•计算机科学与技术•

DOI:10.12454/j.jsuese.202500043

本刊网刊

基于Transformer架构的端到端粗–精网络场景图生成方法

李俊良[1],吕诗融[2],李　炜[1*]

(1.四川大学 空天科学与工程学院,四川 成都 610065;2.西南大学 计算机与信息科学学院,重庆 400715)

摘　要:场景图生成是计算机视觉领域的一个重要任务,旨在对视觉场景有一个全面且深度的理解,着重于识别实体以及实体之间存在的关系,最终要求模型输出一系列三元组<主体,谓词,客体>和一个图结构的场景表示。这对模型的理解能力和推理能力提出了更高的要求。现有的场景图生成方法在现阶段取得了巨大的成功,然而大多数模型存在参数量太大或对谓词(即实体之间的关系)的判断不准确的问题。为了能够解决这些问题,本文提出了一个端到端的粗–精网络(rough-and-refine model, RRM)场景图生成方法,整个模型包括粗网络和精网络两部分。一方面,粗网络负责提取场景中的各种基本信息,包括图像特征、实体特征和谓词特征。该部分利用Transformer架构,通过卷积神经网络和编码器的结合进行特征提取,获取图像特征。紧接着设计了实体解码器和谓词解码器,分别计算得到实体特征和谓词特征。另一方面,精网络利用粗网络中的场景信息,做进一步的信息交互,使模型获得更精确的预测能力。首先,用三元组查询生成模块获得主体查询、客体查询以及谓词查询,然后,分三条路径分别计算三元组的三个不同部分的表示。在不同路径中,设计了特征信息聚合模块、实体信息聚合模块和谓词信息聚合模块,加强模型对场景的理解,并且用路径之间的连接使模型在谓词的表示中更多地考虑主体和客体的信息。本文提出的模型在公开数据集Visual Genome上取得了优秀的成绩,在R@20、R@50、R@100评价指标上达到23.8、29.1、32.5;在mR@20、mR@50、mR@100指标上达到7.7、11.0、12.4。实验数据和可视化结果充分展现了本文提出的模型对场景的理解能力。

关键词:场景图生成;计算机视觉;人工智能;视觉关系检测

中图分类号:TP393.41　　**文献标志码:**A　　**文章编号:**2096-3246(2025)05-0344-11

场景理解旨在从图像或视频中提取并解释场景中的语义信息。它不仅包括对场景中物体的识别,还涉及对象间关系的推断、场景全局特征的分析,以及环境变化的动态理解。作为计算机视觉的关键研究方向之一,场景理解广泛应用于自动驾驶、智能监控、人机交互和机器人导航等领域,对于提升系统的环境感知能力和决策效率具有重要意义。场景理解常通过生成场景图的方式来表示。场景图是一种结构化的表示,以场景中的对象为节点,对象之间的关系为边,形成一个图(graph),并通过三元组的形式<主体,谓词,客体>构建出场景中的语义信息。这种表示方法不仅能够捕捉对象的类别和属性,还能反映它们之间的空间关系、功能关联以及相互作用。通过场景图的生成和分析,计算机视觉系统能够更深入地理解场景的语义层次,实现复杂场景中的推理与决策能力。

场景图生成任务的关键在于准确预测目标之间的关系。为此,许多研究工作采用两阶段[1–2]的检测流程:第一阶段利用目标检测器识别场景中的目标,提取实体信息(如类别、位置等);第二阶段基于这些实体信息和视觉上下文信息,对目标对进行关系预测。在这个过程中,研究人员引入多种方法来提升模型性能。例如,通过添加先验知识(包括统计先验[3–4]、语言先验[5–6]、常识知识[7–8])引入先验信息来增强模型对目标关系的理解能力,并修正某些错误预测。此外,信息传递模块[9]和注意力机制模块[10–11]也在两阶段场景图生成任务中发挥了重要作用。信息传递模块分为局部信息传递[12–13]和全局信息传递[14–15],分别作用于每个三元组和整个场景元素。随着注意力机制在人工智能

收稿日期:2025 – 01 – 20　　**修回日期:**2025 – 04 – 16　　**网络出版日期:**2025 – 06 – 03

基金项目:四川省自然基金项目(2025ZNSFSC0476)

作者简介:李俊良(2000—),男,硕士.研究方向:场景图生成;目标检测. E-mail: 641398265@qq.com

* **通信作者:**李　炜,副教授,E-mail: li.wei@scu.edu.cn

领域的广泛应用，越来越多的研究[16–17]将其引入到场景图生成任务中。注意力机制能够有效提炼特征，增强特征融合能力，帮助模型更准确地捕捉目标间的关系。

两阶段场景图生成方法尽管取得了显著进展，但仍存在一些固有的缺陷。首先，三元组的预测高度依赖于第一阶段目标检测的精度，一旦目标检测存在偏差，将直接影响后续关系预测的准确性。其次，模型需要预测每个目标对的关系，其算法复杂度较高，而大多数目标对并不值得过分关注，因此产生了大量不必要的计算开销。针对这些问题，研究人员从单阶段目标检测模型中获得启发，提出了一系列单阶段场景图生成方法。这些方法借鉴了目标检测模型DETR的结构，采用卷积神经网络CNN（convolutional neural network）与Transformer的结合策略来高效检测三元组。Cong等[18]提出了RelTR，将场景图生成任务视为一个集合预测问题，设计了实体解码器和三元组解码器，用于分别获取三元组中各元素的表示。Li等[19]基于Transformer架构提出了端到端场景图生成方法SGTR，通过实体节点生成器和谓词节点生成器分别生成若干实体和谓词，并利用一个有向二分图模块构建场景图。此外，他们在SGTR的基础上，设计了谓词节点生成器和图组装器[20]，进一步优化关系预测和场景图的构建。Khandelwal等[21]设计了编码器和一系列解码器，直接解码出主体、客体和谓词，以简化生成流程并提升效率。

场景图生成任务中，关系预测的核心在于明确主体与客体之间的角色关系，这是确保语义表达准确性的关键。然而，现有模型在主客体角色区分上常常表现出较大的模糊性。这种不足直接影响了关系预测的精准度，使得模型难以捕捉场景中真实的语义逻辑。此外，主客体角色的混淆还可能导致生成错误的三元组，破坏场景图的结构化表达，从而削弱对场景中复杂语义关系的整体理解。为此，本文提出了一种全新的粗—精网络场景图生成方法（图1），充分挖掘上下文语义信息，强化跨实体与关系的深层语义交互，明确主客体角色分工，提升模型对复杂场景的适应能力。

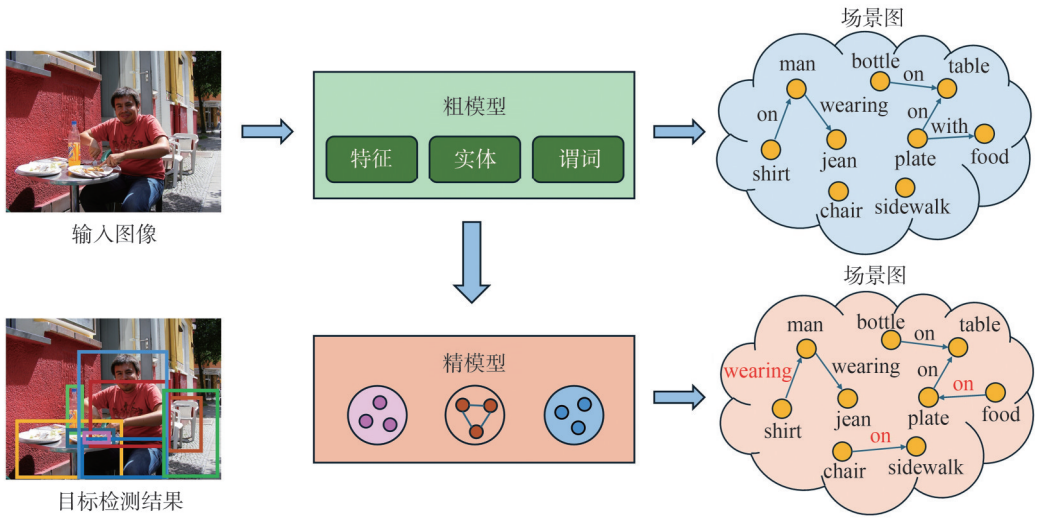

图1　粗–精模型整体框架

Fig. 1　Rough-and-refine framework

本文的主要贡献包括：

1）提出了一种新的场景图生成网络结构，巧妙利用了粗网络的初步预测能力和精网络的优化提升能力，并通过实验验证了这两部分在提升任务性能方面的重要作用。

2）通过设计信息交互路径，使模型能够更深入地理解场景中各部分之间的交互关系，从而生成更为准确的三元组，显著提升了关系预测的准确性和鲁棒性。

3）在Visual Genome数据集上对模型进行了系统全面的实验分析，深入探讨了各模块对整体性能的影响，直观地展示了模型对场景的感知能力和推理能力，证明了该方法在复杂场景语义理解任务中的实际应用潜力。

1　Rough-and-refine模型方法

1.1　问题描述

场景图生成任务旨在识别场景中存在的实体及其之间的关系，并生成一系列三元组<*S*,*P*,*O*>，其中，*S*、*P*、*O*分别表示三元组中的主体、谓词和客体。基于这些预测得到场景图{*V*,*E*}，其中：*V*代表所有实体集合，在图中以节点来表示；*E*代表所有谓词集合，在图中以有向边表示，从主体指向客体。

1.2 模型设计

本文提出的端到端粗–精网络的场景图生成模型如图2所示。

该模型分为两个部分：粗模型(rough part)和精模型(refine part)，分别用于预测和更新实体及其之间的关系。在粗模型中，借鉴DETR方法[22]，利用卷积神经网络和Transformer的编码器来提取图像特征，并将其与实体查询一同输入到实体解码器中进行自注意力计算，得到初步的实体表示。同时，在实体解码器后设计了一个谓词解码器，用于生成谓词的预测。实质上，实体之间的关系预测应该基于实体信息和图像特征信息来综合考虑。

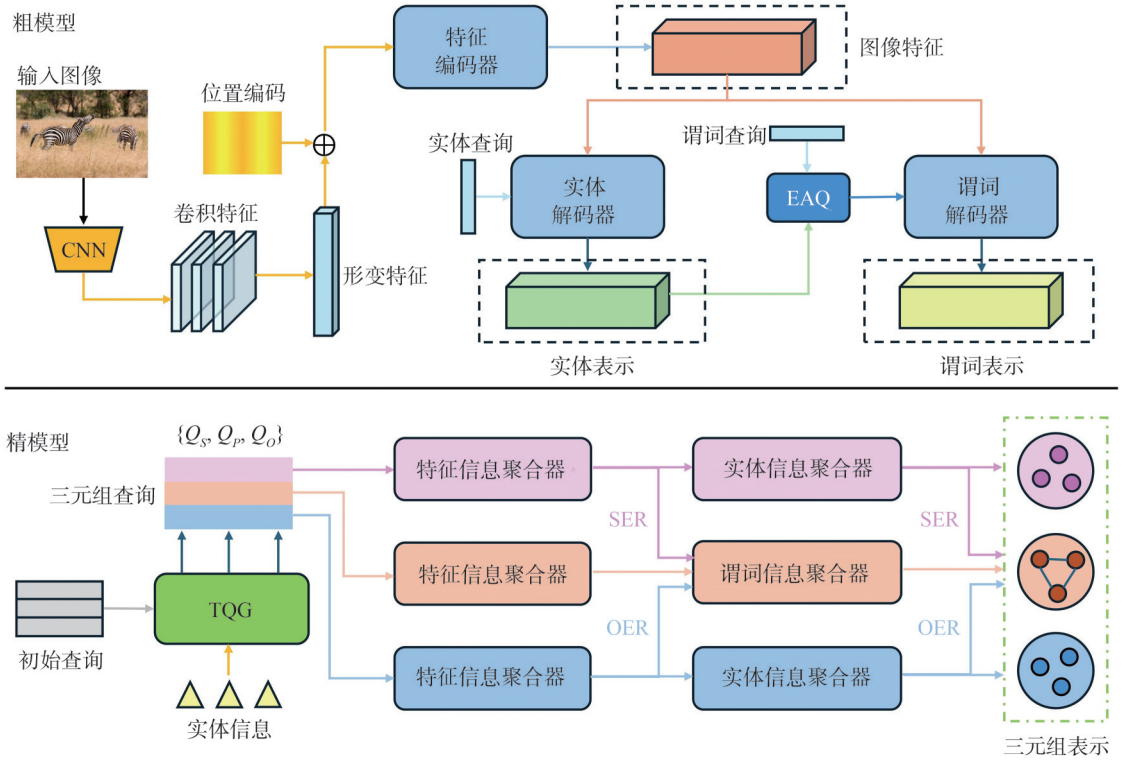

图2 粗–精模型详细网络结构

Fig. 2 Rough-and-refine detailed network

图2中，CNN为卷积神经网络(convolutional neural network)，EAQ为实体感知查询(Entity aware query)，TQG为三元组查询生成模块(triplet query generation)，SER为增强型主体表示(subject-enhanced representation)，OER为增强型客体表示(object-enhanced representation)。

因此，在谓词解码器中，将图像特征、实体表示以及谓词的查询作为输入。具体而言，将实体表示融合到谓词的查询中，之后在谓词解码器中继续做注意力计算，得到谓词的表示。通过粗模型，模型可以对场景有一个初步的感知和预测能力。然而，由于缺乏实体之间的信息交互，这一阶段难以挖掘深层语义信息。此外，模型对主客体的区分仍存在一定模糊性，因此需要设计精模型来进一步提升性能。

在精模型中，首先设计了一个三元组查询生成模块，为后续的三元组预测计算提供支持。由于在一组三元组中需要预测主体、客体和谓词三种不同信息，设计了三条路径：主体路径、客体路径和谓词路径。在每条路径中，模型引入粗模型计算得到的图像特征、实体表示和关系表示，并通过跨注意力计算融合不同信息。同时，将主体和客体的预测结果与谓词信息相融合，增加谓词部分的表示能力。这种设计使得模型在关系预测时能够更充分考虑实体对的状态，从而更深刻地理解主体和客体之间的交互关系。

1.2.1 粗模型设计

在粗模型中，首先使用卷积神经网络提取图像特征 $\boldsymbol{F}\in\mathbb{R}^{W\times H\times d}$，其中 W、H、d 分别代表当前特征图的宽度、高度和通道数。将图像特征输入到Transformer的编码器中，得到增强后的特征表示 $\boldsymbol{F}'\in\mathbb{R}^{W\times H\times d}$。将特征 $\boldsymbol{F}'$ 和实体查询 $\boldsymbol{Q}_S$ 一同输入到实体解码器中进行注意力计算，生成 N_S 个目标表示 $\boldsymbol{F}_{\text{ent}}$。后续从目标表示中进一步提取实体位置 $\boldsymbol{L}_{\text{ent}}\in\mathbb{R}^{N_S\times(1+C_S)}$，其中 C_S 为数据集中所有的实体类别总数。至此，模型的主要网络构成基本与DETR保持一致，完成了从图像到实体的初步检测与分类。

场景图生成任务还需要关系信息。因此，在实体

解码器后面添加一个谓词解码器，用于生成谓词的表示。在谓词解码器中，本文引入动态查询的策略，使用实体感知查询EAQ（Entity aware query）模块将实体信息融合到谓词查询中，使得模型在计算谓词表示时，能够充分地利用实体信息，从而获得更准确的谓词表示。利用初始的谓词查询 $\boldsymbol{Q}_P^0 \in \mathbb{R}^{N_P \times d}$，以及目标表示 $\boldsymbol{F}_{\text{ent}}$ 和实体位置 $\boldsymbol{L}_{\text{ent}}$，计算得到更新的谓词表示 $\boldsymbol{Q}_P^1 \in \mathbb{R}^{N_P \times d}$。具体的计算公式为：

$$\boldsymbol{Q}_P^1 = \text{CrossAtt}(\boldsymbol{q}=\boldsymbol{Q}_P^0, \boldsymbol{k}=\boldsymbol{F}_{\text{ent}} + \text{ReLU}(\boldsymbol{L}_{\text{ent}}\boldsymbol{X}_{\text{trans}}), \boldsymbol{v}=\boldsymbol{F}_{\text{ent}} + \text{ReLU}(\boldsymbol{L}_{\text{ent}}\boldsymbol{X}_{\text{trans}})) \tag{1}$$

式中：CrossAtt代表跨注意力计算；$\boldsymbol{q}$、$\boldsymbol{k}$、$\boldsymbol{v}$分别代表查询、键和值；ReLU表示激活函数；$\boldsymbol{X}_{\text{trans}} \in \mathbb{R}^{4\times d}$为线性变换层，将实体位置 $\boldsymbol{L}_{\text{ent}}$ 的维度从4映射到 d。和实体解码器类似，谓词解码器中用动态的谓词查询 $\boldsymbol{Q}_P^1$ 和图像特征 $\boldsymbol{F}'$ 作为输入，计算得到谓词的表示 $\boldsymbol{F}_{\text{prd}} \in \mathbb{R}^{N_P \times d}$：

$$\boldsymbol{F}_{\text{prd}} = \text{CrossAtt}(\boldsymbol{q}=\boldsymbol{Q}_P^1, \boldsymbol{k}=\boldsymbol{F}'+\boldsymbol{P}, \boldsymbol{v}=\boldsymbol{F}') \tag{2}$$

式中，CrossAtt代表跨注意力计算，$\boldsymbol{P} \in \mathbb{R}^{W\times H\times d}$是图像的位置编码。至此，模型已经计算出了实体和谓词的表示。理论上，只需在这些表示的基础上添加一个前馈神经网络（feed-forward network，FFN）模块，即可对三元组进行预测。

但是，这种直接的方法有一定的局限性：模型未能进一步挖掘实体之间的深层次交互信息，导致语义表达模糊，这限制了模型对复杂场景关系的理解能力，在预测细粒度关系时表现尤为不佳。基于上述观察，本文进一步设计精模型，以增强模型对场景语义和实体关系的深层次理解，从而提升三元组预测的准确性。

1.2.2　精模型设计

精模型对之前初步预测进行细化和更新，从而提升场景图生成的整体精度。如图2的精模型部分所示，包含3条独立的路径，分别对应三元组中主体、客体和谓词的计算。每条路径由几个核心模块组成。

三元组查询生成模块（triplet query generation, TQG）的主要作用是生成高质量的查询，从而提升模型性能并支持后续计算。为此，本文引入了一个复合的三元组查询表示，记为 $\boldsymbol{Q}_T \in \mathbb{R}^{N\times 3d}$。$\boldsymbol{Q}_T$包括主体查询、客体查询和谓词查询，分别记为 $\boldsymbol{Q}_S \in \mathbb{R}^{N\times d}$，$\boldsymbol{Q}_O \in \mathbb{R}^{N\times d}$，$\boldsymbol{Q}_P \in \mathbb{R}^{N\times d}$，即 $\boldsymbol{Q}_T=\{\boldsymbol{Q}_S, \boldsymbol{Q}_O, \boldsymbol{Q}_P\}$。初始的 $\boldsymbol{Q}_T$ 是随机生成的，和粗网络中对初始查询的处理方式相同，通过将实体信息融合到 $\boldsymbol{Q}_T$ 中，获得更新后包含丰富实体信息的 $\boldsymbol{Q}_T$。更新后的 $\boldsymbol{Q}_T$ 被分成3个子部分：$\boldsymbol{Q}_S$、$\boldsymbol{Q}_O$、$\boldsymbol{Q}_P$，分别输入到3条路径中。

实体路径包括主体路径和客体路径，它们的预测相互独立，如图2精模型中的上下路径所示。在实体路径中模型能够更好地区分主客体，生成高质量的主体表示和客体表示，并且为谓词预测提供了丰富信息。具体而言，在实体路径中，本文引入了两个模块：特征信息聚合(feature message integrating, FMI)模块和实体信息聚合(entity message integrating, EMI)模块。期望将图像特征 $\boldsymbol{F}'$ 和实体表示 $\boldsymbol{F}_{\text{ent}}$ 添加到主体路径和客体路径中。在特征信息聚合模块中，采用跨注意力机制实现特征的聚合，使模型关注图像的整体信息。计算过程为：

$$\boldsymbol{M}_{\text{out}}^{\text{sub}} = \text{MHA}(\boldsymbol{q}=\boldsymbol{Q}_S, \boldsymbol{k}=\boldsymbol{F}'+\boldsymbol{P}, \boldsymbol{v}=\boldsymbol{F}') \tag{3}$$

$$\boldsymbol{M}_{\text{out}}^{\text{obj}} = \text{MHA}(\boldsymbol{q}=\boldsymbol{Q}_O, \boldsymbol{k}=\boldsymbol{F}'+\boldsymbol{P}, \boldsymbol{v}=\boldsymbol{F}') \tag{4}$$

式中，$\boldsymbol{M}_{\text{out}}^{\text{sub}}$ 和 $\boldsymbol{M}_{\text{out}}^{\text{obj}}$ 分别代表主体路径和客体路径在FMI模块后的输出结果。MHA代表多头注意力机制。EMI与FMI类似，采用相同的方法聚合实体信息。在这一过程中，实体信息起到指导作用，提供了场景中实体的分布情况和具体的类别。计算过程为：

$$\boldsymbol{E}_{\text{out}}^{\text{sub}} = \text{MHA}(\boldsymbol{q}=\boldsymbol{M}_{\text{out}}^{\text{sub}}, \boldsymbol{k}=\boldsymbol{F}_{\text{ent}}, \boldsymbol{v}=\boldsymbol{F}_{\text{ent}}) \tag{5}$$

$$\boldsymbol{E}_{\text{out}}^{\text{obj}} = \text{MHA}(\boldsymbol{q}=\boldsymbol{M}_{\text{out}}^{\text{obj}}, \boldsymbol{k}=F_{\text{ent}}, \boldsymbol{v}=\boldsymbol{F}_{\text{ent}}) \tag{6}$$

式中，$\boldsymbol{E}_{\text{out}}^{\text{sub}}$ 和 $\boldsymbol{E}_{\text{out}}^{\text{obj}}$ 分别代表主体路径和客体路径在EMI模块后的输出结果。

谓词路径的设计如图2精模型的中间路径所示，是为了在初步谓词表示基础上进一步地更新和优化。该部分和实体路径的结构类似，同样设计了两个模块来聚合信息：特征信息聚合模块和谓词信息聚合（predicate message integrating，PMI）模块。引入粗模型中的图像特征 $\boldsymbol{F}'$ 和谓词表示 $\boldsymbol{F}_{\text{prd}}$，使谓词路径能够学习到更精确的谓词表示。在FMI模块中，输入为谓词查询和图像特征 $\boldsymbol{F}'$，计算过程和实体路径中的FMI相同：

$$\boldsymbol{M}_{\text{out}}^{\text{prd}} = \text{MHA}(\boldsymbol{q}=\boldsymbol{Q}_P, \boldsymbol{k}=\boldsymbol{F}'+\boldsymbol{P}, \boldsymbol{v}=\boldsymbol{F}') \tag{7}$$

式中，$\boldsymbol{M}_{\text{out}}^{\text{prd}}$ 为谓词路径在FMI模块的输出结果。另外，谓词信息聚合模块以谓词表示 $\boldsymbol{F}_{\text{prd}}$ 和FMI模块的输出结果 $\boldsymbol{M}_{\text{out}}^{\text{prd}}$ 作为输入，经过多头注意力计算：

$$\boldsymbol{P}_{\text{out}}^{\text{prd}} = \text{MHA}(\boldsymbol{q}=\boldsymbol{M}_{\text{out}}^{\text{prd}}, \boldsymbol{k}=\boldsymbol{F}_{\text{prd}}, \boldsymbol{v}=\boldsymbol{F}_{\text{prd}}) \tag{8}$$

式中，$\boldsymbol{P}_{\text{out}}^{\text{prd}}$ 为谓词路径在PMI模块的输出结果。

通过整个三元组路径的计算（实体路径和谓词路径），模型能够获得经过信息整合加工后的三元组表示。然而，目前三条路径的计算彼此完全独立。然而，谓词的预测对于主客体对的依赖是非常明显，而这种彼此独立的计算难以有效捕捉到这一依赖。因此，本文提出了增强型谓词表示（enhanced predicate representation，EPR）来解决这一问题。在谓词路径中，本文将主体信息和客体信息加入到该路径的中间过程中，包括增强型主体表示SER和增强型字体表示OER。具体做法为：

$$\boldsymbol{M}_{\text{out}}^{\text{prd}}=\boldsymbol{M}_{\text{out}}^{\text{sub}}+\boldsymbol{M}_{\text{out}}^{\text{obj}}+\boldsymbol{M}_{\text{out}}^{\text{prd}} \tag{9}$$

$$\boldsymbol{P}_{\text{out}}^{\text{prd}}=\boldsymbol{P}_{\text{out}}^{\text{sub}}+\boldsymbol{P}_{\text{out}}^{\text{obj}}+\boldsymbol{P}_{\text{out}}^{\text{prd}} \tag{10}$$

这种方式通过将主体和客体路径的信息注入到谓词路径中，使得谓词预测不仅仅依赖于自身的信息，并且能充分考虑到主体和客体之间的交互关系。

1.2.3 预测网络

整个模型完成了对三元组的表示，最后需要获得具体的预测结果。主体和客体需要预测其类别和用坐标框表示的位置信息，包括归一化的中心坐标(x, y)和坐标框的长度、宽度；而谓词仅需要预测其类别。精模型以三元组的表示$<\boldsymbol{F}_{\text{sub}}, \boldsymbol{F}_{\text{obj}}, \boldsymbol{F}_{\text{prd}}>$作为最终输出，其中：

$$\begin{cases}\boldsymbol{F}_{\text{sub}}=\boldsymbol{E}_{\text{out}}^{\text{sub}},\\ \boldsymbol{F}_{\text{obj}}=\boldsymbol{E}_{\text{out}}^{\text{obj}},\\ \boldsymbol{F}_{\text{prd}}=\boldsymbol{P}_{\text{out}}^{\text{prd}}\end{cases} \tag{11}$$

本文对3个不同部分分别独立地用前馈神经网络完成预测。前馈神经网络包括两个带ReLU激活函数的感知机和一个线性投影层，完成类别的分类和坐标框的回归。

1.3 损失函数计算

整个模型的损失函数可以分为两个部分，一部分为实体损失$\text{Loss}_{\text{entity}}$，另一部分则是来自谓词的损失$\text{Loss}_{\text{predicate}}$。三元组总的损失为两部分相加：

$$\text{Loss}_{\text{triplet}}=\text{Loss}_{\text{entity}}+\text{Loss}_{\text{predicate}} \tag{12}$$

式中，实体部分的损失由主体损失Loss_{sub}和客体损失Loss_{obj}组成：

$$\text{Loss}_{\text{entity}}=\text{Loss}_{\text{sub}}+\text{Loss}_{\text{obj}} \tag{13}$$

使用匈牙利匹配算法来计算真实三元组和预测三元组之间的损失，将所有匹配起来的三元组计算其损失。对于谓词的损失，采用交叉熵损失Loss_{cls}来计算：

$$\text{Loss}_{\text{cls}}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}y_{ij}\log(p_{ij}) \tag{14}$$

式中：N、C分别为样本总数和类别总数；y_{ij}为符号函数，如果样本i的真实类别是j则为1，否则为0；p_{ij}为模型预测样本i属于类别j的概率。对于实体的损失函数，本文需要同时关注其类别和检测框，关于检测框的损失，本文使用L_1损失和GIOU损失来计算：

$$\text{Loss}_{\text{box}}(b_{\text{p}},b_{\text{t}})=\alpha\text{L}_1(b_{\text{p}},b_{\text{t}})+\beta\text{Loss}_{\text{GIOU}}(b_{\text{p}},b_{\text{t}}) \tag{15}$$

式中，α和β为可调整的权重，用来平衡两部分损失，b_{p}和b_{t}表示模型预测的检测框和真实检测框。因此三元组损失函数表示为：

$$\text{Loss}_{\text{triplet}}=\text{Loss}_{\text{sub}}+\text{Loss}_{\text{obj}}+\text{Loss}_{\text{predicate}} \tag{16}$$

式中，Loss_{sub}和Loss_{obj}包括Loss_{cls}和$\text{Loss}_{\text{box}}(b_{\text{p}},b_{\text{t}})$两部分的计算，$\text{Loss}_{\text{predicate}}$仅包含$\text{Loss}_{\text{cls}}$。

2 实验结果

2.1 数据集和任务

本次实验的数据来源是场景图生成领域里面最常用的公开数据集之一——Visual Genome。该数据集专注于图像的高级语义理解，已被广泛应用于多种计算机视觉任务。本文使用从Visual Genome中整理出的VG-150数据集，该数据集包括150种不同的实体类别和50种不同的关系类别。VG-150数据集包含约108 000张图像，提供了丰富的场景图信息，适用于多种与场景理解相关的任务。

场景图生成任务一般包含3种不同的子任务：谓词分类（predicate classification，PredCLS）、场景图分类（scene graph Classification，SGCLS）、场景图检测（scene graph Detection，SGDET），任务的难度依次递增。其中：PredCLS给定场景中的目标位置和类别，要求预测出谓词；SGCLS给出场景中的目标位置，无目标类别信息，要求预测出目标的类别和谓词的类别；SGDET仅给出当前场景，要求预测出场景图的所有信息，包括目标的位置和类别，以及谓词的类别。本次实验选择SGDET任务，模型直接从场景入手，仅考虑视觉图像信息，不涉及目标先验信息和其他的各种先验知识（统计先验、语言先验等）。这种任务设置更贴近实际应用场景，能够考查模型在仅依赖图像信息的情况下，自动推理和生成完整的场景图的能力。

2.2 实现细节

在实验中，所用配备NVIDIA A100 GPU的硬件，GPU显存为40 G，对本文的RRM模型训练150轮次。为了优化训练效率和性能，设置batch大小为8；主干网络选用常见的ResNet-50，该部分的初始学习率设置为10^{-5}；网络的整体初始学习率设置为10^{-4}。为提高模型的泛化性，引入权重衰减技术，将其设置为10^{-3}。另外，模型在预测场景中的实体、谓词、三元组的数量是固定的，本文设置实体数量N_S为100，谓词数量N_P为150，三元组数量N_T为200。

2.3 量化结果和消融实验

2.3.1 量化结果

为了验证RRM在场景图生成任务上的性能表现，本文采用了该研究领域常用的一些评价指标：recall@K（R@K）和mean recall@K（mR@K）。其中，R@K能够反映该模型在数据集上的整体召回率，衡量一个模型预测的前K个三元组在真实标签三元组中是否能找到；而mR@K指标则会对每一个谓词类别计算一个R@K，最后取平均。这种评价指标可以更关注模型对数据集中低频谓词类别学习的能力，尤其

是面对数据集存在的长尾分布问题，能够体现模型对所有谓词类别的学习能力。R@K（记为R_K）的计算公式如下：

$$R_K=\frac{|G\cap X_K|}{|G|} \tag{17}$$

式中：X_K为模型提供的K个置信度最高的三元组，通常K取20、50和100；G为标注的真实三元组。

mR@K指标使得出现频率低的那些谓词和出现频率高的谓词具有相同的重要性。mR@K（记为R_{mK}）的计算公式如下：

$$R_{mK}=\frac{1}{|P'|}\sum_{p\in P'}\frac{\left|G^p\cap X_K^p\right|}{\left|G^p\right|} \tag{18}$$

式中：P是包含所有可能谓词类别的集合；P'是P的子集，包含所有标注里的谓词类别。为了单独计算每个谓词的分数，对于每个谓词$p\in P'$，将所有标注的三元组分成单独的集合G^p，对于模型预测的X_K个三元组也做同样的操作，记为X_K^p。

表1列举出了若干两阶段的方法和单阶段的方法。两阶段的方法包括：IMP[23]、VCTree[24]、RelDN[26]、SQUAT[30]等，它们对每个实体对的关系进行预测。从总体上看，两阶段方法一般有更高的R@K，但往往其参数量也非常巨大。单阶段的方法包括FCSGG[31]、HOTR[32]、RelTR[18]、SGTR[19]、SGTR+[20]。受益于预测更稀疏的关系和更精简的网络结构，其参数量较小，但R@K指标受到影响。

表1　各模型在VG测试集上的实验结果

Tab. 1　Experimental results of each model on the VG test set

方法类别	模型	R@20	R@50	R@100	mR@20	mR@50	mR@100	参数量/M
两阶段方法	IMP(EBM)[23]	18.1	25.9	31.2	2.8	4.2	5.4	322.2
	VCTree(TDE)[24]	14.0	19.4	23.2	6.3	9.3	11.1	360.8
	VTransE[25]	24.5	31.3	35.5	5.1	6.8	8.0	312.3
	RelDN[26]	21.1	28.3	32.7				615.6
	GPS–Net[27]	23.3	31.0	35.8	7.5	10.7	12.6	
	GB–Net[28]		26.4	30.0		6.1	7.3	
	BGT–Net[29]	25.2	32.8	35.2	5.7	7.8	8.9	203.8
	SQUAT[30]		24.5	28.9	10.6	14.1	16.5	
单阶段方法	FCSGG[31]	16.1	21.3	25.1	2.7	3.6	4.2	87.1
	HOTR[32]		23.5	27.7		9.4	12.0	
	RelTR[18]	21.2	27.5		6.3	10.8		63.7
	SGTR[19]		24.6	28.4		12.0	15.2	117.1
	SGTR+[20]		26.4	30.1		13.1	17.0	
	RRM(ours)	23.8	29.1	32.5	7.7	11.0	12.4	76.4

本文提出的方法RRM在单阶段方法中取得了更好的R@K结果，在R@20、R@50、R@100指标上优于其他的单阶段方法。具体而言，本文的R@20=23.8，R@50=29.1，R@100=32.5，分别比其他单阶段方法的最优值高出2.6、1.6、2.4。mR@K指标高于FCSGG、HOTR、RelTR方法以及大多数的两阶段方法，分别达到了mR@20=7.7，mR@50=11.0，mR@100=12.4。纵向比较而言，本文的模型全面优于FCSGG和HOTR；R@K和mR@K六项评价指标优于RelTR，但RelTR参数量更小；对比SGTR和SGTR+模型，本文的模型在R@K、mR@20以及参数量上更好，而SGTR和SGTR+在mR@50和mR@100指标上取得了更好的成绩。这主要是得益于SGTR的谓词节点模块的设计和图组装模块，后者利用生成的实体节点和谓词节点，将场景图生成视为一个二分图匹配问题。

由于数据集的长尾分布问题，模型在mR@K指标的表现有一定的提升空间，引入了logit调整方法（logit adjustment, LA）[35]，用以直接解决谓词类别分布不均的问题。实验结果如表2所示，采用logit调整方法后，模型在mR@K指标上显著提升，并且“body”和“tail”的谓词类别（出现频率中等和较低的类别）明显更多。但R@K和“head”部分的谓词类别（出现频率高的类别）受到影响，有不同程度下降。

表2　解决长尾分布效应的实验结果

Tab. 2　Experimental results for addressing the long tail distribution problem

模型	R@50	R@100	mR@50	mR@100	head	body	tail
RRM	29.1	32.5	11.0	12.4	30.8	15.4	5.7
RRM+LA	26.2	29.3	14.8	16.5	27.5	17.7	10.6

2.3.2　消融实验

在消融实验中本文考虑模型各模块对整体性能的影响。首先是粗模型和精模型的对比，本文将整个精模

型去掉，仅用粗模型来预测三元组。实验结果如表 3 所示，从数据中可以得出：1）粗模型获取了一定的场景理解能力，R@K 和 mR@K 指标上比一些现有的模型更佳；2）第二是精模型的更新作用，使模型对三元组的预测更为准确，在 R@K、mR@K 和平均值评价指标上均有不同程度的提升。具体而言，K 分别取 50、100 时，精模型在 R@K 指标上分别提升了 7.2、6.9，mR@K 分别提升了 3.5、3.3，R@K 与 mR@K 指标平均提升了 5.3（表 3 中平均值为 4 项指标的平均）。两个不同模块所预测的场景图结果详见第 2.4 节。

表 3　粗–精模型和其他模型的对比

Tab. 3　Comparison among rough part, refine part and other models

模型	R@50	R@100	mR@50	mR@100	平均值
FCSGG[31]	21.3	25.1	3.6	4.2	13.6
HOTR[32]	23.5	27.7	9.4	12.0	18.2
SGTR[19]	24.6	28.4	12.0	15.2	20.1
GB–Net[28]	26.4	30.0	6.1	7.3	17.5
TDE[33]	16.9	20.3	8.2	9.8	13.8
INF[34]	23.9	27.1	9.4	11.7	18.0
粗模型（本文）	21.9	25.6	7.5	9.1	16.0
精模型（本文）	29.1	32.5	11.0	12.4	21.3

本文测试了精模型中各个模块对整体性能的影响，消融实验结果如表 4 所示。由表 4 可知，各模块对场景图的预测都起到正向作用，去掉任意一个模块都会造成实验结果下降。其中 FMI 和 EMI 模块对模型的影响较大，在去掉 FMI 或 EMI 后，4 项指标平均值分别下降了 11.7% 和 8.9%，因为这两个模块引入了重要的场景信息。TQG 和 EPR 对模型也有一定程度的提升，去掉 TQG 和 EPR 后，平均值下降了 4.7% 和 5.2%。表 4 中第一排的模型去除了所有 4 个模块，等价于表 2 中单独的 rough part，平均值下降了 24.9%。

表 4　精模型各个模块的消融实验结果

Tab. 4　Ablation study on each module of the refine part

TQG	FMI	EMI	EPR	R@50	R@100	mR@50	mR@100	平均值
×	×	×	×	21.9	25.6	7.5	9.1	16.0
√	√	√	×	27.8	31.5	10.2	11.3	20.2
√	×	√	√	25.4	29.5	9.6	10.6	18.8
√	√	×	√	26.7	30.3	9.9	10.8	19.4
×	√	√	√	28.0	31.6	10.5	11.0	20.3
√	√	√	√	29.1	32.5	11.0	12.4	21.3

最后，本文还研究了不同个数的编码器和解码器对模型的影响，实验结果如表 5 所示。显然，编码器和解码器数量越多，模型性能越好。但过多的解码器会导致参数量的激增，计算负担增大。数量从 6 增加到 9 时，平均值只增长了 0.1，但是参数量和 FLOPs 均增长了超过 30%。说明模型在 6 层时的性能已经趋于饱和。因此最终选择适当的编码器解码器数量（均为 6 个），既能有不错的性能表现，也有较小的参数量和计算量。

表 5　编码器和解码器数量对模型的影响

Tab. 5　Influence of the number of encoders and decoders on the model

编码器数量	解码器数量	评价指标					参数量/M	FLOPs/G
		R@50	R@100	mR@50	mR@100	平均值		
3	3	25.2	28.7	8.7	10.2	18.2	51.2	8.79
3	6	28.3	31.4	10.6	11.9	20.6	72.4	13.31
6	3	27.6	30.8	9.8	11.5	19.9	55.1	8.94
6	6	29.1	32.5	11.0	12.4	21.3	76.4	13.47
9	9	29.4	32.8	11.0	12.2	21.4	101.5	18.14

2.4　可视化实验

为了更直观地感受模型的场景理解能力，本文选取了 3 个不同场景进行分析，分别用粗模型和精模型对图像做场景图预测，并对它们预测出的场景图进行对比，结果如图 3～5 所示。由图 3 可知，在案例 1 中粗模型预测了<man, riding, bicycle>的三元组，但显然模型对谓词的预测不准确，在精模型中该三元组被修正为<man, on, bicycle>，似乎是两者之间更准确的关系，另外<glasses of man>更换为了<man, wearing, glasses>。如图 4 所示，在案例 2 中精模型额外添加了<boy, holding, banana>和<elephant, eating, banana>的三元组，不仅增加了两个有意义的三元组，还建立起了场景中 boy 和 elephant 的联系。虽然数据集中没有合适的谓词来直接描述他们的关系，但精模型仍然通过他们共有的和 banana 之间的关系，提供了间接的联系，而在粗模型中，boy 和 elephant 是完全割裂的。如图 5 所示，案例 3 中粗模型和精模型所生成的场景图相同，同时可以注意到，该场景图的谓词都是“on”，本文推测粗

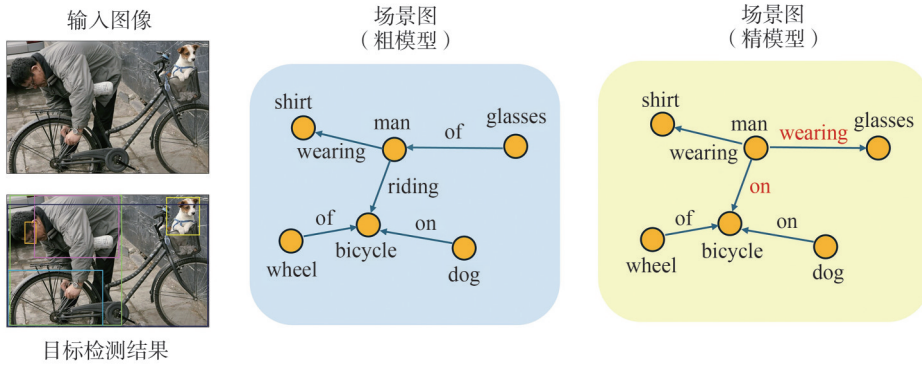

图3　案例1粗–精模型场景图预测

Fig. 3　Rough-and-refine model prediction for Case 1

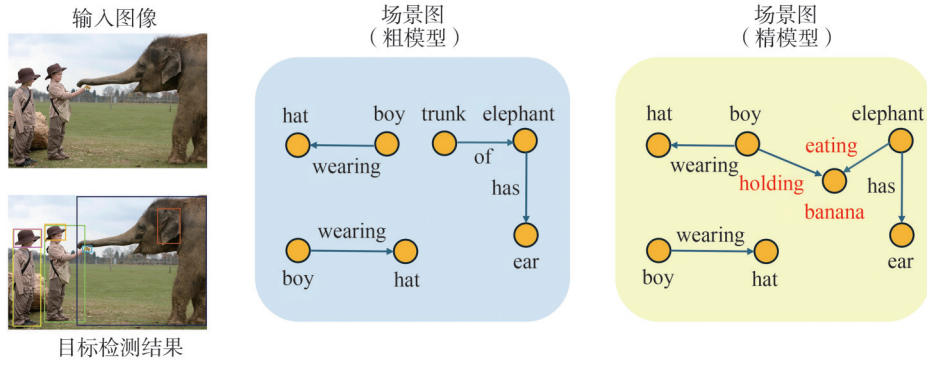

图4　案例2粗–精模型场景图预测

Fig. 4　Rough-and-refine model prediction for Case 2

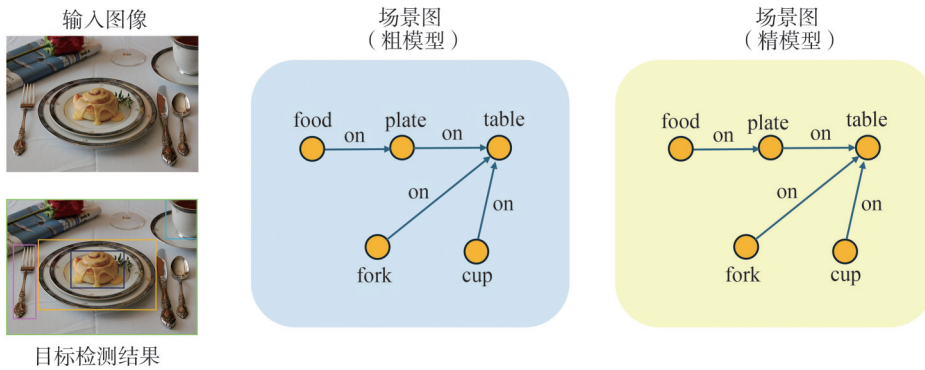

图5　案例3粗–精模型场景图预测

Fig. 5　Rough-and-refine model prediction for Case 3

模型对这种简单的谓词有不错的学习能力，但面对较为复杂的场景，可能会漏掉或者错误地预测三元组。不过，通过评价指标的数据和场景图的可视化结果都显示粗模型可以做出一些正确的认识。

3　结　论

本文针对场景图生成任务中存在的谓词预测不准确、主客体识别不清晰的问题，设计了基于Transformer的端到端粗–精网络研究方法，即rough-and-refine模型。该模型通过两个不同组成部分，从不同层次对场景进行解析。充分利用模型提取到的各种特征，以信息交互为主的计算方式，为场景图生成任务提供了新的思路和方法。并且通过大量的实验结果表明，本文的RRM模型在公开数据集Visual Genome上取得了优异的表现。在可视化实验中，也充分展示了模型针对特定场景生成场景图的能力。

然而从实验数据中可以看出，mR@K指标仍有上升空间，即对于数据集中出现频率低的类别还有待模型进一步提升。

参考文献：

[1] 李林昊,韩冬,董永峰,等. 基于关联信息增强与关系平衡的场景图生成方法[J]. 计算机应用,2025,45(3):953–962.

[2] Duan Jingwen,Min Weidong,Yang Ziyuan,et al.Global semantic information extraction based scene graph generation algorithm[J].Journal of Image and Graphics,2022,27(7):2214–2225.[段静雯,闵卫东,杨子元,等.提取全局语义信息的场景图生成算法[J].中国图象图形学报,2022,27(7):2214–2225.]

[3] Dai Bo,Zhang Yuqi,Lin Dahua.Detecting visual relationships with deep relational networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition.Honolulu:IEEE,2017:3298–3308.

[4] Amodeo F,Caballero F,Díaz–Rodríguez N,et al.OG–SGG:Ontology-guided scene graph generation—A case study in transfer learning for telepresence robotics[J].IEEE Access,2022,10:132564–132583.

[5] Jung J,Park J.Visual relationship detection with language prior and softmax[C]//Proceedings of the 2018 IEEE International Conference on Image Processing, Applications and Systems.Sophia Antipolis:IEEE,2018:143–148.

[6] Liao Wentong,Rosenhahn B,Shuai Ling,et al.Natural language guided visual relationship detection[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach:IEEE,2019:444–453.

[7] Yu J,Chai Y,Wang Y,et al.CogTree: Cognition Tree Loss for Unbiased Scene Graph Generation[C]//International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence Organization, 2021.

[8] Zareian A,Karaman S,Chang S F.Bridging knowledge graphs to generate scene graphs[M]//Computer Vision–ECCV 2020.Cham:Springer International Publishing,2020:606–623.

[9] Wang Lichun,Fu Fangyu,Xu Kai,et al.Scene graph generation method based on dual-stream multi-head attention[J].Journal of Beijing University of Technology,2024,50(10):1198–1205.[王立春,付芳玉,徐凯,等.基于双分支多头注意力的场景图生成方法[J].北京工业大学学报,2024,50(10):1198–1205.]

[10] Zhang Liang, Zhang Shuai, Shen Peiyi, et al. Relationship detection based on object semantic inference and attention mechanisms[C]//Proceedings of the 2019 on International Conference on Multimedia Retrieval. Ottawa: ACM, 2019:68–72.

[11] Dhingra N, Ritter F, Kunz A. BGT-net: Bidirectional GRU transformer network for scene graph generation[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.Nashville:IEEE,2021:2150–2159.

[12] Yin Guojun,Sheng Lu,Liu Bin,et al.Zoom-net:Mining deep feature interactions for visual relationship recognition[C]//Computer Vision–ECCV 2018.Cham:Springer,2018:330–347.

[13] Li Yikang,Ouyang Wanli,Wang Xiaogang,et al.ViP–CNN:Visual phrase guided convolutional neural network[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 7244–7253.

[14] Zellers R,Yatskar M,Thomson S,et al.Neural motifs:Scene graph parsing with global context[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City:IEEE,2018:5831–5840.

[15] Gu Jiuxiang,Zhao Handong, Lin Zhe,et al.Scene graph generation with external knowledge and image reconstruction[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Long Beach:IEEE,2019:1969–1978.

[16] Qi Mengshi,Li Weijian,Yang Zhengyuan,et al.Attentive relational networks for mapping images to scene graphs[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE,2019:3952–3961.

[17] Zhang Liang, Zhang Shuai, Shen Peiyi, et al. Relationship detection based on object semantic inference and attention mechanisms[C]//Proceedings of the 2019 on International Conference on Multimedia Retrieval. Ottawa: ACM, 2019:68–72.

[18] Cong Yuren,Yang M Y,Rosenhahn B.RelTR:Relation transformer for scene graph generation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2023,45(9):11169–11183.

[19] Li Rongjie,Zhang Songyang,He Xuming.SGTR:End-to-end scene graph generation with transformer[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022:19464–19474.

[20] Li Rongjie,Zhang Songyang,He Xuming.SGTR:End-to-end scene graph generation with transformer[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2024,46(4):2191–2205.

[21] Khandelwal S,Sigal L.Iterative scene graph generation[J].Advances in Neural Information Processing Systems,2022,35:24295–24308.

[22] Carion N,Massa F,Synnaeve G,et al.End-to-end object detection with transformers[M]//Computer Vision – ECCV 2020.Cham:Springer International Publishing,2020:213–229.

[23] Xu Danfei,Zhu Yuke,Choy C B,et al.Scene graph generation by iterative message passing[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern

Recognition.Honolulu:IEEE,2017:3097–3106.

[24] Tang Kaihua,Zhang Hanwang,Wu Baoyuan,et al.Learning to compose dynamic tree structures for visual contexts [C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE,2019:6612–6621.

[25] Zhang Hanwang,Kyaw Z,Chang S F,et al.Visual translation embedding network for visual relation detection[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3107–3115.

[26] Zhang Ji,Shih K J,Elgammal A,et al.Graphical contrastive losses for scene graph parsing[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Long Beach:IEEE,2019:11527–11535.

[27] Lin Xin,Ding Changxing, Zeng Jinquan,et al.GPS-net: Graph property sensing network for scene graph generation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Seattle: IEEE,2020:3743–3752.

[28] Zareian A,Karaman S,Chang S F.Bridging knowledge graphs to generate scene graphs[M]//Computer Vision–ECCV 2020.Cham:Springer International Publishing,2020:606–623.

[29] Dhingra N, Ritter F, Kunz A. BGT-net: Bidirectional GRU transformer network for scene graph generation[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville: IEEE,2021:2150–2159.

[30] Jung D,Kim S,Kim W H,et al.Devil's on the edges:Selective quad attention for scene graph generation[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 18664–18674.

[31] Liu Hengyue,Yan Ning,Mortazavi M,et al.Fully convolutional scene graph generation[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Nashville:IEEE,2021:11541–11551.

[32] Kim B,Lee J,Kang J,et al.HOTR:End-to-end human-object interaction detection with transformers[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Nashville:IEEE,2021:74–83.

[33] Tang Kaihua,Niu Yulei,Huang Jianqiang,et al.Unbiased scene graph generation from biased training[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Seattle:IEEE,2020:3713–3722.

[34] Biswas B A,Ji Qiang.Probabilistic debiasing of scene graphs[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver:IEEE,2023:10429–10438.

[35] Menon A K, Jayasumana S, Rawat A S,et al.Long-tail learning via logit adjustment[C]//International Conference on Learning Representations.2021.

End-to-end Rough-and-refine Model for Scene Graph Generation Based on Transformer

LI Junliang[1], *LYU Shirong*[2], *LI Wei*[1*]

(1.School of Aeronautics and Astronautics, Sichuan University, Chengdu 610065,China;
2.College of Computer and Information Science, Southwest University, Chongqing 400715,China)

Abstract:

Objective Scene graph generation is a critical task in computer vision, enabling a comprehensive and deep understanding of visual scenes. It focuses on identifying entities and the relationships between them, ultimately requiring the model to output a series of triplets (subject-predicate-object) and a graph-structured scene representation. This places greater demands on the model's understanding and reasoning capabilities. Although existing scene graph generation methods have achieved substantial success, most models are hindered by either an excessive number of parameters or inaccurate predicate judgments. This study proposes an end-to-end rough-and-refine model (RRM) for scene graph generation to overcome these challenges.

Methods The end-to-end rough-and-refine network model proposed in this study for scene graph generation consisted of two components: the rough part and the refine part, which were responsible for predicting and updating entities and their relationships, respectively. In the rough part, image features were initially extracted using convolutional neural networks and a Transformer encoder. These features were then input alongside entity queries into the entity decoder for self-attention computation, resulting in preliminary entity representations. In addition, a predicate decoder was designed to follow the entity decoder and generate predictions for predicates. Predicting relationships between entities requires considering both entity information and image feature information comprehensively. Therefore, the predicate decoder took image features, entity representations, and predicate queries as inputs. Specifically, entity representations were integrated into the predicate queries, followed by further attention computations in the predicate decoder to obtain predicate representations. Through the rough part, the system gained a preliminary perception and predictive capability regarding the scene. However, due to a lack of information interaction between entities, this stage struggled to excavate deeper semantic information. In addition, ambiguity remained in distinguishing between subjects and objects, necessitating the design of the re-

fine part to enhance performance. In the refine part, a triplet query generation module was first established to support subsequent calculations for triplet prediction. Since a triplet required the prediction of three distinct types of information, subject, object, and predicate, three paths were designed: the subject path, object path, and predicate path. In each path, the model incorporated image features, entity representations, and relationship representations derived from the rough part, utilizing cross-attention computations to integrate different information. In addition, the predictions for subjects and objects were fused with the predicate information to enhance the representational capacity of the predicate component. This design allowed the model to more thoroughly consider the states of entity pairs during relationship prediction, fostering a deeper understanding of the interactions between subjects and objects. After the model completed the representation of triplets, it was required to produce specific prediction results. The subject and object needed to predict their categories along with location information represented by bounding boxes, which included normalized center coordinates (x, y) and the dimensions of the bounding boxes (length and width). In contrast, the predicate only required the prediction of its category. Predictions for the different paths of the triplet were independently executed using feedforward neural networks. Each feedforward neural network consisted of two perceptrons with ReLU activation functions and a linear projection layer, facilitating both category classification and bounding box regression.

Results and Discussions Several commonly used metrics in this research domain were employed, including Recall@K (R@K) and Mean Recall@K (mR@K) to evaluate the performance of RRM in the scene graph generation task. R@K reflected the overall recall rate of the model on the dataset, measuring whether the top-k predicted triplets can be found among the true labeled triplets. In contrast, the mR@K metric calculated an R@K for each predicate category and then computed the average. This evaluation metric placed greater emphasis on the model's ability to learn low-frequency predicate categories within the dataset, ensuring that infrequent predicates received equal importance as frequent ones. This was particularly critical in addressing the long-tail distribution problem present in the dataset, as it demonstrated the model's learning capability across all predicate categories. The proposed method, RRM, achieved superior R@K results among single-stage methods, outperforming other single-stage approaches in the R@20, R@50, and R@100 metrics. Specifically, the RRM model achieved R@20 = 23.8, R@50 = 29.1, and R@100 = 32.5, which were higher than the optimal values of other single-stage methods by 2.6, 1.6, and 2.4, respectively. The mR@K metrics exceeded those of FCSGG, HOTR, RelTR, and most two-stage methods, reaching mR@20 = 7.7, mR@50 = 11.0, and mR@100 = 12.4. In a vertical comparison, the model significantly outperformed FCSGG and HOTR, and also demonstrated better performance across the six evaluation metrics, R@K and mR@K, compared to RelTR, although RelTR has a smaller parameter count. When comparing SGTR and SGTR+, the model performed better in terms of R@K, mR@20, and parameter count, while SGTR and SGTR+ exhibited better results in mR@50 and mR@100. In ablation experiments, the results indicated that each module made a positive contribution to the prediction of scene graphs, with the removal of any single module leading to a decline in the experimental results. The FMI and EMI modules have a significant impact on the model; removing either FMI or EMI resulted in an average decrease of 11.7% and 8.9%, respectively, as these modules introduced crucial scene information. The TQG and EPR modules also provided measurable improvement, with average decreases of 4.7% and 5.2% when removed. The model represented in the first row of the table, which excluded all four modules, was equivalent to the rough part, showing an average decrease of 24.9%.

Conclusions A scene graph generation method based on a rough-and-refine network is proposed to address the challenge of inadequate predicate representation. Experimental results demonstrate that the proposed network model achieves strong performance on public datasets, surpassing existing models across several key evaluation metrics and enabling the accurate extraction of information from images to scene graphs. Visualization experiments conducted in diverse scenarios confirm the model's capability in scene graph generation and highlight the performance improvements provided by the refine model over the rough model.

Key words: scene graph generation; computer vision; artificial intelligence; visual relationship detection

（编辑　陈　雪）

引用格式：Li Junliang, Lyu Shirong, Li Wei.End-to-end rough-and-refine model for scene graph generation based on Transformer[J].Advanced Engineering Sciences,2025,57(5):344–354.[李俊良,吕诗融,李炜 . 基于 Transformer 架构的端到端粗–精网络场景图生成研究方法[J]. 工程科学与技术,2025,57(5):344–354.]