Expanding Scene Graph Boundaries: Fully Open-vocabulary Scene Graph Generation via Visual-Concept Alignment and Retention

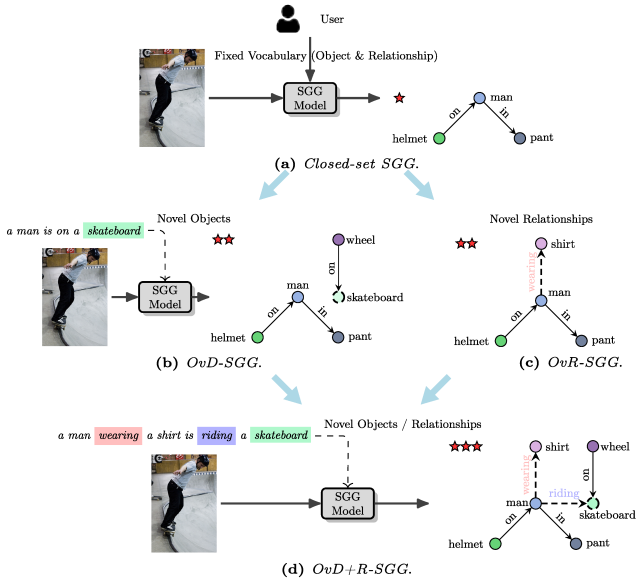
Zuyao (Joseph) Chen[1,2] Jinlin Wu[2,3] Zhen Lei[2,3,4] Zhaoxiang Zhang[2,3,4] Chang Wen Chen[1]

[1]The Hong Kong Polytechnic University [2]Centre for Artificial Intelligence and Robotics, HKISI-CAS

[3]Institute of Automation, CAS [4] University of Chinese Academy of Sciences

Motivation

Scene Graph Generation (SGG) predominantly operates within a **closed-set** setup, constraining its recognition abilities to a predefined set of objects and relationships.



- **Can the model predict unseen objects or relationships?**
- **What if BOTH unseen objects AND unseen relationships are present?**

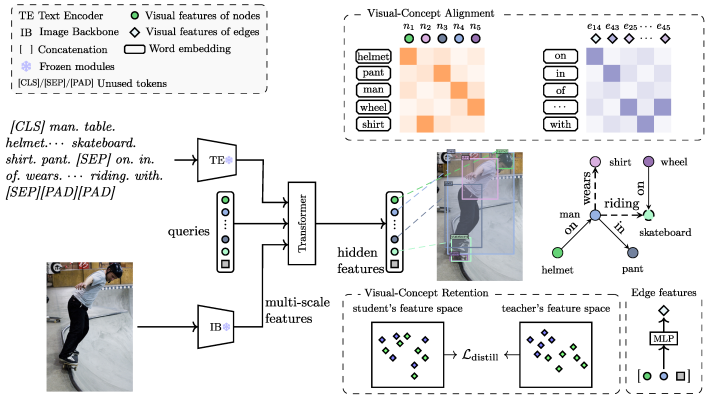
Challenges

- **no relation-aware pre-trained models**
- **distinguish unseen relationships from "background" (OvR-SGG, OvD+R-SGG)**

Highlights

1. Introduced a unified framework for open-vocabulary SGG, addressing limitations of traditional methods
2. Proposed novel visual-concept alignment and retention mechanisms for improved generalization to unseen concepts
3. Achieved state-of-the-art results, demonstrating the effectiveness of OvSGTR

Unified framework: OvSGTR



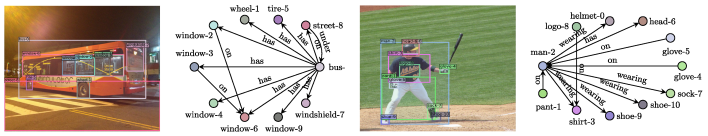
- **DETR-like structure instead of R-CNN**
- **Visual-concept alignment**
- **Visual-concept retention**
- **Relation-aware pre-training**

Experiments

Experimental results of *Closed-set SGG* on VG150 test set.

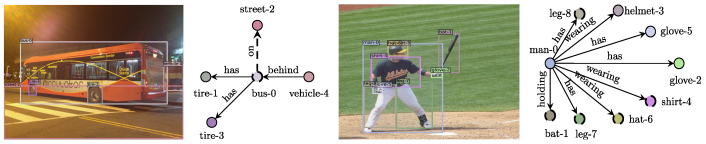
SGG model	Backbone	Detector	Params.	R@20/50/100			mR@20/50/100			Time (s)
IMP	RX-101	Faster R-CNN	146M/308M	17.7	25.5	30.7	2.7	4.1	5.3	0.25
MOTIFS			205M/367M	25.5	32.8	37.2	5.0	6.8	7.9	0.27
VCTREE			197M/358M	24.7	31.5	36.2	-	-	-	0.38
HL-Net			220M/382M	26.0	33.7	38.1	-	-	-	0.10
FCSGG	HRNetW48	-	87M/87M	13.6	18.6	22.5	2.3	3.2	3.9	0.13
SGTR	R-101	DETR	36M/96M	-	24.6	28.4	-	-	-	0.21
VS^3	Swin-T	-	93M/233M	26.1	34.5	39.2	-	-	-	0.16
VS^3	Swin-L	-	124M/432M	27.3	36.0	40.9	4.4	6.5	7.8	0.24
OvSGTR	Swin-T	DETR	41M/178M	**27.0**	**35.8**	**41.3**	**5.0**	**7.2**	**8.8**	0.13
OvSGTR	Swin-B	DETR	41M/238M	**27.8**	**36.4**	**42.4**	**5.2**	**7.4**	**9.0**	0.19

state-of-the-art performance
fewer parameters & lower inference latency



Trained on Closed-set SGG

Experiments



Trained on OvD+R-SGG

Real-world example

Experimental results of *OvD-SGG*.

Method	Base+Novel (Object) PREDCLS	Base+Novel (Object) SGDET	Novel (Object) PREDCLS	Novel (Object) SGDET
VS^3-T	50.10 / 52.05	15.07 / 18.73	46.91 / 49.13	10.08 / 13.65
OvSGTR -T	**60.58 / 62.10**	**18.14 / 23.20**	**59.01 / 60.65**	**12.06 / 16.49**
OvSGTR -B	**60.83 / 62.33**	**21.35 / 26.22**	**59.30 / 60.95**	**15.58 / 19.96**

Experimental results of *OvR-SGG*.

Method	Base+Novel (Relation) R@50	Base+Novel (Relation) R@100	Novel (Relation) R@50	Novel (Relation) R@100
VS^3-T	15.60	17.30	0.00	0.00
OvSGTR-T†	17.71	20.00	0.34	0.41
OvSGTR-T	**20.46**	**23.86**	**13.45**	**16.19**
OvSGTR-B†	18.58	20.84	0.08	0.10
OvSGTR-B	**22.89**	**26.65**	**16.39**	**19.72**

Experimental results of *OvD+R-SGG*.

Method	Joint Base+Novel R@50	Joint Base+Novel R@100	Novel (Object) R@50	Novel (Object) R@100	Novel (Relation) R@50	Novel (Relation) R@100
VS^3-T	5.88	7.20	6.00	7.51	0.00	0.00
OvSGTR-T†	7.88	10.06	6.82	9.23	0.00	0.00
OvSGTR-T	**13.53**	**16.36**	**14.37**	**17.44**	**9.20**	**11.19**
OvSGTR-B†	11.23	14.21	13.27	16.83	1.78	2.57
OvSGTR-B	**17.11**	**21.02**	**17.58**	**21.72**	**14.56**	**18.20**

† refers to *w.o.* distillation

Models and Code are available:

https://github.com/gpt4vision/OvSGTR

Personal homepage: https://josephchenhub.github.io Email: zuyao.chen@connect.polyu.hk Group homepage: https://chenlab.comp.polyu.edu.hk/