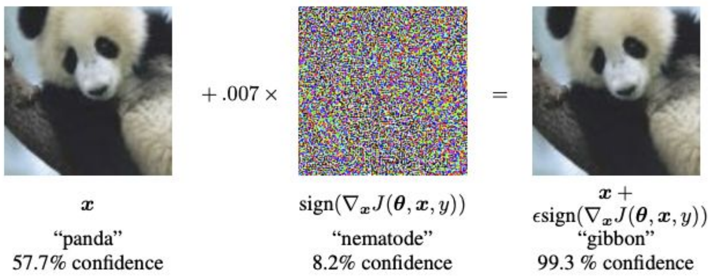
Context-Aware Transfer Attacks for Object Detection

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

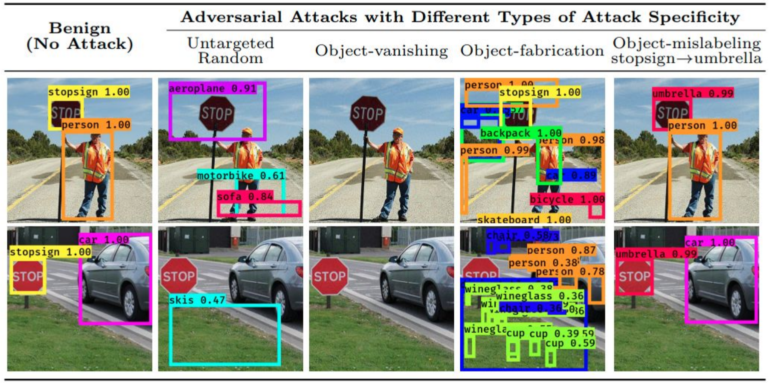
University of California, Riverside

Background

- Adversarial attacks
  - Most methods are developed for classification
  - Attacking object detectors is more challenging



[1] An imperceptible perturbation can fool a classifier

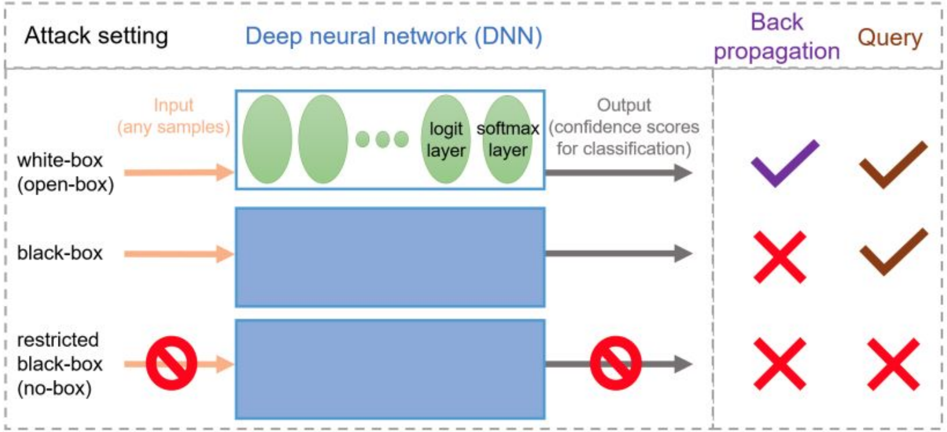


[2] Different types of attacks on object detectors

[1] Goodfellow et al. 2015. Explaining and Harnessing Adversarial Examples. ICLR.
[2] Chow et al. 2020. TOG: Targeted Adversarial Objectness Gradient Attacks on Real-time Object Detection Systems. IEEE TPS.

Background

- Black-box attack approaches
  - Query-based
  - Transfer-based

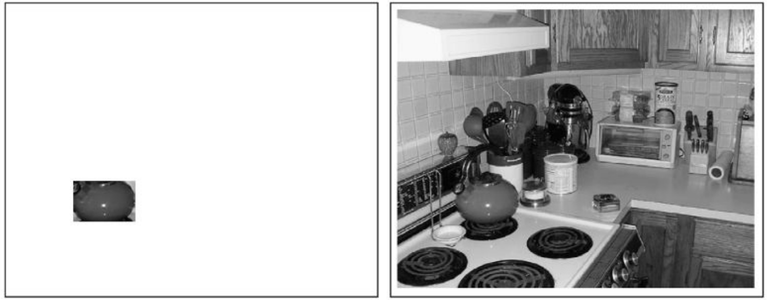


[1] Without access to internal parameters, blackbox attacks are more practical yet challenging

[1] Chen et al. 2017. ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models. ACM AISEC.

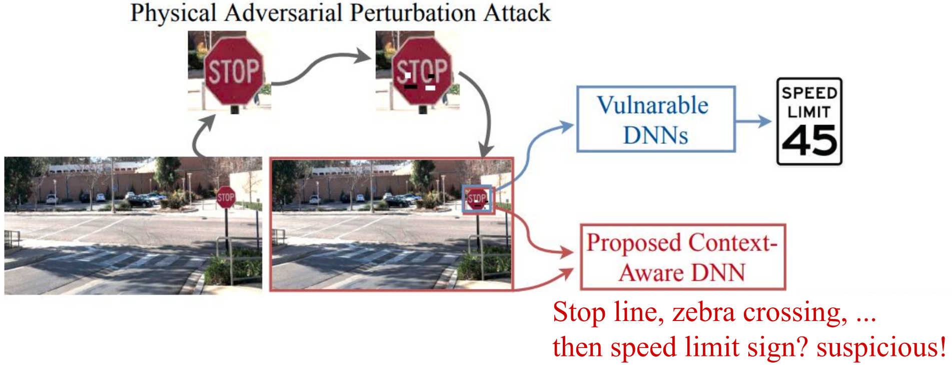
Background

- Contexts in computer vision
  - Context for visual recognition
  - Context-awareness of object detectors



(a) (b)

[1] It is difficult to recognize 'kettle' without its surroundings



[2] Context-awareness in object detection

[1] Galleguillos et al. 2010. Context based object categorization: A critical survey. CVIU.
[2] Li et al. 2020. Connecting the Dots: Detecting Adversarial Perturbations Using Context Inconsistency. ECCV.

Context-aware transfer attacks

Quick overview and key ideas

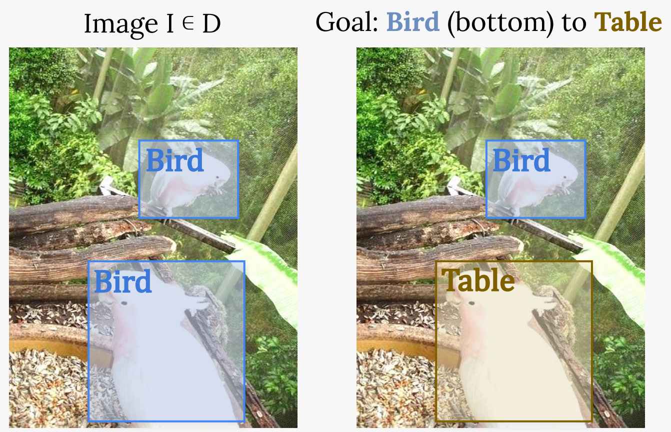
Image $I \in D$

Context-aware transfer attacks

Quick overview and key ideas

- Goal is to misclassify the victim object to a target label

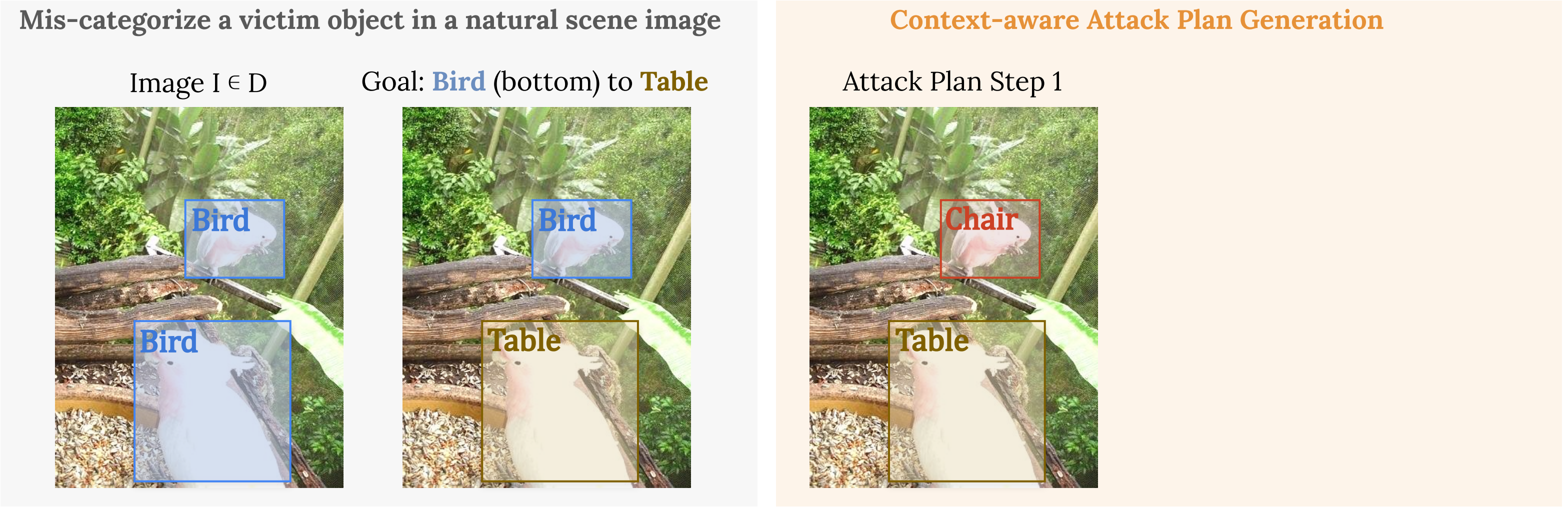
Mis-categorize a victim object in a natural scene image



Context-aware transfer attacks

Quick overview and key ideas

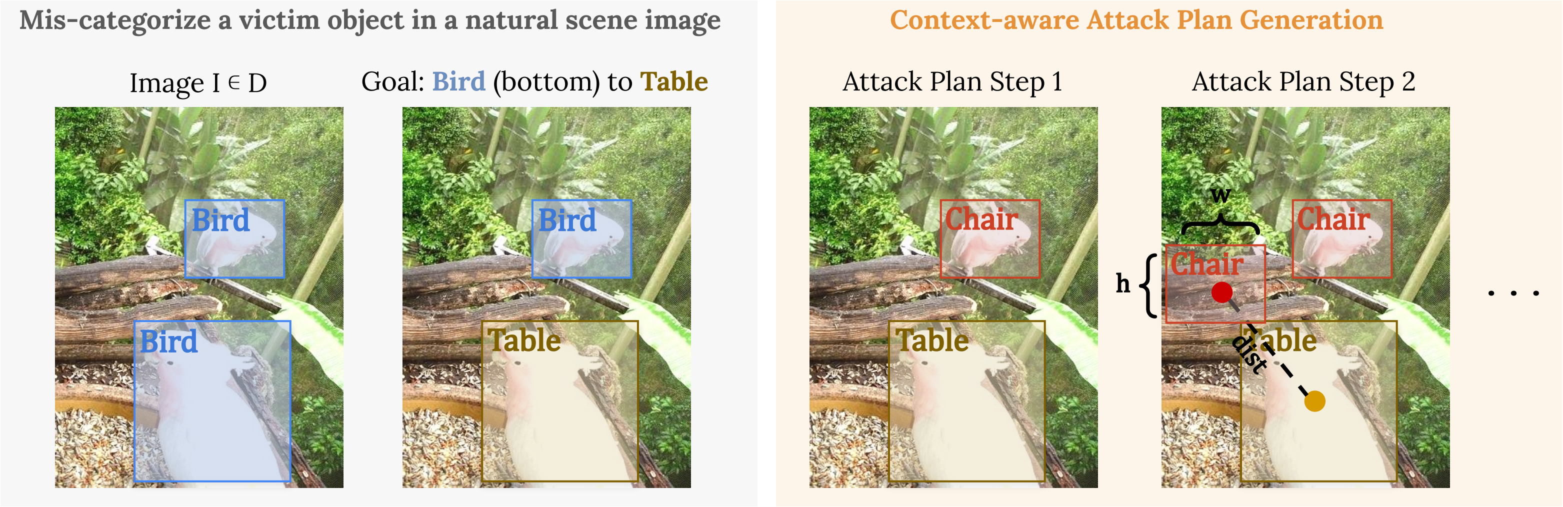
- Goal is to misclassify the victim object to a target label
- To do so, we perturb both the victim object and the "context" associated with the victim object



Context-aware transfer attacks

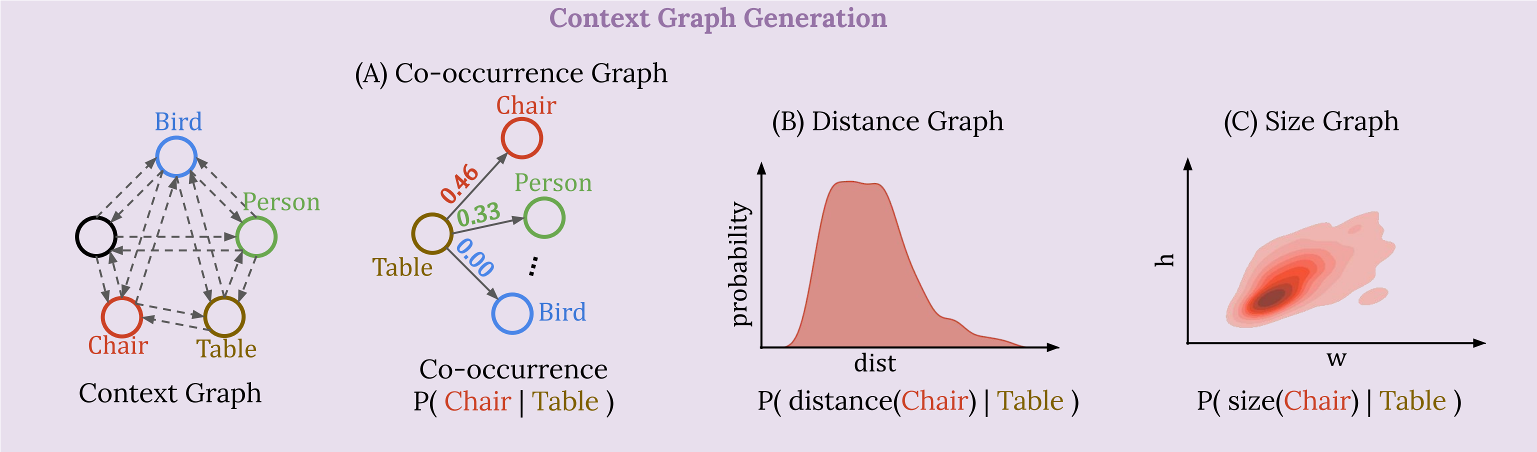
Quick overview and key ideas

- Goal is to misclassify the victim object to a target label
- To do so, we perturb both the victim object and the "context" associated with the victim object
- We keep adding helper objects to enhance the context if necessary



Context-aware transfer attacks

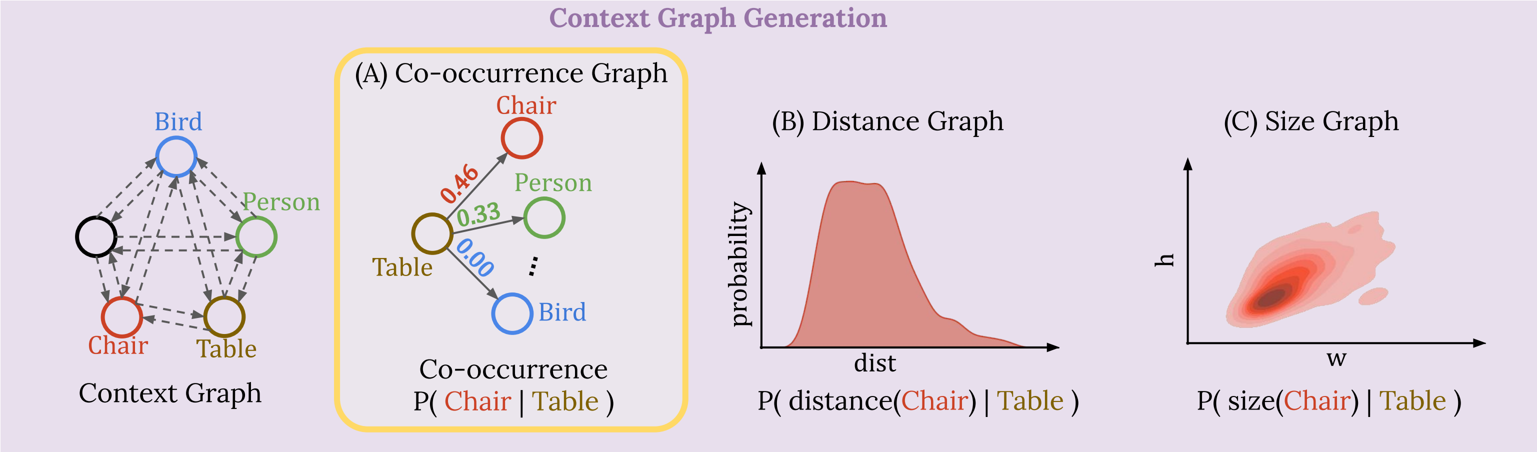
Context Modeling



Context-aware transfer attacks

Context Modeling

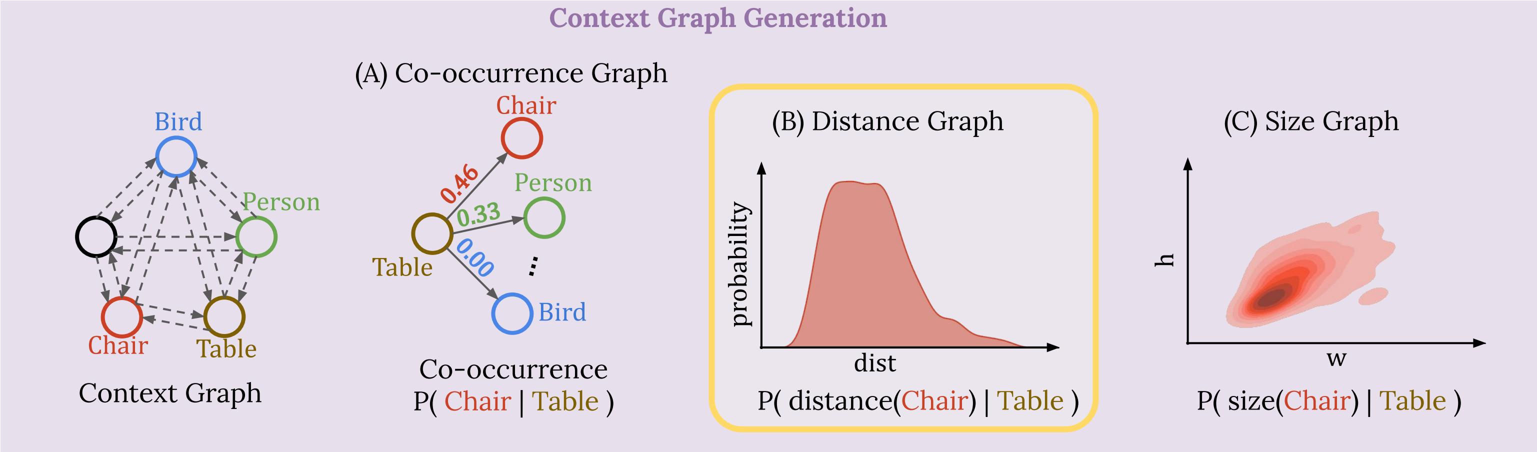
A. Co-occurrence graph: models co-occurrence probability of each pair of instances in images



Context-aware transfer attacks

Context Modeling

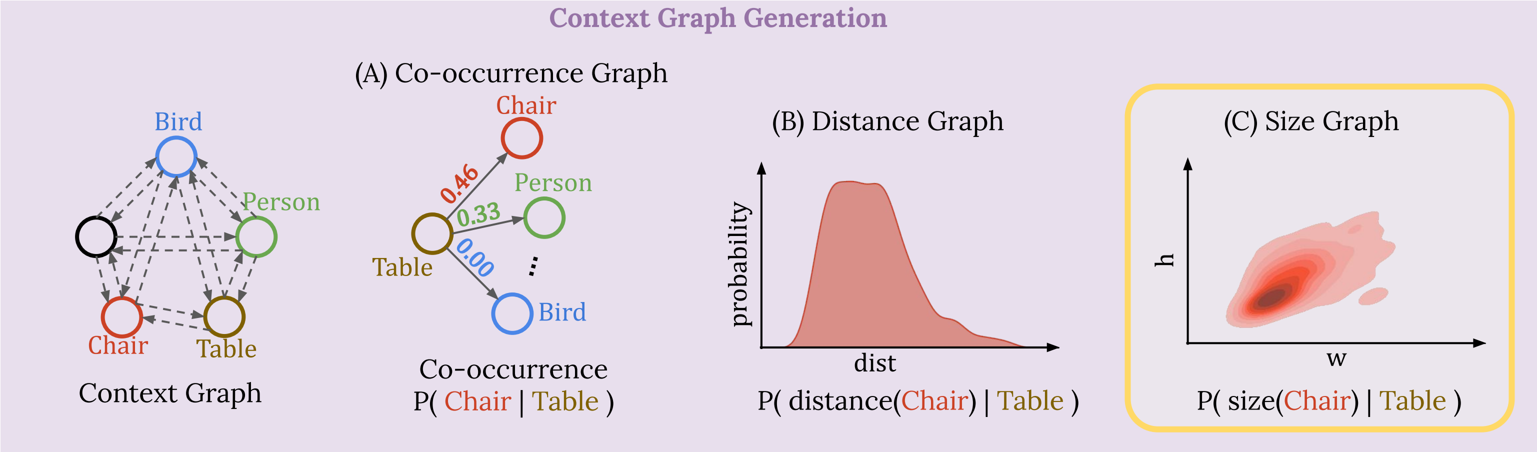
A. Co-occurrence graph: models co-occurrence probability of each pair of instances in images
B. Distance graph: models conditional distance distribution of objects



Context-aware transfer attacks

Context Modeling

A. Co-occurrence graph: models co-occurrence probability of each pair of instances in images
B. Distance graph: models conditional distance distribution of objects
C. Size graph: models the conditional distribution of heights and widths of objects

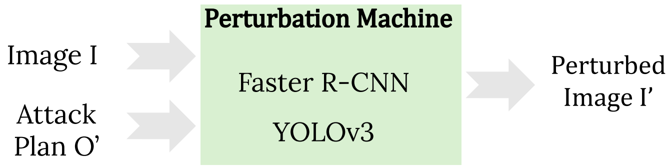


Context-aware transfer attacks

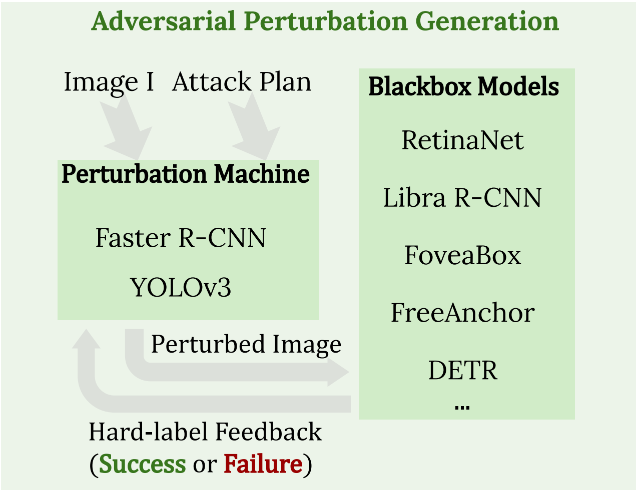
Adversarial perturbation generation

- A diverse set of object detectors (one-stage, two-stages, anchor-free, transformer-based)
  - one-stage, two-stages for perturbation machine, all four types for victim models
- Can work with different types of attack methods (variants of I-FGSM)

Attack Method
Perturbation machine as a surrogate model
to generate perturbed image



I' = argmin_I' Loss(I, O')
Attack using I-FGSM

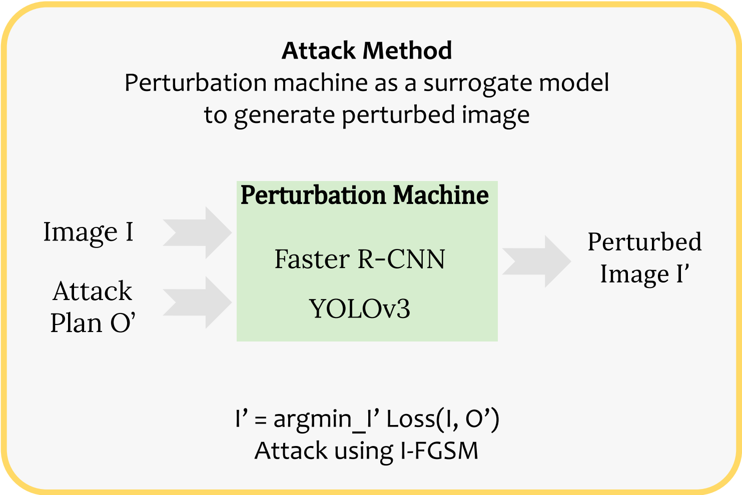


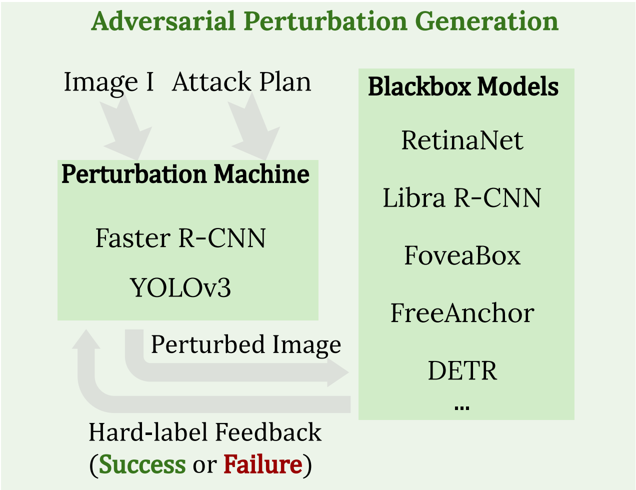
Repeat the attack process for a few iterations
until success or run out of budget

Context-aware transfer attacks

Adversarial perturbation generation

- A diverse set of object detectors (one-stage, two-stages, anchor-free, transformer-based)
  - one-stage, two-stages for perturbation machine, all four types for victim models
- Can work with different types of attack methods (variants of I-FGSM)



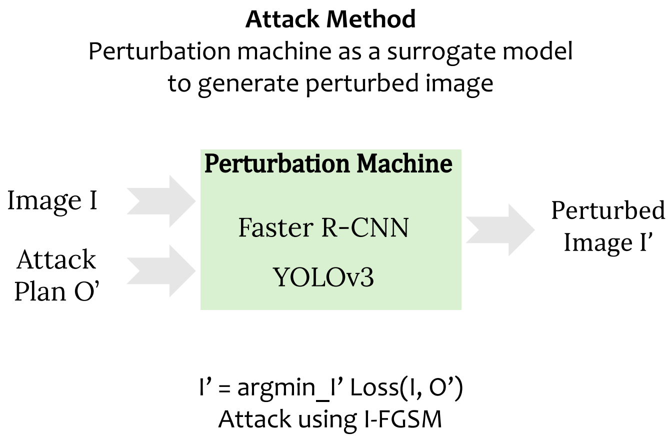


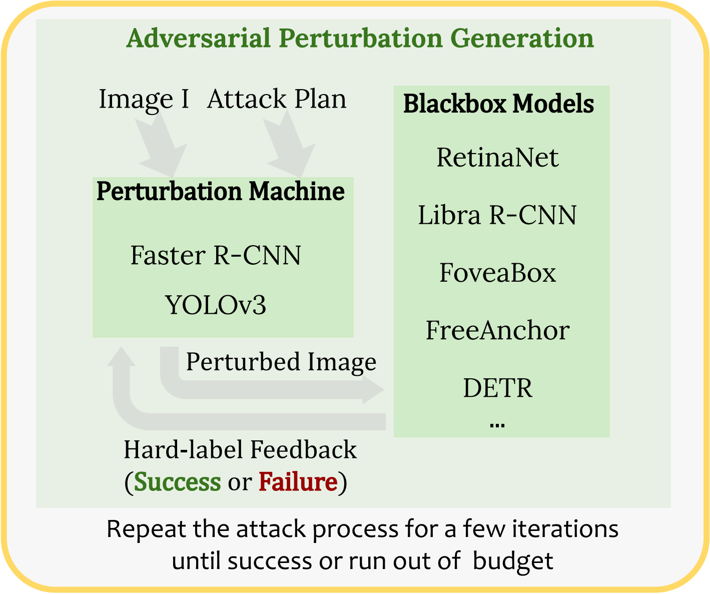
Repeat the attack process for a few iterations
until success or run out of budget

Context-aware transfer attacks

Adversarial perturbation generation

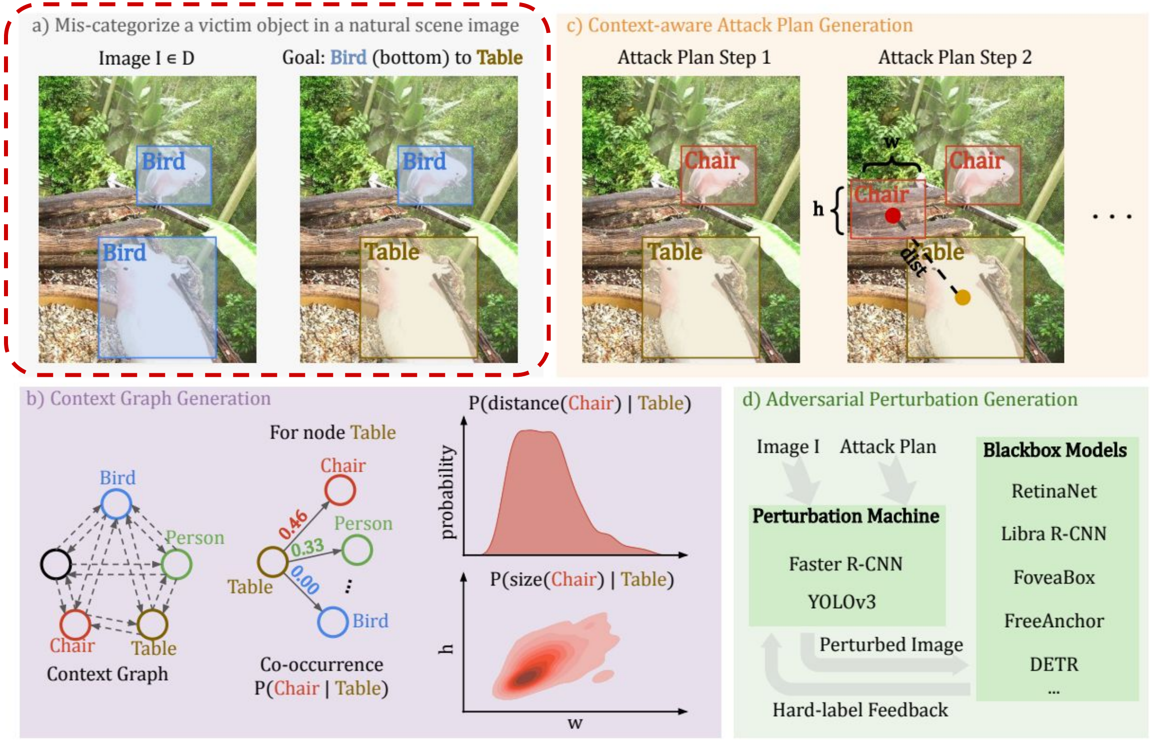
- A diverse set of object detectors (one-stage, two-stages, anchor-free, transformer-based)
  - one-stage, two-stages for perturbation machine, all four types for victim models
- Can work with different types of attack methods (variants of I-FGSM)





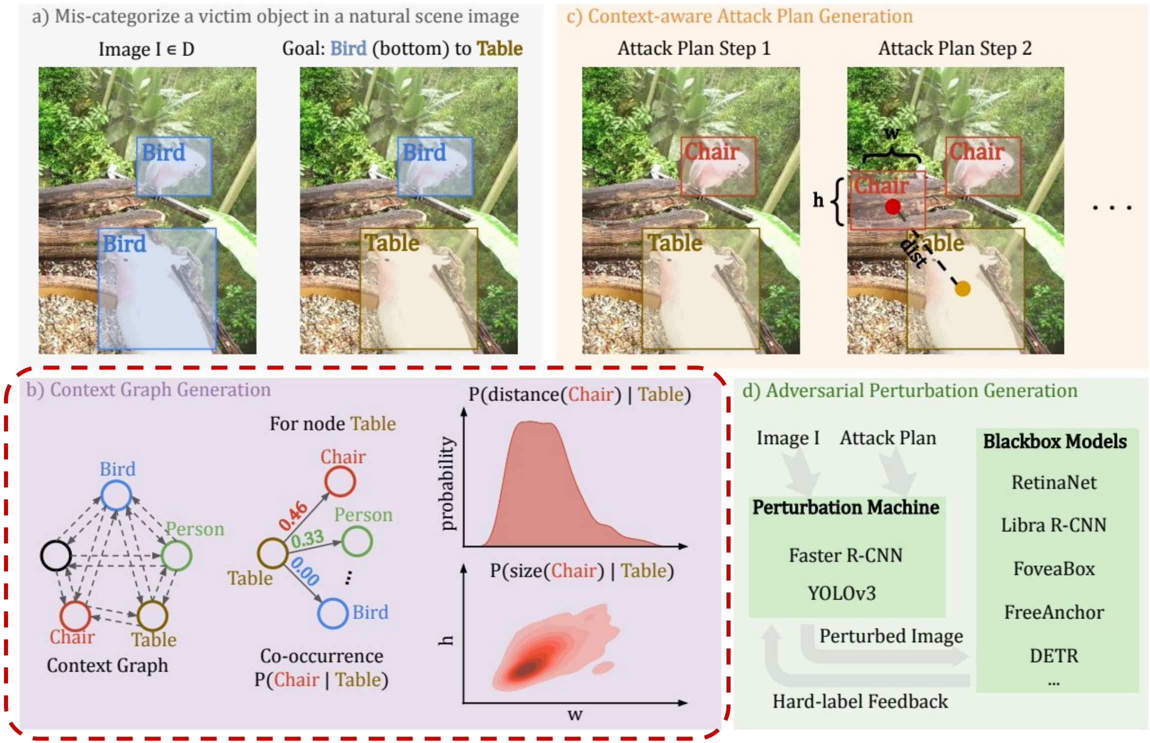
Context-aware transfer attacks

Overall framework



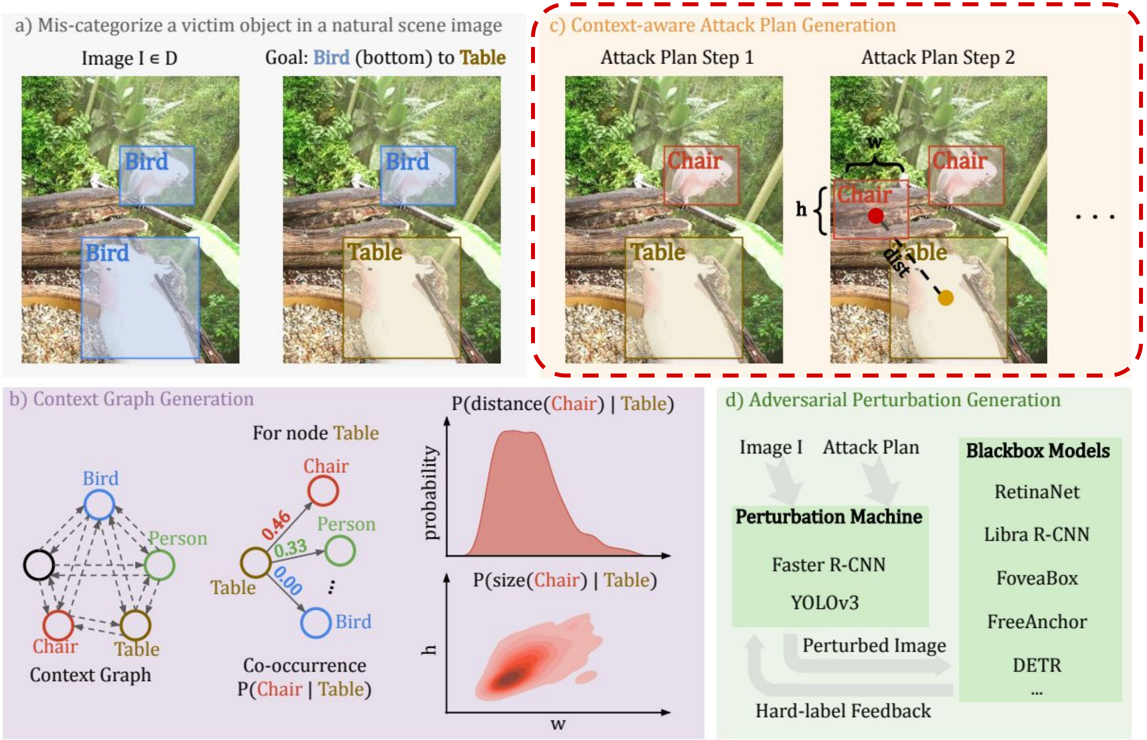
Context-aware transfer attacks

Overall framework



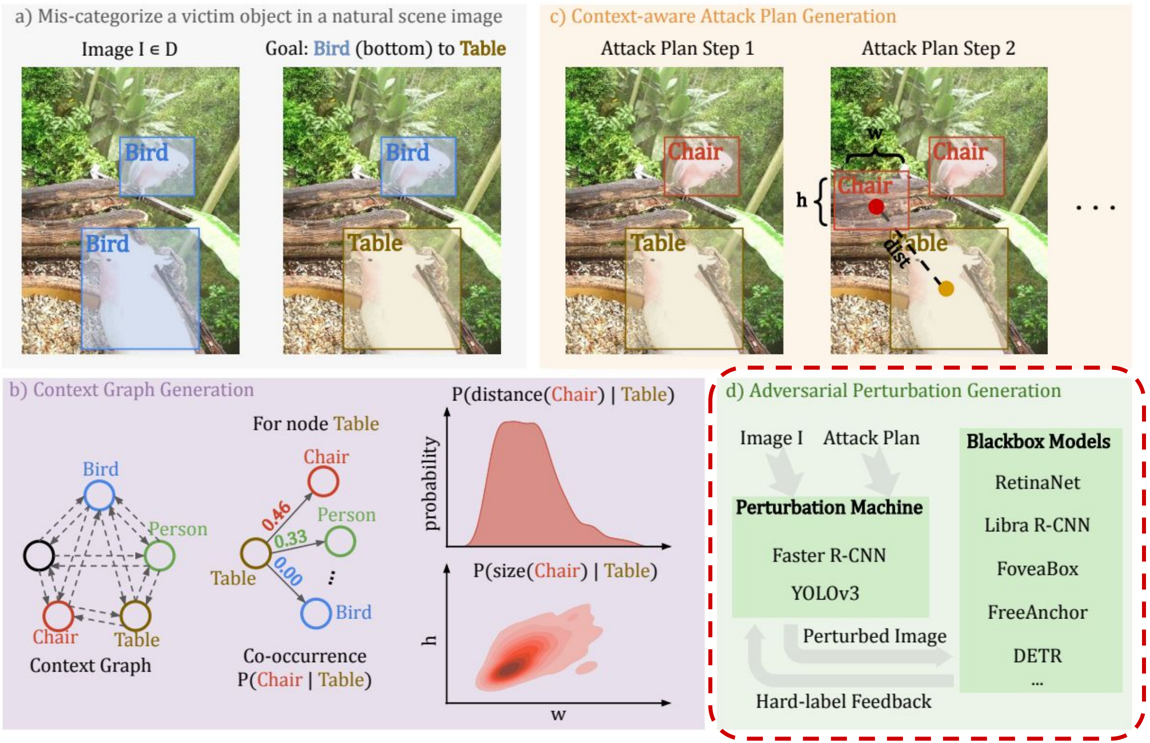
Context-aware transfer attacks

Overall framework



Context-aware transfer attacks

Overall framework



Experimental setup

- Attack type:
  - Mis-categorization attacks at different perturbation levels ($L_\infty \leq \{10,20,30\}$)
- Datasets:
  - PASCAL VOC and MS COCO
  - Evaluated using 500 images that contain multiple $(2-6)$ objects for each dataset
- Object detection models:
  - Surrogate model: Use 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

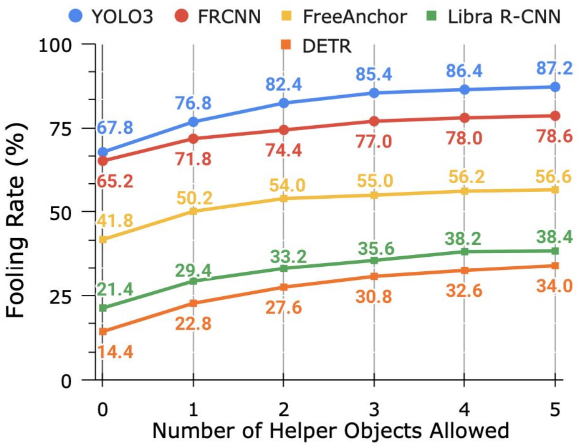
Experimental results

- Mis-categorization fooling rate at different perturbation levels
- Tested on different benchmark datasets and used a large variety of object detectors
- Our approach performs significantly better than context-agnostic and mismatched context approaches

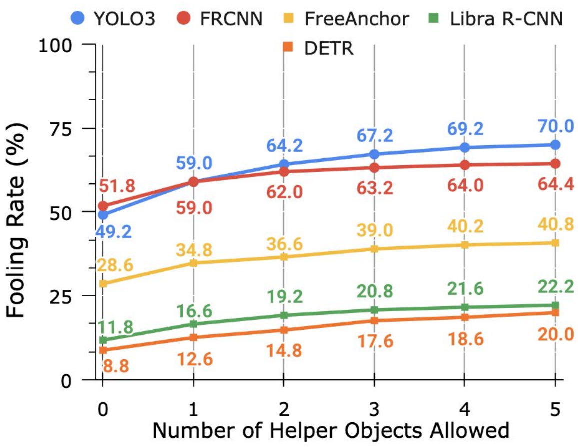
Perturbation Budget	Method	Whitebox FRCNN	Whitebox YOLOv3	Blackbox Retina	Blackbox Libra	Blackbox Fovea	Blackbox Free	Blackbox DETR	Blackbox D-DETR
Results on PASCAL VOC									
$L_\infty \leq 10$	Baseline	40.0	53.8	13.8	9.2	22.2	27.4	9.6	23.2
	Random	52.4	69.2	19.4	17.4	31.6	37.8	17.4	36.8
	Ours	**55.8**	**75.6**	**22.6**	**20.4**	**33.6**	**39.2**	**20.2**	**39.2**
$L_\infty \leq 20$	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**
$L_\infty \leq 30$	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**
Results on MS COCO									
$L_\infty \leq 10$	Baseline	29.0	32.2	7.4	4.8	11.6	16.6	3.4	19.0
	Random	40.2	48.4	11.2	8.0	14.6	20.0	6.2	23.6
	Ours	**41.2**	**54.4**	**12.0**	**11.2**	**18.6**	**25.0**	**10.8**	**27.8**
$L_\infty \leq 20$	Baseline	51.8	49.2	13.4	11.8	22.0	28.6	8.8	26.8
	Random	60.6	66.4	20.6	18.8	31.4	37.2	**20.2**	39.2
	Ours	**64.4**	**70.0**	**20.8**	**22.2**	**35.4**	**40.8**	20.0	**43.2**
$L_\infty \leq 30$	Baseline	57.6	54.4	18.2	15.4	25.6	34.8	8.0	28.8
	Random	65.8	73.6	23.8	21.8	34.8	**47.8**	18.4	42.0
	Ours	**68.6**	**75.4**	**27.2**	**27.2**	**39.2**	46.2	**21.2**	**48.6**

Observations on fooling rate w.r.t. # of helper objects

- Mis-categorization fooling rate at perturbation level $L_\infty \leq 20$
- Dot legends are white-box models in surrogate, square legends are black-box models
- Fooling rate increases with the number of helper objects and plateaus at around #5.



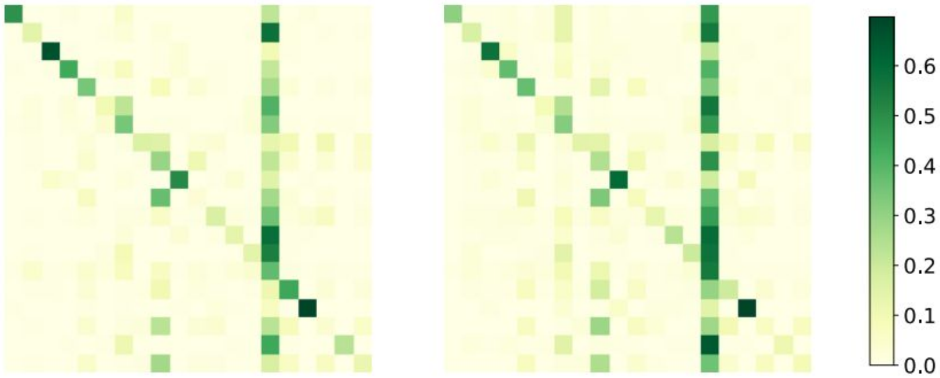
PASCAL VOC



MS COCO

Context graphs of different datasets

- Co-occurrence matrices for VOC and COCO for 20 object categories that are common in both datasets
- The average Pearson correlation of each corresponding row of VOC matrix and COCO matrix is 0.90
- Strong positive correlation between co-occurrence relationships encoded by different context graphs

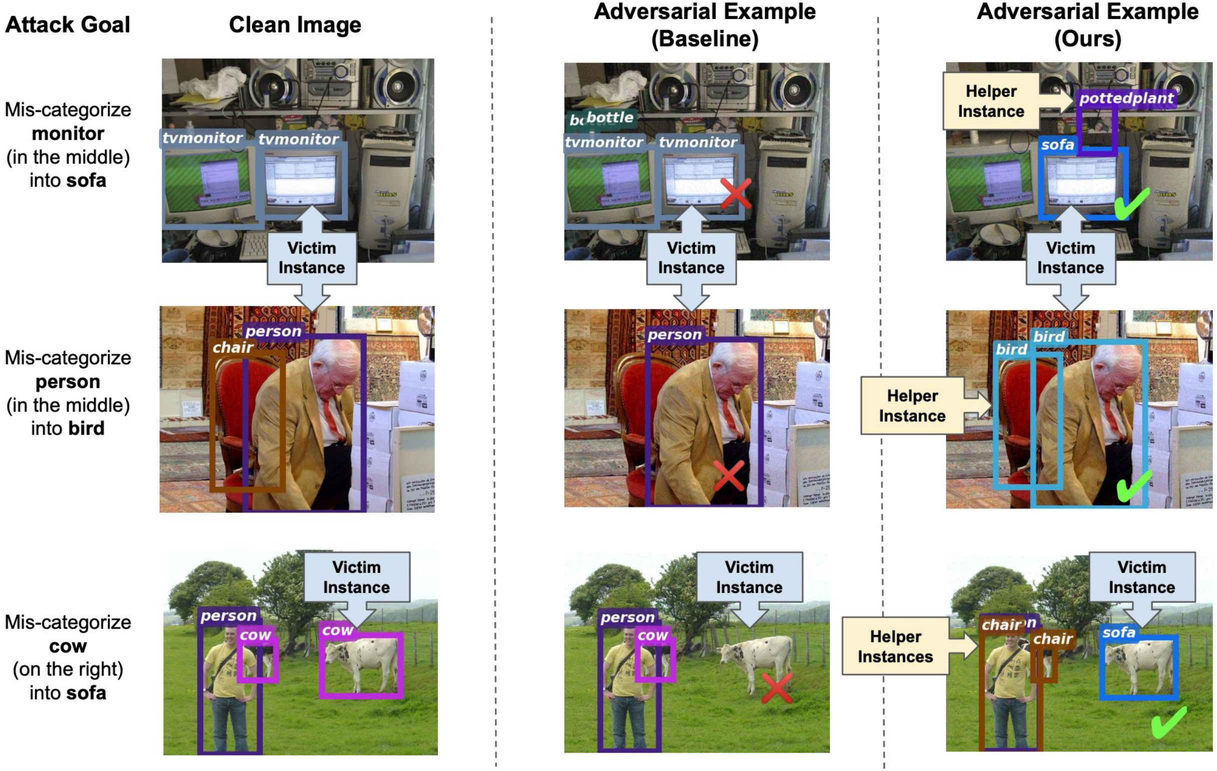


PASCAL VOC

MS COCO

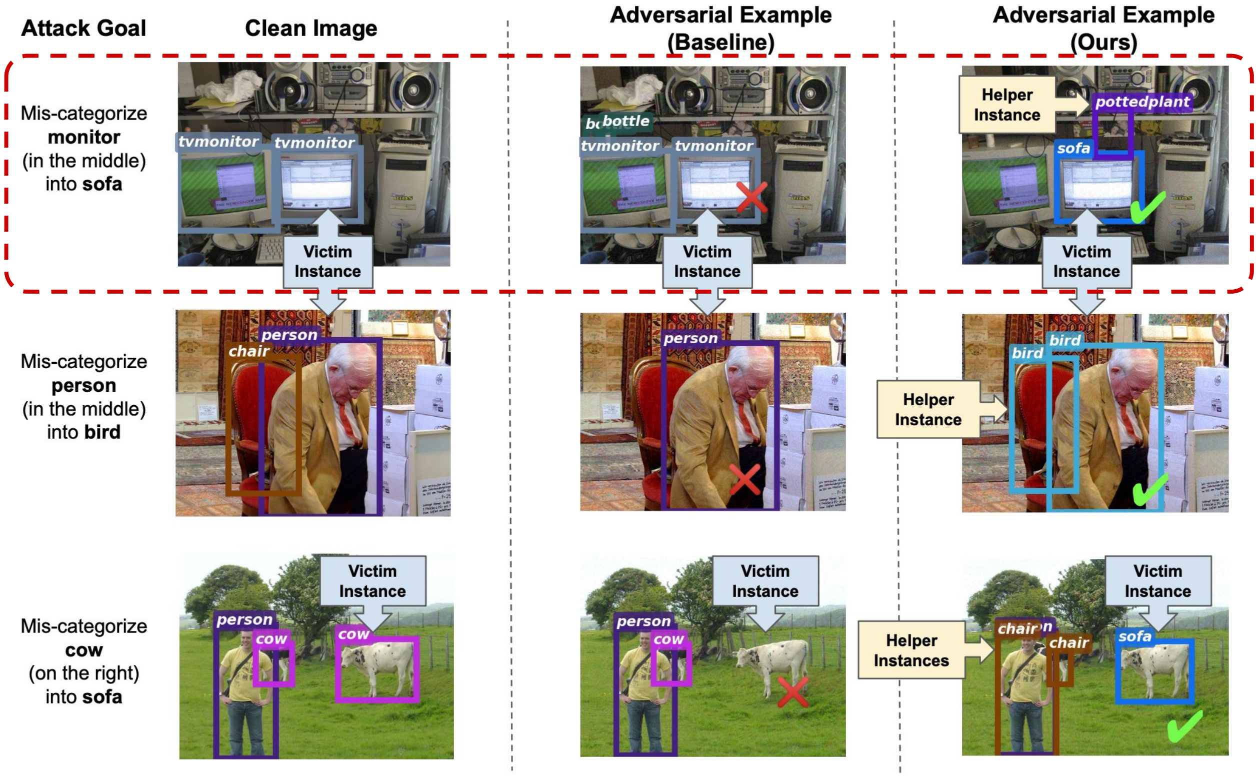
Visualization examples of attacks

- Examples where baseline attack fails but context-aware method succeeds by introducing helper objects in the attack



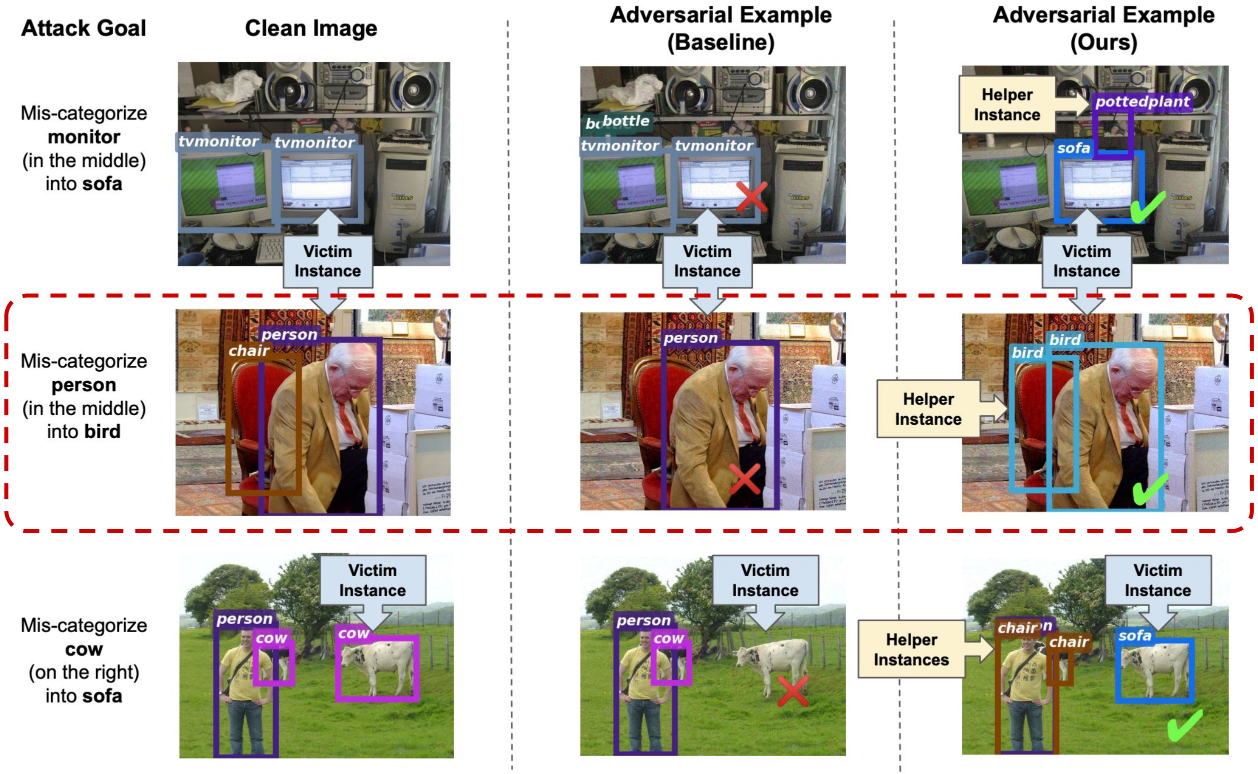
Visualization examples of attacks

- Examples where baseline attack fails but context-aware method succeeds by introducing helper objects in the attack



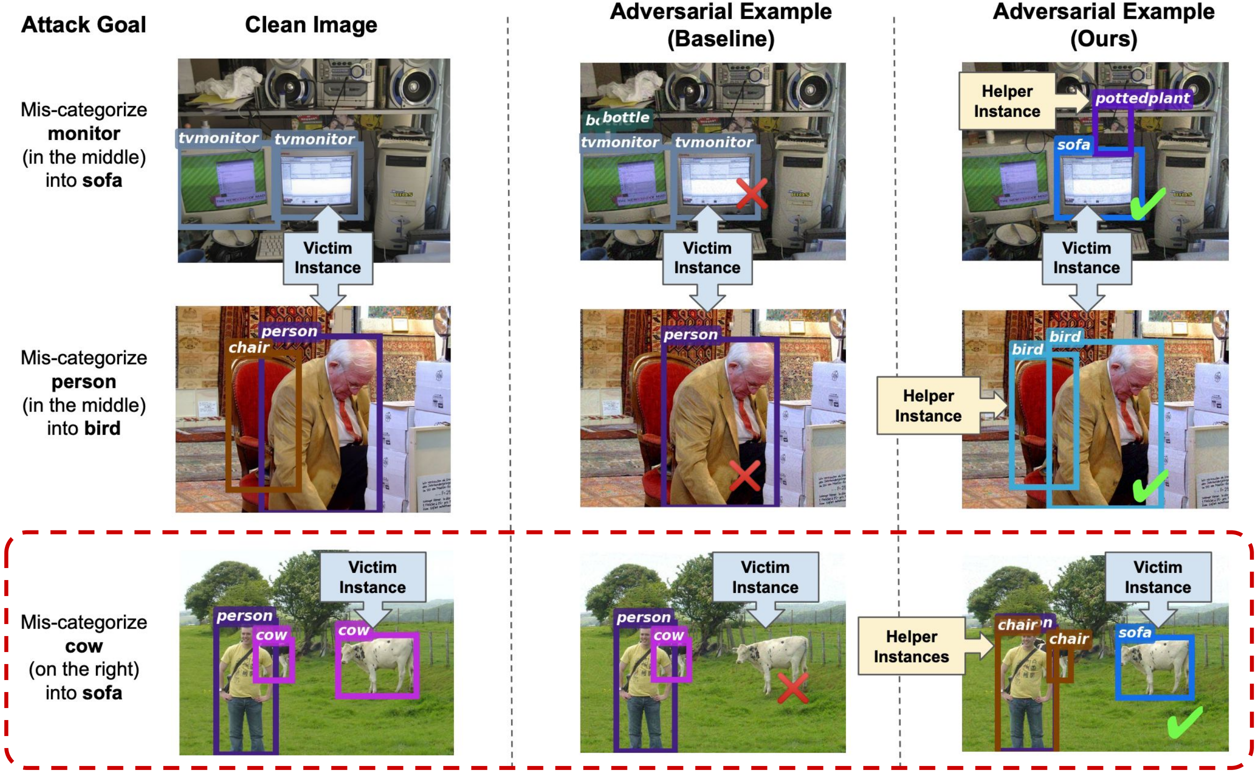
Visualization examples of attacks

- Examples where baseline attack fails but context-aware method succeeds by introducing helper objects in the attack



Visualization examples of attacks

- Examples where baseline attack fails but context-aware method succeeds by introducing helper objects in the attack



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.

Acknowledgements: DARPA under Agreement No. HR00112090096.

More information: Zikui Cai (zcai032@ucr.edu), M. Salman Asif (sasif@ucr.edu)

Code

https://github.com/CSIPlab/context-aware-attacks

Thank you!

Stay safe and healthy!