Learning the visual interpretation of sentences

C. L. Zitnick, D. Parikh, and L. Vanderwende*, ICCV 13

Presenter: Shenlong Wang
CSC 2523

*Many images from Larry Zitnick's ICCV 13 and slides, Coyne SIGGRAPH 01

We will discuss…

- Text to clip arts images
  - *Learning the Visual Interpretation of Sentences*, ICCV 2013 C. L. Zitnick, D. Parikh, and L. Vanderwende
  - *Bringing Semantics Into Focus Using Visual Abstraction*, CVPR 2013 (Oral) C. L. Zitnick and D. Parikh
- Text to 3D scene
  - *WordsEye: an automatic text-to-scene conversion system*, SIGGRAPH 2001, B. Coyne, and R. Sproat.
  - *Learning Spatial Knowledge for Text to 3D Scene Generation,* A. Chang, M. Savva, C. Manning, EMNLP 2014

Brief Review

- Image to Sentence
  - Retrieval
  - Generation
- Sentence to Image
  - Retrieval
  - Generation?

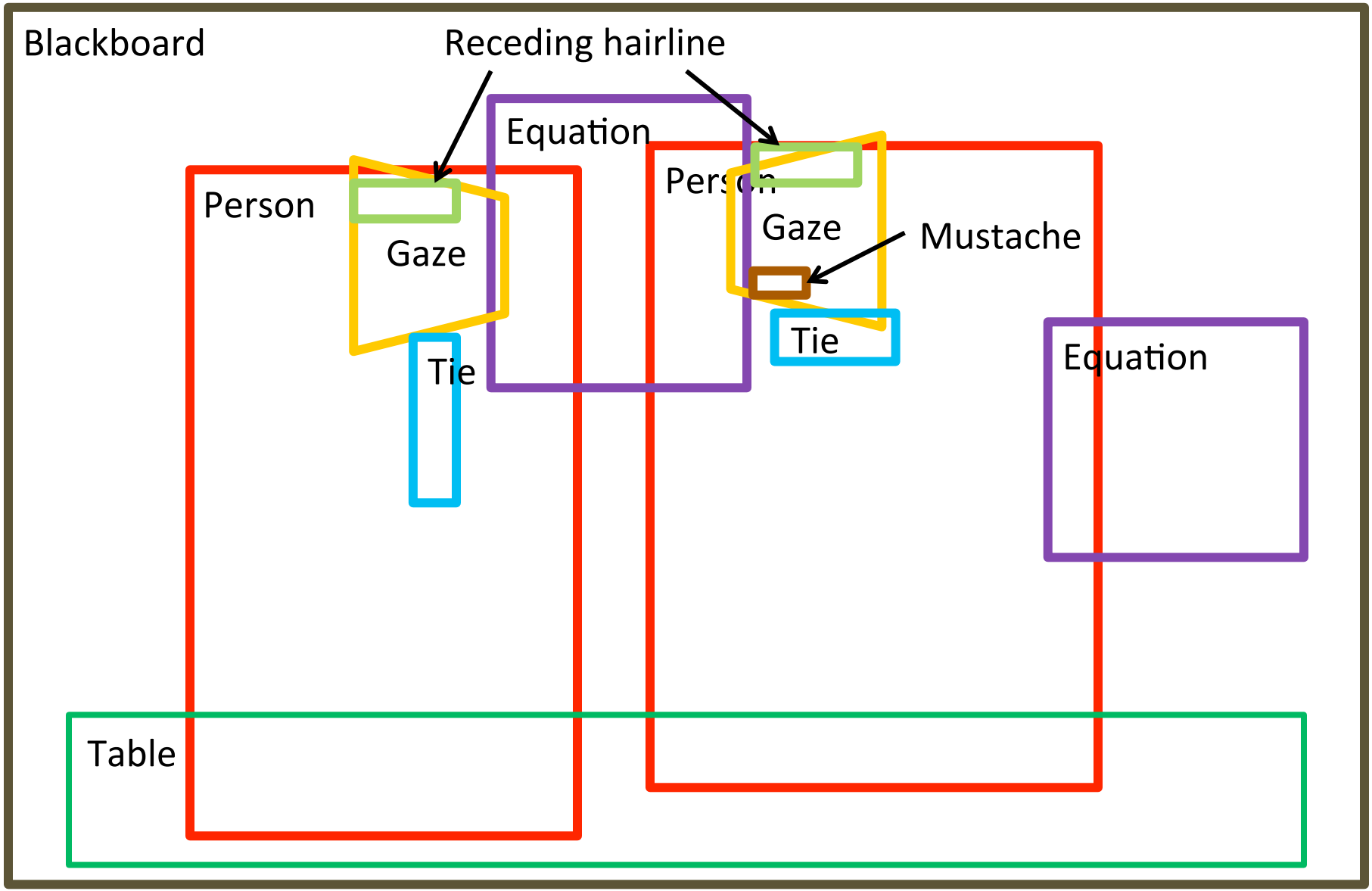
Goal

- To generate semantic meaningful images

Zitnick, 2013

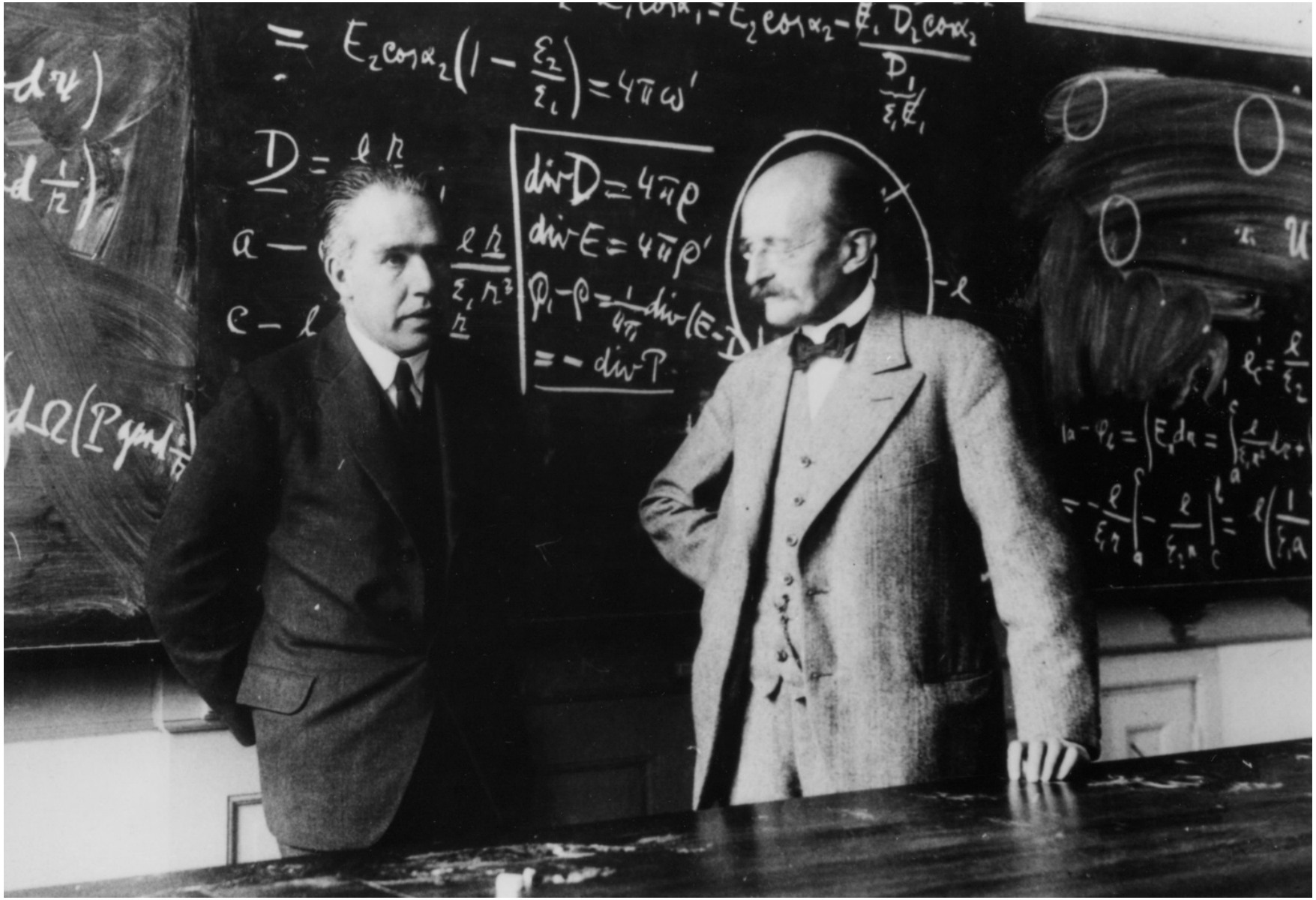
Two professors converse in front of a blackboard.

Two professors converse in front of a blackboard.

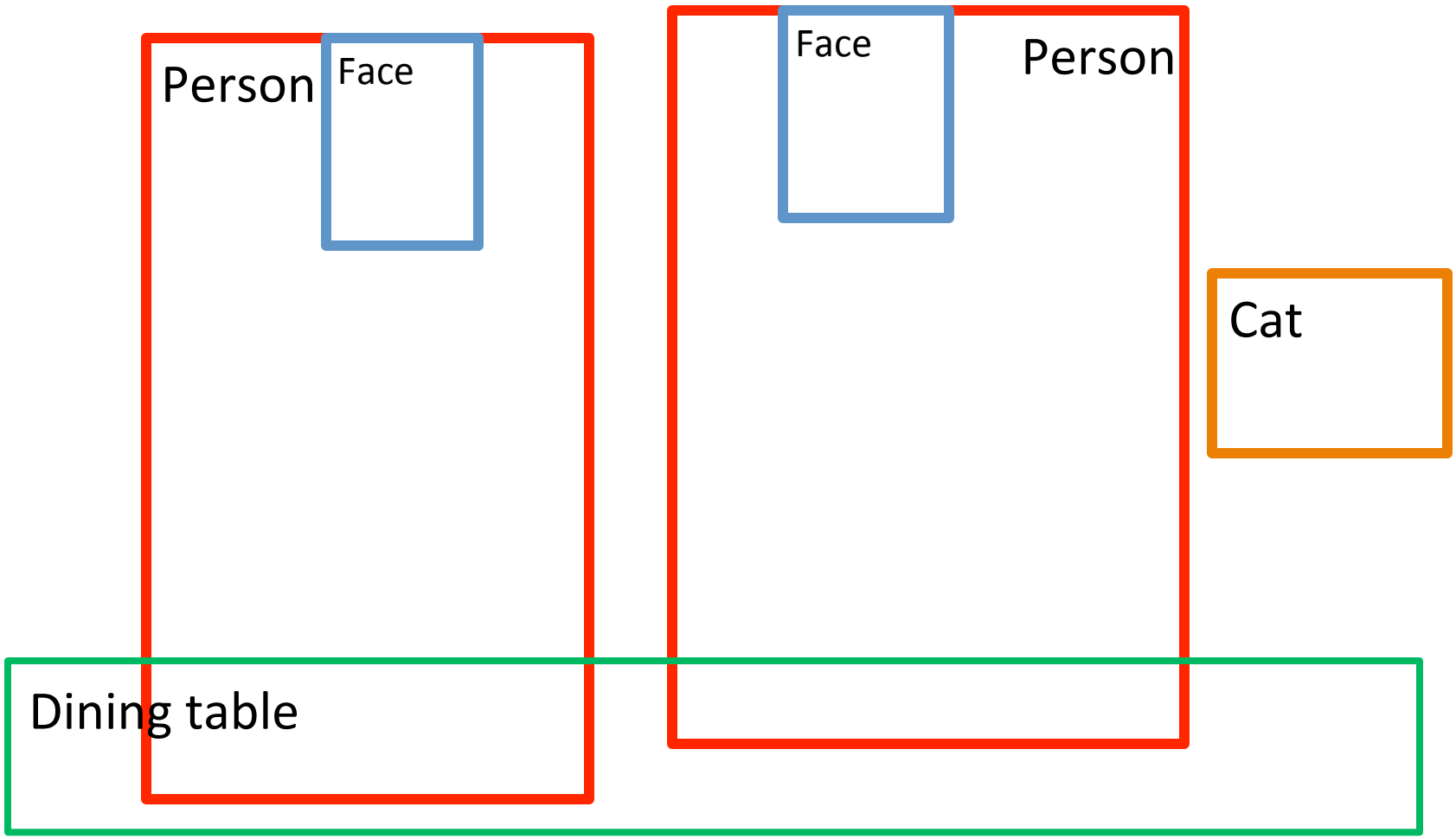


Zitnick, 2013

Two professors converse in front of a blackboard.

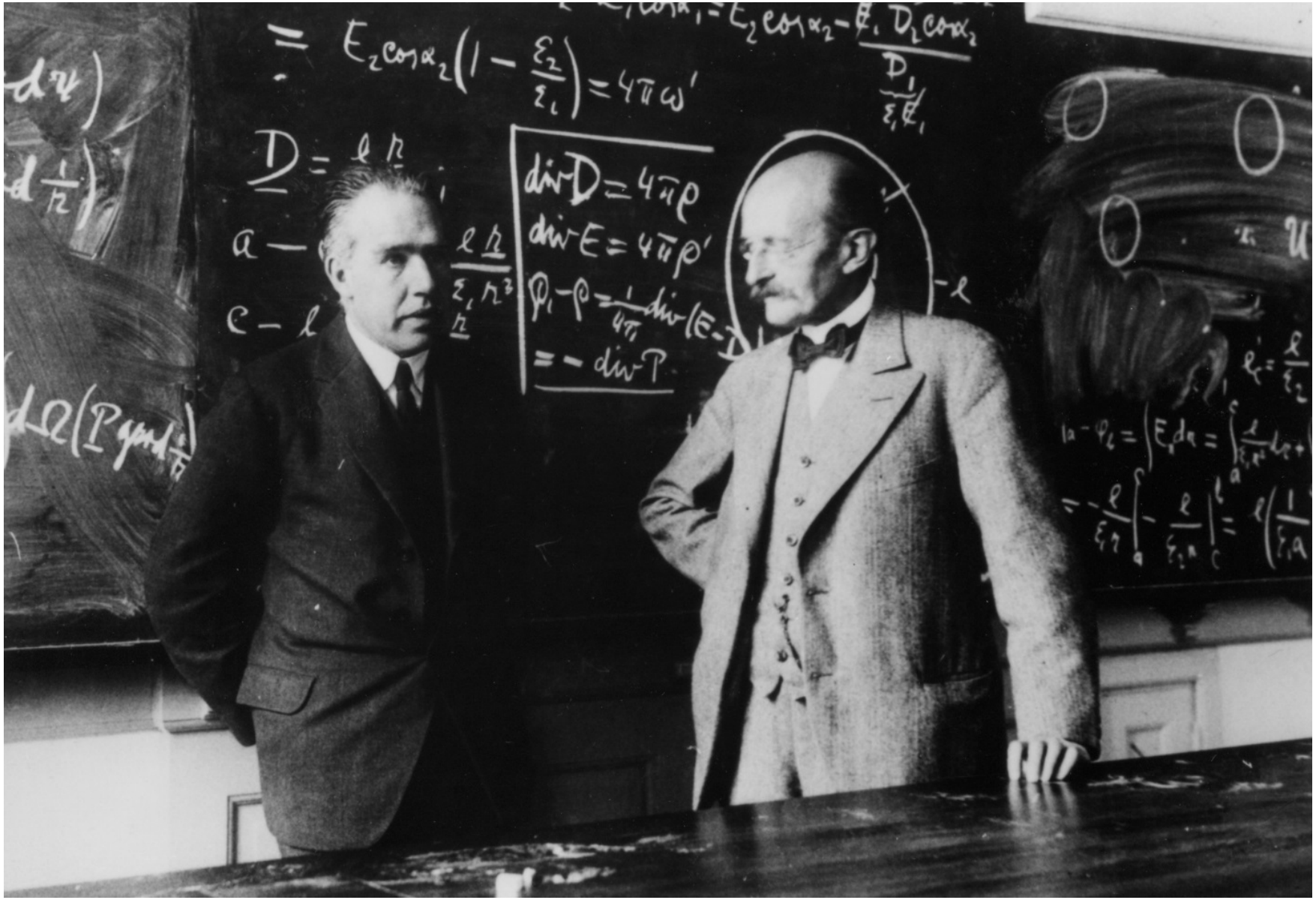


Zitnick, 2013



Felzenszwalb, 2010

Two professors converse in front of a blackboard.

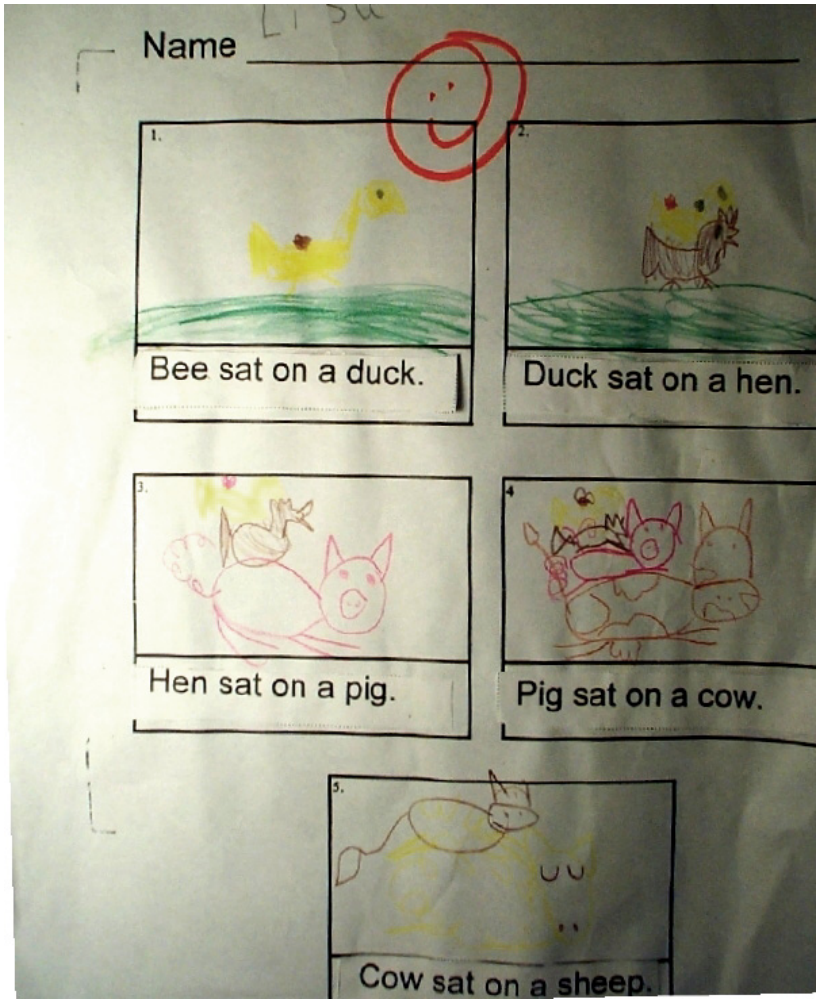


Zitnick, 2013

Two professors converse in front of a blackboard.

Image from 123RF.com

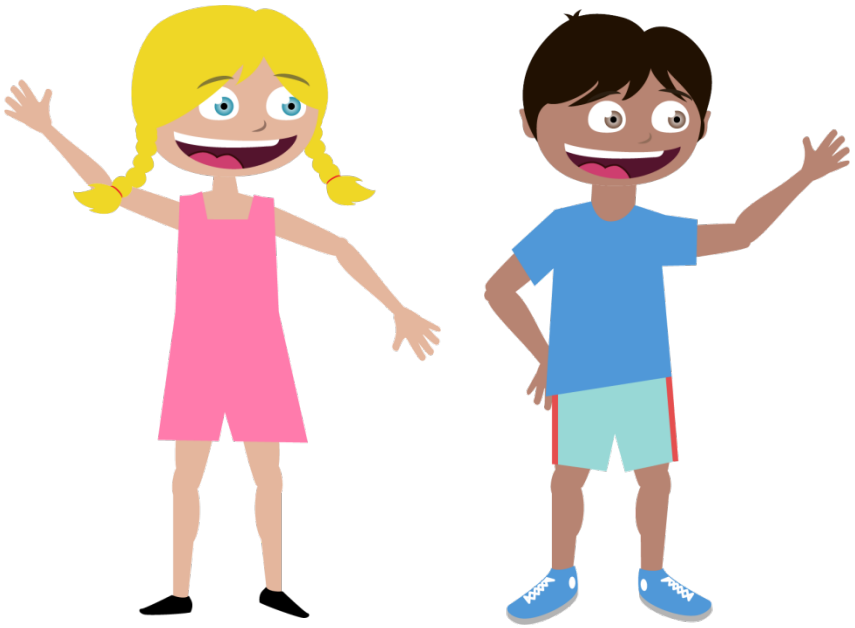
Photorealism is not necessary for learning visual interpretation of semantics



Coyne, 2001

Abstract scenes via 2D Clip Art

- Avoid the challenging vision parts (detection, segmentation, attributes, etc.) for real images.
- Reduce the variations of the real-world images with the same semantic meaning.



Jenny Mike

Zitnick, 2013

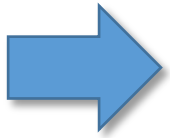
Summary of the dataset

- Clip arts
  - 56 Objects, 80 pieces of clip arts, 10000 scenes
  - 3D location with facing direction
  - Attributes for humans
- MTurker to label the data
  - Image to Sentence
  - Sentence to Image

Zitnick, 2013

Target

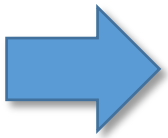
Jenny is catching the ball.
Mike is kicking the ball.
The table is next to the tree.



Sentence Parsing

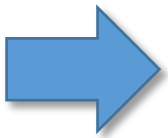
<primary object> <relation> <secondary object>

Jenny is catching the ball.
Mike is kicking the ball.
The table is next to the tree.



<Jenny> <catch> <ball>
<Mike> <kick> <ball>
<table> <next to> <tree>

Jenny and Mike are running from the snake.



<Jenny> <run from> <snake>
<Mike> <run from> <snake>

CRF model

$$\log P(c, \Phi, \Psi | S, \theta) =$$

$$\sum_i \left(\overbrace{\psi_i(c_i, S; \theta_c)}^{occurrence} + \overbrace{\lambda_i(\Phi_i, S; \theta_\lambda)}^{abs.\ location} + \overbrace{\pi_i(\Psi_i, S; \theta_\pi)}^{attributes} \right) +$$

$$\sum_{ij} \overbrace{\phi_{ij}(\Phi_i, \Phi_j, S; \theta_\phi)}^{rel.\ location} - \log Z(S, \theta) \qquad (1)$$

$\Phi_i = \{x_i, y_i, z_i, d_i\}$ Absolute location of object (3D location + facing)

$\Psi_i = \{e_i, g_i, h_i\}$ Attributes of persons (expression, pose, accessory)

c_i Occurrence of object

Learning & Inference

- Learning
  - Noun mapping
  - Update parameters according to empirical probability
- Inference
  - Iterative conditional modes
  - Random selection

Occurrence and Position

Jenny
she
jenny
and
no one

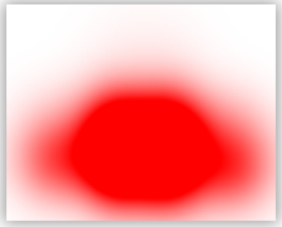
lightning
storm
thunderstorm
lighten
bolt
thunder
weather

helmet
horn
helm
Viking

duck
goose
bird

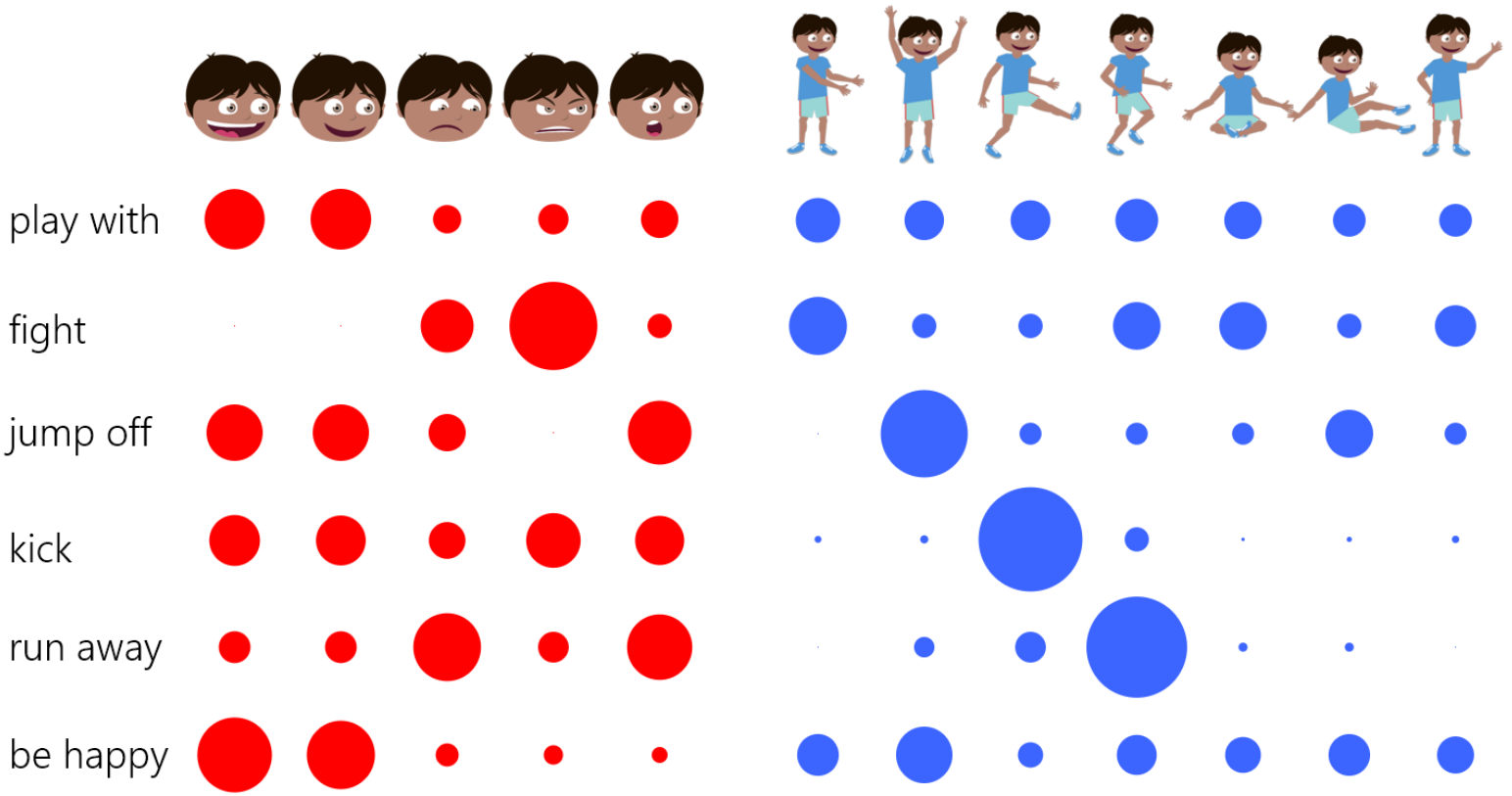
bucket
pail
sand
pale

bonfire
fire
campfire



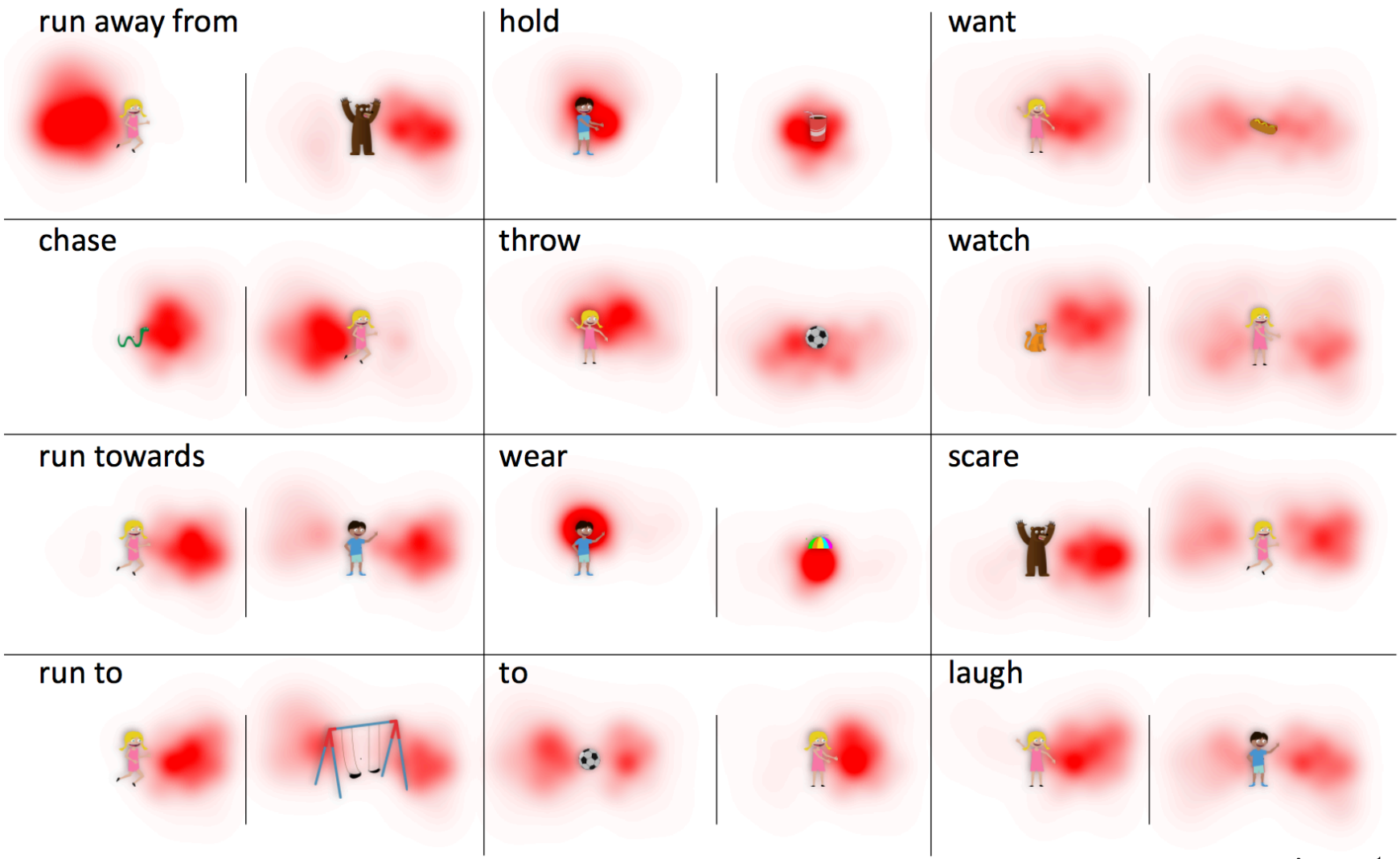
Zitnick, 2014

Attributes



Zitnick, 2013

Relative Location



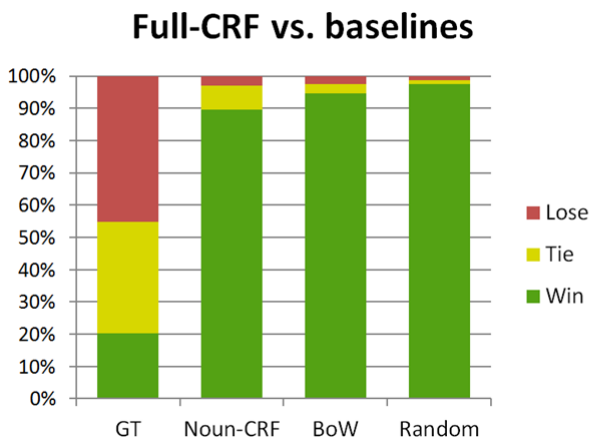
Zitnick, 2014

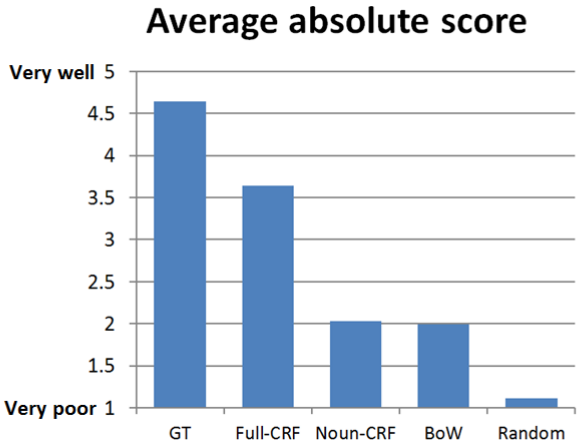
Results

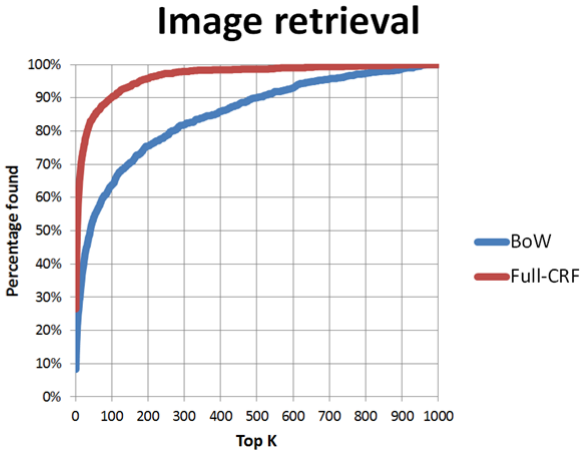
Input Description	Tuples	GT	Full-CRF	BoW	Noun-CRF	Random
Jenny is catching the ball. Mike is kicking the ball. The table is next to the tree.	<<Jenny>, <catch>, <ball>> <<Mike>, <kick>, <ball>> <<table>, <be>, <>>					
Mike is sitting next to Jenny. The catis sitting next to the tree. Jenny is throwing the ball.	<<Mike>, <sit next to>, <Jenny>> <<catis, <sit next to> <tree>> <<Jenny>, <throw>, <ball>>					
Mike is scared of lightning. It is a stormy day. Jenny is standing on the slide.	<<Mike, <be scared>, <>> <<day>, <be, stormy>, <>> <<Jenny>, <stand on>, <slide>>					

Zitnick, 2014

Quantitative Results







Zitnick, 2014

Results

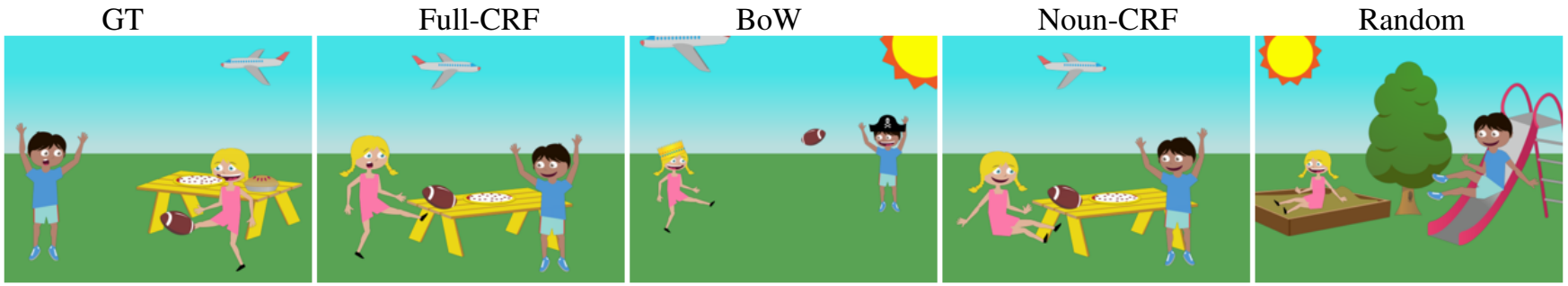


Figure 19: **Input description:** Jenny is kicking the football. The pizza is on the table. The airplane is flying over Jenny. **Tuples:** Jenny kick football; pizza be table; airplane fly:p:over Jenny;

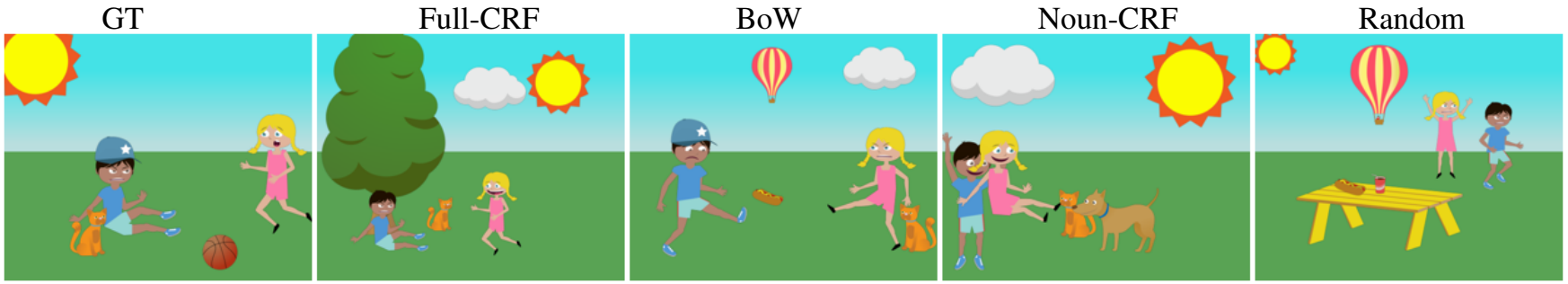


Figure 20: **Input description:** Mike is sitting next to a cat. Mike is angry because he fell down. Jenny is running towards Mike to help him. **Tuples:** Mike sit:p:next to cat; Mike be:pa:angry ; he fall ; Jenny run:p:towards Mike; Jenny help ;

Zitnick, 2014

Failure cases

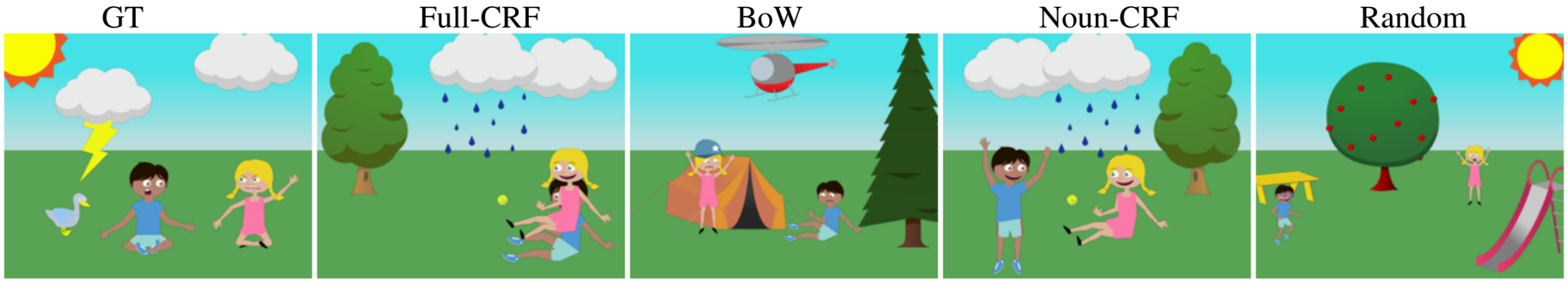


Figure 21: **Input description:** It is lighting out. Mike and Jenny are upset. Mike and Jenny are sitting on the ground with there legs crossed. **Tuples:** it light ; Mike sit ground; Jenny sit ground; ground with leg;

Failed sentence parsing, relative location prior

Figure 43: **Input description:** Mike is mad his ice melted. Jenny is scared of the bear. The bear is wearing a viking hat. **Tuples:** Mike be:pa:mad ; Jenny be:pa:scared ; bear wear hat;

Rare co-occurrence

Zitnick, 2014

Conclusion

- Conclusion
  - New approach for learning “common sense” knowledge about our visual world.
  - Don’t need to wait for object recognition to be solved.
- Future Works
  - Better language model?
  - Larger photorealistic dataset?

Text to 3D Scene

Figure 8: *The bird is in the bird cage*. *The bird cage is on the chair.*

Coyne, 2001

WordsEye

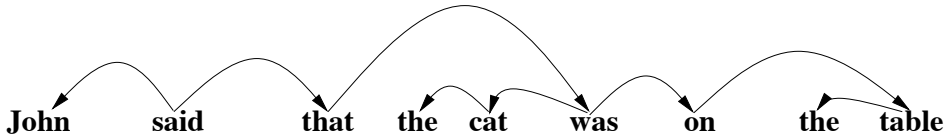


Figure 2: Dependency structure for *John said that the cat was on the table*..

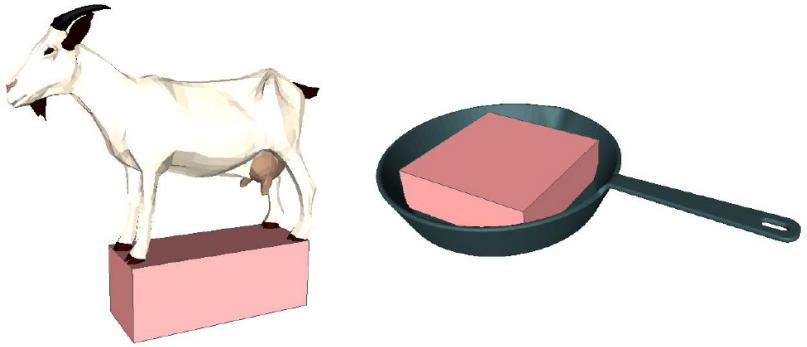


Figure 6: Spatial tags for “base” and “cup”.

Figure 11: *John rides the bicycle*. *John plays the trumpet*.

Coyne, 2001

Objects not depicable

- **Texturalization**
- **Emblematization**
  - *Light bulb* for *idea*, *church* for *religion*
- **Characterization**
  - *Football player* will wear a *football helmet*
- **Convention icon**
  - *Don't think*
- **Degeneralization**
  - *Chair* for *furniture*

Text to 3D Scene

Figure 15: *The blue daisy is not in the army boot*.

Figure 16: *John does not believe the radio is green*.

Coyne, 2001

Text to 3D Scene

Figure 14: *The cat is facing the wall*.

Figure 17: *The devil is in the details*.

Coyne, 2001

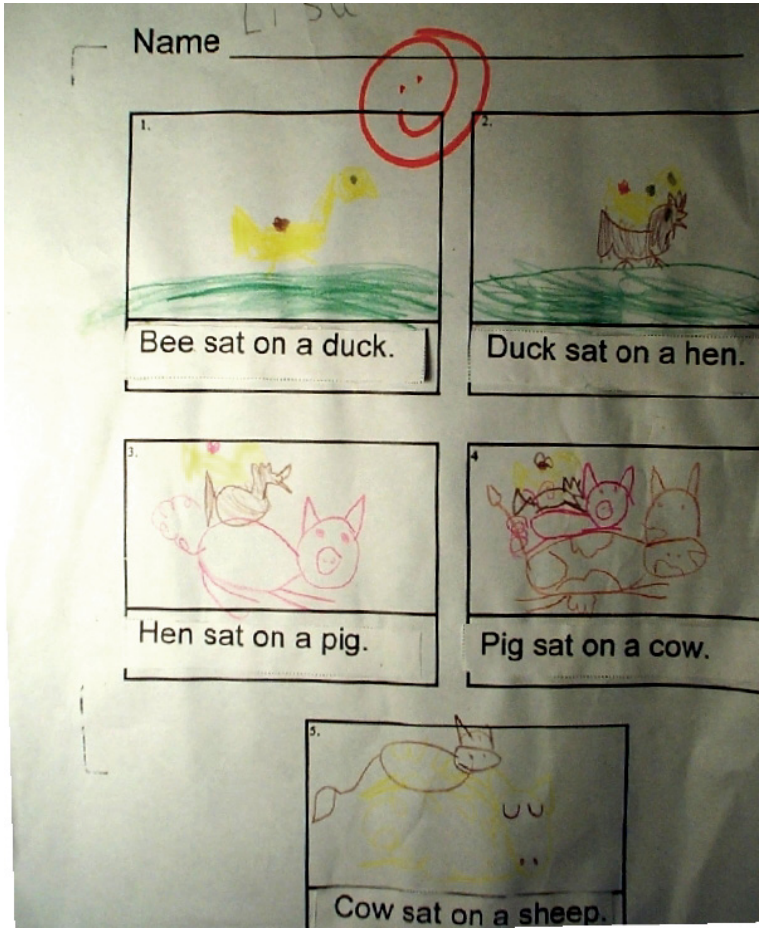
WordsEye 2014

the large radio is on the small car. the large woman is 8 feet behind the car. she is facing the car. the woman is unreflective. the small chair is 2 feet to the east of the car. the small chair is facing the car. the small barn is 5 feet to the left of the woman. the small barn is facing the woman. the large plant is on the chair. the chair is white. the small dog is under the chair. the large pig is .2 feet to the right of the dog. the pig is unreflective. the pig is facing the dog. the man is 1 feet in front of the car. he is facing the car. the man is unreflective. it is sunset. the ground is dark texture. camera-light is red. the light is 5 feet above the plant.

Welcome home! And how are you?

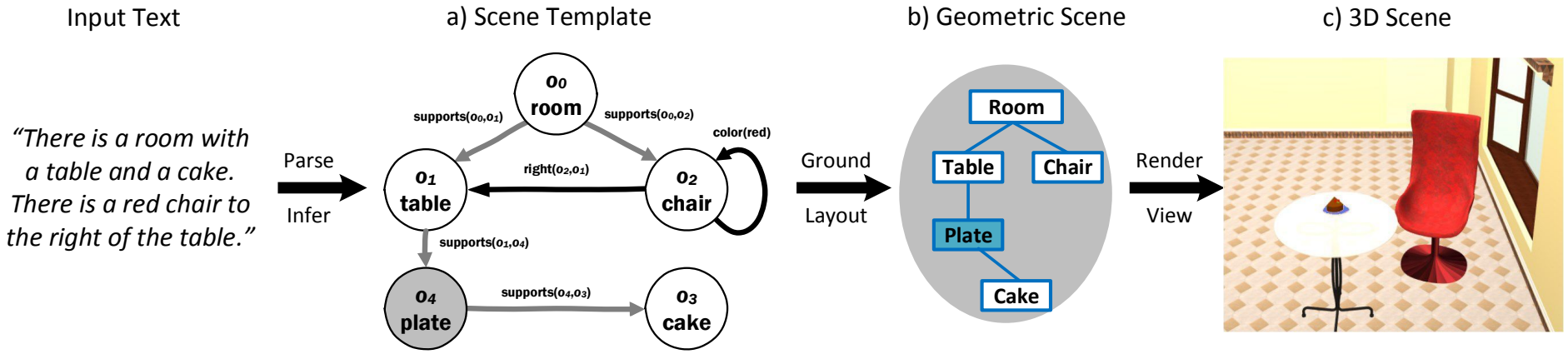
the large radio is on the small car. the large woman is 8 feet behind the car. she is facing the car. the woman is unreflective. the small chair is 2 feet to the east of the car. the small chair is facing the car. the small barn is 5 feet to the left of the woman. the small barn is facing the woman. the large plant is on the chair. the chair is white. the small dog is under the chair. the large pig is .2 feet to the right of the dog. the pig is unreflective. the pig is facing the dog. the man is 1 feet in front of the car. he is facing the car. the man is unreflective. it is sunset. the ground is dark texture. camera-light is red. the light is 5 feet above the plant.

Text to 3D Scene



Coyne, 2001

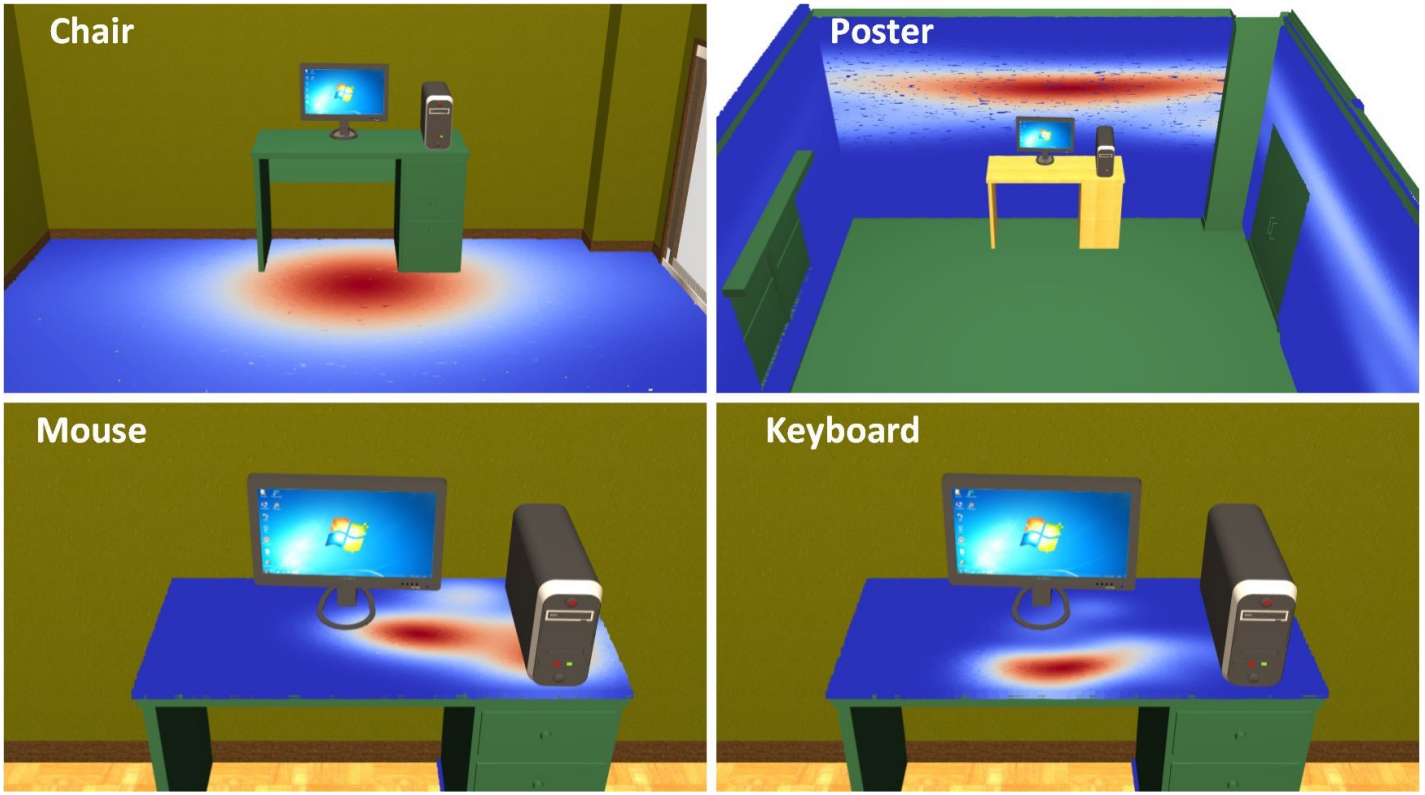
Text to 3D Scene Generation



Learning Spatial Knowledge for Text to 3D Scene Generation A. Chang, M. Savva, C. Manning, EMNLP 2014

Chang, 2014

Learned relative position



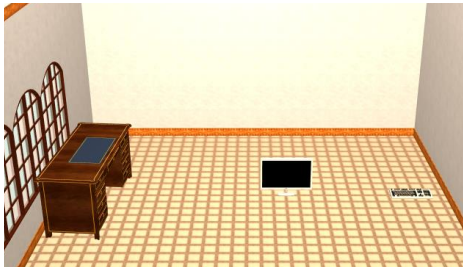
Chang, 2014

Text to 3D Scene Generation

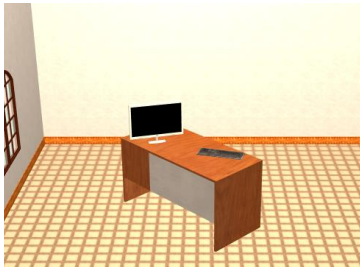
Input Text

"There is a desk and a keyboard and a monitor."

Basic



+Support Hierarchy



+Relative Positions

"There is a coffee table and there is a lamp behind the coffee table. There is a chair in front of the coffee table."

No Relations

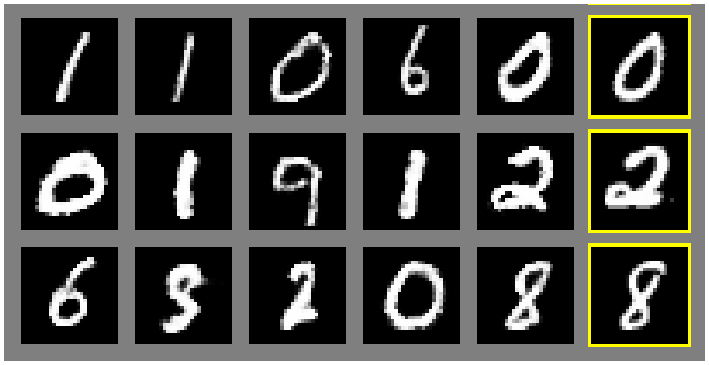
Predefined Relations

Learned Relations

Learning **support relation**, **occurrence**, **spatial relation**, **co-occurrence**

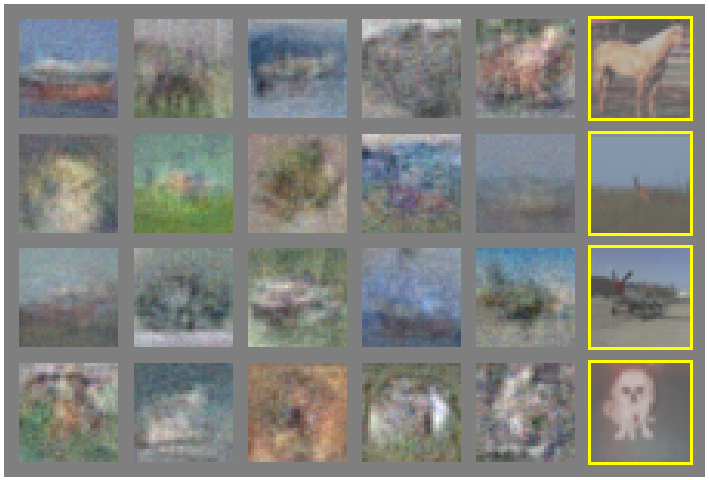
Chang, 2014

Generating Real Image is difficult..

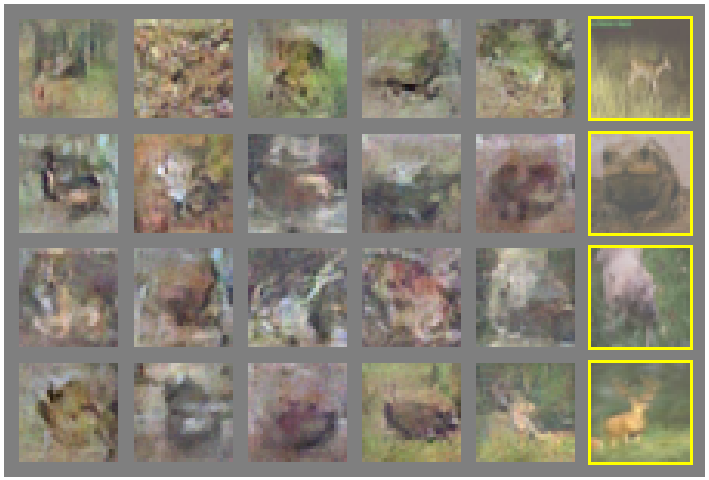


a)

b)



c)



d)

Goodfellow, 2001

Ian Goodfellow, et al. "Generative adversarial nets." Advances in Neural Information Processing Systems. 2014.