Supplementary Material

In this supplementary material, we provide the experimental results on COCO dataset in Section A. In Section B, we show that JPEG defense, one of the most common context-agnostic defense methods, fails against our proposed attack. We also include actual images showing region proposals with detected objects for the zero-query attack, the context-agnostic attack, and the few-query attack. The aim is to demonstrate the visual appearance of the attacked scenes, so that they can ascertain the subjective visual quality of the perturbed scenes, and see examples of cases in which the different attacks succeed or fail in fooling the victim system. Section C.

A. Experimental results on COCO dataset

In this section, we repeat the object detection evaluation experiments for the COCO dataset. The models obtained from `MMDetection` are well trained on COCO2017 training set, and the evaluation results on COCO2017 validation set can be found in Table 4. While the Mean Average Precision (mAP) scores are much lower than those observed for the VOC dataset (Table 1), these values are similar to the officially reported numbers in `MMDetection` repository. This confirms that the object detection algorithm for the COCO dataset – a more challenging dataset than VOC – performs at a level close to the state of the art. The comparison of ZQA and ZQA-PSPM acting on the COCO dataset against our two baseline schemes is shown in Table 5. As for the VOC dataset, the ZQA attack for the COCO dataset outperforms up to 3 attempts of the Few-Query attack (2 rounds of feedback) in the black-box transfer attack setting.

Table 4. Mean average precision (mAP) at IOU (intersection over union) threshold 0.5 of different detectors used in our experiments. Models are evaluated on COCO2017 val set. **Legend:** Faster R-CNN (FRCNN), RetinaNet (Retina), Libra R-CNN (Libra), FoveaBox (Fovea).

Model	FRCNN	Retina	Libra	Fovea
mAP@.50	38.99%	35.13%	40.14%	45.78%

B. Evading context-agnostic defense

We tested against the commonly used context-agnostic JPEG defense and found that our attack is resilient. Our attack can still outperform up to 5 rounds of few-query attacks under the JPEG compression quality of 95, as shown in Table 6, corresponding to the setting in Table 2.

C. Visualization of sample images

In this section, we provide visual examples of scenes before and after perturbation. In doing so, we compare the zero-query scheme, the context-agnostic attack, and the few-query scheme that we developed to benchmark performance. All the results are for a transfer setting, i.e., the attacker creates the perturbations on a surrogate model which is different from the classification model used by the victim system. All the images are generated for the case in which the attacker's perturbation is made using a Faster R-CNN network, while the victim system system uses a RetinaNet model. The perturbation budget used to implement the evasion attack is $\epsilon = 10$.

Figure 5 provides an example in which the context-agnostic attack successfully perturbs the individual objects: chair → dog, chair → bus and chair → bird. However, the resulting list of detected objects (dog, bus, bird) is context-inconsistent according to the co-occurrence matrix. Thus, the attack is detected. In contrast, the ZQA attack perturbs the objects as follows: chair → dog, second chair → second dog, dining table → person. The list of detected objects (dog, dog, person) is context-consistent, which fools the detector. This shows the basic use case of our context-aware approach.

Figure 6 provides an example in which the few-query attack has perturbed the main victim object (sofa → bicycle), as well as one other helper object (chair → bicycle) in the scene. However, the attack fails because the victim system's detector does not detect the main victim object and relegates it to the background. In contrast, the ZQA attack, with the help of the perturbation success probability matrix (PSPM), chooses object perturbations that are most likely to succeed in a single attempt, i.e., sofa → bicycle, chair → person, and leaves the TV monitor unchanged. The perturbation applied to the sofa object is sufficient for it to be detected and misclassified as a bicycle. This attack is context-consistent by construction, and successfully fools the detector. We remark here that the vanishing effect scene above is not unique to the few-query attack. Indeed, evasion attacks which involve perturbing the entire scene while attempting to attack individual objects in the scene are susceptible to the vanishing effect. This occurs when the scene perturbation, constrained by the budget ϵ, is such that it causes one or more objects in the scene to not be detected. As expected, we observe this effect more often at lower perturbation budgets.

Figure 7 shows that, given more rounds of feedback, the few-query detector eventually gets enough information about the detector's decisions, and is able to perturb a large number of objects, thereby fooling the detector. The attack attempts to make the following changes: dog → boat, sofa → boat, cat → boat, person → boat. The victim system misclassifies the dog and the sofa as boats, but does not detect the person and the cat. Even with the vanishing artifact, we deem the few-query attack successful because it has successfully perturbed the victim object (dog → boat)

Table 5. Follow the setting in Table 2 but use 500 images from COCO 2017 test set. Fooling rates (%) of different attack strategies under different L_∞ perturbation $\leqslant \epsilon \in \{50, 40, 30, 20, 10\}$ are as follows.

Method	$\epsilon = 50$				$\epsilon = 40$				$\epsilon = 30$				$\epsilon = 20$				$\epsilon = 10$			
	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3
Context-Agnostic	55.2	23.8	27.2	35.2	60.0	25.8	28.0	31.2	55.4	23.4	21.6	31.8	52.2	18.8	18.6	28.6	39.6	14.2	12.4	15.8
ZQA	82.2	29.8	35.4	43.6	82.8	30.2	35.8	43.2	81.0	**30.4**	31.0	40.0	76.0	23.8	26.4	**37.4**	52.0	14.2	15.6	21.8
ZQA-PSPM	**85.0**	**34.0**	**38.0**	**48.0**	**85.8**	**32.0**	**39.6**	**43.8**	**82.8**	29.8	**32.6**	**46.0**	**79.0**	**27.2**	**29.8**	36.8	**58.2**	**15.6**	**17.6**	**25.0**
Few-Query 0	71.4	27.0	24.0	37.8	73.4	25.6	24.0	35.0	68.4	21.8	18.2	34.0	63.8	21.2	17.6	28.4	47.2	13.2	9.6	18.4
Few-Query 1	80.0	34.2	34.2	46.6	80.0	33.6	34.2	44.0	79.0	29.8	27.2	44.4	72.6	26.0	24.6	36.8	56.4	19.2	15.4	26.0
Few-Query 2	83.4	37.8	41.4	51.4	84.0	39.4	39.0	50.4	84.2	34.2	33.2	49.8	79.2	31.4	30.2	43.6	62.4	21.8	19.0	30.6
Few-Query 3	86.2	40.6	46.2	55.2	86.8	41.8	42.2	54.6	86.2	36.6	39.2	53.6	81.6	33.2	34.8	46.6	66.8	23.6	21.2	34.0
Few-Query 4	88.0	42.8	48.0	57.8	89.8	42.8	45.4	56.6	87.8	37.8	42.0	55.4	84.4	35.2	38.0	49.2	69.0	24.2	23.0	36.2
Few-Query 5	89.2	45.0	52.4	59.2	92.0	44.4	48.0	57.8	89.8	39.6	45.0	57.6	85.8	36.0	40.6	51.6	71.4	25.6	25.4	38.2

Table 6. Follow the setting in Table 2 but under the JPEG compression quality of 95. Fooling rates (%) of different attack strategies under different L_∞ perturbation $\leqslant \epsilon \in \{50, 40, 30, 20, 10\}$ are as follows.

Method	$\epsilon = 50$				$\epsilon = 40$				$\epsilon = 30$				$\epsilon = 20$				$\epsilon = 10$			
	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3	WB	BB1	BB2	BB3
Context-Agnostic	34.6	26.4	30.0	25.6	33.4	23.6	25.0	26.0	34.4	26.4	28.8	27.0	38.4	24.0	23.6	25.8	28.2	9.4	11.0	14.8
ZQA	88.2	41.4	49.4	51.4	86.8	40.0	47.8	47.0	88.2	41.4	49.6	47.4	82.4	35.6	40.6	42.2	49.6	14.2	**16.8**	20.0
ZQA-PSPM	**89.2**	**42.8**	**50.2**	**53.8**	**90.2**	**41.2**	**48.6**	**49.8**	**92.8**	**44.2**	**52.2**	**51.2**	**83.6**	**36.4**	**42.0**	**44.2**	**55.8**	**15.6**	15.2	**21.4**
Few-Query 0	62.2	28.2	28.6	36.0	62.8	26.8	28.6	33.6	64.4	28.8	30.6	33.0	60.6	23.6	24.8	31.2	39.0	10.8	10.8	16.6
Few-Query 1	68.0	37.2	39.6	45.6	70.4	33.2	37.8	41.8	68.8	35.6	39.6	42.8	66.8	31.4	32.4	40.0	46.2	16.6	15.2	22.6
Few-Query 2	78.8	44.0	50.2	55.8	78.2	40.8	49.6	52.2	76.8	42.0	48.0	50.8	76.0	40.0	42.4	47.2	56.2	20.8	19.6	28.4
Few-Query 3	87.4	48.8	57.8	61.6	85.8	48.6	57.4	57.6	84.4	49.8	55.8	58.6	82.8	45.6	49.4	53.4	62.8	23.6	23.6	30.2
Few-Query 4	91.0	52.6	62.4	64.4	90.2	50.8	61.6	61.8	88.8	52.4	61.0	62.8	88.2	48.8	53.0	57.6	68.2	25.8	26.8	33.0
Few-Query 5	93.8	55.8	66.4	66.4	94.6	53.0	65.4	65.8	94.8	55.2	64.4	67.0	90.8	50.6	55.4	60.6	71.2	28.0	28.8	34.8

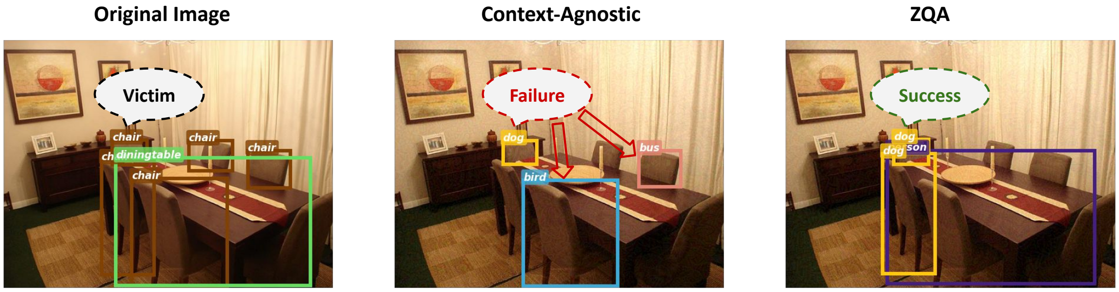

Figure 5. Detections on one original image and images perturbed by the context-agnostic attack and ZQA attack. The goal is to perturb the victim object, which is a chair on the top-left, to a dog. In the transfer attack, both the context-agnostic attack and ZQA attack successfully perturbs the chair to dog, along with some perturbations of surrounding objects. Even though context-agnostic attack is successful in perturbing victim to target, the attack still fails because the surrounding objects (bus and bird) are not context consistent according to the co-occurrence graph.

and it has ensured that the detected objects form a context-consistent list. On the other hand, the ZQA attack intends to leave the person unchanged, while changing the other objects to boats. This attack fails because, at the given perturbation level $\epsilon = 10$, the attack left the person unchanged, altered the sofa to the boat, but caused the cat and the dog vanish into the background. This is a failed attack because the main objective of misclassifying the victim object, i.e., dog $\rightarrow$ boat, was not fulfilled. This shows that the few-query approach – given multiple attempts to enhance the attack – will eventually overwhelm the proposed ZQA attack which is only allowed a single attempt. One disadvantage of the few query-attack, as noted earlier, is that it requires access to the victim system's communication, thus exposing the attacker to the risk of being discovered. The ZQA attack does not have this limitation.

Figure 8 shows one of the failure modes of our approach. (This type of failure is also observed in general perturbation bounded evasion attacks, and in our case, it is also seen in some cases of the few-query attack, and the context-agnostic attack). The goal of the attacker is to perturb the horse to a cat. However, the attack made with the surrogate model does not correctly transfer to the black-box victim model. The detector recognizes the horse as a sheep, which is unintended for our targeted attack.

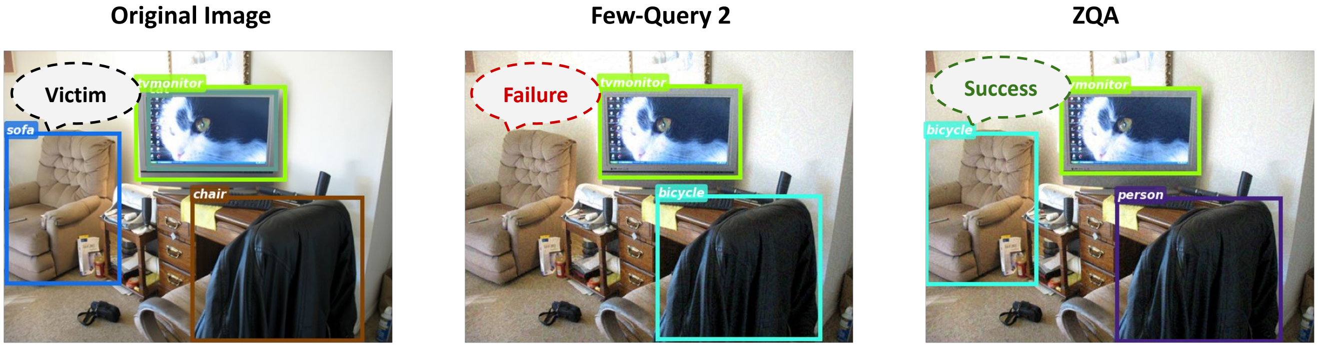

Figure 6. Detections on one original image and images perturbed by the few-query attack and the ZQA attack. The goal is to perturb the victim sofa to a target bicycle. Few-Query attack, building on 2 previous queries, perturbs the sofa to bicycle and the chair to bicycle as well. The TV monitor is not perturbed as it is context consistent. However, the attack failed to transfer to the victim model, in face, not detecting the sofa as a foreground object. Thus, the few-query attack fails. The ZQA attack additionally perturbs the chair to person. Since bicycle, person and TV monitor are all detected and are context-consistent, the attack successfully transfers.

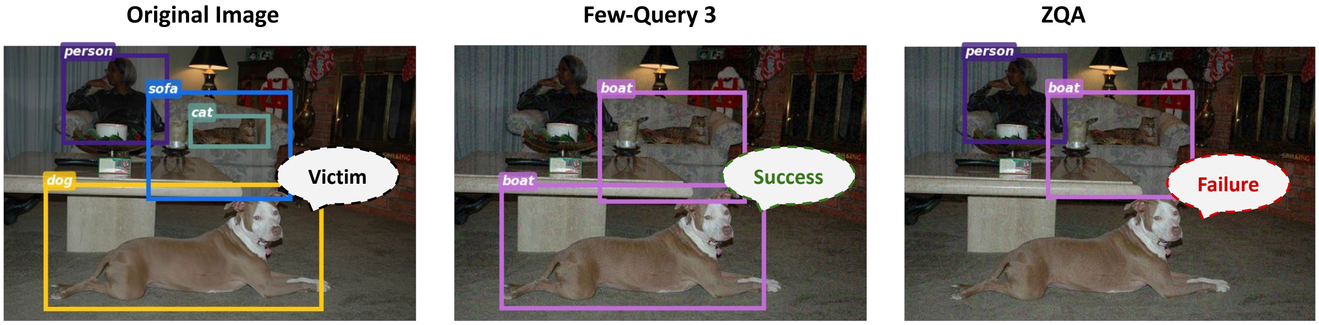

Figure 7. Detections on one original image and images perturbed by few-query attack and ZQA attack. The goal is to perturb the dog to a boat. The few-query attack, building on 3 previous queries, perturbs two objects to boats, and causes the person and the cat to vanish. The result is context-consistent and meets the desired goal. On the other hand, the ZQA attack leaves the person unchanged, perturbs the sofa to a boat, but causes the intended victim object (dog) and another object (cat) to vanish. Even though person and boat are context-consistent in the perturbed scene, the ZQA attack has failed because the intended victim object has vanished.

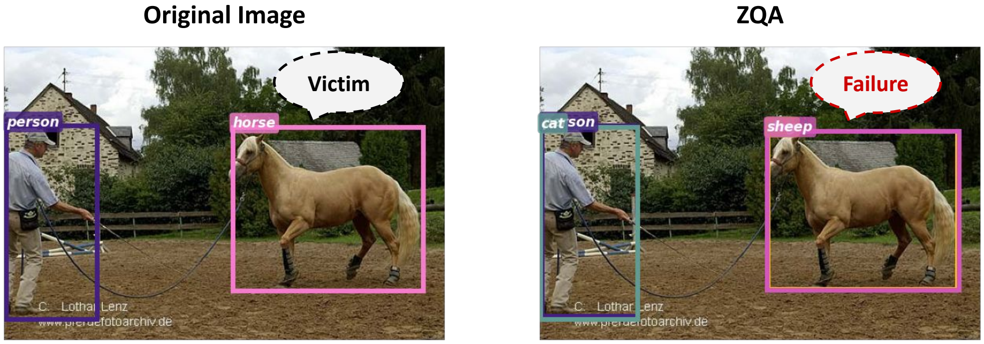

Figure 8. A failure case of ZQA attack. We observe that the perturbation of the victim object (horse → cat) does not succeed. Instead, the victim model classifies the perturbed horse as a sheep.