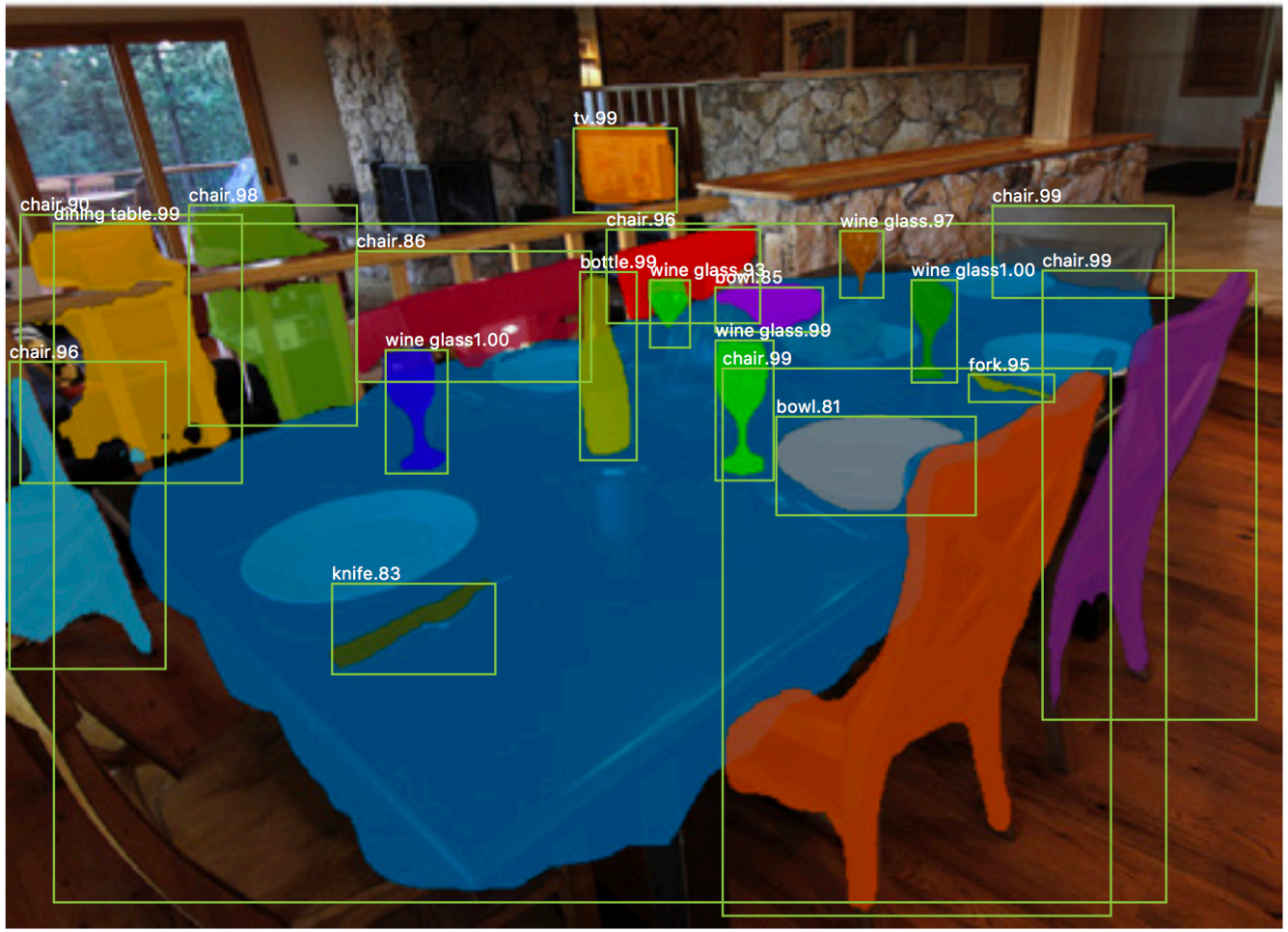
Scene Graphs for Recognition and Generation

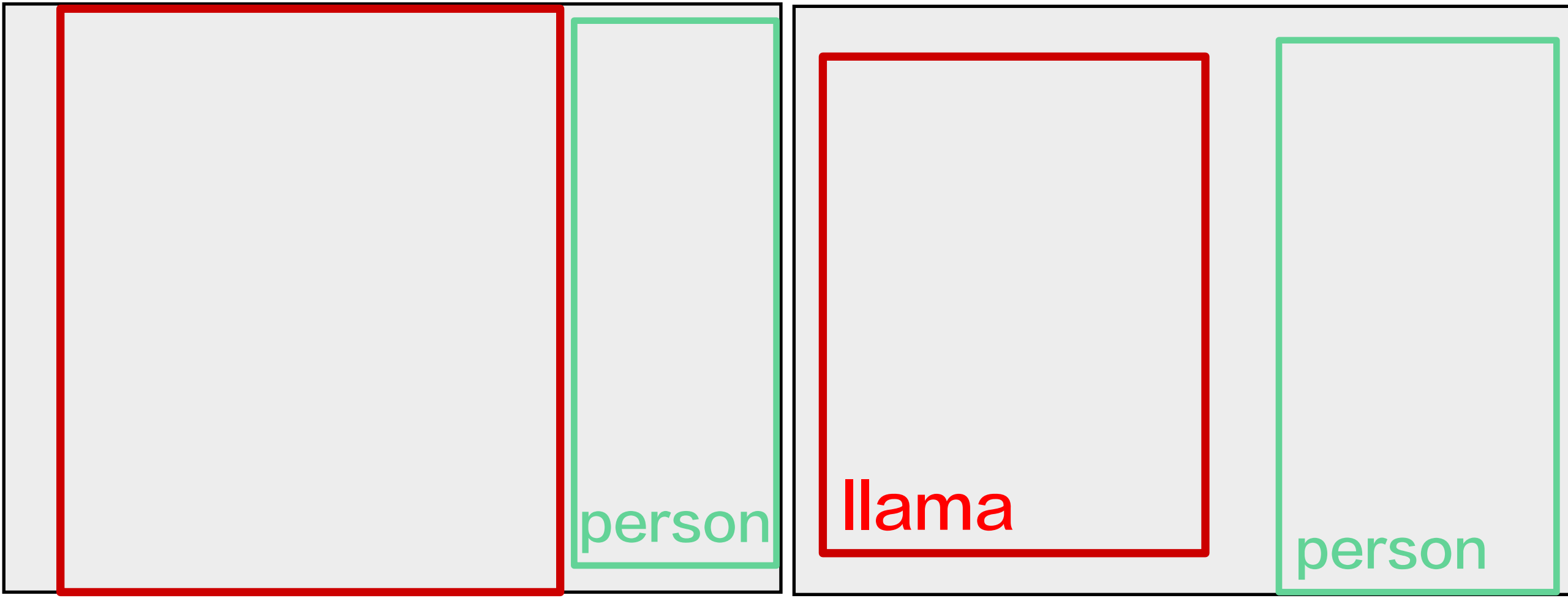
Justin Johnson

6/22/2018

Object Detection: Image -> Objects

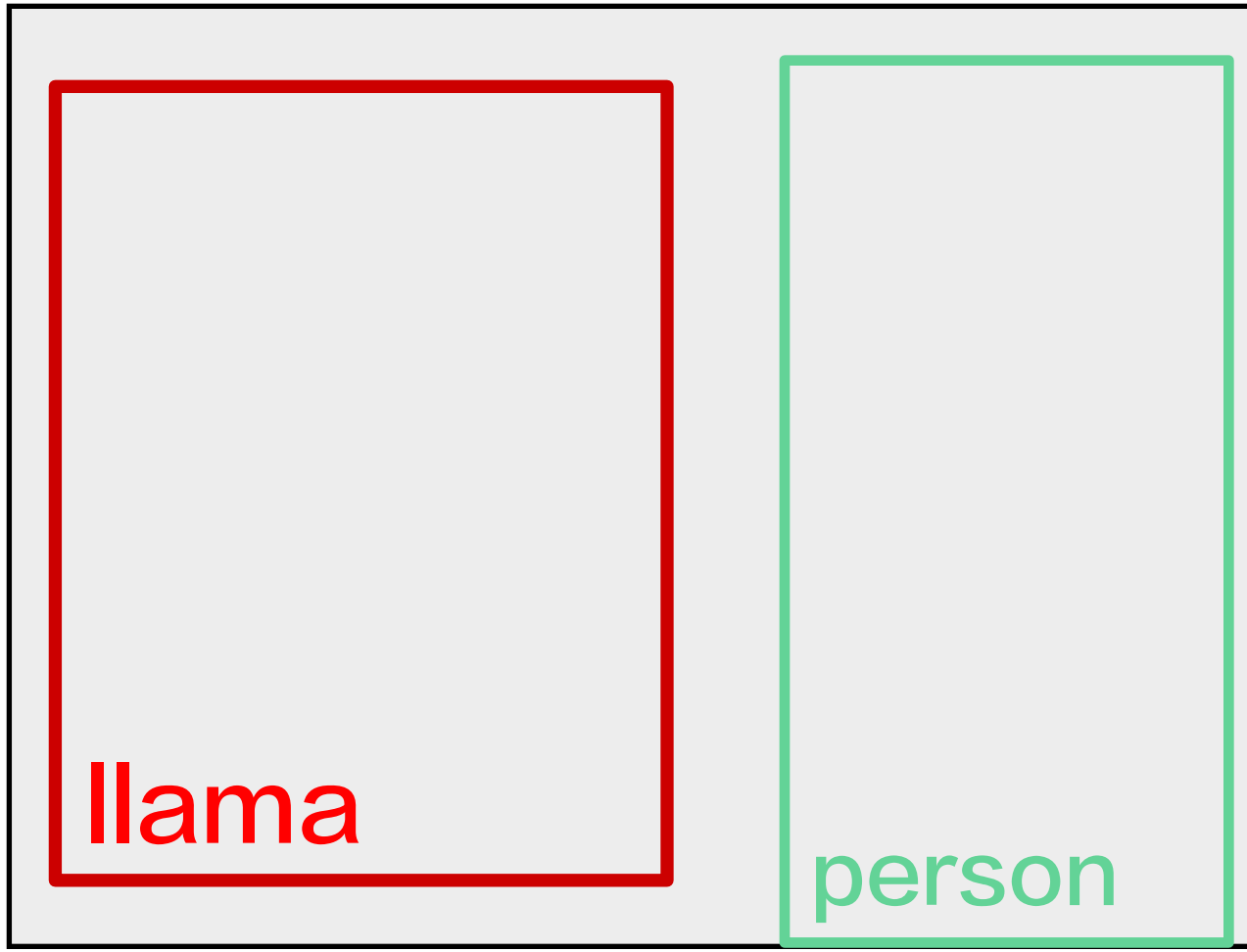


He et al, “Mask R-CNN”, ICCV 2017



Slide credit: Ranjay Krishna

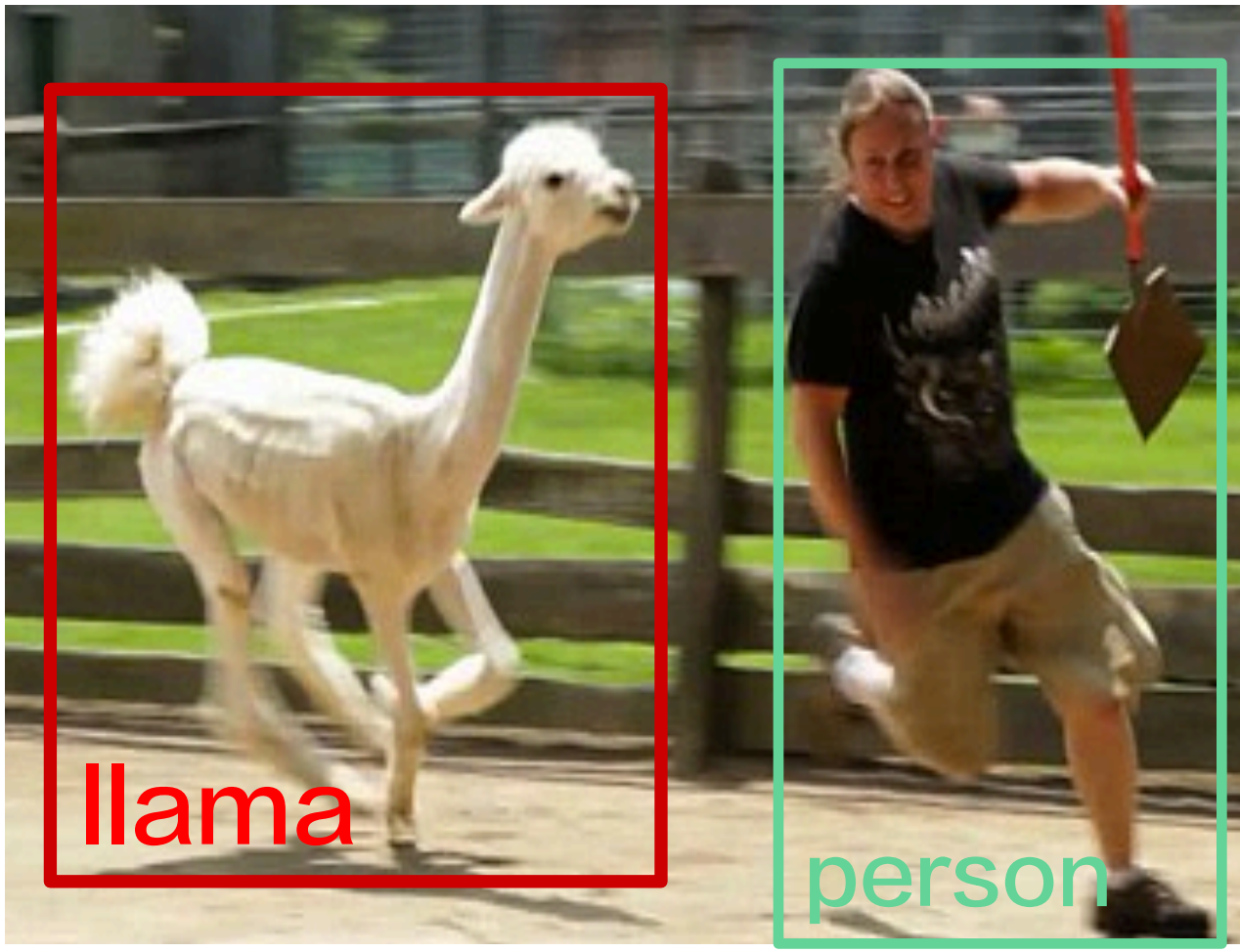
Llama next to person



Slide credit: Ranjay Krishna

Llama next to person

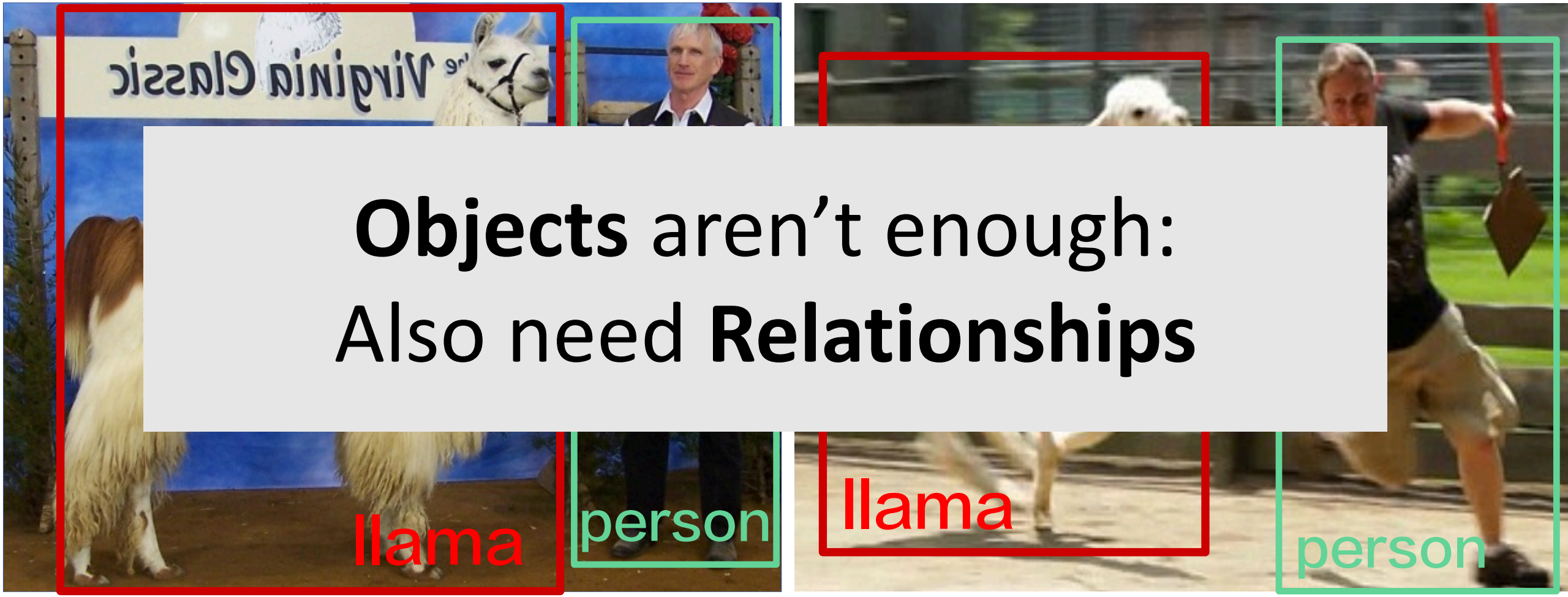
Llama chasing person



Slide credit: Ranjay Krishna

Llama next to person

Llama chasing person

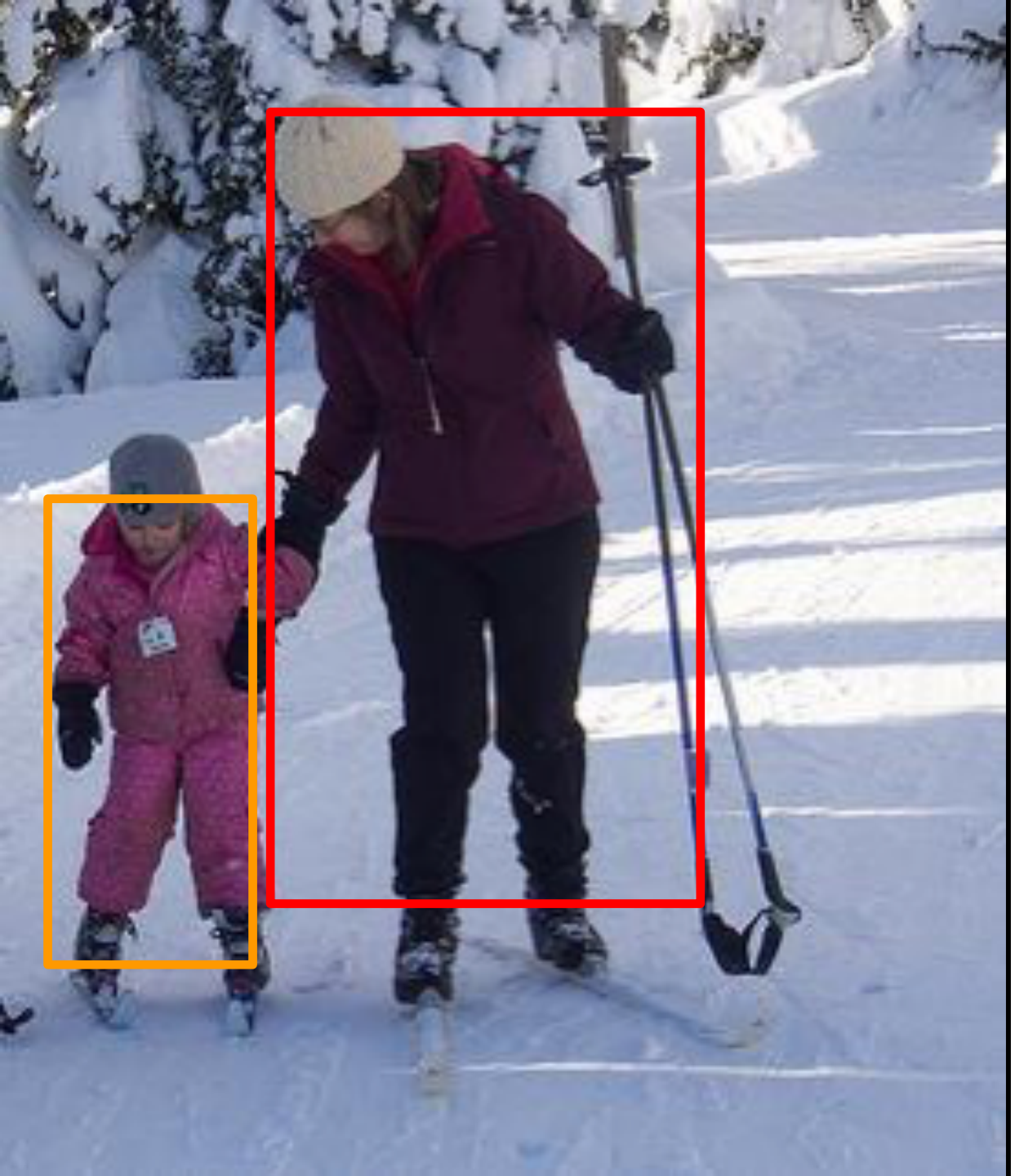


Objects aren't enough:
Also need **Relationships**

Slide credit: Ranjay Krishna

Relationships

Krishna*, Lu*, Bernstein, Fei-Fei, *ECCV* 2016

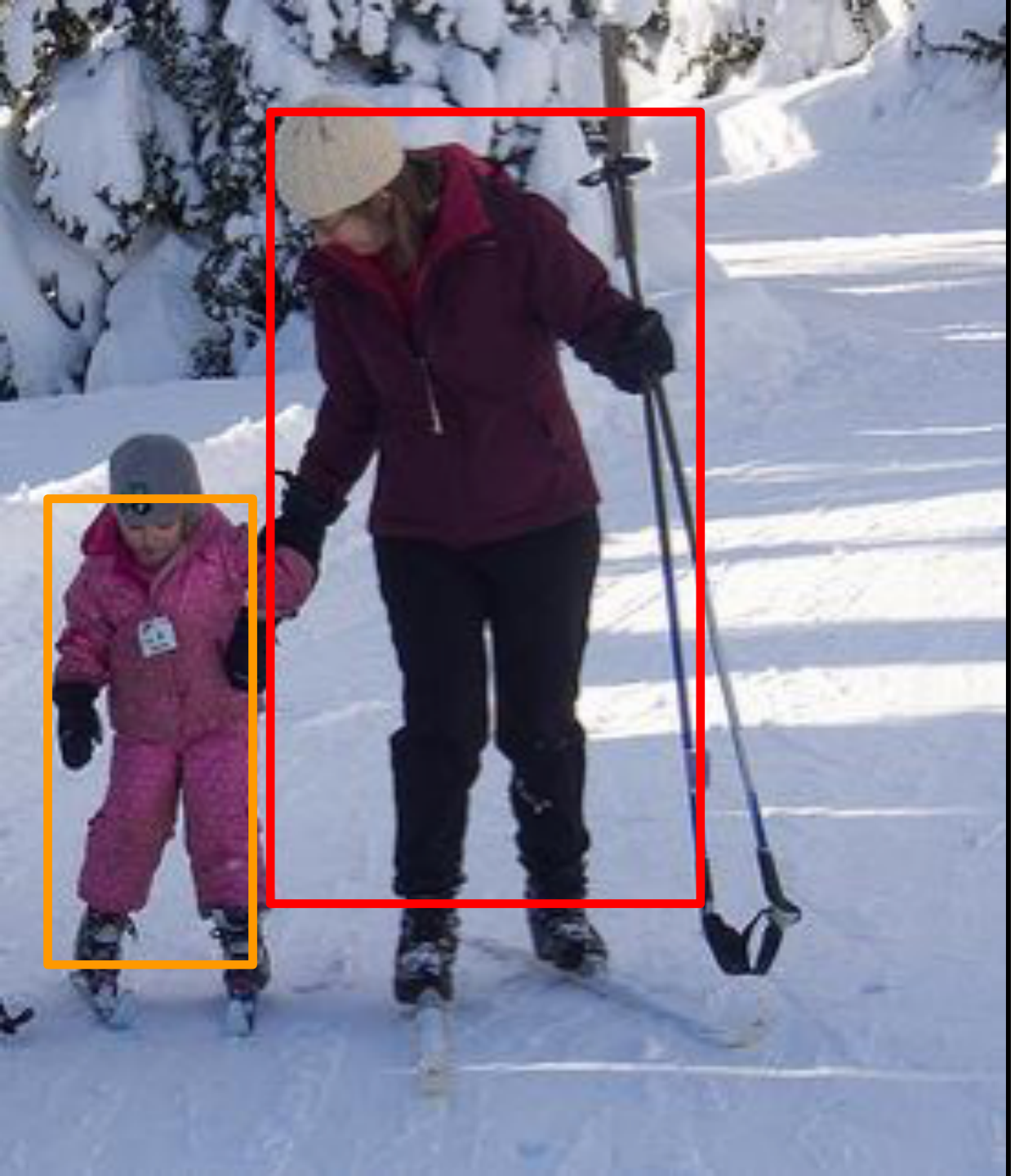


Relationships:

spatial, comparative,
verb, prepositional

person → left of → person

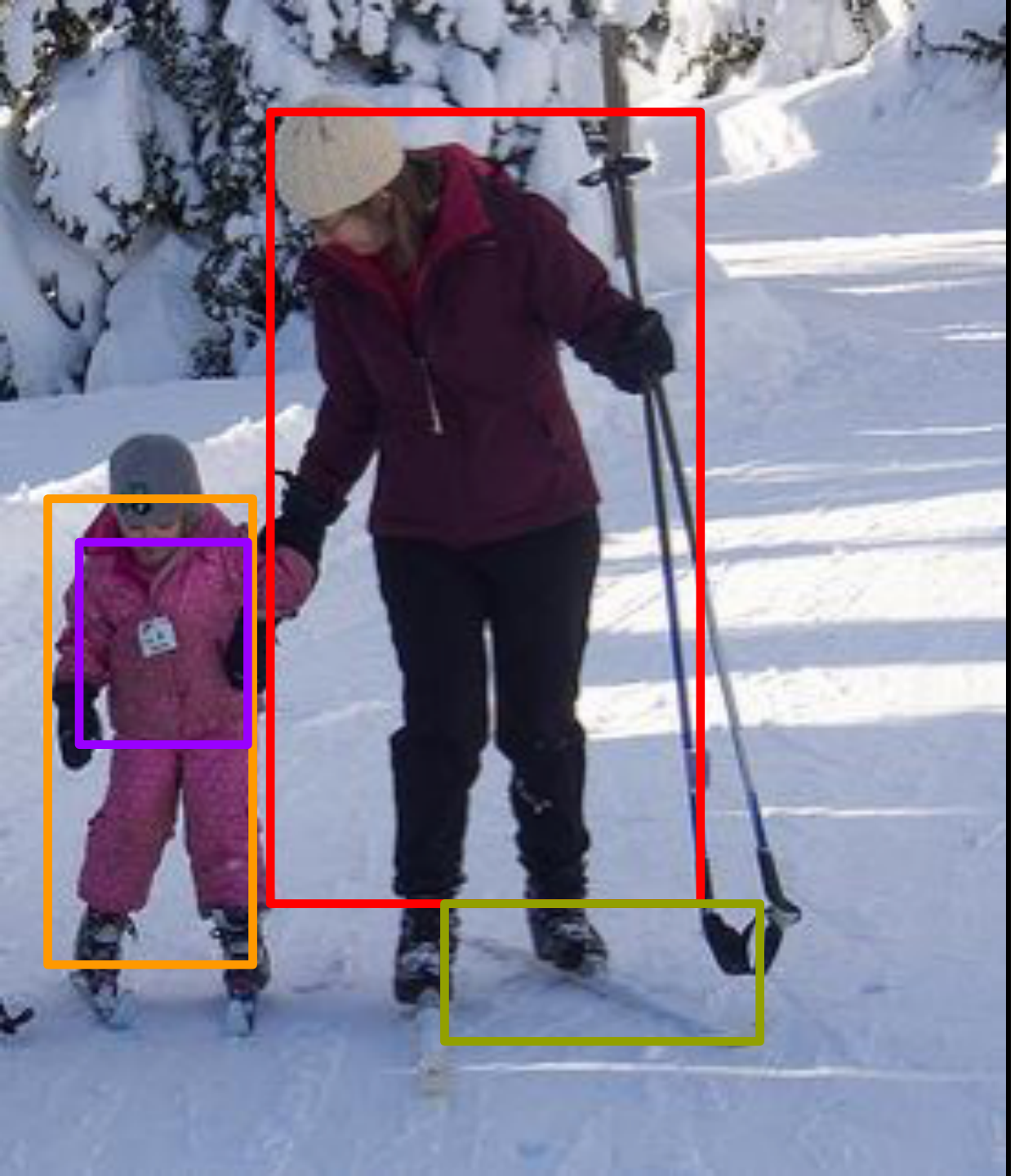
Krishna*, Lu*, Bernstein, Fei-Fei, *ECCV* 2016



Relationships:

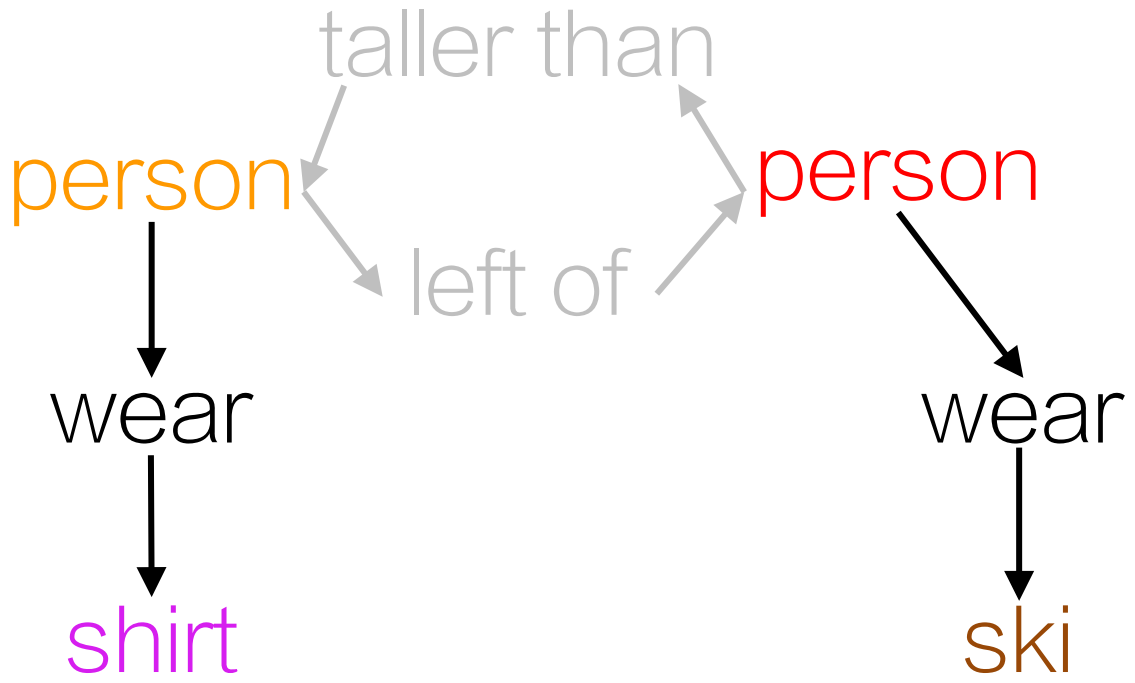
spatial, comparative,
verb, prepositional

Krishna*, Lu*, Bernstein, Fei-Fei, *ECCV* 2016

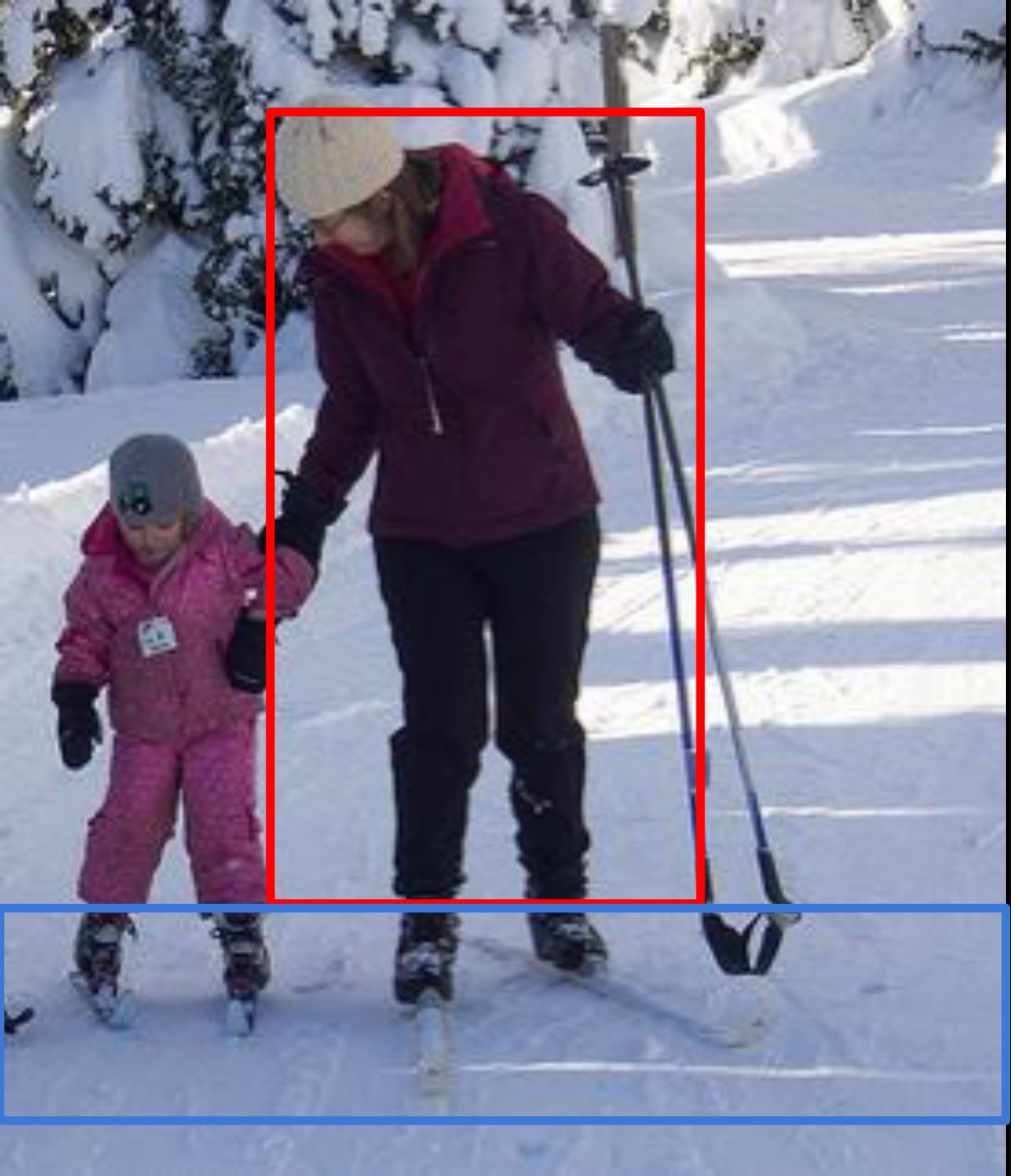


Relationships:

spatial, comparative,
verb, prepositional

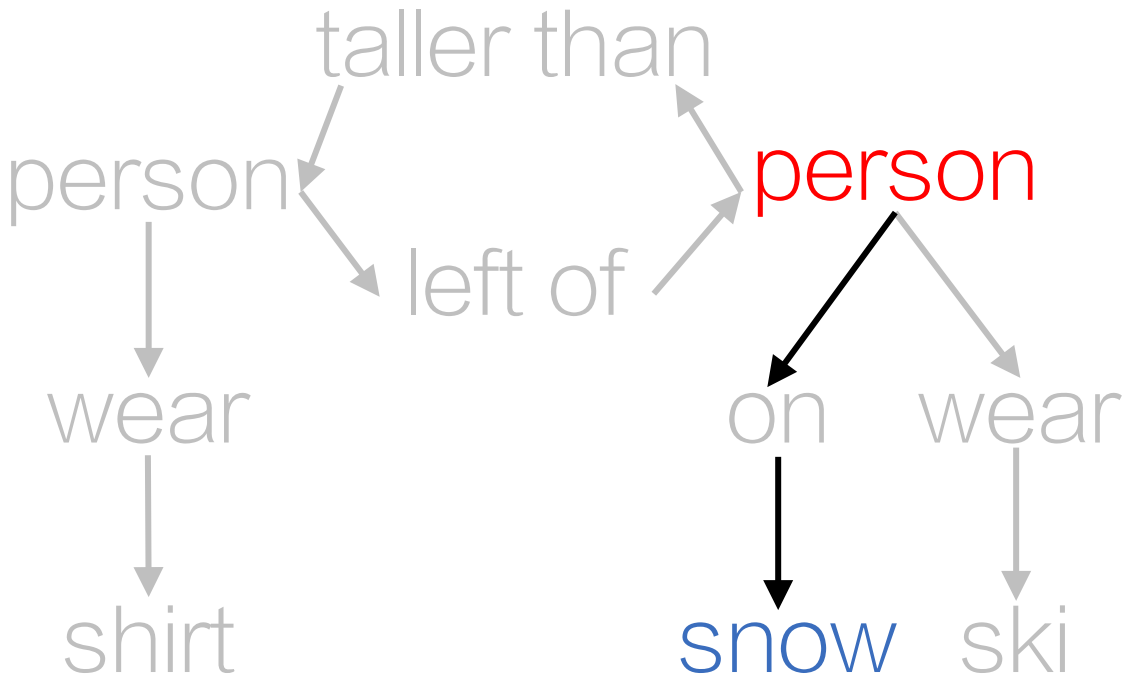


Krishna*, Lu*, Bernstein, Fei-Fei, *ECCV* 2016

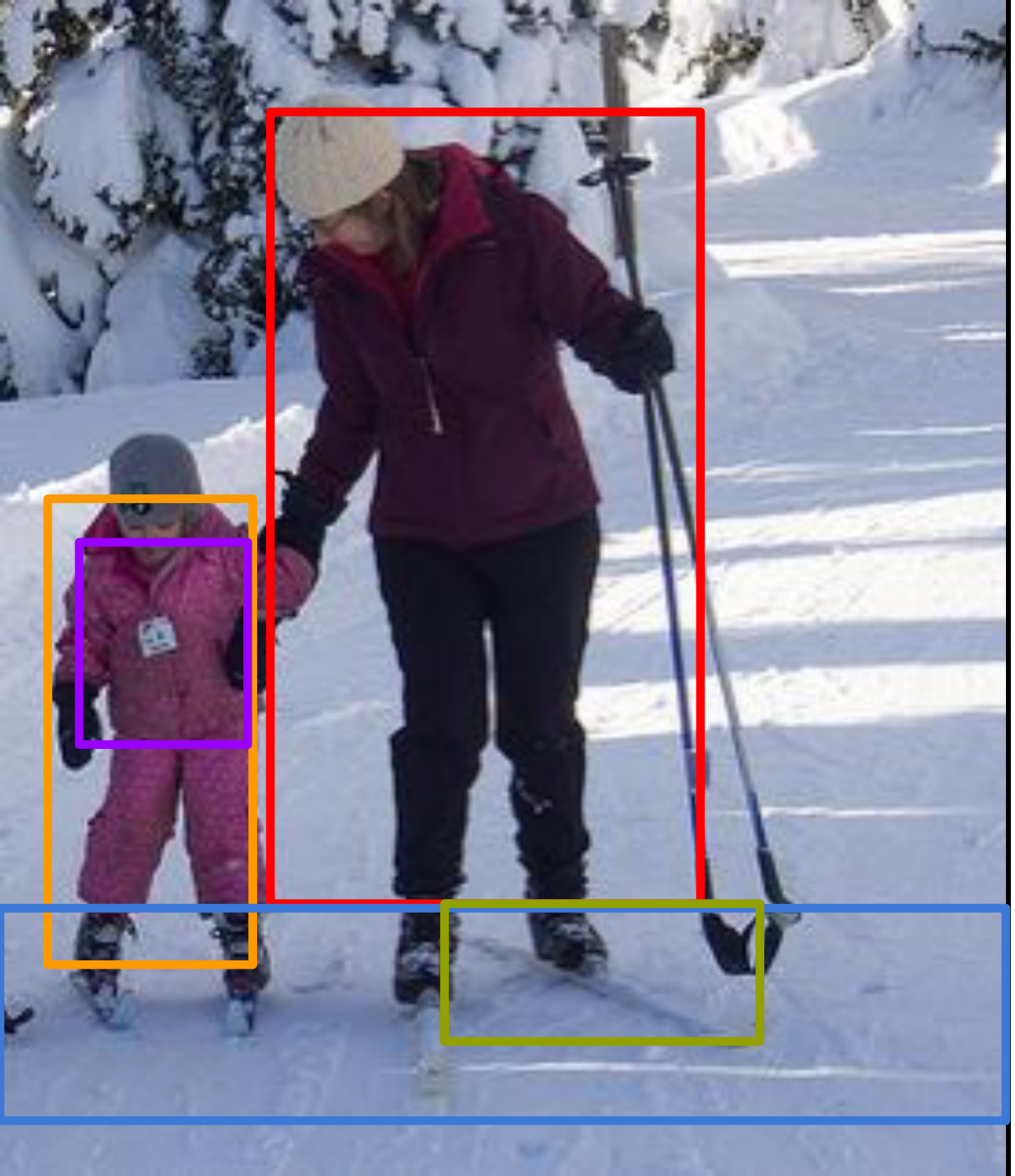


Relationships:

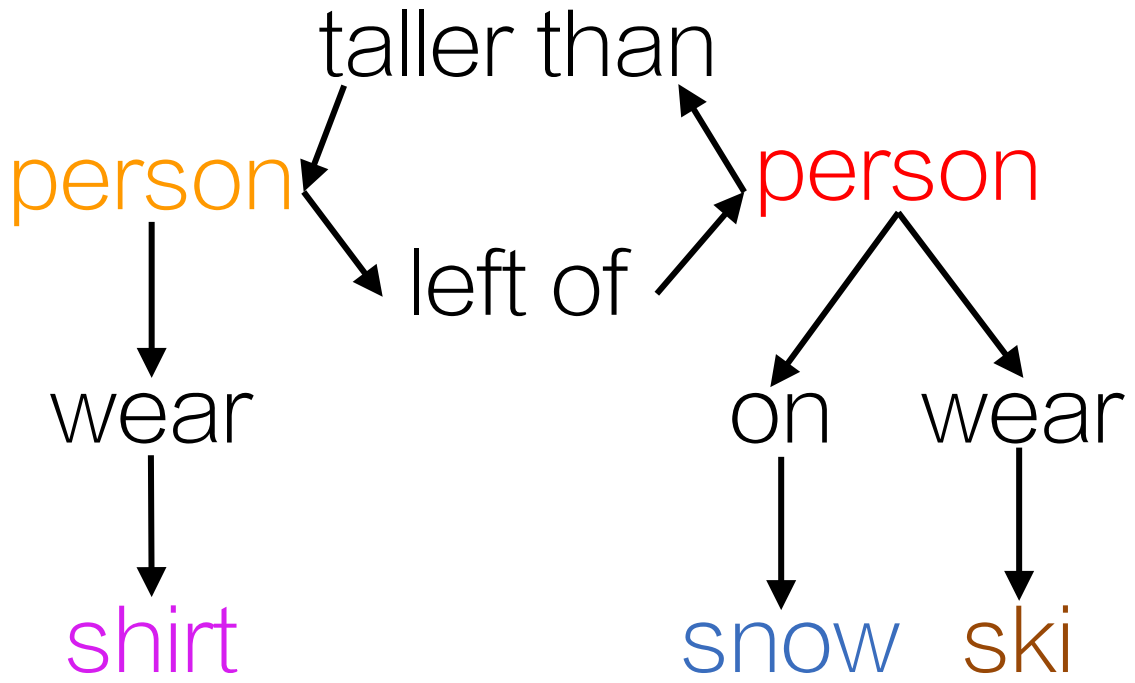
spatial, comparative,
verb, prepositional



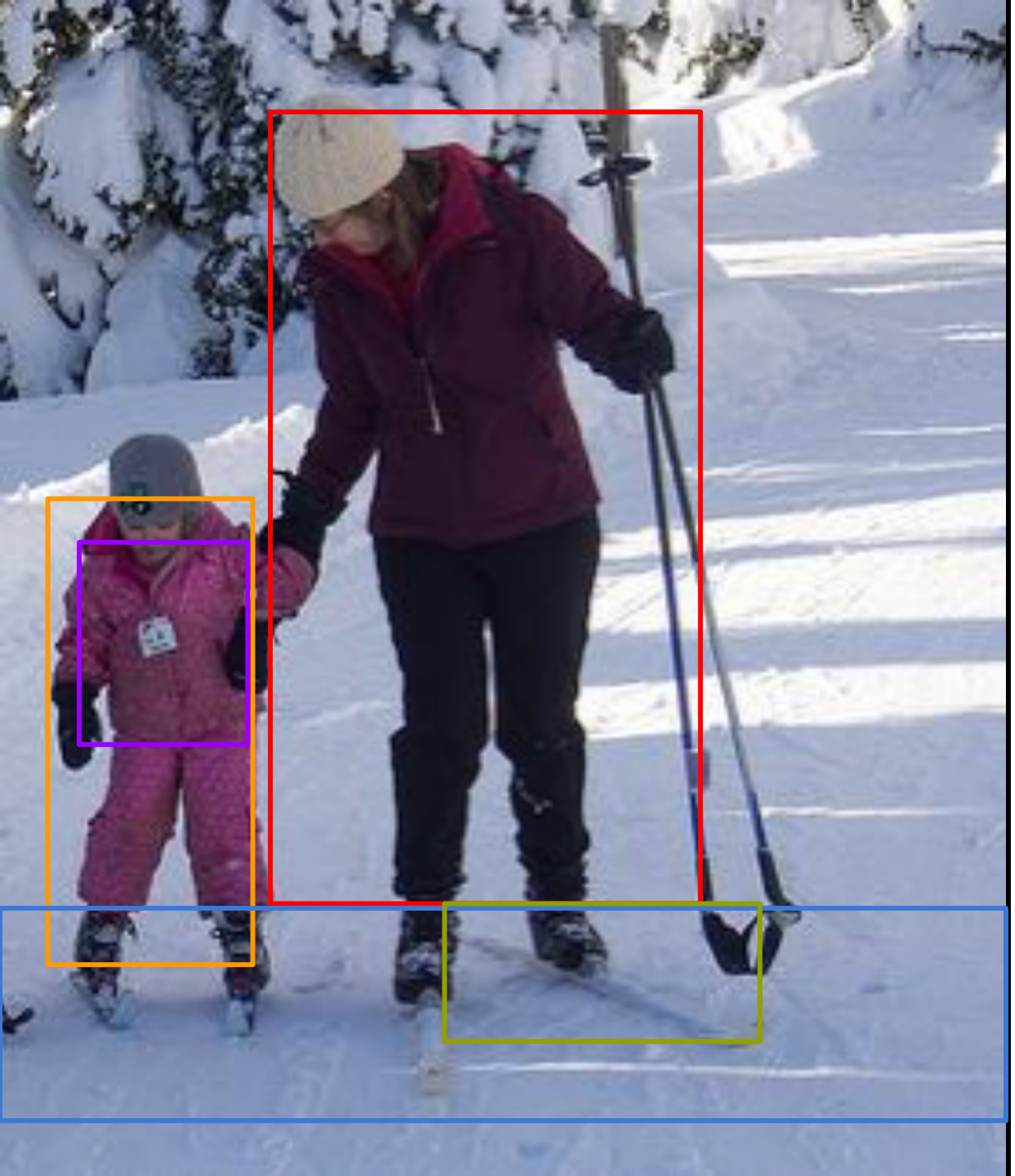
Krishna*, Lu*, Bernstein, Fei-Fei, *ECCV* 2016



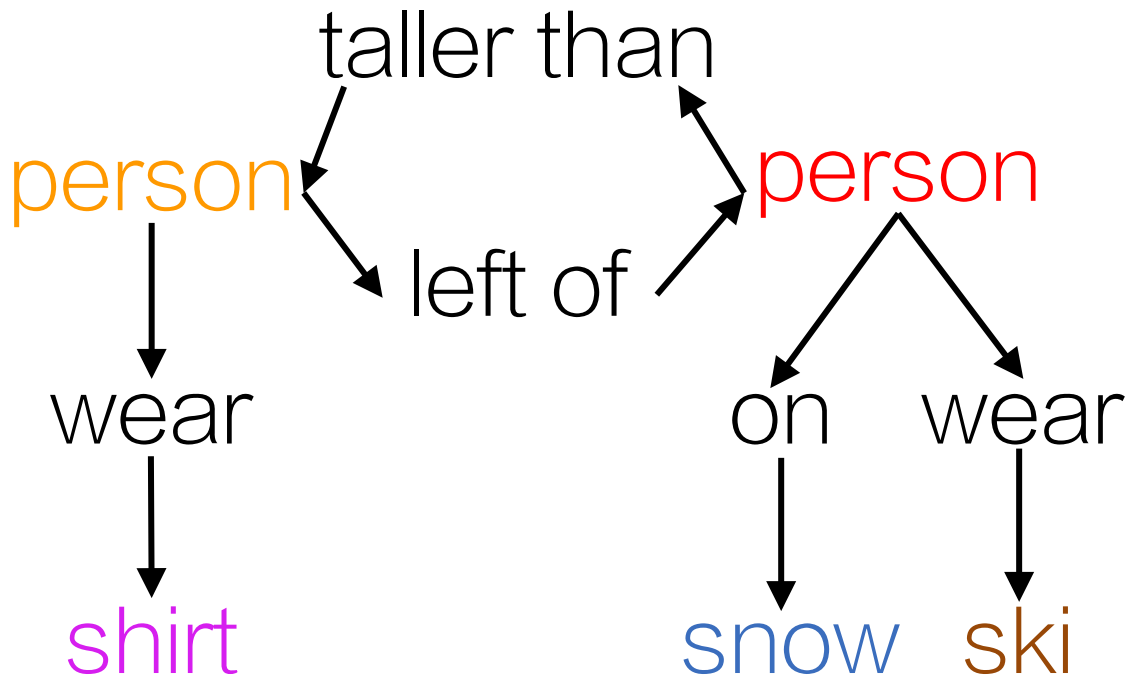
Relationships:
spatial, comparative,
verb, prepositional



Krishna*, Lu*, Bernstein, Fei-Fei, *ECCV* 2016



Objects + Relationships
= **Scene Graphs**

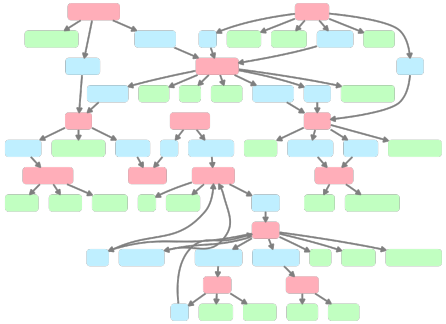


Krishna*, Lu*, Bernstein, Fei-Fei, *ECCV* 2016

Scene Graph 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

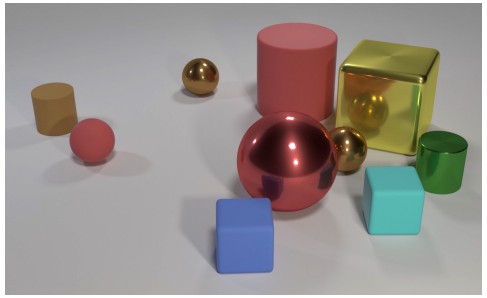
Visual Genome

Krishna et al, IJCV 2017

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

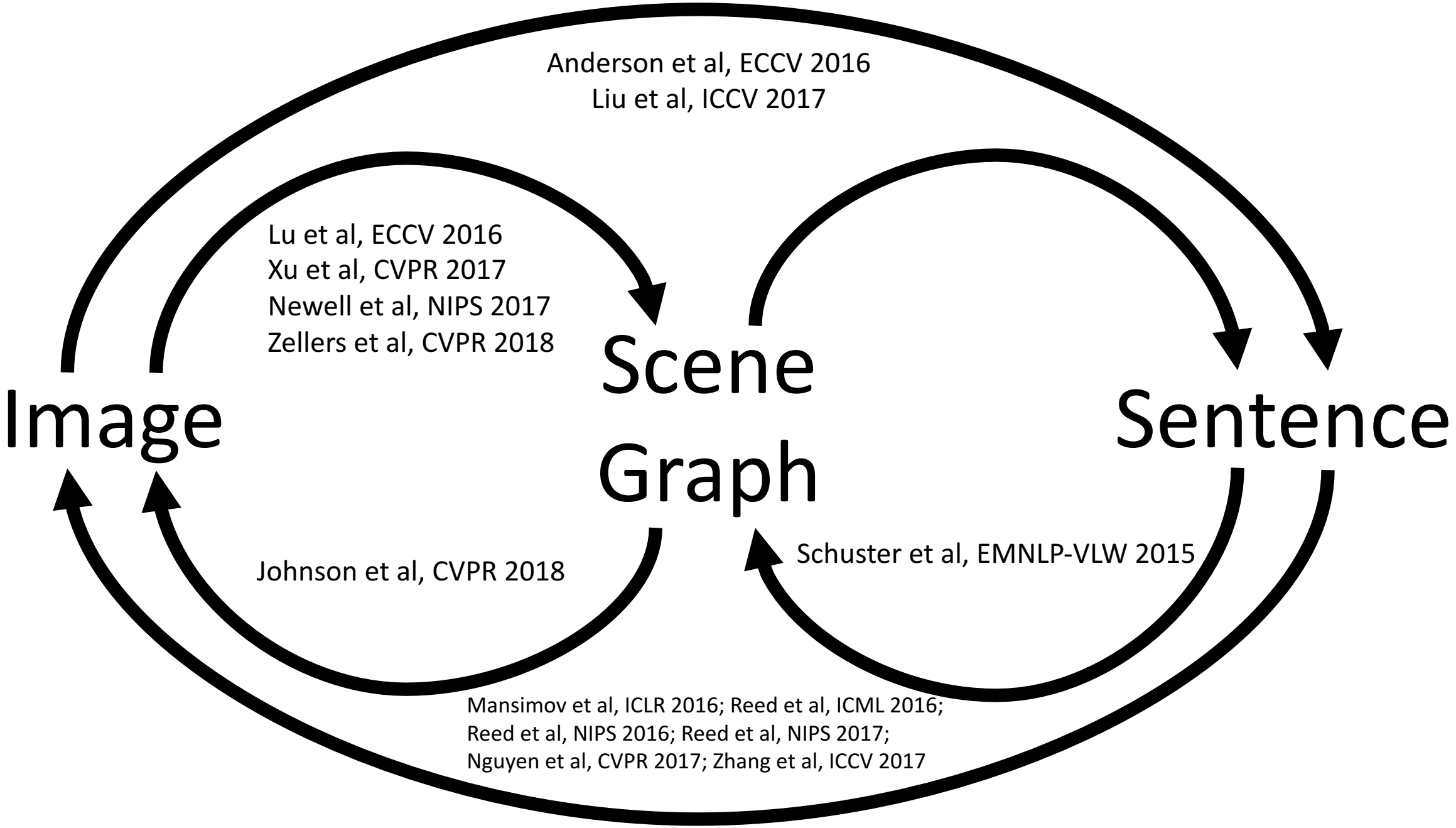
CLEVR

Johnson et al, CVPR 2017



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

What can we do with scene graphs?


Anderson et al, ECCV 2016
Liu et al, ICCV 2017
Lu et al, ECCV 2016
Xu et al, CVPR 2017
Newell et al, NIPS 2017
Zellers et al, CVPR 2018
Image
Scene Graph
Sentence
Johnson et al, CVPR 2018
Schuster et al, EMNLP-VLW 2015
Mansimov et al, ICLR 2016; Reed et al, ICML 2016;
Reed et al, NIPS 2016; Reed et al, NIPS 2017;
Nguyen et al, CVPR 2017; Zhang et al, ICCV 2017


Lu et al, ECCV 2016
Xu et al, CVPR 2017
Newell et al, NIPS 2017
Zellers et al, CVPR 2018
Image
Scene Graph

Image to Scene Graph

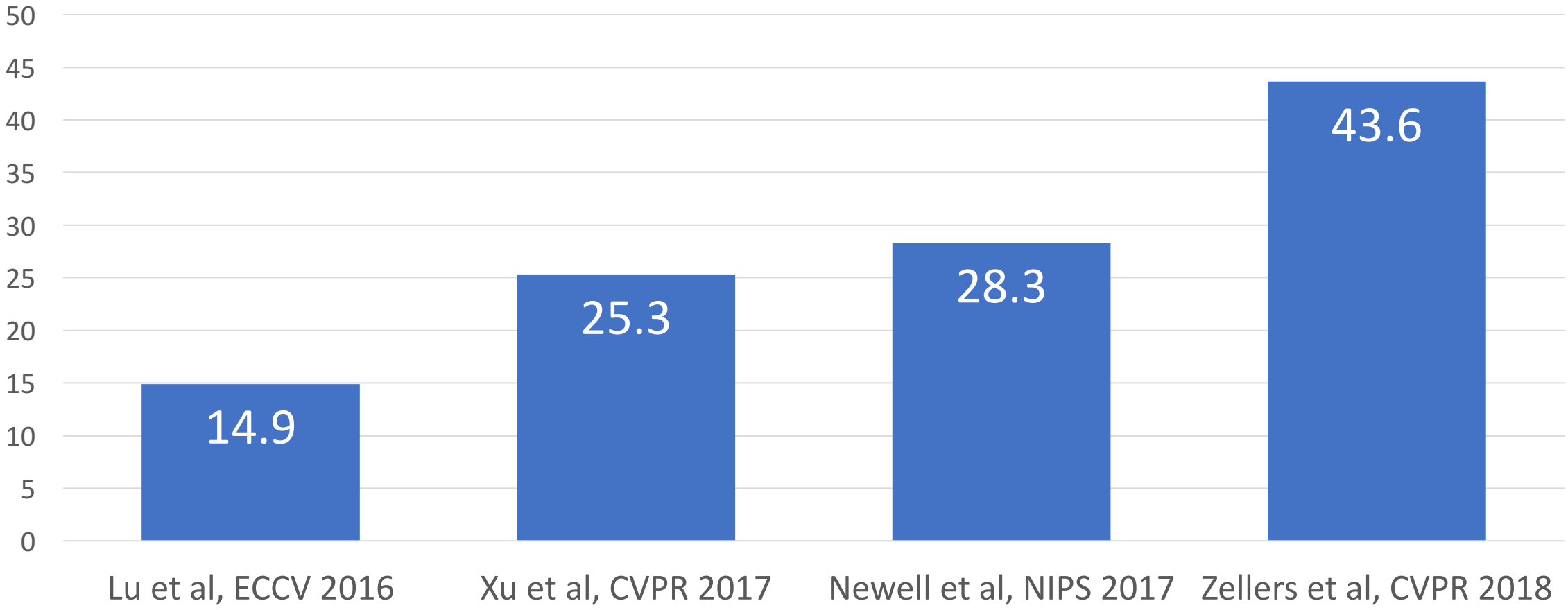
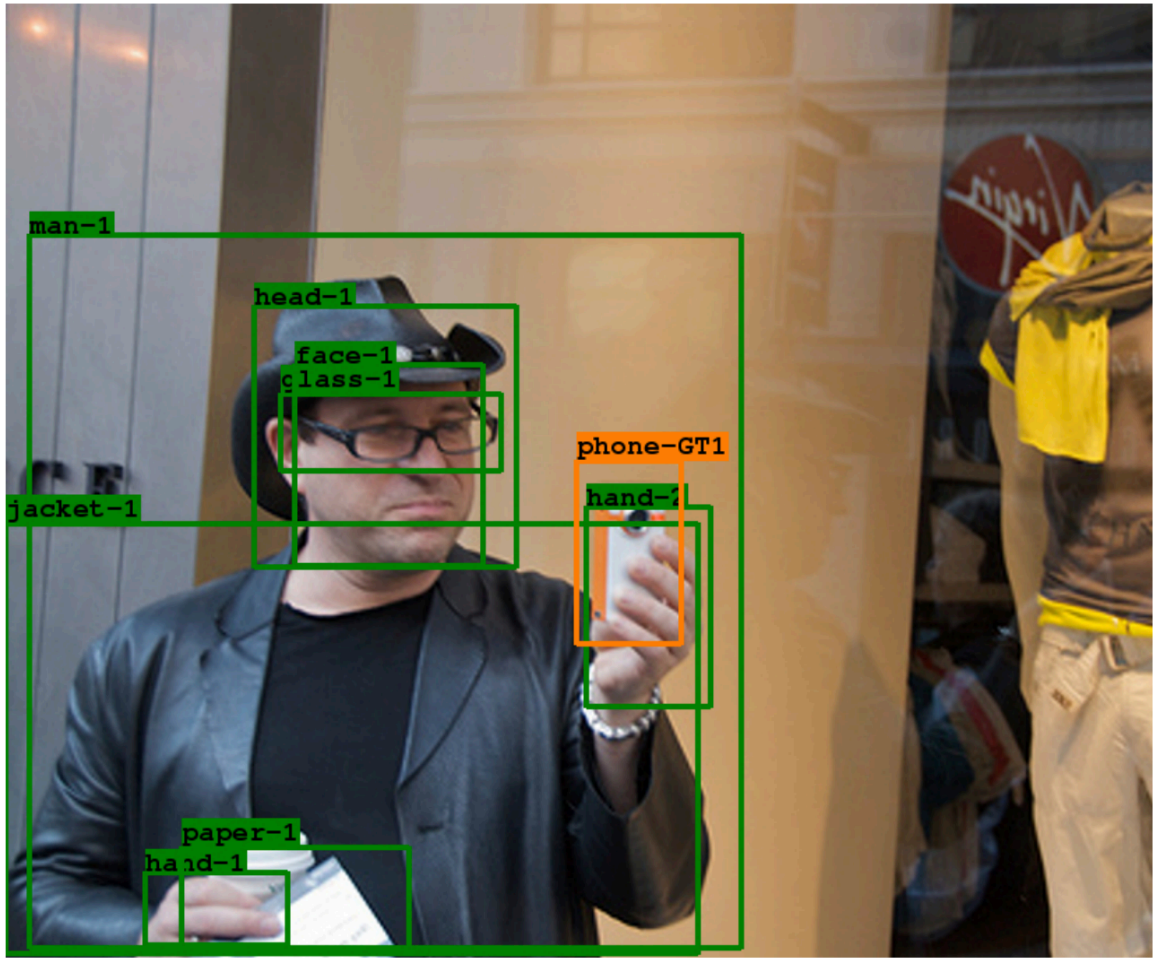
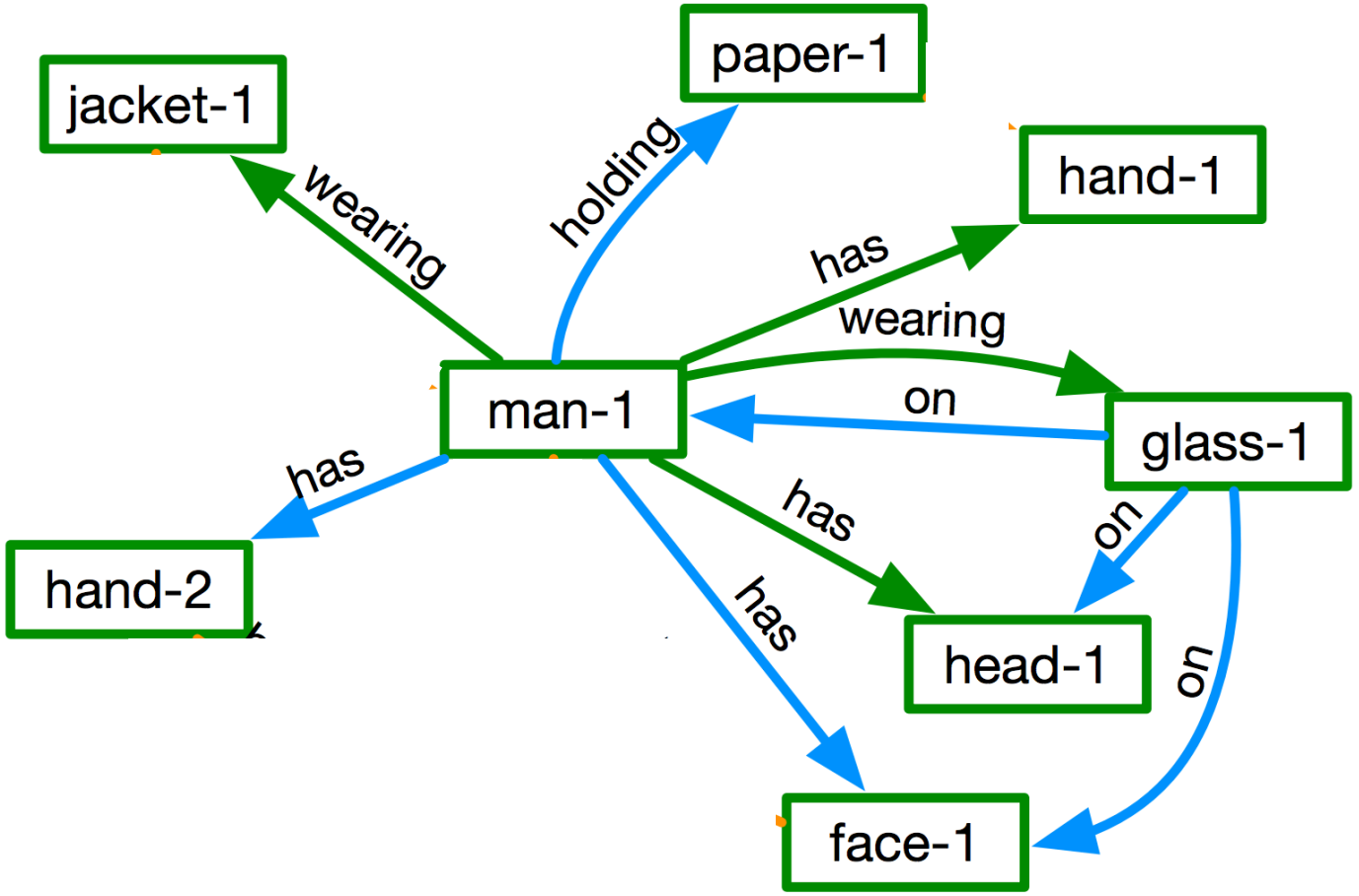


Image to Scene Graph

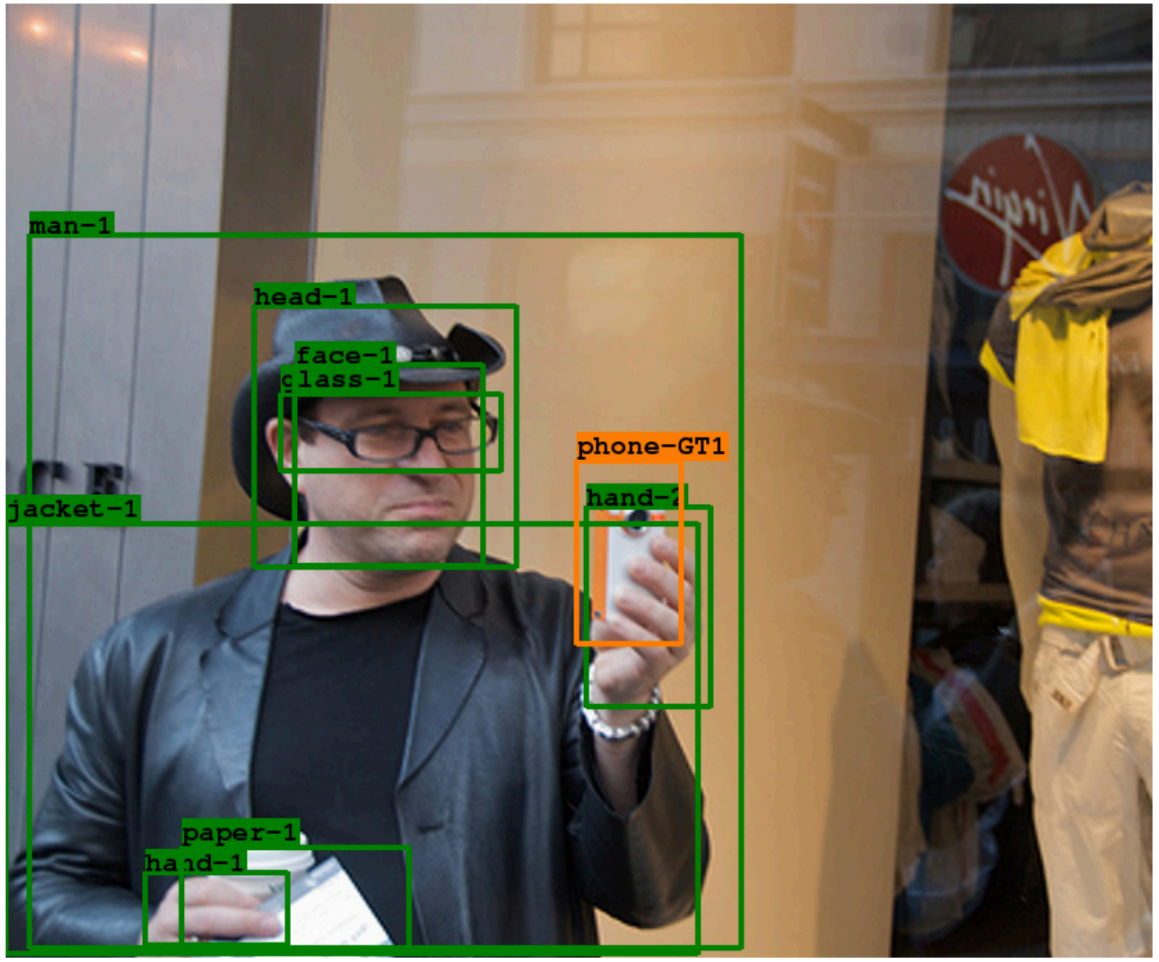


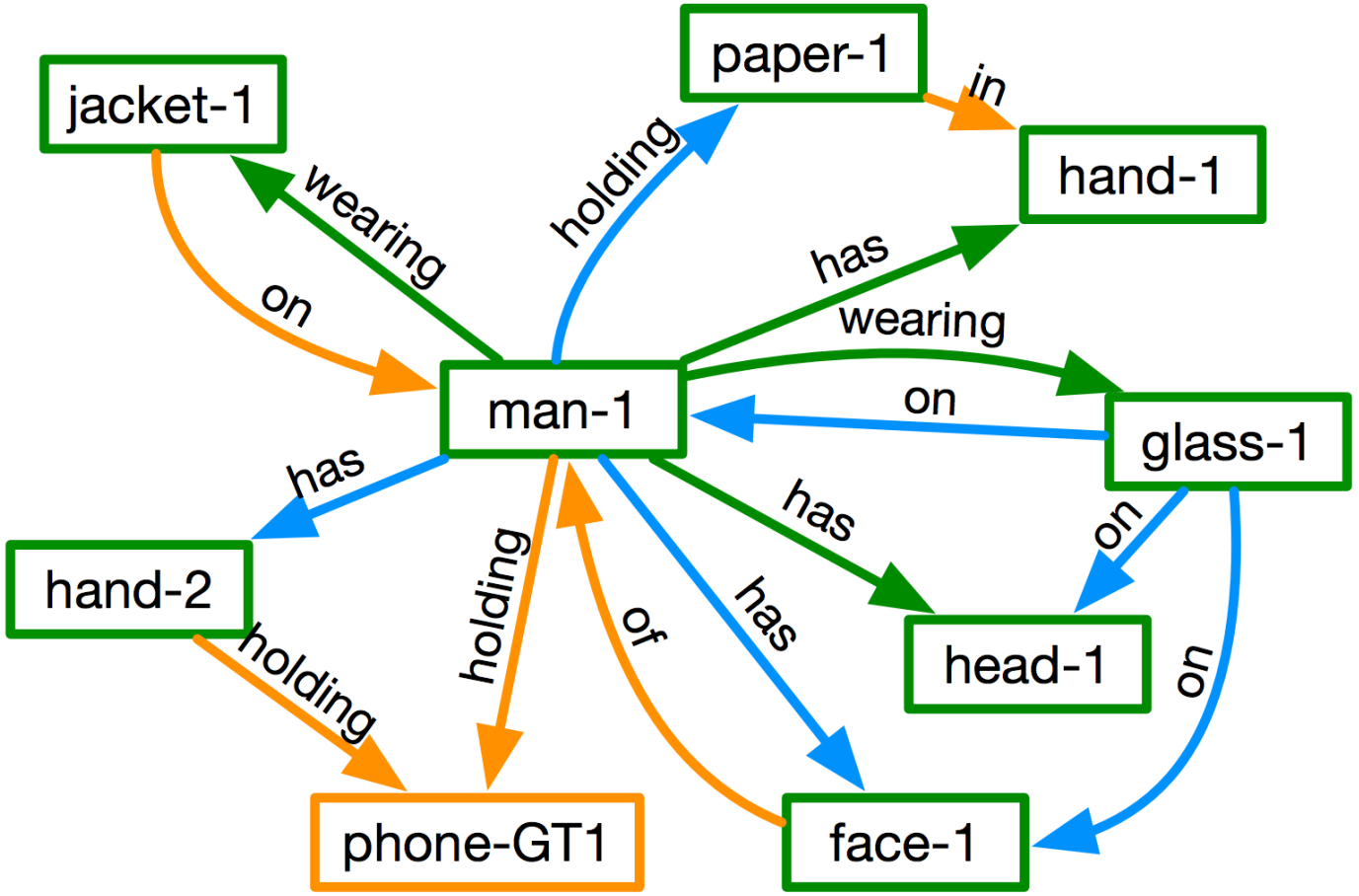


True positive **False Positive**

Zellers et al, CVPR 2018

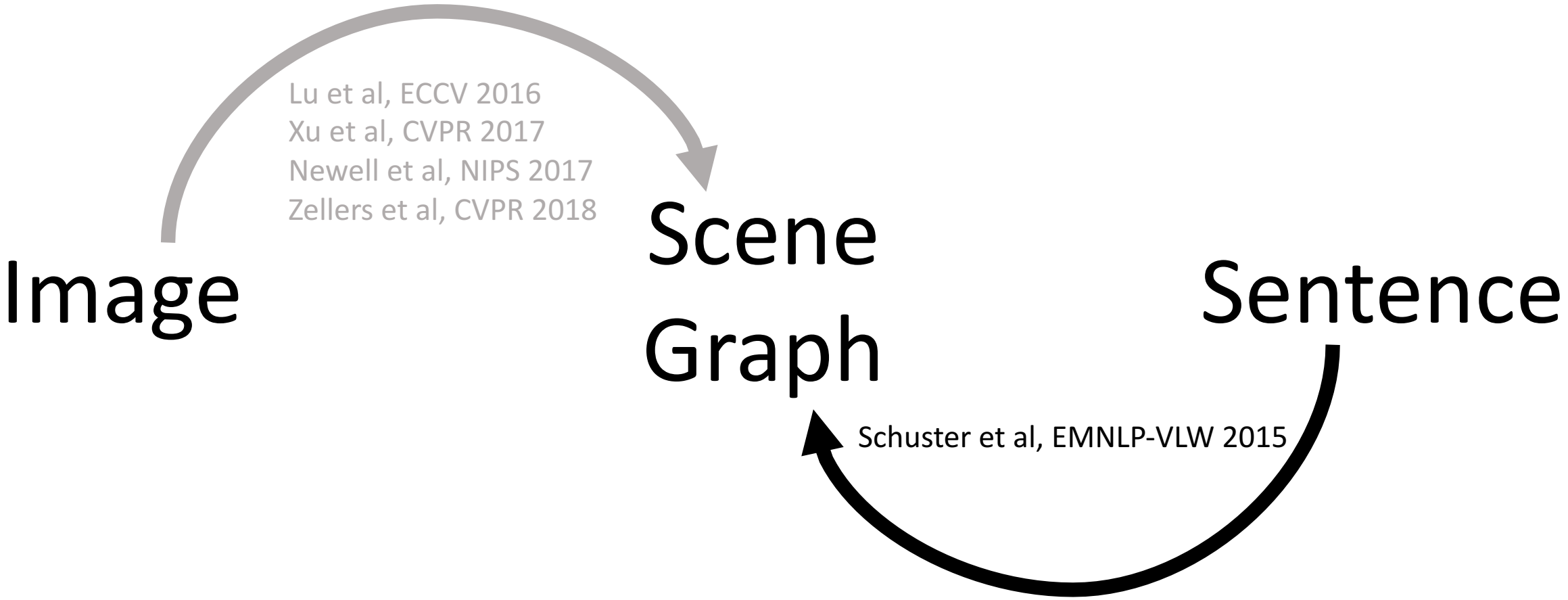
Image to Scene Graph





True positive **False Positive** **False Negative**

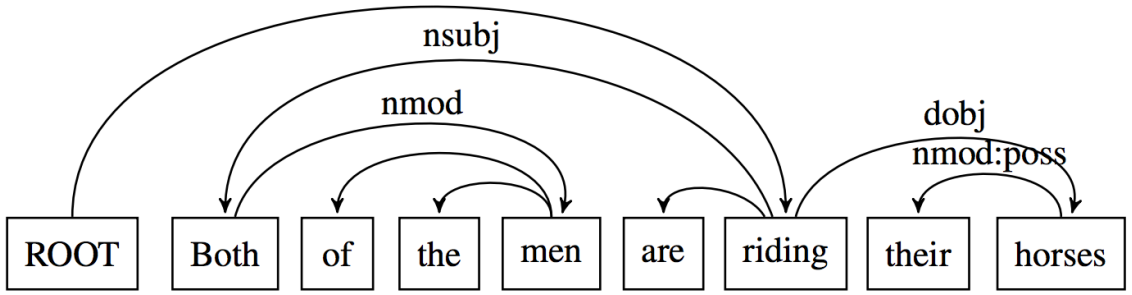
Zellers et al, CVPR 2018


Lu et al, ECCV 2016
Xu et al, CVPR 2017
Newell et al, NIPS 2017
Zellers et al, CVPR 2018
Image
Scene Graph
Sentence
Schuster et al, EMNLP-VLW 2015

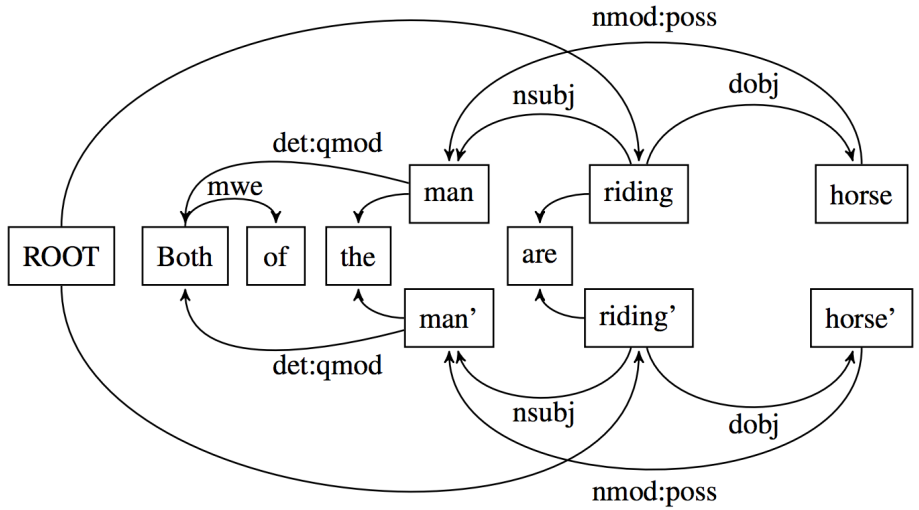
Sentence to Scene Graph

Both of the men are riding their horses

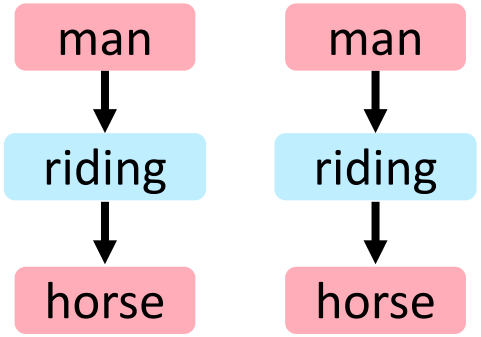
Dependency parser



Tree transforms



Heuristic rules



Schuster et al, EMNLP-VLW 2015

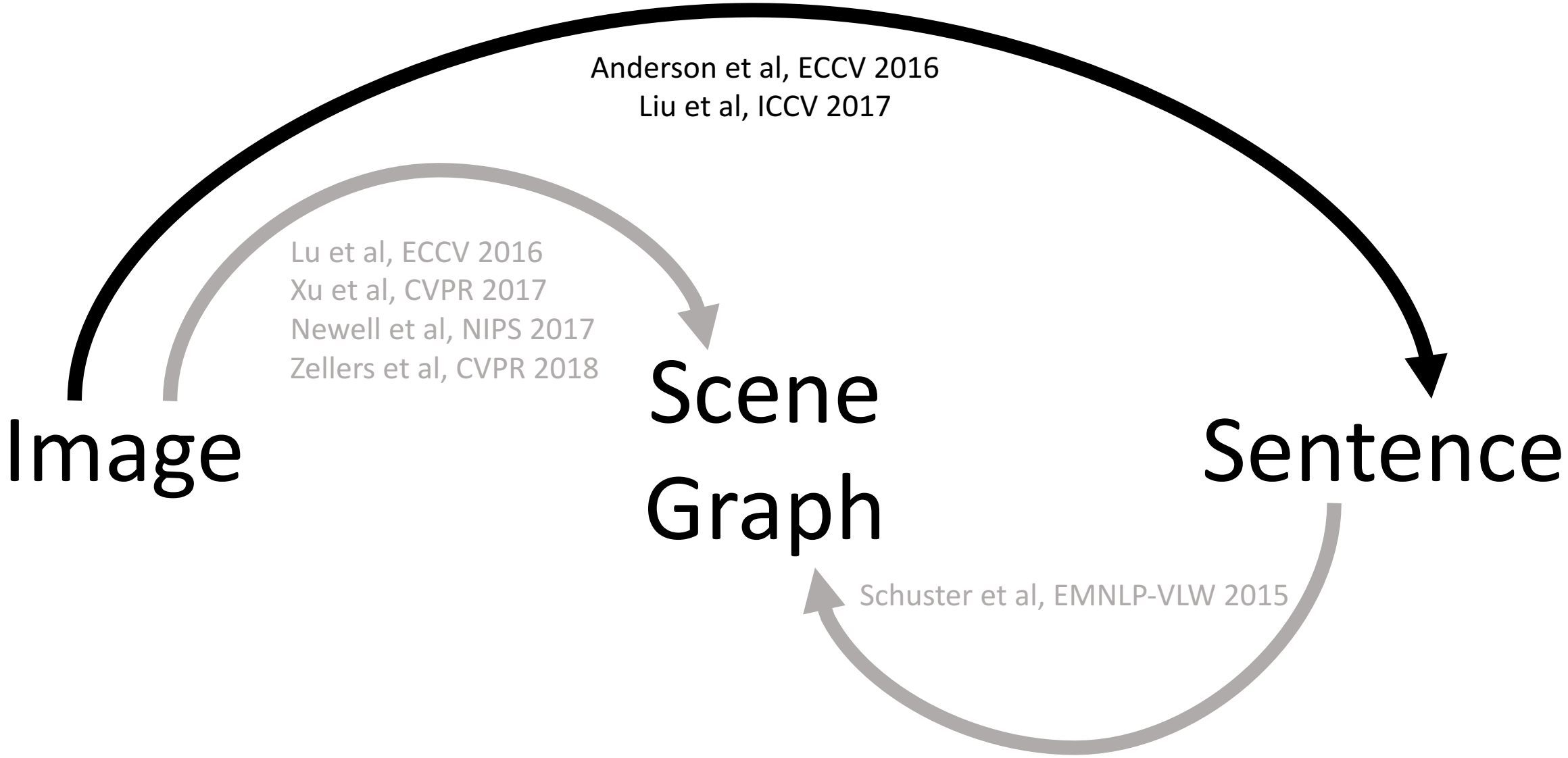

Anderson et al, ECCV 2016
Liu et al, ICCV 2017
Lu et al, ECCV 2016
Xu et al, CVPR 2017
Newell et al, NIPS 2017
Zellers et al, CVPR 2018
Image
Scene Graph
Sentence
Schuster et al, EMNLP-VLW 2015

Image Captioning (with Scene Graphs)
SPICE: Semantic Propositional Image Caption Evaluation

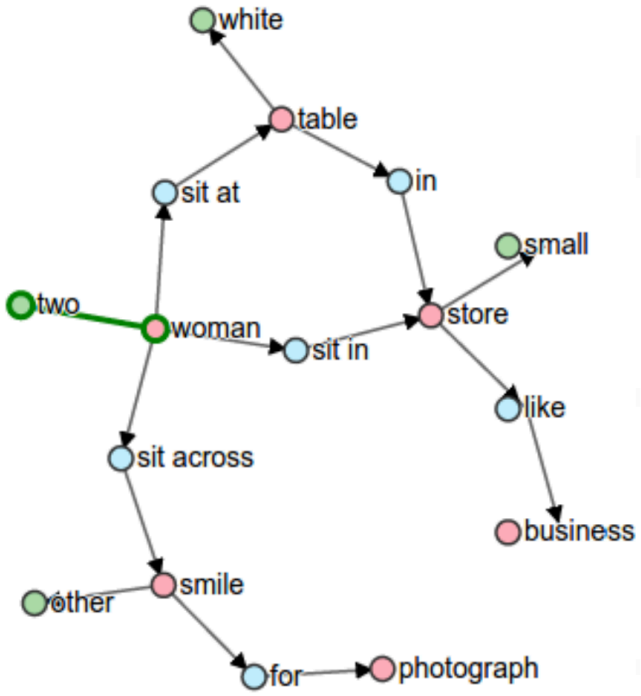
"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

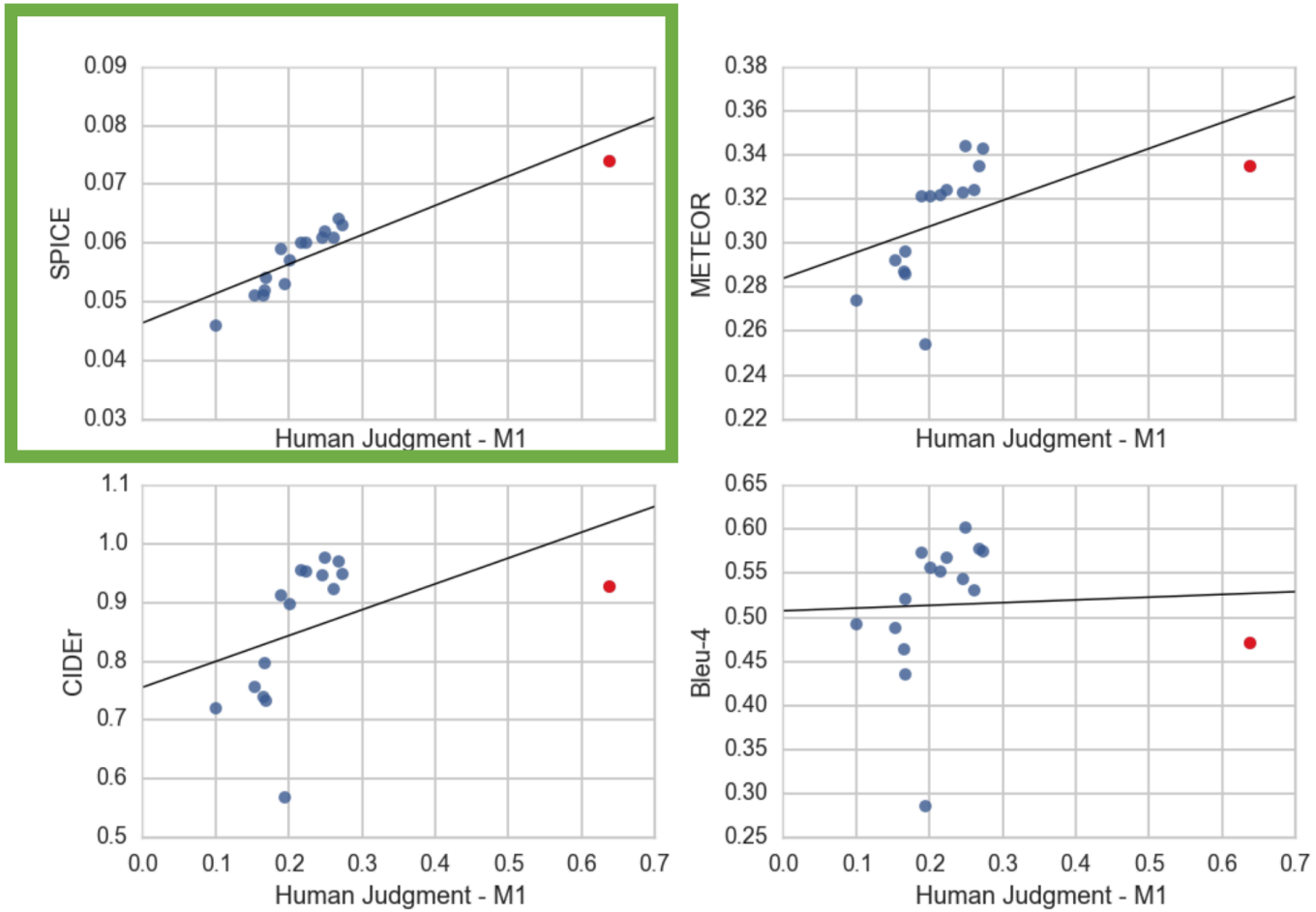
"two woman are sitting at a table"



Evaluate captioning by comparing **scene graphs** instead of comparing **sentences**

Anderson et al, ECCV 2016

Image Captioning (with Scene Graphs)
SPICE: Semantic Propositional Image Caption Evaluation

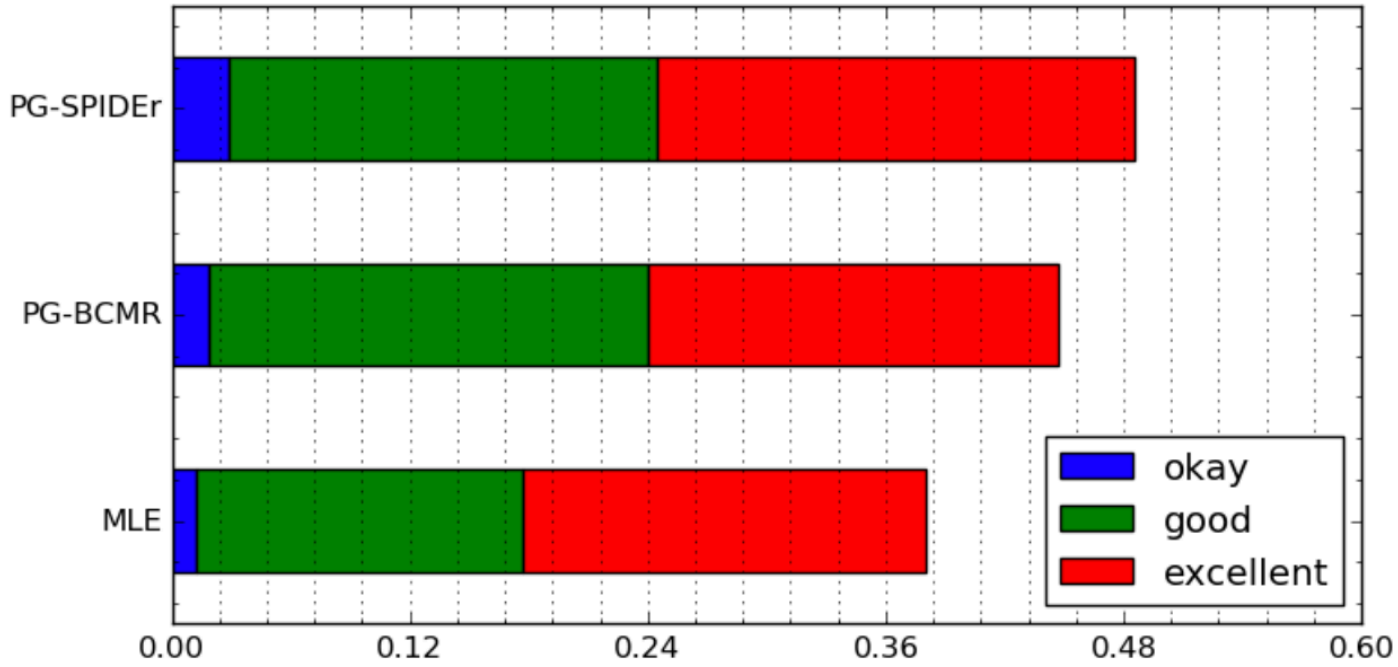


Better correlation with human judgement

Anderson et al, ECCV 2016

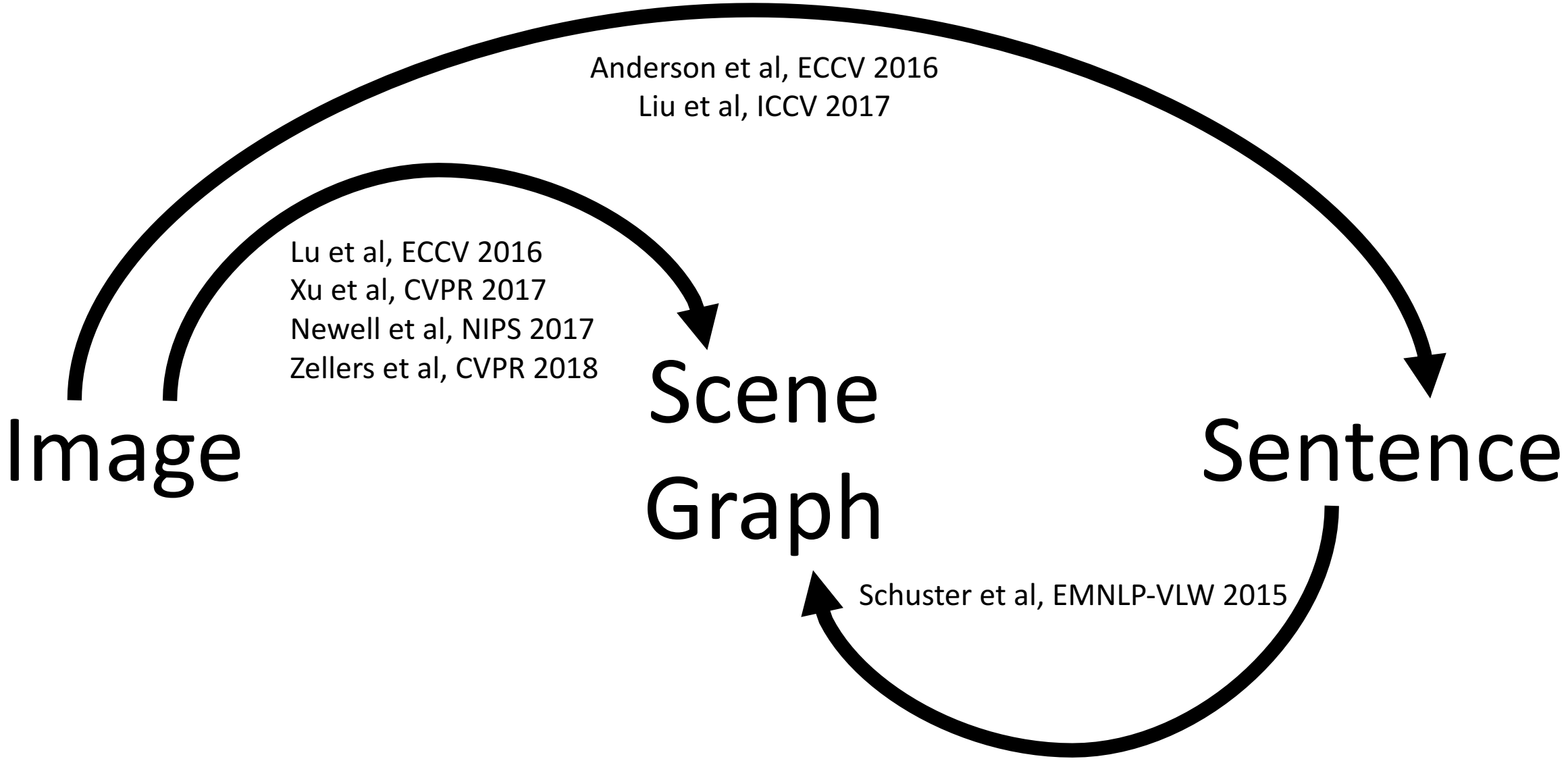
Image Captioning (with Scene Graphs)

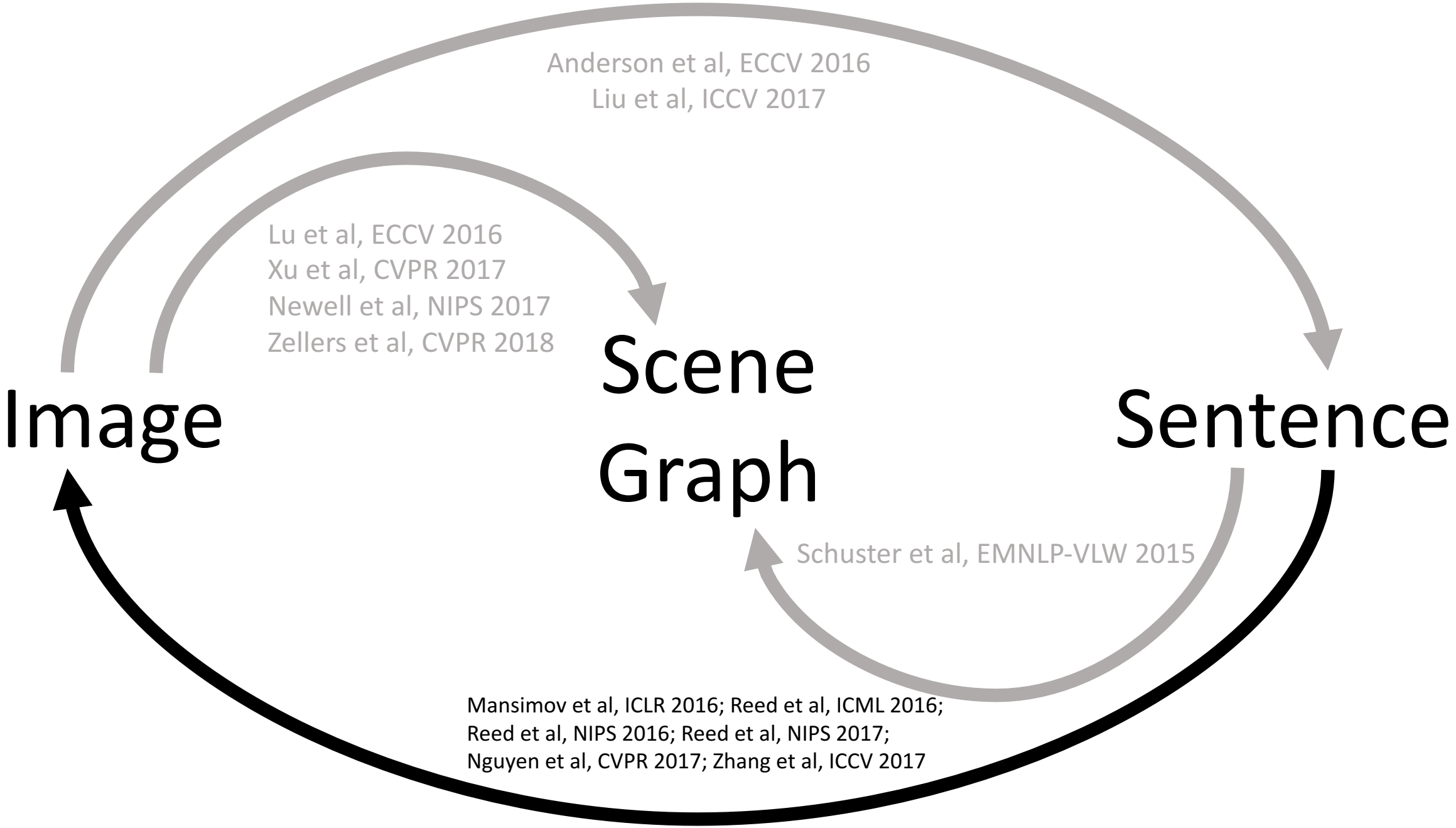
Idea: Train captioning model to maximize
SPIDEr = SPICE + CIDEr using policy gradient



Result: People like the generated captions better!

Liu et al, ICCV 2017


Anderson et al, ECCV 2016
Liu et al, ICCV 2017
Lu et al, ECCV 2016
Xu et al, CVPR 2017
Newell et al, NIPS 2017
Zellers et al, CVPR 2018
Image
Scene Graph
Sentence
Schuster et al, EMNLP-VLW 2015


Anderson et al, ECCV 2016
Liu et al, ICCV 2017
Lu et al, ECCV 2016
Xu et al, CVPR 2017
Newell et al, NIPS 2017
Zellers et al, CVPR 2018
Image
Scene Graph
Sentence
Schuster et al, EMNLP-VLW 2015
Mansimov et al, ICLR 2016; Reed et al, ICML 2016;
Reed et al, NIPS 2016; Reed et al, NIPS 2017;
Nguyen et al, CVPR 2017; Zhang et al, ICCV 2017

Text to Image

A red bird with black wings

Mansimov et al, ICLR 2016; Reed et al, ICML 2016; Reed et al, NIPS 2016; Reed et al, NIPS 2017; Nguyen et al, CVPR 2017; Zhang et al, ICCV 2017

Text to Image

A red bird with black wings

Prior work:
Short description, single object

A man in a purple shirt throwing a red frisbee over a boy in a blue shirt to a man wearing jeans in the backyard

My goal:
Complex description, many objects and relationships

Johnson et al, CVPR 2018

Mansimov et al, ICLR 2016; Reed et al, ICML 2016; Reed et al, NIPS 2016; Reed et al, NIPS 2017; Nguyen et al, CVPR 2017; Zhang et al, ICCV 2017

Text to Image

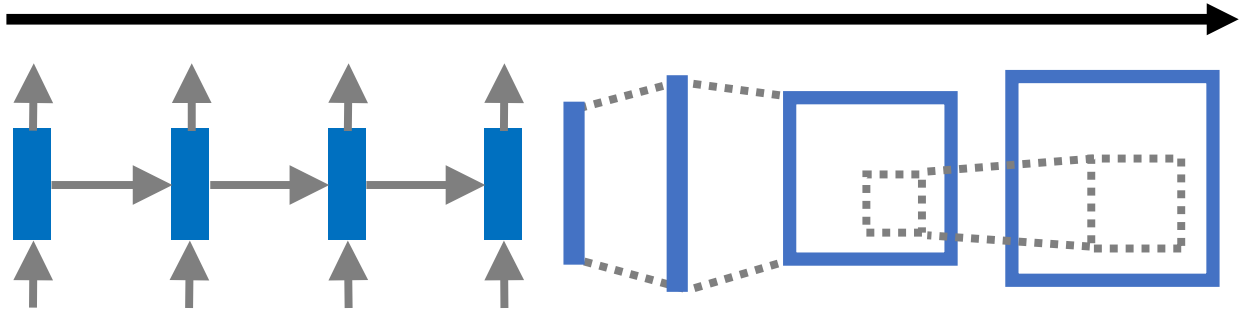
Input x

Model f_w

Generated image

$\hat{y} = f_w(x)$

A red bird with black wings



Recurrent Network

Convolutional Network

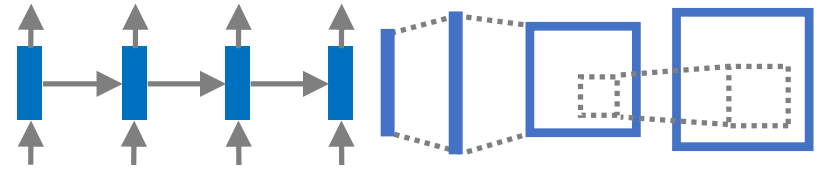
Mansimov et al, ICLR 2016; Reed et al, ICML 2016; Reed et al, NIPS 2016; Reed et al, NIPS 2017; Nguyen et al, CVPR 2017; Zhang et al, ICCV 2017

Text to Image : Loss Function

Input x

A red bird with black wings

Model f_w



Generated image $\hat{y} = f_w(x)$

Loss $\ell(\hat{y}, y)$

$$w^* = \arg\min_w \mathbb{E}_{x,y}\left[\ell(f_w(x), y)\right]$$

Squared Loss $\ell(\hat{y}, y) = \|\hat{y} - y\|_2^2$

Does not capture semantics

Ground-truth image y

Mansimov et al, ICLR 2016; Reed et al, ICML 2016; Reed et al, NIPS 2016; Reed et al, NIPS 2017; Nguyen et al, CVPR 2017; Zhang et al, ICCV 2017

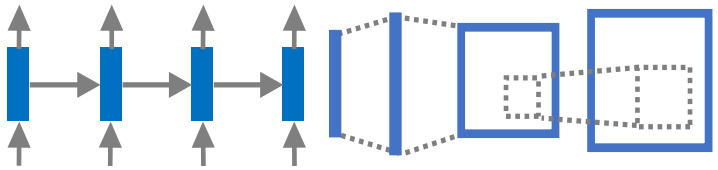
Text to Image: Adversarial Network

Goodfellow et al, NIPS 2014

Input x

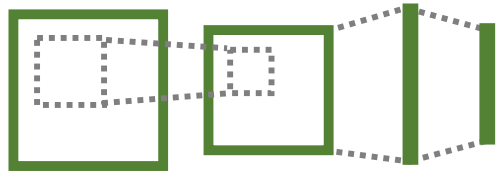
A red bird with black wings

Model f_w



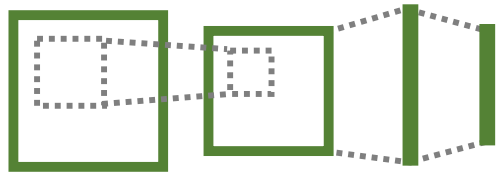
Generated image $\hat{y} = f_w(x)$

Discriminator D_θ



$P(\hat{y} \text{ is real})$

Loss $\ell(\hat{y}, y)$



$P(y \text{ is real})$

Discriminator D_θ

Ground-truth image y

Match pixels

Fool discriminator

$$w^* = \arg\min_w \mathbb{E}_{x,y}\left[\ell(f_w(x), y) + \log D_\theta(f_w(x))\right]$$

$$\theta^* = \arg\max_\theta \mathbb{E}_{x,y}\left[\log D_\theta(y) + \log\left(1 - D_\theta(f_w(x))\right)\right]$$

Classify real images

Classify fake images

Mansimov et al, ICLR 2016; Reed et al, ICML 2016; Reed et al, NIPS 2016; Reed et al, NIPS 2017; Nguyen et al, CVPR 2017; Zhang et al, ICCV 2017

Text to Image

This bird is blue with white and has a very short beak

A white bird with a black crown and yellow beak

Mansimov et al, ICLR 2016; Reed et al, ICML 2016; Reed et al, NIPS 2016; Reed et al, NIPS 2017; Nguyen et al, CVPR 2017; **Zhang et al, ICCV 2017**

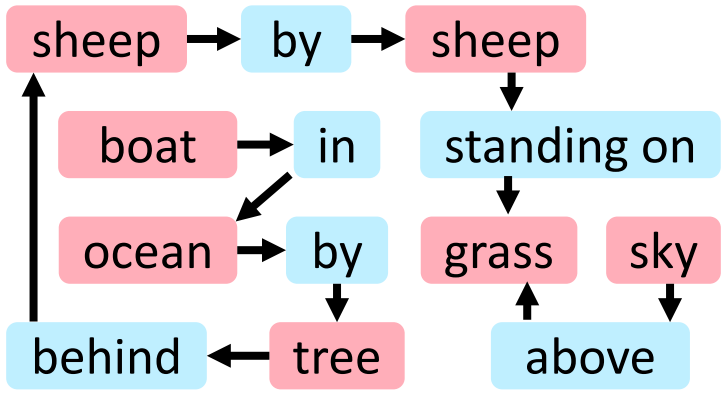
Text to Image: Complex Sentences

Schuster et al, 2015

Natural Language

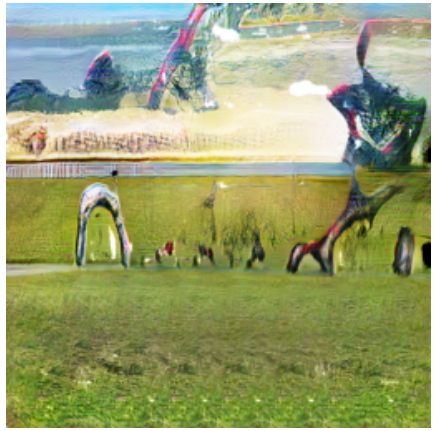
A sheep by another sheep standing on the grass with sky above and a boat in the ocean by a tree behind the sheep

Scene Graph

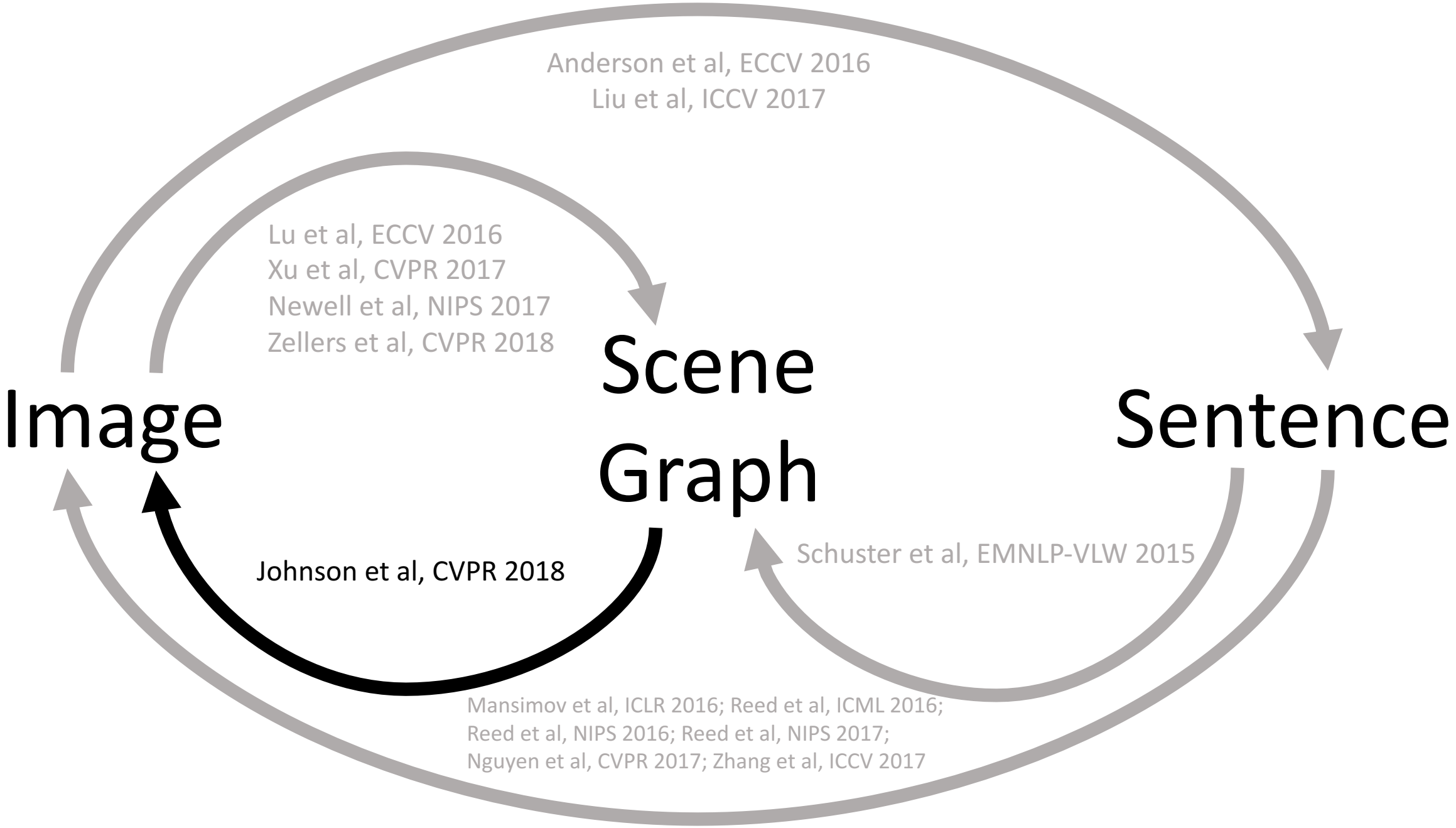


Better results: Zhang et al, Self-Attention Generative Adversarial Networks, arXiv 2018

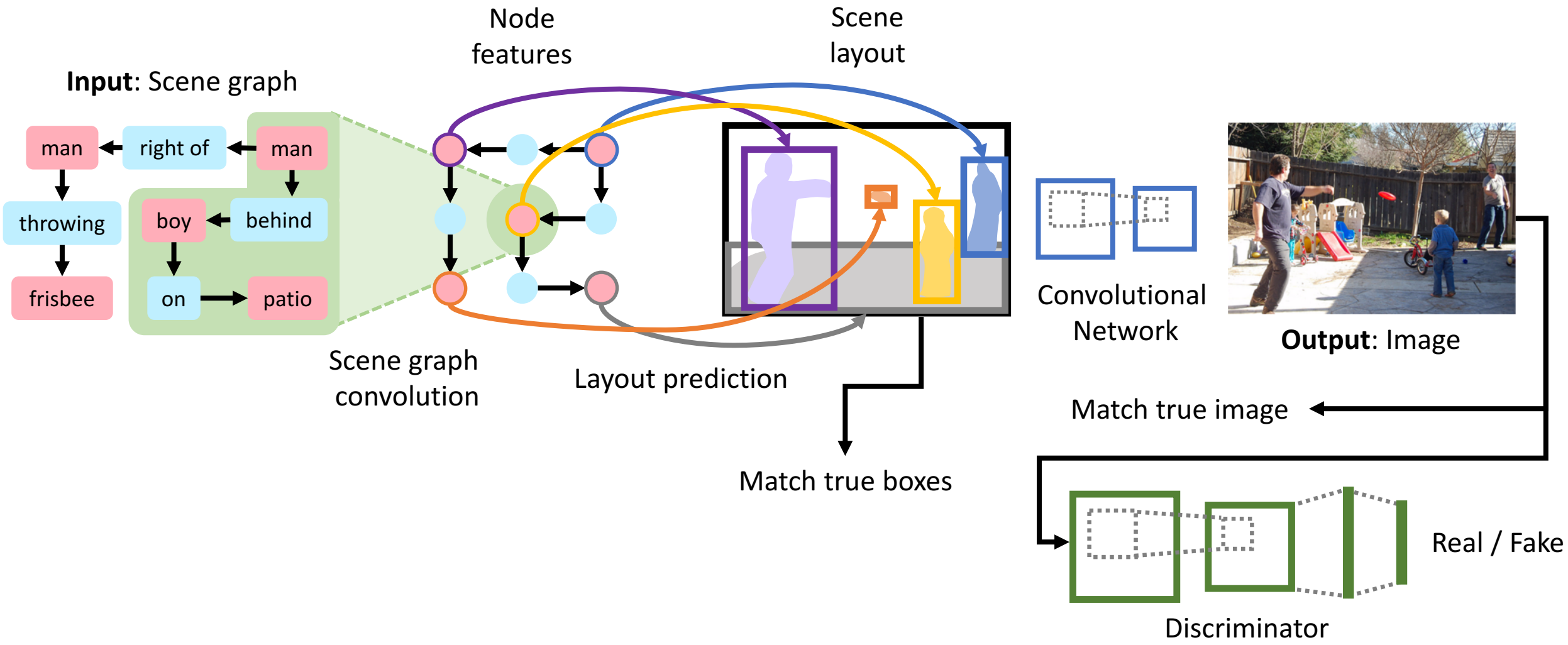
Zhang et al, ICCV 2017



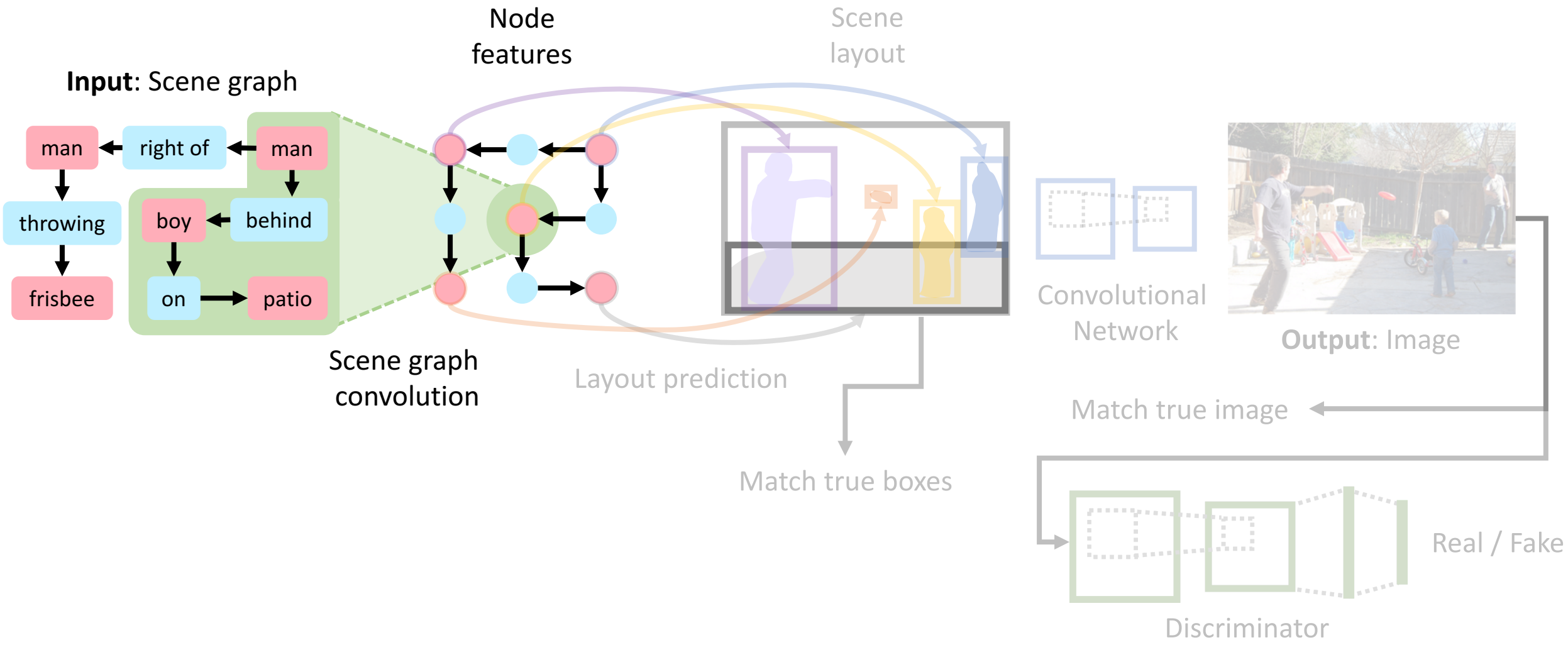
Johnson, Gupta, and Fei-Fei, CVPR 2018


Anderson et al, ECCV 2016
Liu et al, ICCV 2017
Lu et al, ECCV 2016
Xu et al, CVPR 2017
Newell et al, NIPS 2017
Zellers et al, CVPR 2018
Image
Scene Graph
Sentence
Johnson et al, CVPR 2018
Schuster et al, EMNLP-VLW 2015
Mansimov et al, ICLR 2016; Reed et al, ICML 2016;
Reed et al, NIPS 2016; Reed et al, NIPS 2017;
Nguyen et al, CVPR 2017; Zhang et al, ICCV 2017

Scene Graph to Image: Model



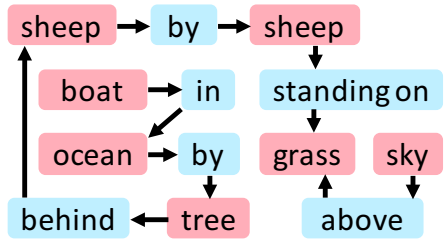
Scene Graph to Image: Model



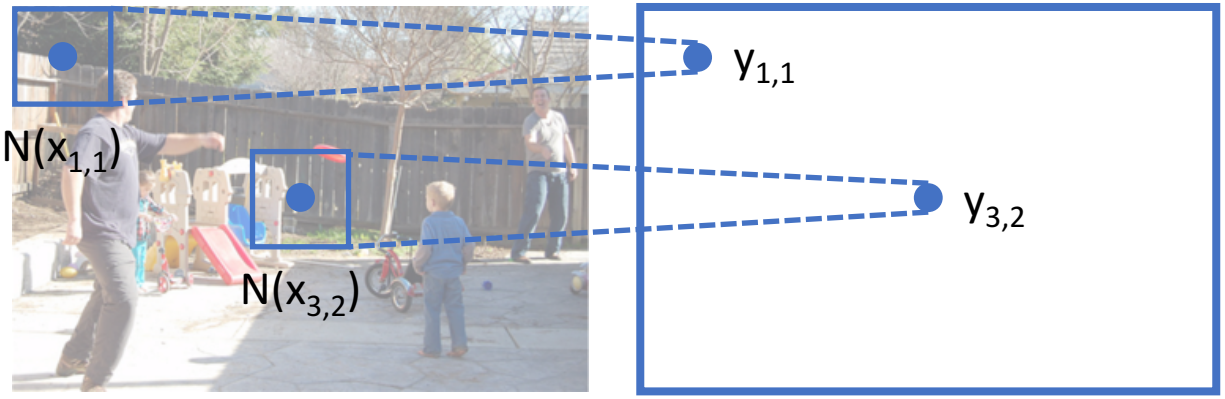
Johnson, Gupta, and Fei-Fei, CVPR 2018

Processing different input types

What about scene graphs?



Images: Convolutional Networks



Input: Grid of vectors

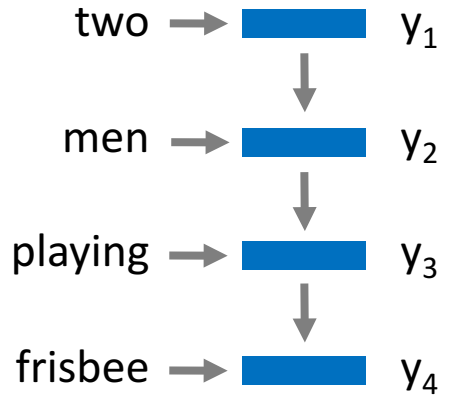
Output: Grid of vectors

Apply same function to all neighborhoods of input:

$$y_i = f_W(\mathcal{N}(x_i))$$

Sequences: Recurrent Networks

Input: Sequence of vectors



Output: Sequence of vectors

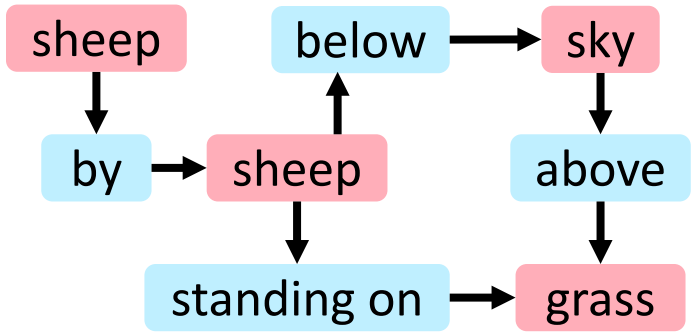
Apply same function to all timesteps of input:

$$y_t = f_W(x_t, y_{t-1})$$

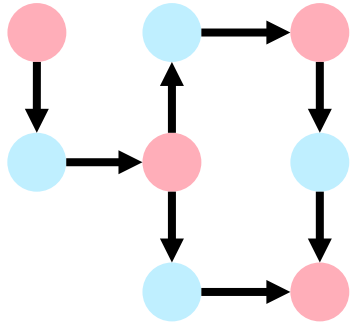
”Fully-connected” neural network: Matrix multiply and nonlinearity

Graph Convolution Networks

Input: Graph of vectors



Output: Graph of vectors



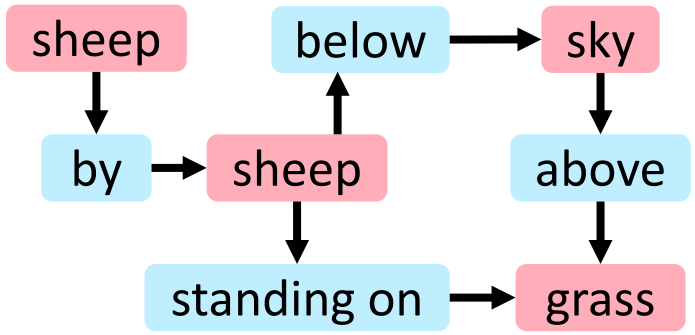
Apply same function to all neighborhoods of input:

$$y_i = f_W(\mathcal{N}(x_i))$$

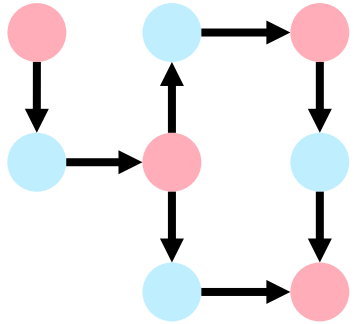
Goller and Kuchler, 1996; Gori et al, 2005; Scarselli et al, 2009; Bruna et al, ICLR 2014; Li et al, ICLR 2015; Henaff et al, arXiv 2015; Duvenaud et al, NIPS 2015; Jain et al, CVPR 2016; Kipf and Welling, ICLR 2017; Wang et al, NIPS 2017

Scene Graph Convolution Networks

Input: Graph of vectors



Output: Graph of vectors



Two types of nodes**: relationships** and **objects**; use different functions for each

Relationship nodes

$$y_r = f_W^r(\mathcal{N}(x_r))$$

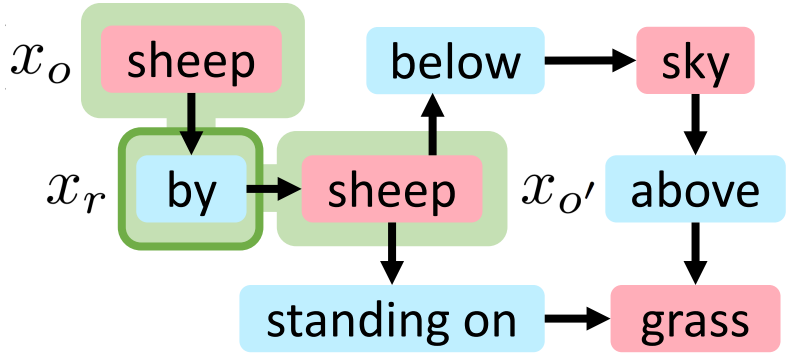
Object nodes

$$y_o = f_W^o(\mathcal{N}(x_o))$$

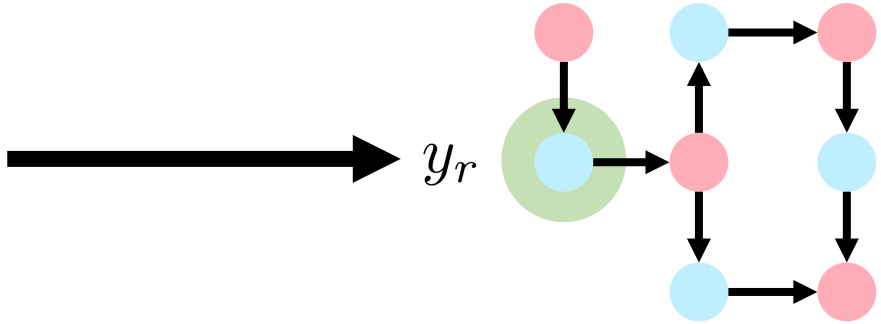
Johnson, Gupta, and Fei-Fei, CVPR 2018

Scene Graph Convolution Networks

Input: Graph of vectors



Output: Graph of vectors



Two types of nodes**: relationships** and **objects**; use different functions for each

Relationship nodes have two neighbors

$$y_r = f_W^r(\mathcal{N}(x_r)) = f_W^r(x_o, x_r, x_{o'})$$

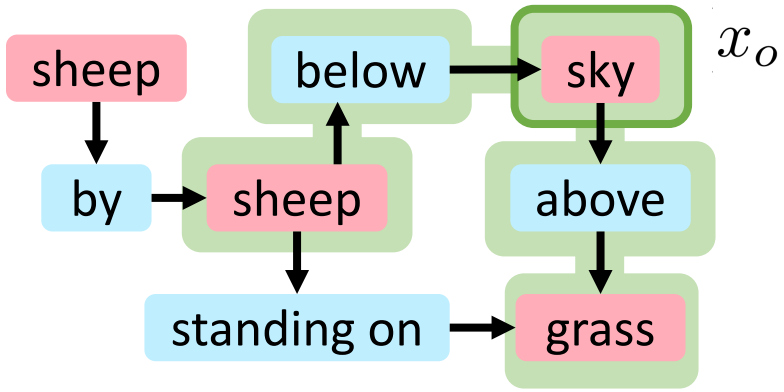
Object nodes

$$y_o = f_W^o(\mathcal{N}(x_o))$$

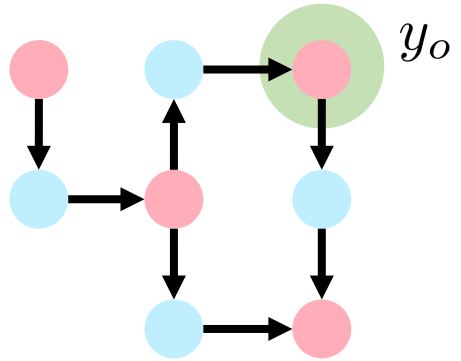
Johnson, Gupta, and Fei-Fei, CVPR 2018

Scene Graph Convolution Networks

Input: Graph of vectors



Output: Graph of vectors



Two types of nodes**: relationships** and **objects**; use different functions for each

Relationship nodes have two neighbors

$$y_r = f_W^r(\mathcal{N}(x_r)) = f_W^r(x_o, x_r, x_{o'})$$

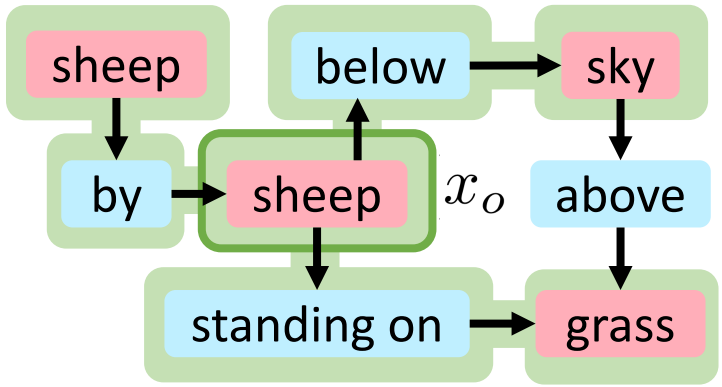
Object nodes have varying neighbors

$$y_o = f_W^o(\mathcal{N}(x_o))$$

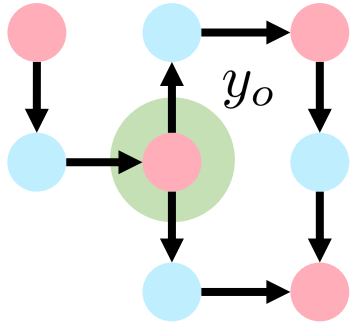
Scene Graph Convolution Networks

Input: Graph of vectors

Three neighbors



Output: Graph of vectors



Two types of nodes**: relationships** and **objects**; use different functions for each

”Fully-connected” neural network:
Matrix multiply and nonlinearity

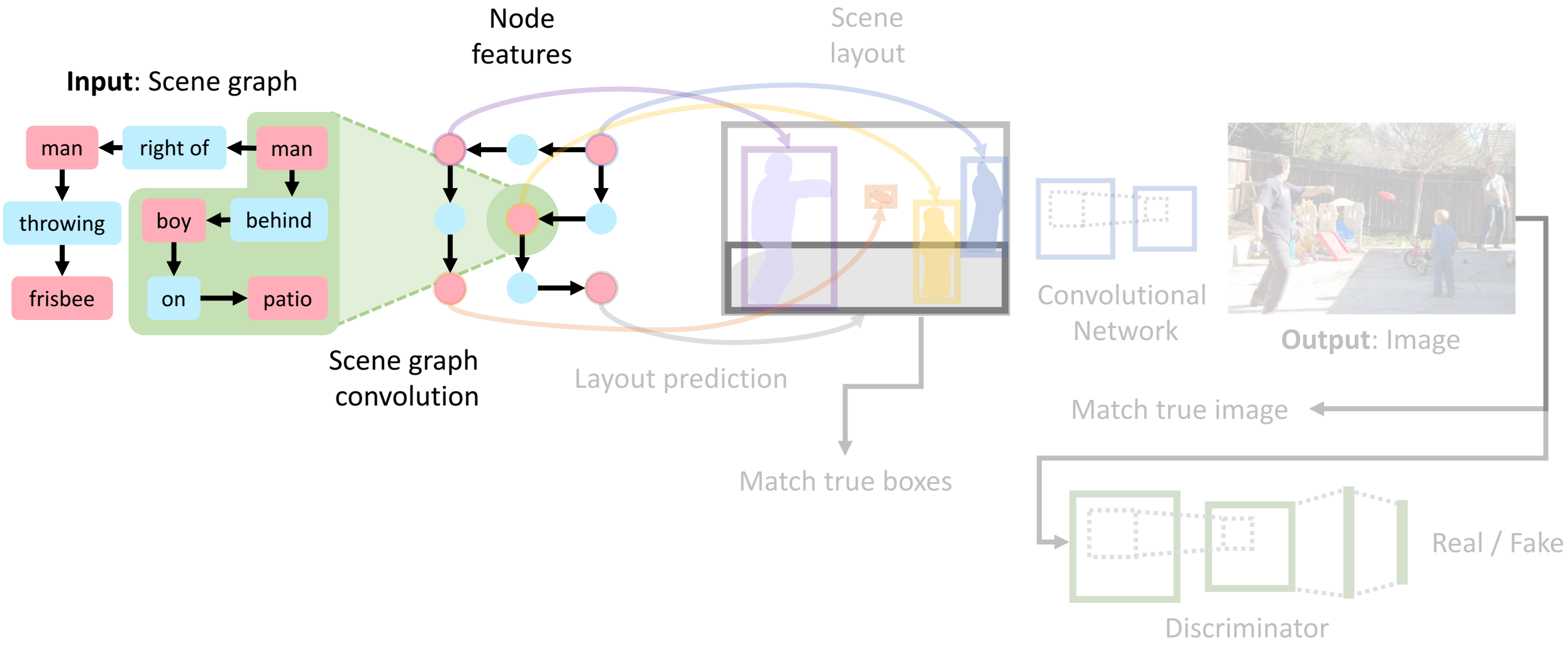
Relationship nodes have two neighbors

$$y_r = f_W^r(\mathcal{N}(x_r)) = f_W^r(x_o, x_r, x_{o'})$$

Object nodes have varying neighbors

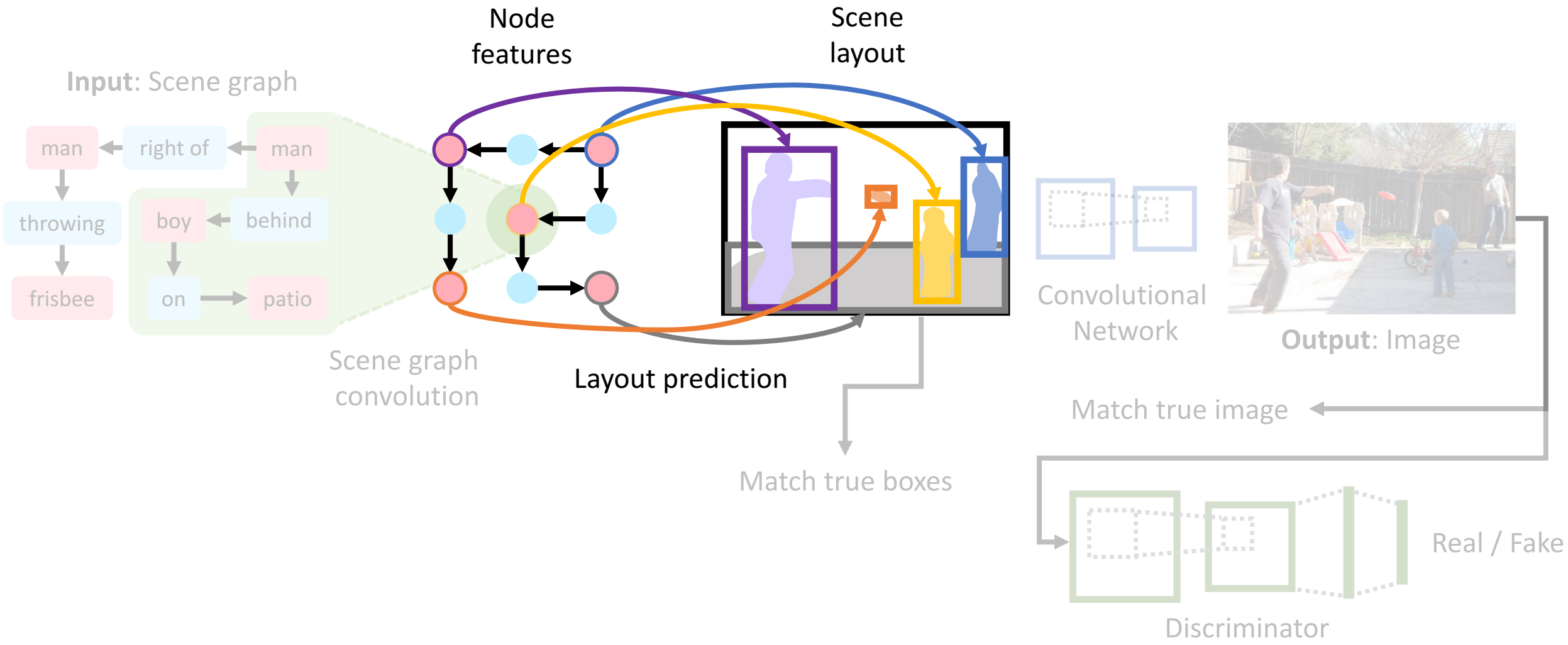
$$y_o = f_W^o(\mathcal{N}(x_o)) \left(\sum_{(o,e,o') \in G} g_W(x_o, x_r, x_{o'}) \right)$$

Scene Graph to Image: Model



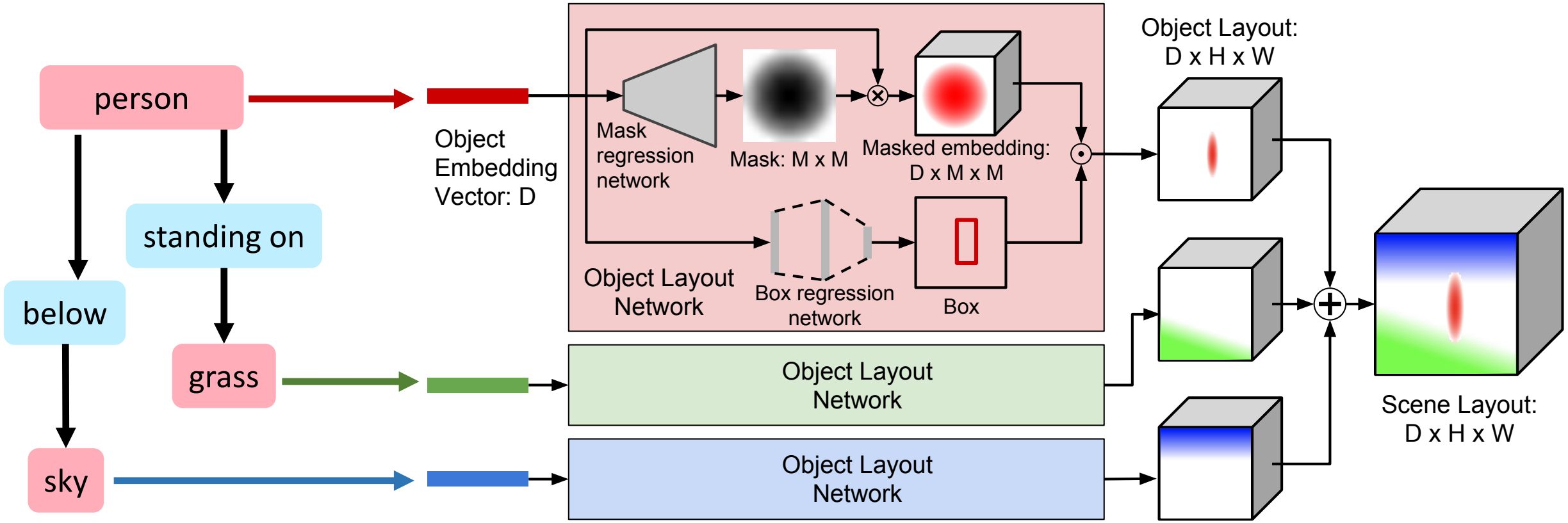
Johnson, Gupta, and Fei-Fei, CVPR 2018

Scene Graph to Image: Model

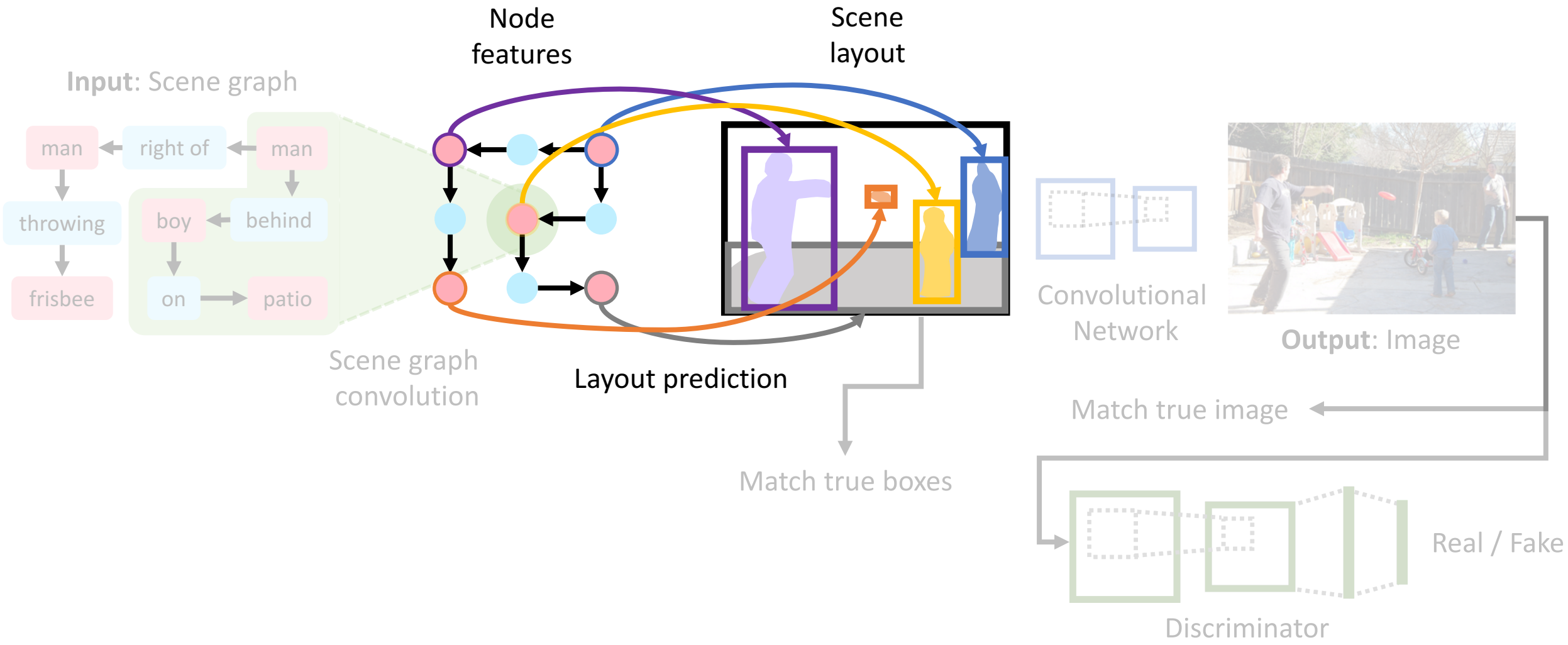


Johnson, Gupta, and Fei-Fei, CVPR 2018

Scene Layout

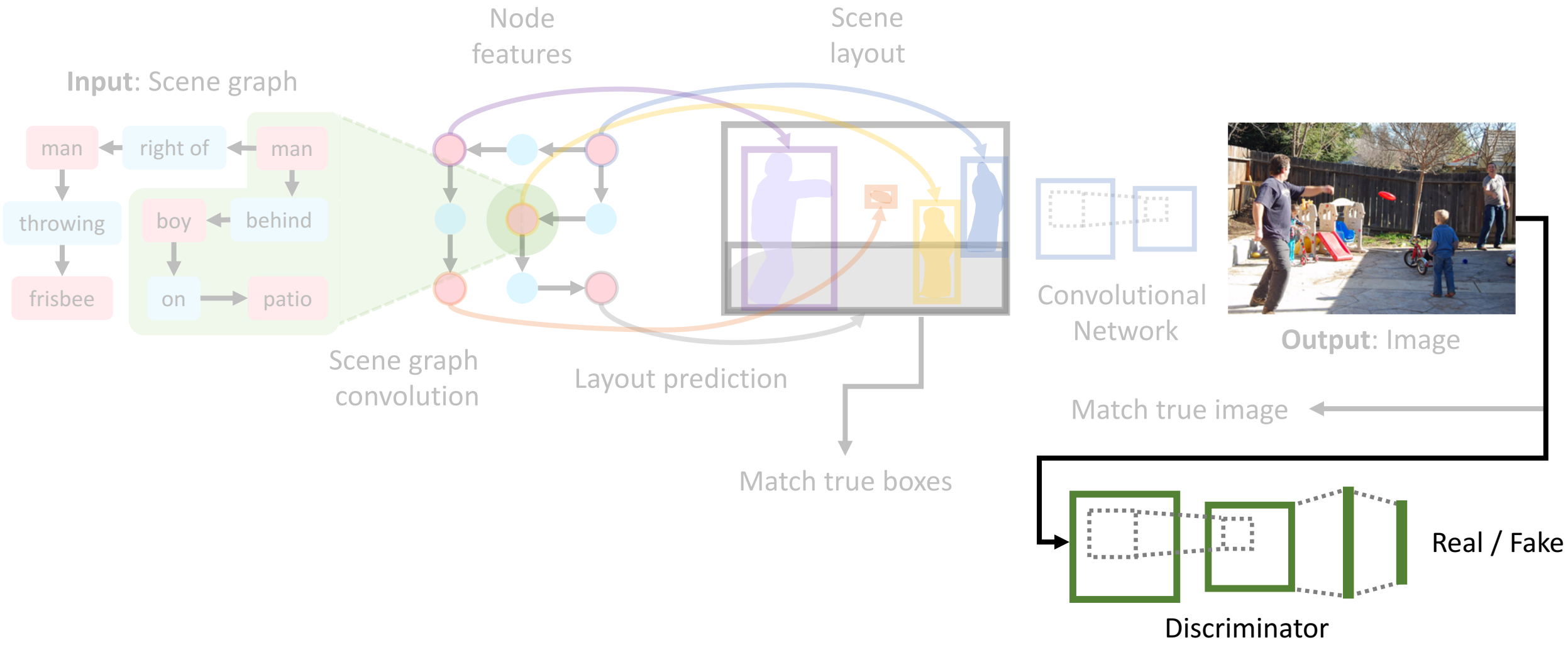


Scene Graph to Image: Model



Johnson, Gupta, and Fei-Fei, CVPR 2018

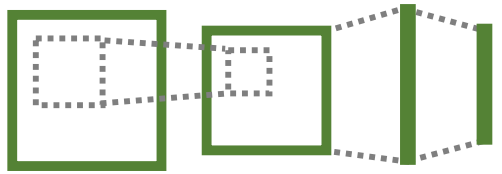
Scene Graph to Image: Model



Johnson, Gupta, and Fei-Fei, CVPR 2018

Scene Graph to Image: Discriminator

Patch discriminator

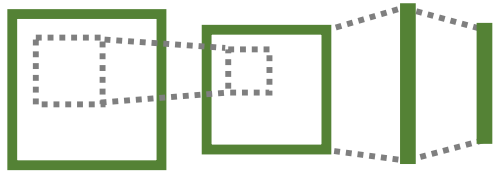


Real / fake

Crop objects

Differentiable bilinear crop

Object discriminator

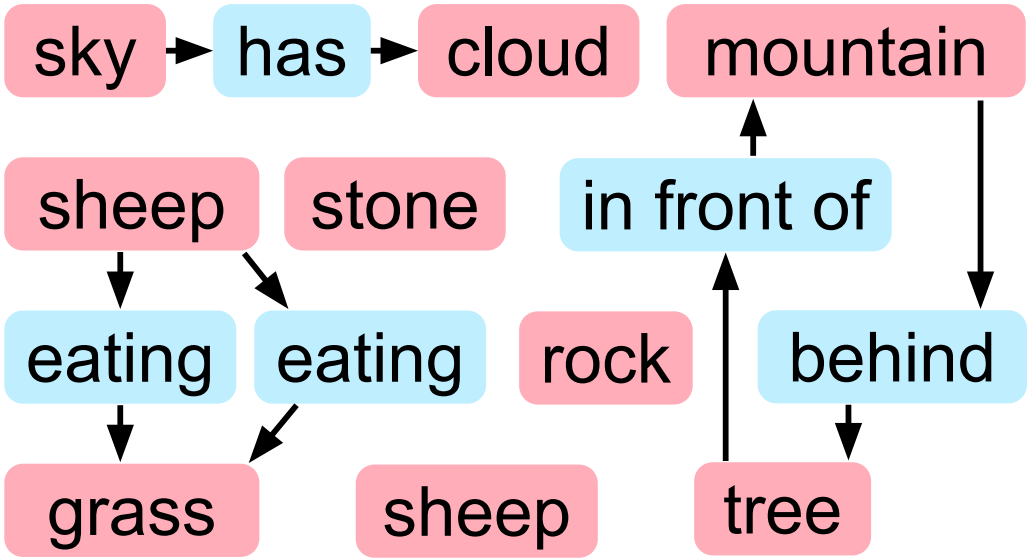


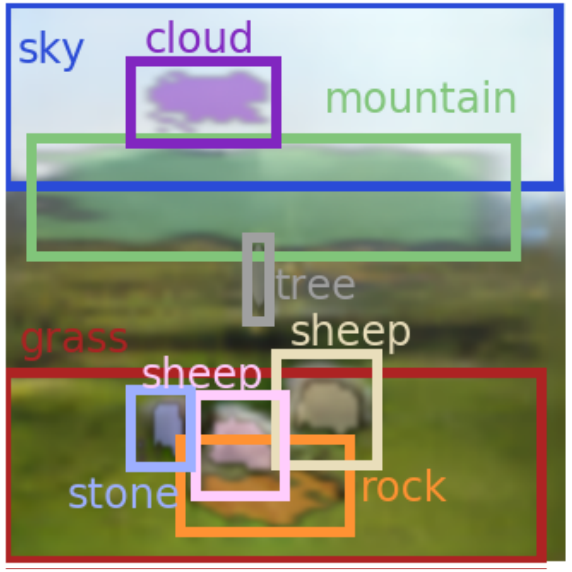
Real / fake
Object category

Johnson, Gupta, and Fei-Fei, CVPR 2018

PatchGAN: Isola et al, CVPR 2017
AC-GAN: Odena et al, ICML 2017

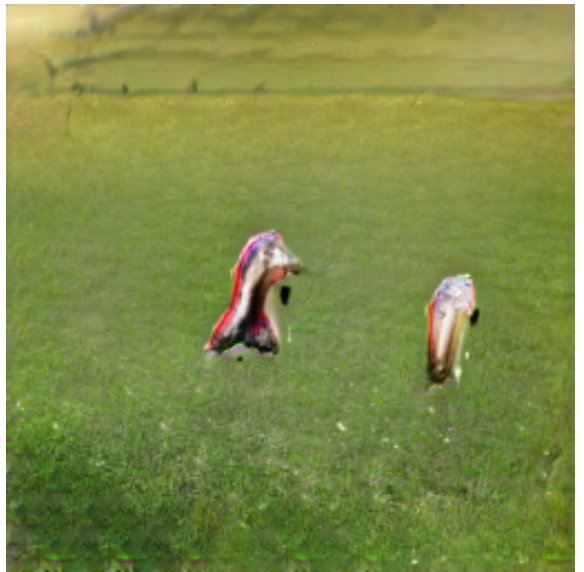
Results on Visual Genome



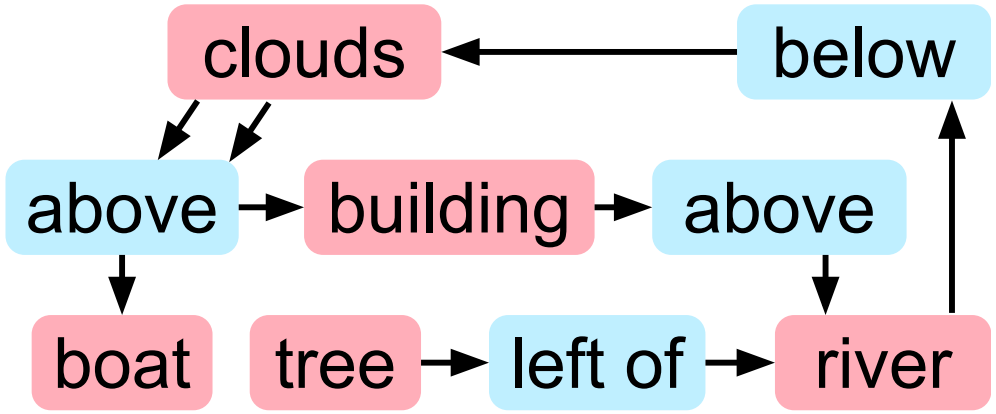


Two sheep, one eating grass with a tree in front of a mountain; the sky has a cloud

StackGAN
Zhang et al, ICCV 2017



Results on COCO-Stuff



Clouds above a boat and a building above a river, with trees left of the river

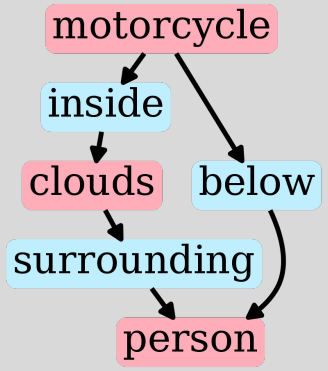
StackGAN
Zhang et al, ICCV 2017

Ablation Studies

Method	Inception	
	COCO	VG
Real Images (64×64)	16.3 ± 0.4	13.9 ± 0.5
Ours (No gconv)	4.6 ± 0.1	4.2 ± 0.1
Ours (No relationships)	3.7 ± 0.1	4.9 ± 0.1
Ours (No discriminators)	4.8 ± 0.1	3.6 ± 0.1
Ours (No D_{obj})	5.6 ± 0.1	5.0 ± 0.2
Ours (No D_{img})	5.6 ± 0.1	$\mathbf{5.7 \pm 0.3}$
Ours (Full model)	$\mathbf{6.7 \pm 0.1}$	5.5 ± 0.1
Ours (GT Layout, no gconv)	7.0 ± 0.2	6.0 ± 0.2
Ours (GT Layout)	$\mathbf{7.3 \pm 0.1}$	$\mathbf{6.3 \pm 0.2}$
StackGAN [59] (64×64)	$\mathbf{8.4 \pm 0.2}$	-

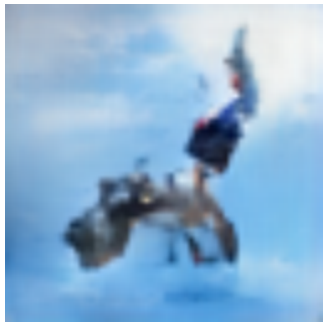
Results: User Study

COCO Annotations



A man flying through the air while riding a bike

Ours

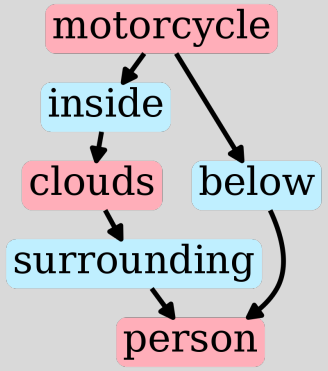


StackGAN

Johnson, Gupta, and Fei-Fei, CVPR 2018

Results: User Study

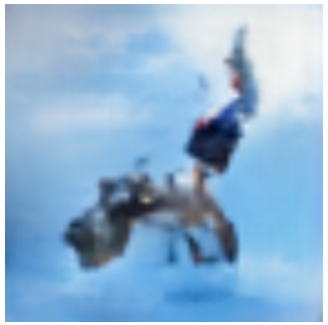
COCO Annotations



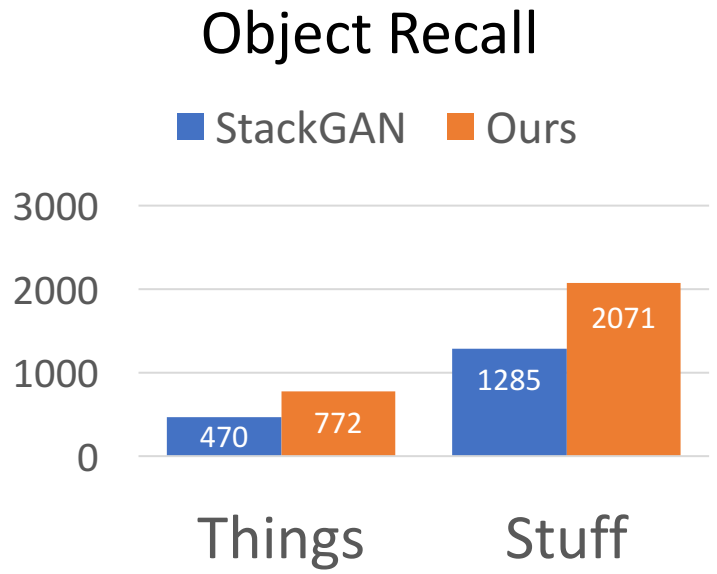
A man flying through the air while riding a bike

Q: Which of the following objects are present?
motorcycle, clouds, person

Ours



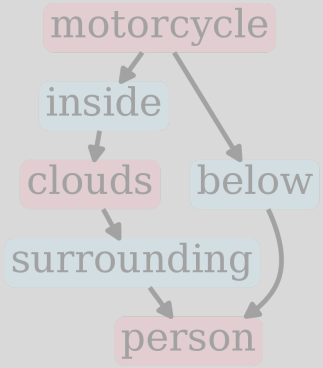
StackGAN



Johnson, Gupta, and Fei-Fei, CVPR 2018

Results: User Study

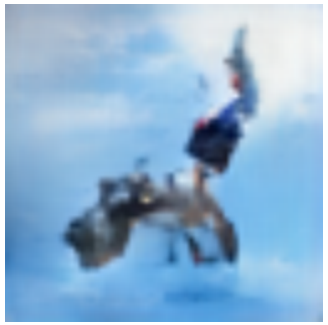
COCO Annotations



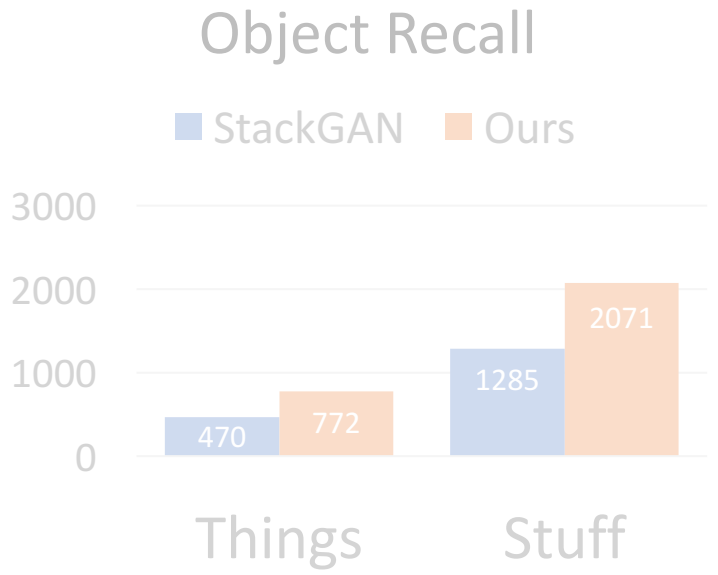
A man flying through the air while riding a bike

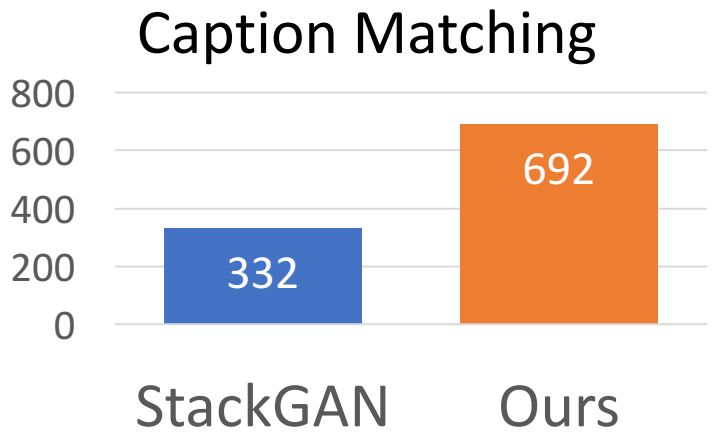
Q: Which image matches the caption better?

Ours →



StackGAN →

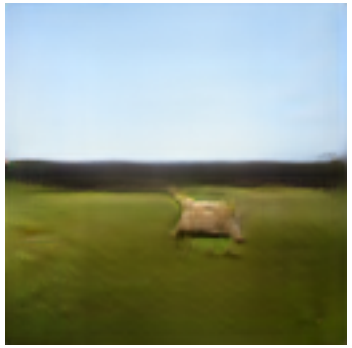




Johnson, Gupta, and Fei-Fei, CVPR 2018

Simple to Complex

Sky above grass; zebra standing on grass



Sky above grass; **sheep** standing on grass

Sky above grass; sheep standing on grass; **sheep2 by sheep**

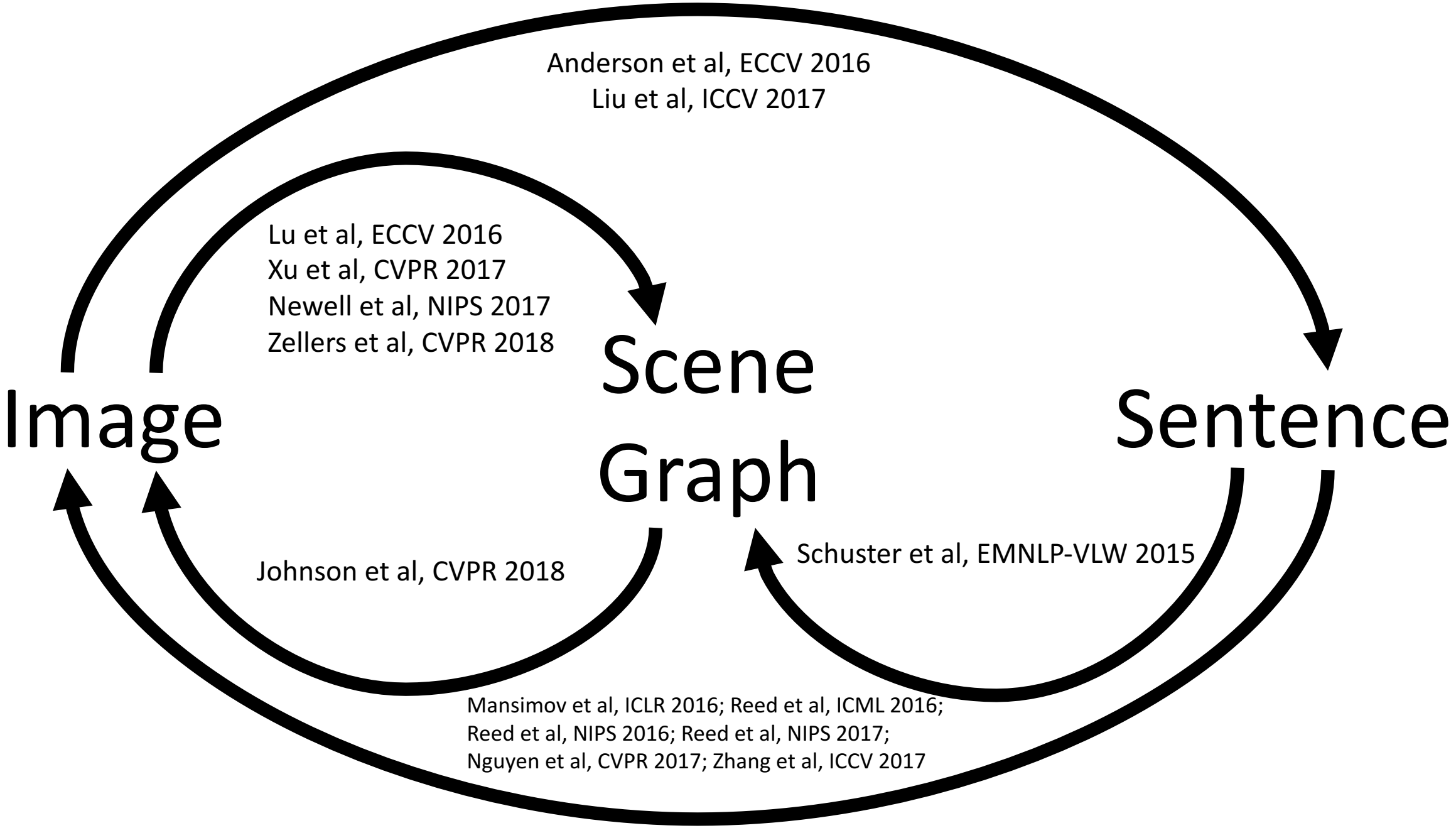
Sky above grass; sheep standing on grass; sheep2 by sheep; **tree behind sheep**

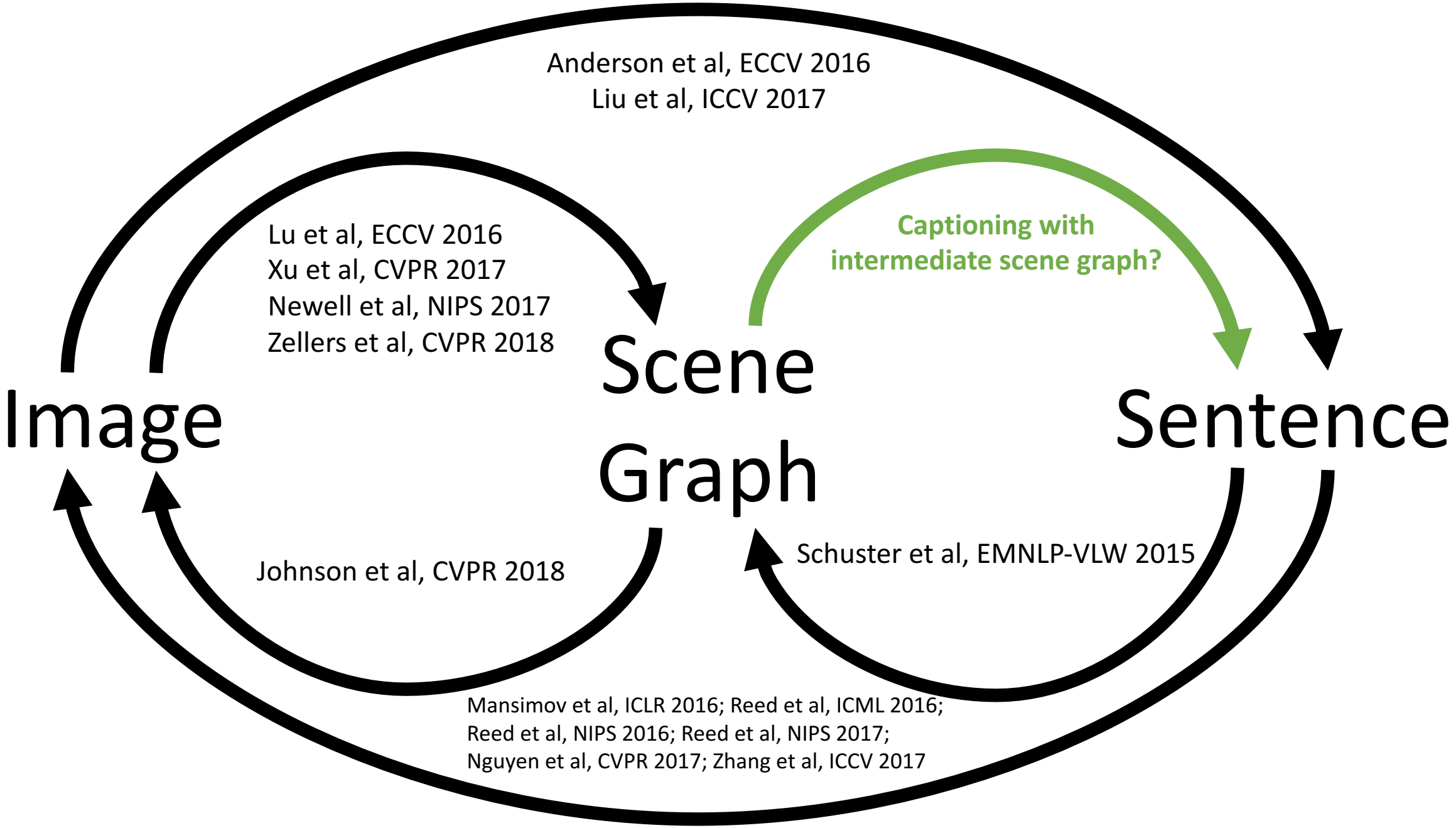
Sky above grass; sheep standing on grass; sheep2 by sheep; tree behind sheep; **ocean by tree**

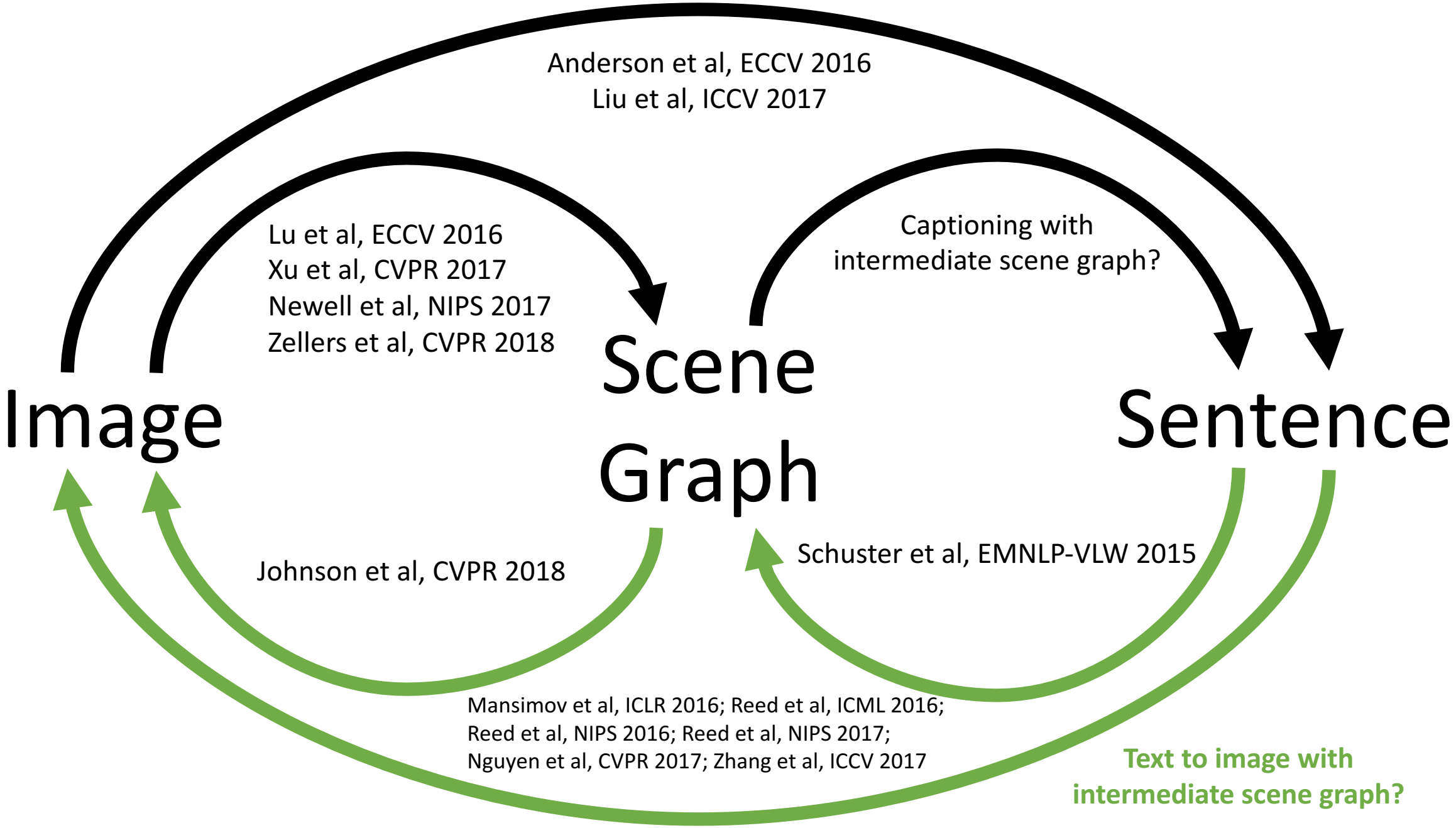
Sky above grass; sheep standing on grass; sheep2 by sheep; tree behind sheep; ocean by tree; **boat in ocean**

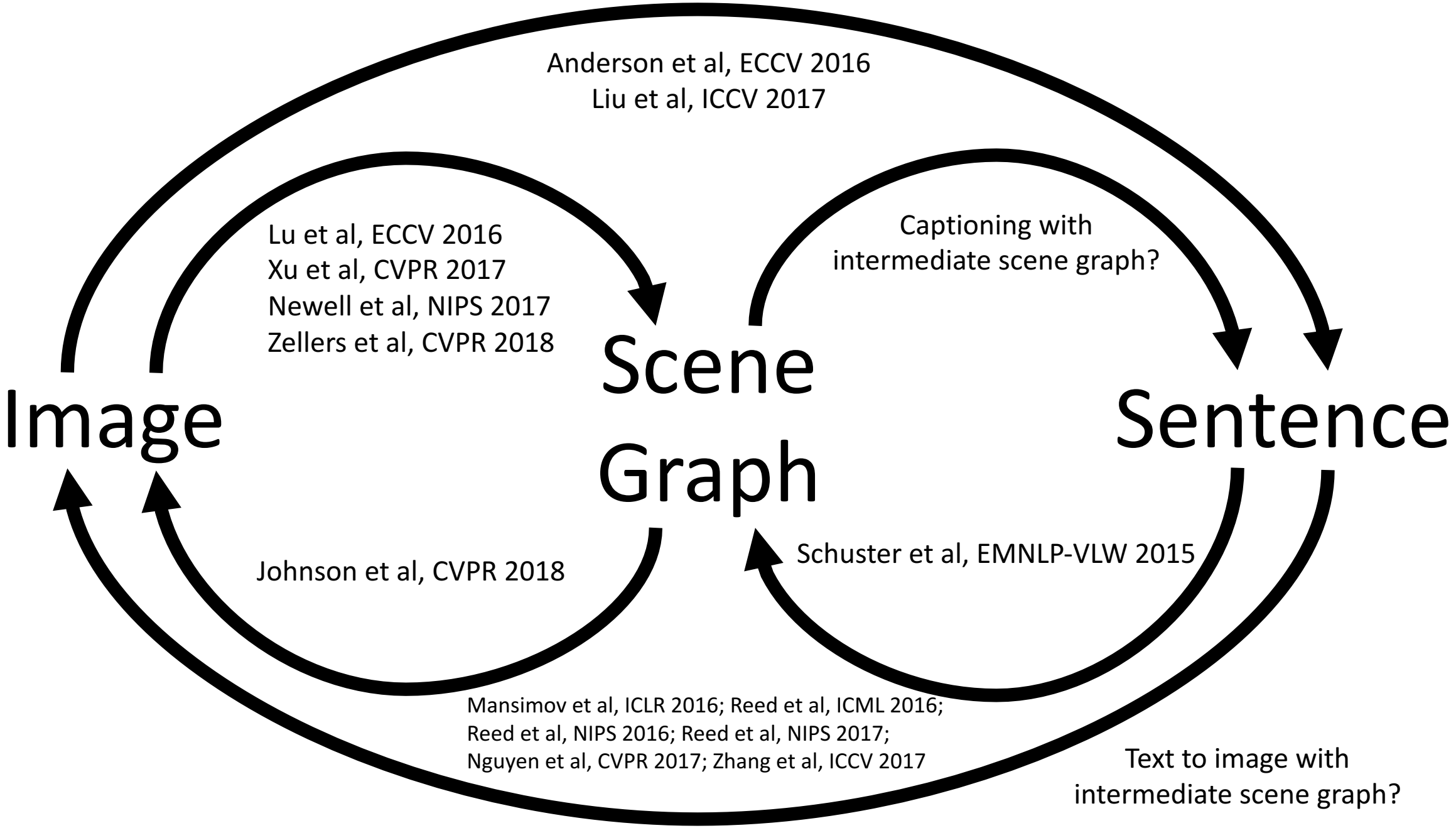
Sky above grass; sheep standing on grass; sheep2 by sheep; tree behind sheep; ocean by tree; **boat on grass**

Johnson, Gupta, and Fei-Fei, CVPR 2018


Anderson et al, ECCV 2016
Liu et al, ICCV 2017
Lu et al, ECCV 2016
Xu et al, CVPR 2017
Newell et al, NIPS 2017
Zellers et al, CVPR 2018
Image
Scene Graph
Sentence
Johnson et al, CVPR 2018
Schuster et al, EMNLP-VLW 2015
Mansimov et al, ICLR 2016; Reed et al, ICML 2016;
Reed et al, NIPS 2016; Reed et al, NIPS 2017;
Nguyen et al, CVPR 2017; Zhang et al, ICCV 2017


Anderson et al, ECCV 2016
Liu et al, ICCV 2017
Lu et al, ECCV 2016
Xu et al, CVPR 2017
Newell et al, NIPS 2017
Zellers et al, CVPR 2018
Captioning with intermediate scene graph?
Image
Scene Graph
Sentence
Johnson et al, CVPR 2018
Schuster et al, EMNLP-VLW 2015
Mansimov et al, ICLR 2016; Reed et al, ICML 2016;
Reed et al, NIPS 2016; Reed et al, NIPS 2017;
Nguyen et al, CVPR 2017; Zhang et al, ICCV 2017


Anderson et al, ECCV 2016
Liu et al, ICCV 2017
Lu et al, ECCV 2016
Xu et al, CVPR 2017
Newell et al, NIPS 2017
Zellers et al, CVPR 2018
Captioning with
intermediate scene graph?
Image
Scene
Graph
Sentence
Johnson et al, CVPR 2018
Schuster et al, EMNLP-VLW 2015
Mansimov et al, ICLR 2016; Reed et al, ICML 2016;
Reed et al, NIPS 2016; Reed et al, NIPS 2017;
Nguyen et al, CVPR 2017; Zhang et al, ICCV 2017
Text to image with
intermediate scene graph?


Anderson et al, ECCV 2016
Liu et al, ICCV 2017
Lu et al, ECCV 2016
Xu et al, CVPR 2017
Newell et al, NIPS 2017
Zellers et al, CVPR 2018
Captioning with intermediate scene graph?
Image
Scene Graph
Sentence
Johnson et al, CVPR 2018
Schuster et al, EMNLP-VLW 2015
Mansimov et al, ICLR 2016; Reed et al, ICML 2016;
Reed et al, NIPS 2016; Reed et al, NIPS 2017;
Nguyen et al, CVPR 2017; Zhang et al, ICCV 2017
Text to image with intermediate scene graph?

Questions?