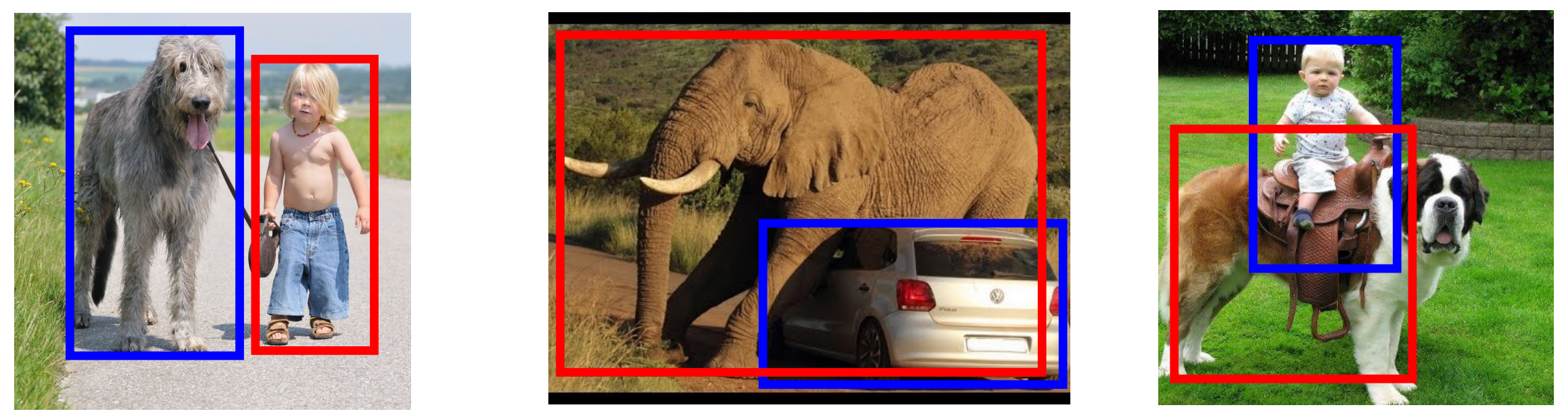
Weakly-supervised learning of visual relations

Julia Peyre[1,2] **Ivan Laptev**[1,2] **Cordelia Schmid**[2] **Josef Sivic**[1,2,3]

[1]DI ENS, École normale supérieure, PSL Research University [2]INRIA [3]CIIRC, CTU in Prague

Goal

- Detect visual relations using only **image-level annotations**
- Consider visual relations of the form (subject, predicate, object)



dog taller than person — car under elephant — dog ride bike

Challenges

- Diversity of visual relations
- Prohibitive cost of exhaustive manual annotation

Contributions

- Learning visual relations from image-level annotations
- UnRel : a new evaluation dataset with clean labels for evaluating the performance and generalization of relation detection

Overview

Training

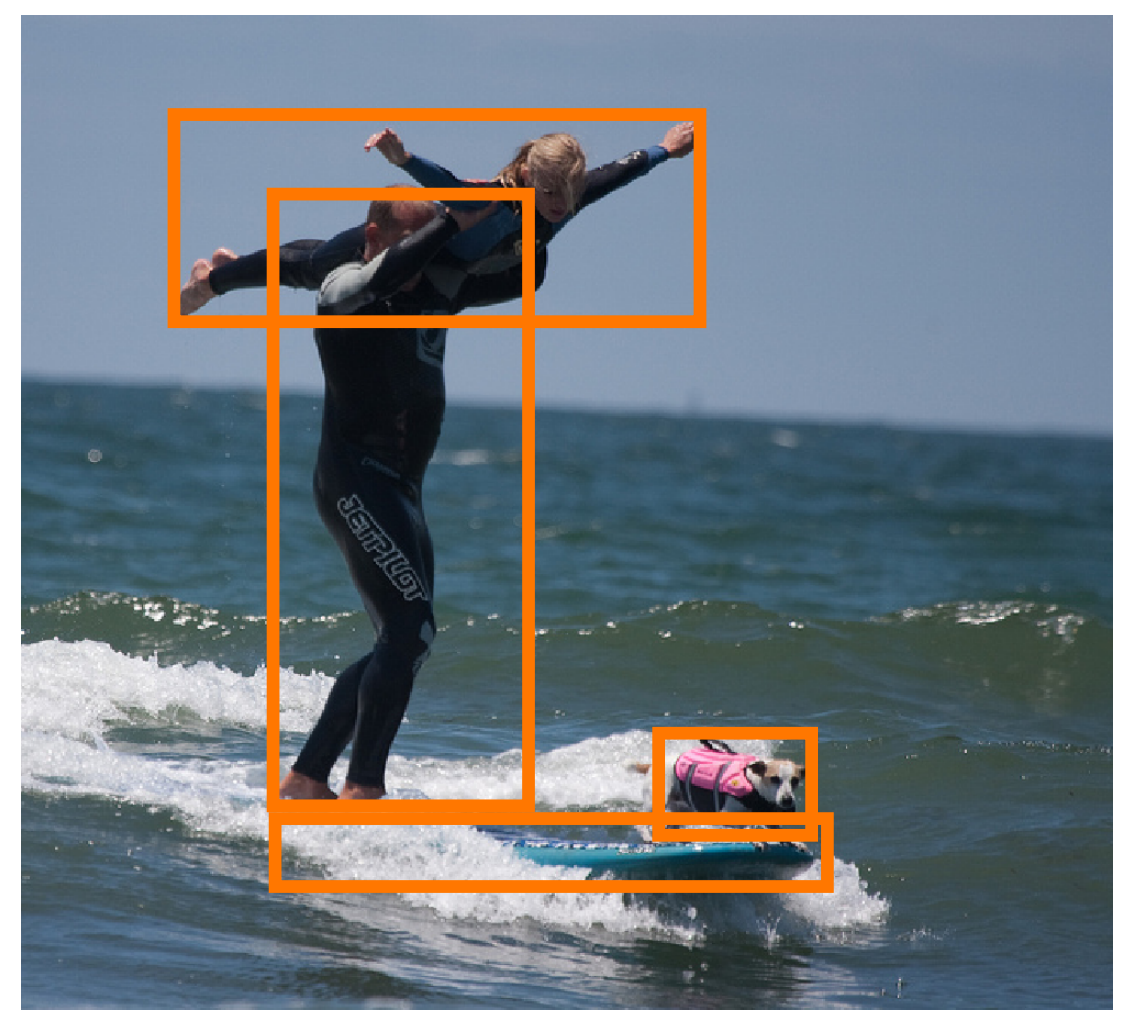
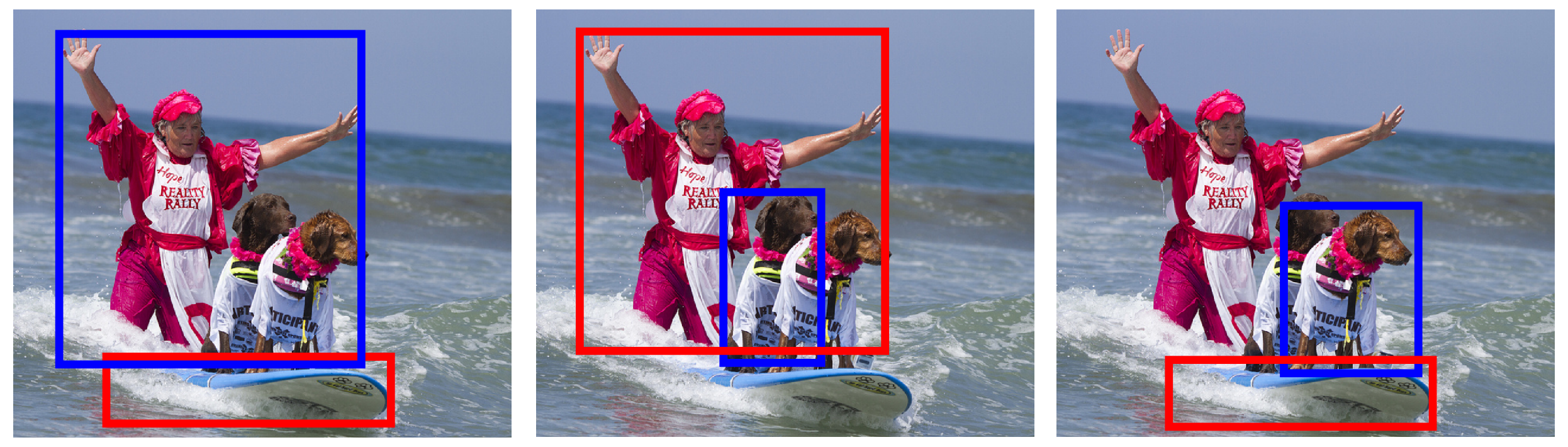


Image-level triplets

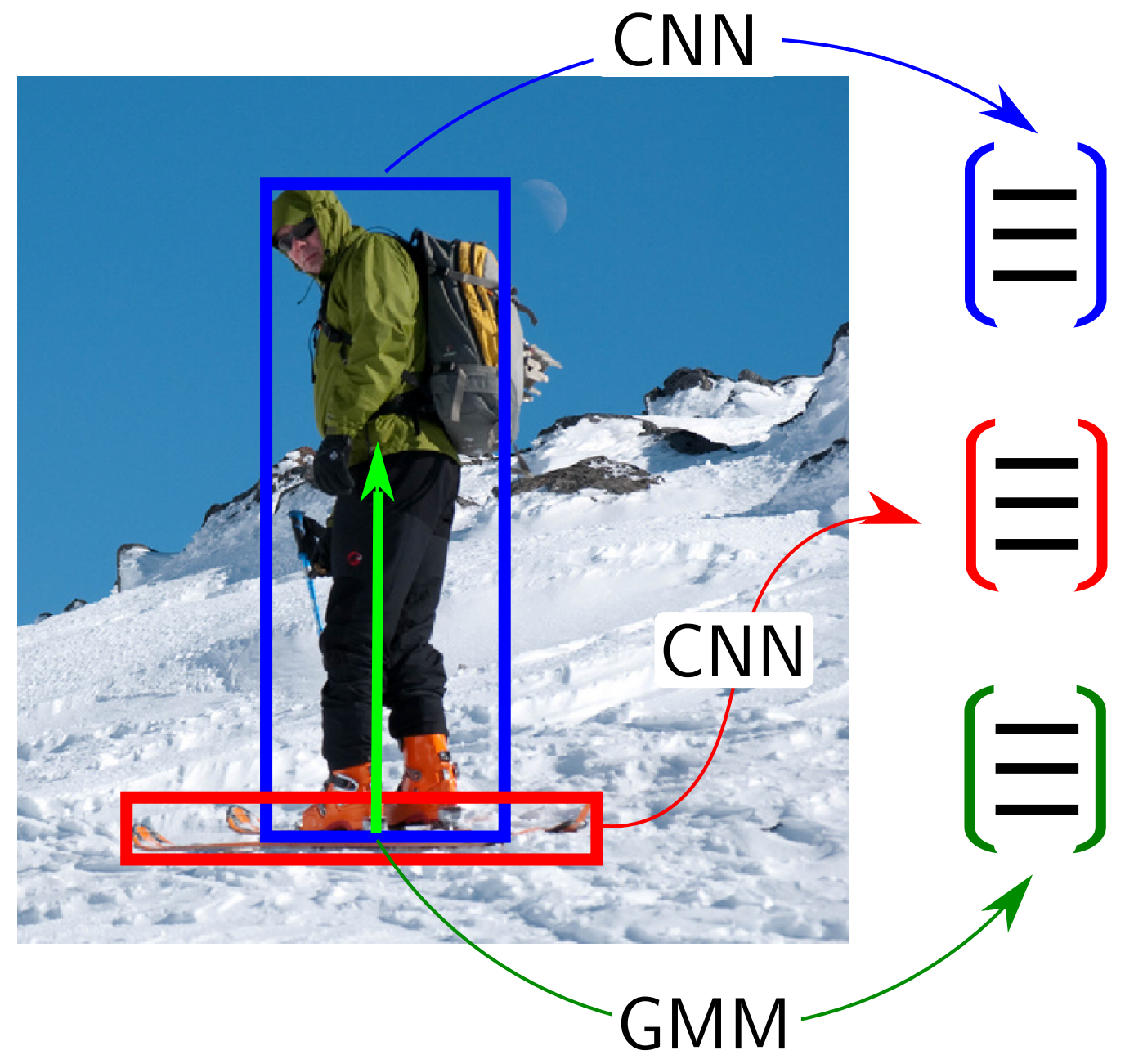
person stand on surfboard
person carry person
person above surfboard
dog on surfboard
person taller than dog

Test



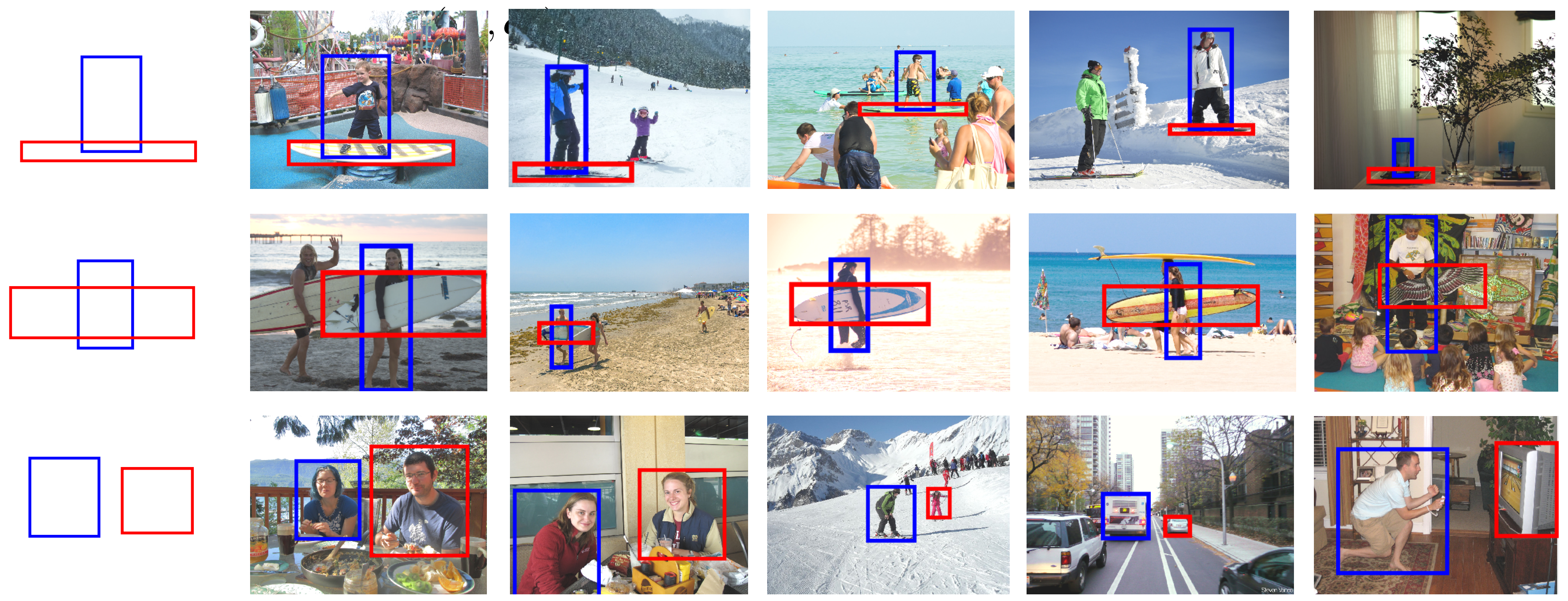
person on surfboard — dog in front of person — dog stand on surfboard

Visual representation of a relation



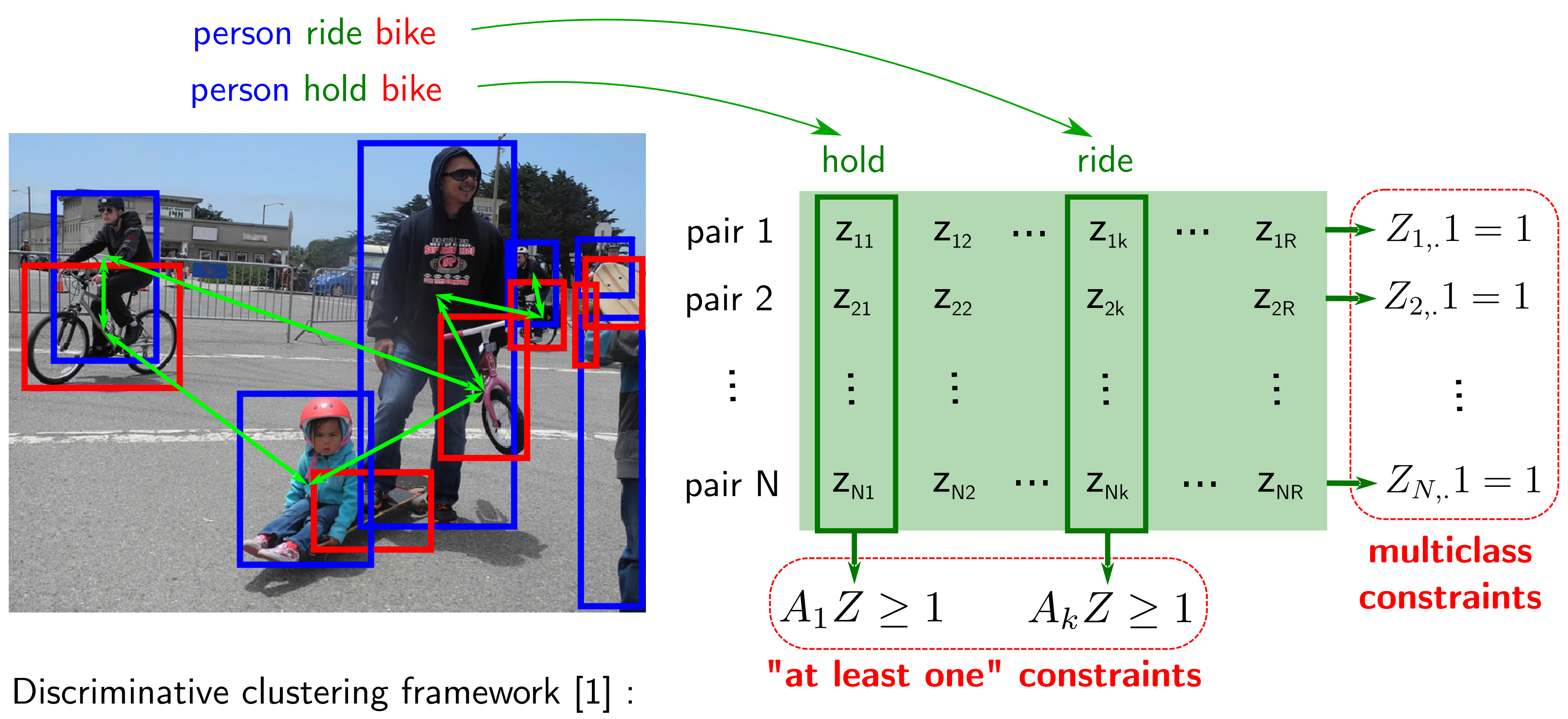
- **Appearance features** for individual objects from fc7 output of Fast-RCNN object detector
- **Quantized spatial configuration** between boxes (with GMM)

$$r(\boldsymbol{o_s},\boldsymbol{o_o}) = [\underbrace{\frac{x_o - x_s}{\sqrt{w_s h_s}}, \frac{y_o - y_s}{\sqrt{w_s h_s}}}_{\text{translation between boxes}}, \underbrace{\sqrt{\frac{w_o h_o}{w_s h_s}}}_{\text{ratio of box sizes}}, \underbrace{\frac{\boldsymbol{o_s} \cap \boldsymbol{o_o}}{\boldsymbol{o_s} \cup \boldsymbol{o_o}}}_{\text{overlap}}, \underbrace{\frac{w_s}{h_s}, \frac{w_o}{h_o}}_{\text{aspect ratio}}]$$



Visualization of the spatial clusters

Learning with image-level labels



Discriminative clustering framework [1] :

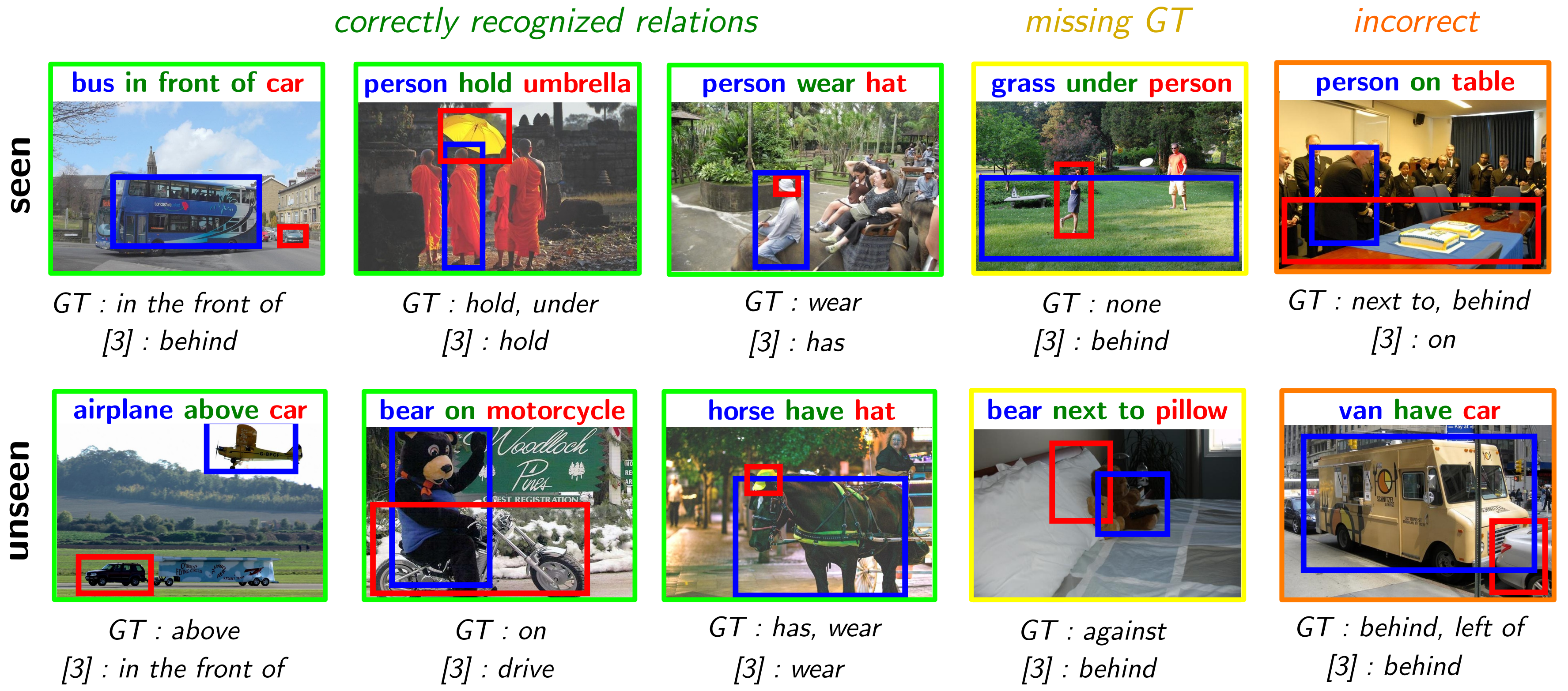
$$\min_{Z \in \mathcal{Z}} \min_{W \in \mathbb{R}^{d \times R}} \underbrace{\frac{1}{N} \| Z - XW \|_F^2}_{\text{matching latent assignments}} + \underbrace{\lambda \| W \|_F^2}_{\text{regularization}}$$

Joint optimization of W and Z with Block-coordinate Frank-Wolfe algorithm [2]

Recall on Visual Relationship Dataset

Metric : recall@k, proportion of ground truth triplets retrieved among the top k detections

	Recall@50	Predicate Det.		Phrase Det.		Relationship Det.	
		All	Unseen	All	Unseen	All	Unseen
full sup.	Visual Phrases [4]	0.9	-	0.04	-	-	-
	Language Prior [3]	47.9	8.5	16.2	3.4	13.9	3.1
	Ours full	**50.4**	**23.6**	**16.7**	**7.4**	**14.9**	**7.1**
weak sup.	Ours weak	46.8	19.0	16.0	6.9	14.1	6.7



correctly recognized relations — *missing GT* — *incorrect*

seen

- bus in front of car — GT : in the front of; [3] : behind
- person hold umbrella — GT : hold, under; [3] : hold
- person wear hat — GT : wear; [3] : has
- grass under person — GT : none; [3] : behind
- person on table — GT : next to, behind; [3] : on

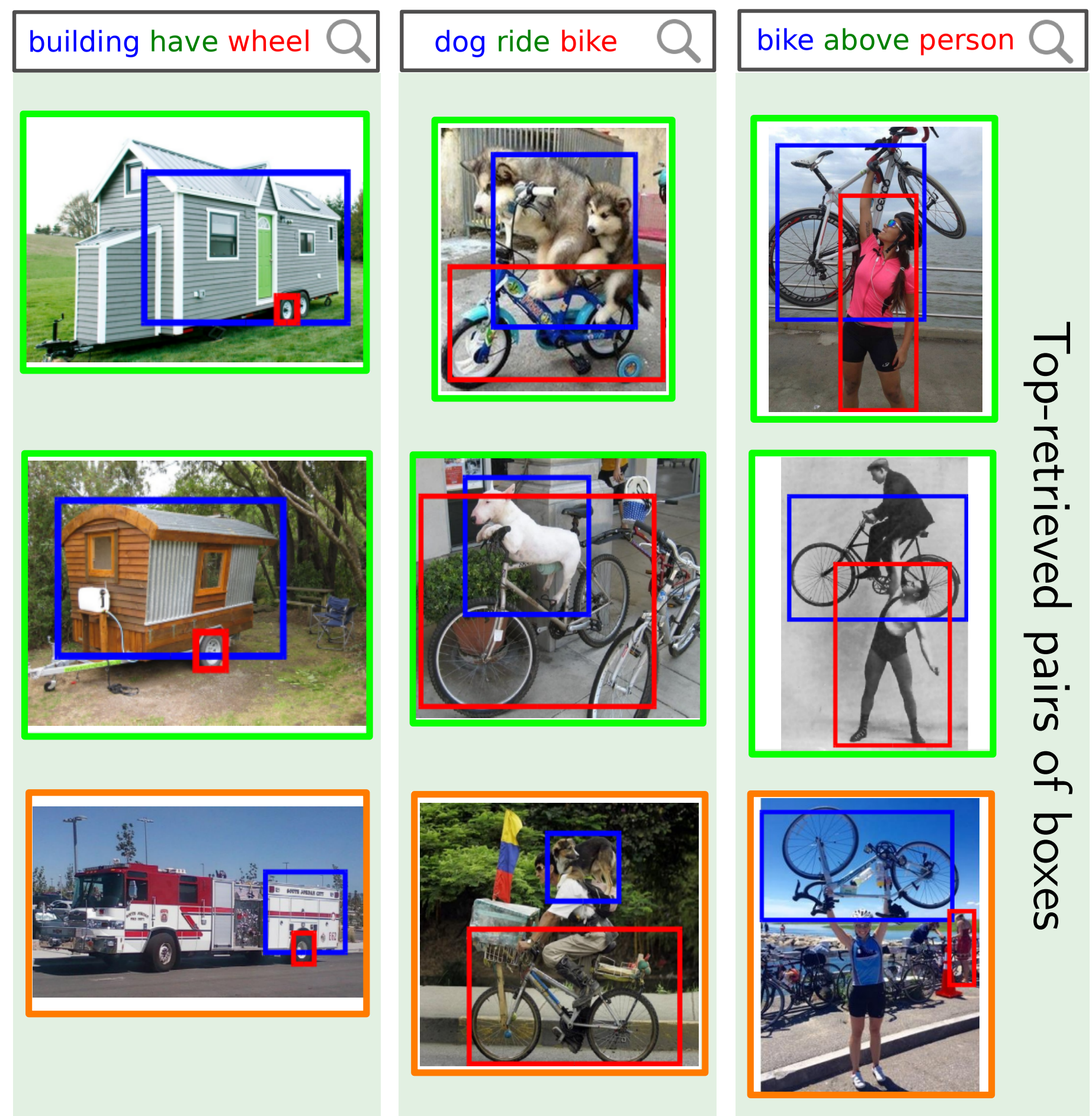
unseen

- airplane above car — GT : above; [3] : in the front of
- bear on motorcycle — GT : on; [3] : drive
- horse have hat — GT : has, wear; [3] : wear
- bear next to pillow — GT : against; [3] : behind
- van have car — GT : behind, left of; [3] : behind

Retrieval for Unusual Relations (UnRel) Dataset

- in the spirit of Out-of-Context dataset
  [Choi et al. Context Models and Out-of-context Objects. In Pattern Recognition Letters, 2012]
- 76 rare queries to test **generalization**
- 1071 images with box-level annotations for relations to evaluate in a **clean setup**
- 1533 relations annotated in total

mAP	With GT	With candidates		
	-	union	subj	subj/obj
DenseCap [5]	-	6.2	6.8	-
Language Prior [3]	50.6	12.0	10.0	7.2
Ours full	**62.6**	**14.1**	**12.1**	**9.9**
Ours weak	58.5	13.4	11.0	8.7



References

[1] F. R. Bach and Z. Harchaoui. Diffrac: a discriminative and flexible framework for clustering. In NIPS, 2007

[2] A. Osokin, J.-B. Alayrac, I. Lukasewitz, P. K. Dokania, and S. Lacoste-Julien. Minding the gaps for block Frank-Wolfe optimization of structured SVMs.

[3] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual relationship detection with language priors. In ECCV, 2016

[4] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In CVPR, 2011

[5] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In CVPR, 2016