Hybrid Reciprocal Transformer with Triplet Feature Alignment for Scene Graph Generation

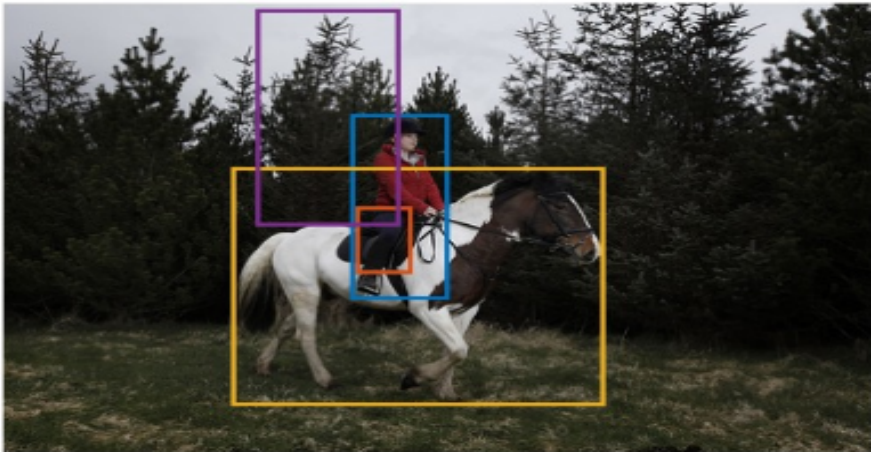
CVPR 2025

Jiawei Fu, Tiantian Zhang, Kai Chen, and Qi Dou

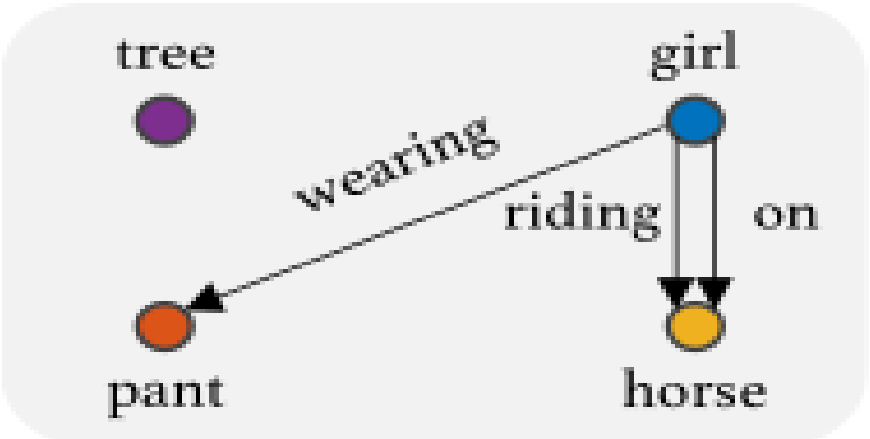
Chinese University of Hong Kong

November 9, 2025

What is a scene graph



What is a scene graph



Applications

Since a scene graph provides structural information of the image, it can be used for various vision tasks that require a higher level of understanding and reasoning about images

- Image Captioning
- Image Retrieval
- Visual Question Answering
- Conditional Image Generation

Priliminary

What is Query?

- Placeholder for anything, including bbox, segmentation mask, etc
- Cross attention with image feature and self attention with other queries
- Learnable

SGG Approaches

Two Main Approaches

- Two-Stage Methods
- One-Stage Methods
  - Object-Triplet Detection Models
  - Triplet Detection Models
  - Relation Extraction Models

Two-Stage Methods

1. Separate object detection model and relation prediction model are trained sequentially
2. They usually detect N objects from the off-the-shelf object detector such as Faster R-CNN
3. Then all possible combinations of the detected objects are fed into the relation prediction model to predict the relations between each object pair.

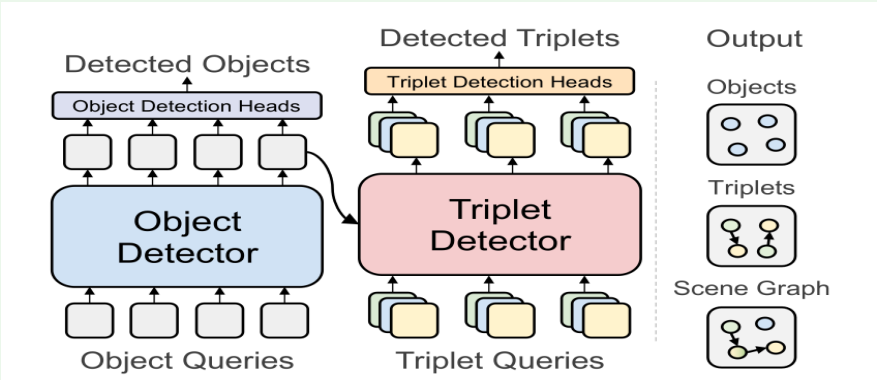
Shortcomings

They have the inherent limitation that the object detector that helps with relation extraction is trained separately, resulting in a significant increase in model complexity.

One-Stage Methods, Object-Triplet Detection Models

Object-triplet detection models are characterized by introducing additional triplet queries and building a triplet predictor on top of the object detector

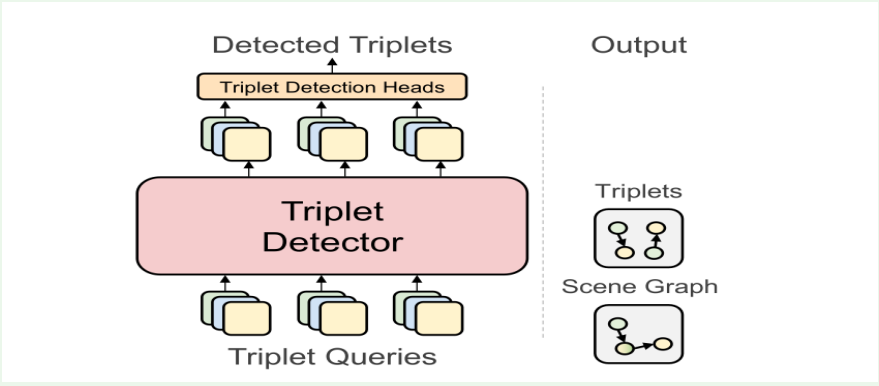
Example



One-Stage Methods, Triplet Detection Models

Triplet detection models directly detect triplets using triplet queries without an object detector

Example



One-Stage Methods, Triplet Detection Models

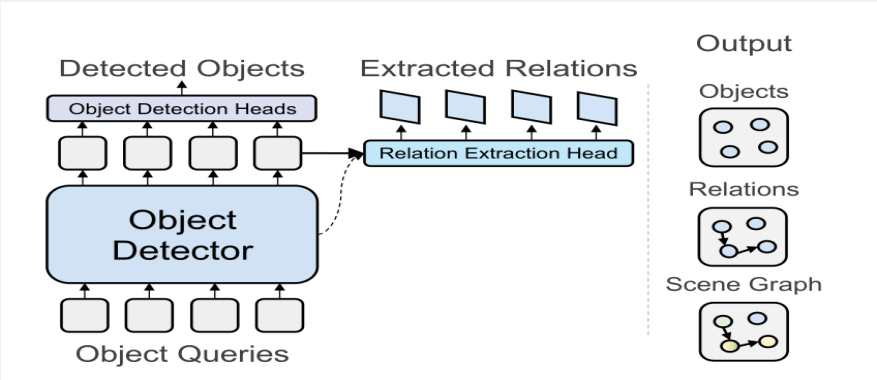
Shortcomings

- The model architecture became even more intricate to detect subjects and objects separately since an explicit object detector is not used.
- The triplet prediction models focus on detecting a sub-graph composed only of objects with relations, overlooking objects in an image that lack explicit relations.

One-Stage Methods, Relation Extraction Models

Relation extraction models extract a scene graph using a lightweight relation predictor without separate triplet queries or triplet detector

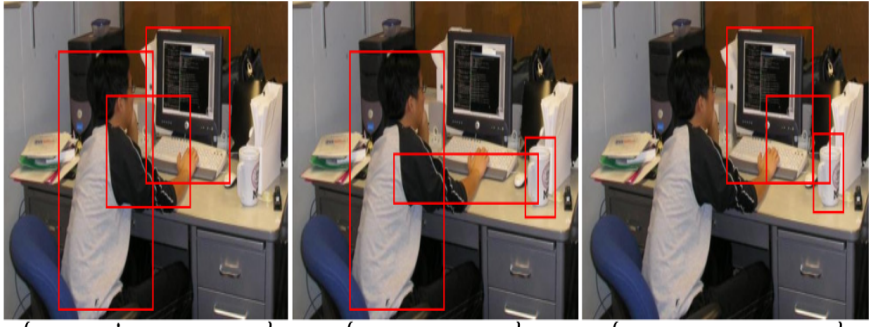
Example



New Taxonomies

Component

1. Class and Bbox of subject
2. Class and Bbox of object
3. Class and Bbox of predicate(Box derived from box of subject and object)



New Taxonomies

Triplet

1. Its image visualization is segmentation mask, indicating a triplet
2. Segmentation masks, derived from components

Why Triplet

Nothing but a special type of augmentation

- Context
- Distinguish multi-role objects better is all this paper aims for

Formulation

Conventional Formulation

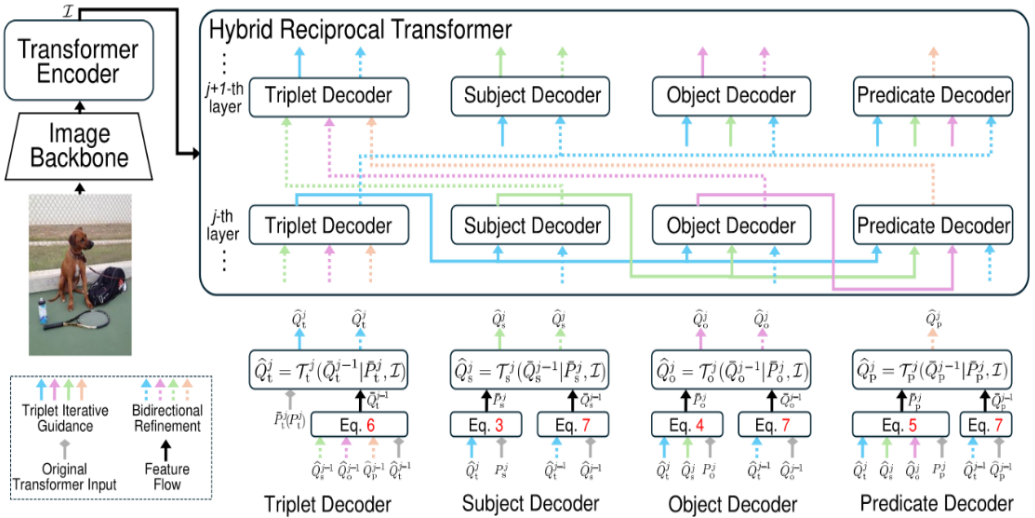
$$G = \{s_k, o_k, p_k\}_{k=1}^{K}$$

Paper Formulation

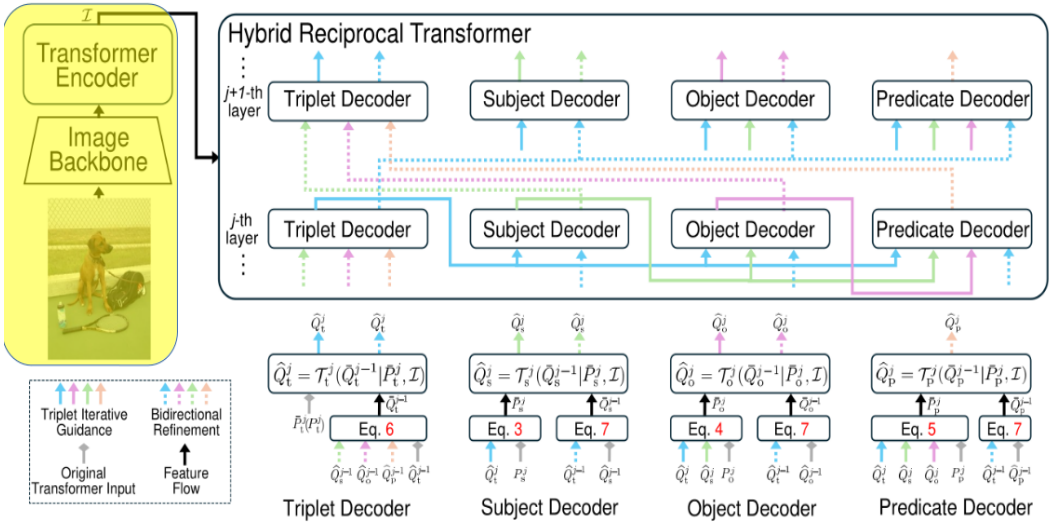
$$G = \{t_k, s_k, o_k, p_k\}_{k=1}^{K}$$

Where t_k represents the binary mask of the triplet of s_k , o_k , p_k with a class label represent predicate class

Proposed Method



Feature Extraction

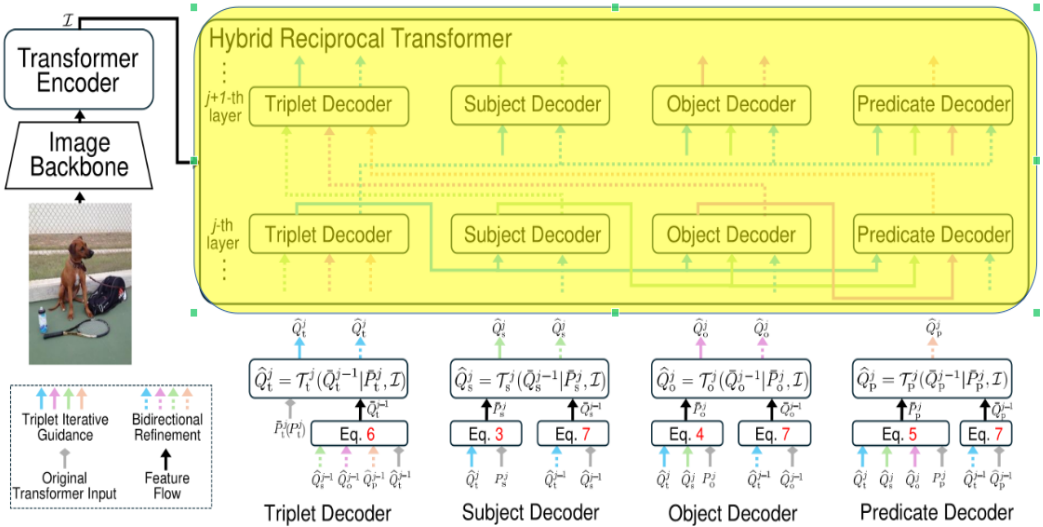


Feature Extraction

Details

- Resnet-101
- Six-layer transformer encoder

Hybrid Reciprocal Transformer



Hybrid Reciprocal Transformer

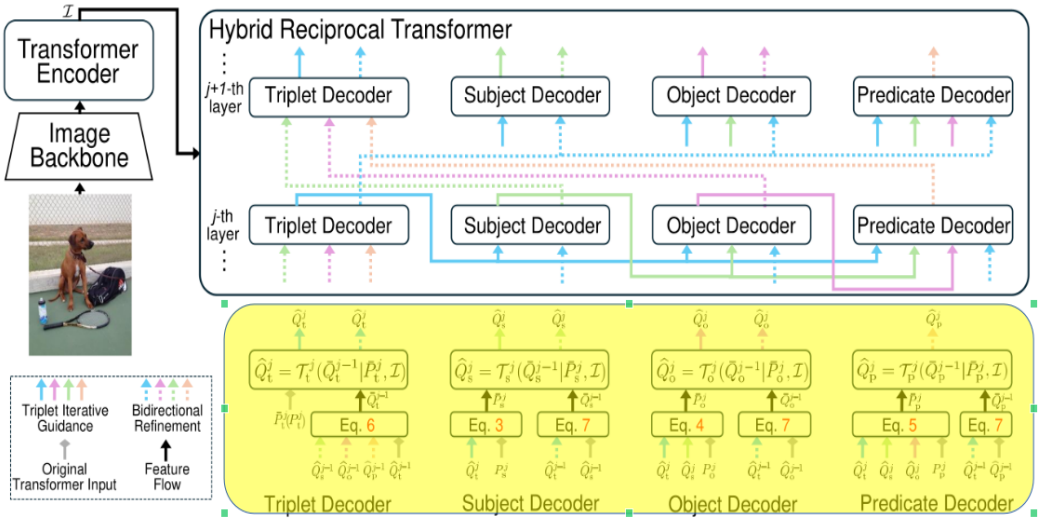
There exists multiple layers of decoders

Decoders

- Triplet
- Subject
- Object
- Predicate

Output of decoders in j-th layer : $\hat{Q}_x^j; x \in \{t, s, o, p\}$

Decoders


$\mathcal{I}$
Transformer Encoder
Image Backbone
Hybrid Reciprocal Transformer
j+1-th layer
j-th layer
Triplet Decoder
Subject Decoder
Object Decoder
Predicate Decoder
Triplet Iterative Guidance
Bidirectional Refinement
Original Transformer Input
Feature Flow
$\hat{Q}_{\mathrm{t}}^{j} = \mathcal{T}_{\mathrm{t}}^{j}(\bar{Q}_{\mathrm{t}}^{j-1} | \bar{P}_{\mathrm{t}}^{j}, \mathcal{I})$
$\hat{Q}_{\mathrm{s}}^{j} = \mathcal{T}_{\mathrm{s}}^{j}(\bar{Q}_{\mathrm{s}}^{j-1} | \bar{P}_{\mathrm{s}}^{j}, \mathcal{I})$
$\hat{Q}_{\mathrm{o}}^{j} = \mathcal{T}_{\mathrm{o}}^{j}(\bar{Q}_{\mathrm{o}}^{j-1} | \bar{P}_{\mathrm{o}}^{j}, \mathcal{I})$
$\hat{Q}_{\mathrm{p}}^{j} = \mathcal{T}_{\mathrm{p}}^{j}(\bar{Q}_{\mathrm{p}}^{j-1} | \bar{P}_{\mathrm{p}}^{j}, \mathcal{I})$
Eq. 6
Eq. 3
Eq. 7
Eq. 4
Eq. 7
Eq. 5
Eq. 7
Triplet Decoder
Subject Decoder
Object Decoder
Predicate Decoder

Decoders

Each Decoer

- Input 1 : Object queries, from previous layer
- Input 2 : Positional Encoding
- Action : Self attention + Cross attention with Image features

Positional Encoding

Original Positional Encoding

- Learnable
- Denoted as $P_x^j; x \in \{t, s, o, p\}$

Triple-Guided Positional Encoding

- Bidirectional Refinement between triplet-level and component level features
- $\bar{P}_x^j; x \in \{t, s, o, p\}$

Positional Encoding

Triplet-Guided Positional Encoding

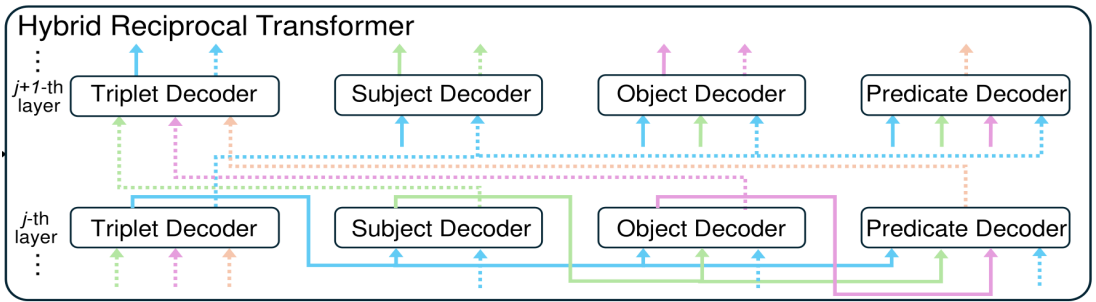
- $\bar{P}_s^j = P_s^j + \text{FFN}(\text{MHA}(P_s^j, \hat{Q}_t^j, \hat{Q}_t^j))$
- $\bar{P}_o^j = P_o^j + \text{FFN}(\text{MHA}(P_o^j, [\hat{Q}_t^j, \hat{Q}_s^j], [\hat{Q}_t^j, \hat{Q}_s^j]))$
- $\bar{P}_p^j = P_p^j + \text{FFN}(\text{MHA}(P_p^j, [\hat{Q}_t^j, \hat{Q}_s^j, \hat{Q}_o^j], [\hat{Q}_t^j, \hat{Q}_s^j, \hat{Q}_o^j]))$

Also for Bidirectional Refinement

Queries from previous layer are cross-conditioned between triplet features and component features

Input to next layer

- $\bar{Q}_t^j = \hat{Q}_t^j + \text{FFN}(\text{MHA}(\hat{Q}_t^j, [\hat{Q}_x^j], [\hat{Q}_x^j])); x \in \{s, o, p\}$
- $\bar{Q}_x^j = \hat{Q}_x^j + \text{FFN}(\text{MHA}(\hat{Q}_x^j, \hat{Q}_t^j, \hat{Q}_t^j)); x \in \{s, o, p\}$



Triplet Feature Alignment Loss

Aligned Query Representations

$$\tilde{Q}_x^j; x \in \{t, s, o, p\}$$

Pseudo Tiplet-Level Representation

$$Q_t^j = \tilde{Q}_s^j + \tilde{Q}_o^j + \tilde{Q}_p^j$$

Similarity Matrix

$$\mathcal{S}^j = Q_t^j (\tilde{Q}_t^j)^T$$

GT

One-hot encoded, derived from datasets annotations.

Results

Method	R@50	R@100	mR@50	mR@100	AvgR@50	AvgR@100	F@50	F@100
Two-Stage, X101-FPN Backbone								
Motifs [52] [CVPR2018]	32.1	36.9	5.5	6.8	18.8	21.9	9.4	11.5
VCTree [43] [CVPR202]	31.8	36.1	6.6	7.7	19.2	21.9	10.9	12.7
TDE [44] [CVPR2020]	19.4	23.2	9.3	11.1	14.4	17.2	12.6	15.0
EBM [42] [CVPR2021]	20.5	24.7	9.7	11.6	15.1	18.2	13.2	15.8
BPLSA [11] [ICCV2021]	21.7	25.5	13.5	15.7	17.6	20.6	16.6	19.4
DT2-ACBS [8] [ICCV2021]	22.0	24.4	15.0	16.3	18.5	20.4	17.8	19.5
BGNN [24] [CVPR2021]	31.0	35.8	10.7	12.7	20.9	24.2	15.9	18.7
PE-Net [57] [CVPR2023]	30.7	35.2	12.4	14.5	21.6	24.9	17.7	20.5
VETO [41] [CVPR2023]	27.5	31.5	8.1	9.5	17.8	20.5	12.5	14.6
DRM [22] [CVPR2024]	34.0	38.9	9.0	11.2	21.5	25.0	14.2	17.4
One-Stage, ResNet-50 Backbone								
RelTR [7] [TPAMI2023]	27.5	30.7	10.8	12.3	19.2	21.5	15.5	17.6
EGTR [14] [CVPR2024]	30.2	34.3	7.9	10.1	19.1	22.2	12.5	15.6
Hydra-SGG [4] [Arxiv2024]	28.4	33.1	16.0	19.7	22.2	26.4	20.5	24.7
One-Stage, ResNet-101 Backbone								
RelDN [53] [CVPR2019]	30.3	34.8	4.4	5.4	17.4	20.1	7.7	9.3
AS-Net [3] [CVPR2021]	18.7	21.1	6.1	7.2	12.4	14.2	9.2	10.7
SGTR [25] [CVPR2022]	25.1	26.6	12.0	14.6	18.6	20.6	16.2	18.9
HOTR [17] [CVPR2021]	22.4	27.1	6.9	9.7	14.7	18.4	10.6	14.3
SpeaQ$_{\text{HOTR}}$ [18] [CVPR2024]	24.7	29.1	9.6	12.7	17.2	20.9	13.8	17.7
ISG$_{\dagger}$ [16] [NIPS2022]	29.5	32.1	7.4	8.4	18.5	20.3	11.8	13.3
SpeaQ$_{\text{ISG}_{\dagger}}$ [18] [CVPR2024]	32.9	36.0	11.8	14.1	22.4	25.1	17.4	20.3
ISG [16] [NIPS2022]	27.2	30.1	15.0	16.6	21.1	23.4	19.3	21.4
SpeaQ$_{\text{ISG}}$ [18] [CVPR2024]	32.1	35.5	15.1	17.6	23.6	26.6	20.5	23.5
Ours	**34.1**(+1.2)	**38.3**(+2.3)	**16.0**(+0.9)	**20.5**(+2.9)	**25.1**(+1.5)	**29.4**(+2.8)	**21.8**(+1.3)	**26.7**(+3.2)

Table 1. **Scene graph generation performance on Visual Genome.** The best results among models with ResNet-101 backbone are marked in bold. † denotes the performance without loss re-weighting proposed in [16].

Qualitive Results

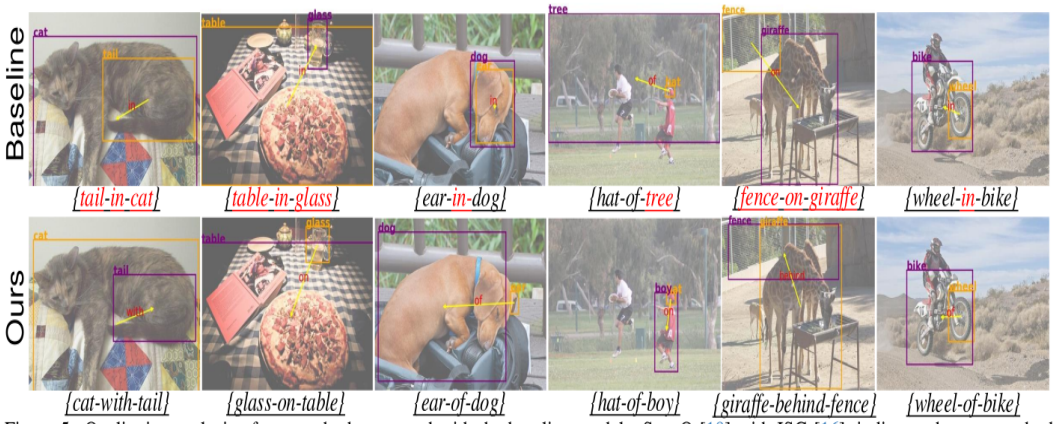


Figure 5. Qualitative analysis of our method compared with the baseline models, SpeaQ [18] with ISG [16], indicates that our method yields significantly enhanced accuracy in the prediction of predicates and object detection. Red marks error.

The end

Thanks for your attention