# ECS 289G: Visual Recognition
# Winter 2018

Yong Jae Lee

Department of Computer Science

# Plan for today

- Questions?
- Sign-up for paper
- Research overview

# Success in modern visual recognition research

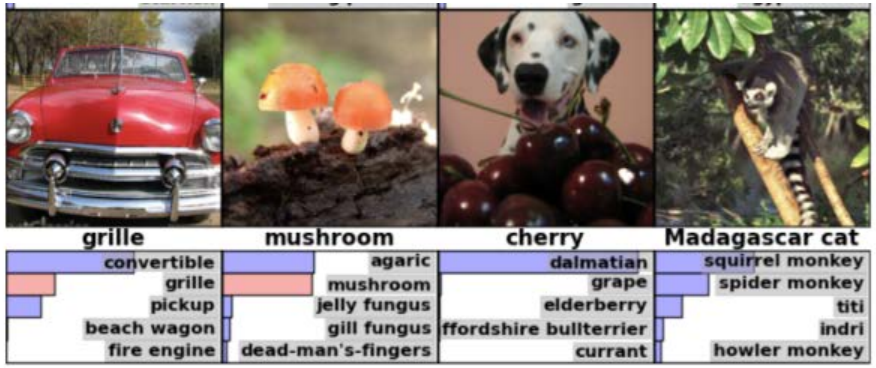
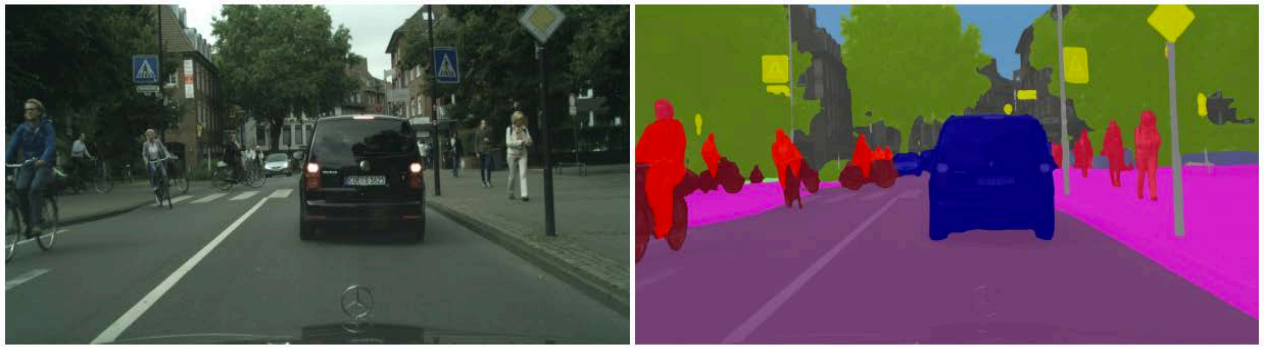
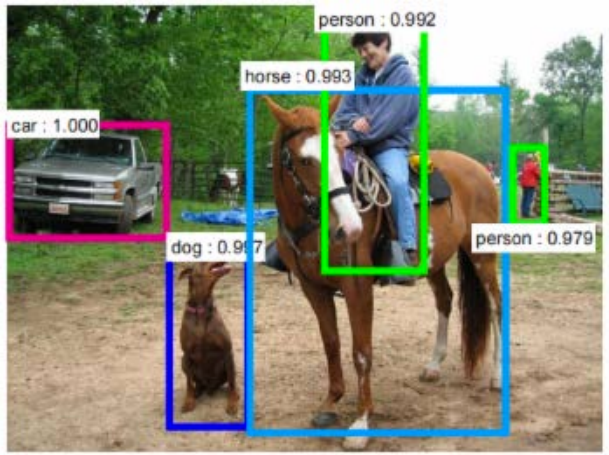



Image classification



Semantic segmentation


Pose recognition



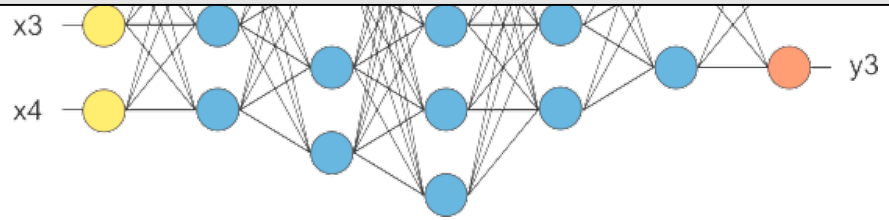

Object detection

… and many more

# Ingredients for success today

1. Fast computation (GPUs)


3. Network architecture




Which ingredient will be the *bottleneck* for tomorrow's success?

2. Large labeled image dataset

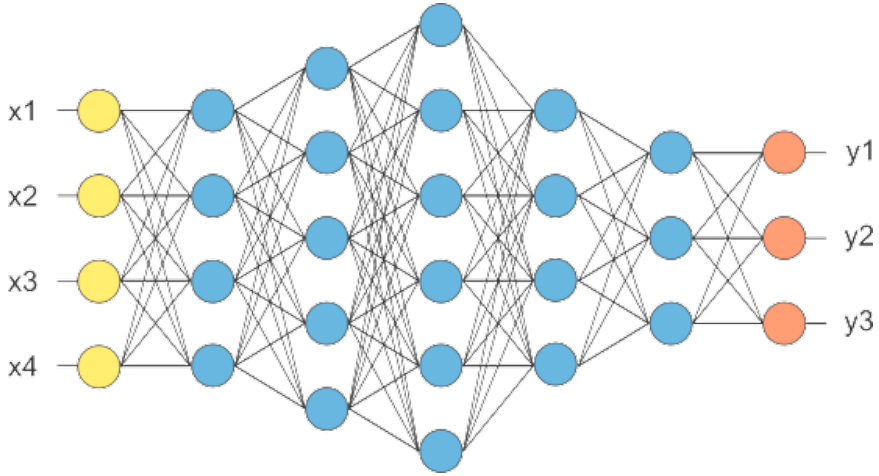
# Ingredients for success today

1. Fast computation (GPUs)

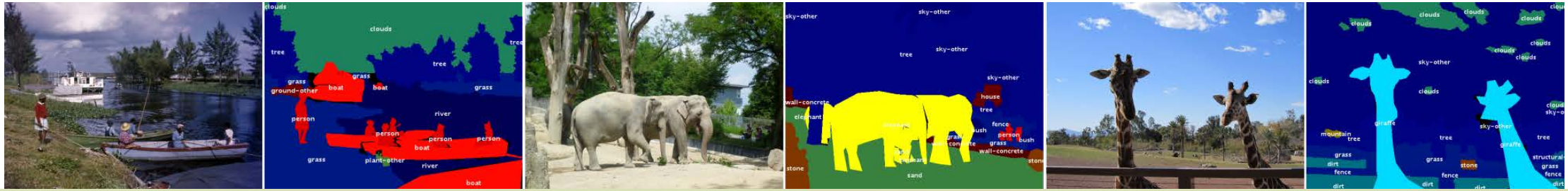

2. Large **labeled** image dataset

Requires *expensive, detailed* human supervision

3. Network architecture



# Take semantic segmentation as an example…



70,000+ annotation hours for 328K images but *only* 80 object categories (MS COCO)

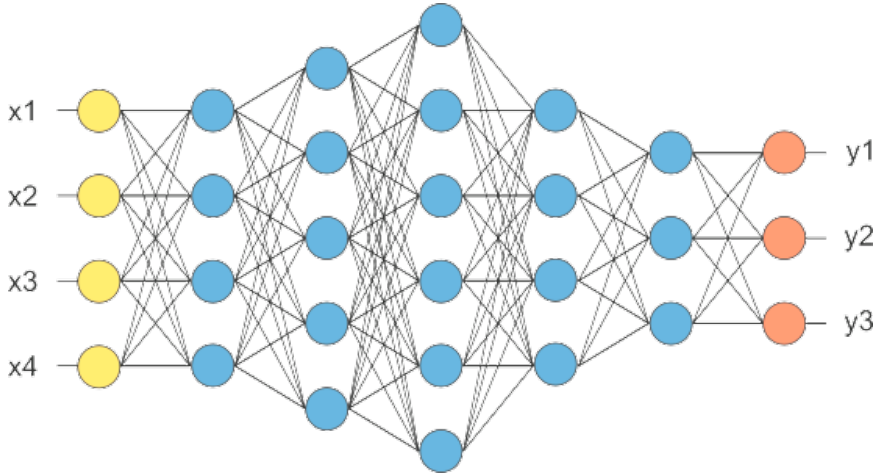
Requires pixel-level semantic labels

# Ingredients for success today

1. Fast computation (GPUs)


2. Large ***weakly-labeled*** image dataset

3. Network architecture



# Fully-supervised vs. Weakly-supervised

[Fully-supervised:](#)
*Pixel-level* labels

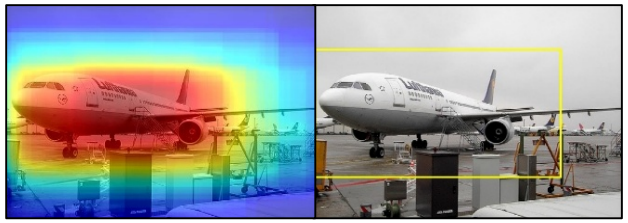

[Weakly-supervised:](#)
*Image-level* labels

{Person, umbrella, hat, … }

# **Overview:** Weakly-supervised visual recognition

1. Learning object detectors with tags

plane

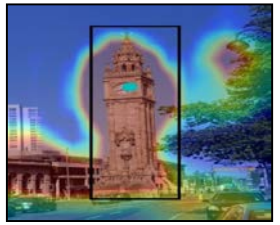


2. Learning visual attributes with pairwise comparisons

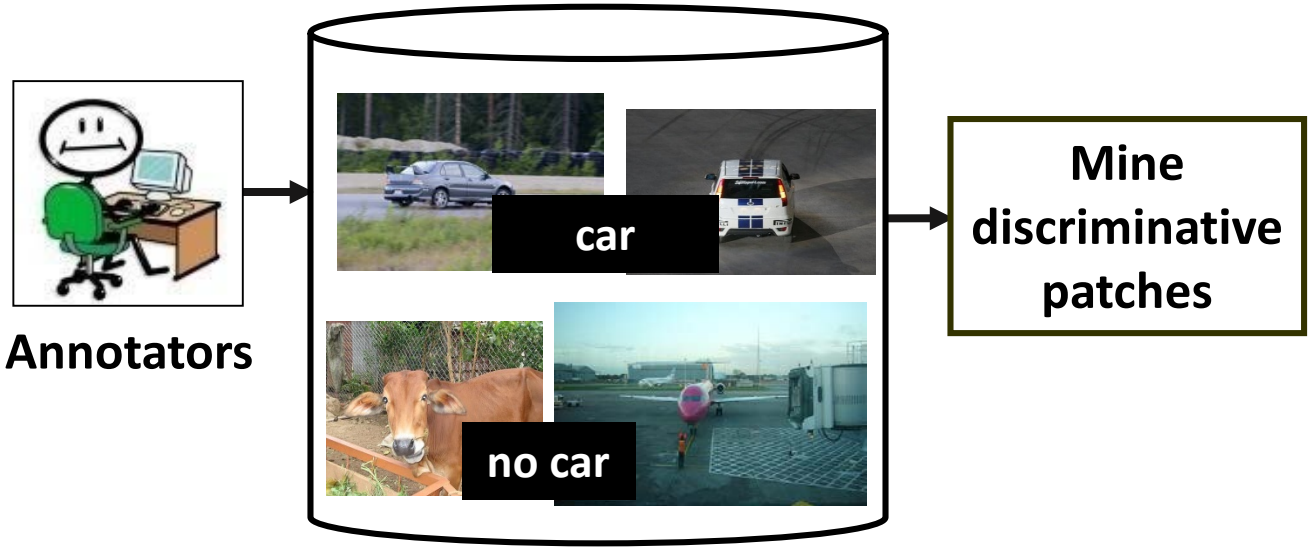
smile

3. Learning visual grounding with captions

“clock tower is tall”

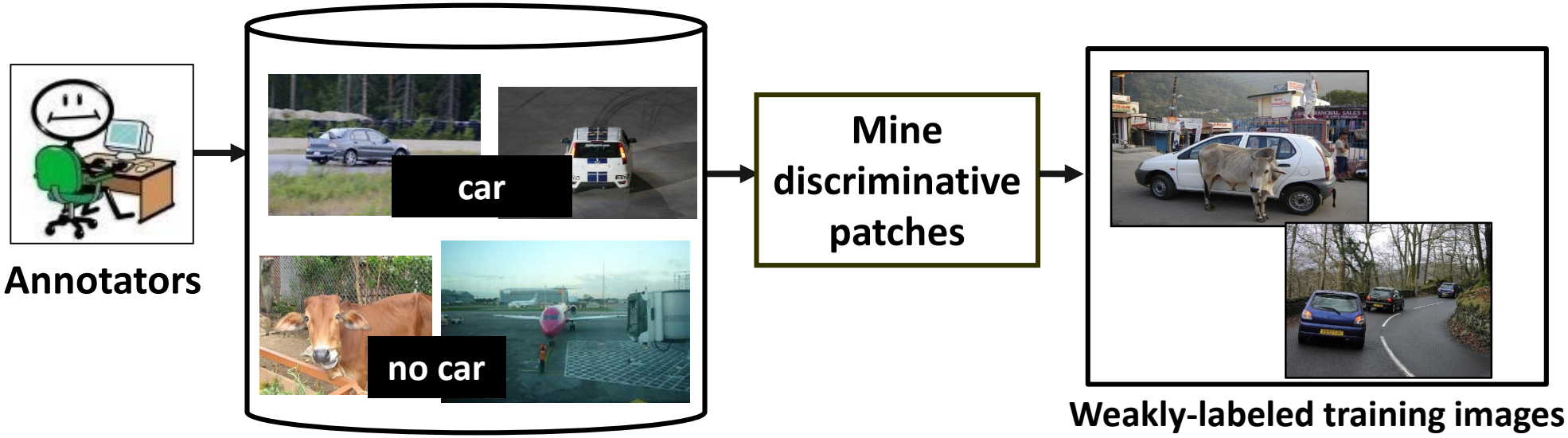


# Weakly-supervised object detection



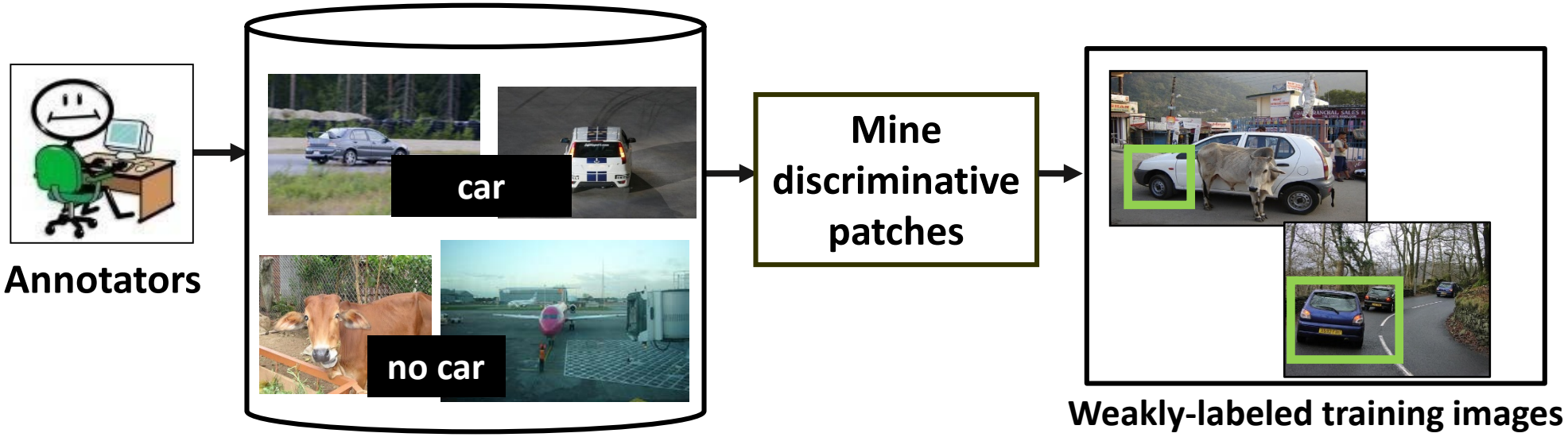

[Weber et al. 2000, Pandey & Lazebnik 2011, Deselaers et al. 2012, Song et al. 2014, …]

# Weakly-supervised object detection



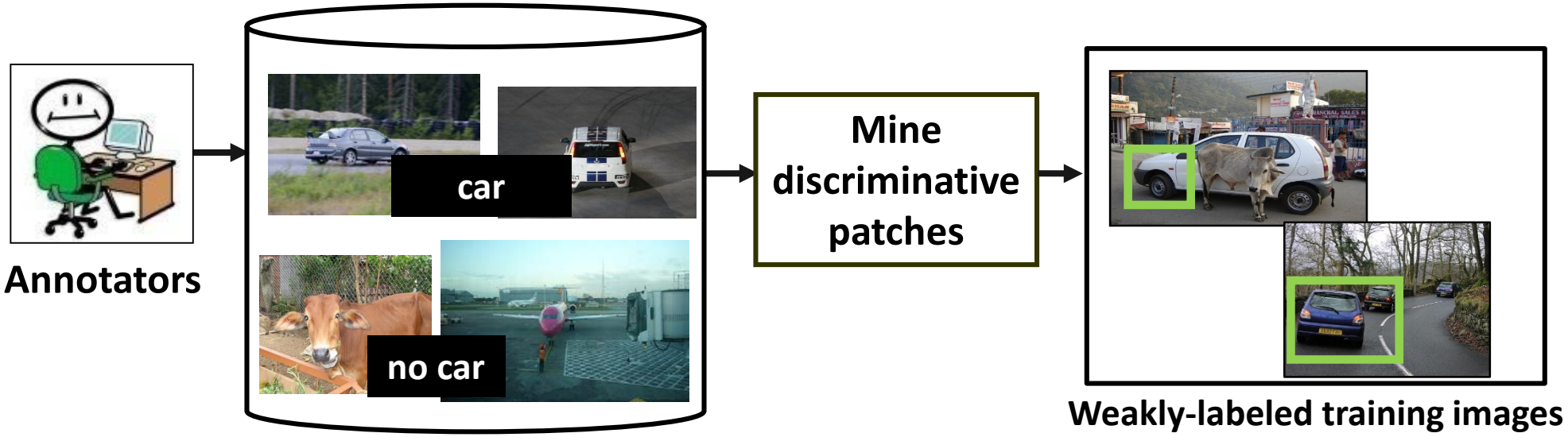

[Weber et al. 2000, Pandey & Lazebnik 2011, Deselaers et al. 2012, Song et al. 2014, …]

# Weakly-supervised object detection




[Weber et al. 2000, Pandey & Lazebnik 2011, Deselaers et al. 2012, Song et al. 2014, …]

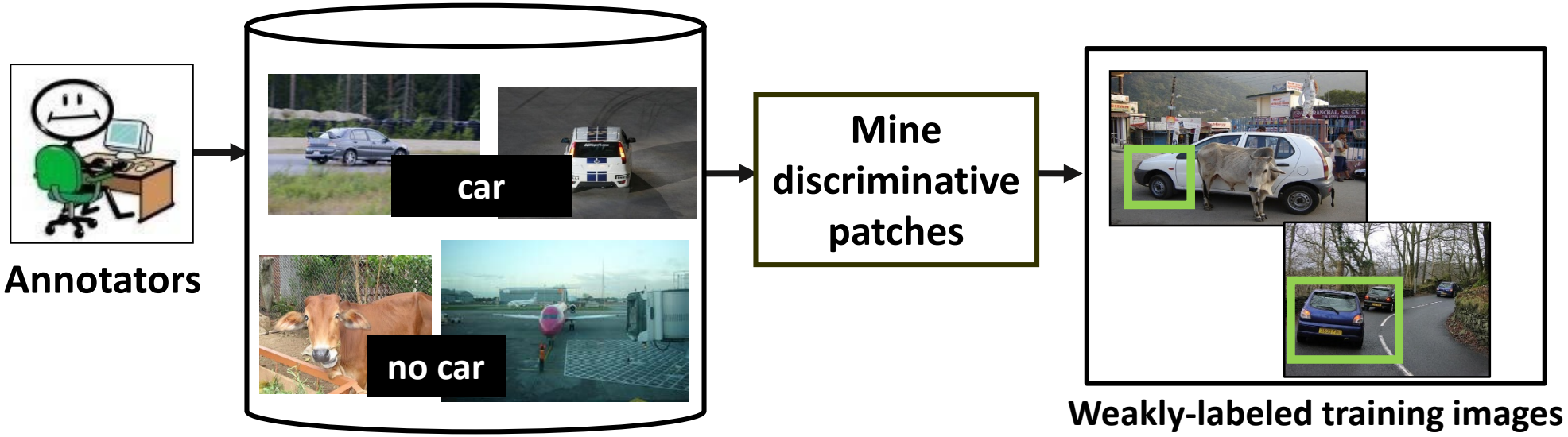
# Weakly-supervised object detection




[Weber et al. 2000, Pandey & Lazebnik 2011, Deselaers et al. 2012, Song et al. 2014, …]

- Supervision is provided at the *image-level* → scalable!

# Weakly-supervised object detection




[Weber et al. 2000, Pandey & Lazebnik 2011, Deselaers et al. 2012, Song et al. 2014, …]

- Supervision is provided at the *image-level* → scalable!
- Due to intra-class appearance variations, occlusion, clutter, mined regions correspond to object-part or include background

# Our idea: Track and transfer

Weakly-labeled *videos* tagged with "`car`"

...

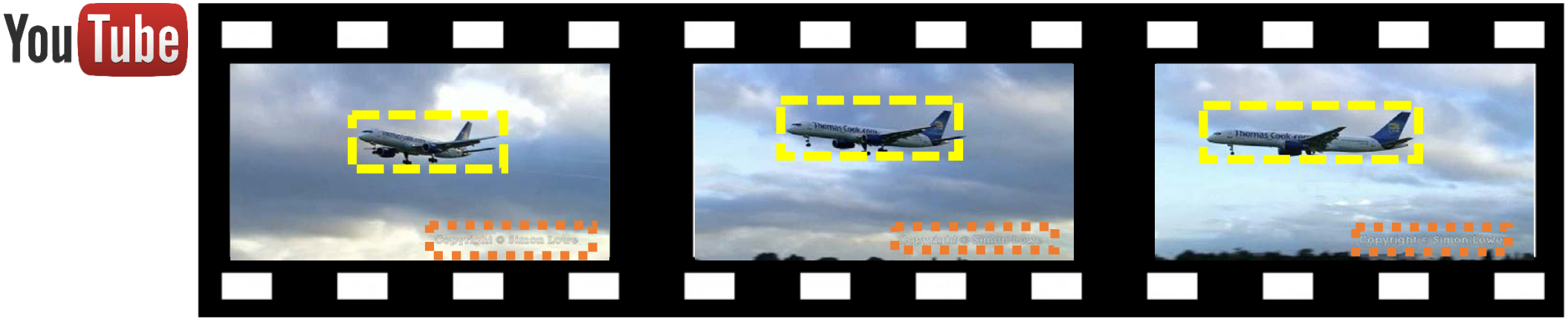
Weakly-labeled *training images* tagged with "`car`"

[Singh, Xiao, Lee, "Track and Transfer: Watching Videos to Simulate Strong Human Supervision for Weakly-Supervised Object Detection", CVPR 2016]

# Transferring tracked object boundaries


Mined positive image region

# Transferring tracked object boundaries

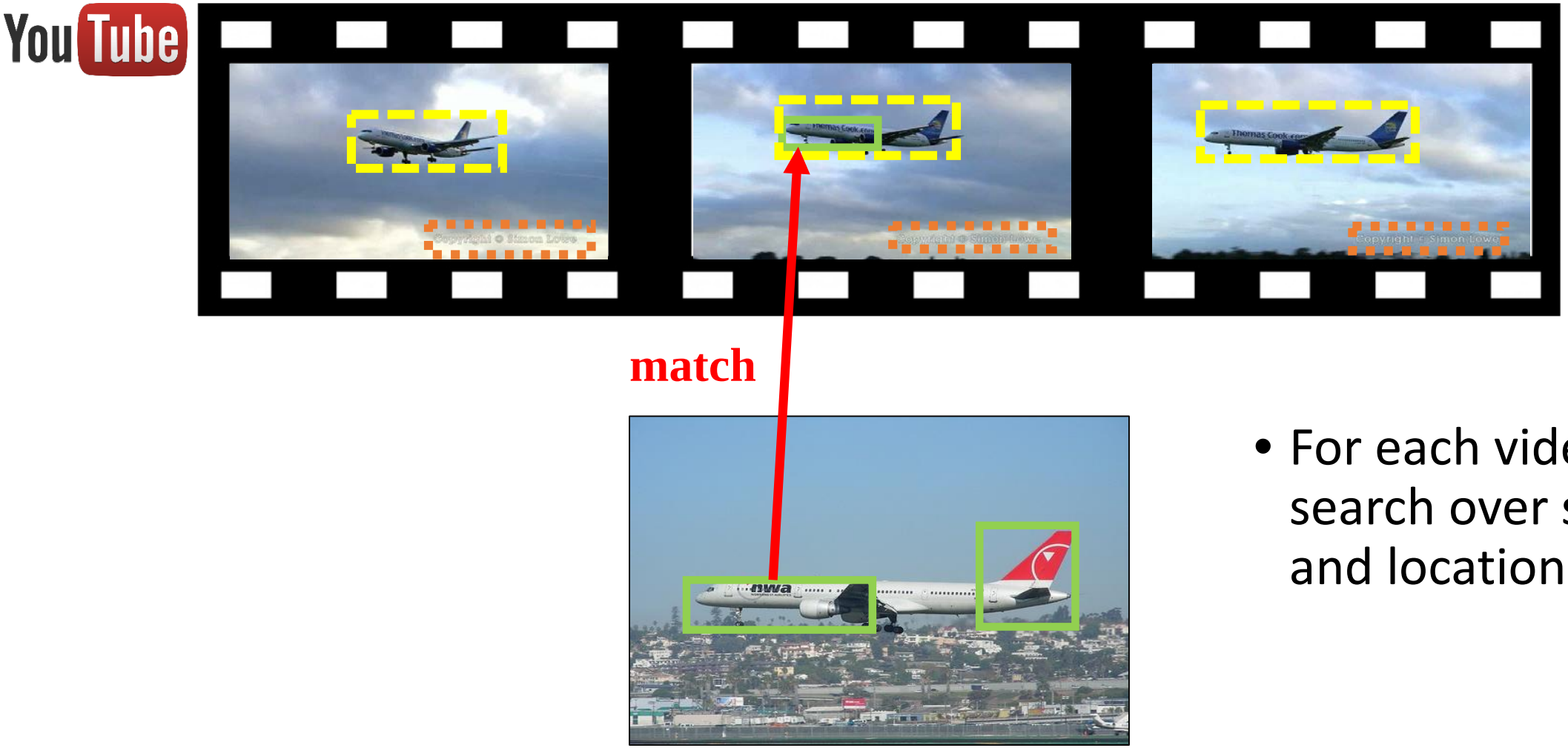


[Unsupervised video object proposals, Xiao and Lee, CVPR 2016]

Mined positive image region

- For each video frame, search over scales and locations

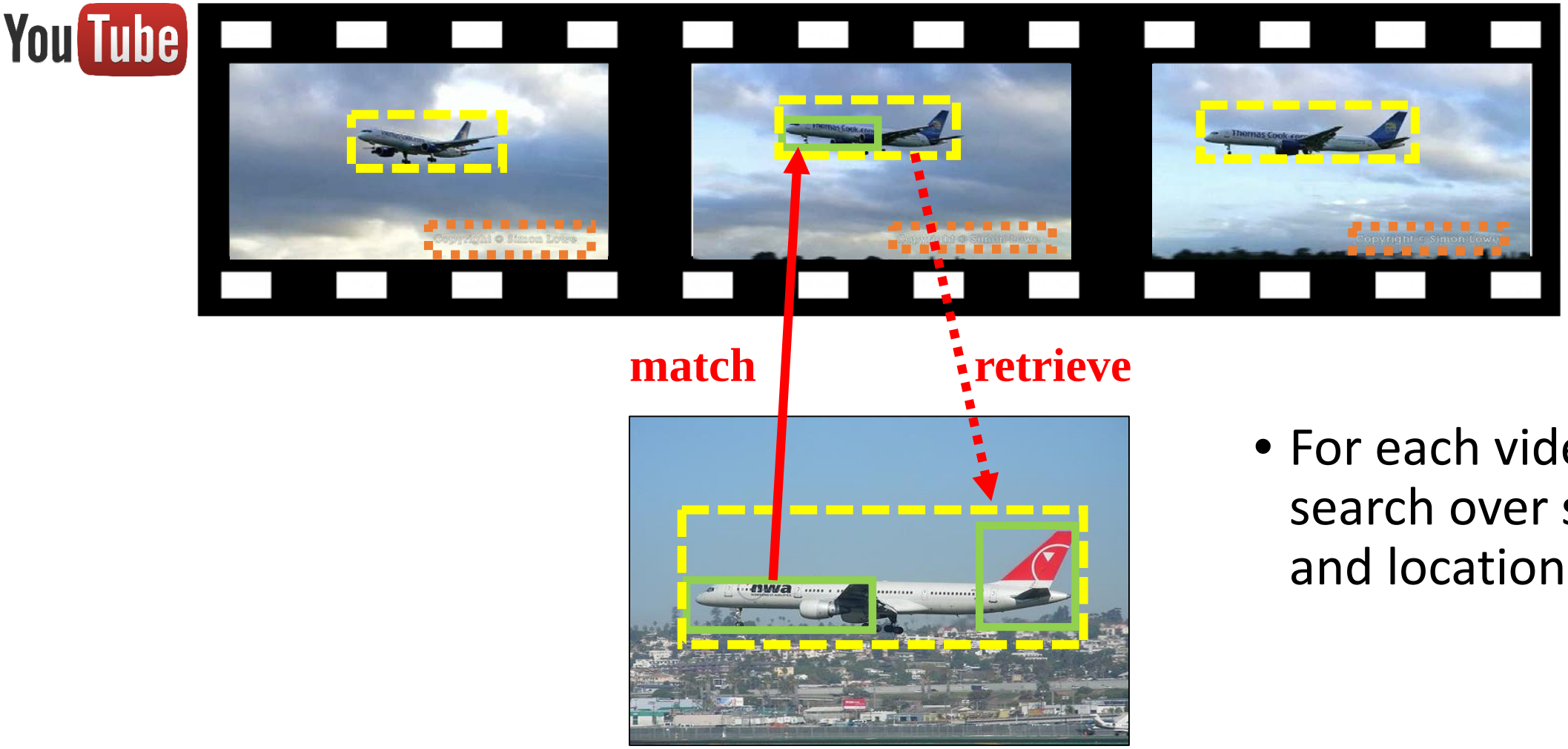
# Transferring tracked object boundaries




Mined positive image region

- For each video frame, search over scales and locations

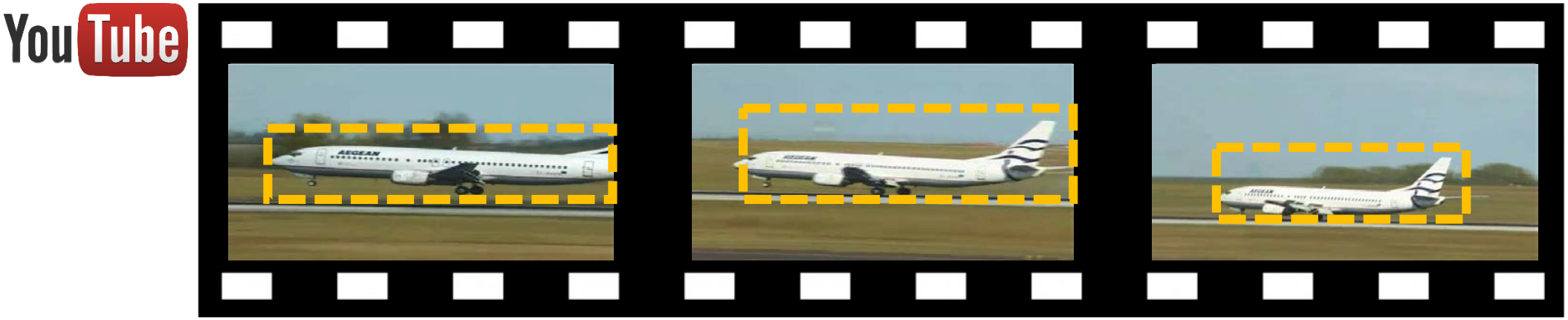
# Transferring tracked object boundaries




Mined positive image region

- For each video frame, search over scales and locations

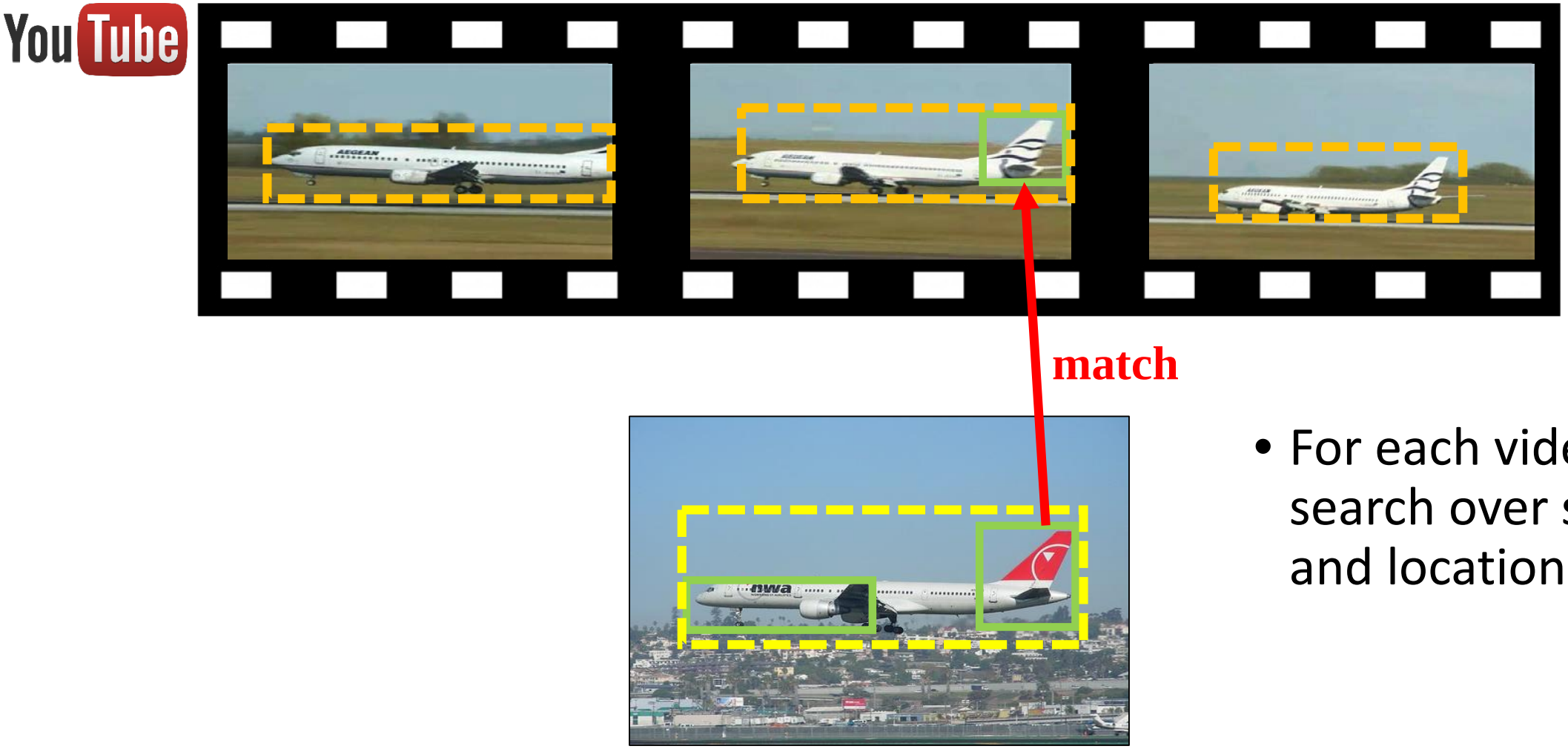
# Transferring tracked object boundaries



Mined positive image region

- For each video frame, search over scales and locations

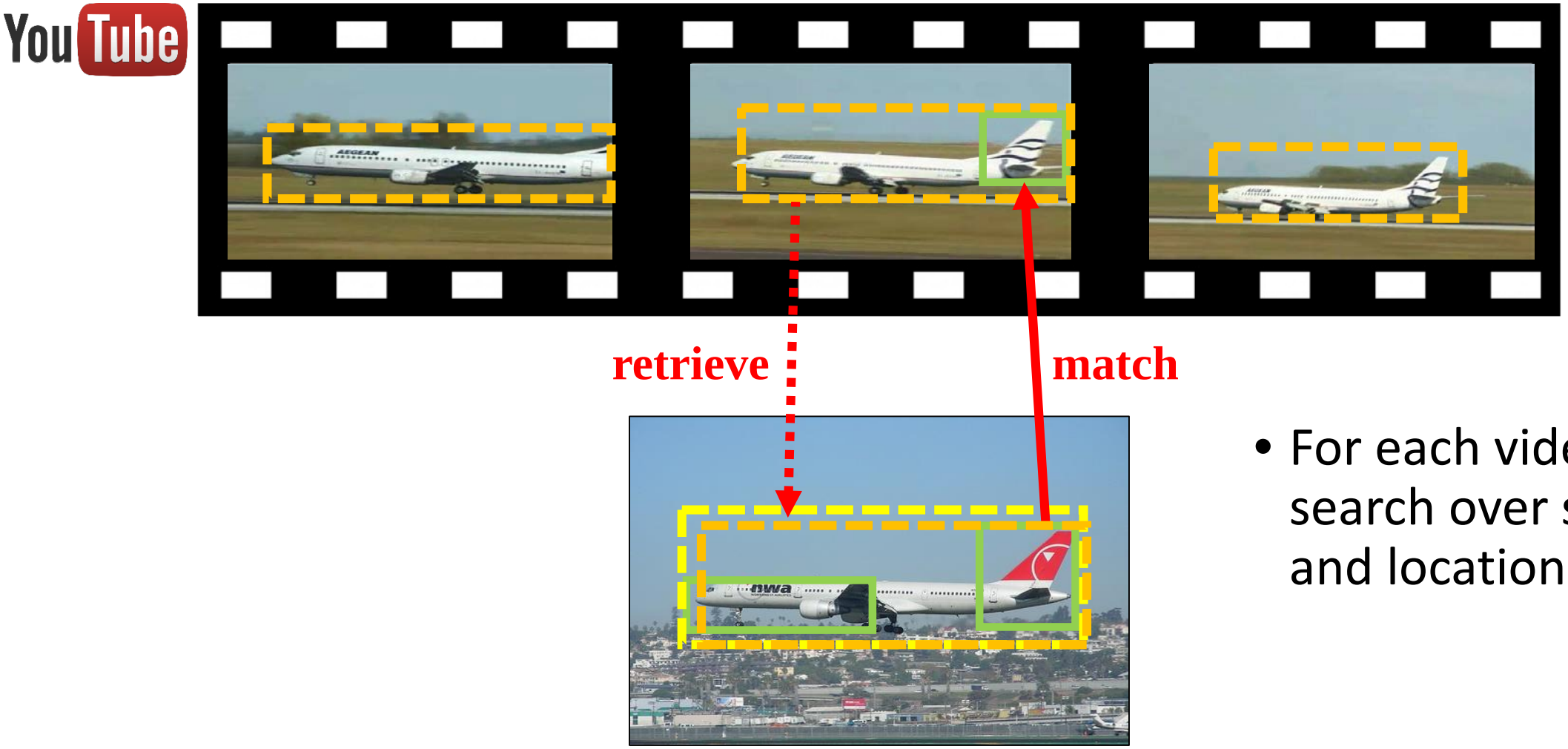
# Transferring tracked object boundaries




Mined positive image region

- For each video frame, search over scales and locations

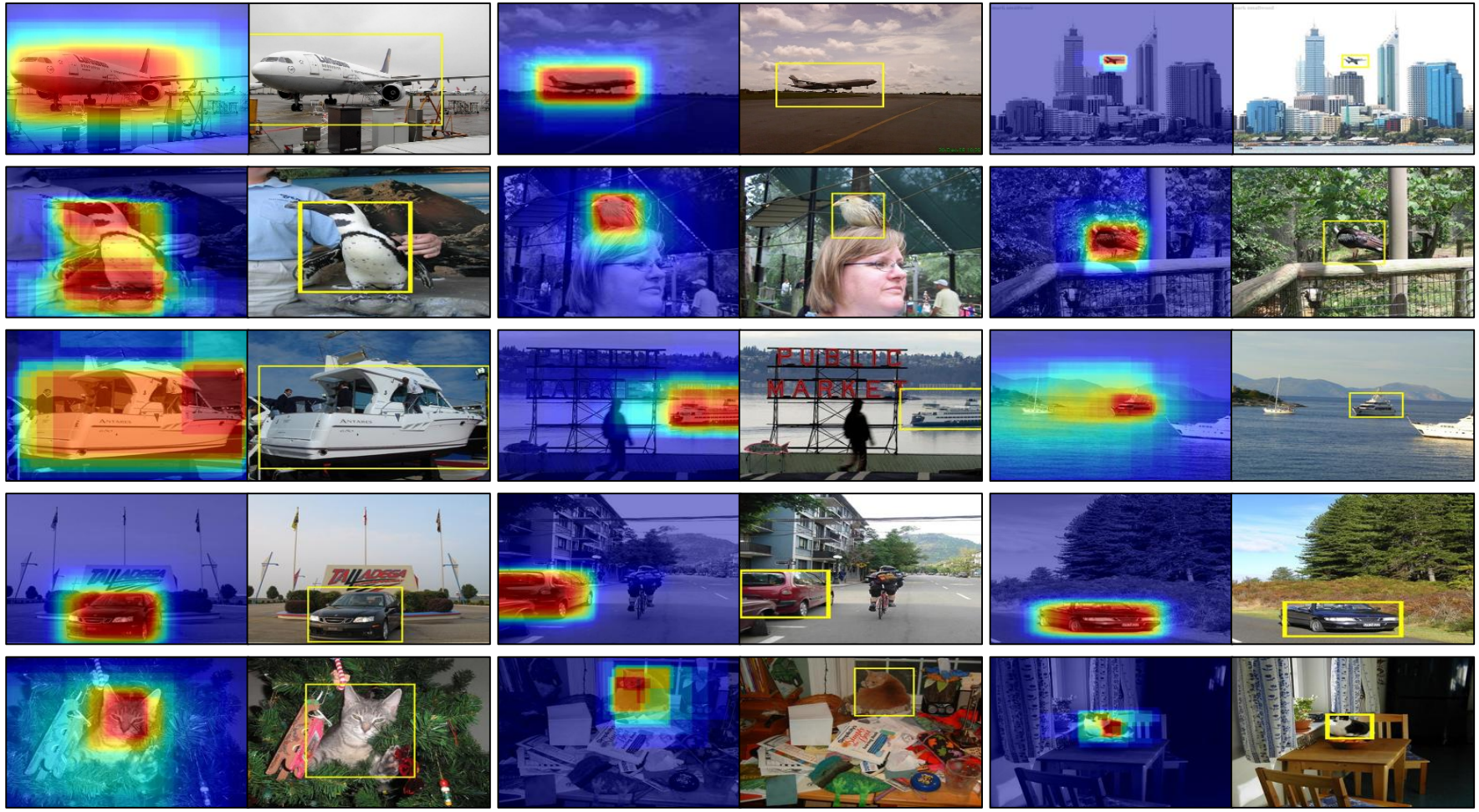
# Transferring tracked object boundaries




Mined positive image region

- For each video frame, search over scales and locations

# Discovered pseudo ground-truth boxes




- Handles atypical pose, partial occlusion, and clutter

# Weakly-supervised object detection accuracy

| VOC 2007 test | aero | bird | boat | car | cat | cow | dog | horse | mbike | train | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Song et al., 2014 [33] | 27.6 | 19.7 | 9.1 | 39.1 | 33.6 | 20.9 | 27.7 | 29.4 | 39.2 | 35.6 | 28.2 |
| Song et al., 2014 [34] | 36.3 | 23.3 | 12.3 | 46.6 | 25.4 | 23.5 | 23.5 | 27.9 | 40.9 | 37.7 | 29.7 |
| Bilen et al., 2014 [1] | 42.2 | 23.1 | 9.2 | 45.1 | 24.9 | 24.0 | 18.6 | 31.6 | 43.6 | 35.9 | 29.8 |
| Wang et al., 2014 [37] | 48.9 | 26.1 | 11.3 | 40.9 | 34.7 | **34.7** | 34.4 | 35.4 | **52.7** | 34.8 | 35.4 |
| Cinbis et al., 2015 [4] | 39.3 | 28.8 | **20.4** | 47.9 | 22.1 | 33.5 | 29.2 | 38.5 | 47.9 | 41.0 | 34.9 |
| Ours | **53.9** | **37.7** | 13.7 | **56.6** | **51.3** | 24.0 | **38.5** | **47.9** | 47.0 | **48.4** | **41.9** |

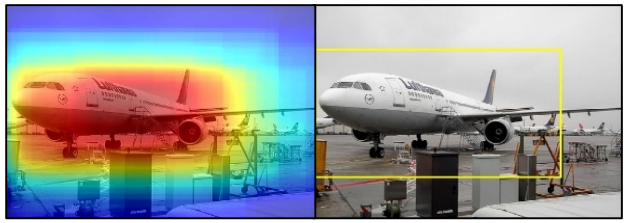
| VOC 2010 test | aero | bird | boat | car | cat | cow | dog | horse | mbike | train | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cinbis et al., 2015 [4] | 44.6 | 25.5 | **14.1** | 36.3 | 23.2 | 26.1 | 29.2 | 36.0 | **54.3** | 31.2 | 32.1 |
| Ours | **53.5** | **37.5** | 8.0 | **44.2** | **49.4** | **33.7** | **43.8** | **42.5** | 47.6 | **40.6** | **40.1** |

- Significantly outperform***ed*** existing weakly-supervised detection methods (at the time)

# **Overview:** Weakly-supervised visual recognition

1. Learning object detectors with tags

plane

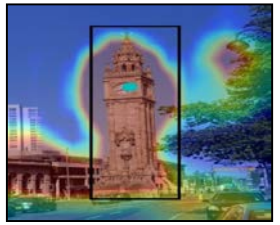


2. Learning visual attributes with pairwise comparisons

smile

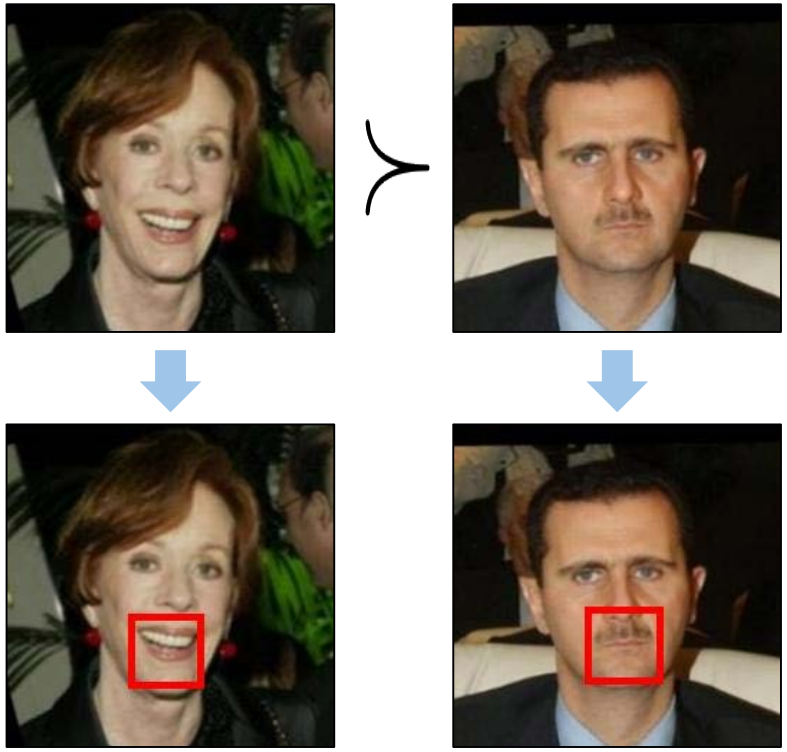
3. Learning visual grounding with captions

“clock tower is tall”



# Goal of weakly-supervised attribute localization

Attribute: `Smile`



- Localizing the relevant region improves attribute modeling
- In the relative setting, we are provided with pairwise comparisons

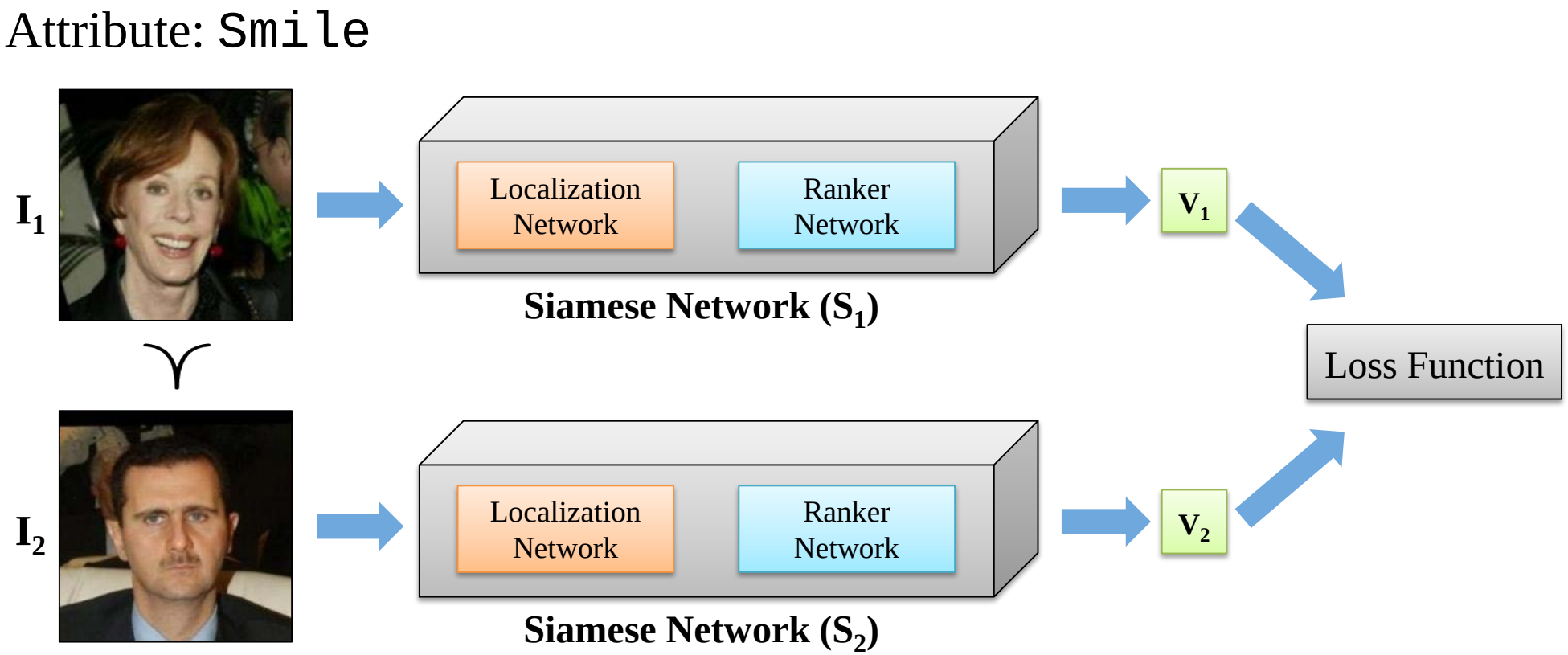
# Previous approaches

- Require pre-trained part detectors or crowdsourcing [Kumar et al. ICCV 2009, Bourdev et al. ICCV 2011, Duan et al. CVPR 2012, Zhang et al. CVPR 2014]
- "Pipeline" where features, localizer, and classifier are trained separately and sequentially; suboptimal and slow [Xiao and Lee, ICCV 2015]

- **Our idea:** *jointly* learn features, localizer, and classifier *end-to-end* in a **deep learning** framework
- We focus on the *relative attribute* setting

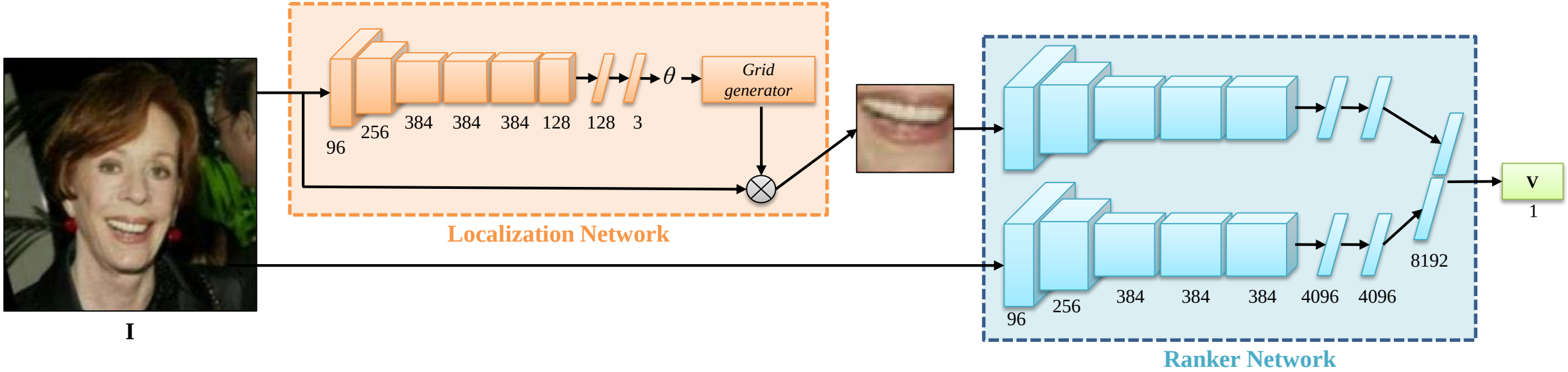
# Overview of our *end-to-end* approach

- **Goal:** Given pairs of ordered training images, simultaneously localize attribute in each image and learn a ranker



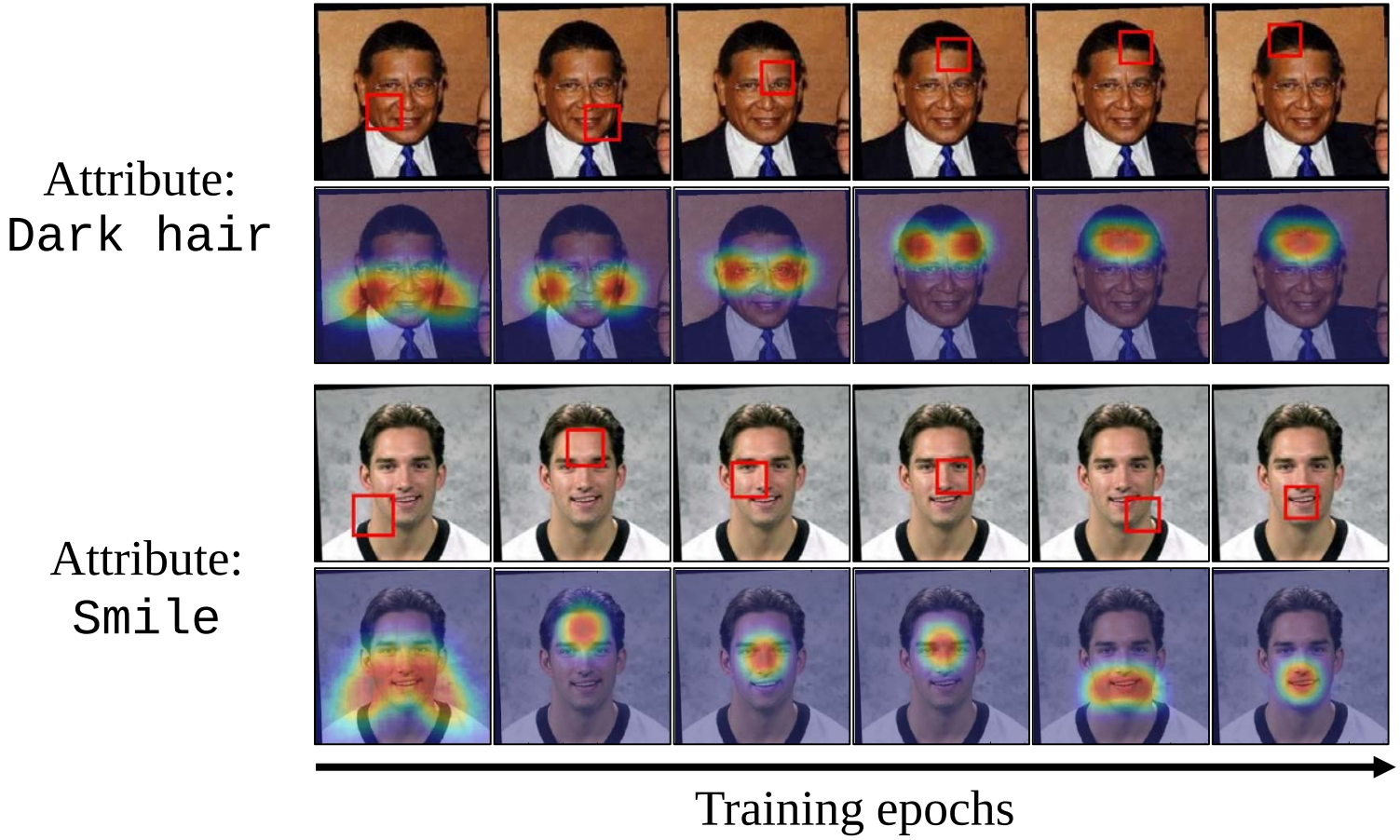

[Singh and Lee, "End-to-End Localization and Ranking for Relative Attributes", ECCV 2016]

# Our *end-to-end* approach



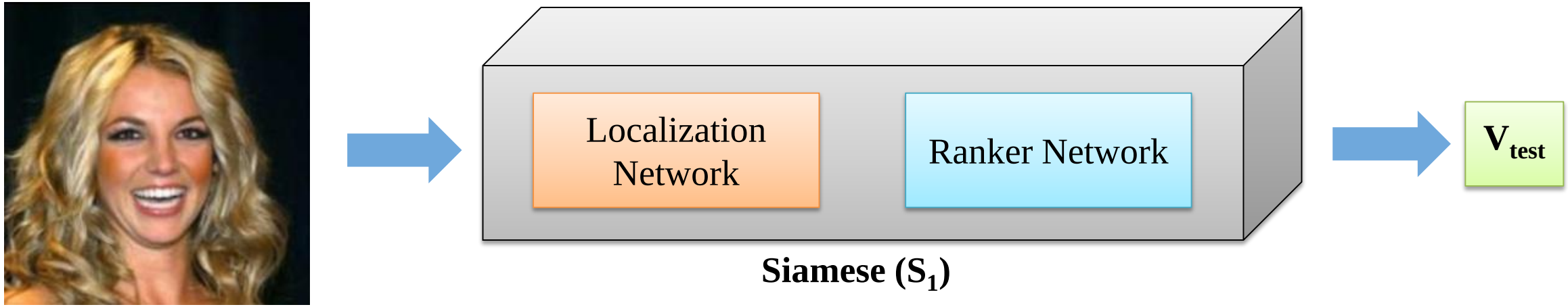

- Ranker network takes the localized region to produce a ranking score
- Combine the global image for global context

# Progression of localized region over training epochs




- Heatmap: distribution of localized region across entire training dataset

# Testing




- Localize the relevant attribute region
- Produce a ranking score for the test image

# Results: Discovered regions and ranking on LFW-10 Faces


- Our network discovers relevant attribute regions
- Leads to accurate rankings

# Results: Discovered regions and ranking UT-Zap50K Shoes

# Results: Image pair ranking accuracy

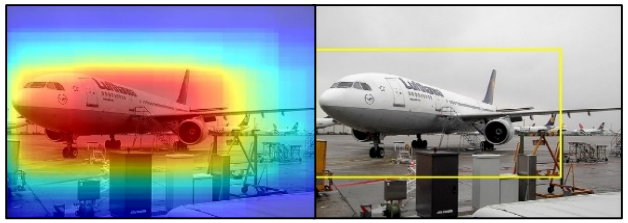
| | BH | DH | EO | GL | ML | MO | S | VT | VF | Y | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Parikh & Grauman [3]+CNN | 78.10 | 83.09 | 71.43 | 68.73 | 95.40 | 65.77 | 63.84 | 66.46 | 81.25 | 72.07 | 74.61 |
| Sandeep et al. [13] | 82.04 | 80.56 | 83.52 | 68.98 | 90.94 | 82.04 | 85.01 | 82.63 | 83.52 | 71.36 | 81.06 |
| Xiao & Lee [14] | 83.21 | 88.13 | 82.71 | **72.76** | 93.68 | 88.26 | **86.16** | 86.46 | **90.23** | 75.05 | 84.66 |
| Ours | **83.94** | **92.58** | **90.23** | 71.21 | **96.55** | **91.28** | 84.75 | **89.85** | 87.89 | **80.81** | **86.91** |

- % of test image pairs whose predicted relative attribute ranking is correct
- State-of-the-art results on LFW-10, UT-Zap50K, OSR, Shoe-with-Attribute

# **Overview:** Weakly-supervised visual recognition

1. Learning object detectors with tags

plane

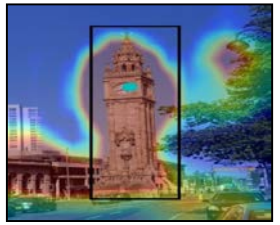


2. Learning visual attributes with pairwise comparisons

smile

3. Learning visual grounding with captions

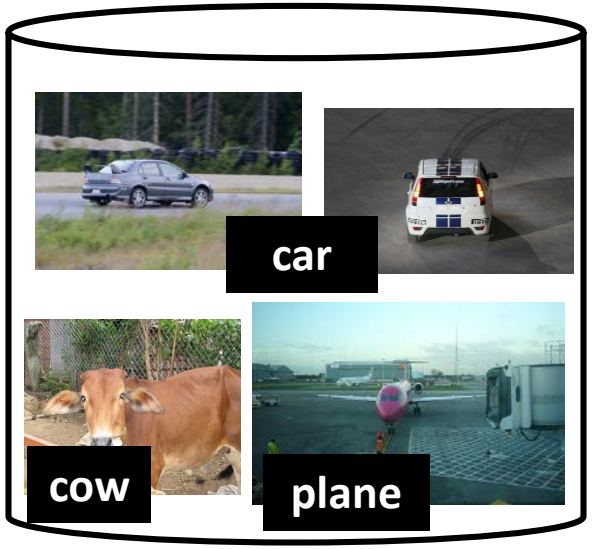
“clock tower is tall”



# Tag-based weakly supervised localization

**Localizing objects**

image + tag

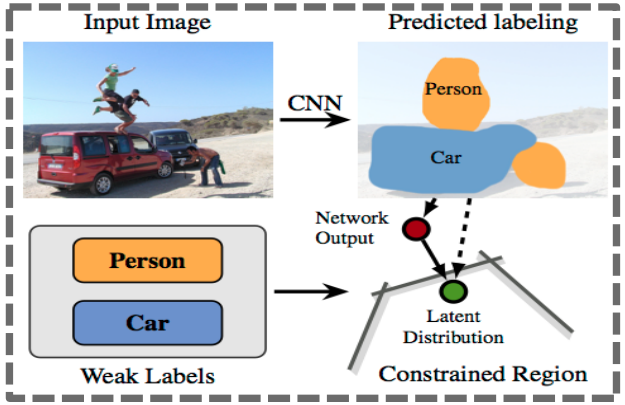


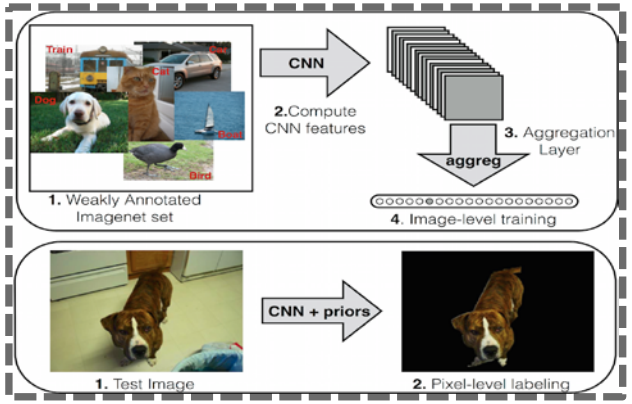

Bounding box detections


Deselaers et al. 2010, Pandey & Lazebnik 2011, Song et al. 2014, Singh et al. 2016, ...

# Tag-based weakly supervised localization

**Semantic segmentation**



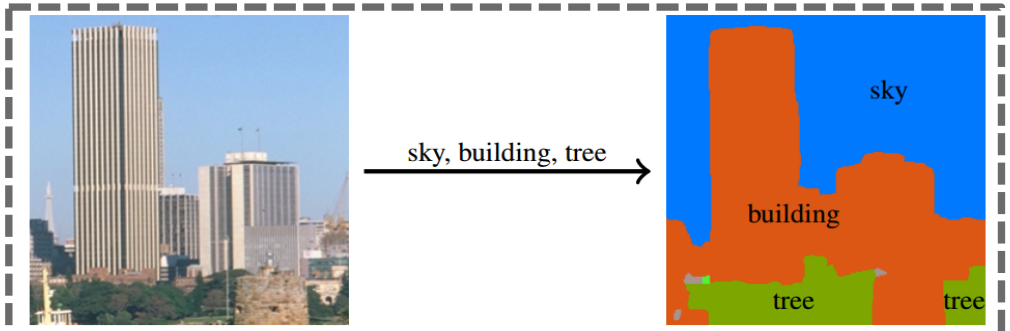




Figure 1. Our approach takes labels in the form of which classes are present in the scene during training, and learns a segmentation model, even though no annotations at the pixel-wise are available.

Xu et al. 2014, Pathak et al. 2015, Pinheiro & Collobert 2015, ...

# Tags vs. natural language

# Tags vs. natural language

Tags: *"cat", "hand", "donut"*

# Tags vs. natural language

Tags: *"cat", "hand", "donut"*

Tags convey very limited amount of information about images

# Tags vs. natural language

Caption: *"A grey cat staring at a hand with a donut."*

# Tags vs. natural language

Caption: *"A grey cat staring at a hand with a donut."*

Much more informative and natural to describe images with caption

# **Our goal:** Leverage structure in natural language for weakly supervised visual grounding

*A grey cat staring at a hand with a donut.*

[Xiao, Sigal, Lee, “Weakly-supervised Visual Grounding of Phrases with Linguistic Structures”, CVPR 2017]

# **Our goal:** Leverage structure in natural language for weakly supervised visual grounding

*A grey cat staring at a hand with a donut.*

[Xiao, Sigal, Lee, “Weakly-supervised Visual Grounding of Phrases with Linguistic Structures”, CVPR 2017]

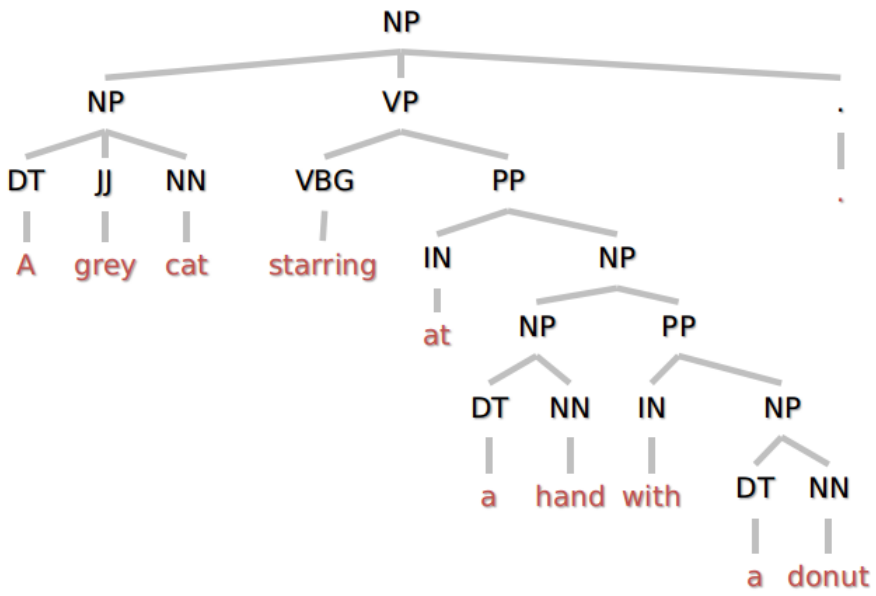
# **Our goal:** Leverage structure in natural language for weakly supervised visual grounding

*A grey cat staring at a hand with a donut.*

*Visual grounding* of *free-from* language with only *image-level supervision*

[Xiao, Sigal, Lee, "Weakly-supervised Visual Grounding of Phrases with Linguistic Structures", CVPR 2017]

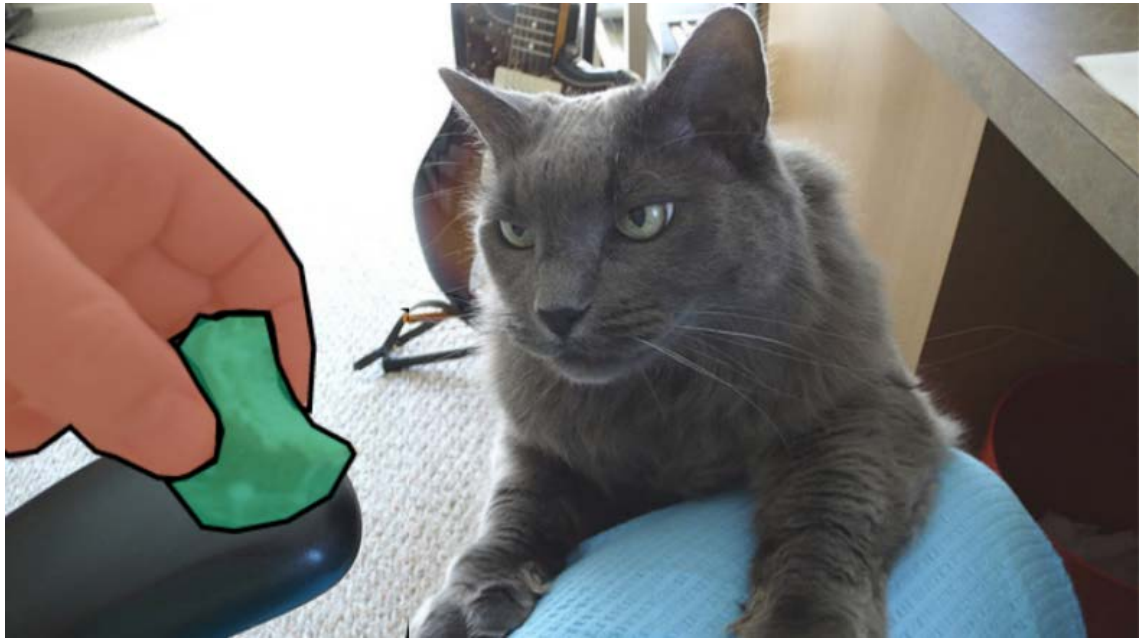
# **Our goal:** Leverage structure in natural language for weakly supervised visual grounding

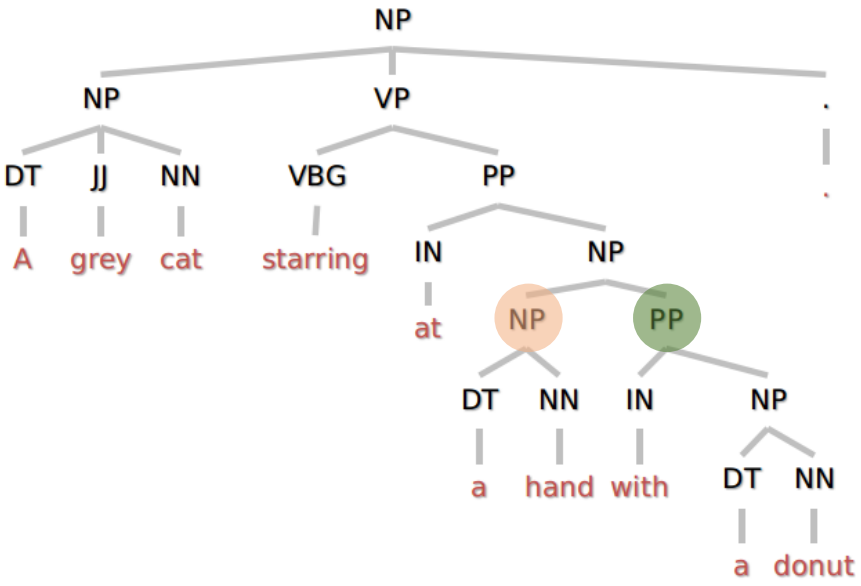



*A grey cat staring at a hand with a donut.*

- Exploit *linguistic structure* in caption
- Parse caption into syntactic parse tree [Socher et al. 2013]

# **Key idea:** Transfer linguistic structure to visual domain

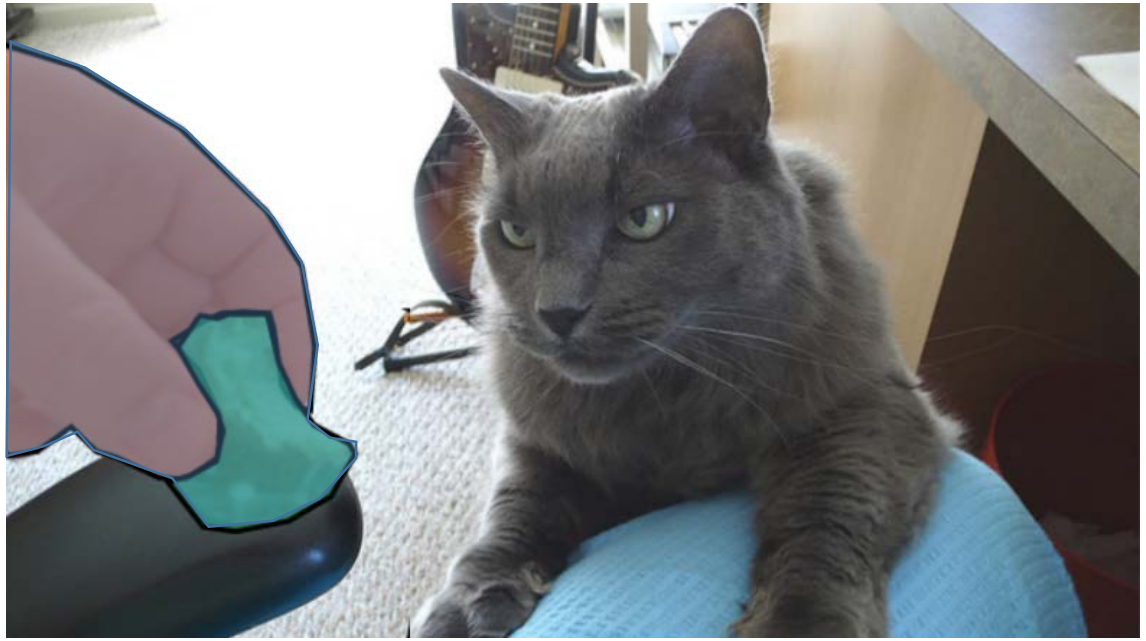


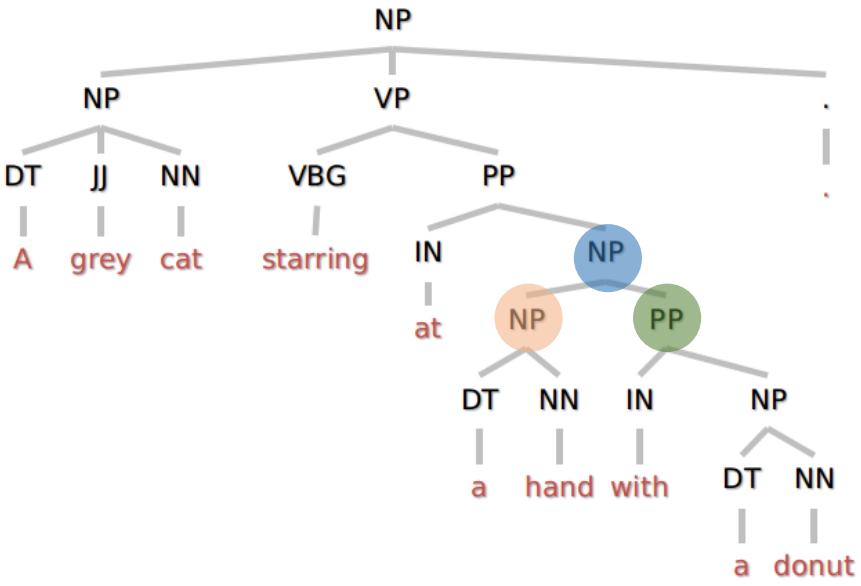



*A grey cat staring at a hand with a donut.*

(1) Implies *sibling exclusivity*:
Image regions for sibling nodes should be *exclusive*

# **Key idea:** Transfer linguistic structure to visual domain

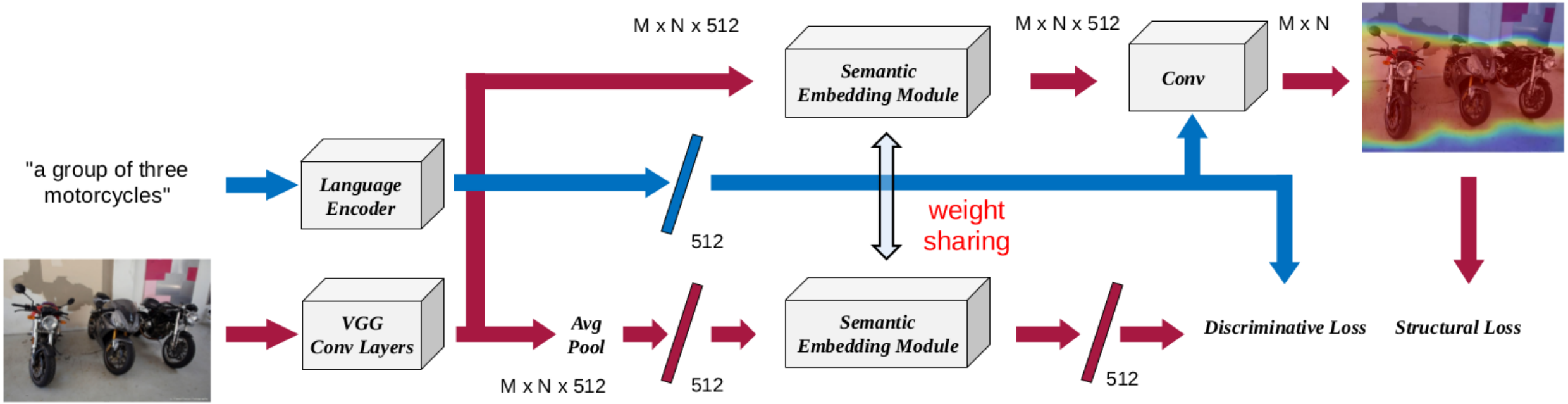





*A grey cat staring at a hand with a donut.*

(2) Implies *parent-child inclusivity*:
Image region of parent node is *union* of children's regions

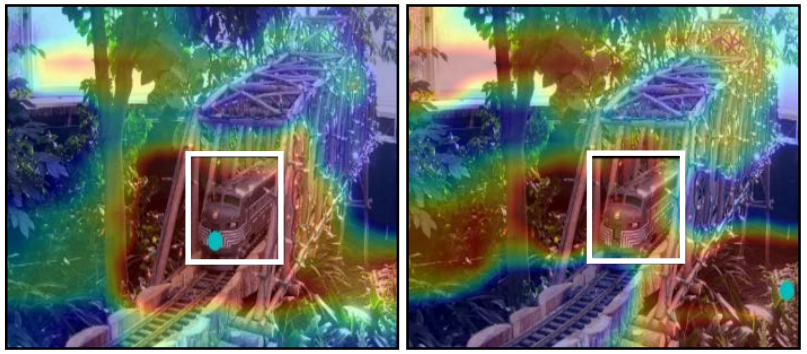
# Network architecture



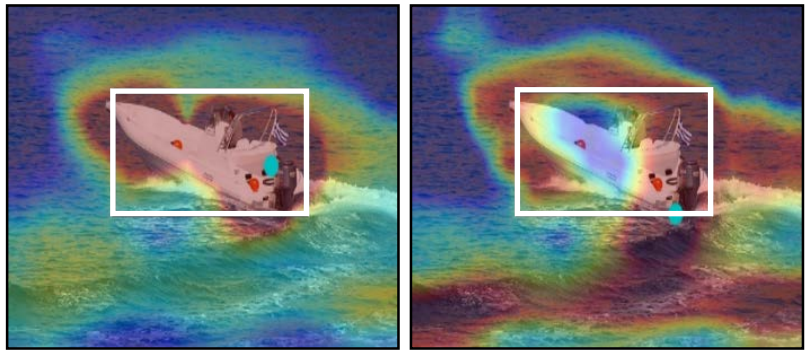

Four submodules: (1) visual encoder, (2) language encoder, (3) semantic embedding module, and (4) loss functions.

# Phrase grounding on Visual Genome

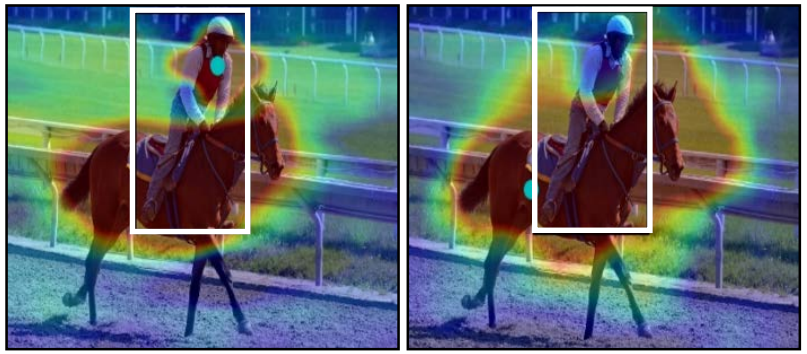
*the train is on the bridge*



*a person driving a boat*



*a jockey riding a horse*



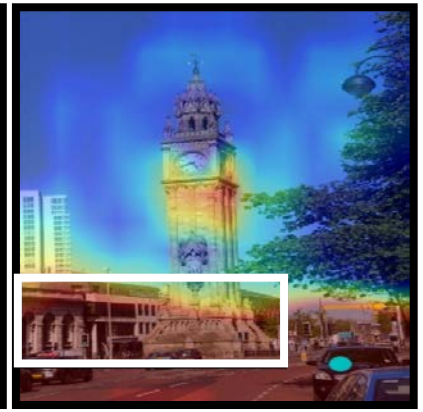
**Left**: our model

**Right**: baseline trained without structural constraints (Discriminative loss only)

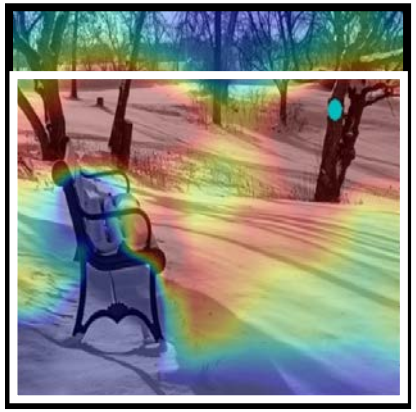
# Multiple queries on Visual Genome

*clock tower is tall*

*buildings by street*

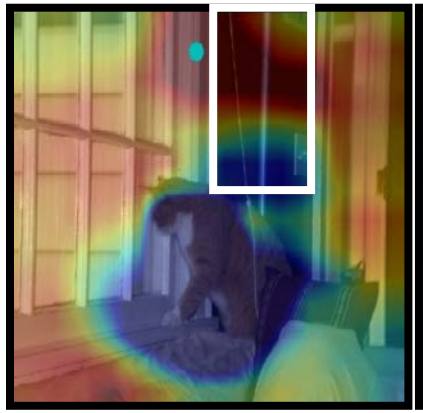


*snow covered park*

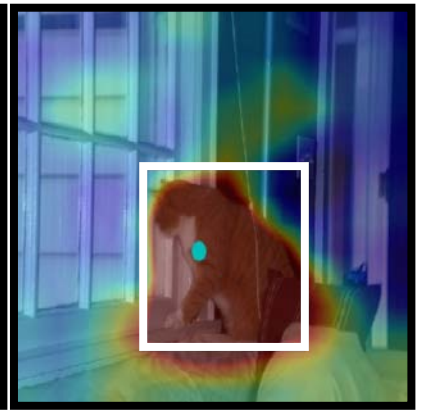


*bench is black and white*

*the walls are dark purple*



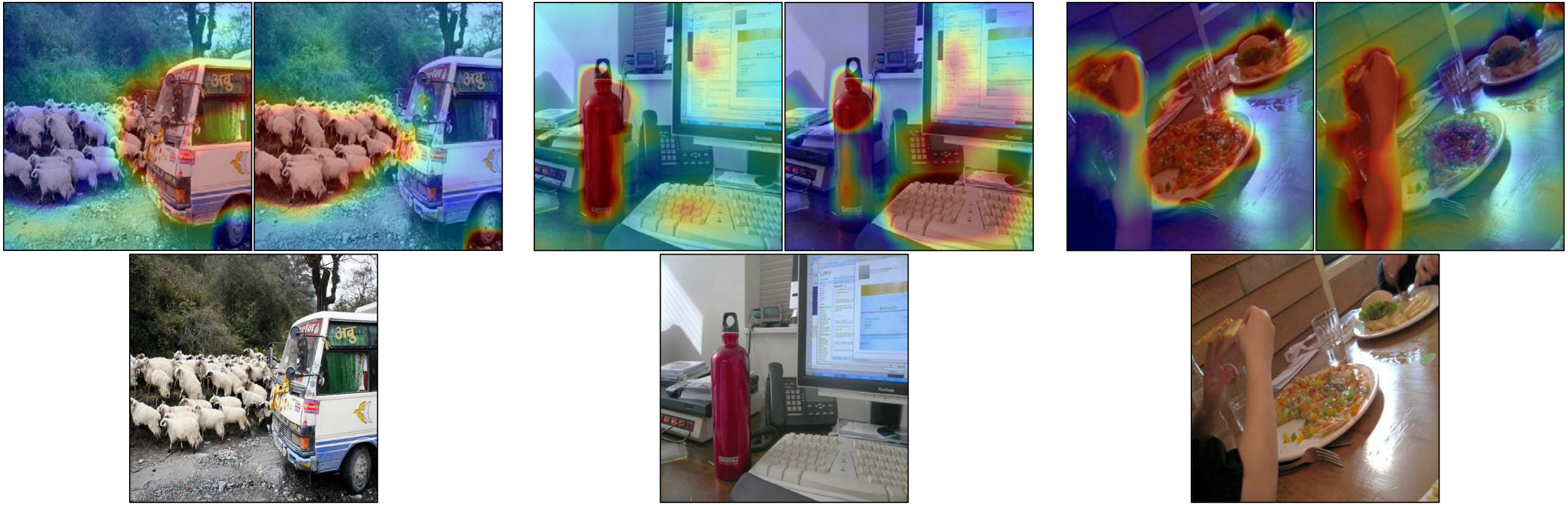
*yellow and orange cat*



Our groundings generated for **different phrases** of the same image

# Multiple queries on MS COCO

*bus* *sheep* *bottle* *keyboard* *pizza* *person*



Our groundings generated for **different tags** of the same image

# Phrase grounding on Visual Genome

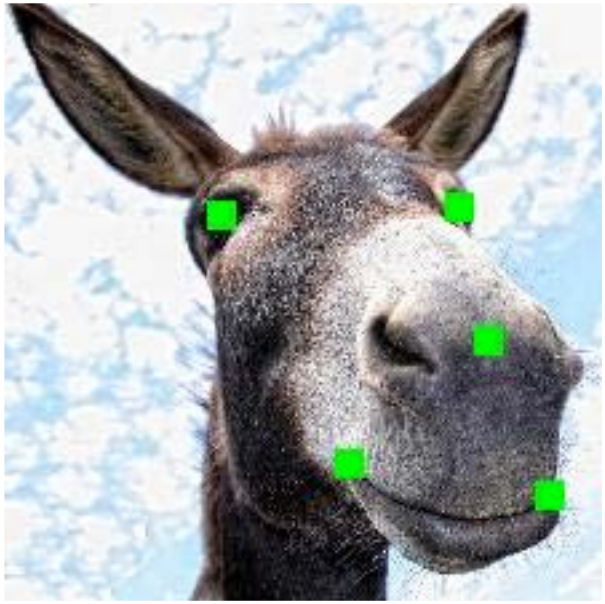
| | Random | Disc-only | Token | PC | SIB | Ours |
|---|---|---|---|---|---|---|
| Accuracy | 0.115 | 0.230 | 0.222 | 0.236 | 0.231 | **0.244** |

Table 1. Localization accuracy as measured by the "pointing game" [52] on Visual Genome. Our model outperforms all baselines, including variants of our method that lack one or more loss terms.

# Animal keypoint detection

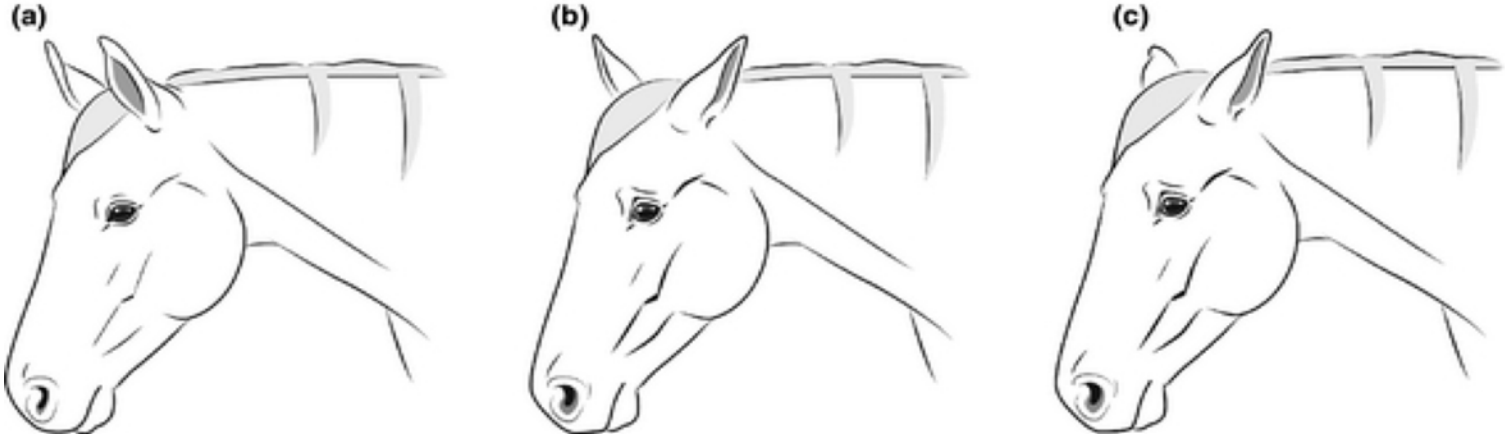
Input

Output



# Motivation

- Veterinary research has found **facial expressions of pain** for horses, sheep, and cows to be a **good indicator of pain**




K. B. Gleerup, B. Forkman, C. Lindegaard, and P. H. Andersen. An equine pain face. Veterinary anesthesia and analgesia, 42(1):103–114, 2015

# Motivation

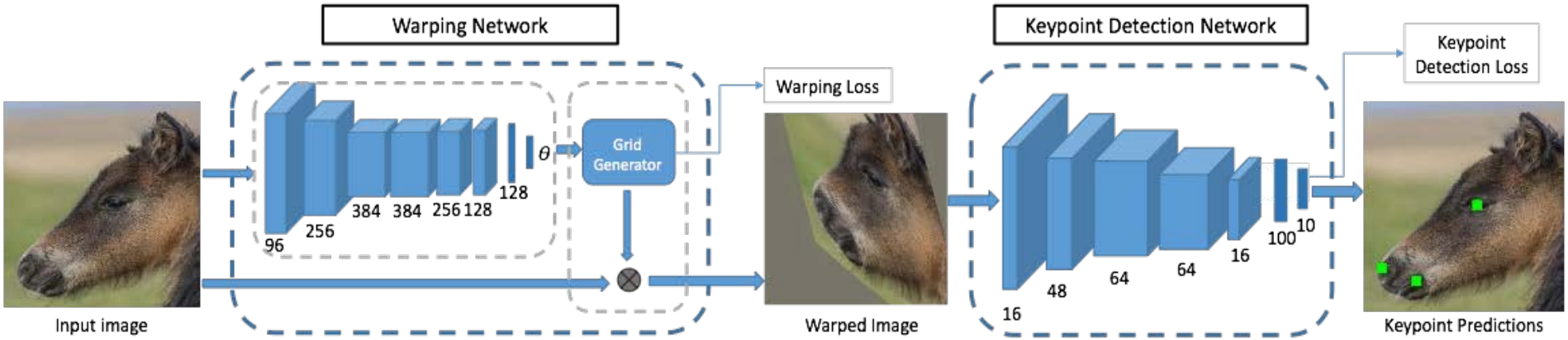
- Automatic pain detection can have a huge impact on animal welfare
  - $0.5 billion spent annually in the US to treat lameness in horses
- Human detection of animal pain is challenging
  - Humans need to be trained to detect pain
  - Livestock may hide expressions of pain near humans

# Key idea

- Leverage existing *human* training data for animal facial keypoint detection (with a small amount of animal training data)
- Warp animal faces to resemble human shape


Maheen Rashid, Xiuye Gu, Yong Jae Lee. Interspecies Knowledge Transfer. CVPR 2017.

# Approach



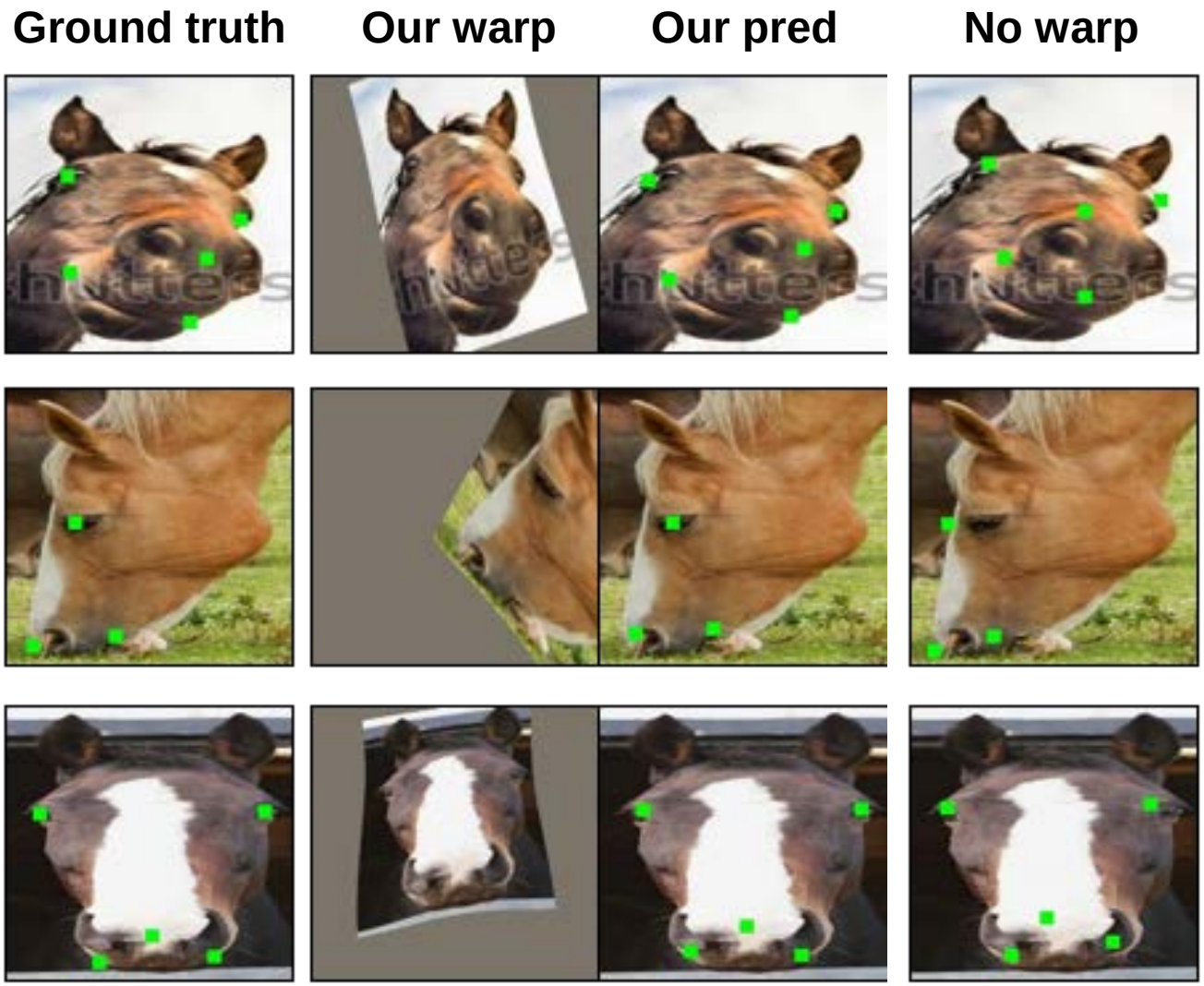

- Warping network warps animal to have human-like shape
- Keypoint network predicts facial keypoints on warped animal face

# Experiments

- Horse Facial Keypoint Dataset
  - 3531 training, 186 testing images
  - Left eye, right eye, nose center, left mouth corner, right mouth corner

- Sheep Facial Keypoint Dataset [Yang et al. 2016]
  - 432 training, 99 testing Images
  - Same 5 keypoints as above

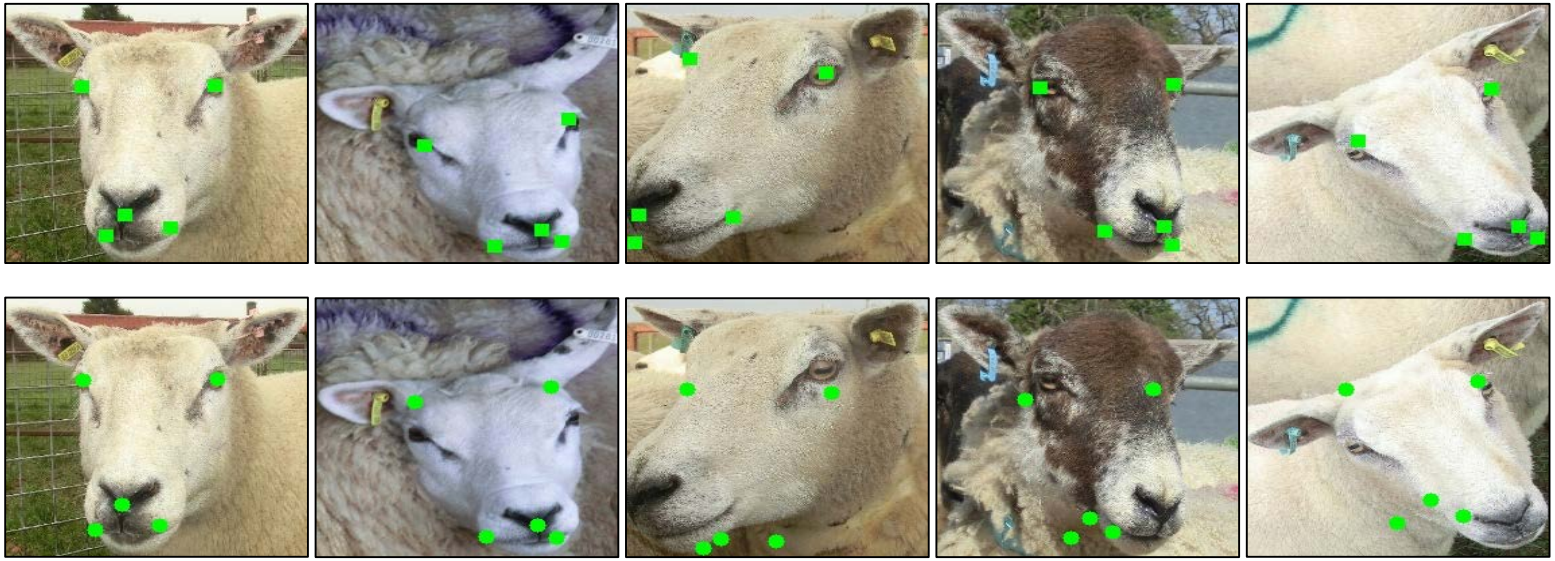
# Qualitative results



# Qualitative results

Ours

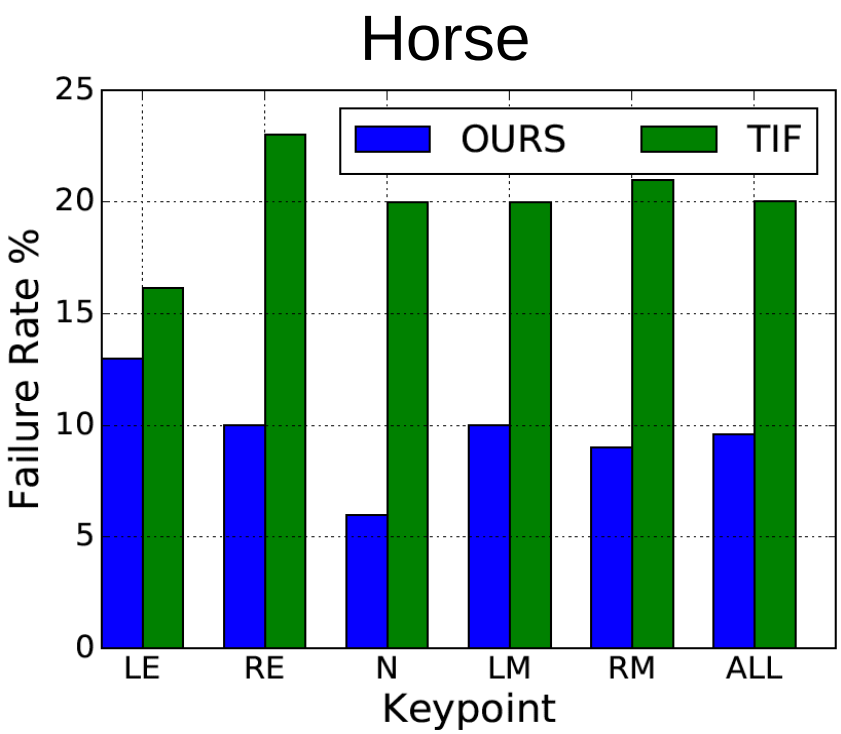
Yang et al. 2016

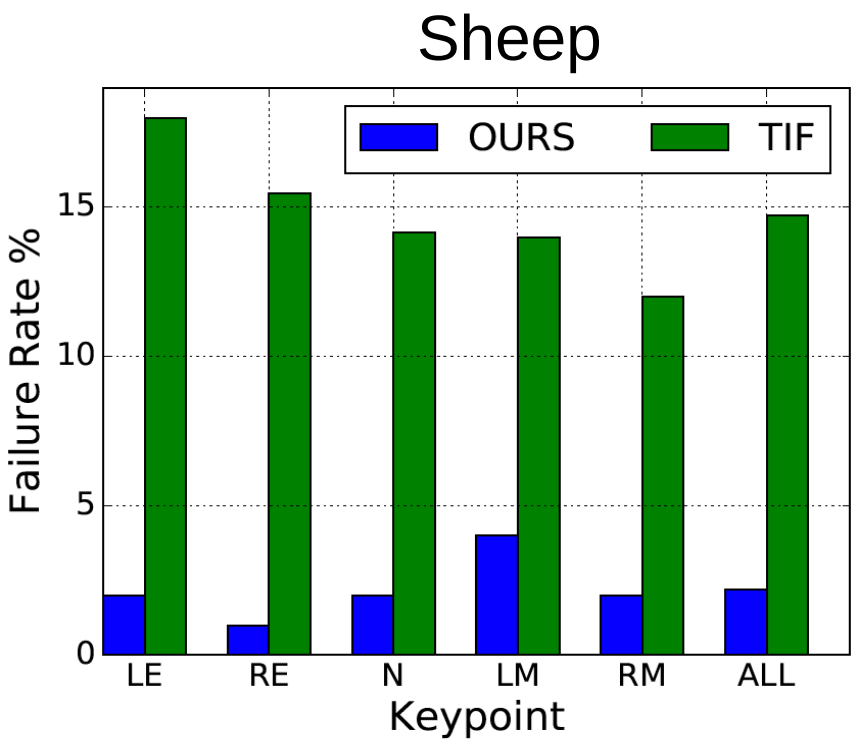


- By leveraging existing human training data, our CNN model performs much better

# Quantitative results

- Triplet Interpolated Features [Yang et al. 2016]



# Conclusions

- Weakly-supervised learning enables *scalable* visual recognition
  - Object detection, attribute recognition, visual grounding

- Still, challenges remain in dealing with
  - Noise in data
  - Limited training data

# Coming up

- Sign-up for paper
- Next class: CNN/PyTorch/Torch tutorial