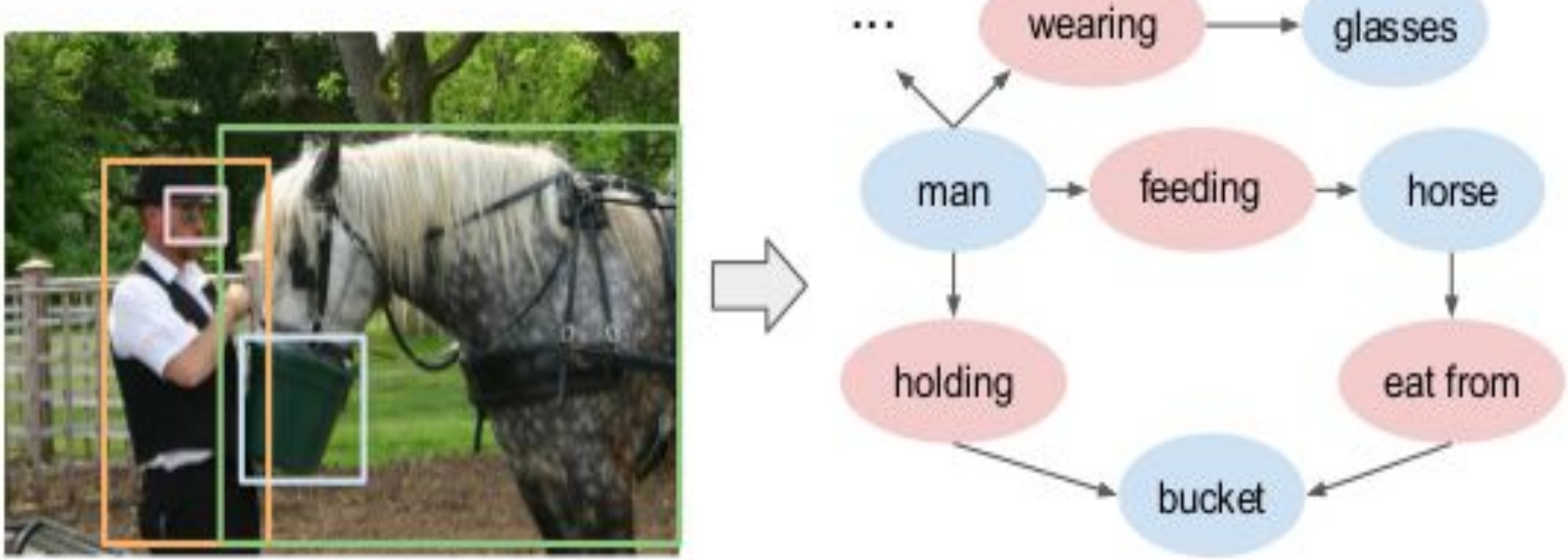
Scene Graph Generation by Iterative Message Passing

Danfei Xu, Yuke Zhu, Christopher B. Choy, Li Fei-Fei (2017)

https://arxiv.org/abs/1701.02426

Scene Graph

In every image, there's more than meets the object detector



Problem statement

Given an image *I* and a set of boxes *B*
from pretrained Region Proposal Network,
we want to identify:

➔ **Object classes**
For each box, the object class

➔ **Box offsets**
For each box, the offset w.r.t. the proposed box coordinates

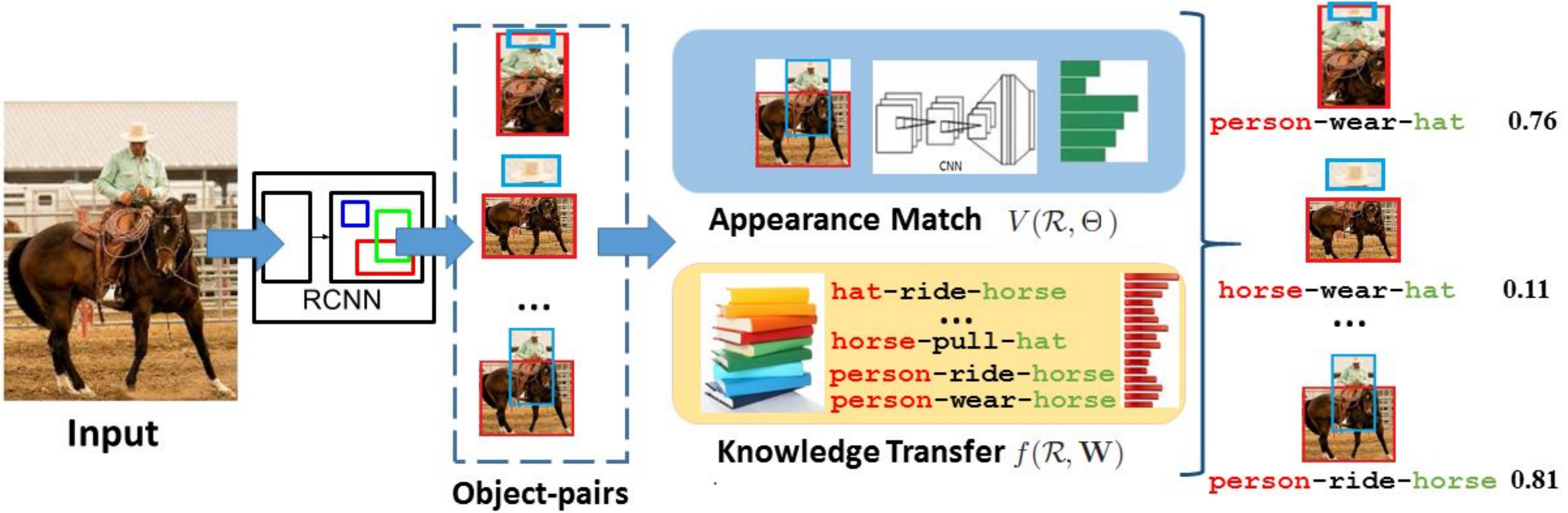
➔ **Pairwise relationships**
For each pair of boxes, the most likely relationship between their objects

The baseline

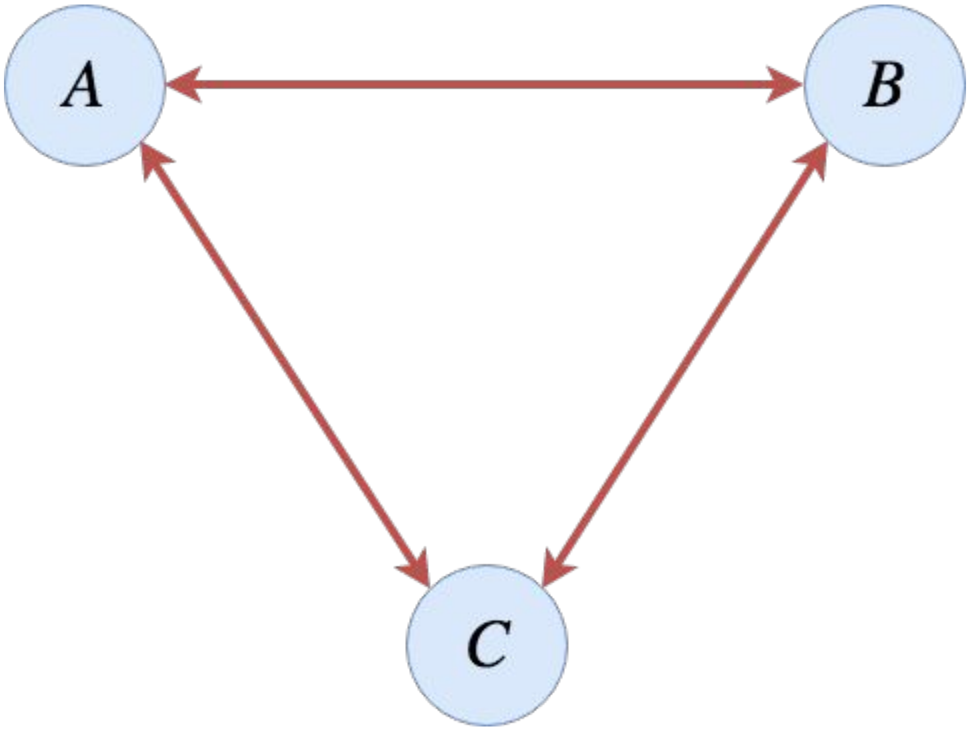
Visual Relationship Detection with Language Priors

Cewu Lu, Ranjay Krishna, Michael Bernstein, Li Fei-Fei (2016)

- Uses visual features from the region containing 2 objects
- Uses language priors to cluster relationships together
- Similar to the approach of this paper, without message passing

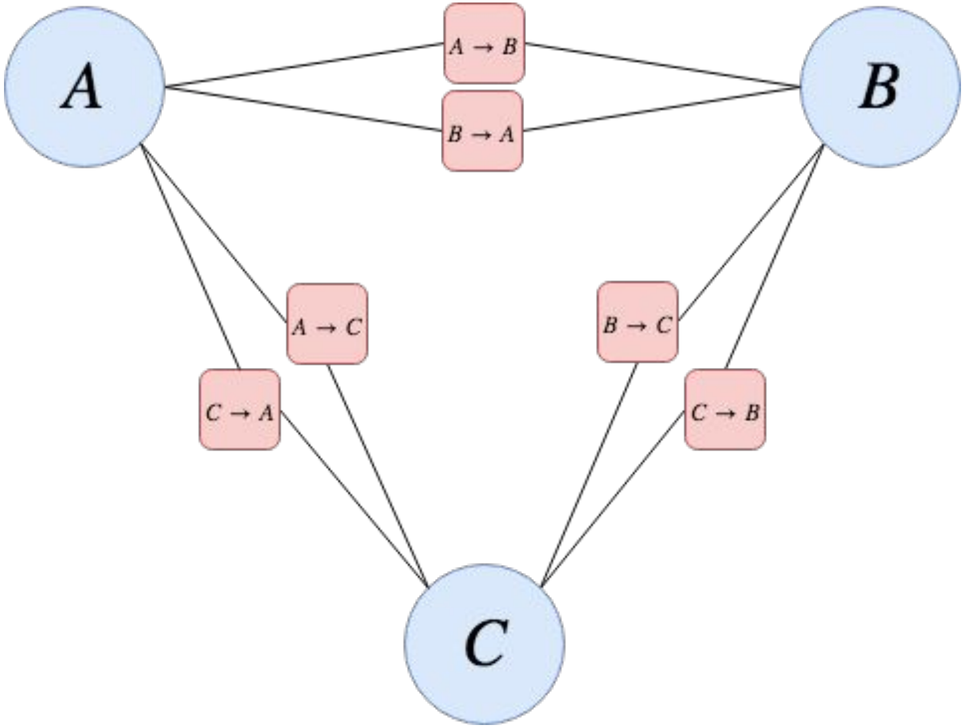


Primal Dual Graph



- **Nodes**
  Represent objects in the scene
- **Edges**
  Represent object relationships

Primal Dual Graph

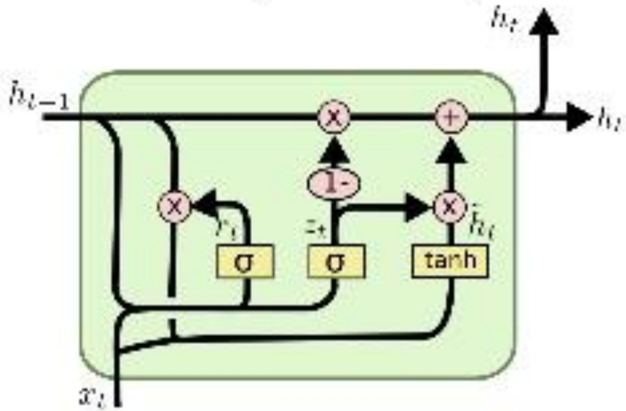


➔ **Object nodes**
Represent objects in the scene

➔ **Relationship nodes**
Represent object relationships

➔ **Edges**
Represent messages exchanged between object and relationship nodes

Object nodes and relationship nodes form a bipartite graph



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

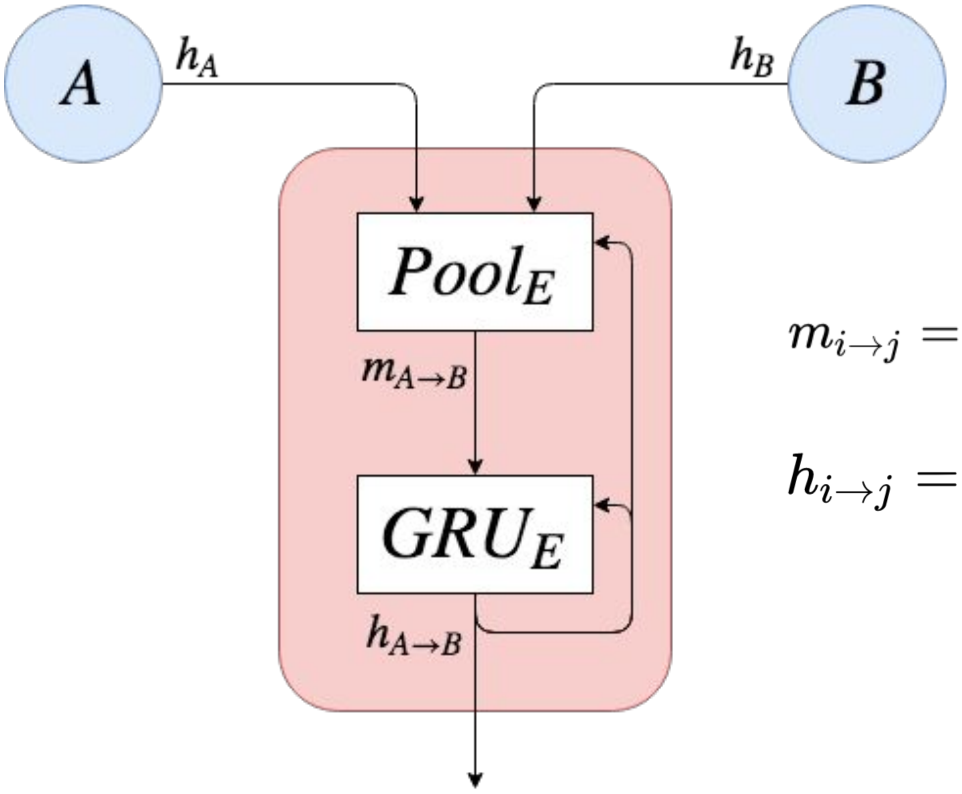
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Gated Recurrent Unit

Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio (2014)

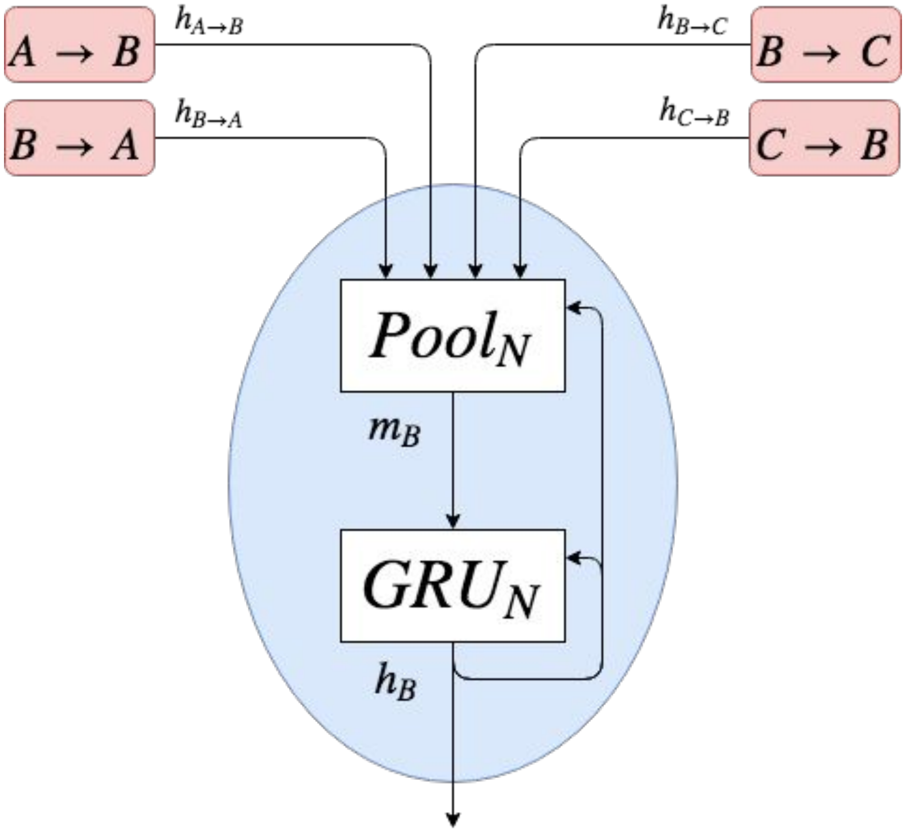
Relationship node A → B



$$m_{i \to j} = \sigma(\mathbf{w}_1^T [h_i, h_{i \to j}]) h_i + \sigma(\mathbf{w}_2^T [h_j, h_{i \to j}]) h_j$$

$$h_{i \to j} = GRU(m_{i \to j}, h_{i \to j})$$

Object node B



$$m_i = \sum_{j:i\rightarrow j} \sigma(\mathbf{v}_1^T[h_i, h_{i\rightarrow j}])h_{i\rightarrow j} + \sum_{j:j\rightarrow i} \sigma(\mathbf{v}_2^T[h_i, h_{j\rightarrow i}])h_{j\rightarrow i}$$

$$h_i = GRU(m_i, h_i)$$

Training procedure

- Pretrained VGG-16 for region proposals and visual feature
- 512-dimensional vectors for state and messages
- For each image, 128 boxes are randomly selected from the top 2.000 proposed boxes
- For each image, 128 labeled relationships are randomly selected from the 8.128 possible object pairs
- For inference, only the top 50 boxes and all their pairs are considered

Visual Genome

- 100k images
- Top 150 object classes
  (avg. 25 per image)
- Top 50 relationships
  (avg. 6.2 per image)

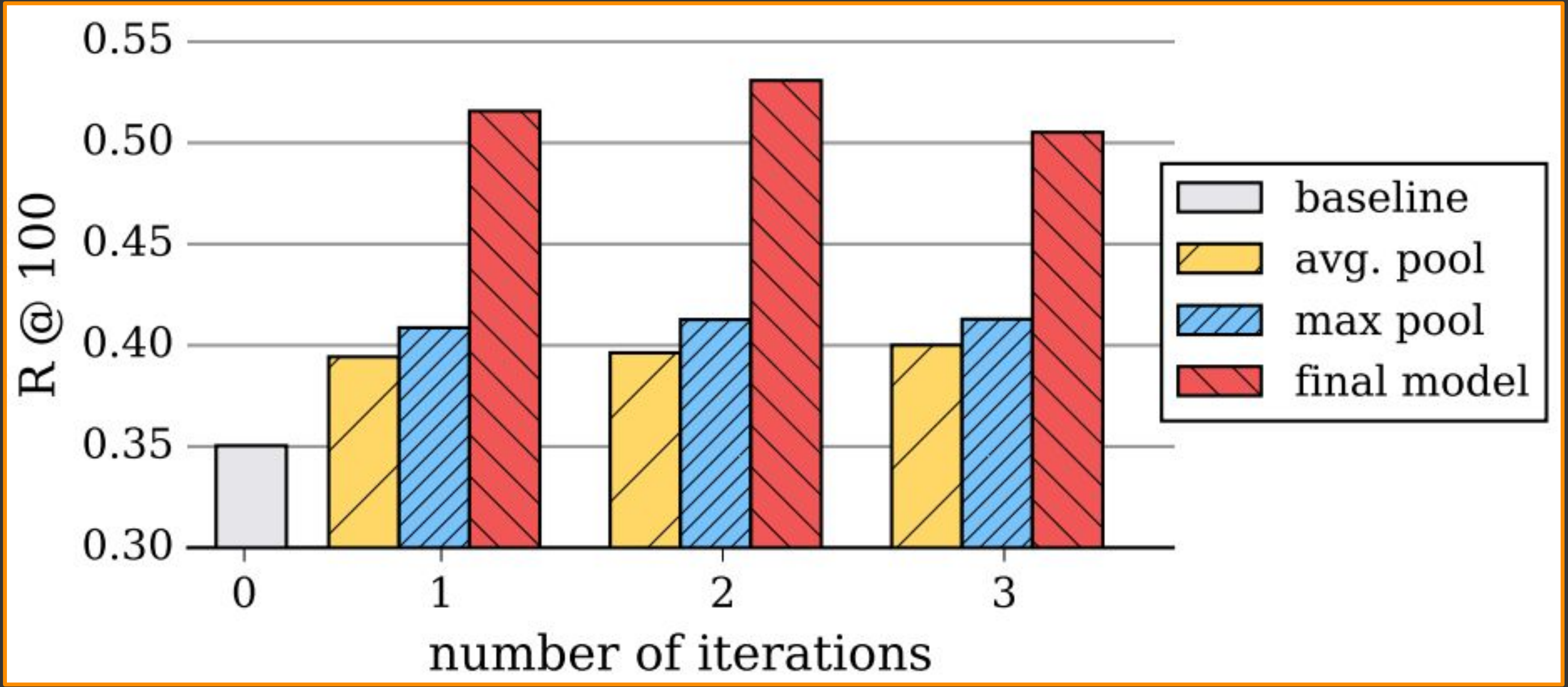
Note

This is a cleaned version of the VG dataset, because the original annotations were of poor quality.

Visual Genome

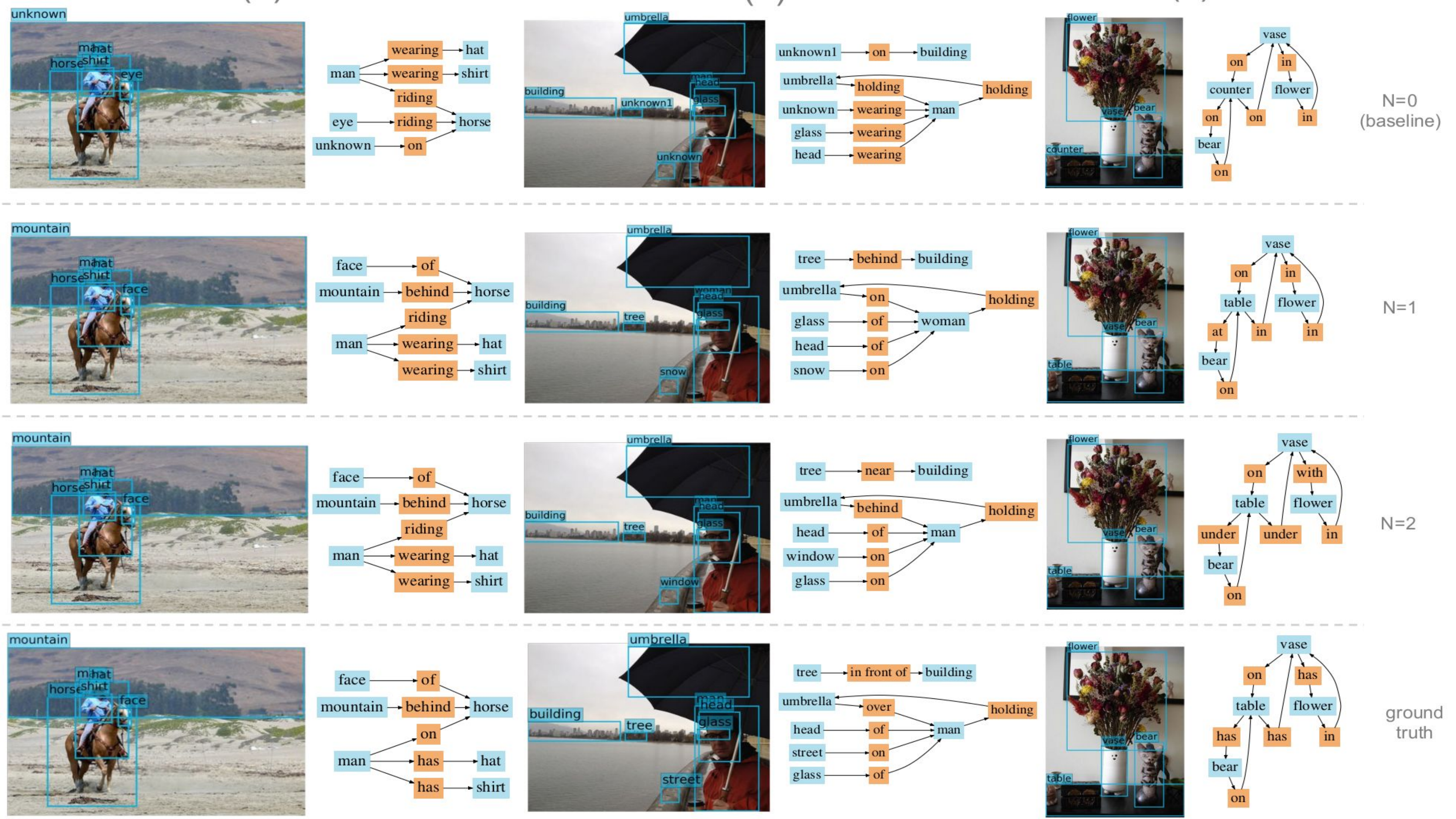
		[26]	avg. pool	max pool	final
PREDCLS	R@50	27.88	32.39	34.33	**44.75**
	R@100	35.04	39.63	41.99	**53.08**
SGCLS	R@50	11.79	15.65	16.31	**21.72**
	R@100	14.11	18.27	18.70	**24.38**
SGGEN	R@50	0.32	2.70	3.03	**3.44**
	R@100	0.47	3.42	3.71	**4.24**

Visual Genome



Visual Genome

predicate	[26]	ours	predicate	[26]	ours
on	**99.71**	99.25	under	28.64	**52.73**
has	**98.03**	97.25	sitting on	31.74	**50.17**
in	80.38	**88.30**	standing on	44.44	**61.90**
of	82.47	**96.75**	in front of	26.09	**59.63**
wearing	**98.47**	98.23	attached to	8.45	**29.58**
near	85.16	**96.81**	at	54.08	**70.41**
with	31.85	**88.10**	hanging from	0.00	0.00
above	49.19	**79.73**	over	**9.26**	0.00
holding	61.50	**80.67**	for	12.20	**31.71**
behind	79.35	**92.32**	riding	72.43	**89.72**


N=0
(baseline)
unknown
man
hat
horse
shirt
eye
wearing
riding
on
umbrella
building
unknown1
head
glass
holding
flower
vase
bear
counter
in
N=1
mountain
face
of
behind
woman
tree
snow
table
at
N=2
near
window
under
with
ground
truth
has
in front of
over
street

NYU Depth v2

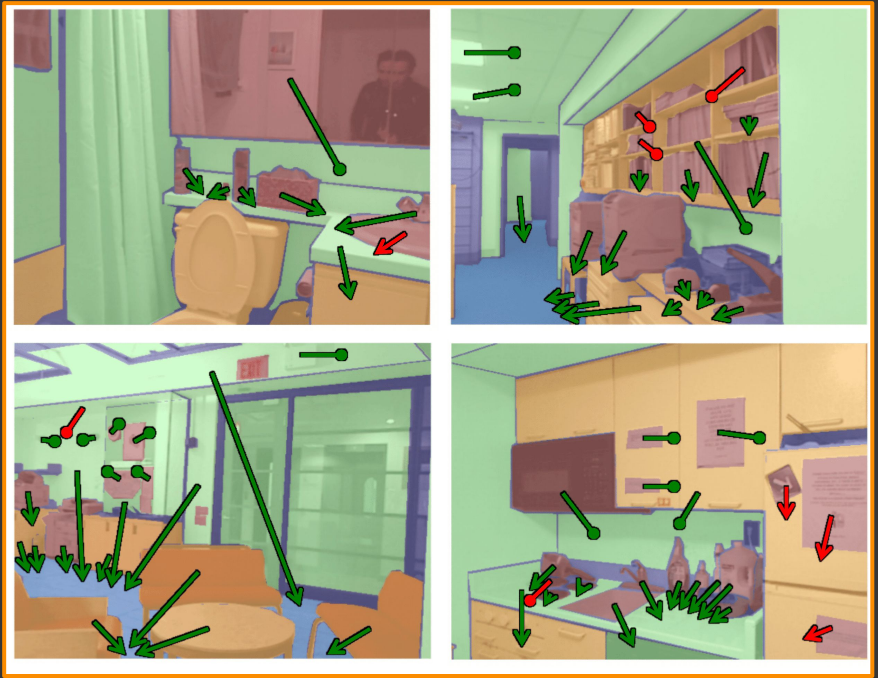
- 1.449 RGB-D images
- 4 object classes
  (floor, structure, furniture, prop)
- 3 support relationships
  (behind, below, hidden)

Notes

The depth channel is not used during the experiments.

Ground-truth object locations are provided as inputs, not predicted.

NYU Depth v2



NYU Depth v2

	Support Accuracy		PredCls	
	t-ag	t-aw	R@50	R@100
Silberman *et al.* [28]	75.9	72.6	-	-
Liao *et al.* [24]	88.4	82.1	-	-
Baseline [26]	87.7	85.3	34.1	50.3
Final model (ours)	**91.2**	**89.0**	**41.8**	**55.5**

Discussion