计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(9): 2285−2301, 2026

# 基于常识引导与多模态特征融合的视觉关系推理

武栩含　孙仕亮
（上海交通大学自动化与感知学院　上海　200240）
（xuhan.wu@sjtu.edu.cn）

# Visual Relationship Reasoning via Commonsense Guide and Multimodal Feature Fusion

Wu Xuhan and Sun Shiliang
（*School of Automation and Intelligent Sensing*, *Shanghai Jiao Tong University*, *Shanghai* 200240）

**Abstract**　Efficient scene graph generation (SGG) serves as a bridge connecting low-level visual perception and high-level semantic reasoning, playing an important role in various vision-language tasks. However, existing dedicated SGG models often produce unreasonable relation predictions due to the lack of semantic constraint mechanisms. Although multimodal large language models (MLLMs) possess rich commonsense knowledge, their inherent black-box characteristics and severe hallucination issues lead to unreliable relationship reasoning results when directly applied to SGG tasks. To address these challenges, we propose KETR (commonsense knowledge embedding Transformer), an end-to-end scene graph generation model that incorporates commonsense knowledge embeddings. The model extracts global and local image features through a multi-scale visual feature fusion encoder and employs a parallel decoder based on learnable queries for entity detection. The core innovation lies in a soft alignment knowledge fusion mechanism that dynamically aggregates knowledge information through probabilistic mapping between visual features and external commonsense knowledge graph embedding spaces, and a commonsense-aware subject-object indicator and knowledge-driven scoring constraints are introduced, integrating the TransE scoring function as a regularization term into the joint training loss to significantly improve the logical consistency of relationship reasoning. Extensive experiments on Visual Genome and Open Images V6 demonstrate that KETR achieves competitive performance on multiple standard SGG metrics. Further ablation studies show that soft-aligned knowledge fusion improves the recognition of infrequent predicates, while the commonsense-aware subject-object indicator and knowledge-driven scoring constraint reduce relation predictions that conflict with entity roles, object affordances, and spatial semantics.

**Key words**　scene graph generation; commonsense knowledge graph; machine learning; relationship reasoning; Transformer

**摘　要**　场景图生成（scene graph generation, SGG）作为连接低级视觉感知和高级语义推理的桥梁，在各种视觉语言任务中扮演着重要角色。然而，现有的专用场景图生成模型常因缺乏语义约束机制而产生不合理的关系预测，存在逻辑幻觉。目前，尽管多模态大模型（MLLMs）具备丰富的常识，但直接将其应用于场景图生成任务时，其固有的黑盒特性与严重的幻觉问题导致关系推理结果缺乏可靠性与可解释性，

**收稿日期**：2026-06-06；**修回日期**：2026-07-17
**基金项目**：国家自然科学基金项目（62576206）；中国商飞—上海交大大飞机研究院（上海商用飞机系统工程科创中心）联合研究基金
This work was supported by the National Natural Science Foundation of China (62576206) and the Joint Research Fund of COMAC-SJTU Large Aircraft Research Institute (Shanghai Center for Systems Engineering of Commercial Aircraft).
**通信作者**：孙仕亮（shiliangsun@gmail.com）

且其参数规模巨大、计算开销高昂，难以满足实际部署需求。为此，提出了 KETR（commonsense knowledge embedding Transformer），一种嵌入常识知识的端到端场景图生成模型。该模型通过多尺度视觉特征融合编码器提取图像全局与局部特征，利用基于可学习查询的并行解码器实现实体检测。核心创新在于提出软对齐知识融合机制，通过视觉特征与外部常识知识图谱嵌入空间的概率映射动态聚合知识信息，同时引入常识感知的主宾语指示器与知识驱动评分约束，将 TransE 模型的评分函数作为正则化项融入联合训练损失，有助于提升关系推理的语义一致性。在 Visual Genome 和 Open Images V6 两个主流基准数据集上的广泛实验表明，KETR 在多项核心指标上取得了大幅度的性能提升。进一步的消融实验表明，软对齐知识融合有助于改善低频关系的识别能力，常识感知主宾语指示器和知识驱动评分约束能够减少与实体角色、物体可供性及空间语义明显冲突的不合理关系预测。

**关键词** 场景图生成；常识知识图谱；机器学习；关系推理；Transformer

**中图法分类号** TP311.13；TP309

**DOI:** 10.7544/issn1000-1239.202660459 **CSTR:** 32373.14.issn1000-1239.202660459

场景图生成（scene graph generation, SGG）是结构化视觉理解中的一项基础任务，旨在将图像表示为一个由实体节点和语义关系边组成的图结构[1-2]。这种三元组（Subject, Predicate, Object）形式的表示提供了丰富的视觉内容抽象，已广泛应用于图像描述生成（Image Captioning）、视觉问答（VQA）以及视觉常识推理（VCR）等下游任务中[3-5]。

目前的 SGG 模型通常采用两阶段或端到端的流程。两阶段模型即先基于目标检测器提取特征再进行关系分类，端到端流程则多为基于 Transformer 的预测框架[6-7]。尽管这些方法在全局上下文建模方面取得了进展，但仍面临严峻的挑战。首先是严重的长尾分布问题：场景图数据集中高频谓词（如“on”“has”）占据主导，导致模型的预测结果严重偏向这些无信息量的关系，掩盖了更具语义价值的低频谓词。其次，现有模型普遍缺乏对外部知识的有效利用，这使得模型在面对数据集中未覆盖或标注稀疏的关系类型时，会产生违反常识的推理结果。针对上述问题，本文的主要贡献如下：

1）提出了 KETR（commonsense knowledge embedding Transformer），一个使模型嵌入常识以符合常识的三元组推理的新型 SGG 架构，实现了视觉特征提取与外部常识推理的端到端整合。

2）设计了软对齐知识融合机制，利用视觉-知识嵌入空间的概率映射取代传统的硬匹配，有效解决了实体映射过程中的灵活性不足问题。

3）引入了常识感知的主语/宾语指示器与评分约束机制，通过在 Transformer 解码过程中注入语义概念并结合常识评分函数，显著提升了关系预测的准确性。

## 1 相关工作

SGG 最初由 Lu 等人提出[1]，此后，SGG 作为连接低级视觉感知和高级语义推理的桥梁，受到了越来越多的关注[2,8]。该领域的早期研究，例如视觉关系检测（VRD）[1]，主要集中于预测检测到的物体候选区域之间的关系。Xu 等人[3] 提出了基于 GNN 的模型 MSGG，建立了一种通过在实体特征和关系特征之间传递信息来进行迭代细化的机制，开创了一系列基于图传播模型的感知方法。MotifNet[4] 强调了实体上下文在关系预测中的重要性，并采用分层 LSTM 来递归顺序预测实体和关系。此后，Graph R-CNN[9] 和 VCTree[5] 等模型也得到了进一步发展。此外，无偏 SGG 方法（如 TDE[10]、BGNN[11]、CogTree[12]）致力于解决长尾分布偏差问题，Zhang 等人[13] 提出了图对比损失来改善场景图解析，Chiou 等人[14] 探索了从有偏场景图中恢复无偏表示的方法。2025 年，Sun 等人[15] 从因果推理的角度出发，将传统 SGG 的因果链结构重构为逆因果结构，通过主动逆向估计（active reverse estimation）与最大信息采样（maximum information sampling）消除检测器输入与最终预测之间的虚假关联，在多个主流基准上取得了当前最优的平均召回率。Li 等人[16] 则将去偏思想推广到视频 SGG 领域，提出了视觉-语义双去偏框架 VISA，通过记忆增强的时序融合和迭代式三元组语义集成，显著提升了视频 SGG 中尾部关系的召回率。尽管上述两阶段方法取得了一定进展，但它们存在 2 个根本性局限：其一，两阶段流程中目标检测与关系预测相互割裂，检测阶段产生的定位误差和类别噪声会不可逆地传播至

关系分类阶段，形成级联错误；其二，基于 GNN/LSTM 的消息传递机制在建模远距离实体对之间的关系时受限于固定的图拓扑或序列顺序，难以捕获跨区域的全局语义依赖。

随着 Transformer 架构的兴起，研究者开始探索端到端的场景图生成范式[6-7]。例如，SGTR[6] 和 RelTR[7] 利用 Transformer 的自注意力机制捕获全局上下文，绕过了传统的实体检测步骤。EGTR[17] 进一步优化了端到端架构，Li 等人[18] 在 SGTR 基础上提出 SGTR+，进一步完善了端到端 Transformer 场景图生成。此外，Suhail 等人[19] 引入了基于能量的学习框架用于场景图生成，Dong 等人[20] 利用堆叠式混合注意力与群组协作学习来缓解偏差问题，Desai 等人[21] 探索了从自然语言监督中学习生成场景图的路径。然而这些端到端方法的关系预测仍然完全依赖于训练集的统计偏好，缺乏从外部知识源中获取支撑。

为了弥补视觉信息的不足，研究者开始探索将外部常识知识融入 SGG 推理过程。KERN[22] 将统计先验知识（如物体对出现的频率）融入图卷积神经网络中，建模物体间的共现规律。GB-Net[23] 则尝试将外部知识图谱的三元组结构直接桥接到场景图推理中。Jiang 等人[24] 提出了一种常识验证模型来改进关系建模。近年来，随着多模态大语言模型（MLLMs）的迅猛发展，研究者开始尝试将大模型（LLM）引入 SGG 任务。LLM4SGG[25] 利用 LLM 的语言理解与思维链（chain-of-thought）推理能力，从图像描述中提取更细粒度的三元组标注，以增强弱监督 SGG 的训练信号。Pix2SG[26] 则将 SGG 建模为图像到文本的序列生成问题，借助视觉-语言预训练模型（VLM）以自回归方式直接输出三元组序列，实现了开放词汇的 SGG。SDSGG[27] 利用 LLM 扮演不同角色（如生物学家、工程师）从多视角分析场景的描述性特征，并据此动态调整文本分类器权重，以适应不同场景上下文。Chen 等人[28] 进一步提出了基于多智能体 LLM 系统的场景图推理框架 SG$^2$，通过推理者与检索者 2 个模块的迭代协作，在图结构信息上进行逐步推理。RelationLMM[29] 将场景图生成、人-物交互检测等多种视觉关系任务统一到单一的多模态大模型框架中，通过将复杂任务拆解为简单子任务并以自然语言形式表达，探索了 1 个模型处理所有任务的通用范式。Ozsoy 等人[30] 提出了面向多模态场景图生成的大视觉-语言模型 MM2SG，验证了多模态输入（RGB-D、语音、文本等）对场景图推理的增益。

然而，上述知识增强方法仍存在一些不足。对于传统的知识注入方法：1）知识注入方式依赖基于精确类别名称的硬匹配来检索知识节点，一旦视觉检测器的类别预测出现偏差或遇到近义实体，便无法获取有效的知识支撑；2）知识与视觉特征的融合深度不足，多数方法仅在最终分类阶段简单拼接或相加知识特征，未能在特征空间中实现视觉语义与知识语义的对齐，导致知识信息在多层推理过程中逐渐被稀释。对于基于大模型的方法，尽管 MLLMs 蕴含丰富的常识知识，但将其直接应用于 SGG 任务仍面临多重挑战：1）大模型固有的黑盒特性使其推理过程缺乏可解释性，关系预测结果难以追溯和验证，且大模型普遍存在的幻觉（Hallucination）问题会导致生成不存在的物体或虚假的关系三元组，严重影响场景图的可靠性[31]；2）大模型的参数规模通常达到数十亿甚至上百亿级别，推理时需要消耗大量的 GPU 显存与计算资源，单张图像的处理延迟远高于专用轻量化模型，难以满足实时性要求较高的下游应用（如自动驾驶、机器人导航）的部署需求；3）基于序列生成的范式将结构化的图预测退化为线性文本输出，丢失了实体间的拓扑结构信息，且输出格式的不稳定性增加了后处理的复杂度。

针对上述问题，本研究提出了 KETR 模型，旨在通过一种软对齐机制动态聚合知识嵌入，并利用混合架构保留局部特征细节。此外，我们提出的常识感知的主语/宾语指示器让实体在推理过程中能结合知识编码，主动识别自己是主语还是宾语角色，并捕获多对实体间的上下文关联，从而避免信息割裂，在维持高效推理的同时实现了更准确、更具逻辑一致性的 SGG。

## 2 KETR：融入常识的端到端 SGG 模型

本文提出了一种嵌入常识知识的端到端 SGG 框架 KETR。本节将详细阐述 KETR 的架构设计，重点介绍如何通过 Transformer 结构实现视觉特征与常识知识的深度耦合。KETR 摒弃了传统 SGG 模型中孤立的特征提取与关系分类步骤，通过构建端到端的推理路径，将外部常识图谱的逻辑约束显式地注入到关系推理过程中。模型的整体框架如图 1 所示。KETR 模型由 3 个阶段构成：多尺度视觉特征融合与实体并行解码、软对齐知识融合和常识知识驱动的关系推理。

### 2.1 多尺度视觉特征融合与实体并行解码

KETR 首先利用深度卷积神经网络（CNN）作为主干网络提取图像的基础特征[32]。给定输入图像

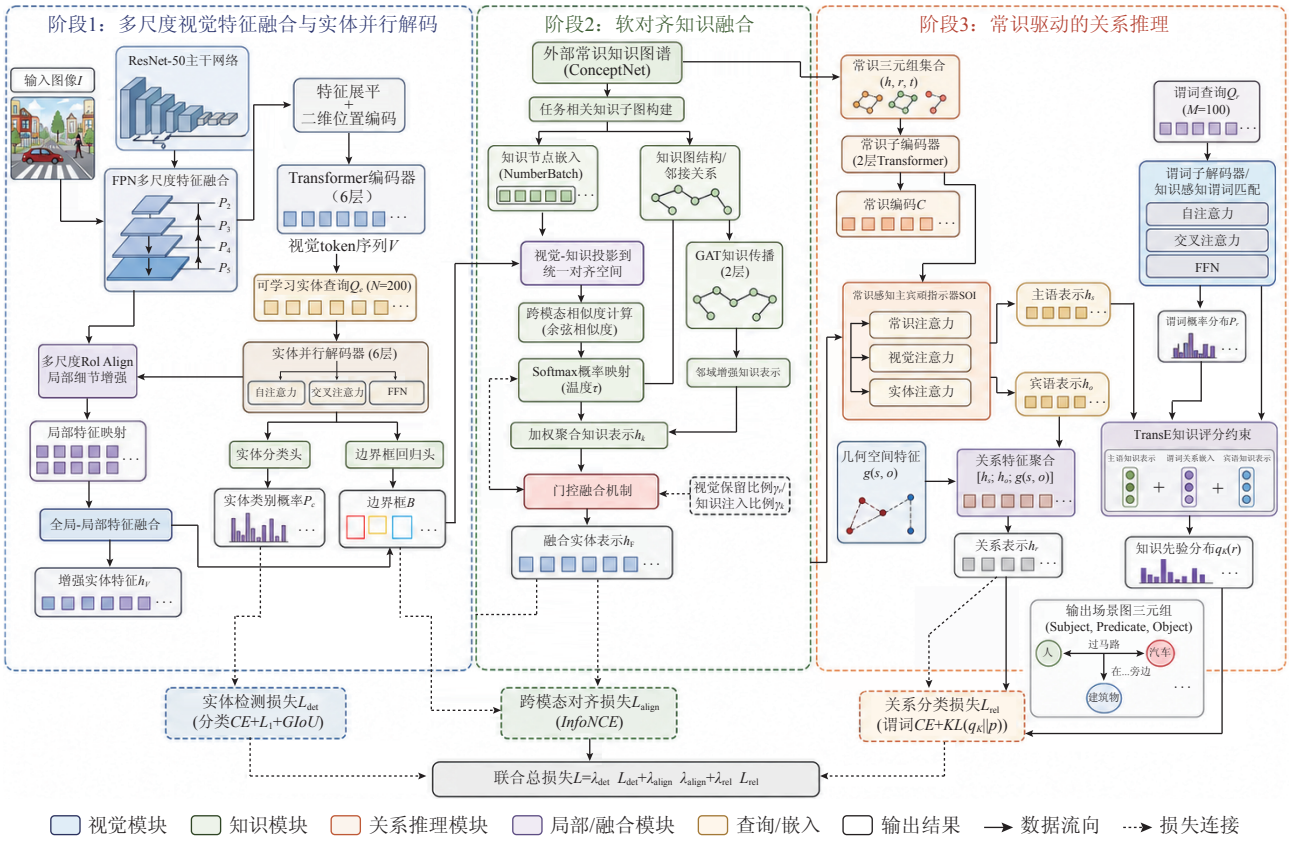



Fig. 1 Overall framework of KETR model

图 1 KETR 模型整体框架

$\boldsymbol{I}\in\mathbb{R}^{3\times H\times W}$，ResNet-50 主干网络输出第 $l$ 层的特征图 $\boldsymbol{C}_l\in\mathbb{R}^{C_l\times H_l\times W_l}$（$l\in\{2,3,4,5\}$）。为了捕获图像中不同尺寸物体的细节，我们利用特征金字塔结构（FPN）[33] 对多层级特征图进行自顶向下的融合与横向连接。具体而言，FPN 通过 $1\times1$ 卷积将各层级通道数统一为 $d_{\text{model}}$ 维，并通过上采样与逐元素相加实现跨尺度特征聚合：

$$\boldsymbol{P}_l=Conv_{1\times1}(\boldsymbol{C}_l)+Up(\boldsymbol{P}_{l+1}),\ l\in\{2,3,4,5\},\tag{1}$$

其中 $\boldsymbol{P}_l\in\mathbb{R}^{d_{\text{model}}\times H_l\times W_l}$ 为第 $l$ 层的金字塔特征。随后，为了建模图像内部的长程依赖关系，我们将各层级特征图 $\boldsymbol{P}_l$ 经过 $3\times3$ 卷积后展平为一维特征序列 $\boldsymbol{F}_l\in\mathbb{R}^{H_l\times W_l\times d_{\text{model}}}$，并将所有层级的特征序列拼接为统一的视觉 token 序列 $\boldsymbol{F}_{\text{vis}}=(\boldsymbol{F}_2;\boldsymbol{F}_3;\boldsymbol{F}_4;\boldsymbol{F}_5)\in\mathbb{R}^{L\times d_{\text{model}}}$，其中 $L=\sum_l H_l\times W_l$ 为 token 总数。为了保留空间位置信息，我们为每个 token 叠加可学习的二维正编码。对于位于特征图 $x,y$ 坐标处的 token，其位置编码定义为：

$$PE(x,y)=\left[\sin\left(x/\omega^0\right),\cos\left(x/\omega^0\right),\cdots,\sin\left(y/\omega^{D/2-1}\right),\cos\left(y/\omega^{D/2-1}\right)\right],\tag{2}$$

其中 $\omega^k=10\,000^{2k/d_{\text{model}}}$。增强位置信息后的特征序列输入到 Transformer 编码器中[34]。编码器包含 $N_e=6$ 层，每层由多头自注意力层（multi-head self-attention, MHSA）和前馈网络层（FFN）组成，并辅以残差连接与层归一化（LN）：

$$\hat{\boldsymbol{F}}^{(n)}=LN\left(\boldsymbol{F}^{(n-1)}+MHSA\left(\boldsymbol{F}^{(n-1)}\right)\right),\tag{3}$$

$$\boldsymbol{F}^{(n)}=LN\left(\hat{\boldsymbol{F}}^{(n)}+FFN\left(\hat{\boldsymbol{F}}^{(n)}\right)\right),\tag{4}$$

其中 $n=1,\cdots,N_e$，$\boldsymbol{F}^{(0)}=\boldsymbol{F}_{\text{vis}}$。每个注意力头的计算过程为 $head_h=softmax\left(\boldsymbol{Q}_h\boldsymbol{K}_h^{\mathrm{T}}/\sqrt{d_k}\right)\boldsymbol{V}_h$，$h=1,\cdots,H$，其中 $H=8$ 为注意力头数，$d_k=d_{\text{model}}/H$。多头注意力的输出通过线性投影合并为 $MHSA(\boldsymbol{F})=Concat(head_1,\cdots,head_H)\boldsymbol{W}_{\text{O}}$。

在实体检测与表征阶段，我们设计了一种基于可学习查询（learnable query）向量的并行解码模块[35]，而非传统的滑动窗口或区域提议方法。该模块初始化了一组固定数量 $N$ 的高维查询向量 $(\boldsymbol{q}_i)_{i=1}^{N}\in\mathbb{R}^{N\times d_{\text{model}}}$，每个向量在空间和语义维度上进行主动的实体检测。解码器同样包含 $N_d=6$ 层，每层由 3 个子层顺序组成。首先，通过多头自注意力机制（self-attention），不同查询之间可以进行信息交换，以避免多个查询关注同一实体：

$$\hat{\boldsymbol{Q}}^{(n)}=LN\left(\boldsymbol{Q}^{(n-1)}+MHSA\left(\boldsymbol{Q}^{(n-1)}\right)\right)。\tag{5}$$

随后，通过交叉注意力（cross-attention），每个查

询向量$\boldsymbol{q}_i$从编码器输出的视觉特征序列$\boldsymbol{F}^{(N_e)}$中选择性地聚合与其对应实体最相关的视觉信息：

$$CrossAttn(\boldsymbol{q}_i,\boldsymbol{F})=softmax\left(\frac{\boldsymbol{q}_i\boldsymbol{W}_{\text{que}}\left(\boldsymbol{F}\boldsymbol{W}_{\text{key}}\right)^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{F}\boldsymbol{W}_{\text{val}}, \tag{6}$$

其中$\boldsymbol{W}_{\text{que}},\boldsymbol{W}_{\text{key}},\boldsymbol{W}_{\text{val}}\in\mathbb{R}^{d_{\text{model}}\times d_k}$为可学习的投影矩阵。交叉注意力之后跟随一个前馈网络子层 FFN。经过$N_{\text{d}}$层解码器的迭代细化，每个查询向量最终演变为一个高维实体嵌入向量$\boldsymbol{e}_i\in\mathbb{R}^{d_{\text{model}}}$。该向量不仅凝聚了目标的视觉外观特征，还隐含了其空间位置信息。这些嵌入向量随后通过 2 个并行的预测分支。分类分支通过线性层输出实体属于预定义 $C$ 个类别（含“无物体”背景类）的概率分布：

$$P_{\text{cls}}(c\mid\boldsymbol{e}_i)=softmax(\boldsymbol{W}_{\text{cls}}\boldsymbol{e}_i+\boldsymbol{b}_{\text{cls}})。 \tag{7}$$

回归分支则通过 3 层多层感知机（MLP）预测归一化的物体边界框坐标$\hat{\boldsymbol{b}}_i=\left(\hat{x}_c,\hat{y}_c,\hat{w},\hat{h}\right)$，其中$(\hat{x}_c,\hat{y}_c)$为中心点坐标，$\hat{w},\hat{h}$为宽和高，均归一化至 [0, 1] 区间：

$$\hat{\boldsymbol{b}}_i=\sigma(\text{MLP}_{\text{box}}(\boldsymbol{e}_i)), \tag{8}$$

其中$\sigma(\cdot)$为 sigmoid 激活函数。

为了弥补全局特征在精细纹理描述上的不足，模型进一步引入了局部特征增强路径。针对解码器预测出的每个实体边界框$\hat{\boldsymbol{b}}_i$，我们利用多尺度 RoI 对齐[36]技术从 CNN 主干网络的多尺度特征图$\{\boldsymbol{C}_l\}_{l=2}^{5}$中精确裁剪出对应的局部特征块$\boldsymbol{r}_i\in\mathbb{R}^{d_r}$。与标准 RoI 对齐不同，多尺度 RoI 对齐为每个采样点学习额外的偏移量$\Delta p_k$，使其能够自适应地关注物体的非规则形状区域。提取的局部特征通过轻量化$1\times1$卷积网络映射为$d_{\text{model}}$维后，与解码器输出的实体嵌入向量进行拼接融合：

$$\boldsymbol{e}_i^{\text{aug}}=LN\left(\boldsymbol{W}_{\text{proj}}[\boldsymbol{e}_i\|\boldsymbol{r}_i]+\boldsymbol{b}_{\text{proj}}\right), \tag{9}$$

其中$[\cdot\|\cdot]$表示向量拼接，$\boldsymbol{W}_{\text{proj}}\in\mathbb{R}^{d_{\text{model}}\times(d_{\text{model}}+d_r)}$为降维投影。这种全局-局部的双重特征表示策略确保了模型在利用全局上下文进行推理的同时，不丢失对小物体和细粒度区分至关重要的局部纹理细节。

### 2.2 软对齐知识融合

在完成实体的视觉特征提取后，为了使模型具备跨模态的常识推理能力，本研究设计了一种软对齐（soft alignment）知识融合机制，旨在将视觉实体特征动态地映射到知识图谱的语义空间中。传统的硬匹配方法往往依赖于分类头输出的单一实体 ID $\hat{c}_i=\arg\max P_{\text{cls}}(\cdot\mid\boldsymbol{e}_i)$来检索知识库中的对应节点嵌入$\boldsymbol{k}_{\hat{c}_i}$，这种方式在面临视觉检测模糊或类别粒度不匹配时，极易引入由于分类错误导致的噪声传播。

本文采用 ConceptNet 作为外部常识实体的基础表示。考虑到通用知识图谱包含大量与视觉场景无关的抽象概念和语言节点，我们首先根据场景图数据集的实体类别构建任务相关知识子图。设数据集实体类别集合为$C=\{c_1,c_2,\cdots,c_C\}$，对于每个类别名称，首先执行小写化、词形还原和复合词规范化，以减少单复数、时态以及多词短语书写差异带来的影响。随后，利用 WordNet 获取其主要词义对应的同义词和上位概念，并在 ConceptNet NumberBatch 词表中检索相应节点。由此得到该类别对应的知识候选集合：

$$\mathcal{A}(c)=\mathcal{A}_{\text{exact}}(c)\cup\mathcal{A}_{\text{syn}}(c)\cup\mathcal{A}_{\text{hyper}}(c), \tag{10}$$

其中，$\mathcal{A}_{\text{exact}}(c)$表示与规范化类别名称直接对应的知识节点，$\mathcal{A}_{\text{syn}}(c)$表示同义概念节点，$\mathcal{A}_{\text{hyper}}(c)$表示用于词汇表外回退的上位概念节点。所有类别候选节点及其与视觉语义相关的邻域共同组成任务相关知识节点集合：

$$\mathcal{V}_{\text{key}}=\bigcup_{c\in C}\mathcal{A}(c),\quad \mathcal{G}_{\text{key}}=\left(\mathcal{V}_{\text{key}},\mathcal{E}_{\text{key}}\right)。 \tag{11}$$

因此，后文所述的 $M$ 表示任务相关知识节点的数量，即 $M=\left|\mathcal{V}_{\text{key}}\right|$。类别集合 $C$ 与知识节点集合$\mathcal{V}_{\text{key}}$之间并非一一对应关系：一个视觉类别可以对应多个同义或不同粒度的知识节点，多个视觉类别也可以共享相同的上位知识概念。软对齐机制的目标正是在这种多对多候选关系中，根据当前实体的视觉特征和图像上下文动态确定不同知识概念的贡献，而不是预先固定唯一映射结果。

设任务相关知识子图中共有 $M$ 个实体节点，其预训练嵌入构成知识矩阵$\boldsymbol{\mathcal{K}}=(\boldsymbol{k}_1,\boldsymbol{k}_2,\cdots,\boldsymbol{k}_M)^{\text{T}}\in\mathbb{R}^{M\times d_k}$。对于增强后的视觉实体特征$\boldsymbol{e}_i^{\text{aug}}$（简记为$\boldsymbol{v}_i$），我们首先通过可学习的线性映射将 2 个模态的特征投影至统一的$d_a$维对齐空间：

$$\tilde{\boldsymbol{v}}_i=\boldsymbol{W}_a^{(v)}\boldsymbol{v}_i,\quad \tilde{\boldsymbol{k}}_j=\boldsymbol{W}_a^{(k)}\boldsymbol{k}_j, \tag{12}$$

其中$\boldsymbol{W}_a^{(v)}\in\mathbb{R}^{d_a\times d_{\text{model}}}$，$\boldsymbol{W}_a^{(k)}\in\mathbb{R}^{d_a\times d_k}$为投影矩阵。在对齐空间中，模型计算视觉特征与每个知识节点之间的余弦相似度：

$$s_{ij}=\frac{\tilde{\boldsymbol{v}}_i^{\text{T}}\tilde{\boldsymbol{k}}_j}{\left\|\tilde{\boldsymbol{v}}_i\right\|\cdot\left\|\tilde{\boldsymbol{k}}_j\right\|}。 \tag{13}$$

为了精确控制语义对齐的锐度（sharpness），我们引入带有温度系数$\tau$的 Softmax 函数对相似度进行归一化处理，得到软对齐权重向量$\boldsymbol{\alpha}_i=(\alpha_{i1},\alpha_{i2},\cdots,\alpha_{iM})$：

$$\alpha_{ij}=\frac{\exp\left(s_{ij}/\tau\right)}{\sum_{m=1}^{M}\exp(s_{im}/\tau)}。 \tag{14}$$

温度系数$\tau$在此起到关键的调节作用：较小的 $\tau$ 值(如 $\tau=0.05$)使概率分布更加尖锐，迫使模型聚焦于最相关的少数知识条目，趋近于硬匹配；较大的 $\tau$ 值则产生更平滑的分布，允许模型考虑更广泛的语义关联。利用该软权重对知识嵌入进行加权聚合，得到该视觉实体在知识空间中的表示：

$$\tilde{\boldsymbol{k}}_i^{\text{agg}}=\sum_{j=1}^{M}\alpha_{ij}\boldsymbol{k}_j。\tag{15}$$

该概率聚合机制能够显式处理视觉类别与知识实体之间的粒度差异。例如，当视觉特征仅支持较粗粒度的类别判断时，权重可以分布在该类别及其相关上位概念之间；当局部外观和场景上下文能够提供更具体的语义信息时，权重会进一步集中于细粒度知识节点。因此，分类头的单次类别错误不会像硬匹配一样直接决定最终知识表示。跨模态对齐损失从视觉-知识正对和批内负对2个方向约束公共空间，使与当前视觉实体语义一致的节点获得更高相似度。该监督与可学习温度共同避免权重分布过度平滑或错误集中。

在获得跨模态聚合的知识嵌入后，模型通过一种基于门控机制(gated control mechanism)的融合策略实现视觉与常识特征的深度集成。该策略通过可学习的投影矩阵$\boldsymbol{W}_v\in\mathbb{R}^{d_{\text{model}}\times d_{\text{model}}}$和$\boldsymbol{W}_k\in\mathbb{R}^{d_{\text{model}}\times d_k}$将2个模态的特征映射到统一的共享语义空间，并利用一个由sigmoid激活函数生成的标量门控信号 $g_i$ 动态平衡不同特征的权重。门控信号由2个模态特征的拼接向量经线性变换后激活生成：

$$g_i=\sigma\left(\boldsymbol{w}_g^{\mathrm{T}}\left[\boldsymbol{v}_i\|\tilde{\boldsymbol{k}}_i^{\text{agg}}\right]+b_g\right),\tag{16}$$

其中 $\boldsymbol{w}_g\in\mathbb{R}^{d_{\text{model}}+d_k}$，$b_g$ 为可学习参数。最终的融合特征为：

$$\boldsymbol{f}_i^{\text{fused}}=g_i\cdot(\boldsymbol{W}_v\boldsymbol{v}_i)+(1-g_i)\cdot\left(\boldsymbol{W}_k\tilde{\boldsymbol{k}}_i^{\text{agg}}\right),\tag{17}$$

该式允许模型根据视觉信号的置信度自动调节知识信息的注入比例：当视觉特征清晰且分类置信度高时，$g_i$趋近于1，模型主要依赖视觉信息；当视觉信号模糊(如遮挡、运动模糊)时，$g_i$减小，更多地依赖常识知识进行补偿推理。外部知识图谱不可避免地包含语义宽泛、上下文不匹配或置信度不足的节点和边，因此本文不将聚合知识表示作为视觉实体表示的直接替代，而是通过门控残差方式进行融合。视觉路径在融合过程中始终被保留，知识表示仅作为可调节的补充信息。当知识表示与视觉证据及任务监督一致时，模型可以提高知识分量；当候选知识节点含有噪声或聚合结果与视觉特征不一致时，训练得到的门控向量会降低该知识分量的影响。因此，知识图谱中的单个错误节点不会以硬规则的形式直接决定实体表示。对于无法通过直接名称、同义词或上位概念找到有效候选节点的词汇表外类别，知识分支使用零向量作为回退表示，此时融合模块保留视觉特征路径，使模型退化为基于视觉表示的关系推理。该回退策略保证了知识覆盖不足时模型仍具有完整的预测能力。

为了捕获知识图谱中的结构化上下文信息，模型进一步利用图注意力网络(GAT)[37]对实体在知识图谱中的邻居节点进行信息聚合。设知识图谱中节点$j$的邻居集合为 $\mathcal{N}(j)$，GAT 的注意力系数通过如下方式计算：

$$e_{jl}=LeakyReLU\left(\boldsymbol{a}^{\mathrm{T}}\left[\boldsymbol{W}_g\boldsymbol{k}_j\|\boldsymbol{W}_g\boldsymbol{k}_l\right]\right),l\in\mathcal{N}(j),\tag{18}$$

$$\beta_{jl}=\frac{\exp\left(e_{jl}\right)}{\sum\limits_{l'\in\mathcal{N}(j)}\exp\left(e_{jl'}\right)},\tag{19}$$

其中$\boldsymbol{W}_g\in\mathbb{R}^{d_g\times d_k}$为共享的线性变换矩阵，$\boldsymbol{a}\in\mathbb{R}^{2d_g}$为注意力向量。聚合后的知识节点表示更新为：

$$\boldsymbol{k}'_j=\mathrm{ELU}\left(\sum_{l\in\mathcal{N}(j)}\beta_{jl}\boldsymbol{W}_g\boldsymbol{k}_l\right)。\tag{20}$$

我们采用 $H_g=4$ 个注意力头并行计算，将结果拼接后通过线性变换得到最终的邻域增强知识表示$\boldsymbol{k}_j^{\text{GAT}}=\boldsymbol{W}_o^{(g)}\|_{h=1}^{H_g}\boldsymbol{k}'_{j,h}\in\mathbb{R}^{d_k}$。GAT 共堆叠2层，使每个知识节点能够感知其二跳邻域范围内的结构化常识信息，如物体用途或常见空间关系等，为后续的关系推理提供坚实的语义支撑。最终的融合特征 $\boldsymbol{f}_i^{\text{fused}}$ 中的知识分量亦经过 GAT 增强。知识子图中的邻接边仅用于提供候选结构上下文，并不被视为等可信的确定性规则。GAT 根据中心节点和邻居节点的当前表示计算归一化注意力权重，使不同邻居对更新结果产生不同影响。与均匀图卷积相比，该机制能够降低与当前实体语义相关性较弱的邻居所造成的干扰。此外，节点自身表示通过残差路径被保留，因此邻域中的噪声关系不会完全覆盖原始 NumberBatch 语义表示。本文仅堆叠2层 GAT，以利用局部常识上下文，同时避免过深传播将远距离无关节点引入当前实体表示。

### 2.3 常识驱动的关系推理

在完成实体特征与常识知识的融合后，KETR 模型进入核心的关系推理阶段，即通过常识感知的主语与宾语指示器(subject-object indicator, SOI)以及知识驱动的关系预测头来生成最终的三元组表示。该

机制的设计初衷是让实体能够自发地感知其在潜在关系中所扮演的角色(主语或宾语)，并利用常识知识库中的先验规则对谓词的分布进行约束，从而避免传统模型中主宾语分类与关系预测相互割裂的问题。

关系推理的第一步是构建常识知识的结构化向量表示。我们设计了一个常识子编码器，以知识图谱中的三元组集合 $\mathcal{T}=\{(h,r,t)\}$ 为输入(其中 $h,t$ 为实体，$r$ 为关系)，将其嵌入为可操作的语义编码。具体而言，对于知识图谱中的每个三元组$(h,r,t)$，我们将头实体嵌入$\boldsymbol{k}_h$、关系嵌入 $\boldsymbol{r}$ 和尾实体嵌入 $\boldsymbol{k}_t$ 拼接后通过多层 Transformer 编码器进行上下文建模：

$$\boldsymbol{z}_{(h,r,t)}=TransEnc([\boldsymbol{k}_h\|\boldsymbol{r}\|\boldsymbol{k}_t])\in\mathbb{R}^{d_{\text{model}}}。\tag{21}$$

子编码器包含 2 层 Transformer，使每个三元组嵌入能够充分融合头尾实体与关系的语义交互。最终的常识编码 $\boldsymbol{Z}=\left(z_1,z_2,\cdots,z_{|\mathcal{T}|}\right)$ 将被用于后续的主宾语指示器。

指示器为主语和宾语分别维护一组查询向量 $(\boldsymbol{s}_i)_{i=1}^N$ 和$(\boldsymbol{o}_i)_{i=1}^N$，初始化为实体融合特征 $\boldsymbol{f}_i^{\text{fused}}$ 的线性变换。指示器通过以下 3 种注意力机制进行迭代更新：

1)常识自注意力。该阶段以常识编码 $\boldsymbol{Z}$ 为键值对，将实体相关的常识信息注入到主语/宾语查询中，赋予其明确的语义概念。对于主语查询 $\boldsymbol{s}_i$，常识注意力的计算为：

$$\boldsymbol{s}_i^{(1)}=\boldsymbol{s}_i+MHCA(\boldsymbol{s}_i,\boldsymbol{Z},\boldsymbol{Z}),\tag{22}$$

其中 $\text{MHCA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$ 表示多头交叉注意力操作。宾语查询 $\boldsymbol{o}_i$ 的更新方式相同。此阶段使主语和宾语查询获得与自身语义最相关的常识三元组信息，例如“人”的主语查询可能与“人-骑-马”、“人-吃-食物”等知识三元组产生较高的注意力权重。同时，不同主语/宾语查询之间通过自注意力进行信息交换，捕获多对实体间的上下文关联：

$$\hat{\boldsymbol{s}}_i^{(1)}=LN\left(\boldsymbol{s}_i^{(1)}+\text{MHSA}\left(\left\{\boldsymbol{s}_j^{(1)}\right\}_{j=1}^N\right)_i\right)。\tag{23}$$

2)视觉注意力。该阶段使主语/宾语查询与编码器输出的视觉特征 $\boldsymbol{F}^{(N_e)}$ 进行交叉注意力交互，从而专注于图像中主语或宾语特定的物理区域：

$$\hat{\boldsymbol{s}}_i^{(2)}=LN\left(\hat{\boldsymbol{s}}_i^{(1)}+\text{MHCA}\left(\hat{\boldsymbol{s}}_i^{(1)},\boldsymbol{F}^{(N_e)},\boldsymbol{F}^{(N_e)}\right)\right)。\tag{24}$$

3)实体注意力。主语/宾语查询与并行解码器输出的全部实体嵌入 $\left(\boldsymbol{e}_i^{\text{aug}}\right)_{i=1}^N$ 进行交叉注意力交互，整合丰富的实体级信息：

$$\hat{\boldsymbol{s}}_i^{(3)}=LN\left(\hat{\boldsymbol{s}}_i^{(2)}+\text{MHCA}\left(\hat{\boldsymbol{s}}_i^{(2)},\boldsymbol{E}^{\text{aug}},\boldsymbol{E}^{\text{aug}}\right)\right),\tag{25}$$

其中 $\boldsymbol{E}^{\text{aug}}=\left(\boldsymbol{e}_1^{\text{aug}},\cdots,\boldsymbol{e}_N^{\text{aug}}\right)^{\text{T}}$。经过 3 种注意力运算之后，每个指示器子层还跟随一个 FFN 子层。宾语查询 $\boldsymbol{o}_i$ 经过完全对称的三阶段更新，最终得到 $\hat{\boldsymbol{o}}_i^{(3)}$。指示器共堆叠 $N_s=3$ 层以充分迭代。

对于任意一对候选实体$(i,j)$，KETR 构建了一个多维特征交互空间来生成关系表示。首先，我们提取该实体对的几何空间特征向量，编码其相对位置信息：

$$\boldsymbol{p}_{ij}=\text{MLP}_{\text{sp}}\left(\left[\frac{\Delta x_{ij}}{w_i},\frac{\Delta y_{ij}}{h_i},\log\frac{w_j}{w_i},\log\frac{h_j}{h_i},\text{IoU}\left(b_i,b_j\right)\right]\right),\tag{26}$$

其中$\Delta x_{ij}=x_j^c-x_i^c$，$\Delta y_{ij}=y_j^c-y_i^c$ 为中心点坐标差，$w_i,h_i$ 为主语边界框的宽和高，$\text{IoU}\left(b_i,b_j\right)$ 为 2 个框的交并比。$\text{MLP}_{\text{spa}}$ 为 2 层感知机，将 5 维几何向量映射为 $d_{\text{model}}$ 维。关系表征通过级联主语指示器输出 $\hat{\boldsymbol{s}}_i^{(3)}$、宾语指示器输出 $\hat{\boldsymbol{o}}_j^{(3)}$ 以及空间特征 $\boldsymbol{p}_{ij}$ 构成：

$$\boldsymbol{u}_{ij}=\boldsymbol{W}_r\left[\hat{\boldsymbol{s}}_i^{(3)}\left\|\hat{\boldsymbol{o}}_j^{(3)}\right\|\boldsymbol{p}_{ij}\right]+\boldsymbol{b}_r,\tag{27}$$

其中 $\boldsymbol{W}_r\in\mathbb{R}^{d_{\text{model}}\times 3d_{\text{model}}}$ 为关系投影矩阵。

需要说明的是，KETR 并不要求场景图数据集中的谓词类别与 ConceptNet 的关系类型建立严格的一一对应映射。Visual Genome(VG)和 Open Images V6(OIV6)中的谓词通常具有更细粒度的视觉含义，例如“wearing”“riding”和“standing on”，而通用知识图谱中的关系类型更加抽象。若强制进行离散关系映射，容易造成语义压缩和错误归并。因此，KETR 始终以数据集定义的谓词集合 $\mathcal{R}=\{r_1,r_2,\cdots,r_R\}$ 作为最终预测空间，并为每个谓词维护独立的可学习关系表示 $\boldsymbol{e}_r^R$。关系表示由规范化后的谓词名称或短语语义表示初始化；对于多词谓词，采用各词语义向量的组合表示；当谓词短语无法在外部词表中直接获得表示时，其关系向量作为普通可学习参数由数据集监督进行优化。外部知识图谱中的关系和三元组用于构建常识子编码，向主语和宾语查询提供物体属性、用途、行为能力及空间关联等上下文信息；数据集谓词表示则在关系交叉熵监督和知识空间正则化下进行学习。两者在连续语义空间中发生交互，而不是通过离散标签进行强制替换。由此，即使 ConceptNet 中缺少与某个视觉谓词完全同名的关系，模型仍能依靠视觉关系特征和训练集标注学习该谓词。

为了提升谓词预测的逻辑一致性，关系特征 $\boldsymbol{u}_{ij}$ 会与常识编码 $\boldsymbol{Z}$ 共同作用以更新谓词查询。具体而言，我们初始化一组谓词查询 $\{\boldsymbol{d}_c\}_{c=1}^R$($R$ 为谓词类别数)，通过谓词子解码器使 $\boldsymbol{u}_{ij}$ 作为条件对谓词查询进行更新：

$$\hat{\boldsymbol{d}}_c^{(ij)}=TransDec\left(\boldsymbol{d}_c,\left[\boldsymbol{u}_{ij}\|\boldsymbol{Z}\right]\right)。\tag{28}$$

最终的谓词分类概率通过 $\boldsymbol{u}_{ij}$ 与更新后的谓词

查询的点积注意力计算：

$$P_{\text{rel}}(r=c\mid i,j)=softmax\left(\frac{\boldsymbol{u}_{ij}^{\mathrm{T}}\hat{\boldsymbol{d}}_c^{(ij)}}{\sqrt{d_{\text{model}}}}\right)。\tag{29}$$

此外，基于 TransE [38] 的知识驱动评分约束被引入以确保三元组的语义合理性。TransE 的核心思想是：对于知识图谱中成立的三元组$(h,r,t)$，头实体嵌入加上关系嵌入应当接近尾实体嵌入。对于任意一个预测的三元组，其在知识空间中的能量函数定义为：

$$E(h,r,t)=\|\boldsymbol{k}_h+\boldsymbol{r}_{\text{emb}}-\boldsymbol{k}_t\|_2,\tag{30}$$

其中$\boldsymbol{k}_h,\boldsymbol{k}_t$为主语和宾语对应的知识嵌入，$\boldsymbol{r}_{\text{emb}}$为预测谓词类别对应的关系嵌入。合理的三元组应具有较低的能量值。该能量函数被用于构建知识驱动的评分约束损失，以惩罚那些在常识语义空间中不合理的关系预测（详见 2.4 节）。需要强调的是，该能量函数并不作为硬规则直接屏蔽某些谓词，而是提供连续的知识先验。真实谓词标签对应的交叉熵仍然是主要监督信号，因此当外部知识缺失、噪声较大或视觉证据与常见知识模式冲突时，模型仍可依赖视觉关系特征和数据集标注完成预测。

### 2.4 损失函数

KETR 框架的训练过程是一个多任务协同优化的过程，旨在同时提升实体检测的精确度、视觉-知识空间的对齐度以及三元组推理的逻辑一致性。为此，我们设计了联合损失函数$\mathcal{L}_{\text{total}}$，由实体检测损失$\mathcal{L}_{\text{det}}$、跨模态对齐损失$\mathcal{L}_{\text{align}}$以及知识驱动的关系分类损失$\mathcal{L}_{\text{rel}}$三部分组成。下面逐一详述各项损失的定义与数学形式。

1）实体检测损失。对于实体检测任务，模型遵循集合预测（set prediction）的范式[35]，将$N$个可学习查询的预测结果视为一个无序集合$\hat{\mathcal{Y}}=\{\hat{y}_1,\cdots,\hat{y}_N\}$，其中每个$\hat{y}_i=(\hat{p}_i,\hat{b}_i)$分别表示实体类别概率分布和预测边界框坐标。训练时，首先利用匈牙利算法（Hungarian algorithm）[39]在$Y$与零填充后的真实标注集合$\mathcal{Y}=\{y_1,\cdots,y_N\}$（其中$y_j=(c_j,b_j)$）之间求解最优二部图匹配。具体地，最优排列$\pi^*$通过最小化全局匹配代价得到：

$$\pi^*=\arg\min_{\pi\in\mathcal{S}_N}\sum_{i=1}^{N}C_{\text{match}}\left(\hat{y}_{\pi(i)},y_i\right),\tag{31}$$

其中，$\mathcal{S}_N$为$N$元排列群，单对匹配代价$C_{\text{match}}$定义为分类代价、$L_1$距离代价与 GIoU 代价的加权和：

$$C_{\text{match}}\left(\hat{y}_j,y_i\right)=-\lambda_{\text{cls}}\hat{p}_j(c_i)+\lambda_{L_1}\left\|\hat{b}_j-b_i\right\|_1+\lambda_{\text{giou}}\mathcal{L}_{\text{GIoU}}\left(\hat{b}_j,b_i\right),\tag{32}$$

其中，$\hat{p}_j(c_i)$表示第$j$个预测对真实类别$c_i$的概率，$\lambda_{\text{cls}}$、$\lambda_{L_1}$、$\lambda_{\text{GIoU}}$为各项代价的平衡系数，实验中分别设置$\lambda_{\text{cls}}=2$、$\lambda_{L_1}=5$和$\lambda_{\text{GIoU}}=2$。GIoU 损失[40]定义为：

$$\mathcal{L}_{\text{GIoU}}\left(\hat{b},b\right)=1-\text{IoU}\left(\hat{b},b\right)+\frac{\left|B_c\setminus\left(\hat{b}\cup b\right)\right|}{|B_c|},\tag{33}$$

其中，$B_c$为$b$和$b$的最小闭包矩形。基于最优匹配$\pi^*$，实体检测损失定义为所有匹配对的加权和：

$$\mathcal{L}_{\text{det}}=\sum_{i=1}^{M}[-\log\hat{p}_{\pi^*(i)}(c_i)+\lambda_{L_1}\left\|\hat{b}_{\pi^*(i)}-b_i\right\|_1+\lambda_{\text{GIoU}}\mathcal{L}_{\text{GIoU}}\left(\hat{b}_{\pi^*(i)},b_i\right),\tag{34}$$

其中，$M$为图像中真实实体数量。对于背景类（$c_i=\varnothing$），分类损失的权重降低至 0.1 以缓解正负样本不均衡的问题。此外，我们在 Transformer 解码器的每一层辅助输出上均施加检测损失，形成深度监督（deep supervision），以加速训练收敛。

2）跨模态对齐损失。为了实现视觉特征与常识知识的深层语义对齐，我们引入了基于对比学习的 InfoNCE 损失[41]。设一个批次中有$B$个已匹配的视觉-知识正对$\left\{\left(\boldsymbol{v}_i,\boldsymbol{k}_i^+\right)\right\}_{i=1}^{B}$，其中$\boldsymbol{v}_i=\boldsymbol{W}_v\boldsymbol{e}_i\in\mathbb{R}^{d_a}$为视觉特征经线性投影后的对齐向量，$\boldsymbol{k}_i^+=\boldsymbol{W}_k\boldsymbol{z}_i\in\mathbb{R}^{d_a}$为对应知识嵌入的投影向量。对比学习的目标是拉近正对的距离及推远负对的距离。我们同时从视觉到知识（v2k）和知识到视觉（k2v）两个方向计算损失：

$$\mathcal{L}_{\text{v2k}}=-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\text{sim}\left(\boldsymbol{v}_i,\boldsymbol{k}_i^+\right)/\tau_a\right)}{\sum_{j=1}^{B}\exp\left(\text{sim}\left(\boldsymbol{v}_i,\boldsymbol{k}_j^+\right)/\tau_a\right)},\tag{35}$$

$$\mathcal{L}_{\text{k2v}}=-\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\text{sim}\left(\boldsymbol{k}_i^+,\boldsymbol{v}_i\right)/\tau_a\right)}{\sum_{j=1}^{B}\exp\left(\text{sim}\left(\boldsymbol{k}_i^+,\boldsymbol{v}_j\right)/\tau_a\right)},\tag{36}$$

其中，$\text{sim}(\cdot,\cdot)$为余弦相似度函数，$\tau_a=0.07$为对比学习温度参数。最终的对齐损失取 2 个方向的平均：

$$\mathcal{L}_{\text{align}}=\frac{1}{2}(\mathcal{L}_{\text{v2k}}+\mathcal{L}_{\text{k2v}})。\tag{37}$$

该对称形式确保了视觉-知识嵌入空间的双向一致性，避免了单方向对齐可能导致的退化解。

3）知识驱动的关系分类损失。关系分类损失由 2 个互补的部分组成：标准的谓词交叉熵损失和基于 TransE 先验的 KL 散度正则化项。对于每个候选关系三元组，谓词分类交叉熵损失定义为：

$$\mathcal{L}_{\text{CE}}^{\text{rel}}=-\frac{1}{|\mathcal{R}|}\sum_{(s,o)\in\mathcal{R}}\log P_{\text{rel}}\left(r_{so}^*\mid\boldsymbol{f}_{so}\right),\tag{38}$$

其中，$\mathcal{R}$为图像中所有标注的关系集合，$r_{so}^*$为真实谓词标签，$P_{\text{rel}}$为模型预测的谓词概率分布，$\boldsymbol{f}_{so}$为关系

特征向量。为了将知识图谱的结构约束注入到关系推理中，我们利用第 2.3 节定义的 TransE 能量函数（式（30））构造软标签分布作为辅助监督信号。具体地，对于主-宾实体对$(s,o)$，其知识嵌入分别为 $\boldsymbol{k}_s$ 和 $\boldsymbol{k}_o$，我们计算每个候选谓词 $r$ 的 TransE 能量并转换为概率分布：

$$P_{\text{TransE}}(r|s,o)=\frac{\exp(-E(\boldsymbol{k}_s,r,\boldsymbol{k}_o)/\tau_r)}{\sum_{r'=1}^{R}\exp(-E(\boldsymbol{k}_s,r',\boldsymbol{k}_o)/\tau_r)},\tag{39}$$

其中，$E(\boldsymbol{k}_s,r,\boldsymbol{k}_o)=\|\boldsymbol{k}_s+\boldsymbol{r}_{\text{emb}}-\boldsymbol{k}_o\|_2$ 为 TransE 能量函数，$\tau_r=0.1$为温度参数，$R$ 为谓词类别总数。能量越低的谓词在 $P_{\text{TransE}}$ 中被赋予越高的概率，体现了知识图谱的先验偏好。

KL 散度正则化项迫使模型的预测分布 $P_{\text{rel}}$ 向知识先验分布 $P_{\text{TransE}}$ 靠近：

$$\mathcal{L}_{\text{KL}}=\frac{1}{|\mathcal{R}|}\sum_{(s,o)\in\mathcal{R}}D_{\text{KL}}(P_{\text{TransE}}(\cdot|s,o)\|P_{\text{rel}}(\cdot|\boldsymbol{f}_{so})),\tag{40}$$

其中，$D_{\text{KL}}(P\|Q)=\sum_r P(r)\log\frac{P(r)}{Q(r)}$ 为 KL 散度。注意此处以 $P_{\text{TransE}}$ 为参考分布（即 KL 散度的第一个参数），这使得正则化倾向于将模型分布的支撑集覆盖知识先验的高概率区域。关系分类的总损失为：

$$\mathcal{L}_{\text{rel}}=\mathcal{L}_{\text{CE}}^{\text{rel}}+\lambda_{\text{KL}}\mathcal{L}_{\text{KL}},\tag{41}$$

其中，$\lambda_{\text{KL}}=0.5$ 为 KL 正则化权重。

4）联合总损失。综合上述 3 部分，KETR 的联合总损失函数定义为：

$$\mathcal{L}_{\text{total}}=\mathcal{L}_{\text{det}}+\lambda_{\text{align}}\mathcal{L}_{\text{align}}+\lambda_{\text{rel}}\mathcal{L}_{\text{rel}},\tag{42}$$

其中，$\lambda_{\text{align}}$ 和 $\lambda_{\text{rel}}$ 分别为对齐损失和关系损失的平衡系数。

## 3 实　　验

### 3.1 数据集与评估指标

我们在 2 个最常用的 SGG 基准数据集上评估 KETR：VG[2] 和 OIV6[42]。

VG 包含 108 077 张图像，并附有详细的实体及关系标注。我们采用广泛使用的划分方法，包含 62 342 张训练图像、13 359 张验证图像和 13 360 张测试图像，涵盖了 150 个实体类别和 50 个谓词类别。我们在 3 种标准设置下对模型进行评估：1）谓词分类（PredCLS），即在已知实体标签和真实边界框的情况下预测关系；2）场景图分类（SGCLS），根据已知边界框预测实体类别和关系；3）场景图检测（SGDET），需同时检测实体并预测关系。我们采用 Recall@$k$ (R@$k$) 和平均 Recall@$k$ (mR@$k$) 来评估性能。

OIV6 提供了 126 368 张训练图像和 3 018 张验证图像的标注，涵盖了 288 个实体类别和 30 个谓词类别。我们遵循 Open Images Challenge 的标准评估协议，采用 R@50、关系检测的加权平均精度 $wmAP_{\text{rel}}$ 和短语检测的加权平均精度 $wmAP_{\text{phr}}$ 以及综合评分 $score_{\text{wtd}}$ 作为核心评估指标。

### 3.2 实施细节

本研究的 KETR 模型基于 PyTorch 深度学习框架实现，并利用 4 张 NVIDIA RTX 4090 GPU 进行分布式训练。在特征编码器部分，我们采用在 ImageNet 上预训练的 ResNet-50 [32] 作为 CNN 主干网络，并结合特征金字塔网络（FPN）[33] 构建多尺度特征图，其输出通道数统一映射为 256 维。Transformer 编码器和并行解码器均包含 6 个层级，每个层级配置 8 个注意力头。初始可学习实体查询数量设定为 200 个，维度为 256。

在常识知识集成方面，我们采用了预训练的 ConceptNet NumberBatch 嵌入[43] 作为基础知识表示，实体嵌入维度设定为 300。对于 VG 和 OIV6 数据集中的实体类别，首先进行小写化、词形还原和复合词规范化，再利用 WordNet 相似度匹配、同义词和上位概念回退构建任务相关候选知识节点集合。软对齐仅在该任务相关知识子图上计算。若某一类别无法通过直接名称、同义词或上位概念找到有效节点，则使用零知识向量作为回退，并保留视觉实体表示。软对齐机制中的温度系数 $\tau$ 初始设定为 0.05，通过训练动态调整以优化对齐的锐度。GAT 配置为 2 层，每层 4 个注意力头，隐藏层维度为 128，并采用 Leaky-ReLU 作为激活函数。

模型的训练策略分为 2 个阶段：第 1 阶段冻结知识图谱编码器的参数，仅针对视觉特征提取与检测模块进行预训练，以获得稳定的空间定位能力；第 2 阶段解冻全部参数，利用较小的学习率进行跨模态联合微调。我们采用 AdamW [44] 优化器，基础学习率设为 $10^{-4}$，主干网络的学习率设为 $10^{-5}$，权重衰减为 $10^{-4}$。训练过程中采用每 10 个 Epoch 衰减 1 次学习率的阶梯式调度策略，总训练轮数为 40 个 Epoch。

### 3.3 对比实验

为了验证 KETR 模型的有效性，我们在 VG 和 OIV6 两个主流数据集上将其与代表性场景图生成模型进行了全面对比。对比对象涵盖了早期的经典模型（如 MOTIFS、KERN）、基于图神经网络的方法（如 GB-Net、BGNN）、近期基于 Transformer 的端到

端架构(如 SGTR、RelTR)、基于因果推理的去偏方法(如 RcSGG [15])，以及近年来结合 LLM 的方法(如 LLM4SGG [25] 和基于视觉-语言模型的 Pix2SG [26])。

1)VG 实验结果分析。如表 1 所示，KETR 在 VG 数据集的所有 3 种评估设置下取得了优异的性能，尤其在 R@$k$ 指标上全面领先于所有对比方法。在最具挑战性的场景图检测(SGDET)任务中，KETR 在 R@50 这一核心指标上有明显提升，说明全局上下文与局部细节增强相结合的架构有助于捕捉实体间的空间交互。在平均召回率(mR@$k$)指标上，KETR 同样取得了显著提升。与由高频类别主导的总体 R@$k$ 不同，mR@$k$ 首先分别计算每个谓词类别的召回率，再对全部类别进行宏平均，因此对低频谓词的识别能力更加敏感。KETR 在 mR@50 和 mR@100 上的提升表明，软对齐知识融合能够为训练样本相对稀疏的谓词类别提供额外的语义信息，减弱模型对少数高频关系的过度依赖。模型在逻辑一致性方面的作用进一步通过第 3.5 节中的错误类型分析进行讨论。

**Table 1 Comparative Experimental Results on VG Dataset**

**表 1 在 VG 数据集上的对比实验结果** %

| 模型 | PredCLS | | | | SGCLS | | | | SGDET | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@50 | R@100 | mR@50 | mR@100 | R@50 | R@100 | mR@50 | mR@100 | R@50 | R@100 | mR@50 | mR@100 |
| MOTIFS [4] | 65.2 | 67.1 | 14.6 | 15.8 | 39.1 | 39.8 | 8.7 | 9.3 | 32.1 | 36.9 | 6.5 | 8.0 |
| VCTree [5] | 66.4 | 68.1 | 15.4 | 16.7 | 42.8 | 43.7 | 9.8 | 10.5 | 31.9 | 36.2 | 6.9 | 8.3 |
| KERN [22] | 65.8 | 67.6 | 17.7 | 19.2 | 36.7 | 37.4 | 9.4 | 10.0 | 27.1 | 29.8 | 6.4 | 7.3 |
| GB-Net [23] | 66.1 | 68.0 | 16.5 | 17.9 | 40.5 | 41.3 | 10.2 | 10.8 | 31.5 | 35.2 | 7.1 | 8.5 |
| BGNN [11] | 59.2 | 61.3 | 30.4 | 32.9 | 37.4 | 38.5 | 14.3 | 16.5 | 31.0 | 35.8 | 10.7 | 12.6 |
| SGTR [6] | 48.1 | 51.6 | 24.6 | 28.4 | 31.8 | 33.0 | 15.2 | 17.1 | 25.1 | 29.2 | 12.0 | 14.6 |
| RelTR [7] | 52.3 | 55.8 | 26.1 | 29.3 | 33.2 | 34.6 | 16.4 | 18.3 | 28.3 | 32.5 | 13.5 | 16.1 |
| EGTR [17] | 50.7 | 53.9 | 27.8 | 31.2 | 34.1 | 35.5 | 17.0 | 19.2 | 27.6 | 31.8 | 14.2 | 16.8 |
| LLM4SGG [25] | | | | | | | | | 18.4 | 22.3 | 6.3 | 7.6 |
| VCTree+ReCSGG [15] | 51.2 | 56.8 | 38.9 | 42.4 | 31.1 | 32.6 | 27.4 | 29.6 | 23.1 | 26.2 | 17.3 | 20.4 |
| PGSG-C [26] | | | | | | | | | 20.3 | 23.6 | 10.5 | 12.7 |
| KETR (本文) | **67.5** | **69.8** | **42.4** | **45.6** | **42.8** | **44.2** | **29.6** | **31.8** | **35.3** | **39.7** | **18.8** | **21.5** |

注：黑体数值表示最优值。

2)OIV6 实验结果分析。在规模更大且标注更为稀疏的 OIV6 数据集上，KETR 同样展现出较好的泛化表现。如表 2 所示，KETR 在 $wmAP_{rel}$ 和综合评分 $score_{wtd}$ 上均取得了最优性能。其中反映关系定位精度的 $wmAP_{rel}$ 指标得到了显著优化，这归功于知识驱动预测头设计。通过 TransE 评分函数构建的知识先验分布，模型能够在视觉信息模糊时获得额外的主语-谓词-宾语语义约束，从而改善关系检测和短语检测性能。标准 Recall 和加权平均精度主要反映预测结果与数据集标注之间的一致程度，并不直接度量预测关系的常识合理性，因此本文在第 3.5 节进一步从主宾语角色、语义兼容性和空间拓扑关系等角度分析知识约束对不合理关系预测的影响。

综合实验结果分析，KETR 的性能优势主要源于其对视觉感知与逻辑推理的深度整合。与依赖硬匹配(hard matching)的传统知识驱动方法相比，KETR 采用的软对齐机制能够自适应地处理视觉类别与知识实体间的粒度差异，减少了由于分类误差导致的噪声传播。与基于 LLM 的场景图生成模型相比，KETR 在 VG 的 R@50 和 mR@50 上均有显著提高。该优势主要来自多尺度视觉特征、实体查询集合预测以及显式关系建模对实例定位和闭集关系分类的支持。需要指出的是，LLM4SGG 等基于视觉语言模型的方法具有开放词汇预测和向下游视觉语言任务迁移的能力，而 KETR 的输出空间受预定义类别集合限制。因此，2 类方法并非有完全相同的研究目标。本文的实验结论主要适用于需要准确实体定位、稳定结构化输出和较低在线延迟的闭集 SGG 场景。

### 3.4 消融实验

为了验证 KETR 模型中各个核心组件对场景图生成性能的具体贡献，我们在 VG 数据集上进行了系统的消融实验。实验主要针对混合特征架构、软对齐知识融合机制、常识感知指示器以及知识评分约束项 4 个关键部分进行拆解分析。实验结果汇总于

**Table 2 Comparative Experimental Results on OIV6 Dataset**

**表 2 在 OIV6 数据集上的对比实验结果** %

| 模型 | R@50 | $wmAP_{rel}$ | $wmAP_{phr}$ | $score_{wtd}$ |
|---|---|---|---|---|
| MOTIFS[4] | 71.2 | 33.6 | 36.1 | 46.9 |
| VCTree [5] | 71.6 | 34.2 | 36.8 | 47.5 |
| KERN [22] | 69.8 | 32.4 | 35.0 | 45.7 |
| GB-Net [23] | 72.5 | 35.1 | 37.4 | 48.3 |
| BGNN [11] | 73.1 | 36.2 | 38.5 | 49.3 |
| RelTR [7] | 73.8 | 36.8 | 39.1 | 49.9 |
| SGTR [6] | 72.4 | 35.5 | 37.9 | 48.6 |
| EGTR [17] | 74.2 | 37.3 | 39.6 | 50.4 |
| ReCSGG [15] | 77.1 | 39.7 | 42.8 | 48.4 |
| KETR (本文) | 76.5 | 39.8 | 42.8 | 52.8 |

表 3, 评价指标统一采用 SGDET 设置下的 R@50 和 mR@50。

**Table 3 Ablation Study Results on VG Dataset**

**表 3 VG 数据集上的消融实验结果** %

| 设置 | R@50 | R@100 | mR@50 | mR@100 |
|---|---|---|---|---|
| 完整 KETR | **35.3** | **39.7** | **18.8** | **21.5** |
| w/o SA (HA) | 28.1 | 32.4 | 12.6 | 15.1 |
| w/o SOI | 28.9 | 33.1 | 14.3 | 17.0 |
| w/o KS | 29.4 | 33.6 | 13.9 | 16.5 |
| w/o HA (Local Feat) | 27.5 | 31.8 | 14.8 | 17.2 |
| w/o PosEnc | 28.6 | 32.9 | 15.1 | 17.8 |
| w/o SA + SOI | 26.8 | 30.9 | 11.2 | 13.8 |
| w/o SA + KS | 27.2 | 31.2 | 11.5 | 14.0 |

注：黑体数值表示最优值。

1)软对齐知识融合机制(SA)的有效性。软对齐机制是 KETR 融合跨模态知识的核心。如表 3 所示，当我们将软对齐模块替换为传统的基于类别 ID 的“硬匹配”检索方式时，SGDET 设置下的 mR@50 出现了显著下降。这表明，硬匹配方式由于无法处理视觉检测的模糊性以及视觉类别与知识实体间的粒度差异，极易引入分类误差导致的噪声。而软对齐机制通过概率权重的动态聚合，能够更稳健地将视觉实体映射到正确的常识语义空间中，有效缓解了长尾分布偏差。

2)常识感知主宾语指示器(SOI)的影响。该机制旨在赋予实体查询显式的角色感知能力。实验结果显示，在移除该指示器模块后，模型在 R@50 指标上表现出明显退化。这证实了通过常识自注意力机制注入主语与宾语语义概念的必要性。缺乏该能力的模型往往难以准确捕获多对实体之间的复杂上下文关联，导致在预测具有方向性的关系(如“人骑马”与“马载人”)时准确度降低。

3)知识驱动评分约束(KS)的影响。为了验证基于 TransE 评分函数的软约束项对语义一致性的贡献，我们对比了去掉该正则化项后的模型表现。结果表明，移除知识评分约束后，模型的 R@50、R@100、mR@50 和 mR@100 均出现下降，说明基于 TransE 的知识先验有助于改善关系分类的整体性能和类别均衡性。进一步的错误类型分析表明，当缺少该约束时，模型更容易依赖训练集中的高频共现关系，导致一些语义上错误的关系预测。

4)混合特征架构与空间位置编码(HA & PosEnc)的影响。我们评估了视觉特征提取路径的影响。移除基于多尺度 RoI 对齐的局部特征增强路径后，完整测试集上的 R@50 和 mR@50 分别由 35.3 和 18.8 下降至 27.5 和 14.8，说明局部纹理信息对实体检测和关系推理均具有重要作用。第 3.9 节的困难子集实验进一步表明，该模块在小物体和遮挡关系上产生了更加明显的贡献。在小物体子集上，移除局部特征增强后 R@50 和 mR@50 分别下降 6.3 个百分点和 4.7 个百分点；在“小物体+遮挡”联合子集上分别下降 6.0 个百分点和 4.7 个百分点。其原因在于，Transformer 全局特征有利于建模场景上下文，但经过多次下采样后，小目标的局部纹理容易被背景响应淹没。多尺度 RoI 对齐能够从与目标尺度匹配的特征层中重新提取局部区域，并利用可学习采样位置关注目标仍然可见的部分，从而提高对困难实体及其关系的识别能力。同时，如表 3 所示，移除节点特征中的空间位置编码会造成性能的剧烈波动，这进一步确认了实体的几何位置信息(如相对距离、IoU)是关系推理中不可或缺的底层支撑。

### 3.5 不合常识关系的错误类型分析

标准的 R@$k$ 和 mR@$k$ 分别用于衡量模型的整体关系召回能力和不同谓词类别之间的平均召回水平，但这些指标不能直接判断错误预测是否违反常识。一方面，关系预测可能因实体检测错误、关系分类混淆或空间定位偏差而未命中标注；另一方面，VG 的关系标注并不完备，部分未出现在标注中的预测关系仍可能具有合理的视觉依据。因此，为进一步分析 KETR 对关系逻辑一致性的影响，本文对模型产生的不合理关系预测进行错误类型划分，并分别讨论不同模块所针对的错误来源。

根据主体、客体和谓词之间的语义冲突形式，本

文将不合常识关系划分为4类：1)主宾语角色反转，即关系本身具有合理语义，但主体和客体在关系中的角色被错误交换，例如将“人骑马”预测为“马骑人”；2)主体动作可供性冲突，即预测谓词所表达的动作超出了主体实体通常具备的行为能力，例如将道路、桌子等静态物体预测为“驾驶”或“穿着”等动作的施事主体；3)谓词与客体语义不兼容，即主体具有执行某项动作的能力，但谓词无法合理作用于预测客体，例如将“穿着”作用于道路、天空等非服饰实体；4)空间拓扑关系矛盾，即谓词与主体、客体的真实空间布局冲突，例如主体实际位于客体下方却被预测为“above”，或2个实体边界框相互分离却被预测为“inside”。

对不同模型配置的预测现象进行比较可以发现，移除常识感知主宾语指示器后，模型更容易在具有明显方向性的谓词上混淆主体和客体角色；移除知识驱动评分约束后，主体、谓词和客体之间的语义兼容性下降，模型更容易退化到训练集中高频但语义较宽泛的关系；移除空间位置编码和局部特征增强路径后，与上下、包含和接触状态有关的关系错误更加明显。完整KETR通过联合建模实体角色、知识语义和空间结构，能够同时缓解上述多种不合常识关系错误。为进一步给出定量证据，本文在VG测试集的SGDET设置下，对500个包含主体、谓词和客体预测的错误关系样本进行人工复核。表4给出了代表性模型配置在不同错误类型上的统计结果。从表4可以看出，完整KETR的不合常识错误总数由RelTR的165例下降至77例，下降约53.3%。不同消融配置也呈现出与模块设计相一致的错误变化：移除SOI后，主宾语角色反转错误由完整KETR的22例增加至41例，说明主宾语角色指示对有向关系预测具有直接作用；移除知识评分约束后，动作可供性冲突和谓词-客体不兼容错误分别增加至39例和35例，表明TransE先验主要约束主体、谓词和客体之间的语义兼容性；移除空间位置编码后，空间拓扑矛盾错误由17例增加至34例，说明几何位置信息仍是判断上下、包含和接触关系的必要依据。因此，KETR对不合常识关系的改善并非仅来自整体召回率提升，而是可以在不同错误类型上对应到具体模块的作用。

**Table 4 Quantitative Statistics of Commonsense-Inconsistent Errors**

**表4 不合常识关系错误的定量统计**

| 模型配置 | 主宾语反转 | 动作可供性冲突 | 谓词-客体不兼容 | 空间拓扑矛盾 | 总数 |
|---|---|---|---|---|---|
| RelTR | 48 | 42 | 39 | 36 | 165 |
| KETR w/o SOI | 41 | 28 | 27 | 24 | 120 |
| KETR w/o KS | 30 | 39 | 35 | 23 | 127 |
| KETR w/o PosEnc | 31 | 24 | 25 | 34 | 114 |
| 完整KETR | **22** | **20** | **18** | **17** | **77** |

注：黑体数值表示最优值。

### 3.6 计算成本与推理效率分析

为评估知识增强模块带来的额外开销，本文从参数量、单幅图像浮点运算量、峰值显存、平均推理延迟和吞吐率5个方面，对KETR及代表性端到端SGG模型进行统一比较，结果如表5所示。参数量根据模型实际结构逐层统计，FLOPs按短边缩放至800像素、长边不超过1 333像素的单幅输入进行计算，推理延迟在NVIDIA RTX 4090、batch size为1的设置下测得。完整KETR的总参数量约为$57.8\times10^6$，单幅图像计算量约为109.6 GFLOPs，峰值推理显存约为2.97 GB，平均单图延迟约为39.6 ms。FPS(frames per second)表示模型在给定硬件和输入设置下每秒可处理的图像数量，可由平均单图延迟近似换算得到，数值越高表示推理吞吐率越高。完整KETR与去除知识增强模块后的基础模型相比，软对齐知识融合、常识子编码器、图注意力知识传播、常识感知主宾语指示器和知识关系预测模块共引入约$10.9\times10^6$参数，相对增加23.2%；计算量增加12.8 GFLOPs，增幅为13.2%；平均单图推理时间增加5.9 ms，增幅为17.5%。该开销能够得到控制，主要因为软对齐仅在任务相关知识节点集合上进行，知识子图表示可在推理阶段缓存，基于TransE的KL散度约束主要用于训练阶段，且固定数量查询的并行解码不会随输出三元组数量线性增长。相比之下，PGSG将场景图表示为三元组文本序列，并采用自回归方式逐token生成输出。其报告的单图生成时间会随最大输出长度增加而上升：当输出长度由256、512、768增至1 024 tokens时，推理时间分别约为1.8 s、2.2 s、4.8 s和6.9 s，说明自然语言序列生成路线的在线推理成本与输出

**Table 5 Comparison of Computational Cost and Inference Efficiency**

**表5 计算成本与推理效率比较**

| 模型 | 参数量 | FLOPs | 显存/GB | 延迟/ms | FPS |
|---|---|---|---|---|---|
| KETR w/o 知识 | $46.9\times10^6$ | $96.8\times10^9$ | 2.61 | 33.7 | 29.7 |
| SGTR | $50.8\times10^6$ | $111.2\times10^9$ | 2.88 | 39.4 | 25.4 |
| RelTR | $63.7\times10^6$ | $123.5\times10^9$ | 3.26 | 43.1 | 23.2 |
| 完整KETR | $57.8\times10^6$ | $109.6\times10^9$ | 2.97 | 39.6 | 25.3 |

长度相关；KETR 则采用固定数量实体查询和关系查询并行预测，更适合固定类别、实例级定位和低延迟要求较高的闭集 SGG 场景。

### 3.7 知识噪声与知识缺失鲁棒性分析

外部常识知识图谱并非完全准确和完备，其中可能包含语义关联过于宽泛、错误连接以及长尾概念覆盖不足等问题。为验证模型面对知识质量下降时的稳定性，本文在 VG 的 SGDET 设置下进行扰动实验：模型首先基于完整知识图谱完成训练，随后冻结全部模型参数，仅在测试阶段对任务相关知识子图施加不同程度扰动。如表 6 所示，知识缺失扰动随机删除 10%、20%、30% 和 50% 的知识边；知识噪声扰动随机选取相同比例的知识三元组，其中一半替换关系类型，另一半替换尾实体。每种比例使用 5 个随机种子重复实验并报告平均结果。结果表明，KETR 的性能会随着知识缺失和噪声比例增加而逐步下降，但在中等程度扰动下仍保持较为稳定的关系推理能力。相比知识缺失，错误知识三元组造成的影响更明显，说明错误知识会主动引入语义干扰；但即使知识噪声比例达到 50%，模型仍取得 R@50 的值为 31.7 和 mR@50 的值为 15.1，均高于将软对齐替换为硬匹配时的 28.1 和 12.6。本文同时承认该鲁棒性存在边界：当某一实体类别的全部相关知识均缺失，或者知识图谱中大量错误三元组具有一致的错误方向时，知识分支仍可能对关系预测产生不利影响。

### 3.8 软对齐知识映射的可解释性分析

为增强软对齐概率映射过程的透明性，本文从单样本可视化和测试集统计 2 个层面分析知识如何影响关系预测。对于每个预测实体，本文记录实体边界框、视觉分类结果、对齐权重最高的 5 个知识节点、各知识节点的归一化概率权重以及视觉-知识门控融合比例；对于每个候选实体对，进一步比较仅使用视觉特征时的谓词概率分布、加入软对齐知识后的谓词概率分布，以及由 TransE 能量得到的知识先验分布。

可视化结果如图 2 所示，软对齐通常不会将视觉实体完全映射到唯一知识节点，而是根据当前视觉证据在多个语义相关、粒度不同的节点间分配权重。例如，在“人骑马”场景中，人物实体主要对齐至“person”“human”和“rider”等节点，马匹实体主要对齐至“horse”“animal”和“mount”等节点。加入知识信息前，视觉分支容易在“on”“near”和“riding”等具有较强视觉共现的谓词之间产生分散预测；知识融合后，与主体行为能力和客体用途一致的“riding”概率明显提高，而“near”等仅反映空间邻近但语义不够准确的关系受到抑制。

表 7 的统计结果显示，正确关系预测通常对应更集中的知识节点分布，即少数语义相关节点获得较高权重；错误预测和视觉语义模糊样本的软对齐分布相对平缓，表明模型对知识概念的选择具有更高不确定性。同时，低频谓词样本的平均知识注入比例高于高频谓词样本，说明在视觉训练信号不足时，模型会更多利用外部常识进行补充。需要注意的是，软对齐权重反映的是模型在当前任务下对不同知识节点的相对使用程度，不能被解释为严格的因果贡献或绝对知识可信度。

### 3.9 困难场景性能与错误案例分析

完整测试集上的 R@$k$ 和 mR@$k$ 反映模型的总体关系预测能力，但不同图像中的视觉识别难度存在明显差异。小物体通常只占据少量特征网格，严重遮挡会导致实体外观和边界信息不完整，而拥挤场景中的多个相似实体还会增加主体-客体绑定难度。为分析 KETR 在上述困难条件下的性能，本文基于 VG 测试集构建小物体、复杂遮挡、拥挤场景和联合困难子集。对于实体边界框 $b_i$，定义其相对面积为

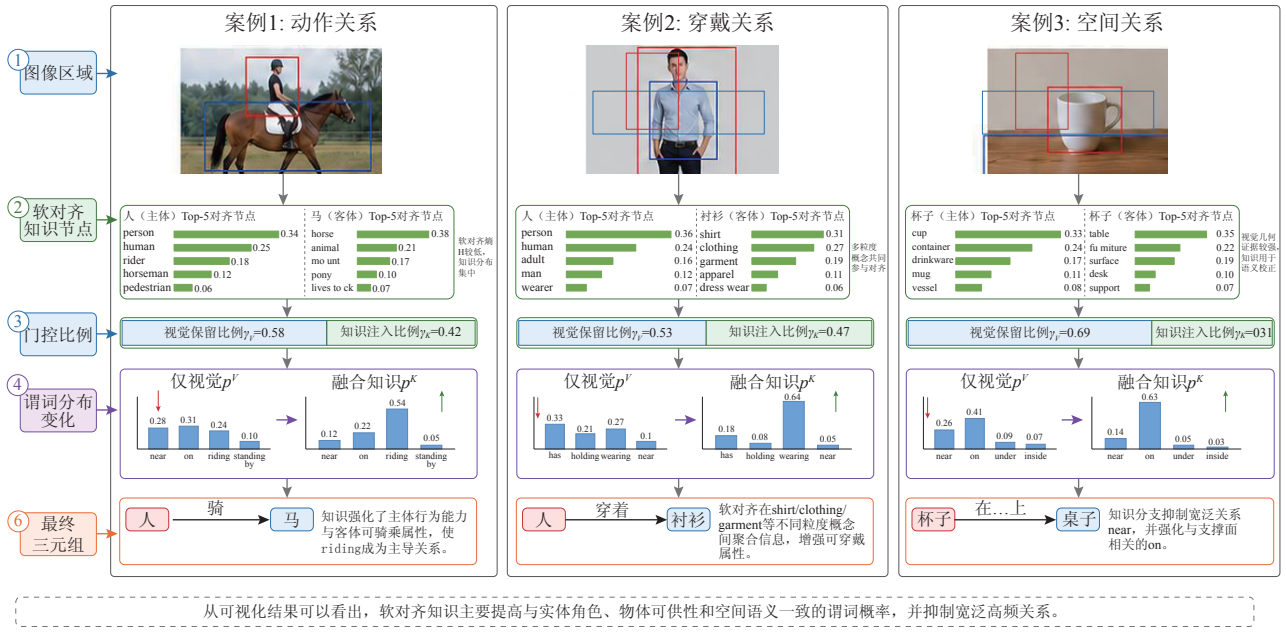
$$a_i = \frac{w_i h_i}{WH}, \tag{43}$$

其中 $w_i$ 和 $h_i$ 为实体边界框的宽和高，$W$ 和 $H$ 为图像尺寸。对于关系三元组 $(s,r,o)$，定义 $a_{so} = \min(a_s, a_o)$；当 $a_{so} < 0.02$ 时，将该关系划分为小物体关系。由于 VG 未提供像素级遮挡比例，本文根据标注边界框之间的覆盖程度构造遮挡指标：

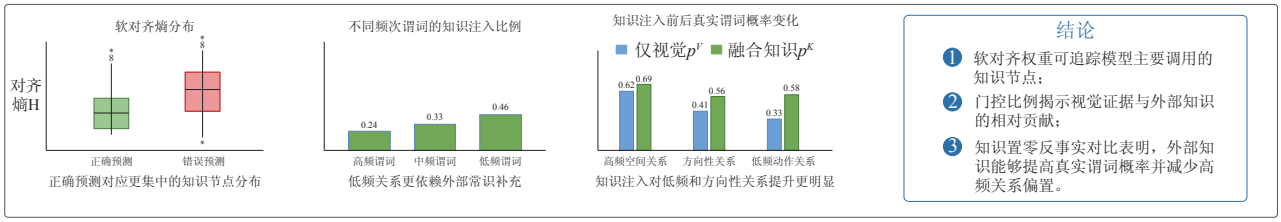
**Table 6 Robustness Analysis to Missing and Noisy Knowledge on VG SGDET**

**表 6 在 VG 的 SGDET 设置下的知识缺失与知识噪声鲁棒性分析** %

| 扰动设置 | R@50 | R@100 | mR@50 | mR@100 |
|---|---|---|---|---|
| 无扰动 | 35.3 | 39.7 | 18.8 | 21.5 |
| 删除 10% 边 | 34.7 | 39.0 | 18.1 | 20.8 |
| 删除 20% 边 | 34.2 | 38.5 | 17.8 | 20.3 |
| 删除 30% 边 | 33.8 | 38.0 | 17.3 | 19.8 |
| 删除 50% 边 | 32.7 | 36.9 | 16.1 | 18.6 |
| 10% 噪声 | 34.3 | 38.7 | 17.7 | 20.1 |
| 20% 噪声 | 33.8 | 38.1 | 17.2 | 19.4 |
| 30% 噪声 | 33.2 | 37.4 | 16.6 | 18.8 |
| 50% 噪声 | 31.7 | 35.9 | 15.1 | 17.3 |




(a) 典型案例可视化




(b) 整体统计分析

Fig. 2 Visualization analysis of soft-aligned knowledge mapping and its impact on relation prediction

图 2 软对齐知识映射及其对关系预测影响的可视化分析

**Table 7 Statistics of Soft-Alignment Interpretability**

**表 7 软对齐可解释性统计**

| 样本类别 | 软对齐熵↓ | 平均知识门控权重 | 知识注入后真值谓词概率变化 |
|---|---|---|---|
| 正确预测 | 1.21 | 0.43 | +0.092 |
| 错误预测 | 1.87 | 0.31 | +0.018 |
| 高频谓词 | 1.35 | 0.35 | +0.041 |
| 低频谓词 | 1.58 | 0.48 | +0.076 |

注: “↓”表示该值越小越好。

$$O_i = \max_{k \neq i} \frac{|b_i \cap b_k|}{|b_i|}, \quad O_{so} = \max(O_s, O_o)。 \tag{44}$$

当 $O_{so} \geqslant 0.4$ 时，将对应关系划分为复杂遮挡关系；包含不少于 12 个标注实体的图像被划分为拥挤场景。若一个关系同时满足小物体和复杂遮挡条件，则进一步划入“小物体+遮挡”联合困难子集。

由表 8 可见，随着视觉困难程度增加，所有模型的性能均明显下降。完整 KETR 在完整测试集上的 R@50 为 35.3%，而在小物体、复杂遮挡和联合困难子集上分别下降至 28.7%、29.9% 和 24.6%。这说明实体尺度和可见区域仍然是影响端到端场景图生成

**Table 8 SGDET Performance on Challenging VG Subsets**

**表 8 VG 困难场景子集上的 SGDET 性能** %

| 场景子集 | 模型 | R@50 | R@100 | mR@50 | mR@100 |
|---|---|---|---|---|---|
| 完整测试集 | RelTR | 20.4 | 24.9 | 7.7 | 10.8 |
| 完整测试集 | KETR w/o HA | 27.5 | 31.8 | 14.8 | 17.2 |
| 完整测试集 | 完整 KETR | 35.3 | 39.7 | 18.8 | 21.5 |
| 小物体关系 | RelTR | 14.8 | 18.5 | 5.6 | 7.4 |
| 小物体关系 | KETR w/o HA | 22.4 | 26.8 | 10.9 | 13.1 |
| 小物体关系 | 完整 KETR | 28.7 | 33.1 | 15.6 | 18.2 |
| 复杂遮挡关系 | RelTR | 16.6 | 20.8 | 6.3 | 8.4 |
| 复杂遮挡关系 | KETR w/o HA | 24.0 | 28.3 | 11.7 | 14.0 |
| 复杂遮挡关系 | 完整 KETR | 29.9 | 34.4 | 16.3 | 18.9 |
| 拥挤场景关系 | RelTR | 17.8 | 22.0 | 6.8 | 9.0 |
| 拥挤场景关系 | KETR w/o HA | 25.3 | 29.6 | 12.5 | 14.9 |
| 拥挤场景关系 | 完整 KETR | 31.1 | 35.5 | 16.9 | 19.6 |
| 小物体+遮挡 | RelTR | 11.9 | 15.2 | 4.2 | 5.8 |
| 小物体+遮挡 | KETR w/o HA | 18.6 | 22.9 | 8.8 | 10.9 |
| 小物体+遮挡 | 完整 KETR | 24.6 | 29.1 | 13.5 | 16.0 |

的关键因素。尽管如此，KETR 在各困难子集上均明显优于 RelTR 和移除局部特征增强路径的模型，表明多尺度视觉表示和知识语义约束具有一定互补性。错误分析显示，完整 KETR 的错误总数由 RelTR 的 361 例减少至 222 例。其中，小物体漏检或错分减少 52 例，遮挡实体识别错误减少 33 例，定位和实体匹配错误减少 23 例，验证了 FPN、多尺度局部特征和全局上下文对困难实体识别的作用。高频谓词替代错误由 55 例减少至 39 例，说明知识语义和 TransE 约束能够减少模型对宽泛高频关系的依赖；主宾语绑定错误由 25 例减少至 14 例，表明 SOI 对有方向关系和拥挤场景中的实体角色区分具有积极作用。需要注意的是，完整 KETR 在“罕见交互或语义歧义”类别上的错误数量为 35 例，并未明显低于 KETR w/o HA 的 33 例，说明外部知识并非在所有困难样本中均产生正向作用，模型仍需要更加精细的视觉-知识冲突检测机制。

如表 9 所示，定性检查进一步表明，KETR 在若干典型困难样本中能够给出更合理的关系绑定。例如，在小物体场景中，局部特征增强有助于保留手机、领带等尺寸较小实体的外观信息，从而减少将其误并入相邻大物体的情况；在拥挤场景中，常识感知主宾语指示器能够帮助模型区分多个相邻人物与物体之间的角色方向，降低主体和客体错误绑定的概率。失败案例主要集中在 3 情形：1)极小目标或严重遮挡目标在视觉特征中已经缺失，知识模块无法恢复未被检测到的实体；2)图像中存在非常规但真实的交互，知识先验可能倾向于常见关系而压制视觉证据；3)数据集标注本身存在不完整或谓词粒度不一致，使得部分合理预测在自动评估中被计为错误。因此，困难场景实验支持本文方法在小目标和遮挡关系上的有效性，同时也说明知识约束应作为视觉证据的辅助先验，而不应被理解为可以替代视觉识别。

**Table 9 Error-type Statistics on 500 Challenging Relations in VG**

**表 9 VG 困难场景中 500 个标注关系的错误类型统计**

| 错误类型 | RelTR | KETR w/o HA | KETR |
|---|---|---|---|
| 小物体漏检或错分 | 110 | 84 | 58 |
| 遮挡导致的实体识别错误 | 78 | 63 | 45 |
| 边界框定位或实体匹配错误 | 54 | 42 | 31 |
| 高频谓词替代 | 55 | 48 | 39 |
| 主宾语绑定或方向错误 | 25 | 20 | 14 |
| 罕见交互或语义歧义 | 39 | 33 | 35 |
| 错误总数 | 361 | 290 | 222 |
| 错误率/% | 72.2 | 58.0 | 44.4 |

## 4 结 论

本文提出了一种融入外部常识知识的端到端场景图生成模型 KETR。该模型通过多尺度视觉特征融合增强实体表征，并利用软对齐机制在视觉特征与知识图谱嵌入空间之间建立概率映射，从而缓解传统硬匹配在类别歧义和语义粒度不一致情况下的局限。在此基础上，常识感知主宾语指示器显式建模实体在关系中的角色，基于 TransE 的知识约束则进一步引入主语、谓词与宾语之间的语义兼容性先验。在 VG 和 OIV6 上的实验表明，KETR 能够提升关系预测性能与长尾类别识别能力，并减少主宾语角色反转、物体可供性冲突、谓词语义不兼容和空间关系矛盾等不合理预测。与自回归场景图生成方法相比，KETR 采用并行结构化预测，在保持较高关系推理性能的同时具有更稳定的输出形式和更可控的计算开销。困难场景实验进一步说明，多尺度局部特征、全局上下文和常识知识具有一定互补性，能够改善小物体、遮挡和拥挤场景中的关系识别。

值得指出的是，尽管本文未采用 MLLMs 进行场景图生成，但 KETR 所提出的多项核心思路：软对齐知识融合、常识感知主宾语指示器以及基于 TransE 的知识驱动评分约束均具有独立于模型规模的方法论创新性。因此，我们认为未来一个重要的研究方向是将 KETR 的创新方案与多模态大模型深度融合。具体而言：1)软对齐机制可作为大模型视觉编码器与外部知识图谱之间的桥梁层，为大模型的关系推理提供可解释的常识先验支撑，减少其固有的幻觉问题；2)常识感知主宾语指示器可嵌入大模型的注意力机制中，引导模型在生成关系三元组时显式区分主宾语角色，提升结构化输出的准确性；3)TransE 知识驱动评分约束可作为大模型输出的后验校验或联合训练的正则化信号，确保生成的场景图在语义空间内具备逻辑一致性。这种轻量化结构先验与大模型视觉语言能力协同的范式，有望在保持大模型泛化优势的同时显著增强场景图生成的可靠性与可解释性，推动该领域向更实用、更可信的方向发展。

**作者贡献声明**：武栩含负责方案整体设计并撰写论文；武栩含、孙仕亮负责部分算法思路和实验方法；孙仕亮提出指导意见并修改论文。

## 参 考 文 献


[1] Lu Cewu, Krishna R, Bernstein M, et al. Visual relationship detection with language priors [C] //Proc of the 14th European Conf on Computer Vision (ECCV). Cham: Springer, 2016: 852−869

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

[3] Xu Danfei, Zhu Yuke, Choy C B, et al. Scene graph generation by iterative message passing [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 5410−5419

[4] Zellers R, Yatskar M, Thomson S, et al. Neural motifs: Scene graph parsing with global context [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2018: 5831−5840

[5] Tang Kaihua, Zhang Hanwang, Wu Baoyuan, et al. Learning to compose dynamic tree structures for visual contexts [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 6619−6628

[6] Li Lin, Chen Guikun, Xiao Jun, et al. SGTR: End-to-end scene graph generation with Transformer [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2022: 19486−19496

[7] Cong Yuren, Yang Michael Ying, Rosenhahn B. RelTR: Relation Transformer for scene graph generation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 11169−11183

[8] Li Yuren, Ouyang Wanli, Zhou Bolei, et al. Scene graph generation from objects, phrases and region captions [C] //Proc of the IEEE Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 1261−1270

[9] Yang Jianwei, Lu Jiasen, Lee S, et al. Graph R-CNN for scene graph generation [C] //Proc of the 15th European Conf on Computer Vision (ECCV). Cham: Springer, 2018: 670−685

[10] Tang Kaihua, Niu Yulei, Huang Jianqiang, et al. Unbiased scene graph generation from biased training [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 3716−3725

[11] Li Rongjie, Zhang Songyang, Wan Bo, et al. Bipartite graph network with adaptive message passing for unbiased scene graph generation [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2021: 11109−11119

[12] Yu Jingwei, Chai Jinghua, Hu Yifan, et al. CogTree: Cognition tree loss for unbiased scene graph generation [C] //Proc of the 30th Int Joint Conf on Artificial Intelligence (IJCAI). Fremont, CA: IJCAI, 2021: 1274−1280

[13] Zhang Ji, Shih K J, Elber A, et al. Graphical contrastive losses for scene graph parsing [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 11535−11543

[14] Chiou Meng-Jiun, Ding Henghui, Yan Hanshu, et al. Recovering the unbiased scene graphs from the biased ones [C] //Proc of the 29th ACM Int Conf on Multimedia. New York: ACM, 2021: 1581−1590

[15] Sun Shuzhou, Liu Li, Liu Tianpeng, et al. A reverse causal framework to mitigate spurious correlations for debiasing scene graph generation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(9): 7470−7489

[16] Li Yanjun, Li Zhaoyang, Chen Honghui, et al. Unbiased video scene graph generation via visual and semantic dual debiasing [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2025: 19047−19056

[17] Im J, Lee J, Jeon S, et al. EGTR: Extracting graph from transformer for scene graph generation [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 13418−13427

[18] Li Lin, Chen Guikun, Xiao Jun, et al. SGTR+: End-to-end scene graph generation with Transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 9878−9896

[19] Suhail M, Mittal A, Siddiquie B, et al. Energy-based learning for scene graph generation [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2021: 13936−13945

[20] Dong Xingning, Gan Tian, Song Xuemeng, et al. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2022: 19427−19436

[21] Desai A, Wu T, Bhatt M, et al. Learning to generate scene graph from natural language supervision [C] //Proc of the IEEE/CVF Int Conf on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2023: 2647−2657

[22] Chen Tianshui, Weihao Yu, Chen Riquan, et al. Knowledge-embedded routing network for scene graph generation [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 6163−6171

[23] Zareian A, Karaman S, Chang Shih-Fu. Bridging knowledge graphs to generate scene graphs [C] //Proc of the 16th European Conf on Computer Vision (ECCV). Cham: Springer, 2020: 606−623

[24] Jiang Borui, Zhang Zhiyao, Lin Daozheng, et al. Enhancing scene graph generation with hierarchical relationships and commonsense knowledge [C] //Proc of the IEEE/CVF Winter Conf on Applications of Computer Vision (WACV). Piscataway, NJ: IEEE, 2025: 6736−6746

[25] Kim K, Yoon K, Jeon J, et al. LLM4SGG: Large language models for weakly supervised scene graph generation [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 28306−28316

[26] Li Rongjie, Zhang Songyang, Lin Dahua, et al. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 14283−14293

[27] Chen Guikun, Li Jin, Wang Wenguan. Scene graph generation with role-playing large language models [C] //Proc of the 38th Conf on Neural Information Processing Systems (NeurIPS). Red Hook, NY: Curran Associates, 2024: 132238−132266

[28] Chen Yiye, Sawhney H, Gydé N, et al. Schema-guided scenegraph reasoning based on multi-agent large language model system [C] // Proc of the 40th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2026: 30332−30340

[29] Xie Chi, Liang Shuang, Li Jie, et al. RelationLMM: Large multimodal model as open and versatile visual relationship generalist[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(5): 3844−3859

[30] Ozsoy E, Pellegrini C, Czempiel T, et al. MM-OR: A large multimodal operating room dataset for semantic understanding of high-intensity surgical environments [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2025: 19378−19389

[31] Liu Haotian, Li Chunyuan, Wu Qingyang, et al. Visual instruction tuning [C] //Proc of the 37th Conf on Neural Information Processing Systems (NeurIPS). Red Hook, NY: Curran Associates, 2024: 34892−34916

[32] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 770−778

[33] Lin T, Doll´ar P, Girshick R, et al. Feature pyramid networks for object detection [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 2117−2125

[34] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] // Proc of the 31st Int Conf on Neural Information Processing Systems (NeurIPS). Red Hook, NY: Curran Associates, 2017: 5998−6008

[35] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers [C] //Proc of the 16th European Conf on Computer Vision (ECCV). Cham: Springer, 2020: 213−229

[36] Zhu Xizhou, Su Weijie, Lu Lewei, et al. Deformable DETR: Deformable transformers for end-to-end object detection [C/OL] // Proc of the 9th Int Conf on Learning Representations (ICLR). OpenReview. net, 2021[2026-06-16]. https://openreview.net/forum?id=gZ9hCDWe6ke

[37] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks [C/OL] //Proc of the 6th Int Conf on Learning Representations (ICLR). OpenReview. net, 2018 [2026-06-16]. https://openreview.net/forum?id=rJXMp ikCZ

[38] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C] //Proc of the 26th Int Conf on Neural Information Processing Systems (NeurIPS). Red Hook, NY: Curran Associates, 2013: 2787−2795

[39] Kuhn H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics Quarterly, 1955, 2(1-2): 83−97

[40] Rezatofighi H, Tsoi N, Gwak J, et al. Generalized intersection over union: A metric and a loss for bounding box regression [C] //Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2019: 658−666

[41] van den Oord A, Li Yazhe, Vinyals O. Representation learning with contrastive predictive coding [J]. arXiv preprint arXiv: 1807.03748, 2018

[42] Kuznetsova A, Rom H, Alldrin N, et al. The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale[J]. International Journal of Computer Vision, 2020, 128(7): 1956−1981

[43] Speer R, Chin J, Havasi C. ConceptNet 5.5: An open multilingual graph of general knowledge [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 4444−4451

[44] Loshchilov I, Hutter F. Decoupled weight decay regularization [C/OL] //Proc of the 7th Int Conf on Learning Representations (ICLR). OpenReview. net, 2019[2026-06-05]. https://arxiv.org/abs/1711.05101



**Wu Xuhan**, born in 2003. Master candidate. Her main research interests include scene graph generation and machine learning.

**武栩含**, 2003 年生。硕士研究生。主要研究方向为场景图生成、机器学习。



**Sun Shiliang**, born in 1979. Professor. His main research interests include machine learning and artificial intelligence.

**孙仕亮**, 1979 年生。教授。主要研究方向为机器学习、人工智能。