Brightness-Restricted Adversarial Attack Patch

Mingzhen Shao
Kahlert School of Computing
University of Utah
Salt Lake City, UT 84108
shao@cs.utah.edu

Abstract

Adversarial attack patches have gained increasing attention due to their practical applicability in physical-world scenarios. However, the bright colors used in attack patches represent a significant drawback, as they can be easily identified by human observers. Moreover, even though these attacks have been highly successful in deceiving target networks, which specific features of the attack patch contribute to its success are still unknown. Our paper introduces a brightness-restricted patch (BrPatch) that uses optical characteristics to effectively reduce conspicuousness while preserving image independence. We also conducted an analysis of the impact of various image features (such as color, texture, noise, and size) on the effectiveness of an attack patch in physical-world deployment. Our experiments show that attack patches exhibit strong redundancy to brightness and are resistant to color transfer and noise. Based on our findings, we propose some additional methods to further reduce the conspicuousness of BrPatch. Our findings also explain the robustness of attack patches observed in physical-world scenarios.

1 Introduction

Deep neural networks (DNNs) have experienced significant success across various domains in recent years, such as image classsification [11, 5], object detection [10, 16], speech recognition [19], and natural language processing [24]. However, the nature of DNNs also makes them vulnerable to adversarial attacks that are crafted by adding carefully designed perturbations on normal examples [2, 9, 13, 18, 20, 21]. Since DNNs have become more critical to some safety-critical applications, such as autonomous driving and biometric authentication, their susceptibility to adversarial attacks raises serious concerns about their safety and reliability [17, 7, 4].

Adversarial attacks in computer vision tasks can be categorized into two domains: digital and physical. In the digital domain, attackers can access the digital values of inputs and make arbitrary pixel-level changes to inputs. However, such ideal conditions are challenging to achieve in the real world. Data security measures in well-designed software are typically difficult to breach. Once an attacker successfully bypasses these measures, further manipulation of the DNNs may become unnecessary. Physical domain attacks, on the other hand, may be more realistic in practice as they assume that only the physical layer objects, such as the environment or objects that the system interacts with, can be manipulated.

Image-dependent and image-independent are the two main types of attacks in the physical domain. Image-dependent attacks require a design of each attack tailored to the specific target image. In most conditions, these attacks need to replace the target with a modified object. For example, to make DNNs misclassify a stop sign, a modified stop sign image needs to be created and substituted for the original one. However, this approach can be highly inefficient and even impossible if the target is not an image. In contrast, image-independent attacks use an additional object (patch) to get rid of this requirement. The patch is trained to create a physical world attack without prior knowledge of the

Preprint. Under review.

other items within the scene. The patch can be put in any environment to launch an attack without replacing the targets.

In spite of the great success the attack patches have shown in deceiving target networks, the bright and vivid coloration of these patches can also be a significant drawback. In many real-world scenarios, attack patches must remain low visibility to human observers, particularly in security applications where an attacker may seek to evade visual surveillance. In order to reduce the visibility of attack patches, a straightforward question to ask is whether their vivid coloration is strictly necessary for a successful attack because reducing this feature can significantly decrease the patch's visibility to human observers.

In this work, we first introduced a brightness-restricted patch (BrPatch) to reduce visibility while maintaining a high attack success rate. By using optical characteristics (brightness) to minimize detectability, the BrPatch can still maintain image independence, which means the BrPatch does not require additional training for specific scenarios. We then conducted an analysis of the impact of various image features (such as color, texture, noise, and size) on the effectiveness of an attack patch in a physical-world deployment. Based on our findings, we proposed a hue mapping method to further reduce the visibility of the BrPatch. Furthermore, we evaluated the performance of the proposed BrPatch in the physical world to demonstrate that the BrPatch still achieves an attack success rate comparable to the original adversarial attack patch in the real world. To the best of our knowledge, this study is the first attempt to use optical characteristics (brightness) to reduce the visibility of an adversarial attack patch and analyze the mechanisms underlying its effectiveness.

The main contributions of this paper can be summarized as follows:

- We propose a brightness-restricted adversarial attack patch (BrPatch) to reduce visibility while preserving image independence.
- Our experiments show that adversarial attack patches exhibit strong redundancy to brightness restrictions.
- We demonstrate that color transfer and random noise in the physical world will not significantly affect the performance of attack patches.

2 Related work

Adversarial attacks for deep learning were first introduced by szegedy *et al.* [18]. Since the publication of their seminal work, numerous researchers have proposed more efficient methods for generating adversarial attacks. These adversarial attacks commonly modify each pixel by only a small amount and can be found using a number of optimization strategies such as the Fast Gradient Sign Method (FGSM) [9], Projected Gradient Descent (PGD) [14], and Skip Gradient (SGD) [22]. Some other attacks seek to modify a small number of pixels in the image [15], or a small patch at a fixed location of the image [17]. However, all these approaches assume that the attacker has digital-level access to the inputs, which limits the range of scenarios in which the attacks can be used.

Therefore, Kurakin *et al.* [12] proposed the first physical-domain attack model by printing digital adversarial examples onto paper. They found that a significant portion of the printed adversarial examples deceived the image classifier. Athalye *et al.* [1] improved on this work by creating adversarial objects that remain effective even when viewed from different angles. They achieved this by modeling small-scale transformations synthetically when generating adversarial perturbations. They demonstrated that the resulting adversarial objects deceived target classifiers. In their paper, they claimed that the algorithm is robust to rotations, translations, and noise as long as the transformation can be modeled synthetically. Eykholt *et al.* [8] also developed an attack algorithm that can generate physical adversarial examples. In contrast to Athalye *et al.*, they modeled image transformations both synthetically and physically. Their target dataset included certain image transformations, such as changes in viewing angle and distance. They synthetically applied other transformations, such as changes in lighting, when generating adversarial examples. Their work suggested that by relying solely on synthetic transformations, subtleties in the physical environment can be overlooked, resulting in a less robust attack. However, all these works require the design of each attack to be tailored to the specific target image, which dramatically limits their applicability in various scenarios.

In order to provide an image-independent adversarial attack model, Brown *et al.* [3] proposed a new approach for creating an adversarial attack patch. This patch can be placed anywhere within the field of view of a classifier and launch an attack. Because this patch is scene-independent, it allows attackers to create a physical-world attack without prior knowledge of the lighting conditions, camera angle, type of classifier being attacked, or even the other items within the scene. Since then, many studies have focused on improving physically feasible attacks aimed at deceiving classifiers or object detectors, such as traffic signs [7], cloaks [23], or vehicles [25]. However, as Brown *et al.*mentioned in their paper, these attack patches are not restricted to imperceptible changes. These patches have striking color, which is very conspicuous to human observers. One approach to reducing the visibility of the attack patch has been proposed by Duan *et al.* [6]. They modified the patch into some natural styles that appear legitimate to human observers. Their approach can be an effective way to evade detection, but it requires generating a new patch for each attack scenario. This regeneration is contrary to the main advantage of using attack patches, which is the ability to be trained once and deployed universally, without being dependent on specific images. Furthermore, in their physical-world experiments, they chose to replace the target object with the modified image rather than using the patch with the original target. These experiments simplified the 3D position relationships between the patch and the target and also suffered the drawback that some targets cannot be replaced with printed images.

3 Method

In this secession, we first provide a method to generate the brightness-restricted attack patch (BrPatch). Then we provide an easy hue mapping method in the RGB color model to further reduce the visibility of attack patches.

3.1 Generating brightness-restricted patch

Given an image $x \in [0,1]^{w \times h \times c}$ with class y, and an attack patch p, let T be a transformation function that can involve location, rotation, and scale. We define $T(p, x)$ as the input image obtained by applying T to patch p to get the transformed patch, and then overlaid onto x.

The adversarial loss for a targeted attack can be formed as follows:

$$L_{adv} = log(softmax(Pr(\hat{y}|T(p,x)))) \tag{1}$$

where the $\hat{y}$ is the target class and $\hat{y} \neq y$, Pr is the prediction of the target model with respect to class $\hat{y}$.

In order to reduce the intensity of vivid colors in an image, it is necessary to manipulate the brightness component in the HSB (Hue, Saturation, Brightness) color model. However, to prevent the need for switching between multiple color models, we introduced a brightness-restricting loss function within the RGB color space:

$$L_b = log(1 - mse(p, p_r)) \tag{2}$$

where $mse(p, p_r)$ calculates the mean square error between the patch p and a reference patch p_r. In printing systems, white is used to represent the absence of any ink. Therefore, we use an all-white patch as the reference patch p_r.

The final loss is a combination of L_{adv} and L_b:

$$L = L_{adv} + \lambda L_b \tag{3}$$

where λ is a parameter that adjusts the strength of the brightness-restricting loss.

3.2 Hue mapping in RGB color model

Given a patch p and a target region on original image x_t, calculate the hue difference between the patch and target region:

$$\Delta H = ave_c(p) - ave_c(x_t) \tag{4}$$

where $ave_c(\cdot)$ computes the average value of each channel (red, green, and blue) of the input patch p and the target region x_t, respectively.

We use a threshold H_t to prevent potential overflows of values in the patch p after the hue mapping process:

$$p' = \begin{cases} p + \Delta H, & \text{if } \Delta H \leq H_t \\ p + H_t, & \text{otherwise} \end{cases} \tag{5}$$

where p' is the patch after hue mapping.

4 Experimental results and analysis

In this section, we first outline the experimental setup. Then we show the performance of the proposed BrPatch in different settings and demonstrate that BrPatch can significantly reduce the visibility of attack patches without compromising their effectiveness in deceiving the target network. The experiments also show that adversarial attack patches have strong redundancy to brightness restrictions. After that, we conduct an analysis of the impact of various image features (such as color, texture, noise, and size) on the effectiveness of an attack patch in the physical world deployment. Based on our findings, we use hue mapping to further reduce the visibility. Finally, we evaluate the robustness of the BrPatch in physical-world attacks.

4.1 Experimental setup

In our experiments, we use a gray-box setting for the threat model: the source and target networks are both ResNet50 networks but were trained separately on ImageNet1K. We perform 40 epochs of training for each setting and select the patch that achieves the best performance as the final output patch. The input images have a size of 224x224, and their values range is $[0, 1]$. In order to test the attack success rate, we randomly select 1,000 images in ImageNet1K as source images, and apply the attack patch to these images.

For a physical-world test, we use an HP laser printer to print adversarial attack patches and a Google Pixel6a to take photos. To further reduce the visibility of the BrPatch during physical deployment, we first capture a photo of the target and apply hue mapping to the BrPatch. Then we print the hue mapped BrPatch **only** on A4 paper. We place the printed BrPatch next to the target and take another photo that includes both the target and BrPatch. All our experiments focus on targeted attacks, as untargeted attacks can be viewed as a special case of targeted attacks where the target can be any valid target value.

4.2 Performance of brightness-restricted patch

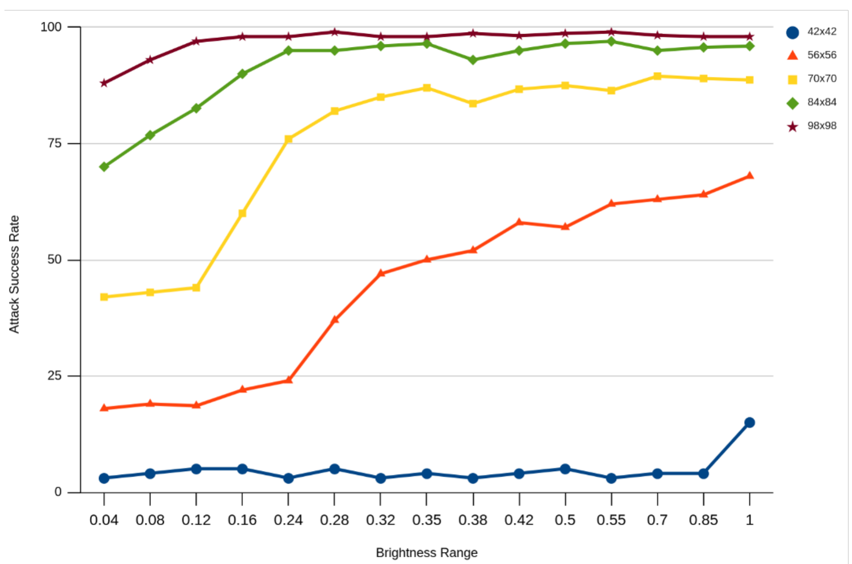

Figure 1: Attack success rate with different patch sizes and brightness ranges.

Figure 1 shows the performance variations for different brightness ranges and patch sizes on ResNet50, where each line corresponds to a specific patch size. The impact of patch size on the performance is apparent from the figure, with larger patches showing better performance. We also find that all patches show redundancy to brightness, with larger patches being more tolerant to stricter restrictions than smaller ones.

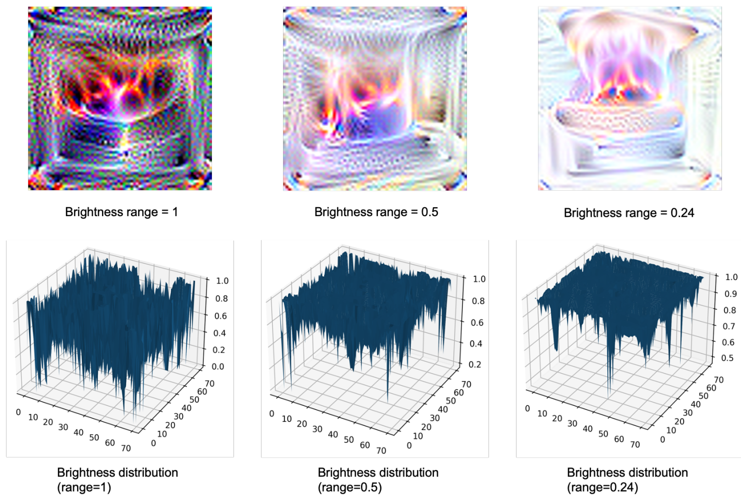

Figure 2: Patches with different brightness restriction and their brightness distribution. (Top: Different brightness-restricted patches; Bottom: Corresponding brightness distributions to each patch.)

Some brightness distributions of different BrPatch are shown in Figure 2. We choose a patch size of 70x70 due to its high sensitivity to brightness restriction and ability to achieve high performance without brightness restriction (original adversarial attack patch). For the following analysis, we will continue to use attack patches of the same size.

Referring to the performance shown in Figure 1, our results indicate that the adversarial patches can achieve a high attack success rate in deceiving the target network with only a small range of brightness values. For instance, the performance remains consistent for the patch sizes of 70x70 even when up to 65% of the brightness is lost. This finding allows us to decrease the brightness of an attack patch without sacrificing its performance, effectively reducing the visibility of the patch.

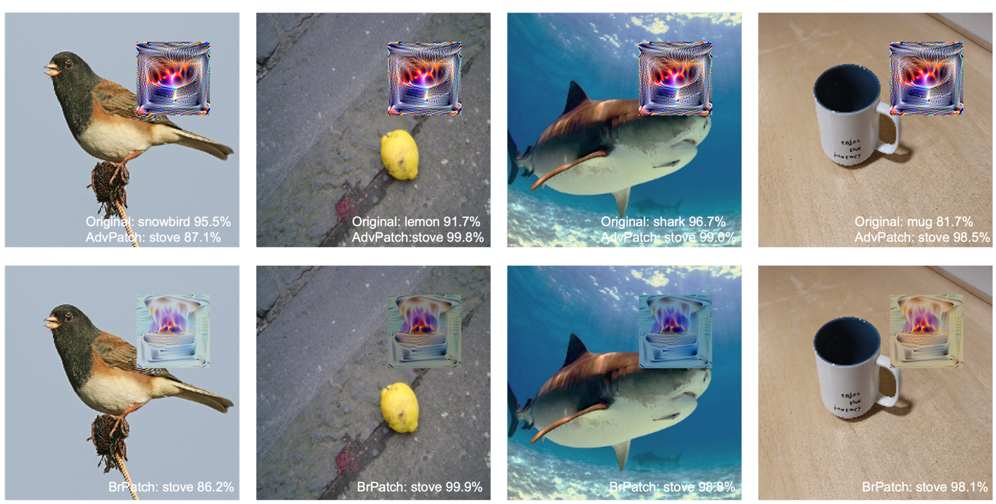

Figure 3: Images with different patches. (Top: AdvPatch; Bottom: BrPatch)

Figure 3 demonstrates a set of images with an original adversarial attack patch (AdvPatch) and a BrPatch (brightness range=0.24). The proposed BrPatch is observed to be less visible to a human observer in all scenarios and does not require any additional training. The performance of the BrPatch is also comparable to that of the AdvPatch.

4.3 Effectiveness of different features in the physical world

4.3.1 Color transfer and texture blurring

Figure 4: Patch with different feature changes. (left: original patch; middle: color transfer patch; right: texture blurring patch)

In physical-world deployments, attack patches are often affected by color transfer due to lighting conditions or blurring caused by camera focus or smudging. To analyze if these changes will affect the performance of the attack patches, we compare the performance of an original patch, a color-adjusted patch, and a local texture adjusted patch. The color-adjusted patch is applied by adding a value (δ) to all values in RGB channels and making sure δ will not lead to an overflow. This color transfer will not change the texture information of the patch. The local texture adjustment is applied by using a 3x3 Gaussian blur. Figure 4 demonstrates the two types of feature adjustments applying to a patch with brightness range=0.24.

Table 1: Performance with color transfer and Gaussian blur

Brightness range	Original	Color transfer	Gaussian blur
1 (AdvPatch)	89.4%	90.8%	47.8%
0.35	89.5%	87.9%	22.7%
0.24	74.2%	75.3%	10.1%

The performances of different patches are shown in Table 1. Regardless of the lightness restriction, the color transfer patch achieves almost the same success rate as the original patch. This performance shows that the patch does not need to maintain a specific color to deceive the target network. On the other hand, the blurred patch exhibits a significant decrease in success rate, suggesting that local texture is the key feature in deceiving target networks. Using these findings, we can apply the proposed hue mapping method to adjust the color of the patch and enhance its integration with the target environment, resulting in further reduced visibility. This process does not require any learning and can be quickly applied when deploying the patch in the physical world.

4.3.2 Random color variations

When printing an attack patch, it is important to consider that normal printers are not able to produce a patch with precisely the same color as the digital version. Therefore, the patch's robustness to random color variations must be evaluated.

To replicate the color drift that commonly occurs during printing, we generate random noise within a restricted range that corresponds to a percentage of the original value. This approach allows us to simulate different levels of drift, and the results are shown in Table 2.

Table 2: Performance with different color drift

Brightness range	Original	10% drift	15% drift	20% drift
1 (AdvPatch)	89.4%	87.6%	85.9%	83.3%
0.35	89.5%	84.2%	77.6%	67.2%
0.24	74.2%	68.6%	43.2%	27.3%

We notice that even with 10% random color drift, the attack patches can still maintain an attack rate as reliable as that for the original clean patches. This performance suggests that the patch is relatively robust and can withstand slight color variations during printing. Figure 5 demonstrates a set of attack patches with different noise levels. As the level of noise increases, the patch experiences certain color deviations. However, adjusting the color of the patch is distinct from this process because it uniformly alters all the pixels to maintain the local texture while achieving the desired color.

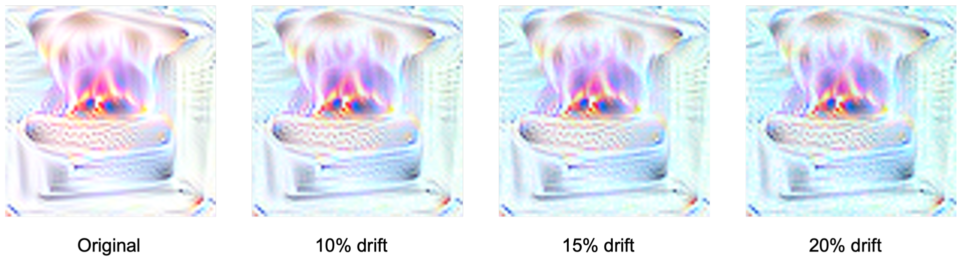

Figure 5: Patches with different color drift.

4.3.3 Patch scaling

In practical deployments, it is often the case that the environment can accommodate larger sized attack patches. As demonstrated in the previous sections, training larger sized patches can result in improved attack performance. However, in such scenarios, there may not be enough time to train a new patch with a custom size. Therefore, we are motivated to investigate whether we can enhance the performance of a pre-existing attack patch by using interpolation.

We use a patch with brightness range=0.24, and its initial attack success rate is 74.2%. Then we use bilinear interpolation to get patches with sizes 84x84, 98x98, and 112x112. The success rates of the different patches are shown in Table 3.

Table 3: Performance across different sizes

Patch size	70x70	84x84	98x98	112x112
Success rate	74.2%	83.4%	87.4%	88.4%

We notice that the attack success rate increases as the patch size grows. However, we also observe a diminishing marginal utility with each increment in size. In fact, we find that the success rate of the bilinear interpolation patch, even when it is the same size as the original patch, is lower. Some images with different scaled patches are shown in Figure 6.

Figure 6: Patch scaling. (left: original 70x70 patch; middle: scaled 98x98 patch; right: original 98x98 patch)

4.4 Physical-world attacks

The physical-world deployability of normal attack patches has been demonstrated in many works. However, a reasonable concern about the proposed BrPatch is whether the restricted brightness reduced the robustness of the patch in physical-world attacks. In order to address this concern, we designed several physical-world deploy instances to show that the proposed attack patch can still work robustly in the physical world.

We have selected two typical scenarios for the experiment: outdoor with natural light and indoor with artificial light. The artificial light has a color temperature of around 3000K (warm white), whereas the natural light has a color temperature of around 6500K (daylight). For each scenario, we took several images from different angles and distances.

Figure 7 shows some figures and the predictions with different patches. It is first observed that although the conspicuousness of the printed BrPatch is not as insignificant as in the digital domain, it still greatly reduces the visibility compared to the normal attack patch (especially with the hue mapping). We can then find that the probability score of the proposed patch does not obviously decrease compared to the original adversarial attack patch. This experiment demonstrates that the proposed BrPatch remains sufficiently robust for physical-world deployment even when losing over 65% of the brightness.

Figure 7: Comparison between AdvPatch and BrPatch in the physical world

5 Conclusion

This paper presents a novel approach to reduce the visibility of adversarial attack patches in the physical world: the brightness-restricted patch (BrPatch). The BrPatch is the first attempt to use optical characteristics (brightness) to minimize detectability. This approach allows BrPatch to maintain image independence and significantly enhances its practical applicability in physical-world attacks. We analyze the impact of various image features (color, texture, noise, and size) on the effectiveness of an attack patch in physical-world deployment and show that attack patches exhibit strong redundancy to brightness and are robust to color transfer and noise. Our experiments also demonstrate the robustness of the BrPatch to different physical world scenarios. The BrPatch presents an advanced camouflage technique, offering a reliable and effective solution for safeguarding objects or individuals from being detected by DNN-based equipment.

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293, 2018.

[2] Yang Bai, Yan Feng, Yisen Wang, Tao Dai, Shu-Tao Xia, and Yong Jiang. Hilbert-based generative defense for adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4784–4793, 2019.

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

[4] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019.

[5] 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.

[6] Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A Kai Qin, and Yun Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1000–1008, 2020.

[7] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2(3):4, 2017.

[8] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[12] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. 2018.

[13] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

[14] Aleksander Madry, Aleksander Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial examples. *arXiv preprint arXiv:1706.06083*, 2017.

[15] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387, 2016.

[16] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[17] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016.

[18] 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.

[19] Yisen Wang, Xuejiao Deng, Songbai Pu, and Zhiheng Huang. Residual convolutional ctc networks for automatic speech recognition. *arXiv preprint arXiv:1702.07793*, 2017.

[20] Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. *arXiv preprint arXiv:2112.08304*, 2021.

[21] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020.

[22] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. *arXiv preprint arXiv:2002.05990*, 2020.

[23] Zuxuan Wu, Ser-Nam Lim, Larry S Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 1–17, 2020.

[24] Min Zeng, Yisen Wang, and Yuan Luo. Dirichlet latent variable hierarchical recurrent encoder-decoder in dialogue generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language*

Processing (EMNLP-IJCNLP), pages 1267–1272, 2019.

[25] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. Camou: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2019.