arXiv:2303.01080v1 [cs.CV] 2 Mar 2023

LANDMARK：语言引导的场景图生成表征增强框架

Xiaoguang Chang[1], Teng Wang[2],
ShaoweiCai[2]和 Changyin Sun[2,3*]

[1]网络科学与工程学院，东南大学，中国南京。[2*]自动化学院，东南大学，中国南京。[3*]人工智能学院，安徽大学，中国合肥。

通讯作者。电子邮箱：cysun@seu.edu.cn；贡献作者：
xg chang@seu.edu.cn；wangteng@seu.edu.cn；
shaoweicai@seu.edu.cn；

摘要

场景图生成（SGG）是一项复杂的任务，既面临复杂视觉特征的挑战，也受数据集长尾问题的影响。近期，研究者通过设计新颖的损失函数和数据平衡策略，提出了多种无偏策略。遗憾的是，从特征细化的角度来看，这些无偏方法未能强调语言先验。受谓词与主客体对中蕴含的语义以及全局上下文高度相关这一事实的启发，我们提出了 LANDMARK（**LAN**guage-gui**D**ed representation enhanceMent fr**A**mewo**RK**），该框架从语言-视觉交互模式、全局语言上下文和主语-宾语对与谓词的相关性中学习与谓词相关的表征。具体而言，我们首先将对象标签映射为三种不同的语义嵌入，以用于不同的表征学习。随后，语言注意力模块（LAM）和经验估计模块（EEM）分别将主客体的词嵌入处理为注意力向量和谓词分布。语言上下文模块（LCM）从每个词嵌入中编码全局上下文，从而避免孤立地从局部信息中学习。最后，利用各模块的输出更新视觉表征和场景图生成模型的预测。所有语言表征均完全生成

自物体类别，因此无需额外知识。该框架与模型无关，并能持续提升现有场景图生成模型的性能。此外，表征层面的无偏策略使 LANDMARK 具备与其他方法兼容的优势。代码可在此处获取：https://github.com/rafa-cxg/PySGG-cxg。

关键词： 场景图生成、无偏方法、视觉-语言表示学习、多语义

1 引言

场景图生成（SGG）是一项至关重要的任务，有助于图像描述生成 [1, 2]，视觉问答 [3–5]，以及视频理解 [6]。然而，大多数生成的场景图都面临预测过于简单的问题，因此迄今为止尚未应用于实际场景。

因此，近期研究一直致力于开发无偏方法，以提高对难以区分的谓词的召回率。一般而言，无偏方法可分为三类：数据重采样（例如 BLS [7], GCL [8]）、谓词感知损失设计（例如 CogTree [9] 和 FGPL [10]）以及逻辑值操控（例如 TDE [11], RTPB [12], FREQ [13]）。然而，这些方法有一个共同的缺点：它们依赖于根据数据集统计信息 [10, 13] 或有偏的模型预测 [10, 11]，显式建模谓词相关性或损失权重，因此对先决条件的变化十分敏感。例如， [11] 在使用无偏模型进行训练时效果不佳，从而限制了场景图生成模型的性能。与基于损失和统计信息的方法相比，语言表示学习更加稳健，因为它能够学习谓词的隐式模式，并避免视觉特征冗余；此前无偏方法尚未重视这一点。

然而，一些基线模型已采用语言表示学习。例如， [14] 使用词嵌入，使注意力聚焦于视觉特征。 [8] 利用交叉注意力（CA）机制进行多模态学习。[15]引入基于Transformer的架构，以弥合图像与文本之间的差距。然而，这些方法中的大多数并非即插即用，而且无论语义上下文如何变化，都仅使用单一的语言表示。因此，它们未能释放语言的力量。

事实上，词语有多种含义，在不同语义层面承载着不同的先验，这些先验能够指导场景图生成。这里，我们在图 1 中给出一个多语义推理示例。给定这张购物图片，首先，人类根据谓词与 女性-球 这一对之间的相关性，以及它们的相对位置，构建谓词分布。这种知识源于经验，其形成过程独立于视觉。接下来，仍以 女性-球 为基础，人类可以建立主客体对与视觉模式之间的关联，例如“女性的手靠近球”，这是一种与“持握”密切相关的强视觉模式。相比之下，女性和女孩的轮廓对于这对客体而言无关紧要

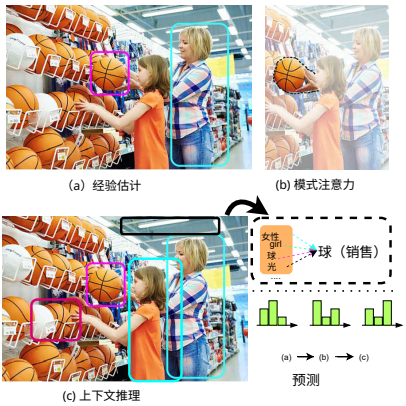

图 1 用于关系预测的多语义语言辅助示例。右下角展示了 3 个候选项可能性更新的过程。**a)经验估计**：人根据谓词与客体对的共现可能性，回忆出一个粗略的谓词分布。**b) 模式注意力**：利用客体对与视觉信息之间的内在关系，定位与谓词相关的视觉模式。**c) 上下文推理**：利用上下文细化主体和客体的含义，避免因孤立地考虑客体对而造成的预测偏差。右下角展示了可能性在 3 个过程中如何变化

对于这对客体而言。最后，根据周围的物体（例如球、女孩、光线），人类可以推断出售卖 上下文，从而避免预测玩耍，因为这不符合该场景。

受这些观察的启发，我们通过启发式方法设计了 3 个即插即用语言模块，以利用物体类别背后不同的语言先验。三个模块以检测到的物体类别为输入，并在不同的语义空间中生成语义嵌入，分别用于从语言-视觉模式相关性、语言上下文和主语-宾语对与谓词的相关性中提取先验。具体而言，**1) 语言注意力模块**将主语和宾语的词嵌入投影到统一的语义矩阵中，然后利用通道注意力为关系视觉特征图提取注意力向量，从而学习客体对与特定谓词相关视觉模式之间的相关性。2)**语言上下文模块**采用基于 Transformer 的编码器，根据实体标签序列将全局语言上下文编码为实体的语义嵌入。与预训练词嵌入相比，该模块能生成符合上下文的语义表示。这两个模块分别用于初始化实体和关系的视觉表示。3) **经验估计模块**以主体-谓词和客体-谓词的边缘概率为监督信号，学习类别与空间感知谓词分布，并将其作为似然偏移。值得一提的是，语言处理从一开始就与视觉特征解耦，因此该框架适用于大多数场景图生成基线模型。

据我们所知，我们是首个利用对象标签中的多语义语言表示来实现无偏场景图生成的研究。主要贡献可概括如下：

1. 我们提出了 LANDMARK，将语言表示学习引入无偏场景图生成，着重探索对象标签中尚未充分研究的多重语义利用。2. 我们设计了三个模块，将对象标签划分到不同的语义空间，进而分别提取语言-视觉交互模式的先验、语义上下文以及主语-宾语对与谓词的相关性。3. 在场景图生成基准上的实验表明，该方法持续提升了基线模型的表现，并且与其他无偏方法兼容，说明从对象标签中诱导出的多语义语言表示是有效的。

2 相关工作

场景图生成： 场景图生成主要有两种主流方法：基于上下文建模或图卷积网络（图卷积网络） [16]。第一种方法侧重于对全局信息进行建模，并广泛采用在所有实体特征之间进行消息传递 [17–21]。其中许多方法 [8, 12, 13, 15, 22] 采用了序列模型，例如长短期记忆网络[23], 门控循环单元 [24] 和 Transformer [25]。Chen 等人 [12] 使用两层 Transformer 堆栈对全局信息进行编码。Zellers 等人 [13] 使用长短期记忆网络编码全局上下文，以辅助关系预测。然而，仅对全局上下文进行建模不足以应对场景图任务。另一种方法 [7, 26–28] 在节点与边特征之间传递消息，并更加关注区域成对信息。 [7] 采用多阶段图消息传播，在候选实体与关系表示之间进行传递。在 [27], Yang等人将图连接剪枝为稀疏形式，然后应用注意力图卷积网络来调节信息流。 [26] 使用图卷积网络将状态表示更新为能量值。Chen等人[29] 在候选实体与所有关系表示之间构建图，并通过门控循环单元聚合消息。然而，这种方法的全局上下文编码不足。我们的方法同时考虑成对上下文和全局上下文，以进行表征细化。

语言先验学习： 实体标签和关系分布是场景图生成中最常用的先验知识。在无偏方法方面，语言知识可以通过损失权重或预测偏置体现。例如，FREQ [13]直接统计数据集中“主语-谓词-宾语”的共现情况，并将其作为预测偏置。Chen 等人 [12] 通过不同方式对数据集中的客体关系进行建模，设计了不同类型的特定偏置。CogTree[9] 基于有偏场景图生成预测中关系的自动构建认知结构，提出了一种损失函数。上述方法在mRecall@k[30], 指标上显著提升了特定基线模型的表现，但当前提条件发生变化时，它们并不稳定。

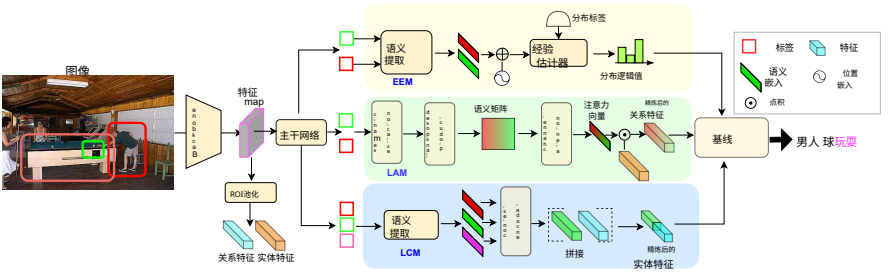

图2 LANDMARK架构。图像首先经过通用目标检测器，获取预测的客体标签和ROI特征。接着，将标签作为三个模块（即 EEM、LAM、LCM）的输入，并计算差异化语义。最后，分别使用 LAM 和 LCM 的输出来优化关系表示和实体表示。EEM 的输出则作为预测偏移量。

在基线模型方面，越来越多的研究致力于多模态表示学习。 [31] 设计了一个语言模块，将表示映射到与视觉模块相同的空间，以缩小相似语义特征之间的距离。 [30] 引入了置信度估计模块，通过在节点特征更新中纳入置信度估计来减轻错误传播。 [32] 通过统一场景图和常识图的表述方式，融入外部常识知识。[33] 对视觉常识进行结构化，并提出级联融合架构以实现融合。然而，现有方法将语言模态视为单一表示，因而丢失了大量信息。

3 方法

问题定义： 给定图像 I，场景图生成旨在预测实体类别集合 $\mathcal{C}_e$、协调集合 $\mathcal{B}_e$ 和关系集合 $\mathcal{C}_r$。通常，现有的场景图生成模型从主干网络获取视觉实体和谓词（或节点和边）的表示。随后，可以构建图 $\mathcal{G} = \{\mathcal{C}_e, \mathcal{B}_e, \mathcal{C}_r\}$ 。

$$\mathcal{G} = P(\mathcal{C}_r, \mathcal{B}_e, \mathcal{C}_e | I) = SGG(N, E) \tag{1}$$

其中， $N = \{e_i\}_{i=1}^{n}$ 和 $E = \{e_{ij}\}$ 分别是实体表示和谓词表示的集合。本文旨在通过引入语义先验来更新 N 和 E 。

框架概述： LANDMARK由三个语义学习模块组成，即语言注意力模块(LAM)、语言上下文模块(LCM)和经验估计模块 (EEM)。框架架构如图2所示。首先，我们从ROI池化中获得 N， E ，并从分类器头中获得 $\mathcal{C}_e$， $\mathcal{B}_e$ 。随后，语义提取操作将标签转换为$\{c_i\}$ 每个模块对应的语义嵌入。对于LAM，主体 c_i 和客体 c_j 的语义嵌入经过转置后与语义矩阵相乘，然后通道注意力将该矩阵转换为注意力向量，并

更新关系表示。LCM对图像中所有实体标签 $\mathcal{C}_e$ 对应的语义嵌入进行编码，生成上下文感知语义实体特征，并将其与视觉实体表示拼接。更新后的实体表示和关系表示将输入基线模型。对于EEM，生成分布标签以监督经验估计器；该估计器将主客体语义嵌入与位置嵌入相结合，从而生成分布逻辑值。最后，使用生成的逻辑值更新最终的谓词似然。

3.1 语义提取

语义提取用于将标签映射到相应的语义空间，并分别应用于三个模块。具体而言，语义提取器包含三个操作：

$$
\begin{aligned}
f_{se}^{sub}(c_i) &= w_s^T c_i \\
f_{se}^{obj}(c_j) &= w_o^T c_j \\
f_{se}^{ent}(c_e) &= w_e^T c_e
\end{aligned} \tag{2}
$$

第一和第二个操作分别用于 LAM 和 EEM 中的主体投影与客体投影。考虑到同一个词作为主体或客体时可能具有对比性的含义（例如，当“男人”是主体时，吃东西可能是谓词；而当“男人”是客体时则不可能），我们使用不同的权重 w_s 和 w_o 将主体和客体映射为语义嵌入。最后一个操作用于 LCM；由于所有标签都被视为客体，因此我们使用统一的w_e 作为语义嵌入权重。

事实上，只要输入是对象标签， f_{se} 可以是任意投影函数。在这里，我们仅使用一个朴素的单层线性函数来证明提取的有效性。

3.2 语言注意力模块

该模块旨在学习客体对与视觉关系表示中视觉谓词相关模式之间的先验 e_{ij}。原始特征提取网络（例如 Reset [34]）保留了空间和语义信息。具体而言，不同通道关注不同的视觉模式。然而，关系视觉特征不可避免地混入大量无关的背景信息，因此需要进行通道选择。根据经验，对于给定的特定主客体对（例如，男孩-篮球 或 男孩-街道），视觉特征应具有不同的激活模式。因此，我们设计了一种标签感知通道注意力机制。具体而言，给定主体 i 和客体 j，我们首先生成一个语义矩阵 x_{ij} ，作为词语-视觉相关性的独特表示。

$$
x_{ij} = f_{se}^{sub}(c_i) \otimes f_{se}^{obj}(c_j)^T \tag{3}
$$

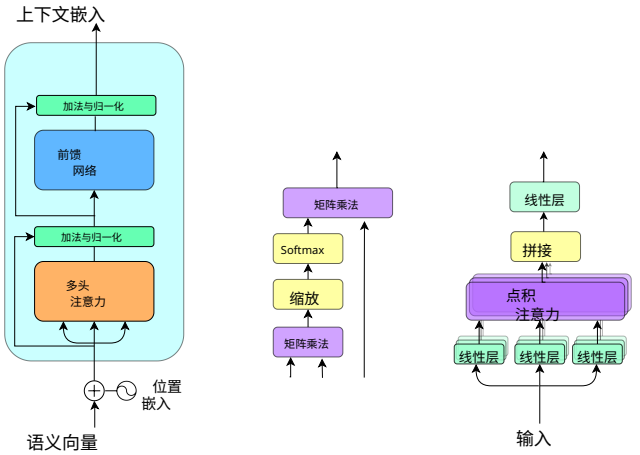

图 3 上下文编码器模块结构（左）。点积注意力（中）。多头自注意力（右）。

其中 $\otimes$ 表示矩阵乘法。我们通过对 x_{ij} 进行空间池化，并使用一系列二维卷积来获得注意力向量 e_{ij}^{c}，从而实现通道注意力：

$$e_{ij}^{c} = \sigma(G_{\text{pooling}}(G_{\text{conv}}^{n_c}...\sigma(G_{\text{conv}}^{1}(x_{ij})))) \in \mathcal{R}^{C,1} \tag{4}$$

其中，C 表示视觉关系特征 e_{ij}的通道数，n_c 表示二维卷积层的数量，G_{pooling} 表示池化操作，σ 表示激活函数。最后，通道权重 e_{ij}^{c} 将用于更新 e_{ij}，从而抑制与关系判别无关的通道。

$$\hat{e}_{ij} = e_{ij} \times e_{ij}^{c} \tag{5}$$

其中，$\hat{e}_{ij}$ 表示精炼关系表示，$\times$ 表示点积算子。

3.3 语言上下文模块

与视觉信息相比，单个词在语义上与句子中的其他成分相互孤立。尽管我们设计了 LAM 和 EEM 用于语义提取，但所使用的成对标签仅限于局部语义，不足以进行全面的语义推理。因此，LCM 旨在解决全局上下文缺失问题。该模块包括语义提取器和上下文编码器。上下文编码器由多层 Transformer 编码器构成，其中包含多头自注意力（MHSA） [25] 和前馈网络（FFN） [25]。其结构如图3所示。具体而言，给定一幅图像，假设其中有 n 个实体，输入序列 X 可表示如下：

$$X = \{s_0, s_1, ..., s_n\} \tag{6}$$

其中

$$s_i = \gamma(f_{se}^{e} c_i + p_i) \in \mathbb{R}^{d} \tag{7}$$

其中， c_i 和 $p_i = \phi[x_i, y_i, w_i, h_i]$ 分别表示第 i个实体的类别标签和实体位置嵌入。 x_i、 y_i、 w_i分别是客体 i 的中心坐标、宽度和高度。 γ 表示可学习线性变换。其中， d 是序列中每个元素的维度。我们首先重述标准的缩放点积注意力 [25]。

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

接下来，多头自注意力定义如下:

$$\begin{aligned} \text{MHSA}(X) = &\ \text{Concat}\ (\text{head}_1, \ldots, \text{head}_\text{h})\, W^O \\ \text{head} = &\ \text{Attention}\ \left(XW_i^Q, XW_i^K, XW_i^V\right) \end{aligned} \tag{9}$$

其中 W_i^Q， W_i^K， W_i^V和 W^O是参数矩阵。第 b 层的输出 X_b 可表示为:

$$X_b' = \text{MHSA}(\text{LN}(X_{b-1})) + X_{b-1}, \tag{10}$$

$$X_b = \text{FFN}(\text{LN}(X_b')) + X_b', \tag{11}$$

其中 X_{b-1}是第 $b-1$层，我们设定 $X_0 = X$。LN 是层归一化。不同于以往的工作 [35][12], ，LCM 以实体标签序列作为输入，因此该模块可以在高层信息的基础上学习语义实体表示。

3.4 经验估计模块:

该模块旨在通过监督学习从主体-客体对中学习关系分布先验，以补偿交叉熵损失。EEM由语义提取器、经验估计器和用于监督的分布标签生成模块组成。由于实体类别和位置都会影响关系判断，该模块利用类别和位置信息来学习精确的分布。首先，我们将实体标签 i、j 和位置嵌入p_{ij} 映射到高维表示空间，然后，经验估计器预测主体 i 与客体 j 之间的谓词分布 d_{ij} 。

$$p_{ij} = \phi_p[x_i, y_i, w_i, h_i, x_j, y_j, w_j, h_j] \tag{12}$$

$$d_{ij} = \varphi\left[\phi_s(f_{se}^{sub}(c_i))\phi_o(f_{se}^{obj}(c_j)), p_{ij}\right] \tag{13}$$

其中， $[\cdot,\cdot]$ 表示拼接操作， ϕ_p， ϕ_s， ϕ_o， φ 是以 ReLU 作为激活函数的全连接层。最后，我们将关系分布 d_{ij} 与基线模型的预测结果合并，可表示如下:

$$\hat{d}_{i,j} = SGG(\hat{n}_{ij}, \hat{e}_{ij}) + d_{ij} \tag{14}$$

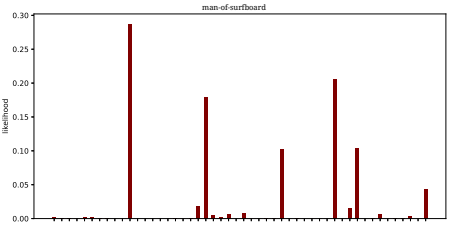

图 4 主体 man 与客体 $surfboard$的联合可能性。真实三元组⟨man,of,surfboard⟩ 采样自零样本划分（其中三元组仅出现在评估集中，而未出现在训练集中）。

其中，$d_{i,\ j}$ 是更新后的预测似然。$\hat{}n_{ij}$ 和 $\hat{e}_{ij}$ 是通过 LAM 和 CAM（第 3.2、3.3节）获得的增强实体和关系特征。

分布标签生成： 数据集标注本质上反映了人类常识，因此我们能够从数据集中生成准确的分布标签。对于主体 i 和客体 j，我们获得 m_i^{sub} m_j^{obj} “主体-谓词”和“谓词-客体”的边缘分布。由于 m_i^{sub}，m_j^{obj} 是相互独立的分布，我们计算联合可能性 p_{ij}^{joint} 如下：

$$p_{ij}^{joint} = m_i^{sub} \odot m_j^{obj} \tag{15}$$

其中 $\odot$ 表示逐元素乘积。尽管 EEM 类似于 FREQ [13]，都是从统计数据生成谓词分布，但 FREQ 直接统计三元组 $\langle subject, predicate, object\rangle$ 的出现次数。然而，一些三元组样本在训练样本中十分稀少，因此很难建立具有信息量的分布先验。相比之下，EEM 使用联合可能性作为标签来推断谓词分布。图4 展示了一个典型的联合可能性生成示例 p_{ij}^{joint}，对应一个未出现在训练集中的三元组。在这种情况下，由于样本数为零，FREQ 无法发挥作用；而我们可以看到，可能性最高的前 5 个结果是合理的，并且包含许多可能的场景。相比之下，谓词“of”具有歧义性。更好的谓词可能是联合可能性中可能性最高的结果之一。

考虑到引入位置信息有助于准确预测谓词，我们设计了一种融合函数，以减轻联合可能性与真实谓词标签的影响。分布标签 l_{ij} 可表示为：

$$l_{ij} = \mu \times p_{ij}^{joint} + (1-\mu) \times \text{onehot}[r_{ij}] \tag{16}$$

其中，μ 是用于调节边缘频率占比的因子。r_{ij} 表示主体 i 与客体 j之间的真实谓词标签。

目标函数：我们选择分布标签 $l_{i,j}$ 与预测分布 $d_{i,j}$之间的均方误差损失，记为：

$$\mathrm{MSE} = \frac{1}{K}\sum_{i=1}^{K}(d_i, l_i)^2 \tag{17}$$

其中，K 是数据集中关系类别的数量。

3.5 基线模型

任何现成的两阶段场景图生成模型都可以用作基线。它可以是序列模型，也可以是图神经网络，只要利用实体和关系特征进行预测即可。

4 实验

本节首先介绍实验中的实验设置。随后，我们在多个场景图生成模型上测试所提框架的有效性，并通过实验分析我们的方法与当前最先进的无偏策略之间的兼容性。最后，我们从不同角度进行详细分析和定量评估，进一步验证 LANDMARK。

4.1 实验设置

数据集 我们采用广泛使用的 Visual Genome [36] 筛选后的子数据集 VG 150 [37]来训练和评估所提框架。VG150 数据集包含 VG 中出现频率最高的 150 个物体类别和 50 个谓词类别。该数据集由超过 10.8 万张图像组成，其中 70% 用于训练，30% 用于测试。在训练集中，5000 张图像用于评估。

任务。 我们考虑场景图生成的三种常见子任务来评估所提框架。1）谓词分类（*PredCls*）在给定真实边界框和类别的情况下，预测每对客体之间的关系。2）场景图分类（*SGCls*）在给定客体真实边界框的情况下，预测客体类别及其关系。3）场景图检测（*SGDet*）需要检测客体类别和边界框，并预测它们之间的关系。

评估指标我们报告了广泛认可的指标平均召回率@k [29]和召回率@k [30]，用于评估模型性能。召回率@k计算在前 k 个关系预测中正确预测关系的比例。然而，R@K 无法反映模型在尾部关系类别上的表现，因此，使用各类别上的召回率（即平均召回率）来评估无偏性能。此外，为了评估 EEM，我们需要一种衡量分布准确率的指标，因此采用了 TOP-N Recall@K 指标，

该指标通过将单次预测中得分最高的 N 个谓词作为候选项来提出，然后使用 $N \times K$ 个候选项计算召回率，即

$$\text{TOP-N Recall@K} = \frac{correct(\{\text{ candidates}\}_{N\times K})}{N_{gt}}$$

当 N=1时，TOP-N Recall@K 等于 Recall@K。TOP 召回率与无图约束的召回率之间的区别在于 [38] 前者会强制保留得分最高的 $N\times K$ 谓词，从而避免候选项因其他预测中的得分更高而被排除。

实现细节 我们使用预训练的 Faster R-CNN [39]，其骨干网络为 ResNeXt-101-FPN [40]，并使用 ROIAlign [40] 作为目标检测器。训练期间，我们冻结其权重。对于我们的框架，我们使用预训练的 GloVe [41]权重作为语义提取器中 w_s的初始 w_o值 w_e。对于 EEM，我们使用 3 层多层感知机 Φ_s，Φ_o，Φ_p 包含 1024 个神经元；φ 是一个隐藏维度为 4096 的 2 层多层感知机。μ 式中提及的16 对于无偏方法设为 0.3，对于基线模型设为 0.7。对于式4，我们选择两个 3×3 卷积层来生成一个 256 通道的注意力向量。对于 LCM，我们选择具有 4 层和 8 个注意力头的上下文编码器，实体维度为 $d = 512$。训练时，每个基线模型大约迭代 10000 次即可。基础学习率为 0.01，批量大小为 16。我们选择随机梯度下降优化器进行优化。

4.2 与基线方法的比较

表 1展示了采用或未采用我们的框架 LANDMARK 时，5 种基线模型的平均召回率和召回率。这些基线模型包括基于图卷积网络的模型：G-RCNN [27], BGNN [7] 以及上下文建模网络：IMP [37],Trans-former [25], Motifs [13]。值得一提的是，我们没有在这些模型上采用任何无偏策略。我们观察到，将所提出的 LANDMARK 纳入所有基线模型后，在三个任务上均能持续提升平均召回率，这证明了我们方法的鲁棒性。对于 mR@100，我们的模型在三个任务上的平均提升分别为 3.66%、2.64% 和 1.37%。这些提升可能是因为多语义语言表征确实有助于视觉表征。三个任务的平均提升依次缩小并不令人意外，这是由于预训练目标检测器预测的类别和位置不准确。此外，召回率也出现了不同程度的下降，这是无偏方法的一个常见特征。

我们还在表1中测量了基线和 LANDMARK 的总参数量（M）及计算开销（GFLOPs）。总体而言，参数量增加了 20M（相对增幅：5.3%），计算量增加了 2.5GFLOPs（相对增幅：1.0%）。与 BGNN+TDE（365.7M 和 213.2GFLOPs）相比，两者的模型规模和计算开销都更小。这得益于轻量化模块设计以及采用低维输入（即语言输入而非图像输入）。

表 1 VG数据集上场景图生成的平均召回率和召回率（%）。* 表示未采用论文中的默认无偏策略。所有基线方法均在我们的代码库上重新实现。

模型	方法	PredCls		SGCls		SGDet		参数(M)/GFLOPs
		mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100	
IMP [37]	基线	16.88/18.02	66.8/68.25	7.57/8.08	38.9/40.17	6.00/7.30	27.24/34.24	336.3/206.4
	LANDMARK	**19.54/21.06**	64.89/66.61	**9.89/10.49**	33.63/34.89	**6.47/8.00**	24.34/29.05	356.5/208.8
Transformer [25]	基线	19.13/20.3	65.59/67.21	10.25/10.72	39.3/40.5	7.99/9.68	26.91/31.26	330.6/205.6
	LANDMARK	**22.19/23.87**	65.48/66.92	**11.94/12.84**	36.62/37.73	**8.23/10.43**	**28.76/32.70**	348.5/207.8
G-RCNN [27]	基线	16.46/17.28	66.27/67.9	9.67/10.17	39.3/40.5	4.89/5.95	30.37/34.47	366.1/207.1
	LANDMARK	**19.24/20.51**	65.76/67.50	**11.14/11.62**	**42.80/43.55**	**5.52/7.21**	24.61/29.69	386.4/209.6
Motifs [13]	基线	18.79/19.69	66.17/67.72	9.39/9.97	41.57/42.7	6.24/7.54	29.82/33.65	367.1/211.5
	LANDMARK	**22.37/23.79**	59.34/61.16	**14.36/15.25**	33.57/35.02	**7.87/10.56**	24.65/28.87	389.8/214.4
BGNN [7]	基线	17.26/18.29	66.14/67.65	10.30/10.83	39.94/41.17	6.46/8.22	30.90/35.36	341.9/205.0
	LANDMARK	**20.35/22.63**	55.10/57.16	**11.64/12.79**	33.3/34.89	**7.34/8.97**	24.34/29.05	360.1/207.2

4.3 与无偏方法的兼容性

无偏场景图生成策略最棘手的问题是，大多数策略对应用场景要求较高。因此，我们通过将两种方法叠加，测试框架与其他无偏方法的兼容性；平均召回率和召回率列于表2中。这些策略属于不同类型，例如，数据重采样：BLS [7], 逻辑值操控：例如 TDE [11], 以及特征细化：LANDMARK（本文方法）。根据该表，可以得出以下几点发现：

• 将 LANDMARK 应用于其他方法是可行的。例如，将 BLS 与 LANDMARK 结合应用于 BGNN，可获得新的最先进性能。原因有两点：1）LANDMARK 具有独特的语义特征增强策略，不会与其他方法冲突。2）大多数无偏方法旨在从有偏前提（例如基线、长尾数据集）中获取先验，而 LANDMARK 与基线无关，也就是说，基线不会影响 LANDMARK 的推理。

• 使用 BLS 和 TDE 的组合仅带来了微小提升，甚至导致 mR@K 下降。例如，Motif+BLS+TDE 会使 mR@K 明显下降。这表明这些方法对外部环境的变化（例如采样分布、基线模型能力）较为敏感，

y).

• 我们的网络不会大幅牺牲召回率。例如，在 Motifs 上，BLS+LANDMARK 的召回率高于不使用 LANDMARK 的情况。然而，使用 BLS+TDE 会显著损害召回率性能。我们推测，现有的无偏方法并未探索谓词之间真实的差异，而只是在提高

y 尾部谓词的 g 可能性。

4.4 分析每个谓词：

如图 5所示，我们展示了 VG 数据集的类别分布，以及 Predcls 任务上 Recall@100 的提升（BGNN+BLS，使用或不使用我们的方法）。在全部 50 个谓词中，基线预测结果仅有 11 个谓词优于 LAND- MARK。此外，LANDMARK 还召回了 6 个难以区分的谓词（即 bitam、属于、在……上行走、安装在……上、由……制成、说）。

表 2 % 下无偏方法与 LANDMARK 的兼容性测试。* 表示该基线模型和无偏方法是在同一篇论文中提出的。所有方法均在我们的代码库上重新实现。

方法	PredCls		SGCls		SGDet	
	mR@50/100	R@50/100	mR@50/100	R@50/100	mR@50/100	R@50/100
$\text{Motifs}_{\text{TDE}}$	22.87/25.90	35.28/41.11	13.76/15.36	23.88/27.42	7.87/9.55	9.83/13.06
$\text{Motifs}_{\text{BLS}}$	30.64/32.76	54.56/56.27	20.21/21.09	34.11/35.00	13.20/15.97	25.27/28.88
$\text{Motifs}_{\text{BLS+TDE}}$	26.19/30.84	17.01/18.89	17.31/19.28	19.63/21.43	9.18/11.89	8.34/10.27
$\text{Motifs}_{\text{TDE+LANDMARK}}$	24.59/29.27	39.24/45.25	14.43/17.03	28.70/32.28	8.98/11.35	9.92/12.85
$\text{Motifs}_{\text{BLS+LANDMARK}}$	**33.62/35.69**	**55.29/56.80**	**21.43/22.05**	**36.02/36.85**	**13.86/16.31**	25.10/**29.12**
BGNN_{TDE}	19.23/21.59	32.75/36.45	13.09/14.18	28.66/31.20	8.65/11.01	20.36/26.00
$\text{BGNN}^{*}_{\text{BLS}}$	30.64/32.76	55.39/56.8	16.69/17.83	35.70/36.94	13.17/15.56	23.54/27.56
$\text{BGNN}_{\text{BLS+TDE}}$	31.88/34.89	30.37/33.33	16.97/18.70	21.62/23.52	11.60/14.32	19.45/24.11
$\text{BGNN}_{\text{TDE+LANDMARK}}$	20.32/22.78	50.21/53.85	14.56/15.58	35.31/**37.38**	9.68/12.41	**25.72**/27.98
$\text{BGNN}_{\text{BLS+LANDMARK}}$	**34.41/36.43**	53.99/55.42	**18.27/19.12**	34.55/35.80	**14.08/18.34**	24.87/**29.41**

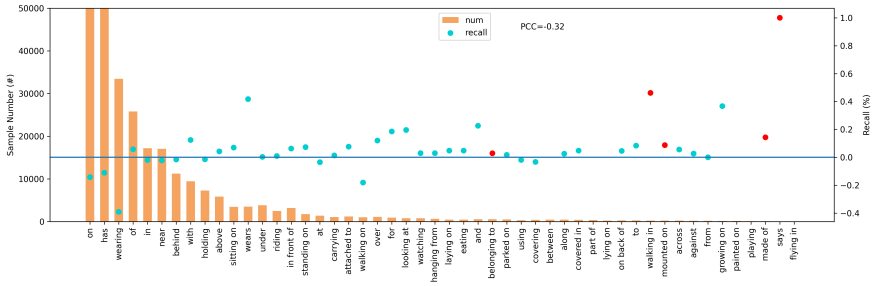

图5 LANDMARK在BGNN+BLS上各谓词的采样分布和Recall@100提升（Predcls任务）。红点表示BGNN+BLS未能……。PCC是皮尔逊相关系数的缩写

为探究 Recall 提升与数据分布之间的相关性，我们使用了皮尔逊相关系数（PCC）。PCC=0.32 表明数据集偏差与 LANDMARK 的提升之间存在弱负相关。值得注意的是，一些无偏方法会对尾部类别进行过采样，因此呈现出较强的负相关（例如，TDE：PCC=-0.56）。此外，BGNN 上数据分布与 LANDMARK 或 BLS 之间的 PCC（-0.1 对比 -0.3）证明，两者的提升和预测对长尾数据均具有鲁棒性。

4.5 消融研究

我们通过逐步向 BGNN+BLS 添加 EEM、LAM 和 LCM，来研究 LANDMARK 的各个组成部分。表 3 中的结果表明：1）每个组成部分都有助于整个框架，且彼此之间不存在冲突。这证明这三个模块能够从相同的标签输入中提取差异化语义。2）EEM模块对 Predcls 的提升主要优于其他任务，这可能是因为位置信息和物体标签预测不够准确。3）LCM 和 LAM 持续提升各项任务的性能。这是因为语言上下文提供的先验，以及词语-视觉模式之间的相关性，对于被错误分类但语义相近的语义物体标签而言相对稳健。

表3 三项任务中框架结构的消融研究

方法			PredCls	SGCls	SGDet
EEM	LAM	LCM	mR@50/100	mR@50/100	mR@50/100
			30.64/32.76	16.69/17.83	13.17/15.56
✓			31.75/33.00	17.27/18.18	13.54/15.92
✓	✓		32.06/33.36	17.75/18.89	14.34/16.42
	✓	✓	34.10/36.07	18.07/19.01	13.89/17.00
✓	✓	✓	**34.41**/**36.43**	**18.27**/**19.12**	**14.08**/**18.34**

表4 EEM模块与FREQ模型在基线Motifs上的比较。

方法	PredCls	SGCls	SGDet
	Top 1/5 R@100	Top 1/5 R@100	Top 1/5 R@100
FREQ [13]	14.72/**42.47**	9.71/27.56	7.66/18.32
EEM	**22.74**/40.38	**14.76**/**29.14**	**14.13**/**23.43**

4.6 经验估计模块分析：

如前所述，语言注意力模块会独立输出谓词预测，这与频率基线（FREQ）类似 [13]。因此，我们评估了在 BGNN+BLS+上训练的 EEM 和 FREQ 在评估数据集划分上的表现，结果见表 4。除 Predcls 任务上的 Top-5 外，EEM 的所有性能指标均不及 FREQ；随着任务难度增加，两者之间的差距也随之扩大。这归因于边际概率提供的监督缓解了稀有主客体样本带来的不足。此外，EEM 引入的位置信息使其在推断相同客体对时更加准确。

4.7 语言注意力模块分析：

为了验证 LAM 模块的有效性，我们将基于 BGNN$_{ij}$ （式 5）训练得到的关系表示 ˆ 的热力图可视化，+BLS+LANDMARK 在图 6 中使用真实或随机生成的主客体对（红色文字）进行可视化。直观来看，注意力区域与给定的主体和客体相关。例如，在最左侧的两幅图像中，当客体从 鞋子变为衬衫时，注意力区域从脚部转移到男人身体的中部。而在最右侧两幅图像中，给定的词语并不相关，例如 时钟-雪，模块似乎关注左上方和底部区域，这表明 LAM 无需给定坐标，就能将词语与相关视觉模式关联起来。

4.8 EEM 因子评估

图 7 展示了 4 个模型在 SGDet 任务上采用不同 EEM 因子时的 mRecall@100 结果 μ （见公式 16）。我们发现，对于基线模型，因子=0.7 更为合适；对于无偏方法，0.3 则更好。这表明，计算得到的边缘概率主要记录了多样化的可能谓词，而真实标签更可能是高频谓词。

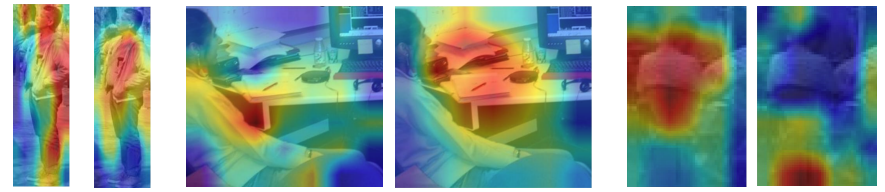

图 6 LAM 评估。我们将给定的真实标注主客体对与随机泛化对（红色文字）的并集区域进行可视化，展示词对与特定视觉特征之间的联系。

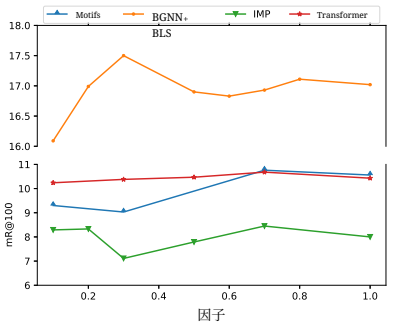

图 7 在 SGDet 任务上，具有不同 EEM 因子的场景图生成模型的 mR@100 性能。

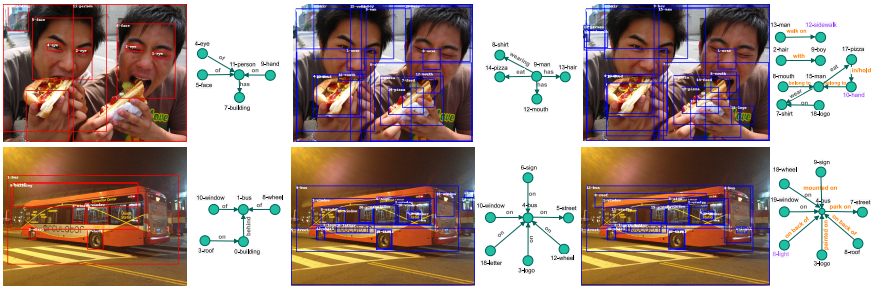

图 8 可视化结果：我们分别在三列中展示由标注、采用 BLS 的 BGNN（无偏模型）和 LANDMARK 生成的场景图。标注和基线中均未出现的关系和实体以紫色和橙色标出。

4.9 定性研究

我们在图 中可视化了来自标注、采用 BLS 的 BGNN（无偏模型）以及 BGNN+BLS+LANDMARK 在 Predcls 任务上的场景图生成结果。图 8 直观来看，标注和无偏模型倾向于预测“信息量较少的配对”之间的关系（例如 人-眼睛、屋顶-建筑物 和灯-公交车），这些关系相当常见。LANDMARK 还能进一步检测 男人-人行道 或 屋顶-公交车之间的关系。此外，LANDMARK 侧重于

高语义性和位置关系。例如，“手拿披萨”和“公交车后部的灯”，这证明我们的框架成功学习了物体之间的位置关系。

5 结论

本文首先指出，以往场景图生成方法对语言模态的利用不足。因此，受语言多义性的启发，我们为场景图生成任务提出了一种表征增强框架（LANDMARK），其特点是从对象标签中进行多语义提取。该插件网络从词嵌入中挖掘词语-视觉相关模式和语言上下文，并根据带有位置信息的主客体对学习谓词分布。与其他无偏方法相比，我们的框架从表征细化的角度提供了一种新方法。实验和分析表明，该框架能够持续提升基线模型的性能，并与其他无偏方法兼容。

声明

竞争性利益：

作者未报告任何潜在利益冲突

数据和材料的可用性：

支持本研究结论的数据可向第一作者索取。研究结果可通过我们的 GitHub 仓库复现。

代码可用性：

代码已公开发布于 [PySGG-cxg]，网址为 [https://github.com/rafa-cxg/PySGG-cxg]。

参考文献

[1] Gu, J., Joty, S., Cai, J., Zhao, H., Yang, X., Wang, G.：通过场景图对齐实现无配对图像描述。载于：IEEE/CVF国际计算机视觉会议论文集，第10323–10332页（2019年）[2] Xu, N., Liu, A.-A., Liu, J., Nie, W., Su, Y.：场景图描述生成器：基于结构化视觉表示的图像描述。《视觉通信与图像表示杂志》**58**，477–485（2019年）[3] Zhang, C., Chao, W.-L.,Xuan, D.：利用场景图进行视觉问答的实证研究。arXiv预印本 arXiv:1907.12133（2019年）

[4] Shi, J., Zhang, H., Li, J.：基于场景图的可解释显式视觉推理。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第 8376–8384 页（2019年）[5] Teney, D., Liu, L., van Den Hengel, A.：用于视觉问答的图结构表示。载于：IEEE计算机视觉与模式识别会议论文集，第 1–9 页（2017年）[6] Teng, Y., Wang, L., Li, Z., Wu, G.：用于视频场景图生成的目标自适应上下文聚合。载于：IEEE/CVF国际计算机视觉会议论文集，第 13688–13697 页（2021年）[7] Li, R., Zhang, S., Wan, B., He, X.：用于无偏场景图生成的自适应消息传递二分图网络。载于：IEEE/CVF计算机视觉与模式识别会议论文集（CVPR），第 11109–11119 页（2021年）[8] Dong, X., Gan, T., Song, X., Wu, J., Cheng, Y., Nie, L.：用于无偏场景图生成的堆叠式混合注意力与群组协同学习。arXiv 预印本 arXiv:2203.09811（2022年）[9] Yu,J., Chai,Y., Wang,Y., Hu,Y., Wu,Q.: Cogtree：用于无偏场景图生成的认知树损失。arXiv 预印本 arXiv:2009.07526（2020年）[10] Lyu, X., Gao, L., Guo, Y., Zhao, Z., Huang, H., Shen, H.T., Song, J.：用于场景图生成的细粒度谓词学习。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第 19467–19475 页（2022年）[11] Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.：从有偏训练中进行无偏场景图生成。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第 3716–3725 页（2020年）[12] Chen, C., Zhan, Y., Yu, B., Liu, L., Luo, Y., Du, B.：利用先验偏差进行抵抗训练：迈向无偏场景图生成。arXiv 预印本 arXiv:2201.06794（2022年）[13] Zellers, R., Yatskar, M., Thomson, S., Choi, Y.：神经模体：结合全局上下文的场景图解析。载于：IEEE计算机视觉与模式识别会议论文集，第 5831–5840 页（2018年）[14] Gkanatsios, N., Pitsikalis, V., Koutras, P., Maragos, P.：用于可扩展场景图生成的注意力-翻译-关系网络。载于：IEEE/CVF国际计算机视觉会议研讨会论文集，第 0–0 页（2019年）

[15]Zhong, Y., Shi, J., Yang, J., Xu, C., Li, Y.：在自然语言监督下学习生成场景图。载于：IEEE/CVF国际计算机视觉会议论文集，第1823–1834页（2021年）[16] Kipf, T.N., Welling, M.：利用图卷积网络进行半监督分类。arXiv预印本，arXiv:1609.02907（2016年）[17] Lin, X., Ding, C., Zeng, J., Tao, D.：Gps-net：用于场景图生成的图属性感知网络。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第3746–3753页（2020年）[18] Wang, W., Wang, R., Shan, S., Chen, X.：探索场景图生成中的关系上下文与视觉模式。载于：IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集（2019年）[19] Woo, S., Kim, D., Cho, D., Kweon, I.S.：Linknet：面向场景图的关系嵌入。神经信息处理系统进展 **31**（2018年）[20] Li, Y., Ouyang, W., Zhou, B., Shi, J., Zhang, C., Wang, X.：可因子分解网络：一种用于场景图生成的高效基于子图的框架。载于：欧洲计算机视觉会议（ECCV）论文集，第335–351页（2018年）[21] Liu, Y., Wang, R., Shan, S., Chen, X.：结构推理网络：利用场景级上下文和实例级关系进行目标检测。载于：IEEE计算机视觉与模式识别会议论文集，第6985–6994页（2018年）[22] Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.：从物体、短语和区域描述生成场景图。载于：IEEE国际计算机视觉会议论文集，第1261–1270页（2017年）[23] Hochreiter, S., Schmidhuber, J.：长短期记忆。神经计算 **9**（8），1735–1780（1997年）[24] Cho, K., Van Merri¨enboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.：使用RNN编码器-解码器学习用于统计机器翻译的短语表示。arXiv预印本，arXiv:1406.1078（2014年）[25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, ŁL., Polosukhin, I.：注意力就是你所需要的一切。载于：神经信息处理系统进展

载于《神经信息处理系统进展》，第5998–6008页（2017年）[26] Suhail, M., Mittal, A., Siddiquie, B., Broaddus, C., Eledath, J., Medioni, G., Sigal, L.：基于能量的学习在场景图生成中的应用。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第13936–13945页（2021年）[27] Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.：用于场景图生成的 Graph r-cnn。载于：欧洲计算机视觉会议（ECCV）论文集，第670–685页（2018年）[28] Dai, B., Zhang, Y., Lin, D.：利用深度关系网络检测视觉关系。载于：IEEE计算机视觉与模式识别会议论文集，第3076–3086页（2017年）[29] Chen, T., Yu, W., Chen, R., Lin, L.：用于场景图生成的知识嵌入路由网络。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第6163–6171页（2019年）[30]Tang, K., Zhang, H., Wu, B., Luo, W., Liu, W.：学习组合动态树结构以构建视觉上下文。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第6619–6628页（2019年）[31] Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.：利用语言先验进行视觉关系检测。载于：欧洲计算机视觉会议，第852–869页（2016年）。Springer[32] Zareian, A., Karaman, S., Chang, S.-F.：连接知识图谱以生成场景图。载于：欧洲计算机视觉会议，第606–623页（2020年）。Springer[33] Zareian, A., Wang, Z., You, H., Chang, S.-F.：学习视觉常识以实现稳健的场景图生成。载于：欧洲计算机视觉会议，第642–657页（2020年）。Springer[34] He, K., Zhang, X., Ren, S., Sun, J.：用于图像识别的深度残差学习。载于：IEEE计算机视觉与模式识别会议论文集，第770–778页（2016年）[35] Lu, Y., Rai, H., Chang, J., Knyazev, B., Yu, G., Shekhar, S., Taylor, G.W., Volkovs, M.：利用序列到序列变换器进行上下文感知的场景图生成[36] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen,

S., Kalantidis, Y., Li, L.-J., Shamma, D.A., 等人：Visual Genome：利用众包密集图像标注连接语言与视觉。《国际计算机视觉杂志》**123**(1)，32–73（2017年）[37]Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.：通过迭代消息传递生成场景图。载于：IEEE计算机视觉与模式识别会议论文集，第5410–5419页（2017年）[38] Newell, A., Deng,J.：通过关联嵌入将像素转换为图。神经信息处理系统进展**30** （2017年）[39]Ren, S., He, K., Girshick, R., Sun, J.：Faster r-cnn：利用区域提议网络实现实时目标检测。神经信息处理系统进展**28** （2015年）[40] He, K., Gkioxari, G., Doll´ar, P., Girshick, R.：掩码区域卷积神经网络。载于：IEEE国际计算机视觉会议论文集，第2961–2969页（2017年）[41]Pennington, J., Socher, R., Manning, C.D.：GloVe：用于词表示的全局向量。载于：2014年自然语言处理实证方法会议（EMNLP）论文集，第1532–1543页（2014年）