Context-Aware Transfer Attacks for Object Detection

Zikui Cai, Xinxin Xie, Shasha Li, Mingjun Yin, Chengyu Song,
Srikanth V. Krishnamurthy, Amit K. Roy-Chowdhury, M. Salman Asif
University of California, Riverside

AAAI-22

Introduction

- Deep neural networks (DNNs) are vulnerable to adversarial attacks. A large volume of work has been devoted to explore the vulnerabilities of DNNs, however the vast majority of work has been on image classification, and the more challenging task, object detectors, has been underexplored.
- Black-box attack is a practical yet challenging setting. Transfer-based attacks often suffer from low attack success rate. Query-based attacks have high success rates but require an overwhelmingly large number (often hundreds or thousands) of queries. We explore a more stringent case where only a very small number of model calls is allowed.
- Object detectors take a holistic view of the image and the detection of one object (or lack thereof) depends on other objects in the scene. This is why object detectors are inherently context-aware and adversarial attacks are more challenging than those targeting image classifiers.
- We focus on the problem of generating context-aware adversarial attacks on images to affect the performance of object detectors.

Proposed Method

Four Main Components of our Framework

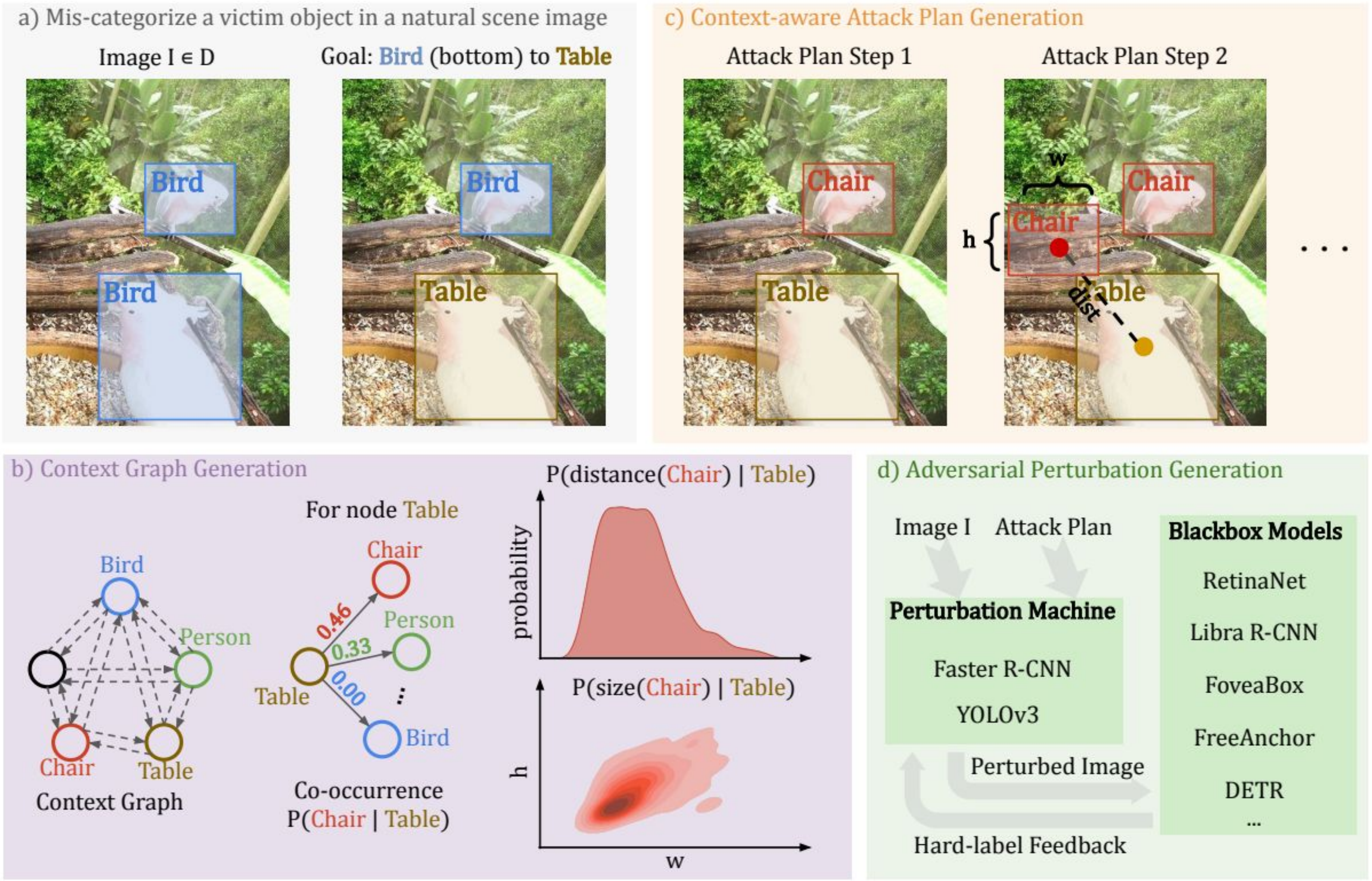
A. Given a natural image, our goal is to trick an object detector to assign the victim object a given target label (e.g., bird to table);

B. We construct a context graph that encodes the co-occurrence probability, distance, and relative size distribution relating pairs of objects (e.g., the edge from table to chair represent they co-occur with probability 0.46);

C. Given the attack goal and context graph, we generate a context-aware attack plan that has a small number of steps. In each step, we assign target labels for existing objects and introduce new helper objects if needed. For example, co-occurrence of chair with table is most probable, we change the bird to a chair for stronger context consistency (depicted in Attack Plan Step 1). We may need to add a phantom chair around the table (as depicted in Attack Plan Step 2)

D. Given the attack plan and the victim image, we generate perturbations using I-FGSM on the surrogate whitebox models in our perturbation machine. We test the perturbed image with the given blackbox model and based on the hard-label feedback, we either stop (when the attack is successful or when we exhaust our budget of the helper objects) or craft new attack based on the next steps and repeat the process.

Framework of Context-aware Attacks



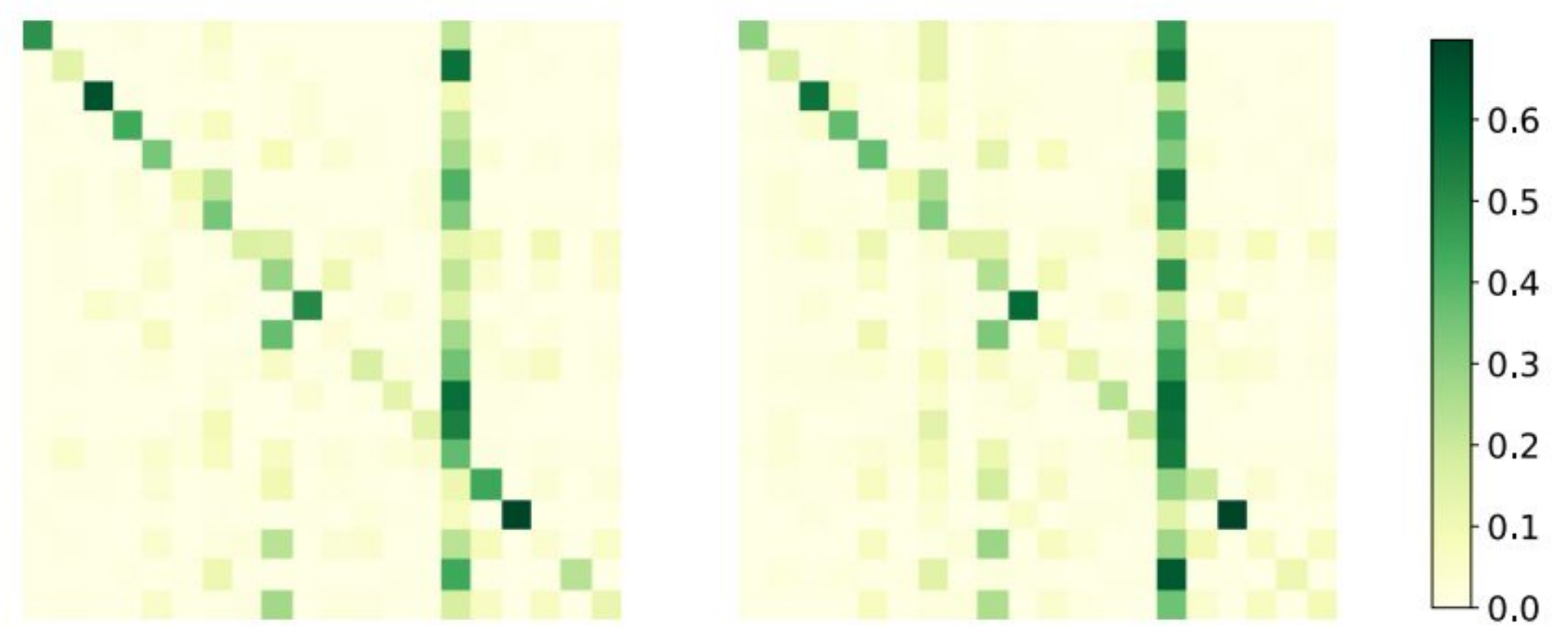
Experimental Setup

- **Attack type:**
  - Mis-categorization attack at different perturbation levels ($L_\infty \le \{10,20,30\}$)
- **Datasets:**
  - PASCAL VOC and MS COCO.
  - Evaluated using 500 images that contain multiple (2–6) objects for each dataset
- **Object detectors:**
  - Surrogate model: an ensemble of Faster R-CNN and YOLOv3
  - Victim models: different two-stage, one-stage, anchor-free, and transformer-based detectors
- **Comparisons:**
  - Baseline is where no helper object is added
  - Random is where the helper objects are added in a randomized fashion (mismatched context)
- **Evaluation metric:**
  - Use attack success rate (or fooling rate) to evaluate the adversarial attack performance on any victim object detector

Results and Analysis

- Our approach performs significantly better than context-agnostic and mismatched context approach.
- Fooling rate increases with the number of helper objects and plateaus at around #5.
- Strong positive correlation between co-occurrence relationships encoded by different context graphs.

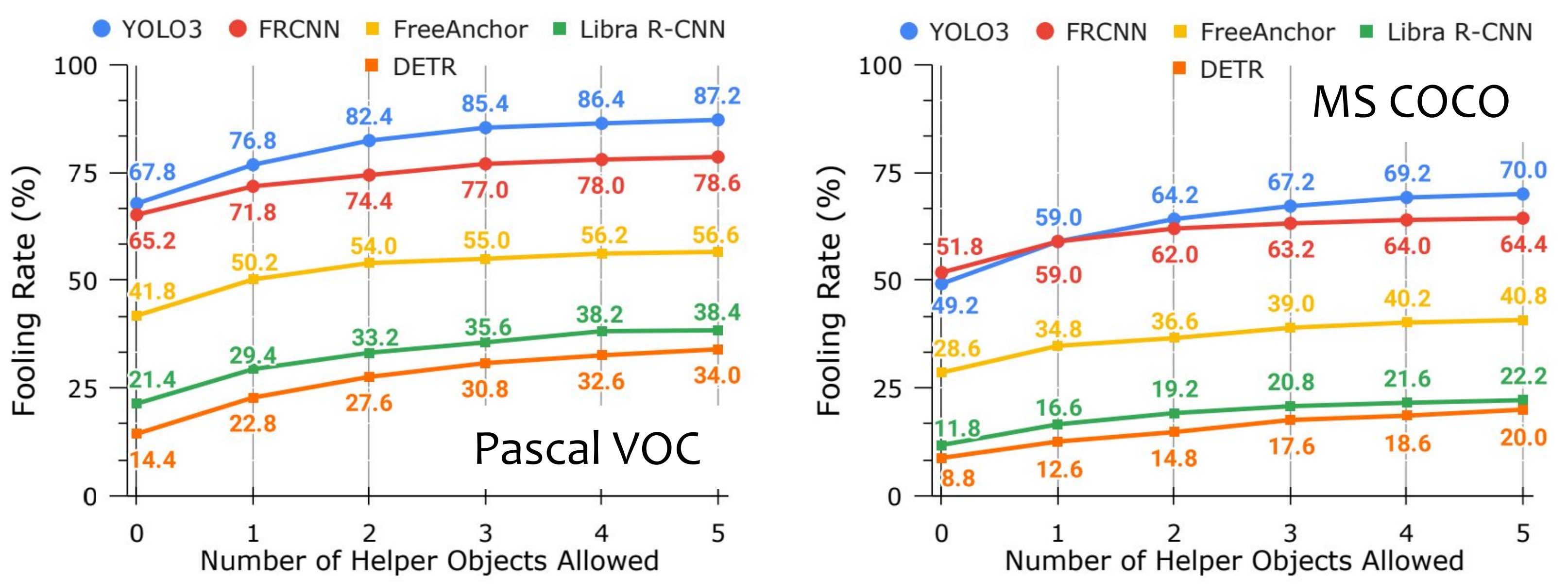
Co-occurrence matrices



Pascal VOC

MS COCO

Mis-categorization attack fooling rate w.r.t. number of helper objects



Mis-categorization attack fooling rate on different models

Perturbation Budget	Method	Whitebox		Blackbox					
		FRCNN	YOLOv3	Retina	Libra	Fovea	Free	DETR	D-DETR
				Results on PASCAL VOC					
$L_\infty \le 10$	Baseline	40.0	53.8	13.8	9.2	22.2	27.4	9.6	23.2
	Random	52.4	69.2	19.4	17.4	31.6	37.8	17.4	36.8
	Ours	**55.8**	**75.6**	**22.6**	**20.4**	**33.6**	**39.2**	**20.2**	**39.2**
$L_\infty \le 20$	Baseline	65.2	67.8	24.0	21.4	34.4	41.8	14.4	37.6
	Random	74.4	83.8	31.0	29.6	46.2	54.4	28.0	52.6
	Ours	**78.6**	**87.2**	**35.2**	**38.4**	**51.6**	**56.6**	**34.0**	**58.4**
$L_\infty \le 30$	Baseline	70.6	70.4	29.8	28.6	41.6	48.0	20.4	38.6
	Random	79.2	82.6	37.8	36.8	53.4	59.8	34.4	52.8
	Ours	**80.6**	**88.0**	**42.0**	**44.2**	**56.8**	**63.6**	**40.2**	**59.0**

Summary

- Our context-aware adversarial attack method exploits rich object co-occurrence relationships plus location and size information;
- Our method can effectively improve mis-categorization attack fooling rate against a large variety of blackbox object detectors;
- The attack performance significantly improves and gradually plateaus as we add around 5 helper objects;
- The contextual relationships modeled by our method holds true in different datasets within natural image domain, making our methods applicable to a wide range of datasets.

Acknowledgment: This material is based upon work supported by DARPA under Agreement No. HR00112090096.