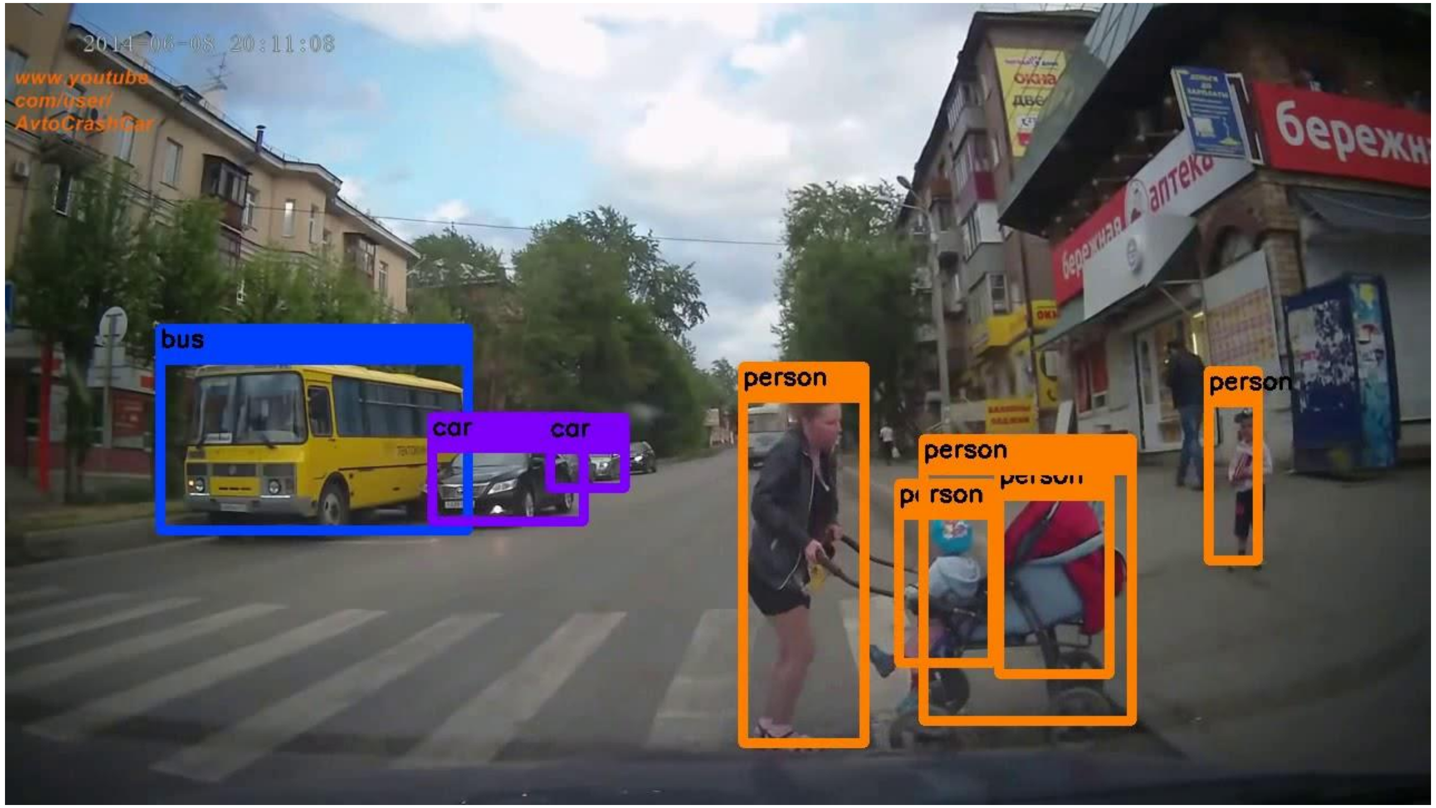
Composite Association Fields with Supervised Deformable Convolutions for Scene Graph Generation

George Adaimi

Sven Kreiss

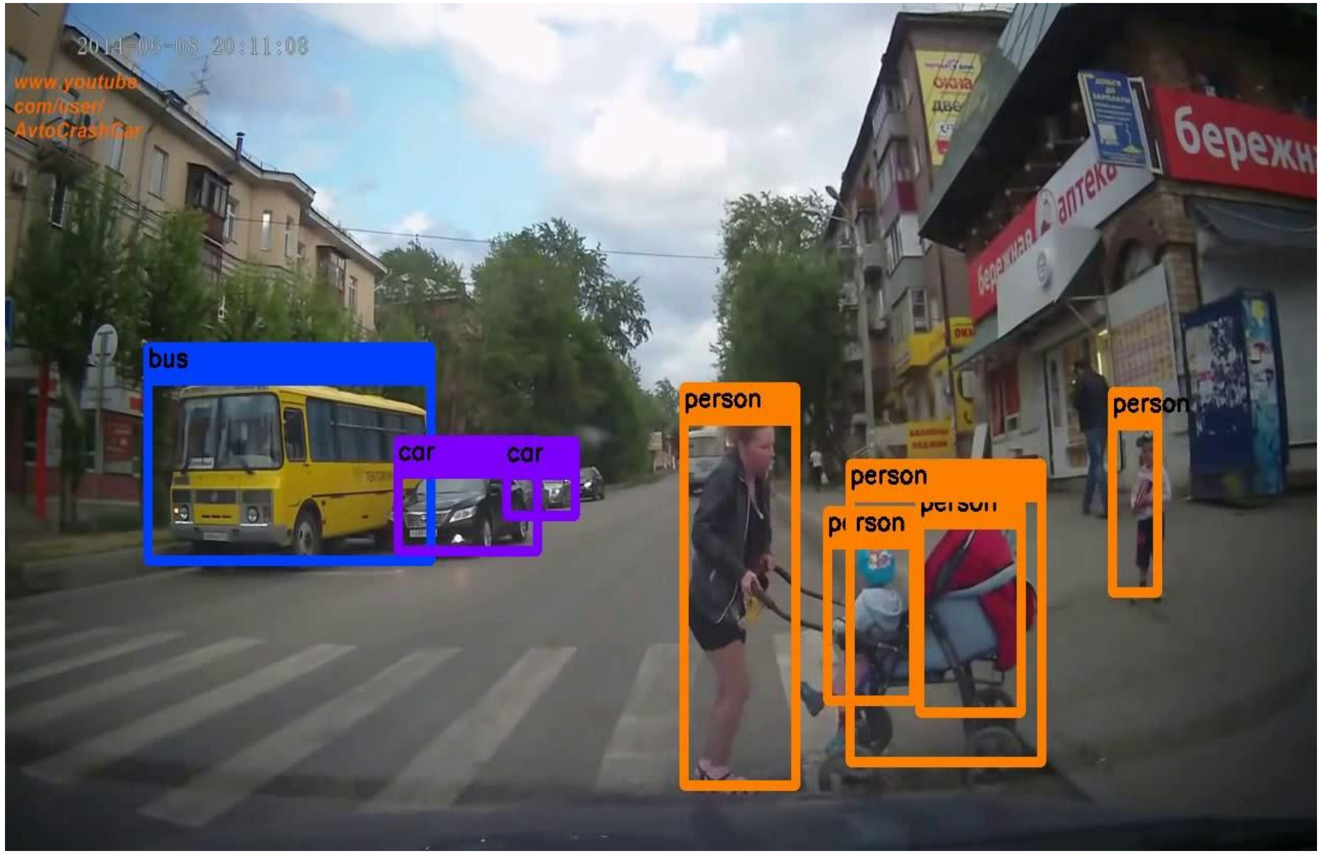
Alexandre Alahi

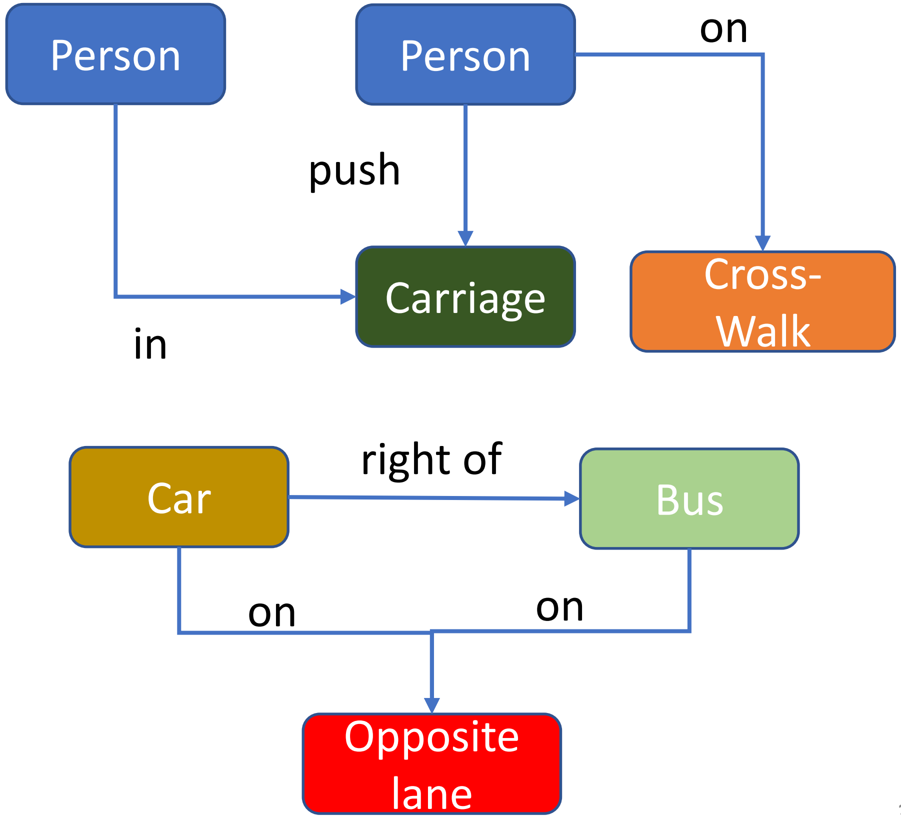
Object Detection



What information do we use to make a decision?

Object Detection –> Scene Graph





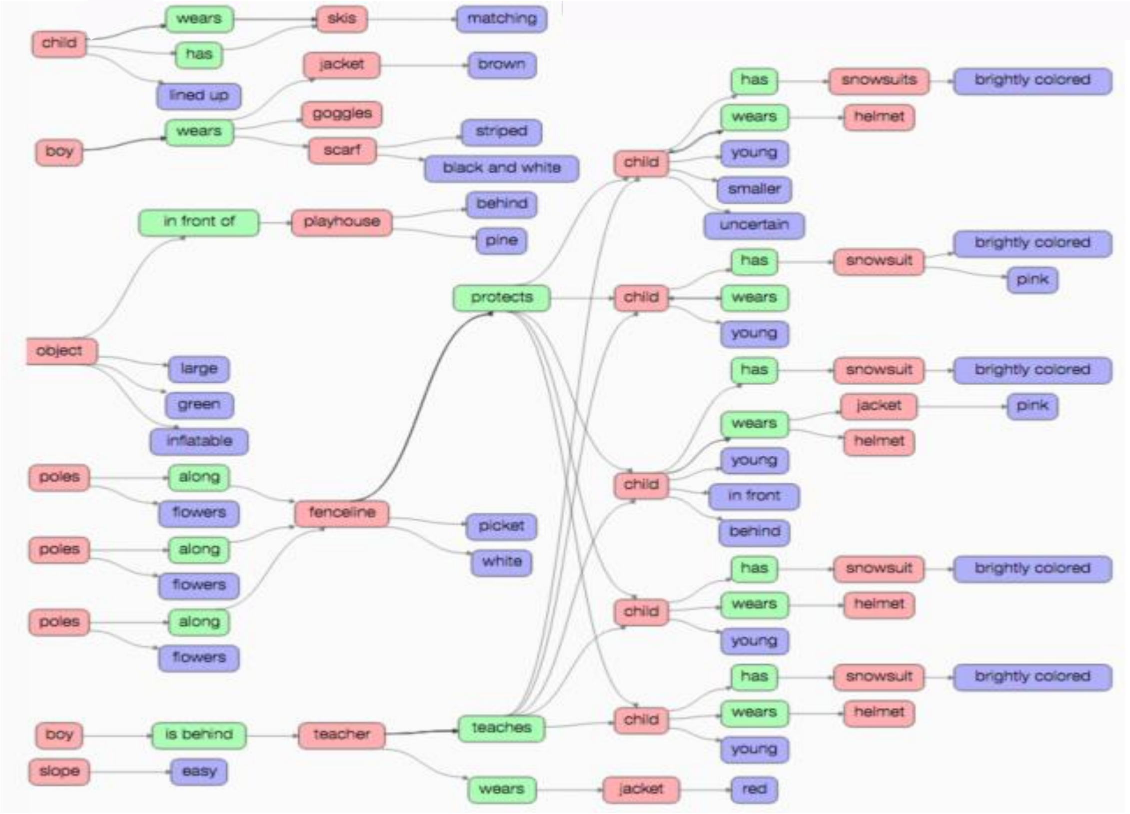
Problem Formulation

Input:

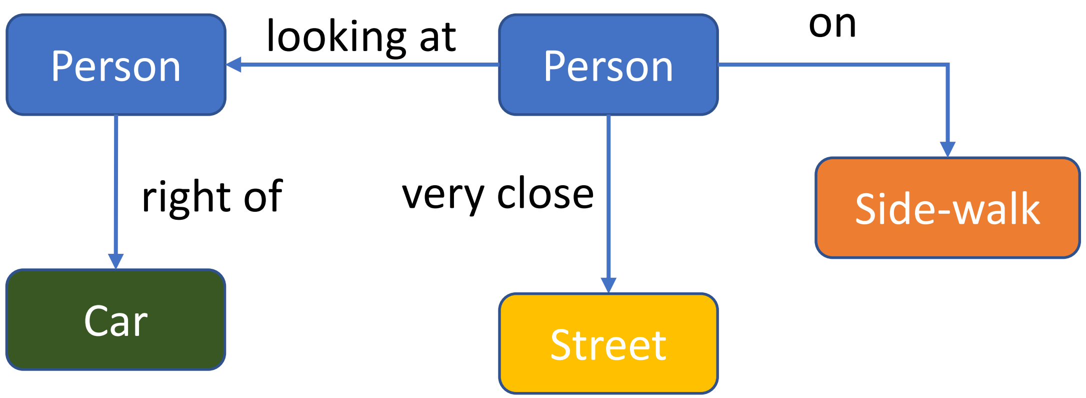
An Image

Output:

Scene Graph <subject, predicate, object>



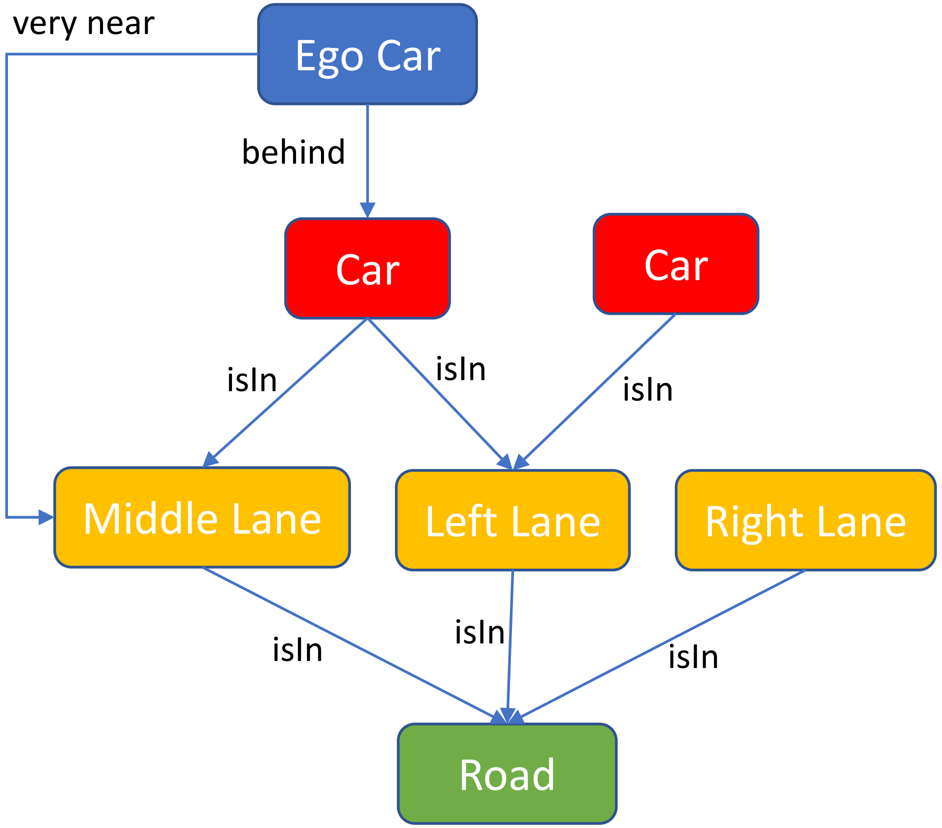
Action/Intention Prediction



Is it enough to detect the people?

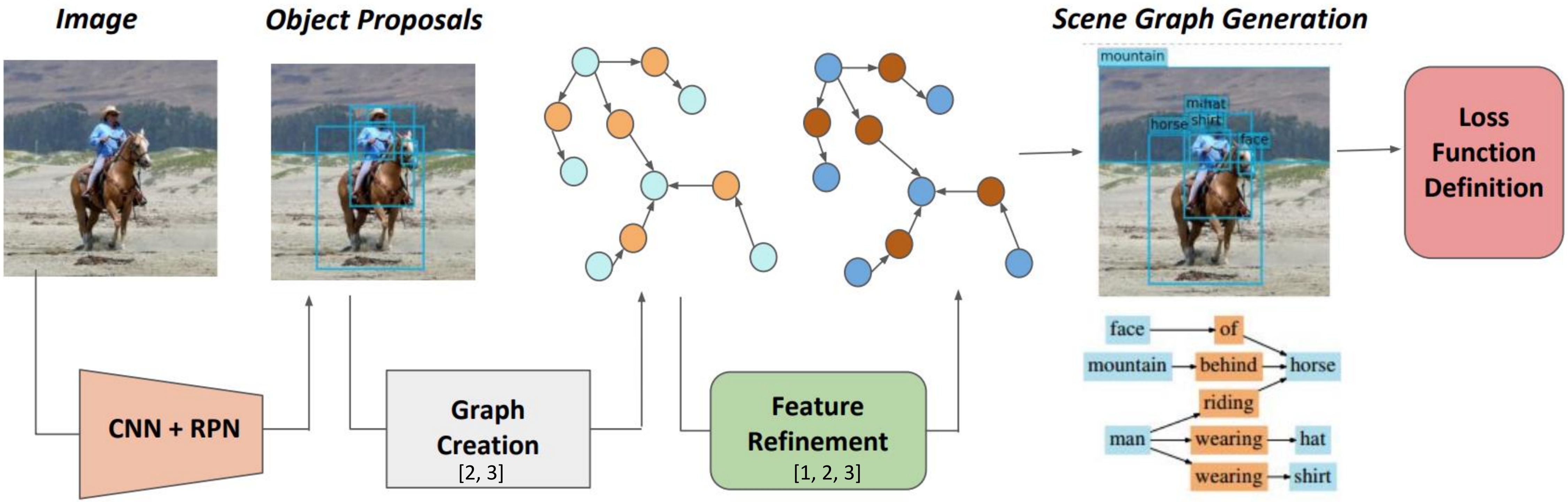
Risk Assessment

Is lane change risky?



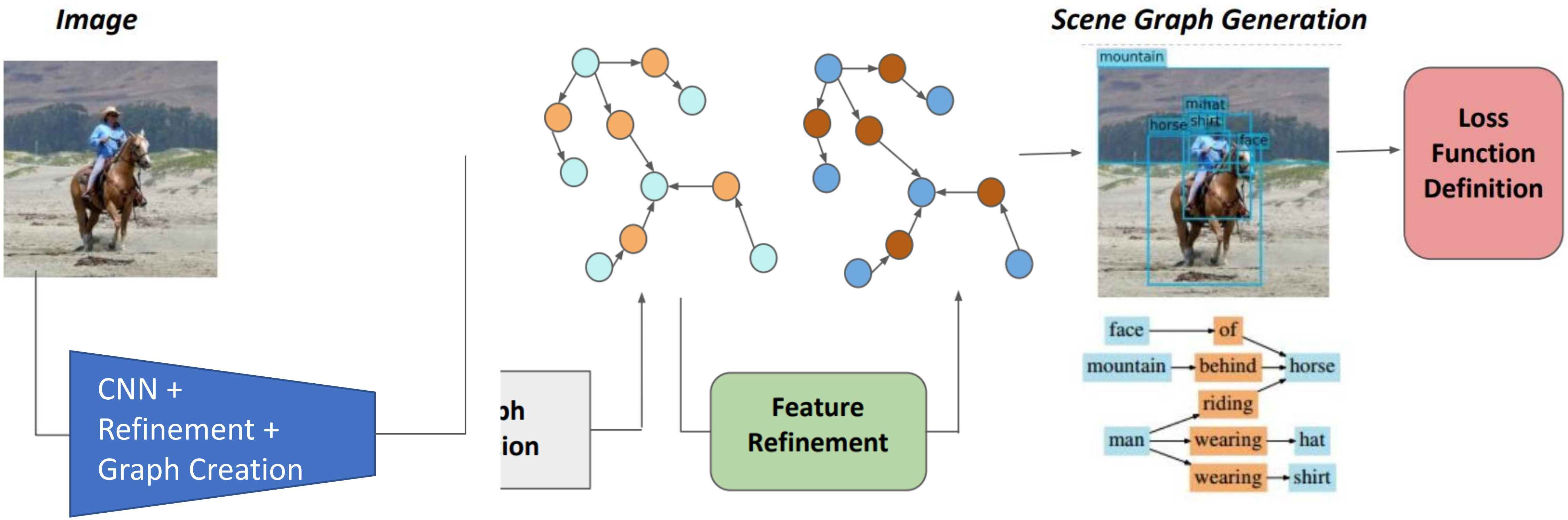
Yu, Shih-Yuan, et al. "Scene-Graph Augmented Data-Driven Risk Assessment of Autonomous Vehicle Decisions." *arXiv preprint arXiv:2009.06435* (2020).

Previous Work

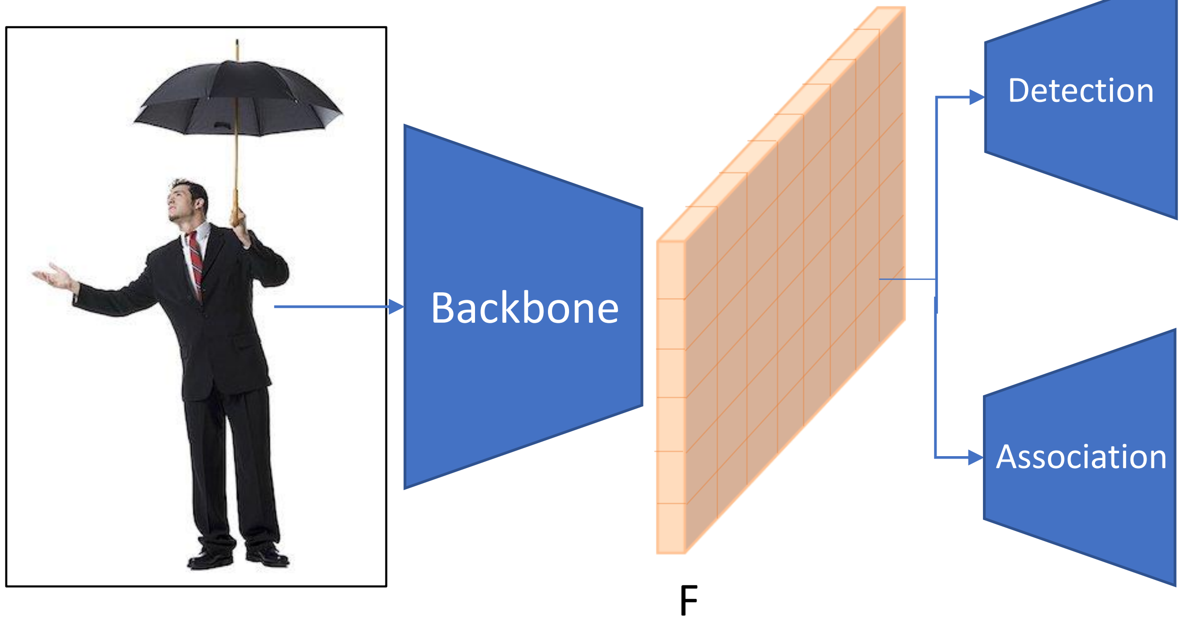


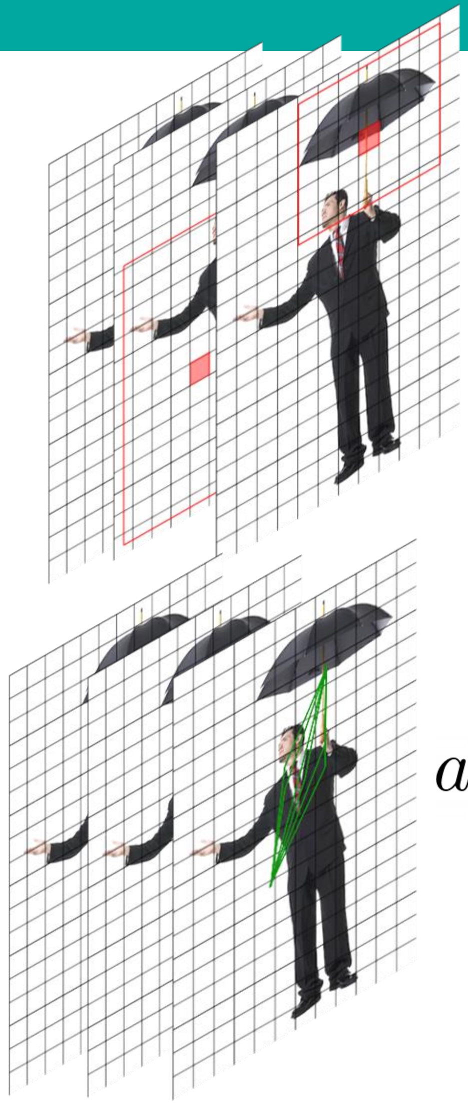
[1] Dai, Bo, Yuqi Zhang, and Dahua Lin. "Detecting visual relationships with deep relational networks." *Proceedings of the IEEE conference on computer vision and Pattern recognition*. 2017.
[2] Li, Yikang, et al. "Factorizable net: an efficient subgraph-based framework for scene graph generation." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
[3] Yang, Jianwei, et al. "Graph r-cnn for scene graph generation." *Proceedings of the European conference on computer vision (ECCV)*. 2018.

Proposed Implementation: Bottom Up



Proposed Implementation: Bottom Up





Object Detection

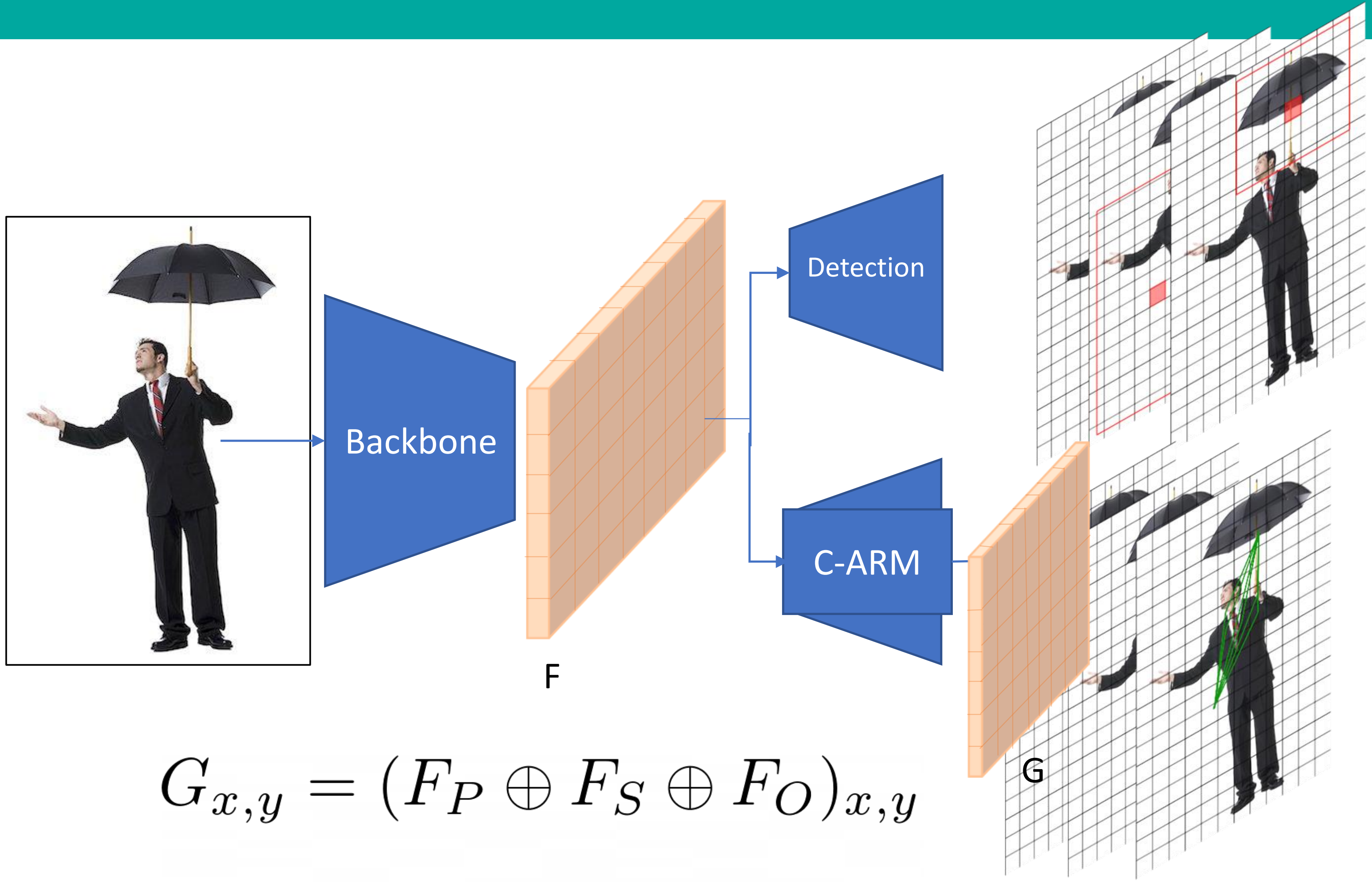
$$f_{ij}^{c} = \{p_{ij}^{c}, x_{ij}^{c}, y_{ij}^{c}, w_{ij}^{c}, h_{ij}^{c}\}$$

CAF [1]

$$a_{ij}^{r} = \{p_{ij}^{r}, x_{ij}^{s}, y_{ij}^{s}, x_{ij}^{o}, y_{ij}^{o}\}$$

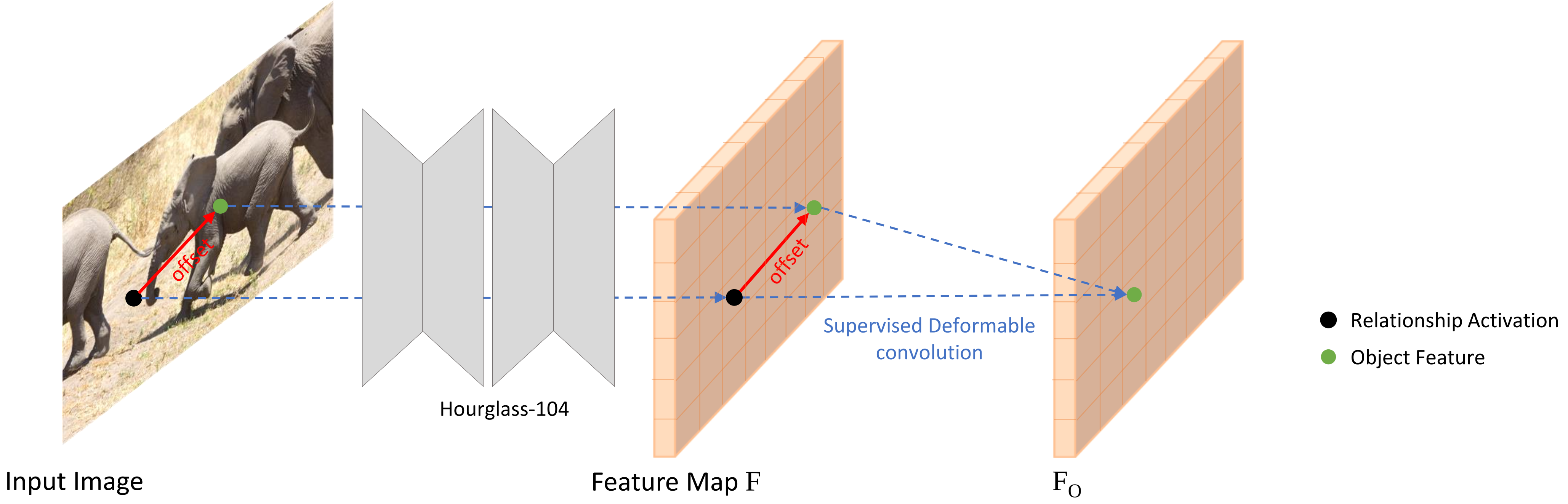
[1] Kreiss, Sven, Lorenzo Bertoni, and Alexandre Alahi. "OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association." *arXiv preprint arXiv:2103.02440* (2021).

Proposed Implementation: Refinement



$$G_{x,y} = (F_P \oplus F_S \oplus F_O)_{x,y}$$

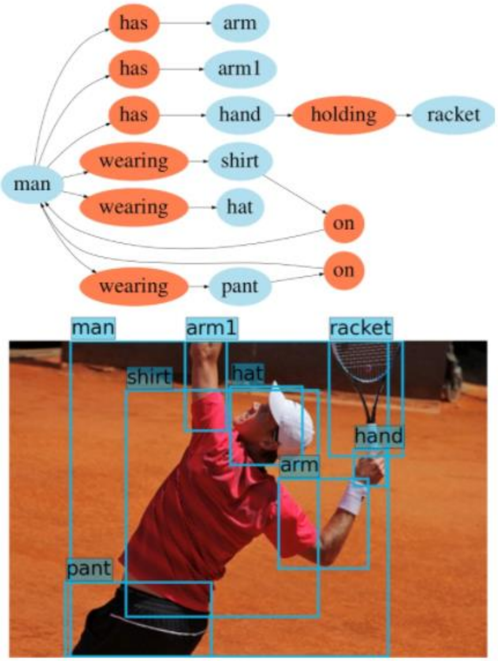
Proposed Implementation: Refinement



$$G_{x,y} = (F_P \oplus F_S \oplus F_O)_{x,y} = \underbrace{(W_r \cdot F_{x,y})}_{\text{predicate}} \oplus \underbrace{(W_s \cdot F_{x_s,y_s})}_{\text{subject}} \oplus \underbrace{(W_o \cdot F_{x_o,y_{so}})}_{\text{object}}$$

Datasets & Experiments

- **Visual Genome**
  - 108,249 images
  - 33,877 object categories
  - 42, 374 Relationship Categories
  - Full Scene Graph



Evaluation Metrics

- Predicate Classification (PredCls)
- Scene Graph/Phrase Classification (SGCls)
- Scene Graph Detection (SGDet)

Ablation Study

Table 3: Ablation study on the effect of C-ARM

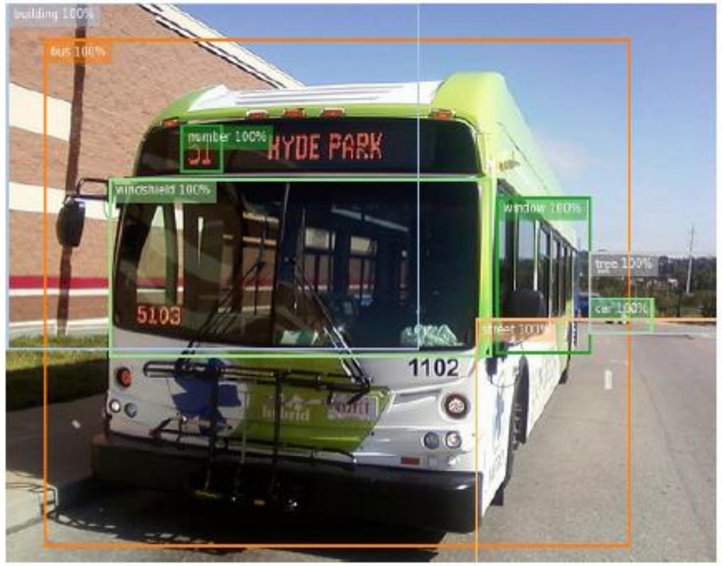
		PredCls		SGCls		SGDet	
	$AP_{0.5}$	R@50	ng-R@50	R@50	ng-R@50	R@50	ng-R@50
Baseline	18.1	44.57	56.86	17.15	19.86	14.58	17.21
+ C-ARM (Ours)	**19.7**	**45.79**	**58.20**	**18.31**	**21.48**	**15.99**	**18.47**

Quantitative Results

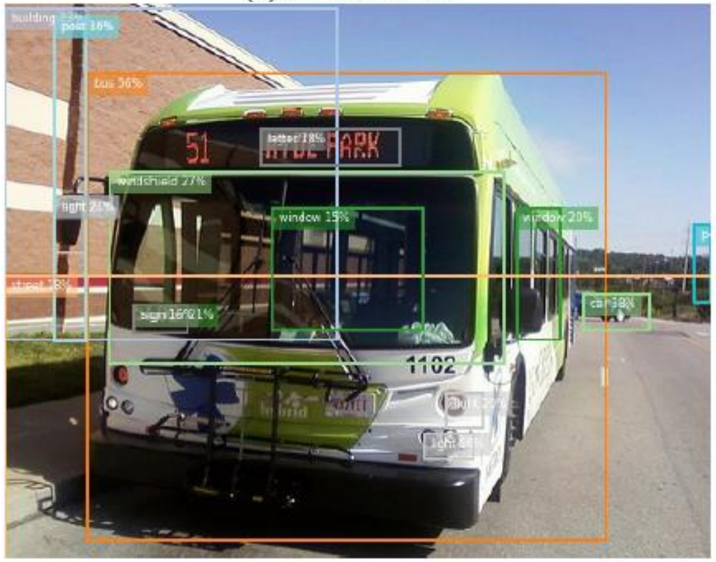
Table 1: Recall@50 for graph and no-graph constraint on Visual Genome [43]. $\star$ indicates that [9] trained a different model for each metric whereas all non-italic methods used the same model for all metrics. f indicates using frequency bias. RPN = Region Proposal Network [11].

			PredCls		SGCls		SGDet	
		$AP_{0.5}$	R@50	ng-R@50	R@50	ng-R@50	R@50	ng-R@50
Top-down	IMP [12]	–	44.8	–	21.7	–	3.4	–
	Graph R-CNN [7]	23.0	54.2	–	29.6	–	11.4	–
	VRF [8]	–	56.7	–	23.7	–	13.2	–
	CISC [18]	–	53.2	–	27.8	–	11.4	–
	LinkNet [19]	–	67.0	–	41	–	27.4	–
Bottom-up	Px2Graph* [9]	–	–	*68.0*	–	*26.5*	–	*9.7* (RPN)
	Px2Graph$^{\star}_{new}$ [9]	–	–	*82.0*	–	*35.7*	–	*15.5* (no RPN)
	FCSGG$_{W32}$ [10]	21.6	34.9	46.3	15.5	19.3	15.1	18.2
	FCSGG$_{W48}$ [10]	25.0	31.0	40.3	17.1	19.6	15.5	18.3
	Ours	19.7	44.83	57.22	17.96	21.09	15.83	17.97
	Ours$_f$	19.7	**45.79**	**58.20**	**18.31**	**21.48**	**15.99**	**18.47**

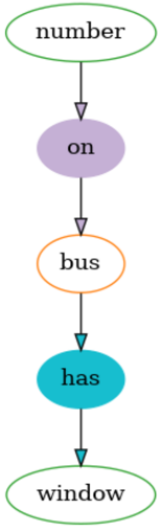
Qualitative Results



(a) GT detections



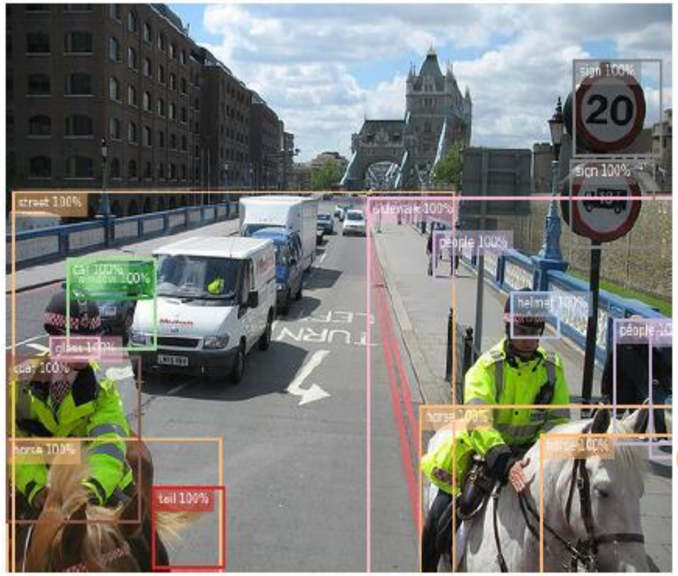
(b) Predicted detections



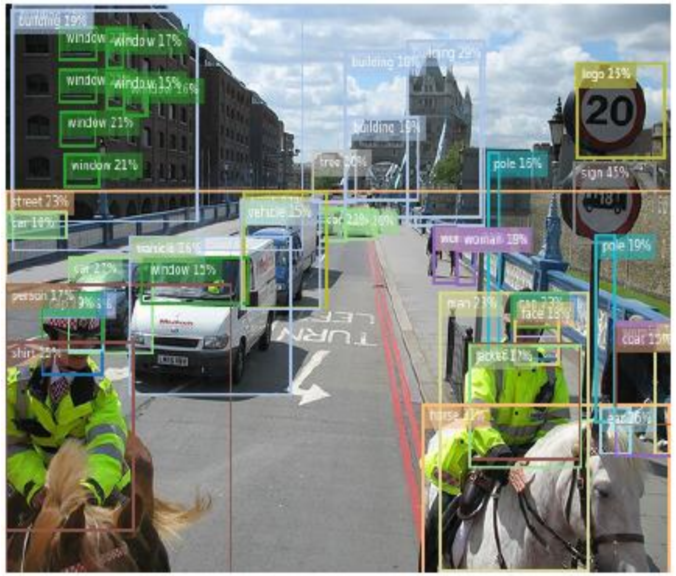
(c) GT Scene Graph

(d) Composite Association Fields for different predicates

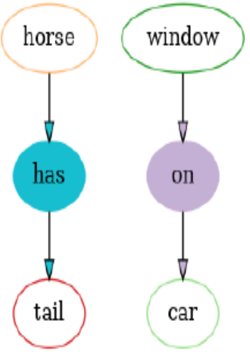
Qualitative Results



(a) GT detections



(b) Predicted detections



(c) GT Scene Graph

(d) Composite Association Fields for different predicates

Thank you!