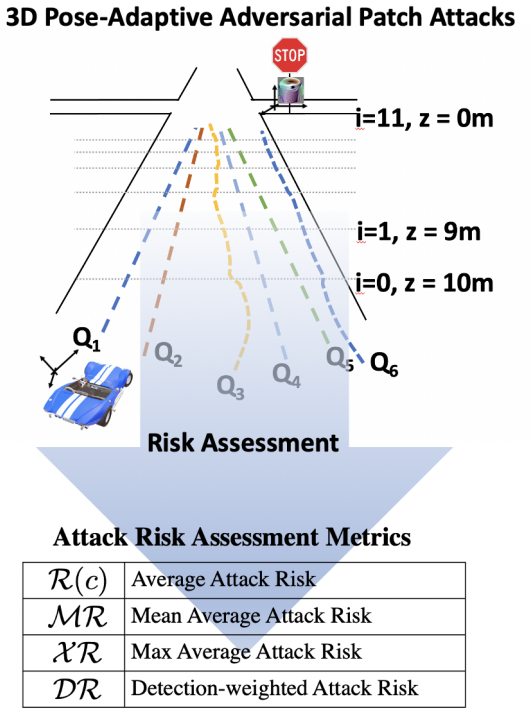
2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 978-1-6654-9346-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/WACV56688.2023.00143

# Do Adaptive Active Attacks Pose Greater Risk Than Static Attacks?

Nathan Drenkow*, Max Lennon*, I-Jeng Wang, Philippe Burlina
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road, Laurel, Maryland 20723
first.last@jhuapl.edu

## Abstract

*In contrast to perturbation-based attacks, patch-based attacks are physically realizable, and are therefore increasingly studied. However, prior work neglects the possibility of adaptive attacks optimized for 3D pose. For the first time, to our knowledge, we consider the challenge of designing and evaluating attacks on image sequences using 3D optimization along entire 3D kinematic trajectories. In this context, we study a type of dynamic attack, referred to as "adaptive active attacks" (AAA), that takes into consideration the pose of the observer being targeted. To better address the threat and risk posed by AAA attacks, we develop several novel risk-based and trajectory-based metrics. These are designed to capture the risk of attack success for attacking earlier in the trajectory to derail autonomous driving systems as well as tradeoffs that may arise given the possibility of additional detection. We evaluate performance of white-box targeted attacks using a subset of ImageNet classes, and demonstrate, in aggregate, that AAA attacks can pose threats beyond static attacks in kinematic settings in situations of predominantly looming motion (i. e., a prevalent use case in automated vehicular navigation). Results demonstrate that AAA attacks can exhibit targeted attack success exceeding 10% in aggregate, and for some specific classes, up to 15% over their static counterparts. However, taking into consideration the probability of detection by the defender shows a more nuanced risk pattern. These new insights are important for guiding future adversarial machine learning studies and suggest researchers should consider defense against novel threats posed by dynamic attacks for full trajectories and videos.*




**Attack Risk Assessment Metrics**

| | |
|---|---|
| $\mathcal{R}(c)$ | Average Attack Risk |
| $\mathcal{MR}$ | Mean Average Attack Risk |
| $\mathcal{XR}$ | Max Average Attack Risk |
| $\mathcal{DR}$ | Detection-weighted Attack Risk |

Figure 1. Our concept focuses on two foci a) developing new pose-optimized adaptive dynamic patch attacks and b) considering different threat models/requirements for measuring the resulting risk compared to static patch attacks.

## 1. Introduction

### 1.1. Motivation

Deep learning (DL) systems [19] have recently increased the promise of higher performance and greater autonomy in a range of AI application areas including image interpretation, vehicular navigation, robotics, and medicine [18, 15, 28, 27, 25, 29, 7, 4, 20, 32]. Concerns about trust in AI systems including ethical concerns such as explainability, privacy [30, 24, 17], fairness [5, 6], and – the focus of this paper – adversarial vulnerabilities (adversarial machine learning or AML) [14, 8] have recently put into question the deployment of certain autonomous systems. We investigate new types of AML attacks on image classifiers with the ultimate objective that such approaches could be used beneficially to audit the resilience of AI and autonomous systems toward these types of adversarial approaches.

Most work thus far in AML has focused on perturba-

*equal contribution

2642-9381/23/$31.00 ©2023 IEEE
DOI 10.1109/WACV56688.2023.00143

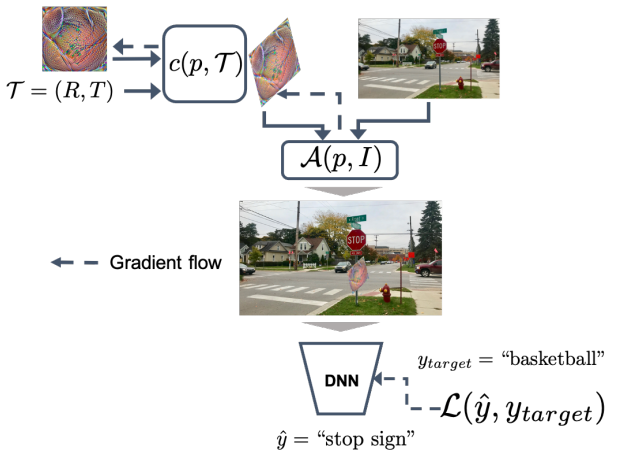



Figure 2. 3D patch optimization uses a loss based on an expectation taken over an ensemble of images and 3D poses that correspond those expected in a trajectory. Then the process does gradient descent in patch space so as to maximize the expected value of the probability for the targeted class.

tion attacks which are not physically implementable. Some work has developed patch-based attacks which are implementable, but these methods typically perform optimization in 2D image domains and solely on static patch/background image attack settings. By contrast, this study's foci (Figure 1) are a) to consider 3D settings and developing novel adaptive attacks along dynamic trajectories, and b) establishing measures of their associated risk based on different requirements to characterize success (such as preferences for attacking earlier in the trajectory).

### 1.2. Prior Work

AML has led to a wide variety of work in the past few years, with a continuous arms race between novel methods for both attacks and defenses. Most adversarial efforts targeting machine perception have focused on techniques that optimize attacks for specific full images while constraining attack perturbations to be imperceptible to humans (e. g., [33], [14], [22], [21]). These attacks are typically the main focus of AML research in the area of visual perception, despite the clear downside that perturbation-based attacks are only feasible if the adversary has electronic access to the vision pipeline. Patch-based attacks are characterized by local confinement of the perturbation to a specific area of the image and are intended to generalize to an ensemble of images. In contrast to perturbation methods, these attacks can be realized in a physical setting and this category includes both patch- and occlusion-based attacks [3, 11, 9, 31, 35, 16]. We focus here on these attacks since they are more easily implemented (e. g., by printing and placing optimized patches or occlusions in a scene). Given their potential impact on autonomous vehicular systems, patch attacks are of interest for many applications in robotics/vehicular autonomy, and for other AI applications such as software medical devices and robotics.

Examples of work focusing on such attacks includes [1, 3] which implemented a loss function featuring an expectation taken over an ensemble of images and a set of possible geometric transformations including 2D rotation, translation, and scale [1], black-box attacks [12, 36, 10], or the study of physical adversarial attacks for vehicle detectors using a simulator (Carla) as in [34, 23].

### 1.3. Contributions

In contrast to past studies, we take several steps toward considering threats posed by new adversarial strategies using physically realizable, kinematic-informed, active attacks that are adaptive. Specifically: a) we seek adversarial patch attacks that are optimized for 3D observers' pose; and that can b) adapt dynamically based on external geometric conditions. We focus on the specific case of looming translations of the observer being attacked by the patch (i. e., translations with a non-zero component along the camera optical axis). We also c) design new metrics to measure attack effectiveness over a kinematic trajectory; and d) compare performance of baseline and adaptive attack systems to assess their expected risk. Our new metrics allow specification and measurement of explicit preferences for attacks that pose maximal risk to autonomous systems by causing early-on failure to their navigation systems. Each metric captures a different perspective of attack risk allowing defenders to better understand the specific vulnerabilities of their vision system to patch attacks along trajectories.

## 2. Methods

### 2.1. Patch Optimization

We describe a framework and method for developing AAAs. As illustrated in Figure 2, a patch is optimized by applying backpropagation all the way to the patch/image domain to find maximally attacking perturbations. Formally, for a targeted, white-box attack, optimization of patches is performed such that the patch is optimal in expectation over a range of pose transformations and scene/image contexts, as in:

$$\arg\max_{p} \mathbb{E}_{I\sim\mathcal{I},T\sim\mathcal{T}} P(y_t|\mathcal{A}_T(p, I); \mathbf{w}) \quad (1)$$

for patch $p$, image $I$, transformation $T$ (including both rotation and translation), application method $\mathcal{A}_T(.)$, targeted label $y_t$, and predictive model $P(.)$ with parameters $\mathbf{w}$.

In the case of attacks against dynamic observers (e. g., autonomous vehicles, entities like Unmanned Aerial/Ground/etc. Vehicles), we consider how a patch attack might be produced to account for aspects of the vehicle

trajectory. To be more precise, we define a single probabilistic trajectory as a sequence:

$$Q = \{q_i = (I_i, \mathbf{q_i}) = (I_i, t_i^1, \ldots, t_i^M) \mid (t_i^1, \ldots, t_i^M) \sim \mathcal{T}(T^1, \ldots, T^M | i)\}_{i=1}^N, \quad (2)$$

where $I_i \sim \mathcal{I}$ is the scene image at index $i$ and $T(\cdot|i)$ represents the joint distribution over M types of patch transformations $t^1, \ldots, t^M$ at index $i$ collected over $N$ steps. The transformations may be standard rigid motion (i. e., in SE(3)) or may include other appearance-related transformations (e. g., colorimetric, textural, etc.).

We assume a common indexing convention across trajectories to ensure that for any trajectory $Q$, $Q[i]$ is consistent in some real-world sense. Without loss of generality, we assume $i$ is determined by partitioning a continuous trajectory along steps of time or distance.

Given this definition, we can now collect a set of trajectories from a fixed real-world scene (where $Q_j \in \Psi$ for $j \leq |\Psi|$). We consider scenes where some aspects of the scene are static (e. g., roads, road signs, etc.) while others are dynamic (e. g., cars on the road, pedestrians, etc.). For our purposes, we assume that patch attacks will be placed on static components in the scene (although the following methods generalize to placement on dynamic elements as well).

As illustrated in Figure 1, to allow for adaptation, we instead seek to optimize:

$$\arg\max_{p_i} \mathbb{E}_{I \sim \mathcal{I}_i, \mathbf{t} \sim \mathcal{T}_i} P(y_t | \mathcal{A}_{\mathbf{t}}(p_i, I_i)) \quad (3)$$

where we optimize a discrete set of patches $\mathcal{P}$, with each patch $p_i$ corresponding to a specific interval of the trajectory. This enables creating a dynamic patch sequence $\mathcal{P} = [p_0, p_1, \ldots, p_N]$ which changes as a function of the interval index (e. g., $p_i = \mathcal{P}[i]$). Intuitively, the interval index may correspond to something like the distance of the patch from the camera.

The transformation $\mathbf{t}$ itself is sampled from the joint distribution over transformations consistent with a fixed, specific interval $i$ of the trajectory (i. e., $\mathcal{T}_i$). To perform the optimization, we sample batches of background images of size $n_b$ and optimize patches according to the $i^{th}$ interval. The patch is optimized in expectation to perform well over scene contexts and transformations consistent with specific intervals of the trajectory.

We compare opposing approaches along the spectrum, on one end static attacks which are obtained by considering a single unique interval and patch for the entire trajectory (i. e., $M = 1$), and on the other, very dynamic attacks with multiple intervals (i. e., $M \geq 2$).

## 2.2. Metrics

In order to characterize the overall success of patch attacks along trajectories, and to impart a notion of risk measurement to the autonomous system, we introduce the following novel metrics:

**Average Attack Risk ($\mathcal{R}(c)$)** Taken across the entire attack trajectory:

$$\mathcal{R}(c) = \mathbb{E}_{l \sim \mathcal{U}(0, l_{max})} \Big[ A_c(l) \Big] \quad (4)$$

This is defined for a given target class $c$ in the set of classes $\mathcal{C}$, where $l \in [0, l_{max}]$, denotes the time or spatial interval (or cell index as we shall use) that the trajectory spans, and where $A_c(l)$ denotes attack accuracy or success rate for target class $c$ at cell distance $l$. In practice, we integrate over all the trajectory intervals to compute this metric. Note that $\mathcal{R}$ can be interpreted as a type of area under the curve measure where the x-axis is a distance/time index.

**Mean Average Attack Risk ($\mathcal{MR}$)**

We also measure aggregate performance computed over a set $C$ of targeted classes.

$$\mathcal{MR} = \mathbb{E}_{c \in C} \Big[ \mathcal{R}(c) \Big] \quad (5)$$

here the expectation is taken over all classes $c \in \mathcal{C}$, where $l \in [0, l_{max}]$ for all intervals of the trajectory, and where $A_c(l)$ the accuracy for class $c$ at point $l$.

**Max Average Attack Risk ($\mathcal{XR}$)**

Since the attacker is opportunistic and may have the pick of target class label, we also measure risk via maximal performance over all classes as an better measure assessing risk of a targeted attack:

$$\mathcal{XR} = \arg\max_{c \in C} \Big[ \mathcal{R}(c) \Big] \quad (6)$$

**Re-Weighted Risk Metrics ($\mathcal{R}_r$, $\mathcal{XR}_r$ and $\mathcal{MR}_r$)**

Further recognizing that there is increased value for the attacker (and risk for the attacked observer) to successfully and persistently attack earlier in the trajectory, we consider the re-weighted version of the previous metrics. To make this risk measurement approach more concrete, consider the fact that there is a distance past which braking effectiveness is reduced or non-beneficial to bring the car to a complete stop. Therefore, considering an attack that changes a stop sign detection into a 60 mph speed limit sign (or another class that would not lead the navigation system to apply breaking), then attacking early and consistently would disallow early braking which would have maximal benefit for the attacker and risk for the attacked vehicle.

Therefore, we redefine the re-weighted average attack accuracy taken across the entire attack trajectory as now:

$$\mathcal{R}_r(c) = \mathbb{E}_{l\sim\mathcal{U}(0,l_{max})}\Big[w(l)A_c(l)\Big] \tag{7}$$

for a weighting function $w(l)$ designed to emphasize greater distances and which is normalized to integrate to 1 over $l \in (0, l_{max}]$ ($w(l)$ is assumed here linearly decreasing in range $[1, 0]$ throughout the trajectory).

Likewise we can generalize $\mathcal{MR}$ and $\mathcal{XR}$ from before as:

$$\mathcal{MR}_r = \mathbb{E}_{c\in C}\Big[\mathcal{R}_r(c)\Big] \tag{8}$$

and

$$\mathcal{XR}_r = \arg\max_{c\in C}\Big[\mathcal{R}_r(c)\Big] \tag{9}$$

Changing the expression for $w(l)$ gives flexibility to tune the metric to particular trajectories or scenarios of interest.

**Detection-weighted Risk ($\mathcal{DR}_r(c)$)**

Lastly, to account for the possibility that a defender can detect patch attacks along the trajectory, the detection-weighted attack accuracy can be formulated as:

$$\mathcal{DR}_r(c) = \mathbb{E}_{l\sim\mathcal{U}(0,l_{max})}\Big[w(l)P_{\neg D}(l)A_c(l)\Big] \tag{10}$$

where $P_{\neg D} = 1 - P_D$ is the probability that the defender does not detect the patch attack. If $P_D(l) = \beta$ is constant for all positions along the trajectory (an optimistic case for the defender), then the risk reduces to $\mathcal{DR}_r(c) = (1-\beta)\mathcal{R}_r(c)$. It is then obvious that in this setting this equation only re-scales the original risk. Intuitively, any non-zero detection rate will naturally decrease the attack risk as we would expect.

## 2.3. Optimization

To determine the merits of dynamic patch attacks optimized for a trajectory in 3D space, we perform experiments with the following common characteristics. As in [13], we evaluate the patch attack effectiveness over a range of training/testing transformations. However, unlike that work, we do not vary the range of transformations introduced throughout the training time of the overall attack. Rather, we investigate the effects of subdividing the attack into increasingly narrow intervals, with a separate phase of the attack optimized for each interval.

We define a range of camera distances $[a, b]$ over which to consider the effectiveness of the patch at train and test time. The trajectory for which the attacks are optimized is defined along the z-axis within this range of distances. This distance range is intended to cover all patch scales that might be of interest: from the closest attack that might reasonably occur, to a distance great enough that developing salient features in the limited pixels allocated to the patch becomes difficult. At sufficient distances, roll, yaw, and pitch rotations become negligible.

In order to investigate the role of increased subdivision of a dynamic patch attack, we divide the range into $N$ intervals, and define the train and test time support for the patch (as in Eq. 3), $\mathcal{T} \sim \mathcal{U}_{\psi,\theta,z}$, where the set of transformations is drawn uniformly random from ranges $\psi \in [\pm 0°]$, $\theta \in [\pm 0°]$, and $z \in [(a + \frac{b-a}{N} \times (i-1)), (a + \frac{b-a}{N} \times (i))]$ for $i \in \{1, 2, \ldots, N\}$ (for yaw, roll, and distance respectively). Here, $\mathcal{U}_{\psi,\theta,z_i}$ denotes a multivariate uniform distribution sampled according to the intervals specified by $\psi, \theta, z_i$. Thus, each $i^{th}$ patch is unique to a specific interval of the trajectory. Taken as a group, the $N$ patches constitute a dynamic attack capable of acting over the entire camera distance range.

# 3. Experiments

For each experiment setting, we run the optimization and evaluation in a white-box setting against a ResNet-50 model trained on ImageNet, which represents a best case scenario from the patch attack scenario (i. e., where full knowledge of the architecture and weights is available to the attacker).

In our case, the range of camera distances is chosen to mimic the realism of a vehicle approaching an intersection. Although we may refer to distances in terms of meters, this is merely a convenience unit where the interpretation of a given distance measure is internal to the PyTorch3D library [26].

After optimization, each patch is tested against the same range of transformations under which it was trained, with each test transformation applied to patches inserted into a test set consisting of a random sample of images drawn from the same source as the train set. Patch coordinates are randomized during optimization and at attack time unless specified otherwise. Field of view (FoV) of the camera is kept at a constant $60.0\,°$.

## 3.1. Experiment Setting 1: Attacking Over An Ensemble of Classes

**Question:** In this experiment, we seek to assess the efficacy of a generalized dynamic patch attack vs. a static one. Therefore we ask: *what is the impact on risk of adaptive attacks compared to that of a single static patch attacks?*

**Approach:** This question relates to whether the patch optimization process produces patterns that depend heavily on scale or can be effective from a variety of apparent distances. To address this question, we optimized a set of patches following our core experimental protocol with patch roll fixed at $0\,°$, patch yaw fixed at $0\,°$, and location randomized. Patches were trained for 8 epochs, each epoch consisting of 200 batches of 32 images sampled from

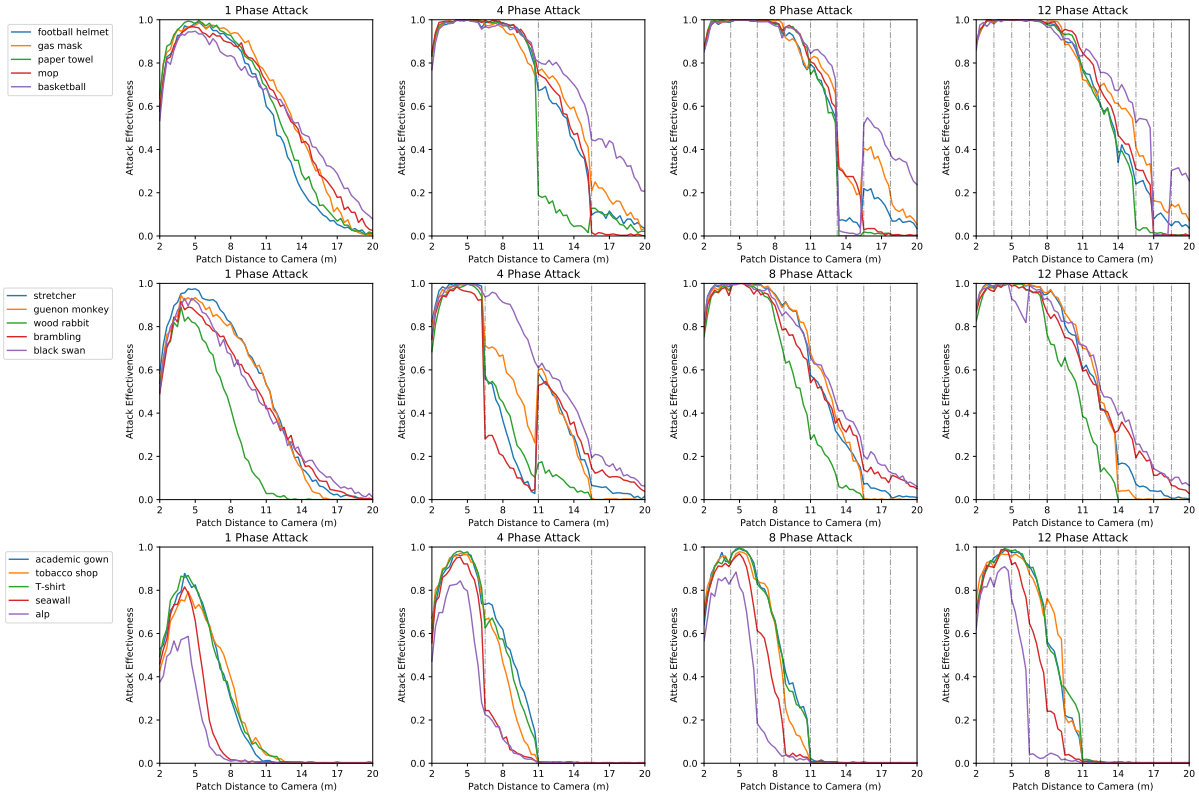



Figure 3. Performance results for setting 1 and obtained with attacks subdivided into different numbers of intervals (separated by the dotted lines shown). Performance is shown on the Y axes and is computed as attack effectiveness as a function of the camera distance testing was performed at (X axes). Top, middle and bottom rows show aggregation of results grouping together the high-, mid-, and low- performing target classes.

the ILSVRC 2012 validation set. We performed experiments for four levels of subdivision with respect to the attack; namely, we optimize patches with train time support $\mathcal{T} \sim \mathcal{U}_{\psi,\theta,z_{i,j}}$, where $\psi \in [\pm 0°]$, $\theta \in [\pm 0°]$, and $z_{i,j} \in [(a + \frac{b-a}{N_j} \times (i-1)), (a + \frac{b-a}{N_j} \times (i))]$ for $i \in \{1, 2, \ldots, N_j\}$, where $N_j \in \{1, 4, 8, 12\}$. All experiments were evaluated over the distance range $z \in [2, 20]$, approximated by 61 test points 0.3 meters apart. At each test point, the attack was evaluated using the patch whose train time support contained the point.

**Results:** Results are captured in Figures 3 and 4 and Tables 1 and 2.

**Discussion:** Tables 1 and 2 suggest that both for the weighted and unweighted metrics, adaptive patch attacks outperform their static counterpart. Drilling down to the specific metrics $\mathcal{R}(c)$ for all classes $c$, adaptation clearly confers a benefit, but the benefit varies depending on the number of intervals used. In aggregate (i. e., $\mathcal{MR}$), however, more adaptation produces a more successful attack. In these results, in Figure 3, we can see a common increase in overall attack success as the number of attack intervals increases. The attackness increases later in the trajectory when looming closer to the patch. The presence of certain discontinuities ("ravine") exist for specific intervals in some of the plots, in which the effectiveness of the optimized path over said interval is significantly lower than those on either side. This is likely due to the fact that the optimization is

Table 1. Metrics $\mathcal{R}(c)$, $\mathcal{XR}$ and $\mathcal{MR}$ for Experiment 3.1

| Number of Attack Steps | 1 | 4 | 8 | 12 |
|---|---|---|---|---|
| $\mathcal{R}$(football helmet) | 0.53 | 0.62 | 0.6 | **0.64** |
| $\mathcal{R}$(gas mask) | 0.61 | 0.66 | 0.66 | **0.69** |
| $\mathcal{R}$(paper towel) | 0.56 | 0.52 | 0.57 | **0.61** |
| $\mathcal{R}$(mop) | 0.6 | 0.62 | 0.61 | **0.66** |
| $\mathcal{R}$(basketball) | 0.6 | **0.74** | 0.69 | 0.73 |
| $\mathcal{R}$(stretcher) | 0.48 | 0.41 | **0.56** | **0.56** |
| $\mathcal{R}$(guenon monkey) | 0.45 | 0.46 | **0.56** | 0.55 |
| $\mathcal{R}$(wood rabbit) | 0.27 | 0.34 | **0.45** | **0.45** |
| $\mathcal{R}$(brambling) | 0.43 | 0.39 | 0.56 | **0.58** |
| $\mathcal{R}$(black swan) | 0.44 | 0.6 | **0.61** | **0.61** |
| $\mathcal{R}$(academic gown) | 0.24 | 0.34 | **0.36** | 0.35 |
| $\mathcal{R}$(tobacco shop) | 0.24 | 0.3 | 0.33 | **0.36** |
| $\mathcal{R}$(T-shirt) | 0.25 | 0.33 | 0.35 | **0.36** |
| $\mathcal{R}$(seawall) | 0.16 | 0.23 | **0.28** | **0.28** |
| $\mathcal{R}$(alp) | 0.11 | 0.19 | **0.2** | 0.19 |
| $\mathcal{XR}$ | 0.61 | **0.74** | 0.69 | 0.73 |
| $\mathcal{MR}$ | 0.4 | 0.45 | 0.49 | **0.51** |

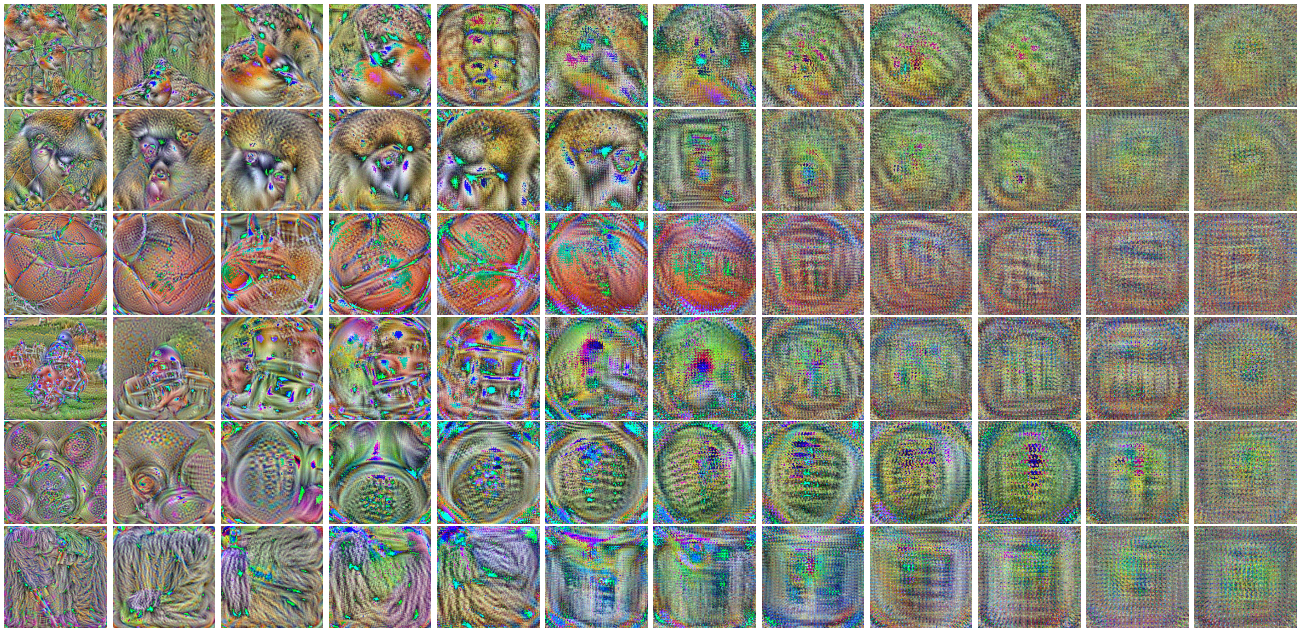


Figure 4. AAA patches shown from closest to furthest location for 12-step attacks and for classes Brambling, guenon monkey, basketball, football, gas mask and mop target classes for experiment setting 1. Distance flows left to right (and time in reverse of that).

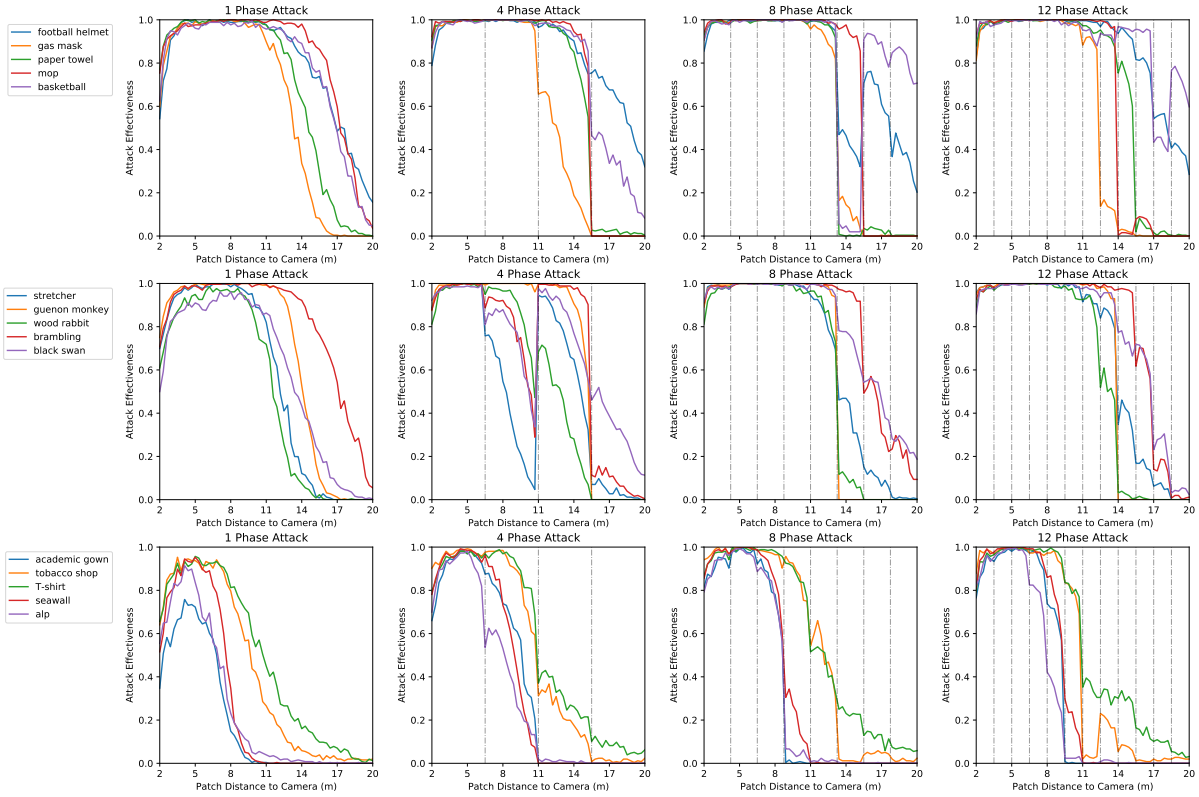



Figure 5. Performance results obtained for setting 2 and obtained for adaptive attacks subdivided into different numbers of intervals (separated by the dotted lines shown). Performance is shown on the Y axes and is computed as attack effectiveness as a function of the camera distance testing was performed at (X axes). Top, middle and bottom rows show aggregation of results grouping together the high-, mid-, and low- performing target classes.

Table 2. Metrics $\mathcal{R}_r(c)$, $\mathcal{XR}_r$ and $\mathcal{MR}_r$ for Experiment 3.1

| Number of Attack Steps | 1 | 4 | 8 | 12 |
|---|---|---|---|---|
| $\mathcal{R}_r$(football helmet) | 0.36 | 0.45 | 0.42 | **0.48** |
| $\mathcal{R}_r$(gas mask) | 0.45 | 0.50 | 0.51 | **0.54** |
| $\mathcal{R}_r$(paper towel) | 0.40 | 0.33 | 0.37 | **0.43** |
| $\mathcal{R}_r$(mop) | 0.46 | 0.44 | 0.42 | **0.49** |
| $\mathcal{R}_r$(basketball) | 0.48 | **0.63** | 0.56 | 0.60 |
| $\mathcal{R}_r$(stretcher) | 0.31 | 0.26 | **0.38** | 0.37 |
| $\mathcal{R}_r$(guenon monkey) | 0.29 | 0.29 | **0.38** | 0.36 |
| $\mathcal{R}_r$(wood rabbit) | 0.13 | 0.18 | **0.26** | **0.26** |
| $\mathcal{R}_r$(brambling) | 0.28 | 0.27 | 0.40 | **0.41** |
| $\mathcal{R}_r$(black swan) | 0.29 | **0.45** | **0.45** | **0.45** |
| $\mathcal{R}_r$(academic gown) | 0.11 | 0.17 | **0.18** | **0.18** |
| $\mathcal{R}_r$(tobacco shop) | 0.12 | 0.14 | 0.16 | **0.18** |
| $\mathcal{R}_r$(T-shirt) | 0.12 | 0.16 | **0.18** | **0.18** |
| $\mathcal{R}_r$(seawall) | 0.06 | 0.10 | **0.13** | **0.13** |
| $\mathcal{R}_r$(alp) | 0.04 | **0.08** | **0.08** | 0.07 |
| $\mathcal{XR}_r$ | 0.48 | **0.63** | 0.56 | 0.60 |
| $\mathcal{MR}_r$ | 0.26 | 0.30 | 0.32 | **0.34** |

done independently for each interval and that in some cases the patch optimization may be locked into a local maximum and is likely interval and class-dependent since we are aiming for targeted attacks. In general thought, it appears that this effect is limited, and performance curves, by and large, are smooth.

### 3.2. Experiment Setting 2: Attacking Over a Specific Class (Traffic Signs)

**Question:** In this experiment, we seek to better understand the the performance of dynamic patch attacks under more realistic attack conditions. As such, we ask: *what is the risk of a dynamic patch attack trained to attack specifically against a single class (here street signs)? And what effect does transforming the background realistically (changing the apparent distance of the surroundings in accordance with the distance of the patch) have on the efficacy of the attack?*

**Approach:** These questions address whether the patch optimization process allows patches to specialize against a specific ground truth class to be defeated, as well as whether a "zoomed out" background image is easier to attack at higher camera distances. To address these settings, we optimized a set of patches following the experimental protocol of Section 2, with patch roll fixed at $0\,^\circ$, patch yaw fixed at $0\,^\circ$, and location randomized. Patches were trained for 8 epochs, each epoch consisting of 32 batches of 32 images labeled as "street sign" from the ILSVRC 2012 training set. When each patch was inserted into an image, in addition to the patch, the image itself was also rendered using PyTorch3D at a camera distance corresponding to that of the patch. Similarly to Experiment 3.1, we conducted experiments for four levels of subdivision with respect to the attack; namely, we optimized patches with train time support $\mathcal{T} \sim \mathcal{U}_{\psi,\theta,z_{i,j}}$, where $\psi \in [\pm 0°]$, $\theta \in [\pm 0°]$, and $z_{i,j} \in [(a+\frac{b-a}{N_j}\times(i-1)), (a+\frac{b-a}{N_j}\times(i))]$ for $i \in \{1,2,\ldots,N_j\}$, where $N_j \in \{1,4,8,12\}$. All experiments were evaluated over the distance range $z \in [2,20]$ meters, approximated by $61$ test points $0.3$ meters apart. At each test point, the attack was evaluated using the patch whose train time support contained the point.

**Results:** Results are captured in Figure 5 and Tables 3 and 4 for the metrics both unweighted and weighted.

Table 3. Metrics $\mathcal{R}$, $\mathcal{XR}$ and $\mathcal{MR}$ for Experiment 3.2

| Number of Attack Steps | 1 | 4 | 8 | 12 |
|---|---|---|---|---|
| $\mathcal{R}$(football helmet) | 0.8 | 0.86 | 0.8 | **0.87** |
| $\mathcal{R}$(gas mask) | **0.61** | 0.58 | 0.62 | 0.57 |
| $\mathcal{R}$(paper towel) | 0.68 | **0.71** | 0.62 | **0.71** |
| $\mathcal{R}$(mop) | **0.83** | 0.72 | 0.73 | 0.66 |
| $\mathcal{R}$(basketball) | 0.78 | 0.8 | 0.84 | **0.9** |
| $\mathcal{R}$(stretcher) | 0.55 | 0.53 | 0.66 | **0.68** |
| $\mathcal{R}$(guenon monkey) | 0.66 | **0.71** | 0.62 | 0.65 |
| $\mathcal{R}$(wood rabbit) | 0.5 | 0.56 | **0.61** | 0.6 |
| $\mathcal{R}$(brambling) | **0.81** | 0.69 | **0.81** | 0.8 |
| $\mathcal{R}$(black swan) | 0.59 | 0.71 | **0.8** | 0.78 |
| $\mathcal{R}$(academic gown) | 0.2 | **0.38** | 0.34 | 0.37 |
| $\mathcal{R}$(tobacco shop) | 0.42 | 0.51 | **0.54** | 0.49 |
| $\mathcal{R}$(T-shirt) | 0.48 | 0.55 | **0.58** | 0.57 |
| $\mathcal{R}$(seawall) | 0.28 | 0.37 | 0.38 | **0.4** |
| $\mathcal{R}$(alp) | 0.25 | 0.32 | **0.34** | 0.33 |
| $\mathcal{XR}$ | 0.83 | 0.86 | 0.84 | **0.90** |
| $\mathcal{MR}$ | 0.56 | 0.6 | 0.62 | **0.63** |

Table 4. Metrics $\mathcal{R}_r$, $\mathcal{XR}_r$ and $\mathcal{MR}_r$ for Experiment 3.2

| Number of Attack Steps | 1 | 4 | 8 | 12 |
|---|---|---|---|---|
| $\mathcal{R}_r$(football helmet) | 0.71 | 0.79 | 0.70 | **0.80** |
| $\mathcal{R}_r$(gas mask) | **0.43** | 0.38 | **0.43** | 0.37 |
| $\mathcal{R}_r$(paper towel) | 0.52 | **0.54** | 0.43 | **0.54** |
| $\mathcal{R}_r$(mop) | **0.73** | 0.56 | 0.57 | 0.48 |
| $\mathcal{R}_r$(basketball) | 0.68 | 0.68 | 0.79 | **0.85** |
| $\mathcal{R}_r$(stretcher) | 0.37 | 0.38 | 0.48 | **0.51** |
| $\mathcal{R}_r$(guenon monkey) | 0.49 | **0.55** | 0.42 | 0.46 |
| $\mathcal{R}_r$(wood rabbit) | 0.32 | 0.37 | **0.42** | 0.41 |
| $\mathcal{R}_r$(brambling) | **0.72** | 0.55 | 0.69 | 0.67 |
| $\mathcal{R}_r$(black swan) | 0.44 | 0.60 | **0.68** | 0.66 |
| $\mathcal{R}_r$(academic gown) | 0.09 | **0.20** | 0.16 | 0.18 |
| $\mathcal{R}_r$(tobacco shop) | 0.25 | 0.32 | **0.35** | 0.30 |
| $\mathcal{R}_r$(T-shirt) | 0.31 | 0.38 | **0.41** | 0.40 |
| $\mathcal{R}_r$(seawall) | 0.13 | 0.19 | 0.19 | **0.21** |
| $\mathcal{R}_r$(alp) | 0.11 | 0.16 | **0.17** | 0.15 |
| $\mathcal{XR}_r$ | 0.73 | 0.79 | 0.79 | **0.85** |
| $\mathcal{MR}_r$ | 0.42 | 0.44 | 0.46 | **0.47** |

**Discussion:** Tables 3 and 4 suggest that for the weighted and unweighted risk metrics, adaptive patch attacks outperform their static counterpart. Similar to the previous experiments we see improvements with increased adaptation (intervals). This is also echoed in Tables 3 and 4 albeit the fact that some targeted classes do not exhibit an improvement with adaptation (e. g. gas mask and mop), but the majority of the targeted classes do. The fact that the targeted classes share little semantic/visual similarity with the attacked class only reinforces the strength of the attack. Using

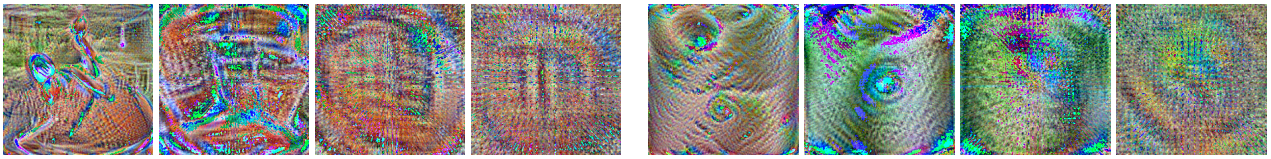


Figure 6. Comparison of patch classes "basketball" (left set) and "paper towel" (right set) over the 4 phase trajectory attack, which are notable for their significantly different behavior when optimized over the [11,15.5] distance range.

targeted classes with visual/semantic features shared with the attacked class (e.g., other street signs) would have made it easier to increase the dynamic attack's effectiveness. We also note a discontinuous behavior similar with the previous experiment setting. The pattern of increasing overall effectiveness is somewhat present in the low and medium-ranked effectiveness patch tiers (second and third rows of Figure 5, but not for the top ranked patches. Overall, however, the $\mathcal{R}(c)$ and $\mathcal{MR}$ scores for this experimental setting are higher than those for the first, with a "broadening to the left" of the time/distance interval for which the attack succeeds suggesting that the consistent truth class of "street sign" and / or the introduction of backgrounds transformed in a manner similar to the patch are conducive to a more effective attack.

## 4. Additional Discussion, Limitations and Future Work

While experiments demonstrate clear improvements of dynamic patch attacks over static ones, we do so only for rigid transformations in SE(3) and excluding other transformations which lend additional realism to the patch insertion (e. g., lighting or colorimetric variation). We only evaluate our method digitally, and for looming conditions, therefore future work should consider more extensive evaluation in real-world and extended kinematic settings.

We recognize that fabricating these dynamic attacks for real-world testing comes with technical challenges and costs, and this work provides quantitative justification for pursuing testing beyond simulation. Since the attack itself is assumed to be stationary relative to the attack observer, this allows the use of dynamic monitors (e.g., electronic billboard), a mosaic of static patches, or other view-dependent display types (e.g., auto-stereoscopic/holographic images). Additionally, by aligning the attack intervals with distance, the physical implementation does not require highly-precise or complex coordination with the observer's motion along the trajectory. Even so, new methods for monocular depth estimation [2] now achieve sufficient accuracy that a single device with a camera and display could be enough for both estimating the vehicle/observer's distance/orientation and choosing the appropriate patch attack to display.

Also, future work should broaden risk evaluation to consider more stringent threat settings from the attacker perspective. For example, one such setting could entail that a) the attacker be required to be consistently successful along the entire trajectory in order to disrupt the navigation system; or b) the defender has the additional opportunity to detect patches via some type of change detection and that any such detection would trigger a defense from the defender to adopt a more cautious navigation policy. In that case the advantage of the attacker relative to the defender could be expressed as the following risk:

$$\mathcal{RA} = P(\textit{attack} \ \wedge \ \textit{not detected}) = (P_A) * (P_{\neg D})^{N-1} \tag{11}$$

where $N$ is the number of unique patches along the trajectory. This establishes a trade-off for the attacker regarding using more dynamic attack patches vs. a single static one. Considering our results and even a modest success rate of change detection (e. g., $0.05$), the relative advantage shifts away from a dynamic patch back towards a static one. This indicates that in urban scenarios with much change (pedestrians, other movers) and changing illumination conditions, AAA may pose more risk since overall success of change detection mechanisms may be more muted. Also, just as a dynamic patch requires additional optimization requirements, the defender must also then develop a change-based defense to realize these benefits. In sum, future work should carefully consider threat settings to assess the actual risks of attacks, and consider how development/computational budgets for dynamic attacks and multi-layer defenses weigh on the overall viability of both approaches.

## 5. Conclusion

This study takes several steps towards designing and assessing risks of novel types of dynamic patch attacks set in kinematic settings, and optimized/adaptive to changing geometric scenarios. We conduct a sensitivity analysis on attack effectiveness as a function of the pose of the patch with respect to the camera and introduce new metrics to measure risk-based attack effectiveness over a kinematic trajectory. We demonstrate improvements of greater than 10% in attack success over baseline static patch attacks. Critically, we show that evaluation of risk should be nuanced and depends on assumptions made and threat settings, opening the door and informing future development of perception systems robust to dynamic attacks.

## References


[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

[2] Amlaan Bhoi. Monocular depth estimation: A survey. *arXiv preprint arXiv:1901.09402*, 2019.

[3] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[4] Phillippe Burlina, Neil Joshi, Katia D Pacheco, David E Freund, Jun Kong, and Neil M Bressler. Utility of deep learning methods for referability classification of age-related macular degeneration. *JAMA ophthalmology*, 136(11):1305–1307, 2018.

[5] Philippe Burlina, Neil Joshi, William Paul, Katia D Pacheco, and Neil M Bressler. Addressing artificial intelligence bias in retinal disease diagnostics. *arXiv preprint arXiv:2004.13515*, 2020.

[6] Philippe Burlina, William Paul, Philip Mathew, Neil Joshi, Katia D Pacheco, and Neil M Bressler. Low-shot deep learning of diabetic retinopathy with potential applications to address artificial intelligence bias in retinal diagnostics and rare ophthalmic diseases. *JAMA ophthalmology*, 138(10):1070–1077, 2020.

[7] Philippe M Burlina, Neil J Joshi, Elise Ng, Seth D Billings, Alison W Rebman, and John N Aucott. Automated detection of erythema migrans and other confounding skin lesions via deep learning. *Computers in biology and medicine*, 105:151–156, 2019.

[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.

[9] Aran Chindaudom, Prarinya Siritanawan, Karin Sumongkayothin, and Kazunori Kotani. Adversarialqr: an adversarial patch in qr code format. In *2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pages 1–6. IEEE, 2020.

[10] Francesco Croce, Maksym Andriushchenko, Naman D Singh, Nicolas Flammarion, and Matthias Hein. Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks. *arXiv preprint arXiv:2006.12834*, 2020.

[11] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

[12] Alhussein Fawzi and Pascal Frossard. Measuring the effect of nuisance variables on classifiers. In *British Machine Vision Conference (BMVC)*, number CONF, 2016.

[13] Neil Fendley, Max Lennon, I-Jeng Wang, Philippe Burlina, and Nathan Drenkow. Jacks of all trades, masters of none: Addressing distributional shift and obtrusiveness via transparent patch attacks. 2019.

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7848–7857, 2021.

[17] Bo Hui, Yuchen Yang, Haolin Yuan, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao. Practical blind membership inference attack via differential comparisons. *arXiv preprint arXiv:2101.01341*, 2021.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[19] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[20] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[21] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[22] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[23] Federico Nesti, Giulio Rossolini, Saasha Nair, Alessandro Biondi, and Giorgio Buttazzo. Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. *arXiv preprint arXiv:2108.06179*, 2021.

[24] William Paul, Yinzhi Cao, Miaomiao Zhang, and Phil Burlina. Defending medical image diagnostics against privacy attacks using generative methods. *arXiv preprint arXiv:2103.03078*, 2021.

[25] Mike Pekala, Neil Joshi, TY Alvin Liu, Neil M Bressler, D Cabrera DeBuc, and Philippe Burlina. Deep learning based retinal oct segmentation. *Computers in Biology and Medicine*, 114:103445, 2019.

[26] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[27] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2015.

[28] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region

proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[30] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[31] Abdul Jabbar Siddiqui and Azzedine Boukerche. Adversarial patches-based attacks on automated vehicle make and model recognition systems. In *Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, pages 115–121, 2020.

[32] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[34] Tong Wu, Xuefei Ning, Wenshuo Li, Ranran Huang, Huazhong Yang, and Yu Wang. Physical adversarial attack on vehicle detector in the carla simulator. *arXiv preprint arXiv:2007.16118*, 2020.

[35] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European Conference on Computer Vision*, pages 665–681. Springer, 2020.

[36] Chenglin Yang, Adam Kortylewski, Cihang Xie, Yinzhi Cao, and Alan Yuille. Patchattack: A black-box texture-based attack with reinforcement learning. In *European Conference on Computer Vision*, pages 681–698. Springer, 2020.