Intriguing Properties of Diffusion Models: An Empirical Study of the Natural Attack Capability in Text-to-Image Generative Models

Takami Sato†[1], Justin Yue†[1], Nanze Chen†[2], Ningfei Wang[1], Qi Alfred Chen[1]
[1]University of California, Irvine
[2]University of Cambridge
{takamis, jpyue, ningfei.wang, alfchen}@uci.edu, nc630@cam.ac.uk

Abstract

Denoising probabilistic diffusion models have shown breakthrough performance to generate more photo-realistic images or human-level illustrations than the prior models such as GANs. This high image-generation capability has stimulated the creation of many downstream applications in various areas. However, we find that this technology is actually a double-edged sword: we identify a new type of attack, called the Natural Denoising Diffusion (NDD) attack based on the finding that state-of-the-art deep neural network (DNN) models still hold their prediction even if we intentionally remove their robust features, which are essential to the human visual system (HVS), through text prompts. The NDD attack shows a significantly high capability to generate low-cost, model-agnostic, and transferable adversarial attacks by exploiting the natural attack capability in diffusion models. To systematically evaluate the risk of the NDD attack, we perform a large-scale empirical study with our newly created dataset, the Natural Denoising Diffusion Attack (NDDA) dataset. We evaluate the natural attack capability by answering 6 research questions. Through a user study, we find that it can achieve an 88% detection rate while being stealthy to 93% of human subjects; we also find that the non-robust features embedded by diffusion models contribute to the natural attack capability. To confirm the model-agnostic and transferable attack capability, we perform the NDD attack against the Tesla Model 3 and find that 73% of the physically printed attacks can be detected as stop signs. Our hope is that the study and dataset can help our community be aware of the risks in diffusion models and facilitate further research toward robust DNN models.

1. Introduction

Denoising diffusion probabilistic models (DDPM) [29], or simply diffusion models, have shown breakthrough performance in image generation. Once the DDPM demonstrates the generation capability of photo-realistic images and human-level illustrations, numerous diffusion models, such as DALL-E 2 [44], Stable Diffusion [47], and Firefly [5], are actively released and widely available through APIs or as open-source models. This technical breakthrough has facilitated many applications in various fields such as the arts [57], medicine [34], and autonomous driving [32]. While diffusion models have brought significant benefits to these areas, recent studies have also raised concerns regarding the new security and privacy risks introduced by diffusion models. Chen et al. [15] show that the diffusion models can generate more transferable and imperceptible adversarial attacks. Chen et al. [17] generate more natural and stealthy perturbations with guidance from diffusion models. Carlini et al. [14] demonstrate that the diffusion models memorize training images and can emit them.

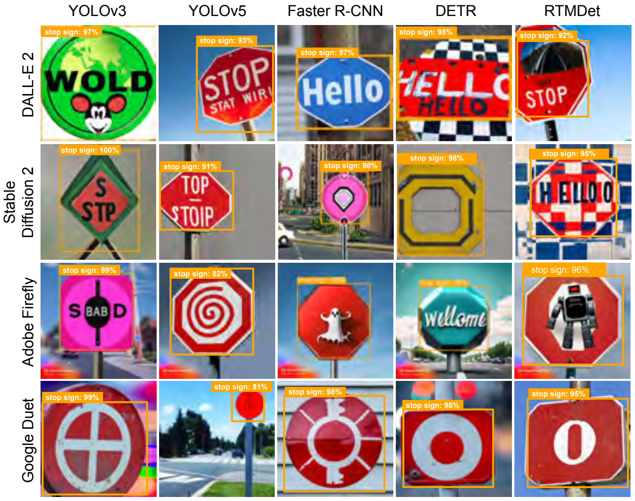


Figure 1. Examples of the natural attack capability in diffusion models (row). The images are generated with prompts that intentionally remove essential features to humans while keeping "stop sign" in the prompt. Even without these essential features, object detectors (column) still detect these objects with high scores.

These recent studies motivate us to further investigate the security risks of diffusion models. In this work, we discover simple but intriguing properties of the images generated by

[†]denotes co-first authors

text-to-image diffusion models, in which text prompts guide the image generation process via a contrastive image-text supervision model, such as OpenAI CLIP [43]. Fig. 1 shows representative examples that motivate this work. We generate the stop sign images using state-of-the-art diffusion models with text prompts that intentionally break the fundamental properties that humans use to identify stop signs (e.g., red color and octagonal shape) in the human visual system (HVS) [23, 26, 31], while the text prompt still contains the object name "stop sign".

As shown, the diffusion models faithfully follow our instructions and generate images that should not be recognized as stop signs since legitimate stop signs should not be blue or rectangular. However, many state-of-the-art Deep Neural Network (DNN)-based object detectors still recognize these examples as stop signs with surprisingly high confidence. These results suggest that these object detectors can be highly affected by imperceptible features embedded by diffusion models. We recognize this phenomenon as a new type of adversarial attack because the fundamental properties of the target object should be removed by maliciously designed text prompts. We name it the Natural Denoising Diffusion (NDD) attack and pursue the following question in this study:

Do images generated by diffusion models have natural attack capability against DNN models?

The rest of this paper is structured to validate the question. We first construct our dataset, named the Natural Diffusion Denoising Attack (NDDA) dataset, to systematically understand the natural attack capability in diffusion models (§3.1). We use 3 state-of-the-art diffusion models to collect the images with and without robust features that play essential roles in the HVS. Following prior works [23, 26], we define 4 robust features: shape, color, text, and pattern.

Secondly, we conduct an attack capability analysis of the NDD attack against state-of-the-art object detection models (§3.2) and image classification models (§3.3) on the NDDA dataset. We find that object detectors and classifiers typically maintain their detection, even though the text prompt guides them to remove robust features entirely or partially. For example, 32% of the generated stop signs are still detected as stop signs even though all robust features are guided to be removed. This result means that text-to-image diffusion models embed intriguing features that are imperceptible to humans but generalizable to DNN models if the subject word (e.g., stop sign) is in the prompt. We confirm that all diffusion models we evaluate have the natural attack capability to enable the NDD attack.

Third, we conduct a large-scale empirical study to further evaluate the validity and quantify the natural attack capability in diffusion models by answering 6 novel research questions (§4). We conducted a user study to evaluate the stealthiness of the NDD attacks because valid adversarial attacks need to be not only effective against the DNN models but also stealthy against humans. For example, humans will not be fooled if the robust features are not correctly removed since it should remain a legitimate stop sign. As a result, we identify the high stealthiness of the NDD attacks: The stop sign images generated by altering their "STOP" text have 88% detection rate against object detectors while 93% of humans do not regard it as a stop sign. Furthermore, we evaluate the impact of the non-robust features that are predictive but incomprehensible to humans on the natural attack capability in diffusion models with an analysis inspired by Ilyas et al. [31], which proposes a methodology to train "robustified" classifiers against the non-robust features. By comparing the robustified and normal classifiers, we illustrate that the non-robust features play a meaningful role in the natural attack capability in diffusion models.

Finally, we discuss our findings and the limitations (§5). In summary, the contributions of this work are as follows:

- We discover a new security threat, the NDD attack, which exploits the natural attack capability of the diffusion model to generate model-agnostic and transferable adversarial attacks via simple text prompts that are designed to remove robust features.
- We construct a new large-scale dataset, named the NDDA dataset, to systematically evaluate the natural attack capability in diffusion models. We cover all 4 robust features essential to the HVS: shape, color, text, and pattern.
- We performed a large-scale empirical study to systematically evaluate the natural attack capability by answering 6 research questions. The NDD attack can achieve an attack success rate of 88%, while being stealthy for 93% of human subjects in the stop sign case. We also find that the sensitivity to the non-robust features has a high correlation with the natural attack capability.
- We confirm the model-agnostic and transferable attack capability of the NDD attack on a Tesla Model 3, which identifies 8 out of 11 (73%) printed attacks as stop signs.

Dataset release. NDDA dataset is on the our website [4].

2. Related Work

2.1. Denoising Diffusion Model

DDPM [29] is a generative model that exploits the intuition behind nonequilibrium thermodynamics. Training a diffusion model involves both forward and reverse diffusion processes. In the forward process, the model perturbs a clean image with Gaussian noise. In the reverse process, the diffusion model learns to remove this noise for the same number of time steps. In short, diffusion models are image denoisers that learn to reconstruct the original images from noisy images. This procedure is simple but shows remarkable performance in producing high-quality images. Dhariwal et al. [21] shows that diffusion models achieve much higher quality metrics, such as FID, inception score, and precision, than prior works such as GANs [9].

The text-to-image diffusion model [44, 47, 48] is a variant of diffusion models that can flexibly control the output image via text prompts. Due to its easy usability, major diffusion models (e.g., DALL-E 2, Stable Diffusion, and FireFly) adopt this text-based guidance. To inform the text information in the generation process, contrastive image-text supervision models, such as CLIP [43] and OpenCLIP [30], are integrated into their training and inference processes.

2.2. Adversarial Attacks

DNN models are known to be generally vulnerable to adversarial attacks [10, 22, 24, 39, 52, 58], which can alter the predictions of DNN models by adding small changes that are not noticeable to humans. The early-stage works [24, 52] originally use a subtle imperceptible perturbation on the entire input image as their attack vector. Recent research has identified that adversarial attacks can be achieved by broader attack vectors that are any stealthy changes to the human perception such as putting small patches [49, 50, 54, 56] or placing stickers on the target [22]. Natural adversarial examples [28] demonstrate that even clean natural images can be used as adversarial attacks. To this extent, the NDD attack is similar to the natural adversarial examples, but the natural adversarial examples do not have any guarantees that can generate an attack against targeted scenarios (e.g., stop sign) as they just find out-of-distribution samples in the existing images. Furthermore, we find that the non-robust features [31] play a large role in the NDD attack (RQ4 in §4). Ilyas et al. [31] report that the adversarial attacks are not caused by a bug in DNNs but instead by non-robust features that are predictive but incomprehensible to humans. We thus consider that the NDD attack is enabled by similar root causes as the traditional adversarial attacks, i.e., enabled by non-robust features, rather than by out-of-distribution samples.

3. Attack Capability Analysis

To systematically evaluate the natural attack capability of the diffusion models, we first construct a new large-scale dataset, called the Natural Denoising Diffusion Attack (NDDA) dataset. With the NDDA dataset, we then confirm the effectiveness of the NDD attack against state-of-the-art object detectors and image classifiers.

3.1. NDDA Dataset Design

The design goal for the NDDA dataset is to systematically collect images both with and without robust features upon which human perception relies. For this sake, the major challenge is to identify what kinds of robust features are essential in our human visual system (HVS) for object recognition. Although the complete mechanism of the HVS is not fully understood, prior works [23, 26] identify that shape, texture, and color are the most important features for the HVS to identify objects. Therefore, in this study, we follow the prior works and define them as robust features for object recognition. To further explore the motivated examples in Fig. 1, we decompose the texture into text and pattern because text has a special meaning for human perception. For example, people may not consider a sign to be a stop sign if it does not have the exact text “STOP” on it.

Table 1 lists the templates of text prompts and examples for the “stop sign” label to remove or alter the 4 robust features. In this case, we consider 16 different combinations with and without robust features and generate 50 images for each combination. The 3 object classes are selected from the classes of the COCO dataset [37]: a stop sign for an artificial sign, a fire hydrant for an artificial object, and a horse for a natural object. We select these 3 classes because of their relatively higher detection rates than others in our preliminary experiments on generated images. We adopt the COCO’s classes to make the experiments easier as we can utilize many existing pretrained models on the COCO dataset. Fig. 2 shows an overview of our datasets. More details are in the supplementary materials.

3.2. Attack Capability against Object Detectors

We evaluate the attack capability of the NDD attack against object detectors with the NDDA dataset to validate the generality of the motivated examples shown in Fig. 1.

Experimental setup. We obtain the inference results of all images in the NDDA dataset with 5 popular object detectors: YOLOv3 [45], YOLOv5 [33], DETR [11], Faster R-CNN [46], and RTMDet [38]. For YOLOv5, we use their official pretrained model; For the others, we use the pretrained models in MMDetections [16]. All models are trained on the COCO dataset [37]. We use 0.5 as the confidence threshold for all models.

Results. Table 2 shows the detection results of the stop sign images in the Natural Diffusion Attack dataset generated by 3 diffusion models. The detection rate is calculated by whether one or more stop signs are detected in the input image. As shown, the majority of the images are still detected as stop signs even though we remove a robust feature. While YOLOv5 shows slightly higher robustness, all object detectors still detect stop signs in the majority of images: On average, the detection rate for all object detectors is $\geq$37%, meaning that 37% of the images generated by the diffusion models have the potential to be used as adversarial attacks. We also observed similar results on the other labels. More detailed results are in the supplementary materials.

3.3. Attack Capability against Image Classifiers

We also observe that the NDD attack is highly effective against image classifiers. For example, $\geq$47% of the generated stop sign images are still classified as stop signs even though all 4 robust features are removed. More details on the setup and results are in the supplementary materials.

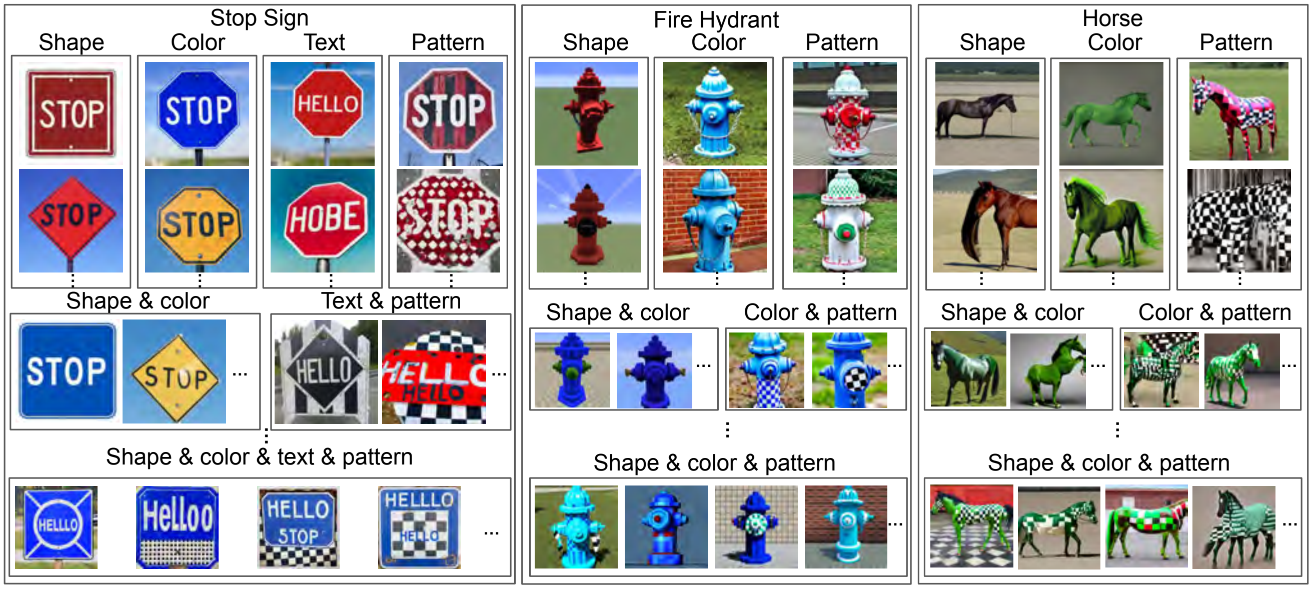


Figure 2. Overview of the Natural Denoising Diffusion Attack (NDDA) dataset. We alter or remove the 4 types of robust features partially or entirely. For the stop sign, we alter the text on it considering its importance to be recognized as a stop sign. For each set of robust features, we generate images with 3 diffusion models for 3 object classes.

Table 1. Templates of text prompts and examples for the "stop sign" object to remove the 4 robust features.

Removed Robust Features		Text prompt to remove/alter robust features	
		Format	Example: Stop sign
Benign prompt		**[Subject]**	**Stop sign**
Shape		**[Shape]** **[Subject]**	**Square** **stop sign**
Color		**[Color]** **[Subject]**	**Blue** **stop sign**
Texture	Text	**[Subject]** with "**[Text]**" on it	**Stop sign** with "**hello**" on it
	Pattern	**[Subject]** with a **[Pattern]** paint on it	**Stop sign** with a **checkerboard pattern** paint on it

3.4. Implications of Attack Capability Analysis

The NDD attack shows quite high attack effectiveness against both object detectors and image classifiers. These results imply that diffusion models can be utilizable to generate model-agnostic and highly transferable adversarial attacks with significantly less effort than prior works [12, 18, 40] that need iterative attack optimization processes.

However, the current analysis is not sufficient to fully conclude the vulnerability of DNN models against NDD attacks because adversarial attacks must satisfy 2 requirements: effectiveness against DNN models and stealthiness to humans. For example, diffusion models may ignore the text prompts and just merely generate legitimate stop signs. In the next section, we systematically evaluate the stealthiness of the NDD attack and explore potential root causes.

4. Empirical Analysis

We perform an extensive empirical study to further explore the characteristics and root causes of the NDD attack by answering 6 research questions (RQs).

RQ1: Does the natural attack capability exist in the previous image generation models?

We first investigate whether the natural attack capability exists in the previous image generation models.

Experimental setup. As the state-of-the-art image generation methods before diffusion models, we evaluate BigGAN [9]. To generate images guided by text prompts, we use BigSleep [6] that guides the image generation process of BigGAN by OpenAI CLIP [43]. We evaluate the 5 object detectors in Table 2. We use a detection threshold of 0.5. For each model, we evaluate 6 combinations of partial or complete removal of the robust features.

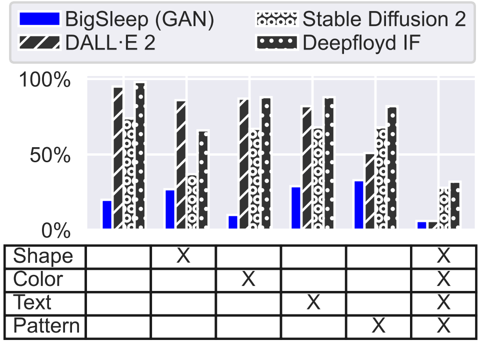
Results. Fig. 3 shows the average detection rate of stop sign images generated by each image generation model over the 5 object detectors. As shown, all diffusion models have significantly higher detection rates than the BigSleep GAN model. In particular, Stable Diffusion 2 and Deepfloyd IF have detection rates of $\geq$28% even if all robust features are removed or altered. Meanwhile, BigSleep has much lower detection rates (7.5%). We thus consider that *the natural attack capability has existed slightly in the prior image generation models, but it becomes significantly severe in diffusion models.* Diffusion models have higher image generation capabilities than GANs, but it may also enhance their ability to inject more non-robust features into generated images.

RQ2: How stealthy is NDD attack against humans?

To be a valid adversarial attack, the NDD attack should satisfy the two requirements: effectiveness against DNN models-

Table 2. Detection rates of 5 object detectors on the stop sign images in the NDDA dataset generated by the 3 diffusion models. **Bold** and underline denote highest and lowest scores in each row.

	Removed Robust Features				Object Detectors					
	Shape	Color	Text	Pattern	YOLOv3	YOLOv5	DETR	Faster R-CNN	RTMDet	Avg.
DALL-E 2					**100%**	76%	**100%**	**100%**	**100%**	95%
	✓				**98%**	36%	**98%**	**98%**	**98%**	86%
		✓			**98%**	48%	94%	**98%**	**98%**	87%
			✓		82%	32%	**100%**	**100%**	94%	82%
				✓	52%	10%	50%	52%	**90%**	51%
	✓	✓	✓	✓	**12%**	0%	6%	4%	8%	6%
				Avg.	74%	34%	75%	75%	**81%**	68%
Stable Diffusion 2					74%	50%	84%	**86%**	76%	74%
	✓				30%	24%	46%	**60%**	26%	37%
		✓			**78%**	40%	**78%**	72%	66%	67%
			✓		58%	56%	**90%**	70%	66%	68%
				✓	56%	48%	**90%**	76%	72%	68%
	✓	✓	✓	✓	30%	6%	**48%**	30%	24%	28%
				Avg.	54%	37%	**73%**	66%	55%	57%
Deepfloyd IF					**100%**	88%	**100%**	**100%**	**100%**	98%
	✓				68%	52%	70%	**84%**	56%	66%
		✓			**100%**	58%	92%	94%	94%	88%
			✓		84%	78%	94%	**96%**	88%	88%
				✓	80%	64%	88%	**90%**	88%	82%
	✓	✓	✓	✓	**60%**	0%	34%	34%	32%	32%
				Avg.	82%	57%	80%	**83%**	76%	76%



Shape		X				X
Color			X			X
Text				X		X
Pattern					X	X

Figure 3. Average detection rate of stop sign images over the 5 object detectors for 4 models. The "x" mark means the removed robust features.

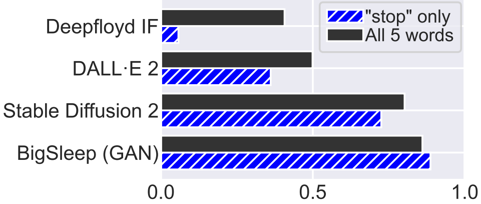


Figure 4. Averaged normalized Levenshtein distances between the given word and the detected text by OCR. The black bar is the averaged distance of all 5 words; the blue bar is only for the "stop" word.

els and stealthiness against humans. We so far have confirmed the effectiveness against DNN models as in Table 2, but it does not mean that the attack is stealthy. For example, the diffusion models may ignore the text prompts and just generate legitimate objects. To answer the questions, we perform a user test to investigate the stealthiness against humans and the validity of generated images.

Experimental setup. We recruited 82 human subjects on Prolific [1], a crowdsourcing platform specialized for research purposes. Human subjects are asked to answer yes or no to whether the object of interest (e.g., stop sign and fire hydrant) is in the presented image or not. Considering the reasonable experimental time for human subjects to maintain their concentration, image generation models were limited to the following: Deepfloyd IF, the diffusion model with the highest detection rates in Table 2, and BigSleep, a state-of-the-art GAN-based model. We generate 3 images per text prompt for each image generation model; for the baseline, 3 real images are presented. The evaluation images were chosen from a pool that can fool at least one object detector in Table 2, i.e., all the images are valid attacks. More details are in our questionnaire form [3].

Results. Table 3 lists the results of our user study. The detection rate is the proportion of users who answer "yes", i.e., they identify the target object in the presented image. As shown, DeepFloyd IF's images on the benign prompts have similarly high detection rates as the real images. On the other hand, the BigSleep images are not detected as the target objects as the detection rates are $\leq$4%. This indicates that the images generated by the state-of-the-art GAN-based model are not only not perceived by humans but also are not effective attacks against object detectors, i.e., the generated images are far from realistic. For the images without robust features, their detection rates are much lower than the real images and the images of benign prompts. We observe that the different objects have different sensitivities to each robust feature. For example, the text is very important to the stop sign as the human detection rate is 7% when the text on it is altered. DeepFloyd IF thus can easily generate effective adversarial attacks for stop signs because 93% of the human subjects did not see the images as stop signs even though 88% of the generated images are detected by stop signs as in Table 2. This result indicates that *the natural attack capability in diffusion models can indeed generate valid adversarial attacks that are stealthy to humans.*

RQ3: Does the incapability of text generation correlate with the natural attack capability?

We observed that the NDD attack has a high stealthiness to humans through our user study and found that the different objects have different sensitivities to different robust features. In particular, the text on a stop sign shows high importance for humans to identify a stop sign as in Table 3. Meanwhile, the object detectors are influenced by the shape or pattern rather than the text as in Table 2. Motivated by this, we further evaluate the impact of text generation capability in the natural attack capability.

Experimental setup. We generate the images with the 3

Table 3. Results of our user study. The detection rate is the ratio that the human subjects identify the targeted object in the presented image. Cell color means the magnitude of the detection rates: Greener means that more people think the object is in the image, and reddish means that fewer people think so.

	Real Image	Benign Prompt		Removed Robust Feature	Detection Rate
Stop sign	65%	BigSleep	4%	Shape	3%
				Color	1%
				Text	2%
				Pattern	1%
				All	1%
		Deepfloyd	56%	Shape	19%
				Color	11%
				Text	7%
				Pattern	25%
				All	8%
Horse	82%	BigSleep	1%	Shape	1%
				Color	2%
				Pattern	4%
				All	4%
		Deepfloyd	59%	Shape	7%
				Color	14%
				Pattern	2%
				All	7%
Fire hydrant	88%	BigSleep	3%	Shape	0%
				Color	2%
				Pattern	1%
				All	2%
		Deepfloyd	57%	Shape	1%
				Color	31%
				Pattern	48%
				All	2%

diffusion models (DALL-E 2, Stable Diffusion 2, and DeepFloyd IF) and on the GAN-based model (BigSleep). We use the text prompt format: "text of [word]", with "[word]" being one of the following: hello, welcome, goodbye, script, and stop; 20 images per word are generated with different seeds. Given a "[word]", we apply an optical character recognition (OCR) method, specifically the pipeline provided by the Keras-OCR package [41], for generated images and calculate the normalized Levenshtein (edit) distance between [word] and the recognized sentence(s), which are joined with a space if multiple are recognized. We expect the Levenshtein distance to be 0 for identical pairs of text and 1 for completely different pairs of text.

Results. Fig. 4 shows the averaged normalized Levenshtein (edit) distances of all 5 words and only "stop". As shown, the Deepfloyd IF can generate the most accurate text; DALL-E 2 and Stable Diffusion 2 are the second and third while BigSleep is the worst. This order is the same as the order of average detection rates of the object detectors in Table 2. In particular, the Deepfloyd IF shows a high capability to generate "stop" texts. This result is consistent with the observation in RQ2 that the stealthiness against humans is significantly improved on the DeepFloyd IF when the robust feature of the text is removed. In summary, the experimental results show that the text generation capability of image generation models has a certain similarity to their natural attack capability. The capability to generate a complex pattern such as the alphabet may correlate with the capability to generate non-robust features that are too subtle to the HVS. We may use this characteristic to design a simple defense against the NDD attack, i.e. we can differentiate the NDD attack by checking if the text "stop" is in it for the stop sign attack. Although these empirical results are still not enough to fully support the edit distance-based method as a metric to measure the natural attack capability, *the metric can be used as a simple sanity check for the image generation models and as a simple defense against NDD attacks on stop signs and other objects with text.*

RQ4: Are non-robust features responsible for the natural attack capability?

This study is motivated by the concept of robust and non-robust features proposed in [31] that the adversarial attack is not a bug but instead caused by non-robust features that are predictive but incomprehensible to humans. In the user study, we have observed that diffusion models introduce some features that are imperceptible to humans but important to the DNN models. While we may call these features non-robust features by definition of the concept, we perform a further evaluation to directly evaluate the effect of non-robust features by following the methodology in [31].

Experimental setup. According to [31], we first create the "robustified" dataset to train robust classifiers. Since our NDDA dataset uses the same labels as the COCO dataset [37], we convert the COCO dataset into a dataset for the multiclass classification task by cropping the images within their bounding boxes, randomly selecting 500 images for each class, and "robustify" the images. We train not only the robustified classifier with the dataset but also a normal classifier with the original, non-robust dataset, again with 500 random images per class. ResNet50 [27] is the model architecture for both classifiers.

Results. Table 4 shows the accuracy of the robust and normal classifiers on the stop sign images in the NDDA dataset. As shown, both the normal and robustified classifiers have higher accuracy when the robust features exist, but the robust classifier's accuracy decreases more than the normal classifier's when all robust features are removed. This means that there is a clear correlation between the sensitivity to the non-robust features and the natural attack capability. We thus consider that *the non-robust features play an important role in enabling the natural attack capability in the diffusion models.* For the other classes, we include the results in the supplementary materials.

RQ5: Is the natural attack capability caused by sharing training datasets?

We evaluate the impact of sharing the training dataset on the natural attack capability. If the evaluating object detectors and diffusion models may share the training dataset,

Table 4. Accuracy of the robust and normal classifiers on the stop sign images in the NDDA dataset. The benign means the images generated with the benign prompts; The NDD means the NDD attack that removes all robust features.

	Robustified classifier			Normal classifier		
	Benign	NDD	Diff.	Benign	NDD	Diff.
DALL-E 2	1.00	0.34	0.66	1.00	0.70	0.30
Stable Diffusion 2	0.78	0.58	0.20	0.80	0.66	0.14
DeepFloyd IF	1.00	0.92	0.08	1.00	0.98	0.02
Avg.	0.93	0.61	**0.31**	0.93	0.78	**0.15**

Table 5. Accuracy of the classifiers on the images generated by the DDPM. The rows indicate the splits used to train the DDPM. The columns indicate the splits used for training the classifier.

	MNIST		Fashion MNIST		CIFAR-10	
DDPM\Classifer	Split 1	Split 2	Split 1	Split 2	Split 1	Split 2
Split 1	1.00	1.00	0.98	0.94	0.71	0.64
Split 2	1.00	1.00	0.99	0.93	0.67	0.67
Abs. Diff.	0.00	0.00	0.01	0.01	0.04	0.03

the natural attack capability can be just due to the non-generalizable features in the dataset, i.e., both may just fit into particular noises in the dataset. The large diffusion models, such as DALL-E 2 and Stable Diffusion 2, are likely to use famous, publicly available datasets and may use the COCO dataset [37], which the 5 object detectors in Table 2 use for training. In this RQ, we thus assess the impact of the dataset sharing to see if it can be a major source of the natural attack capability.

Experimental setup. We randomly split the training datasets of MNIST [36], FashionMNIST [55], and CIFAR-10 [35] into 2 splits, respectively. We train simple CNN classifiers and conditional DDPM model [42] for all splits. For each conditional DDPM, we generate 100 images for each class. We then evaluate the difference in the accuracies of the generated images between when the diffusion model and the classifier use the same split and when they use different splits. If the hypothesis is true, these accuracies should have a large gap.

Results. Table 5 shows the accuracy of each pair of conditional DDPMs and classifiers. The row means which split is used to train the conditional DDPM model; the column means which split is used to train the CNN model. The accuracy is the percentage of the correct classification on the 100 generated images with the corresponding DDPM model. As shown, there is no significant difference between when the diffusion model and the classifier use the same split and when they use different splits. We thus think that *the natural attack capability of diffusion is not due to data sharing but rather due to the non-robust features embedded by diffusion models.*

RQ6: Is the natural attack capability general enough to attack against real-world systems?

To demonstrate the model-agnostic and transferable attack capability of the NDD attack, we evaluate the NDD attacks against the real-world traffic sign detection system. We adopt the threat model of appearing attack [59], which causes a false positive detection of a stop sign.

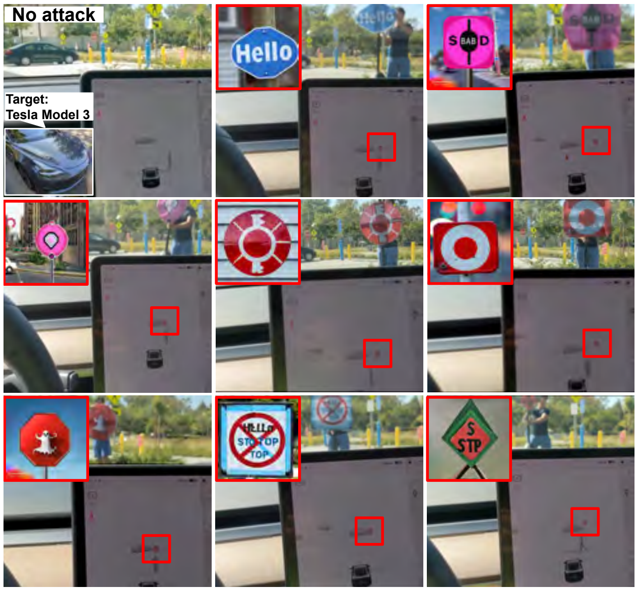


Figure 5. Successful NDD attacks against Tesla Model 3. We demonstrate that 8 out of 11 printed attacks can successfully fool the commercial traffic sign detection system in Tesla. The left-top images are the original images generated from diffusion models.

Experimental setup. We printed 11 images of the NDD attack as in Fig. 5, showed them to the windshield cameras of the Tesla Model 3, and checked the detection results displayed on the monitor in front of the driver's seat. We select the 11 attack images based on the confidence scores of object detectors in the digital space, especially for YOLOv5, which shows the highest robustness as in Table 2.

Results. Fig. 5 shows the successful NDD attacks against Tesla. As shown, 8 of the 11 printed attacks can successfully fool Tesla's commercial traffic sign system as the detected stop signs appear on the driver's monitor. This means that the *NDD attack has 73% of the attack success rate, which is a surprisingly high success rate considering that we do not perform any optimization processes to attack the Tesla.* All attacks are simply generated by diffusion models with simple text prompts. Furthermore, we did not have any special considerations when printing the attacks. We just use a commodity printer, roughly stick papers together with transparent tape, and use normal printing paper. More demos and details are on our project website: https://sites.google.com/view/cav-sec/ndd-attack.

5. Discussion and Limitation

We discuss the implications of our findings and the limitations of this work, especially the potential negative societal impacts raised by this work.

Safety Implications: We demonstrate the attack effectiveness of the NDD attack against the Tesla Model 3 but also find that these attacks were not as robust at different distances. Thus, this vulnerability may not pose an immediate threat to fast-driving Tesla or other autonomous vehicles. However, the possibility of affecting a driving vehicle cannot be completely ruled out. We note that the effect of the attack is unlikely to be a coincidence, as these attacks are detected as stop signs even though they do not resemble similar to legitimate stop signs at all. If the attack is reddish and hexagonal, it could be a coincidence. However, it cannot be considered a coincidence that such an attack with blue, purple, green, or non-hexagonal shapes is detected as a stop sign at such a high rate. Furthermore, the current NDD attack does not have any special considerations to attack the Tesla Model 3. As in prior work [8, 15], we may improve the attack by integrating diffusion models into attack generation processes. We hope that our study can facilitate further research to assess the security threat of diffusion models. We have performed a responsible vulnerability disclosure to Tesla before the public release of this work.

Countermeasures: A naive but possible defense for the stop sign attack is the OCR-based attack detection as discussed in RQ3. For the NDD attack removing the text feature, we found that it can achieve 92% true positive rate with 4% false positive rate on the images generated by Deepfloyd IF. However, the true positive rates drop to ≥40% if one of the other robust features is removed. Another approach is the "robustified" training [31] as observed in RQ4, but it remains a mitigation measure. So far, no generic defense against adversarial attacks has been reported [7, 13, 53]. Further research efforts are needed in this area.

Definition of Attack Success: As discussed in RQ2, valid adversarial attacks should satisfy the two requirements: effectiveness against DNN models and stealthiness against humans. From this aspect, the stealthiness of all images in the NDD dataset has not been fully confirmed. However, human annotation for all images did not yield a feasible choice due to the cost, which motivated us to pursue RQ2. We further note that we use state-of-the-art 3 diffusion models, which generally follow the text prompt very faithfully as shown in Fig. 2.

Root Causes of Natural Attack Capability: Through this study, we empirically explore possible root causes of the natural attack capability and identify potential clues that the natural attack capability is due to the non-robust features injected by diffusion models. However, our empirical study is not sufficient to fully conclude the root causes. For example, the methodology used in RQ4 to extract (non-)robust features proposed by Ilyas et al. [31] is not designed to analyze the natural capability of artificially generated images. Considering the impact on safety-critical applications such as autonomous driving, we are presenting our current best-effort empirical analysis along with our dataset. We hope our study can facilitate further theoretical or more large-scale empirical studies to identify the root causes of the natural attack capability in diffusion models.

Evaluation with More Models and Categories: In this work, we focus on 3 popular diffusion models and 3 object classes with relatively high detection rates to perform deeper analysis on each RQ. Meanwhile, we keep updating the dataset with more diffusion models and object classes for the sake of the dataset's comprehensiveness. The latest version of the NDDA dataset includes 7 diffusion models (Dall-E 2 [19], Dall-E 3 [20], Stable Diffusion 2 [47], Deepfloyd IF [51], Stable Diffusion 1.5 [47], MidJourney [2], and Google Duet [25]) with 15 object classes (stop sign, car, dog, hot dog, traffic light, zebra, fire hydrant, frog, horse, bird, boat, air plane, bicycle, cat, and carrot) to benefit future studies. In the supplementary materials, we evaluate the detection rates of the 15 classes and find that the majority of classes have certain levels of vulnerability against the NDD attack. Current candidates for robust features are chosen to ensure removal (e.g., blue for stop sign); future updates will include more variations for the sake of dataset comprehension on our website. In total, the NDDA dataset has 45,820 images. Each class has ≥50 images for the models with API access and ≥20 images for other models.

Ethical Considerations: We have gone through the IRB process for the user study. In the experiment on the Tesla Model 3, we ensured that the attacks were not visible to other Tesla vehicles driving on public roads.

6. Conclusion

In this study, we identify a new security threat of the NDD attack that leverages the natural attack capability in the diffusion models. To systematically evaluate the characteristics and root causes of the NDD attack, we conduct a large-scale empirical study using our newly constructed dataset, the NDDA dataset, in which we generate images with state-of-the-art diffusion models while intentionally removing the robust features that are essential to the HVS. We demonstrate that the images without robust features are still detected as the original object and are stealthy to humans. For example, the stop signs with altered text are still detected as stop signs in 88% of cases but are stealthy to 93% of humans. We find that the non-robust features contribute to the natural attack capability. To evaluate the realizability of the NDD attack, we demonstrate the attack against the Tesla Model 3 and confirm that 73% of the NDD attacks are detected as stop signs. Finally, we discuss the implications and limitations of our research. We hope that our study and dataset can help the community to be aware of the risk of the natural attack capability of diffusion models and facilitate further research to develop robust DNN models.

Acknowledgements

This research was supported in part by the NSF CNS-2145493, CNS-1929771, CNS-1932464, and USDOT under Grant 69A3552348327 for the CARMEN+ UTC.

References

[1] Prolific. https://researcher-help.prolific.co/hc/en-gb, 2014. 5

[2] MidJourney. https://www.midjourney.com/, 2023. 8

[3] Our User Study Form: AI Images' Realism Survey. https://drive.google.com/file/d/1ifbOT8YE-iHfLvss_h4nzNsq6PZwOIwo/view?usp=drive_link, 2023. 5

[4] Our Proejct Website. https://sites.google.com/view/cav-sec/ndd-attack, 2024. 2

[5] Adobe. Adobe Firefly - Generative AI for creatives. https://www.adobe.com/sensei/generative-ai/firefly.html, 2023. 1

[6] Adverb. BigSleep. https://github.com/lucidrains/big-sleep, 2021. 4

[7] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *International Conference on Machine Learning (ICML)*, 2018. 8

[8] Tao Bai, Jun Zhao, Jinlin Zhu, Shoudong Han, Jiefeng Chen, Bo Li, and Alex Kot. AI-GAN: Attack-Inspired Generation of Adversarial Examples. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2543–2547. IEEE, 2021. 8

[9] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2019. 3, 4

[10] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. Invisible for Both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving under Physical-World Attacks. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 176–194. IEEE, 2021. 3

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020. 3

[12] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *IEEE symposium on security and privacy (SP)*, pages 39–57. IEEE, 2017. 4

[13] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On Evaluating Adversarial Robustness. *arXiv preprint arXiv:1902.06705*, 2019. 8

[14] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting Training Data from Diffusion Models. *arXiv preprint arXiv:2301.13188*, 2023. 1

[15] Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion Models for Imperceptible and Transferable Adversarial Attack. *arXiv preprint arXiv:2305.08192*, 2023. 1, 8

[16] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 3

[17] Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. AdvDiffuser: Natural Adversarial Example Synthesis with Diffusion Models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4562–4572, 2023. 1

[18] Minhao Cheng, Simranjit Singh, Patrick H. Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-OPT: A Query-Efficient Hard-label Adversarial Attack. In *International Conference on Learning Representations (ICLR)*, 2020. 4

[19] DALL-E-2. DALL-E-2. https://openai.com/dall-e-2, 2022. 8

[20] DALL-E-3. DALL-E-3. https://openai.com/dall-e-3, 2023. 8

[21] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis, 2021. 2

[22] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust Physical-World Attacks on Deep Learning Visual Classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[23] Yunhao Ge, Yao Xiao, Zhi Xu, Xingrui Wang, and Laurent Itti. Contributions of Shape, Texture, and Color in Visual Recognition. In *European Conference on Computer Vision (ECCV)*, pages 369–386. Springer, 2022. 2, 3

[24] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *ICLR*, 2015. 3

[25] Google. Introducing Duet AI for Google Workspace. https://workspace.google.com/blog/product-announcements/duet-ai, 2023. 8

[26] Kalanit Grill-Spector and Rafael Malach. The Human Visual Cortex. *Annu. Rev. Neurosci.*, 27:649–677, 2004. 2, 3

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 6

[28] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, 2021. 3

[29] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 1, 2

[30] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. https://doi.org/10.5281/zenodo.5143773, 2021. 3

[31] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 3, 6, 8

[32] Zou Jiayu, Zhu Zheng, Ye Yun, and Wang Xingang. DiffBEV: Conditional Diffusion Model for Bird's Eye View Perception. *arXiv preprint arXiv:2303.08333*, 2023. 1

[33] Glenn Jocher. YOLOv5 by Ultralytics. `https://github.com/ultralytics/yolov5`, 2020. 3

[34] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion Models in Medical Imaging: A Comprehensive Survey. *Medical Image Analysis*, page 102846, 2023. 1

[35] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images, 2009. 7

[36] Yann LeCun and Corinna Cortes. MNIST Handwritten Digit Database. http://yann.lecun.com/exdb/mnist/, 2010. 7

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 3, 6, 7

[38] Chengqi Lyu, Wenwei Zhang, Haian Huang, Yue Zhou, Yudong Wang, Yanyi Liu, Shilong Zhang, and Kai Chen. RTMDet: An Empirical Study of Designing Real-Time Object Detectors. *arXiv preprint arXiv:2212.07784*, 2022. 3

[39] Chen Ma, Ningfei Wang, Qi Alfred Chen, and Chao Shen. SlowTrack: Increasing the Latency of Camera-Based Perception in Autonomous Driving Using Adversarial Examples. *AAAI*, 2024. 3

[40] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representation (ICLR)*, 2018. 4

[41] Fausto Morales. A Packaged and Flexible Version of the CRAFT Text Detector and Keras CRNN Recognition Model. `https://github.com/faustomorales/keras-ocr`, 2022. 6

[42] Tim Pearce. Conditional Diffusion MNIST. `https://github.com/TeaPearce/Conditional_Diffusion_MNIST`, 2022. 7

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 2, 3, 4

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with Clip Latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3

[45] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *International Conference on Neural Information Processing Systems (NIPS)*, 2015. 3

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 3, 8

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[49] Takami Sato, Junjie Shen, Ningfei Wang, Yunhan Jia, Xue Lin, and Qi Alfred Chen. Dirty Road Can Attack: Security of Deep Learning based Automated Lane Centering under Physical-World Attack. In *USENIX Security Symposium*, 2021. 3

[50] Junjie Shen, Ningfei Wang, Ziwen Wan, Yunpeng Luo, Takami Sato, Zhisheng Hu, Xinyang Zhang, Shengjian Guo, Zhenyu Zhong, Kang Li, et al. Sok: On the Semantic AI Security in Autonomous Driving. *arXiv preprint arXiv:2203.05314*, 2022. 3

[51] StabilityAI. DeepFloyd IF. `https://github.com/deep-floyd/IF`, 2023. 8

[52] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. In *International Conference on Learning Representations (ICLR)*, 2014. 3

[53] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On Adaptive Attacks to Adversarial Example Defenses. *arXiv preprint arXiv:2002.08347*, 2020. 8

[54] Ningfei Wang, Yunpeng Luo, Takami Sato, Kaidi Xu, and Qi Alfred Chen. Does Physical Adversarial Example Really Matter to Autonomous Driving? Towards System-Level Effect of Adversarial Object Evasion Attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4412–4423, 2023. 3

[55] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-Mnist: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 7

[56] Ping yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studor, and Tom Goldstein. Certified Defenses for Adversarial Patches. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[57] Lvmin Zhang and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv preprint arXiv:2302.05543*, 2023. 1

[58] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. Interpretable Deep Learning under Fire. In *USENIX Security Symposium*, 2020. 3

[59] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen. Seeing isn't Believing: Practical Adversarial Attack Against Object Detectors. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2019. 7