







# Benchmarking Robustness Beyond $l_p$ Norm Adversaries

Akshay Agarwal<sup>1,2</sup> , Nalini Ratha<sup>1,2</sup> , Mayank Vatsa<sup>1,2</sup> ,  
and Richa Singh<sup>1,2</sup> 

<sup>1</sup> University at Buffalo, Buffalo, NY, USA

{aa298,nratha}@buffalo.edu, {mvatsa,richa}@iitj.ac.in

<sup>2</sup> IIT Jodhpur, Karwar, India

**Abstract.** Recently, a significant boom has been noticed in the generation of a variety of malicious examples ranging from adversarial perturbations to common noises to natural adversaries. These malicious examples are highly effective in fooling almost ‘any’ deep neural network. Therefore, to protect the integrity of deep networks, research efforts have been started in building the defense against these anomalies of the individual category. The prime reason for such individual handling of noises is the lack of one unique dataset which can be used to benchmark against multiple malicious examples and hence in turn can help in building a true ‘universal’ defense algorithm. This research work is an aid towards that goal that created a dataset termed “wide angle anomalies” containing 19 different malicious categories. On top of that, an extensive experimental evaluation has been performed on the proposed dataset using popular deep neural networks to detect these wide-angle anomalies. The experiments help in identifying a possible relationship between different anomalies and how easy or difficult to detect an anomaly if it is seen or unseen during training-testing. We assert that the experiments in seen and unseen category attack training-testing reveals several surprising and interesting outcomes including possible connection among adversaries. We believe it can help in building a universal defense algorithm.

## 1 Introduction

To protect the integrity of deep neural networks, defense algorithms against modified images are proposed; although, the majority of them deal with a unique category of malicious examples and have shown tremendous success in that [1, 5]. It is seen in existing research that the defense algorithms targeting specific attacks fail in identifying the malicious data coming from the same or different malicious examples categories [12, 51, 54]. This ineffectiveness can be seen as a blind spot that leaves a space for an attacker to attack the system and perform undesired tasks. Therefore, looking at the severeness of this existing limitation, we have studied numerous adversaries intending to develop universal security. In this research, we divide the malicious examples into three broad categories: (i) common corruptions [21, 41], (ii) adversarial perturbations [3, 4, 33], and (iii) natural

adversary [8, 22, 40, 53]. These different categories follow different rules in crafting the perturbation and hence have significant distribution differences among each other. For instance, in the common corruptions, the noises are uniformly distributed, adversarial perturbations affect the critical regions, and natural examples might occur due to cluttered background or low foreground region [31, 32].

To study such broad robustness, a unique dataset covering such malicious examples is a necessity, and therefore, we have first curated the “*wide angle anomalies*” examples dataset covering the three broad categories mentioned above. In total, the dataset contains approximately 60,000 images belonging to 20 classes including real and various malicious generation algorithms. Surprisingly, the majority of the malicious examples generation algorithm belonging to the above categories are not explored in existing research, and hence security against them is still a serious concern. Once the wide angle anomalies dataset is prepared we have performed an extensive experimental evaluation using deep convolutional networks to identify these malicious examples. Henceforth, the analysis presented in this paper is an act of benchmarking robustness against such a broad umbrella of malicious examples and in turn helps in building a universal robustness system. We find that there is a connection among the different anomalies coming from the same broad group and can also be used to detect other groups’ adversaries. In brief,

- We present a large-scale malicious examples dataset covering 19 different attack generation algorithms. The images consist of a wide distribution shift among the malicious examples due to contradictory ways of generation;
- A benchmark study is presented using a deep convolutional network for the detection of such broad malicious examples categories. The experimental results corresponding to both seen and unseen malicious category in training and testing reveals several interesting and thoughtful insights. We assert the presence of such wide-angle malicious examples dataset and the benchmark study can significantly boost the development of universal robustness.

## 2 Related Work

In this section, a brief overview of the existing works developed to counter the malicious examples and protect the integrity of the deep neural networks is presented. As mentioned earlier, the majority of the defense work is focused on defending one specific type of malicious example. We first provide a brief overview of the existing defense work tackling artificial adversarial perturbations followed by the defense algorithms countering common corruptions. To the best of our knowledge, no work so far has been presented to build a defense against natural adversarial examples.

The defense algorithms against artificial adversarial perturbations are broadly grouped into (i) detection based, (ii) mitigation based, and (iii) robustness. In the mitigation case, a denoising algorithm is presented to map the noisy data to the clean counterpart. The aim is to reduce the impact of the adversarial perturbation so that the accuracy of the classifier can be restored

[15, 18, 44]. Robustness-based defenses are one of the most popular defense techniques to make the classifier resilient against noisy test data. The robustness in this category of algorithms is achieved by training the network by utilizing data augmentation techniques including adversarial training [9, 46, 47]. Adversarial training is one of the powerful defense techniques where the classifiers are either trained or fine-tuned using the adversarial images. However, the probable limitations of the techniques are the computational cost and generalizability against the unseen perturbations [42, 50, 54]. To address this issue several general-purpose data augmentation techniques are also proposed [6, 7, 10]. These data augmentation-based defenses also overcome another limitation of adversarial training which is maintaining the performance on the clean images that significantly drops in adversarial training. Another popular and most effective defense against artificial adversarial perturbation is the development of a binary classifier. Recently, several generalized adversarial perturbation detection algorithms are proposed that either utilize the handcrafted features, deep classifiers or a combination of both [1, 5, 20, 30, 52].

In contrast to the defense against adversarial perturbations, limited work has been done so far to protect against common corruption. Similar to increasing the robustness against adversarial perturbations, data augmentation is one of the favorite defenses against common corruptions as well [34, 36]. In other forms of defense, recently, Schneider et al. [45] have proposed to mitigate the covariate shift by replacing the batch normalization statistics computed over the clean data with the statistics computed over corrupted samples. Another recent work utilizes the mixture of two deep CNN models biased towards low and high-frequency features of an image [43]. The reason might be that noise signals are considered high-frequency information and the author aims to improve the robustness against high-frequency features. The major limitation of the defenses proposed so far is the generalization against the unseen corruptions [14, 17, 35], computational cost, and degradation performance on the clean or in-distribution images. Interestingly, very limited work has tried to identify/detect the corrupted examples and the majority of them tried to improve the robustness of the model directly which in turn leads to the degradation performance on clean images. The issue of common corruption has recently been explored in other computer vision tasks or models as well such as semantic segmentation [27] and transformers [39]. Therefore, looking at the severity of both common corruptions and artificial perturbations, a resilient defense is critical. Apart from handling these well-known malicious examples, detecting the advanced or recently explored natural adversarial examples [8, 22, 31] is also important in building a universal defense mechanism. To the best of our knowledge, recent work [2] is the only work which has started an effort in building a unified defense system.

### 3 Proposed *Wide Angle Anomalies* Dataset

In this research, we have selected different malicious examples generation algorithms which we have broadly grouped into two groups: (i) common corruptions



**Fig. 1.** Samples from our proposed wide angle anomalies dataset cover a broad spectrum of malicious examples. The covariate shift between the different classes including real is evident which makes the generalizability a tough task in handling a broad spectrum of anomalies.

**Table 1.** Parameters of the common corruption used.

Noise	GN	UN	SPN	SN	SPKL
Param.	0.08	0.1	0.1	60	0.15

and (ii) adversarial examples. The adversarial examples consist of two broad categories (i) artificial perturbations and (ii) natural adversary.

### 3.1 Common Corruptions

For the generation of common corruption-induced malicious examples five different popular variants are selected namely Gaussian noise (GN), salt & pepper noise (SPN), uniform noise (UN), shot noise (SN), and speckle noise (SPKN). Each of the selected corruption is applied with low severity with a twofold aim: first is to fool the deep classifiers but at the same time keep the perceptibility of noise pattern minimal. The parameters used with individual common corruption are given in Table 1.

### 3.2 Adversarial Examples

In contrast to the common corruption, adversarial examples contain the perturbation generated using the classifier itself utilizing its ingredients including image gradient and decision probabilities. For adversarial examples, both artificial perturbation optimization algorithms and natural adversarial examples generation algorithms are selected.

**Artificial Perturbations.** For artificial adversarial examples, five benchmark algorithms namely fast gradient sign method (FGSM) [19], basic iterative method (BIM) also known as iterative FGSM (IFGSM) [28], projected gradient descent (PGD) [33], DeepFool [38], and Universal perturbation [37] are adopted. FGSM is one of the simplest and most effective adversarial perturbation generation algorithms. It works on the manipulation using the gradient information

computed over an image to maximize the loss over the modified image. Mathematically, it can be described as follows:  $X^* = X + \eta \cdot \text{sign}(\nabla_X J(X, Y_{true}))$ .  $X$  and  $X^*$  represent the clean and FGSM adversarial images, respectively.  $\eta$  controls the strength of added perturbation optimized through the loss function  $J$  computed over image  $X$  and its associate true label  $Y_{true}$ .  $\nabla_X$  is the gradient concerning  $X$  and  $\text{sign}$  represents the sign function. Even though  $\eta$  tries to control the perturbation visibility it is still highly perceptible with a naked eye and hence to improve that the iterative variants are proposed by Kurakin and Goodfellow [28]. It can be described as follows:

$$X_0^* = X$$

$$X_N^* = \text{Clip}_{X,\epsilon}(X_{N-1}^* + \alpha \cdot \text{sign}(\nabla_X J(X_{N-1}^*, Y_{true})))$$

where,  $\text{Clip}_{X,\epsilon}$  represents the scaling of an image in the range of  $[X - \epsilon, X + \epsilon]$ . Due to its nonlinearity in the gradient direction, several iterations are required to optimize the perturbation. It makes the generated perturbation more effective as compared to FGSM. Compared to the above gradient-based perturbation, the deepfool attack is based on the minimization of the  $L_2$  norm and aims to make sure the adversarial examples jump the decision hyperplane developed by the classifier. The idea is to perturb an image iteratively and with each iteration, the closest decision surface is assumed to be fooled by the updated image. Madry et al. [33] have proposed the PGD attack which is also considered the strongest first-order universal adversary. The optimization iteratively searches for a perturbation vector that minimizes a  $l_p$  norm ball around the clean image. The above artificial perturbation generation algorithms whether simple or complex, generate the noise vector individually for each image. To optimize a single perturbation vector that can be applied to multiple images, a universal perturbation is also presented in the literature [37]. The above selected perturbation reflects the wide variety in the generation of adversarial perturbation and hence makes the study of universal defense interesting and a thoughtful step.

**Natural Adversary.** Recently, several researchers have explored the natural way of crafting adversarial examples. One such way is proposed by Hendrycks et al. [22]. The authors have downloaded the natural images of 200 classes from multiple image hosting websites that can fool the ResNet-50 classifier. In total, the dataset contains 7,500 adversarial images and is termed Imagenet-A. Later, Li et al. [31] identify several bottlenecks of the natural adversarial examples in ImageNet-A. It is found that the background in the adversarial examples of ImageNet-A is more cluttered due to the presence of multiple objects that might be a possible reason for distribution shift and leads to misclassification. Another possible drawback of the images is that the foreground region in an image occupies a small part as compared to the background. The authors show that removing these limitations can significantly boost the performance of several ImageNet trained models and hence the need for an intelligent way of crafting a natural adversary is highlighted. For that, the authors have presented an Imagenet-A-Plus dataset by minimizing the background information in an image and leaving

only one salient object region covering a large portion of an image. The dataset is generated from the images of the ImageNet-A dataset by first filtering out the images containing object proportion 8 times less than the object proportion in the ImageNet images. The filtered images are passed through the ResNet-50 model and the background is clipped from the selected adversarial images so that the target object proportion in images can be increased. Hosseini et al. [24] have proposed another way of generating natural adversarial examples by shifting the color information in images. The assumption of image generation is based on the assertion that the human visual system is biased towards shape while classifying an object as compared to the color information [29]. The authors have utilized the HSV color space due to its closeness to the human visual system as compared to the RGB space. The authors have modified the hue and saturation components of an image keeping the value component intact. The adversarially color shifted images are generated by solving the following optimization problem:

$$\begin{aligned} & \min |\delta_S|, \quad s.t. \\ & \begin{cases} X_H^* = (X_H + \delta_H) \bmod 1 \\ X_S^* = \text{Clip}(X_S + \delta_S, 0, 1) \\ X_V^* = X_V \end{cases} \end{aligned}$$

where,  $\delta_H$  and  $\delta_S$  represent the shift introduced in the hue  $X_H$  and saturation  $X_S$  components of an image  $X$ , respectively.  $X_V$  is the value component of an image. Both  $\delta_H$  and  $\delta_S$  are scalar values only. The authors have used the 1000 iteration to perturb the color components or till the modified image is misclassified by the VGG classifier, whichever happens first. The authors have reported the success of the attack on the CIFAR10 dataset only. In this research, to keep the adversarial examples of one dataset, we have trained the VGG classifier on the selected ImageNet subset and generated the color shift semantic adversarial examples.

In contrast to the above natural examples which either utilize the limitation of the classifier of not being trained on the kind of images that occur in the adversarial dataset such as ImageNet-A and ImageNet-A-Plus or color shifting the images, Agarwal et al. [8] have utilized the noise inherited at the time image acquisition. When the images are captured from the camera they passed through several intermediate steps, the authors assert that these steps induced some form of manipulation or the environment can itself be noisy. The intuition of the authors is that can this inherited noise be used as an adversarial noise. The inherited noise vector is extracted using several image filtering techniques such as Laplacian and Gaussian. The adversarial examples generation process can be described as follows:

$$\begin{aligned} \text{Noise} &= X - \phi(X) \\ X^* &= \text{Clip}(X \circledast (\psi \cdot \text{Noise}), 0, 1) \end{aligned}$$

where,  $\text{Noise}$  is generated by subtracting the acquired clean image ( $X$ ) from its filtered version obtained by applying any image filtering technique  $\phi$ .  $\psi$  is a scalar value controlling the strength of the noise.  $\circledast$  represents the noise manipulation operator that is either added ( $\oplus$ ) or removed ( $\ominus$ ) from the image with the

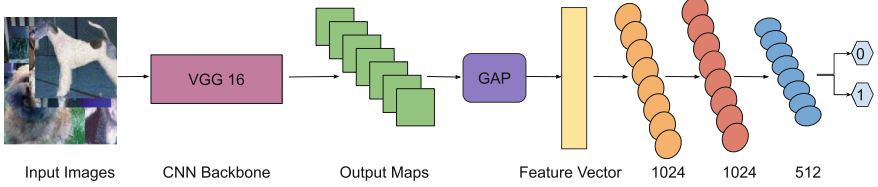
following assertion: (i) added (referred to as -P) the subtracted component by assuming it as a noise vector and (ii) removed (referred as -S) on the fact that the noise is a high-frequency feature and removing that feature can fool the classifiers which are found highly biased towards shape and texture [16, 23, 43]. In this research, we have used the Integral, laplacian, and laplacian of Gaussian as three image filtering techniques due to their high effectiveness as compared to the other filtering methods used in the paper [8]. The adversarial examples generated using the above technique are referred to as camera induced perturbation (CIPer) as termed in the original paper. In brief, various classes along with the number of images covered in the proposed research can be summarized as follows:

- Real/Clean Images (3,000)
- Common Corruptions (15,000)
  - Gaussian Noise (*GN*) (3,000)
  - Uniform Noise (*UN*) (3,000)
  - Salt & Pepper Noise (*SPN*) (3,000)
  - Shot Noise (*SN*) (3,000)
  - Speckle Noise (*SPKN*) (3,000)
- Adversarial Images
  - Artificial Perturbations (15,000)
    - \* FGSM
    - \* IFGSM
    - \* PGD
    - \* DeepFool (*DF*)
    - \* Universal (*Univ.*)
  - Natural Examples (8,763)
    - \* Subset of ImageNet-A (*IN-A*)
    - \* Subset of ImageNet-A-Plus (*IN-A-P*)
    - \* Semantic Color-Shift Examples (*CS*)
  - Camera Induced (18,000)
    - \* Integral filtering (*Int-P* and *Int-S*)
    - \* Laplace filtering (*Lap-P* and *Lap-S*)
    - \* Laplace of Gaussian filtering (*LoG-P* and *LoG-S*)

To generate the malicious images, 3,000 clean images from the validation set of the ImageNet dataset are first selected [13]. Later, each malicious examples generation algorithm is applied to the selected images except ImageNet-A and ImageNet-A-Plus. The images in this category are directly taken from the images provided by the original contributors. In total, the proposed wide angle anomalies dataset consists of 3,000 clean images and 56,763 malicious images. For the experimental purpose, the first 1500 images of each class are used for training, and the last 1500 images are used for evaluation.

## 4 Experimental Results and Analysis

In this research, the aim is to study the universal robustness by detecting these wide-angle adversaries, i.e., classifying the images into either real or modified



**Fig. 2.** Malicious examples detection architecture used in this research.

classes. For that, a binary classifier is developed using VGG-16 [48] as a backbone architecture<sup>1</sup>. The robustness performance acts as a benchmark for the study of universal defense by identifying a possible connection between different groups of adversaries. The proposed malicious examples detection architecture is shown in Fig. 2. We have set the gradient of the first 10 layers of VGG equal to zero and finetuned the remaining layers along with training the newly added dense layers from scratch. The architecture is trained for 30 epochs using Adam optimizer where the batch size is set to 32 and initial learning used is  $1e^{-4}$ .

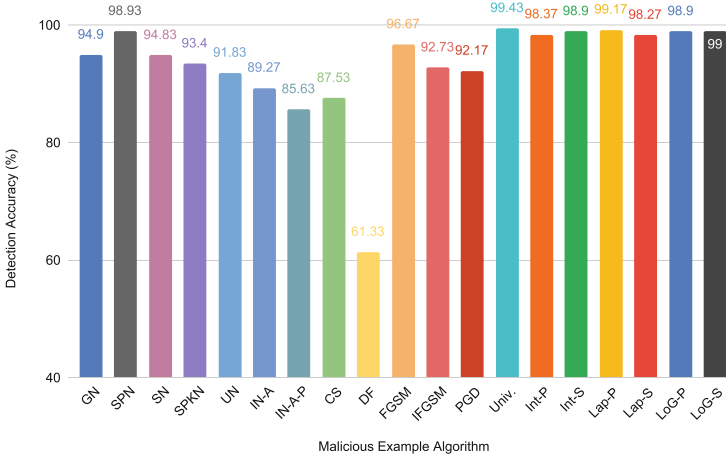
#### 4.1 Results and Analysis

We have performed an extensive experimental analysis on the collected wide angle anomalies dataset. The experiments are performed in several generalized settings: (i) seen attack generation algorithm such as PGD vs. PGD, (ii) unseen attack generation algorithm such as PGD vs. FGSM, and (iii) unseen attack types such as Natural adversary vs. common corruptions. First, we will present the results and analysis of the seen attack training and testing scenarios followed by the experimental observations on unseen attack settings. In the end, the connection between different malicious examples is established by testing the malicious examples detection algorithm trained on entirely different malicious examples.

**Seen Attack Results.** In total, in the proposed dataset there are 20 classes belonging to one real and 19 attack class. Therefore, in the seen attack setting, a total of 19 classifiers are trained individually on each attack training data and tested on the same attack type using the testing set. The results of these experiments are shown in Fig. 3. The analysis can be broken down into the following ways: (i) global analysis and (ii) local analysis. In the global analysis, it can be seen that the DF (DeepFool) attack is found highly challenging to detect, i.e., yielding the lowest detection accuracy value of 61.33%. Whereas, the remaining

<sup>1</sup> While the results are reported using VGG, similar evaluation analysis (with  $\pm 1$ –12% as shown in Table 5) is observed across wide range of backbone networks including Xception [11], InceptionV3 [49], DenseNet121 [26], and MobileNet [25]. However, VGG tops each network in the majority of the cases and is hence chosen for detailed study in the paper.





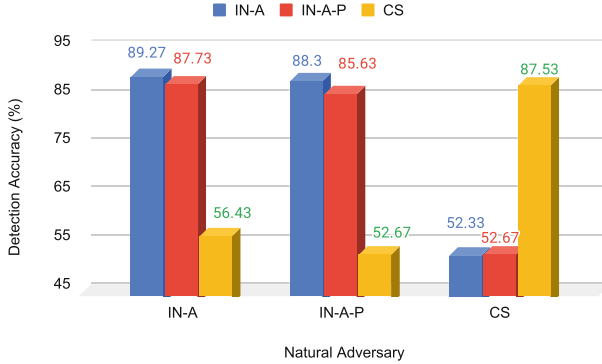
**Fig. 3.** Malicious examples detection in the seen attack setting, i.e., where training and testing attack generation algorithm is same.

**Table 2.** Common corruption detection performance in unseen noise training testing conditions. – represents the seen noise training-testing setting and the results are reported in Fig. 3.

Train ↓ Test →	GN	SPN	SN	SPKN	UN	Avg. $\pm$ SD
GN	–	<b>97.90</b>	94.80	93.00	85.27	<b>92.74 <math>\pm</math> 5.37</b>
SPN	57.10	–	<b>58.30</b>	55.63	51.67	<b>55.67 <math>\pm</math> 2.88</b>
SN	94.17	<b>95.83</b>	–	94.47	87.47	<b>92.98 <math>\pm</math> 3.74</b>
SPKN	92.80	<b>95.57</b>	93.67	–	86.73	<b>92.19 <math>\pm</math> 3.82</b>
UN	91.87	91.90	<b>92.00</b>	91.97	–	<b>91.93 <math>\pm</math> 0.06</b>

attack detection yields a high accuracy value of at least 85.63% reflecting that defending against adversarial attacks even coming from a variety of algorithms is not difficult even from a simple classification architecture. However, this might give us a false sense of security as in reality all possible attacks might not be known beforehand.

In the case of local analysis, an observation concerning the different classes of malicious examples can be described. For instance, when the common corruptions are aimed to detect, it is found that *uniform noise* (UN) is one of the toughest corruption to detect. We went ahead to identify the potential reason for such lower performance and found that the perceptibility of the uniform noise is low (last column of common corruption in Fig. 1) as compared to other perturbations and it is approximately similar to artificial adversarial perturbations. It can be seen from another point that the SPN noise is found highly perceptible and hence yield a higher detection performance. In the case of natural



**Fig. 4.** Unseen natural adversarial examples detection.

adversarial examples, ImageNet-A-Plus images are found less detectable as compared to the other natural adversary. The ImageNet-A-Plus (IN-A-P) can be seen as an advanced version of ImageNet-A (IN-A) where the cluttered background and foreground object region is enhanced. It might be the possible reason for the lower detection performance of these examples. The proposed binary classification algorithm is also found effective in detecting the color shift (CS) semantic natural adversarial examples. The detection performance on each artificial adversarial perturbation except DF is significantly high where the lowest detection accuracy observed is 92.17%. The universal perturbation images are found easiest to detect and demonstrated approximately perfect detection accuracy (99.43%). In comparison to the other malicious attack classes, the detection performance across each variety of CIPer noise is at least 98.27% which makes it less effective in terms of its detection. We want to highlight that such high detection performance observed is in the case where the detection algorithm has seen each malicious class while optimizing the network parameters and therefore, might not be a true indicator of the complexity of the malicious class.

**Unseen Attack Detection.** From the experimental analysis in seen attack setting, we have seen the impression that it might not be difficult to identify the malicious examples; however, such an impression can be dangerous until the evaluation has been performed under an unseen attack testing setting. Therefore, we have extensively evaluated the generalizability concern and showcased how easy or difficult the detection of malicious examples is in open-world settings. On the common corruption, five-fold cross-validation experiments are performed where at each time one noise type is used for training, and the remaining are used for testing. From the global view, each attack is found similarly generalized in detecting the unseen noise variation except Salt & Pepper noise (SPN). The training on SPN noise is found highly ineffective in identifying other noise variations. While the UN corruption performs similar to other corruptions, the variation in its detection performance is the smallest. Through a multi-fold

**Table 3.** Artificial adversarial perturbation detection performance in unseen noise training testing conditions. – represents the seen perturbation training-testing setting and reported in Fig. 3.

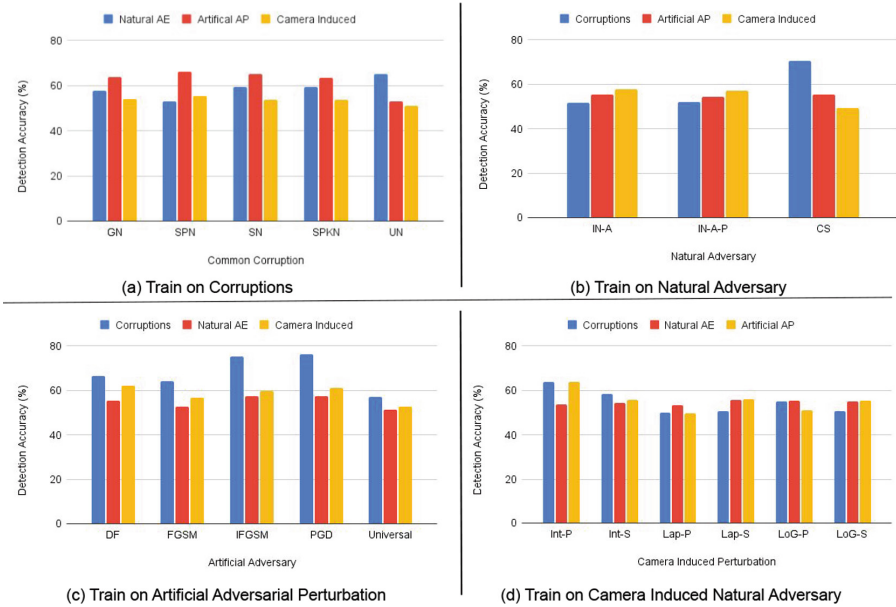
Train ↓ Test →	DF	FGSM	IFGSM	PGD	Univ.	Avg. ± SD
DF	–	89.60	77.03	77.00	<b>94.47</b>	<b>84.52 ± 8.90</b>
FGSM	57.63	–	71.20	71.33	<b>97.80</b>	<b>74.49 ± 16.82</b>
IFGSM	59.10	94.37	–	92.87	<b>95.30</b>	<b>85.41 ± 17.57</b>
PGD	59.10	94.43	92.17	–	<b>95.37</b>	<b>85.27 ± 17.50</b>
Univ.	55.00	<b>62.07</b>	51.40	51.30	–	<b>54.94 ± 5.05</b>

**Table 4.** Camera induced adversarial perturbation detection performance in unseen noise training testing conditions.

Train ↓ Test →	Int-P	Int-S	Lap-P	Lap-S	LoG-P	LoG-S
Int-P	–	58.77	50.00	56.07	<b>70.63</b>	64.87
Int-S	54.33	–	49.27	54.23	52.60	<b>63.60</b>
Lap-P	49.63	49.53	–	49.50	<b>86.47</b>	49.50
Lap-S	54.70	56.20	49.10	–	49.10	<b>98.87</b>
LoG-P	58.40	50.67	<b>95.23</b>	49.43	–	49.43
LoG-S	52.43	55.17	49.53	<b>96.63</b>	49.53	–

cross-validation experiment it is observed that the generalizability of the unseen corruption detection is somewhat better; however, needs further attention and evaluation against other categories. In an interesting observation, it can also be noticed that the detectors trained on each noise yield the lowest performance on the uniform noise, and the performance on SPN noise is highest. The results of this experimental analysis are reported in Table 2.

In the second unseen attack setting, natural adversary examples are chosen where a three-fold cross-validation experimental evaluation has been performed. Similar to common corruption only one attack is used for training and others are used for testing. This kind of setting represents the worst-case performance where it might be possible that only one type of attack is available for training or in other words limited knowledge about the adversary is known. In contrast to generalizability in common corruption, the robustness in the detection of the natural adversary is poor. The ImageNet-A (IN-A) and ImageNet-A-Plus (IN-A-P) being the same category of natural adversary yields similar and high detection performance; whereas the color shift semantic adversary is contradictory to these adversaries and yields lower accuracy. It can also be seen from the detection performance: (i) when IN-A or IN-A-P adversary is used for training it yields low detection performance on the color shift adversary (CS) and (ii) when the CS



**Fig. 5.** Unseen malicious type examples detection. Each attack of four malicious examples category (common corruption, natural adversary, artificial adversary, and camera induced perturbation) are used to train the detection algorithm and evaluated on the unseen malicious attack category.

adversary is used in training and IN-A and IN-A-P adversary used for evaluation, the performance is closed to random chance value. In between IN-A and IN-A-P, the IN-A is found robust in detecting the other natural adversaries. The quantitative finding of these experimental evaluations is reported in Fig. 4.

Another unseen attack detection performance is performed on the artificial adversarial perturbation images. We have noticed in the seen attack training-testing setting that the DF is highly challenging to detect as compared to other malicious examples including artificial adversarial perturbations. In the unseen attack setting as well it is observed that the generalization performance on the DF perturbation is the lowest among all the artificial perturbations. Whereas, the universal perturbation is found easily detectable and yields at least 94.47% detection accuracy even if the detector has not seen the perturbation while training. In surprising observations, the universal perturbation which was found highly detectable shows poor generalization in detecting other perturbations; whereas, the perturbation (DF) which is found complex in detection found significantly generalized in detecting unseen adversarial perturbations. The numerical analysis related to the above observations is reported in Table 3.

In final unseen attack detection experiments, the camera-induced noises are used for training and testing. The noise is extracted using three image filtering operations and applied in two forms (addition and subtraction). Therefore, a

total of *six fold* cross-validation experiments are performed to evaluate the generalizability of malicious examples detection networks. The camera-induced perturbations which were found almost perfectly detectable in seen attack training-testing are found complex in the unseen testing setting. In brief, when the detector is trained on the noisy examples obtained using addition operation, it yields higher performance in the detection of unseen attack images obtained using addition operation and yields poor performance generated using subtraction operation even same image filtering is applied. For instance, *Lap-P* trained detector yields more than 86% detection performance on the *LoG-P* images; whereas, it yields random chance accuracy (49.50%) on the *LoG-S* adversarial images. A similar interesting observation is observed in the adversarial examples obtained using subtraction of camera noise. For instance, *Int-S* trained malicious examples detector yields 11% better accuracy when the adversarial examples are obtained using LoG filtering with subtraction operation as compared to the addition operation. The quantitative results are shown in Table 4.

**Unseen Malicious Type Detection.** In the final generalizability analysis (shown in Fig. 5), we have evaluated the anomaly examples detection in unseen malicious type training-testing scenarios. In the proposed research, four different malicious categories are used which in turn contain several attack generation algorithms. To extensively study the malicious examples detection and pave a way for future research to enhance the robustness, four-fold cross-validation experiments are performed. In each fold, one malicious examples category is used for training, and remaining are the used for testing. We have earlier observed that the accuracy is lower in the unseen attack setting in comparison to the seen attack setting. The drop in detection performance is further observed when the malicious attack category is changed in the training and testing set. Except in a few cases, the majority of the detection performance lies close to 60% only which shows that the detection of malicious examples demands careful attention, especially in extremely generalized and open-world settings.

Let us dig deeper towards understanding the detection of individual attacks of the malicious category used for evaluation. When the common corruptions are used to train the detection algorithm, across each corruption it is found that the detection of CS attack, universal perturbation, and Int-P attack belonging to the natural adversary, artificial adversarial perturbation, and camera noise, respectively, is the highest. In natural adversary, color shift shows the highest correlation with other malicious categories and yields the highest detection performance. Interestingly, artificial adversarial perturbations are found more effective in detecting the Salt & Pepper noise (SPN) common corruption in comparison to other common corruptions and remaining unseen malicious categories. Whereas, the camera-induced noises yield better performance on universal artificial perturbation along with SPN corruption in comparison to other unseen malicious example categories. In quantitative terms, the detection performance of SPN and universal perturbation is at least 17% and 30% better when artificial adversary and camera-induced noises are used in training, respectively. We believe that

**Table 5.** Ablation study utilizing different backbone architectures for seen and unseen common corruption detection. VGG yields the best performance across each network.

Corrup.	Model	Test					Avg.
		GN	SPN	SN	SPKN	UN	
GN	VGG	94.93	97.90	94.83	93.00	85.27	<b>93.19</b>
	DenseNet	96.70	98.33	95.77	91.50	73.63	91.19
	MobileNet	95.33	97.80	93.90	87.40	61.83	87.25
	InceptionV3	94.30	96.53	91.17	84.80	65.60	86.48
	Xception	97.40	98.40	96.10	89.03	66.23	89.43
SN	VGG	94.17	95.83	94.83	94.47	87.47	<b>93.35</b>
	DenseNet	94.63	98.33	95.10	91.10	69.00	89.63
	MobileNet	95.30	98.57	95.63	90.40	63.30	88.64
	InceptionV3	92.97	98.20	93.87	85.33	58.13	85.70
	Xception	94.70	99.30	95.97	86.73	60.20	87.38
SPKN	VGG	92.80	95.57	93.67	93.40	86.73	<b>92.36</b>
	DenseNet	94.43	96.83	95.17	93.37	77.30	91.42
	MobileNet	92.30	93.67	93.30	92.00	71.70	88.59
	InceptionV3	88.17	92.83	89.47	86.10	64.83	84.28
	Xception	96.60	96.97	97.23	94.50	69.00	90.86
UN	VGG	91.87	91.90	92.00	91.97	91.83	<b>91.91</b>
	DenseNet	92.03	93.17	92.00	90.83	86.23	90.85
	MobileNet	90.13	90.33	90.00	88.67	78.00	87.43
	InceptionV3	84.10	84.87	84.13	83.67	80.23	83.40
	Xception	82.37	82.53	82.43	82.17	80.70	82.04

there is a connection between different malicious examples categories and can be exploited further to build a universal defense system.

**Impact of CNN Backbone.** We have extensively analyzed the impact of different CNN architectures as a backbone network in the malicious examples detection pipeline. In brief, the VGG architecture yields the best average malicious examples detection performance whether evaluated in seen or unseen attack image settings. The detailed results are added in Table 5. When the VGG architecture is trained using SN, it shows the highest average generalization performance; although the performance does not shows significant degradation even if trained on other corruptions. In terms of corruption, the detection of uniform noise corrupted images is complex as compared to the other corruptions. For instance, when the VGG architecture is trained on GN corruption and tested on UN, it shows at least 7.73% lower performance in comparison to the detection performance on other corruption images. On the other hand, SPN corruption

is found the easiest to be defended, i.e., out of all corruptions used, each CNN architecture yields the best detection accuracy of the SPN images.

## 5 Conclusion

In this research, we put a strong step towards developing a universal robustness system by building the ‘*first-ever wide angle*’ anomalies dataset. The said dataset covers 19 different attack generation algorithms and contains approximately 60,000 images. An experimental evaluation setup shows that the detection of these malicious examples categories is easy when they are seen at the time of training the attack detection algorithm. However, several generalization experimental protocol reveals that we should not fall prey to such high detection accuracies as performance can significantly drop if an unseen attack or unseen malicious category comes for evaluation. In the real world, we can expect such unseen training-testing scenarios; therefore, we demand careful attention while developing the defense algorithms and their evaluation in several generalized settings. The experimental results also reveal a potential connection between different malicious categories which can be effectively used in building a *universal* detection algorithm. In the future, newer malicious attack images can be added to the proposed dataset and a sophisticated detection algorithm will be built to get universal robustness.

## References

1. Agarwal, A., Goswami, G., Vatsa, M., Singh, R., Ratha, N.K.: Damad: database, attack, and model agnostic adversarial perturbation detector. *IEEE Trans. Neural Netw. Learn. Syst.* **33**, 1–13 (2021). <https://doi.org/10.1109/TNNLS.2021.3051529>
2. Agarwal, A., Ratha, N., Vatsa, M., Singh, R.: Exploring robustness connection between artificial and natural adversarial examples. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 179–186 (2022)
3. Agarwal, A., Ratha, N.K.: Black-box adversarial entry in finance through credit card fraud detection. In: *CIKM Workshops* (2021)
4. Agarwal, A., Ratha, N.K.: On the robustness of stock market regressors. In: *ECML-PKDD Workshops* (2022)
5. Agarwal, A., Singh, R., Vatsa, M., Ratha, N.: Image transformation-based defense against adversarial perturbation on deep learning models. *IEEE Trans. Depend. Secure Comput.* **18**(5), 2106–2121 (2021). <https://doi.org/10.1109/TDSC.2020.3027183>
6. Agarwal, A., Vatsa, M., Singh, R., Ratha, N.: Cognitive data augmentation for adversarial defense via pixel masking. *Pattern Recogn. Lett.* **146**, 244–251 (2021)
7. Agarwal, A., Vatsa, M., Singh, R., Ratha, N.: Intelligent and adaptive mixup technique for adversarial robustness. In: *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 824–828 (2021). <https://doi.org/10.1109/ICIP42928.2021.9506180>
8. Agarwal, A., Vatsa, M., Singh, R., Ratha, N.K.: Noise is inside me! generating adversarial perturbations with noise derived from natural filters. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3354–3363 (2020)

9. Andriushchenko, M., Flammarion, N.: Understanding and improving fast adversarial training. *Adv. Neural Inf. Process. Syst.* **33**, 16048–16059 (2020)
10. Chhabra, S., Agarwal, A., Singh, R., Vatsa, M.: Attack agnostic adversarial defense via visual imperceptible bound. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 5302–5309 (2021). <https://doi.org/10.1109/ICPR48806.2021.9412663>
11. Chollet, F.: Xception: deep learning with depthwise separable convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
12. Chun, S., Oh, S.J., Yun, S., Han, D., Choe, J., Yoo, Y.: An empirical evaluation on robustness and uncertainty of regularization methods. *arXiv preprint [arXiv:2003.03879](https://arxiv.org/abs/2003.03879)* (2020)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
14. Dodge, S., Karam, L.: Quality resilient deep neural networks. *arXiv preprint [arXiv:1703.08119](https://arxiv.org/abs/1703.08119)* (2017)
15. Esmaeilpour, M., Cardinal, P., Koerich, A.L.: Cyclic defense gan against speech adversarial attacks. *IEEE Signal Process. Lett.* **28**, 1769–1773 (2021)
16. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint [arXiv:1811.12231](https://arxiv.org/abs/1811.12231)* (2019)
17. Geirhos, R., Temme, C.R., Rauber, J., Schütt, H.H., Bethge, M., Wichmann, F.A.: Generalisation in humans and deep neural networks. *Adv. Neural Inf. Process. Syst.* **31**, 1–13 (2018)
18. Goel, A., Singh, A., Agarwal, A., Vatsa, M., Singh, R.: Smartbox: benchmarking adversarial detection and mitigation algorithms for face recognition. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–7. IEEE (2018)
19. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. *arXiv preprint [arXiv:1412.6572](https://arxiv.org/abs/1412.6572)* (2014)
20. Goswami, G., Agarwal, A., Ratha, N., Singh, R., Vatsa, M.: Detecting and mitigating adversarial perturbations for robust face recognition. *Int. J. Comput. Vision* **127**(6), 719–742 (2019)
21. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint [arXiv:1903.12261](https://arxiv.org/abs/1903.12261)* (2019)
22. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15262–15271 (2021)
23. Hermann, K., Chen, T., Kornblith, S.: The origins and prevalence of texture bias in convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **33**, 19000–19015 (2020)
24. Hosseini, H., Poovendran, R.: Semantic adversarial examples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1614–1619 (2018)
25. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. *arXiv preprint [arXiv:1704.04861](https://arxiv.org/abs/1704.04861)* (2017)
26. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)



27. Kamann, C., Rother, C.: Benchmarking the robustness of semantic segmentation models with respect to common corruptions. *Int. J. Comput. Vision* **129**(2), 462–483 (2021)
28. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: *Artificial Intelligence Safety and Security*, pp. 99–112. Chapman and Hall/CRC (2018)
29. Landau, B., Smith, L.B., Jones, S.S.: The importance of shape in early lexical learning. *Cogn. Dev.* **3**(3), 299–321 (1988)
30. Li, F., Liu, X., Zhang, X., Li, Q., Sun, K., Li, K.: Detecting localized adversarial examples: a generic approach using critical region analysis. In: *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pp. 1–10. IEEE (2021)
31. Li, X., Li, J., Dai, T., Shi, J., Zhu, J., Hu, X.: Rethinking natural adversarial examples for classification models. *arXiv preprint [arXiv:2102.11731](https://arxiv.org/abs/2102.11731)* (2021)
32. Ma, X., et al.: Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recogn.* **110**, 107332 (2021)
33. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. *arXiv preprint [arXiv:1706.06083](https://arxiv.org/abs/1706.06083)* (2017)
34. Mikołajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. In: *2018 International Interdisciplinary PhD Workshop (IIPhDW)*, pp. 117–122. IEEE (2018)
35. Mintun, E., Kirillov, A., Xie, S.: On interaction between augmentations and corruptions in natural corruption robustness. *Adv. Neural Inf. Process. Syst.* **34**, 1–13 (2021)
36. Modas, A., Rade, R., Ortiz-Jiménez, G., Moosavi-Dezfooli, S.M., Frossard, P.: Prime: a few primitives can boost robustness to common corruptions. *arXiv preprint [arXiv:2112.13547](https://arxiv.org/abs/2112.13547)* (2021)
37. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1765–1773 (2017)
38. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582 (2016)
39. Morrison, K., Gilby, B., Lipchak, C., Mattioli, A., Kovashka, A.: Exploring corruption robustness: inductive biases in vision transformers and mlp-mixers. *arXiv preprint [arXiv:2106.13122](https://arxiv.org/abs/2106.13122)* (2021)
40. Pedraza, A., Deniz, O., Bueno, G.: Really natural adversarial examples. *Int. J. Mach. Learn. Cybern.* **13**, 1–13 (2021)
41. Pei, Y., Huang, Y., Zou, Q., Zhang, X., Wang, S.: Effects of image degradation and degradation removal to cnn-based image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(4), 1239–1253 (2019)
42. Raghunathan, A., Xie, S.M., Yang, F., Duchi, J.C., Liang, P.: Adversarial training can hurt generalization. *arXiv preprint [arXiv:1906.06032](https://arxiv.org/abs/1906.06032)* (2019)
43. Saikia, T., Schmid, C., Brox, T.: Improving robustness against common corruptions with frequency biased models. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10211–10220 (2021)
44. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-gan: protecting classifiers against adversarial attacks using generative models. *arXiv preprint [arXiv:1805.06605](https://arxiv.org/abs/1805.06605)* (2018)
45. Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., Bethge, M.: Improving robustness against common corruptions by covariate shift adaptation. *Adv. Neural Inf. Process. Syst.* **33**, 11539–11551 (2020)

46. Shafahi, A., et al.: Adversarial training for free! *Adv. Neural Inf. Process. Syst.* **32** (2019)
47. Shafahi, A., Najibi, M., Xu, Z., Dickerson, J., Davis, L.S., Goldstein, T.: Universal adversarial training. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 5636–5643 (2020)
48. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)* (2014)
49. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826 (2016)
50. Taheri, H., Pedarsani, R., Thrampoulidis, C.: Asymptotic behavior of adversarial training in binary classification. *arXiv preprint [arXiv:2010.13275](https://arxiv.org/abs/2010.13275)* (2020)
51. Tramer, F.: Detecting adversarial examples is (nearly) as hard as classifying them. *arXiv preprint [arXiv:2107.11630](https://arxiv.org/abs/2107.11630)* (2021)
52. Wang, J., et al.: Smsnet: a new deep convolutional neural network model for adversarial example detection. *IEEE Trans. Multimedia* **24**, 230–244 (2021)
53. Xue, M., Yuan, C., He, C., Wang, J., Liu, W.: Naturalae: natural and robust physical adversarial examples for object detectors. *J. Inf. Secur. Appl.* **57**, 102694 (2021)
54. Zhang, H., Chen, H., Song, Z., Boning, D., Dhillon, I.S., Hsieh, C.J.: The limitations of adversarial training and the blind-spot attack. *arXiv preprint [arXiv:1901.04684](https://arxiv.org/abs/1901.04684)* (2019)