# Supplementary Materials for Towards Benchmarking and Assessing Visual Naturalness of Physical World Adversarial Attacks

Simin Li [1], Shuning Zhang[2], Gujun Chen[2], Dong Wang[2], Pu Feng[1],
Jiakai Wang[3], Aishan Liu[1], Xin Yi[2, 3†], Xianglong Liu[1, 3, 4†]
[1]SKLSDE Lab, Beihang University [2]Tsinghua University [3]Zhongguancun Laboratory
[4]Institute of data space, Hefei Comprehensive National Science Center

In this supplementary material, we provide additional details that we do not include in our main paper.

## 1. A brief survey of physical world attacks

In this section, we give a brief survey of existing physical world attacks. We do not consider this survey as a separate contribution of our paper, but rather an summary of how naturalness are recognized and evaluated in current attacks. We limit the surveyed work in our paper to the scope of existing survey [48], while this survey [48] do not summarize attack naturalness. We carefully report if existing literature claimed naturalness or stealthiness, and their methods to evaluate naturalness. The survey is given in Table. 1 and 2. A rise is shown in the focus of naturalness and stealthiness. In 22 papers before 2020, 7 paper claimed to be natural or stealthy (21%), while in 26 papers after 2020, this number raise to 13 (50%), showing a rapid increase. The increasing literature in this field motivates us to study the naturalness evaluation and understanding naturalness of physical world attacks.

## 2. Ethical discussions

**Ethical discussions of our research.** We briefly discuss how our works can be applied for social good. Firstly, the main motivation of our work is that physical world adversarial attacks are highly harmful for applications that rely on DNNs in physical world, yet its naturalness is still unknown, confusing defenders of possible defense strategies. In early stage of physical world attacks, physical world attacks are highly unnatural, allowing human to recognize and handle this attack easily. However, with large volume of works claiming their attack is more natural than baselines, it is unknown for defenders if human can still identify and remove them easily, since many of them do not compare directly with *clean* scenario with no attacks. In our work, we find while most physical world attacks are still unnatural, it can be more harmful under certain environmental configurations, or with certain semantic variations. Our findings remind defenders that adversarial attacks can be more harmful at certain occasions, while attackers might make it even more harmful by guiding human gaze directions.

Secondly, despite adversarial attacks, another line of works tries to generate *defensive* patterns in physical world for DNNs to recognize more confidently, as well as getting rid of adversarial attacks and extreme weathers [37, 50]. While these line of works serve as benign purpose, their method and generated defensive pattern is similar to physical world adversarial attacks, which is unnatural and not preferred by human. Our work helps to make their defensive patterns more natural in daily lives.

Thirdly, natural physical world adversarial examples offers a promising way to understand and test the robustness of DNNs. While attacks can be difficult in physical world because of environmental variations, natural physical world attacks seeks to find a physical noise with maximum attack capability under environmental variations, while achieving minimal perturbations, thus probing the most brittle and harmful part of DNNs. Being able to design natural physical world attack will help the design of robust DNNs.

Finally, while adversarial attacks can be harmful to DNNs, it can also be used as privacy protection methods against malicious surveillance technologies [6, 39]. In this case, ordinary people are able to leverage easy-to-get physical world adversarial attack methodologies to prevent them from being tracked by a malicious company. In such circumstances, a natural looking attack can encourage people to use this kind of technology more frequently, and discourage the use of AI technologies in malicious surveillance.
**Ethical discussions of experiments.** We detail the ethical concerns that might raise during the experiments. First, all participants only viewed the pictures for 2.5 seconds. Experiments were controlled to no more than 35 minutes so as not to cause aesthetic fatigue. In fact, most of the participants have commented that viewing these cars would not have any influence on their aesthetic value, although they

† Corresponding author

| Methods | Year | Natural? | Evaluation Methods |
|---|---|---|---|
| Sharif *et al.* [40] | 2016 | Yes | |
| AdvPatch [2] | 2017 | No | - |
| Hendrik Metzen et al. [13] | 2017 | No | - |
| CAMOU [59] | 2018 | No | - |
| DPATCH [30] | 2018 | No | - |
| EOT [1] | 2018 | Yes | - |
| $RP_2$ [10] | 2018 | Yes | - |
| $RP_2 +$ [43] | 2018 | Yes | - |
| ShapeShifter [5] | 2018 | No | - |
| ILLC [25] | 2018 | No | - |
| Rogue Signs [42] | 2018 | No | - |
| Zhao et al. [60] | 2019 | No | - |
| Thys et al. [46] | 2019 | No | - |
| Lee and Kolter [26] | 2019 | No | - |
| ShapeShifter [4] | 2019 | No | - |
| advPattern [51] | 2019 | No | - |
| D2P [21] | 2019 | No | - |
| MeshAdv [55] | 2019 | Yes | Conduct a user study where participants are asked to recognize those adversarial object to the ground-truth class or the adversarial target class. |
| PS-GAN [27] | 2019 | Yes | - |
| AGN [41] | 2019 | No | - |
| Morgulis *et al.* [33] | 2019 | Yes | - |
| D2P [22] | 2019 | No | - |
| AdvCam [8] | 2020 | Yes | Conduct a human perception study and ask human evaluators to choose whether a shown image is “natural and realistic” or “not natural or realistic”. |
| PhysGAN [24] | 2020 | Yes | - |
| Wu *et al.* [54] | 2020 | No | - |
| Adv T-shirt [56] | 2020 | No | - |
| Dynamic Adversarial Patch [14] | 2020 | No | - |
| UPC [20] | 2020 | Yes | - |
| Bias-based Attack [28] | 2020 | Yes | - |
| Wu et al. [53] | 2020 | No | - |
| Nakka and Salzmann [34] | 2020 | No | - |
| Yang *et al.* [57] | 2020 | No | - |
| Nguyen *et al.* [35] | 2020 | No | - |
| DAS [49] | 2021 | Yes | Conduct human perception studies on recognition and naturalness. Participants are asked to assign each of the camouflages to one of the 8 classes, from ”ground-truth” to ”I cannot tell what it is”, and score the naturalness of the camouflages from 1 to 10. |
| LAP [45] | 2021 | Yes | - |
| Meta-Attack [11] | 2021 | Yes | - |
| Zolfi et al. [62] | 2021 | No | - |

Table 1. An overview of physical world adversarial examples, including year of publication, claimed natural or not, and evaluation methods for naturalness. In 22 papers before 2020, 7 papers claimed to be natural (21%); while in 26 papers after 2020, this number raise to 13 (50%), showing a rapid increase.

| Methods | Year | Natural? | Evaluation Methods |
|---|---|---|---|
| Hu *et al.* [17] | 2021 | Yes | Conduct a subjective survey and ask participants to rank the naturalness of each patch. |
| SLAP [31] | 2021 | No | - |
| Sayles *et al.* [38] | 2021 | No | - |
| Adversarial Mask [61] | 2021 | Yes | - |
| AdvHat [23] | 2021 | No | - |
| CAC [9] | 2022 | No | - |
| DTA [44] | 2022 | No | - |
| FCA [47] | 2022 | Yes | - |
| TC-EGA [18] | 2022 | No | - |
| TnTs [7] | 2022 | Yes | Conduct two user studies. In Study 1 participants are asked to choose the most natural patch from given patches, and in Study 2 they are asked to compare the naturalness of adversarial patches and real images. |
| SPAA [19] | 2022 | Yes | Compare perceptual color distance and SSIM with other baselines. |
| Adversarial Sticker [52] | 2022 | Yes | - |

Table 2. (Table 1 continued) An overview of physical world adversarial examples, including year of publication, claimed natural or not, and evaluation methods for naturalness.

revealed that some of the pictures were really unnatural. Second, as only pictures in the real scenario were shown, participants do not appear to grow aversion for paintings on the car, nor do they grow the aversion towards the scenarios themselves. In fact, some participants (Participant 28 and Participant 100, denoted P28 and P100) commented that the environment was natural and beautiful to some extent indeed. Third, pictures we selected do not appear to affect participants' perception towards these semantic meanings. Participants only commented about the weird way how pikachu were attached on the screen, but they do not think pikachu would affect their own prior viewings of that semantic pictures.

**Ethical discussions of data collection.** We detail the ethical concerns that might arise in dataset collection. All participants in our experiments are clearly informed contents in our experiments and signed a consent that they agree their subjective ratings and gaze signals to be used for non-commercial research. Each participants are properly compensated $15 for their time. The experiment do not contain visually inappropriate, or sensitive contents (since only vehicles are contained), while we carefully control experiment design to avoid visual fatigue. Participants are able to quit whenever they feel inappropriate, albeit we do not observe such issue during experiment.

We took multiple efforts to ensure participants' anonymity. First, we highlight that our data *do not contain personally identifiable data* that uniquely defines an entity (fingerprints, face, iris, *etc*), since only subjective ratings and gaze fixations are collected. While information such as sex and age are not considered as Personal Identifiable Information, we do not disclose detailed information in our dataset and these information are used to report participants statistics only. To better protect anonymity, we also release rating distributions and gaze distributions averaged across multiple participants, leaving individual gaze data and ratings intact.

## 3. Experiment details

### 3.1. Implementation details

We initialize all compared baselines using their own implementations, network architecture and hyperparameters. To unify drastically different optimization methods used by baselines, for fair comparison, we use an Adam optimizer with learning rate $3 \times 10^{-5}$ for all baselines. For our own DPA, we randomly initialize prototype vector $z_\ell$ to a 1000-dimensional trainable vector, same as dimensions in ResNet50. For attention alignment, GradCam of model attention and human gaze are all calculated at size $224 * 224$ so as to ensure the same size as ResNet50. Hyperparameters $\lambda$ and $\gamma$ are empirically set to 8 and 3 respectively, which is determined by grid search. The train/valid/test split in all experiments are set to 8/1/1. We train all methods by 20 epochs except 100 epochs for WaDIQaM to ensure its convergence.

### 3.2. Evaluation metrics

We use Spearman Rank Order Correlation Coefficient (SROCC) and Pearson's Linear correlation coefficient (PLCC) as our evaluation metric, both is widely used in IQA literature [58]. SROCC measures the correlation between the rank order of predicted scores and ground truth scores. The higher SROCC is, the higher monotonic rela-

tionship is achieved between ground truth score and proposed IQA method. Specifically, given N distorted images, SROCC is computed as:

$$SROCC = 1 - \frac{6\sum_{n=1}^{N}(v_n - p_n)^2}{N(N^2-1)}, \tag{1}$$

where $v_n$ is the rank of ground truth MOS score $y_n$, $p_n$ is the rank of ground truth MOS score $\hat{y}_n$.

Similarly, Pearson's Linear correlation coefficient (PLCC, also called linear correlation coefficient, LCC in some papers) measures the *linear* correlation between predicted scores and ground truth scores, calculated as:

$$PLCC = \frac{\sum_{n=1}^{N}(y_n - \bar{y})(\hat{y_n} - \bar{\hat{y}})}{\sqrt{\sum_{n=1}^{N}(y_n - \bar{y})^2}\sqrt{\sum_{n=1}^{N}(\hat{y}_n - \bar{\hat{y}})^2}}, \tag{2}$$

where $\bar{y}$ is the mean of ground truth score, $\bar{\hat{y}}$ is the mean of predicted scores, respectively.

Additionally, we use cosine similarity (denoted as $S_C$) [32] to reflect the effect of attentive prior alignment loss. Specifically, we reshape model attention and human gaze into a one-dimensional vector, and calculate their cosine similarity as:

$$S_C = \frac{\tilde{\mathcal{S}} \cdot \tilde{\mathcal{A}}}{||\tilde{\mathcal{S}}|| \cdot ||\tilde{\mathcal{A}}||}, \tag{3}$$

in which $\tilde{\mathcal{A}}$ and $\tilde{\mathcal{S}}$ denotes the flattened one-dimensional vector of model attention $\mathcal{A}$ and human gaze $\mathcal{S}$. We find cosine similarity as a better measure of similarity, since we find the area of gaze signal is relatively small. If $\mathcal{L}_A$ was used as similarity, methods such as LPIPS can cheat to achieve "higher" alignment by having uniformly zero model attention $\mathcal{A}$, such that it achieves small $\mathcal{L}_A$, yet not focusing on vehicle at all.

### 3.3. Overall training algorithm

The overall training algorithm of DPA can be seen in Algorithm. 1. To align model behavior with human gaze and human rating distribution, we design Dual Prior Alignment (DPA) to imitate the naturalness assessment process of human. Without bells and whistles, we use a ResNet50 backbone as feature extractor. Then, we calculate pseudo probability $p_\ell(x, z)$ of each image and align it with human rating distribution by rating prior alignment loss (RPA) loss $\mathcal{L}_R$. Next, we calculate the modified attention map $\mathcal{A}(x, p)$ and then align it with human attention by attentive prior alignment loss $\mathcal{L}_A$.

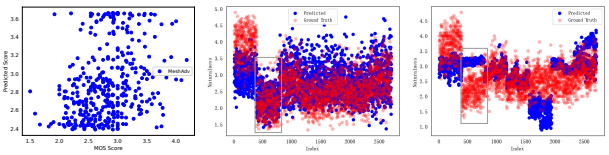
### 3.4. Generalization settings

In this section, we detail reasons to use our current generalization setting. While generalization towards unseen attacks can be tested by holding out one group of attacks and testing results within this group individually, such as Fig. 1a, such method do not consider generalization between different groups. It is possible for a method to achieve high correlation within its own group, yet having very low correlation outside its group. Since the overall goal of attack naturalness evaluation is to determine which method is more natural, in generalization settings, correlations between groups (between attack baselines) are considered more important. For example, in Fig. 1c (using MANIQA as baseline, each image cluster represents an attack baseline), the generalization result of the second cluster of images (in gray box) achieves high correlation within its own group, yet are rated much higher than its ground truth and higher than other predicted results. This violation of rank order is very harmful since MANIQA gives a false sense of ratings that the second cluster of images used have higher naturalness, yet these baseline itself is far less natural than imagined. Our DPA is free of this problem. Shown in Fig. 1b, all image clusters well represents its relative magnitude, thus better represents the quality of each attack baselines.

But how to represent and test this property? One feasible way is to test the generalization result (the attack baseline hold out for test only) with training data together. However,

**Algorithm 1** Training of Dual Prior Alignment Network

**Input:** Image database $\mathcal{D}$, human rating distribution $r$, human gaze saliency map $\mathcal{S}$, human MOS score $y$, backbone network $f_\theta$.
**Output:** Predicted naturalness score $\hat{y}$.
Randomly initialize prototype $z_\ell$ for each rating levels.
**for** Minibatch $x$ in dataset $\mathcal{D}$ **do**
  Calculate $p_\ell(x, z)$ by Eqn. 1.
  Calculate $\mathcal{L}_R$, $\mathcal{L}_S$ by Eqn. 2 and $\mathcal{L}_S = \frac{1}{N}\sum_{n=1}^{N}||\hat{y}_n - y_n||_2^2$.
  Calculate $\mathcal{A}(x, p)$ by Eqn. 5.
  Calculate $\mathcal{L}_A$ using $\mathcal{A}(x, p)$ and $\mathcal{S}$ by Eqn. 6.
  Update $\theta$ and $z$ by backpropagating Eqn. 7.
**end for**



(a) (b) (c)

Figure 1. Understanding our generalization setting. (a) Generalization do not well capture the correlation within class. (b) Results of DPA well captures the rank order between different classes. (c) Baseline (MANIQA in this illustration) failed to capture the rank orders. In (b) and (c), grouped ratings corresponds to different baselines.

we argue using training data in evaluating test capability is improper. To get rid of this problem and evaluate our generalization fairly, we propose for all seven groups (*i.e.*, attack patterns), we hold one group out for testing, and use the rest six groups for training. This process is repeated for seven times, each time holding out a distinct attack pattern. After all seven trainings are over, we got testing result from all seven groups from seven trained models. However, at this time, all images in testing set are not seen a priori to these seven trained models, respectively. Finally, correlation coefficients are calculated by the result of concatenating all seven testing set. This results in testing on the whole PAN dataset, yet all evaluated images are not seen in test set, which is consistent with generalization setting. The process is similar to a seven-fold cross validation, while holding out one group each time and reporting the merged result across each training as the final result.

# 4. Introduction of statistical tests

We provide an introduction of statistical tests here, while a more detailed tutorial can be seen in http://depts.washington.edu/acelab/proj/ps4hci/index.html. The overall goal of statistical test is to determine if there are enough reasons to "reject" a hypothesis, with randomness in evaluation process. For example, to validate a claim made in insight 1 in our main text, "distance has significant effect on naturalness", the corresponding hypothesis is "distance have no effect on naturalness". While the mean value of naturalness differs under different distances, it is unclear if this difference arise because of randomness in human ratings, or if there is indeed a significant difference. Statistical test calculates the probability $p$ this hypothesis holds: a $p < .05$ means there exists sufficient reason to reject the hypothesis, or "distance have an effect on naturalness", while $p > .05$ means there is no sufficient reason to reject the hypothesis (which means, distance has no effect on naturalness). Oftentimes we are interested in rejecting the hypothesis, with statistical significance reported under two levels: $p < .05$, meaning the probability that the difference is because of randomness is lower than .05 and $p < .001$, meaning the probability that difference is due to randomness is less than .001. If the hypothesis is not rejected, $p$ value are directly reported.

## 4.1. Analysis of variance

ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and generalizes the t-test beyond two means.

## 4.2. One-way analysis of variance

One way ANOVA is a technique that can be used to compare whether two sample's means are significantly different or not (using F distribution). This technique can be used only for numerical response data, and numerical or categorical input data. The ANOVA tests the null hypothesis, which states that samples in all groups are drawn from populations with the same mean values. If the gorup means are drawn from populations with the same mean values, the variance between group means shoule be lower than variance of the samples, following central limit theorem. A higher ratio implies that samples were drawn from populations with different mean values.

## 4.3. Two-way analysis of variance

Two way ANOVA is an extension of one-way ANOVA that examines the influence of two different categorical independent variables on one continuous dependent variable. Two-way ANOVA not only aims at assessing the main effect of each independent variable but also if there is any interaction between them.

## 4.4. Tukey's honest significance test

Also known as Tukey's tets, it is a isngle-step multiple comparison procedure and statistical test, which is used to find means that are significantly different from each other. It applies simultaneously to the set of all pairwise comparisons $\mu_i - \mu_j$ and identifies any difference between two means greater than the expected standard error.

## 4.5. Dunnett's test

Dunnett's test is a multiple comparison procedure. While Tukey's and Scheffe's methods allow any number of comparisons among a set of sample means, Dunnett's test only compares one group with the others, addressing a special case of multiple comparisons problem, doing pairwise comparisons of multiple treatment groups with a single control group.

## 4.6. Mann-Whitney U test

Mann-Whitney U test is a nonparametric test of the null hypothesis that for randomly selected values X and Y from two populations, the probability of X being greater than Y is equal to the probability of Y being greater than X.

## 4.7. Levene's Test

Levene's test is an inferential statistic used to assess the quality of variances for a variable calculated for two or more gorups. Some commomn statistical procedures assume that variances of populations from which different samples are drawn are equal. Levene's test assesses this assumption. It tests the null hypothesis that the population variances are

| Variation | Values |
|---|---|
| Background | Crossroads, Parking Lot |
| Illuminance | Day, Night |
| Yaw Angle | $0°, 45°, 90°, 135°, 180°, -45°, -90°, -135°$ |
| Pitch Angle | $22.5°, 45°, 67.5°, 90°$ |
| Distance | 5m, 8m, 10m |

Table 3. Values for environmental variations in our PAN dataset.

equal. If the resulting p-value of Levene's test is less than some significance level, the obtained differences in sample variances are unlikely to have occurred based on random sampling from population with equal variances. Thus, null hypothesis of equal variances would be rejected.

# 5. More details of PAN dataset

## 5.1. Variations

**Environmental variations.** We survey the common environmental variations in current physical world attacks and thus consider backgrounds, illuminance, yaw angles, pitch angles and distances for each baseline in PAN dataset. All values taken for each variation can be seen in Table. 3.

**Diversity variations.** For fair comparison, we select 10 diversity for each baselines, which is assembled from model diversity and semantic diversity. Model diversity refers to generating attack pattern based on 10 different recognition models, including 7 classification models and 3 object detection models. Semantic diversity contains 10 different images to constrain attack patterns, which is illustrated in Fig. 2 in our main text. Note that not all attack methods support all forms of variations. For example, CAMOU [59] and MeshAdv [55] do not constrain their attack by a natural image. Their diversity thus contains 10 model diversity only. For painting baseline, no attack is added, thus their diversity contains 10 semantic diversity only. For DAS [49] and AdvCam [8], we combine their diversity effect by 2 model diversity (ResNet50, DenseNet161) and five semantic diversity (smile, cat, dog, bird, pikachu). For UPC [20], attacks are generated on FasterRCNN, following their default settings, while we select 10 semantic diversity for it. Details about the model diversity and semantic diversity can be seen in Table. 4.

**Baseline reproduction.** To examine the correctness of our code, we test the attack capability of all our reproduced baselines following the code of DAS [49], following their default settings. We compare our reproduced code with DAS since we all test on CARLA environment, such that the results are comparable. The reported result in DAS and our reproduced result are given in Table. 5. The result of Clean, MeshAdv and DAS follows the same trend with the result in original DAS paper, while our reproduced results are uniformly 3% higher than original DAS, which could be explained by randomness, or difference in experimental minutiae. For the higher attack capability of CAMOU and UPC, in DAS paper, they used the released attack pattern by CAMOU and UPC directly, while we train them in CARLA environment from scratch. For painting and AdvCam, their attack results are not compared in DAS, thus also unavailable in Table. 5.

**Impact of different variations.** We divide our PAN dataset by different variations, and calculate the performance of the images on each subset respectively to find the tradeoff between attack capability and naturalness, and the impact of each variation. We use Mean Opinion Score (MOS) to denote naturalness and Attack Success Rate (ASR) to denote attack capability (calculated by attacking ResNet152, a black-box model to all our used models). For each variation, the line graph for each baseline at different values of that variation is given in its corresponding plot. See Fig. 2 for all baselines' performance under each variation. While an inverse relationship between attack capability and naturalness can be found in general, some levels are typically more natural and having higher attack capability, which might be more harmful in reality.

## 5.2. Data properties

**Human ratings.** To aggregate subjective human ratings, we use Mean Opinion Score (MOS) after outlier rejection to calculate the naturalness of each image. Following prior works [12], MOS is calculated by averaging the individual opinion scores from multiple subjects.

**Gaze saliency map.** After obtaining all human fixations, the saliency map of human gaze distribution can be generated by applying a Gaussian mask of the same shape to each human fixation points [29]. The saliency map of each images can be calculated as:

$$\mathcal{S}_i(k,l) = \frac{1}{C}\sum_{j=1}^{|\mathcal{J}_i|}\sum_{t=1}^{T}\exp\left[-\frac{(k-f_k^{j,t})^2+(l-f_l^{j,t})^2}{\sigma^2}\right], \tag{4}$$

where $(k,l)$ is the coordinates of saliency map $\mathcal{S}_i$, $i$ refers to the $i^{th}$ image in dataset. $j \in \mathcal{J}_i$ is the participant who rated image $i$. $f_k^{j,t}$ and $f_l^{j,t}$ is the fixation point of participant $j$ at time $t$, at coordinate $k$ and $l$. $\sigma$ is the standard deviation of Gaussian (*i.e.*, $\sigma$=0.33 is recommended for our eye tracker [①]). Finally, $C$ is the normalization constant that normalize the sum of $\mathcal{S}_i$ to 1.

**Participants.** A total of 126 participants (57 female, 69 male, age=$22.2\pm3.3$) were recruited from campus. All with normal(corrected) eyesights. None of them were familiar with image quality assessment experiments.

[①] See https://github.com/TobiasRoeddiger/GazePointHeatMap for more details.

| Baseline | Model Diversity | Semantic Diversity |
|---|---|---|
| Clean | - | - |
| CAMOU [59] | ResNet50, DenseNet161, VGG16, Inception-v3, MobileNet-v2, EfficientNet-b0, MnasNet, YOLOv4, Faster R-CNN, Mask R-CNN | - |
| MeshAdv [55] | ResNet50, DenseNet161, VGG16, Inception-v3, MobileNet-v2, EfficientNet-b0, MnasNet, YOLOv4, Faster R-CNN, Mask R-CNN | - |
| DAS [49] | ResNet50, DenseNet161 | Smile, Cat, Dog, Bird, Pikachu |
| UPC [20] | Faster RCNN | Smile, Cat, Dog, Bird, Pikachu, Flower, Bird2, Dog2, Flower2, Hello Kitty |
| AdvCam [8] | ResNet50, DenseNet161 | Smile, Cat, Dog, Bird, Pikachu |
| Painting | - | Smile, Cat, Dog, Bird, Pikachu, Flower, Bird2, Dog2, Flower2, Hello Kitty |

Table 4. Model and semantic diversity of all baselines.

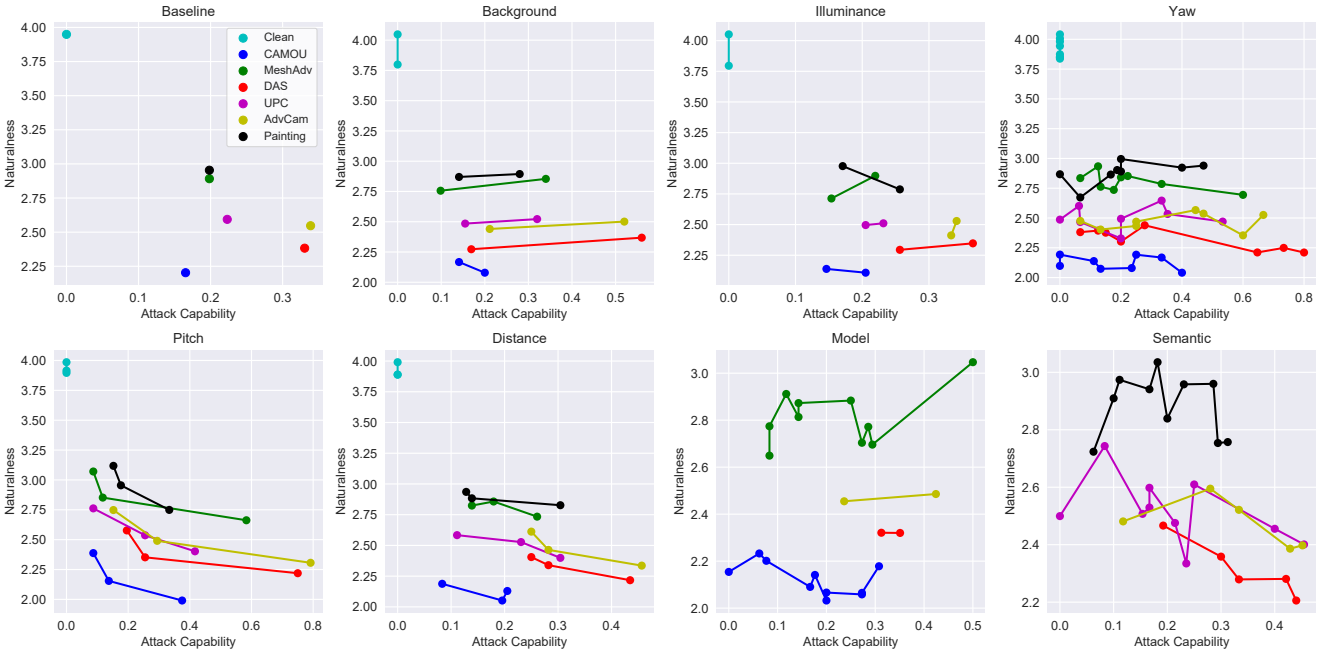



Figure 2. The naturalness and attack capability of the images at different settings in PAN dataset. For pitch angles, we remove 90° because ResNet152 has a 0% classification success rate for clean images in that case.

| Baseline | Accuracy(%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Clean | Painting | CAMOU | MeshAdv | DAS | UPC | AdvCam |
| Original | 73.51 | - | 48.93 | 35.33 | 32.49 | 48.18 | - |
| Ours | 77.33 | 66.75 | 41.73 | 38.27 | 35.87 | 41.73 | 40.80 |

Table 5. Test results of DAS (original) and ours.

**Labortary setup.** The images were displayed on a 16-inch(2560*1600) resolution integrated screen for experiment. A low light room was used with an approximate viewing distance about 70cm. A Tobii Eye Tracker 5 (equipped in front of the screen) was adapted for eye gaze tracking. It records eye gaze points at about 60 GP/sec. A calibration process was done before the experiment.

**Stimuli presentation.** We adopt a single stimulus continuous procedure for naturalness evaluation [36] and ask the participant to focus on and evaluate the naturalness of the image. To avoid the interference of rating naturalness on gaze signal collection, evaluation of each image follows a 2 phase process. In phase 1, an image was selected from PAN dataset and participants were asked to look at this image for 2.5 seconds, with eye tracker activated to collect eye gaze points. The time is determined by our pilot study to ensure eye gaze coverage and prevent fatigue. In phase 2, a dark screen was shown, while participants were asked to rate the image by a 5-point Absolute Category Rating (ACR) scale [15], *i.e.*, bad (1), poor (2), fair (3), good (4) and excellent (5). They were guided to rate from the perspective that whether the picture was natural and beautiful

in physical world setting. To minimize the interference of rating process on gaze signal, participants were asked to press 1-5 on keyboard to give the rating and press enter to evaluate the next image, without needing to actually look at the keyboard.

**Experiment process.** Each participants were asked to evaluate 320 images in the 2 phase process, resulting in a total of $(320 * 126 =)40320$ image ratings. We shuffle the dataset such that each participant view no repeated image, whereas each image will be rated by 15 participants (totally $(15 * 2688 =)40320$ images) when all experiments are over. The images were divided into 8 sessions, each session containing 40 images, with a warmup session (40 images) at the beginning, images used in warmup session is randomly selected for each participant. To reduce fatigue, there is a rest session with at least 20 seconds between two sessions. It takes participants no more than 35 minutes to finish all experiments so as to avoid fatigue [3]. Each participant was compensated $15.

**Quality control.**

**Line clickers** During the period of scoring, we have observed from the data that 3 of the participants (2 female, 1 male) have chosen over 80% the same score. We judge from the setting similar to Koniq10k [16] that they were line-clickers. We remove all their ratings and gaze data from PAN dataset.

**Outliers** Similar to Koniq10k [16], we used SROCC as a metric to judge whether a worker is an outlier. We removed the pictures from the workers with the least SROCC score (6.54%) calculated with the full-volume data.

**Quality control of scoring** During the examination of scores, we find that some of the scores were unfit for the certain participant as well as the certain picture. We further used SROCC as an metric, from the dimension of participant as well as picture to judge whether a picture has high congruence with the whole dataset. We removed the ones with the least score (9.36%), resulting in an increase on Intraclass Correlation Efficient (ICC) of 0.12 (from 0.22 to 0.34). We shall note that, while the images we evaluate are from different domains comparing with KonIQ-10k, the ICC results of our PAN dataset and KonIQ-10k are not comparable. However, the improvement of ICC in our method is similar to KonIQ-10k (+0.12 in our paper, +0.13 in KonIQ-10k). We further calculated SROCC score using bootstraping following KonIQ-10k [15] and the results were shown in Fig. 3. We observe similar SROCC improvement trend of KonIQ-10k and our PAN dataset. Note that since KonIQ-10k are collected via crowdsourcing, they have larger number of observers. However, our PAN dataset additionally provide high-quality gaze signal which supports human behavioral analysis, which is not possible in KonIQ-10k.

**Quality control of gaze** We considered the quality of gaze important since (1) it could reflect the engagement of the certain participant during experiment; (2) gazing data is an essential part to be adopted and analyzed throughout our work. Unqualified eye gaze data mainly consist of the following situations: (1) eye gaze data of the pictures were unable to be collected due to problems of eye tracker; (2) participants were unable to concentrate effectively on the picture due to their own reason; (3) participants who concentrate on too small an area that we considered it not a valid gaze. We only removed these unqualified eye gaze data (27.58%) according to corresponding criteria.

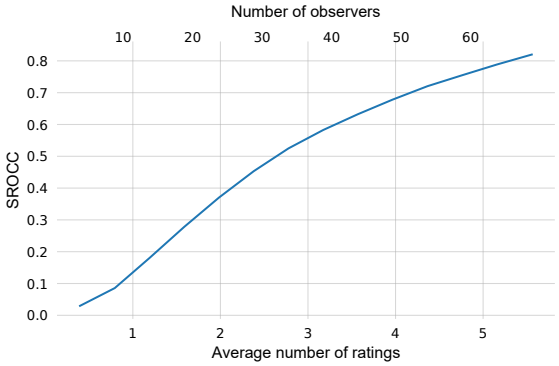



Figure 3. Testing inner coherence of our PAN dataset by bootstrapping participants' scoring data. Result shows similar trend with KonIQ-10k dataset.


Figure 4. Exemplar images of PAN-Phys in physical world settings. PAN-Phys contains 504 images collected in real world, which includes variance in pitch angles, yaw angles and backgrounds. Similar to PAN, we collect human ratings and human gaze for all images.

## 5.3. PAN-Phys dataset: PAN dataset in physical world

To verify the generalization capability of PAN in real world, we propose PAN-Phys dataset, which contains 504 real world images with 8 pitch angles (0°, 45°, 90°, 135°, 180°, 225°, 270°, 325°), 3 yaw angles (0°, 45°, 90°) and 3 backgrounds (shot in three different places in our lab). Since painting adversarial patterns in real vehicles is costly, we use a toy vehicle and stick paper-printed adversarial patterns by HP Color LaserJet Pro MFP M281fdw on the toy vehicle to mimic real world scenarios. We photo all images by a Huawei Mate 40 Pro camera. Exemplar images are shown in Fig. 4. After acquiring all images, we collect human gaze and human ratings using the same protocol and

procedures as our PAN dataset.

As a dataset photoed in real world, the environment setting and vehicle of PAN-Phys is significantly different from PAN, thus acts as a good testbed for (1) testing domain generalization. The model trained on PAN should also work well in PAN-Phys. (2) testing domain adaptation, where algorithms trained on PAN can leverage PAN-Phys to adapt their assessment to real world scenarios. (3) understanding the domain gap between real world and CARLA simulation environment. Statistical tests can be applied to gain further insight of human decisions and human behaviors via analyzing human ratings and human gaze in two domains. We defer these experiments as future work.

# 6. Physical naturalness assessment protocol

To solve these problems and avoid bias mentioned in Insight 2, we design Physical Naturalness Assessment Protocol (PNAP) to standardize subjective naturalness assessment in physical world attack. While PAN tests on large number of images that might be burdensome to researchers, PNAP picks a minimal subset of factors in PAN (10 images for each attack pattern), while maximally maintaining consistency with results in PAN by keeping statistical significance of ranks and values between baselines. In order to facilitate fast and automatic process and assessment of image naturalness, we proposed Physical Naturalness Assessment Protocol, a suggestion for which factor and which picture should be considered during naturalness evaluation. We revealed the exploration and execution steps in a sequence.

## 6.1. Factor pruning criterion

Factor pruning of PNAP follows two criterion. Overall, a pair of factors should be kept if baselines has different rankings under these pairs, and the ranking difference is not attributed to randomness. First, we prune a factor if the rank order of different baselines is consistent in different levels of this factor, such that testing on any level of factor provides identical result (*i.e.*, prune the factor if $p < .05$). Second, if the first criterion does not hold, we prune the factor if the value of baselines is not statistically significant in different levels of that factor, such that violations in rank order can be attributed to randomness (*i.e.*, prune the factor if $p > .05$). If both criterion are not satisfied, we run a pairwise post-hoc test between each factor levels and determine the smallest subset of levels of this factor that satisfy the two criterion above.

## 6.2. Evaluation process

**Experiment process.** Based on criterion discussed above, the environmental factors left for evaluation are: *pitch angle*: 67.5°, 90°; *yaw angle*: -90°, -45°, 45°, 90°, 135°. All levels of distance, background and illuminance are pruned (*i.e.*, select any level yields equal result), so we use *distance*-5m, *background*-crossroads, *illuminance*-light for simplicity. As a result, for each attack pattern, 10 images with different environmental factors are used for evaluation. We recommend using Absolute Categorical Rating (ACR) [15] to collect human rating.

**Data analysis.** We recommend using one-way ANOVA for data analysis. Improvement in naturalness could be claimed if significant effect are found ($p < .05$) and the mean value of newly proposed attack is higher than all baselines.

## 6.3. Discussion

While PNAP struggles to alleviate cherry picking, we acknowledge that PNAP might not rule out all bias, since new attacks is not contained in PAN and might have unique visual characteristics. However, PNAP still provides a valuable first step to alleviate the bias arise from the disparate impact caused by environmental variations, and we suggest using PNAP for a more solid and comprehensive evaluation.

# 7. Investigation of the correlation between semantic pictures and cars

Through the analysis of different semantic pictures' MOS, we found that semantic factors may significantly affect the naturalness of the pictures. Hence we further conducted a user study to investigate whether cars affect peoples' naturalness perception of the semantic pictures, namely whether the car would affect people's perception of pictures' naturalness.

## 7.1. Design, apparatus and participants

We used a between group design to investigate the effect of cars on the semantic pictures in order to minimize the inner influence of cars. We recruited (14*2=)28 participants (16 female, 12 male, age=21.3, SD=4.1) from campus. Each participant was presented $10 * 9/2$ =45 questions using a questionnaire and all participants have filled in the questionnaire without dropping.

## 7.2. Procedure

Each participant was asked to rate 10 image in pair, resulting in 45 pairs. The order of the image pairs were randomized and predetermined. For each pair of picture per participant, they were asked to judge which picture is more natural using a True of False question. We thus collected (45*14=)630 judges in total. Each participant took about 5 minutes to complete the judging process. For those who need to rate which semantic pictures attached to the cars were more natural(Group 1, G1), we separated the semantic pictures and the cars, then asked the participants to perceive "Attaching which semantic picture to the car would be more natural". For those who need to rate which semantic pictures themselves would be more natural(Group 2, G2), we

only present semantic pictures and asked the participants to perceive "which semantic picture would seem more natural". Thus, we could judge whether the presence of cars would affect the naturalness of semantic pictures.

### 7.3. Result and analysis

We firstly analyzed which picture gained the most preference by participants. For image pair (Image A and Image B), if the participant thought Image A is more natural than Image B, we would add one point to Image A and add zero point to Image B. Hence using this method, the naturalness of images could be reflected from the total points a image earned. For G1, the top-3 favorite semantic pictures were cat, dog and pikachu1 while for G2 the top-3 favorite semantic pictures were flower, bird and cat. Participants preference showed significance (Mann Whitney test was used between independent samples, $Z = -3.724$, $p < .001$). Further interview with participants shown that participants tended to more focused on the "semantic meanings of the picture" when asked "Attaching which semantic picture to the car would be more natural"(G1, P9). However, participants tended to be more focused on the "structural meanings of the picture" when asked "Which semantic picture would seem more natural"(G2, P11). Through the interview process with the participants, we arrived at the conclusion that the existence of car caused the change of perception attention, resulting in different naturalness assessment scores.

## 8. Improve naturalness with DPA

While initially proposed as a naturalness evaluation method, in this section, we explore the possibility to use our DPA method as an optimization metric to improve naturalness of existing physical world attacks (*i.e.*, DAS). To ensure comparison consistency, we first train DAS to minimize its own original loss function:

$$\min \mathcal{L}_d + \lambda \mathcal{L}_e + \mathcal{L}_s. \tag{5}$$

To tradeoff attack capability and naturalness, we select $\lambda$ to 0, 1e-5, 1e-4 and 1e-3, respectively. Note that 1e-4 is used in their official code. In subsequent sections, we refer these results as DAS, while $\lambda$=1e-4 referred as original DAS.

To improve naturalness of DAS, we add DPA on DAS by directly subtracting the naturalness ratings of the images in the loss function to optimize naturalness:

$$\min \mathcal{L}_d + \lambda \mathcal{L}_e + \mathcal{L}_s - \gamma \mathcal{L}_n, \tag{6}$$

where $\mathcal{L}_n$ is the naturalness ratings of the images assessed by DPA. Example of the generated camouflages are shown in Fig. 5. To quantify the effect of DPA, we kept $\lambda$ to 1e-4 used in their official code, while changing the magnitude of $\gamma$ as 0, 1, 2 and 3 to evaluate its impact. We call settings with $\gamma$= 1, 2, 3 as DAS+DPA. DAS+DPA reduces to original DAS when $\gamma$=0.

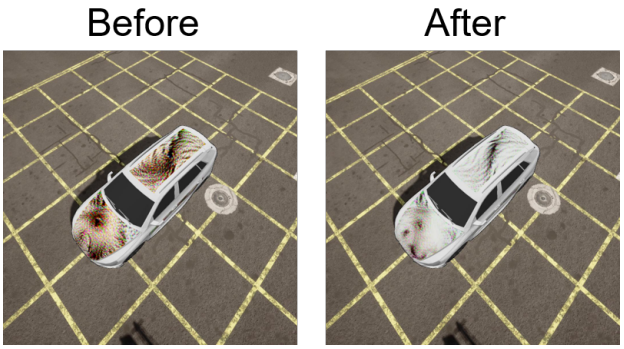



Figure 5. Comparison of the textures before and after subtracting the naturalness ratings assessed by DPA in the loss function of DAS.

To subjectively evaluate naturalness ratings, we recruited 11 participants (9 male, 2 female, age=22.91, SD=3.12) from campus to evaluate the naturalness of our attack, none of these participants took part in our prior experiments. We strictly follow our proposed PNAP protocol, generating 10 images for each attack patterns and show each images in a predetermined random sequence, resulting in a total of 70 images (3 DAS, 3 DAS+DPA, 1 original DAS). The experiment follows the same setting and interface as PAN dataset collection using 5-point Absolute Categorical Rating (ACR), the only difference is we do not collect gaze signal for simplicity. The experiment takes approximately 5 minutes, each participant was compensated $1 for their time.

### 8.1. Comparing naturalness with PNAP

We first evaluate the impact of DPA on improving naturalness of DAS. Following our proposed PNAP protocol, we compare DPA+DAS setting with original DAS. Adding DPA improves naturalness of DAS by a significant margin, *i.e.*, $\gamma$ = 2 (One way ANOVA, $F_{1,110} = 12.05$, $p < .001$, +18.93%) and 3 (One way ANOVA, $F_{1,110} = 3.82$, $p < .05$, +18.93%). When $\gamma$ = 1, DPA was too small have significant effect on naturalness (One way ANOVA, $F_{1,110} = 0.02, p = 0.874, n.s.$) or attack capability.

While the comparison demonstrates naively adding DPA to DAS improves naturalness, it is possible that this improvement results in sacrificing attack capability. To address this concern, we compare the attack-naturalness tradeoff in the subsequent section. Statistical tests are not available in this setting since requires prohibitively large computation resources to tune attack hyperparameters in order to ensure different attacks have exactly the same attack capability. Moreover, it might be impossible for some old attacks to achieve high attack capability as SOTA methods.

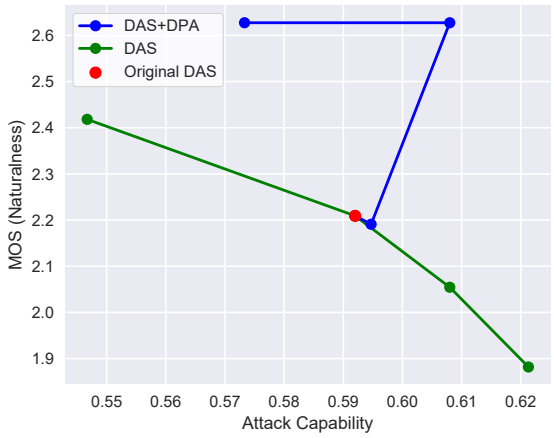



Figure 6. Improving attack naturalness by DPA. Green line refers to DAS with different naturalness adjustments. Blue line requires DAS+DPA, with different magnitude of DPA. Red dot refers to the original parameter used by DPA.

| Category | Method | SROCC (↑) | PLCC (↑) | $S_C$ (↑) |
|---|---|---|---|---|
| FR-IQA | PSNR | 0.3560 | 0.3685 | - |
| | SSIM | 0.4573 | 0.3968 | - |
| | LPIPS | 0.0256 | 0.0128 | 0.0019 |
| | E-LPIPS | 0.1515 | 0.4173 | 0.0461 |
| Others | GIQA(KNN) | 0.0900 | 0.0637 | - |
| | GIQA(GMM) | 0.1231 | 0.0839 | - |
| NR-IQA | BRISQUE | 0.0593 | -0.0130 | - |
| | ResNet50 | 0.4925 | 0.4461 | 0.2235 |
| | WaDIQaM | 0.3242 | 0.3187 | 0.2470 |
| | RankIQA | 0.2920 | 0.2834 | 0.0336 |
| | DBCNN | 0.3539 | 0.3171 | 0.2229 |
| | HyperIQA | 0.4790 | 0.4403 | 0.1871 |
| | Paq2Piq | 0.3639 | 0.3274 | 0.1540 |
| | MANIQA | 0.1550 | 0.1863 | 0.0790 |
| NR-IQA | DPA (Ours) | **0.5468** | **0.4856** | **0.7630** |

Table 6. Generalization results of DPA and other baselines on PAN dataset. By imitating human behavior, DPA also gets strong generalization capability comparing with baseline methods.

### 8.2. Generalizing to unseen attacks

For DPA to reliably estimate attack naturalness, it must be ensured new attacks keep the same order with others. To solve this problem, we follow a 7-fold cross validation setting, taking out one evaluated method from training set each time. Next, we concatenate all testing set of 7-fold cross validation and report it as our final result, such that the generalization result of all evaluated methods are fairly evaluated. *See detailed illustration in supplementary materials*. As listed in Table. 6, we can draw several conclusions as follows:

(1) DPA improves generalization over the best baseline (ResNet50) by **0.0543 (+11.02%)** in SROCC and **0.0395 (+8.85%)** in PLCC. While human has high generalization capability, we are surprising to find in DNNs, aligning with human behaviors also leads to better generalization.

(2) Shown in Fig. **??**, during generalization, model attention of DPA keeps aligned with human gaze, achieving 208.90% higher $S_C$ comparing with the best baseline.

### 8.3. Naturalness-attack tradeoff with DPA

In this section, we illustrate the tradeoff between naturalness and attack capability by comparing DAS with DAS+DPA. As illustrated in Fig. 6, DAS+DPA achieves higher attack capability as well as higher naturalness comparing with DAS. Perhaps surprisingly, we even find using DPA improves the attack capability of original DAS sometimes. Comparing with the most natural variation of DAS, DAS+DPA improves naturalness by 8.65% and attack capability by 11.21%, respectively. Moreover, comparing with the original version of DAS, DAS+DPA improves naturalness by 18.93% and attack capability by 2.70%, respectively. The experiment results points out our DPA indeed helps improve naturalness, while still maintaining reasonable attack capability.

Due to the time limit, we conduct experiments of using DPA to improve attack naturalness on DAS only, leaving experiments to improve other attacks by DPA as our future work. Overall, in this paper, we take a first step to evaluate naturalness of physical world attacks, we deem improving attack naturalness by environment modulation, guiding human gaze to enhance naturalness or using DPA more smartly and robust to adversarial attacks as our future research direction.

## References

[1] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 2

[2] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 2

[3] S. Chakravarty. Methodology for the subjective assessment of the quality of television pictures. 1995. 8

[4] S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. volume 11051, pages 52–68. 2019. 2

[5] S.-T. Chen, C. Cornelius, J. Martin, and D. H. P. Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 52–68. Springer, 2018. 2

[6] V. Cherepanova, M. Goldblum, H. Foley, S. Duan, J. Dickerson, G. Taylor, and T. Goldstein. Lowkey:

Leveraging adversarial attacks to protect social media users from facial recognition. *arXiv preprint arXiv:2101.07922*, 2021. 1

[7] B. G. Doan, M. Xue, S. Ma, E. Abbasnejad, and D. C. Ranasinghe. Tnt attacks! universal naturalistic adversarial patches against deep neural network systems. *IEEE Transactions on Information Forensics and Security*, 2022. 3

[8] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1000–1008, 2020. 2, 6, 7

[9] Y. Duan, J. Chen, X. Zhou, J. Zou, Z. He, J. Zhang, W. Zhang, and Z. Pan. Learning coated adversarial camouflages for object detectors. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 891–897, 2022. 3

[10] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. 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. 2

[11] W. Feng, B. Wu, T. Zhang, Y. Zhang, and Y. Zhang. Meta-attack: Class-agnostic and model-agnostic physical adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7787–7796, October 2021. 2

[12] D. Ghadiyaram and A. C. Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 6

[13] J. Hendrik Metzen, M. Chaithanya Kumar, T. Brox, and V. Fischer. Universal adversarial perturbations against semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2755–2764, 2017. 2

[14] S. Hoory, T. Shapira, A. Shabtai, and Y. Elovici. Dynamic adversarial patch for evading object detection models. *arXiv preprint arXiv:2010.13070*, 2020. 2

[15] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 7, 8, 9

[16] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 8

[17] Y.-C.-T. Hu, B.-H. Kung, D. S. Tan, J.-C. Chen, K.-L. Hua, and W.-H. Cheng. Naturalistic physical adversarial patch for object detectors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7848–7857, October 2021. 3

[18] Z. Hu, S. Huang, X. Zhu, F. Sun, B. Zhang, and X. Hu. Adversarial texture for fooling person detectors in the physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13307–13316, June 2022. 3

[19] B. Huang and H. Ling. Spaa: Stealthy projector-based adversarial attacks on deep image classifiers. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 534–542. IEEE, 2022. 3

[20] L. Huang, C. Gao, Y. Zhou, C. Xie, A. L. Yuille, C. Zou, and N. Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 6, 7

[21] S. T. Jan, J. Messou, Y.-C. Lin, J.-B. Huang, and G. Wang. Connecting the digital and physical world: Improving the robustness of adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):962–969, Jul. 2019. 2

[22] S. T. Jan, J. Messou, Y.-C. Lin, J.-B. Huang, and G. Wang. Connecting the digital and physical world: Improving the robustness of adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 962–969, 2019. 2

[23] S. Komkov and A. Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 819–826. IEEE, 2021. 3

[24] Z. Kong, J. Guo, A. Li, and C. Liu. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[25] A. Kurakin, I. J. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 2

[26] M. Lee and Z. Kolter. On physical adversarial patches for object detection, 2019. 2

[27] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao. Perceptual-sensitive gan for generating adversarial patches. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):1028–1035, Jul. 2019. 2

[28] A. Liu, J. Wang, X. Liu, B. Cao, C. Zhang, and H. Yu. Bias-based universal adversarial patch attack for automatic check-out. In *European conference on computer vision*, pages 395–410. Springer, 2020. 2

[29] H. Liu and I. Heynderickx. Visual attention in objective image quality assessment: Based on eye-tracking data. *IEEE transactions on Circuits and Systems for Video Technology*, 21(7):971–982, 2011. 6

[30] X. Liu, H. Yang, Z. Liu, L. Song, H. Li, and Y. Chen. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299*, 2018. 2

[31] G. Lovisotto, H. Turner, I. Sluganovic, M. Strohmeier, and I. Martinovic. {SLAP}: Improving physical adversarial examples with {Short-Lived} adversarial perturbations. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1865–1882, 2021. 3

[32] H. Mitchell. Image similarity measures. In *Image Fusion*, pages 167–185. Springer, 2010. 4

[33] N. Morgulis, A. Kreines, S. Mendelowitz, and Y. Weisglass. Fooling a real car with adversarial traffic signs. *arXiv preprint arXiv:1907.00374*, 2019. 2

[34] K. K. Nakka and M. Salzmann. Indirect local attacks for context-aware semantic segmentation networks. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 611–628, Cham, 2020. Springer International Publishing. 2

[35] D.-L. Nguyen, S. S. Arora, Y. Wu, and H. Yang. Adversarial light projection attacks on face recognition systems: A feasibility study. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 814–815, 2020. 2

[36] M. H. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on broadcasting*, 50(3):312–322, 2004. 7

[37] H. Salman, A. Ilyas, L. Engstrom, S. Vemprala, A. Madry, and A. Kapoor. Unadversarial examples: Designing objects for robust vision. *Advances in Neural Information Processing Systems*, 34:15270–15284, 2021. 1

[38] A. Sayles, A. Hooda, M. Gupta, R. Chatterjee, and E. Fernandes. Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14666–14675, 2021. 3

[39] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*, pages 1589–1604, 2020. 1

[40] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 2

[41] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. A general framework for adversarial examples with objectives. *ACM Transactions on Privacy and Security (TOPS)*, 22(3):1–30, 2019. 2

[42] C. Sitawarin, A. N. Bhagoji, A. Mosenia, P. Mittal, and M. Chiang. Rogue signs: Deceiving traffic sign recognition with malicious ads and logos. *arXiv preprint arXiv:1801.02780*, 2018. 2

[43] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, and T. Kohno. Physical adversarial examples for object detectors. In *12th USENIX workshop on offensive technologies (WOOT 18)*, 2018. 2

[44] N. Suryanto, Y. Kim, H. Kang, H. T. Larasati, Y. Yun, T.-T.-H. Le, H. Yang, S.-Y. Oh, and H. Kim. Dta: Physical camouflage attacks using differentiable transformation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2022. 3

[45] J. Tan, N. Ji, H. Xie, and X. Xiang. Legitimate adversarial patches: Evading human eyes and detection models in the physical world. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 5307–5315, New York, NY, USA, 2021. Association for Computing Machinery. 2

[46] S. Thys, W. V. Ranst, and T. Goedeme. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 49–55, Long Beach, CA, USA, 2019. IEEE. 2

[47] D. Wang, T. Jiang, J. Sun, W. Zhou, Z. Gong, X. Zhang, W. Yao, and X. Chen. Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2414–2422, 2022. 3

[48] D. Wang, W. Yao, T. Jiang, G. Tang, and X. Chen. A survey on physical adversarial attack in computer vision. *arXiv preprint arXiv:2209.14262*, 2022. 1

[49] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8565–8574, June 2021. 2, 6, 7

[50] J. Wang, Z. Yin, P. Hu, A. Liu, R. Tao, H. Qin, X. Liu, and D. Tao. Defensive patches for robust recognition in the physical world. In *Proceedings of the*

*IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2456–2465, 2022. 1

[51] Z. Wang, S. Zheng, M. Song, Q. Wang, A. Rahimpour, and H. Qi. advpattern: physical-world attacks on deep person re-identification via adversarially transformable patterns. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8341–8350, 2019. 2

[52] X. Wei, Y. Guo, and J. Yu. Adversarial sticker: A stealthy attack method in the physical world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[53] T. Wu, X. Ning, W. Li, R. Huang, H. Yang, and Y. Wang. Physical adversarial attack on vehicle detector in the carla simulator, 2020. 2

[54] Z. Wu, S.-N. Lim, L. S. Davis, and T. Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 2

[55] C. Xiao, D. Yang, B. Li, J. Deng, and M. Liu. Meshadv: Adversarial meshes for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 6, 7

[56] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European conference on computer vision*, pages 665–681. Springer, 2020. 2

[57] X. Yang, F. Wei, H. Zhang, and J. Zhu. Design and interpretation of universal adversarial patches in face detection. In *European Conference on Computer Vision*, pages 174–191. Springer, 2020. 2

[58] G. Zhai and X. Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63(11):1–52, 2020. 3

[59] Y. Zhang, H. Foroosh, P. David, and B. Gong. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2019. 2, 6, 7

[60] Y. Zhao, H. Zhu, R. Liang, Q. Shen, S. Zhang, and K. Chen. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1989–2004, London United Kingdom, 2019. ACM. 2

[61] A. Zolfi, S. Avidan, Y. Elovici, and A. Shabtai. Adversarial mask: Real-world adversarial attack against face recognition models. *arXiv preprint arXiv:2111.10759*, 2021. 3

[62] A. Zolfi, M. Kravchik, Y. Elovici, and A. Shabtai. The translucent patch: A physical and universal attack on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15232–15241, 2021. 2