论文复现 VL-SAT: Visual-Linguistic Semantics Assisted Training for 3D Semantic Scene Graph Prediction in Point Cloud

摘要

点云中的 3D 语义场景图 (3DSSG) 预测任务具有挑战性，因为 (1) 与 2D 图像相比，3D 点云仅捕获语义有限的几何结构，(2) 长尾关系分布本质上阻碍了学习的无偏预测。由于 2D 图像提供了丰富的语义，而场景图本质上是与语言相关的，因此在本研究中，我们提出了视觉语言语义辅助训练（VL-SAT）方案，该方案可以显着增强 3DSSG 预测模型对长尾和模糊语义关系的区分能力。关键思想是训练一个强大的多模态预言机模型来辅助 3D 模型。该预言机基于视觉、语言和 3D 几何的语义学习可靠的结构表示，其好处可以在训练阶段异构地传递给 3D 模型。通过在训练中有效利用视觉语言语义，我们的 VL-SAT 可以显着增强常见的 3DSSG 预测模型，例如 SGFN 和 SGGpoint，仅在推理阶段使用 3D 输入，特别是在处理尾部关系三元组时。对 3DSSG 数据集的综合评估和消融研究验证了所提出方案的有效性。

关键词： 3D 场景图理解；点云

1 引言

从结构上理解 3D 几何场景对于需要与现实世界环境交互的任务尤其重要，例如 AR/VR 和导航。作为该领域的一个重要课题，在点云中预测 3D 语义场景图（3DSSG）近年来受到了越来越多的关注。具体来说，给定与类无关的 3D 实例蒙版关联的 3D 场景的点云，3DSGG 预测任务希望构建一个有向图，其节点是 3D 实例的语义标签，边识别连接的 3D 实例之间的方向语义或几何关系。

然而，除了场景图预测面临的常见困难之外，3DSSG 预测任务还面临一些挑战。(1) 3D 数据（例如点云）仅捕获每个实例的几何结构，并且可以通过相对方向或距离表面地定义关系。(2) 最近的 3DSSG 谓词数据集非常小，并且存在长尾谓词分布，其中语义谓词通常比几何谓词少见。尽管此后提出了一些尝试，但点云数据的固有局限性在一定程度上阻碍了这些方法的有效性。

由于 2D 图像提供了丰富且有意义的语义，并且场景图预测任务本质上与自然语言一致，因此我们探索使用视觉语言语义来辅助训练，作为从根本上增强常见 3DSSG 预测模型能力的另一种途径上述挑战。

如何利用视觉语言语义辅助 3D 结构理解仍然是一个悬而未决的问题。先前的研究主要集中在利用 2D 语义来增强实例级任务，例如对象检测、视觉基础和密集字幕。其中大多数在训练和推理中都需要视觉数据，但其中一些，例如 SAT 和 X Trans2Cap ，将 2D 语义视为辅助训练信号，因此仅使用 3D 数据提供更实用的推理。但这些方法都是针对实例级任务的，需要精心设计的网络来提供有效的帮助，因此它们对于我们的结构预测问题不太理想。由于最近像 CLIP 这样的大规模跨模态预训练的成功，图像中的 2D 语义可以与自然语言中的语言语义很好地结合起来，并且视觉语言语义已被应用于缓解相关任务中的长尾问题。到 2D 场景图和人与物体交互。但如何将视觉语言语义的类似帮助应用于 3D 场景仍不清楚。

在本研究中，我们提出了视觉语言语义辅助训练（VL-SAT）方案，使基于点云的 3DSSG 预测模型（称为 3D 模型）能够充分区分长尾和模糊的语义关系三元组。在这个方案中，我们同时训练一个强大的多模态预测模型作为预言机（称为预言机模型），它与 3D 模型异构对齐，通过来自视觉的额外数据、来自语言的额外训练信号捕获可靠的结构语义，如以及 3D 模型的几何特征。这些引入的视觉语言语义已通过 CLIP 进行了调整。因此，预言机模型的优点，尤其是多模态结构语义，可以通过反向传播梯度流有效地嵌入到 3D 模型中。在推理阶段，3D 模型只需 3D 输入即可实现卓越的 3DSSG 预测性能。据我们所知，VL-SAT 是第一个应用于点云中 3DSSG 预测的视觉语言知识转移工作。此外，VL-SAT 可以成功增强 SGFN 和 SGGpoint，验证该方案可以推广到常见的 3DSSG 预测模型。

2 相关工作

2.1 点云中的场景图预测

近年来，基于图像的语义场景图预测已被广泛研究 [6, 14, 15, 18, 20, 21, 26]，但只有少数工作尝试预测点云中的 3D 语义场景图。阿梅尼等人 [2] 提出了第一个 3D 场景图数据集，它将 3D 建筑物映射为分层结构。沃尔德等人 [17] 构建了一个基于点云的语义场景图数据集，即 3DSSG 以及名为 SGPN 的基于 GNN 的基线模型。后续工作 SGFN [19] 从 RGB-D 序列增量预测 3DSSG。近年来，人们提出了一些方法来改进基于 GNN 的基线。SGGpoint [27] 使用面向边缘的图卷积网络来利用多维边缘特征进行关系建模。张等人 [28] 提出了一种图自动编码器网络，可以预先自动学习每个类的一组嵌入，然后执行 3DSSG 预测，从预先学习的知识中识别可信的关系三元组。

2.2 使用 2D 语义理解 3D 场景

一系列方法已采用 2D 语义来帮助 3D 实例级任务，例如 3D 对象检测、分割、视觉基础和密集字幕 [3, 10, 11, 13, 16, 23, 31]。它们可以粗略地分为两类，即将图像特征与每个 3D 点 [3–5, 10, 16, 23, 30] 连接起来，并将对象检测结果投影到 3D 空间中。大多数方法在训练和推理阶段都需要二维语义。最近，SAT [22] 和 X -Trans2Cap [24] 探索仅在训练中使用 2D 语义来辅助 3D 视觉基础和密集字幕。他们都可以学习仅使用 3D 输入进行推理的增强模型。但这些方法仅限于实例级任务，并且必须仔细设计网络。我们遵循与类似的想法 [22, 24]，仅在训练中使用 2D 语义，但我们希望增强需要结构理解而不是实例级感知的 3DSSG 预测。

2.3 场景图预测中的知识插入方法

泽勒斯等人 [26] 和陈等人 [6] 表明对象对和关系之间的统计共现对于关系预测很有用。此外，[12, 29] 从所有先前的感知输出中生成了类级原型表示作为先验知识。这些方法将数据先验显式编码到模型中，[1, 7–9] 尝试将语言先验与场景图预测结合起来。扎雷安等人 [25] 提出了一种图桥接网络来在场景图和知识图之间传播消息。我们的 VL-SAT 方案使用 CLIP 来编码语言语义，从而更好地与 2D 语义，甚至训练阶段所需的 3D 结构语义保持一致。

3 本文方法

3.1 问题表述

假设我们有一个点云 ¶ $\in R^{N\times 3}$ 有 N 个点，和一组与类无关的实例掩码 $M = \{M_1, ..., M_K\}$ 将点云 P 与 K 个语义实例相关联，如 SGPN 所示，我们的目标是将 3D 语义场景图预测为有向图 $G = \{O, R\}$。对象集 $O = \{o_i\}_{i=1}^{K}$ 是由实例掩码 M 指定的所有命名对象实例。每条边 r_{ij} 在 R 中描述了关系三元组 < 主语，谓语，宾语 > 中的谓词，其中这条边的头节点 o_i 是主语，尾节点 o_j 是宾语。具体来说，o_j 表示来自 N_{obj} 语义类的对象标签。r_{ij} 是 N_{rel} 谓词类的谓词标签。

3.2 3D 预测模型

如图1所示，我们采用的 3D 预测模型与基于 GNN 的场景图预测方法（例如 SGFN 和 SGGpoint）具有相似的网络结构，主要由节点编码器、边缘编码器和场景图推理模块。

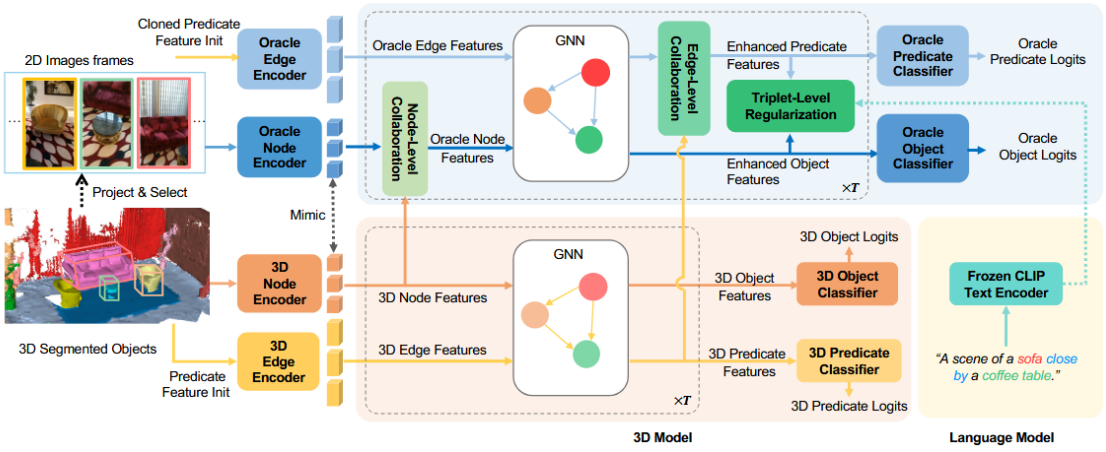

图 1. 方法示意图

节点编码器。基于一个与类无关的实例掩码 M_i 以及输入点云 P，我们可以提取对应于一个语义实例的一组点 P_i。我们采用简单的 PointNet 来提取实例级特征。基于 GNN 的场景图推理之前的节点特征 $o_i^{3\text{d}} \in R^D$ 由这些实例级特征给出。

边编码器。我们遵循与 SGFN 相同的做法来编码基于 GNN 的边缘特征场景图推理。它需要计算链接实例之间的多个属性之间的差异。对于每个实例，这些属性包括 3D 点的平均值 和标准差 、大小 $b = (b_x, b_y, b_z)$、体积 $v = b_x b_y b_z$ 以及最大边长 $l = \max(b_x, b_y, b_z)$ 的边界框。

因此，边缘特征 ri3jd RD 通过多层感知器（MLP）层投影两个实例之间这些属性的串联差异来编码，即

$$r_{ij}^{3\mathrm{d}} = \texttt{MLP}(\texttt{cat}(u_i - u_j, \sigma_i - \sigma_j, b_i - b_j, \ln \frac{l_i}{l_j}, \ln \frac{v_i}{v_j})), \tag{1}$$

其中下标 i 表示头节点中的实例 P_i，j 表示尾节点中的实例 P_j。

场景图推理。

在我们的实验中，我们应用了与 SGFN 类似的 GNN 结构，它利用特征注意力（FAT）模块在节点和边之间传递消息，然后获取更新的节点和边特征。每个 GNN 模块都与多头自注意力（MHSA）模块配对，重复 T 次以提取最终的节点和边缘特征 $\{\tilde{o}_i^{3\mathrm{d}}\}_{i=1,\dots,K}$ 和 $\{\tilde{r}_{ij}^{3\mathrm{d}}\}_{i\neq j,i,j=1,\dots,K}$。此后，对象分类器和谓词分类器将根据三元组特征 $\{\tilde{o}_i^{3\mathrm{d}}, \tilde{r}_{ij}^{3\mathrm{d}}, \tilde{o}_j^{3\mathrm{d}}\}$ 来预测每个可能的关系三元组的元素 $\{o_i, r_{ij}, o_j\}$。这些关系三元组最终构造语义场景图 $G = \{Q, R\}$。

3.3 视觉语言语义辅助训练

在本小节中，我们将详细介绍视觉语言语义辅助训练（VL-SAT）方案如何使 3D 预测模型能够充分区分长尾和模糊语义关系三元组。关键思想是，这种判别能力来自于辅助学习强大的多模态预测模型，该模型接收来自视觉和语言的结构语义，以及来自 3D 预测模型的 3D 几何形状。多模态语义有望在节点和边缘级别与 3D 语义异构对齐，并且预言机模型的好处可以在训练过程中被 3D 预测模型有效吸收。具体来说，我们首先介绍与 3D 预测模型具有异构协作的应用多模态预测模型，然后是提高预言机模型性能并最终增强 3D 预测的辅助训练策略。

该预言机模型的特征与 3D 预测模型中的特征在节点和边缘级别上异构协作，这些特征在场景图推理模块中的每个 GNN 层之前和之后进行。前者是节点级协作，后者是边缘级协作。具体来说，这些协作操作是通过多头交叉注意（MHCA）模块 [33] 实现的，其中键和值是来自 3D 模型的节点/边缘特征，查询是来自多模态的对应项模型。节点级协作具有距离感知屏蔽策略，以消除相距较远、没有有效关系的实例之间不必要的关注。两个实例 P_i 和 P_j 之间的掩码值通过以下方式学习：

$$D_{ij}^{\mathrm{node}} = \texttt{MLP}(\texttt{cat}(u_i - u_j, \|u_i - u_j\|_2)), \tag{2}$$

相对于点云实例 P_iP_i 和 P_j 的平均坐标 u_i 和 u_j。边缘级协作不使用距离感知掩蔽策略，因为边缘之间的距离很难定义，因此将所有边缘纳入注意计算会更安全。

请注意，异构协作从 3D 模型到预言机模型是单向的，而预言机模型的好处通过反向传播的梯度流传递到 3D 模型。它有利于在推理阶段，预测 3D 语义场景图不需要来自其他模态的额外数据。

辅助训练策略。

由于预言机多模态模型希望从视觉和语言角度感知场景图，因此很自然地使用 CLIP 捕获的视觉语言知识来增强预言机模型。具体来说，我们可以为每个真实关系三元组生成 CLIP 文本嵌入 e_{ij}^{text}，并在场景图推理模块的每个 GNN 层末尾对相应的三元组特征 $\{\tilde{o}_i^{\mathrm{oracle}}, \tilde{r}_{ij}^{\mathrm{oracle}}, \tilde{o}_j^{\mathrm{oracle}}\}$ 进行正则化。CLIP 文本嵌入通过模板 “A scene of a/an [subject][predicate] a/an [object]” 为

每个真实关系离线提取。因此，正则化变得最小化文本嵌入 e_{ij}^{text} 和融合的三元组特征为 t_{ij}^{oracle} ，即

$$L_{\text{tri-emb}} = \sum_{i=1}^{K} \sum_{j=1, j\neq i}^{K} \rho(t_{ij}^{\text{oracle}}, e_{ij}^{\text{text}}) \cdot I_{[e_{ij}^{\text{text}} \text{ is from GT triplet}]} \tag{3}$$

其中 $t_{ij}^{\text{oracle}} = \texttt{MLP}(\texttt{cat}(\tilde{o}_i^{\text{oracle}}, \tilde{r}_{ij}^{\text{oracle}}, \tilde{o}_j^{\text{oracle}}))$ 是串联特征 $\tilde{o}_i^{\text{oracle}}, \tilde{r}_{ij}^{\text{oracle}}$, 和 $\tilde{o}_j^{\text{oracle}}$ 的融合嵌入。$\rho(\cdot,\cdot)$ 是一个距离度量，我们可以应用 ℓ_1 范数或负余弦距离。$I_{[\cdot]}$ 是指示函数，当参数为 true 时等于 1，否则等于 0。因此方程 (3) 仅对三元组具有真实关系的节点和边特征进行正则化。

此外，在放入场景图推理模块之前，可以将来自 3D 模型的 3D 节点特征 oi3d 和来自 oracle 模型的 2D 节点特征 oi2d 进行对齐。我们应用与等式相同的距离测量。

$$L_{\text{node-init}} = \sum_{i=1}^{K} \rho(o_i^{\text{3d}}, o_i^{\text{2d}}). \tag{4}$$

为了增强初始化 2D 节点特征的表示能力，2D 实例编码器是固定的 CLIP 预训练视觉编码器。此外，为了增强两个模型的对象分类器，我们使用 CLIP 对象嵌入来初始化 3D 预测模型和 oracle 多模态预测模型中对象分类器的权重。

3.4 损失函数

整个网络的训练目标定义为:

$$L = \lambda_{\text{obj}}(L_{\text{obj}}^{\text{3d}} + L_{\text{obj}}^{\text{oracle}}) + \lambda_{\text{pred}}(L_{\text{pred}}^{\text{3d}} + L_{\text{pred}}^{\text{oracle}}) + \lambda_{\text{aux}}(L_{\text{tri-emb}} + L_{\text{node-init}}) \tag{5}$$

L_{obj} 表示对象分类损失，通过交叉熵损失实现。$L_{\text{obj}}^{\text{3d/oracle}}$ 应用于 3D/oracle 对象分类器。L_{pred} 表示谓词分类损失，并被公式化为每类二元交叉熵损失。$L_{\text{pred}}^{\text{3d/oracle}}$ 应用于 3D/oracle 谓词分类器。$\lambda_{\text{node}}, \lambda_{\text{edge}}, \lambda_{\text{aux}}$ 是超参数，用于平衡相同尺度下的各个损失。

4 复现细节

4.1 与已有开源代码对比

本文是有开源代码的（https://github.com/wz7in/CVPR2023-VLSAT）。

在开源代码的基础上，我们对图神经网络加入 Transformer 进行改进，具体为讲三元组中的主语作为 query，全局节点作为 key，之后我们再在所得到的结果与边特征（即谓词）进行点乘之后再经过 softmax，为了让模型学习到更多的局部和全局特征。

4.2 实验环境搭建

conda create -n vlsat python=3.8

conda activate vlsat

pip install -r requirement.txt

pip install torch==1.12.1+cu113 torchvision==0.13.1+cu113 torchaudio==0.12.1 –extra-index-url https://download.pytorch.org/whl/cu113

pip install torch-scatter -f https://pytorch-geometric.com/whl/torch-1.12.1+cu113.html

pip install torch-sparse -f https://pytorch-geometric.com/whl/torch-1.12.1+cu113.html

pip install torch-spline-conv -f https://pytorch-geometric.com/whl/torch-1.12.1+cu113.html

pip install torch-geometric

pip install git+https://github.com/openai/CLIP.git

4.3 创新点

往图神经网络里加入了 Transformer。

5 实验结果分析

我们得到的结果在场景图分析中，具体如图2所示

	Object			Predicate						Triplet			
	A@1	A@5	A@10	A@1	A@3	A@5	mA@1	mA@3	mA@5	A@50	A@100	mA@50	mA@100
VL-SAT	**55.66**	**78.66**	**85.91**	89.81	**98.45**	**99.53**	**54.03**	**77.67**	**87.65**	**90.35**	**92.89**	**65.09**	**73.59**
3D-VLAP	50.04	74.02	83.54	87.52	95.34	97.99	35.19	48.39	62.32	87.15	90.04	42.70	52.82
修改	53.13	76.21	84.76	**90.91**	98.42	99.44	50.30	70.67	82.01	89.46	92.14	55.08	68.42

	Triplet			
	Unseen		Seen	
	R@50	R@100	R@50	R@100
VL-SAT	**27.79**	**43.61**	**72.55**	**81.44**
修改	24.31	39.77	71.76	80.65

图 2. 实验结果示意

同时我们发现在谓词预测中，body 部分的数据集有所提升，如图3所示。

model	Predicate								
	Head			Body			Tail		
	mA@1	mA@3	mA@5	mA@1	mA@3	mA@5	mA@1	mA@3	mA@5
VL-SAT	77.66	**96.31**	99.21	51.34	80.03	93.64	**30.07**	**52.38**	**66.13**
3dvlap	-	83.3	95.6	-	31.9	60.2	-	31.9	39.1
ours	**80.96**	96.50	**99.39**	**56.58**	**87.84**	**96.16**	28.70	43.19	60.59

图 3. 实验结果示意

参考文献

[1] Sherif Abdelkarim, Aniket Agarwal, Panos Achlioptas, Jun Chen, Jiaji Huang, Boyang Li, Kenneth Church, and Mohamed Elhoseiny. Exploring long tail visual relationship recognition with large vocabulary. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15921–15930, 2021.

[2] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5664–5673, 2019.

[3] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1090–1099, 2022.

[4] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16464–16473, 2022.

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020.

[6] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2019.

[7] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1347–1356, 2017.

[8] Wentong Liao, Bodo Rosenhahn, Ling Shuai, and Michael Ying Yang. Natural language guided visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 0–0, 2019.

[9] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li FeiFei. Visual relationship detection with language priors. In *European Conference on Computer Vision (ECCV)*, pages 852–869. Springer, 2016.

[10] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4404–4413, 2020.

[11] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018.

[12] Sahand Sharifzadeh, Sina Moayed Baharlou, and Volker Tresp. Classification by attention: Scene graph classification with prior knowledge. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 5025–5033, 2021.

[13] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282. IEEE, 2019.

[14] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020.

[15] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6619–6628, 2019.

[16] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020.

[17] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3961–3970, 2020.

[18] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. *Advances in neural information processing systems*, 31, 2018.

[19] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021.

[20] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017.

[21] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018.

[22] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1856–1866, 2021.

[23] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021.

[24] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8563–8573, 2022.

[25] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *European Conference on Computer Vision (ECCV)*, pages 606–623. Springer, 2020.

[26] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018.

[27] Chaoyi Zhang, Jianhui Yu, Yang Song, and Weidong Cai. Exploiting edge-oriented reasoning for 3d point-based scene graph analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9705–9715, 2021.

[28] Shoulong Zhang, Aimin Hao, Hong Qin, et al. Knowledge-inspired 3d scene graph prediction in point cloud. *Advances in Neural Information Processing Systems*, 34:18620–18632, 2021.

[29] Shoulong Zhang, Aimin Hao, Hong Qin, et al. Knowledge-inspired 3d scene graph prediction in point cloud. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 18620–18632, 2021.

[30] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvgtransformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2928–2937, 2021.

[31] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021.