EGTR：从 Transformer 中提取图以生成场景图

摘要

场景图生成 (SGG) 是检测对象和预测对象之间关系的一项具有挑战性的任务。DETR 开发出来后，基于一级目标检测器的一级 SGG 模型得到了积极的研究。然而，复杂的建模用于预测对象之间的关系，而在对象检测器的多头自注意力中学习到的对象查询之间的内在关系却被忽略了。本文提出了一种轻量级的单阶段 SGG 模型，该模型从 DETR 解码器的多头自注意力层中学习到的各种关系中提取关系图。通过充分利用自注意力副产品，可以使用浅关系提取头有效地提取关系图。考虑到关系提取任务对对象检测任务的依赖性，本文提出了一种新颖的关系平滑技术，该技术可以根据检测到的对象的质量自适应地调整关系标签。通过关系平滑，模型按照连续课程进行训练，在训练开始时以目标检测任务为重点，随着目标检测性能逐渐提高而进行多任务学习。此外，本文提出了一种连通性预测任务，该任务预测对象对之间是否存在关系，作为关系提取的辅助任务。本文展示了本文提出的方法在 Visual Genome 和 Open Image V6 数据集上的有效性和效率。本论文的代码可在 https://github.com/naver-ai/egtr 上公开获取。

关键词： 场景图生成；Transformer；DETR

1 引言

场景图生成 (SGG) [6] 旨在生成一个场景图，将对象表示为节点，将对象之间的关系表示为图像中的边缘，如图 2 所示。SGG 是一项具有挑战性的任务，因为它不仅需要检测对象，还需要基于对场景的全面理解来预测它们之间的关系。由于场景图提供了图像的结构信息，因此它可用于需要对图像进行更高水平的理解和推理的各种视觉任务，例如图像字幕 [4, 8, 36]、图像检索 [6, 24]、和视觉问答 [13, 38]。

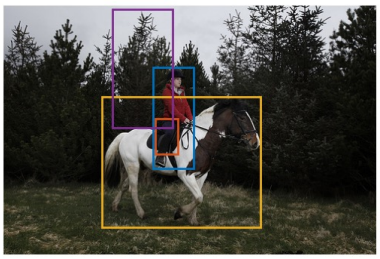

图 1. 图片示例

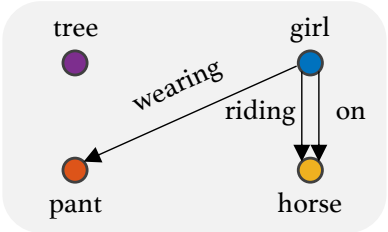

图 2. 场景图

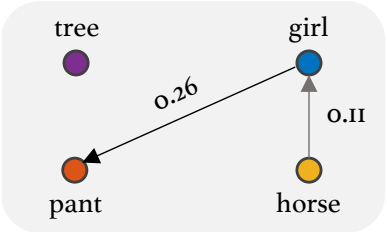

图 3. 注意力图

之前的大多数研究 [15, 16, 30, 34, 37] 采用两阶段 SGG 方法，首先检测对象，然后预测它们的关系。然而，这些方法带来了高昂的计算成本和对象检测中错误传播的风险。为了解决

两阶段方法的缺点，最近研究了利用一阶段对象同时执行对象检测和关系预测的一阶段 SGG 模型探测器，例如 DETR [1]。由于场景图中的边可以表示为主谓宾三元组，因此许多研究采用了基于三元组的方法，如图 4 和 5 所示。然而，对象三元组检测模型 [2,14] 需要复杂的三元组检测器来从对象检测器获取三元组查询所需的信息。此外，三元组检测模型 [7,31] 缺乏检测没有关系的对象的能力，例如图 2 中的 “树”，只关注三元组检测而没有对象检测器。考虑到 Visual Genome [10] 数据中没有关系的对象占 42% 以上，它们的优先级在于检测子图而不是完整的场景图。

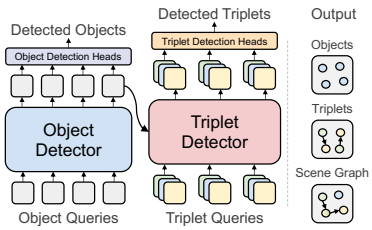

图 4. 对象-三元组检测模型

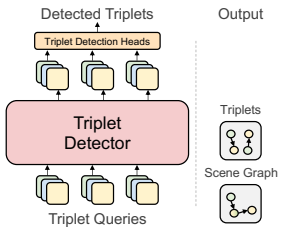

图 5. 三元组检测模型

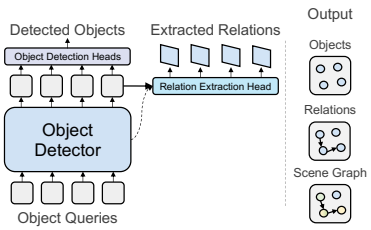

图 6. 关系提取模型

为了解决现有一阶段 SGG 模型的缺点，我们关注对象检测器中固有的对象之间的关系。如图 1 所示，对象是相互关联的。例如，当场景中出现一匹马时，场景中很可能会出现一个人，帽子、夹克、裤子等服装往往描绘出这个人当前的情况。根据这种直觉，人们长期以来一直相信对对象之间的关系或上下文进行建模将有利于对象检测任务。因此，最近的一级对象检测器 [1,28,41] 已经合并了自注意力层 [32] 来隐式地建模对象查询之间的关系。我们假设在单级对象检测器中学习的对象查询之间的自注意力可能包含用于预测三元组输出的有价值的信息。在我们的初步研究中，我们能够通过简单地连接来自预训练 DETR 的两个具有高注意力权重的对象查询来提取合理的注意力图，如图 3 所示。它显示了对象查询之间的注意力权重可以解释为它们之间的关系的潜力。

根据这些发现，我们提出了一种轻量级的单阶段场景图生成器 EGTR，它代表从 TRansformer 中提取图。我们设计的模型是为了全面利用目标检测器的副产品，从而无需单独的三元组检测器，如图 6 所示。从目标检测器的多头自注意力层中，我们将注意力查询和键（它们的关系在注意力权重中学习）分别视为主体实体和客体实体。随后，我们利用浅层分类器来预测它们之间的关系。由于来自所有自注意力层的副产品中存在有关对象之间关系的丰富信息，我们可以有效地提取场景图。

由于关系提取任务依赖于对象检测任务，因此我们推测在没有充分学习对象查询表示的情况下执行关系提取可能是有害的。因此，我们设计了一种新颖的自适应平滑技术，可以根据对象检测性能来平滑地面实况关系标签的值。通过自适应平滑，该模型通过连续课程进行训练，最初侧重于目标检测，并逐渐进行多任务学习。此外，我们提出了连通性预测任务作为关系提取的辅助任务，旨在预测主体实体和客体实体之间是否存在任何关系。该辅助任务有助于获取关系提取的表示。

为了验证所提出方法的有效性，我们在两个代表性的 SGG 数据集：Visual Genome [10] 和 Open Images V6 [11] 上进行了实验。通过积极利用物体检测器的副产品，EGTR 显示出最佳的物体检测性能和可比的三元组检测性能，并且具有最少的参数和最快的推理速度。

本文的主要贡献可概括如下：

- 本文提出了 EGTR，它利用从目标检测器导出的多头自注意力副产品，高效且有效地生成场景图。
- 本文提出了自适应平滑，为对象检测和关系提取提供有效的多任务学习。此外，所提出的连通性预测为关系提取提供了线索。
- 综合实验表明了所提出的模型框架的优越性和所设计的训练技术的有效性。

2 相关工作

SGG 模型可以分为两类：两阶段模型和一阶段模型。对于两阶段模型 [3, 9, 12, 15, 16, 19–21, 26, 30, 33, 34, 37, 40]，单独的对象检测模型和关系预测模型按顺序进行训练。他们通常从现成的目标检测器（例如 Faster R-CNN [23]）中检测 N 个目标，然后将检测到的对象的所有可能组合输入关系预测模型以预测每个对象对之间的关 系。尽管它们表现出较高的关系提取性能，但它们具有固有的局限性，即帮助关系提取的对象检测器是单独训练的，导致模型复杂性显着增加。

对于一阶段模型 [2, 7, 14, 17, 22, 25, 31]，对象检测和关系预测以端到端的方式进行训练。早期研究提出了全卷积 SGG 模型并采用基于像素的方法。在 DETR [1] 作为基于 Transformer [32] 的单级目标检测器取得巨大成功之后，许多单级 SGG 研究都是基于单级目标检测器。他们通过引入对象查询或三元组查询来有效地对 SGG 进行建模。我们将它们分为三个不同的组：(a) 对象三元组检测模型、(b) 三元组预测模型和 (c) 关系提取模型。

对象三元组检测模型。对象三元组检测模型的特点是引入额外的三元组查询并在对象检测器之上构建三元组预测器，如图 4 所示。RelTR [2] 引入了配对的主语查询和宾语查询，SGTR [14] 引入了分解为主语、宾语和谓语的组合查询。由于引入的查询是在没有来自对象检测器的先前提示的情况下初始化的，因此三元组预测器需要合并来自对象检测器的输出的信息的模块和用于三元组查询的模块以相互交换信息。它导致三元组预测器具有复杂的结构。相反，我们避免引入额外的查询，并将注意力查询和注意键（它们的关系在对象检测器中学习）分别视为主题和对象查询。

三元组检测模型。三元组检测模型使用三元组查询直接检测三元组，无需对象检测器，如图 5 所示。迭代 SGG [7] 引入了主语、宾语和谓语查询，并使用单独的主语、宾语和谓语多层 Transformer 解码器对它们之间的条件依赖关系进行建模。受稀疏 R-CNN 的启发，结构化稀疏 R-CNN (SSR-CNN) [31] 设计了由主语框、宾语框、主语、宾语和谓语查询组成的三元组查询。由于仅使用稀疏三元组注释来训练模型很困难，因此 SSR-CNN 引入了一些复杂的训练细节：它使用 Siamese Sparse RCNN 模型和辅助查询来检测对象对，并使用检测到的对象对作为伪标签来执行额外的三元组匹配。尽管它们表现出出色的三元组检测性能，但由于未使用显式对象检测器，因此模型架构变得更加复杂，无法分别检测主体和对象。最后但并非最不重要的一点是，三元组预测模型专注于检测仅由具有关系的对象组成的子图，忽略图像中缺乏明确关系的对象。

关系提取模型。关系提取模型使用轻量级关系预测器提取场景图，无需单独的三元组查询或三元组检测器，如图 6 所示。Relationformer [25] 添加了一个特殊的 “[rln]” 标记来捕获与对象查询结合的全局信息。他们将对象查询对和关系标记的最终隐藏表示连接起来，然后

是用于关系预测的浅层全连接网络。在这项工作中，结合最终的隐藏表示，我们使用了在对象检测器的多头自注意力层中学习到的对象查询之间的内在关系信息。此外，我们提出了促进多任务学习的培训技术，从而通过架构简单性显着提高性能。

3 本文方法

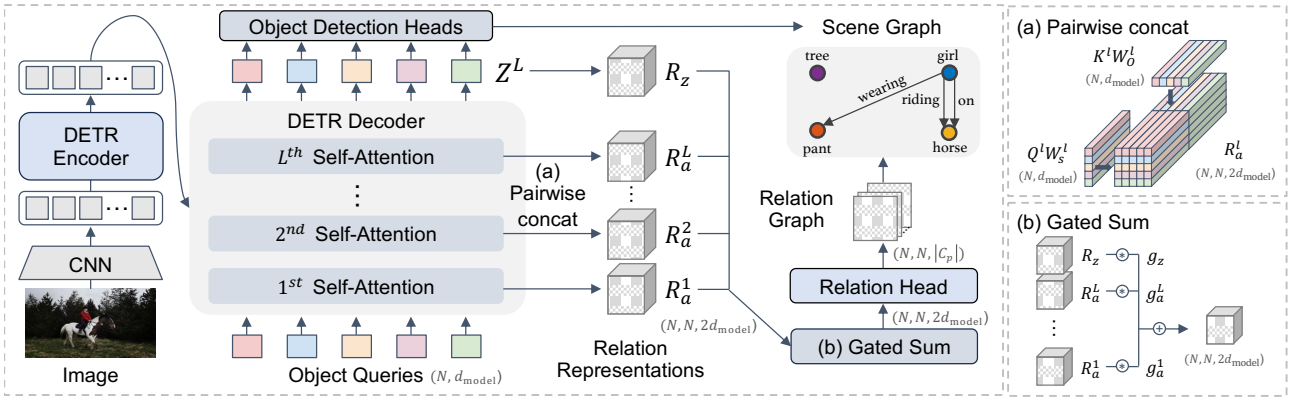

图 7. EGTR 整体架构

3.1 EGTR

我们提出了一种新颖的轻量级关系提取器 EGTR，它利用 DETR 解码器的自注意力，如图 7 所示。我们的关系提取器旨在通过将注意力查询和键分别视为主语和宾语，从整个 L 层中的自注意力权重中提取谓词信息。为了保留在自注意力层中学到的丰富信息，我们设计了注意力查询和键之间的另一个关系函数 f，而不是等式中的点积注意力。由于串联完全保留了表示，因此我们将所有 N × N 对注意查询和键的表示连接起来，如图 3（a）所示，以获得第 l 层的关系表示 $R_a^l \in \mathbb{R}^{N\times N\times 2d_{\text{model}}}$。在成对串联之前，为了帮助查询和键发挥主语和宾语的作用，添加了线性投影，如下所示：

$$R_a^l = [Q^l W_S^l; K^l W_O^l], \tag{1}$$

其中 $Q^l \in \mathbb{R}^{N\times d_{\text{model}}}$ 和 $K^l \in \mathbb{R}^{N\times d_{\text{model}}}$ 指的是第 l 层的注意力查询和键，并且 $[\cdot;\cdot]$ 表示成对连接。W_S^l 和 W_O^l 是形状为 $\mathbb{R}^{d_{\text{model}}\times d_{\text{model}}}$ 的线性权重。

我们还利用对象查询的最后一层表示 $Z^L \in \mathbb{R}^{N\times d_{\text{model}}}$，它们以相同的方式用于对象检测：

$$R_z = [Z^L W_S; Z^L W_O], \tag{2}$$

其中 W_S, W_O 是形状 $\mathbb{R}^{d_{\text{model}}\times d_{\text{model}}}$ 的线性权重。为了有效地利用所有层中学到的各种信息，我们引入了如下门控机制：

$$g_a^l = \sigma(R_a^l W_G), g_z = \sigma(R_z W_G), \tag{3}$$

其中 g_a^l 和 $g_z \in \mathbb{R}^{N\times N\times 1}$ 分别表示通过单个线性层 $W_G \in \mathbb{R}^{2d_{\text{model}}\times 1}$ 对于 R_a^l 和 R_z 获得的门值。最后，我们从所有层的关系表示的门控求和中提取关系图，如下所示：

$$\hat{G} = \sigma(\text{MLP}_{\text{rel}}(\Sigma_{l=1}^{L}(g_a^l * R_a^l) + g_z * R_z)), \tag{4}$$

其中 $\hat{G} \in \mathbb{R}^{N \times N \times |\mathcal{C}_p|}$ 表示预测关系图，$\mathrm{MLP}_{\mathrm{rel}}$ 是具有 ReLU 激活的三层感知器。请注意，我们使用 sigmoid 函数 σ，以便对象之间可以存在多种关系。

3.2 学习与推理

为了训练 EGTR，我们进行多任务学习。除了对象检测和关系提取之外，我们还设计了连通性预测，这是关系提取的辅助任务。框架的总体损失如下：

$$\mathcal{L} = \mathcal{L}_{\mathrm{od}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}} + \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}}, \tag{5}$$

其中 $\mathcal{L}_{\mathrm{od}}$ 是物体检测损失。$\mathcal{L}_{\mathrm{rel}}$ 和 $\mathcal{L}_{\mathrm{con}}$ 分别是关系提取和连通性预测的损失函数。由于使用了显式对象检测损失，因此 EGTR 能够检测场景图中的所有节点。下面提供了每项任务的损失的详细信息。

3.2.1 关系提取

我们使用二元交叉熵损失来进行关系提取。匹配预测图 $\hat{G} \in \mathbb{R}^{N \times N \times |\mathcal{C}_p|}$ 和地面实况三元组 $\mathcal{E}$，我们首先将 $\mathcal{E}$ 编码为单热地面实况图 $G \in \mathbb{R}^{N \times N \times |\mathcal{C}_p|}$ 通过将与地面实况对象之间的关系不对应的区域填充为零。然后，我们使用对象检测中发现的排列来排列预测图的索引。根据置换图 $\hat{G}'$，关系提取的损失计算为 $\mathcal{L}_{\mathrm{rel}} = \mathcal{L}_r(\hat{G}', G)$。

然而，由于图的大小与 N 的平方成正比，因此真实图 G 的稀疏性过于严重。例如，当 Visual Genome 验证数据集的 N 设置为 200 时，密度仅为 10^{-14}。因此，我们将 G 分为三个区域：(1) GT 区域，(2) 负区域，(3) 非匹配区域，如图 8 所示。GT 区域表示 G 值为 1 的地面真值三元组。负区域由受试者和对象由真实对象组成，但它们之间不存在任何关系。非匹配区域表示补零区域。它与一个由候选对象组成的区域配对，这些候选对象与地面实况对象不匹配，但与等式中的 ϕ 匹配。对于每个区域，我们应用不同的技术来有效地训练具有对象检测的关系提取，如下所述。

自适应平滑。我们为 GT 区域提出了一种新颖的自适应平滑。对于属于 GT 区域的 G_{ijk}，训练模型来预测主语实体 v_i 和宾语实体 v_j 之间的第 k 个谓词类别。然而，由于分别匹配 v_i 和 v_j 的 $\hat{v}'_i$ 和 $\hat{v}'_j$ 在训练开始时对相应的地面实况对象没有足够的表示，因此将谓词的概率预测为 1 可能并不合适此外，即使当目标检测性能得到合理保证时，检测性能对于各个候选目标仍然可能有所不同。因此，我们通过自适应平滑反映关系标签上每个候选对象的检测性能。我们首先用相应的二分匹配成本来衡量每个候选对象的不确定性。对于候选对象 $\hat{v}'_i$，我们将不确定性定义如下：

$$u_i = \sigma(\mathrm{cost}_i - \mathrm{cost}_{\min} + \sigma^{-1}(\alpha)), \tag{6}$$

其中 cost_i 表示匹配成本，$\mathrm{cost}_{\min}$ 表示 $\hat{v}'_i$ 完美匹配 v_i 时的匹配成本。α 是表示最小不确定性的非负超参数。考虑到不确定性，我们将 G_{ijk} 的值设置为 $(1-u_i)(1-u_j)$。通过使用不确定性调整的关系标签，根据检测对象的质量动态调节对象检测和关系提取的多任务学习。

负采样和不匹配采样。我们不是使用所有负数，而是从负数区域中对它们进行采样。受到 Liu [18] 等人引入的硬负挖掘的启发，我们根据预测的关系得分 $\hat{G}'_{ijk}$ 对所有负数进行排序，并选择顶部的 $k_{\mathrm{neg}} \times |\mathcal{E}|$ 最具挑战性的底片。类似地，我们提取 $k_{\mathrm{non}} \times |\mathcal{E}|$ 来自非匹配区域的硬样本。由于非匹配区域通常包含图 G 的大部分，因此这种方法显着减少了稀疏性。

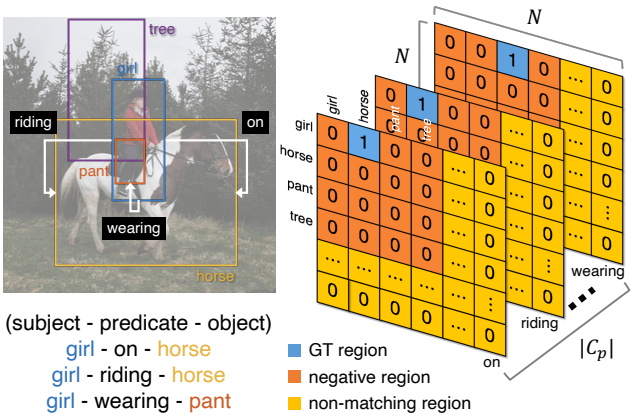

图 8. 图 G 区域示例

3.2.2 连通性预测

受预测修剪对象对相关性的 Graph-RCNN [35] 的影响，我们提出了一种连接预测，预测两个对象节点之间是否存在至少一条边以进行关系提取。我们以类似的方式得到一个连通图 $\hat{E} \in \mathbb{R}^{N\times N\times 1}$ 来得到式（1）中的关系图。我们使用另一个 $\mathrm{MLP}_{\mathrm{con}}$ 进行二元分类，而不是用于多标签预测的 $\mathrm{MLP}_{\mathrm{rel}}$。我们从置换连通图计算二元交叉熵损失如下：$\mathcal{L}_{\mathrm{con}} = \mathcal{L}_r(\hat{E}', E)$。

3.2.3 推理

对于模型推理，我们通过将谓词得分 $\hat{G}_{ijk}$ 乘以 $\hat{v}_i^c$ 和 $\hat{v}_j^c$ 的相应类得分来获得三元组得分。请注意，我们将 $\hat{G}_{iik}$ 设置为 0，以防止主语和客体是同一实体的自连接。此外，我们通过利用连通性得分 $\hat{E}_{ijk}$. 来增强三元组得分。通过乘以连通性分数，我们可以有效地过滤掉主语和宾语之间不太可能有关系的三元组。

3.3 损失函数定义

4 复现细节

4.1 与已有开源代码对比

本次实验中我直接使用了作者在 github 上开源的源代码，在源代码的基础上，根据自己的理解将代码各个模块重新梳理并且编写一遍，使代码结构更加清晰并且易懂，便于后续的学习与改进。

4.2 实验环境搭建

实验的 Pytorch 版本为 2.1.2，cuda 版本为 11.8。

4.3 创新点

在源代码的基础上，使用在主训练阶段保存的最佳模型权重，以此为起点进行训练。最终结果得到小部分提升。

	Model	# params (M)	AP50	R@20	R@50	R@100	mR@20	mR@50	mR@100
8* two-stage	IMP (EBM) [27,34]	*322.2*	28.1	18.1	25.9	31.2	2.8	4.2	5.4
	VTransE [39]	*312.3*	-	24.5	31.3	35.5	5.1	6.8	8.0
	Motifs [37]	*369.9*	28.1	25.1	32.1	36.9	4.1	5.5	6.8
	VCTree [30]	*361.5*	28.1	24.8	31.8	36.1	4.9	6.6	7.7
	VCTree (TDE) [29,30]	*361.3*	28.1	14.0	19.4	23.2	6.9	9.3	11.1
	VCTree (EBM) [27,30]	*372.5*	28.1	24.2	31.4	35.9	5.7	7.7	9.1
	GPS-Net [16]	-	-	-	31.1	35.9	-	6.7	8.6
	BGNN [15]	341.9	29.0	23.3	31.0	35.8	7.5	10.7	12.6
11* one-stage	FCSGG [17]	87.1	<u>28.5</u>	16.1	21.3	25.1	2.7	3.6	4.2
	RelTR [5]	<u>63.7</u>	26.4	21.2	27.5	-	6.8	10.8	-
	SGTR [14]	*117.1*	25.4	-	24.6	28.4	-	12.0	15.2
	Relationformer [25]	92.9	26.3	22.2	28.4	31.3	4.6	9.3	10.7
	Iterative SGG [7]	*93.5*	*27.7*†	-	29.7	32.1	-	8.0	8.8
	SSR-CNN [31]	*274.3*	*23.8*†	**25.8**	**32.7**	**36.9**	6.1	8.4	10.0
	SSR-CNN [31] $_{\text{LA},\tau=0.3}$	*274.3*	*23.8*†	18.4	23.3	26.5	**13.5**	**17.9**	<u>21.4</u>
2-11	**EGTR** (Ours)	**42.5**	**30.8**	<u>23.5</u>	<u>30.2</u>	<u>34.3</u>	5.5	7.9	10.1
	EGTR (Ours) $_{\text{LA},\tau=0.7}$	**42.5**	**30.8**	15.7	18.7	20.5	<u>12.1</u>	<u>17.8</u>	**21.7**
	EGTR (Ours) $_{\text{LA},\tau=0.5}$	**42.5**	**30.8**	19.7	24.2	26.7	11.0	17.1	<u>21.4</u>
	EGTR (Ours) $_{\text{LA},\tau=0.3}$	**42.5**	**30.8**	22.4	28.2	31.7	8.8	14.0	18.3

表 1. **Visual Genome 数据集测试集上的图约束结果**。

5 实验结果分析

我们提出了定量结果，并对我们提出的框架与代表性 SGG 模型进行了比较分析。

Visual Genome. Visual Genome 的实验结果展示在表 1。我们提出的方法展示了与当前单阶段 SGG 模型的竞争性能，该模型具有 1.5 至 6.5 倍大的参数和最快的推理速度。特别是，与最先进的方法 SSR-CNN [35] 相比，我们的方法实现了最高的目标检测性能，并且在三元组检测中显示出可比的性能。结果表明，我们的方法可以通过利用对象检测器的副产品以高效且有效的方式生成场景图。通过应用 logit adjustment [35]，一种可以很好地预测尾部谓词类的技术，我们的方法在 R@k 和 mR@k 之间的权衡中表现良好，显示出优于现有最先进模型的优越性。请注意，对于没有显式对象检测器的三元组检测模型 [9, 35]，AP50 是通过将非极大值抑制 (NMS) 应用于预测主体和对象集的并集来测量的。

Open Image V6. 如表 2 所示，在 Open Image V6 数据集上进行的实验也展示了具有竞争力的性能，强调了我们的方法在不同数据集上的有效性和鲁棒性。

6 总结与展望

本文介绍了一种名为 EGTR（Extracting Graph from TRansformer）的轻量级单阶段场景图生成模型。该模型的核心思想是充分利用 DETR（DEtection TRansformer）解码器中多头自注意力层学到的对象查询之间的关系，从而有效地提取场景图中的关系图。

Model	score	micro-R@50	wmAP_{rel}	wmAP_{phr}
Motifs [37]	38.9	71.6	29.9	31.6
VCTree [30]	40.2	74.1	34.2	33.1
GPS-Net [16]	41.7	74.8	32.9	34.0
BGNN [15]	42.1	75.0	33.5	34.2
RelTR [5]	43.0	71.7	34.2	37.5
SGTR [14]	42.3	59.9	37.0	38.7
SSR-CNN [31]	**49.4**	**76.7**	<u>41.5</u>	**43.6**
EGTR (Ours)	<u>48.6</u>	<u>75.0</u>	**42.0**	<u>41.9</u>

表 2: **Open Image V6 上的实验结果**。

EGTR 通过充分利用对象检测器的自注意力机制，为场景图生成任务提供了一种新的思路。未来的研究可以从以下几个方面进行拓展和改进：

- **更复杂的关系建模**：虽然 EGTR 在提取对象之间的关系方面取得了一定的成效，但场景图中的关系可能更加复杂多样，例如涉及多个对象之间的关系或时序关系等。未来可以探索更复杂的关系建模方法，以更准确地捕捉场景中的丰富语义.
- **跨模态信息融合**：场景图生成任务不仅依赖于视觉信息，还可以结合文本、音频等其他模态的信息进行更全面的理解和生成。例如，结合图像描述或语音指令来生成更准确的场景图，或者利用场景图来辅助图像描述生成等.
- **实时性与轻量化**：尽管 EGTR 在推理速度上表现较好，但在实际应用中，如移动设备或嵌入式系统上，对模型的实时性和轻量化要求更高。可以进一步研究模型压缩、加速等技术，使其在保持性能的同时，能够更好地适应资源受限的环境.

参考文献

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[2] Yuren Cong, Michael Ying Yang, and Bodo Rosenhahn. Reltr: Relation transformer for scene graph generation. *IEEE TPAMI*, 2023.

[3] Naina Dhingra, Florian Ritter, and Andreas Kunz. Bgt-net: Bidirectional gru transformer network for scene graph generation. In *CVPR*, pages 2150–2159, 2021.

[4] Lizhao Gao, Bo Wang, and Wenmin Wang. Image captioning with scene-graph based semantic concepts. In *ICMLC*, pages 225–229, 2018.

[5] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *CVPR*, pages 3588–3597, 2018.

[6] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, pages 3668–3678, 2015.

[7] Siddhesh Khandelwal and Leonid Sigal. Iterative scene graph generation. In *NeurIPS*, 2022.

[8] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *CVPR*, pages 6271–6280, 2019.

[9] Rajat Koner, Suprosanna Shit, and Volker Tresp. Relation transformer network. *ECCV*, pages 422–439, 2022.

[10] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017.

[11] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981, 2020.

[12] Lin Li, Guikun Chen, Jun Xiao, Yi Yang, Chunping Wang, and Long Chen. Compositional feature augmentation for unbiased scene graph generation. In *ICCV*, pages 21685–21695, 2023.

[13] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *ICCV*, pages 10313–10322, 2019.

[14] Rongjie Li, Songyang Zhang, and Xuming He. Sgtr: End-to-end scene graph generation with transformer. In *CVPR*, pages 19486–19496, 2022.

[15] Rongjie Li, Songyang Zhang, Bo Wan, and Xuming He. Bipartite graph network with adaptive message passing for unbiased scene graph generation. In *CVPR*, pages 11109–11119, 2021.

[16] Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *CVPR*, pages 3746–3753, 2020.

[17] Hengyue Liu, Ning Yan, Masood Mortazavi, and Bir Bhanu. Fully convolutional scene graph generation. In *CVPR*, pages 11546–11556, 2021.

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37, 2016.

[19] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016.

[20] Yichao Lu, Himanshu Rai, Jason Chang, Boris Knyazev, Guangwei Yu, Shashank Shekhar, Graham W Taylor, and Maksims Volkovs. Context-aware scene graph generation with seq2seq transformers. In *ICCV*, pages 15931–15941, 2021.

[21] Yukuan Min, Aming Wu, and Cheng Deng. Environment-invariant curriculum relation learning for fine-grained scene graph generation. In *ICCV*, pages 13296–13307, 2023.

[22] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. *NeurIPS*, 30, 2017.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015.

[24] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80. Association for Computational Linguistics, 2015.

[25] Suprosanna Shit, Rajat Koner, Bastian Wittmann, Johannes Paetzold, Ivan Ezhov, Hongwei Li, Jiazhen Pan, Sahand Sharifzadeh, Georgios Kaissis, Volker Tresp, et al. Relationformer: A unified framework for image-to-graph generation. In *ECCV*, pages 422–439, 2022.

[26] Gopika Sudhakaran, Devendra Singh Dhami, Kristian Kersting, and Stefan Roth. Vision relation transformer for unbiased scene graph generation. In *ICCV*, pages 21882–21893, 2023.

[27] Mohammed Suhail, Abhay Mittal, Behjat Siddiquie, Chris Broaddus, Jayan Eledath, Gerard Medioni, and Leonid Sigal. Energy-based learning for scene graph generation. In *CVPR*, pages 13936–13945, June 2021.

[28] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse r-cnn: End-to-end object detection with learnable proposals. In *CVPR*, pages 14454–14463, June 2021.

[29] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, pages 3716–3725, 2020.

[30] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *CVPR*, pages 6619–6628, 2019.

[31] Yao Teng and Limin Wang. Structured sparse r-cnn for direct scene graph generation. In *CVPR*, pages 19437–19446, 2022.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[33] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. *NeurIPS*, 31, 2018.

[34] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, pages 5410–5419, 2017.

[35] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *ECCV*, pages 670–685, 2018.

[36] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, pages 10685–10694, 2019.

[37] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, pages 5831–5840, 2018.

[38] Cheng Zhang, Wei-Lun Chao, and Dong Xuan. An empirical study on leveraging scene graphs for visual question answering. In *BMVC*, 2019.

[39] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, page 3107–3115, July 2017.

[40] Chaofan Zheng, Xinyu Lyu, Yuyu Guo, Pengpeng Zeng, Jingkuan Song, and Lianli Gao. Learning to generate scene graph from head to tail. In *ICME*, pages 1–6. IEEE, 2022.

[41] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021.