JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

POSTER ID
THU-AM-209

# Devil's on the Edges:

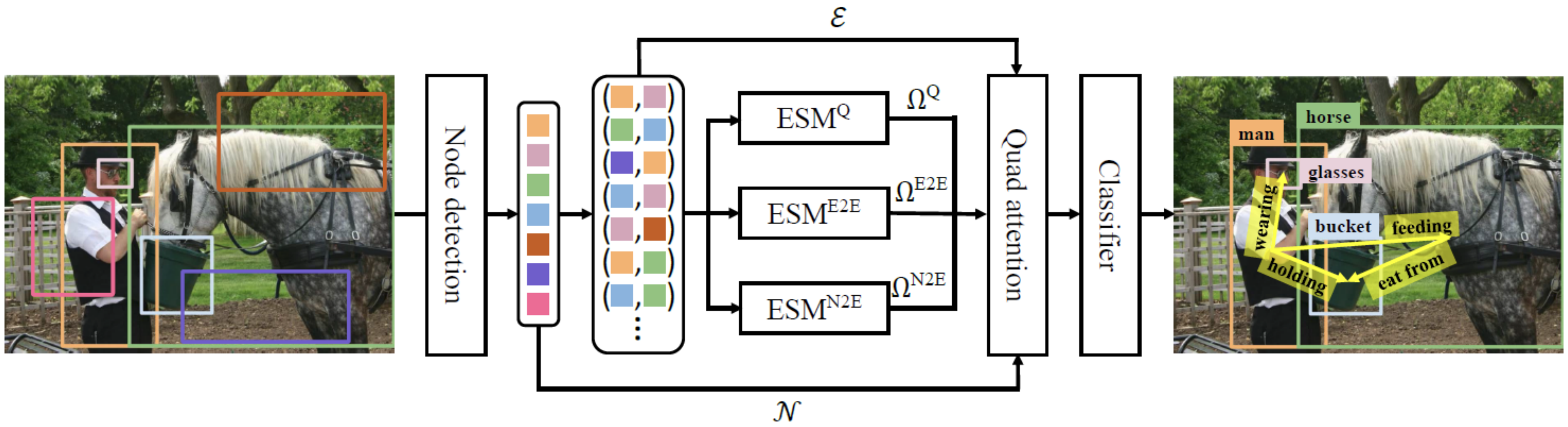
## Selective Quad Attention for Scene Graph Generation

Deunsol Jung     Sanghyun Kim     Won Hwa Kim     Minsu Cho

Pohang University of Science and Technology (POSTECH)

POSTECH

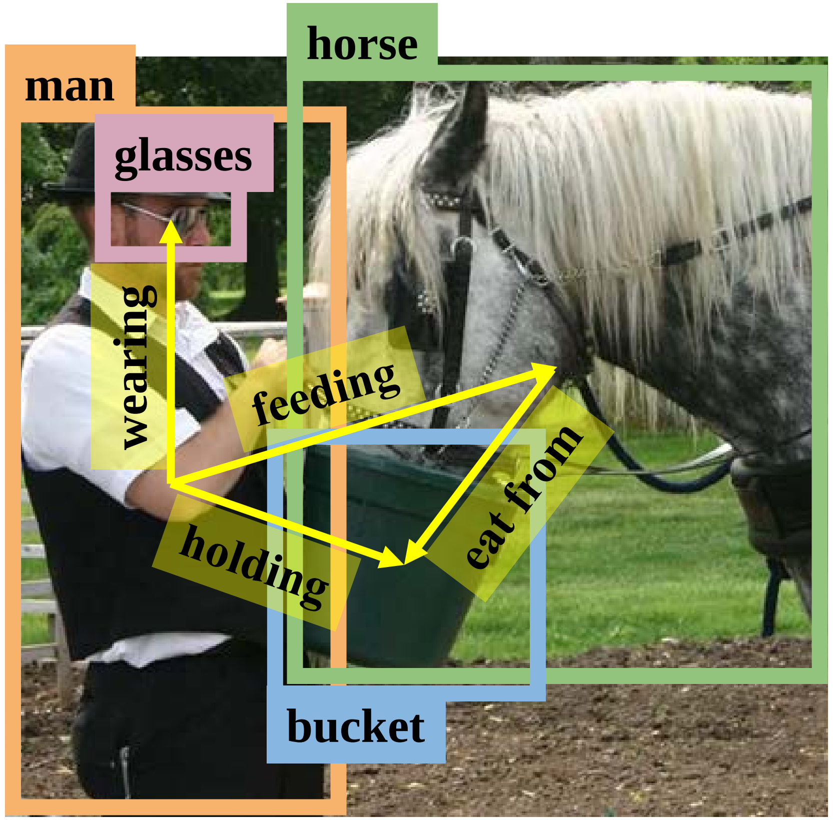
# Selective Quad Attention Networks



- ✓ Edge selection module: selecting relevant edges for contextual reasoning.
- ✓ Quad attention module: updating node and edge features via diverse interactions.
- ✓ Selective Quad Attention Networks (SQUAT)
  - Edge selection module + Quad attention module = achieving state-of-the-art

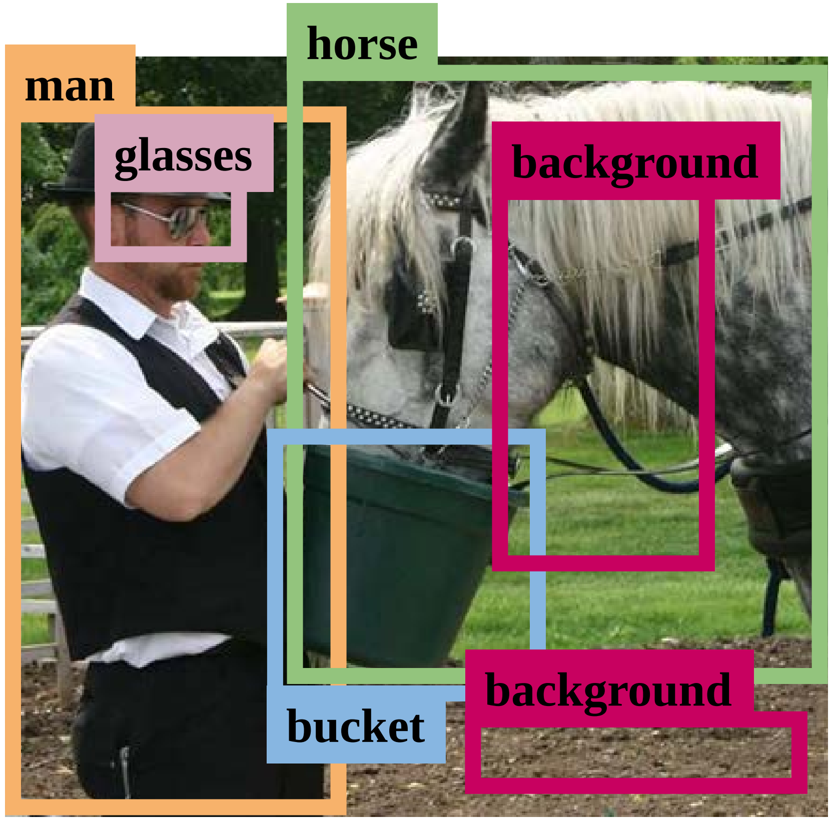
# Preliminary: scene graph generation

Predicting the objects $\mathcal{O}$ and their semantic relationships $\mathcal{R}$ given an image.



# Preliminary: scene graph generation

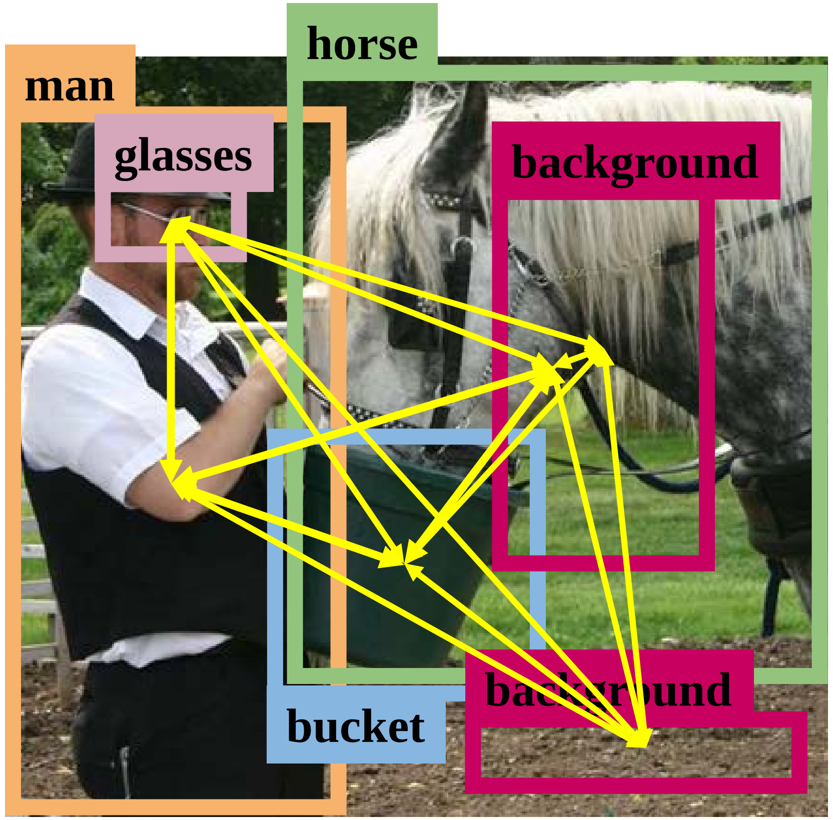
1. Using a pre-trained object detector [1], extract the object bounding boxes and labels.



Girshick, Ross. "Fast r-cnn.", *ICCV*. 2015.

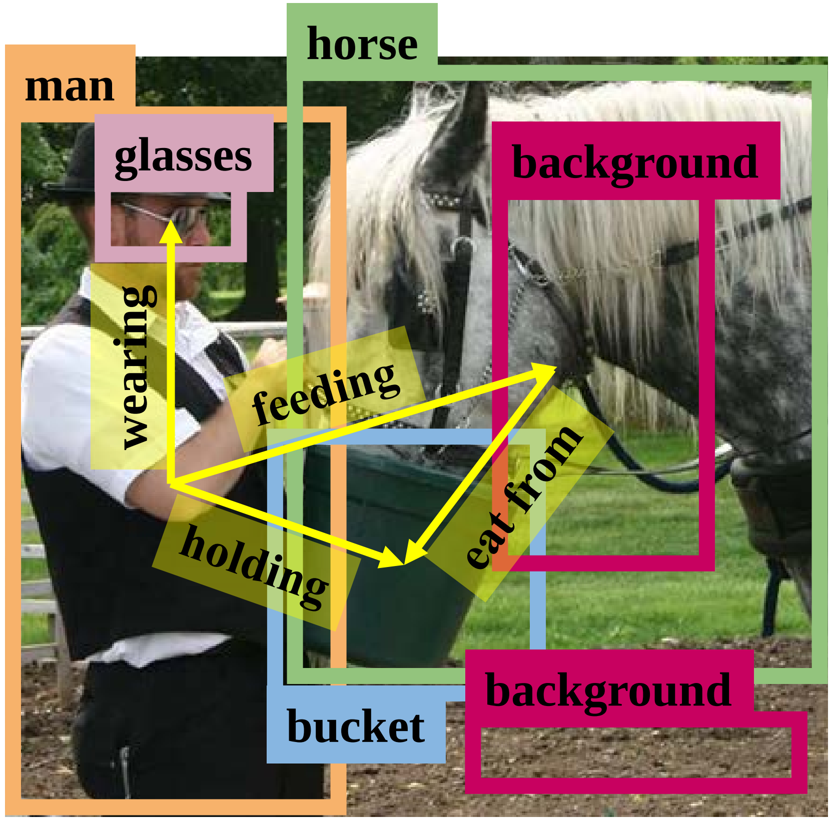
# Preliminary: scene graph generation

2. Node and edge features are updated by contextual reasoning through fully-connected graph.



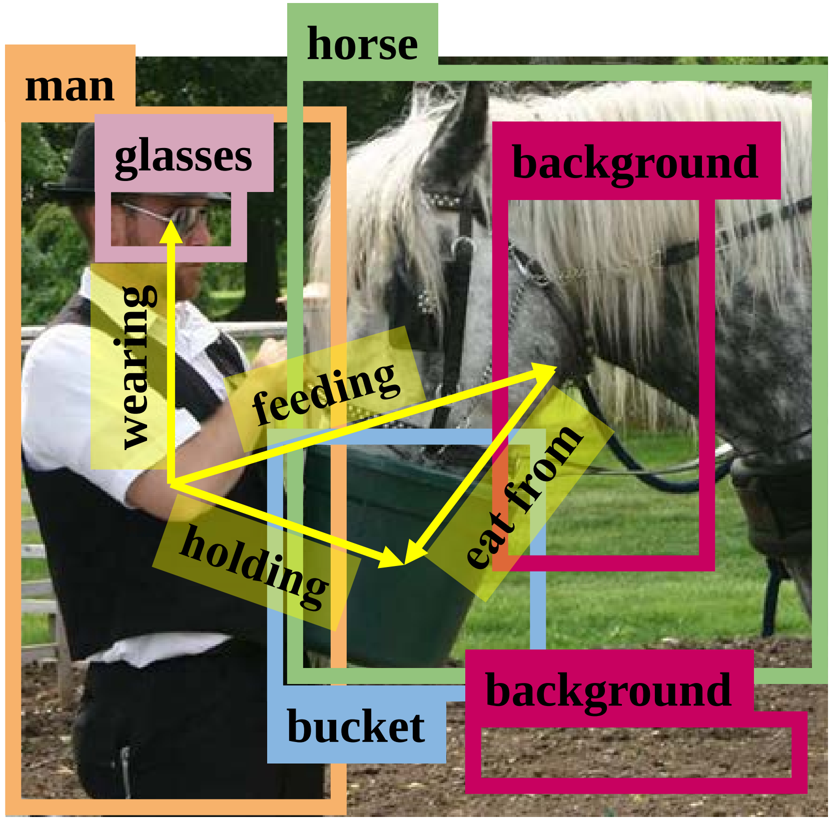
# Preliminary: scene graph generation

3. The relationships between objects are classified based on the updated edge features.

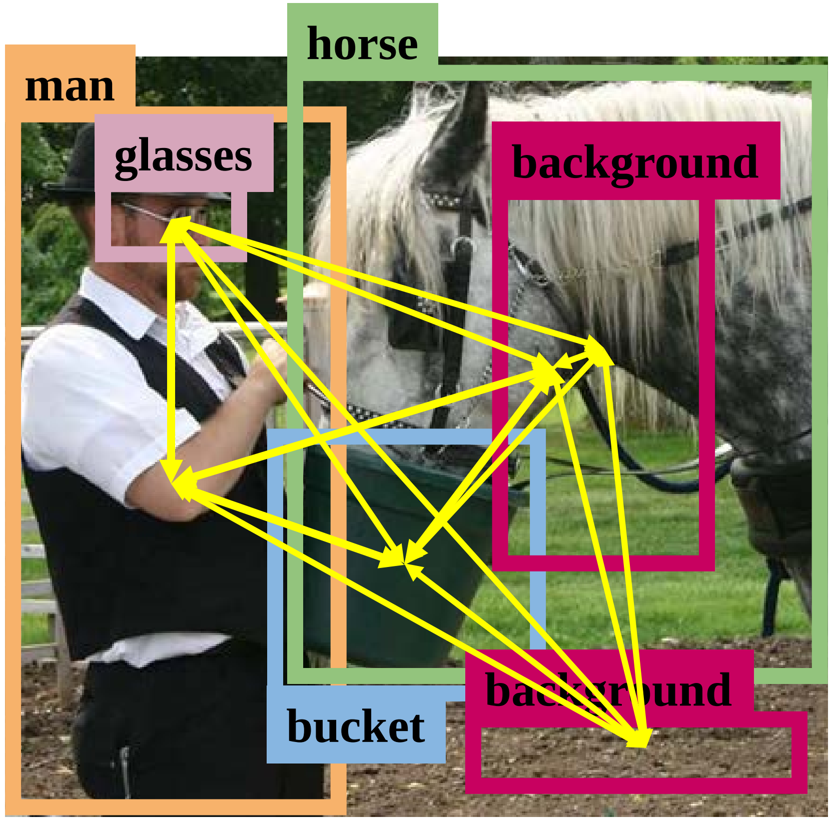


# Limitation of existing methods

The contextual reasoning is largely distracted by irrelevant objects and their relationship pairs.



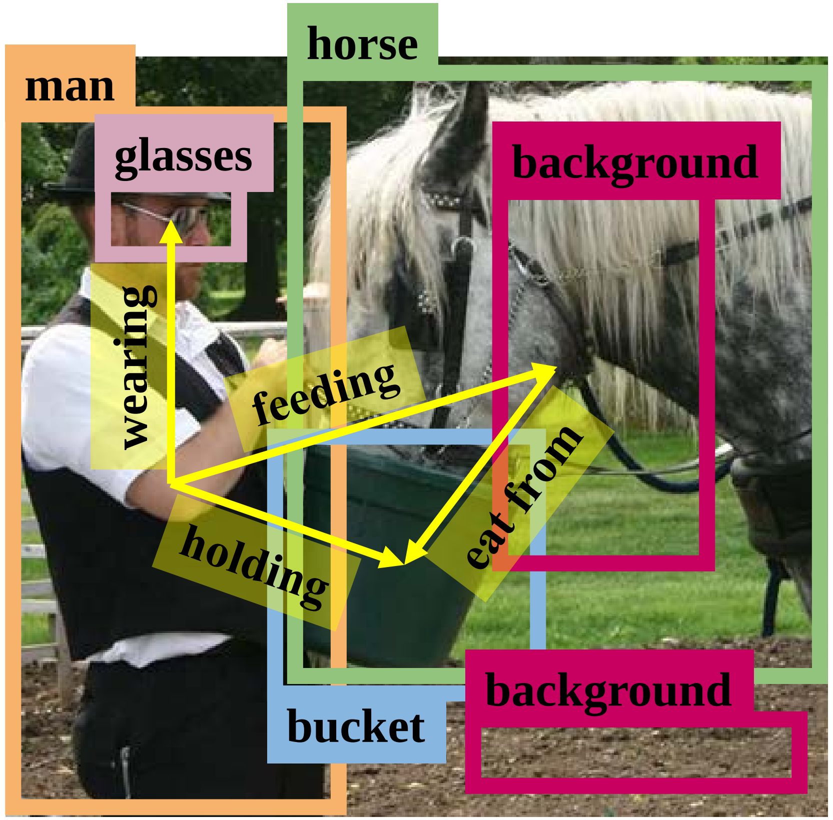
(a) ground-truth scene graph



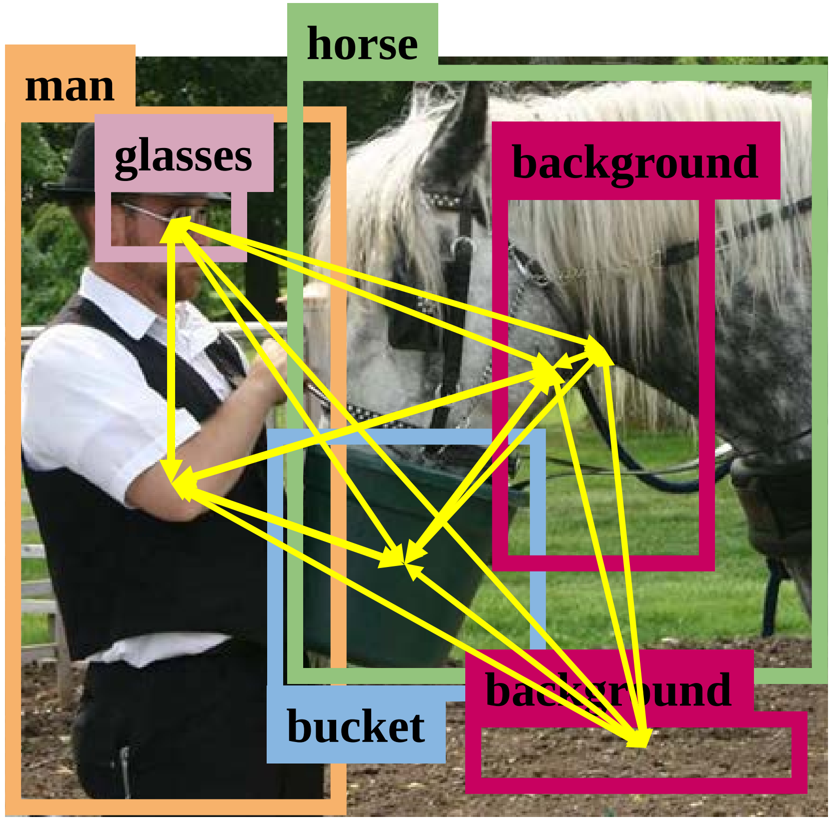
(b) fully-connected initial graph

# Limitation of existing methods

Node-to-node or node-to-edge interactions are limited in capturing such relations between edges.



(a) ground-truth scene graph



(b) fully-connected initial graph

# Limitation of existing methods

The contextual reasoning is largely distracted by irrelevant objects and their relationship pairs.

→ **Edge selection module**

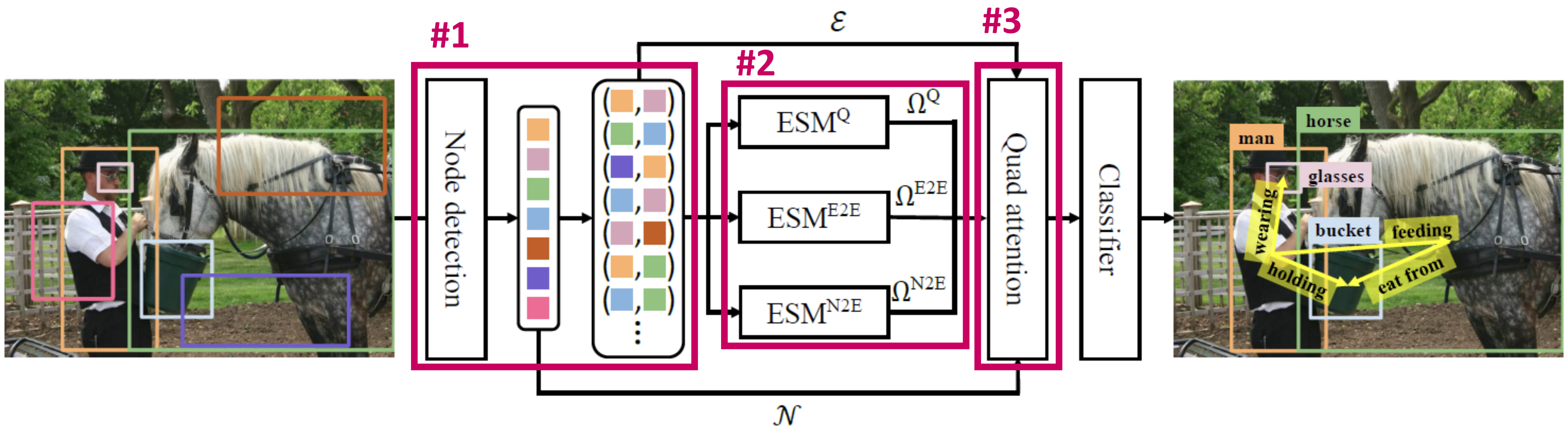
selecting relevant edges for contextual reasoning.

Note-to-node or node-to-edge interactions are limited in capturing such relations between edges.

→ **Quad attention module**

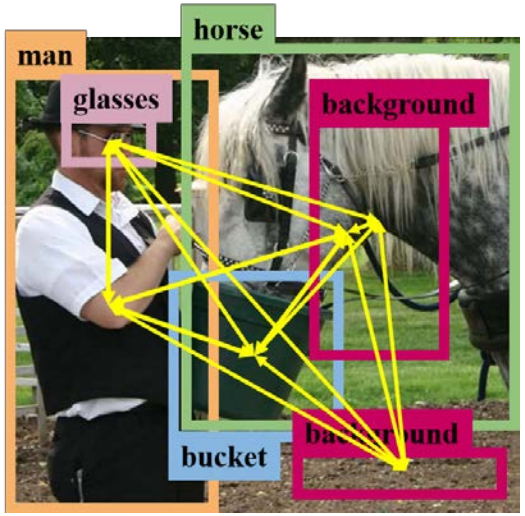
updating node and edge features via diverse interactions.

# Selective Quad Attention Network



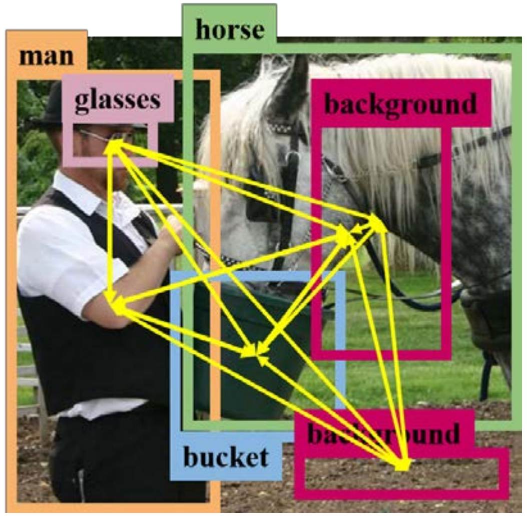
1. Node detection for object candidates
2. Edge selection for relevant object pairs
3. Quad attention for contextual reasoning

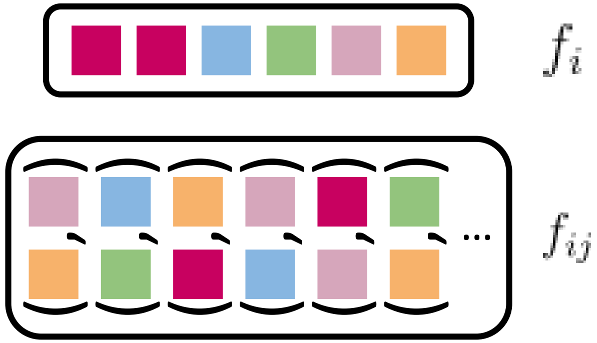
# Node detection module



1. Node detection for object candidates
   - Using pre-trained object detectors, extract a set of object bounding boxes.
   - Construct a initial node and edge features of fully-connected graph.

# Node detection module

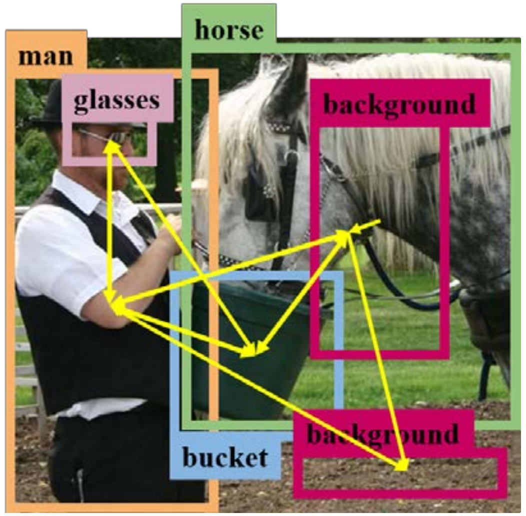


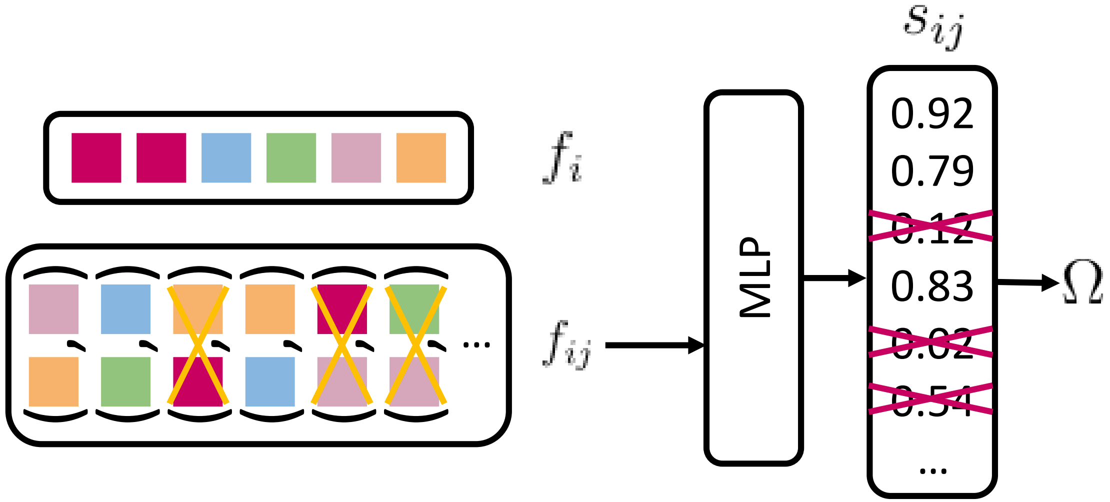


1. Node detection for object candidates
   - Using pre-trained object detectors, extract a set of object bounding boxes.
   - Construct a initial node and edge features of fully-connected graph.

$$f_i = W_o[W_v v_i; W_g b_i] \qquad f_{ij} = W_p[f_i; f_j]$$

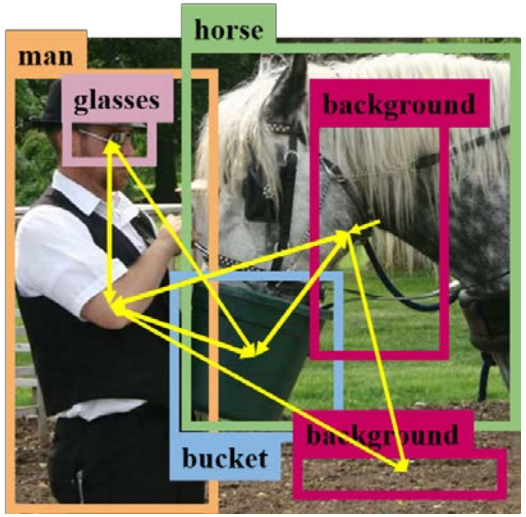
# Edge selection module

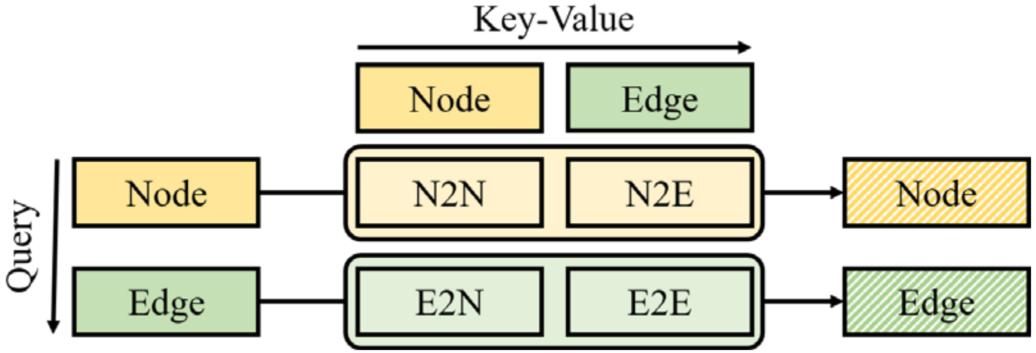




2. Edge selection for relevant object pairs
    - Predicts a relatedness score for the edges with a simple MLP.
    - Choose the edges with top-$\rho$% highest relatedness scores.

# Quad attention module





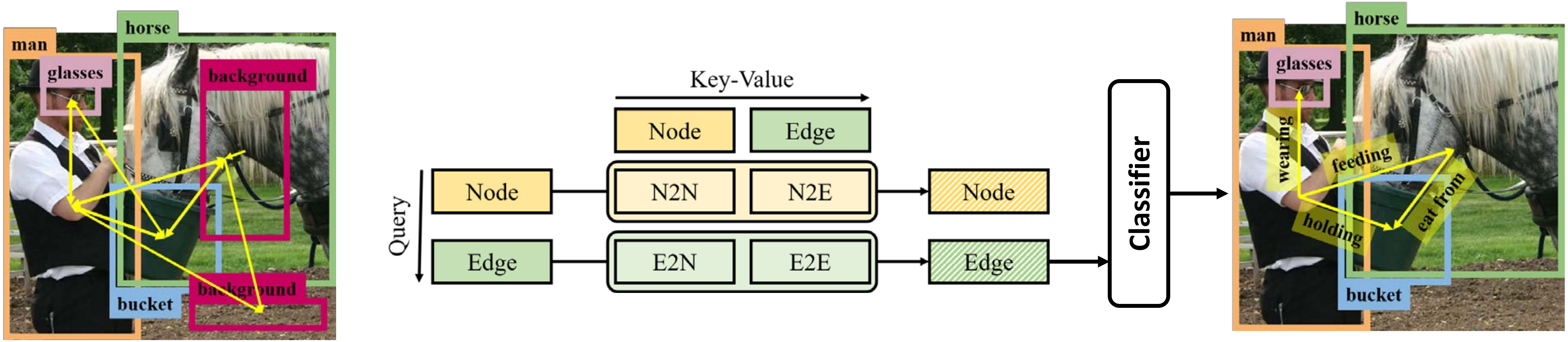
3. Quad attention for contextual reasoning
   - Update the node and selected edge features via four types of attention:

node-to-node(N2N), node-to-edge(N2E), edge-to-node(E2N), and edge-to-edge(E2E)

$$N_t' = \mathrm{LN}(N_t + \underbrace{\mathrm{MHA}(N_t, N_t, N_t)}_{\text{node-to-node attention}} + \underbrace{\mathrm{MHA}(N_t, E_t^{\mathrm{N2E}}, E_t^{\mathrm{N2E}})}_{\text{node-to-edge attention}})$$

$$E_t^{\mathrm{Q}\prime} = \mathrm{LN}(E_t^{\mathrm{Q}} + \underbrace{\mathrm{MHA}(E_t^{\mathrm{Q}}, N_t, N_t)}_{\text{edge-to-node attention}} + \underbrace{\mathrm{MHA}(E_t^{\mathrm{Q}}, E_t^{\mathrm{E2E}}, E_t^{\mathrm{E2E}})}_{\text{edge-to-edge attention}})$$

# Quad attention module

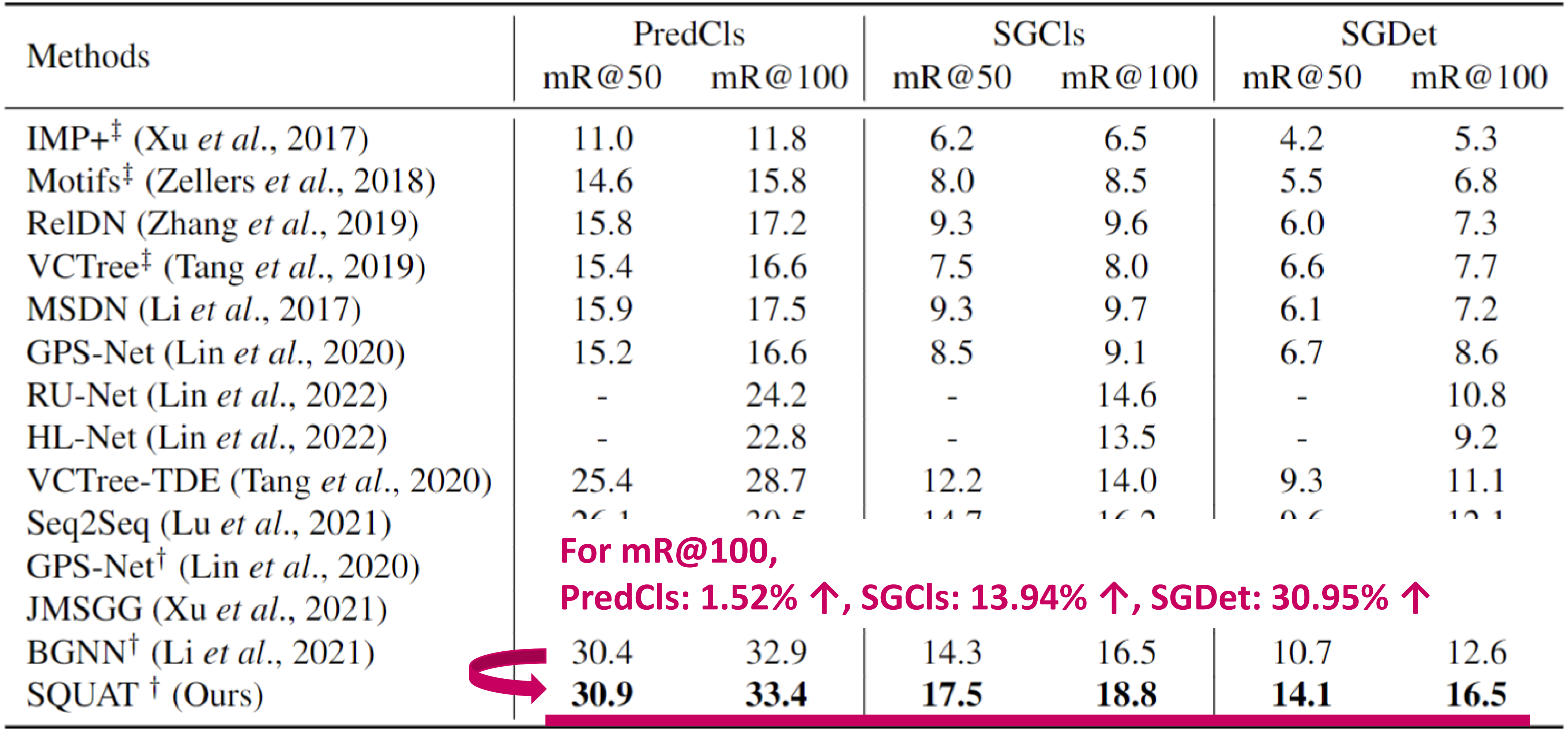


3. Quad attention for contextual reasoning
   - Update the node and selected edge features via four types of attention.
   - Produce sets of predicate probabilities using the output edge features.

# Datasets: Visual Genome

- The most popular dataset which is composed of 108k images with 150 object classes and 50 predicate classes.
- Tasks
  - Predicate Classification (PredCls): to predict the relationships given the G.T. bounding boxes and object labels.
  - Scene Graph Classification (SGCls): to predict the object labels and the relationships given the G.T. bounding boxes only.
  - Scene Graph Detection (SGDet): to predict the object bounding boxes, object labels, and the relationships.
- Evaluation Metrics
  - R@*K*: measure the fraction of G.T. relationship triplets that appear among the top most K confident predicates.
  - mR@*K*: retrieves each predicate separately and then averages R@*K* for all predicates.

# Experiments



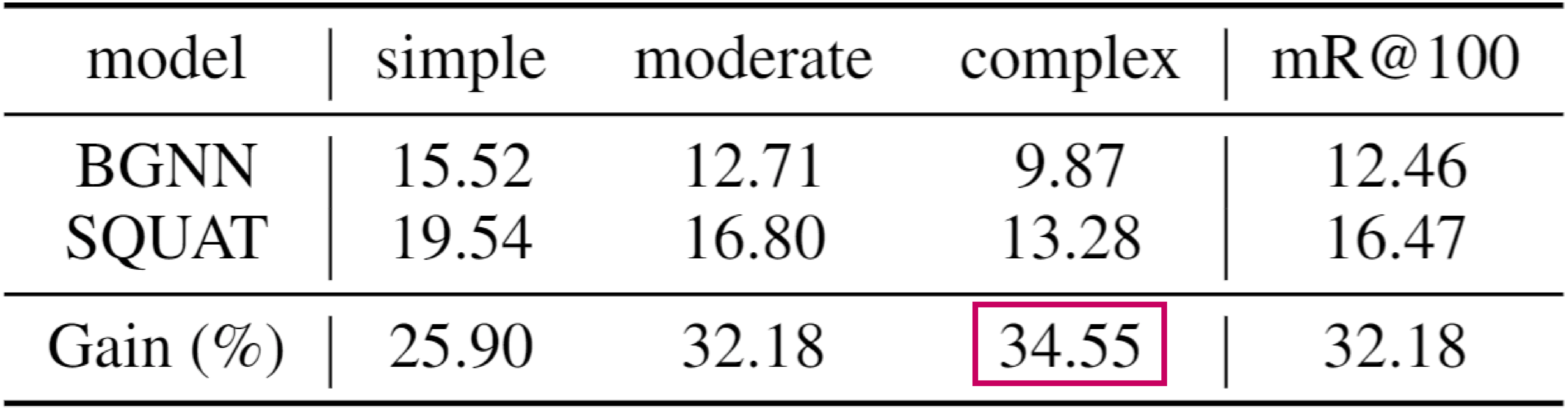
| Methods | PredCls |  | SGCls |  | SGDet |  |
|---|---|---|---|---|---|---|
|  | mR@50 | mR@100 | mR@50 | mR@100 | mR@50 | mR@100 |
| IMP+‡ (Xu *et al.*, 2017) | 11.0 | 11.8 | 6.2 | 6.5 | 4.2 | 5.3 |
| Motifs‡ (Zellers *et al.*, 2018) | 14.6 | 15.8 | 8.0 | 8.5 | 5.5 | 6.8 |
| RelDN (Zhang *et al.*, 2019) | 15.8 | 17.2 | 9.3 | 9.6 | 6.0 | 7.3 |
| VCTree‡ (Tang *et al.*, 2019) | 15.4 | 16.6 | 7.5 | 8.0 | 6.6 | 7.7 |
| MSDN (Li *et al.*, 2017) | 15.9 | 17.5 | 9.3 | 9.7 | 6.1 | 7.2 |
| GPS-Net (Lin *et al.*, 2020) | 15.2 | 16.6 | 8.5 | 9.1 | 6.7 | 8.6 |
| RU-Net (Lin *et al.*, 2022) | - | 24.2 | - | 14.6 | - | 10.8 |
| HL-Net (Lin *et al.*, 2022) | - | 22.8 | - | 13.5 | - | 9.2 |
| VCTree-TDE (Tang *et al.*, 2020) | 25.4 | 28.7 | 12.2 | 14.0 | 9.3 | 11.1 |
| Seq2Seq (Lu *et al.*, 2021) | 26.1 | 30.5 | 14.7 | 16.2 | 9.6 | 12.1 |
| GPS-Net† (Lin *et al.*, 2020) |  |  |  |  |  |  |
| JMSGG (Xu *et al.*, 2021) |  |  |  |  |  |  |
| BGNN† (Li *et al.*, 2021) | 30.4 | 32.9 | 14.3 | 16.5 | 10.7 | 12.6 |
| SQUAT † (Ours) | **30.9** | **33.4** | **17.5** | **18.8** | **14.1** | **16.5** |

**For mR@100,**
**PredCls: 1.52% ↑, SGCls: 13.94% ↑, SGDet: 30.95% ↑**

Li, R., Zhang, S., Wan, B., & He, X. Bipartite graph network with adaptive message passing for unbiased scene graph generation. CVPR 2021.

# Experiments

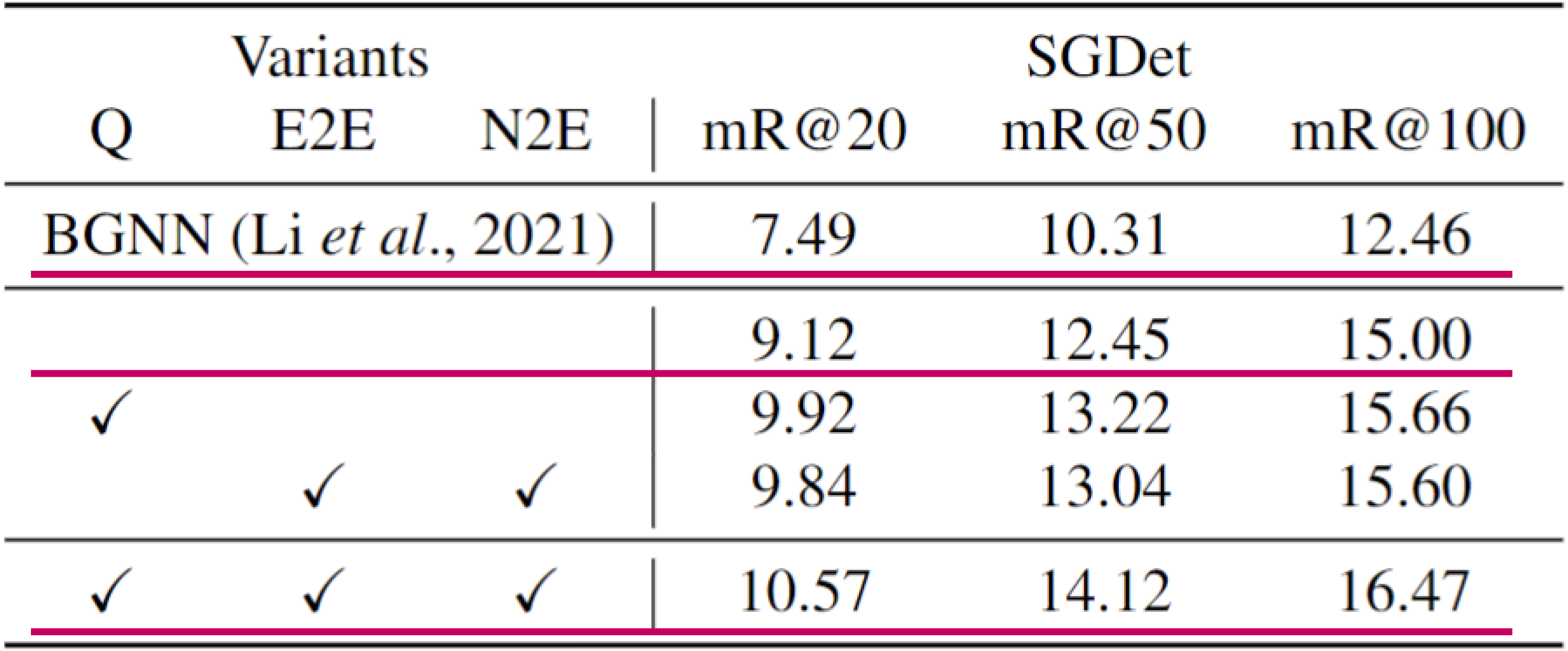
- Divide Visual Genome according to the number of objects in the scene: simple (≤ 9), moderate (10∼16), and complex (≥ 17)



| model | simple | moderate | complex | mR@100 |
|---|---|---|---|---|
| BGNN | 15.52 | 12.71 | 9.87 | 12.46 |
| SQUAT | 19.54 | 16.80 | 13.28 | 16.47 |
| Gain (%) | 25.90 | 32.18 | 34.55 | 32.18 |

Li, R., Zhang, S., Wan, B., & He, X. Bipartite graph network with adaptive message passing for unbiased scene graph generation. CVPR 2021.

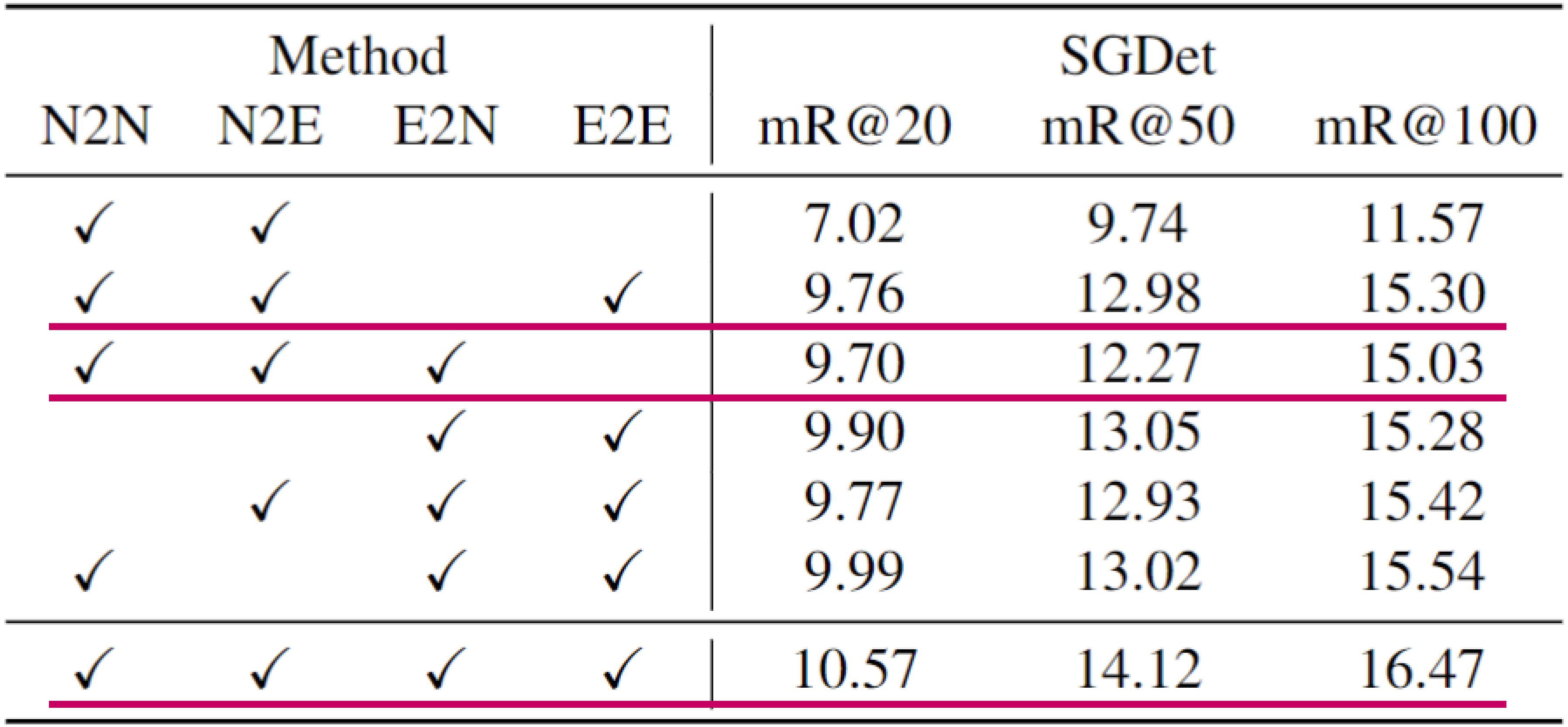
# Ablation study on Edge Selection



| Variants | | | SGDet | | |
|---|---|---|---|---|---|
| Q | E2E | N2E | mR@20 | mR@50 | mR@100 |
| BGNN (Li *et al.*, 2021) | | | 7.49 | 10.31 | 12.46 |
| | | | 9.12 | 12.45 | 15.00 |
| ✓ | | | 9.92 | 13.22 | 15.66 |
| | ✓ | ✓ | 9.84 | 13.04 | 15.60 |
| ✓ | ✓ | ✓ | 10.57 | 14.12 | 16.47 |

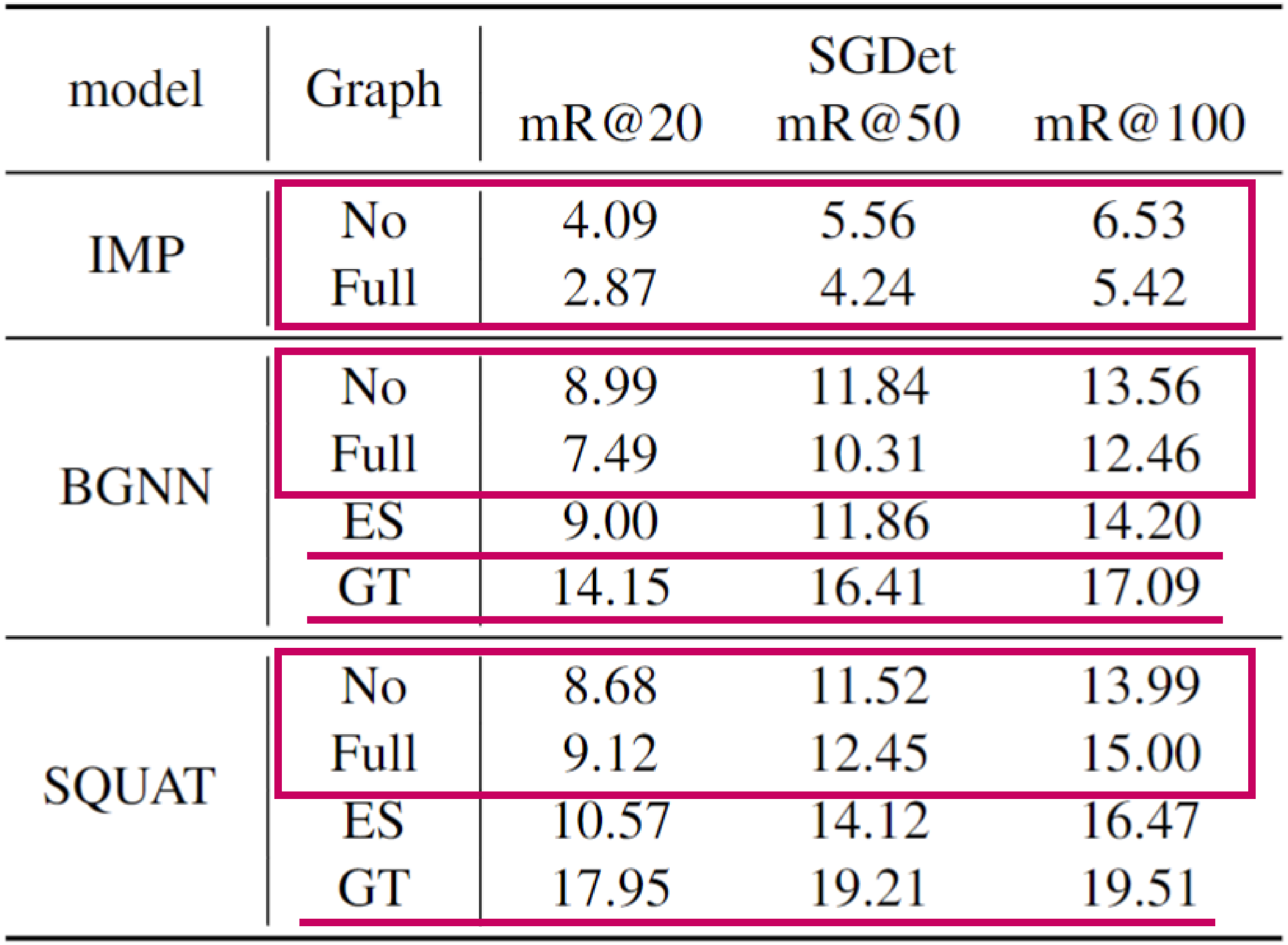
Li, R., Zhang, S., Wan, B., & He, X. Bipartite graph network with adaptive message passing for unbiased scene graph generation. CVPR 2021.

# Ablation study on Quad attention



| Method | | | | SGDet | | |
|---|---|---|---|---|---|---|
| N2N | N2E | E2N | E2E | mR@20 | mR@50 | mR@100 |
| ✓ | ✓ | | | 7.02 | 9.74 | 11.57 |
| ✓ | ✓ | | ✓ | 9.76 | 12.98 | 15.30 |
| ✓ | ✓ | ✓ | | 9.70 | 12.27 | 15.03 |
| | | ✓ | ✓ | 9.90 | 13.05 | 15.28 |
| | ✓ | ✓ | ✓ | 9.77 | 12.93 | 15.42 |
| ✓ | | ✓ | ✓ | 9.99 | 13.02 | 15.54 |
| ✓ | ✓ | ✓ | ✓ | 10.57 | 14.12 | 16.47 |

# Ablation study on message passing



| model | Graph | SGDet | | |
|---|---|---|---|---|
| | | mR@20 | mR@50 | mR@100 |
| IMP | No | 4.09 | 5.56 | 6.53 |
| | Full | 2.87 | 4.24 | 5.42 |
| BGNN | No | 8.99 | 11.84 | 13.56 |
| | Full | 7.49 | 10.31 | 12.46 |
| | ES | 9.00 | 11.86 | 14.20 |
| | GT | 14.15 | 16.41 | 17.09 |
| SQUAT | No | 8.68 | 11.52 | 13.99 |
| | Full | 9.12 | 12.45 | 15.00 |
| | ES | 10.57 | 14.12 | 16.47 |
| | GT | 17.95 | 19.21 | 19.51 |

Xu, D., Zhu, Y., Choy, C. B., & Fei-Fei, L., Scene graph generation by iterative message passing. CVPR 2017.
Li, R., Zhang, S., Wan, B., & He, X., Bipartite graph network with adaptive message passing for unbiased scene graph generation. CVPR 2021.

# Thank you

## Takeaways

✓ Edge selection module: selecting relevant edges for contextual reasoning.

✓ Quad attention module: updating node and edge features via diverse interactions.

✓ Selective Quad Attention Networks (SQUAT)

## Visit us @ CVPR’23

➢ Paper id #4074

➢ THU-AM-209

➢ 22nd June, 209