Benchmarking Image Classifiers for Physical Out-of-Distribution Examples Detection

Ojaswee[1], Akshay Agarwal[1], and Nalini Ratha[2]
[1]IISER Bhopal, India, [2]University at Buffalo, USA
{ojaswee19, akagarwal}@iiserb.ac.in, nratha@buffalo.edu

Abstract

The rising popularity of deep neural networks (DNNs) in computer vision has raised concerns about their robustness in the real world. Recent works in this field have well-demonstrated the vulnerability of these networks to carefully crafted adversarial attacks which yield out-of-distribution (OOD) samples. Interestingly, the majority of the existing literature focuses on adversarial attacks crafted for the digital domain only. Physical adversarial attacks are easier to deploy in the real world and yield higher attack success than digital perturbations. The prime limitation of such a dearth of studies handling physical out-of-distribution images is the lack of benchmark datasets. To overcome this limitation, this research proposes a novel out-of-distribution dataset using adversarial patches of different variations to advance the robustness of deep networks against such stealthy out-of-distribution images. We have also conducted extensive experiments both under seen and unseen patch settings and observed that unseen adversarial patches are hard to defend. By conducting this study and delving into the complexities of defending against patch attacks, we believe it will serve as inspiration for future researchers to incorporate physical OOD attacks into their defense strategies.

1. Introduction

While CNNs show tremendous success, their vulnerability against out-of-distribution (OOD) samples is a major concern. One popular form of getting an out-of-distribution sample is the addition of an adversarial pattern whether it is an imperceptible perturbation or visible patches [8,9,20]. Adversarial attacks can be classified into three broad categories based on the amount of perturbation: (i) imperceptible adversarial perturbations [34], (ii) universal perturbations [27], and (iii) physical adversarial patches [13, 22]. While novel adversarial attacks ensure that the DNNs are secure from any possible vulnerability, the defense algorithms tackling them independently might be a severe concern [1,6,7]. It is seen that both imperceptible and universal adversarial perturbations have limitations in their transferability and applicability to the physical world, especially when compared to adversarial patch attacks. However, despite the real-world effectiveness of adversarial patch attacks, the majority of adversarial defense algorithms primarily focus on countering imperceptible adversarial perturbations (disturbing a very small number of pixels to induce misclassification). *It leaves a gap in developing a true defense algorithm that can counter the wide variety of adversaries that exist in the real world.*

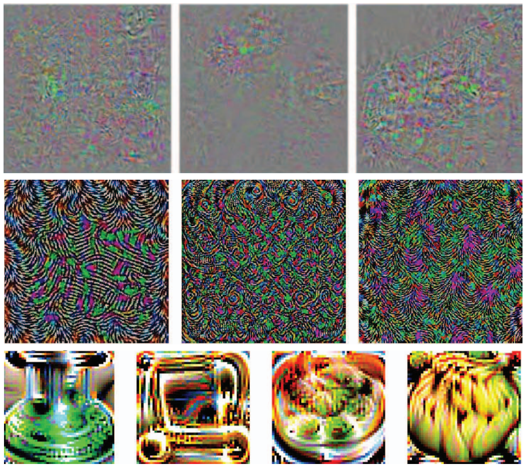

Figure 1. Distribution shift among different adversarial attacks. The first row is imperceptible adversarial perturbation noise, the second is the universal perturbation vectors, and the third is the physical adversarial attacks.

We assert that ignoring the impact of adversarial patch attacks can be dangerous, especially when aiming to deploy these state-of-the-art DNNs in the unconstrained physical world. Figure 1 shows the adversarial noises, which when incorporated into the images yield a broad spectrum of out-of-distribution examples. The reason is that imperceptible and universal perturbations do not occlude any region of the images; however, the adversarial patches can occlude a small or significant portion of an image, depending on its size. We want to mention that while several benchmark studies are proposed in the literature to tackle the issue of adversarial defense, no work has included adversarial patch attacks. For example, Hendrycks and Dietterich [21] showcase the impact of several common corruptions, such as Gaussian noise, blur, and fog on deep neural networks (DNNs), but no defense has been proposed in this study. Dong et al. [17] proposed a benchmark study to tackle only the imperceptible adversarial perturbations. Recently, Agarwal et al. [2, 4] have performed several benchmark studies to counter adversarial perturbations and common corruptions; however, still, the OOD examples obtained physical patches are missing from the literature. It is also to be noted that several defense works also exist that can effectively detect imperceptible adversarial attacks in several generalized settings such as unseen datasets, unseen perturbation, and unseen threat model [1, 5, 6]. The evaluation of these defenses for physical patch adversaries is still an open-research direction.

As discussed, no studies have benchmarked the defense against adversarial patch attacks; therefore, in this research, we have performed the study by creating patch attack datasets. The OOD patch detection is performed using several deep image classification networks, including the network architecture search (NAS) method [40] and the vision transformer [19]. Neural Architecture Search (NAS) automates the process of designing optimal neural network architectures. Instead of relying on manual design, NAS employs algorithms to search and discover architectures that maximize performance for specific tasks. On the other hand, Vision Transformers (ViTs) are a recent breakthrough in computer vision. ViTs adopt the Transformer architecture, originally designed for natural language processing, and apply it directly to images. The prime reason for conducting a benchmark study on adversarial patches can also be understood from the distribution shift among the attacks and out-of-distribution handling limitations of the DNNs. We assert that the presence of the dataset and benchmark evaluation can help advance the research in this direction and make comparisons with new novel algorithms. In brief, the contributions of this research are:

- A novel adversarial patch attack dataset has been developed. The dataset contains images of multiple variations of patches. The presence of different style patches will ensure that the defense algorithms are not biased;
- A benchmark evaluation has also been conducted. For that, several real-world evaluations and protocols are developed to handle seen patches and unseen patches. A defined protocol can help make fair comparisons in future works, which is often missing in imperceptible adversarial detection literature.

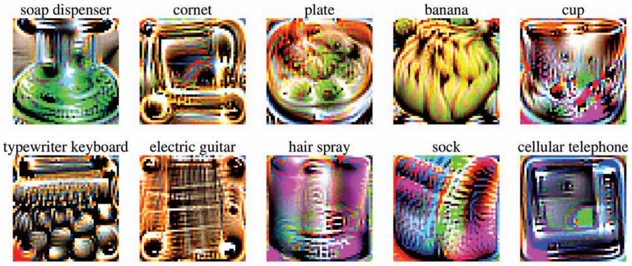

Figure 2. Samples of ten adversarial patches (arranged in an order, i.e., patch 0 is top left and patch 9 is bottom right) along with their target labels that are used to generate the proposed OOD datasets.

2. Related Work

The adversarial attacks on DNNs using crafted imperceptible perturbations are first introduced by Szegedy et al. [34]. They found an imperceptible adversarial perturbation can help in generating effective out-of-distribution samples which can trigger the misclassification of a deep neural network. Since then numerous methods to generate imperceptible adversarial perturbations have been proposed [25]. Surprisingly, the attacks are effective in fooling the deep neural networks working in different domains such as image classification [12, 24], geoscience [10, 38], segmentation and autonomous driving [28, 30].

While these imperceptible perturbations are found effective in fooling deep neural networks and are applicable in various vision tasks, some limitations halt their deployment in the real world. A few such critical limitations are transferability to unseen deep neural networks [36] and ineffectiveness in the physical world. To tackle the limitations of imperceptible adversarial patterns, an "adversarial patch attack" is introduced which contains malicious information which is visible but can result in stealthy out-of-distribution images that have higher physical world practicality [35, 37]. Adversarial patch attack on image classifiers is first proposed by Brown et al. [11], presenting a universal and targeted attack on real-world physical object detectors. Another approach to reducing the suspiciousness of adversarial patches to human eyes involved the creation of an adversarial QR patch [14]. Similar to the digital adversarial perturbations, adversarial patch-enhanced out-of-distribution images are effective in several domains such as point clouds [18], biometrics recognition [23, 33], and traffic sign recognition [39]. From the above review, it is observed that several adversarial patch generation algorithms have been proposed; however, one prime limitation of the literature is the inexistence of a benchmark out-of-distribution image dataset containing adversarial patch examples. The limita-

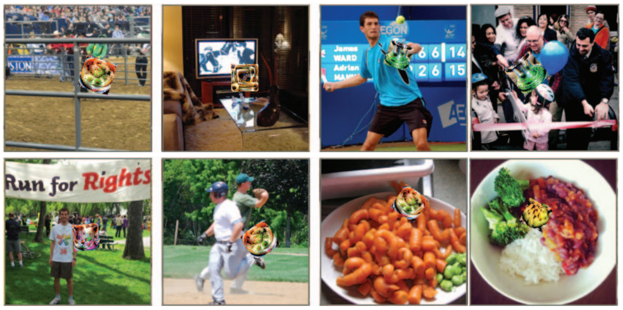

Figure 3. Samples of the different OOD adversarial patch images developed as part of our dataset. It can be seen that these patches can be blended with the image content and hence increase the complexity of its detection, especially when the detection network has not seen them in the training.

tion impacts the development of defense algorithms that can protect the integrity of deep networks against stealthy out-of-distribution adversarial patch images. We assert that this research is a first step towards that goal which presents a benchmark study in defending deep neural networks from out-of-distribution physical world adversaries.

3. Proposed Out-of-Distribution Dataset

Given the limited research in the area of detecting physical out-of-distribution images consisting of adversarial patches, this study aims to address the gap by introducing two novel adversarial patch datasets. These datasets have been curated using images of well-known datasets, namely ImageNet [16] and COCO (Common Objects in Context) [26]. We have used 2000 randomly selected images from the ImageNet dataset and treated them as the real subset of the proposed dataset. For the patched subset, on another 2000 randomly selected real images, ten different adversarial patches [29] as shown in Figure 2 are applied. Each of these patches is designed to be targeted toward a specific class, resulting in a diverse range of target classes for misclassification. Hence these patches not only have variations in texture and style but can also misclassify the images into different categories. *'Thus, we have generated a large-scale adversarial patch dataset containing* $20,000$ *patched images along with additional* 2000 *real images using the subset of ImageNet'*.

To extensively benchmark and understand the robustness of adversarial patch OOD examples detection, we have also utilized the COCO dataset. Similar to ImageNet, 4000 images are randomly selected from the dataset, out of which 2000 images are kept as real images, and on the remaining 2000 images each of the adversarial patches has been applied. **In total, the proposed dataset contains** $\mathbf{40,000}$ **(**$\mathbf{20,000}$ **of ImageNet and** $\mathbf{20,000}$ **of COCO) adversarial patched images, and** $\mathbf{4,000}$ **real images.** Figure 3 shows some of the samples from the proposed dataset reflecting the challenge in detecting the adversarial attack not only due to *significant style change of the patches but also their blending nature with the complex image regions*.

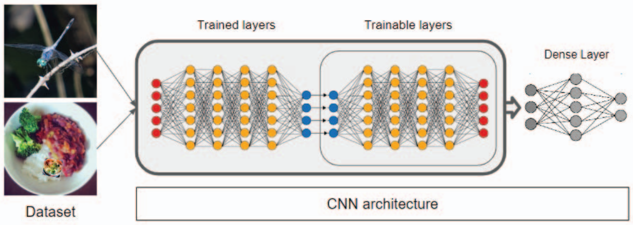

Figure 4. Schematic diagram of the adversarial patch detection network.

4. Benchmarking Adversarial Patch Detection Results and Analysis

Architecture and Experiments: This research aims to overcome the limitation of the existing adversarial defense literature and benchmark state-of-the-art (SOTA) image classifiers for physical OOD examples detection. Henceforth, we have used several CNN-based architectures and vision transformers which varied in terms of basic architecture, the number of layers, the connection between layers, and their formation. The selected classifiers are as follows: XceptionNet [15], MobileNetv2 [31], NASMobileNet [40], VGG16 [32], and Vision Transformer [19]. The reason for using these architectures is that they are heavily popular for image classification. Further, NAS-style architectures and ViTs have not been explored comprehensively for adversarial patch detection. Therefore, benchmarking their robustness can pave the way to incorporate the adversarial nature of the images while crafting network architecture. Figure 4 shows the schematic diagram of the physical out-of-distribution adversarial patch detection framework.

These networks are fine-tuned by keeping approximately 60 percent of the network pre-trained on ImageNet and training the rest 40 percent of the network using the ImageNet patch subset. On top of that, two dense layers are added to extract features along with the classification layer to perform binary classification (real vs. patch). We split the ImageNet patch OOD subset into train and test sets in the ratio of 3:2 and trained all the models using the training set. Thus, each model has been trained on 1200 patched images of one patch class at a time and 1200 real images. In total, we have trained five different models on each of the ten patches and then tested each of these models on seen and unseen real and patch images of the ImageNet (seen OOD dataset) and COCO (unseen OOD dataset).

Results and Analysis: As described above, we have performed a comprehensive set of experiments in seen patches and unseen patches settings to effectively evaluate the per-

Table 1. Adversarial patch detection accuracy of the different architectures on ImageNet subset. The results are reported in terms of mean and standard deviation (SD), where the trained on one patch is tested on all the patches. Here X, NAS, and M represent Xception, NASNet, and MobileNet, respectively. The best mean and SD value is highlighted and underlined across each network, respectively.

Models	Matrices	Patch 0	Patch 1	Patch 2	Patch 3	Patch 4	Patch 5	Patch 6	Patch 7	Patch 8	Patch 9
X	Mean	75.04	75.41	77.28	76.22	**80.81**	71.36	70.55	79.39	79.83	74.61
	SD	10.60	03.73	07.11	09.59	06.95	08.92	03.65	10.33	11.55	11.92
VGG16	Mean	71.06	75.77	69.39	65.10	**80.35**	74.89	74.35	68.68	70.78	73.04
	SD	16.20	14.91	18.16	18.39	16.04	16.63	14.93	19.88	19.72	17.25
M	Mean	**85.38**	84.05	68.76	73.94	84.27	77.65	83.04	76.45	82.83	79.50
	SD	09.46	04.06	00.67	00.68	04.39	04.65	04.45	17.62	10.00	12.39
NAS	Mean	59.34	57.73	59.27	60.45	59.34	55.15	59.90	**65.83**	62.08	58.63
	SD	01.66	00.44	00.68	01.37	01.00	00.44	01.09	04.62	01.84	02.67
ViT	Mean	72.88	75.36	**77.28**	65.05	76.46	74.98	66.62	72.76	69.06	70.73
	SD	15.31	14.58	16.73	16.84	17.04	14.99	15.04	20.26	20.84	20.11

Table 2. Adversarial patch detection accuracy of the different architectures on COCO subset where the networks are trained on the ImageNet OOD subset. The results are reported in terms of mean and standard deviation (SD), where the trained on one patch is tested on all the patches. The best mean and SD value is highlighted and underlined across each network, respectively.

Models	Metric	Patch 0	Patch 1	Patch 2	Patch 3	Patch 4	Patch 5	Patch 6	Patch 7	Patch 8	Patch 9
X	Mean	73.39	75.32	76.51	75.25	**79.69**	70.93	70.94	77.41	78.94	73.39
	SD	10.66	03.58	07.63	09.52	06.72	09.41	03.67	09.97	11.53	11.96
VGG16	Mean	70.36	74.12	70.01	66.11	**79.61**	74.84	73.23	67.63	70.10	72.63
	SD	16.31	15.47	18.58	18.47	15.64	17.07	15.83	19.87	19.30	17.21
M	Mean	83.46	83.29	68.67	73.96	**84.86**	77.40	82.42	76.50	81.82	79.46
	SD	10.17	03.91	00.77	00.82	04.81	05.29	04.53	17.46	09.47	11.98
NAS	Mean	57.83	56.15	55.87	58.10	57.39	54.28	56.46	**62.79**	58.18	57.18
	SD	01.66	03.44	00.75	01.11	00.87	00.31	00.67	03.81	01.61	02.12
ViT	Mean	72.78	74.98	**77.46**	64.09	76.47	75.16	66.32	73.08	69.36	70.98
	SD	15.23	15.14	16.80	17.15	17.34	15.02	15.50	19.86	20.83	19.61

formance and robustness of the image classifiers for adversarial patch detection. The analysis of the experiments can be broadly performed based on the following factors: (i) robustness of the classification model and (ii) effectiveness of the training patch. The brief experimental results on ImageNet and COCO datasets in terms of average classification accuracy are given in Tables 1 and 2, respectively. It can be seen that the ViT model which is pre-trained on the large-scale dataset, is shown state-of-the-art performance for OOD examples detection. Interestingly, the VGG model shows comparable performance to ViT and is not pre-trained on large-scale datasets on which ViT is trained. On the other other hand, the NASMobileNet architecture shows the lowest OOD examples detection performance on both datasets. However, it is to be noted that the standard deviation of the performance of NASNet is the lowest among all the networks. From the results of both datasets, it is interesting to note that, for pre-defined network structures such as VGG and XceptionNet, patch 4 is found the most robust (unseen dataset and patch) architecture to detect OOD adversaries. Whereas ViT found patch-2 as the most effective and NASMobileNet found patch 7 as the most effective patch detector both under seen and unseen patch evaluation settings. Out of all the networks, MobileNet archived the highest accuracy when trained on patch 0 (Table 1) and patch 4 under unseen dataset setting (Table 2). However, the performance of MobileNet is inconsistent when different types of patches are used for training. It can also be analyzed from the heatmaps (Figure 5) where the MobileNet shows the best performance majority of the time when it is trained on the individual patches of both datasets.

Tables 3 and 4 provide the detailed performance of each network and how they have performed when they are trained on a specific patch and evaluated on the same and unseen patches. Here the analysis can be broken into generalized conditions: (i) seen patch detection on the same training-testing dataset and (ii) unseen patch detection on the same training-testing dataset. The diagonal elements of these tables represent the performance of the models on seen patches. As expected, the models perform best on the seen patches but are found vulnerable against unseen patches. Out of all the networks, ViT and VGG show high performance in detecting the patches which are seen at the time of training. As mentioned that the biggest real-world challenge for a defense algorithm is that at the time of testing, out-of-distribution attack (patch) images come for classification. Henceforth, the desired goal for an ideal defense algorithm is the robustness in handling these unseen attack

Table 3. Adversarial patch detection accuracy of the different architectures on **ImageNet** subset under seen and unseen patch detection scenarios. Seen settings results are highlighted and the best unseen-test patch performance is underlined.

Train →	Patch 0	Patch 1	Patch 2	Patch 3	Patch 4	Patch 5	Patch 6	Patch 7	Patch 8	Patch 9
Test ↓	Xception									
Patch 0	**87.12**	72.06	82.56	83.00	82.19	61.06	64.25	75.38	83.25	75.25
Patch 1	64.81	**81.19**	76.50	72.87	73.62	79.19	73.81	72.37	68.88	67.06
Patch 2	85.75	78.87	**88.31**	88.50	86.19	68.37	72.00	80.44	84.56	76.81
Patch 3	79.56	73.75	84.31	**89.81**	78.94	73.69	71.13	70.19	72.19	59.25
Patch 4	84.00	76.69	83.31	81.19	**89.00**	69.94	70.50	86.69	89.69	81.81
Patch 5	58.38	78.00	68.88	73.06	69.13	**86.12**	73.81	60.75	59.25	57.06
Patch 6	60.31	78.62	71.94	77.06	72.31	83.44	**76.25**	75.75	69.00	65.38
Patch 7	71.56	70.31	70.56	63.19	84.44	64.44	68.75	**92.31**	89.63	85.75
Patch 8	79.75	71.25	77.81	71.75	87.44	62.44	66.50	91.12	**92.81**	87.44
Patch 9	79.12	73.37	68.62	61.75	84.81	64.94	68.50	88.94	89.06	**90.31**
VGG16										
Patch 0	**98.00**	71.00	86.12	79.37	89.63	69.38	60.06	61.44	70.44	73.31
Patch 1	54.50	**96.81**	59.75	56.63	61.62	92.69	91.44	54.75	53.81	63.56
Patch 2	89.63	85.56	**97.94**	93.94	95.44	63.44	67.81	52.94	63.63	65.69
Patch 3	83.50	87.06	95.56	**98.44**	74.31	91.69	88.38	50.38	50.38	51.44
Patch 4	71.69	63.25	76.81	57.88	**95.75**	61.50	63.56	67.31	76.38	72.75
Patch 5	52.50	82.31	53.06	56.19	57.19	**95.06**	88.38	52.13	51.06	55.31
Patch 6	51.94	89.44	53.12	56.00	58.31	93.69	**95.63**	58.13	52.56	60.25
Patch 7	59.75	53.37	50.75	50.13	84.75	55.06	59.94	**97.87**	97.00	94.69
Patch 8	74.19	55.31	59.75	50.75	94.38	55.37	57.50	97.44	**98.12**	96.13
Patch 9	74.94	73.62	61.06	51.63	92.12	71.00	70.75	94.38	94.38	**97.31**
MobileNet										
Patch 0	**92.25**	73.12	69.00	74.00	83.06	68.56	71.81	63.81	81.44	68.50
Patch 1	79.37	**86.81**	68.94	74.12	85.06	82.06	86.44	70.81	82.56	82.44
Patch 2	92.81	84.81	**69.13**	74.44	86.87	76.50	84.50	72.62	86.94	80.56
Patch 3	89.88	84.87	69.13	**74.44**	85.62	80.75	86.19	65.75	82.50	71.56
Patch 4	93.25	86.12	69.06	74.31	**87.06**	78.56	84.06	92.56	90.50	90.38
Patch 5	66.94	82.63	67.19	72.19	74.31	**82.94**	85.37	52.56	61.44	59.50
Patch 6	71.81	83.88	67.87	73.56	78.87	82.44	**87.06**	56.19	69.94	65.69
Patch 7	88.19	85.87	69.06	74.25	87.12	73.19	81.81	**96.56**	90.94	91.81
Patch 8	92.44	85.56	69.13	74.31	87.31	74.62	82.00	97.06	**90.94**	92.31
Patch 9	86.81	86.87	69.06	73.81	87.37	76.88	81.19	96.56	91.06	**92.25**
NASMobileNet										
Patch 0	**60.94**	57.38	59.75	61.62	59.13	55.31	59.00	65.25	61.81	59.87
Patch 1	58.00	**58.50**	59.00	60.31	58.69	55.56	60.56	63.69	61.12	57.81
Patch 2	60.50	58.06	**60.12**	62.06	59.94	55.62	60.81	67.94	63.19	59.50
Patch 3	59.87	58.38	59.56	**62.31**	59.31	55.19	61.44	62.69	61.44	54.81
Patch 4	60.37	57.38	59.75	61.00	**60.31**	55.37	59.62	67.94	63.06	60.00
Patch 5	56.50	57.50	57.88	58.38	57.81	**55.62**	59.62	58.56	58.63	55.25
Patch 6	56.63	57.25	58.38	59.19	57.69	54.75	**60.56**	59.62	59.81	55.19
Patch 7	59.87	57.75	59.31	60.12	60.25	54.31	60.12	**71.25**	64.06	61.00
Patch 8	60.31	57.69	59.56	60.81	60.19	54.87	59.69	70.63	**63.94**	61.31
Patch 9	60.37	57.44	59.38	58.69	60.12	54.87	57.56	70.75	63.69	**61.56**
ViT										
Patch 0	**98.25**	57.63	90.19	60.75	72.62	68.69	51.56	68.00	60.25	60.12
Patch 1	56.56	**97.50**	73.81	67.62	58.06	93.19	86.19	57.38	51.31	58.56
Patch 2	86.81	79.81	**98.81**	92.12	84.31	72.56	62.50	57.75	56.50	56.31
Patch 3	73.12	89.44	97.50	**98.75**	61.81	80.00	72.25	54.19	50.94	51.69
Patch 4	86.31	70.25	92.50	63.19	**98.44**	77.50	59.87	92.31	85.37	90.25
Patch 5	53.81	81.44	58.00	52.75	58.06	**96.19**	70.00	52.19	50.69	51.63
Patch 6	53.12	87.94	57.88	54.81	55.94	89.44	**96.50**	55.25	50.19	54.62
Patch 7	71.94	58.06	60.00	52.19	89.44	56.00	54.00	**98.44**	94.69	87.63
Patch 8	80.19	56.06	80.37	53.81	96.69	54.37	50.63	98.31	**99.25**	97.62
Patch 9	68.75	75.50	63.75	54.56	89.25	61.87	62.75	93.81	91.50	**98.94**

Table 4. Adversarial patch detection accuracy on **COCO** subset under seen and unseen patch detection scenarios. Here the evaluation has been performed under 'duly' generalized settings, including where the dataset and patches for testing are also unknown to the adversarial detection training models. Seen settings results are highlighted and the best unseen-test patch performance is underlined.

Train →	Patch 0	Patch 1	Patch 2	Patch 3	Patch 4	Patch 5	Patch 6	Patch 7	Patch 8	Patch 9
Test ↓	Xception									
Patch 0	**87.44**	73.87	82.88	83.50	83.00	62.12	66.62	76.25	84.31	75.44
Patch 1	64.31	**80.75**	77.56	73.75	73.50	79.69	74.94	70.88	69.81	65.81
Patch 2	84.19	78.44	**87.88**	87.19	84.44	70.06	71.81	79.31	83.94	74.81
Patch 3	76.75	71.13	82.69	**88.25**	75.88	72.75	69.13	67.25	70.50	58.44
Patch 4	82.63	76.81	83.44	79.50	**87.50**	69.25	71.13	84.44	89.38	81.81
Patch 5	56.75	77.81	67.56	72.94	69.19	**86.69**	74.62	60.00	58.25	56.88
Patch 6	58.88	78.56	68.31	73.81	70.63	82.63	**76.81**	71.50	66.81	61.69
Patch 7	69.25	70.81	70.25	63.44	83.13	62.06	70.38	**90.19**	87.44	84.06
Patch 8	77.50	71.63	77.00	70.56	85.87	59.44	65.81	88.44	**91.44**	85.37
Patch 9	76.19	73.37	67.56	59.56	83.75	64.63	68.12	85.87	87.56	**89.56**
VGG16										
Patch 0	**97.94**	69.50	88.63	82.19	89.63	69.00	59.25	59.69	69.19	73.44
Patch 1	53.44	**96.06**	61.69	58.44	63.25	93.00	90.75	54.62	53.87	64.88
Patch 2	88.38	83.44	**98.06**	94.00	94.44	62.88	66.44	52.00	63.44	64.63
Patch 3	82.56	86.75	96.63	**98.94**	72.94	93.37	89.13	50.13	49.94	50.88
Patch 4	70.88	62.25	77.06	60.81	**95.13**	62.12	62.81	64.94	76.31	73.06
Patch 5	52.25	81.38	52.75	57.81	55.81	**95.88**	88.75	52.25	50.88	54.56
Patch 6	51.50	88.50	52.63	56.81	58.81	93.37	**95.38**	55.56	51.88	58.88
Patch 7	57.94	50.63	50.25	49.88	81.69	54.12	57.00	**97.62**	95.44	93.69
Patch 8	75.31	53.37	61.37	50.69	94.13	55.87	55.50	96.88	**97.69**	95.88
Patch 9	73.44	69.31	61.06	51.50	90.25	68.81	67.25	92.62	92.31	**96.44**
MobileNet										
Patch 0	**91.87**	72.62	68.75	73.69	84.62	67.25	72.12	63.38	80.81	69.56
Patch 1	76.44	**85.50**	68.88	74.12	85.25	82.25	85.37	71.69	82.06	82.44
Patch 2	90.62	84.69	**69.13**	74.62	87.25	76.63	84.00	71.25	85.31	79.62
Patch 3	88.50	84.62	69.13	**74.62**	87.06	80.69	86.19	66.69	81.06	71.25
Patch 4	92.69	85.37	69.13	74.37	**88.00**	78.94	83.94	91.94	89.31	90.00
Patch 5	64.94	81.75	66.69	71.94	73.50	**83.63**	86.00	53.12	61.56	59.19
Patch 6	68.06	83.56	68.00	73.44	79.25	82.50	**86.75**	56.50	69.56	67.37
Patch 7	85.94	84.50	68.94	74.44	87.94	71.94	79.00	**96.81**	89.56	91.56
Patch 8	91.31	85.06	69.06	74.56	87.94	73.31	80.56	97.44	**89.50**	91.62
Patch 9	84.25	85.25	68.94	73.81	87.75	76.81	80.31	96.19	89.44	**91.94**
NASMobileNet										
Patch 0	**59.62**	55.87	56.50	59.25	57.25	54.31	55.94	61.94	58.19	57.94
Patch 1	56.50	**56.63**	55.50	57.94	56.31	54.44	57.13	59.87	56.69	56.25
Patch 2	59.75	56.69	**56.81**	59.44	58.06	54.37	57.56	64.31	59.19	57.81
Patch 3	57.63	55.94	56.44	**59.56**	57.63	54.44	56.88	60.00	57.63	53.75
Patch 4	58.19	56.19	56.31	58.25	**57.88**	54.44	55.94	65.00	58.75	58.63
Patch 5	55.31	56.19	54.62	56.56	55.94	**54.69**	55.87	57.56	55.31	54.87
Patch 6	55.31	56.50	54.62	58.25	56.38	54.50	**57.00**	58.00	56.50	54.75
Patch 7	58.25	56.00	56.00	57.25	58.06	53.69	56.19	**67.56**	59.81	59.13
Patch 8	58.19	55.81	56.12	58.13	58.25	53.81	56.63	66.81	**59.94**	59.06
Patch 9	59.56	55.75	55.81	56.44	58.19	54.12	55.50	66.94	59.87	**59.69**
ViT										
Patch 0	**98.12**	56.56	90.44	62.19	71.94	70.31	52.06	67.37	61.75	59.81
Patch 1	56.38	**97.87**	74.37	66.50	58.13	94.31	86.81	58.13	51.50	60.06
Patch 2	86.12	80.37	**98.19**	93.06	86.62	74.31	61.50	58.31	56.06	57.13
Patch 3	72.69	88.69	97.44	**99.25**	60.19	78.69	71.75	54.25	50.94	51.75
Patch 4	86.12	69.19	93.37	62.81	**98.69**	76.75	60.37	91.44	86.06	89.94
Patch 5	54.69	83.25	58.56	53.31	58.38	**96.50**	71.88	53.62	51.13	53.12
Patch 6	53.00	87.69	57.56	54.56	55.75	88.69	**96.69**	56.56	50.38	55.12
Patch 7	71.06	57.25	60.00	51.69	88.13	56.19	52.38	**98.44**	94.44	86.50
Patch 8	81.63	54.81	81.63	53.12	96.75	53.87	50.19	98.50	**99.31**	97.69
Patch 9	68.00	74.19	63.13	54.50	90.13	62.00	59.62	94.19	92.12	**98.75**

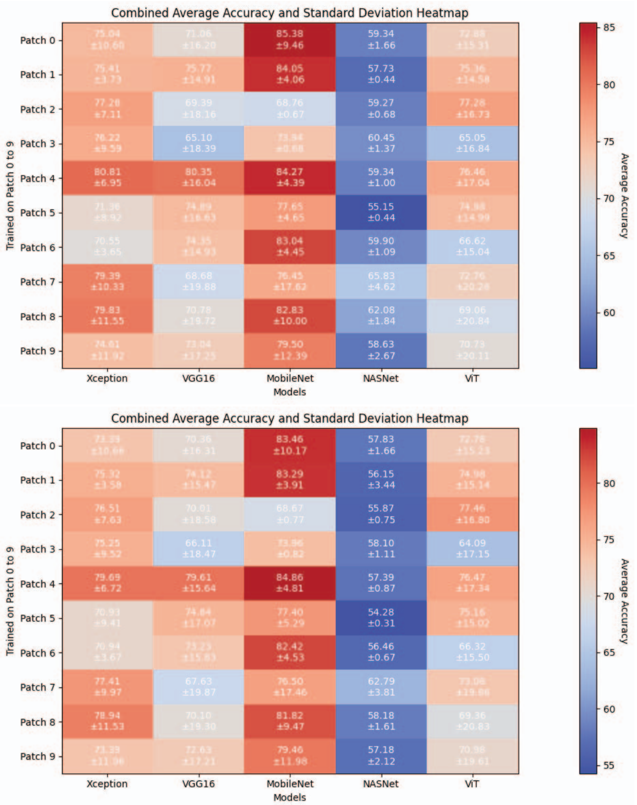

Figure 5. Average and standard deviation performance of each model when trained on individual adversarial patches of datasets (ImageNet: top and COCO: bottom). Here the average accuracies are from red to blue where red being the high effectiveness of a model and blue representing the ineffectiveness of a model in detecting patches.

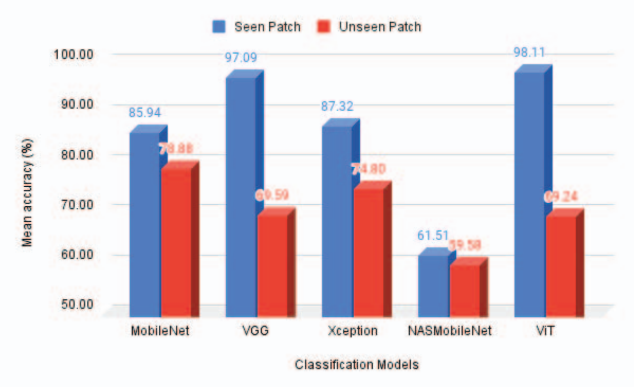

Figure 6. Average adversarial patch attack detection performance on the ImageNet subset under seen and unseen patches evaluation setting. The results reflect that when unseen patches come for classification, the performance of the networks drops drastically.

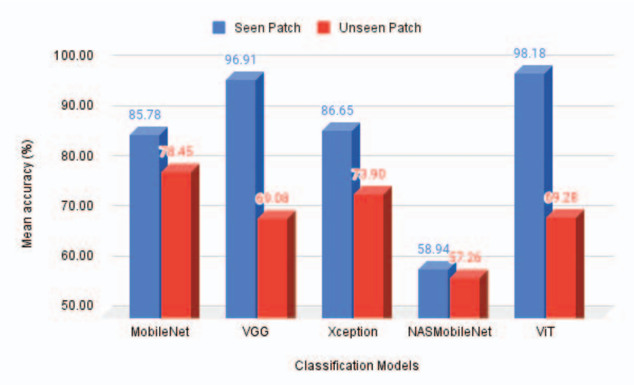

Figure 7. Average adversarial patch attack detection performance on the COCO subset under seen and unseen patches evaluation setting. The results reflect that when unseen patches come for classification, the performance of the networks drops drastically.

distributions. *In such real-world defense challenges, the MobileNet outperforms the other networks by a significant margin.* In terms of the performance of individual patches, it is observed that when the models are trained on patches 0 and 3, they are found most resilient in handling patch 2. Similarly, when the models are trained on patch 9, they are not only effective in handling patch 9 but can also detect patch 8. Patch 6 trained models are effective in handling patch 1.

Figure 6 and Figure 7 illustrates the robustness of each classifier when trained and tested on both seen and unseen attack patches of ImageNet and COCO dataset, respectively. The classifiers observed significant drops in performance when attempting to classify adversarial patches that were not seen during training. This decline in performance can be attributed to two main factors: first, the shift in style texture distribution among the patches, and second, the challenging blending of these patches with complex image regions.

An interesting finding is that if we only consider the percentage drop in accuracy, we can see that NASMobilenet demonstrates the highest level of generalizability across different patches. Although, it is also found to have the lowest network capacity, referring to its performance in seen patches training and testing conditions. For instance, NASMobileNet's performance drops by 3.1% in the unseen patches setting compared to the seen patches setting. Nevertheless, it still lags behind other architectures by at least 28.4% in the seen evaluation setting. Despite these challenges, NASMobileNet presents promising prospects for the development of a robust adversarial patch detection classifier. This could be achieved by intelligently incorporating adversarial patch information during the architecture search process, thereby enhancing its ability to detect attacks effectively.

5. Conclusion

The vulnerability of different convolutional neural networks, including vision transformers [3], to adversarial attacks is a significant concern when considering their real-world deployment. Out of several adversarial perturbations, an adversarial patch is one of the most complex and highly effective physical world attacks, yet there has been limited focus on developing defenses against it. To address this gap, we have created a comprehensive dataset containing $44,000$ images, comprising both real and adversarially patched examples. These adversarial patch examples can be seen as out-of-distribution (OOD) examples due to having significantly different distribution than the images on which the networks are trained. Using this dataset, we conducted an extensive benchmark study to understand whether the current image classification models are sufficient enough to detect these OOD samples. Our experimental analysis revealed that the image classifiers are effective in detecting the adversarial patch attack; however, the catch they must be seen during the training of the detector. The challenge is exacerbated when dealing with images from out-of-distribution settings, meaning datasets that were not seen or considered during training, testing, and evaluation. In the future, we aim to expand the dataset and develop a sophisticated and robust architecture for detecting out-of-distribution adversaries.

References

[1] Akshay Agarwal, Gaurav Goswami, Mayank Vatsa, Richa Singh, and Nalini K. Ratha. Damad: Database, attack, and model agnostic adversarial perturbation detector. *IEEE TNNLS*, 33(8):3277–3289, 2022.

[2] Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. Benchmarking robustness beyond l_p norm adversaries. In *ECCVW*, 2022.

[3] Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. Crafting adversarial perturbations via transformed image component swapping. *IEEE TIP*, 31:7338–7349, 2022.

[4] Akshay Agarwal, Nalini Ratha, Mayank Vatsa, and Richa Singh. Exploring robustness connection between artificial and natural adversarial examples. In *IEEE CVPRW*, pages 179–186, 2022.

[5] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini Ratha. Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? In *IEEE BTAS*, pages 1–7, 2018.

[6] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Nalini Ratha. Image transformation-based defense against adversarial perturbation on deep learning models. *IEEE TDSC*, 18(5):2106–2121, 2021.

[7] Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Cognitive data augmentation for adversarial defense via pixel masking. *PRL*, 146:244–251, 2021.

[8] Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini K Ratha. Noise is inside me! generating adversarial perturbations with noise derived from natural filters. In *IEEE/CVF CVPRW*, pages 774–775, 2020.

[9] Divyam Anshumaan, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Wavetransform: Crafting adversarial examples via input decomposition. In *ECCVW*, pages 152–168. Springer, 2020.

[10] Tao Bai, Hao Wang, and Bihan Wen. Targeted universal adversarial examples for remote sensing. *Remote Sensing*, 14(22):5833, 2022.

[11] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[12] Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, and Changick Kim. Improving the transferability of targeted adversarial examples through object-based diverse input. In *IEEE CVPR*, pages 15244–15253, 2022.

[13] Zhiyuan Cheng, James Liang, Hongjun Choi, Guanhong Tao, Zhiwen Cao, Dongfang Liu, and Xiangyu Zhang. Physical attack on monocular depth estimation with optimal adversarial patches. In *ECCV*, pages 514–532, 2022.

[14] Aran Chindaudom, Prarinya Siritanawan, Karin Sumongkayothin, and Kazunori Kotani. Adversarialqr: An adversarial patch in qr code format. In *IEEE icIVPR*, pages 1–6, 2020.

[15] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE CVPR*, pages 1251–1258, 2017.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255, 2009.

[17] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *IEEE CVPR*, pages 321–331, 2020.

[18] Yinpeng Dong, Jun Zhu, Xiao-Shan Gao, et al. Isometric 3d adversarial examples in the physical world. *NeruIPS*, 35:19716–19731, 2022.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[20] Gaurav Goswami, Akshay Agarwal, Nalini Ratha, Richa Singh, and Mayank Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. *IJCV*, 127(6-7):719–742, 2019.

[21] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[22] Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *IEEE CVPR*, pages 7848–7857, 2021.

[23] Zhanhao Hu, Siyuan Huang, Xiaopei Zhu, Fuchun Sun, Bo Zhang, and Xiaolin Hu. Adversarial texture for fooling person detectors in the physical world. In *IEEE/CVF CVPR*, pages 13307–13316, 2022.

[24] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[25] Hongshuo Liang, Erlu He, Yangyang Zhao, Zhe Jia, and Hao Li. Adversarial attack and defense: A survey. *Electronics*, 11(8):1283, 2022.

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[27] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *IEEE CVPR*, pages 1765–1773, 2017.

[28] Federico Nesti, Giulio Rossolini, Saasha Nair, Alessandro Biondi, and Giorgio Buttazzo. Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks. In *IEEE/CVF WACV*, pages 2280–2289, 2022.

[29] Maura Pintor, Daniele Angioni, Angelo Sotgiu, Luca Demetrio, Ambra Demontis, Battista Biggio, and Fabio Roli. Imagenet-patch: A dataset for benchmarking machine learning robustness against adversarial patches. *PR*, 134:109064, 2023.

[30] Jérôme Rony, Jean-Christophe Pesquet, and Ismail Ben Ayed. Proximal splitting adversarial attack for semantic segmentation. In *IEEE/CVF CVPR*, pages 20524–20533, 2023.

[31] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE CVPR*, pages 4510–4520, 2018.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[33] Inderjeet Singh, Toshinori Araki, and Kazuya Kakizaki. Powerful physical adversarial examples against practical face recognition systems. In *IEEE/CVF WACV*, pages 301–310, 2022.

[34] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[35] Donghua Wang, Wen Yao, Tingsong Jiang, Guijiang Tang, and Xiaoqian Chen. A survey on physical adversarial attack in computer vision. *arXiv preprint:2209.14262*, 2022.

[36] Futa Waseda, Sosuke Nishikawa, Trung-Nghia Le, Huy H Nguyen, and Isao Echizen. Closer look at the transferability of adversarial examples: How they fool different models differently. In *IEEE/CVF WACV*, pages 1360–1368, 2023.

[37] Xingxing Wei, Bangzheng Pu, Jiefan Lu, and Baoyuan Wu. Physically adversarial attacks and defenses in computer vision: A survey. *arXiv preprint arXiv:2211.01671*, 2022.

[38] Yonghao Xu and Pedram Ghamisi. Universal adversarial examples in remote sensing: Methodology and benchmark. *IEEE TGRS*, 60:1–15, 2022.

[39] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *IEEE CVPR*, pages 15345–15354, 2022.

[40] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *IEEE CVPR*, pages 8697–8710, 2018.