Zero-Query Transfer Attacks on Context-Aware Object Detectors

Zikui Cai[1], Shantanu Rane[2], Alejandro E. Brito[2], Chengyu Song[1],
Srikanth V. Krishnamurthy[1], Amit K. Roy-Chowdhury[1], M. Salman Asif[1] *
[1]University of California, Riverside, USA [2]Palo Alto Research Center, Palo Alto, USA
{zcai032}@ucr.edu, {srane, abrito}@parc.com
{csong, krish}@cs.ucr.edu, {amitrc, sasif}@ece.ucr.edu

Abstract

Adversarial attacks perturb images such that a deep neural network produces incorrect classification results. A promising approach to defend against adversarial attacks on natural multi-object scenes is to impose a context-consistency check, wherein, if the detected objects are not consistent with an appropriately defined context, then an attack is suspected. Stronger attacks are needed to fool such context-aware detectors. We present the first approach for generating context-consistent adversarial attacks that can evade the context-consistency check of black-box object detectors operating on complex, natural scenes. Unlike many black-box attacks that perform repeated attempts and open themselves to detection, we assume a "zero-query" setting, where the attacker has no knowledge of the classification decisions of the victim system. First, we derive multiple attack plans that assign incorrect labels to victim objects in a context-consistent manner. Then we design and use a novel data structure that we call the perturbation success probability matrix, which enables us to filter the attack plans and choose the one most likely to succeed. This final attack plan is implemented using a perturbation-bounded adversarial attack algorithm. We compare our zero-query attack against a few-query scheme that repeatedly checks if the victim system is fooled. We also compare against state-of-the-art context-agnostic attacks. Against a context-aware defense, the fooling rate of our zero-query approach is significantly higher than context-agnostic approaches and higher than that achievable with up to three rounds of the few-query scheme.

1. Introduction

Despite achieving significant performance gains on a variety of vision and language tasks, deep neural networks (DNNs) are vulnerable to adversarial attacks [49]. One of the most popular adversarial approaches is the class of perturbation-bounded evasion attacks [9, 19, 35, 41]. Here, an attacker can make a model yield arbitrarily wrong classification results by adding imperceptible perturbations to the input image. These attacks are quite practical and can be performed at test time without needing access to the training data. The vast majority of work in this area has focused on attacking classifiers trained on datasets like ImageNet, MNIST, CIFAR-10, and CIFAR-100, where the classifier attempts to recognize one dominant object in a given image. In contrast, we are primarily concerned with object detectors [24, 29, 39, 42, 44] that localize and recognize multiple objects in an image, which is the case in most natural images. Such detectors often take a holistic view of an image, rather than considering it as a collection of arbitrary objects [55, 57]. The objects in natural scene images form a context that can help identify the scene or given the scene we are likely to find the objects that conform to the scene

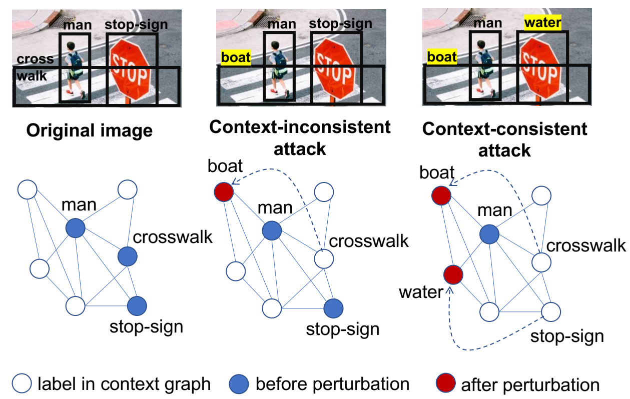

Figure 1. For natural scenes containing multiple objects, applying an evasion attack on an individual object (e.g., crosswalk → boat) violates the context: A boat and a stop-sign rarely occur together. A context-aware detector can detect this attack. In this work, we perturb multiple objects in a context-consistent way (e.g., crosswalk → boat, stop-sign → water) in a single attempt. The combination (person, boat, water) does not violate context and thus fools even a context-aware detector.

*Corresponding author

context. For instance, a boat is unlikely to co-occur with a stop sign, and much more likely to co-occur with water. Leveraging this observation, some recent attack [7] and defense [28, 62] mechanisms have been proposed that take image context into account. Context-aware defense methods in [28, 62] can detect attacks that are inconsistent with the scene context, as shown in Figure 1. To evade these defenses, changing one object in the image is insufficient. The context-aware attack method in [7] uses the knowledge of co-occurrence between different object classes in complex images to generate a sequence of transferable attacks for black-box object detectors; however, this method needs a few queries to test which attack plan is successful. *We present, for the first time, a zero-query attack algorithm that changes multiple objects simultaneously in a context-consistent manner thereby creating a holistic adversarial scene that can overcome context-aware defenses.*

In this work, we consider "zero-query" attacks (ZQA) that refer to a setting in which the attacker has no feedback channel to access the classification decisions of the victim system. This setting is extremely useful in practice because in many applications the victim system is inaccessible to the attacker; even if the victim system is accessible, the attacker's communications can be monitored, and thus draw suspicion. ZQA, on the other hand, is a truly stealthy attack. The attacker can only implement an attack plan once by perturbing multiple objects in a given scene and submitting the perturbed image to the victim system. Furthermore, we assume that the victim system is explicitly context-aware; that is, it will examine the list of detected objects and determine whether that list is "context-consistent" or not. If yes, then the detector will not suspect an attack. If not, it will suspect that the image has been perturbed by an attacker. Our ZQA approach is able to subvert more sophisticated multi-label object detectors that either implicitly or explicitly take context relationships across objects into account while performing their inference. In fact, accounting for context is what makes it possible to achieve high success rates in a single attempt.

Several approaches exist for scene context modeling [2, 17, 22, 38, 50]. In this paper, we restrict our attention to object co-occurrence, which is the most fundamental approach to modeling semantic context. The context model is represented by a co-occurrence graph (or equivalently the co-occurrence matrix) that is computed for a given set of images. We consider a list of objects as context-consistent only if the corresponding labels form a fully connected subgraph within the co-occurrence graph.

The main contributions of this paper are as follows.

- We develop an architecture for designing attacks on multi-object scenes that fool context-aware object detectors. Our detectors explicitly use object co-occurrence to model scene context.
- We propose an approach for zero-query context-aware attacks that generate adversarial scenes to fool a context-aware detector in a single step.
- We introduce the concept of a perturbation success probability matrix (PSPM) that models the probability of successfully perturbing a given target object to a given victim object in the white box setting. We use the PSPM to refine our attack plans, essentially choosing the one which is most likely to succeed. We show that the PSPM-guided attacks improve the fooling rate even in a black-box setting.
- We show experimentally that the fooling rate of ZQA is significantly higher than that achieved by a context-agnostic black-box attack. Furthermore, we compare our results against a possible "few-query" strategy [7] that repeatedly enhances the attack plan, while observing the detector output, until the detector is fooled. For the Pascal VOC dataset [18], the ZQA attacks provide fooling rates comparable to 5-query and 3-query attacks in the white-box and black-box settings, respectively.

2. Background and Preliminaries

2.1. Context in Object Detection

The role of context in improving visual recognition tasks has been well studied [37, 46, 61]. The state-of-the-art object detectors locate and detect several objects in the scene based on holistic information in the image [42,44]. Many of these methods explicitly utilize context information to improve the performance of object detectors [3–5, 14, 34, 52]. Co-occurrence-based contextual information derived from image pixels [28], object labels [7], and language models [62] have also been used to build a context-aware object detector. Different from the above contributions, to the best of our knowledge, this is the first paper that proposes a context-aware attack to fool a context-aware detector with zero queries.

2.2. Black-Box Attacks

In a black-box attack, the attacker has no access to the internal parameters of the model; thus, instead of generating the perturbed image by backpropagation, the attacker can only test a perturbed image on the victim model. In some cases, the attacker can observe the output but in many cases even that is not possible [27]. This renders query-based attacks inapplicable, which usually take an overwhelmingly large number (often hundreds or thousands) of queries [6, 11, 13, 20, 25, 53]. In this paper, we explore the most stringent case where no model queries are allowed. Such attacks will be extremely hard to detect, free from suspicion of repeated queries, and thus will be a more viable option for subversion. Several papers [16, 26, 33, 41] have examined the phenomenon of transfer attacks where the adversarial examples generated using a surrogate network can fool a black-box victim network. A large body of work exists on designing (perturbation-bounded) evasion attacks for images containing one predominant object and evaluating how well they transfer in a black-box setting [8, 10, 19, 40, 48]. In this work, our goal is to fool object detectors for general scenes. This is considered a harder problem because of the need to perturb multiple objects [55, 57]. This difficulty is exacerbated by the need to preserve contextual consistency during the attack [28]. A "few-query" strategy for context-aware attacks was proposed in [7] that repeatedly enhances the attack plan, while observing the detector output, until the detector is fooled. Our goal in this paper is to develop context-aware attacks with zero queries.

2.3. Attacks against Object Detectors

Attacking object detectors is harder than attacking classifiers, since the attack must obfuscate the category as well as the location of one or more objects [55,57]. Object detectors [42, 44] implicitly use contextual information – for instance, relationships between object pixels and background pixels – to increase the speed of inference. Researchers have exploited this fact by developing various kinds of adversarial patches [23, 32, 45] which do not overlap with the victim objects, but can still fool the detectors. Some other attacks [15, 54, 57, 63], perturbing the image globally, are also successful in fooling object detectors in a white-box setting. A recently proposed method has demonstrated the ability to transfer attack black-box object detectors by exploiting context information, but the victim models do not explicitly check for context-consistency and also the attack needs multiple queries [7].

2.4. Defense methods

Some representative defense mechanisms [43] for mitigating adversarial attacks include enhancing adversarial robustness of the model intrinsically through adversarial training [1, 35, 51] or strengthening model architectures [31, 58]; and destroy adversary externally through input transformation [21, 56] or denoising [36, 47, 60]. These defenses are context-agnostic. Some recent papers consider context-aware object detectors operating on natural multi-object scenes [28, 62]. Although these works use different notions of context from our work, they confirm that attacking a context-aware detector is more difficult.

3. Context-Aware Zero-Query Attacks

We describe ZQA for natural scenes. A high-level diagram of zero-query context-aware black-box attacks is shown in Figure 2. We first present a high-level description and later elaborate on the building blocks. The attacker determines a list of objects detected in the scene and uses the co-occurrence-based context model to derive several context-aware attack plans that perturb one or more target objects to their respective victim labels. Then, given the perturbation budget, the attacker refines the list using a pre-computed PSPM, which we will discuss in detail later. The result is an attack plan that consists of a list of (victim label, target label) pairs that are most likely to succeed in fooling the victim object detector. The attacker then uses an evasion attack algorithm to generate the perturbed scene according to the refined attack plan. The perturbed image is sent to a black-box classification / detection machine equipped with an explicit context-consistency detection mechanism. The attack is considered successful only if the victim object is successfully perturbed to the target label *and* the victim system's object detector does not find any context inconsistency in its list of detected objects. We now describe the building blocks of the attack in detail.

3.1. Context Model

The context model is used by the attacker and the victim system's object detector to determine whether a given list of objects is context-consistent or not. Let us first define what is considered as context-consistent and what is context-inconsistent. Intuitively, combinations of objects detected in natural images from the training data should be considered as context-consistent, since these objects appear together in such scenes. On the other hand, combinations of objects should be considered as context-inconsistent if some of the objects have never appeared together in the training data. Thus, object co-occurrence is a fundamental cue in determining context-consistency.

Co-occurrence matrix/graph. We build a matrix, called the co-occurrence matrix $\mathbf{G}$, to model the co-occurrence relationships between objects as follows. Given a training data set with N labels and a label set $\mathcal{N} = \{\ell_1, \ldots, \ell_N\}$, a co-occurrence matrix $\mathbf{G}$ is an $N \times N$ matrix whose entries $\mathbf{G}(i, j)$ represent the number of unique pairs of objects with labels $\ell_i, \ell_j \in \mathcal{N}$ appearing together in the images. This matrix is symmetric before normalization. After normalizing each entry in $\mathbf{G}$ with the sum of elements in its row, we obtain $\bar{\mathbf{G}}$, where $\bar{\mathbf{G}}(i, j)$ indicates the conditional probability $p_{j|i}$ which is the probability of observing label ℓ_j if label ℓ_i is observed. In general, $\bar{\mathbf{G}}$ is not symmetric. An example of $\bar{\mathbf{G}}$ for the Pascal VOC Dataset is illustrated in Figure 3. The co-occurrence matrix can also be interpreted as a context graph (which we will also denote by $\mathbf{G}$) with N nodes, where the weight of the edge between nodes i and j represents the number of times that those two labels appear together in the training data.

Context consistency. If two nodes in a context graph do not have an edge connecting them, then these two la-

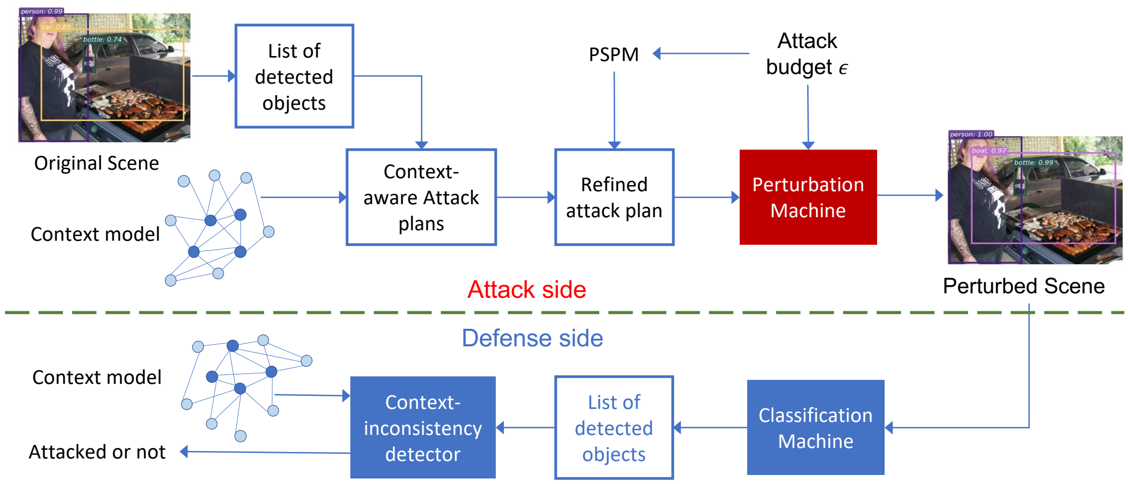

Figure 2. High-level diagram of zero-query context-aware black-box attacks. Given a victim image to be attacked, the attacker first finds a list of detected objects in the image and then consults the semantic context associated with the detector to design a context-aware attack plan that perturbs a victim object to a target label. To improve attack success rate, the attacker checks the PSPM corresponding to the perturbation machine with a certain perturbation budget ϵ and refines the original attack plan. With the refined attack plan, the attacker perturbs the image within bound ϵ, using the perturbation machine. The attacker's action is now complete. The perturbed image is then sent to a black-box classification / detection machine equipped with an explicit context-consistency detection mechanism. The attack is considered successful only if the victim object is successfully perturbed to the target object and the context-inconsistency detector does not find any inconsistency in the list of all detected objects.

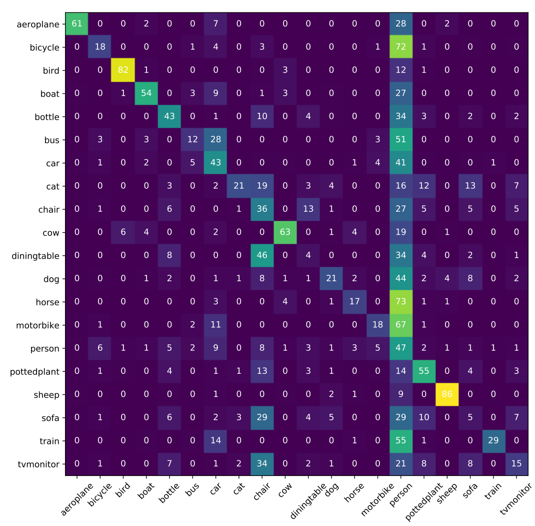

Figure 3. Co-occurrence matrix (conditional probability form) for the Pascal VOC07 training data set. Each cell indicates the probability as a quantized integer percentage (%).

bels never appear together in an image. Using this notion of co-occurrence-based context, we can define context-consistency and context-inconsistency as follows. A list of objects is considered as *context-consistent* if all the nodes representing the labels of those objects form a fully connected subgraph of the context graph $\mathbf{G}$. All the natural images in the data set satisfy this condition. A list of objects is considered as *context-inconsistent* if there are at least two nodes, representing the labels of two objects, that do not have an edge between them in the context graph.

Generalized context-consistency. Suppose we threshold the entries of $\mathbf{G}$ to obtain a matrix $\mathbf{H}_\eta$, where $\mathbf{H}_\eta(i,j) = \mathbf{G}(i,j)$ if $\mathbf{G}(i,j) > \eta$, for some threshold η. Otherwise $\mathbf{H}_\eta(i,j) = 0$. Build a context graph using $\mathbf{H}_\eta$, which we will also denote by the same symbol $\mathbf{H}_\eta$. Then the co-occurrence based notion of context consistency can be readily generalized as follows. A list of objects is considered as context-consistent up to a threshold η if all the nodes representing the labels of those objects form a fully connected subgraph of the context graph $\mathbf{H}_\eta$. Generalized context-inconsistency is defined similarly.

We describe the notion of context consistency for a list of object labels in Figure 1. Let us take a clean image containing three objects as an example. Suppose we have a person, a crosswalk and a stop-sign in the image as indicated by the dark blue nodes. These three nodes form a fully connected subgraph indicating that these three objects are context consistent. Suppose we want to perturb the crosswalk in the image to a boat. An example of a context-inconsistent attack involves just perturbing the victim object (crosswalk) to the target label (boat denoted as a red node). The perturbed object list (person, boat, stop-sign) is no longer context-consistent because these three nodes of boat, person and stop-sign do not form a fully connected graph, indicating that the combination never appears in the training data. Intuitively, it would be unlikely to see a boat with a stop-sign in the natural images. An example of a context-consistent attack is to perturb the crosswalk to a boat but also perturb the stop-sign to water. That combination (person, water, boat) does appear in the training data, as seen from the fact that the label nodes form a fully connected sub-graph.

3.2. Perturbation Success Probability Matrix

Perturbation Success Probability Matrix (PSPM) is an $N \times N$ matrix denoted as $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}$ that is defined for an ensemble of classification models $\mathcal{C}$, perturbation budget ϵ, and an object perturbation algorithm α. PSPM is defined for a specific training data set with labels $\mathcal{N} = \{\ell_1, \ldots, \ell_N\}$. $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}(i, j)$ encodes the probability that an object with label ℓ_i can be successfully perturbed to label ℓ_j by the perturbation algorithm α using the perturbation budget ϵ.

The PSPM expresses an attacker's ability to perturb *individual* objects in a scene. The utility of the PSPM can be explained as follows. Suppose the list of objects in the given scene is $\mathcal{A} = \{A_1, \ldots, A_S\}$. Suppose the list $\mathcal{A}$ is perturbed to a list $\mathcal{B}$ using an evasion attack algorithm α with perturbation budget ϵ. A context-agnostic attack would choose the labels in $\mathcal{B}$ at random from the label set $\mathcal{N}$. A context-aware attack would make sure that the labels in $\mathcal{B}$ are context-consistent as determined by the co-occurrence matrix $\bar{\mathbf{G}}$. Then, there are in general one or more possible perturbation assignments $\mathcal{A} \rightarrow \mathcal{B}$ that are context-consistent. Depending upon the training set (e.g., the presence of objects in different poses, sizes, illuminations) some object perturbations, $A_i \rightarrow B_j$, are likely to be more successful than others, even in a white-box setting. Thus, each perturbation assignment $\mathcal{A} \rightarrow \mathcal{B}$ suggested by an examination of the co-occurrence matrix has a different likelihood of success. The PSPM enables us to select the assignment that is most likely to succeed. An example of a PSPM for the Pascal VOC07 dataset is shown in Figure 4.

Some ways to choose an assignment amongst many assignments permitted by the co-occurrence matrix include

1. Choose the assignment $\mathcal{A} \rightarrow \mathcal{B}$ that maximizes each $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}(i, j)$ with $i \in \mathcal{A}, j \in \mathcal{B}$.
2. Choose the assignment $\mathcal{A} \rightarrow \mathcal{B}$ that maximizes the average of all $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}(i, j)$ with $i \in \mathcal{A}, j \in \mathcal{B}$.
3. Choose the assignment $\mathcal{A} \rightarrow \mathcal{B}$ that maximizes the minimum of all $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}(i, j)$ with $i \in \mathcal{A}, j \in \mathcal{B}$.

For simplicity, we use the first approach in the ensuing development. Admittedly, the PSPM considers the success of perturbing some $A_i \in \mathcal{A}$ to $B_j \in \mathcal{B}$ in a white-box setting; that is, for one or more classification models specified in the given ensemble $\mathcal{C}$. In a black-box setting, the attacker does not know which model is being used at the victim side. Thus, attack plans based on the PSPM are, at best, approximations. We hypothesize that these approximations are still better than choosing, at random, one of many possible assignments $\mathcal{A} \rightarrow \mathcal{B}$ suggested by the co-occurrence matrix.

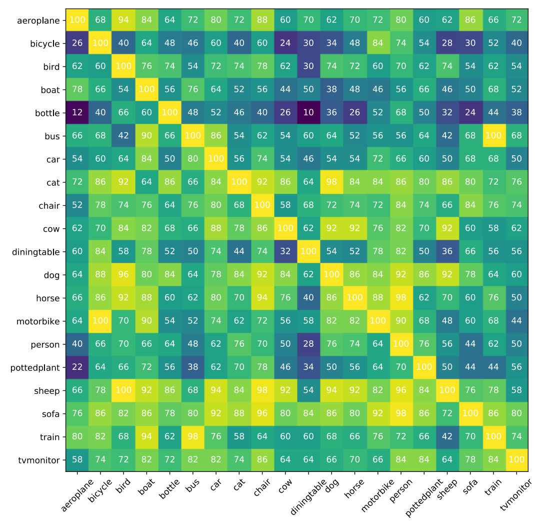

Figure 4. PSPM of Pascal VOC07 data set for $\mathcal{C} = \{$"Faster R-CNN"$\}$, $\epsilon = 10$, $\alpha = \{$"PGD"$\}$. Each cell indicates the probability as a quantized integer percentage (%).

Our experiments, described in Section 4, indicate that refining the attack plan using the PSPM indeed improves the fooling rate compared to choosing an attack plan at random from the context-consistent candidate attacks suggested by the co-occurrence matrix.

3.3. Context-Consistent Attack Plan Generation

We now describe the zero-query method for deriving context-consistent attack plans using the co-occurrence matrix $\bar{\mathbf{G}}$ and the PSPM matrix $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}$. The overall procedure is described in Algorithm 1. We assume that there is a desired target assignment for one object in the scene (e.g., change one of the horses to a bicycle). As discussed, just perturbing one object can result in a context-inconsistent list of objects in the perturbed scene. Thus, other objects in the scene may need to be perturbed so that the resulting list is context-consistent. The procedure described below ensures not only the attack plan is context-consistent, but it is also the plan that is most likely to succeed in a white-box setting. As we mentioned earlier, for a black-box setting this plan may not always be the most likely to succeed. Experimentally, we found that the attack plan guided by PSPM also increases the fooling rate for black-box models.

3.4. Implementation of Attack Plan

To generate the adversarial scene, evasion attacks can be implemented using a single or multiple surrogate model(s). Our attack generation method with a single surrogate detector is based on projected gradient descent (PGD) [35] within a ℓ_∞ ball, which can be considered as a powerful multi-step variant of FGSM [19]. We initialize a zero per-

Algorithm 1 Zero-Query Context-Consistent Attack Plan

Inputs: A list of objects in the scene $\mathcal{A} = \{A_1 = \ell_v, A_2, \ldots, A_S\}$ where S denotes the total number of objects in the scene, the desired target assignment consisting of (victim label $\ell_v \in \mathcal{A}$, target label ℓ_t) pair for one object, co-occurrence matrix $\mathbf{G}$ computed over the training data, co-occurrence threshold η, perturbation budget ϵ, perturbation success probability matrix $\mathbf{M}_{\mathcal{C},\epsilon,\alpha}$ generated in advance from the training data.

Output: Attack plan $\mathcal{B}$ that is context-consistent with $\mathbf{G}$ and most likely to succeed in a white-box setting.

Initialization: $\mathcal{B} = \{\ell_t\}$

Procedure:

1. Obtain a set of labels that co-occur with ℓ_t, denoted as $\mathcal{T} \subset \mathcal{N}$ such that $\mathbf{G}(j,t) > \eta$ for all $\ell_j \in \mathcal{T}$.
2. Set running counter $k = 2$.
3. While $k \leqslant S$, compute

$$\ell^* = \operatorname*{argmax}_{\ell_b \in \mathcal{T}} \mathbf{M}_{\mathcal{C},\epsilon,\alpha}(a,b), \tag{1}$$

where a is the index of label A_k

$$\mathcal{B} \leftarrow \mathcal{B} \bigcup \ell^*$$

$$k \leftarrow k+1$$

turbation $\delta^0 = 0$, and update it in each iteration as

$$\delta^{t+1} = \Pi_{\mathcal{S}} \left(\delta^t - \lambda \operatorname{sgn}(\nabla_\delta \mathcal{L}(x+\delta^t, y)) \right), \tag{2}$$

where $\mathcal{L}$ is the loss function, $x \in \mathbb{R}^d$ is the input image and y is the target label. We generate the desired output y based on our attack plan, which includes object categories, locations, and confidence scores. We take a step size λ at each iteration t, and project (clip) $\Pi_{\mathcal{S}}$ the perturbation δ^t to the feasible set $\mathcal{S}$ which satisfies the following two criteria

$$\begin{cases} \|\delta^t\|_\infty \leqslant \epsilon, \\ x + \delta^t \in [0, 255]^d. \end{cases} \tag{3}$$

We use PGD for its simplicity in our experiments, but we can easily modify our method to use other (more advanced) perturbation methods such as MIM [16] and DIM [59] etc. without losing generalizability.

4. Experiments

We performed extensive experiments on two large-scale object detection datasets to evaluate the proposed context-aware zero-query attack strategy. We construct a query-based baseline scheme and evaluate the fooling rate of our proposed zero-query approach against it.

Datasets: We conduct our experiments using images from both PASCAL VOC [18] and MS COCO [30] datasets. VOC contains 20 common classes of objects, and COCO contains 80 classes which is a super-set of the categories in VOC. We randomly selected 500 images from `voc2007test` and `coco2017val` respectively, which contain multiple $(2-6)$ objects. This manuscript contains results for various models on the `voc2007test`. The results for the `coco2017val` are in the supplementary.

Attack models: To mimic a realistic black-box setting, we pick a variety of object detectors, including two-stage detectors: Faster RCNN [44] and Libra R-CNN [39]; one-stage detector: RetinaNet [29]; and anchor-free detector: FoveaBox [24]. We use implementations of the aforementioned models from the `MMDetection` [12] code repository. The models in `MMDetection` are trained on `coco2017train`; therefore, while testing the detectors on VOC images, we only return the object labels available in VOC. The models under such adaptation still get good detection performance, as shown in Table 1.

Table 1. Mean average precision (mAP) at IOU (intersection over union) threshold 0.5 of different detectors used in our experiments. Models are evaluated on VOC07 test set. **Legend:** Faster R-CNN (FRCNN), RetinaNet (Retina), Libra R-CNN (Libra), FoveaBox (Fovea).

Model	FRCNN	Retina	Libra	Fovea
mAP@.50	78.30%	78.51%	79.01%	77.68%

Zero-Query attacks (ZQA and ZQA-PSPM): We evaluate two variants of our zero-query attack. The first variant (ZQA) ensures that the attack plan is chosen at random from the available set of context-consistent attacks and is determined using Equation (4). The second variant (ZQA-PSPM) generates a context-consistent attack plan based on the PSPM matrix that was pre-computed for the given classification model, perturbation budget and evasion attack algorithm. See Equation (1). ZQA-PSPM is the key contribution of this paper, and our results demonstrate that that ZQA-PSPM provides better fooling success rate compared to ZQA and baselines.

Baselines and comparisons: We compare the ZQA and ZQA-PSPM schemes against two relevant baselines. The first is the context-agnostic zero-query attack, which we call "Context-Agnostic". In this attack, the attack plan that drives the scene perturbation is chosen randomly (i.e., without explicitly enforcing co-occurrence-based context). This means that some attack plans in the Context-Agnostic scheme may be context-consistent by accident, while others are context-inconsistent. Comparison against Context-

Agnostic is performed with the aim of investigating the benefits of exploiting context while designing the attack plan.

We also compare with a second, more powerful baseline, which we refer to as the "Few-Query" approach [7]. In this scheme, the attacker is equipped with the co-occurrence matrix $\mathbf{G}$ but doesn't have the PSPM matrix. More importantly, the few-query attacker can query the victim system to find out whether the attack succeeded. Because of this, they don't need to perturb all the objects in the scene in one step. The few-query attacker proceeds as described in Algorithm 2. The few-query attack is denoted as "Few-Query q" in Table 2 and Table 3 where q is the number of previous queries that the current attack is built on, $q \in \{0, 1, 2, 3, 4, 5\}$, thus "Few-Query 0" is identical to ZQA in terms of queries.

Algorithm 2 Few-Query Context-Consistent Attack Plans

Inputs: Same as the inputs to Algorithm 1 but with no PSPM matrix, number of victim system queries allowed $q \leqslant S$ where S is the number of objects in the scene.

Output: A sequence of attack plans $\mathcal{D}_k$ consistent with $\mathbf{G}$, where $k = 1, 2, ..., q$.

Initialization: $\mathcal{D}_1 = \{\ell_t\}$

Procedure:

1. Obtain the label set $\mathcal{T} \subset \mathcal{N}$ such that $\mathbf{G}(j, t) > \eta$ for all $\ell_j \in \mathcal{T}$.
2. Set running counter $k = 2$.
3. While $k \leqslant q$, compute

$$\begin{aligned} &\widehat{\ell} \overset{iid}{\sim} U(\mathcal{T}) \quad \triangleright \text{ uniformly sample from } \mathcal{T} \qquad (4)\\ &\mathcal{D}_k \leftarrow \mathcal{D}_{k-1} \bigcup \widehat{\ell} \\ &k \leftarrow k + 1 \end{aligned}$$

The attacker implements $\mathcal{D}_1$ and queries the victim system. If the attack succeeded, they stop; else, they implement $\mathcal{D}_2$ and query the victim system, and so on, until the attack succeeds.

Attack generation: We use the PGD-based method to generate a perturbation on the whole image (as discussed in Equation (2)). We experiment with L_∞ perturbation budget $\epsilon \in \{10, 20, 30, 40, 50\}$. The step size $\lambda = 2$, and maximum number of iterations is 50. We observe that when an object is very close to or overlaps with the victim object, perturbing that object to any label different from the victim's target label reduces the success rate; thus, we map all objects whose regions have IOU $\geqslant 0.3$ with the victim object to the victim's target label.

Evaluation metrics: We use the metric "context-consistent attack success rate" (or fooling rate) to evaluate the attack performance on a victim object detector. For an attack to be regarded as a successful context-consistent attack, it must (1) successfully perturb the victim object to the target label, and (2) pass the context-consistency check described in Figure 1. We define the fooling rate as the percentage of the number of test cases for which the above two conditions are satisfied.

Comparing zero-query attacks with few-query attacks: The fooling rates of the few-query attacks with different numbers of queries (Few-Query 0 to Few-Query 5) and the zero-query attack under white-box and black-box settings at different perturbation budgets are shown in Table 2 and Table 3. The settings for both tables are detailed in the captions. The fooling rates for the few-query attacks are cumulative; that is, the values reported for few-query-k accounts for successful attacks with $0, 1, \ldots, k$ queries.

We observe that ZQA can achieve higher fooling rates than the few-query attack for up to 4 queries in the white-box setting and up to 2 queries in the black-box setting. When PSPM is used to refine the zero-query attack plan (ZQA-PSPM), the fooling rate increases, outperforming up to 5 queries of the few-query attack in the white-box setting and up to 3 queries in the black-box setting. These results are consistent across several detector models tested. As ϵ reduces, the perturbation is not always enough to carry out the evasion attacks, and thus the fooling rates fall from $\epsilon = 50$ to $\epsilon = 10$.

The results clearly demonstrate the advantage of simultaneous context-aware perturbation of all objects in the scene. In many cases, using the PSPM to refine the context-aware attack further improves the fooling rate. While the few-query approach eventually outperforms the ZQA attack, recall that the former requires the attacker to communicate with the detector, which is either not always possible, or might expose the attacker.

5. Discussion

Limitations and Future Work: We have assumed that the data distribution at the victim system is known. We also assume that the context is consistent across surrogate and victim systems. In practice, this is rarely the case. The attacker and the victim system may have different data distributions, overlapping but non-identical label sets, and thus similar but non-identical context models. A useful avenue for future work is to introduce controlled discrepancies between the distributions, context models, and label sets and examine their effect on the fooling rate of the ZQA attack.

While co-occurrence is a fundamental notion of context, it does not capture key properties such as relative size and

Table 2. Fooling rates (%) of different attack strategies under different L_∞ perturbation $\leqslant \epsilon \in \{50, 40, 30, 20, 10\}$. We compare ZQA and ZQA-PSPM with Context-Agnostic ZQA, and Few-Query attacks where feedback from blackbox (BB) models is allowed. The white-box (WB) is **Faster R-CNN** and three black-box models (BB1, BB2, BB3) are **RetinaNet, Libra R-CNN and FoveaBox** respectively. Fooling rate is counted as the percentage of attacks where victim is perturbed to a target label and all detected labels satisfy context consistency. Tested on 500 images from VOC 2007 test set which contain multiple (2-6) objects. Shaded cell indicates up to which few-query step, ZQA or ZQA-PSPM has better performance than few-query attack. Lighter shades are for ZQA, and darker shades are for ZQA-PSPM.

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	34.0	29.0	30.0	25.4	36.8	26.2	30.0	29.6	35.4	27.4	31.2	27.8	35.2	24.4	30.8	27.6	30.4	13.8	15.6	17.8
ZQA	90.0	46.6	52.2	54.0	92.0	48.0	55.0	51.8	91.6	46.0	57.0	52.2	87.4	39.6	50.4	51.0	65.2	21.0	23.8	24.2
ZQA-PSPM	**92.6**	**51.2**	**61.6**	**56.8**	**92.0**	**51.8**	**55.4**	**54.4**	**93.0**	**49.2**	**57.2**	**54.0**	**88.2**	**44.0**	**51.4**	**53.4**	**70.6**	**23.2**	**27.4**	**28.2**
Few-Query 0	60.0	29.8	29.8	34.8	64.2	34.6	34.2	39.8	66.2	34.2	35.0	37.8	61.2	29.2	30.8	35.8	48.2	14.8	14.0	20.2
Few-Query 1	64.4	35.8	40.4	43.0	68.0	41.0	44.2	49.4	69.6	41.8	45.2	47.6	68.2	36.0	40.2	43.4	58.4	21.6	23.8	27.8
Few-Query 2	77.6	48.0	56.2	59.0	80.0	50.0	52.6	57.2	78.4	47.0	54.2	54.6	77.6	43.8	50.0	49.8	70.4	27.0	29.6	34.2
Few-Query 3	86.8	55.4	65.0	65.8	89.6	55.4	58.6	62.0	86.2	54.8	62.0	61.2	86.4	49.6	57.2	55.4	77.0	31.4	34.0	39.8
Few-Query 4	91.6	60.0	71.8	69.4	95.2	61.4	63.8	66.0	91.2	58.0	68.2	67.2	89.6	53.2	60.6	59.8	81.4	34.2	37.8	43.2
Few-Query 5	95.0	61.8	75.0	73.4	97.2	62.8	68.0	69.4	96.2	61.2	71.0	70.6	93.0	56.6	65.2	63.4	85.2	35.8	40.2	46.2

Table 3. Follow the setting in Table 2 but use **Libra R-CNN** as WB and use **Faster R-CNN, RetinaNet and FoveaBox** as BB1, BB2, BB3 respectively. 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	30.6	26.2	18.2	21.4	34.6	29.4	24.2	24.4	34.8	28.8	20.4	23.4	37.6	24.8	16.8	19.4	29.2	15.8	10.2	13.0
ZQA	90.4	47.8	33.0	43.4	91.8	48.4	32.2	40.0	88.6	48.6	33.0	41.8	**86.6**	39.4	25.6	35.0	64.4	23.8	13.4	21.6
ZQA-PSPM	**92.2**	**51.0**	**34.2**	**45.0**	**92.4**	**52.8**	**34.4**	**44.0**	**92.8**	**48.8**	**35.2**	**42.0**	86.2	**42.8**	**27.2**	**37.6**	**67.2**	**25.4**	**14.8**	**23.4**
Few-Query 0	63.0	35.6	25.8	34.0	64.2	37.0	27.0	32.2	63.2	37.0	25.2	33.0	62.2	33.2	23.2	32.2	44.0	19.0	11.0	19.6
Few-Query 1	66.8	43.6	32.4	41.0	68.2	43.6	32.4	39.2	67.8	43.8	32.0	39.4	68.0	42.8	28.8	38.2	58.6	28.0	17.8	27.4
Few-Query 2	77.4	51.8	38.2	47.8	76.8	51.4	37.6	47.2	78.2	51.4	37.4	45.6	76.6	47.4	33.2	43.0	67.6	33.8	21.4	33.0
Few-Query 3	87.8	57.8	45.0	55.2	87.4	59.4	43.0	54.0	85.4	56.4	42.8	50.4	85.8	53.0	36.8	50.2	75.4	36.4	24.0	34.6
Few-Query 4	93.4	61.0	47.4	59.4	93.4	63.6	47.0	58.8	91.4	60.6	45.4	55.8	92.4	57.4	39.8	54.0	80.8	40.2	25.8	37.6
Few-Query 5	96.8	64.8	49.6	61.8	97.0	67.0	49.6	61.0	95.8	63.2	47.0	58.0	95.2	59.6	41.8	57.0	84.6	41.6	27.2	39.6

location of the objects or the relationship between the object and the background. Extending the ZQA attack to more sophisticated context models is a topic for future research. Furthermore, evaluating the ZQA in situations where the attacker uses a different notion of context than the victim system – e.g., attacker uses semantic context, detector uses context learned from pixels – would help researchers understand the broader applicability of this work.

Another limitation is that it is expensive to precompute the PSPM. Unlike the context graph, which only depends on the dataset, PSPM is a function of the dataset, attack model and perturbation level. We need to measure the attack success rate for each surrogate model at each perturbation level, and for all possible perturbations of a given object.

Potential negative societal impact and mitigation: This paper, as any other work that investigates an attack method, may be used maliciously to generate attacks against victim systems. Our goal with this work is to reduce technical surprise, and to fuel the development of defenses against powerful attacks. This work already makes the case for using a context-aware detector to thwart simple attacks. To thwart ZQA attacks described in this paper, the detector can attempt to constantly update its training data and label sets, and to develop increasingly sophisticated context models.

6. Conclusions

In this paper, we craft a novel zero-query attack by exploiting "a context graph" that captures co-occurrence relations of objects in a natural image. Against context-aware detectors, the fooling rate is significantly higher than that achieved by a context-agnostic attack. Unlike prior query-based attacks, our attack is extremely hard to detect since it hinges on using a single attempt. It achieves fairly good fooling rates by choosing an attack plan (i.e., perturbing multiple objects simultaneously to ensure context consistency) which is likely to succeed. The key innovation is a PSPM that provides this information offline. We observe that the use of PSPM not only boosts fooling rates in white-box settings, but also carries over to the black-box setting (i.e., when the detector model is different from that of the attacker) consistently for different attacker-defender pairs.

Acknowledgments. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under agreement number HR00112090096. Approved for public release; distribution is unlimited.

References

[1] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *International Joint Conference on Artificial Intelligence*, 2021. 3

[2] Ehud Barnea and Ohad Ben-Shahar. Exploring the bounds of the utility of context for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7412–7420, 2019. 2

[3] Ehud Barnea and Ohad Ben-Shahar. Exploring the bounds of the utility of context for object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. 2

[4] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context R-CNN: Long term temporal context for per-camera object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020. 2

[5] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. 2

[6] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *International Conference on Learning Representations*, 2018. 2

[7] Zikui Cai, Xinxin Xie, Shasha Li, Mingjun Yin, Chengyu Song, Srikanth V Krishnamurthy, Amit K Roy-Chowdhury, and M Salman Asif. Context-aware transfer attacks for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 2, 3, 7

[8] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017. 3

[9] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 1

[10] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018. 3

[11] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pages 1277–1294. IEEE, 2020. 2

[12] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[13] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017. 2

[14] Zhe Chen, Shaoli Huang, and Dacheng Tao. Context refinement for object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 71–86, 2018. 2

[15] Ka-Ho Chow, Ling Liu, Margaret Loper, Juhyun Bae, Mehmet Emre Gursoy, Stacey Truex, Wenqi Wei, and Yanzhao Wu. Adversarial objectness gradient attacks in real-time object detection systems. In *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 263–272. IEEE, 2020. 3

[16] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting Adversarial Attacks with Momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 2, 6

[17] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018. 2

[18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 6

[19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1, 3, 5

[20] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493. PMLR, 2019. 2

[21] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2017. 3

[22] Andrew Hollingworth. Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, 127(4):398, 1998. 2

[23] Shengnan Hu, Yang Zhang, Sumit Laha, Ankit Sharma, and Hassan Foroosh. Cca: Exploring the possibility of contextual camouflage attack on object detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7647–7654. IEEE, 2021. 3

[24] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020. 1, 6

[25] Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. Qeba: Query-efficient boundary-based blackbox attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1221–1230, 2020. 2

[26] Maosen Li, Cheng Deng, Tengjiao Li, Junchi Yan, Xinbo Gao, and Heng Huang. Towards transferable targeted attack.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 641–649, 2020. 2

[27] Qizhang Li, Yiwen Guo, and Hao Chen. Practical no-box adversarial attacks against dnns. *Advances in Neural Information Processing Systems*, 33:12849–12860, 2020. 2

[28] Shasha Li, Shitong Zhu, Sudipta Paul, Amit Roy-Chowdhury, Chengyu Song, Srikanth Krishnamurthy, Ananthram Swami, and Kevin S Chan. Connecting the dots: Detecting adversarial perturbations using context inconsistency. In *European Conference on Computer Vision*, pages 396–413. Springer, 2020. 2, 3

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 6

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[31] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 369–385, 2018. 3

[32] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. Dpatch: An adversarial patch attack on object detectors. *AAAI Workshop on Artificial Intelligence Safety*, 2019. 3

[33] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017. 2

[34] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6985–6994, 2018. 2

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 3, 5

[36] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 135–147, 2017. 3

[37] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. 2

[38] Aude Oliva, Antonio Torralba, Monica S Castelhano, and John M Henderson. Top-down control of visual attention in object detection. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 1, pages I–253. IEEE, 2003. 2

[39] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 6

[40] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016. 3

[41] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017. 1, 2

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2, 3

[43] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020. 3

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2, 3, 6

[45] Aniruddha Saha, Akshayvarun Subramanya, Koninika Patil, and Hamed Pirsiavash. Role of spatial context in adversarial robustness for object detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 3

[46] Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. Learning to share visual appearance for multiclass object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011. 2

[47] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018. 3

[48] Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*, 2020. 3

[49] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 1

[50] Antonio Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2):169–191, 2003. 2

[51] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018. 3

[52] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2020. 2

[53] Yajie Wang, Yu an Tan, Wenjiao Zhang, Yuhang Zhao, and Xiaohui Kuang. An adversarial attack on DNN-based black-box object detectors. *Journal of Network and Computer Applications*, 161, 2020. 2

[54] Xingxing Wei, Siyuan Liang, Ning Chen, and Xiaochun Cao. Transferable adversarial attacks for image and video object detection. In *IJCAI International Joint Conference on Artificial Intelligence*, 2019. 3

[55] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 1, 3

[56] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018. 3

[57] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1369–1378, 2017. 1, 3

[58] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 501–509, 2019. 3

[59] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 6

[60] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Annual Network and Distributed System Security Symposium*, 2017. 3

[61] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012. 2

[62] Mingjun Yin, Shasha Li, Zikui Cai, Chengyu Song, M Salman Asif, Amit K Roy-Chowdhury, and Srikanth V Krishnamurthy. Exploiting multi-object relationships for detecting adversarial attacks in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7858–7867, 2021. 2, 3

[63] Hantao Zhang, Wengang Zhou, and Houqiang Li. Contextual Adversarial Attacks for Object Detection. *International Conference on Multimedia and Expo (ICME)*, 2020. 3