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

Abstract

Physical world adversarial attack is a highly practical and threatening attack, which fools real world deep learning systems by generating conspicuous and maliciously crafted real world artifacts. In physical world attacks, evaluating naturalness is highly emphasized since human can easily detect and remove unnatural attacks. However, current studies evaluate naturalness in a case-by-case fashion, which suffers from errors, bias and inconsistencies. In this paper, we take the first step to benchmark and assess visual naturalness of physical world attacks, taking autonomous driving scenario as the first attempt. First, to benchmark attack naturalness, we contribute the first Physical Attack Naturalness (PAN) dataset with human rating and gaze. PAN verifies several insights for the first time: naturalness is (disparately) affected by contextual features (i.e., environmental and semantic variations) and correlates with behavioral feature (i.e., gaze signal). Second, to automatically assess attack naturalness that aligns with human ratings, we further introduce Dual Prior Alignment (DPA) network, which aims to embed human knowledge into model reasoning process. Specifically, DPA imitates human reasoning in naturalness assessment by rating prior alignment and mimics human gaze behavior by attentive prior alignment. We hope our work fosters researches to improve and automatically assess naturalness of physical world attacks. Our code and dataset can be found at https://github.com/zhangsn-19/PAN.

1. Introduction

Extensive evidences have revealed the vulnerability of deep neural networks (DNNs) towards adversarial attacks [5, 17, 27, 37, 53, 72–74] in digital and physical worlds. Different from digital world attacks which make pixelwise perturbations, physical world adversarial attacks are especially dangerous, which fail DNNs by crafting specifically designed daily artifacts with adversarial capability [2,13,31,48,51,59,70]. However, physical world attacks are often conspicuous, allowing human to easily identify and remove such attacks in real-world scenarios. To sidestep such defense, in 48 physical world attacks we surveyed[①], 20 papers (42%) emphasize their attack is natural and stealthy to human [9, 11, 22, 31, 49, 55].

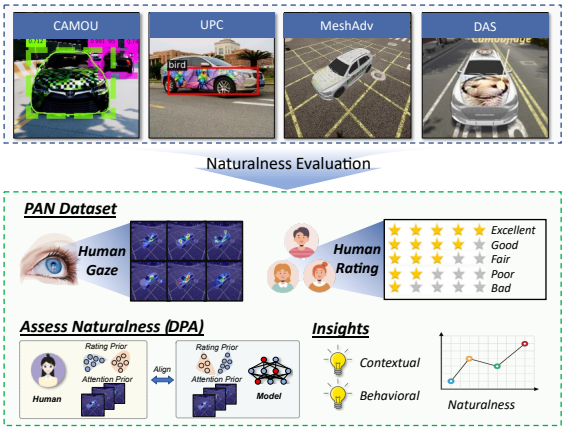


Figure 1. Overview of our work. To solve the problem in physical world attack naturalness evaluation, we provide PAN dataset to support this research. Based on PAN, we provide insights and naturalness assessment methods of visual naturalness.

Despite the extensive attention on visual naturalness, studies on natural attacks follow an inconsistent and case-by-case evaluation. In 20 surveyed papers claimed to be natural or stealthy[②], (1) 11 papers perform no experiment to validate their claim. (2) 11 papers claim their attack closely imitates natural image, but it was unclear if arbitrary natural image indicates naturalness. (3) 5 papers validate naturalness by human experiments, yet follow very different evaluation schemes and oftentimes neglect the gap between existing attacks and natural images. These problems raise our question: how natural indeed are physical world attacks?

† Corresponding author

[①] See this survey in supplementary materials.
[②] A work can have multiple limitations.

In this paper, we take the first attempt to evaluate visual naturalness of physical world attacks in autonomous driving [24], a field of attack with increasing attention [11, 22, 55, 58, 70]. Since the factors and methods studied in our work are common in physical world attacks and not limited to autonomous driving, our methods and findings also have the potential to be applied to other scenarios. The overview of our work is summarized in Fig. 1. To benchmark attack naturalness, we contribute Physical Attack Naturalness (PAN) dataset, the first dataset to study this problem. Specifically, PAN contains 2,688 images in autonomous driving, with 5 widely used attacks, 2 benign patterns (*i.e.*, no attacks) for comparison, 5 types of environmental variations and 2 types of diversity enhancement (semantic and model diversity). Data was collected from 126 participants, containing their subjective ratings as an indicator of naturalness, and their gaze signal for all images as an indicator of the selective attention area of human when they make naturalness ratings [66].

PAN provides a plethora of insights for the first time. First, we find contextual features have significant effect on naturalness, including semantic variations (using natural image to constrain attack) and environmental variations (illumination, pitch/yaw angles, *etc*). Properly selecting environmental and semantic factors can improve naturalness up to 34.73% and 8.09%, respectively. Second, we find contextual features have disparate impact on naturalness of different attacks, some attacks might look more natural under certain variations, which can lead to biased subjective evaluation even under identical settings. Third, we find naturalness is related to behavioral feature (*i.e.*, human gaze). Specifically, we find attacks are considered less natural if human gaze are more centralized and focus more on vehicle (with statistical significance at $p < .001$). This correlation suggests modelling and guiding human gaze can be a feasible direction to improve attack naturalness.

Finally, since manually collecting naturalness ratings requires human participation and can be laborious as well as costly, based on PAN dataset, we propose Dual Prior Alignment (DPA), an objective naturalness assessment algorithm that gives a cheap and fast naturalness estimate of physical world attacks. DPA aims to improve attack result by embedding human knowledge into the model. Specifically, to align with human reasoning process, rating prior alignment mimics the uncertainty and hidden desiderata when human rates naturalness. To align with human attention, attentive prior alignment corrects spurious correlations in models by aligning model attention with human gaze. Extensive experiments on PAN dataset and DPA method shows training DPA on PAN dataset outperforms the best method trained on other dataset by 64.03%; based on PAN dataset, DPA improves 3.42% in standard assessment and 11.02% in generalization compared with the best baseline. We also make early attempts to improve naturalness by DPA.

Our **contributions** can be summarized as follows:

- We take the first step to evaluate naturalness of physical world attacks, taking autonomous driving as a first attempt. Our methods and findings have the potential to be applied to other scenarios.
- We contribute PAN dataset, the first dataset that supports studying the naturalness of physical world attacks via human rating and human gaze. PAN encourage subsequent research on enhancing and assessing naturalness of physical world attacks.
- Based on PAN, we unveil insights of how contextual and behavioral features affect attack naturalness.
- To automatically assess image naturalness, we propose DPA method that embeds human behavior into model reasoning, resulting in better result and generalization.

2. Related Works

Adversarial Attacks and Naturalness. Adversarial attacks are elaborately designed attacks to fool DNNs. Based on attack domains, adversarial attacks could be categorized as *digital world attacks* and *physical world attacks*. Digital world attacks [5, 17, 37, 53] add oftentimes imperceptible pixelwise adversarial perturbations on images, *its naturalness are well characterized.* Laidlaw *et al.* [28] find LPIPS [65] well correlates with naturalness. E-LPIPS [25] improves robustness over adversarial attacks by adding transformations to input image. To generate naturalness adversarial attack, approaches has been made based on color space [71], LPIPS [7] or frequency analysis [36].

However, digital world attack fail in physical world, where diverse environmental variations exists. This motivates physical world attacks, which create adversarial artifacts robust to real world uncertainty. Volumes and scenarios of physical world attacks are growing rapidly. Widely known studies includes adversarial glass [48], 3D adversarial objects [2], road sign classification [13, 31, 51], vehicle camouflage [11, 22, 55, 58, 70] and adversarial t-shirt [21, 54, 59]. *While physical world attacks are practical and robust, its visual appearance are usually unnatural.* Mainstream works in naturalness enhancement hide attack patterns in a suitable image that well fits with attack scenario [9, 22, 31, 49, 55] or hide attack with natural styles [11].

Image Quality Assessment and Gaze. The aim of Image Quality Assessment (IQA) is to automatically evaluate the visual quality of an image. Base on training process, IQA can be categorized as full-reference (FR) IQA, reduced reference (RR) IQA and no-reference (NR) IQA [62]. Specifically, FR-IQA [8, 25, 43, 63, 65] compare image naturalness based on reference image and distorted image; RR-IQA [3, 45, 46, 56] evaluate naturalness extracts partial information from reference image; NR-IQA

Datasets	Distortion	Image Source	Property
LIVE [50]	Artificial	Kodak Test Set	Quality
TID2008 [42]	Artificial	Kodak Test Set	Quality
CSIQ [29]	Authentic	Kodak Test Set	Quality
LIVE-itW [16]	Authentic	Daily Scenes	Quality
TID2013 [41]	Artificial	Kodak Test Set	Quality
KADID-10k [30]	Artificial	Social Media	Quality
KonIQ-10k [20]	Authentic	MultiMedia	Quality
PAN (Ours)	**Adversarial**	**Autonomous Driving**	**Naturalness**

Table 1. Comparisons between existing IQA datasets and PAN dataset. PAN differs from existing IQA database from type of distortion, image source and the property of assessed images.

[4, 35, 39, 52, 60, 61, 69, 76] directly evaluate visual naturalness without reference. In our work, we consider physical world attack as a novel type of distortion, and assess its naturalness in the pipeline of NR-IQA. We do not use FR-IQA since with noise in environment, an exact reference required by FR-IQA is hard to get.

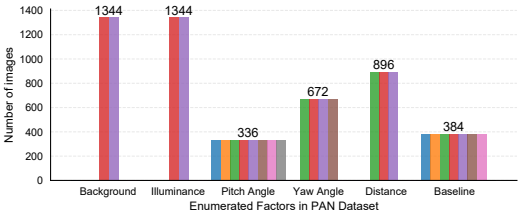
As an indicator of human attention, gaze was studied for gaining better IQA accuracy. There has been works that collect gaze fixations for existing IQA datasets [1, 12, 33, 38, 44], yet their datasets contain at most 160 images with gaze. In contrast, PAN contains all 2,688 images with accompanied gaze. To leverage collected gaze, one line of works use the collected gaze as a weighting metric [32, 34, 38, 68], while other line of works use human gaze as an additional quality indicator [64, 67]. In our work, we embed human gaze directly into IQA reasoning process.

3. Physical Attack Naturalness (PAN) Dataset

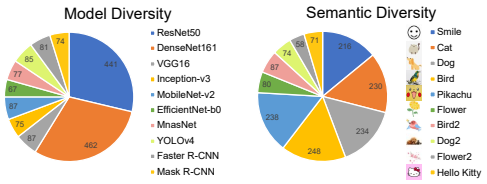
We define the task of evaluating naturalness of physical world adversarial attacks as a particular instance of No-Reference image quality assessment (NR-IQA) [①]. As illustrated in Table. 1, PAN differs from existing IQA database from three aspects: type of distortion, image source and the property of image assessed. As for distortion type, IQA databases mainly consider artificial (*e.g.*, Gaussian noise, JPEG compression) or authentic (*e.g.*, motion blur) distortions, while physical world attacks are maliciously generated patterns, unexplored in these two distortion types.

Thus, we contribute physical attack naturalness (PAN) dataset, the first dataset to understand naturalness of physical world attacks in autonomous driving. While other attack scenarios exists (*e.g.*, road sign [13, 31, 51], T-shirt [21, 54, 59], *etc*), the factors we study are commonly used in physical world attacks, making it possible for our methods and findings to be extended to other attack scenarios.

[①] See more details of IQA in related work section.



(a) Environment variations and baselines



(b) Diversity

Figure 2. Distribution of data variations. (a) number of images contained by each factor in PAN dataset. (b) distribution of diversity factors, including semantic and model diversity.

3.1. Construction Process

3.1.1 Image Generation

Evaluated baselines. We generate all test images using CARLA [10], an open source 3D virtual simulator based on Unreal Engine 4, which was widely used for autonomous driving [24] as well as physical world adversarial attacks [55]. As a first-step study, we use CARLA to first disentangle the impact of each variate on naturalness by controlling views, urban layouts, illuminations, *etc*in simulator. We discuss how PAN can be applied in real world in Section. 5.4. We evaluate naturalness on 7 distinct baselines, including 2 clean baselines: (1) *clean*, no perturbations exists. (2) *painting*, vehicle has benign car paintings, a common motivation for many physical world attacks [11, 22, 55]. We select 5 widely compared physical world attacks on autonomous driving with diverse naturalness enhancement methods and naturalness evaluation protocols, including CAMOU [70], MeshAdv [58], AdvCam [11], UPC [22] and DAS [55]. *We carefully reproduce attack results of these baselines [55], with detailed results in supplementary materials.*

Variations. We simulate images in PAN with possible real world variations. For *environmental* variations, following prior arts [55, 70], we consider 2 backgrounds, 2 illuminance, 8 pitch angles, 4 yaw angles and 3 distances for each baselines, resulting in 7 (baselines)$\times$2$\times$2$\times$8$\times$4$\times$3 = 2,688 images, with details of each enumerated factors in Fig. 2a. For *diversity* variations, we improve semantic diversity by constraining attack patterns to 10 natural images with different semantic meaning. Additionally, we improve model diversity by generating attack patterns on 7 classification models and 3 object detection models. Database distributions for semantic and model diversity are shown in Fig. 2b. The distribution is not balanced since semantic or model diversity is not supported on certain evaluated methods. *See*

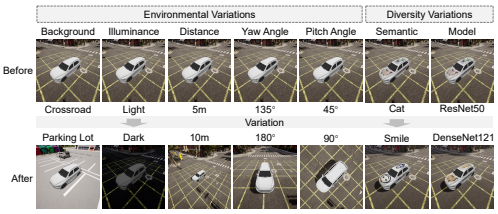


Figure 3. Environmental and diversity variations in PAN dataset.

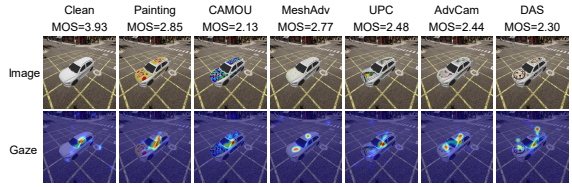


Figure 4. Illustration of data contained in PAN dataset. We provide raw image, corresponding human gaze and MOS score.

more details in supplementary materials. Examples of applying each variations to PAN are given in Fig. 3.

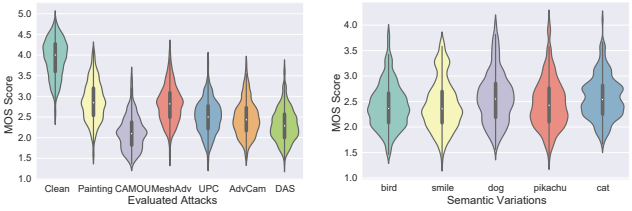
Data properties. For all images in PAN, we release their subjective naturalness ratings by Mean Opinion Score (MOS), calculated by averaging all human ratings [16], with rating distribution of each images given correspondingly. We also release the gaze saliency map $\mathcal{S}$, calculated by applying a Gaussian mask for all raw human fixations, following [34]. Exemplar images, corresponding human gaze and MOS score are illustrated in Fig. 4.

3.1.2 Human Assessment

Participants and apparatus. We recruit 126 participants (57 female, 69 male, age=22.2±3.3) from campus, all with normal (corrected) eyesight. Each participant is compensated $15. Images are displayed on a 16-inch screen with a resolution of 2560*1600 and an approximate viewing distance about 70cm. A Tobii Eye Tracker 5 (equipped in front of the screen) is adapted for eye gaze tracking. It records eye gaze points at about 60 GP/sec. A gaze calibration process is done before the experiment[①].

Experiment process. We adopt a single stimulus continuous procedure [40] and ask participants to evaluate the naturalness of image. For each image, participants first view it for 2.5 seconds, with eye tracker activated. The time is determined by our pilot study to ensure eye gaze coverage and prevent fatigue. Next, participants rate the image by a 5-point Absolute Category Rating (ACR) scale [20]. Each participants are asked to evaluate 320 images which are divided into 8 sessions. A warmup session is given at the beginning, with a 20 seconds' rest between two sessions. Participants take no more than 35 minutes to finish all experiments. We follow quality control process of [20], enabling each image to contain ratings and gaze of at least 10 subjects. *Due to the space limit, we defer more details of image generation, human assessment and quality control of PAN dataset to supplementary materials.*

[①]PAN does not contain Personally Identifiable Information (PII). Additional ethical concerns are discussed in supplementary materials.



(a) Naturalness of existing attacks (b) Impact of semantic factors

Figure 5. Visualization of factors that affect naturalness. Violin plot indicates MOS score distribution across all images.

3.2. Insights

We first provide an overview of PAN dataset, categorized by evaluated baselines. As shown in Fig. 5a, even for the most natural attack (*MeshAdv*), its MOS score is still much lower than *clean* baseline (2.77 vs 3.93). This suggests that, at least in autonomous driving scenario and using CARLA simulation environment, while *AdvCAM* and *DAS* claim their attacks to be more natural than others, they still remain far less natural than *clean* images. But what affects naturalness? How can we improve naturalness? Based on PAN, we find naturalness are disparately affected by contextual features, and are related to behavioral factors. We also offer pragmatic advice on improving naturalness below. *We defer tradeoff between attack capability and naturalness; the impact of environmental factors and diversity factors to supplementary materials. Results and analysis below are reported using proper statistical tests [57] (e.g., one-way ANOVA) and post-hoc pairwise comparisons (e.g., Tukey's Honest Significant Difference (HSD) test.) to analyze our PAN dataset. We report significant findings at $p < .05$.*

Insight ❶: *Naturalness is affected by contextual features, including semantic diversity and environmental variations; Naturalness can be improved by selecting proper contextual features.*

For diversity, as shown in Fig. 5b, we find semantic factors, *i.e.*, natural image used to constraining attacks (*i.e.*, method used by *UPC*, *AdvCAM* and *DAS*) have significant effect on naturalness ($p < .001$, One-way ANOVA)[②]. Based on this observation, we find replacing the natural image used by *UPC* (bird), *AdvCam* (pikachu) and *DAS* (smile)[③] by the most natural image (cat) improves their naturalness by 8.09%, 6.01% and 5.04%, respectively. We hypothesis the semantic relations between vehicle and natural image affects naturalness, which was verified by an additional user study ($p < .001$, Mann Whitney Test), *details*

[②]We do not find significant effect of model diversity on naturalness ($p = .717$, One-way ANOVA).

[③]Original images are not provided by *UPC* and *AdvCam*, so we replace images with similar appearance and semantics on internet.

deferred in supplementary materials.

Besides, almost all environmental factors, (*i.e.*, illumination, pitch/yaw angle, distance), except background, has significant effect on naturalness ($p < .001$ for all factors except $p = .588$ for background, One-way ANOVA). Post-hoc analysis shows that nearly all levels are significantly different ($p < .001$, post-hoc Tukey HSD test, variance uniformity satisfied with $p = .17$). Specifically, higher naturalness can be achieved at farther distance, pitch angle $0°$, yaw angle $90°$ and higher luminance. By simply changing environment condition, we find an improvement up to 34.73% on naturalness. This reminds defenders physical world attacks can be more stealthy in certain occasions.

Insight ❷: *Contextual features have disparate impact on naturalness of different attacks, which can lead to biased evaluation even under identical settings.*

Additionally, we find contextual features do not affect attacks equally (*i.e.*, some attacks look more natural under certain contextual features). This may lead to biased naturalness evaluation, even under same contextual features. For example, while *UPC* is overall more natural than *AdvCam* in PAN dataset ($p < .001$, independent samples t-test), at certain conditions (*e.g.*, yaw angle $135°$, $180°$, or distance of $10m$), *AdvCam* can be more natural than *UPC* ($p < .001$, independent samples t-test). This bias is not reported in any previous work of physical world attack.

While this inconsistency can explained by the interaction of perceptual characteristics of attacks and contextual features, the bias nonetheless pose a threat to subjective naturalness evaluation. To solve this problem, we suggest subsequent research to report their attack naturalness on multiple contextual features. *In supplementary materials, we also suggest a naturalness evaluation setting which is consistent with result in PAN and requires minimal number of testing.*

Insight ❸: *Naturalness is correlated with behavioral feature (i.e., human gaze). Manipulation of human gaze can be a feasible direction to improve naturalness.*

Besides contextual features, we find behavioral feature (*i.e.*, human gaze) correlates with naturalness: attacks are considered less natural if gaze are more *centralized* ($p < .05$, one-way ANOVA), or *focus more on vehicle* ($p < .001$, one-way ANOVA). Specifically, *centralize* measures how much human concentrates, calculated as the standard deviation of gaze saliency map, while *focused* measures how much human pay attention to vehicle, calculated by sum of dot product between gaze saliency map and vehicle area.

This correlation suggests a feasible direction to improve naturalness: optimizing attack patterns that guides human gaze to be less centralized, or focus less on vehicle. This is possible via the prior work of Gatys *et al.* [15], which tries to guide human gaze by optimized visual patterns. Additionally, we note that our finding shares similar motivation with *DAS*, which aims to improve naturalness by evading human attention. However, we do not find *DAS* triggers distinctive gaze behavior comparing with other attacks ($p = 0.967$ on average, Post-hoc Tukey HSD test).

4. Assess Naturalness by Dual Prior Alignment

While the procedures in PAN offers a feasible way to assess naturalness, collecting human ratings can be costly and laborious. In this section, we propose Dual Prior Alignment (DPA), a quality assessment algorithm to automatically evaluate the naturalness of physical world attacks.

4.1. Motivation

The goal of IQA is to design algorithms with objective naturalness predictions well correlated with human subjective ratings [35, 52, 60, 61, 76]. Thus, it is reasonable to assume that better modelling and imitating human behavior leads to better IQA result. With rich human behaviors offered in PAN dataset, we propose Dual Prior Alignment (DPA) network that aligns human behavior with model decisions. As shown in Fig. 6, DPA network consists of two modules, *i.e.*, rating prior alignment module and attentive prior alignment module, which enables DPA to align with human reasoning process and human attention process.

To align model behavior with human reasoning process, participants reveal their decisions contain uncertainty and are based on vague cognitive criterion. To reflect uncertainty in human ratings, we remodel IQA as a classification problem instead of regression, and align model output with the distribution of human ratings. To capture hidden criterion of human, we use a prototype vector which learns the hidden knowledge of each level during training. *To align model attention with human attention*, as shown in Fig. 7, we find existing methods cheat by exploiting spurious correlations between naturalness ratings and irrelevant areas. Intuitively, such biased model have weak generalization capability on unseen test images. To mitigate such bias, we design an IQA-specific visual grounding criterion that aligns model attention with human gaze.

4.2. Rating Prior Alignment

To understand how human reasons naturalness, we give an interview to participants after experiment. Participant 21 (P21) and P47 noted their ratings contains uncertainty when they find both rating levels are appropriate. P6, P40 and P47 also noted that they developed a vague criterion, or believes, of each rating levels. Based on such criterion, they select ratings that fits best with current image.

Based on participants' feedbacks, we explicitly mimic the decision process of human. To represent the hidden judgement rules of human, we initiate a *prototype* vector z_ℓ for each rating levels $\ell \in \{1, 2, 3, 4, 5\}$, with values updated during training. For image x and backbone DNN f_θ parameterized by θ, we assume the representations $f_\theta(x)$

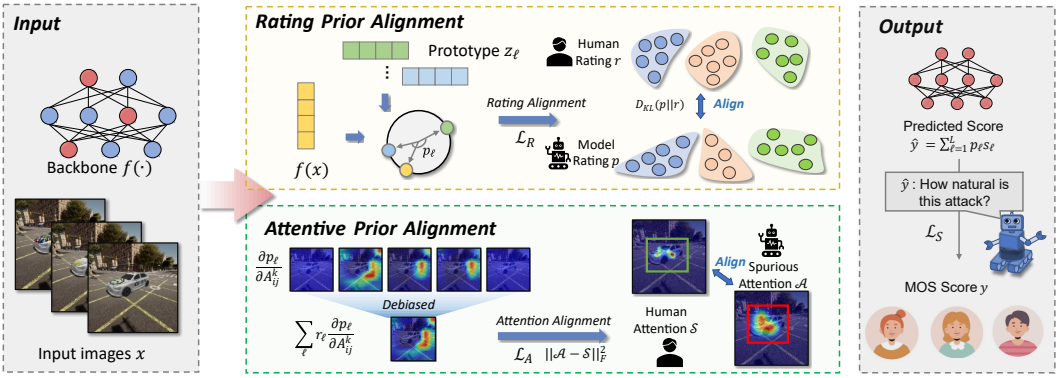


Figure 6. Framework of DPA. Rating prior alignment mimics the uncertainty and hidden desiderata in human naturalness rating process. Attentive prior alignment corrects spurious correlations in models by aligning model attention with human gaze.

captures the relevant information of x for naturalness assessment. To represent decision uncertainty of human and avoid overfitting to a continuous value, we model NR-IQA as a classification problem instead of regression. Specifically, the likelihood p_ℓ that image x belongs to each levels is calculated by the cosine similarity between image representations $f_\theta(x)$ and the prototype of each levels z_ℓ, followed by a softmax function:

$$p_\ell(x,z) = \frac{\exp(f_\theta(x)\cdot z_\ell/||f_\theta(x)||\cdot||z_\ell||)}{\sum_{j=1}^{L}\exp(f_\theta(x)\cdot z_j/||f_\theta(x)||\cdot||z_j||}, \quad (1)$$

where L is level of ratings, set to 5 in our experiment.

With likelihood p_ℓ calculated, we propose rating prior alignment (RPA) loss $\mathcal{L}_R$ to address human rating uncertainty by aligning p_ℓ with human rating distributions r_ℓ:

$$\mathcal{L}_R = KL(p(x,z)||r) = \sum_{\ell=1}^{L} p_\ell(x,z)\log\left(\frac{p_\ell(x,z)}{r_\ell}\right), \quad (2)$$

where KL is the Kullback-Leibler divergence.

4.3. Attentive Prior Alignment

While training models to fit subjective MOS score y can yield low error, as shown in Fig. 7, models can cheat by exploiting spurious correlations between background minutiae and predictions. To mitigate this bias, we leverage gaze signal as a guidance to correct attention of IQA model such that model align its intrinsic attention with human gaze.

To capture model attention, visual attention techniques [6, 47, 75] explain and visualize the attention of DNNs by back-propagating to neurons of the last convolutional layer:

$$\mathcal{A}(x,\hat{y}) = \frac{1}{Z}ReLU\left(\sum_{i,j,k}\frac{\partial\hat{y}}{\partial A_{ij}^k}A_{ij}^k\right), \quad (3)$$

where $\mathcal{A}$ is the attention map, Z is a normalizing constant, A_{ij}^k denotes the value in position (i,j) of feature map k.

However, Eqn.3 is biased to emphasize higher ratings:

$$\frac{\partial\hat{y}}{\partial A_{ij}^k} = \sum_{\ell=1}^{L}\frac{\partial\hat{y}}{\partial p_\ell}\cdot\frac{\partial p_\ell}{\partial A_{ij}^k} = \sum_{\ell=1}^{L} s_\ell\cdot\frac{\partial p_\ell}{\partial A_{ij}^k}. \quad (4)$$

As a result, naively applying Grad-CAM bias the backward gradient $\partial p_\ell/\partial A_{ij}^k$ by s_ℓ, the score of current rating level. To correct this bias, we modify the backpropagation step of Grad-CAM as a weighted average of the gradients back-propagated from rating likelihood $\partial p_\ell/\partial A_{ij}^k$, using p_ℓ:

$$\mathcal{A}(x,p) = \frac{1}{Z}ReLU\left(\sum_{i,j,k}\sum_{\ell} p_\ell\cdot\frac{\partial p_\ell}{\partial A_{ij}^k}A_{ij}^k\right). \quad (5)$$

Finally, we propose attentive prior alignment loss $\mathcal{L}_A$ to align model attention $\mathcal{A}$ with human gaze $\mathcal{S}$:

$$\mathcal{L}_A = ||\mathcal{A}(x,p)-\mathcal{S}||_F^2. \quad (6)$$

4.4. Overall Training

We have discussed how to align human rating prior by $\mathcal{L}_R$ and human attentive prior by $\mathcal{L}_A$. To get the final IQA result, the predicted MOS score $\hat{y}$ was calculated by the expectations over scores for each levels s_ℓ, *i.e.*, $\hat{y} = \sum_{\ell=1}^{L} p_\ell(x,z)s_\ell$. We also add a standard mean square error loss $\mathcal{L}_S = \frac{1}{N}\sum_{n=1}^{N}||\hat{y}_n - y_n||_2^2$ between $\hat{y}_n$ and ground truth subjective MOS ratings y_n, where n is the image index in a minibatch of size N. Overall, DPA learns to assess image naturalness by jointly optimizing $\mathcal{L}_R$, $\mathcal{L}_A$ and $\mathcal{L}_S$:

$$\min_{\theta,z}\mathcal{L}_S + \lambda\mathcal{L}_R + \gamma\mathcal{L}_A, \quad (7)$$

where λ and γ are hyperparameters to control the strength of $\mathcal{L}_R$ and $\mathcal{L}_A$, respectively. θ is the parameters of the backbone network, $z = \{z_\ell\}$ are the set of prototypes for each rating levels. *Overall training algorithm of DPA can be find in supplementary materials.*

Category	Method	SROCC (↑)	PLCC (↑)	S_C (↑)
FR-IQA	PSNR	0.3560	0.3685	-
	SSIM	0.4573	0.3968	-
	LPIPS	0.1056	0.1395	0.0583
	E-LPIPS	0.3990	0.3694	0.0727
Others	GIQA(KNN)	0.1382	0.1133	-
	GIQA(GMM)	0.1537	0.1392	-
NR-IQA	BRISQUE	0.1029	0.0494	-
	ResNet50	0.1149	0.1682	0.1692
	WaDIQaM	-0.0704	-0.1078	0.1821
	RankIQA	0.1809	0.1992	0.0095
	DBCNN	0.1409	0.1167	0.0876
	HyperIQA	0.1639	0.1285	0.2188
	Paq2Piq	0.0320	0.0504	0.2791
	MANIQA	0.2741	0.2717	0.0810
NR-IQA	DPA+PAN (Ours)	**0.7501**	**0.7727**	**0.7178**

Table 2. Validating necessity of PAN dataset. All baselines are trained without using PAN, with DPA trained on PAN.

Category	Method	SROCC (↑)	PLCC (↑)	S_C (↑)
FR-IQA	PSNR	0.3560	0.3685	-
	SSIM	0.4573	0.3968	-
	LPIPS	0.0994	0.1114	0.0089
	E-LPIPS	0.4082	0.4064	0.0136
Others	GIQA(KNN)	0.1428	0.1132	-
	GIQA(GMM)	0.0838	-0.0366	-
NR-IQA	BRISQUE	0.4753	0.3777	-
	ResNet50	0.6916	0.7453	0.2066
	WaDIQaM	0.6998	0.6841	0.2130
	RankIQA	0.7227	0.7564	0.1134
	DBCNN	0.6800	0.6621	0.3947
	HyperIQA	0.7253	0.7265	0.1955
	Paq2Piq	0.6044	0.6089	0.2003
	MANIQA	0.7129	0.7331	0.0861
NR-IQA	DPA (Ours)	**0.7501**	**0.7727**	**0.7178**

Table 3. Validating the effectiveness of DPA using PAN dataset. DPA outperform other baselines by aligning with human rating prior and human attention prior.

5. Experiments

In this section, we use experiments to verify: (1) do we need PAN dataset? (2) can DPA better assess naturalness? (3) can DPA generalize in real world scenarios?

5.1. Experimental Settings

5.1.1 Datasets and Baselines

We conduct experiments on our proposed PAN dataset. To evaluate the effectiveness of image quality assessment, we compare with 13 state-of-the-art methods, including four widely used FR-IQA methods: PSNR, SSIM, LPIPS [65] and E-LPIPS [25]; one IQA method for GAN: GIQA [18]; eight NR-IQA methods: including vanilla ResNet50 [19], BRISQUE [39], WaDIQaM [4], RankIQA [35], DBCNN [69], HyperIQA [52], Paq2Piq [61] and MANIQA [60].

5.1.2 Implementation Details and Evaluation Metrics

Two evaluation metrics are selected to compare the performance of different IQA algorithms: Spearman Rank Order Correlation Coefficient (SROCC) and Pearson's Linear correlation coefficient (PLCC). We also measure attention alignment by cosine similarity S_C between model attention and human gaze. Results are averaged for all baselines (distortions). For implementations, we use a ResNet50 backbone for our DPA method, with hyperparameters λ and γ empirically set to 8.0 and 3.0, respectively. For fair comparison, we train all methods for 20 epochs using an Adam optimizer [26] with learning rate 3×10^{-5}. *See additional experiment settings in supplementary materials.*

5.2. Do We Need PAN Dataset?

In this section, we answer the following question: can existing IQA database solve the problem of naturalness assessment, so PAN is not needed? Specifically, we test the result of existing methods on PAN with its released models, compared with training our DPA directly on PAN. For methods without released model, we train them on TID2013 dataset [41] using their default conditions. From results in Table. 2, we can draw several conclusions as follows:

(1) Collecting our PAN dataset is vital for assessing naturalness of physical world attacks. Our DPA+PAN achieves **0.2928 (+64.03%)** higher SROCC and **0.3759 (+94.73%)** higher PLCC than SSIM, the best existing method. This clearly shows existing methods and datasets are insufficient to evaluate the naturalness of physical world attacks.

(2) Since existing methods are ineffective, we do not recommend using SSIM and LPIPS as naturalness indicators in physical world, as opposed to digital world [7, 28]. However, if our DPA is not applicable, SSIM provides a best estimate (**0.0583** higher in SROCC and **0.0274** higher in PLCC than second best baseline E-LPIPS).

5.3. Can DPA Better Assess Naturalness?

Next, based on PAN dataset, we ask the question: with human priors incorporated, can DPA better assess naturalness of physical world attacks? For non-learning methods PSNR and SSIM, we evaluate them on all PAN dataset and the result is thus identical to Table. 2. From results listed in Table. 3, we can draw several conclusions as follows:

(1) Aligning the behavior of DNNs with human improves naturalness assessment. Trained on PAN, our DPA outperform the best baseline by **0.0248 (+3.42%)** in SROCC and **0.0163 (+2.15%)** in PLCC.

(2) Using S_C as a measure of alignment between model and human attention, under attentive prior alignment loss, DPA gains 81.86% higher alignment compared with the best baselines, which provides significantly better alignment between model attention and human gaze. We also illustrate model attentions and corresponding human gaze in Fig. 7: while almost all baselines rely on spurious areas

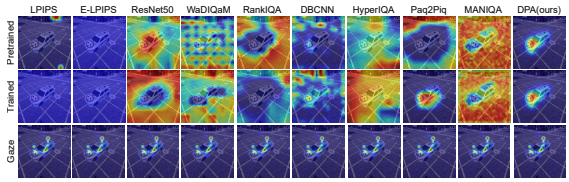


Figure 7. Grad-CAM visualization of DPA and baselines.

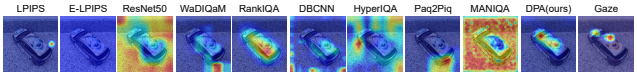


Figure 8. Grad-CAM visualization of real world images.

for prediction, DPA base its decision on correct areas.

(3) The ineffectiveness of FR-IQA and GIQA methods could be explained by adversarial feature [14,23]: adversarial attacks are effective because they are not just noise, but meaningful features from other domains for DNNs. While FR-IQA and GIQA methods keep backbone parameters unchanged, the extracted features might be polluted by adversarial features, thus unable to give reliable results.

5.4. Can DPA Generalize?

Finally, we ask the question: can DPA generalize to unseen real world images? To verify this, we manually collected 504 real world images, called *PAN-phys* with 8 pitch angles, 3 yaw angles and 3 backgrounds. *See details of this dataset in supplementary materials.* Next, we collect human rating and gaze signal using the same approach as PAN dataset. Finally, we fix the parameters of all methods and evaluate their result on *PAN-phys*. From results listed in Table. 4, we can draw several conclusions as follows:

(1) Through aligning model behaviors with human, DPA also achieves stronger generalization capability when evaluating images drawn from unseen real world scenario, outperforming the best baseline by **0.0332 (+8.40%)** in SROCC and **0.0236 (+5.34%)** in PLCC.

(2) Our DPA is able to align its attention with human attention even under unseen images, achieving 12.73% higher S_C than best performing baseline. As shown in Fig. 8, the attention area of DPA keeps aligned with human gaze during generalization, while most baselines yields predictions on spurious correlations.

(3) The domain gap between real world and simulation environment harms the naturalness assessment accuracy, calling an urgent need to further improve naturalness assessment methods via domain generalization.

5.5. Ablation Studies

In this section, we conduct ablation studies to verify the effect of different loss terms, namely rating prior alignment loss $\mathcal{L}_R$ and attentive prior alignment loss $\mathcal{L}_A$. We argue that $\mathcal{L}_R$ and $\mathcal{L}_A$ jointly improves alignment with human ratings, while $\mathcal{L}_A$ also improves alignment with human gaze. Shown in Table. 5, $\mathcal{L}_A$ and $\mathcal{L}_R$ contributes to SROCC and PLCC individually, while combining them shows further improvement. For attention alignment S_C, while $\mathcal{L}_A$ significantly improves alignment with gaze, we surprisingly find aligning human rating prior by $\mathcal{L}_R$ also partly enhance S_C. Additionally, the effect of $\mathcal{L}_R$ on S_C is enhanced with the presence of $\mathcal{L}_A$ (+0.0314 w/o $\mathcal{L}_A$, +0.1071 w/ $\mathcal{L}_A$). We hypothesize that aligning human behavior from one aspect might have an synergy effect on another aspect. We left detailed study of this phenomenon for future work.

Category	Method	SROCC (↑)	PLCC (↑)	$\mathbf{S_C}$ (↑)
FR-IQA	PSNR	0.3163	0.3009	-
	SSIM	0.3594	0.3558	-
	LPIPS	-0.2659	-0.3540	0.0163
	E-LPIPS	-0.3778	-0.3589	0.1658
Others	GIQA(KNN)	0.0075	0.0275	-
	GIQA(GMM)	0.0747	0.0809	-
NR-IQA	BRISQUE	0.0261	0.0245	-
	ResNet50	0.2874	0.3282	0.1935
	WaDIQaM	-0.1362	-0.1375	0.0329
	RankIQA	-0.1313	-0.1368	0.2942
	DBCNN	0.3907	0.4144	0.3028
	HyperIQA	0.3951	0.4416	0.3645
	Paq2Piq	0.3752	0.3905	0.2244
	MANIQA	0.3673	0.3839	0.2502
NR-IQA	DPA (Ours)	**0.4283**	**0.4652**	**0.4109**

Table 4. Generalization results of DPA and other baselines on real world image dataset, *PAN-phys*.

Method	SROCC (↑)	PLCC (↑)	$\mathbf{S_C}$ (↑)
ResNet50	0.6916	0.7453	0.2066
$\mathcal{L}_A$	0.7121	0.7586	0.6107
$\mathcal{L}_R$	0.7154	0.7673	0.2380
$\mathcal{L}_R + \mathcal{L}_A$	**0.7501**	**0.7727**	**0.7178**

Table 5. Ablation study for different loss terms when evaluating human ratings. All terms in DPA achieved their desired goal.

6. Conclusion

In this paper, we study to evaluate naturalness of physical world adversarial attacks. Specifically, we contribute PAN, the first dataset to benchmark and evaluate naturalness of physical world attacks. Besides, we propose DPA, an automatic naturalness assessment algorithm which offers higher alignment with human ratings and better generalization. Our work fertilizes community by (1) contributing PAN, which enables research on evaluating naturalness of physical world attacks by human rating and high-quality, large scale gaze signals; (2) encouraging new research on natural physical world attacks via analysis of contextual and behavioral features; (3) encouraging new research to design better IQA algorithms for physical world attacks.

Acknowledgement. This work was supported by the National Key Research and Development Plan of China (2021ZD0110601), the National Natural Science Foundation of China (62022009, 62132010 and 62206009), and the State Key Laboratory of Software Development Environment.

References

[1] H. Alers, J. A. Redi, H. Liu, and I. Heynderickx. Studying the effect of optimizing image quality in salient regions at the expense of background content. *Journal of Electronic Imaging*, 22(4):043012, 2013. 3

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

[3] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik. Speed-qa: Spatial efficient entropic differencing for image and video quality. *IEEE signal processing letters*, 24(9):1333–1337, 2017. 2

[4] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017. 2, 7

[5] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 1, 2

[6] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 6

[7] 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. 2, 7

[8] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[9] 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. 1, 2

[10] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 3

[11] 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. 1, 2, 3

[12] U. Engelke, A. Maeder, and H.-J. Zepernick. Visual attention modelling for subjective image quality databases. In *2009 IEEE International Workshop on Multimedia Signal Processing*, pages 1–6. IEEE, 2009. 3

[13] 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. 1, 2, 3

[14] L. Fowl, M. Goldblum, P.-y. Chiang, J. Geiping, W. Czaja, and T. Goldstein. Adversarial examples make strong poisons. *arXiv preprint arXiv:2106.10807*, 2021. 8

[15] L. A. Gatys, M. Kümmerer, T. S. Wallis, and M. Bethge. Guiding human gaze with convolutional neural networks. *arXiv preprint arXiv:1712.06492*, 2017. 5

[16] 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. 3, 4

[17] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2

[18] S. Gu, J. Bao, D. Chen, and F. Wen. Giqa: Generated image quality assessment. In *European conference on computer vision*, pages 369–385. Springer, 2020. 7

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[20] 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. 3, 4

[21] 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*, pages 13307–13316, 2022. 2, 3

[22] 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*, pages 720–729, 2020. 1, 2, 3

[23] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019. 8

[24] J. Janai, F. Güney, A. Behl, A. Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020. 2, 3

[25] M. Kettunen, E. Härkönen, and J. Lehtinen. E-lpips: robust perceptual image similarity via random transformation ensembles. *arXiv preprint arXiv:1906.03973*, 2019. 2, 7

[26] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[27] 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. 1

[28] C. Laidlaw, S. Singla, and S. Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *arXiv preprint arXiv:2006.12655*, 2020. 2, 7

[29] E. C. Larson and D. M. Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006, 2010. 3

[30] H. Lin, V. Hosu, and D. Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 3

[31] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao. Perceptual-sensitive gan for generating adversarial patches. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1028–1035, 2019. 1, 2, 3

[32] H. Liu, U. Engelke, J. Wang, P. Le Callet, and I. Heynderickx. How does image content affect the added value of visual attention in objective image quality assessment? *IEEE Signal Processing Letters*, 20(4):355–358, 2013. 3

[33] H. Liu and I. Heynderickx. Studying the added value of visual attention in objective image quality metrics based on eye movement data. In *2009 16th IEEE international conference on image processing (ICIP)*, pages 3097–3100. IEEE, 2009. 3

[34] 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. 3, 4

[35] X. Liu, J. Van De Weijer, and A. D. Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*, pages 1040–1049, 2017. 2, 5, 7

[36] C. Luo, Q. Lin, W. Xie, B. Wu, J. Xie, and L. Shen. Frequency-driven imperceptible adversarial attack on semantic similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15315–15324, 2022. 2

[37] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2

[38] X. Min, G. Zhai, Z. Gao, and K. Gu. Visual attention data for image quality assessment databases. In *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 894–897. IEEE, 2014. 3

[39] A. Mittal, A. K. Moorthy, and A. C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 2, 7

[40] M. H. Pinson and S. Wolf. A new standardized method for objectively measuring video quality. *IEEE Transactions on broadcasting*, 50(3):312–322, 2004. 4

[41] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30:57–77, 2015. 3, 7

[42] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10(4):30–45, 2009. 3

[43] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen. Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2018. 2

[44] J. Redi, H. Liu, R. Zunino, and I. Heynderickx. Interactions of visual attention and quality perception. In *Human Vision and Electronic Imaging XVI*, volume 7865, pages 267–277. SPIE, 2011. 3

[45] J. A. Redi, P. Gastaldo, I. Heynderickx, and R. Zunino. Color distribution information for the reduced-reference assessment of perceived image quality. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(12):1757–1769, 2010. 2

[46] A. Rehman and Z. Wang. Reduced-reference image quality assessment by structural similarity estimation. *IEEE transactions on image processing*, 21(8):3378–3389, 2012. 2

[47] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6

[48] 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. 1, 2

[49] 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. 1, 2

[50] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Processing*, 15(11):3440–3451, 2006. 3

[51] 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. 1, 2, 3

[52] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020. 2, 5, 7

[53] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2

[54] S. Thys, W. Van Ranst, and T. Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2, 3

[55] 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*, pages 8565–8574, 2021. 1, 2, 3

[56] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik. Quality-aware images. *IEEE transactions on image processing*, 15(6):1680–1689, 2006. 2

[57] J. O. Wobbrock. Practical statistics for human-computer interaction: An independent study combining statistics theory and tool know-how. In *Annual Workshop of the HCI Consortium (HCIC'11)*, 2011. 4

[58] 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*, pages 6898–6907, 2019. 2, 3

[59] 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. 1, 2, 3

[60] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1191–1200, 2022. 2, 5, 7

[61] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3585, 2020. 2, 5, 7

[62] G. Zhai and X. Min. Perceptual image quality assessment: a survey. *Science China Information Sciences*, 63(11):1–52, 2020. 2

[63] L. Zhang, Y. Shen, and H. Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10):4270–4281, 2014. 2

[64] L. Zhang, Y. Shen, and H. Li. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10):4270–4281, 2014. 3

[65] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 2, 7

[66] R. Zhang, A. Saran, B. Liu, Y. Zhu, S. Guo, S. Niekum, D. Ballard, and M. Hayhoe. Human gaze assisted artificial intelligence: A review. In *IJCAI: Proceedings of the Conference*, volume 2020, page 4951. NIH Public Access, 2020. 2

[67] W. Zhang and H. Liu. Learning picture quality from visual distraction: Psychophysical studies and computational models. *Neurocomputing*, 247:183–191, 2017. 3

[68] W. Zhang and H. Liu. Toward a reliable collection of eye-tracking data for image quality research: challenges, solutions, and applications. *IEEE Transactions on Image Processing*, 26(5):2424–2437, 2017. 3

[69] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018. 2, 7

[70] 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*, 2018. 1, 2, 3

[71] Z. Zhao, Z. Liu, and M. Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1039–1048, 2020. 2

[72] Z. Zhao, S. Xu, C. Zhang, J. Liu, J. Zhang, and P. Li. Didfuse: Deep image decomposition for infrared and visible image fusion. In *IJCAI*, pages 970–976. ijcai.org, 2020. 1

[73] Z. Zhao, S. Xu, J. Zhang, C. Liang, C. Zhang, and J. Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Trans. Circuits Syst. Video Technol.*, 32(3):1186–1196, 2022. 1

[74] Z. Zhao, J. Zhang, S. Xu, Z. Lin, and H. Pfister. Discrete cosine transform network for guided depth map super-resolution. In *CVPR*, pages 5687–5697. IEEE, 2022. 1

[75] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 6

[76] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi. Metaiqa: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14143–14152, 2020. 2, 5