使视觉关系（尤其是不常见的视觉关系）的识别受益，因此我们进一步使用 mean Recall@K（mR@K）来减轻关系类别的不平衡分布的影响。mR@K 指标计算每个关系类别的 R@K 分数的平均值。一些模型[64] 只考虑每个目标对之间的单一关系，而其他工作[65] 允许对一对目标进行多次关系预测，从而获得更高的分数。在这项工作中，我们在单个和多个预测的设定下评估我们的模型。

4.3.2　与已有方法进行对比

数值结果：我们在 VG 数据集上比较了现有方法和带有指导图 ($S^c \rightarrow S^t$) 实例化的 PKT 模型的性能。R@K 和 mR@K 的评估结果如表4.2和表4.3所示。可以看到，PKT 模型在 mR@K 指标上的提升更为显著，验证了 PKT 方法的有效性以及其检测不频繁关系的能力，证明了我们方法在使用有限的训练样本预测不频繁关系的优越性。VG 数据集上关系类别的分布非常不均匀，R@K 得分常常被频繁类的性能主导，而 mR@K 倾向于鼓励模型能够检测到更多不频繁类的样本，PKT 引入的外部指导图可以更多地辅助模型对不常见关系类进行预测预测。由于 VG 数据集的人工标注噪声较大，且关系标注本身主观性较强，此处实验评测中我们仅考虑为每对目标预测一个关系的设定。我们还使用三个数据集（包括 VRD，VG-MSDN 和 VG-DR-Net）将我们的方法的 R@K 分数与 PhrDet 任务上的现有方法进行比较，如表4.4所示，我们的方法明显在优于以前的方法，尤

Method	SGCls			PredCls			SGGen		
	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
IMP [64]	5.6	6.8	7.2	9.2	11.9	12.9	2.7	4.2	5.2
MOTIFS [17]	6.3	7.7	8.2	10.8	14.0	15.3	4.2	5.7	6.6
KERN [18]	-	9.4	10.0	-	17.7	19.2	-	6.4	7.3
VCTree [15]	8.2	10.1	**10.8**	14.0	17.9	19.4	**5.2**	6.9	8.0
PKT (Ours)	**8.3**	**10.2**	10.6	**15.3**	**19.6**	**21.2**	5.1	**7.2**	**8.5**

表 4.3 VG 数据集上单关系预测结果 mean Recall@K 的对比

Table 4.3 Comparison of the single prediction results (mean Recall@K) on VG dataset

Method	VRD		VG-MSND		VG-DR-Net	
	R@50	R@100	R@50	R@100	R@50	R@100
IMP[64]	27.7	28.2	20.9	23.3	-	-
Vip-CNN[66]	22.8	27.9	-	-	-	-
DR-Net[63]	19.9	23.5	-	-	24.0	27.6
MSDN[62]	-	-	19.9	24.9	-	-
MOTIFS[17]	26.8	28.1	28.3	30.8	33.4	33.7
LKD[19]	26.5	29.8	-	-	-	-
FNet[14]	26.0	30.8	22.8	28.8	26.9	32.6
KB-GAN[20]	27.4	**34.4**	23.5	30.0	-	-
PKT(Ours)	**28.2**	29.3	**30.1**	**31.8**	**35.1**	**35.3**

表 4.4 在 PhrDet 任务上的 Recall@K 对比

Table 4.4 Comparison on VG-MSDN dataset on the Phrase Detection task

其是在 R@50 指标上，这表明我们的方法能够从模糊的关系候选中识别更多正确的视觉关系。

模型容量和时间代价：如表4.5所示，PKT 在保持最佳性能的同时，使用的网络参数少于其他基线方法。这里我们在不考虑检测器主干和 RoiAlign 模块（它们是这些模型中的共享体系结构）的情况下对参数进行计数。PKT 可以通过在引入指导图的同时以高度并行的方式参与最相关的功能来生成关系感知的特征表示，实验结果表明，PKT 相比于已有视觉关系检测模型有着更高的训练效率。与 MOTIFS[17] 类似，PKT 模型需要两个阶段的训练。首先使用真实目标框训练模型，对应子任务 SGCls，然后使用目标预选训练微调结果模型，对应子任务 SGDet。

Method	Params (MB)	Time (ms/img)		
		PredCls	**SGCls**	**SGDet**
MOTIFS[17]	160	0	74	322
KERN[18]	147	229	236	403
VCTree[15]	153	930	849	725
PKT(Ours)	**142**	**0**	**72**	**252**

表 4.5 VG 数据集上模型参数和训练时间代价对比

Table 4.5 The comparison of model parameters and time cost on VG dataset

PredCls 任务在 SGCls 任务所得的同个模型上执行测试。相对地，KERN[18] 和 VCTree[15] 在模型训练过程中还需要多余的步骤，在 SGCls 训练阶段之前，他们需要使用真实的目标框和类别标签训练模型，对应于子任务 PredCls。如表4.5所示，PKT 不仅需要更少的训练步骤，而且在每个步骤中花费的训练时间也更少，不同阶段的平均训练时间（ms/image）均在一块 NVIDIA GeForce GTX 1080Ti GPU 上测试得到。

可视化结果：从图4.6中显示的可视化示例中可以得出以下结论：（1）PKT 可以对复杂和混乱场景中的视觉关系产生正确的预测，从而证明我们方法的有效性。（2）由于常识性知识的指导，PKT 可以预测标注中不存在但是合理的视觉关系（蓝色箭头）。（3）遗漏的关系（橙色箭头）通常是由遗漏的物体（橙色框）引起的。在未来的工作中，我们将探索如何进一步提高目标检测器的性能。

4.3.3　消融研究

PKT 可以与许多不同形式的常识知识图兼容，无论是有向边或无向边，浮点权值的边或整数权值。我们在表4.6中从三个方面评估用不同的引导图实例化的 PKT。为了减轻嘈杂的目标区域候选和标签的影响，我们在 PredCls 任务上使用具有多关系预测设定的 mR@K 指标评估模型。为了简化对比，我们以 PKT^{obj} 为例。

指导图个数：我们将 PKT^{obj} 中的指导图的数量从四个减少到两个，因此 PKT 的最后两层的多头注意力无需外部监督。表4.6的第一行显示，当删除两个指导图时，性能会显著下降，这表明在外部知识的帮助和引导下，PKT 模型的视觉关系检测性能得到了提升。

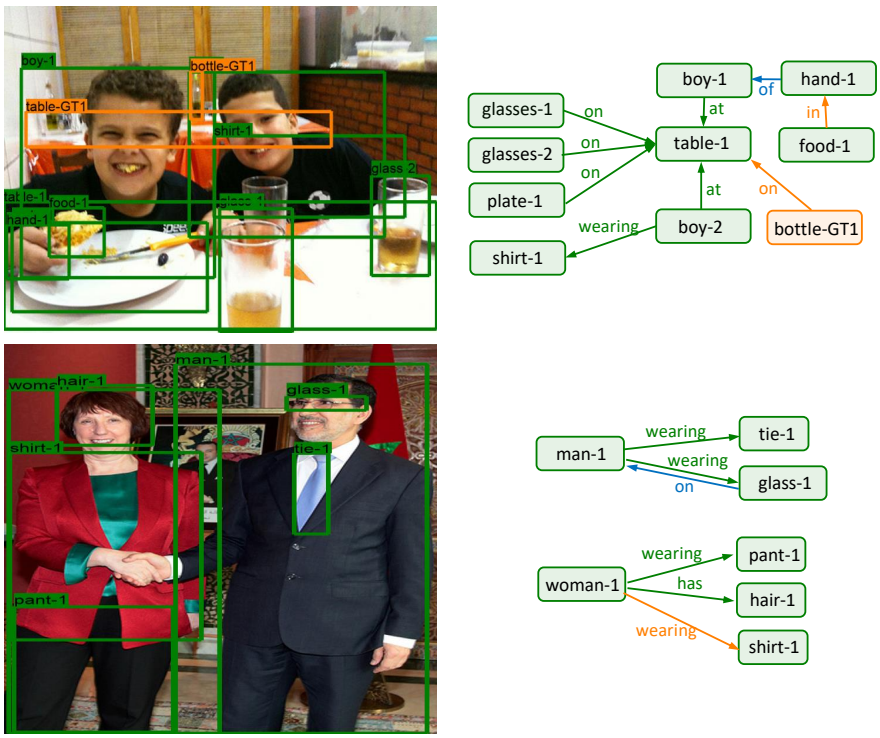


图 4.6 视觉关系检测结果示例

Figure 4.6 Visualization examples of visual relationship detection

不同类型的知识：表4.6的第二行显示 PKT^{obj} 的结果，并以实际关系图 S^t 作为指导图。使用目标真实连接图的性能比关系图 S^r 差，因为很难在视觉场景之间共享此类特定于图像的知识。从第三行可以看到，配备有先验共现图 S^c 的 PKT^{obj} 的性能略逊于先验关系图 S^r 的性能。原因可能是图像中同时出现的目标类别可能未指示与视觉关系检测任务中所需的特定语义相关的关系。

混杂的常识知识：当 PKT^{obj} 使用混合的知识图作为引导图时，在表4.6的第四行，可以看到使用引导图序列 (S^c, S^c, S^r, S^r) 的模型的 mR@K 结果低于使用 (S^c, S^c, S^c, S^c) 和 (S^r, S^r, S^r, S^r) 的模型，因为 S^c 中所蕴涵的关系是无向且含糊的，因此与有向的且指向语义关系的 S^r 并不兼容。相对的，从第五行我们可以看到使用 (S^r, S^r, S^t, S^t) 的模型的 mR@K 的分高于 (S^t, S^t, S^t, S^t) 的并低于 (S^r, S^r, S^r, S^r) 的。图像特定的引导图 S^t 实际上可以看成是 S^r 其中的一次采样，因此 S^r 可以提供更加一般化的知识来帮助模型提升性能。从以上研究可以观测到，使用重复的 S^r 作为引导图序列效果较好，并且在我们其他实验中都沿用该设定。

Guiding graphs	PredCls		
	mR@20	mR@50	mR@100
(S^r, S^r, N, N)	11.1	20.8	31.2
(S^t, S^t, S^t, S^t)	13.9	23.0	35.3
(S^c, S^c, S^c, S^c)	14.0	25.1	36.0
(S^c, S^c, S^r, S^r)	14.0	23.2	34.3
(S^r, S^r, S^t, S^t)	13.8	25.1	35.8
(S^r, S^r, S^r, S^r)	**14.5**	**26.1**	**38.5**

表 4.6 VRD 数据集上关于引导图的消融研究

Table 4.6 Ablation studies on different kind of guiding graphs in different orders

4.3.4　知识迁移

常识知识通常指示主体-客体对的各种语义关联，这是相当普遍的，并且可以在不同的任务和数据集上自然地转移。我们建立了三种有关视觉对象之间语义关系频率的知识：KG_S，KG_M 和 KG_L，分别收集自 VRD 数据集的 30,355 个关系，VG 数据集的 439,063 个关系，以及 GQA 数据集（来自视觉问答任务）的 3,795,907 个关系，基于更多关系建立的知识将更接近现实世界场景中的视觉常识。我们使用 L1 距离 $d_{l1}(\cdot,\cdot)$ 来评估不同知识之间的相似性：$d_{l1}(KG_M, KG_L) = 299.7 > d_{l1}(KG_S, KG_L) = 196.8 > d_{l1}(KG_S, KG_M) = 143.4$，这说明跨任务迁移知识比跨数据集迁移知识更难。然后我们从两个方面进行知识迁移实验，如表4.7所示。

跨数据集知识迁移：与来自 VG 的知识 KG_M 结合使用时，在 VRD 数据集上训练的 PKT 的性能更高（在 mR@K 上提升约 1%），因为来自更多视觉关系样本的 KG_M 更接近现实世界场景中的视觉常识，从而有利于 VRD 的性能提升。

跨任务知识迁移：为了进一步验证 PKT 是否甚至与来自另一个任务域的知识也兼容，我们使用源自 VQA 任务的 KG_L 在 VRD 和 VG 数据集上训练 PKT。在 VRD 和 VG 数据集上使用 KG_L 的 PKT 的性能与其基线（VRD+KG_S 和 VG+KG_M）相当，从而验证了 PKT 指导图的可移植性。

Dataset	Knowledge	SGCls		PredCls	
		mR@50	mR@100	mR@50	mR@100
VRD	KG_S	6.5	6.7	12.1	12.4
	KG_M	7.5	7.7	13.5	13.8
	KG_L	6.8	7.0	13.0	13.3
VG	KG_M	9.6	10.3	18.3	19.6
	KG_L	9.4	10.2	18.2	19.3

表 4.7 在 VRD 和 VG 上的跨任务和跨数据集知识迁移

Table 4.7 Knowledge transfer experiments on VRD and VG dataset

4.4 本章小结

针对不频繁类别样本量稀少难以学习的问题，本章提出了渐进式知识驱动的变换器（PKT），该模块通过从常识知识中提取多样化的指导图来生成关系感知的特征表达，从而辅助不频繁类别的识别。PKT 借助引导图来学习约束视觉区域之间的语义关联，生成了更加紧凑的关系特征。该关系特征更多地关注与当前目标区域最相关的区域上下文信息，从而促进区域间视觉语义关系的识别。PKT 可以灵活地整合多种形式的知识，并且具有很强的泛化性。实验表明，PKT 能够更加准确地识别不常见的关系类别，同时显著地减少了模型训练时间。

第 5 章　基于可配置图推理的视觉关系检测

在引入先验知识辅助不频繁关系类的识别时，模型基于目标对的类别标签来推断关系，它们遵循固定的关系推理路径，即 {主体（subject），客体（object）→ 谓词（predicate）}，缩写为 {s, o → p}，利用相关的先验知识来辅助不常见关系类别的识别。它们的知识是从特定数据集的统计数据中总结出来的，以提供初始的关系预选[18,63] 或关系分类器[17,67]。这些工作无法从带噪声的知识中选择性地接收信息，因此无法利用更广泛更真实的视觉关系常识知识。另外一些工作从包含许多冗余连接的知识图中蒸馏[19] 或检索知识[20]。但由于其关系推理路径是固定的，当目标类别标签成倍增长时，可能的目标对组合会迅速增长。这需要外部知识能够尽可能全面地覆盖不同的组合来提供足够的信息辅助关系推理，因此模型性能将大大受限于知识图的质量和容量。

5.1　模型概述

为了使视觉关系模型能够灵活地兼容更通用的常识知识，本章提出可配置图推理 CGR（Configurable Graph Reasoning），该技术使用一系列的子推理路径：{s → p}, {o → p}, {s → o} 和 {o → s}，以检索和匹配知识连接，并且学习如何动态地组合这些知识增强的推理路径以辅助视觉关系的预测。具体地，有向关系图是根据实体（目标和谓词标签）之间的语义相关性进行连接而构建的。遵循多个子路径的推理路径，我们首先投影目标或关系的特征以匹配知识图中的实体节点，接着再通过相应实体之间的知识连接在全局视角下更新这些特征，最后再投影回到节点特征空间。通过将知识联系整合到不同子路径的特征中，图推理模块可以在常识知识的指导下为每个视觉元素生成丰富而紧凑的特征表示形式。CGR 的配置模块（充当信息流动开关）学习做出离散决策，决定是否要在当前的推理子路径上使用知识增强的特征。

推理模块和配置模块协同工作，对不同的子推理路径进行知识增强，并根据图片中的实际情况动态地组合推理路径，从而获取对关系推理真正有用的知识，并辅助识别关系类别。CGR 可与来自不同数据集合的知识图兼容，只要这些图包含有关实体之间相关性的基本共识，可以用于子路径上的推理即可。如图5.1所

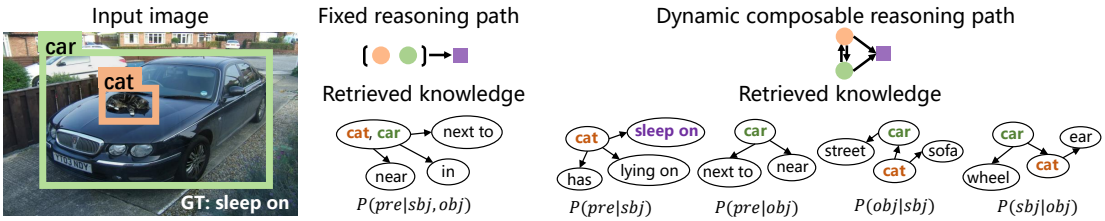


图 5.1 固定和动态的知识推理路径对比

Figure 5.1 Comparison of the reasoning over a fixed path and dynamic path

示，从固定推理路径中检索到的知识仅限于描述 {cat，car} 和 “sleep on” 之间的先验关联，这在当前知识库中并不常见。相反，CGR 选择关于 $P(pre|sbj = \text{“}cat\text{”})$ 的知识，对子路径 {s → p} 上的关系推理提供有力的证据。

CGR 的优势主要在以下三个方面：（1）相比于固定的多依赖推理路径，CGR 考虑对依赖关系更加通用且稳定的多个子路径执行推理。（2）CGR 选择性地为视觉关系的推理引入知识连接，实现了对不频繁和未知关系类别识别的泛化能力。（3）我们的方法与来自任何数据集合中的知识图都可兼容，只要它们包含有关子路径的基本常识即可。在 VRD 数据集和 VG 数据集上进行的大量实验表明，CGR 在视觉关系检测的多个基准上显著优于已有方法的性能。在不频繁关系和未知的关系上的优良表现验证了我们的方法能够从外部知识图中提取补充信息从而改进视觉关系检测。知识泛化实验证明，CGR 不仅可以从通用关系中选择知识，而且还可以抵抗噪声和冗余连接。

5.2 模型定义与实现

基于可配置图推理（CGR）的视觉关系检测框架可以分为以下步骤：（1）为目标预选和关系预选生成特征表示；（2）将视觉常识知识沿不同的子路径引入到关系推理中；（3）为每个视觉关系的推理选择知识增强的子路径；（4）基于知识增强的特征预测目标和谓词的标签。

5.2.1 特征表示

我们首先为视觉关系所包含的视觉元素生成初始的特征表示，作为在视觉常识知识图上进行的关系推理时的子路径节点特征。

目标区域特征：对于图像 I，我们使用 Faster R-CNN[24] 检测器来提取目标框 $B = \{b_i\}_{i=1}^{\hat{N}}$，其中 $b_i = [x_i, y_i, w_i, h_i]$ 是框的坐标，其对应的区域特征为 v_i。

关系特征：对于任意两个不同的目标 $[b_i, b_j]$，存在 $\hat{N}(\hat{N}-1)$ 种可能的有向关系。考虑到模型的计算效率，我们定义 $N = 512$ 为每幅图片中的目标对的最大值。我们构建一个单独的网络分支来提取主体-客体的目标对 $[b_i, b_j]$ 的联合区域特征 v_{ij}^{p}，作为初始的谓词特征，该网络分支和检测网络中的 ROI-Align 层有着相同的结构。这个分支的目的是聚焦主体-客体的互动区域，从而为关系（谓词）生成视觉表达。

子路径的节点特征：对于目标对 (b_i, b_j) 之间的视觉关系 $\{s, p, o\}$ 的推理，CGR 首先将推理路径解耦为一系列的子路径，其中包含两种路径，一种是同质节点的路径，如 $\{s \to o\}$ 和 $\{o \to s\}$，另一种是异质节点的路径，如 $\{s \to p\}$ 和 $\{o \to p\}$。在路径 $\{s \to o\}$ 和 $\{o \to s\}$ 中，s 和 o 的节点特征为目标框 b_i 和 b_j 的区域特征 v_i^s 和 v_j^o，而对于路径 $\{s \to p\}$ 和 $\{o \to p\}$ 中的 s 和 o，其节点特征为 $v_{ij}^{sp} = \text{concat}(v_i^s, v_{ij}^p)$ 和 $v_{ij}^{op} = \text{concat}(v_i^o, v_{ij}^p)$，其中包含了谓词的上下文信息。

5.2.2　基于常识知识的图推理

我们使用 CGR 为每个子推理路径引入高级语义和语言信息，视觉元素之间的信息传播是基于常识知识图的语义连接来执行的，从而形成可用于关系识别的丰富紧凑的表示形式。定义 $X = \{x_i\}_{i=1}^{N}, x_i \in \mathbb{R}^{D^d}$ 表示每个图像中相同类型子路径的起始节点的特征。用于更新每个子路径的输入特征的图推理模块可以表示为 $Y = g(X)$。在本节中，为了简化数学表达，我们只描述了基于子路径的通用版本的图推理。进行视觉关系检测时，图推理模块 $g(\cdot)$ 根据知识的不同语义相关性进行实例化，以实现在不同子路径上的推理。

知识编码：常识知识图通常用于描述实体之间显式的相关性，该图可以定义为 $\mathcal{G}^k = (\mathcal{V}^k, \mathcal{E}^k)$，其中 $\mathcal{V}^k$ 和 $\mathcal{E}^k$ 分别表示图的节点集合和边集合。图 $\mathcal{G}^k$ 是一个异质图，其中的节点包含两种不同类型的实体，即目标标签（如图5.2中的圆形节点）和谓词标签（如图5.2中的方形节点）。两个节点之间的边表示从现实场景中的视觉关系统计而来的的节点共现频率。我们定义知识图 $\mathcal{G}^k$ 的邻接矩阵为 $A \in \mathbb{R}^{M \times M}$，其中 $M = C + K$，C 和 K 分别为目标和谓词类别的数目。其中 $A(i, j)$ 代表给定节点 i 的情况下出现节点 j 的先验频次，因此我们可以通过边 $i \to j$ 来推理节点 j 的标签。邻接矩阵的不同部分代表不同类型的边的权值，例如 $A_{s\to o}$ 对应 $s \to o$，$A_{s\to o}^T$ 对应 $o \to s$，$A_{s\to p}$ 对应 $s \to p$，以及 $A_{p\to o}^T$ 对应 $o \to p$。I_p 是一个单位矩阵。在进行图推理之前，通过除以相应行的总和来归一化每个

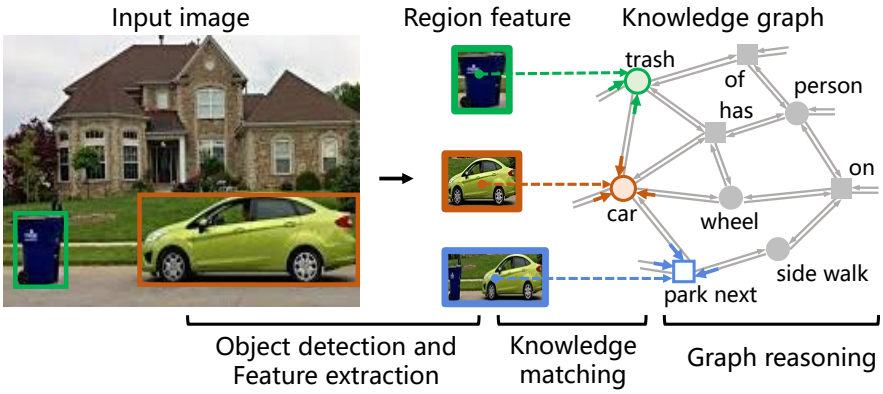


图 5.2 图推理模块示意图

Figure 5.2 Illustration of the graph reasoning module

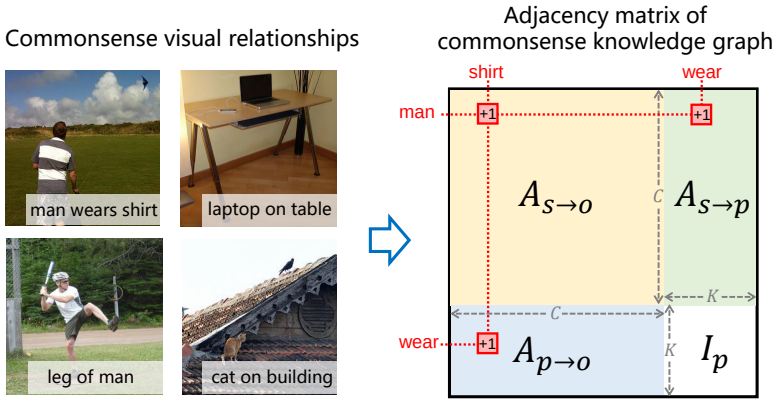


图 5.3 常识知识图的构建

Figure 5.3 Construction of the commonsense knowledge graph

子推理路径的子邻接矩阵。从含有丰富的关系标注的数据集中，我们可以收集包含两个目标及其之间的关系的知识三元组，例如图5.3中的 {man, wear, shirt}。三元组中的每两个元素可记共现频率一次，对于每个视觉关系 $\{s,p,o\}$，我们对 $A(s,p)$，$A(p,o)$，$A(s,o)$ 增加 1。为了构建节点编码，我们使用现有的词向量特征[68] 作为知识图 $\mathcal{G}^k$ 中每个实体的语义编码，定义为 $S=\{s_n\}, s_n\in\mathbb{R}^L$。

知识匹配：为了将每个子路径的推理与知识图相结合，CGR 将路径中节点的特征表示投影到由知识图中的 M 个语义实体组成的统一语义空间中。CGR 通过汇总每个图像的局部视觉特征 X 的语义投票，生成知识图中实体的视觉表示 H^{lt}：

$$H^{lt}=(\sigma(XW^l))^T XW^a, \tag{5.1}$$

其中 $\sigma(\cdot)$ 是 softmax 激活函数用以归一化不同节点的概率值使其和为 1。为了

将输入特征映射到包含 M 个标签的语义空间，我们定义一个可学习的转换权值 $W^l \in \mathbb{R}^{D^d \times M}$。为了降低输入特征的维度以防止过拟合，我们定义 $W^a \in \mathbb{R}^{D^d \times D^c}$。

图推理：在建立了知识图和每个视觉关系之间的语义联系之后，以结构化知识为指导的推理被用来生成实体节点的表示，生成过程受人类常识知识语义的约束。CGR 通过矩阵乘法形式在所有实体节点的表示 H^{lt} 上进行图传播，从而得到进化的特征 H^{st}：

$$H^{st} = \sigma(A[H^{lt}, S]W^s), \tag{5.2}$$

其中 $W^s \in \mathbb{R}^{(D^c+L)\times D^c}$ 是一个可学习的权值矩阵。直接使用原始的邻接矩阵 A 进行运算会改变特征的尺度，为了解决这个问题，我们参考 GCN[69] 将 A 归一化使其所有行的和分别为 1，即 $Q^{-\frac{1}{2}}AQ^{-\frac{1}{2}}$，其中 Q 是 A 的对角的节点度矩阵，因此可得新的传播法则：

$$H^{st} = \sigma(\hat{Q}^{-\frac{1}{2}}\hat{A}\hat{Q}^{-\frac{1}{2}}[H^{lt}, S]W^s), \tag{5.3}$$

其中 $\hat{A} = A + I$ 是图 $\mathcal{G}^k$ 的邻接矩阵并增加了自连接。

特征更新：集成了相关知识和语言信息的更新表示 H^{st} 可被用于提高每种视觉关系的不同子路径的特征表示的能力。由于特征在匹配知识图的实体后已发生更改，因此我们需要再映射回节点语义空间，将有关常识的消息从 H^{st} 传递回视觉特征 X：

$$\begin{aligned} H^{rt} &= \sigma([\hat{X}, \hat{H}^{st}]W^q)H^{st}W^p, \\ Y &= X + H^{rt}, \end{aligned} \tag{5.4}$$

其中 $W^q \in \mathbb{R}^{(D^d+D^c)\times 1}$ 是一个可学习的权重矩阵，用来估计 H^{st} 和子路径节点特征 X 的相容性。为了实现上述目的，我们将 X 扩展成 $\hat{X} \in \mathbb{R}^{N\times M\times D^d}$，将 H^{st} 扩展成 $\hat{H}^{st} \in \mathbb{R}^{N\times M\times D^c}$。$W^p \in \mathbb{R}^{D^c\times D^d}$ 将特征映射回和输入特征同维度的向量，得到 H^{rt}。把 H^{rt} 作为 X 的一个残差连接，得到更新后的子路径节点特征 $Y = \{y_i\}_{i=1}^n, y_i \in \mathbb{R}^{D^d}$。

如上所述，CGR 通过三个步骤执行图形推理：（1）使每个图像中的视觉元素与知识图的实体相匹配，如公式5.1；（2）在实体的视觉表示中传播语义信息，并在全局视角下更新特征，如公式5.3；（3）使全局更新后的特征适应图像中每个关系的表示，如公式5.4所示。公式5.1和公式5.4共同作用以弥合知识图和每个实际场景图之间的语义鸿沟，如图5.4所示。

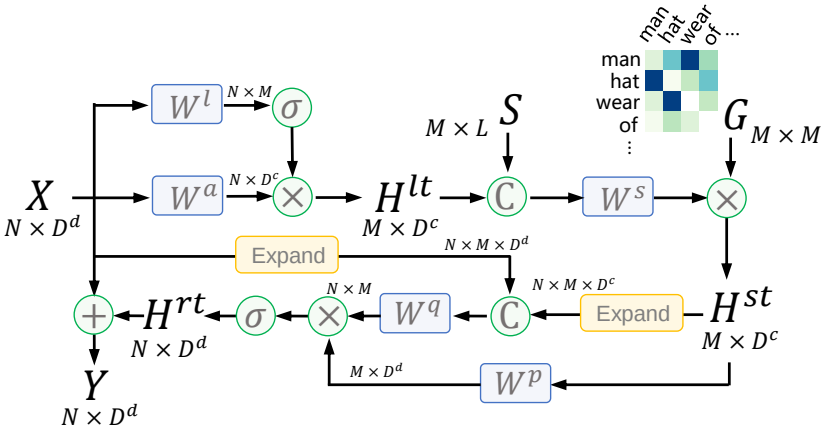


图 5.4 基于常识知识图的图推理过程

Figure 5.4 Illustration of the graph reasoning module with visual commonsense knowledge

常识知识图 $\mathcal{G}^k$ 是一个异构图，包含两种节点（目标标签和谓词标签）以及这些节点之间的不同类型的边（例如，主体 → 客体，谓词 → 客体，以及客体 → 谓词等）。用于不同子路径的图推理模块可以激活 $\mathcal{G}^k$ 中的不同类型的连接边。通过对知识图推理和路径配置的联合学习，我们的 CGR 能够选择常识知识中有效的语义连接，从而促进视觉关系预测。我们将综合知识图用于图推理模块的原因有三个方面：（1）关于场景图的常识知识图自然包含从目标到目标或谓词的异构语义信息。邻接矩阵 A 的不同部分描述了可以在图推理过程中独立激活的不同子路径的语义相关性，并且对应的配置模块将忽略无效的知识增强。因此，没有必要将整个图分成具有同构连接的子图。（2）在知识匹配过程中，图推理模块的输入被投影到新的特征语义空间，向知识图的实体节点进行投票和匹配。在异构图中，允许输入元素对目标节点和谓词节点进行投票，从全局语义理解的角度执行子路径上的图形推理。（3）我们进行了实验以比较这两种设置，使用异构图的 CGR 性能与使用一组同类图的 CGR 性能相当。

5.2.3 图推理路径配置

我们不能确保来自开放域的常识性知识中的所有图连接都可以有益于视觉关系的推理，因此需要在进行关系推理时舍弃无效的知识，以避免在推理中引入噪声分布。我们设计图推理配置模块，以探索如何做出关于是否将知识图推理纳入每个子推理路径的决策。

配置模块：我们希望可以通过对端到端训练的模块配置的离散决策来反向传播梯度。因此，我们选择基于非线性变换 $f(\cdot)$ 的每个路径的知识图推理，然后

使用 Gumbel-Max 及其连续的 softmax 松弛[70–73]。样本 $\boldsymbol{z}$ 可以采样自类别分布 $\boldsymbol{\pi} = \{\pi_1, ..., \pi_k\}$，如：

$$\boldsymbol{z} = \text{one\_hot}\Big(\text{argmax}_{i\in\{1,...,k\}}(log(\pi_i) + o_i)\Big), \tag{5.5}$$

其中 $\boldsymbol{o} = -log(-log(\boldsymbol{u}))$，且 $\boldsymbol{u} \sim \text{Uniform}(0, 1)$。类别变量 $\boldsymbol{z}$ 在这里是一个 k 维的 one-hot 向量。在 CGR 中，是否要使用常识知识的决策是一个二元的离散值，因此我们设定 $k = 2$。$\boldsymbol{z}$ 是一个 2 维向量表示离散决策结果，z_0 作为一个布尔值表示是否要使用常识知识更新输入特征。Gumbel-Max 的 softmax 松弛方法是用连续的 softmax 函数代替公式5.5中的不可微分 argmax 操作：

$$\hat{\boldsymbol{z}} = \text{softmax}((log(\pi_k) + \boldsymbol{o})/\tau), \tag{5.6}$$

其中 Softmax 函数的温度值 τ 在我们实验中设为 1。在训练过程中，配置模块根据公式5.5输出 $\boldsymbol{z}$ 以继续前向传播，并且根据公式5.6计算基于 $\boldsymbol{\pi}$ 的梯度以进行反向传播。在测试阶段，我们不进行 Gumbel 采样，直接取输出概率的最大值。配置不同子推理路径的最简单方法是直接决定是否在使用每个基于常识的推理路径的输出 $y_i \in Y$ 来代替原始输入特征，换句话说，即通过 Gumbel-Softmax（GS）为每个 y_i 估算 e_i：

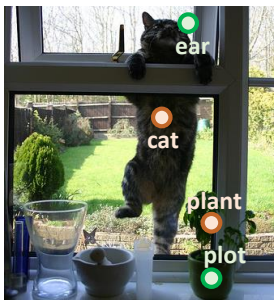
$$e_i = GS(W^{gp} y_i), \tag{5.7}$$

其中 $W^{gp} \in \mathbb{R}^{D^d \times 2}$ 代表可学习的参数，e_i 是一个 2 维向量用以决策是否要使用知识增强后的特征 y_i 进行最终的预测。我们定义配置模块为函数 $\phi(\cdot)$，其输入为图推理模块的输出 y_i，输出为 y_i 或 x_i：

$$\phi(y_i) = e_{i,0} x_i + e_{i,1} y_i. \tag{5.8}$$

如果 $e_{i,1} = 1$，我们使用 y_i 作为分类器的输入特征，否则的话，我们使用 x_i。y_i 它集成了视觉常识知识，可以促进视觉关系检测。CGR 总共包含四个配置模块，对应于不同的子推理路径。这些模块都被插入在图推理模块之后，用以决策是否引入知识增强的特征。这样的路径级配置模块确定知识图推理对不同子路径的贡献，以用于视觉元素之间的关系预测，从而可为每个视觉关系实例动态地集成知识。

相关探讨：图5.5中展示了不同的配置模块的输出以及其选择集成的常识知识连接。可以看到，CGR 能够为视觉关系中的元素动态地选择更具有决定性的路



Outputs of the configuration modules

ϕ_{sp}	ϕ_{op}	ϕ_{so}	ϕ_{os}	Reasoning Paths
1	1	1	0	cat → ear, cat → has
0	1	0	0	plant → plot, plot → in

图 5.5 配置模块的输出结果示例

Figure 5.5 Examples of the outputs of each configuration module

径依赖。在第一行，每个子路径的配置模块输出依次为 $\phi_{sp}=1$，$\phi_{op}=1$，$\phi_{so}=1$ 和 $\phi_{os}=0$，说明我们的模型选择了引入常识知识连接 $\{cat \rightarrow ear\}$，$\{cat \rightarrow has\}$ 和 $\{ear \rightarrow has\}$ 来辅助增强推理，以促进对视觉关系 $\{cat, has, ear\}$ 的识别。在第二行，只有知识 $\{plot \rightarrow in\}$ 被选择，而其余路径都被抛弃。

配置模块在三个方面为基于 CGR 的知识图推理带来了优势：（1）该模块在引入知识之前和之后的特征表达中进行选择，确保有效地增强知识并丢弃有偏见和嘈杂的常识知识连接；（2）它探索每种关系的个性化图形推理路径，并以自适应方式选择高度可预测的推理证据；（3）由于该模块能够从常识知识图中过滤出关键信息，因此 CGR 可以很好地兼容从不同领域收集的知识。

5.2.4 视觉关系检测

视觉关系检测致力于检测图像中的目标以及预测每对目标之间的语义关系，CGR 用于视觉关系检测的模型框架图如图5.6所示。目标区域特征 $V=\{v_i\}_{i=1}^{N}, v_i \in \mathbb{R}^{4096}$ 会被两两配对，得到主体特征 $V^s=\{v_i^s\}_{i=1}^{N}$ 和客体特征 $V^o=\{v_i^o\}_{i=1}^{N}$，他们对应的谓词特征为 $V^p=\{v_i^o\}_{i=1}^{N}$。每个子路径上图推理模块可定义为函数 $g(\cdot)$，其输入为目标或关系的区域特征，通过引入知识连接辅助更新特征，再输出更新后的特征。如图5.6所示，我们为每个子路径实例化一个图推理模块 g，即 $g_{s\rightarrow o}$，$g_{s\rightarrow p}$，$g_{o\rightarrow p}$，和 $g_{o\rightarrow s}$。这些模块采用不同的特征作为输入，并从常识图中提取不同类型的语义相关性，以辅助对应的子路径上的推理。

目标分类： 主体和客体的特征经过子路径 $s \rightarrow o$ 和 $o \rightarrow s$ 上的图推理模块 $g_{s\rightarrow o}$ 和 $g_{o\rightarrow s}$ 更新。在推理模块之后紧接着是路径配置模块 ϕ_{so} 和 ϕ_{os}，它们决策是否要使用图推理模块输出的知识增强后的特征。我们使用多层感知器（MLP）作为目标分类器，为主体和客体目标预测类别标签 $L^s=\{l_i^s\}_{i=1}^{N}, L^o=\{l_i^o\}_{i=1}^{N}$ 其

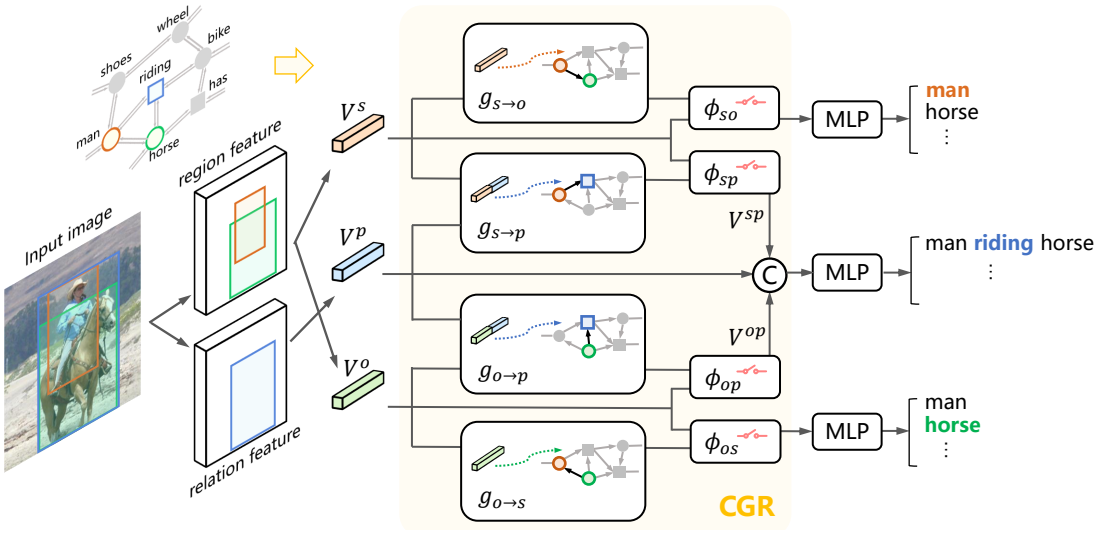


图 5.6 基于可配置图推理的视觉关系检测框架示意图

Figure 5.6 Overview of Configurable Graph Reasoning (CGR) for visual relationship detection

中 $l_i^s, l_i^o \in \mathbb{R}^C$:

$$\begin{aligned} L^s &= \text{MLP}(\phi_{so}(g_{s\to o}(V^s))), \\ L^o &= \text{MLP}(\phi_{os}(g_{o\to s}(V^o))). \end{aligned} \tag{5.9}$$

类似地，我们实例化另外两个图推理模块 $g_{s\to p}$ 和 $g_{o\to p}$，用以向路径 $\{s \to p\}$ 和 $\{o \to p\}$ 集成对应的视觉常识知识。相应地，我们使用两个配置模块 ϕ_{sp}, ϕ_{op} 来决策是否使用更新后的特征，以及一个 MLP 分类器来预测谓词标签 $L^p = \{l_i^p\}_{i=1}^N, l_i^p \in \mathbb{R}^K$，如:

$$\begin{aligned} V^{sp} &= \phi_{sp}(g_{s\to p}(\text{concat}(V^s, V^p))), \\ V^{op} &= \phi_{op}(g_{o\to p}(\text{concat}(V^o, V^p))), \\ L^p &= \text{MLP}(\text{concat}(V^{sp}, V^p, V^{op})). \end{aligned} \tag{5.10}$$

除了视觉特征以外，我们还考虑目标区域的空间信息以及主体和客体目标的标签。基于此我们可以预测最终的关系类别标签如下:

$$l_i^p = \text{softmax}(l_i^p + l_i^{spt} + l_i^{emb}),\ i = 1, ...N. \tag{5.11}$$

一个视觉关系的预测结果包含:（1）主体和客体目标的标签 l_i^s, l_i^o;（2）主体和客体目标的框坐标 b^s, b^o;（3）主体-客体对之间的关系谓词标签 l_i^p。

5.3 实验结果及分析

本节我们在常用的基准上评估 CGR 的性能，首先简单介绍评估方法细节，接着将 CGR 的性能与现有方法进行对比，最后进一步测试 CGR 对常识知识的兼容性。

5.3.1 实验设定介绍

数据集：我们使用两个视觉关系检测数据集 VRD（Visual Relationship Detection）[59] 和 VG（Visual Genome）[61]，来评估我们的方法。VRD 包含 5000 张图片（4000 张训练图片，1000 张测试图片），有 100 个目标类别和 70 个谓词类别。VG 包含 89,189 张图片（其中 62,723 张用于训练，26,446 张用于测试），有 150 个目标类别和 50 个谓词类别。

评测任务和指标：我们使用经典的四种视觉关系评测子任务来进行模型评测，**SGGen**，**PhrDet**，**SGCls**，和 **PredCls**。我们使用 Recall@K（R@K）和 mean Recall@K（mR@K）作为评测指标。子任务和指标的详细介绍见第4章。

5.3.2 常识知识分析

统计信息：在实验中我们使用三种常识知识图：KG_S、KG_M 和 KG_L（其中 S、M 和 L 表示知识图的规模：小、中、大)。三个知识图分别统计来自 VRD、VG 和 GQA 数据集，分别包含 30,355、439,063、3,795,907 个关系，100、150、1,073 个目标类别，以及 70、50、311 个谓词类别。

知识先验的质量：在图5.7中，我们研究从不同子路径上的常识知识图中可以获取多少有价值的信息。具体地，我们直接使用标签的统计作为先验去猜测其连接的可能的主体 s、客体 o 和谓词 p 的标签。我们在 VRD 和 VG 数据集上评估 Top-K 猜测的精度，使用该数据集对应的知识 VRD-KG_S 和 VG-KG_M 以及来自不同域的知识 VRD-KG_L 和 VG-KG_L，Accuracy 值更高的曲线说明该路径对于预测元素标签更加可依赖。从图中可以看出，对于知识 VG-KG_M 和 VRD-KG_S 路径 $s, o \rightarrow p$ 的猜测精度更高，此时知识和测试样本来自同源数据集。然而，当使用跨域的知识进行猜测时，例如 VG-KG_L 和 VRD-KG_L，路径 $s, o \rightarrow p$ 的猜测精度相比于其他四种子路径显著降低，这说明了我们的子推理路径具有一定的稳定性和泛化性。

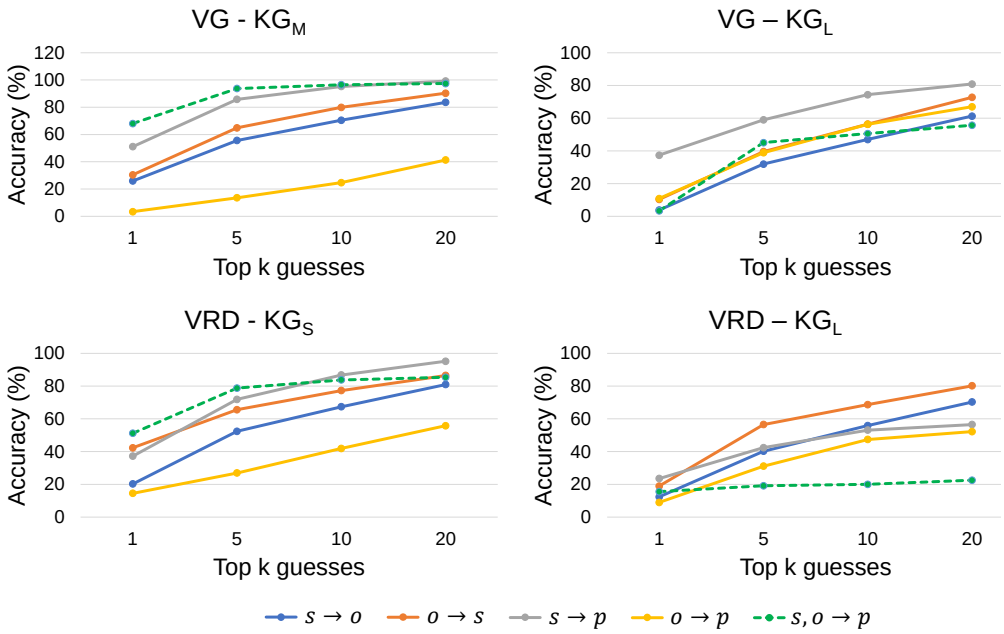


图 5.7 Top-K 猜测精度统计图

Figure 5.7 Accuracy (%) of the top-k guessing results

5.3.3　与已有方法进行对比

VRD 和 VG 的性能：表5.1和表5.3中的结果显示，CGR 在 VRD 和 VG 数据集上的评测结果均优于现有方法。我们还在 VRD 数据集上评估了使用不同的 Faster-RCNN 检测器基网的性能，包括使用 VGG-16[26] 作为基网分别在 ImageNet 和 COCO 上预训练的网络，对于 VG 数据集，我们使用 VGG-16 和 ResNeXt-101-FPN[33,74] 作为基网。CGR 在 PredCls 和 SGCls 任务上的提升比较明显，说明了我们方法挑选常识知识的能力。相对地，在 SGDet 和 PhrDet 上提升较小，因为性能受到了目标检测器基网的限制。显然，更强大的检测器将在 SGDet 和 PhrDet 上实现明显更好的性能。在将来的工作中，我们将考虑通过联合优化检测器主干和关系检测网络来解决此问题。在表5.3中，我们展示模型在 VG 数据集上的 mean Recall@K（mR@K）得分。可以看到 CGR 在 SGCls 和 PredCls 上的性能表现优于现有方法，证明了我们方法在不频繁关系类别上的性能优势。

耗时分析：在表5.2中我们还比较了不同方法的耗时，为了获得最终的检测结果，MOTIFS[17] 需要两个阶段的训练：首先使用真实标注框训练模型，然后使用目标检测器预测的目标框对整体模型进行微调。相比之下，KERN[18] 在模型训练过程中还要多出一步，在进行训练之前，KERN 使用真实目标框和标签训练模型。VCTree[15] 首先训练关系分类器，然后优化目标连接的树结构，随后针对

Pretrain	Method	SGCls			PredCls			SGDet		PhrDet	
		R@20	R@50	R@100	R@20	R@50	R@100	R@50	R@100	R@50	R@100
VRD	IMP[64]	24.0	25.4	25.6	45.2	48.6	49.0	13.4	15.2	27.7	28.2
	MOTIFS[17]	24.2	25.9	26.1	49.5	53.5	54.1	15.4	16.8	26.8	28.1
	CGR (Ours)	**25.8**	**27.6**	**27.8**	**51.3**	**55.3**	**56.0**	**15.8**	**16.9**	**28.2**	**29.3**
ImageNet	RelDN[67]	34.8	34.8	34.8	52.6	52.6	52.6	21.5	26.4	28.2	**35.4**
	CGR (Ours)	**35.8**	**37.3**	**38.1**	**55.5**	**57.0**	**57.7**	**21.7**	**26.5**	**28.7**	**35.4**
COCO	RelDN[67]	34.7	34.7	34.7	52.4	52.4	52.4	28.2	33.2	**34.5**	42.1
	CGR (Ours)	**36.6**	**37.5**	**38.0**	**55.9**	**57.4**	**58.1**	**28.4**	**34.6**	**34.5**	**42.2**

表 5.1 CGR 和已有方法在 VRD 数据集上的 Recall@K 对比

Table 5.1 Comparison of our CGR with existing methods on the VRD dataset

不同的评估任务训练不同的模型。使用 Faster-RCNN 主干检测到的目标区域直接训练 RelDN 和我们的关系模型，如表5.2中所示，该模型的训练时间比 VCTree 和 RelDN 少。

Method	SGCls			PredCls			SGDet				
	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100	Mean	Time
IMP[64]	31.7	34.6	35.4	52.7	59.3	61.3	14.6	20.7	24.5	37.2	-
MOTIFS[17]	32.9	35.8	36.5	58.5	65.2	67.1	21.4	27.2	30.3	41.7	0.4
KERN[18]	-	36.7	37.4	-	65.8	67.6	-	27.1	29.8	-	0.9
VCTree[15]	35.2	38.1	38.8	60.1	66.4	68.1	22.0	27.9	31.3	43.1	2.5
RelDN[67]	36.1	36.8	36.8	66.9	68.4	68.4	21.1	28.3	**32.7**	43.9	4.3
CGR (Ours)	**36.2**	**38.9**	**40.8**	**67.8**	**69.3**	**70.3**	**21.3**	**28.4**	**32.7**	**45.1**	1.6
RelDN (X-101-FPN)[67]	38.2	38.9	38.9	67.2	68.7	68.8	22.5	31.0	36.7	45.7	-
CGR (X-101-FPN)	**38.5**	**40.0**	**41.1**	**68.3**	**69.7**	**70.8**	**22.9**	**31.3**	**36.8**	**46.6**	-

表 5.2 CGR 和已有方法在 VG 数据集上的 Recall@K 对比

Table 5.2 Comparison of results and time costs for CGR and existing methods on VG

零次命中学习：在真实场景中，视觉关系检测模型需要能够预测未见过的关系，因为训练数据不会涵盖所有可能的关系组合类型。Lu 等人[59] 使用词嵌入将相似的关系投影到未知关系上，Liang 等人[75] 使用了一个大型语义动作图来学习共享节点上的相似关系。在表5.3中报告的结果表明，CGR 有助于预测不频繁出现的关系。表5.4将 CGR 与零次命中学习设置中的最新方法进行比较，以验证该方法检测未知关系的能力（例如，测试集的主体-谓词-客体组合，但不会出现在训练集中）。VRD 数据集包含 37,993 个关系实例，其中 1,168 个关系组合仅在测试集中而不在训练集中出现。对于 VG 数据集，我们对 7601 个未见关系进行

Method	SGCls		PredCls	
	mR@50	mR@100	mR@50	mR@100
IMP[64]	6.8	7.2	11.9	12.9
MOTIFS[17]	7.7	8.2	14.0	15.3
KERN[18]	9.4	10.0	17.7	19.2
VCTree[15]	10.1	10.8	17.9	19.4
RelDN[67]	11.0	11.0	22.2	22.3
CGR (Ours)	**12.4**	**13.0**	**23.2**	**23.5**
RelDN (X-101-FPN)[67]	11.7	11.7	22.5	22.5
CGR (X-101-FPN)	**13.4**	**13.8**	**24.4**	**24.6**

表 5.3 CGR 和已有方法在 VG 数据集上的 mean Recall@K 对比

Table 5.3 Comparison of mean Recall results for CGR and existing methods on VG

Dataset	Method	SGCls	PredCls
VRD	RelDN[67]	35	113
	CGR (KG_L)	**94**	**220**
VG	RelDN[67]	80	471
	CGR (KG_L)	**434**	**1613**

表 5.4 零次命中学习设定下正确预测的未知关系

Table 5.4 Number of correctly detected relationships under the zero-shot setting

评估，如表5.4所示，在 SGCl 和 PredCl 上我们方法比 RelDN 检测到更多关系。

可视化分析：配置模块为每个场景中的子推理路径选择知识增强表示，同时激活常识中的对应图形连接。在图5.8中我们统计总结了被各个配置模块激活的知识连接。图5.8展示路径 $\{s \rightarrow p\}$ 上的配置模块 ϕ_{sp} 频繁激活的知识。其中 $woman$ 和 of 常常被激活，因为它们往往比固定的多依赖路径更加通用和稳定。知识连接 $\{woman \rightarrow lying\ on\}$ 频繁地被 ϕ_{sp} 激活，在不关注具体目标标签的情况下促进视觉关系推理。当不考虑变化的目标标签时，如 $sofa$ 或 bed，路径 $\{woman \rightarrow lying\ on\}$ 是更加通用和稳定的。由此 CGR 动态地组成了高度可预测的推理路径，用于识别每个视觉关系。

图5.10展示我们方法的一些可视化示例,绿色框和边缘代表正确的预测(True positive)。预测中遗漏了蓝色框和边缘（False negative)。橙色边缘表示我们模型

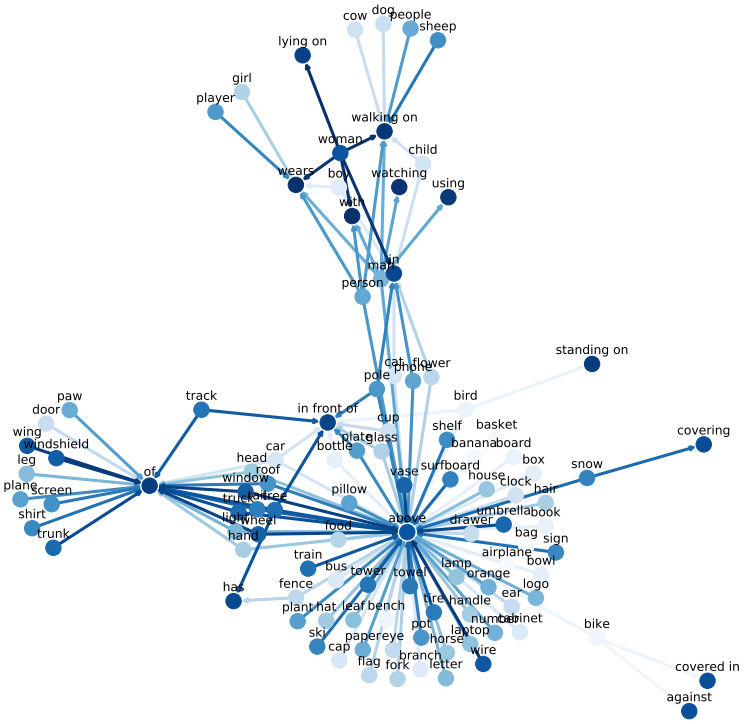


图 5.8 ϕ_{sp} 在路径 *subject* → *predicate* 上激活的知识图

Figure 5.8 Learned knowledge graph w.r.t *subject* → *predicate* (selected by ϕ_{sp})

的合理预测但未被标注（False Possitive）。从示例中可看出，我们的模型可以识别复杂和混乱场景中的各种目标和关系。此外，根据场景图示例中的橙色边缘，我们发现 CGR 可以做出注释中未包含的合理的关系预测。

如图5.10中的第一个图片，我们的方法正确地预测了 “wheel-1 on truck-1” 和 “sky-1 above truck-1”，此外，还预测了相似的关系，如 “sky-1 above building-1”，其对应的标注为 “sky-1 in the front of building-1”。这是因为 CGR 可以动态地合并常识知识，以发现场景中其他的合理的关系。

5.3.4 消融研究

不同的推理子路径：如表5.5所示，我们删除了 VRD 数据集上不同子路径的配置模块，并对模型进行评估。我们使用在 COCO 上预训练的检测器基网。第一行显示从 CGR 中删除所有配置模块的结果，这样所有子路径都集成了常识知识，而没有丢弃任何无效和嘈杂的知识连接。第二行到第五行显示了 CGR 使用单个配置模块来动态预测是否对每个样本在不同的子推理路径上使用知识增强

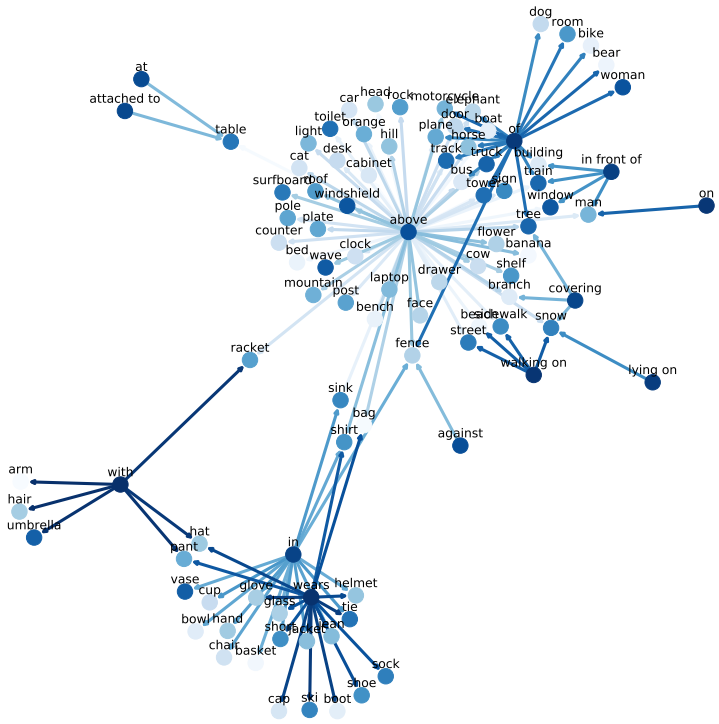


图 5.9 ϕ_{op} 在路径 *object* → *predicate* 上激活的知识图

Figure 5.9 Learned knowledge graph w.r.t *subject* → *predicate* (selected by ϕ_{sp})

功能，结果显示 CGR 在 SGDet 子任务上获得的一致性提升。在最后一行，CGR 具有用于所有子路径的配置模块，并且能够从知识图中选择语义相关性，从而有利于视觉关系检测任务。

Gumbel-softmax 和 Softmax： CGR 使用 Gumbel-softmax 来学习离散决策，选择是否接受知识增强。为了研究 CGR 选择性接受知识的有效性，我们直接用 softmax 代替 Gumbel-softmax 来预测概率概率权重，以结合图推理之前和之后的特征。如表5.5的第六行所示，CGR（softmax）的性能略低于 Gumbel-softmax，这证明了我们方法的挖掘有利于视觉关系预测的知识连接的能力。

5.3.5　知识兼容性

视觉元素之间语义相关性的常识知识是非常普遍且通用的，并且可以在不同的任务和数据集上自然地转移。我们用从不同数据集和不同域中总结的常识性知识图对图推理模块进行实例化，并评估对应的 CGR 模型的性能。为了减轻目标检测结果的影响，我们报告了 PredCls 任务上的 R@K 结果，该任务要

图 5.10 CGR 视觉关系检测示例图

Figure 5.10 Qualitative results of the proposed CGR on the VRD dataset

求使用真实标注的目标框坐标和标签来预测该主体-客体对之间的关系类别。我们使用 L1 距离 $d_{l1}(\cdot,\cdot)$ 来评估不同知识图之间的相似性。计算出的知识距离为 $d_{l1}(KG_S, KG_L) = 196.8$，$d_{l1}(KG_S, KG_M) = 143.4$。$KG_L$ 和 KG_S 之间的差异大于 KG_M 和 KG_S 之间的差异，我们从跨数据集和跨任务的角度，使用知识图在 VRD 上进行泛化实验。

跨数据集知识：我们首先验证了 CGR 从知识图谱中选择有用信息并不受冗余知识干扰的能力。从表5.6的前两行中，我们可以看到，使用 KG_M 的 CGR 在 SGCls 和 PredCls 上的性能要高于使用 KG_S 的 CGR，因为 KG_M 是从 VG 总结而来，包含更多的关系，可以帮助 CGR 探索更完整的常识，从而有利于视觉关

Model	SGDet	
	R@50	R@100
CGR (w/o $\phi(\cdot)$)	27.1	33.4
CGR (only ϕ_{so})	28.1	34.2
CGR (only ϕ_{os})	27.9	34.1
CGR (only ϕ_{sp})	28.3	34.3
CGR (only ϕ_{op})	27.6	33.9
CGR (softmax)	28.0	34.2
CGR (ours)	**28.4**	**34.6**

表 5.5 图推理配置模块的消融研究

Table 5.5 Ablation study of the configuration modules

KG	SGDet		PhrDet	
	R@50	R@100	R@50	R@100
CGR (KG_S)	27.8	33.7	33.7	41.3
CGR (KG_M)	27.6	34.0	33.4	41.6
CGR (KG_L)	**28.4**	**34.6**	**34.5**	**42.2**

表 5.6 VRD 数据集上跨任务和跨数据集的知识泛化性结果

Table 5.6 Cross-task and cross-dataset knowledge generalization on VRD dataset

系识别。

跨任务的知识: 为了进一步验证该方法是否能与其他任务领域的知识一起很好地工作，我们使用源自 VQA 任务的 KG_L 在 VRD 数据集上训练了 CGR。从表5.6的最后一行可以看出，KG_L 的 CGR 的性能明显优于 KG_S 和 KG_M 的 CGR 的性能，说明 CGR 在为每个视觉关系的自适应推理的子路径提取通用和稳定的知识联系方面具有显著的能力。

5.4　本章小结

基于固定的关系推理路径引入先验知识，对知识图的质量和数量有着较严苛的要求，使得模型对于不同知识图的的泛化性受限。针对此问题，本章提出了

将固定的关系推理路径进行拆分和动态组合，以兼容更通用的视觉常识知识。我们设计了用于检测视觉关系的可配置图推理（CGR）方法。CGR 使用多个子路径探索每种视觉关系的自适应推理路径，并有选择地将常识知识纳入关于视觉关系推理的不同子路径。CGR 与从其他数据集和域中总结的一般关系知识能够很好地兼容，且不会受到大量无关知识的干扰。

第 6 章　基于场景图攻击的少样本视觉关系检测

不频繁的视觉关系类别的训练样本非常稀少，当可用的训练样本只有几个时，我们需要基于少样本学习训练视觉关系检测模型。该任务非常具有挑战性，一个视觉关系的类别往往涉及多种不同的目标组合，在标注示例十分有限的情况下，模型难以学习到准确的关系分类器。因此，模型常常面临训练不充分、鲁棒性较差的问题。一些工作，例如，Visual Relationships as Functions[76] 方法，通过引入辅助数据以促进少样本学习。为了减小辅助数据的标注代价，Image-Agnostic[77] 方法引入大量的未标记数据。该方法基于已标记数据为无标记数据分配伪标签，然后使用增广后的数据集训练传统的场景图生成模型[17]。但是这种生成伪标记的方式会产生大量嘈杂的监督信息，从而增加了训练数据偏差，降低了模型泛化能力。

6.1　模型概述

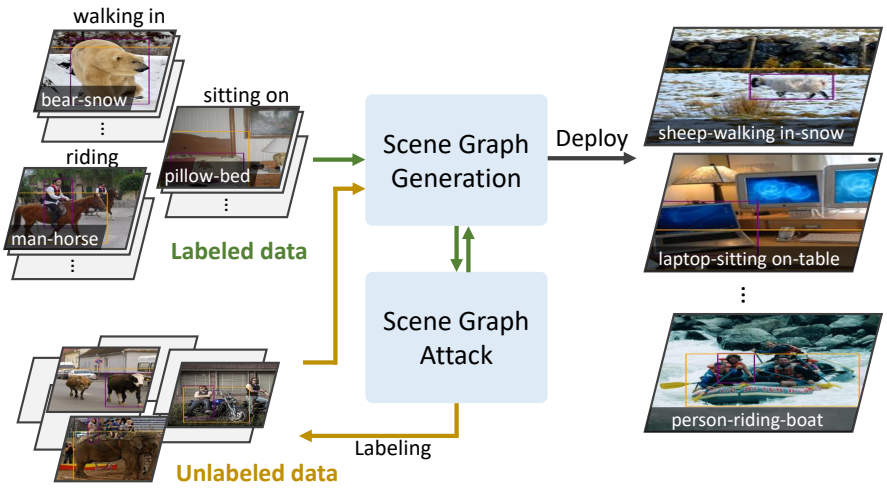


图 6.1 基于场景图攻击（SGA）的视觉关系检测

Figure 6.1 Scene Graph Attack for visual relationship detection

在本章，我们提出了一种场景图攻击 SGA（Scene Graph Attack）方法，其目的是学习一种对训练数据中的标记噪声更加鲁棒的视觉关系检测模型。我们的模型以攻击-防御-标记的方式进行学习，如图6.1所示。具体地，SGA 在场景图上对节点进行语义攻击，使得少量目标节点的预测标签发生改变。常规图攻击

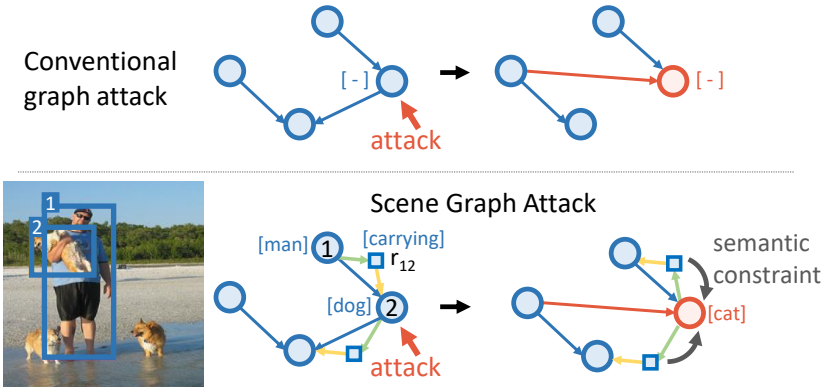


图 6.2 传统图攻击和场景图攻击（SGA）方法的对比

Figure 6.2 Comparison of the conventional graph attack and our SGA

方法致力于对目标节点进行错误分类，SGA 则考虑了场景图的全局语义一致性，还利用场景图中视觉元素之间的语义关联对攻击进行约束，如图6.2所示。在攻击过程中，模型学习预测新的节点标签并与原始场景图保持兼容，从而保证场景图的全局语义一致性。通过迭代的攻击和防御，关系检测模型可以学习通用且健壮的特征表达（例如，something that can be ride），而忽略与关联谓词（如 “ride”）有关的特定对象（例如 “person” 和 “bike”）的琐碎特征。SGA 产生的新关系三元组可被用于动态更新未标记数据的伪标签并减少嘈杂的标签，从而提供更多真实有效的训练样本。我们将 SGA 与场景图生成模型集成在一起，可在极少量带标记的样本上训练视觉关系检测模型。大量的实验表明，SGA 与最新方法相比，其性能明显更高。

6.2 场景图攻击方法介绍

在本节中，我们首先描述基于带标记的数据生成场景图的基本模型。然后，我们在生成的场景图上引入场景图攻击 SGA，为无标记数据生成伪标签。最后，我们提出了一种基于带噪声标记的关系检测方法。基于场景图攻击的少样本关系检测框架如图6.3所示。

6.2.1 场景图生成

我们使用基于 GCN（Graph Convolutional Network）[69] 构建的 Graph R-CNN[13] 模型来构建初始的场景图。给定一个节点特征集合 X，GCN 的逐层转播法则可

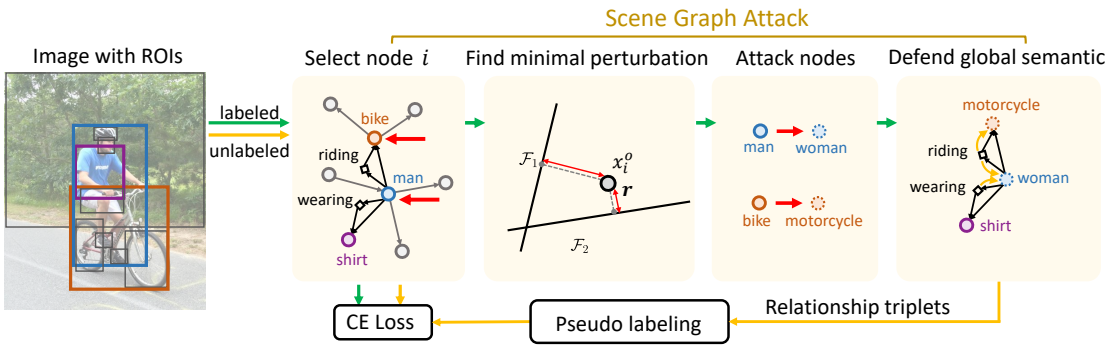


图 6.3 基于场景图攻击的少样本关系检测框架示意图

Figure 6.3 Overview of Scene Graph Attack (SGA) for few-shot visual relationship detection

被定义为：

$$\tilde{X} = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}XW\right), \tag{6.1}$$

其中邻接矩阵 A 被归一化为 $\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$，$\hat{A} = A + I$，其中 I 为单位矩阵，$\hat{D}$ 为 A 中对角节点的度矩阵。σ 为激活函数，W 代表模型参数。一个有向语义图 $\mathcal{G} = (V, E)$ 可基于一幅场景图像 I 构建而来。V 代表目标节点集合，E 代表连接每对目标节点的谓词集合。在我们基于 Graph R-CNN 所生成的场景图中，谓词既可被看作是边，也可被看作是节点。通过 GCN 传播，场景图 $\mathcal{G}$ 中的目标和谓词的特征表达将被更新。

用 $X^o = \{x_i^o\}_{i=1}^n, x_i^o \in \mathbb{R}^{d_o}$ 表示目标节点的特征，$X^r = \{x_i^r\}_{i=1}^m, x_i^r \in \mathbb{R}^{d_r}$ 表示谓词节点的特征，其中 n 和 m 代表目标和谓词节点数目，d_o 和 d_r 代表它们的特征维度。x_i^o 的初始值是由 Faster R-CNN[78] 计算的目标检测框基于 ROI-Align 层提取的区域特征，x_i^r 由每个谓词所连接的目标对的联合框的区域特征进行初始化。根据公式6.1，目标节点的特征表达被其所连接的谓词节点特征逐层地更新，而谓词节点特征由其相连的主体和客体节点特征更新，如：

$$\tilde{X}^o = \sigma(A^o W^o X^o + A^{sr} W^{sr} X^r + A^{or} W^{or} X^r), \tag{6.2}$$

$$\tilde{X}^r = \sigma(X^r + A^{rs} W^{rs} X^o + A^{ro} W^{ro} X^o), \tag{6.3}$$

其中，不同的邻接矩阵 $(A^{sr}, A^{or} \in \mathbb{R}^{n\times m}, A^{rs}, A^{ro} \in \mathbb{R}^{m\times n}, A^o \in \mathbb{R}^{n\times n})$ 和它们所对应的线性转换矩阵 W 捕获不同类型的有向连接：主体和谓词，客体和谓词，主体和客体。s，o，和 r 分别代表主体（subject），客体（object）和谓词（predicate）。n 和 m 代表场景图中目标和谓词节点的数目。A^o 基于 RePN[13] 计算而来，表示

每对目标之间的连接置信度，用以在初始的全连接场景图中剪除冗余的边，从而提升模型运算效率。其他的邻接矩阵都基于 A^o 计算而来。在图传播之后，更新的目标节点和谓词节点被输入到全连接层，以预测对应的类别标签。

6.2.2 场景图攻击

SGA 旨在通过考虑全局语义一致性来攻击场景图以更改一些节点的标签，从而迫使模型归纳学习与谓词相关的目标原型，从而为视觉关系生成更加稳定的特征表达。SGA 在 A^o 上添加扰动，对扰动进行变换从而影响基于 A^o 构建的其他邻接矩阵。图传播过程中，模型通过受到攻击的邻接矩阵更新节点特征以欺骗目标分类器。我们使用 $A = A^o$ 来表示邻接矩阵，$\tilde{X} = X^o$ 表示目标节点特征。$f(A, X)$ 表示基于图的目标分类模型，$g(A, X)$ 表示基于图的谓词分类模型。$f(A, X)$ 和 $g(A, X)$ 共享同样的基础 GCN 模型，使用不同的全连接层来分类目标节点和谓词节点。

选择目标节点： 考虑到具有较高的度的节点具有更多的连接，这些节点对场景的全局语义来做出更多贡献，因此我们基于场景图中节点的度来选择攻击节点。传统的图攻击方法只是在不考虑与其他节点的语义一致性的情况下对目标节点进行了错误分类，而 SGA 致力于为该节点找到微小的语义扰动，同时在全局语义范围内与其他节点保持兼容。SGA 没有直接在目标节点上添加扰动噪声，而是翻转 A 的少量边并使用更改后的邻接矩阵通过图传播[79] 修改节点特征。关系检测模型学习防御这样的语义攻击，从而提高了模型的泛化能力。

我们定义目标节点 i 的预测标签为 $\hat{l}(A, x_i) = \arg\max f(A, X)_i$，定义目标节点 i 和他的邻居节点 $j \in \mathcal{N}_i$ 之间的预测谓词标签为 $\hat{q}(A, x_i) = \arg\max g(A, X)_i$。SGA 致力于寻找一个扰动 $\boldsymbol{r} \in \mathbb{R}^n$ 以将邻接矩阵 A 改变为 A' 并且欺骗分类器 $f(\cdot)$ 以修改目标标签，并且保持其相连的谓词标签不变：

$$\begin{aligned} &\Delta(x_i; f) = \min_{\boldsymbol{r}} \|\boldsymbol{r}\|_2 \\ &\text{s.t. } \hat{l}(A', x_i) \neq \hat{l}(A, x_i) \text{ and } \hat{q}(A', x_i) = \hat{q}(A, x_i), \end{aligned} \tag{6.4}$$

其中，A' 的计算如下：

$$A' = h(A, i, r) = (\mathbb{1} - R) \circ A + R \circ (\mathbb{1}_0 - A), \tag{6.5}$$

其中 R 是来源于 $\boldsymbol{r}$ 的一个 $n \times n$ 矩阵，R 的初始值为 0，我们设置 $R_{i,:} = \boldsymbol{r}$。$R_{ij} = 1$

表示连接节点 i 和 j 的边被翻转。$\mathbb{1}$ 表示一个元素值全为 1 的矩阵，$\mathbb{1}_0$ 表示对角元素皆为 0 的全 1 矩阵。

寻找扰动：为了求解公式6.4，我们将多分类 DeepFool[80] 算法扩展到带有语义约束的版本。直观上，SGA 找到使目标分类器改变目标标签并将目标节点 i 送过另一个类的决策边界的最小扰动。我们使用 l_2 范数来衡量可能的扰动方向，以确定最小扰动：

$$k = \arg\min_{c \neq \hat{l}(A,x_i)} \frac{|f'_c|}{||w'_c||_2} \cdot ||e_c - e_l||_2^2, \tag{6.6}$$

其中，$w'_c = \nabla f(A', X)_{i,c} - \nabla f(A', X)_{i,l}$ 和 $f'_c = f(A', X)_{i,c} - f(A', X)_{i,l}$，简化起见，$l$ 代表 $\hat{l}(A, x_i)$。e_c 和 e_l 是对目标标签提取的语义编码词向量[21]。除了仅根据从节点到最近决策边界的距离来确定最小扰动，SGA 还在语义级别上加权了攻击前后节点标签之间的距离（$||e_c - e_l||_2^2$）进一步减小了对场景图全局语义的影响。确定目标类 k 之后，扰动 $\boldsymbol{r}$ 更新为：

$$\boldsymbol{r} \leftarrow \boldsymbol{r} + \Delta\boldsymbol{r}, \quad \Delta\boldsymbol{r} = \frac{|f'_k|}{||w'_k||_2^2} w'_k, \tag{6.7}$$

直到 $(1+\eta)\boldsymbol{r}$ 能够成功地攻击节点 i 使得其预测标签发生改变。我们使用 $\eta = 0.02$ 作为一个小的因子值，用以保证该节点能够穿越决策边界。我们还对新的邻接矩阵 A' 进行裁剪以保证稳定性。

全局语义一致性：关系检测模型通过保持连接的谓词标签不变来学习防御攻击，从而从全局角度生成与原始场景图在语义上一致的新的攻击后的关系三元组。全局语义防御通过 SGA 损失实现为：

$$\mathcal{L}_{SGA} = -\sum_{(x_i^r, t_i) \in D_l} t_i \, log(\hat{q}(A', x_i^r)), \tag{6.8}$$

其中 D_l 表示带标签的集合，每个关系样本 x_i^r 都有其对应的真实标注标签 t_i。

6.2.3　少样本关系检测

生成伪标签：现有工作[77] 在进行监督训练之前，对未标记数据进行预处理，得到伪标记用于模型训练。我们的模型在训练过程中迭代更新未标记样本的标签。每次成功的攻击将目标标签从 s 变为 s'，其相关的关系三元组 $\langle s - p - o \rangle$ 会相应地变为 $\langle s' - p - o \rangle$，其中 s，p 和 o 表示主体（subject）、谓词（predicate）和客体（object）的标签。在每个训练迭代中，关系三元组会被动态地收集到一个

知识矩阵 $M \in \mathbb{R}^{C\times C\times P}$ 中，该矩阵表示谓词标签和目标标签之间的统计分布。C 和 P 表示目标和谓词的类别数目。M 被迭代地更新并且给未标记的样本预测伪标签。伪标签 $M(s_i, o_i, \cdot)$ 是一个 P 维的归一化的概率向量，是使用主体和客体标签（s_i 和 o_i）从知识矩阵检索而来。我们的模型基于伪监督信息在未标记数据上进行训练：

$$\mathcal{L}_{PSE} = -\sum_{(x_i^r)\in D_u} M_{s_i,o_i,\cdot}\ log(\hat{q}(A, x_i^r)), \tag{6.9}$$

其中 D_u 表示未标记集合。随着越来越多的三元组的产生，知识 M 的容量和多样性得到了提升，可以收集谓词的各种示例，从而产生更准确的伪标记。我们还进行了人工评估，以检查新的三元组在有关关系的人类常识方面是否合理。另外我们的实验中也评估了伪标签的质量。

模型训练目标函数：我们模型的训练过程包括三个阶段：目标检测、场景图生成和场景图攻击。对于目标检测，与区域预选网络（RPN）一样，我们对目标预选使用二进制交叉熵（BCE）损失，对锚点使用回归损失函数。为了生成场景图，我们使用二元交叉熵损失来学习关系预选，如图 R-CNN[13] 所示，并且分别使用两个多类交叉熵损失进行目标和谓词分类。我们将场景图生成模型的损失表示为 $\mathcal{L}_{SGG}$，损失仅在带标记的集合上计算。对于场景图攻击，我们引入了另外两个损失函数，一个是 $\mathcal{L}_{SGA}$，用于限制对目标节点的攻击必须考虑与原场景图的全局语义兼容，如公式6.8所示。关系检测模型被迫学习通用且鲁棒的特征表达，同时忽略相关谓词类的特定目标的琐碎特征，同时收集 SGA 产生的新关系三元组以迭代更新伪标签。另一个损失函数是 $\mathcal{L}_{PSE}$，用于引入未标记数据以及带噪声的标签来训练更加鲁棒的模型，如公式6.9所示。最后一个阶段的模型训练所使用的损失函数为：

$$\mathcal{L} = \mathcal{L}_{SGG} + \mathcal{L}_{SGA} + \mathcal{L}_{PSE}. \tag{6.10}$$

关于 SGA 如何执行少样本关系检测的详细过程见算法2。

6.3 实验结果及分析

在本节，我们首先介绍实验设定以及对比的基线方法，然后将 SGA 模型与现有方法进行对比，并进一步研究上界和少样本学习的设定，最后我们评估生成的伪标签的质量，并对攻击过程和结果进行可视化。

算法 2 基于场景图攻击（SGA）的少样本关系检测

Input: 标记数据 D_l 和未标记数据 D_u。
Output: 网络参数，知识矩阵 M。
repeat
　对每个来自 D_l 的图片 I 生成场景图 $\mathcal{G}$。
　根据节点度的值对 $\mathcal{G}$ 中的节点进行排序。
　repeat
　　从 $\mathcal{G}$ 选择一个节点 x_i，其标签为 $\hat{l}(A, x_i)$。
　　初始化一个扰动向量 $\boldsymbol{r} \leftarrow 0$。
　　更新节点特征表示，如公式6.2和公式6.3所示。
　　计算被攻击后的邻接矩阵 A'，如公式6.5所示。
　　网络前向传播。
　　找到攻击后的新的节点类别标签 k，如公式6.6所示。
　　计算 $\Delta \boldsymbol{r}$ 以更新 $\boldsymbol{r}$，如公式6.7所示。
　until $i \geq n_{atk}$
　前向传播，计算损失，如公式6.10所示，更新 M。
　网络反向传播。
until 学习收敛

6.3.1　实验设定

数据集：我们在 VG 数据集[62] 上评估我们的方法，使用与[77] 中相同的数据集划分方式，即训练集 8045 张图像，验证集 1173 张图像和测试集 2358 张图像。VG 的少量标记样本包含 150 个对象类和 20 个谓词类，对于 N-shot 视觉关系检测，我们对每个谓词类别随机抽取 N 个关系示例作为标记数据。对于[77] 中的 10-shot 划分，D_l 包含 190 个图像和 20 个谓词类的 200 个关系。D_u 包含 7,875 张图像和 18,303 个未标记的关系，未标记的数据只保留目标级别的标注（边界框和类标签）。

评测任务和指标：我们在视觉关系检测的标准评估任务 SGDet，SGCls 和 PredCls 上评估模型，评价指标采用经典的 Recall@K（K=20，50，100）指标。

基线方法介绍：我们将 SGA 与基线方法在 VG 数据集上 10-shot 划分的测试集上进行比较。Freq 基线方法直接使用目标共现频率来预测测试集的关系标签。如果目标对的边界框重叠，则 Freq + Overlap 基线会为它们计算关系标签。

Transfer Learning 基线[81,82] 使用 VG 数据集中前 50 个关系中的源域，该域与先前选择的关系集合并不重叠。Decision Tree 基线[83] 基于图像无关特征[77] 拟合单个决策树。Label Propagation 基线[84] 采用一种半监督方法将类信息从已标记数据传播到未标记数据。图像无关特征基线[77] 丢弃图像视觉特征，仅根据用于学习预测未标记集的概率标记。

训练细节：我们使用以 ResNet101 为骨干的 Faster R-CNN 检测器，并使用随机梯度下降算法作为优化器在 VG 数据集上预训练检测器。我们的模型首先在没有 SGA 的情况下进行了 15,000 次迭代训练，并基于预先训练的 Faster-RCNN 权重使用两个交叉熵损失对目标和谓词进行分类。在此期间，知识矩阵 $\boldsymbol{M}$ 基于标记的集合 $\boldsymbol{D}_1$ 构造并固定不变，学习率设置为 0.005。随后，我们使用验证集选择最佳模型以进行场景图攻击。使用公式6.10中的 $\mathcal{L}_{SGA}$ 对模型进行额外的 15,000 次迭代训练，学习率设置为 0.001。根据公式6.7更新 $\boldsymbol{r}$ 的最大迭代次数设置为 10。模型训练期间 batch size 大小为 8，在四块 Titan 2080ti GPU 设备上花费 20 个小时（攻击过程花费约 6 个小时），在模型测试期间不进行场景图攻击。

6.3.2 数值结果

与已有方法进行对比：如表6.1所示，我们在 VG 的少样本数据集的测试集上评估 SGA 模型在 SGDet，SGCls 以及 PredCls 上的性能，基于指标 Recall@20，50 和 100。可以看到，SGA 模型比目前最佳性能在 SGDet 上高出 1.28%，1.84%，和 2.08%，证明了 SGA 在少样本关系检测任务上的优越性。模型在 PredCls 子任务上的提升更多，分别为 1.31%，3.72% 和 5.24%，进一步验证了我们模型分类仅有少量标注样本的谓词类别的能力。

性能上界分析：我们在表6.2中研究了 SGA 的几种上界方法，与每个谓词类仅使用 10 个标记示例的 SGA 相比，使用 18,344 个未标记关系的真实标签训练的 GT Labeling 模型将 SGDet 的结果大幅提高。我们基于 GQA 数据集[85] 中的 3,795,907 个视觉关系，收集目标和谓词标签的共现频率作为关系统计信息，并建立一个 GQA 标签模型，该模型使用从 GQA 计算出的固定知识矩阵 $\boldsymbol{M}$ 对未标记数据生成伪标签。实验结果表明，SGA 的性能与 GQA 标签模型的性能相当，这表明我们通过攻击防御所挖掘的知识矩阵中目标与谓词的分布接近 GQA 中更加通用更加泛化的常识知识。

人工评估： SGA 欺骗目标分类器来以改变目标预测标签，同时保持连接的

Model	SGDet			SGCls			PredCls		
	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
Freq	9.01	11.01	11.64	11.10	11.08	10.92	20.98	20.98	20.80
Freq + Overlap	10.16	10.84	10.86	9.90	9.91	9.91	20.39	20.90	22.21
Transfer Learning	11.99	14.40	16.48	17.10	17.91	18.16	39.69	41.65	42.37
Decision Tree[83]	11.11	12.58	13.23	14.02	14.51	14.57	31.75	33.02	33.35
Label Propagation[84]	6.48	6.74	6.83	9.67	9.91	9.97	24.28	25.17	25.41
Image-agnostic Feat.[77]	17.69	18.69	19.28	**20.91**	21.34	21.44	45.49	47.04	47.53
SGA (Ours)	**18.97**	**20.53**	**21.36**	19.76	**22.13**	**23.25**	**46.80**	**50.76**	**52.77**

表 6.1 SGA 与已有方法在少样本视觉关系检测任务上的性能对比

Table 6.1 Comparison of SGA with existing methods on few-shot visual relationship detection

Models	SGDet		
	R@20	R@50	R@100
GT Labeling	24.82	27.90	29.34
GQA Labeling	20.14	22.93	23.92
SGA (Ours)	18.97	20.53	21.36

表 6.2 SGA 模型性能上界分析

Table 6.2 Upper bound analysis of SGA results

谓词标签不变。为了查看新收集的关系三元组是否与人类关于视觉关系的常识知识相一致，我们进行了人工评估，要求三个人对攻击过程中收集的新三元组中的正确合理的关系进行投票，评估所得的平均接受率为 83.8%，这表明我们的模型很好地捍卫了全局语义，并产生了基于原始关系三元组的合理的变体。

6.3.3　消融研究

***N*-shot 设定：**在表6.3中，我们研究每个谓词类别的标记示例数量 N 对于模型学习的影响，在 N-shot 视觉关系检测设置下训练该模型，其中 $N = 3, 5, 7, 10$，并在测试集上进行评估。实验结果显示，通过为每个谓词类提供更多带标签的示例，SGA 模型在测试集上实现了更高的性能。同时，我们还比较了引入 SGA 训练前后模型的性能。可以看到，无论每个谓词类有多少示例可用，这些模型的性能在经过场景图攻击训练后都能实现一致的提升。

知识矩阵 *M*：我们研究不同的知识矩阵 M 对于未标记数据生成的伪标签，以及对模型学习的影响，如表6.4所示。首先，我们直接删除 M 并禁用伪标记，

N	After Attack			Before Attack		
	R@20	R@50	R@100	R@20	R@50	R@100
3	11.53	12.47	12.87	8.80	10.46	11.23
5	16.28	18.08	19.12	13.59	14.72	15.35
7	17.36	20.03	20.81	15.57	17.02	17.96
10	**18.97**	**20.53**	**21.36**	**17.95**	**19.93**	**20.62**

表 6.3 不同 N-shot 设定下攻击前后的 SGDet 性能对比

Table 6.3 SGDet results before and after attacking under different N-shot settings

Models	**SGDet**		
	R@20	R@50	R@100
without M	15.94	17.98	18.99
Use M without SGA	17.95	19.93	20.62
Simulate SGA	17.94	19.65	20.74
SGA (Ours)	**18.97**	**20.53**	**21.36**

表 6.4 关于知识矩阵 M 的消融研究

Table 6.4 Ablation studies about the knowledge matrix M

仅使用标记集训练模型而忽略未标记集上的损失函数。进一步地，我们还训练了没有 SGA 且仅使用基于标记集的固定知识矩阵的模型。在不迭代更新的情况下，知识矩阵将提供嘈杂且有偏见的伪标签。实验结果表明，使用 SGA 训练的模型可以动态更新 M 以生成更加准确的伪标记，并且比上述模型具有更好的性能。这证明了 SGA 模型能够学习更紧凑的关系特征并产生更准确的伪标签的能力，从而促进少样本关系检测。为了验证 SGA 不仅仅将目标的标签攻击改变为同义词类，我们禁用了整个攻击机制，并模拟了新的三元组关系以生成知识矩阵。如果两个目标标签的词向量[21] 之间的 L2 距离小于 0.5，则我们将其视为成功的攻击并生成对应的新的三元组。SGDet 的结果低于 SGA，这表明我们的攻击不仅将目标标签更改为同义词类，如 “man” → “boy”，而且能改变为相似的类别，如 “dog” → “cat”，甚至是共享相同谓词的模式，如 “kite” → “bird”。

生成伪标签：为了验证 SGA 生成的伪标签的质量，我们在未标记的集合上评估我们的模型。为了减轻目标检测模型的影响，我们选择 PredCls 任务，该任

N	Before SGA			After SGA		
	R@20	R@50	R@100	R@20	R@50	R@100
3	22.81	28.87	32.80	30.34	33.09	34.40
5	34.15	38.32	39.92	39.06	43.64	45.66
7	32.01	35.82	38.04	43.67	48.47	50.98
10	42.17	46.60	48.85	43.97	48.40	50.86

表 6.5 未标记集在攻击前后的 PredCls 结果对比

Table 6.5 The PredCls results of the unlabeled set before and after attacking

务为目标提供标注的真实标签和边界框。根据表6.5中的结果所示，我们可以得出以下结论：（1）攻击后 N-shot 设置的性能不断提高，这表明我们可以通过引入 SGA 为未标记数据生成更准确的伪标记；（2）在攻击之前，基于 7-shot 集合生成的伪标记的性能低于 5-shot，这表明更多标记示例的谓词类别可能不会带来更准确的伪标记，因为标记集和未标记集之间的数据分布存在一定的差距。

6.3.4　可视化分析

关系特征：在图6.4中，我们使用 t-SNE[86] 绘制了来自不同谓词类的关系样本的特征表示。图6.4(a) 可视化了攻击前场景图中关系节点的特征，而图6.4(b) 可视化了攻击后更新的关系特征。该模型是在 10-shot 设置下训练所得。SGA 学习扰动场景图中的目标节点并欺骗目标分类器以更改预测的类别标签。为了抵御该攻击，谓词分类器将学习去维持场景图的全局一致性和兼容性。我们观察到攻击后的关系特征比以前更具有判别性，例如 “carrying”，“sitting on” 和 “covering”，这表明我们的方法可以为少样本关系类别学习丰富而紧凑的关系表示的能力。此外，我们可以看到具有相似语义的关系特征点越来越近，例如 “covered in” 和 “covering”，说明我们的方法可以编码用于关系特征表示的高级语义信息。

攻击过程：在图6.5中，我们展示两个成功的场景图攻击案例，“cow” → “horse” 和 “hill” → “mountain”。可以看到，SGA 通过翻转邻接矩阵的边成功地欺骗了目标分类器并更改了目标标签。在考虑全局语义约束的情况下，关系检测模型通过连接谓词和整个场景图来学习防御攻击。SGA 产生的新目标标签不仅在语义上与连接的谓词相容，而且与场景图全局语义兼容。

攻击结果分析：在图6.6中我们对攻击前后的目标预测标签的变化进行分析

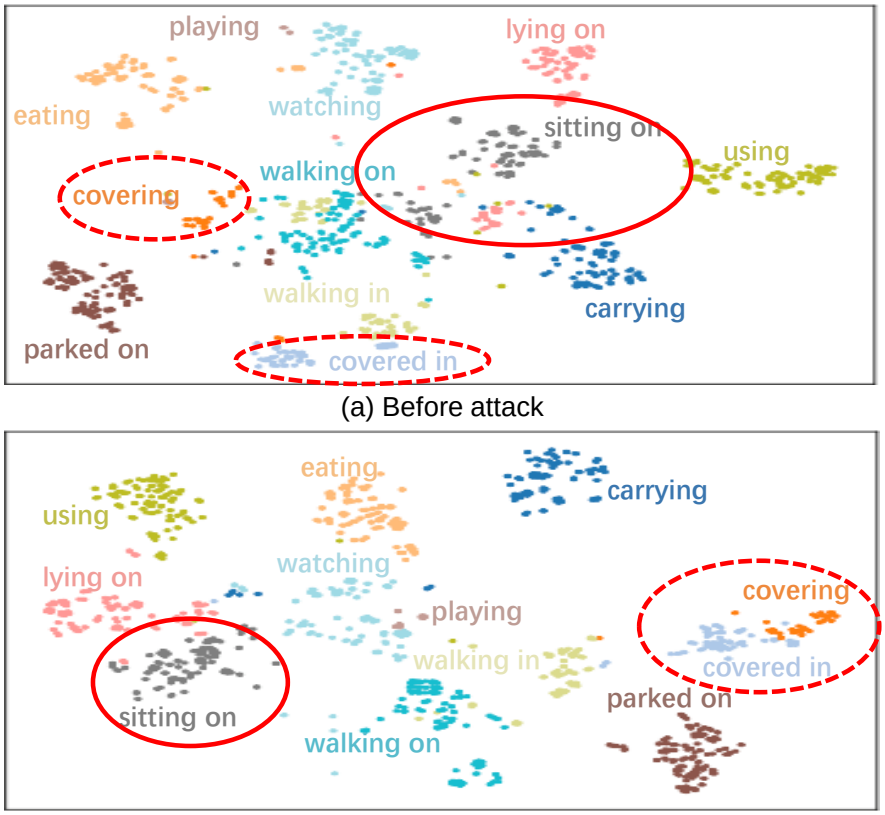


图 6.4 关系特征的 T-SNE 可视化效果

Figure 6.4 T-SNE visualization of relationship features

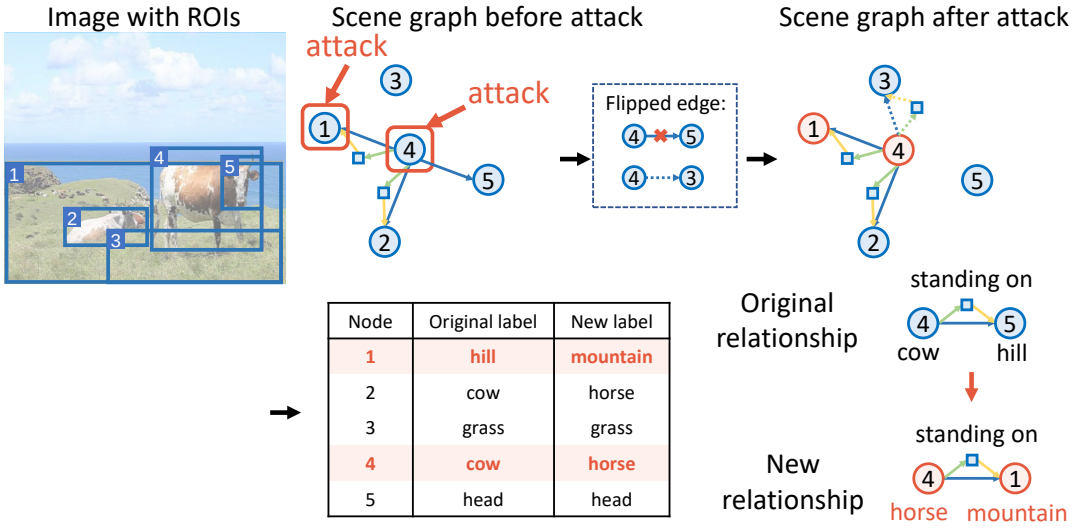


图 6.5 场景图攻击（SGA）示例图

Figure 6.5 Examples of Scene Graph Attack (SGA)

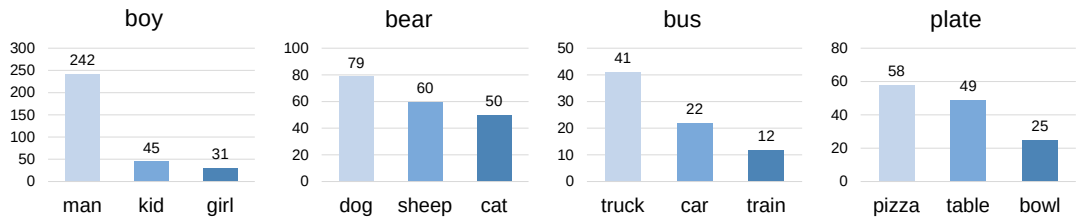


图 6.6 攻击前后目标标签变化统计

Figure 6.6 Statistics about the object labels before and after attacking

统计，图中展示了每个类攻击后经常产生的前三个新标签。可以看到，SGA 学习三种不同的攻击模式：（1）可根据词向量距离确定的同义词类别，如 “boy” → “man”；（2）属于同一个超类的相似子类，不仅考虑词向量距离，而且还考虑了全局语义防御，如 “bear” → “dog” 和 “bus” → “car”；（3）具有相同关系模式的类，这些关系模式主要由场景图的全局语义约束确定，如 “plate” → “pizza” 和 “kite” → “bird”。

6.4　本章小结

当不频繁关系类别的训练样本及其稀少时，基于少样本训练的视觉关系检测模型无法充分学习不同的类别特征，且鲁棒性较差。针对此问题，本章提出了一种场景图攻击（SGA）方法。该方法尝试将对抗攻击技术引入视觉关系检测模型，从未标记数据中挖掘更多示例，使得该模型在使用少量标记数据训练的情况下变得更加鲁棒。SGA 在场景图中的目标节点进行攻击时还关注其全局语义一致性，使得模型能够学习到更加紧凑的视觉关系特征。这些特征更加关注目标间的互动关系，而非具体的目标外观信息。SGA 可基于初始视觉关系来生成新的关系三元组，可用于辅助更新未标记数据的伪标签，从而为模型提供更加准确且多样化的训练数据。我们的方法为后续的少样本视觉关系检测相关研究提供了一个完善的框架，通过该框架我们可以引入大量未标记数据，并对其进行动态地在线标注，从而为关系检测模型提供充足的训练样本。

第 7 章　基于跨模态记忆的视觉语言导航

视觉场景理解包含对目标的识别与定位，以及对目标间语义关系的推理。较为基础的视觉任务只需要依赖视觉信号即可完成，而在真实世界中，当智能体面临更加复杂的任务时，常常需要和人类进行互动合作，并结合多模态的信息来完成任务。视觉对话导航任务要求智能体在真实环境中导航寻找目标物体，在导航过程中，允许智能体以自然语言方式向人类求助，获取指示性信息。具体地，为了实现导航目标（即找到目标），代理向用户询问问题（例如，我应该向左还是向右走?），用户回应（例如，留在卧室）后代理即采取操作，如图7.1所示。为了更好地理解当前对话的内容，Seq2seq[87] 方法将对话历史和当前对话进行连接并编码。对话历史为当前对话提供了丰富的上下文信息，可辅助解析当前对话中的指代性信息。

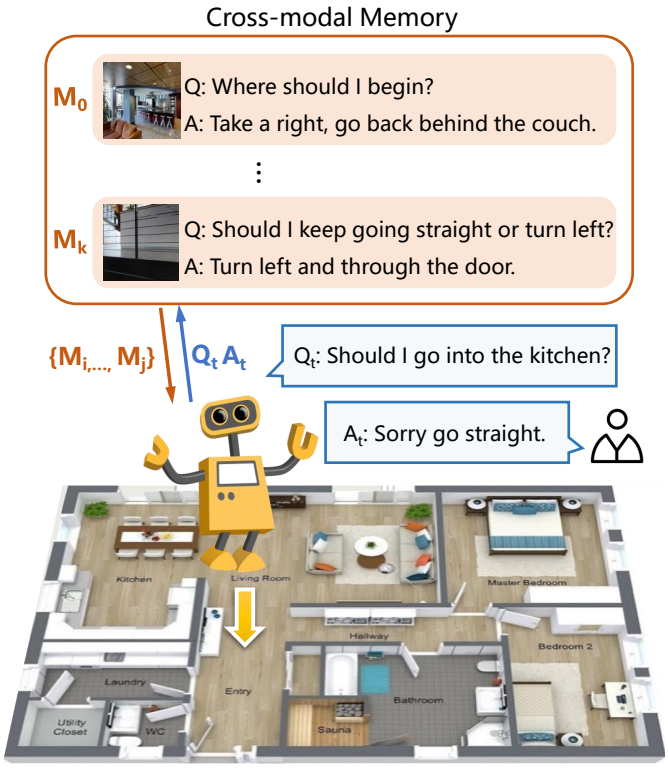


图 7.1 视觉对话导航任务示意图

Figure 7.1 Illustration of the Vision-Dialog Navigation task

7.1 模型概述

本章提出跨模态记忆网络 CMN（Cross-modal Memory Network）以探索代理在导航过程中所产生的语言和视觉记忆。CMN 包含两种记忆模块，第一个模块是语言记忆模块（L-mem），该模块基于代理和用户之间的对话历史来解析当前对话中的指令。L-mem 模块的目标是更好地理解用户提供的关于下一步动作的指令。第二个模块是视觉记忆模块（V-mem），该模块旨在为代理按时序方式编码导航过程中关于视觉场景的记忆。视觉记忆特征将根据 L-mem 生成的记忆感知特征进行融合，编码代理的导航历史。在 CMN 中，L-mem 和 V-mem 协同工作来探索关于历史导航动作决策每一步的记忆信息，并为当前时刻的动作预测提供了跨模态的上下文信息，从而得到更加准确的动作预测结果。

7.2 跨模态记忆网络

7.2.1 问题定义

根据 NDH 任务的定义，在开始对话导航前，用户首先提供一个不清晰的模糊的指令（包含要到达的目标）给代理。一个对话提示信息（S, t_o, p_0, G_j）包含一个房间场景的扫描 S，一个待寻找的目标 t_o，一个起始位置 p_0 以及目标区域 G_j。在每一轮对话过程中，代理提出问题 Q，用户给出回答 A，代理根据对话和当前场景预测下一步动作 N。每次视觉对话导航任务包含 k 轮重复的提问-回答-移动序列 $< N_0, Q_1, A_1, N_1, ..., Q_k, A_k, N_k >$。对于每个（$S, t_o, p_0, G_j$），一个视觉对话导航实例可以基于 $0 \leq i \leq k$ 生成。输入包含一个关于目标 t_o 的提示，一个第 t 轮的对话历史为 $H_t = \{D_1, ..., D_{t-1}\}$，其中 $D_i = (Q_i, A_i)$。根据上述问题定义，CMN 框架可看成是一个编码器-解码器体系结构：（1）编码器用于探索有关代理与用户之间的历史对话 H_t 的语言记忆，从而生成 D_t 的上下文表示；（2）解码器用于回顾导航的历史信息以帮助解析当前对话，然后通过跨模态记忆增强的特征预测导航动作 A_t。CMN 模型的框架图如图7.2所示。

7.2.2 特征表示

语言特征：我们首先使用预训练的 GloVe 词向量[68] 对当前对话 D_t 中的每个单词进行编码，得到$\{w_{t,1}, ..., w_{t,T}\}$，其中 T 代表 Q_t 和 A_t 中的单词数。接着使用两层的 LSTM 来编码生成隐状态序列 $\{h_{t,1}, ..., h_{t,T}\}$。将最后一个隐状态的

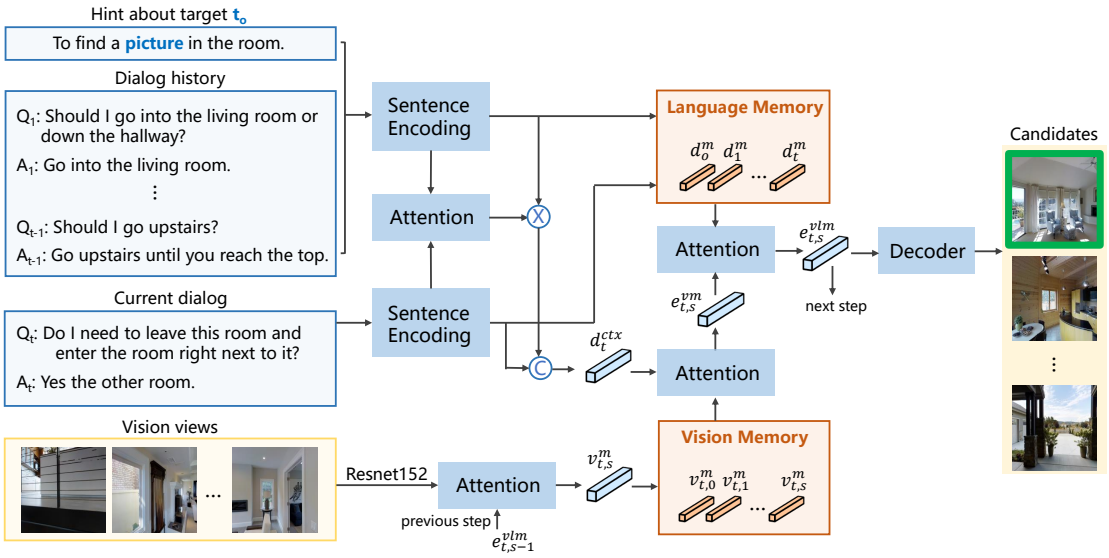


图 7.2 基于跨模态记忆的视觉对话导航模型框架

Figure 7.2 Overview of the Cross-modal Memory Network for Vision-Dialog Navigation

特征 $h_{t,T}$ 作为每个对话 D_t 的特征，$d_t \in \mathbb{R}^L$，其中 L 是对话句子最大长度。

$$\begin{aligned} \{h_{t,1}, ..., h_{t,N}\} &= \text{LSTM}(\{w_{t,1}, ..., w_{t,N}\}) \\ d_t &= h_{t,N} \end{aligned} \tag{7.1}$$

类似的，将对话历史 H_t 进行编码，得到 $\{d_i\}_{i=0}^{t-1} \in \mathbb{R}^{t \times L}$。

图片特征：对于每个视频帧的场景图片，我们使用全景特征表示，每个全景图被划分为 36 个不同的图片，对应不同的视角，每个视角对应的特征为 $V_{t,s} = \{v_{t,s,i}\}, v_{t,s,i} \in \mathbb{R}^{2048}$，其中 $v_{t,s,i}$ 代表对于视角 i 预提取的卷积神经网络特征。

7.2.3　视觉记忆

我们希望代理通过记忆先前对环境的视觉感知来做出当前的动作决策，V-mem 模块用于在导航过程中还原先前所见的视觉场景，以帮助生成当前视觉感知的记忆感知。首先，我们使用上一步 $s-1$ 的交叉模态记忆编码 $e_{t,s-1}^{vlm}$ 来逐步关注全景特征 $V_{t,s}$，生成的特征 $V_{t,s}^m$ 描述了上一步决策和当前视图之间的相关性。我们首先将 $e_{t,s-1}^{vlm}$ 和 $V_{t,s}$ 映射到 c 维并且计算注意力权值 A^{vis}：

$$\begin{aligned} X &= f_v(e_{t,s-1}^{vlm}) \odot f_{vlm}(v_{t,s,i}) \\ A^{vis}(e_{t,s-1}^{vlm}, v_{t,s,i}) &= \sigma(X)/\sqrt{c}, \end{aligned} \tag{7.2}$$

其中 $f_v(\cdot)$ 和 $f_{vlm}(\cdot)$ 表示两层的多层感知机，可将输入特征映射到 c 维。σ 代表 softmax 激活函数。$\odot$ 代表逐元素相乘操作。接着计算记忆感知的表达，其中包

含上一步的动作决策信息：

$$v_{t,s}^{mem} = \sum_{i=1}^{s} A_{s,i}^{vis} v_{t,s,i}. \tag{7.3}$$

通过上一步决策与当前视角的注意力权值，可以得到 V-mem 的输出 $v_{t,s}^{m} \in \mathbb{R}^{K}$。

7.2.4 语言记忆

给定当前的问题答案 D_t 和对话历史特征，L-mem 模块旨在提取历史上与当前轮次对话相关的历史中最相关的记忆。首先计算多头注意力，定义 d_t 和 $M_t = \{h_i\}_{i=0}^{t-1}$ 分别表示问题和对话历史的特征向量。使用两个可学习权值矩阵 $W_n^d \in \mathbb{R}^{L\times c}$ 和 $W_n^M \in \mathbb{R}^{L\times c}$ 将特征 d_t 和 M_t 分别映射到 $c = 512$ 维。接着计算 d_t 和对话记忆 M_t 之间的注意力 A_n^{lan}：

$$\begin{aligned} A_n^{lan}(d_t, h_i) &= \mathrm{softmax}((d_t W_n^Q)(h_i W_n^K)^T)/\sqrt{c}, \\ \hat{d}_t &= \mathrm{concat}_{n=1}^{N} \left\{ \sum_{i=0}^{t} A_n^{lan}(d_t, h_i) W_n^V h_i \right\}, \\ \hat{d}_t &= \mathrm{LayerNorm}(\hat{d}_t + d_t), \end{aligned} \tag{7.4}$$

其中每个注意力头的输出会被连接起来，当前对话 $\hat{d}_t$ 的上下文特征将被用作残差连接，随后是 LayerNorm 层：

$$\begin{aligned} \hat{d}_t &= \mathrm{LayerNorm}(f_{lan}(\hat{d}_t) + \hat{d}_t), \\ d_t^{ctx} &= \mathrm{concat}\{\hat{d}_t, d_t\}. \end{aligned} \tag{7.5}$$

通过连接上下文特征 $\hat{d}_t$ 和原始对话特征 d_t，可以得到记忆感知的对话特征表示，$d_t^{ctx} \in \mathbb{R}^{2L}$。L-mem 建立在多头注意力机制的基础上，可以堆叠为多层，以获得对话历史记录上下文的高级抽象。

7.2.5 跨模态记忆

在分别探讨视觉感知的记忆和语言与注意力模块的交互作用之后，我们进一步介绍跨模态注意力，以探讨语言与视觉记忆之间的语义相关性。首先通过使用最后一个轮对话的内存感知 d_t^{ctx} 表示来计算从语言到视觉的注意，计算其对视觉记忆 $V_{t,s}^{m}$ 的注意力为：

$$e_{t,s}^{vm} = \mathrm{Attention}(d_t^{ctx}, \{v_{t,0}^{m}, ..., v_{t,s}^{m}\}), \tag{7.6}$$

视觉记忆提供有关先前所见场景的补充信息，从而可以使代理更好地了解场景。接着计算视觉到语言的注意力，以生成最终的跨模态记忆编码 $e_{t,s}^{vlm}$：

$$e_{t,s}^{vlm} = \text{Attention}(e_{t,s}^{vm}, \{d_0^m, ..., d_t^m\}), \tag{7.7}$$

这里语言记忆被使用了两次。第一次是在公式7.5中，第二次是在公式7.7中。两种合并方式之间的区别在于三方面：（1）d_t^{ctx} 是最后一轮对话特征 d_t 和先前对话历史记录 $\boldsymbol{H}_t$ 的注意力加权特征的串联，因此其主要语义来自 d_t，因为 d_t^{ctx} 提供 d_t 的上下文信息。相比之下，跨模式记忆感知表示 $e_{t,s}^{vlm}$ 可以发现视觉记忆与所有现有对话之间的相关性。（2）计算 d_t^{ctx} 的目的是帮助代理更好地理解用户的回答，而 $e_{t,s}^{vlm}$ 的目的是学习先前视觉和语言之间的对应关系，捕获时序上的相关性以获得更好的视觉理解。（3）公式7.6中的语言到视觉注意和公式7.7中的视觉到语言注意构成了视觉和语言上下文之间的闭环推理路径，为行动预测提供了丰富的跨模式记忆信息。

7.2.6　动作解码器

通过在语言指令和视觉视图的配合下执行内存感知推理，导航代理能够更好地理解对话历史和先前场景之间在时序上的关联，从而为当前步骤 s 的动作预测提供丰富的上下文信息：

$$\begin{aligned} \hat{a}_{t,s} &= \sigma(f_m(e_{t,s}^{vlm})), \\ a_{t,s} &= \text{softmax}(f_a(\hat{a}_{t,s})), \end{aligned} \tag{7.8}$$

其中 $f_m(\cdot)$ 和 $f_a(\cdot)$ 是单层的线性变换器，用以将 $e_{t,s}^{vlm}$ 从 $\boldsymbol{K}+\boldsymbol{L}$ 维映射到 $\boldsymbol{K}$ 维，和将 $\hat{a}_{t,s}$ 从 $\boldsymbol{K}$ 维映射到 $\boldsymbol{M}$，$\boldsymbol{M}$ 是预测的动作的类别数目。我们使用基于全景图特征的全景动作空间[88]，代理需要在多个视角的图片特征中挑选下一步的候选。

7.3　实验结果及分析

7.3.1　实验设定

数据集：我们在 CVDN 数据集上评估模型，该数据集在 83 个 MatterPort 房间[89] 中收集了 2050 个人与人的导航对话和 7,000 条轨迹，每条轨迹对应于几个问答轮次，数据集包含 81 种目标，每次导航均以模糊的指令开头，随后代理与用户之间的问答交互将引导导航代理找到目标。

评价指标：我们使用四个流行的指标从不同方面评估模型：(1) 成功率 (SR)，最终停止位置距离目标位置少于 3m 的导航的百分比。(2) Oracle 成功率 (OSR)，即代理可以沿着其轨迹在距离目标最近的位置停止时的成功率。(3) 目标进度 (GP)，代理向目标位置的平均进度。(4) Oracle 路径成功率 (OPSR)，代理可以沿着最短路径在距目标最近的点处停止的成功率。请注意，如果最短路径未用于监督（即混合路径或代理路径），则此路径可能与 OSR 不同。

不同的监督信息： CVDN 数据集中的导航路径（Navigator Path）是从充当代理的人员那里收集的，而用户已知的路径（Oracle Path）是由最短路径规划生成的。导航任务中代理的监督信息由最短路径定义，但是，现实情况中即使是人类标注也可能是不完美的，因此，CVDN 数据集还提供了一种新的监管形式，称为混合监管路径（Mixed Path）。当导航路径和用户路径的末端节点相同时，将混合监督路径定义为导航路径，否则将其定义为用户路径。

7.3.2 数值结果和可视化示例

与已有方法对比：基线方法包括：(1) Shortest Path Agent 在推断时使用最短路径作为监督，是导航代理性能的上界；(2) Random Agent[89] 随机选择下一步的方向并且向前走五步；(3) Vision Only 基线方法中，导航代理只考虑视觉信息。(4) Dialog Only 基线方法中，导航代理只考虑语言信息。(5) Sequence to sequence 模型[87]，将对话历史联结成一条简单的指令。如表7.1所示，CMN 在不同监督路径下的 Goal Progress 性能指标优于已有方法，证明了我们方法的有效性。

Method	Val Seen			Val Unseen			Test Unseen		
	Oracle	Navigator	Mixed	Oracle	Navigator	Mixed	Oracle	Navigator	Mixed
Baseline (Shortest Path Agent)	8.29	7.63	9.52	8.36	7.99	9.58	8.06	8.48	9.76
Baseline (Random Agent)	0.42	0.42	0.42	1.09	1.09	1.09	0.83	0.83	0.83
Baseline (Vision Only)	4.12	5.58	5.72	0.85	1.38	1.15	0.99	1.56	1.74
Baseline (Dialog Only)	1.41	1.43	1.58	1.68	1.39	1.64	1.51	1.20	1.40
Sequence-to-sequence model[87]	4.48	5.67	5.92	1.23	1.98	2.10	1.25	2.11	2.35
CMN (Ours)	**5.47**	**6.14**	**7.05**	**2.68**	**2.28**	**2.97**	**2.69**	**2.26**	**2.95**

表 7.1 在 Goal Progress (m) 指标上的性能对比

Table 7.1 Comparison of the performance on Goal Progress (m)

消融研究：我们对 V-mem 和 L-mem 模块分别在校验集和测试集上进行消融研究，结果如表7.2和表7.3所示。第一行直接使用串联的对话历史作为语言输入

Method	Val Seen				Val Unseen			
	GP (m)	OSR (%)	SR (%)	OPSR (%)	GP (m)	OSR (%)	SR (%)	OPSR (%)
Seq-to-seq[87]	5.92	63.8	36.9	72.7	2.10	25.3	13.7	33.9
VLN Baseline[88]	6.15	58.9	33.0	69.4	2.30	35.5	19.7	45.9
CMN w/o V-mem	6.33	61.3	30.9	72.3	2.52	36.7	20.5	48.4
CMN w/o L-mem	6.47	58.6	31.9	68.6	2.64	39.1	20.5	50.4
CMN (Ours)	**7.05**	**65.2**	**38.5**	**76.4**	**2.97**	**40.0**	**22.8**	**51.7**

表 7.2 在校验集上的 L-mem 和 V-mem 模块的消融研究

Table 7.2 Ablation study about the L-mem and V-mem module on CVDN val. set

V-mem	L-mem	VL-mem	Goal Process (m)
✓	✓	✓	2.95
	✓	✓	2.74
✓		✓	2.04
✓	✓		2.54

表 7.3 在测试集上的 L-mem 和 V-mem 模块的消融研究

Table 7.3 Ablation study about the L-mem and V-mem module on CVDN test set

以构建 VLN 基线。我们通过直接平均每个全景图的视觉特征来禁用 V-mem 模块，从表7.2和表7.3中可以看出，当模型丢失关于先前导航的视觉记忆后，导航性能下降。为了删除 L-mem 模块，我们用最后一个问答句中的单词级上下文替换语言交互的记忆感知特征表示。从7.2和表7.3中可以看出，当禁用语言存储模块（L-mem）时，我们方法的性能急剧下降，这表明语言记忆对于理解用户指令以获得更好的导航目标至关重要。在表7.3中，我们还将编码器的输出 $e^{vlm}_{t,s-1}$ 设为 0 以消除跨模态记忆上下文（VL-mem），VL-mem 可以看作是历史导航信息的高级抽象，它表示有关代理先前做出的动作决定的丰富信息。结果表明，当丢失交叉模式记忆时，我们模型的性能有所下降。

7.4　本章小结

本章探索如何将视觉场景理解与多模态信息进行有效融合以辅助更加复杂的视觉任务。在视觉对话导航的任务设定下，我们提出跨模式内存网络（CMN），该网络通过按时序编码代理的跨模态记忆来辅助视觉对话导航任务。语言记忆

可以帮助代理基于对话历史更好地理解来自用户的指令。视觉记忆关注历史导航路径上的场景信息，并与语言信息进行互相关联。在 CMN 中，视觉和语言记忆模块与网络协同训练，辅助代理预测更加准确的导航动作，成功地接近最终的导航目标。实验证明，我们的方法在视觉对话导航任务上相比于 Seq2seq 模型取得了显著的提升。当代理处于未知的环境中时，依然能够通过自然语言对话获取用户指令信息，准确地导航至目标地点。

第 8 章　总结与展望

8.1　本文工作总结

本文研究有限监督条件下场景中的目标识别与定位问题，以及目标间的语义关系检测问题。本文针对现实世界场景中标注数据数量或质量有限的情况，提出了系统的方法和算法框架，总结如下：

（1）提出了一种目标概率预选模块，用于弱监督目标定位任务。该模块通过捕捉目标区域内的语义和空间关联计算目标置信度，引导网络通过学习来激活更加完整的目标区域。在 PASCAL VOC、COCO 和 ImageNet 多个数据集上的测试结果表明，所提出模型在弱监督目标定位方面的性能优于现有方法。

（2）提出了一种实例激活图技术，用于弱监督实例分割任务。该技术基于实例响应和目标预选掩膜来学习目标一致性，将稀疏且局部的实例响应填充为实例分割掩膜，从而得到更加准确的预测结果。实验表明，该技术的实例分割性能优于已有方法，且大大缩短了推断时间。此外，实验验证了该技术所学的类无关填充权值可以直接应用到其他数据集上，无需进行训练和微调。

（3）提出了一种渐近知识驱动变换器，用于视觉关系检测。该变换器通过渐近地使用外部知识引导模型来学习目标区域之间的语义关联，约束冗余的连接，使得目标区域特征更多地被关联区域更新。在多个数据集上的实验表明，该变换器能够使得视觉关系检测模型对不常见关系类别预测更加准确的标签。

（4）提出了一种可配置图推理技术，该技术将现有方法中固定的关系推理路径进行分解，学习从海量的知识库中挑选有用的知识先验用于辅助模型预测。实验表明，该模型在视觉关系预测任务上达到 state-of-the-art 性能，且可以和来自不同域的知识很好地兼容。

（5）提出了一种场景图攻击技术，以学习更加鲁棒的视觉目标检测模型。该技术在只有极少量的标注样本时，能够从大量的未知数据中挖掘训练样本，从而提升视觉关系检测模型的预测精度。

（6）提出了一种跨模态记忆网络，用于视觉语言导航任务。该网络模型从时序角度将视觉场景信息和语言信息进行编码，并进行跨模态的信息关联和匹配。实验结果显示，相比于 Seq2seq 方法，该模型能够辅助代理做出更准确的导航决

策，并且可以更好地泛化到未知的场景中。

8.2 未来工作展望

有限监督条件下的目标定位与关系推理，是真实世界场景理解的重要研究内容。然而因为监督信息有限导致的一系列问题，如标注噪声、模型训练不充分、以及方法泛化性较差等，一定程度上限制了模型方法在实际任务中的应用。未来可以研究如何有效地利用已有先验知识或结合自监督思想挖掘样本间关系来解决现实世界场景识别面临的问题，后续的研究和改进方向包括：

（1）域迁移学习：本文已研究以常识知识图的形式引入先验辅助模型学习，我们可以进一步考虑结合域迁移学习的实现，小数据集上训练的模型，在实际应用时，再进行简单的微调与适配。该算法可以适用于开集关系识别问题，为关系识别模型在真实场景的部署有着非常重要的意义。

（2）图结构上的对比学习：自监督对比学习与推理任务的结合时当前计算机视觉研究的一个重要分支方向。基于本文中关于图传播模型的探索，后续可以针对图结构中的节点和边之间的关联关系设计对比学习，即根据不同传播法则或推理路径得到不同的场景图。该算法可从大量无标记样本中挖掘样本信息，有这重要的研究意义。

参考文献

[1] Uijlings J R, Sande K E, Gevers T, et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013, 104(2):154-171.

[2] Zitnick C L, Dollár P. Edge boxes: Locating object proposals from edges[C]//European Conference on Computer Vision. 2014: 391-405.

[3] Pont-Tuset J, Arbelaez P, T.Barron J, et al. Multiscale combinatorial grouping for image segmentation and object proposal generation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(1):128-140.

[4] Bilen H, Pedersoli M, Tuytelaars T. Weakly supervised object detection with posterior regularization[J]. Proceedings BMVC 2014, 2014:1-12.

[5] Cinbis R G, Verbeek J, Schmid C. Weakly supervised object localization with multi-fold multiple instance learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(1):189-203.

[6] Bilen H, Vedaldi A. Weakly supervised deep detection networks[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2846-2854.

[7] Kantorov V, Oquab M, Cho M, et al. Contextlocnet: Context-aware deep network models for weakly supervised localization[C]//European Conference on Computer Vision (ECCV). 2016: 350-365.

[8] Sun C, Paluri M, Collobert R, et al. Pronet: Learning to propose object-specific boxes for cascaded neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 3485-3493.

[9] Oquab M, Bottou L, Laptev I, et al. Is object localization for free? - weakly-supervised learning with convolutional neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 685-694.

[10] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2921-2929.

[11] Khoreva A, Benenson R, Hosang J, et al. Simple does it: Weakly supervised instance and semantic segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 1665-1674.

[12] Zhou Y, Zhu Y, Ye Q, et al. Weakly supervised instance segmentation using class peak response [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 3791-3800.

[13] Yang J, Lu J, Lee S, et al. Graph r-cnn for scene graph generation.[J]. European Conference on Computer Vision (ECCV), 2018:690-706.

[14] Li Y, Ouyang W, Zhou B, et al. Factorizable net: An efficient subgraph-based framework for scene graph generation[J]. European Conference on Computer Vision (ECCV), 2018:346-363.

[15] Tang K, Zhang H, Wu B, et al. Learning to compose dynamic tree structures for visual contexts [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 6619-6628.

[16] Chen L, Zhang H, Xiao J, et al. Counterfactual critic multi-agent training for scene graph generation[C]//IEEE International Conference on Computer Vision (ICCV). 2019: 4613-4623.

[17] Zellers R, Yatskar M, Thomson S, et al. Neural motifs: Scene graph parsing with global context[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 5831-5840.

[18] Chen T, Yu W, Chen R, et al. Knowledge-embedded routing network for scene graph generation.[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6163-6171.

[19] Yu R, Li A, Morariu V I, et al. Visual relationship detection with internal and external linguistic knowledge distillation[C]//IEEE International Conference on Computer Vision (ICCV). 2017: 1068-1076.

[20] Gu J, Zhao H, Lin Z, et al. Scene graph generation with external knowledge and image reconstruction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1969-1978.

[21] Liu H, Singh P. Conceptnet —a practical commonsense reasoning tool-kit[J]. Bt Technology Journal, 2004, 22(4):211-226.

[22] Zhou B, Khosla A, Lapedriza A, et al. Object detectors emerge in deep scene cnns[C]// International Conference on Learning Representations (ICLR). 2015.

[23] Newman M E. The mathematics of networks[J]. The new palgrave encyclopedia of economics, 2008, 2(2008):1-12.

[24] REN S, HE K, GIRSHICK R B, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Conference on Neural Information Processing Systems (NeurIPS). 2015: 91-99.

[25] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: Delving deep into convolutional nets[C]//British Machine Vision Conference 2014. 2014.

[26] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C]//International Conference on Learning Representations (ICLR). 2015.

[27] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 1-9.

[28] Zhang J, Lin Z L, Brandt J, et al. Top-down neural attention by excitation backprop[C]// European Conference on Computer Vision (ECCV). 2016: 543-559.

[29] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps[C]//International Conference on Learning Representations (ICLR Workshop). 2013.

[30] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//European Conference on Computer Vision (ECCV). 2014: 818-833.

[31] Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.[J]. PLOS ONE, 2015, 10(7):130140.

[32] Bency A J, Kwon H, Lee H, et al. Weakly supervised localization using deep feature maps [C]//European Conference on Computer Vision (ECCV): volume 9905. 2016: 714-731.

[33] Lin T Y, Dollar P, Girshick R, et al. Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 936-944.

[34] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 770-778.

[35] Maninis K K, Pont-Tuset J, Arbelaez P, et al. Convolutional oriented boundaries: From image segmentation to high-level tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4):819-833.

[36] Teichmann M T T, Cipolla R. Convolutional crfs for semantic segmentation[C]//British Machine Vision Conference (BMVC). 2018: 142.

[37] Burtsev S V, Kuzmin Y P. An efficient flood-filling algorithm[J]. Computers & Graphics, 1993, 17(5):549-561.

[38] Everingham M, Eslami S M, Gool L, et al. The pascal visual object classes challenge: A retrospective[J]. International Journal of Computer Vision, 2015, 111(1):98-136.

[39] Hariharan B, Arbelaez P, Bourdev L, et al. Semantic contours from inverse detectors[C]//IEEE International Conference on Computer Vision (ICCV). 2011: 991-998.

[40] Pont-Tuset J, Gool L V. Boosting object proposals: From pascal to coco[C]//IEEE International Conference on Computer Vision (ICCV). 2015: 1546-1554.

[41] Zhu Y, Zhou Y, Ye Q, et al. Soft proposal networks for weakly supervised object localization [C]//IEEE International Conference on Computer Vision (ICCV). 2017: 1859-1868.

[42] Wan F, Wei P, Jiao J, et al. Min-entropy latent model for weakly supervised object detection [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 1297-1306.

[43] Wah C, Branson S, Welinder P, et al. The caltech-ucsd birds-200-2011 dataset[J]. California Institute of Technology, 2011.

[44] Zhang X, Wei Y, Feng J, et al. Adversarial complementary learning for weakly supervised object localization[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 1325-1334.

[45] Zhang X, Wei Y, Kang G, et al. Self-produced guidance for weakly-supervised object localization[C]//European Conference on Computer Vision (ECCV). 2018: 610-625.

[46] Cheng M M, Mitra N J, Huang X, et al. Salientshape: Group saliency in image collections[J]. The Visual Computer, 2014, 30(4):443-453.

[47] Liu T, Yuan Z, Sun J, et al. Learning to detect a salient object[J]. IEEE Transactions on Pattern Analysis and Machine intelligence, 2011, 33(2):353-367.

[48] YAN Q, XU L, SHI J, et al. Hierarchical saliency detection[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013: 1155-1162.

[49] Hou Q, Cheng M M, Hu X, et al. Deeply supervised salient object detection with short connections[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(4): 815-828.

[50] Luo Z, Mishra A, Achkar A, et al. Non-local deep features for salient object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 6593-6601.

[51] Li G, Yu Y. Deep contrast learning for salient object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 478-487.

[52] Zhang D, Han J, Zhang Y. Supervision by fusion: Towards unsupervised learning of deep salient object detector[C]//IEEE International Conference on Computer Vision (ICCV): volume 1. 2017: 3.

[53] Zhang J, Zhang T, Dai Y, et al. Deep unsupervised saliency detection: A multiple noisy labeling perspective[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 9029-9038.

[54] Jiang H, Wang J, Yuan Z, et al. Salient object detection: A discriminative regional feature integration approach[C]//IEEE conference on computer vision and pattern recognition (CVPR). 2013: 2083-2090.

[55] Zhu W, Liang S, Wei Y, et al. Saliency optimization from robust background detection[C]// IEEE conference on computer vision and pattern recognition (CVPR). 2014: 2814-2821.

[56] Li X, Lu H, Zhang L, et al. Saliency detection via dense and sparse reconstruction[C]//IEEE International Conference on Computer Vision (ICCV). 2013: 2976-2983.

[57] Hudson D, Manning C. Learning by abstraction: The neural state machine for visual reasoning [C]//Thirty-third Conference on Neural Information Processing Systems. 2019.

[58] Zhang H, Kyaw Z, Chang S F, et al. Visual translation embedding network for visual relation detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3107-3115.

[59] Lu C, Krishna R, Bernstein M S, et al. Visual relationship detection with language priors[J]. European Conference on Computer Vision (ECCV), 2016:852-869.

[60] Bengio Y, Louradour J, Collobert R, et al. Curriculum learning[C]//Proceedings of the Annual International Conference on Machine Learning. 2009: 41-48.

[61] Krishna R, Zhu Y, Groth O, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision (IJCV), 2017, 123(1):32-73.

[62] Li Y, Ouyang W, Zhou B, et al. Scene graph generation from objects, phrases and region captions[C]//IEEE International Conference on Computer Vision (ICCV). 2017: 1270-1279.

[63] Dai B, Zhang Y, Lin D. Detecting visual relationships with deep relational networks[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 3298-3308.

[64] Xu D, Zhu Y, Choy C B, et al. Scene graph generation by iterative message passing[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 3097-3106.

[65] Newell A, Deng J. Pixels to graphs by associative embedding[J]. Conference on Neural Information Processing Systems (NeurIPS), 2017:2171-2180.

[66] Li Y, Ouyang W, Wang X, et al. Vip-cnn: Visual phrase guided convolutional neural network [C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 7244-7253.

[67] Zhang J, Shih K J, Elgammal A, et al. Graphical contrastive losses for scene graph parsing [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[68] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]// Conference on Empirical Methods in Natural Language Processing. 2014.

[69] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks[J]. International Conference on Learning Representations (ICLR), 2017.

[70] Gumbel E J. Statistical theory of extreme values and some practical applications: A series of lectures[J]. 1954.

[71] Jang E, Gu S, Poole B. Categorical reparameterization with gumbel-softmax[J]. arXiv preprint:1611.01144, 2016.

[72] Maddison C J, Mnih A, Teh Y W. The concrete distribution: A continuous relaxation of discrete random variables[C]//International Conference on Learning Representations. 2017.

[73] Niu Y, Zhang H, Zhang M, et al. Recursive visual attention in visual dialog[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[74] Xie S, Girshick R, Dollar P, et al. Aggregated residual transformations for deep neural networks [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[75] Liang X, Lee L, Xing E P. Deep variation-structured reinforcement learning for visual relationship and attribute detection[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 4408-4417.

[76] Dornadula A, Narcomey A, Krishna R, et al. Visual relationships as functions:enabling few-shot scene graph prediction[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019: 1730-1739.

[77] Krishna R, Chen V, Varma P, et al. Scene graph prediction with limited labels[C]//IEEE International Conference on Computer Vision (ICCV). 2019: 2580-2590.

[78] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6):1137-1149.

[79] Zang X, Xie Y, Chen J, et al. Graph universal adversarial attacks: A few bad actors ruin graph learning models[J]. arXiv preprint arXiv:2002.04784, 2020.

[80] Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 2574-2582.

[81] Donahue J, Jia Y, Vinyals O, et al. Decaf: A deep convolutional activation feature for generic visual recognition[C]//International Conference on Machine Learning (ICML). 2014: 647-655.

[82] Yosinski J, Clune J, Bengio Y, et al. How transferable are features in deep neural networks [C]//Conference on Neural Information Processing Systems (NeurIPS). 2014: 3320-3328.

[83] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1):81-106.

[84] Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation[J]. Center for Automated Learning and Discovery, CMU: Carnegie Mellon University, USA., 2002.

[85] Hudson D A, Manning C D. Gqa: A new dataset for real-world visual reasoning and compositional question answering[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 6700-6709.

[86] VAN DER Maaten L, Hinton G. Visualizing data using t-sne[J]. Journal of Machine Learning Research, 2008, 9:2579-2605.

[87] Thomason J, Murray M, Cakmak M, et al. Vision-and-dialog navigation[J]. arXiv preprint arXiv:1907.04957, 2019.

[88] Fried D, Hu R, Cirik V, et al. Speaker-follower models for vision-and-language navigation [C]//Conference on Neural Information Processing Systems (NeurIPS). 2018: 3314-3325.

[89] Anderson P, Wu Q, Teney D, et al. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 3674-3683.

致　谢

五年多的硕博连读生涯即将结束，仔细回首，曾经的那些面对现实选择的纠结，实验论文结果不如人意的焦虑，奋力争取后依然不得的不甘与无奈，都已经消散，此刻只留下满满的开心和感激。很幸运当初选择攻读计算机专业的学士、博士学位，学习与研究最最前沿的知识和问题。很幸运在学生时代见证互联网的蓬勃发展、深度学习技术带来的变革、人工智能的巨大浪潮，以及体验科技进步对人类生活的颠覆性改变。这一路的收获和成长太多，严格的科研训练对思维与心智的磨练，完成博士学业巨大的成就感和自信心，来自亲朋好友老师同学善意的帮助和鼓励，都将成为我未来人生一笔宝贵的精神财富，支撑我在任何困难的时候都不放弃积极努力。

感谢我的导师焦建彬教授一直以来对我的指导和支持。焦老师是我们强大的后盾，为大家提供了宽松的科研环境和良好的科研指导，让我们可以自由地探索前沿课题，研究自己真正感兴趣的问题，并有机会在领域顶级会议和刊物上发表成果，与世界一流学者探讨前沿问题。焦老师对我们的工作一向是高标准严要求，引导我们逐渐形成把事情做到极致的态度和习惯。焦老师还常常关心学生生活，为我们解决生活中的烦恼，让我们能够全身心投入科研。我这样不是特别让人省心的学生，常常有一些奇奇怪怪的想法，焦老师总是有极大的包容和耐心给出建议，从未否定和苛责。

感谢实验室的叶齐祥教授。我博士期间的每一个工作，都离不开叶老师细心的指导。叶老师对待科研工作十分严谨认真，这种踏实细致的作风对我影响颇深。叶老师常常关心我们的身心健康，每当天气好或者节假日时，总是让我们出去多玩玩不要闷在实验室里。临近毕业的这段时间，叶老师也为我的学业规划和去向选择提供了非常多宝贵的意见和建议。感谢韩振军副教授和秦飞副教授对我的关怀和指导，尤其是在我刚开始科研非常迷茫的时候，两位老师的指导和建议让我受益匪浅。

感谢我本科母校中山大学的林倞教授和梁小丹副教授对我的支持和帮助，两位老师认真严谨、精益求精的工作作风对我影响颇深。在我的科研工作面临瓶颈十分迷茫时，林老师热心地提供实习访问的机会，大力支持我们的研究工作。梁

老师是我在实习访问期间的直接指导老师，为我的科研工作和生活提供了非常多的帮助和建议。梁老师有着十分开阔的学术视野，对前沿课题有着非常敏锐的嗅觉，在她的细心指导下，我进一步了解和阅读了非常多其他领域的前沿工作，扩宽了学术视野和思路，直到我回到实验室后梁老师还一直无私地帮助和指导我的研究工作。

感谢实验室的师兄师姐师弟师妹们，感谢你们陪我走过这段人生旅程，大家平时一起学习生活互相帮助，夏天一起打球冬天一起滑雪，整个楼层恐怕就我们屋每天笑声最大，这些对我而言都是非常美好的回忆。感谢中山大学的林冰倩、梁曦文和翁跃同学，还有莫纳什大学的朱峰达同学，那些一起拼命调实验赶论文的日子，现在回想起来依然热血又中二。感谢从高中开始友谊一直延续到现在的几位小伙伴，感谢大学毕业后一起来北京的几位同学，感谢他们在我焦虑失眠的时候鼓励我开导我，在我感到孤单的时候带我回家用美食治愈我。

感谢我的家人，感谢他们一直以来的呵护和陪伴、支持和鼓励，让我能够有信心和勇气克服困境、坚定向前。感谢我的外公外婆，学前启蒙教我识字背诗，总说我以后要当博士，在我幼小的心中埋下了这颗种子，冥冥之中影响着我的每个选择。感谢我的父母，直到我独自离家求学这些年经历一些人和事，才渐渐意识到在家的时候被父母保护的太好了，能够无忧无虑的长大，以自己的意愿选择工作和生活。感谢我的弟弟，这位优秀的同学常常以超出年龄的成熟包容安抚我的心情，在我焦头烂额忙于自己的事情而忽略家人时，他即使学业同样繁忙也会更多地陪伴父母。

感谢参加开题、中期和毕业答辩的各位指导老师和专家，他们丰富的经验和细致的指导和对论文方向及研究进度的指点给我的研究工作带来了巨大的帮助。

最后，再次向所有关心、帮助、支持、鼓励过我的老师、同学、家人、朋友表示最诚挚的谢意！

朱　艺

2020 年 11 月

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历:

2009 年 09 月–2013 年 07 月 中山大学 工学学士

2015 年 09 月–2020 年 12 月 中国科学院大学 工学博士

已发表 (或正式接受) 的学术论文:

一作论文:

[1] **Yi Zhu**, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, Jianbin Jiao. Soft Proposal Networks for Weakly Supervised Object Localization[C]. *In: Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017: 1841-1850. (EI).

[2] **Yi Zhu**, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, Jianbin Jiao. Learning Instance Activation Maps for Weakly Supervised Instance Segmentation[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019: 3116-3125. (EI).

[3] **Yi Zhu**, Fengda Zhu, Zhaohuan Zhan, Bingqian Lin, Jianbin Jiao, Xiaojun Chang, Xiaodan Liang. Vision Dialogue Navigation by Exploring Cross-modal Memory[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020: 10730-10739. (EI).

[4] **Yi Zhu**, Xiwen Liang, Bingqian Lin, Qixiang Ye, Jianbin Jiao, Liang Lin, Xiaodan Liang. Configurable Graph Reasoning for Visual Relationship Detection[J]. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2020. (SCI).

合作论文:

[1] Yanzhao Zhou, **Yi Zhu**, Qixiang Ye, Qiang Qiu, Jianbin Jiao. Weakly Supervised Instance Segmentation using Class Peak Response[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR, Spotlight)*, 2018: 3791-3800. (EI).

[2] Yao Ding, Yanzhao Zhou, **Yi Zhu**, Qixiang Ye, Jianbin Jiao. Selective Sparse

Sampling for Fine-Grained Image Recognition[C]. *In: Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2019: 6599-6608. (EI).

[3] Fengda Zhu, **Yi Zhu**, Xiaojun Chang, Xiaodan Liang. Vision-Language Navigation with Self-Supervised Auxiliary Reasoning Tasks[C]. *In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR, Oral)*, 2020: 10012-10022. (EI).

[4] Pengxu Wei, Fei Qin, Fang Wan, **Yi Zhu**, Jianbin Jiao, Qixiang Ye. Correlated Topic Vector for Scene Classification[J]. *IEEE Transactions on Image Processing (TIP)*, vol 26, 2017. 3221-3234 (SCI).

在审论文:

[1] **Yi Zhu**, Xiaodan Liang, Qixiang Ye, Jianbin Jiao. Progressive Knowledge-driven Transformer for Visual Relationship Detection[J]. *Pattern Recognition, Elsevier*, 2020. (SCI).

[2] **Yi Zhu**, Xiwen Liang, Xiaodan Liang, Qixiang Ye, Jianbin Jiao. Scene Graph Attack for Few-shot Relationship Detection[C]. *In: Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, 2021. (EI).

授权国家发明专利:

[1] 焦建彬，朱艺，叶齐祥，韩振军，张如飞，基于弱监督和深度响应重分配的 X 光图片违禁品定位方法，201811582841.2，中国发明专利，2020 年 6 月授权。

参加的研究项目及获奖情况:

研究项目:

基于 X 光图片的危险品目标检测，北京市科学基金委，No.61671427。

获奖情况:

- 中国科学院院长奖，2020 年
- 硕士国家奖学金，2017 年
- 航天星图杯高分软件大赛汽车检测第一名、飞机检测第二名，2017 年
- 中国科学院大学，三好学生，2017 年