CVPR 2019

Long Beach, CA

June 16th - June 20th

Scene Graph Generation and Its Application to Vision-and-Language Tasks

Jianwei Yang @ Georgia Tech

Georgia Tech

06/16/2019

What is scene graph?

Image as a single label

Image Source: ImageNet

Image as an object set

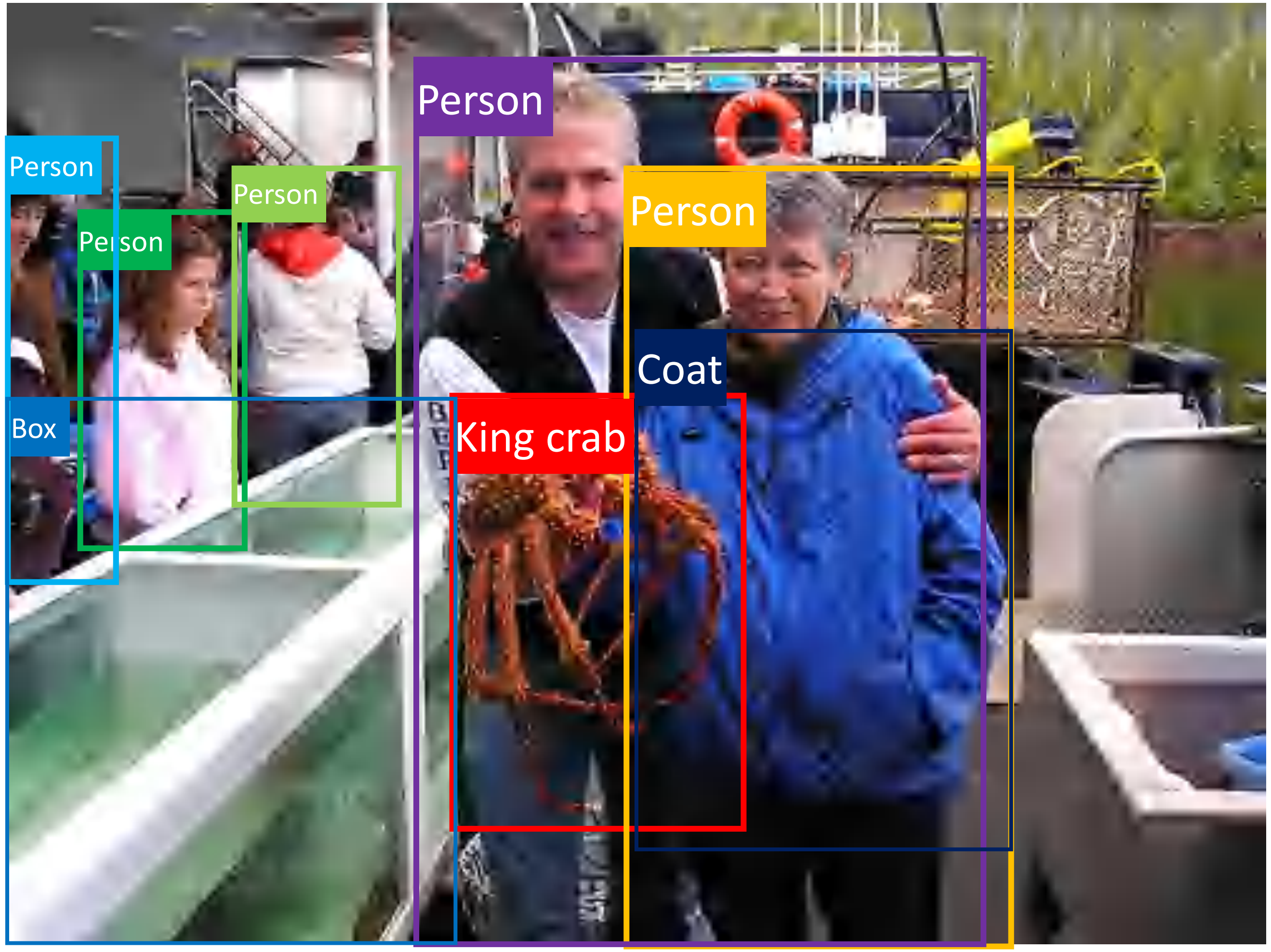
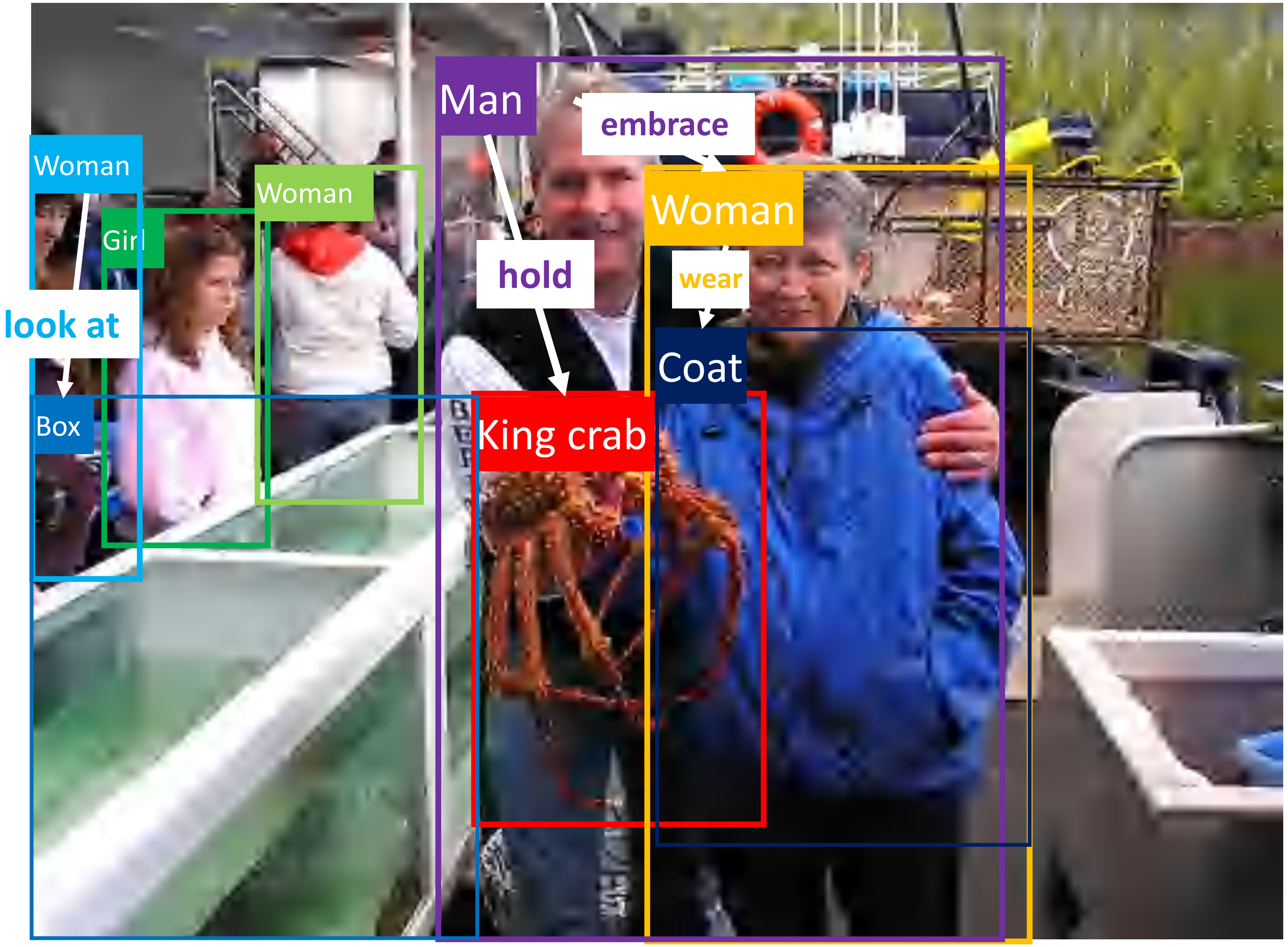


Image Source: ImageNet

Image as a scene graph

“Woman look at box”

“Man hold king crab”



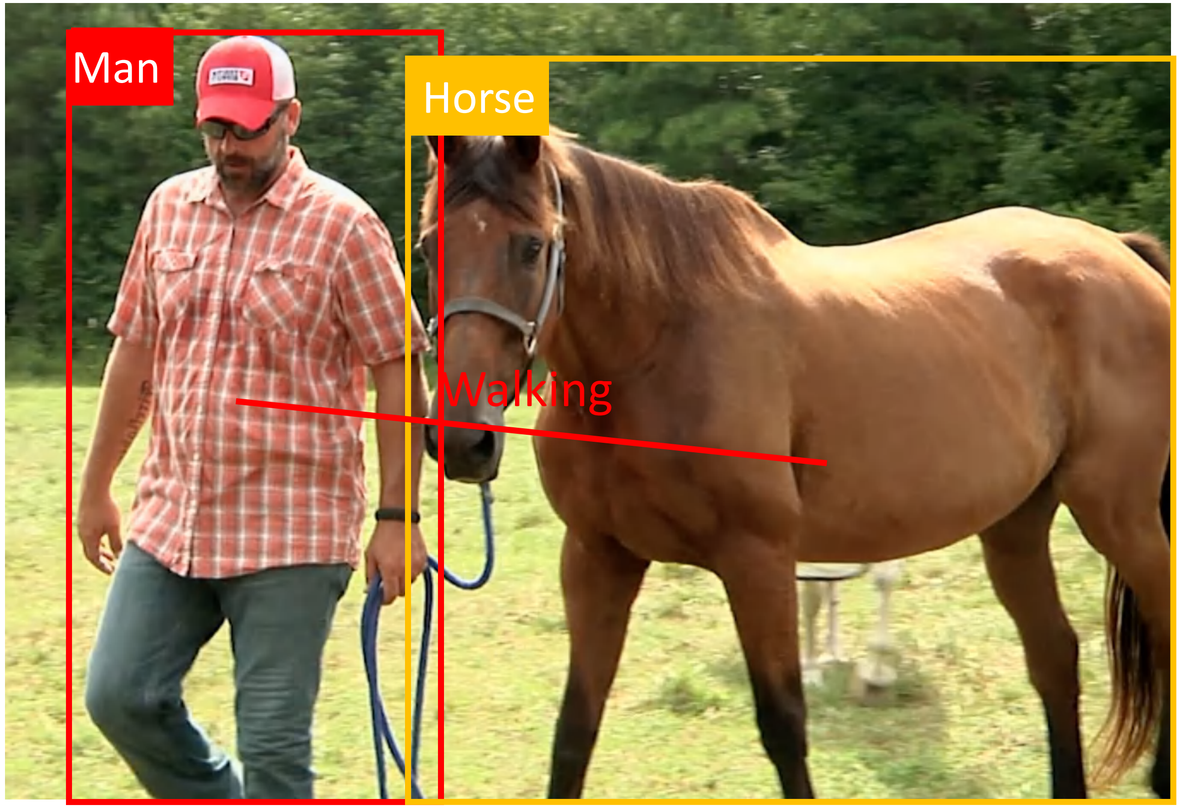
“Woman wear coat”

“Man embrace woman”

Image Source: ImageNet

Why we need scene graph?

Distinguish images more accurately

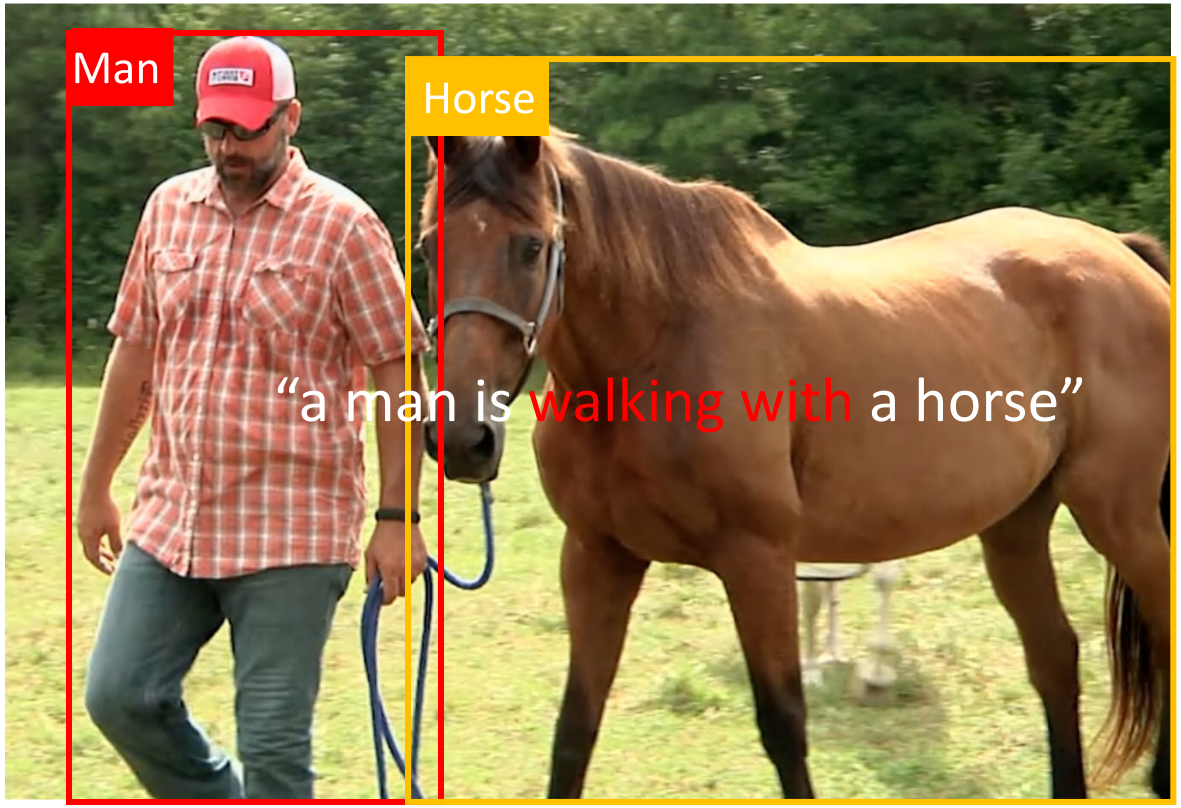


[1] Image Retrieval using Scene Graphs. Johnson et al. CVPR 2015

Left: https://cals.ncsu.edu/wp-content/uploads/2016/08/horse-1500x931.png
Rigth: https://www.videoblocks.com/video/the-man-in-hat-feed-a-brown-horse-with-flowers-on-the-meadow-supmox_3xj0tvkb67

Why we need scene graph?

Describe images more grounding

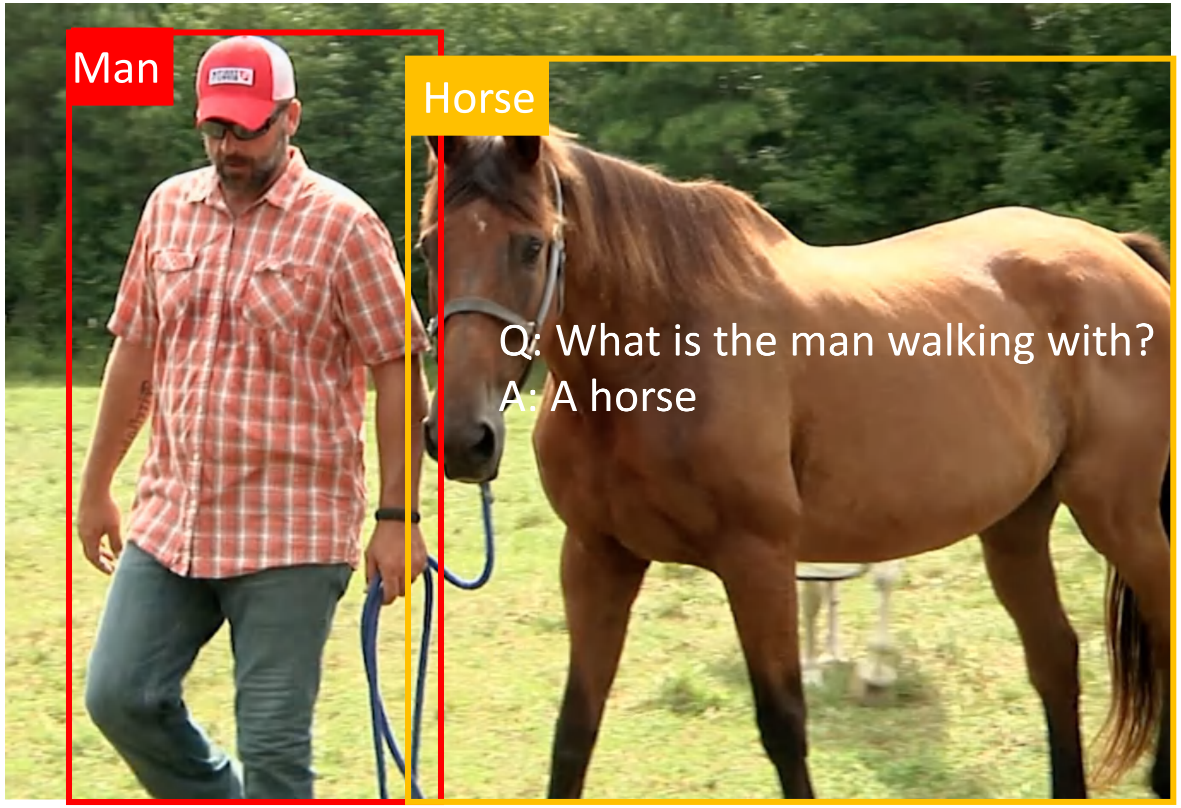


[1]. Auto-Encoding Scene Graphs for Image Captioning. Yang et al. arXiv 2018
[2]. Exploring Visual Relationship for Image Captioning. Yao et al. ECCV 2018

Left: https://cals.ncsu.edu/wp-content/uploads/2016/08/horse-1500x931.png
Rigth: https://www.videoblocks.com/video/the-man-in-hat-feed-a-brown-horse-with-flowers-on-the-meadow-supmox_3xj0tvkb67

Why we need scene graph?

Answer question more precisely



[1] Graph-Structured Representations for Visual Question Answering. Teney et al. CVPR 2017
[2] Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. Yi et al. Neurips 2018

Left: https://cals.ncsu.edu/wp-content/uploads/2016/08/horse-1500x931.png
Rigth: https://www.videoblocks.com/video/the-man-in-hat-feed-a-brown-horse-with-flowers-on-the-meadow-supmox_3xj0tvkb67

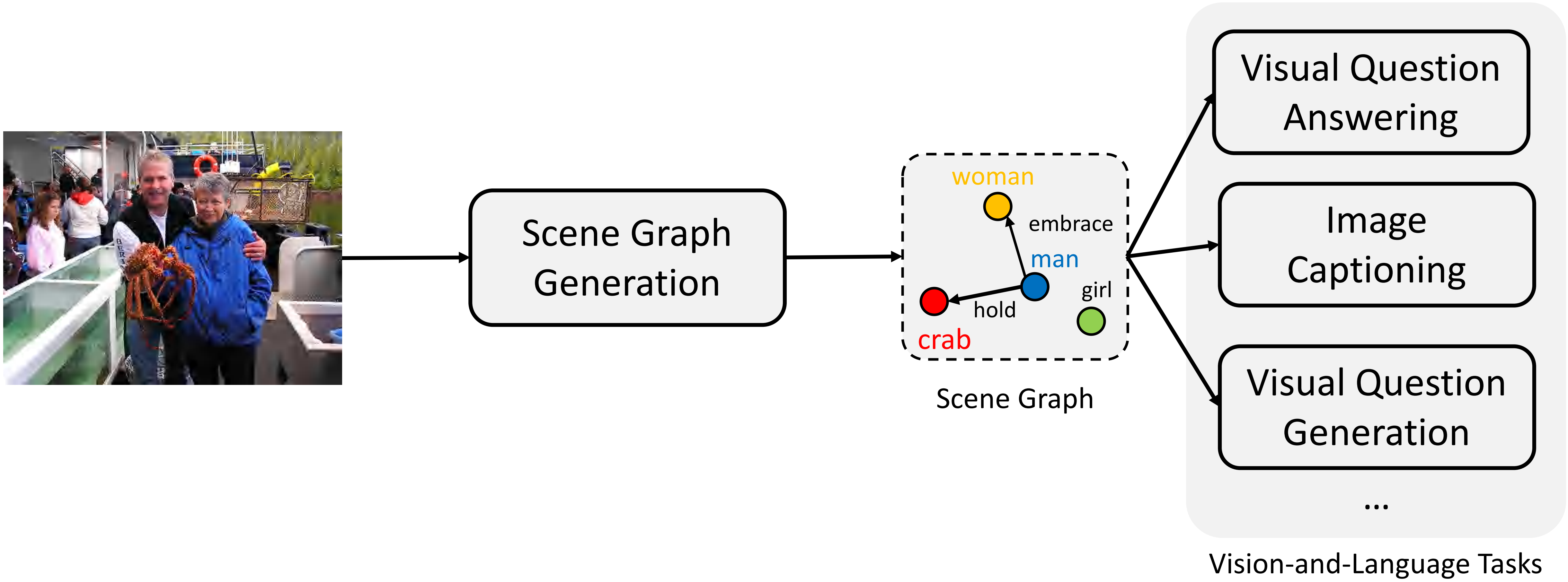
Why we need scene graph?

Generate questions more grounding

[1] Visual Curiosity: Learning to Ask Questions to Learn Visual Recognition. Yang et al. CoRL 2018

Left: https://cals.ncsu.edu/wp-content/uploads/2016/08/horse-1500x931.png
Rigth: https://www.videoblocks.com/video/the-man-in-hat-feed-a-brown-horse-with-flowers-on-the-meadow-supmox_3xj0tvkb67

In this tutorial

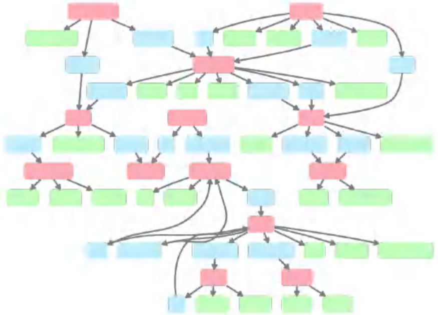


Part 1: Scene Graph Generation

Datasets

Scene Graphs 5K

Johnson et al, CVPR 2015



- 5000 images
- 6745 object categories
- 1310 relationship types
- Long-tailed

Visual Relationships

Lu et al, ECCV 2016

- 5000 images
- 100 object categories
- 70 relationship types
- Fully-annotated

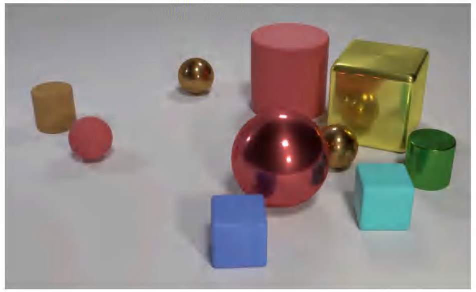
Visual Genome

Krishna et al, IJCV 2017

- 108K images
- 33K object categories
- 42K relationship types
- Long-tailed

CLEVR

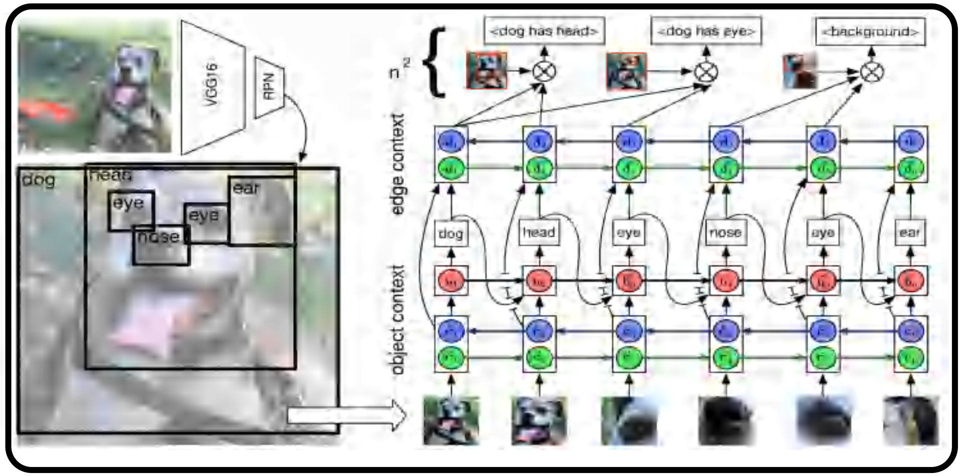
Johnson et al, CVPR 2017



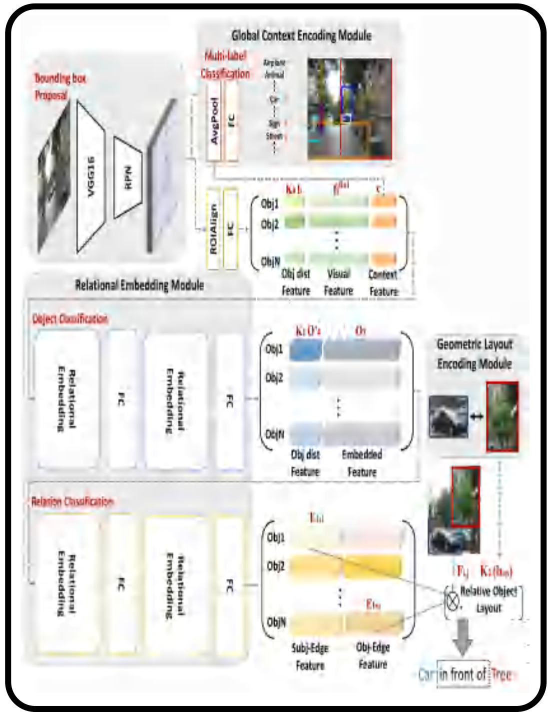
- 100K images
- 3 object categories
- 8 relationship types
- Fully-annotated

Slide Credit: Justin Johnson

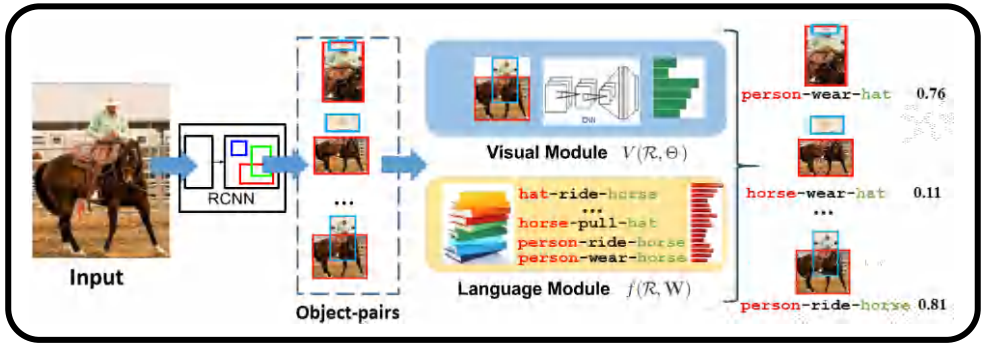
Models



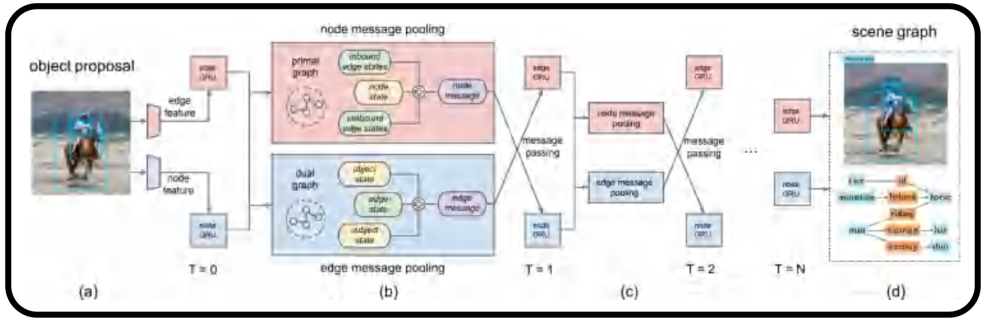
Neural Motif Network, Zellers et al. CVPR 2018



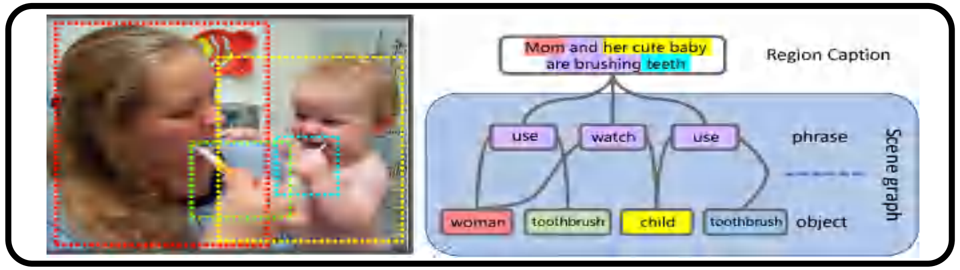
LinkNet, Woo et al. Neurips 2018



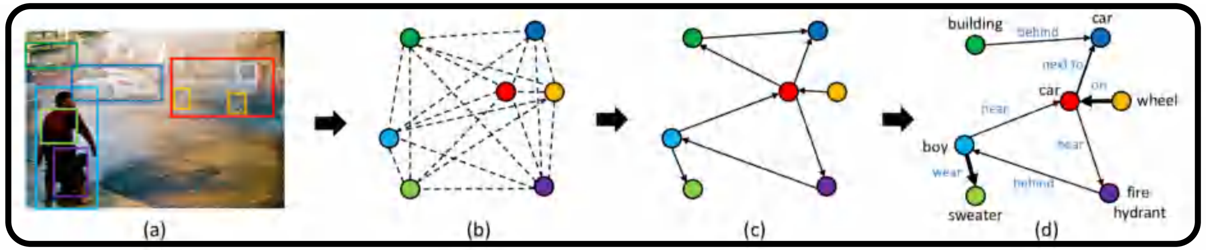
Language Prior, Lu et al. ECCV 2016



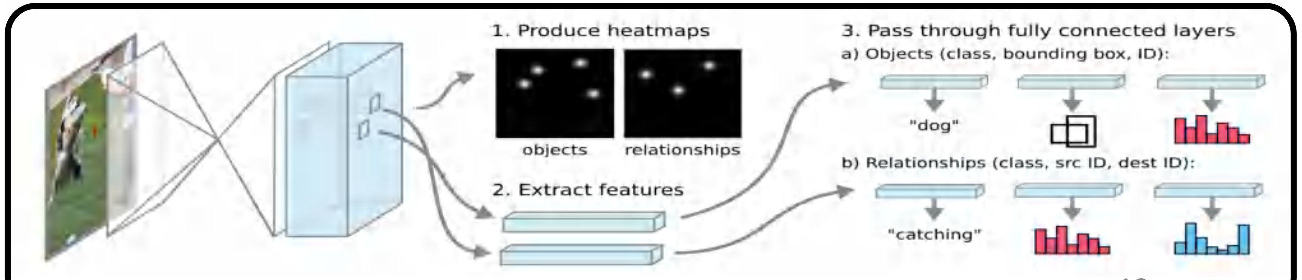
IMP, Xu et al. CVPR 2017



MSDN. Li et al. ICCV 2017



Graph R-CNN. Yang et al. ECCV 2018



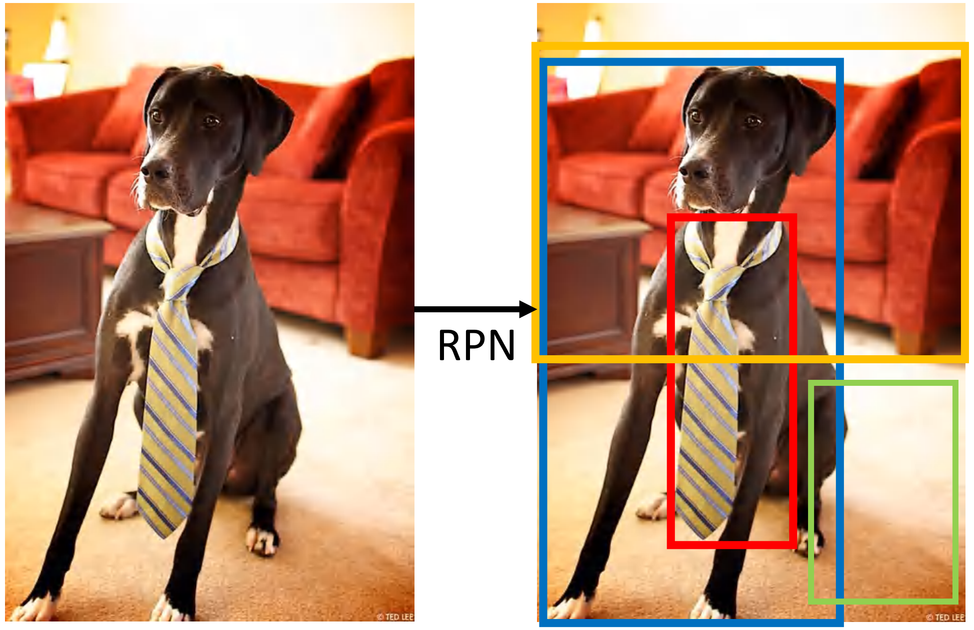
Pixel2Graph. Newell et al. Neurips 2018

Base Model

Base Model

Input

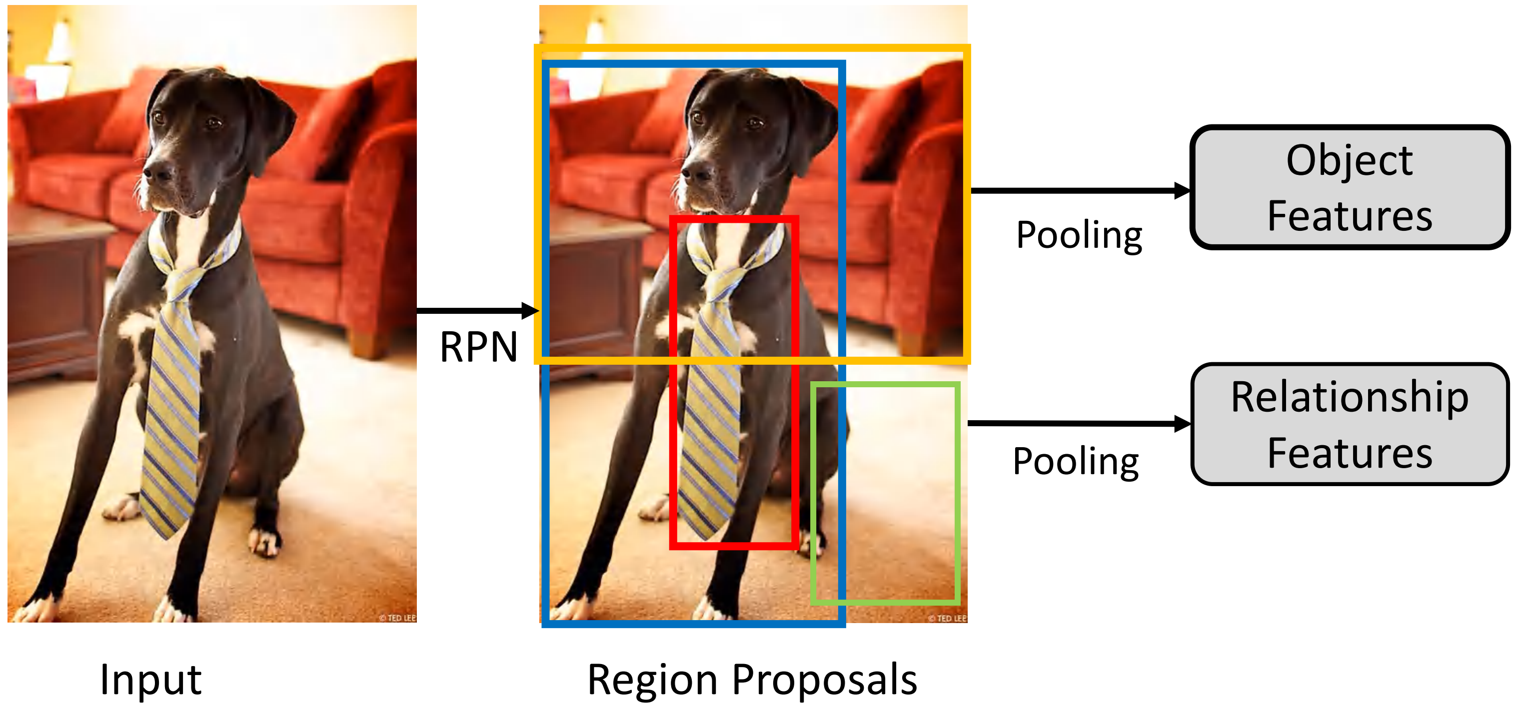
Base Model



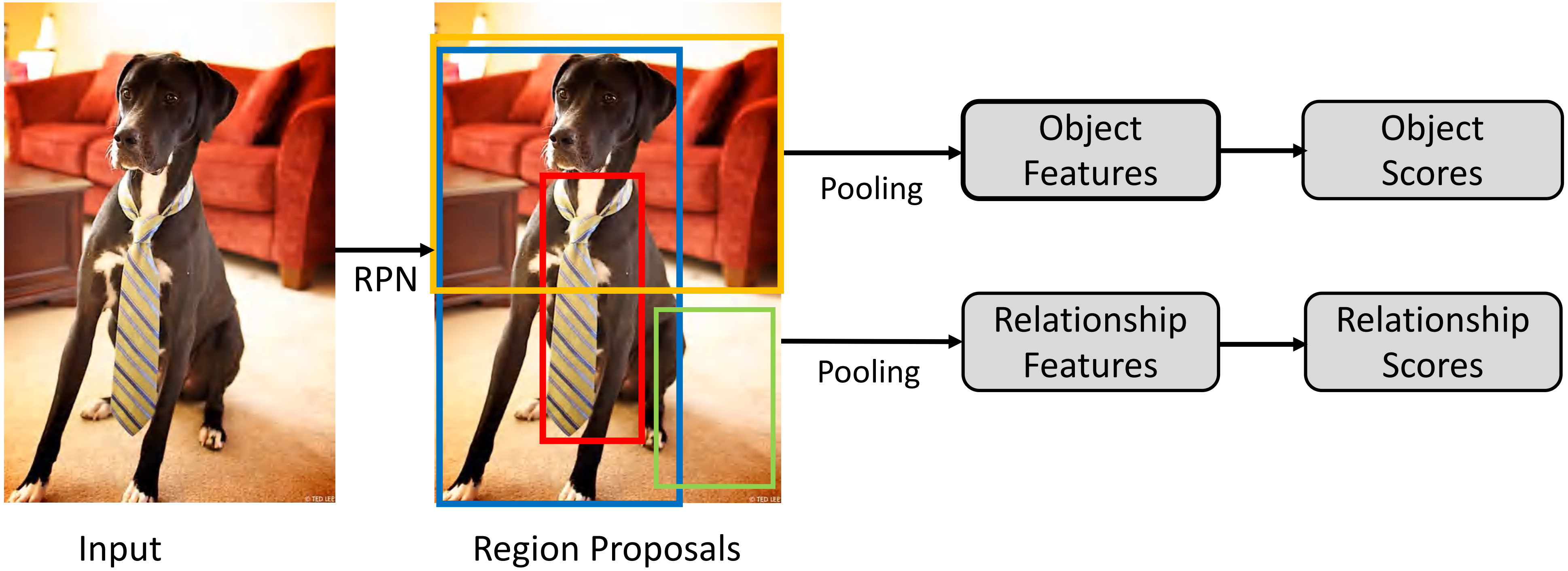
Input

Region Proposals

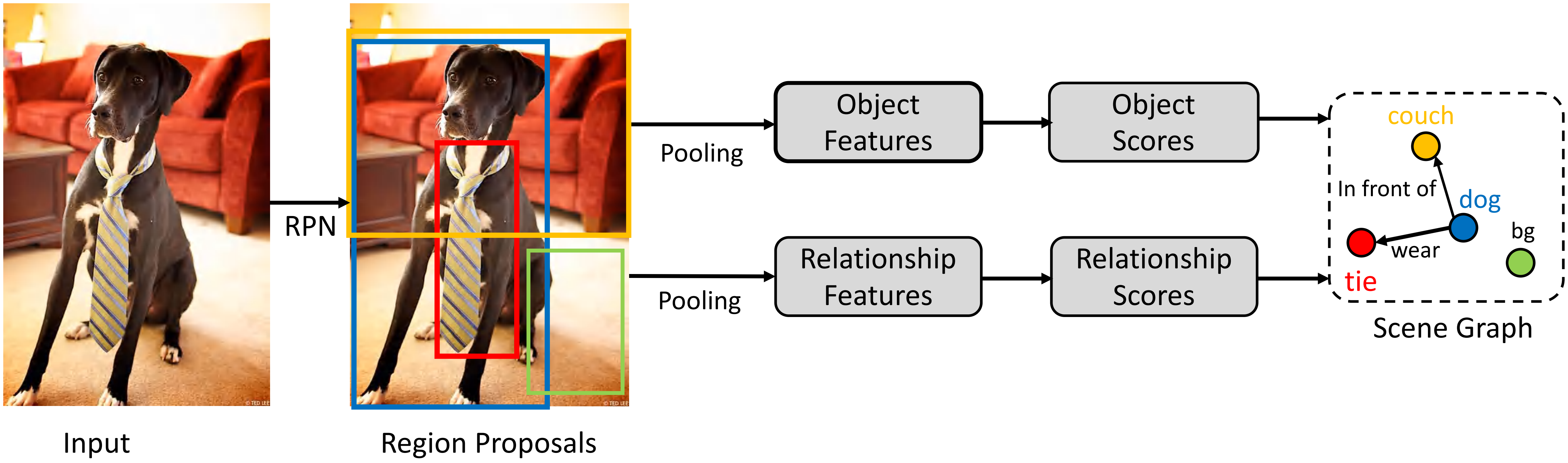
Base Model


RPN
Pooling
Object Features
Pooling
Relationship Features
Input
Region Proposals

Base Model

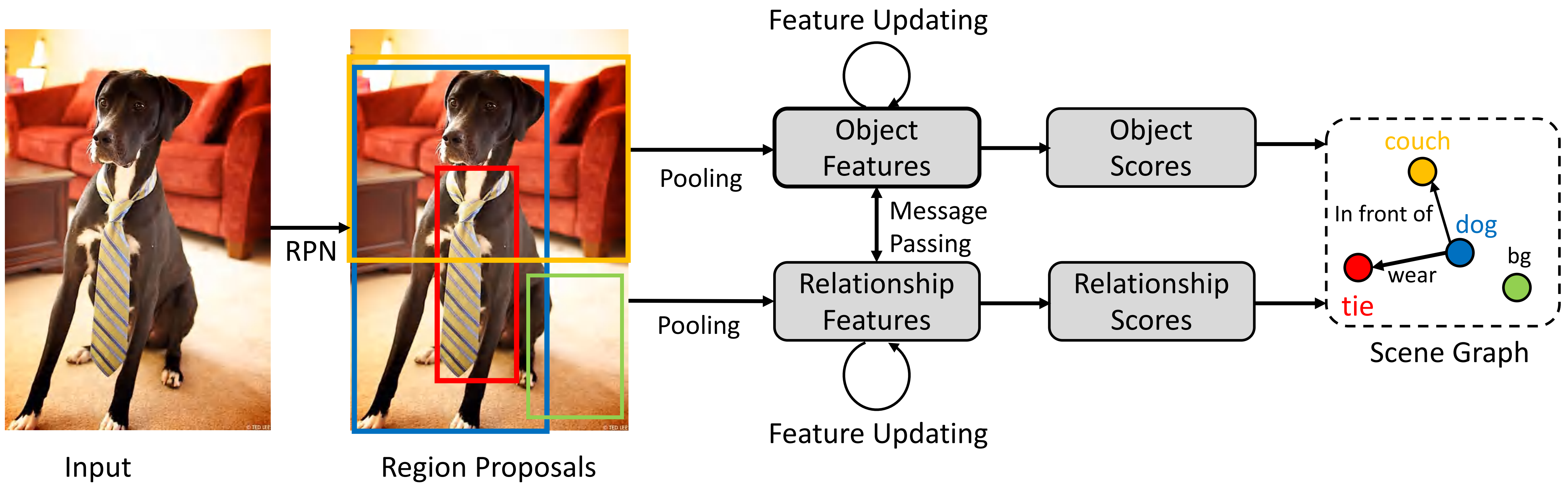

RPN
Pooling
Object Features
Object Scores
Pooling
Relationship Features
Relationship Scores
Input
Region Proposals

Base Model



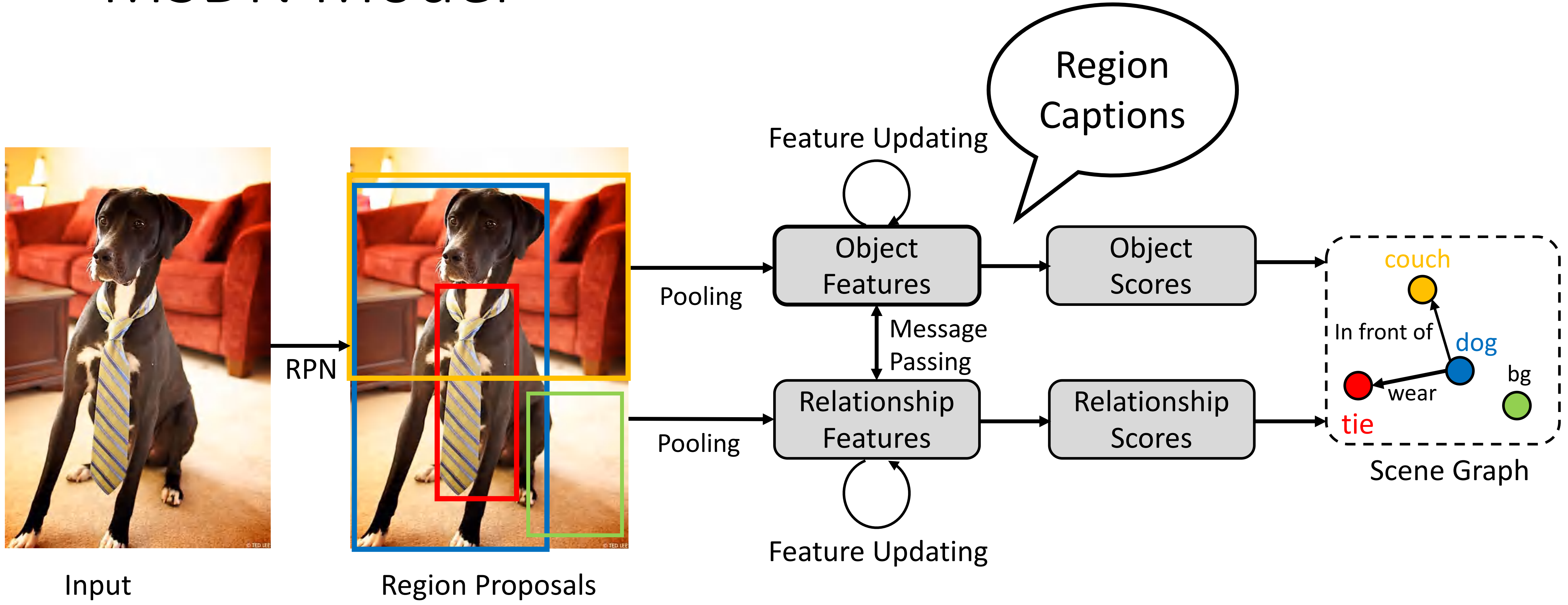
RPN
Pooling
Pooling
Object Features
Object Scores
Relationship Features
Relationship Scores
couch
In front of
dog
bg
wear
tie
Scene Graph
Input
Region Proposals

IMP Model



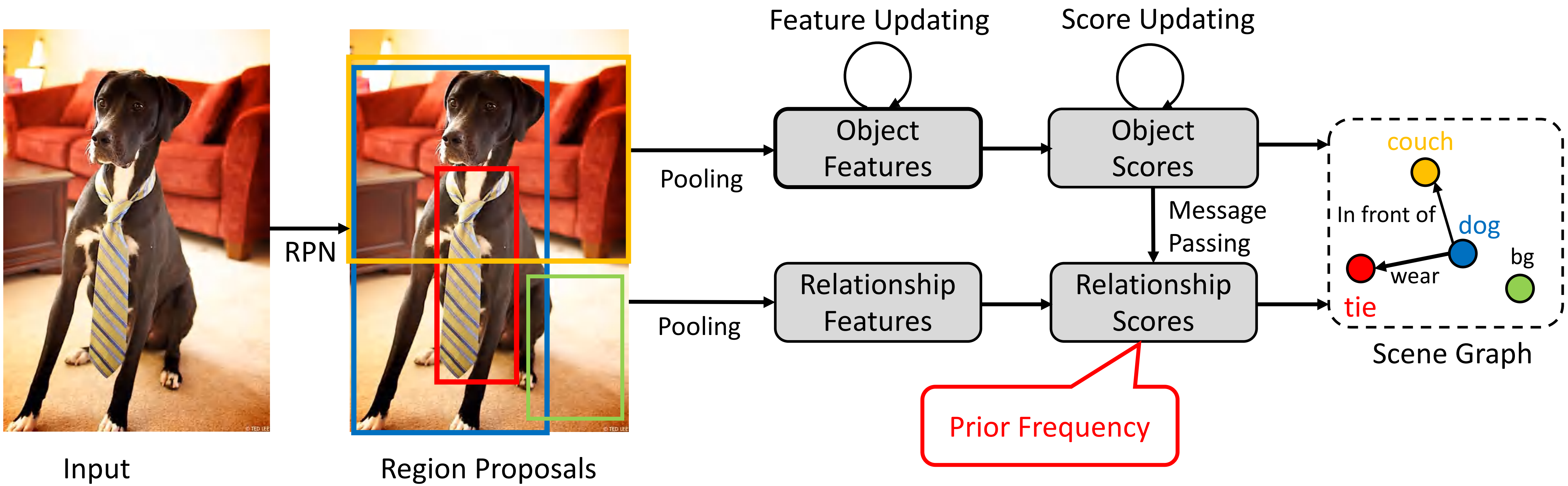
Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

MSDN Model



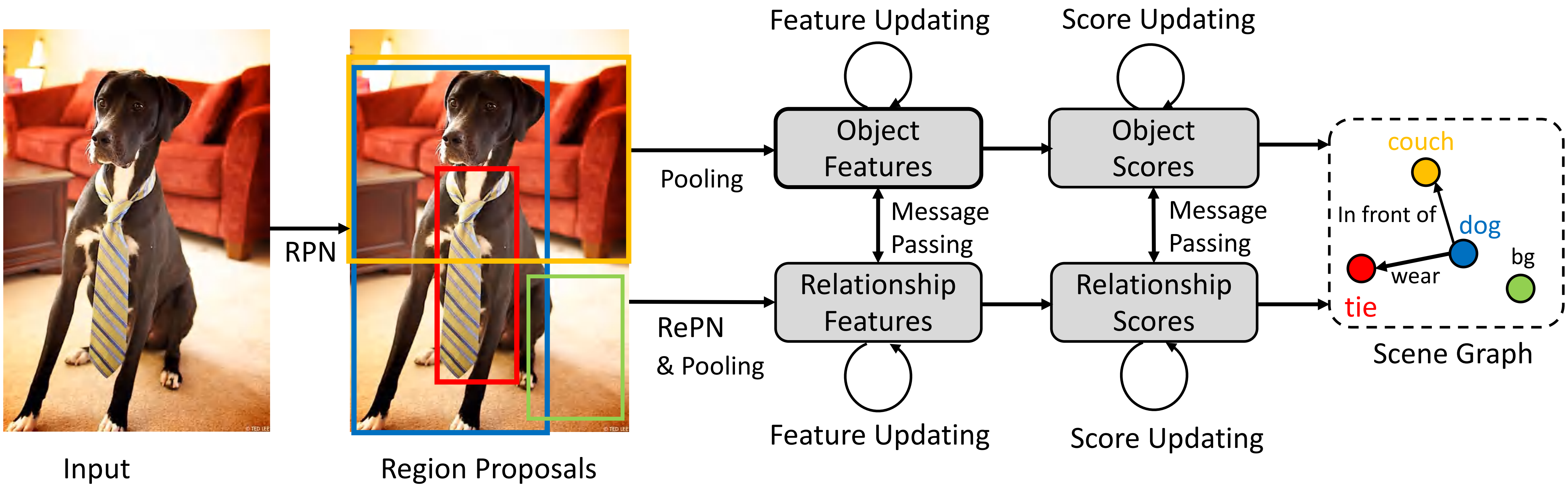
Scene Graph Generation from Objects, Phrases and Region Captions. Li et al. ICCV 2017

Neural Motif Network



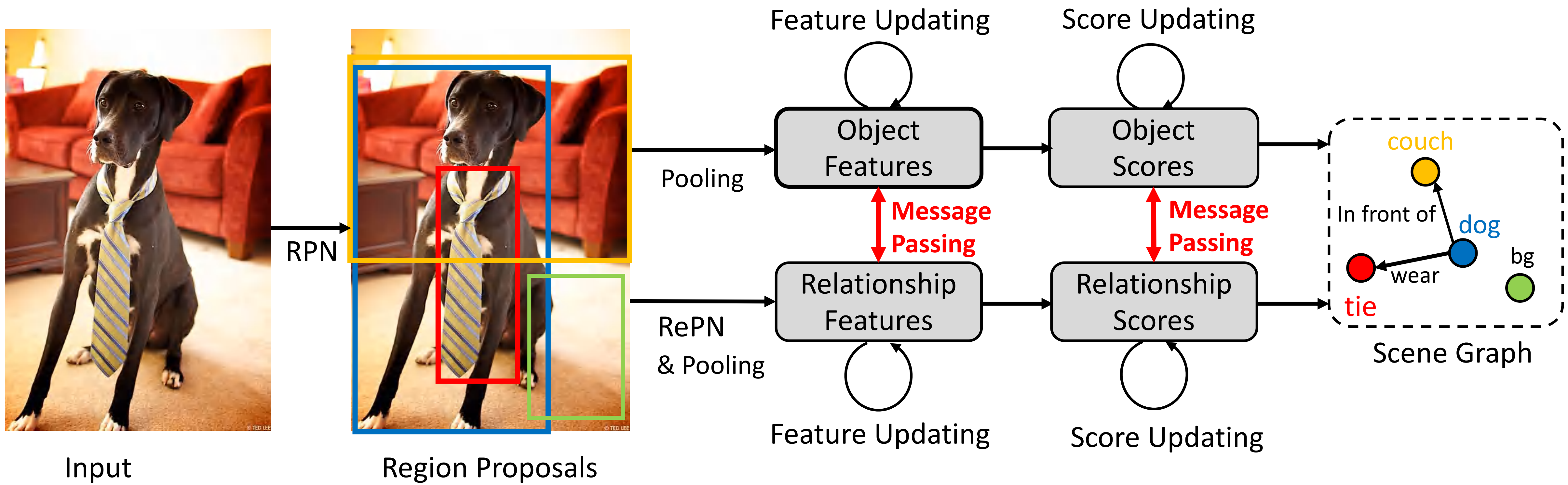
Neural Motifs: Scene Graph Parsing with Global Context. Zellers et al. CVPR 2018

Our model: Graph R-CNN



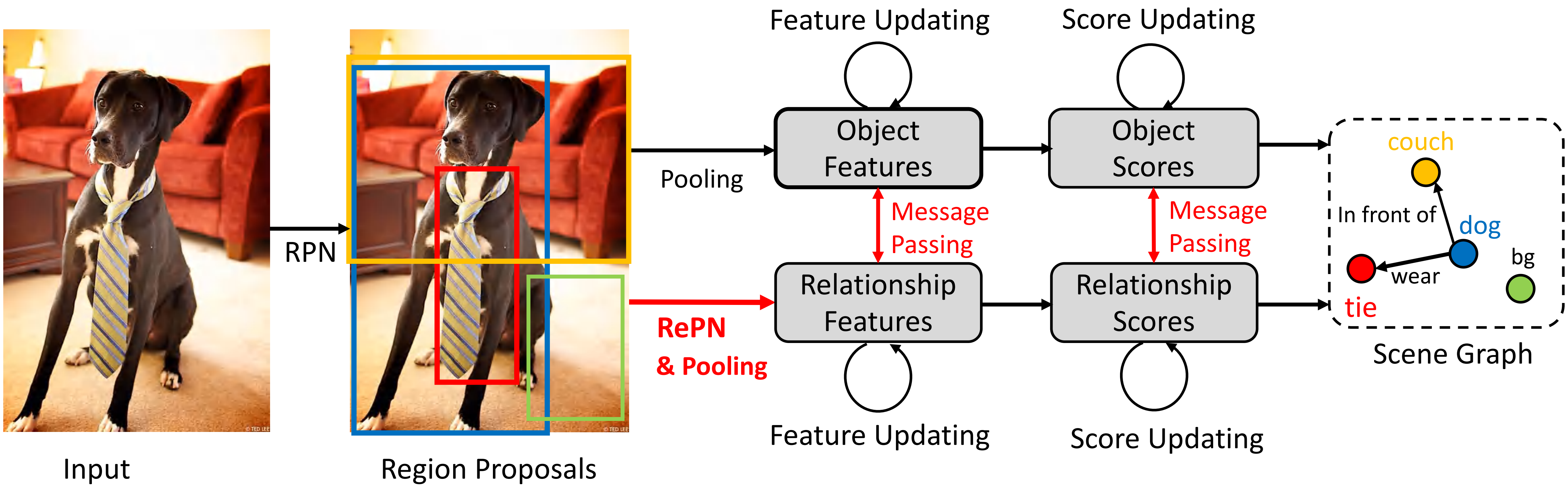
Jianwei Yang*, Jiasen Lu*, Stefan Lee, Dhruv Batra, Devi Parikh. Graph R-CNN for Scene Graph Generation. ECCV 2018.

Our model: Graph R-CNN



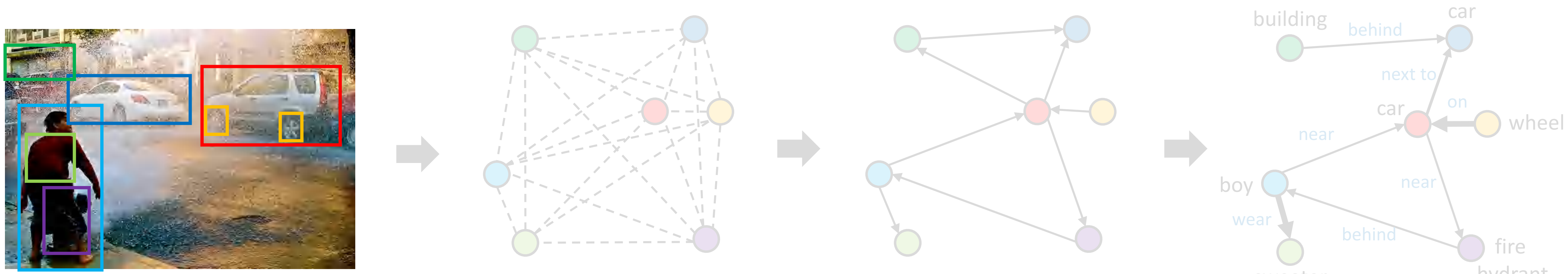
Jianwei Yang*, Jiasen Lu*, Stefan Lee, Dhruv Batra, Devi Parikh. Graph R-CNN for Scene Graph Generation. ECCV 2018.

Our model: Graph R-CNN



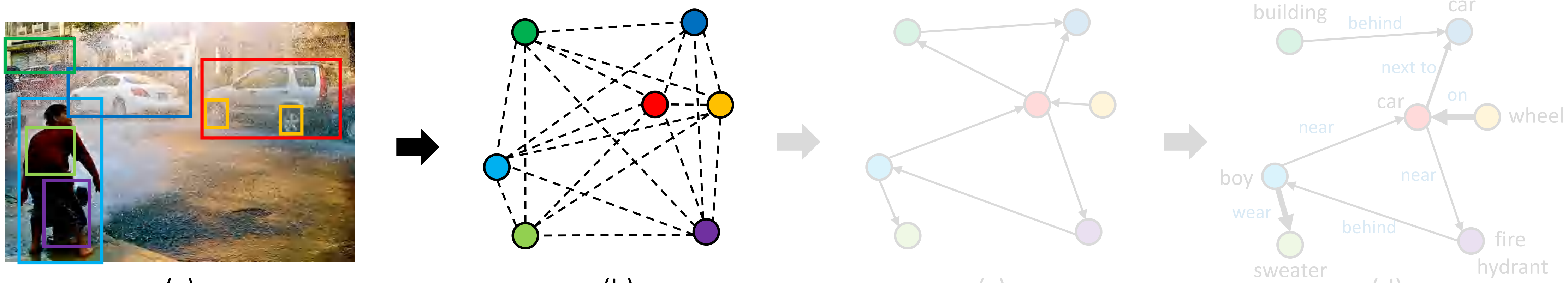
Jianwei Yang*, Jiasen Lu*, Stefan Lee, Dhruv Batra, Devi Parikh. Graph R-CNN for Scene Graph Generation. ECCV 2018.

Motivations



(a) (b) (c) (d)

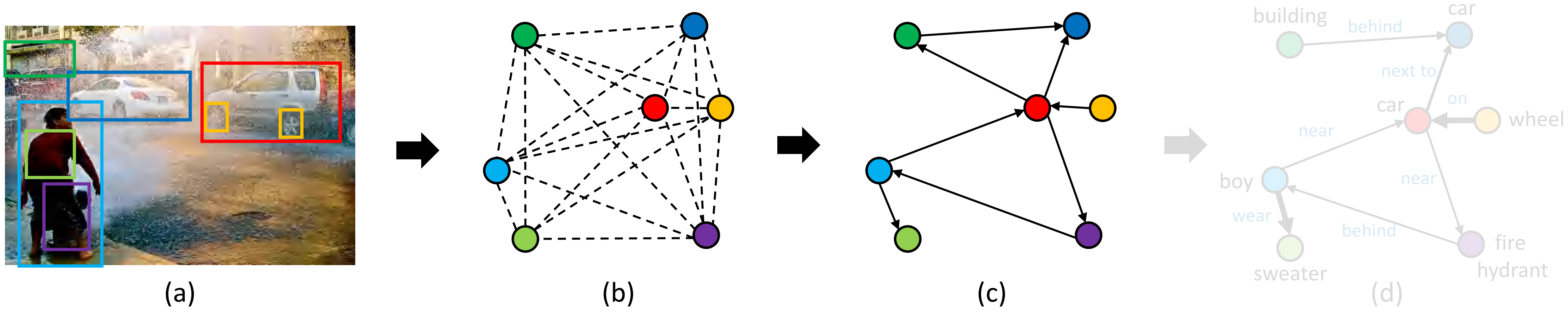
Motivations



(a) (b) (c) (d)

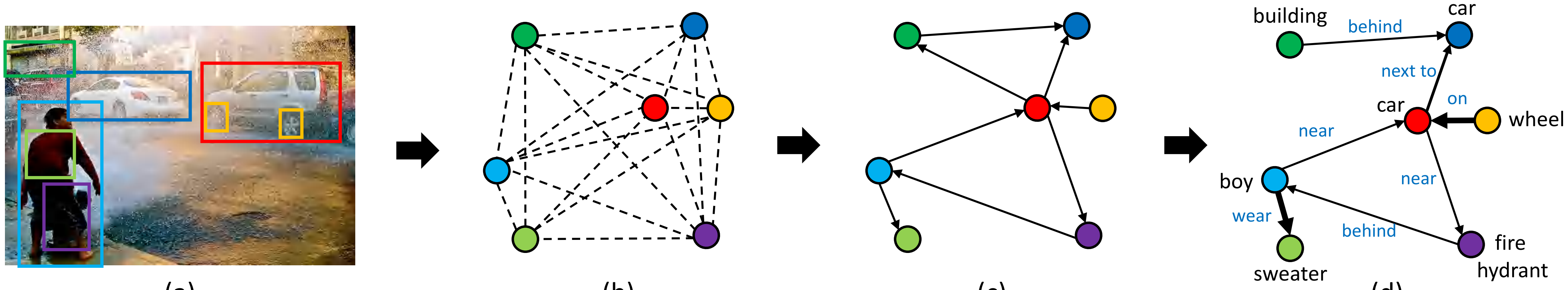
1. Objects in a scene usually have relationships with others;

Motivations



1. Objects in a scene usually have relationships with others;
2. Not all object pairs have relationships, the scene graph is usually sparse,

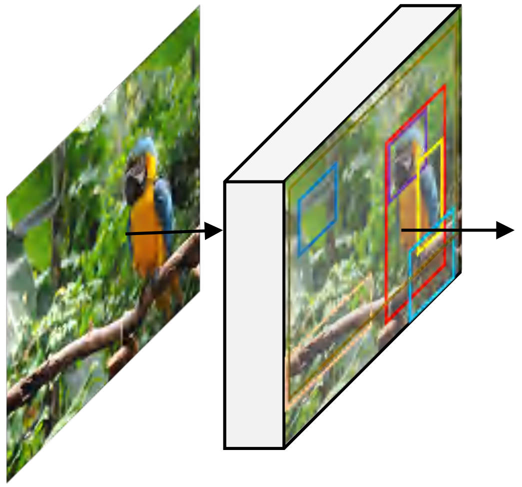
Motivations



(a) (b) (c) (d)

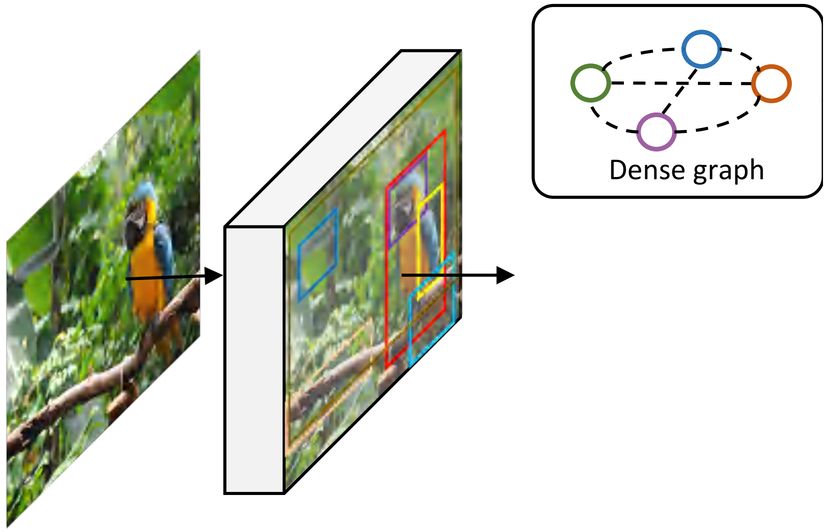
1. Objects in a scene usually have relationships with others;
2. Not all object pairs have relationships, the scene graph is usually sparse,
3. Existence of relationships highly depends on the object categories, and type of relationships highly depends on the context.

Framework



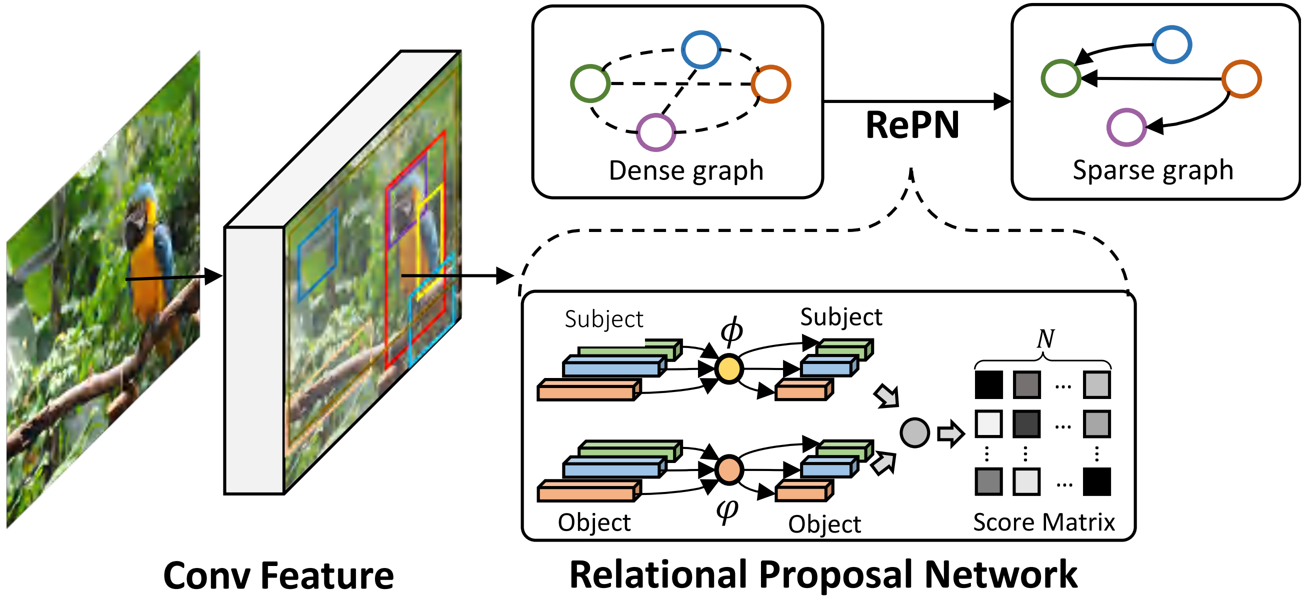
Conv Feature

Framework



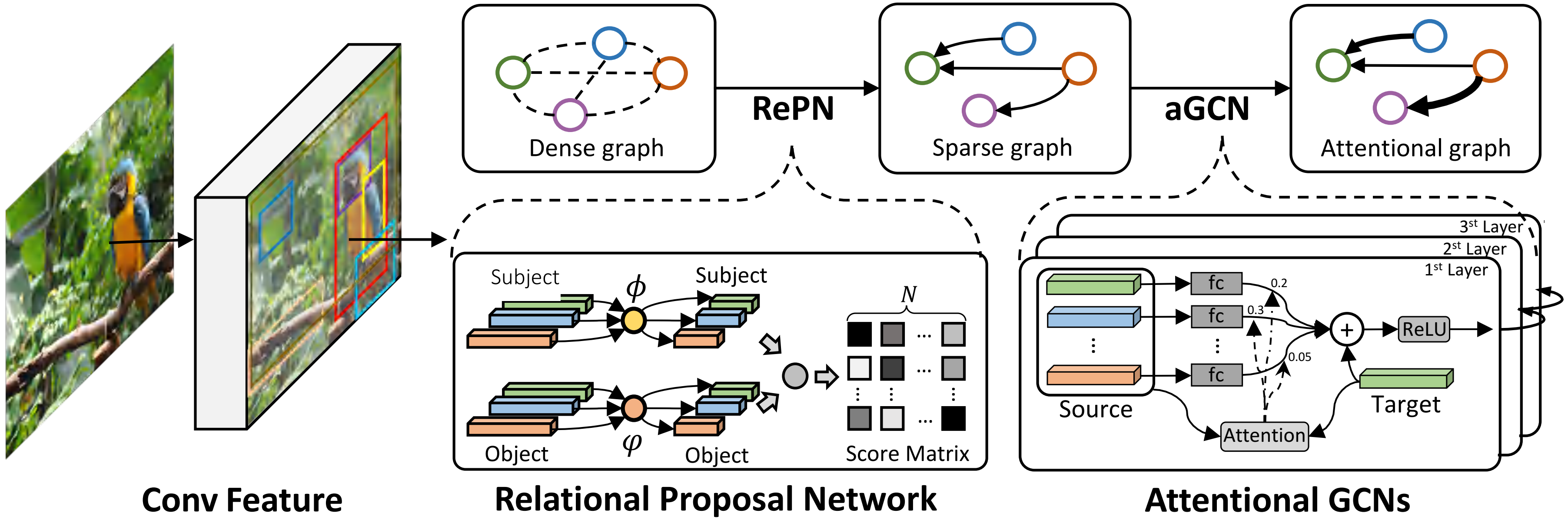
Conv Feature

Framework



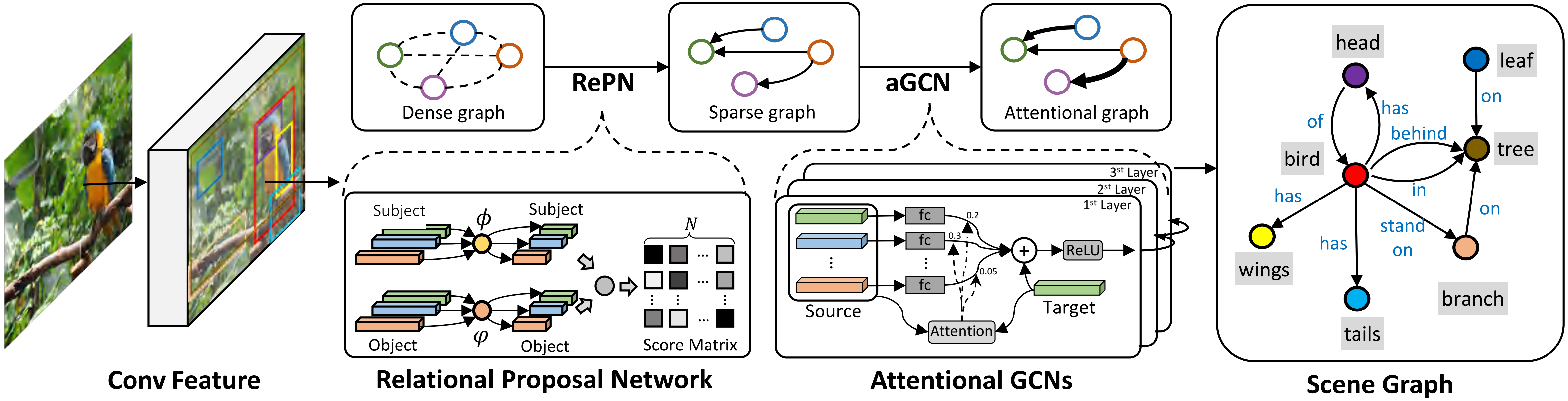
1. Relation proposal network (RePN) to learn to prune the densely connected scene graph;

Framework



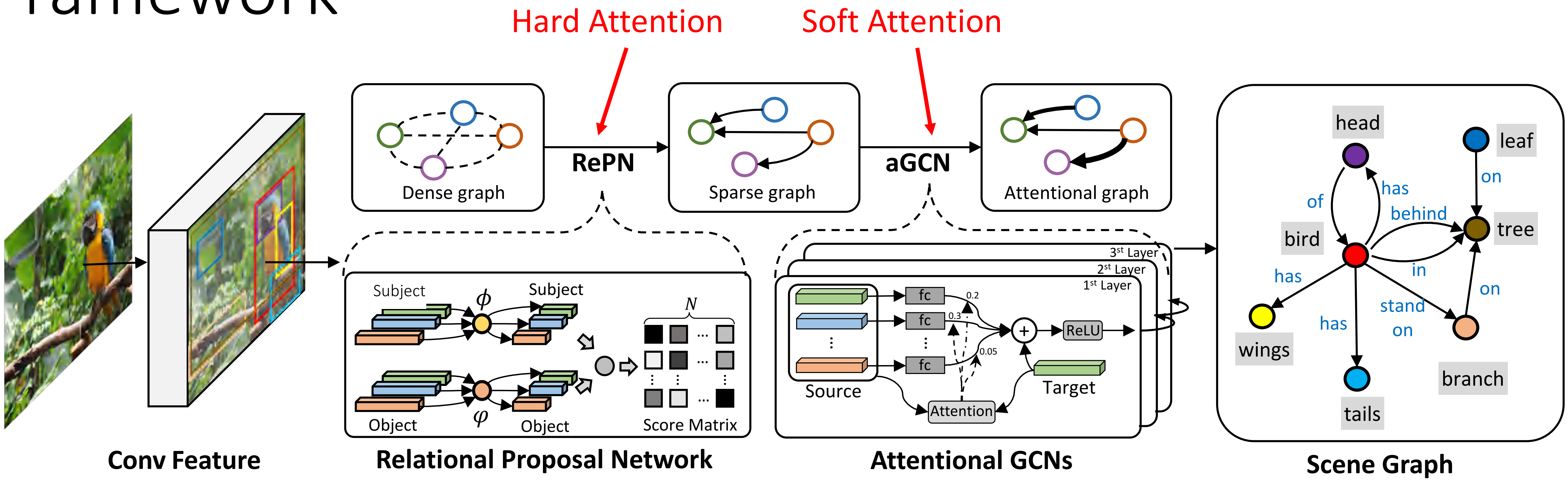
1. Relation proposal network (RePN) to learn to prune the densely connected scene graph;
2. Attentional graph convolutional networks (aGCN) to incorporate the contextual information.

Framework



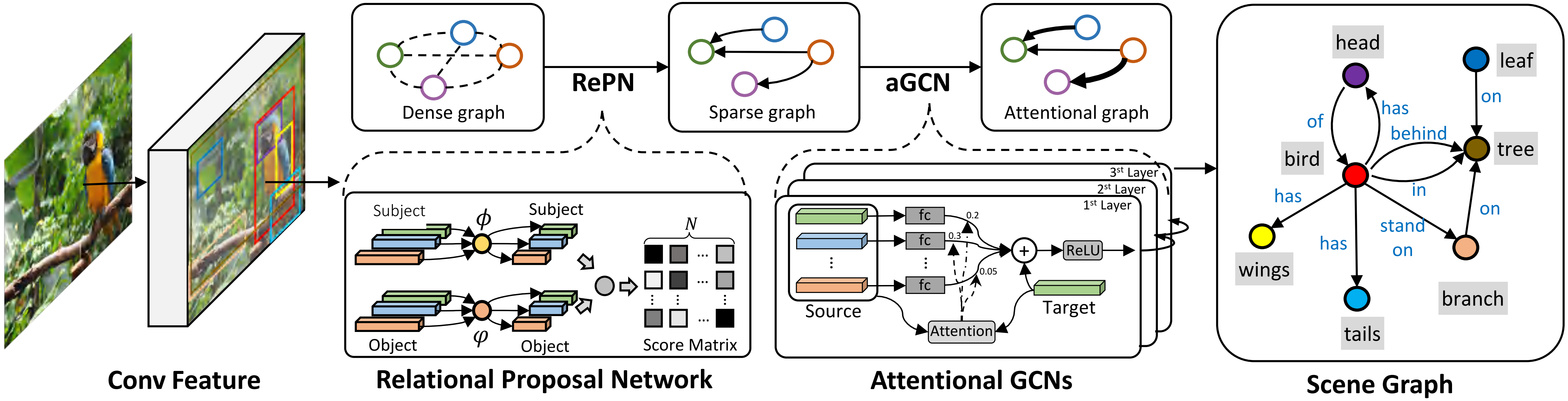
1. Relation proposal network (RePN) to learn to prune the densely connected scene graph,
2. Attentional graph convolutional networks (aGCN) to incorporate the contextual information.

Framework



1. Relation proposal network (RePN) to learn to prune the densely connected scene graph,
2. Attentional graph convolutional networks (aGCN) to incorporate the contextual information.

Framework



I – Input Image; S: Scene graph

V –Scene graph vertices (object)

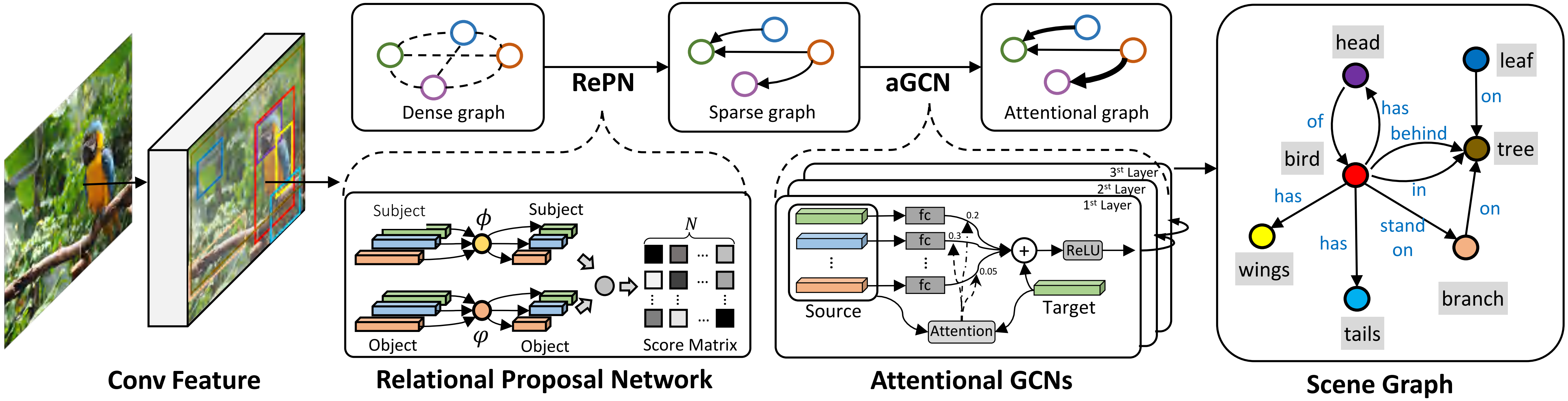
E – Scene graph edges (relationship)

O – Scene graph object labels

R – Scene graph relationship labels

$$\overbrace{P(V|I)}^{\text{Region Proposal}}$$

Framework



I – Input Image; S: Scene graph

V –Scene graph vertices (object)

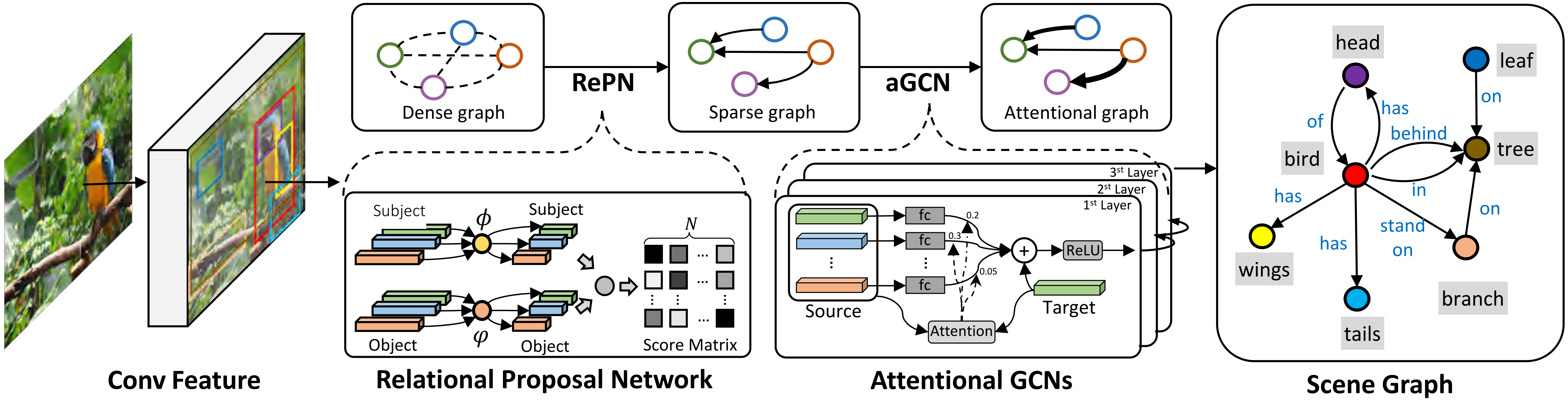
E – Scene graph edges (relationship)

O – Scene graph object labels

R – Scene graph relationship labels

$$\overbrace{P(V|I)}^{\text{Region Proposal}} \quad \underbrace{P(E|V,I)}_{\text{Relation Proposal}}$$

Framework

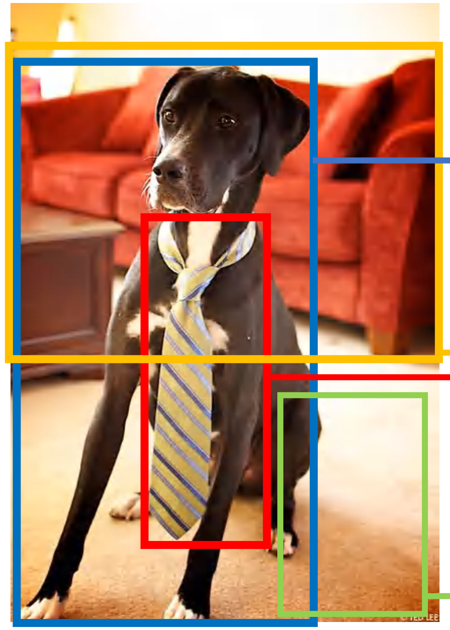


I – Input Image; S: Scene graph
V –Scene graph vertices (object)
E – Scene graph edges (relationship)
O – Scene graph object labels
R – Scene graph relationship labels

$$\overbrace{P(V|I)}^{\text{Region Proposal}} \; \underbrace{P(E|V,I)}_{\text{Relation Proposal}} \; \overbrace{P(R,O|V,E,I)}^{\text{Graph Labeling}} = P(S|I)$$

Relation Proposal Network

Inspired by Region Proposal Network[1]:



Step 1: Compute Relationship-ness between subject and object:

Subj. and obj. rep.   Kernel functions[2]

$$R(m,n) = f([x_m^o, x_n^o]) =< \phi(x_m^o), \varphi(x_n^o) >$$

Here, we use object prediction scores as the representation.

$$R(p,q) = f\left(\left[x_p^o, x_q^o\right]\right) =< \phi\left(x_p^o\right), \varphi(x_q^o) >$$

Step 2: NMS for object pairs based on pair-wise IoU:

$$IoU\left(\{r_m^o, r_n^o\}, \{r_p^o, r_q^o\}\right) = \frac{I\left(r_m^o, r_p^o\right) + I(r_n^o, r_q^o)}{U\left(r_m^o, r_p^o\right) + U\left(r_n^o, r_q^o\right)}$$

[1]. Faster R-CNN. Ren et al. Neurips 2016.
[2]. Non-local Networks. Want et al. CVPR 2018.

Attentional GCN

GCN layer with residual connection[1]:

$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{i \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)}\right)$$

[1]. Semi-Supervised Classification with Graph Convolutional Networks. Kipf et al. ICLR 2017

Attentional GCN

GCN layer with residual connection[1]:

$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{i \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)}\right)$$

Matrix Computation $$z_i^{(l+1)} = \sigma\left(W Z^{(l)} \alpha_i\right)$$

Nonlinear function Learnable parameters Inputs from last layer

[1]. Semi-Supervised Classification with Graph Convolutional Networks. Kipf et al. ICLR 2017

Attentional GCN

GCN layer with residual connection[1]:

$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{i \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)}\right)$$

Matrix Computation $$z_i^{(l+1)} = \sigma\left(WZ^{(l)}\boxed{\alpha_i}\right)$$

$\boxed{\alpha_i}$ ← Predetermined Affinities

σ: Nonlinear function; W: Learnable parameters; $Z^{(l)}$: Inputs from last layer

[1]. Semi-Supervised Classification with Graph Convolutional Networks. Kipf et al. ICLR 2017

Attentional GCN

GCN layer with residual connection[1]:

$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{i \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)}\right)$$

Matrix Computation

$$z_i^{(l+1)} = \sigma\left(W Z^{(l)} \alpha_i\right)$$

Nonlinear function Learnable parameters Inputs from last layer

Learning the affinities!

$$u_{ij} = w_h^T \sigma\left(W_a \left[z_i^{(l)}, z_j^{(l)}\right]\right)$$

$$\alpha_i = \mathrm{softmax}(u_i)$$

[1]. Semi-Supervised Classification with Graph Convolutional Networks. Kipf et al. ICLR 2017
[2]. Graph Attention Networks. Veličković et al. ICLR 2018

Attentional GCN

GCN layer with residual connection[1]:

$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{i \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)}\right)$$

Matrix Computation $$z_i^{(l+1)} = \sigma(W Z^{(l)} \alpha_i)$$

Nonlinear function Learnable parameters Inputs from last layer

Learning the affinities!

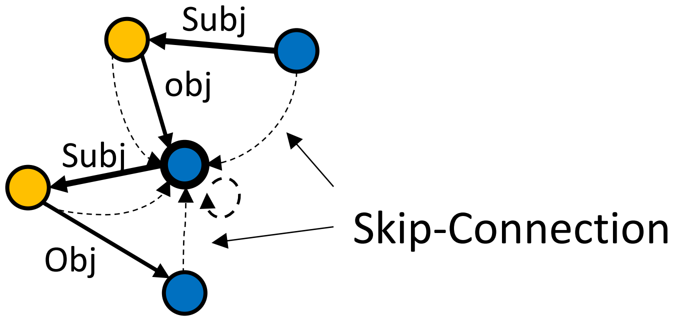
$$u_{ij} = w_h^T \sigma\left(W_a \left[z_i^{(l)}, z_j^{(l)}\right]\right)$$

$$\alpha_i = \mathrm{softmax}(u_i)$$

Attentional GCNs (aGCN) on scene graph:

Update object representations:

$$z_i^o = \sigma\left(W^{\mathrm{skip}} Z^o \alpha^{rs} + W^{sr} Z^r \alpha^{sr} + W^{or} Z^r \alpha^{or}\right)$$



[1]. Semi-Supervised Classification with Graph Convolutional Networks. Kipf et al. ICLR 2017
[2]. Graph Attention Networks. Veličković et al. ICLR 2018

Attentional GCN

GCN layer with residual connection:

$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{i \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)}\right)$$

Matrix Computation $$z_i^{(l+1)} = \sigma(W Z^{(l)} \alpha_i)$$

Nonlinear function Learnable parameters Inputs from last layer

Learning the affinities!

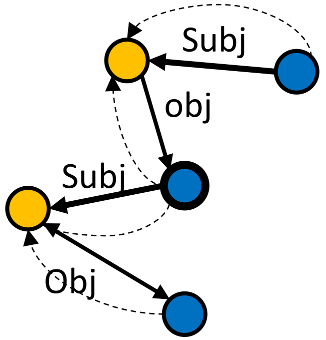
$$u_{ij} = w_h^T \sigma\left(W_a \left[z_i^{(l)}, z_j^{(l)}\right]\right)$$

$$\alpha_i = \text{softmax}(u_i)$$

Attentional GCNs (aGCN) on scene graph:

Update predicate representations:

$$z_i^r = \sigma(z_i^r + W^{rs} Z^o \alpha^{rs} + W^{ro} Z^o \alpha^{ro})$$



[1]. Semi-Supervised Classification with Graph Convolutional Networks. Kipf et al. ICLR 2017
[2]. Graph Attention Networks. Veličković et al. ICLR 2018

Training

$$P(V|I) \quad P(E|V,I) \quad P(R,O|V,E,I) = P(S|I)$$

Training

Binary Cross Entropy Loss

Region Proposal Network

$$P(V|I) \quad P(E|V,I) \quad P(R,O|V,E,I) = P(S|I)$$

Training

Binary Cross Entropy Loss

Region Proposal Network

$$P(V|I) \quad P(E|V,I) \quad P(R,O|V,E,I) = P(S|I)$$

Relation Proposal Network

Binary Cross Entropy Loss

Training

Two Cross Entropy Losses,
one for node and one for edge

Binary Cross Entropy Loss

Region Proposal Network

Graph Labeling Network

$$P(V|I) \quad P(E|V,I) \quad P(R,O|V,E,I) = P(S|I)$$

Relation Proposal Network

Binary Cross Entropy Loss

Metrics

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Metrics

Assume there are N objects extracted from an image, then $N * (N - 1)$ edges

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Metrics

Assume there are N objects extracted from an image, then $N * (N - 1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Metrics

Assume there are N objects extracted from an image, then $N * (N-1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

Step 2: Compute relationship scores: $Rel(i,j) = Subj(i) * Obj(j) * Pred(i,j)$

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Metrics

Assume there are N objects extracted from an image, then $N * (N-1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

Step 2: Compute relationship scores: $Rel(i,j) = Subj(i) * Obj(j) * Pred(i,j)$

Step 3: Sort the relationship triplets in a descending order:

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Metrics

Assume there are N objects extracted from an image, then $N * (N-1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

Step 2: Compute relationship scores: $Rel(i,j) = Subj(i) * Obj(j) * Pred(i,j)$

Step 3: Sort the relationship triplets in a descending order:

Step 4: Compute the triplet recalls (Recall@50, Recall@100) based on the ground-truth

SGGen: $Recall = \dfrac{C(T_{pred}\ and\ T_{gt})}{N(T_{gt})}$ IoU > 0.5

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Metrics

Assume there are N objects extracted from an image, then $N * (N-1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

Step 2: Compute relationship scores: $Rel(i,j) = Subj(i) * Obj(j) * Pred(i,j)$

Step 3: Sort the relationship triplets in a descending order:

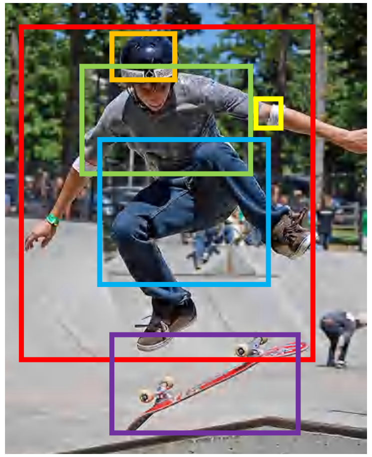
Step 4: Compute the triplet recalls (Recall@50, Recall@100) based on the ground-truth

SGGen: $Recall = \dfrac{C(T_{pred}\ and\ T_{gt})}{N(T_{gt})}$ IoU > 0.5

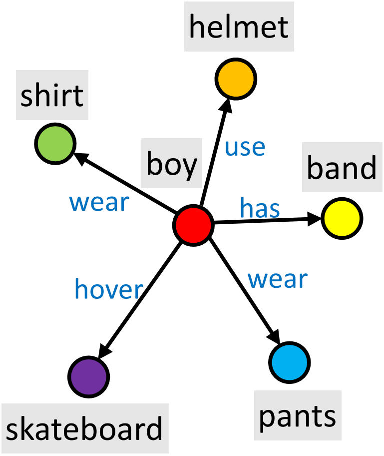
PhrCls: all object locations are known **PredCls:** all object locations and labels are known

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

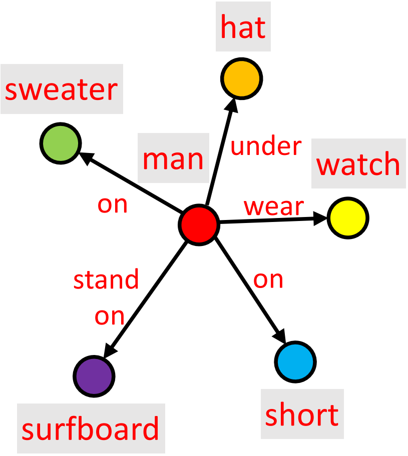
SGGen+: A new metric



Ground-Truth

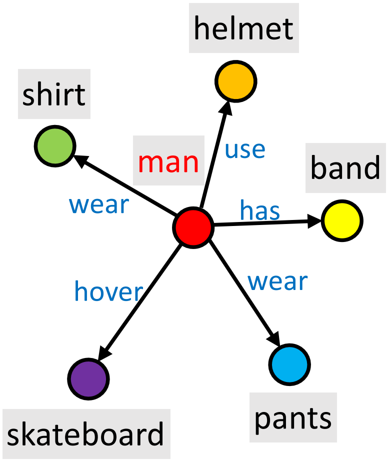


Prediction-1:



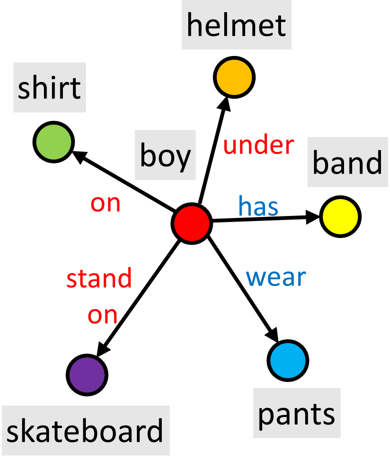
All predictions
are wrong

Prediction-2:



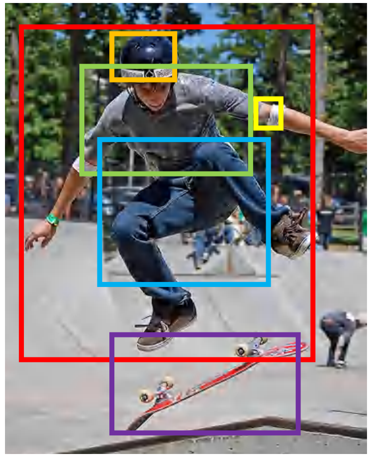
One node is
wrong

Prediction-3:

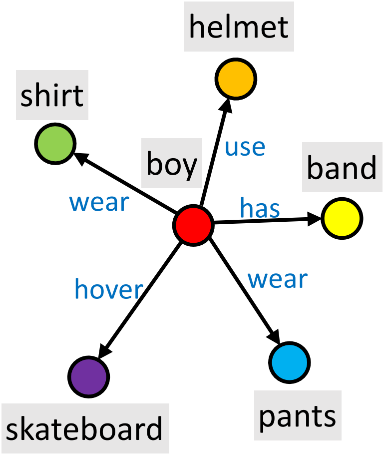


Three predicates
are wrong

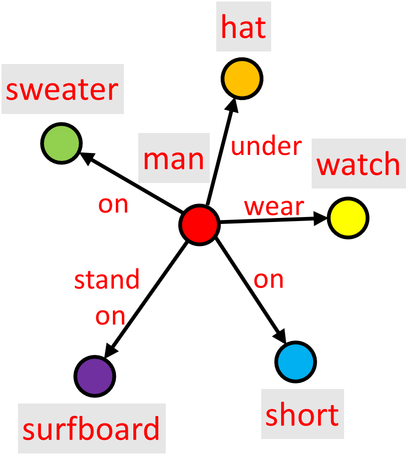
SGGen+: A new metric



Ground-Truth

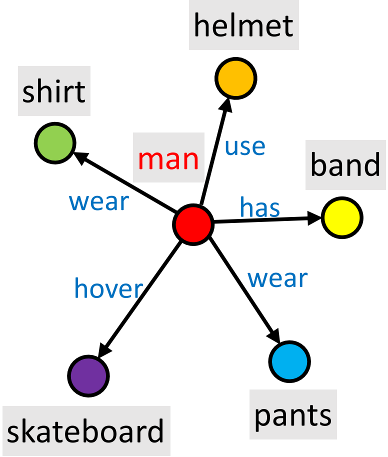


Prediction-1:



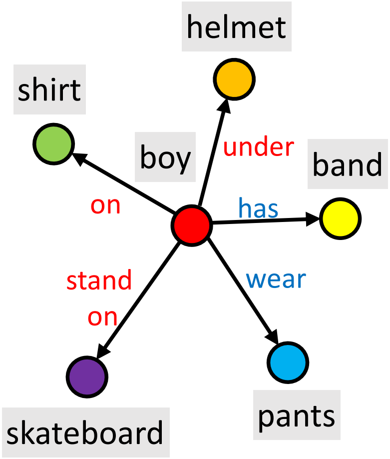
All predictions are wrong

Prediction-2:



One node is wrong

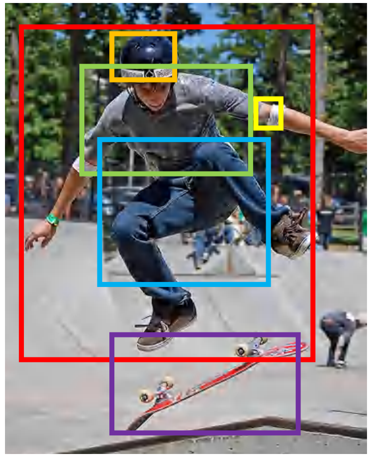
Prediction-3:



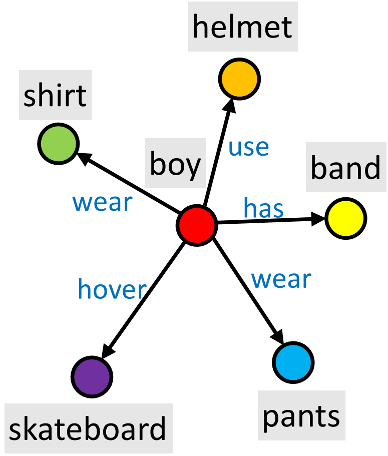
Three predicates are wrong

$$SGGen = \frac{C(T_{pred}\ and\ T_{gt})}{N(T_{gt})}$$

SGGen+: A new metric

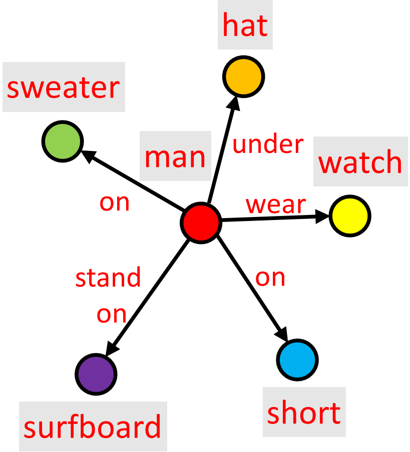


Ground-Truth



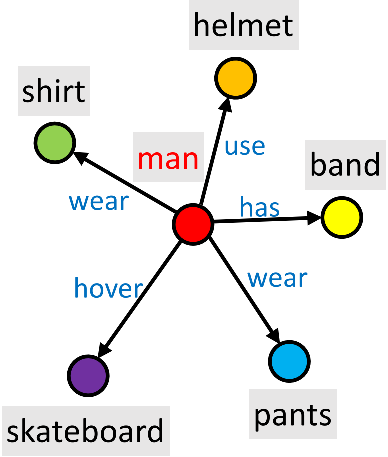
SGGen = 5

Prediction-1:



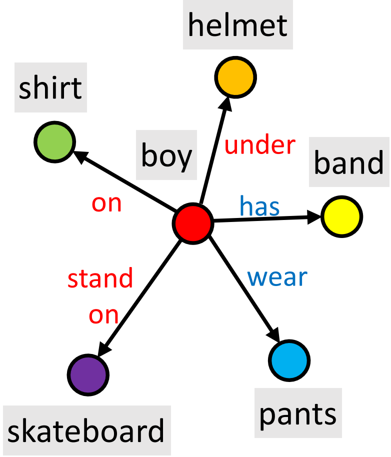
SGGen = 0

Prediction-2:



SGGen = 0

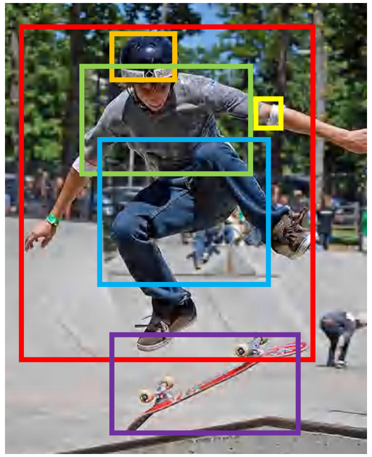
Prediction-3:



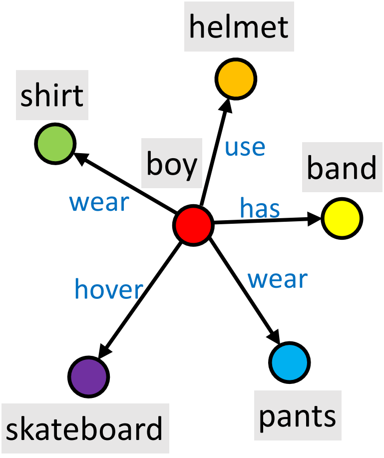
SGGen = 2

$$SGGen = \frac{C(T_{pred}\ and\ T_{gt})}{N(T_{gt})}$$

SGGen+: A new metric

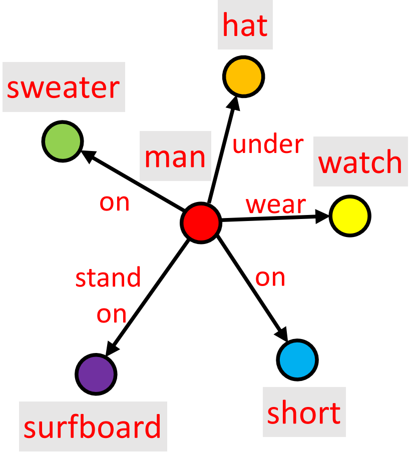


Ground-Truth



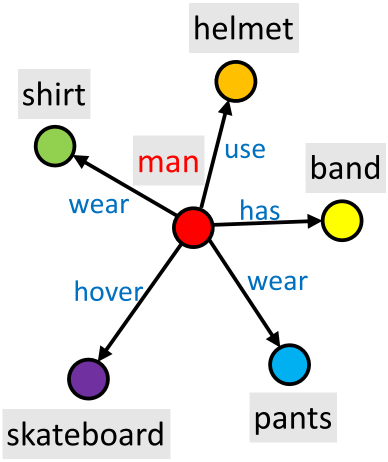
SGGen = 5

Prediction-1:



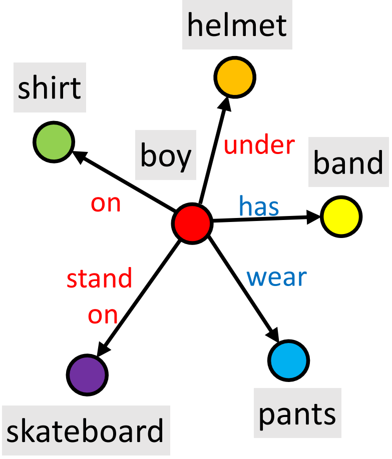
SGGen = 0

Prediction-2:



SGGen = 0

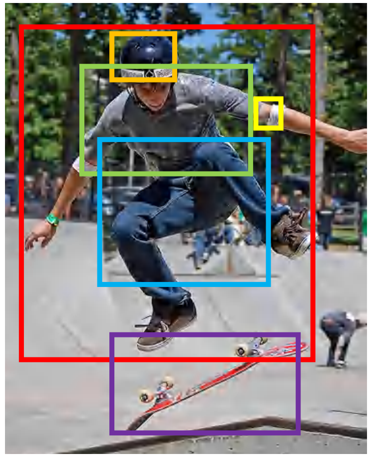
Prediction-3:



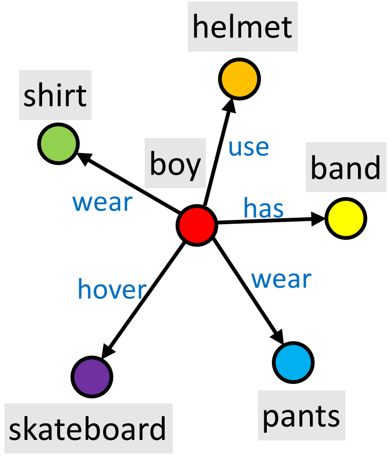
SGGen = 2

$$SGGen = \frac{C(T_{pred}\ and\ T_{gt})}{N(T_{gt})}$$

SGGen+: A new metric

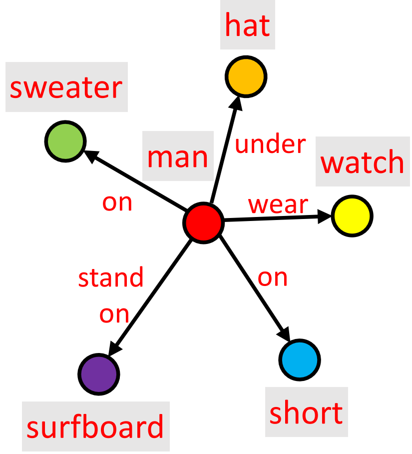


Ground-Truth



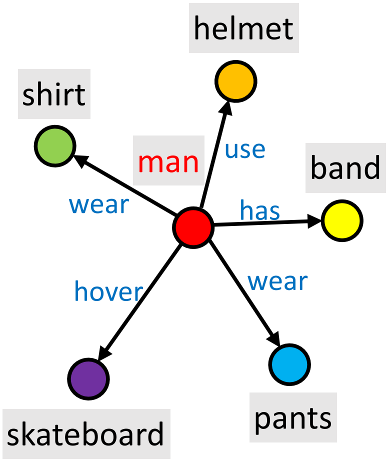
SGGen = 5

Prediction-1:



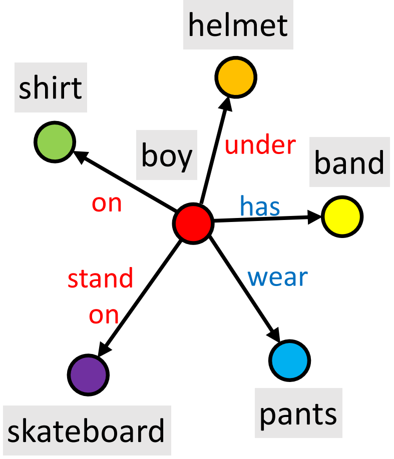
SGGen = 0

Prediction-2:



SGGen = 0

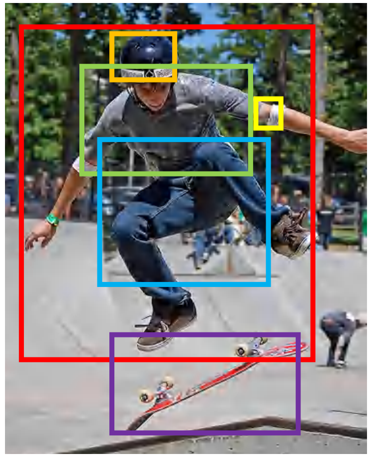
Prediction-3:



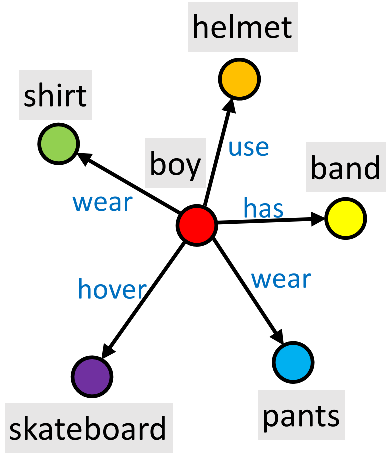
SGGen = 2

$$SGGen = \frac{C(T_{pred}\ and\ T_{gt})}{N(T_{gt})} \longrightarrow SGGen += \frac{C(O) + C(P) + C(T)}{N(O_{gt}) + N(P_{gt}) + N(T_{gt})}$$

SGGen+: A new metric



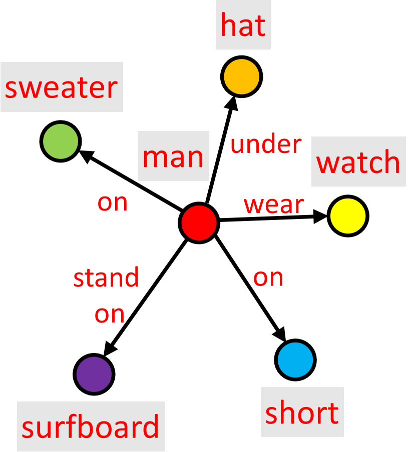
Ground-Truth



SGGen = 5

SGGen+ = 16

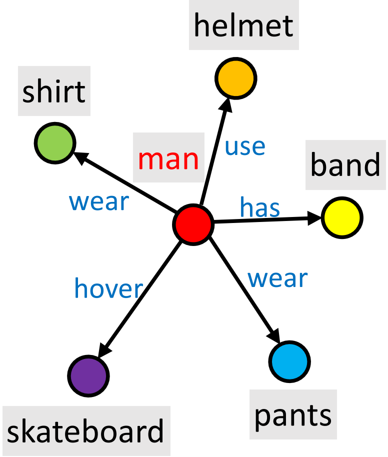
Prediction-1:



SGGen = 0

SGGen+ = 0

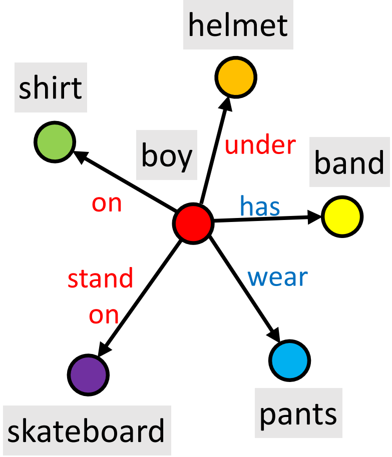
Prediction-2:



SGGen = 0

SGGen+ = 10

Prediction-3:



SGGen = 2

SGGen+ = 9

$$SGGen = \frac{C(T_{pred}\ and\ T_{gt})}{N(T_{gt})} \longrightarrow SGGen += \frac{C(O) + C(P) + C(T)}{N\left(O_{gt}\right) + N\left(P_{gt}\right) + N(T_{gt})}$$

Experiments

Table. Implementation Details.

Dataset	Backbone	#objects	#predicates	Metrics
Visual Genome Train: 75,651 Test: 32,422	VGG-16 Faster R-CNN[1]	150	50	PredCls,SGCls, SGGen,SGGen+, mAP

[1] A Faster Implementation of Faster R-CNN. Yang and Lu et al.

Comparing SGGen+ with SGGen

Perturbation: change the node labels in ground-truth scene graphs

Perturb on	Node w/o relationship			Nodes w/ relationship			Both		
Perturb ratio	20%	50%	100%	20%	50%	100%	20%	50%	100%
SGGen	100.0	100.0	100.0	54.1	22.1	0.0	62.2	24.2	0.0
SGGen+	94.5	89.1	76.8	84.3	69.6	47.9	80.1	56.6	22.8

1. *SGGen* is **completely insensitive** to the perturbation on objects w/o rel.

Comparing SGGen+ with SGGen

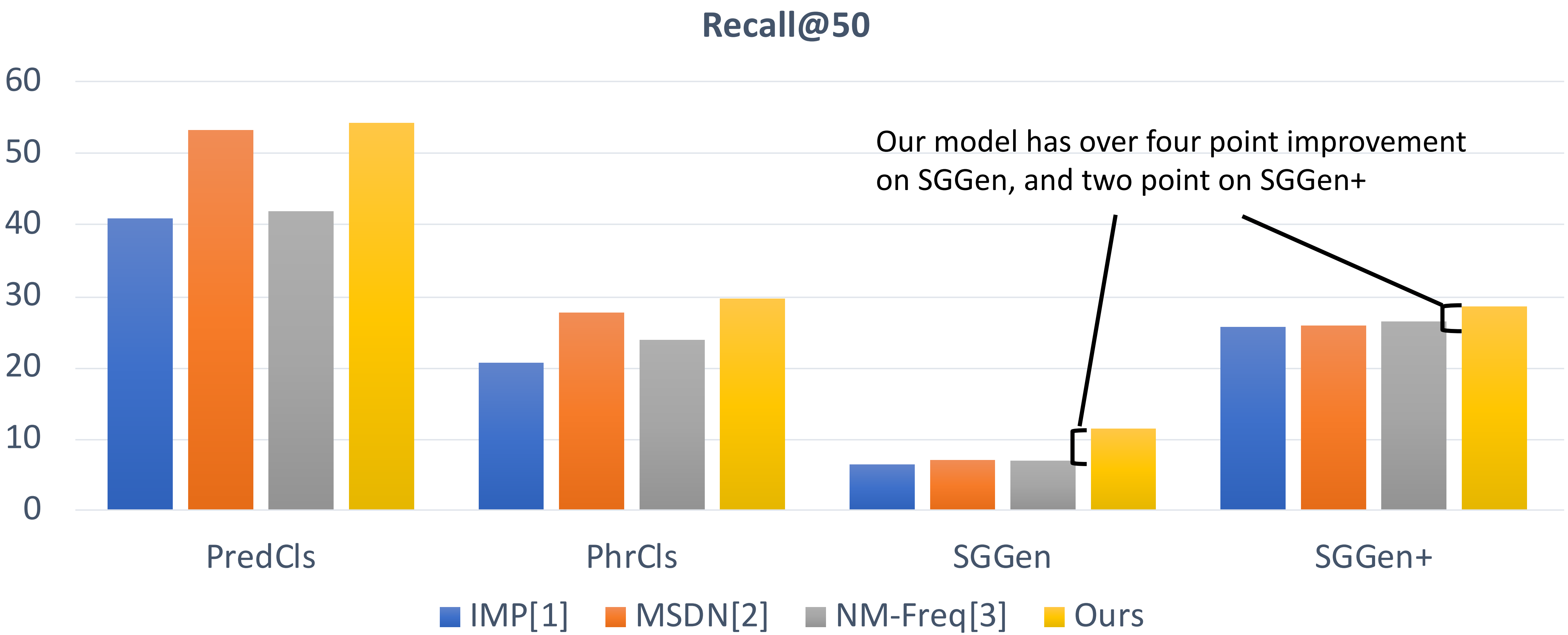
Perturbation: change the node labels in ground-truth scene graphs

Perturb on	Node w/o relationship			Nodes w/ relationship			Both		
Perturb ratio	20%	50%	100%	20%	50%	100%	20%	50%	100%
SGGen	100.0	100.0	100.0	54.1	22.1	0.0	62.2	24.2	0.0
SGGen+	94.5	89.1	76.8	84.3	69.6	47.9	80.1	56.6	22.8

1. *SGGen* is **completely insensitive** to the perturbation on objects w/o rel.

2. *SGGen* is **over sensitive** to perturbations on objects with rel.

Comparing with Previous Work

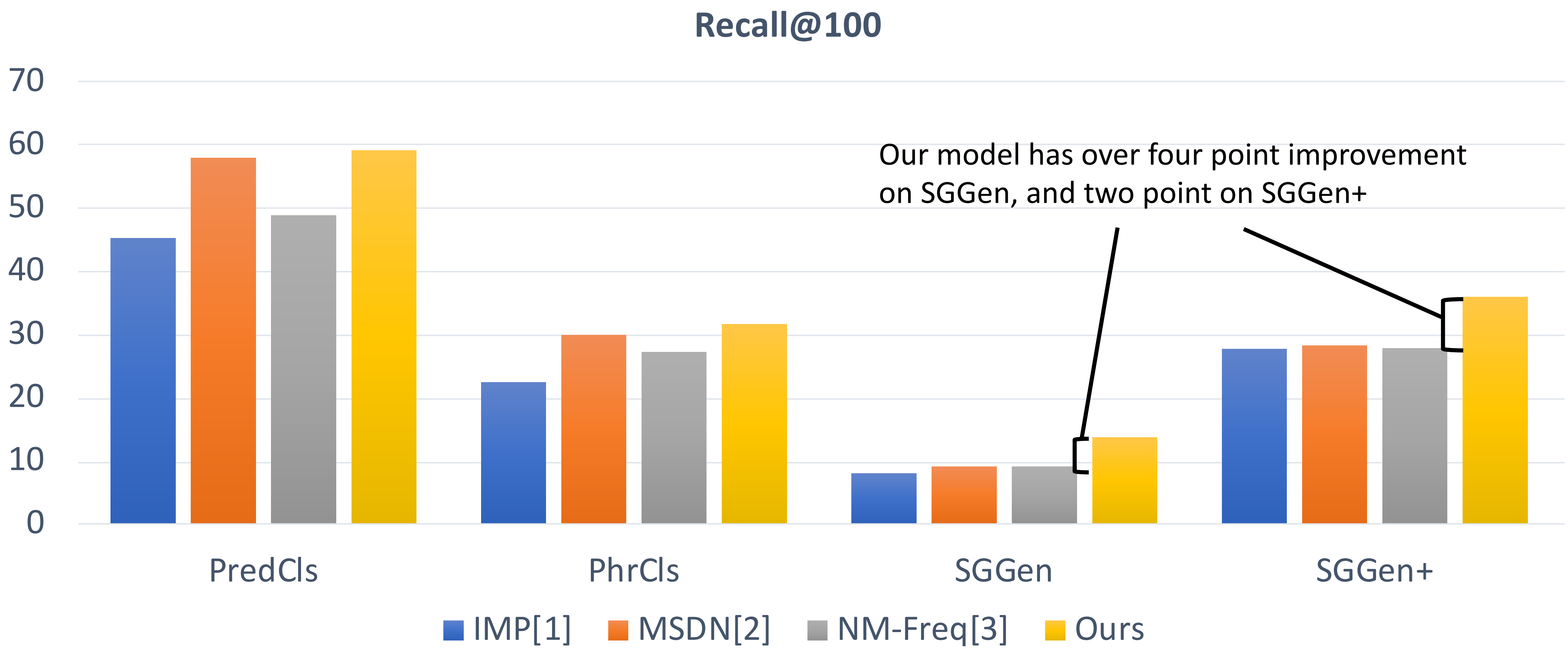


[1] Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

[2] Scene Graph Generations from Objects, Phrases and Captions. Li et al. ICCV 2017

[3] Neural Motif: Scene Graph Parsing with Global Context. Zellers et al. CVPR 2018

Comparing with Previous Work

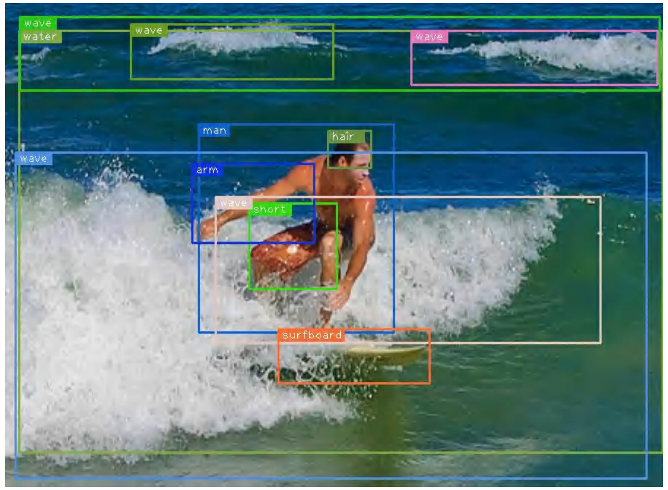


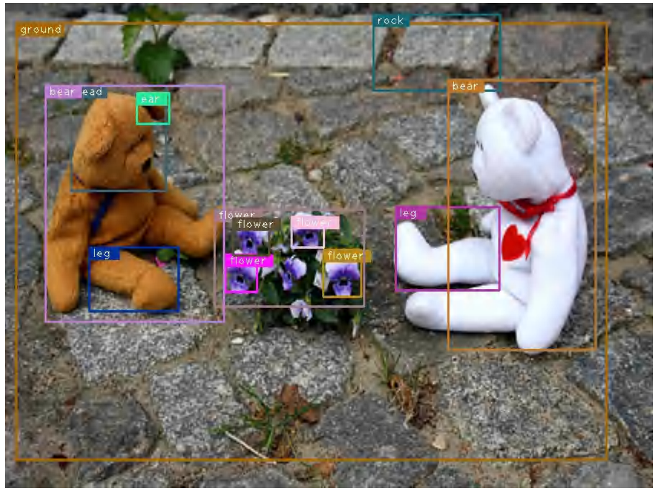
[1] Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

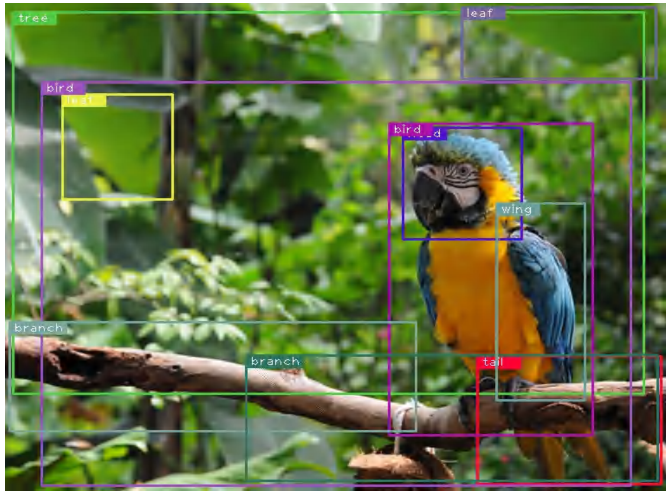
[2] Scene Graph Generations from Objects, Phrases and Captions. Li et al. ICCV 2017

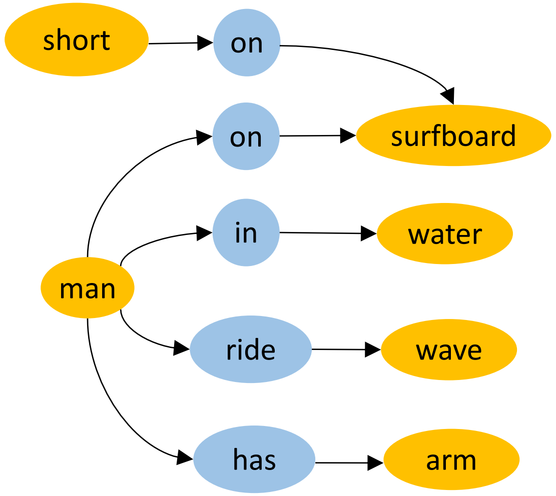
[3] Neural Motif: Scene Graph Parsing with Global Context. Zellers et al. CVPR 2018

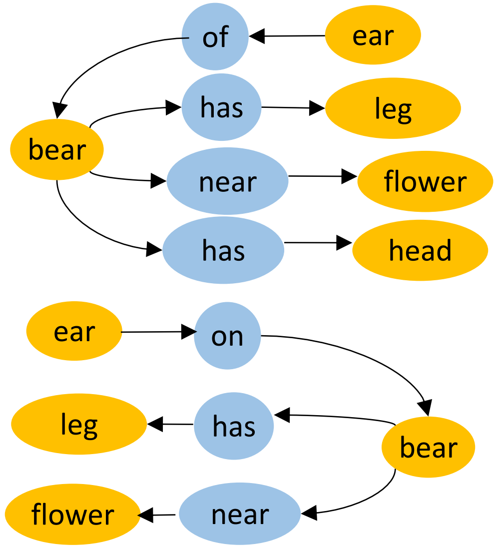
Qualitative Results

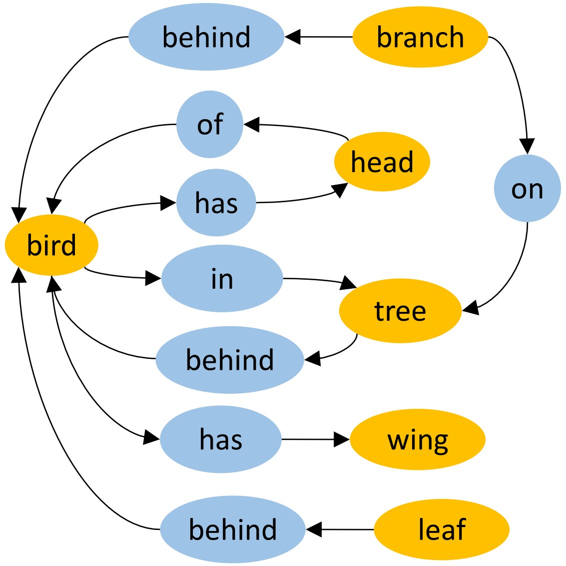












Common sense emerges

We extract the weights in the score-level aGCN layer, and sort it in descending order.

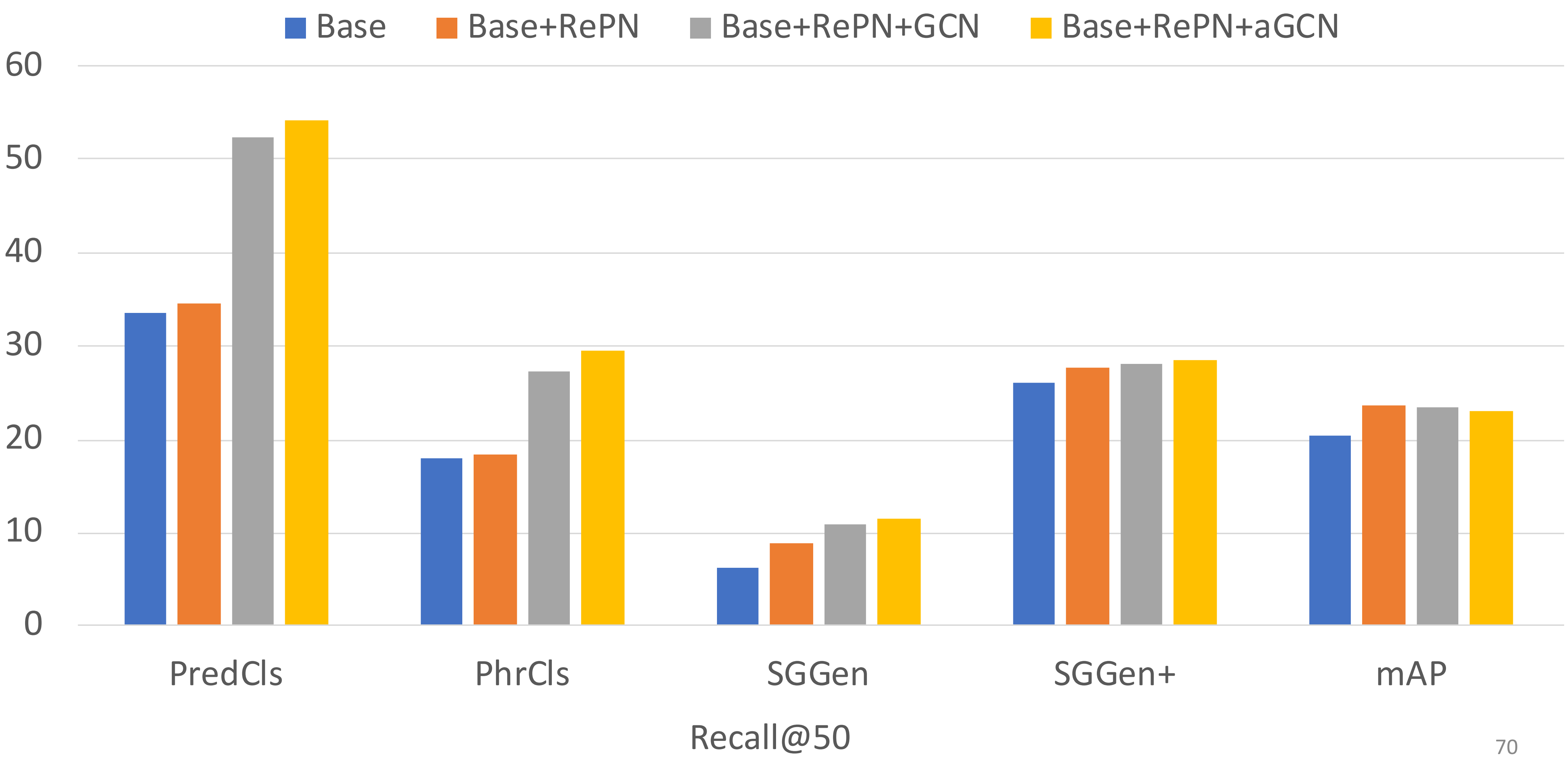
Object-Object Co-Occurrence

Object	Top-1	Top-2	Object	Top-1	Top-2
boat	water	beach	**girl**	woman	hair
plane	wing	tail	**cow**	horse	dog
clock	building	root	**sidewalk**	street	bus
bottle	cup	glass	**handle**	plate	food
bus	truck	vehicle	**snow**	pole	ski

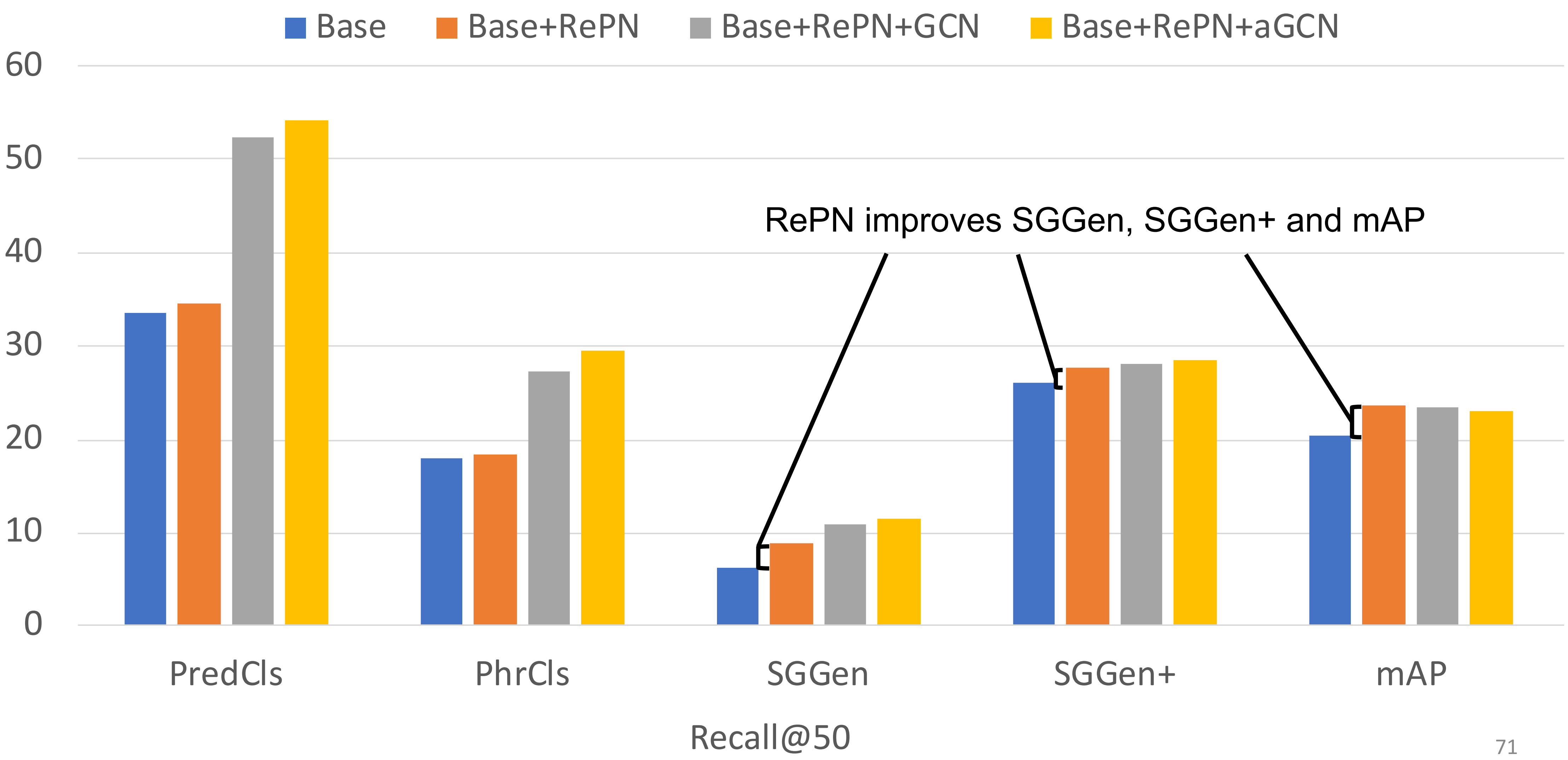
Object-Predicate Co-Occurrence

Object	Top-1	Top-2	Object	Top-1	Top-2
hat	hold	wear	**kite**	watch	look at
boat	in	sit in	**girl**	look at	watch
umbrella	carry	hold	**jacket**	wear	with
track	with	on	**stripe**	on	has
sidewalk	at	walk on	**snow**	on	near

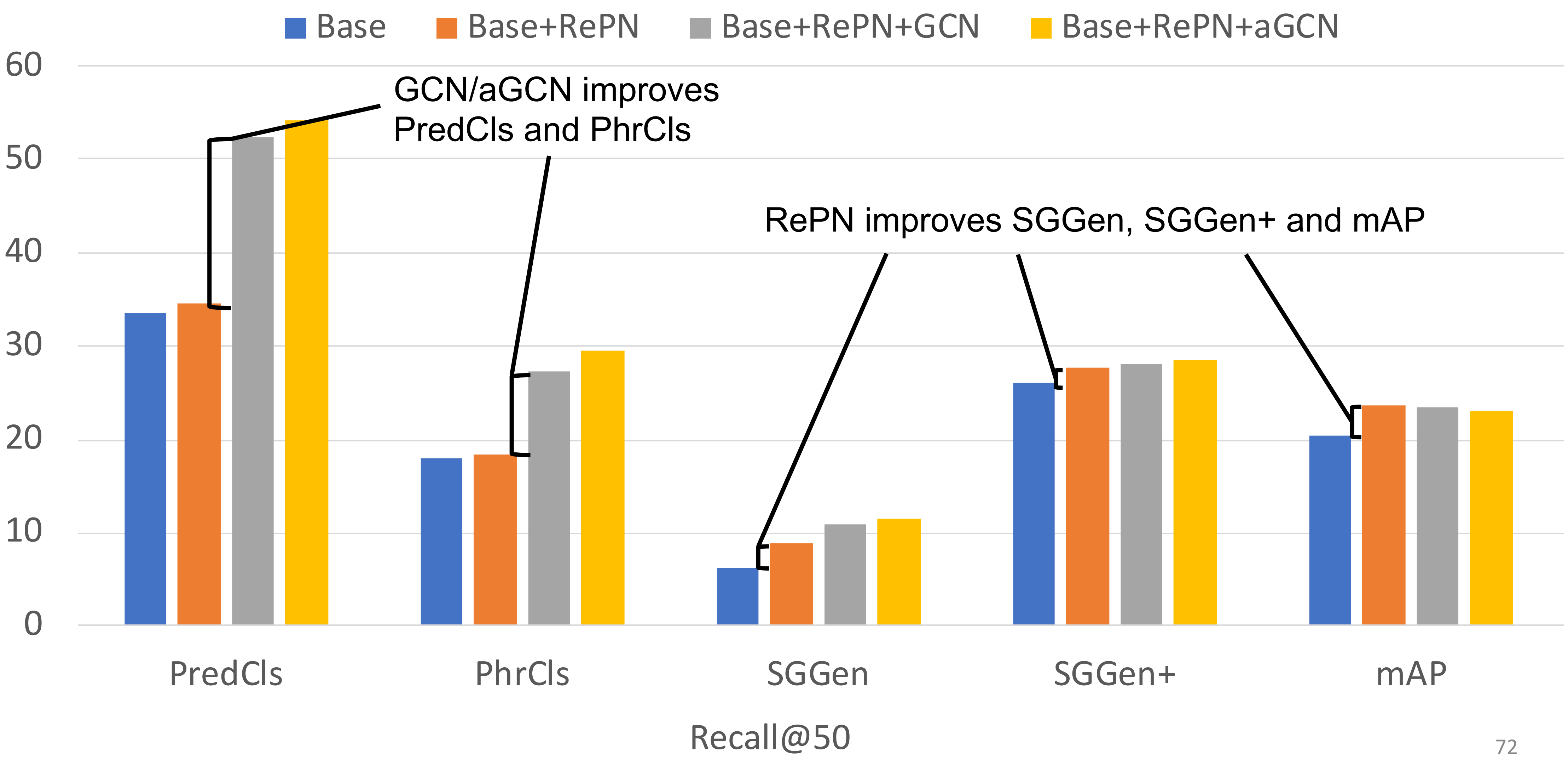
Ablation Study



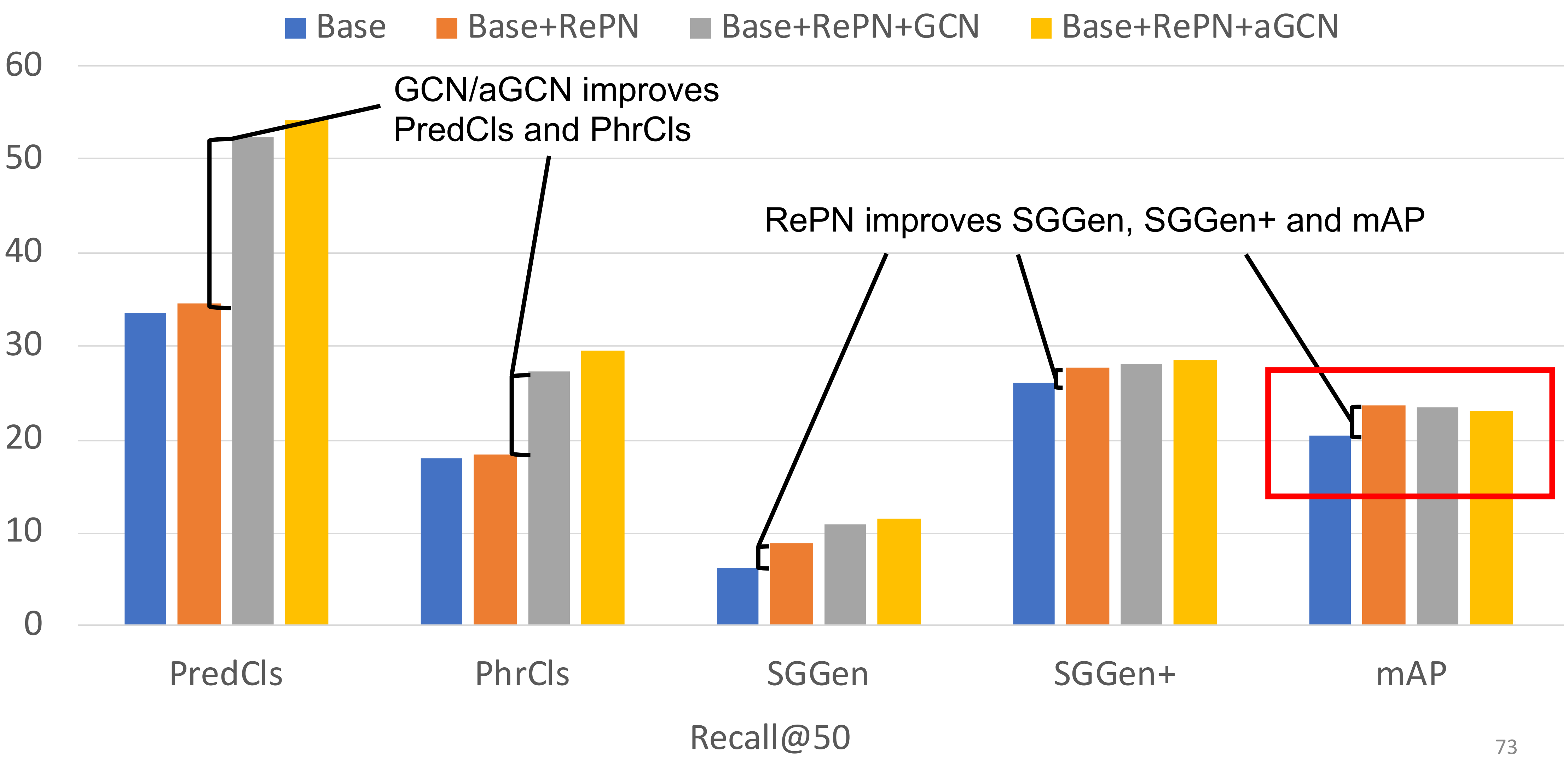
Ablation Study



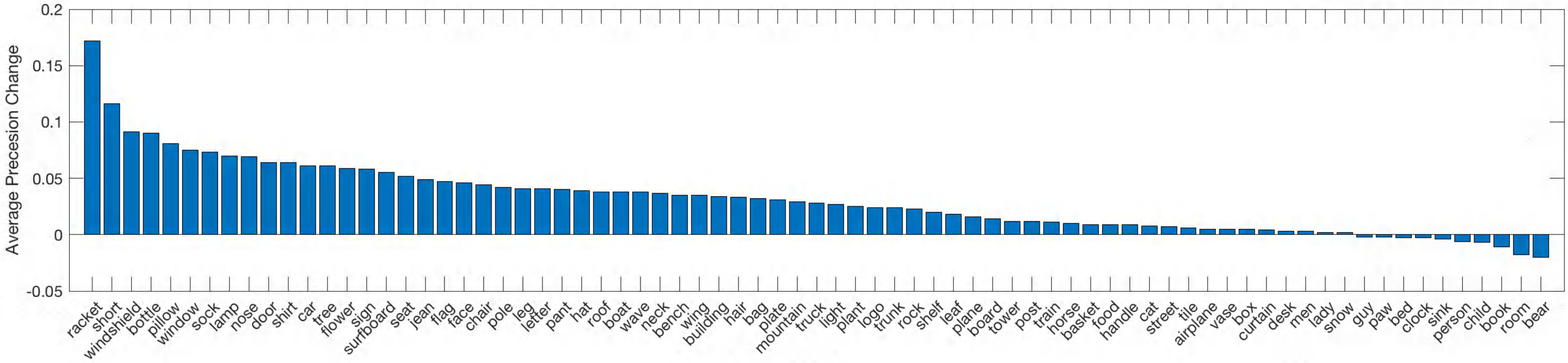
Ablation Study



Ablation Study

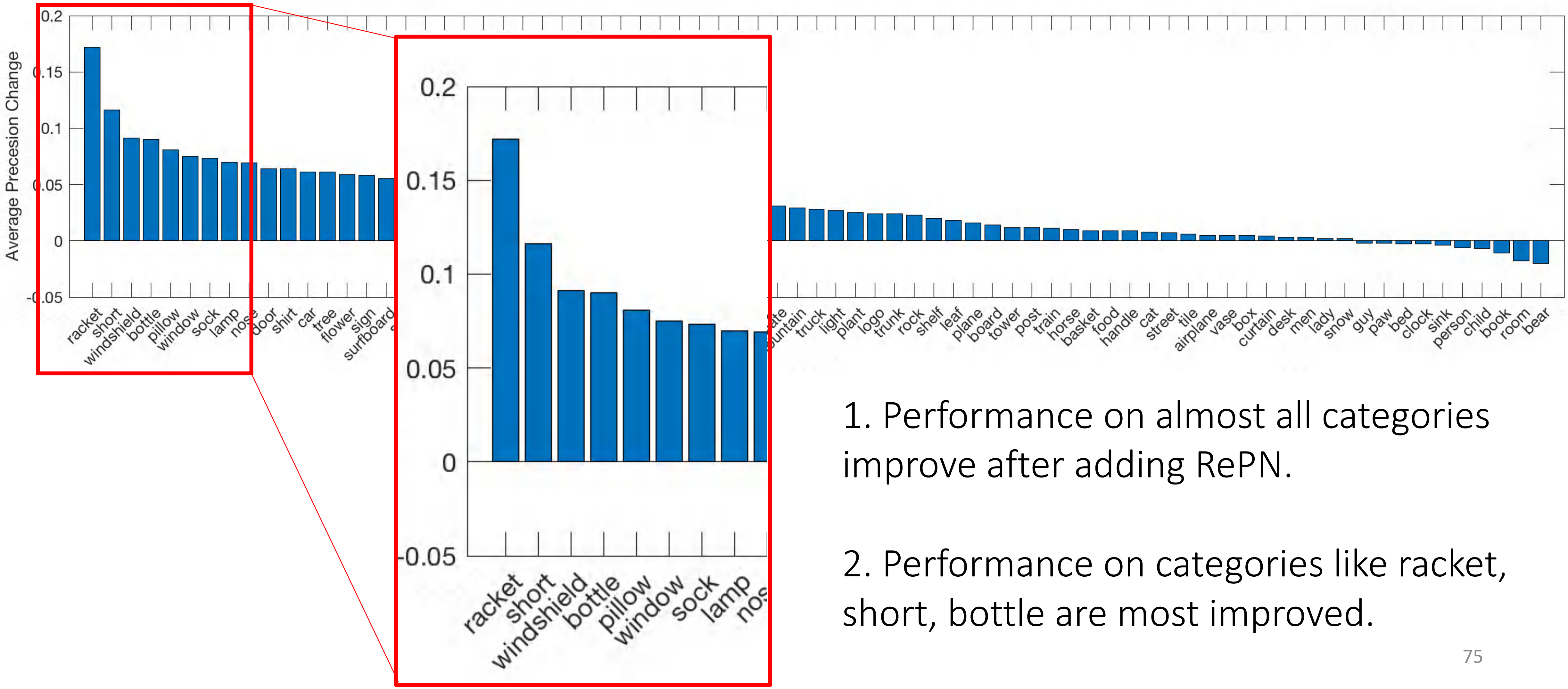


Object Detection Investigation



1. Performance on almost all categories improve after adding RePN.

Object Detection Investigation



1. Performance on almost all categories improve after adding RePN.

2. Performance on categories like racket, short, bottle are most improved.

Part I: Summary

- Take aways:
  - Introducing a general base model for scene graph generation
  - Pruning the fully-connected graph is important for scene graph generation
  - Exploiting the context across objects and predicates is crucial
  - Scene graph generation helps to improve object detection

- Challenges:
  - The dataset is **noisy** (incomplete and inconsistent annotations)
  - Relationships need more fine-grained categorizing (spatial, semantic, etc)
  - Rare/novel relationship is hard to detect

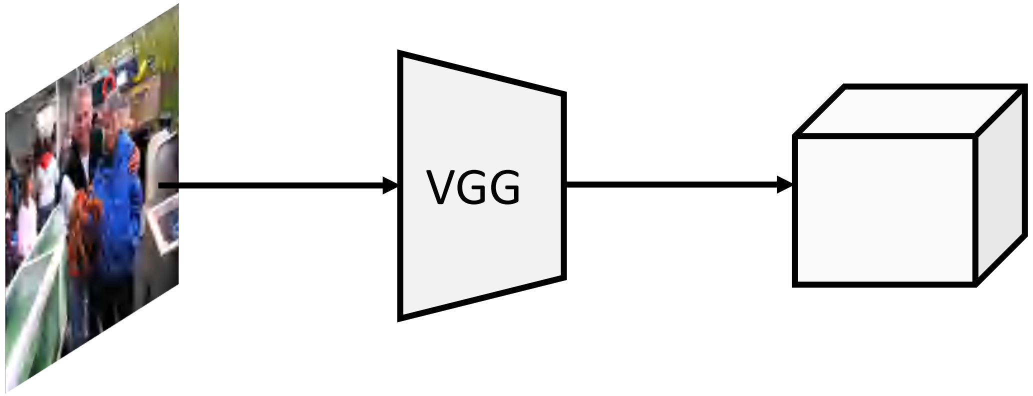
Part 2: Scene Graph for Vision-and-Language Tasks

How we can use scene graph?

How we can use scene graph?

Scene Graph as Feature Representation

Image Representations for Vision-and-Language Tasks



Performance on VQA 1.0

58.05%

Image Representations for Vision-and-Language Tasks

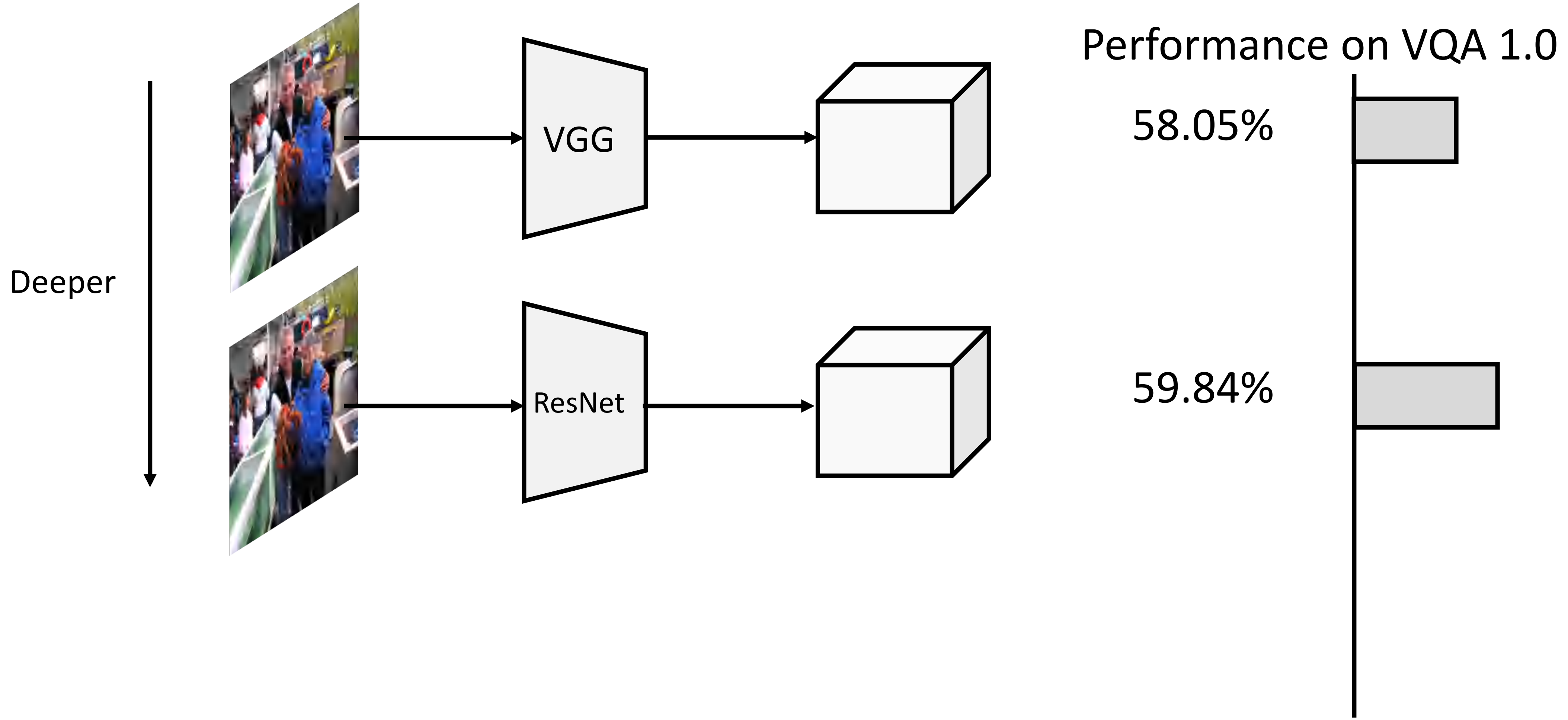
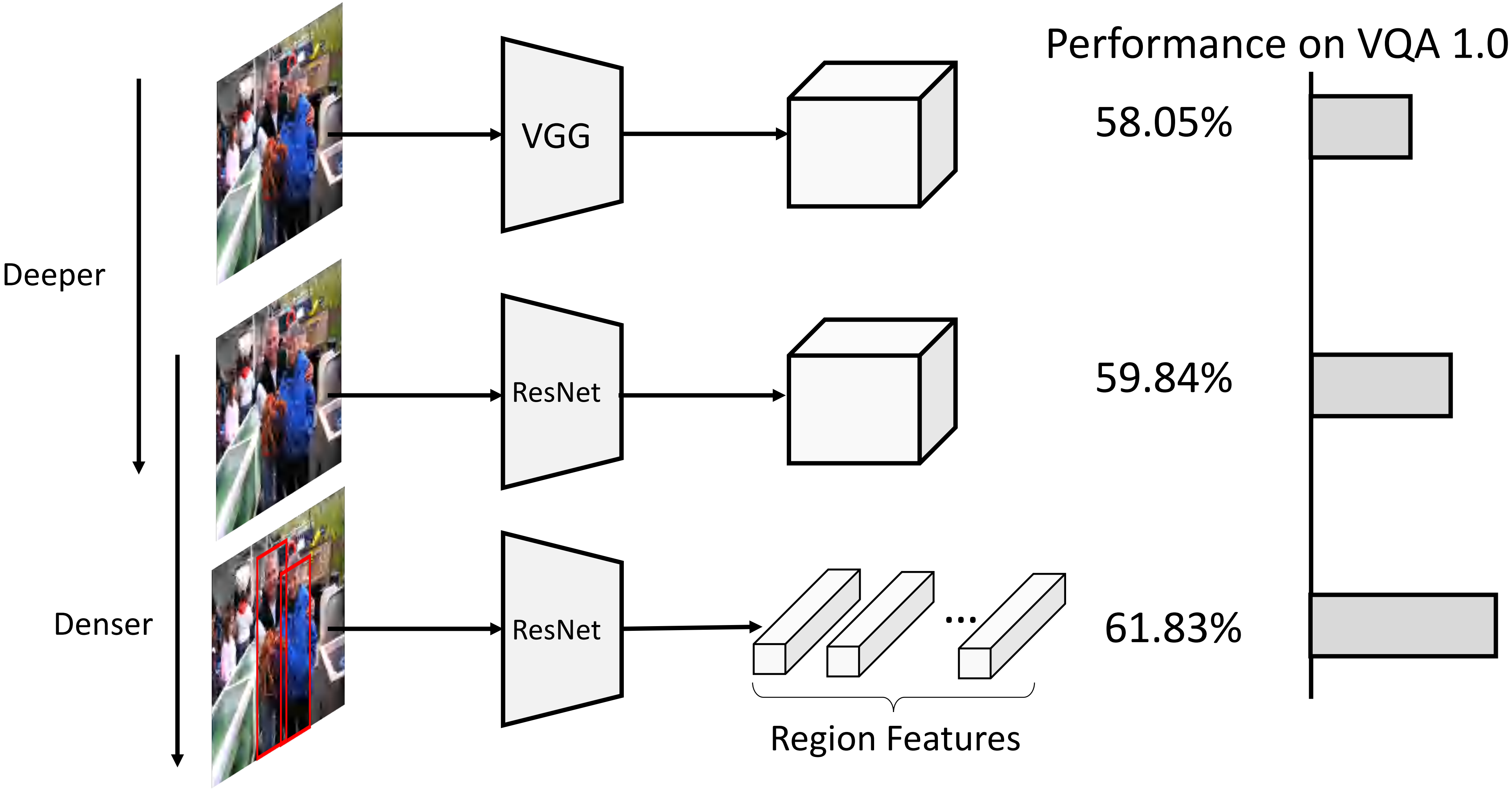


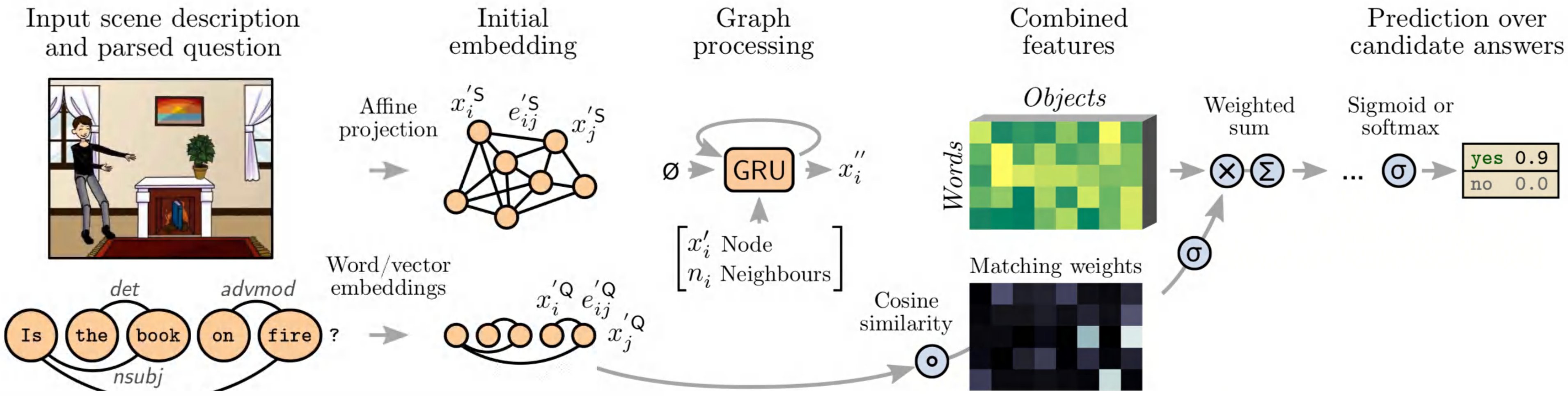
Image Representations for Vision-and-Language Tasks



[1] Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. Anderson et al. CVPR 2018

Visual Question Answering on Clipart

Graph-Structured Representations for Visual Question Answering. Teney et al. CVPR 2017



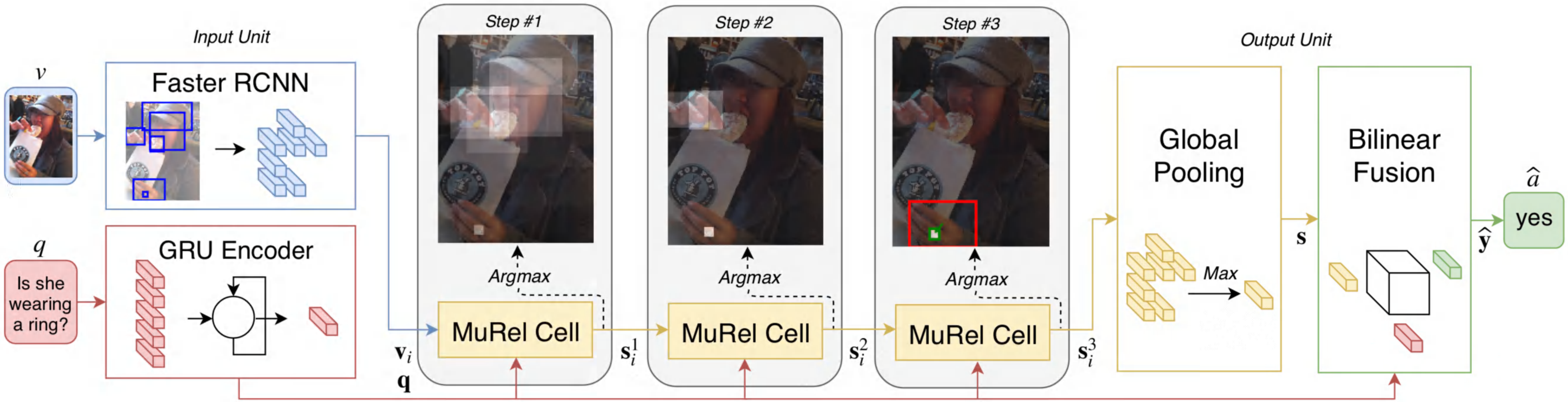
	Multiple choice				Open-ended			
Method	Overall	Yes/no	Other	Number	Overall	Yes/no	Other	Number
LSTM blind [4]	61.41	76.90	49.19	49.65	57.19	76.88	38.79	49.55
LSTM with global image features [4]	69.21	77.46	66.65	52.90	65.02	77.45	56.41	52.54
Zhang *et al.* [30] (yes/no only)	35.25	79.14	—	—	35.25	79.14	—	—
Multimodal residual learning [13]	67.99	79.08	61.99	52.57	62.56	79.10	48.90	51.60
U. Tokyo MIL (ensemble) [22, 1]	71.18	79.59	67.93	56.19	69.73	80.70	**62.08**	58.82
Graph VQA (full model)	**74.37**	**79.74**	**68.31**	**74.97**	**70.42**	**81.26**	56.28	**76.47**

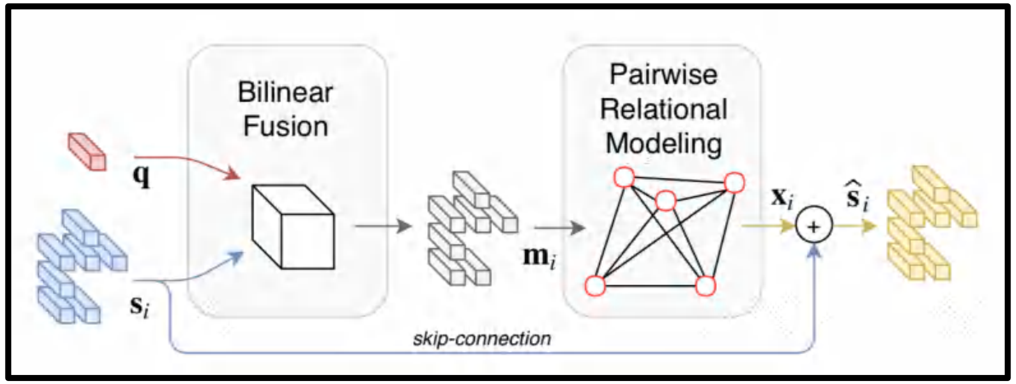
Table 2. Results on the test set of the "abstract scenes" dataset (average scores in percents).

VQA: Visual Question Answering. Antol et al. ICCV 2015.

Visual Question Answering on Realistic Data

MUREL: Multimodal Relational Reasoning for Visual Question Answering. Cadene et al. CVPR 2019





MuRel Cell

Model	*test-dev* Yes/No	Num.	Other	All	*test-std* All
Bottom-up [3]	81.82	44.21	56.05	65.32	65.67
Counter [41]	83.14	**51.62**	**58.97**	**68.09**	**68.41**
MuRel	**84.77**	49.84	57.85	68.03	**68.41**

Table 3. **State-of-the-art comparison on the VQA 2.0 dataset.**

Compositional Reasoning VQA Dataset

GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. Hudson et al. CVPR 2019

Pattern: What|Which <type> [do you think] <is> <dobject>, <attr> or <decoy>?

Program: **Select:** <dobject> ⟶ **Choose** <type>: <attr>|<decoy>

Reference: The food on the red object left of the small girl that is holding a hamburger

Decoy: brown

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger ⟶ **Relate:** girl,holding ⟶ **Filter** size: small ⟶ **Relate:** object, left ⟶ **Filter** color: red ⟶ **Relate:** food,on ⟶ **Choose** color: yellow | brown

Graph Normalization
- Ontology construction
- Edge Pruning
- Object Augmentation
- Global Properties

Question Generation
- Pattern Collection
- Compositional References
- Decoy Selection
- Probabilistic Generation

Sampling and Balancing
- Distribution Balancing
- Type-Based Sampling
- Deduplication

Entailment Relations
- Functional Programs
- Entailment Relations
- Recursive Reachability

New Metrics
- Consistency
- Validity & Plausibility
- Distribution
- Grounding

Compositional Reasoning VQA Dataset

GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. Hudson et al. CVPR 2019

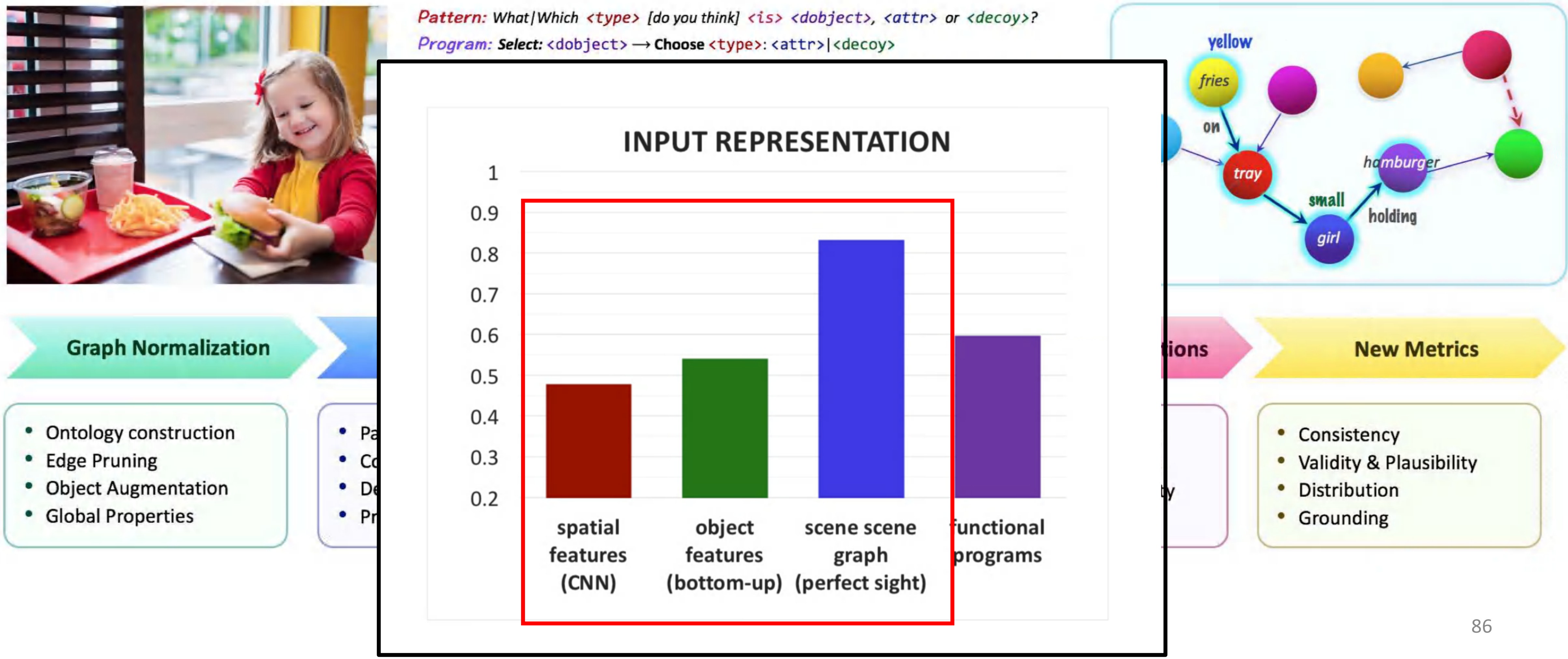
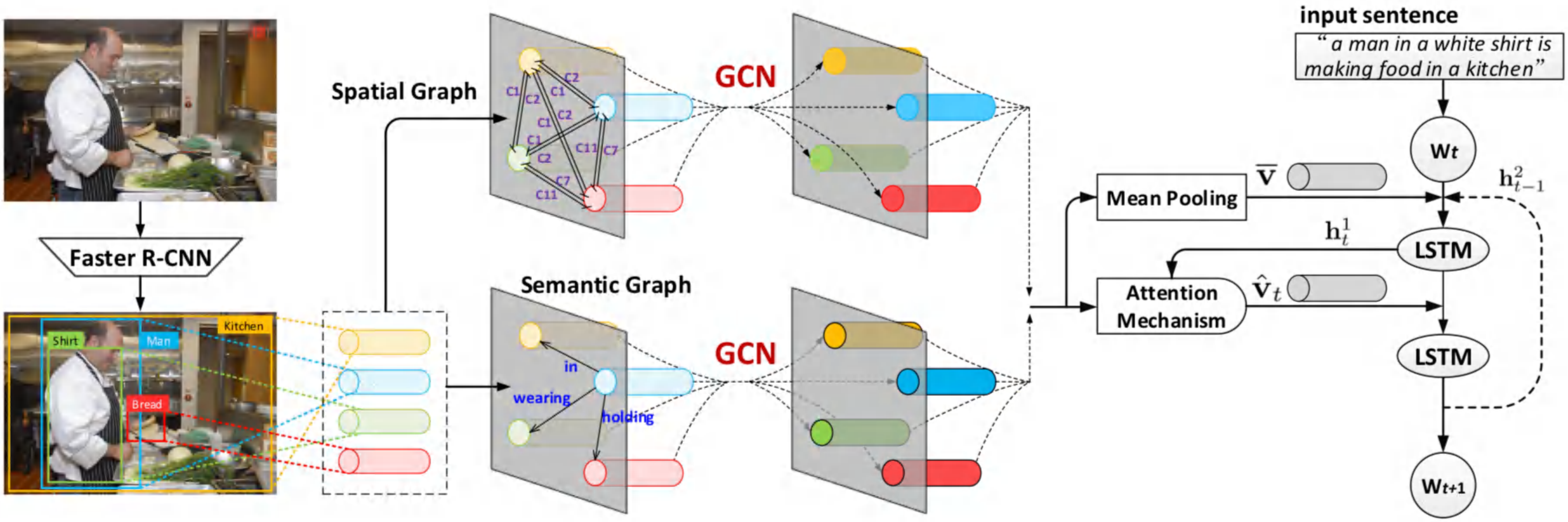


Image Captioning given Relationship

Exploring Visual Relationship for Image Captioning. Yao et al. ECCV 2018



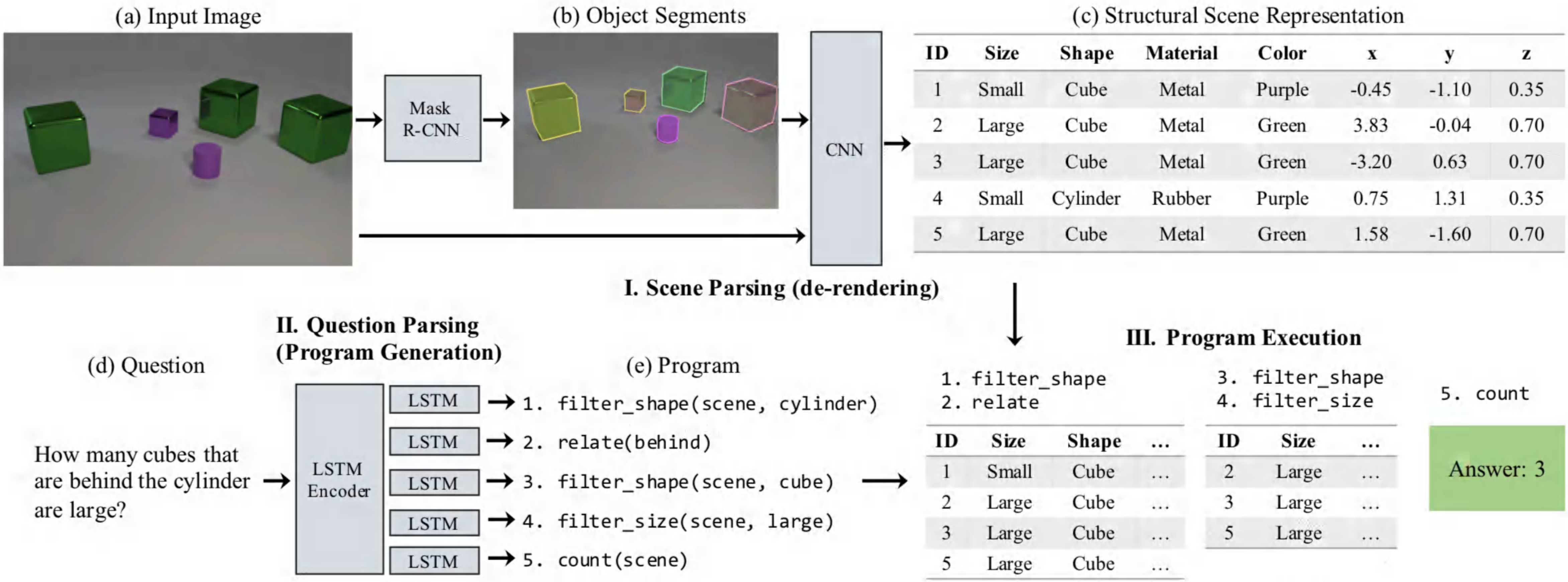
[1] Microsoft coco: Common objects in context. Lin et al. ECCV 2014

How we can use scene graph?

Scene Graph as Symbolic Representation

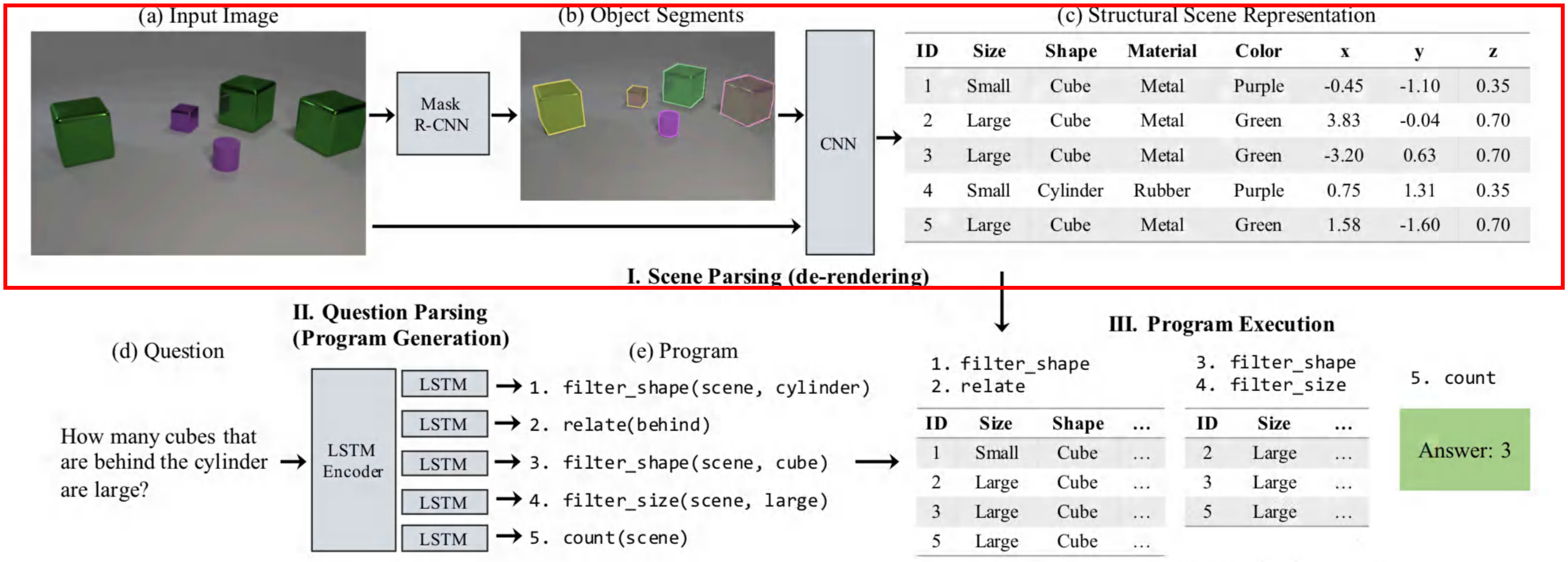
Neural-Symbolic VQA

Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. Yi et al. NeurIPS 2018



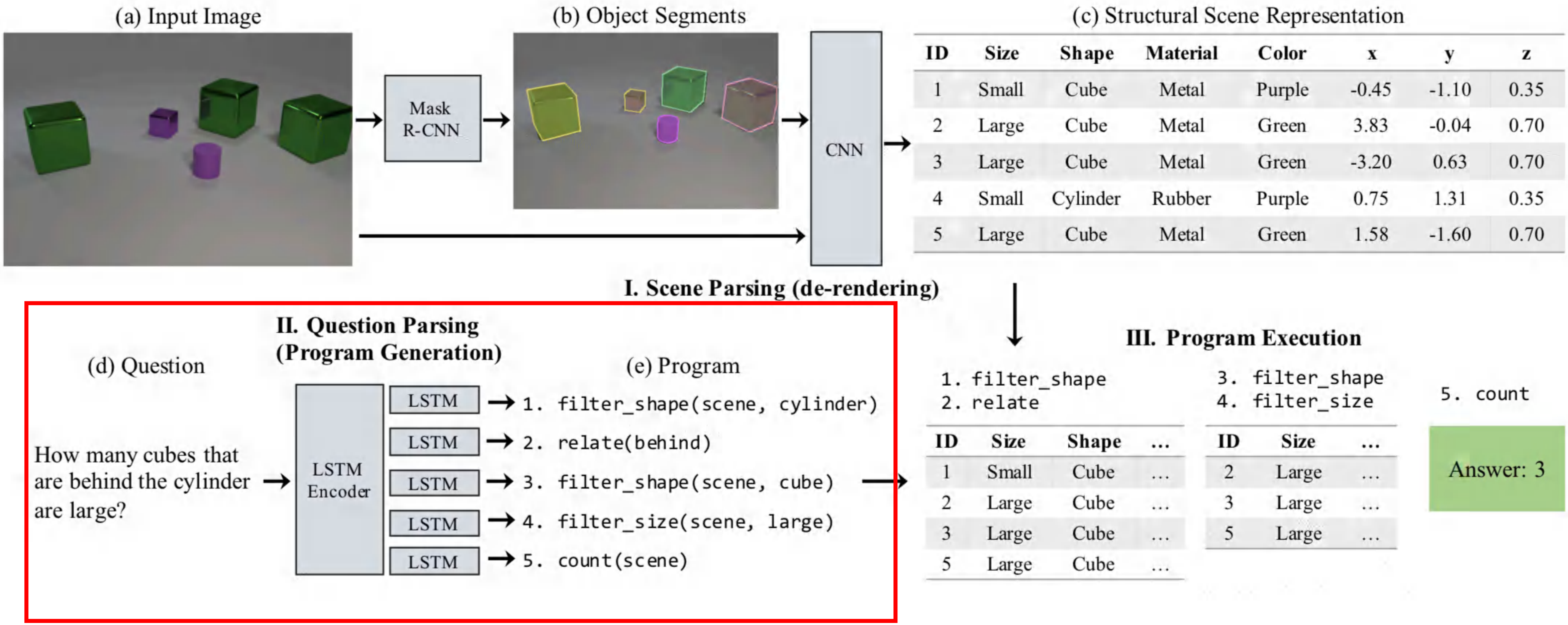
Neural-Symbolic VQA

Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. Yi et al. NeurIPS 2018



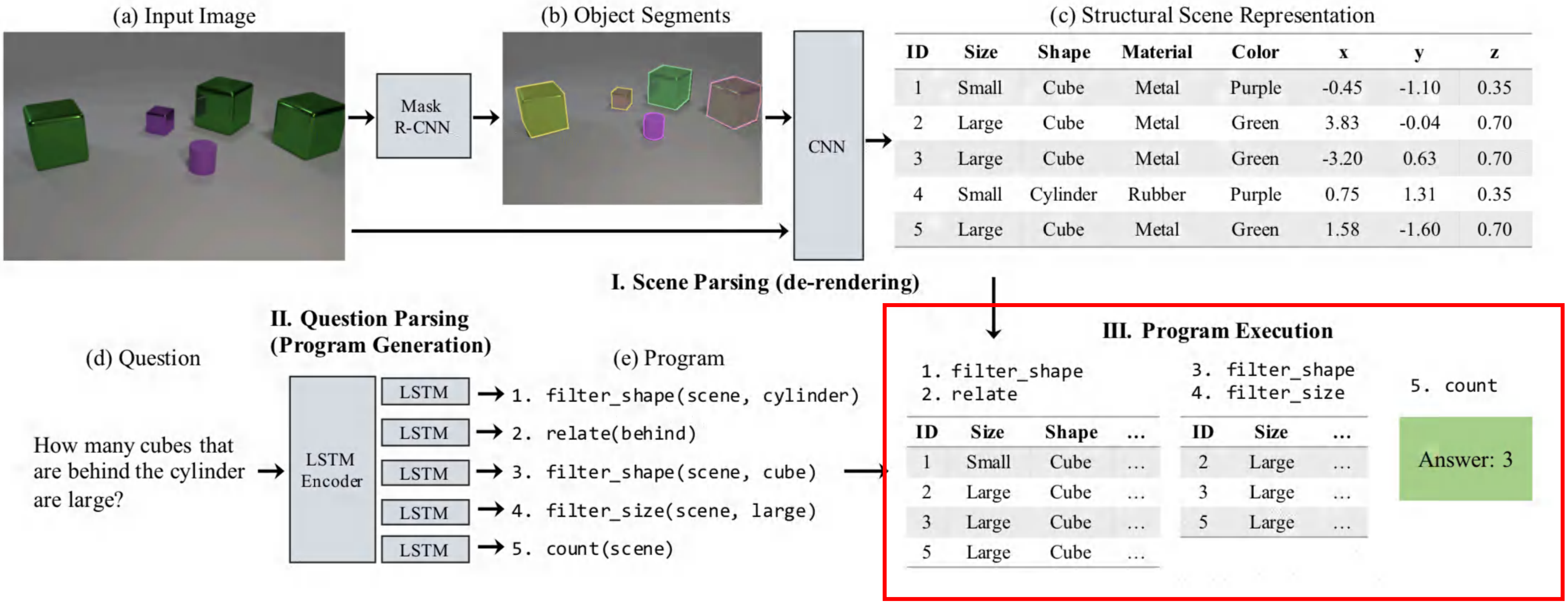
Neural-Symbolic VQA

Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. Yi et al. NeurIPS 2018



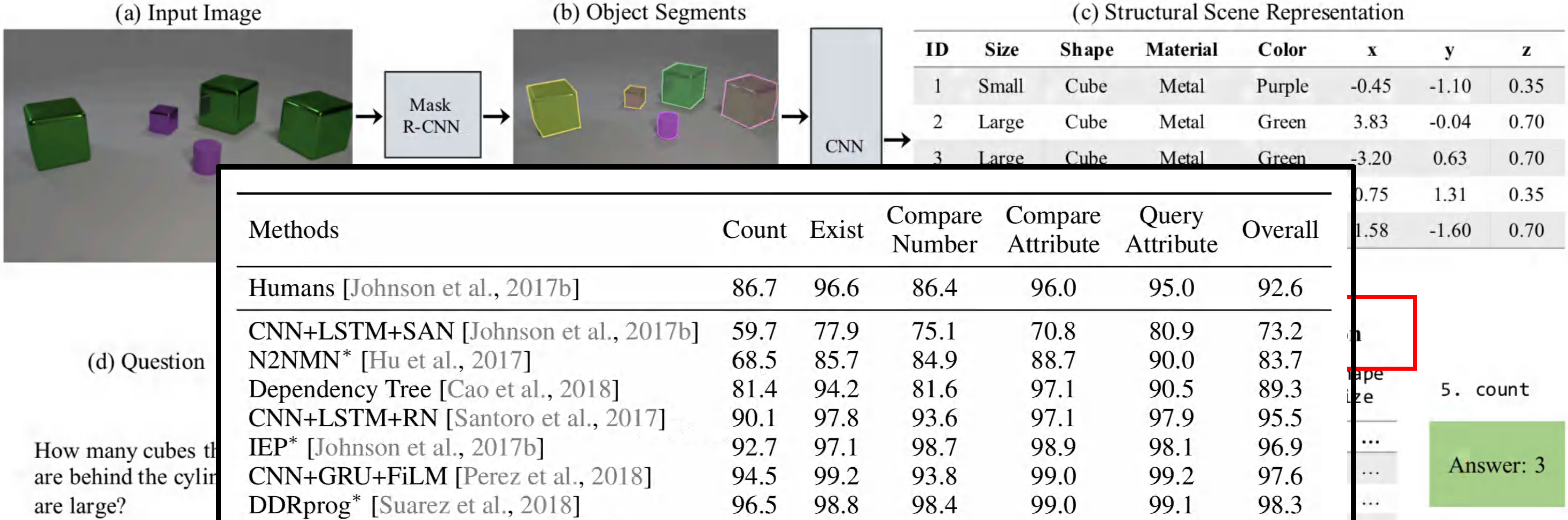
Neural-Symbolic VQA

Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. Yi et al. NeurIPS 2018



Neural-Symbolic VQA

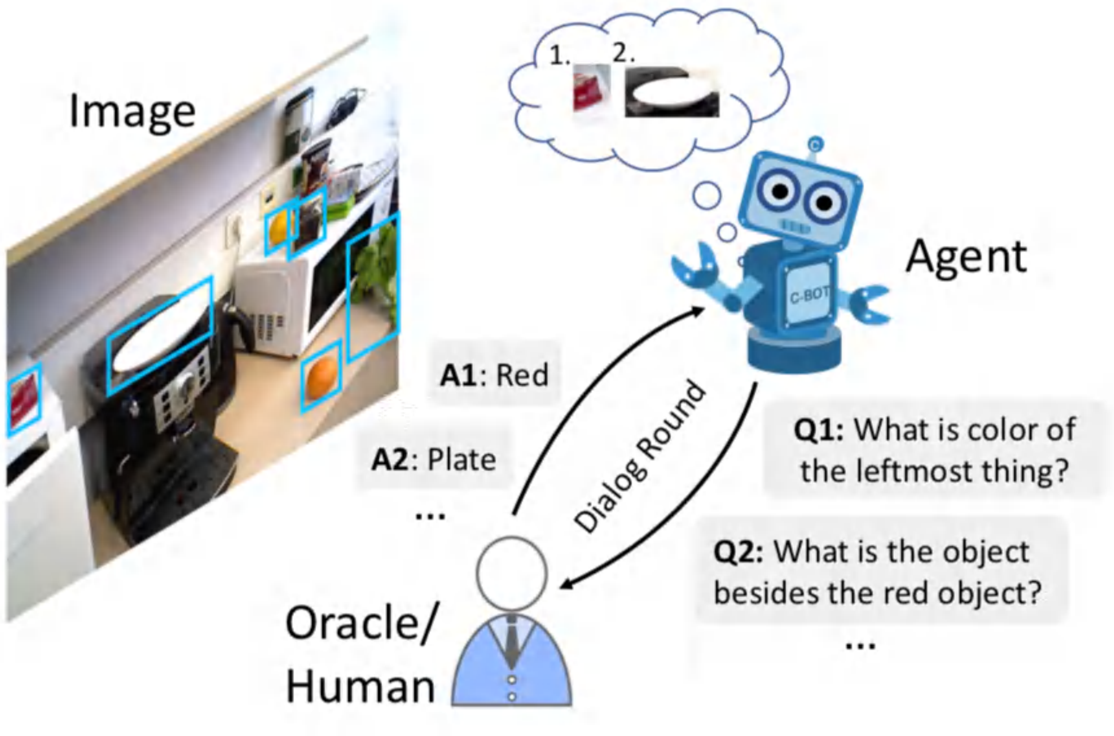
Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. Yi et al. NeurIPS 2018

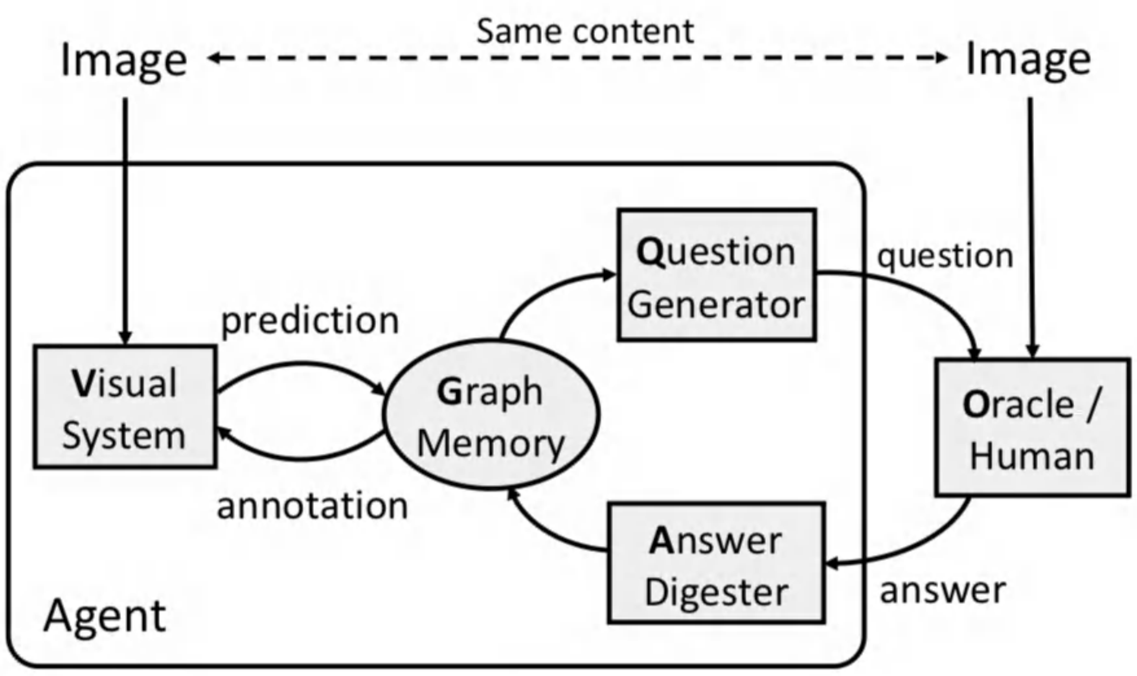


Methods	Count	Exist	Compare Number	Compare Attribute	Query Attribute	Overall
Humans [Johnson et al., 2017b]	86.7	96.6	86.4	96.0	95.0	92.6
CNN+LSTM+SAN [Johnson et al., 2017b]	59.7	77.9	75.1	70.8	80.9	73.2
N2NMN* [Hu et al., 2017]	68.5	85.7	84.9	88.7	90.0	83.7
Dependency Tree [Cao et al., 2018]	81.4	94.2	81.6	97.1	90.5	89.3
CNN+LSTM+RN [Santoro et al., 2017]	90.1	97.8	93.6	97.1	97.9	95.5
IEP* [Johnson et al., 2017b]	92.7	97.1	98.7	98.9	98.1	96.9
CNN+GRU+FiLM [Perez et al., 2018]	94.5	99.2	93.8	99.0	99.2	97.6
DDRprog* [Suarez et al., 2018]	96.5	98.8	98.4	99.0	99.1	98.3
MAC [Hudson and Manning, 2018]	97.1	99.5	99.1	99.5	99.5	98.9
TbD+reg+hres* [Mascharka et al., 2018]	97.6	99.2	99.4	99.6	99.5	99.1
NS-VQA (ours, 90 programs)	64.5	87.4	53.7	77.4	79.7	74.4
NS-VQA (ours, 180 programs)	85.0	92.9	83.4	90.6	92.2	89.5
NS-VQA (ours, 270 programs)	**99.7**	**99.9**	**99.9**	**99.8**	**99.8**	**99.8**

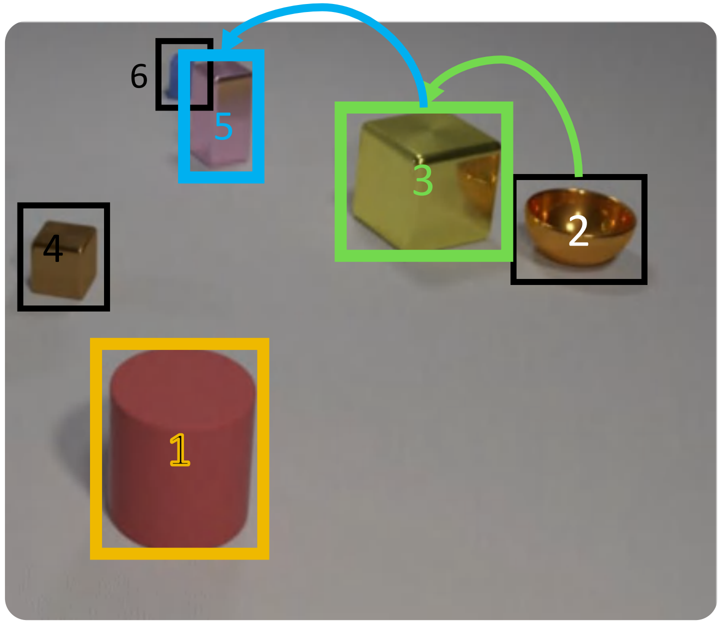
Learning to Generate Questions

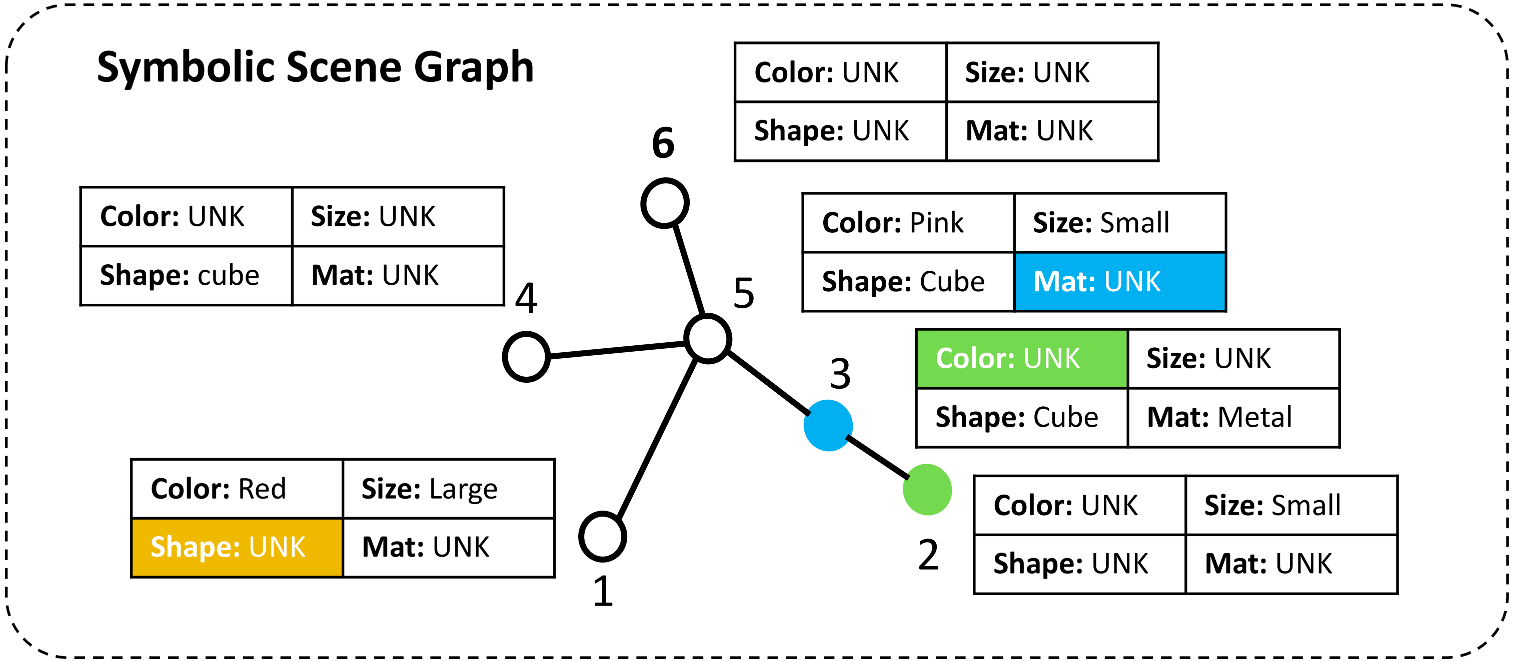
Visual Curiosity: Learning to Ask Questions to Learn Visual Recognition. Yang and Lu et al. CoRL 2018





Learning to Generate Questions

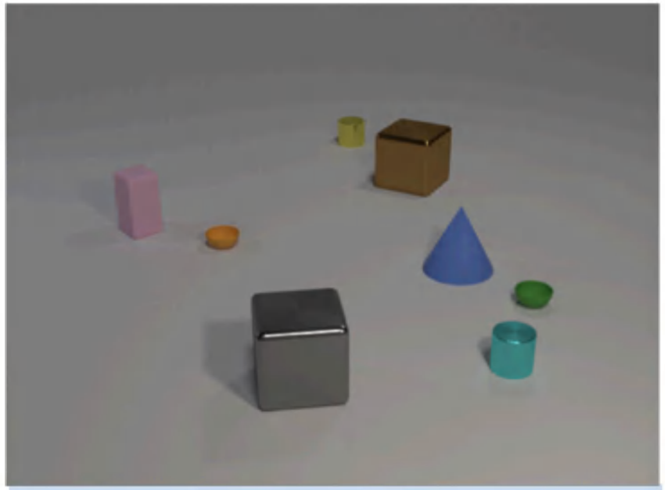




Target	1	3	5
Attribute	Shape	Color	Material
Reference	None	2	3
Question	What is the shape of the front most large red object?	What is the color of the metal cube on the left side of a small object?	What is the material of object at left side of metal cube?

Visual Curiosity: Learning to Ask Questions to Learn Visual Recognition. Yang and Lu et al. CoRL 2018

Learning to Generate Questions

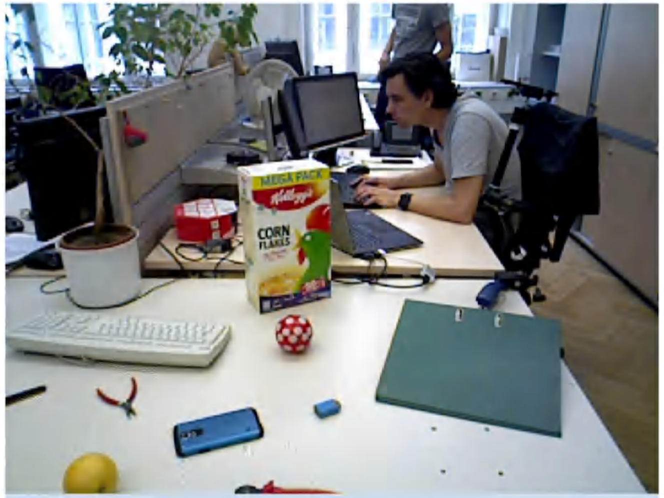


Q1: What is the closest thing made of?
A1: metal
Q2: What shape is the closest object?
A2: cylinder

Q7: What shape is the rightmost thing?
A7: bowl
Q8: The rightmost thing has what color?
A8: green
Q9: What is the size of the rightmost thing?
A9: small
Q10 : There is a closest object to the left of the tiny green matte bowl; what is its material? **A10**: rubber
Q11: What shape is the closest matte object to the left of the tiny green matte

Q1: What is the shape of the farthest thing?
A1: ball
Q2: What material is the farthest object?
A2: plastic

Q7: The leftmost object is what color?
A7: brown
Q8: What is the closest thing that is in front of the yellow plastic ball made of?
A8: paper
Q9: What shape is the closest thing that is in front of the yellow plastic ball?
A9: cereal
Q10: The closest paper cereal in front of the yellow plastic ball is what color?
A10: red



Q1: What is the rightmost thing made of?
A1: plastic
Q2: There is a rightmost object; what shape is it? **A2**: stapler

A6: __AMBIGUOUS__
Q7: What material is the leftmost thing?
A7: food
Q8: The leftmost thing is what shape?
A8: orange
Q9: The leftmost thing is what color?
A9: yellow
Q10: What material is the closest object right of the yellow food orange?
A10: plastic
Q11: There is a closest plastic thing that is

Part II: Summary

- Take away messages:
  - Scene graph can be used as feature or symbolic representation of image
  - Scene graph improves vision-language tasks like VQA and image captioning
  - Scene graph make the models more interpretable

- Potential Directions:
  - Leverage scene graph for explicit and effective reasoning on **realistic** data
  - Language context dependent scene graph generation
  - Combine scene graph and knowledge graph for common sense reasoning

Thanks! Questions?