CVPR July 21-26 2017 HONOLULU

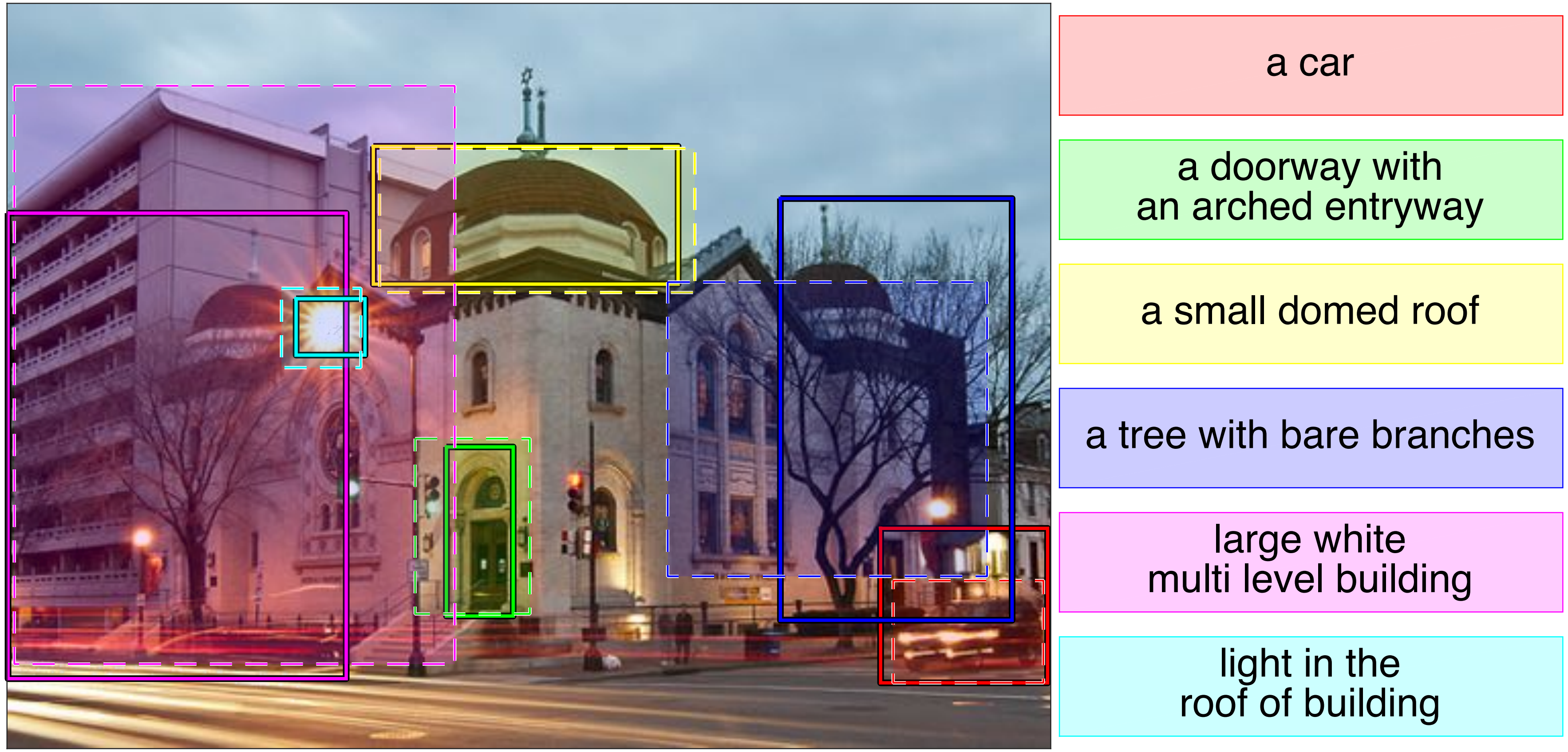
University of Michigan

Discriminative Bimodal Networks for Visual Localization and Detection with Natural Language Queries

Yuting Zhang, Luyao Yuan, Yijie Guo,
Zhiyuan He, I-An Huang, Honglak Lee

University of Michigan, Ann Arbor

Detection with natural language queries



a car

a doorway with an arched entryway

a small domed roof

a tree with bare branches

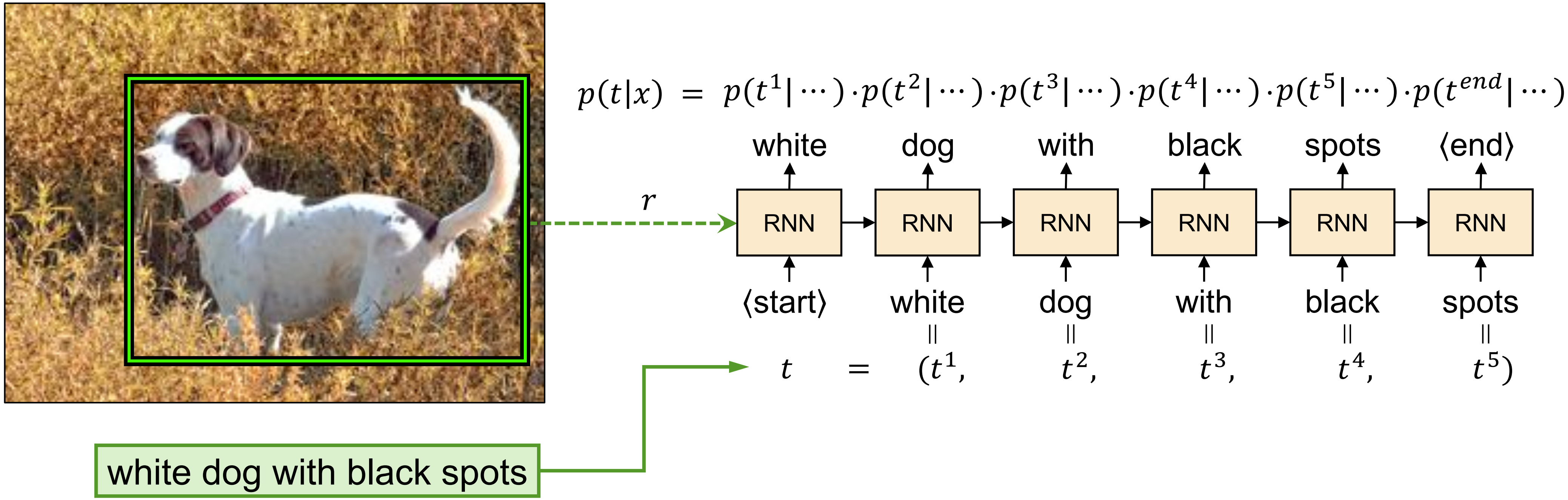
large white multi level building

light in the roof of building

Detection results from our work.

Detection:	Boxes with SOLID edges.
Ground truth:	Semi-transparent boxes with DASHED edges.

Typical previous works (based on captioning)



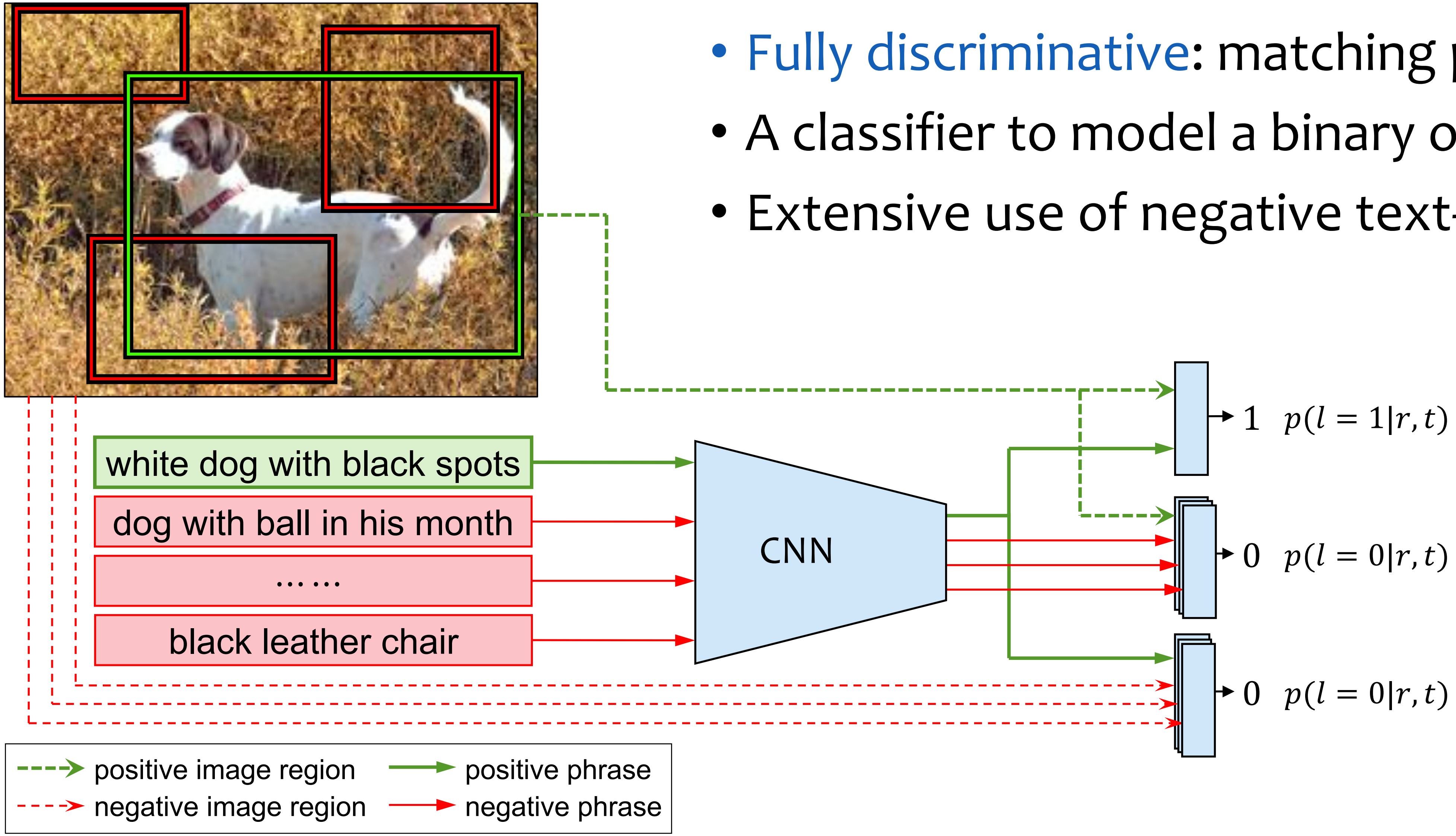
$$p(t|x) = p(t^1|\cdots)\cdot p(t^2|\cdots)\cdot p(t^3|\cdots)\cdot p(t^4|\cdots)\cdot p(t^5|\cdots)\cdot p(t^{end}|\cdots)$$

$$t = (t^1, t^2, t^3, t^4, t^5)$$

- Based on generative models for image captioning.
- The posterior probability in the huge language space is hard to model.
- Only positive training samples (matched box and text)
- Or a limited amount of negative training samples (mismatched box and text)

Discriminative Bimodal Networks (DBNet)

- Fully discriminative: matching probability
- A classifier to model a binary output
- Extensive use of negative text-box pairs



Discriminative Bimodal Networks (DBNet)

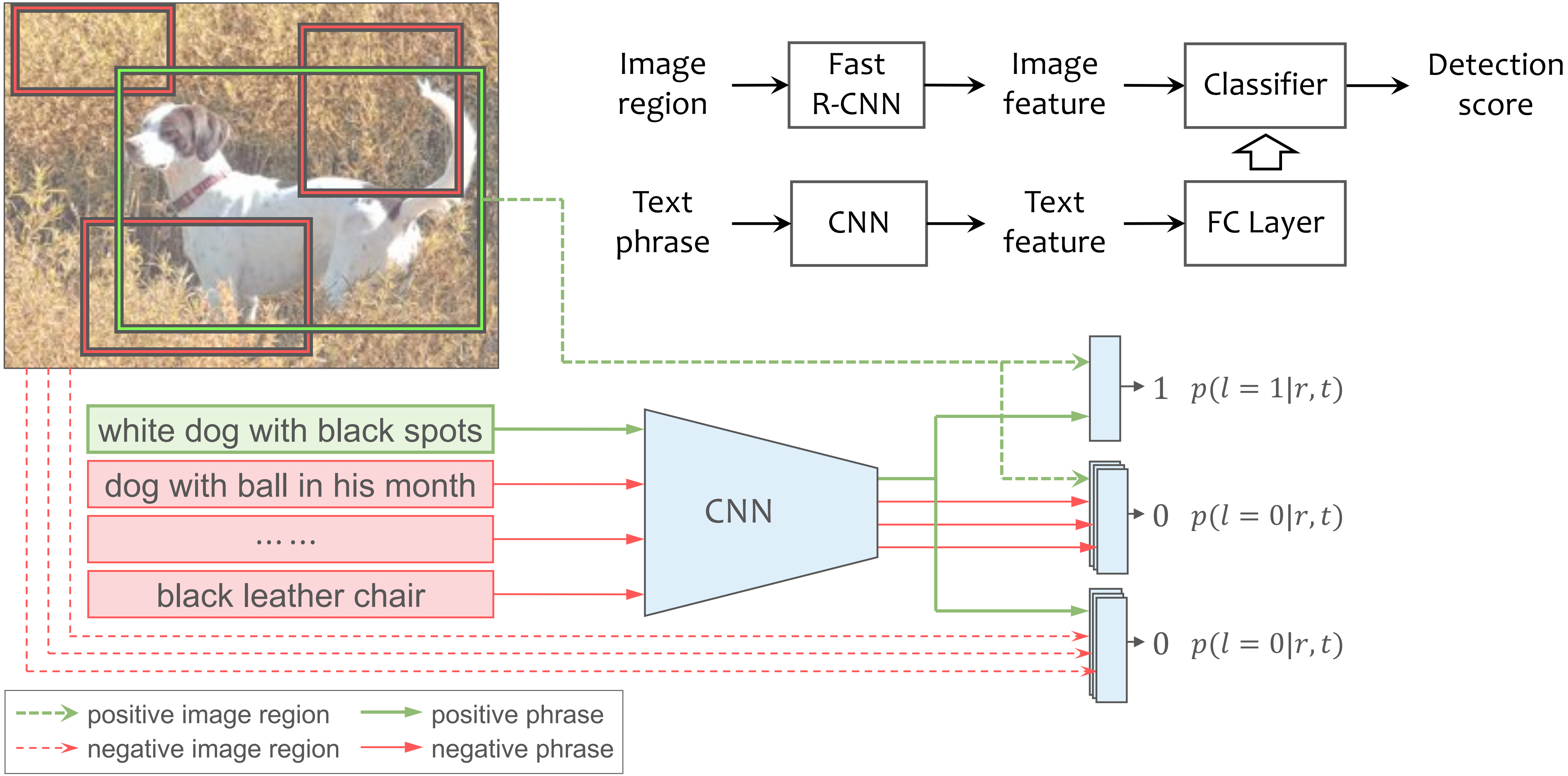


Image region → Fast R-CNN → Image feature → Classifier → Detection score

Text phrase → CNN → Text feature → FC Layer → Classifier

white dog with black spots

dog with ball in his month

… …

black leather chair

CNN

1 $p(l = 1|r, t)$

0 $p(l = 0|r, t)$

0 $p(l = 0|r, t)$

positive image region

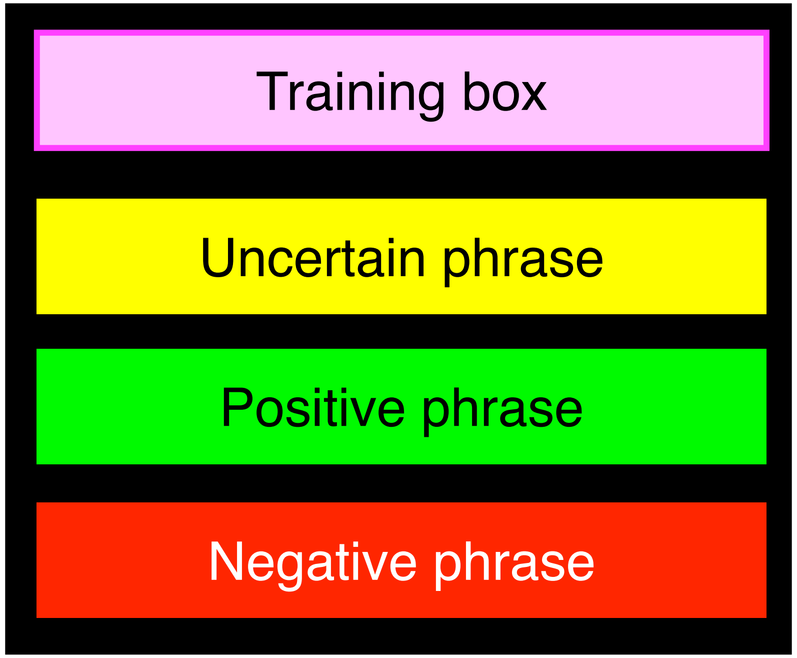
negative image region

positive phrase

negative phrase

DBNet: Training labels for text-box pairs

- Spatial overlapping based labeling

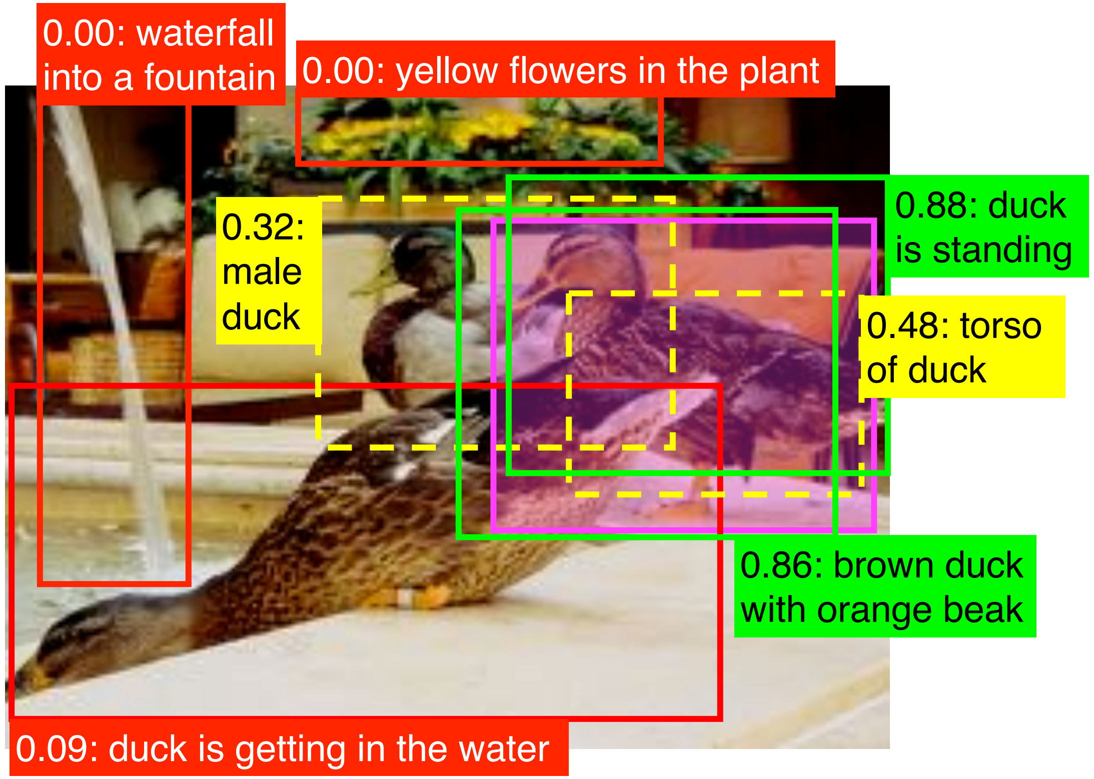


Training box

Uncertain phrase

Positive phrase

Negative phrase


0.00: waterfall into a fountain
0.00: yellow flowers in the plant
0.88: duck is standing
0.32: male duck
0.48: torso of duck
0.86: brown duck with orange beak
0.09: duck is getting in the water

- Text similarity based augmentation of uncertain phrases

Uncertain phrases:

- torso of duck
- male duck
- a male duck
- …

Experiments: Localization in Single Images

- Visual Genome dataset
- VGGNet is the default backbone image network

Method	Accuracy/% for IoU@			Median IoU	Mean IoU
	0.3	0.5	0.7		
DenseCap	25.7	10.1	2.4	0.092	0.178
SCRC	27.8	11.0	2.5	0.115	0.189
DBNet	**38.3**	**23.7**	**9.9**	**0.152**	**0.258**
DBNet (ResNet)	**42.3**	**26.4**	**11.2**	**0.205**	**0.284**

Experiments: Detection in Multiple Images

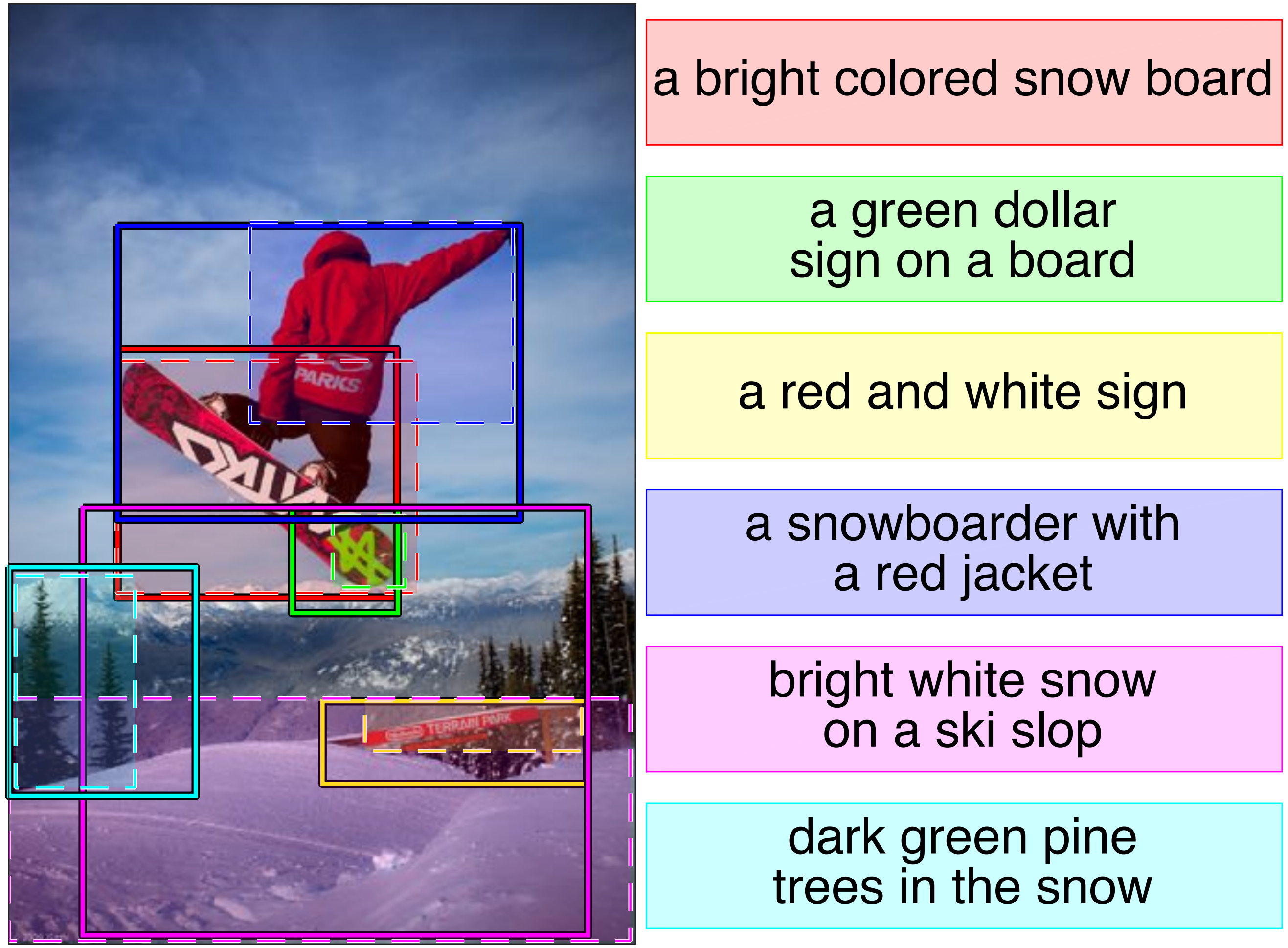
- We propose a new evaluation protocol for detection with text queries
    - 3 difficulty levels: increasing numbers of negative images per phrase
- Mean AP (mAP): each phrase has its own decision threshold
- Global AP (gAP): all phrases share the same decision threshold (requires scores to be calibrated over phrases)

Difficulty level:	**0**		**1**		**2**	
AP / %	mAP	gAP	mAP	gAP	mAP	gAP
DenseCap	15.7	0.5	10.0	0.3	1.7	0.0
SCRC	16.5	0.5	16.3	0.4	12.8	0.2
DBNet	**30.0**	**10.8**	**28.8**	**9.9**	**17.7**	**3.9**
DBNet (ResNet)	**32.6**	**11.5**	**31.2**	**10.7**	**19.8**	**4.3**

Thank you!

Data, Code & Models:

http:// DBNet.link


a bright colored snow board
a green dollar sign on a board
a red and white sign
a snowboarder with a red jacket
bright white snow on a ski slop
dark green pine trees in the snow

Detection results from our work.

Detection:	Boxes with SOLID edges.
Ground truth:	Semi-transparent boxes with DASHED edges.