

fakeWeather: Adversarial Attacks for Deep Neural Networks Emulating Weather Conditions on the Camera Lens of Autonomous Systems

Original

fakeWeather: Adversarial Attacks for Deep Neural Networks Emulating Weather Conditions on the Camera Lens of Autonomous Systems / Marchisio, A; Caramia, G; Martina, M; Shafique, M. - ELETTRONICO. - 1:(2022), pp. 1-9. (Intervento presentato al convegno International Joint Conference on Neural Networks (IJCNN) tenutosi a Padua, Italy nel 18-23 July 2022) [10.1109/IJCNN55064.2022.9892612].

Availability:

This version is available at: 11583/2977296 since: 2023-03-22T11:38:20Z

Publisher:

IEEE

Published

DOI:10.1109/IJCNN55064.2022.9892612

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

IEEE postprint/Author's Accepted Manuscript

©2022 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

fakeWeather: Adversarial Attacks for Deep Neural Networks Emulating Weather Conditions on the Camera Lens of Autonomous Systems

Alberto Marchisio^{1,*}, Giovanni Caramia^{2,*}, Maurizio Martina², Muhammad Shafique³

¹Technische Universität Wien, Vienna, Austria ²Politecnico di Torino, Turin, Italy ³New York University, Abu Dhabi, UAE

Email: alberto.marchisio@tuwien.ac.at, giovanni.caramia@studenti.polito.it

maurizio.martina@polito.it, muhammad.shafique@nyu.edu

Abstract—Recently, Deep Neural Networks (DNNs) have achieved remarkable performances in many applications, while several studies have enhanced their vulnerabilities to malicious attacks. In this paper, we emulate the effects of natural weather conditions to introduce plausible perturbations that mislead the DNNs. By observing the effects of such atmospheric perturbations on the camera lenses, we model the patterns to create different masks that fake the effects of rain, snow, and hail. Even though the perturbations introduced by our attacks are visible, their presence remains unnoticed due to their association with natural events, which can be especially catastrophic for fully-autonomous and unmanned vehicles. We test our proposed *fakeWeather* attacks on multiple Convolutional Neural Network and Capsule Network models, and report noticeable accuracy drops in the presence of such adversarial perturbations. Our work introduces a new security threat for DNNs, which is especially severe for safety-critical applications and autonomous systems.

Index Terms—Deep Neural Networks, Adversarial Attacks, Weather, Rain, Snow, Hail.

I. INTRODUCTION

In the last decade, Deep Neural Networks (DNNs) have lifted groundbreaking advancements in several fields, including object recognition [1], autonomous systems [2], video, image, and signal processing [3], and achieving the human-level or even more classification accuracy for certain tasks [4]. However, despite their great success [5][6], DNNs have been proven to be vulnerable to adversarial attacks, which undermine their security since they maliciously subvert the DNN predictions [7]. While several works of adversarial machine learning have been proposed earlier [8][9], their first application to DNN-based algorithms was conducted by Szegedy et al. [10], who demonstrated that DNNs can easily be fooled by injecting imperceptible perturbations into the input images.

The usage of DNN-based algorithms for safety-critical applications requires that the DNNs do not fail their predictions under challenging conditions [11][12]. Such situations can appear in many different forms, including process variations that induce hardware faults, input pollution, or poisoning that induce a misbehavior [13]. For instance, for vision applications in smart transportation systems, the DNNs should correctly work under different lights and atmospheric

phenomena. Hence, an image captured in such conditions represents a plausible image that can be processed by the DNN-based algorithm.

A. Target Research Problem and Associated Challenges

The key objective for an adversarial attack and its applicability in practical use-cases consists of not being recognized as adversarial, but rather as common/plausible. The most intuitive approach is to inject a very limited amount of perturbations, with the goal of making the differences between the clean images and the adversarial images imperceptible to the human eye. This approach has been adopted by several works, including Luo et al. [14], Croce et al. [15], and Marchisio et al. [16]. However, the attacker needs to have access to a set of information, including DNN model architecture and parameters, inputs, and outputs (i.e., in *white-box* settings), or only inputs and outputs (in *black-box* settings). Even the most advanced decision-based black-box attacks such as HopSkipJumpAttack [17] and FaDeC [18] still have access to the DNN predicted output class for each image. However, in practical cases, it may be very complicated to obtain such information, due to the protection mechanisms applied by the DNN-based system developers [19]. Moreover, another key limitation resides in the fact that even the most query-efficient algorithms [20] need to perform a certain number of queries (i.e., inference passes) to generate the adversarial perturbation, which may not be practical in case of stringent real-time constraints, because of the latency overhead caused by the queries.

Due to these limitations of the adversarial attacks that aim at introducing imperceptible perturbations compared to the original images, our approach follows a different strategy (see Figure 1). *Our novel idea is to introduce perturbations to the input image in such a way that it is not considered as adversarial, since it resembles a natural condition captured by the camera.* While the differences between the clean image and the adversarial image can be noticed, the adversarial image itself is hardly categorized as “adversarial”, since it simply captures a plausible natural condition. The reason is based on the fact that traditional adversarial machine learning takes into account the comparison between the adversarial image and the original image. However, in real-world practice,

*These authors contributed equally to this work.

it is impossible to obtain such a comparison, since the only accessible image is the one recorded by the camera. Noticeably, our approach is advantageous compared to previous works, since it is conducted in what we call a *true black-box setting*, i.e., in a scenario in which the attacker has no information about the DNN model architecture and parameters, nor its outputs. The only information required is the size of the input image, for generating an adversarial mask of that size. Moreover, our attack does not require any query. Thus it can easily be applied at run time.

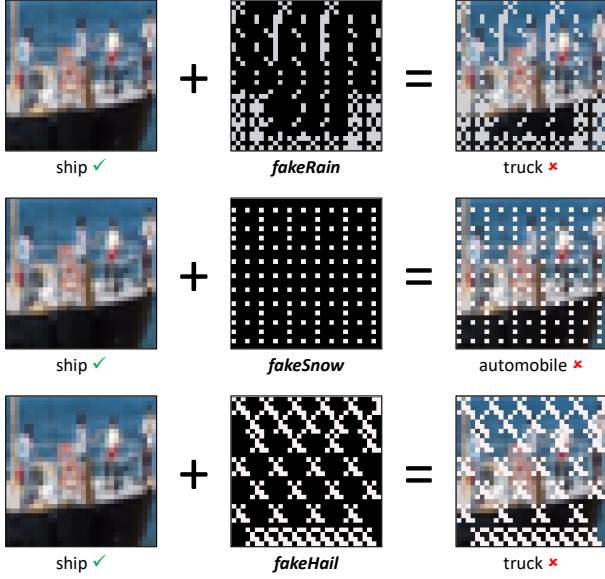


Fig. 1. *fakeWeather* attacks functionality.

B. Our Novel Contributions

Towards this, we observe how natural weather conditions, such as rain, snow, and hail, are perceived by the camera. We exploit this observation by designing *fakeWeather* attack algorithms that emulate these effects on the camera lens. An overview of its functionality is shown in Figure 2. Our methodology can be used not only as an adversarial attack to mislead the DNN, but also as a data augmentation approach for reinforcing the DNN training under these conditions. Our contributions can be summarized as follows:

- We observe several images of natural weather events that affect the camera (**Section III-B**), and identify the patterns that are more commonly present in such images (**Section III-C**).
- By only knowing the image size, we design three *fakeWeather* masks that fake the effect of such weather conditions on the camera lenses (**Sections III-D, III-E, and III-F**).
- We evaluate the *fakeWeather* attacks on multiple DNN models (LeNet-5, ResNet-32, CapsNet) for the CIFAR-10 dataset, and obtain a success rate of the attacks varying between 30% and 82.5% (**Section IV**).

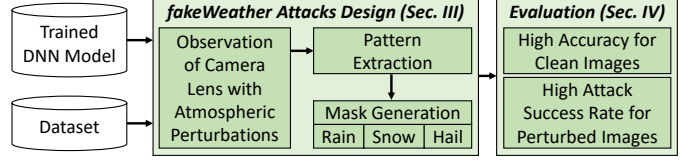


Fig. 2. Overview of our Novel Contributions.

Before proceeding to the technical sections, we provide an overview of the adversarial attacks and the related works in Section II.

II. BACKGROUND AND RELATED WORKS

The purpose of the most common adversarial attack algorithms, such as gradient-based attacks [21], is to introduce some perturbations that induce a decision boundary cross in the DNNs, and therefore lead to a misclassification. Examples of such attacks include the Fast-Gradient Sign Method (FGSM) [22], DeepFool [23], the Projected Gradient Descent (PGD) [24], and the Carlini & Wagner attack [25]. Other classes of attacks in which the perturbations were inserted only in a small set of pixels or only in one pixel were proposed by Narodytska et al. [26] and Su et al. [27], respectively. Concurrently, Moosavi-Dezfooli et al. [28] proposed image-agnostic universal perturbations that are applied to every sample, and Zhang et al. [29] generated different adversarial perturbations for each target class.

In *black-box* settings, several works were conducted. Kurakin et al. [22] proposed a method that crafts adversarial examples in the physical world by taking the images from a cell-phone camera. Moreover, taking into account the high-saliency and low-distortion path, Gragnaniello et al. [30] introduced an attack that improves the perceptual quality of the adversarial image.

Several attacks have been designed for *real-world* settings that incorporate so-called environmental perturbations. Brown et al. [31] generated adversarial patches, i.e., image-independent patches, to be placed anywhere inside the original image to mislead the DNNs. Following a similar approach, Eykholt et al. [32] added stickers to road signs to fool the traffic sign recognition system, while Sharif et al. [33] added glasses to faces to fool the face recognition system. Man et al. [34] proposed GhostImage attacks, in which the adversarial patterns are inserted into the camera systems through a projector.

Focusing on more closely related approaches to our work, other methods in which DNN models are fooled due to *atmospheric phenomena* were proposed. Temel et al. [35] analyzed several challenging conditions, including snow and rain, for traffic sign detection systems, and collected them into their proposed CURE-TSD-Real dataset. Zhai et al. [36] simulated various rainy situations using a gradient-based rain generation process. However, there are clear differences compared to our *fakeWeather* attacks. *Both these two related works inject perturbations in the long-range, i.e., relatively*

far from the camera, while our methodology introduces perturbations in the close proximity of the camera lens. Unlike other methods in the related works, our approach is non-invasive, since it does not require any modification in the real world, but it only modifies some pixel intensities of the images without interfering with the underlying DNN processing.

III. FAKEWEATHER ATTACK DESIGN

A. Problem Formulation and Assumptions

Taking into account the previous discussions, we propose the *fakeWeather* methodology. An overview of its functionality is shown in Figure 3. The final goal is to generate a finite set of perturbations with certain patterns which resemble the effect of natural weather events. Hence, such patterns are crafted by faking that the camera lens is dirty due to atmospheric conditions (such as rain, snow, and hail). After observing their effects on several examples in the real world, the common patterns are extracted and reproduced to generate the perturbation masks. The attack is conducted in what we call a *true black-box setting*, i.e., assuming that:

- the adversary has no information about the DNN model architecture, its parameters, and its output;
- the only information available for the attacker is the size of the input images.

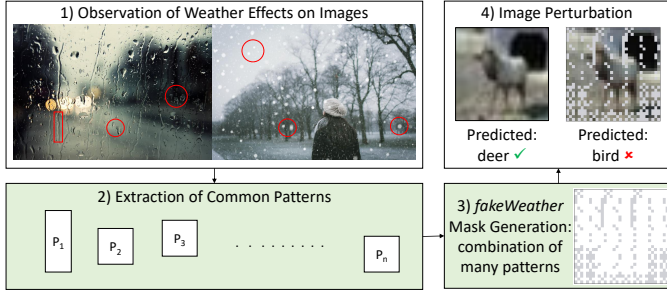


Fig. 3. Overview of the *fakeWeather* attack methodology.

B. Observation of Weather Conditions

The *fakeWeather* attacks are performed through the introduction of drops of water and snowflakes. A typical water drop has a spherical shape, while a snowflake has a hexagonal shape. However, in practical use-cases, these weather conditions do not represent the main focus of the camera. A camera captures the effects of rain and snow in a different way, which results into a set of blurry dots that are overlapped to the image. For instance, if we consider the use case of vision for smart mobility, the camera can be placed either outside of the vehicle (and hence exposed to the weather conditions), or inside the vehicle but in close proximity to the window. Without loss of generality, we can model a drop or a snowflake as a single pixel w.r.t. the image of $h \cdot l$ pixels, where h and l represent the height and length, respectively.

C. Pattern Extraction and Mask Generation

According to the previous considerations, the *fakeWeather* methodology extends the formulation of the One Pixel Attack [27], in which the perturbation of a *single pixel* is defined as a tuple of 5 elements (x, y, r, g, b) where:

- (x, y) represent the coordinates of the pixel to be modified;
- (r, g, b) indicate the color of the pixel in RGB format.

Therefore, an adversarial pattern combines multiple pixel attacks, in which the perturbation introduced on the pixel i can be written as in Equation 1. An example of the corresponding mask of an adversarial pattern is shown in Figure 4.

$$pixel_i = (x_i, y_i, r_i, g_i, b_i) \quad (1)$$

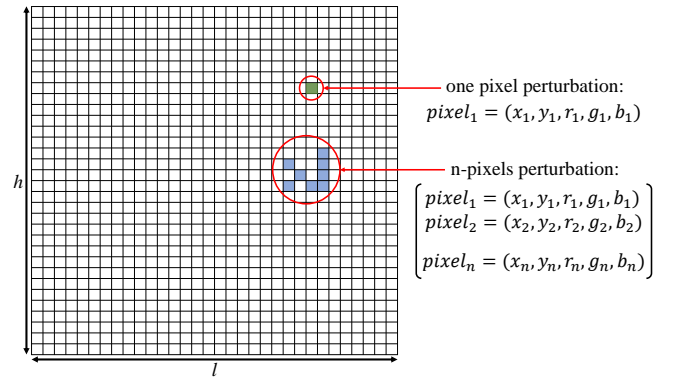


Fig. 4. Encoding of pixel perturbations that form the adversarial pattern.

The colors, i.e., the values assumed by (r_i, g_i, b_i) , are determined according to the weather condition:

- **rain:** $(r_r, g_r, b_r) = (208, 209, 214)$
- **snow and hail:** $(r_s, g_s, b_s) = (249, 242, 242)$

For each type of *fakeWeather* attack (i.e., rain, snow, and hail), specific patterns are generated. Common patterns are extracted from real images and reproduced to form the set of pixel coordinates (x_i, y_i) that belongs to the attack mask. Once generated, the same mask is applied to all the images under attack.

D. fakeRain Attack

The mask employed in the *fakeRain* attack is designed based on the combination of several water drops. In the real world, the camera lens can be soiled due to the rain, where the water droplets make up different patterns. It is possible to recognize three real-case scenarios, which can be categorized as agglomerate of drops, drop patch, and drop lines. As shown in Figure 5, the next step consists of modeling these patterns in terms of pixel coordinates that are perturbed.

The *Agglomerate Pattern* can be modeled by combining together 5 pixels to form a cross sign, according to the sketch in Figure 6a and Algorithm 1. The *Patch Pattern* (see Figure 6b) can have three different shapes, namely the *vertical patch*, which can be modeled as two consecutive pixels that

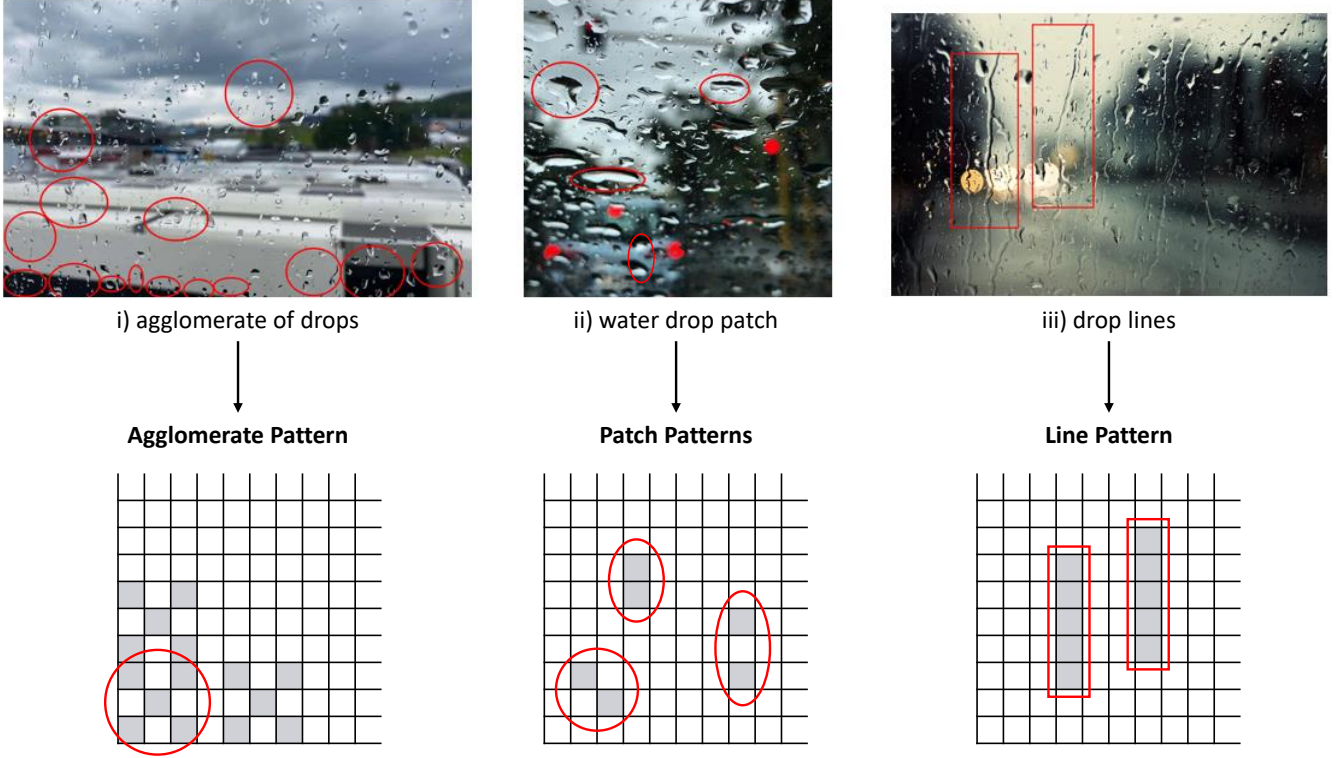


Fig. 5. Several patterns of water drops observed from the real environment. i) agglomerate of drops, ii) water drop patch, iii) drop lines.

share the same x coordinate (lines 4-6 of Algorithm 2), the *diagonal patch*, modeled as two pixels arranged to form a diagonal (lines 10-17 of Algorithm 2, and the *two dots patch*, in which two pixels are separated by a blank space (lines 20-22 of Algorithm 2). The *Line Pattern*, shown in Figure 6c, is modeled as a vertical line of n pixels (see Algorithm 3).

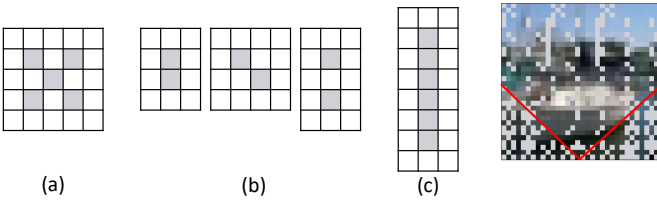


Fig. 6. A graphical representation of (a) Agglomerate Pattern, (b) Patch Patterns, and (c) Line Pattern.

Moreover, in rainy conditions, we can notice that the water drops tend to concentrate in the bottom corners of the image. Hence, to emulate this effect, in the *fakeRain* attack, a *V-shape* is created to divide the image into two regions (see the example in Figure 7). Below the *V*, several agglomerate patterns are densely concentrated. Above the *V*, path and line patterns are more sparsely distributed. Algorithm 4 describes the procedure for generating the *fakeRain* mask. Note that it is a three-step process in which (i) several agglomerate patterns are added (see line 2 of Algorithm 4), (ii) other agglomerate patterns are added if the coordinate is below the *V* (line 5 of

Fig. 7. V-shaped fakeRain attack.

Algorithm 1: Agglomerate Pattern

input : Coordinate (x_0, y_0)
output: Agglomerate Pattern P_a

```

1  $P_a = \emptyset$ 
2  $k = 0$ 
3 for  $i \leftarrow 0$  to 2 do
4   for  $j \leftarrow 0$  to 2 do
5     if  $(i + j = 0 \vee i + j = 2 \vee i + j = 4)$  then
6        $P_a \leftarrow pixel_k = (x_0 + i, y_0 + j, r_r, g_r, b_r)$ 
7        $k \leftarrow k + 1$ 
8     end
9   end
10 end

```

Algorithm 4), and (iii) patch patterns of different types and line patterns are added above the *V* (line 7 of Algorithm 4).

E. fakeSnow Attack

The design of the *fakeSnow* attack is based on the assumption that a snowflake can be modeled as a single pixel, since the dimension of each snowflake is relatively small, as observed in Figure 8. According to these considerations, the snow pattern P_s consists of a single pixel, which can be modeled as in Equation 2, where (x_0, y_0) represents the coordinate in which the snow pattern is constructed.

Algorithm 2: Patch Pattern

input : Coordinate (x_0, y_0) , Type t
output: Patch Pattern P_p

```

1  $P_p = \emptyset$ 
2 switch  $t$  do
3   case 0 do // Vertical Patch
4     for  $j \leftarrow 0$  to 1 do
5        $P_p \leftarrow pixel_j = (x_0, y_0 + j, r_r, g_r, b_r)$ 
6     end
7   end
8   case 1 do // Diagonal Patch
9      $k = 0$ 
10    for  $i \leftarrow 0$  to 1 do
11      for  $j \leftarrow 0$  to 1 do
12        if  $(i + j = 1)$  then
13           $P_p \leftarrow pixel_k =$ 
14             $(x_0 + i, y_0 + j, r_r, g_r, b_r)$ 
15           $k \leftarrow k + 1$ 
16        end
17      end
18    end
19    case 2 do // Two Dots Patch
20      for  $j \leftarrow 0$  to 1 do
21         $P_p \leftarrow pixel_j = (x_0, y_0 + 2 \cdot j, r_r, g_r, b_r)$ 
22      end
23    end
24 end

```

Algorithm 3: Line Pattern

input : Coordinate (x_0, y_0) , Length n
output: Line Pattern P_l

```

1  $P_l = \emptyset$ 
2 for  $j \leftarrow 0$  to  $n - 1$  do
3    $P_l \leftarrow pixel_j = (x_0, y_0 + j, r_r, g_r, b_r)$ 
4 end

```

$$P_s \leftarrow pixel_0(x_0, y_0, r_s, g_s, b_s) \quad (2)$$



Fig. 8. Several snowflakes observed, which can be modeled as single dots.

Another key feature noticed from the observation of real images is that the snowflakes are more densely concentrated in close proximity to the horizon line. In practice, this effect can be modeled by cutting the image into three parts through

Algorithm 4: fakeRain Mask Generation

input : Image size: length l and height h
output: fakeRain Mask M_r

```

1  $M_r = \emptyset$ 
2  $M_r \leftarrow P_a(\{0, \dots, l - 3\}, 0)$ 
   // use many agglomerate patterns in
   // the first line
3 for  $(i, j) \in (\{0, \dots, l - 3\}, \{0, \dots, h - 3\})$  do
4   if  $(i + j < \frac{h+l}{4}) \vee (l - i + j < \frac{h+l}{4})$  then
5      $M_r \leftarrow P_a(i, j) \vee \{\}$ 
     // sparsely add agglomerate
     // patterns below the V
6   else
7      $M_r \leftarrow P_p(i, j, t) \vee P_l(i, j, n) \vee \{\}$ 
     // sparsely add patch patterns
     // or line patterns above the V
8 end

```

two horizontal lines, as shown in Figure 9, and placing more dense snow patterns in the middle region, while maintaining the top and the bottom of the image relatively less populated by snow patterns. The generation of the mask for the *fakeSnow* attack is described in Algorithm 5. It proceeds in different ways based on the vertical coordinate j . In the middle part of the image, equally-spaced dense snow patterns are added to the *fakeSnow* mask (line 9 of Algorithm 5). In the upper part and lower part of the image, alternating rows of dense and sparse (i.e., largely spaced) snow patterns are added (lines 4-7 of Algorithm 5).

Algorithm 5: fakeSnow Mask Generation

input : Image size: length l and height h
output: fakeSnow Mask M_s

```

1  $M_s = \emptyset$ 
2 for  $j, \in \{0, 2, 4, \dots, h - 2\}$  do
3   if  $(j < \frac{h}{3} - 1) \vee j > \frac{2h}{3} - 1)$  then
     // upper and lower parts
4     if  $j \equiv 0 \pmod{4}$  then
5        $M_s \leftarrow P_s(\{0, 3, 6, 9, \dots, l - 2\}, j + 1)$ 
6     else // skip some snow patterns
7        $M_s \leftarrow P_s(\{0, 6, 12, \dots, l - 2\}, j + 1)$ 
8   else // middle part
9      $M_s \leftarrow P_s(\{0, 3, 6, 9, \dots, l - 2\}, j + 1)$ 
     // add dense snow patterns
10 end

```

F. fakeHail Attack

Compared to the snow, a hail scenario produce relatively larger ice balls perceived by the camera, as shown in Figure 10. Hence, the hail pattern is not modeled as a single pixel, but as an agglomerate of 8 pixels, as described in Algorithm 6.

Since the hail patterns appear irregularly, the *fakeHail* mask can be generated through a collection of hail patterns, as



Fig. 9. Mask for the fakeSnow attack, divided into three parts.



Fig. 10. Observation of hail conditions, which lead to the design of the hail pattern.

described in Algorithm 7. Note that the hail patterns are sparsely added, since, for each coordinate of the mask, the hail pattern can be added to the mask or not (line 3 of Algorithm 7).

IV. EVALUATING THE WEATHER ATTACK

A. Experimental Setup

We conducted the experiments on three different DNN models, which are the LeNet-5 [37], the ResNet-32 [4] and the CapsNet [38], trained for the CIFAR-10 dataset [39]. It is a collection of 50,000 training images and 10,000 testing images of size $32 \times 32 \times 3$, divided into 10 classes. An overview of the setup and tool-flow employed for conducting the experiments is shown in Figure 11.

Algorithm 6: Hail Pattern

input : Coordinate (x_0, y_0)
output: Hail Pattern P_h

```

1  $P_h = \emptyset$ 
2  $k = 0$ 
3 for  $i \leftarrow 0$  to 3 do
4   for  $j \leftarrow 0$  to 3 do
5     if
6        $(i = j \wedge i < 2) \vee (i + j = 3) \vee (i = 2 \wedge j \neq 2)$ 
7       then
8          $P_h \leftarrow pixel_k = (x_0 + i, y_0 + j, r_s, g_s, b_s)$ 
9          $k \leftarrow k + 1$ 
10    end
11  end
12 end

```

Algorithm 7: fakeHail Mask Generation

input : Image size: length l and height h
output: fakeHail Mask M_h

```

1  $M_h = \emptyset$ 
2 for  $(i, j) \in (\{0, \dots, l-4\}, \{0, \dots, h-4\})$  do
3    $M_h \leftarrow P_h(i, j) \vee \{ \}$ 
4 end

```

The LeNet, which is composed of two convolutional layers and two fully-connected layers followed by a softmax layer, has been trained for 200 epochs, using a batch size of 128, weight decay 0.0001, and a learning rate scheduler that progressively reduces its value from 0.05 to 0.0004. The 32-layer ResNet has been trained for 200 epochs, using a batch size of 128, weight decay 0.0001, and a learning rate scheduled to decrease from 0.1 to 0.001. The CapsNet, composed of a convolutional layer, a primary capsule layer, and a dynamic routing layer, has been trained for 200 epochs with a batch size of 64 and a learning rate equal to 0.001. For clean test images, we measure the accuracy values of 74.88%, 92.31%, and 79.82%, for the LeNet, ResNet, and CapsNet, respectively.

Afterwards, the *fakeWeather* masks have been applied to 200 testing samples and the attack success rate has been evaluated for every attack type (i.e., *fakeRain*, *fakeSnow* and *fakeHail*) and every DNN model. The training, as well as the implementation of the *fakeWeather* attacks and their evaluation, has been carried out using the Keras framework [40] with the TensorFlow [41] back-end, and executed on an ML-workstation equipped with two Nvidia GeForce RTX 2080 Ti GPUs.

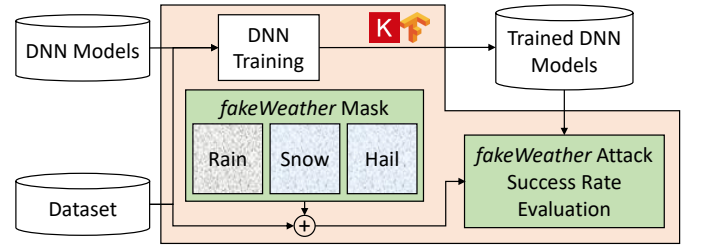


Fig. 11. Experimental setup and tool-flow for conducting our experiments.

B. fakeWeather Attacks Evaluation

Table I reports the results for the *fakeRain*, *fakeSnow* and *fakeHail* attacks in terms of Adversarial Success Rate (ASR), which corresponds to the ratio between the misclassified examples and all the tested examples. The results are compared with the state-of-the-art 1-pixel, 3-pixel, and 5-pixel attacks proposed by Su et al. [27]. Moreover, Figure 12 shows a collection of adversarial examples generated with the *fakeWeather* attacks.

fakeRain Evaluation

The *fakeRain* attack is successful for the LeNet and the ResNet, since their ASRs are 72% and 67%, respectively. The

TABLE I
EVALUATION OF THE ADVERSARIAL SUCCESS RATE (ASR) FOR THE THE
LENET, THE RESNET, AND THE CAPSNET ON THE CIFAR-10 DATASET.
OUR PROPOSED *fakeWeather* ATTACKS HAVE BEEN COMPARED TO THE
1-PIXEL, 3-PIXEL, AND 5-PIXEL ATTACKS [27].

ASR on Attack	LeNet	ResNet	CapsNet
1-pixel [27]	63%	34%	19%
3-pixel [27]	92%	79%	39%
5-pixel [27]	93%	79%	36%
fakeRain (ours)	72%	67%	36%
fakeSnow (ours)	75.5%	79.5%	30%
fakeHail (ours)	82.5%	78.5%	63%

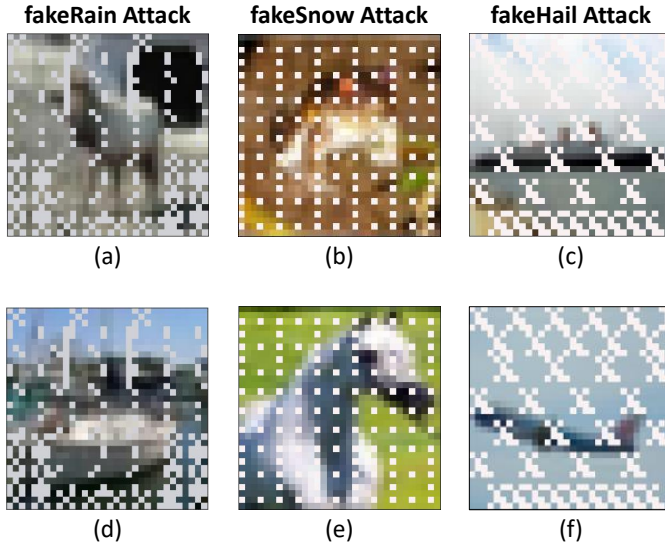


Fig. 12. Examples of a few images of the CIFAR-10 dataset on which the *fakeWeather* attacks are applied. (a) and (d): *fakeRain* adversarial examples. (b) and (e): *fakeSnow* adversarial examples. (c) and (f): *fakeHail* adversarial examples.

ResNet results slightly more robust than the LeNet, due to its deeper structure. The ASR falls to 36% for the CapsNet, since its architecture that groups the neurons into capsules, along with the dynamic routing, helps to better encode the spatial relations between features of the images. The example in Figure 12a shows the image of a deer on which the *fakeRain* mask is applied. All the three DNN models erroneously classify it as a “bird”, while its clean version is correctly classified as a “deer”. Similarly, the image in Figure 12d is incorrectly classified as a “truck” by the LeNet and the ResNet, while its clean version is correctly classified as a “ship”. However, the CapsNet still classifies this adversarial example as a “ship”.

fakeSnow Evaluation

For the *fakeSnow* attack, the relations between the ASRs of the three DNN models are similar to the observations made for the *fakeRain* attack, in which the CapsNet is more robust than the other CNNs. However, the ASR results are higher for the

ResNet, compared to the LeNet. The example in Figure 12b showing a frog with the *fakeSnow* mask is correctly classified by the CapsNet, while it is incorrectly classified as a “cat” by the ResNet and as a “truck” by the LeNet. Its clean version is correctly classified as a “frog” by all the DNNs. The horse in Figure 12e is correctly classified by the ResNet and the CapsNet, while the LeNet classifies it as a “deer”.

fakeHail Evaluation

The ASR relative to the *fakeHail* attack is significantly higher than the previous attacks, in particular for the CapsNet. Due to the relatively large perturbations imposed by the hail patterns (i.e., 8-pixel perturbations), the *fakeHail* mask can break the spatial relations learned by the CapsNet and lead to many misclassified samples. The example in Figure 12c represents a ship with the *fakeHail* mask that is incorrectly classified as an “airplane” by the LeNet and CapsNet, and as a “truck” by the ResNet. The image in Figure 12f is incorrectly classified as a “cat” by the LeNet, as a “deer” by the ResNet, and as a “frog” by the CapsNet, despite showing an airplane.

C. Case Studies: Output Probability Variations under *fakeWeather* attacks.

Towards a more comprehensive evaluation, we analyze the output probability variations when different types of *fakeWeather* attacks are applied to the LeNet, ResNet, and CapsNet models. For reference, the 10 classes of the CIFAR-10 dataset are associated with a digit 0 – 9 according to the convention in Table II.

TABLE II
CLASS LABELS FOR THE CIFAR-10 DATASET [39].

#	Class
0	airplane
1	automobile
2	bird
3	cat
4	deer
5	dog
6	frog
7	horse
8	ship
9	truck

Figure 13 shows how the image of a “truck” of the CIFAR-10 dataset is classified, for different *fakeWeather* attacks and different DNN models. The clean image is correctly classified as the class 9, i.e., “truck” by the LeNet, despite having a relatively low confidence (see pointer ① in Figure 13b). When each of the *fakeWeather* masks is applied, the LeNet predicts the image as a “frog” with quite high confidence (see pointer ② in Figure 13b). The probability variations for the ResNet assume a different behavior. While the clean image is correctly classified with high confidence (see pointer ③

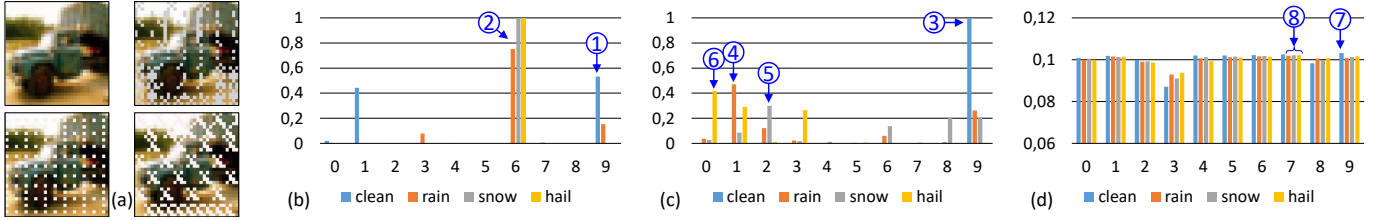


Fig. 13. Example showing a “truck” to which the *fakeWeather* attacks are applied. (a) Clean image, *fakeRain* image, *fakeSnow* image, and *fakeHail* image. (b) Output probabilities for the LeNet. (c) Output probabilities for the ResNet. (d) Output probabilities for the CapsNet.

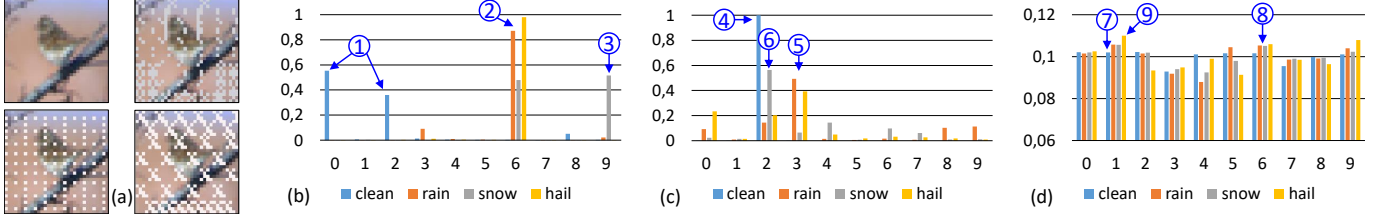


Fig. 14. Example showing a “bird” to which the *fakeWeather* attacks are applied. (a) Clean image, *fakeRain* image, *fakeSnow* image, and *fakeHail* image. (b) Output probabilities for the LeNet. (c) Output probabilities for the ResNet. (d) Output probabilities for the CapsNet.

in Figure 13c), the *fakeWeather* attacks produce different outcomes. With the *fakeRain* mask the image is classified as an “automobile” by the ResNet (see pointer ④ in Figure 13c), with the *fakeSnow* mask the highest probability belongs to the class “bird” (see pointer ⑤ in Figure 13c), and the adversarial *fakeHail* image is classified as an “airplane” by the ResNet (see pointer ⑥ in Figure 13c). The output probabilities for the CapsNet, while they are more concentrated in the middle values, i.e., 1/10, show that the clean image is correctly classified (see pointer ⑦ in Figure 13d), while for all the *fakeWeather* attacks, the highest probability belongs to the class “horse” (see pointer ⑧ in Figure 13d).

Figure 14 shows the output probability variations associated to a “bird” image of the CIFAR-10 dataset. The clean image is already incorrectly classified as an “airplane” by the LeNet (see pointer ① in Figure 14b). With the *fakeRain* or the *fakeHail* mask, the LeNet classifies the adversarial image as a “frog” (see pointer ② in Figure 14b), while the adversarial *fakeSnow* image is classified as a “truck” (see pointer ③ in Figure 14b). The ResNet correctly classifies the clean image as a “bird” with high confidence (see pointer ④ in Figure 14c). The *fakeRain* and *fakeHail* adversarial images are classified as a “cat” (see pointer ⑤ in Figure 14c), while the *fakeSnow* is unsuccessful, since the image is still correctly classified by the ResNet (see pointer ⑥ in Figure 14c), even though with lower confidence than the clean image. The CapsNet correctly classifies the clean image with very narrow difference w.r.t. the other classes (see pointer ⑦ in Figure 14d). The *fakeRain* and *fakeSnow* attacks produce adversarial images that are classified as a “frog” by the CapsNet (see pointer ⑧ in Figure 14d), while the image with the *fakeHail* mask is correctly classified by the CapsNet (see pointer ⑨ in Figure 14d).

D. Results Discussion and Comparison

To summarize, given the above-discussed results, we can make the following considerations:

- All the *fakeWeather* attacks produce a high ASR for the LeNet and ResNet ($ASR > 65\%$).
- The *fakeHail* attack is the strongest, since it achieves an ASR equal to 63% for the CapsNet and higher for the other DNNs.

Compared to the methods of Su et al. [27], our *fakeWeather* methods have higher ASR than the 1-pixel attack for every DNN model (see Table I). However, the 3-pixel and 5-pixel attacks have higher ASR than our methods. Note that the approach used by Su et al. is based on an evolutionary algorithm that requires several queries, while our methodology does not require any query. Yet, the ASR relative to the CapsNet for the *fakeHail* attack is 27% higher than the 5-pixel attack.

E. Future Outlooks and Applicability

From another perspective, our contributions, other than a methodology for generating adversarial attacks in real-time without queries, can be viewed as a data augmentation methodology for generating synthetic samples of weather conditions. We envision the possibility of enlarging the dataset with images that contain *fakeWeather* masks and train DNN-based classifiers more robustly to such atmospheric phenomena, in a similar way as the adversarial training’s functionality [24]. Since the only information required is the image size, its high scalability makes our *fakeWeather* attack methodology suitable to any vision-based outdoor application.

V. CONCLUSION

In this paper, we presented *fakeWeather* attacks, adversarial attacks for DNNs that emulate the natural weather conditions. Our methodology consists of observing a series of images that capture the effects of such conditions perceived by the camera lens, and modeling a set of patterns to create dedicated *fakeRain*, *fakeSnow*, and *fakeHail* masks as a collection of these patterns. Hence, these sets of perturbations make the adversarial image a plausible input to the DNN. Our proposed attack is conducted in *true black-box* settings, in which the adversary has no access to the DNN model, its parameters, and its output. The evaluation of *fakeWeather* attacks on different DNN models (Convolutional Neural Networks and Capsule Networks) highlights noticeable adversarial success rates.

ACKNOWLEDGMENT

This work has been supported in part by the Doctoral College Resilient Embedded Systems, which is run jointly by the TU Wien's Faculty of Informatics and the UAS Technikum Wien. This work was also supported in parts by the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001, Center for CyberSecurity (CCS), funded by Tamkeen under the NYUAD Research Institute Award G1104, and Center for Artificial Intelligence and Robotics (CAIR), funded by Tamkeen under the NYUAD Research Institute Award CG010.

REFERENCES

- [1] L. Jiao *et al.*, "A survey of deep learning-based object detection," *IEEE Access*, 2019.
- [2] S. Kuutti *et al.*, "A survey of deep learning applications to autonomous vehicle control," *arXiv*, 2019.
- [3] R. Lee *et al.*, "Deep neural network-based enhancement for image and video streaming systems: A survey and future directions," *ACM Comput. Surv.*, 2021.
- [4] K. He *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [5] M. Capra *et al.*, "Hardware and software optimizations for accelerating deep neural networks: Survey of current trends, challenges, and the road ahead," *IEEE Access*, 2020.
- [6] M. Capra *et al.*, "An updated survey of efficient hardware architectures for accelerating deep convolutional neural networks," *Future Internet*, 2020.
- [7] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," in *CCS*, 2018.
- [8] N. Dalvi *et al.*, "Adversarial classification," in *KDD*, 2004.
- [9] D. Lowd and C. Meek, "Adversarial learning," in *KDD*, 2005.
- [10] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *ICLR*, 2014.
- [11] M. Shafique *et al.*, "Towards energy-efficient and secure edge AI: A cross-layer framework ICCAD special session paper," in *ICCAD*, 2021.
- [12] S. Dave *et al.*, "Special session: Towards an agile design methodology for efficient, reliable, and secure ML systems," in *VTS*, 2022.
- [13] M. Shafique *et al.*, "Robust machine learning systems: Challenges, current trends, perspectives, and the road ahead," *IEEE Design & Test*, 2020.
- [14] B. Luo *et al.*, "Towards imperceptible and robust adversarial example attacks against neural networks," in *AAAI*, 2018.
- [15] F. Croce and M. Hein, "Sparse and imperceptible adversarial attacks," in *ICCV*, 2019.
- [16] A. Marchisio *et al.*, "Capsattacks: Robust and imperceptible adversarial attacks on capsule networks," *arXiv*, 2019.
- [17] J. Chen *et al.*, "Hopskipjumpattack: A query-efficient decision-based attack," in *SP*, 2020.
- [18] F. Khalid *et al.*, "Fadec: A fast decision-based attack for adversarial machine learning," in *IJCNN*, 2020.
- [19] M. Xue *et al.*, *DNN Intellectual Property Protection: Taxonomy, Attacks and Evaluations (Invited Paper)*. 2021.
- [20] D. Willmott *et al.*, "You only query once: Effective black box adversarial attacks with minimal repeated queries," *arXiv*, 2021.
- [21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.
- [22] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv*, 2016.
- [23] S.-M. Moosavi-Dezfooli *et al.*, "Deepfool: A simple and accurate method to fool deep neural networks," in *CVPR*, 2016.
- [24] A. Madry *et al.*, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.
- [25] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," *arXiv*, 2016.
- [26] N. Narodytska and S. Kasiviswanathan, "Simple black-box adversarial attacks on deep neural networks," in *CVPRW*, 2017.
- [27] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *arXiv*, 2017.
- [28] S.-M. Moosavi-Dezfooli *et al.*, "Universal adversarial perturbations," in *CVPR*, 2017.
- [29] C. Zhang *et al.*, "CD-UAP: class discriminative universal adversarial perturbation," in *AAAI*, 2020.
- [30] D. Gragnaniello *et al.*, "Perceptual quality-preserving black-box attack against deep learning image classifiers," *arXiv*, 2019.
- [31] T. B. Brown *et al.*, "Adversarial patch," *arXiv*, 2017.
- [32] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning visual classification," in *CVPR*, 2018.
- [33] M. Sharif *et al.*, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *CCS*, 2016.
- [34] Y. Man, M. Li, and R. Gerdes, "Ghostimage: Remote perception domain attacks against camera-based image classification systems," in *USENIX RAID*, 2020.
- [35] D. Temel, M.-H. Chen, and G. AlRegib, "Traffic sign detection under challenging conditions: A deeper look into performance variations and spectral characteristics," in *IEEE TITS*, 2019.
- [36] L. Zhai *et al.*, "It's raining cats or dogs? adversarial rain attack on DNN perception," *arXiv*, 2020.
- [37] Y. Lecun *et al.*, "Gradient-based learning applied to document recognition," *IEEE*, 1998.
- [38] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *NeurIPS*, 2017.
- [39] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Tech. Report*, 2009.
- [40] F. Chollet *et al.*, "Keras," *GitHub*, 2015.
- [41] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *OSDI*, 2016.