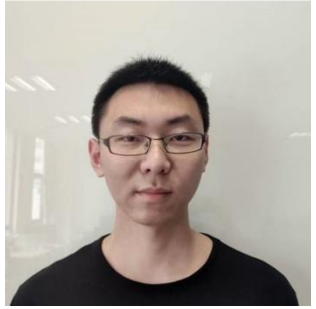
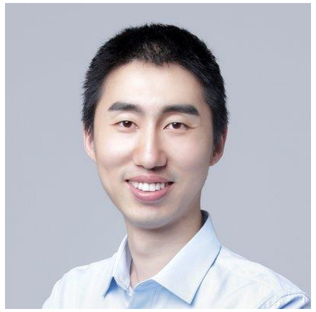
ACM
Multimedia 2022
Lisbon, Portugal | 10-14 October

#2754

Heterogeneous Learning for Scene Graph Generation

Yunqing He[1] **Tongwei Ren**[*,1] **Jinhui Tang**[2] **Gangshan Wu**[1]

1 State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

2 School of Computer Science, Nanjing University of Science and Technology, Nanjing, China

南京大學 NANJING UNIVERSITY

南京理工大学 NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY

Scene Graph Generation

Goals

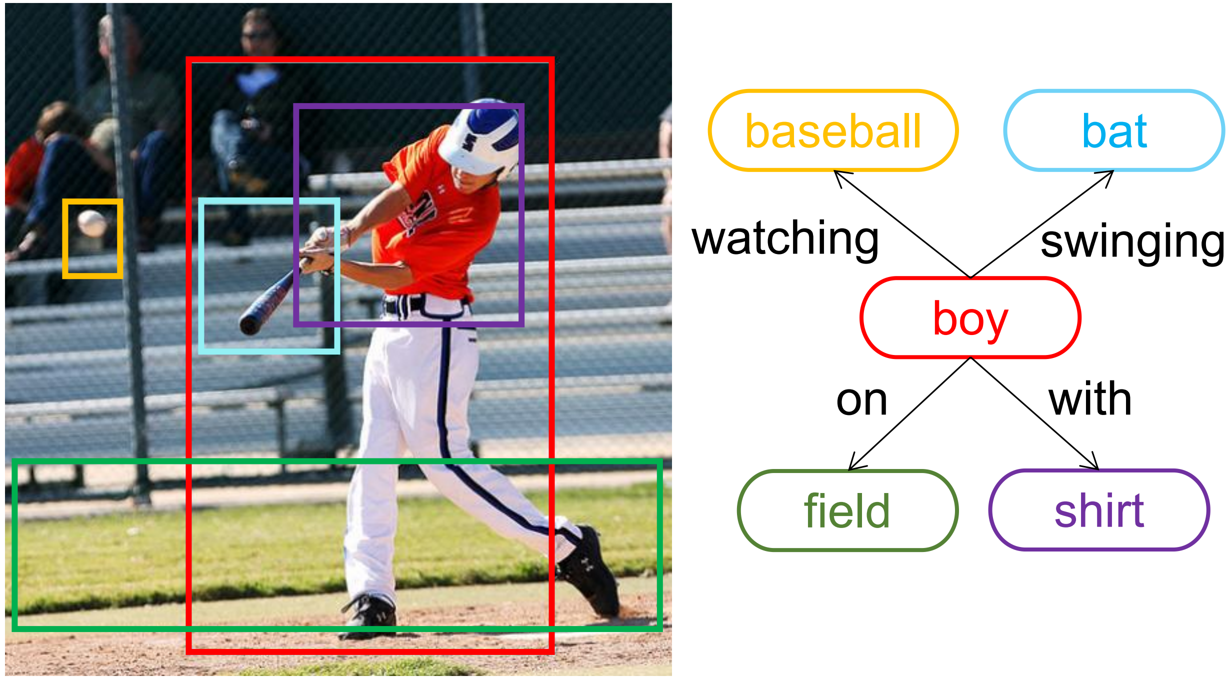
- localize holistic object instances
- recognize their relationships

Challenges

- long-tail data distribution
- sparse samples on triplet categories
- large intra-class variation and high inter-class similarity

Application

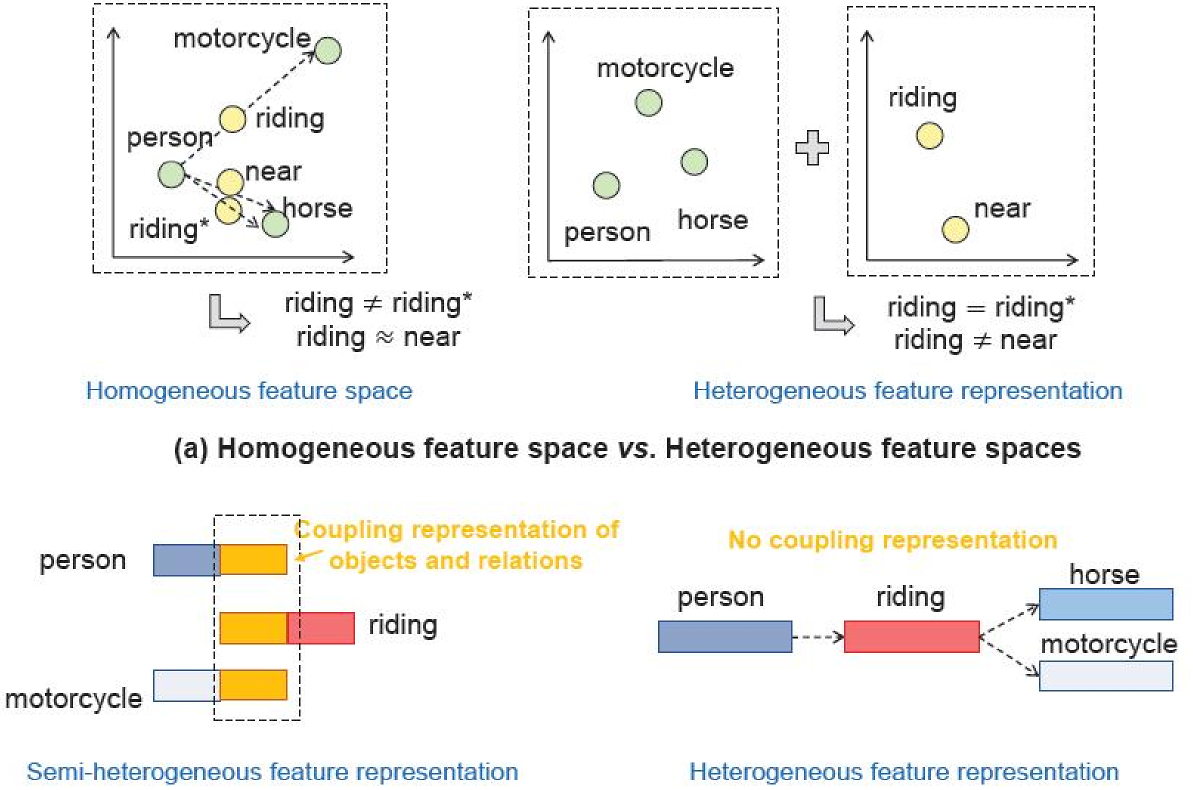
- captioning
- retrieval
- visual question answering
- multi-modal dialog



An example of scene graph

Motivation

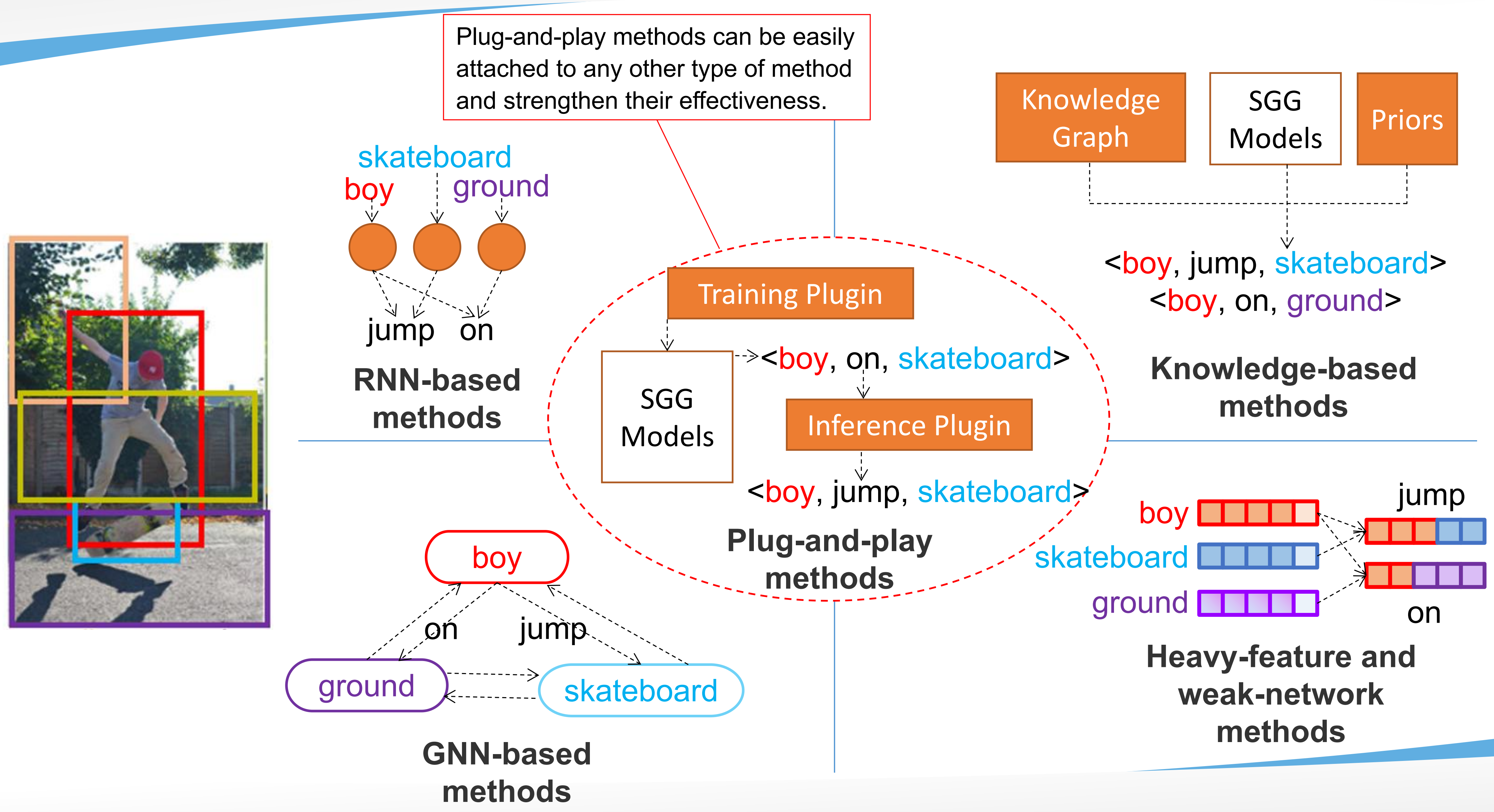
- Heterogeneity between objects and relationships has not been discussed yet.
- Heterogeneous objects and relation feature spaces can alleviate the large intra-class variation and inter-class ambiguity problem.



(a) Homogeneous feature space *vs.* Heterogeneous feature spaces

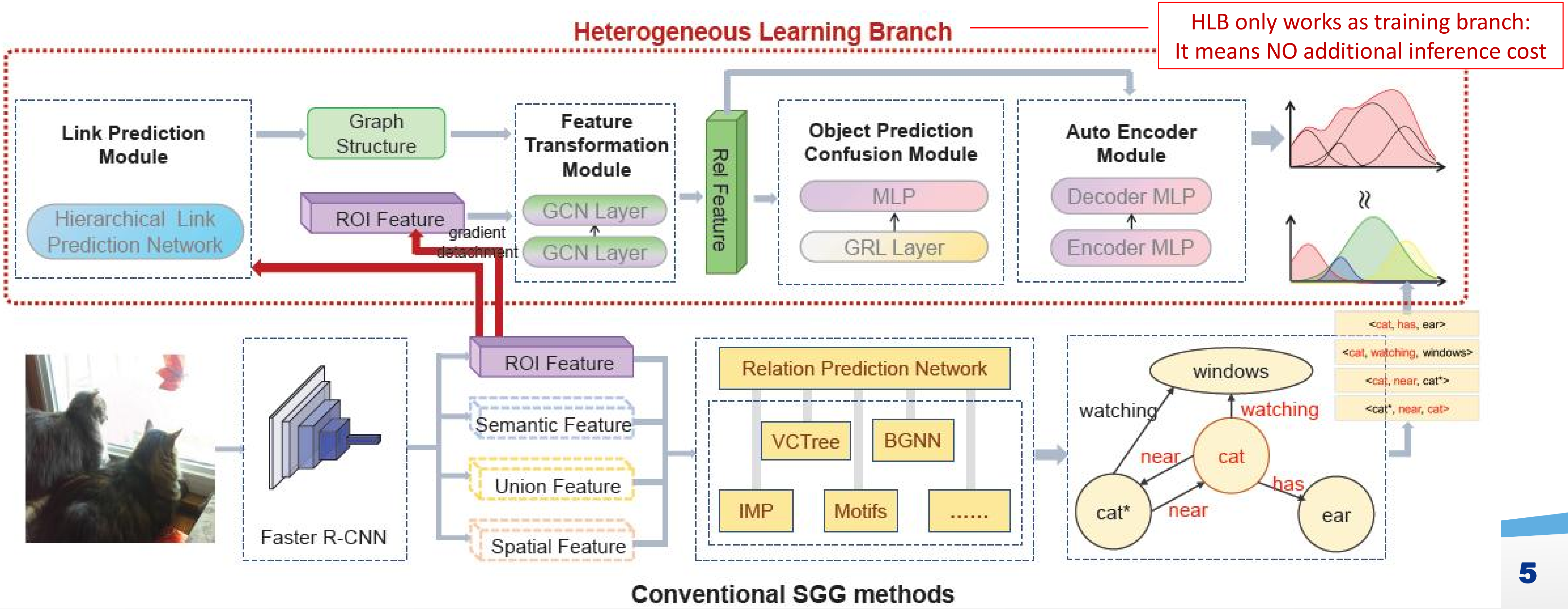
(b) Semi-heterogeneous feature representation *vs.* Heterogeneous feature representation

Related Work



Our Method – Framework

- Initialize relation representation with Feature Transformation Module
- Find possible relation proposals with Link Prediction Module
- Construct heterogeneous object and relation features spaces with Object Prediction Confusion Module
- Propagate the heterogeneity to arbitrary SGG relation predictors with Auto Encoder Module



Our Method – Feature Transformation

MAGUS
MediA recoGnition and UnderStanding

- GNN usually suffers from over-smooth problem
- Severe long-tail problem in VG dataset exacerbates the over-smooth problem
- Alleviate over-smooth problem by enhancing each node's original feature

$$x_i' = \sigma(\omega_1 \cdot F_{j\in N(i)}(x_j)) \quad \begin{cases} F(x): mean(x), & in\ HLB \\ F(x): \sum \frac{e_{j,i}}{\sqrt{d_j d_i}} x, & in\ GCN \\ F(x): \sum a_{i,j} x, & in\ GAT \end{cases}$$

$$x_i' = \sigma(\omega_1 \cdot F_{j\in N(i)}(x_j) + \omega_2 \cdot x_i)$$

over-smooth-proof item

	SGDet					
	mR@20	mR@50	mR@100	R@20	R@50	R@100
GCN	3.14	4.17	4.83	23.05	29.58	33.54
GCN+	**4.02**	**5.45**	**6.32**	**24.53**	**31.47**	**35.52**
GAT	3.96	5.35	6.21	**24.55**	**31.41**	**35.50**
GAT+	**4.06**	**5.50**	**6.40**	24.38	31.28	35.29
HLB-	3.96	5.27	6.09	24.36	31.08	35.13
HLB	**4.34**	**5.87**	**6.84**	**24.78**	**31.79**	**35.91**

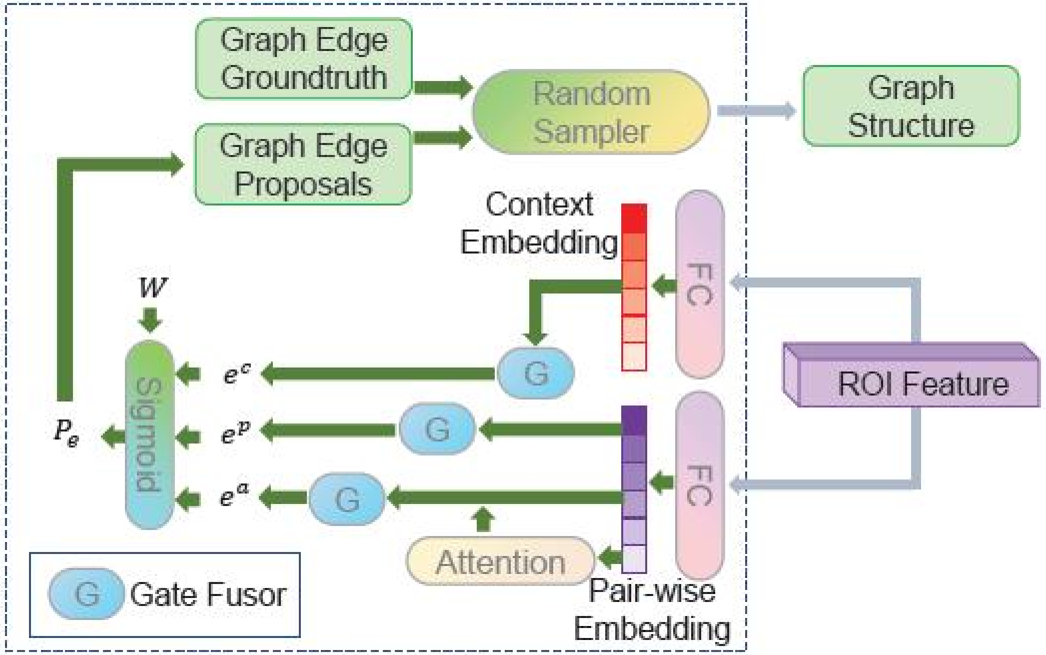
Our Method – Link Prediction

- **Hierarchical Link Prediction Module**
  - Probability between two isolated objects ($P_{(i,j)}$)
  - Probability between two objects with consideration of context ($P_{(i,j)|context}$)
  - Probability of all possible existing relations ($P_{R|context}$)

What is the relationship between these two object?

In current scene, what is the relationship between these two object?

What could happen in current scene?



Our Method – Object Prediction Confusion

Relevance between high-dimension tensors is difficult in quantification

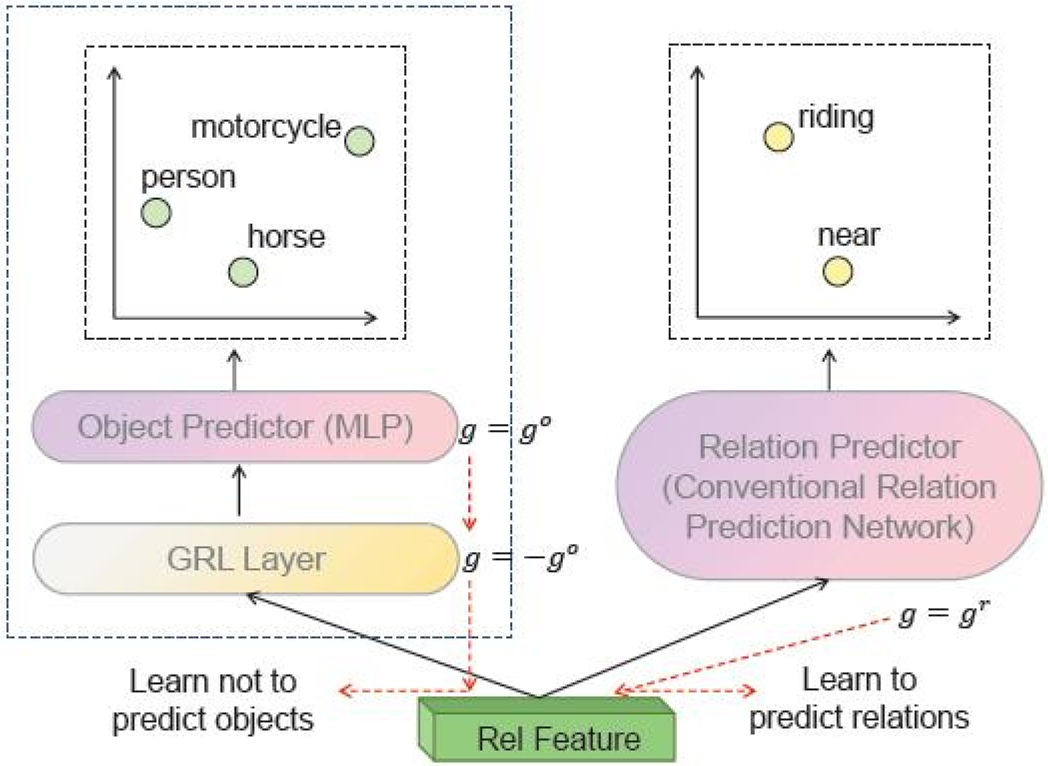
- make the object and relation features less relative

Problem ↓ Reduction

- make relation features contain minimum object information

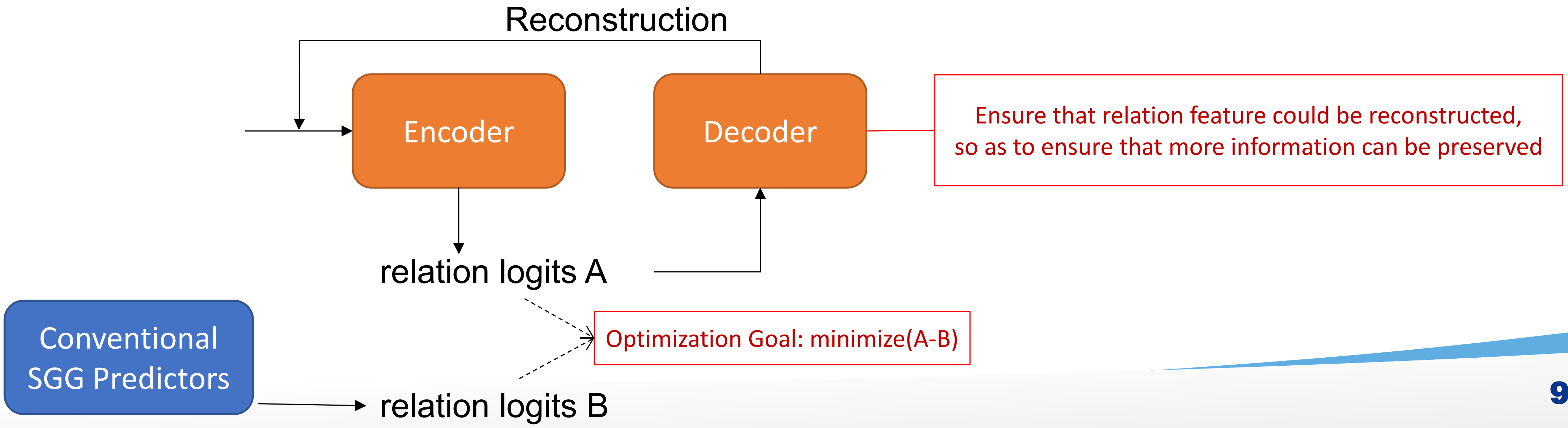
Problem ↓ Reduction

- make relation features cannot be used in object recognition



Our Method – Auto Encoder

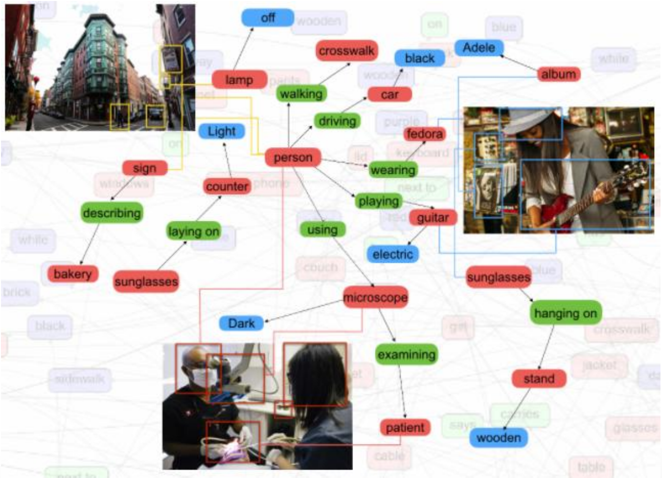
- Defect of a classifier (only Encoder):
  - Since training two different classifiers simultaneously (classifier from conventional SGG method and that from our HLB method) is difficult, a possible dilemma is both the classifiers tend to generate logits that close to zero (sparse logits) to make the loss value seems to decease.
- Advantage of an Auto Encoder (Encoder + Decoder):
  - The better the prediction logits can be reconstructed, the more information is preserved in the logits. It means that the classifier/Encoder tend to generate dense/non-zero logits.



Experiment Settings

Datasets: Visual Genome (VG-150)

- 108,077 images
- 1,366,673 object instances
- 1,531,448 relation instances
- 108,249 isolated scene graphs
- 150 object categories
- 50 relation categories



Tasks

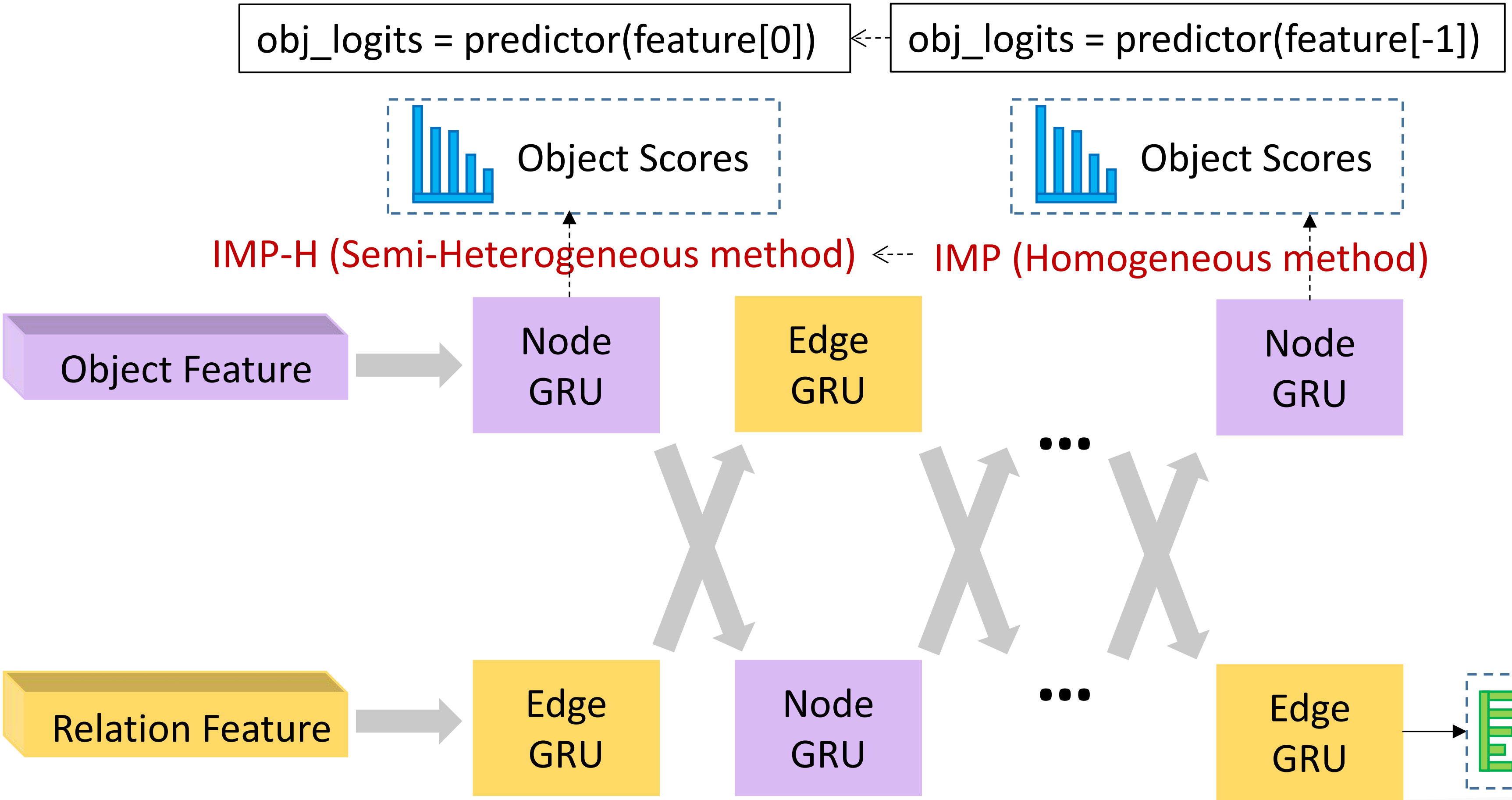
- Scene Graph Detection
- Scene Graph Classification
- Predicate Classification

Evaluation metrics

- R@N (recall in top-N results)
- mR@N (mean recall over classes in top-N results)
- ng-R@N (no graph-constraint recall in top-N results)
- zR@N (zero-shot recall in top-N results)

Experimental Results – Toy Experiments

Restricted fusion between object and relation features can significantly improves the performance, i.e., heterogeneity between object and relation features is worth noting.



SGDet	mR		R	
	@50	@100	@50	@100
IMP	3.8	4.8	20.7	24.5
IMP-H	**5.37**	**6.30**	**31.21**	**35.36**

Experimental Results – Comparison Results

MAGUS
MediA recoGnition and UnderStanding

		PredCls				SGCls				SGDet			
	Model	mR@50	R@50	mR@100	R@100	mR@50	R@50	mR@100	R@100	mR@50	R@50	mR@100	R@100
	GBNet-β [34]	22.1	66.6	24.0	68.2	12.7	37.3	13.4	38.0	7.1	26.3	8.5	29.9
	Graph R-CNN [32]	16.4	54.2	17.2	59.1	9.0	29.6	9.5	31.6	5.8	11.4	6.6	13.7
	RelDN [38]	15.8	68.7	17.2	68.8	9.3	38.9	9.6	38.9	6.0	31.0	7.3	36.7
	FCSGG [17]	6.3	41.0	7.1	45.0	3.7	23.5	4.1	25.7	3.6	21.3	4.2	25.1
	GPS-Net [31]	19.2	69.7	21.4	69.7	11.7	42.3	12.5	42.3	7.4	28.9	9.5	33.2
RNN-based	IMP [29]	9.8	59.3	10.5	61.3	5.8	34.6	6.0	35.4	3.8	20.7	4.8	24.5
	IMP+HLB	**10.63**	**60.91**	**11.37**	**62.95**	**6.62**	**38.10**	**6.98**	**39.01**	**4.19**	**26.67**	**5.23**	**31.85**
	IMP-H	10.17	58.89	10.97	61.31	6.05	34.89	6.47	36.59	5.37	31.21	6.30	35.36
	IMP-H+HLB	**10.44**	**59.43**	**11.17**	**61.52**	**7.07**	**38.21**	**7.47**	**39.09**	**5.87**	**31.79**	**6.84**	**35.91**
Heavy-Feature	VTransE [37]	14.7	65.7	15.8	67.6	8.2	38.6	8.7	39.4	5.0	29.7	6.0	34.3
	VTransE+HLB	**15.26**	65.68	**16.40**	67.60	8.24	**39.72**	8.74	**40.61**	**5.14**	29.74	**6.22**	**34.47**
Knowledge-based + GNN	KERN [3]	**17.7**	**65.8**	**19.2**	**67.6**	**9.4**	36.7	**10.0**	37.4	6.4	27.1	7.3	29.8
	KERN+HLB	15.89	61.17	17.15	64.17	9.01	**38.16**	9.69	**39.37**	**7.11**	**28.70**	**8.58**	**33.41**
Knowledge-based + RNN	MOTIFS [36]	14.0	**65.2**	15.3	**67.1**	7.7	35.8	8.2	36.5	5.7	27.2	6.6	30.3
	MOTIFS+HLB	**15.39**	64.91	**16.74**	66.80	**8.90**	**39.48**	**9.44**	**40.32**	**7.19**	**32.57**	**8.43**	**37.01**
Tree-RNN	VCTree-SL [24]	17.0	**66.2**	18.5	**67.9**	9.8	**37.9**	10.5	**38.6**	6.7	27.7	7.7	31.1
	VCTree-SL+HLB	**17.47**	65.73	**18.79**	67.35	**11.98**	36.95	**12.73**	38.50	**7.46**	**32.04**	**8.75**	**36.34**
GNN-based	BGNN [13]	**30.4**	59.2	**32.9**	61.3	14.3	**37.4**	16.5	**38.5**	10.7	**31.0**	12.6	**35.8**
	BGNN+HLB	28.20	**61.06**	30.43	**63.22**	**16.72**	35.27	**18.09**	36.64	**12.57**	27.80	**15.03**	32.28
	On Average	+1.45%	-0.21%	+0.96%	-0.02%	+11.73%	+4.08%	+10.74%	+4.39%	+12.63%	+8.83%	+14.20%	+10.48%
		+0.54%				+7.73%				+11.53%			

Comparison with the state-of-the-arts methods

Experimental Results – Quantitative Analysis

- **Component Analysis**
  - AD: remove decoder from Auto-Encoder
  - LP: remove Link Prediction Module
  - GE: remove over-smooth-proof item from GNN
- **Feature Representation Analysis**

higher intra-class similarity

less inter-class ambiguity

Feature representation analysis

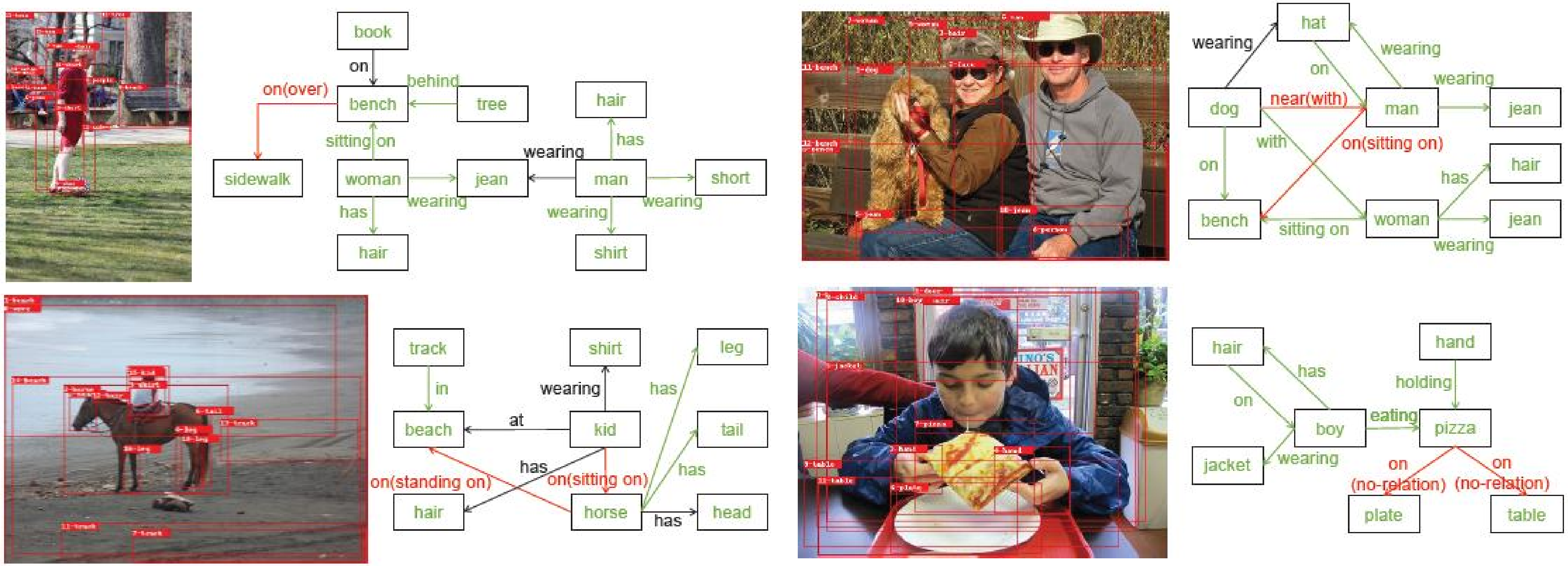
	PredCls		SGCls		SGDet	
	mR @20 @50 @100	R @20 @50 @100	mR @20 @50 @100	R @20 @50 @100	mR @20 @50 @100	R @20 @50 @100
IMP-H	8.04 10.17 10.97	51.47 58.89 61.31	5.01 6.05 6.47	30.37 34.89 36.59	3.98 5.37 6.30	24.45 31.21 35.36
IMP-H-AD	7.78 9.67 10.43	51.67 58.95 61.38	4.83 5.80 6.21	30.49 35.00 36.68	3.90 5.25 6.14	24.49 31.32 35.49
IMP-H-LP	7.72 9.54 10.24	51.61 58.94 61.36	4.76 5.69 6.09	30.42 34.91 36.56	4.02 5.38 6.23	24.44 31.26 35.39
IMP-H-GE	7.76 9.82 10.66	50.74 58.28 60.90	4.83 5.85 6.27	29.85 34.20 35.80	3.87 5.27 6.23	23.17 30.03 34.36
IMP-H-HLB	**8.50 10.44 11.17**	**52.73 59.43 61.52**	**5.84 7.07 7.47**	**34.89 38.21 39.09**	**4.34 5.87 6.84**	**24.78 31.79 35.91**

Component analysis

Experimental Results – Qualitative Analysis

MAGUS
MediA recoGnition and UnderStanding

- The words marked with green denote the correctly detected objects and relations
- The red words and lines represent the wrongly predicted ones with notated labels in brackets
- The **words marked with black** color refer to the predicted relations which are considered **positive but unlabeled**



Qualitative results of the proposed method

Thank You

heyq@smail.nju.edu.cn

南京大學
NANJING UNIVERSITY
南京理工大学
NANJING UNIVERSITY OF SCIENCE & TECHNOLOGY