Visual Phrases

CSE 590V

Supasorn Suwajanakorn

Visual Phrases

A person riding a horse

Objects + Interactions

A woman drinks from a water bottle

Visual Phrases

Dog Jumping

Object + Activity

Why do we care?

- So that we understand the scene better
- Help detect individual objects!

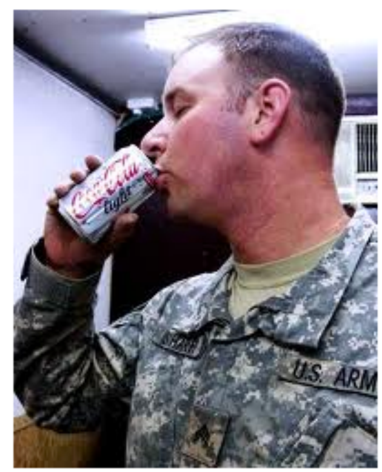
  (... if we have an accurate visual phrase detector)

Design a Visual Phrase Detector

- Say, we want to **detect people** as well as **describe activity** in these pictures





Design a Visual Phrase Detector

Let's look at what our detectors are good at

Find a person like this

Find a horse like this

So, we can combine these two detectors then try to model the relationship

Design a Visual Phrase Detector

Using that method, we can excel at finding person in pictures like these

Can we find a person in this picture with good precision?

Maybe

Design a Visual Phrase Detector

VS

Person riding a horse usually has:

Change in Appearance
A few postures
One leg not visible
...

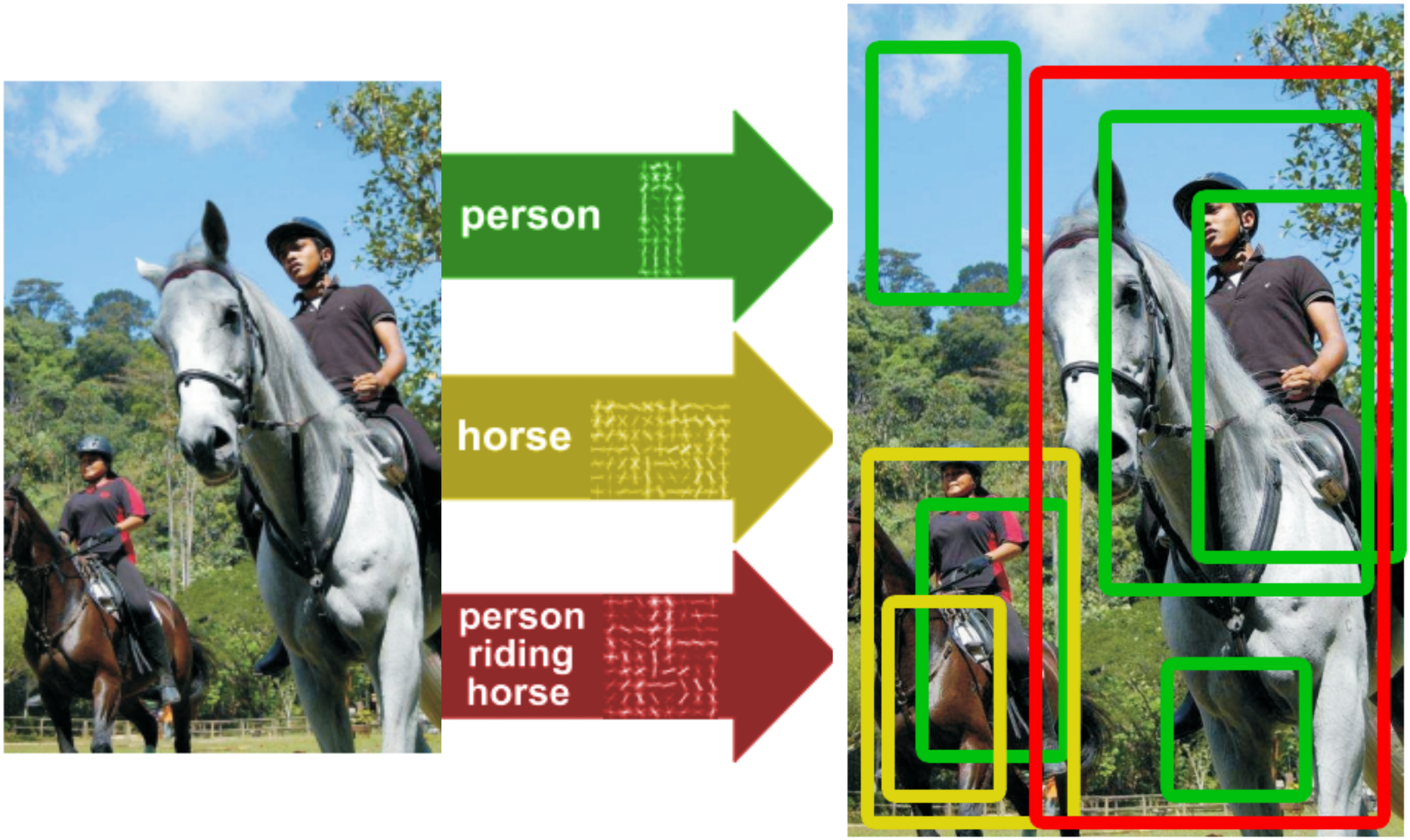
How do we take advantage of this?

Design a Visual Phrase Detector

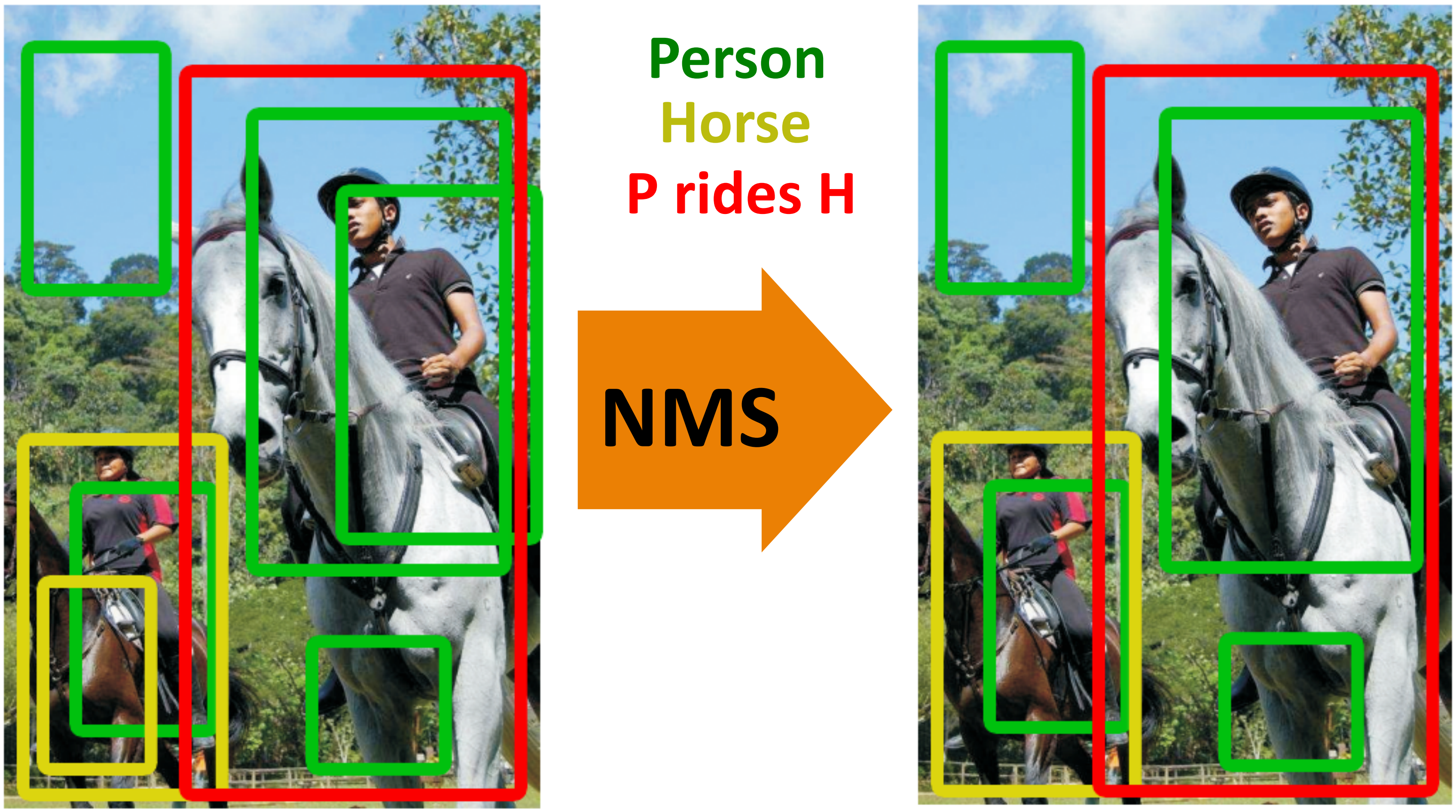
  

- A simple solution
  - Add one more class “person riding horse”, in addition to “person” and “horse”
  - Train a classifier to detect “person riding horse” using some training examples
  - Done?

Design a Visual Phrase Detector



Design a Visual Phrase Detector



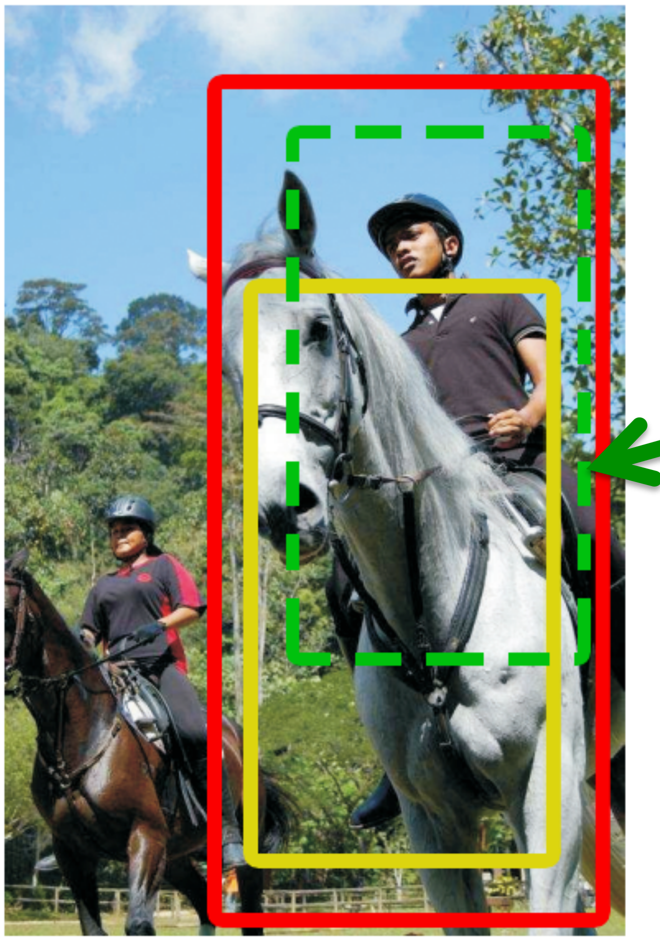
Non-maximum suppression

What’s wrong with NMS

We could have done better if visual phrase plays a role

Maybe remove this because some person is riding a horse and there shouldn’t be another person under the horse

What's wrong with NMS



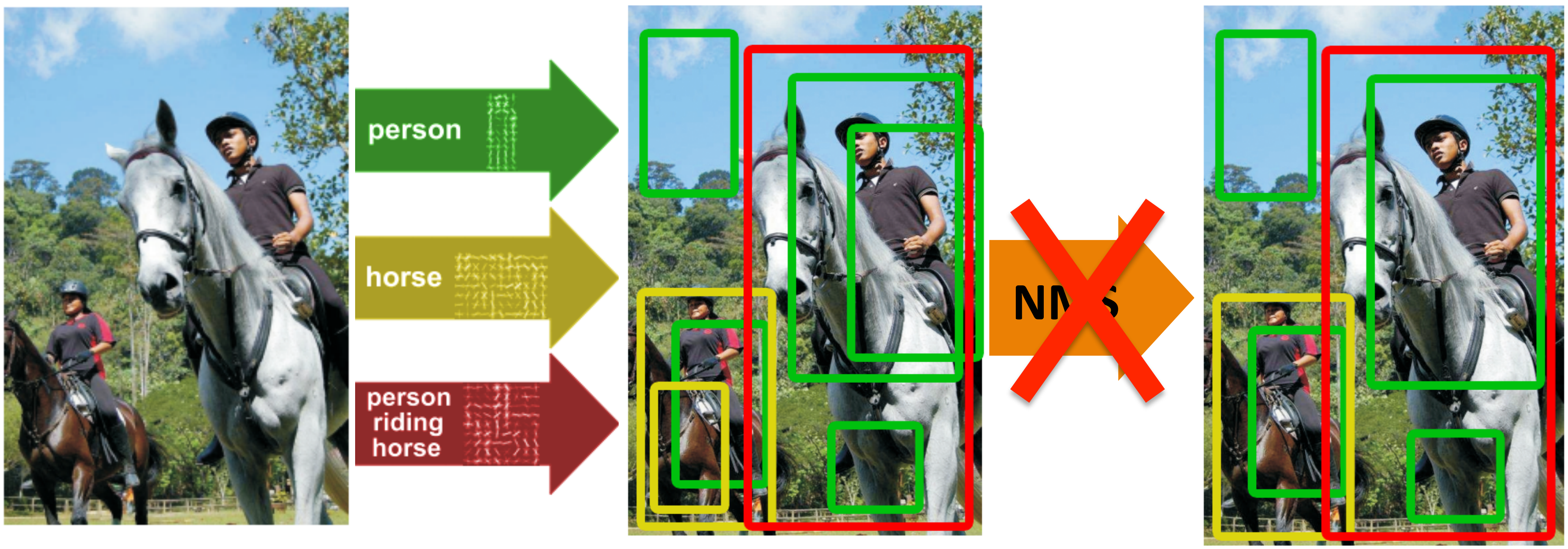
We could have done better if visual phrase plays a role

If person detector gives a low confidence, but we are pretty sure there are horse and person riding it, confidence for this person should go up

Need a better method that take into account the relationship between objects

NMS to Decoder

Our current pipeline

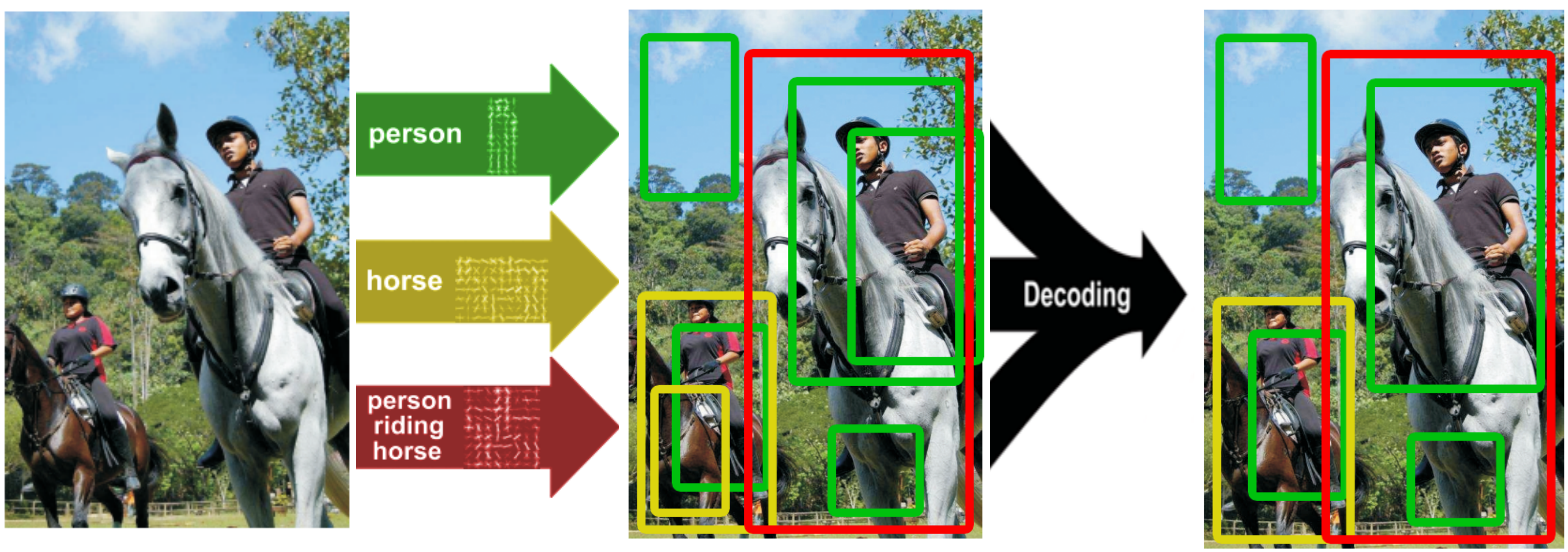


Novel decoding procedure

"Recognition Using Visual Phrases"
Mohammad Sadeghi, Ali Farhadi

NMS to Decoder

Our current pipeline

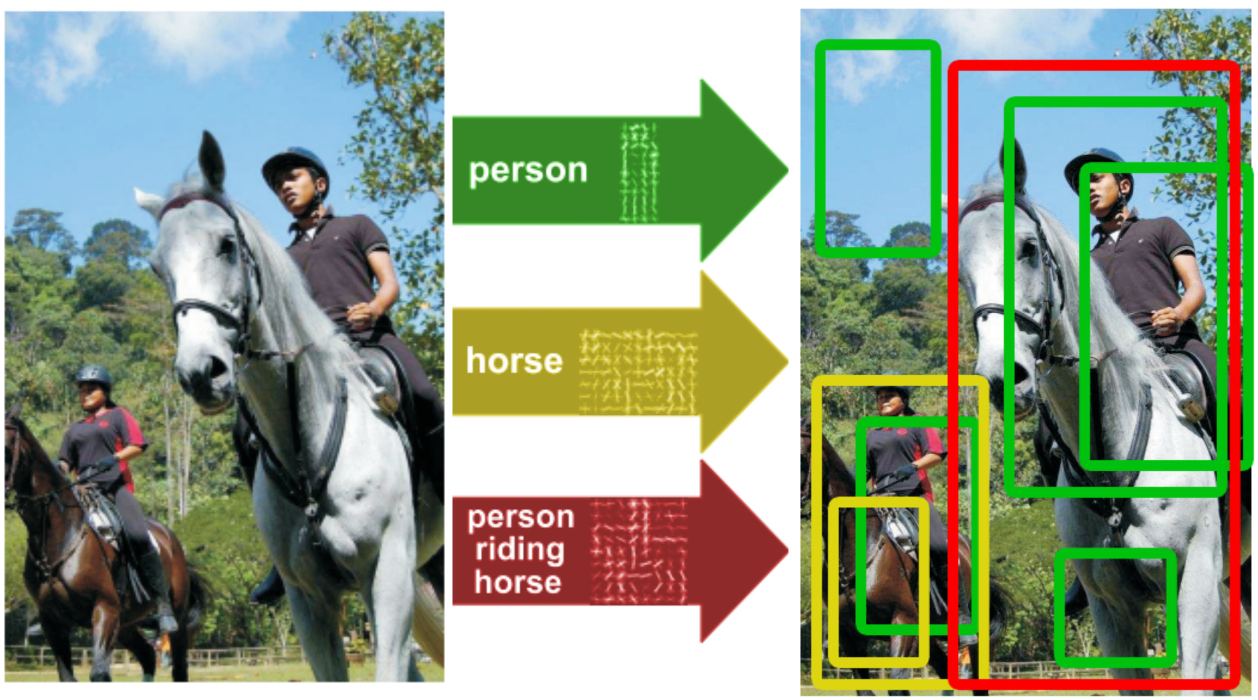


Novel decoding procedure

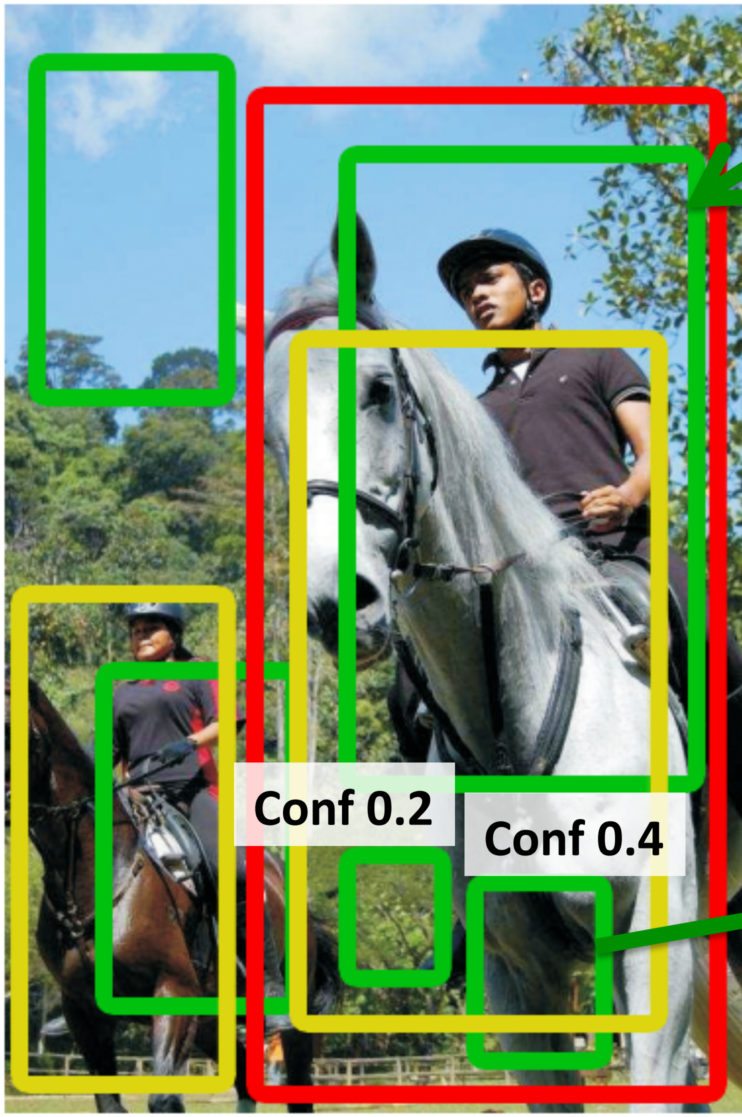
"Recognition Using Visual Phrases"
Mohammad Sadeghi, Ali Farhadi

Redefine Feature

- Decoding needs more info from features
- Goal: a new representation of feature that is aware of the surrounding features



Representation of Feature x_1

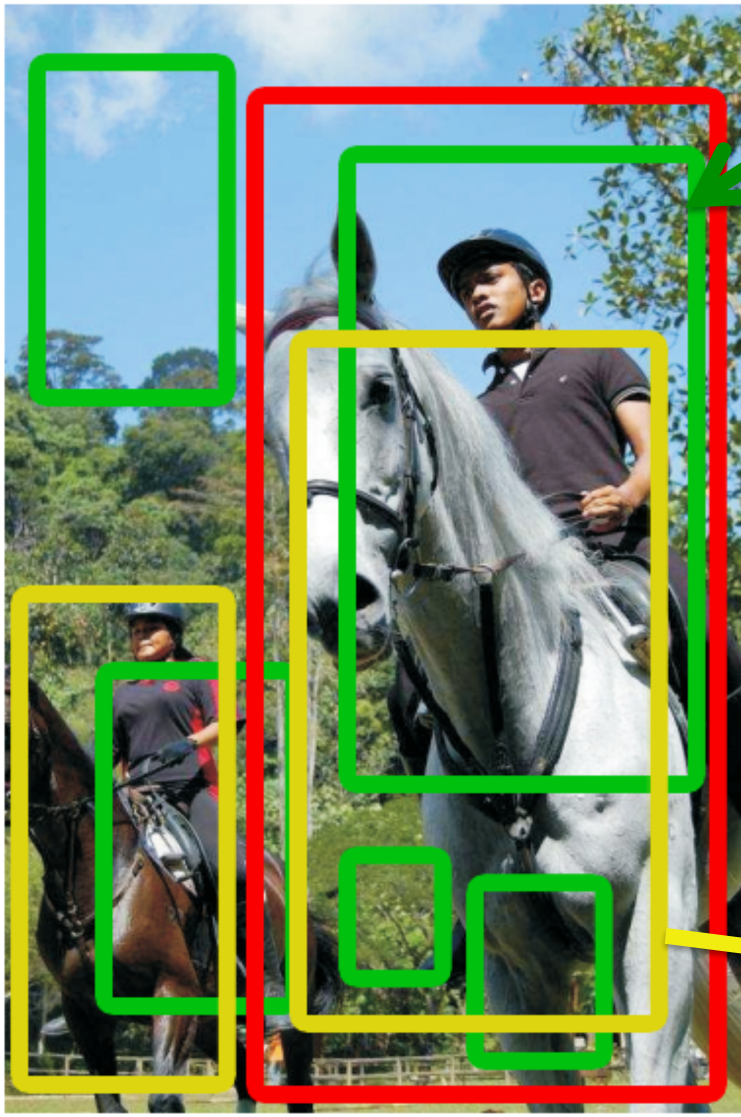


Consider this “person”-bounding box
Suppose this is feature x_1

Now let’s consider x_1 in relation with other surrounding “person”

	Confidence	Overlap	Size ratio
Above	0	0	0
Below	0.4	0	0.2
Overlap	0	0	0

Representation of Feature x_1

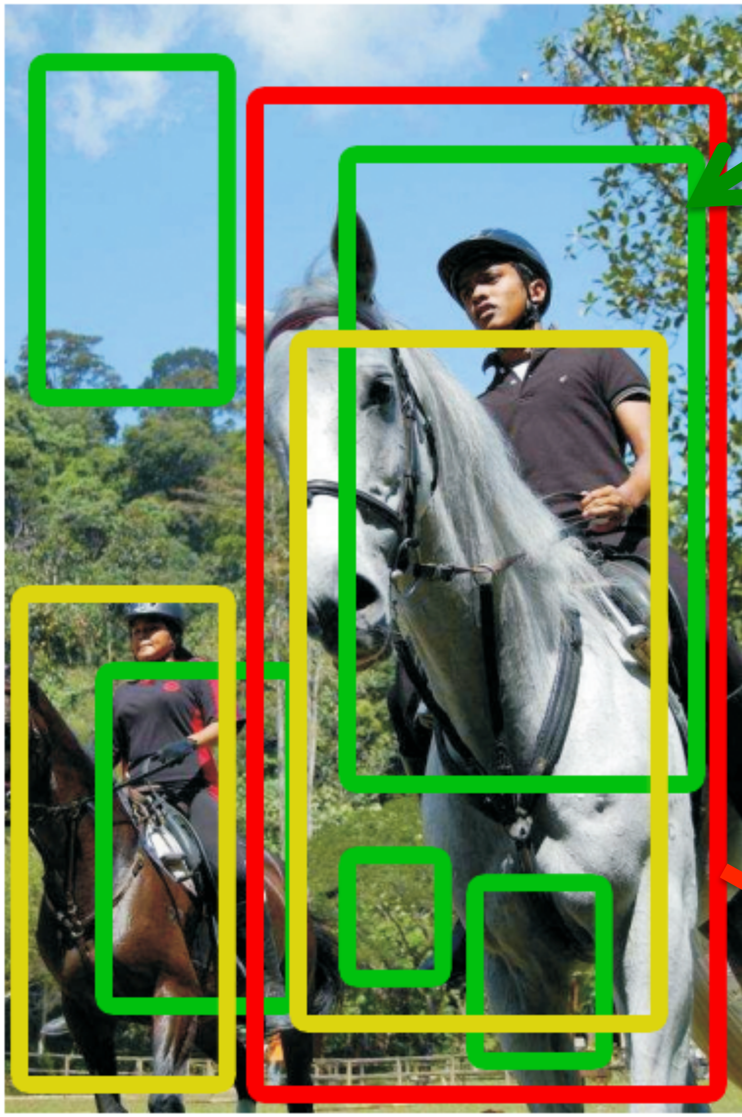


Consider this "person"-bounding box
Suppose this is feature x_1

Now let's consider x_1 in relation with other surrounding "horse"

	Confidence	Overlap	Size ratio
Above	0	0	0
Below	0	0	0
Overlap	0.8	0.7	1.2

Representation of Feature x_1

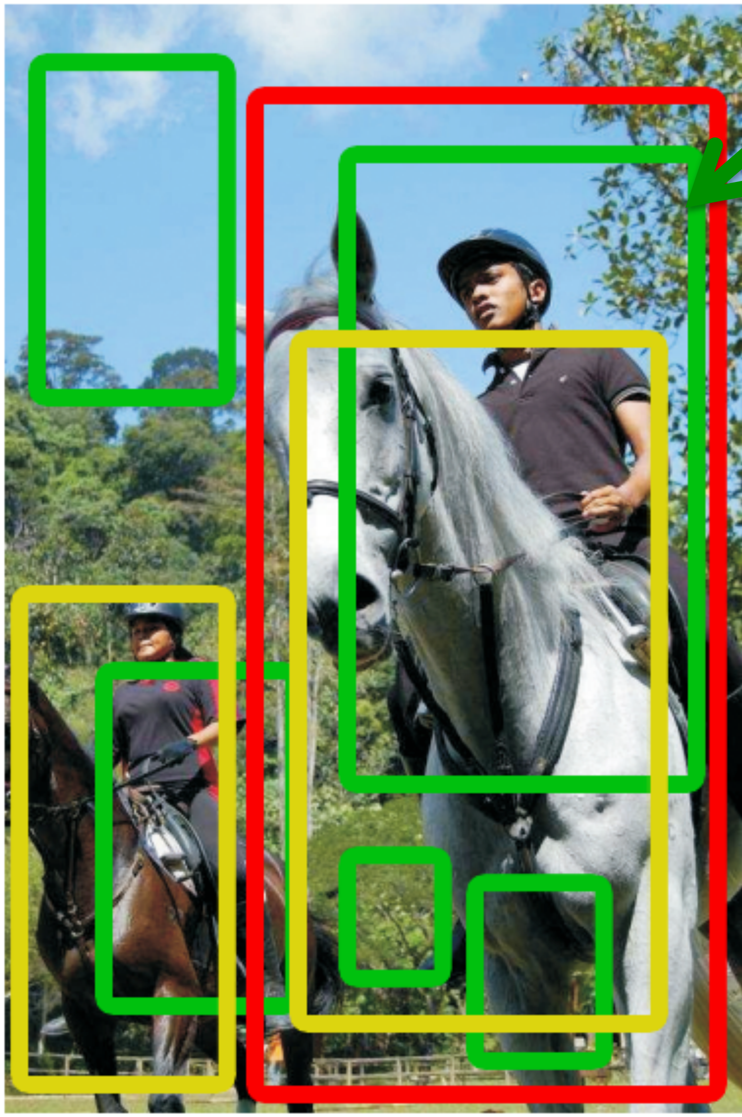


Consider this "person"-bounding box
Suppose this is feature x_1

Now let's consider x_1 in relation with other surrounding "P rides H"

	Confidence	Overlap	Size ratio
Above	0	0	0
Below	0	0	0
Overlap	0.9	0.6	1.8

Representation of Feature x_1



feature vector x_1 (class "person")

0	0	0
0.4	0	0.2
0	0	0

Interaction of x_1 with "person"

0	0	0
0	0	0
0.8	0.7	1.2

Interaction of x_1 with "horse"

0	0	0
0	0	0
0.9	0.6	1.8

Interaction of x_1 with "P rides H"

Representation of Feature x_1

"person"

0	0	0
0.4	0	0.2
0	0	0

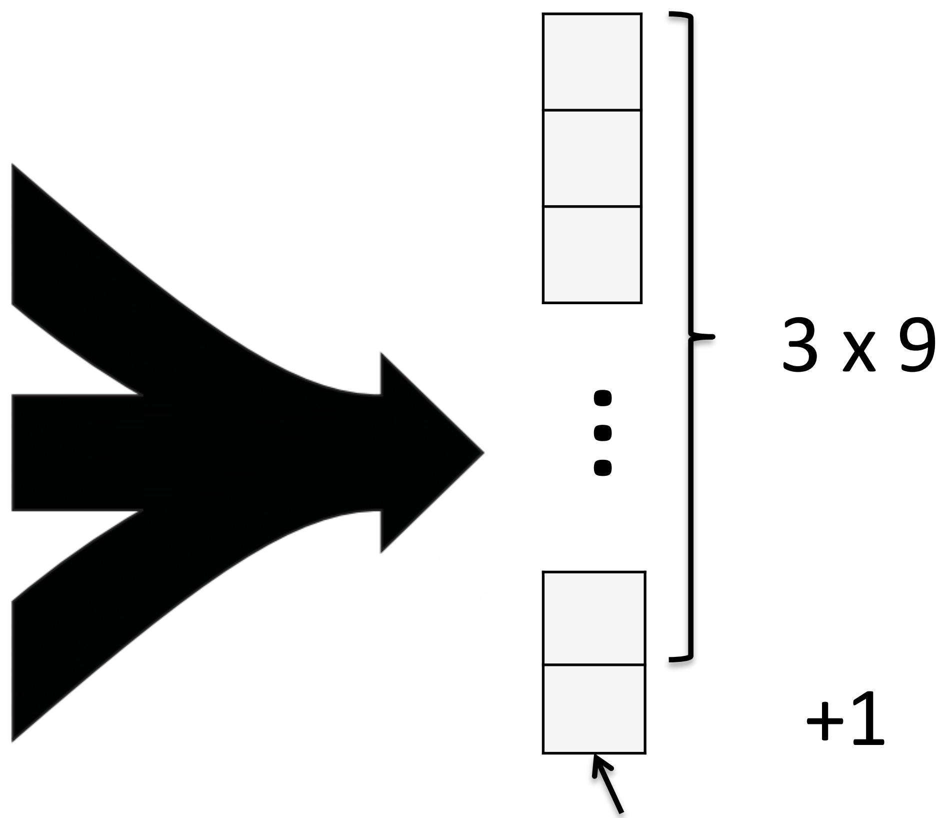
"horse"

0	0	0
0	0	0
0.8	0.7	1.2

"P rides H"

0	0	0
0	0	0
0.9	0.6	1.8

feature vector x_1



Confidence of this bounding box

More generally (K x 9) + 1, K=# of classes

Inference (Decoder)

Goal: Decides whether x_i should be in final response

$$Y^* = \{y_1^*, y_2^*, \ldots, y_M^*\}$$

$$y_i^* = \arg\max_{y_i} \boxed{w_{c_i}^T} x_i y_i$$

Max margin structure learning

$X = \{x_1, x_2, \ldots, x_M\}$: M bounding boxes / features

$Y = \{y_1, y_2, \ldots, y_M\}$: $y_i \in \{0, 1\}$ if x_i should be in final response

$c_i \in \{1, 2, \ldots, K\}$: class of i^{th} bounding box.

w_{c_i}: the set of weights corresponding to the class of c_i

Comparing Methods

This paper

Sadeghi & Farhadi

$$S(X,Y) = \sum_i w_{c_i}^T x_i$$

Related Method

Discriminative models for multi-class object layout (C. F. C. Desai, D. Ramanan)

$$S(X,Y) = \sum_{i,j} \boxed{w_{y_i,y_j}^T} d_{ij} + \sum_i w_{y_i}^T \boxed{x_i}$$

Pairwise term

No info about surrounding

Problem?

Inference is hard. Need to guess labels (greedily search)

Fix (Sadeghi & Farhadi)

No need to guess labels. Labels directly from detectors

Infer y_i only (0 or 1)

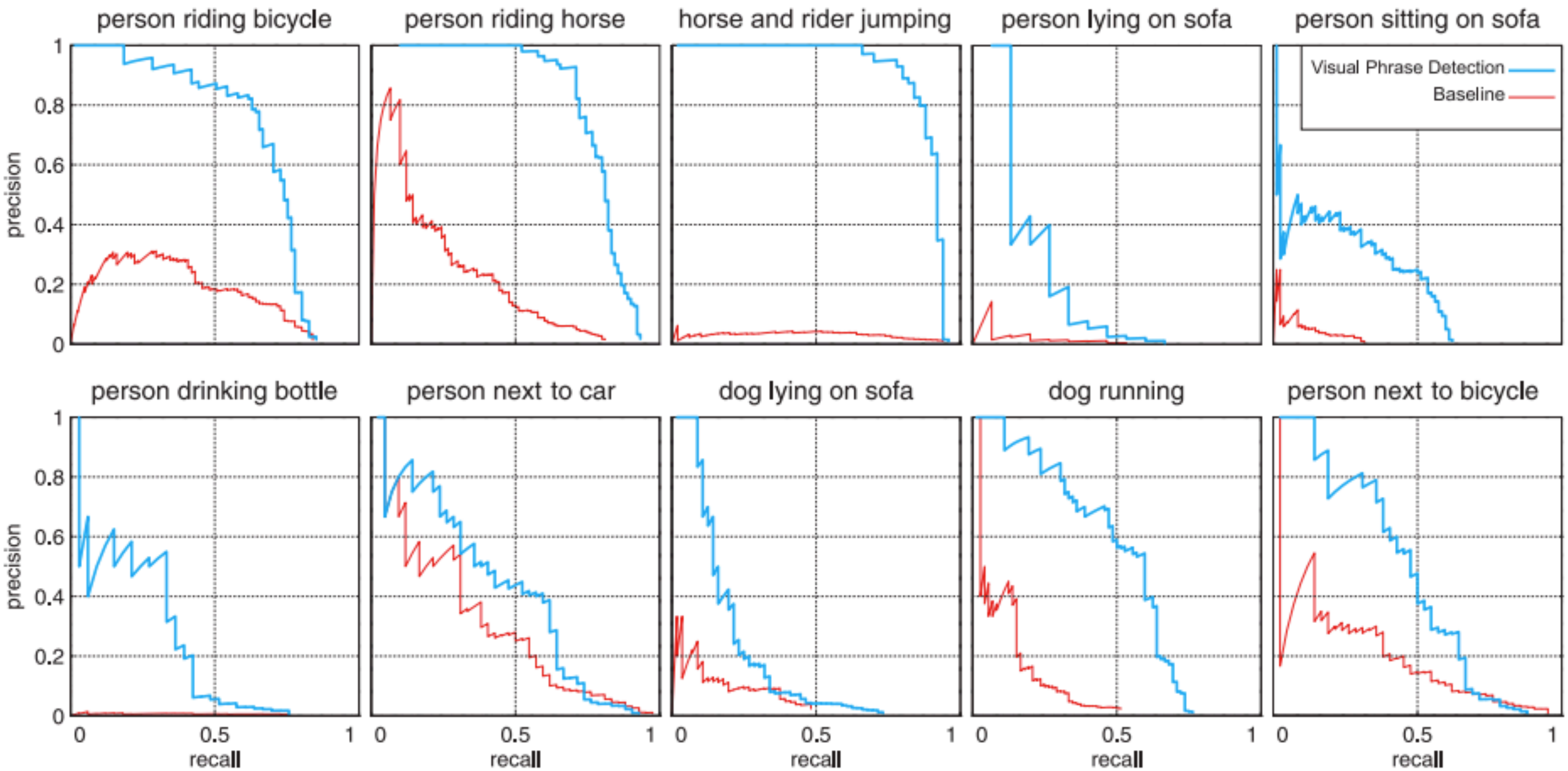
Get exact inference

Results

Phrases (Trained with 50 positive images)	Phrase (AP)	Baseline (AP)	Gain (AP)
Person next to bicycle	0.466	0.252	**0.214**
Person lying on sofa	0.249	0.022	**0.227**
Horse and rider jumping	0.870	0.035	**0.835**
Person drinking from bottle	0.279	0.010	**0.269**
Person sitting on sofa	0.262	0.033	**0.229**
Person riding horse	0.787	0.262	**0.525**
Person riding bicycle	0.669	0.188	**0.481**
Person next to car	0.443	0.340	**0.103**
Dog lying on sofa	0.235	0.069	**0.166**
Bicycle next to car	0.448	0.461	**-0.013**
Dog Jumping	0.072	0.134	**-0.062**
Person sitting on chair	0.201	0.141	**0.060**
Person running	0.718	0.484	**0.234**
Person lying on beach	0.179	0.140	**0.039**
Person jumping	0.317	0.036	**0.281**
Person next to horse	0.351	0.287	**0.064**
Dog running	0.504	0.160	**0.344**

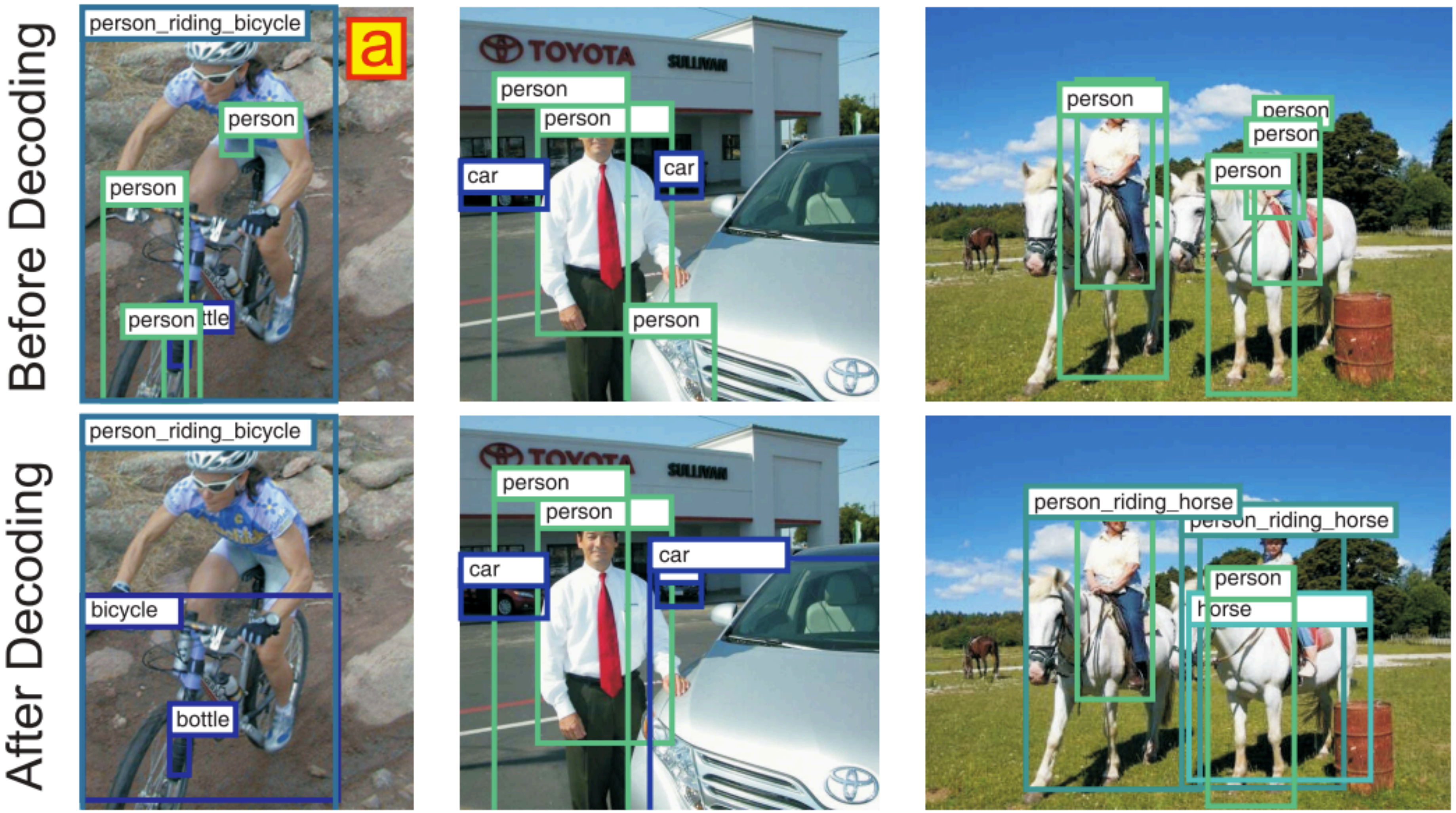
Baseline:

Optimistic upper-bound on how well one can detect visual phrases by individually detecting participating objects then Modeling the relation.



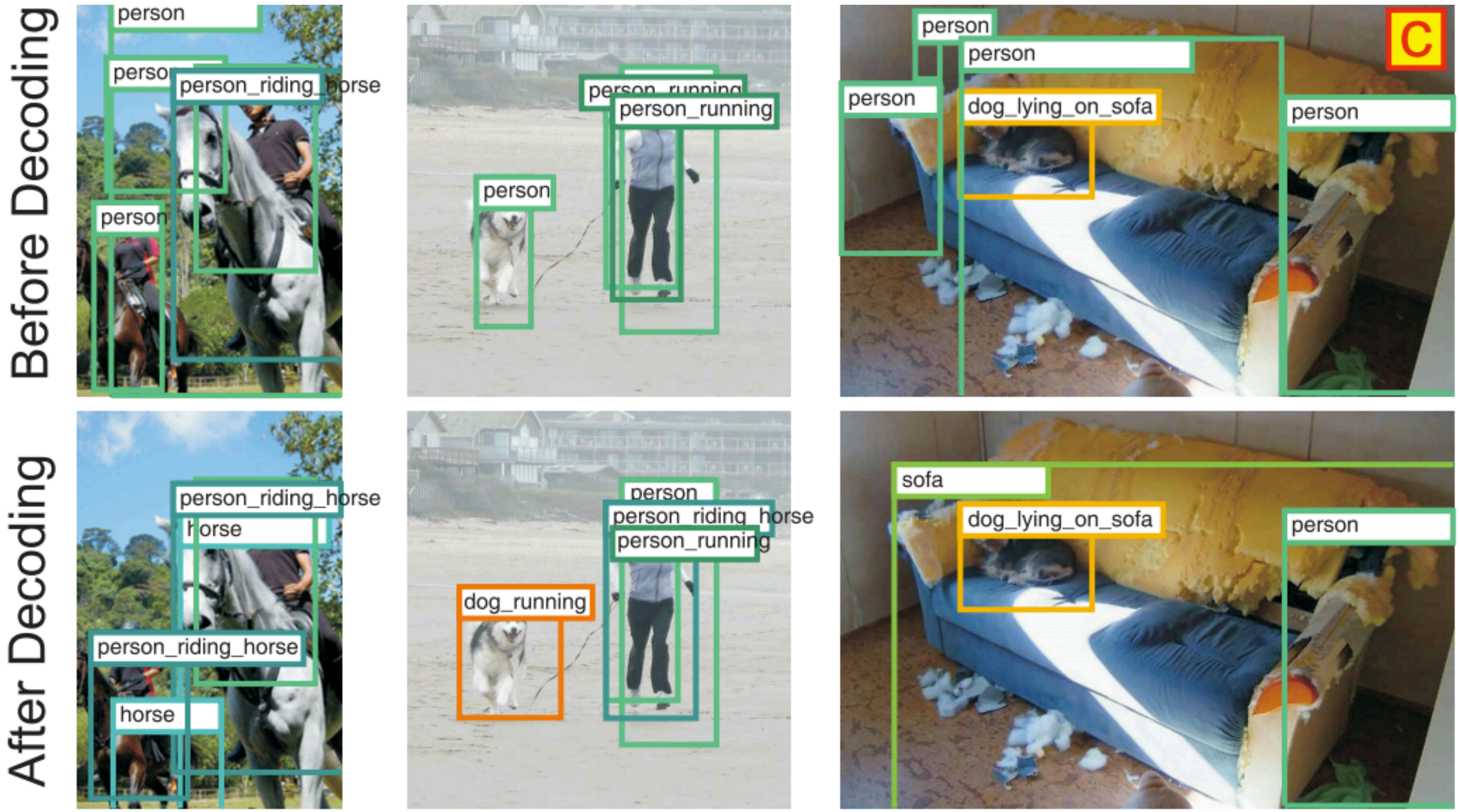
Significant gain in detecting visual phrases compared to detecting objects and describing their relations.

Results



Before Decoding
After Decoding
a
person_riding_bicycle
person
bottle
bicycle
car
person_riding_horse
horse
TOYOTA
SULLIVAN

Results



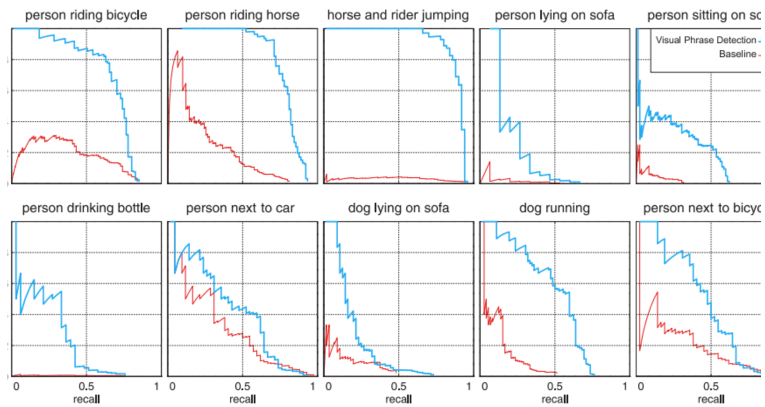
Before Decoding
After Decoding
person
person_riding_horse
person_running
dog_lying_on_sofa
C
horse
dog_running
sofa

Results

	bicycle	bottle	car	chair	dog	horse	person	sofa
detectors of [8]	0.434	0.429	0.329	0.213	0.316	0.438	0.295	0.204
[2] without phrases	0.431	0.425	0.191	0.225	0.297	0.475	0.204	0.167
[2] with phrases	0.449	**0.435**	0.228	0.217	0.316	0.462	0.286	0.204
Our decoding without phrases	0.437	0.434	0.330	0.216	0.329	0.440	0.297	0.218
Our decoding with phrases	**0.457**	**0.435**	**0.344**	**0.227**	**0.335**	**0.485**	**0.302**	**0.260**

This method outperforms state-of-the-art object detector+NMS and state-of-the-art multiclass recognition method of C. F. C. Desai, D. Ramana.

Discussion



- Negative examples do not contain participating objects. If we detect person riding horse with a picture of person next to horse, false positive might rise, precision might fall
- Visual phrases in practice, limitations