DropAttack: Leveraging Realistic Templates for Enhanced Universal Adversarial Attacks on Detectors

Subrat Kishore Dutta (7028082) Somrita Ghosh (7028737) Soham Roy (7028704)

1 Abstract

Physical adversarial attacks involve manipulating the real-world environment, such as adding patches or changing objects, to deceive machine learning models, particularly computer vision systems. These attacks have gained popularity in recent years, yet traditional approaches demand direct access to the target object for patch application, that is placing the patch on the object instances. Consequently, covering multiple objects necessitates applying individual adversarial patches to each one of them. We introduce a physical patch that is fixed to the camera lens to deceive cutting-edge object detectors. Our primary objective is to conceal all instances of a specified target class and achieve high attack ability. As a secondary objective, we prioritize the inconspicuousness of the applied patch by introducing templates that are homogeneous to the context of the image. Our experiments, conducted on state-of-the-art object detection models that mimic the employed systems in autonomous driving, examine the impact of the patch on selected target classes. Our attack has demonstrated a notable enhancement, surpassing the current state-of-the-art by achieving around a 6% improvement in average precision and around a 5% increase in the fooling rate.

2 Introduction

Convolutional neural networks(CNN) have been the most prominent architectural choice when it comes to computer vision tasks capturing almost all its various realms [8, 3, 2, 15, 16, 13, 17]. Even though deep neural networks have been able to achieve near human-level performance due to their ability to capture the underlying multivariate data distribution, these models are notoriously susceptible to attacks from specifically curated perturbation which makes the DNN model to misclassify or to give specifically targeted predictions as desired by the attacker. These attacks have been extensively studied in computer vision specifically due to their applicability in the digital as well as the real world [1, 12]. Works like [6, 14] consider the question of making adversarial attacks possible in the real world while maintaining a high degree of stealth. Works that tried to attack object detector models like [4] generated adversarial stop signs by incorporating distinct background patterns through a printing process. [9] attacked the objects represented in an image by placing very small pixel-level perturbations that are highly imperceptible.

Recent research illustrates evasion attacks, where objects are misclassified using physical patches. Instances include hiding individuals with cardboard plates [19], wearable attire [20, 21, 10], and more. However, these physical attacks often necessitate direct access to the object for patch application, limiting scalability, especially when concealing multiple objects. In the domain of detection tasks and autonomous driving, recent strides in attack methodologies have demonstrated remarkable success, but require direct object accessibility, limiting their applicability to specific instances[9][19]. Often, numerous objects necessitate multiple patches, posing challenges when access to the objects is restricted. Our approach draws inspiration from the study of Zolfi et. al[23], which aims to address this problem by focusing on attacking specific classes without direct object interaction. This paper generates a carefully constructed pattern containing multiple shapes of different orientations on a translucent film that is placed on the camera lens. The paper mentions that as the number of shapes increases the attack success rate increases but after a certain point it gets saturated due to the overlapping of shapes. Also, with the increase in the number of shapes, the patches become more perceptible. Unlike the original study's technique, our method diverges by leveraging natural-

looking patch templates, such as droplet images, to counteract the perceptibility of patches to the human eye. This innovation aims to boost patch effectiveness even as the number of shapes increases, circumventing the saturation point caused by shape overlap highlighted in prior research. By employing these templates, we ensure that the shapes do not overlap, thereby yielding superior outcomes in targeted attacks.

Our main contributions are :

1. We employ templates tailored to the image context, enhancing the inconspicuous nature of our attack.
2. The challenge of overlapping shapes as mentioned in Zolfi et. al[23] is addressed through the introduction of realistic templates which prevent the overlapping of objects.
3. Human evaluation is conducted to quantify the imperceptibility and impact of our patch on human observers.

3 Related works

Brown et. al [1] described the notion of the universal adversarial patch which are perturbations localized to a small region but are not bounded by an ϵ value for targeted attacks and applies to any image. The study also highlighted its applicability in the real world as stickers. Karmon et al.[12] in their work LAVAN described a similar methodology but kept it limited to digital attacks. Hayes[7] demonstrated that a sparse adversarial patch spread across a larger area over the original image can achieve better results even after keeping the percentage of perturbed pixels the same. The newly constructed attack could easily surpass the current level of defenses against adversarial patch attacks because the defenses back then were not equipped to deal with large perturbations.

Eykholt et al. [6] proposed centralized physical perturbations, such as applying black and white stickers on stop signs, to deceive image classifiers. The approach also restricts the perturbations applied to these stickers so that imperceptibility and inconspicuousness can be attained. The idea can be extended to detectors as well. Chen and colleagues [4] created adversarial stop signs adorned with particular background designs to circumvent object detectors, while Sitawarin and collaborators [18] fabricated deceptive signs similar in appearance to authentic traffic signals to outsmart autonomous vehicle systems. Thys et al. [19] proposed an attack on detection models where they generated adversarial patches specifically targeting objects with significant intra-class diversity. Huang et. al [9] highlighted that large perceptible patches are not required for attacking detector models, and introduced a patch selection and refining scheme which successfully fooled YOLOv4 and Faster RCNN with 100% missed detection rate with just 0.32% of its pixels perturbed.

[14] introduced the idea of context homogeneity in achieving inconspicuousness. The central idea is drawn from the consideration that although the adversarial perturbations are visible to the observer, due to their prior knowledge about the original image it does not raise any suspicion to the observer. Zolfi et. al [23] highlighted the accessibility issues associated with targeting multiple objects and proposed a contactless translucent physical patch that can be applied physically over a camera for the attack. The patch, designed with a specific pattern, effectively conceals instances of a targeted class, showcasing a 42.27% success rate in preventing the detection of stop signs while maintaining robust detection for other classes. In our work, we are mainly motivated by the works of [14, 7] and [23] where we intend to place realistic templates to achieve inconspicuousness while not accessing individual objects by spreading the perturbation all across the input.

4 Methodology

Our approach describes the generation of an adversarial patch that can be applied to a physical camera lens. We consider the digital setting for the experiments which consists of a 2D image with four channels for the perturbations: RGB color channels and an alpha channel denoting pixel opacity that comes from the template image. The application process of the patch to the original image involves alpha blending so that the contents of the original image are not replaced resulting in a natural and realistic look of the perturbed image.

4.1 Constructing the structure of the patch

In the context of creating adversarial patches for an image, a reference image (such as a screen with droplets) serves as a template for the location where perturbations would be added. The template is first converted into a grayscaled form, resized to fit the input size, and processed through binary thresholding to isolate the specific locations, likely droplets. Erosion operation is applied so that the locations captured are pronounced in the formed mask. Subsequently, contours are identified on the processed template, and the regions enclosed by contours with an area surpassing a defined threshold are chosen for the final mask M. These selected regions indicate potential locations where adversarial patches can be strategically placed for a successful attack. In the selected locations located by the mask M, we intend to mimic the appearance of a droplet for which we utilize alpha blending between the original image and the perturbation tensor. This considers the weighted sum of the RGB color channels and the alpha channel at each pixel (i, j). The grayscaled image of the template serves as the alpha map for the alpha blending as its values are between 0 and 1. This very process also incorporates the appearance of droplets in each region thereby contributing to the inconspicuous nature of the final perturbed image. This approach aims to generate convincing adversarial patches that seamlessly blend into the visual context of the original scene. The formulation for the perturbation application is as follows:

$$\hat{x} = M \circ x + (1 - M) \circ (\gamma * x \circ \alpha + \beta * (1 - \alpha) \circ \delta) \tag{1}$$

where M represents the mask for the location of the perturbation, x represents the input image, α represents the α map, and δ is the perturbation tensor. γ and β represent the balancing terms for the perturbation and the original image in the alpha blending process where we used γ as 1 and β as 1.2. $\circ$ represents an element-wise dot product.

To initialize δ we used two approaches where in the first method we initialize δ with random values drawn from a normal distribution and in the second method we initialize each channel of δ with a normalized grey value of 0.25 which gives the final tensor a grey color. Figure 1 represents examples of perturbed images with different parameter settings. The central intentions of the two methods are as follows:

- Considering random values results in a final perturbation with iridescent coloration. This mimics the dispersion effect caused by normal droplets when white light passes through it, thus maintaining context homogeneity.
- Grey color initialization paired with alpha blending and static color update results in a perturbed image where the saliency is as minimal as possible.

4.2 Patch optimization:

We optimize the δ to reduce the custom loss function that we represent below in 2.

$$\mathrm{L_T} = w_1 \cdot \mathrm{L_{tc}} - w_2 \cdot \mathrm{L_{uc}} + w_3 \cdot \mathrm{L_v} + w_4 \cdot \mathrm{L_{norm}} \tag{2}$$

The loss function employed for the adversarial attack is designed to manipulate the model's predictions in a targeted manner while preserving the visual coherence of the generated perturbations. The loss function is composed of four components that attain three specific objectives. The targeted confidence loss represented by L_{tc} encourages the model to push down the confidence for the target class in each of the cells in the detection grid. We consider a high value for the w_1 weight so that the contribution to the loss is maximized hence updates for perturbation can cater more to the effectiveness of the attack. The untargeted confidence Loss denoted by L_{uc} attempts to preserve the untargeted classes by predicting high confidence for them. It is to be noted however, preserving the untargeted classes is not the topmost priority of this attack as the observer will expect some drop in the performance of the system when they see droplets on the camera lens. We intend to achieve enough confidence for the untargeted class so that it gives the observer a false sense of safety and raises no suspicion. Consequently, we set a relatively small value for the w_2 weight. The variation loss and regularization term represented by L_v and L_{norm} respectively encourage the perturbation to be smooth and also restrict the perturbation amount discouraging extreme values in the perturbation so that the saliency can be reduced. The weights w_3 and w_4 regulate the restrictions on the perturbation.

We varied the update rule for the perturbation in the optimization process where for the first approach we updated the randomly initialized perturbation using Adam optimizer with a learning rate of 0.1.

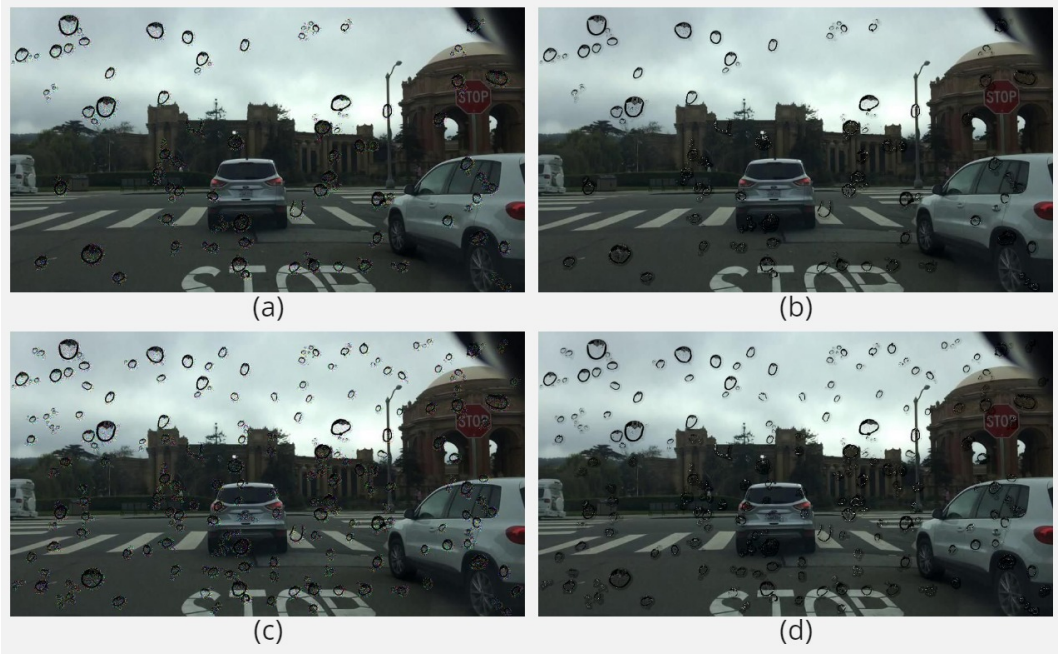

Figure 1: Examples of an image with our attack template having parameter values (a) area threshold of 400 and grey initialization (b) area threshold of 400 and Gaussian initialization (c) area threshold of 200 and grey initialization (d) area threshold of 200 and Gaussian initialization.

In the second approach, we followed a new update rule which does not change the base color of the pixel but only alters the saturation and brightness of the pixel. We achieve this by keeping the rate of change the same for each channel of the perturbation tensor. The update rule is a modification of the rule mentioned by [5] and that is given as follows:

$$\delta = \delta - \alpha * \overline{\nabla_\delta L_T} * \delta \tag{3}$$

where $\overline{\nabla_\delta L_T}$ represents the mean of the gradient over the three channels and α represents the learning rate which we set to 0.1. For both optimization rules, we updated the patch using 1000 iterations for each image.

4.3 Evaluation

Model and Dataset: We evaluate our attack against the YOLOv5 [11] model in a white box setup. We use pretrained weights for YOLOv5 trained on the MS-COCO dataset which is trained to recognize 80 object categories encompassing several domains. We select two classes namely traffic light and stop sign as our target classes while person, bicycle, car, motorbike, airplane, bus, train, truck, and boat are taken as untargeted classes. We use Berkeley DeepDrive(BDD)'s [22] train dataset to train the patch which contains 7000 images whereas we used the test set containing both instances of traffic light and stop signs to evaluate the formed patches.

Evaluation: We evaluate the attack on two aspects. Firstly we evaluate the attack ability of the attack by considering the per-class average precision(AP) and per-class fooling rate(FR)[23]. Secondly, we conducted a human-based evaluation where we showed participants a total of 40 images each containing the template however 20 of them contained perturbation while the other twenty did not. The observer as a secondary task needs to pick out the images that they think are perturbed and which they think are not while doing a primary task of operating a simulation of driving a car. The experimental setup for this is demonstrated in Figure3 The intention behind parallelly playing a driving game is to simulate an environment where the observer is driving a car and the video feed is seen on the dashboard. We kept the size of the images shown similar to that of a normal dashboard size. We then evaluate the Inconspicuous Score as:

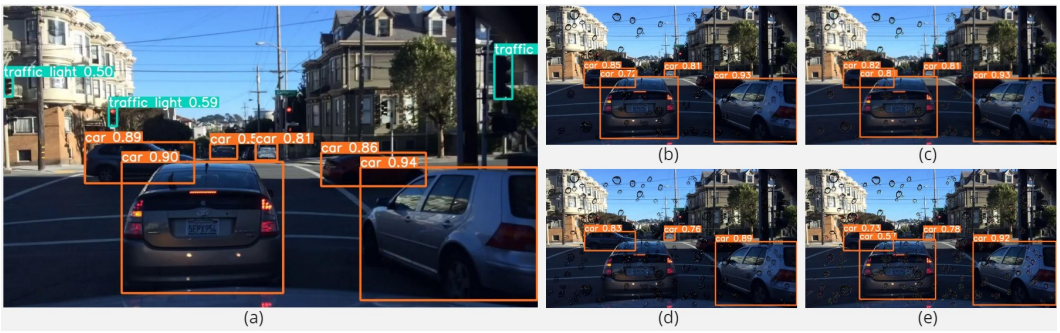

Figure 2: Illustrating the results of our attack on the YOLOv5 model through visual representations of detection on the (a) original Image (b) attacked image with grey initialized template and area threshold of 400 (c) attacked image with Gaussian initialized template and area threshold of 400 (d) attacked image with grey initialized template and area threshold of 200 (e) attacked image with gaussian initialized template and area threshold of 200.

$$IS = \frac{N_{IC}}{N_T} \tag{4}$$

Where N_{IC} is the number of incorrectly identified samples and N_T is the total number of samples. The added normal samples add a level of complexity for the observer.

5 Experiments and results

We conducted two sets of experiments to evaluate the attack success in a white box setup.

Attack performance on varying parameters: We consider two key settings and analyze the attack by varying these two aspects namely the area coverage of the patch and the initialization of the perturbation. To analyze the proportion of coverage the template needs to conduct a successful attack we varied the area threshold while creating the mask for the perturbations and analyzed the attack performance. For each area threshold value, we assessed two different initializations: Gaussian normal initialization and grey value-based initialization as described in the previous section. We evaluated the attack success based on per class Average Precision as described earlier with a confidence threshold of 0.4 mainly to follow the same methodology as [23]. We used the detector output on the original image as the ground truth to evaluate the Average Precision. We also evaluated the impact of the perturbation on the attack by analyzing a controlled setting where we only overlay the template with no perturbation and studied its attack ability for the class "traffic sign". We used this setting as the baseline for our attack. We also studied the fooling rate of the attack where we used traffic light and stop sign class as the target class. All the results are summarized in the 1 and 2. Figure 2 shows the detection on the perturbed examples.

Area	Initialisation	Baseline	"traffic light" Class	"stop sign" Class	Other Class
200	random	47.3	25.24	33.71	58.66
400	random	60.01	46.87	49.81	65.95
200	grey	47.3	41.25	42.31	66.61
400	grey	60	47.61	50.45	71.52

Table 1: Average Precision as a function of area threshold and initialization.

From Table 1 and Table 2 we can see that from all of our attacks even though the patches with Gaussian initialization achieve the best attack capabilities they also compromise the model's detection of the other classes significantly compared to other form of initialization. In both initializations, we observed that an area threshold of 200 achieves better attack abilities even though it covers a larger area of the input image. Considering both the criteria of keeping the attack capabilities high and preserving the other class's confidence we propose that the grey initialization with an area threshold of 200 maintains a good balance between both the criteria. It is to be noted that this particular selection

Area	Initialisation	"traffic light" Class	"stop sign" Class	Other Class
200	random	75.20	53.62	41.2
400	random	40.49	35.71	27.76
200	grey	52.81	47.81	27.2
400	grey	48.76	35.71	32.32

Table 2: Fooling rate as a function of area threshold and initialization.

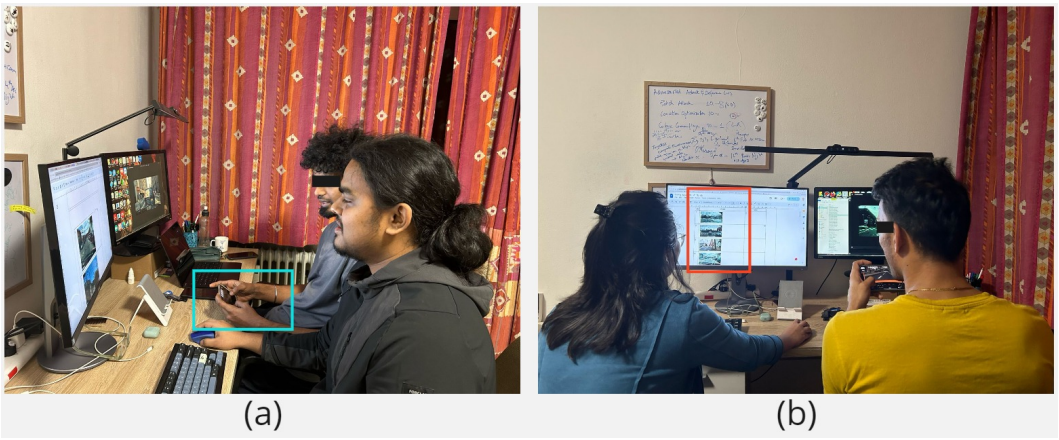

Figure 3: Demonstration of human evaluation where the blue rectangle in (a) shows the primary task of operating a driving simulation and the red rectangle in (b) shows the secondary task of detecting abnormal images

of attack has exhibited a significant improvement, surpassing the existing state-of-the-art[23] with approximately a 6% enhancement in average precision and a 5% rise in the fooling rate.

Human Evaluation Results: During the human evaluation, participants were explicitly instructed to identify any abnormalities present in the images. It's crucial to note that, in our experimental setting, individuals anticipated the existence of patches or anomalies in the images shown to them. However, in real-life situations, such as a person driving a car, the expectation of encountering abnormalities in camera images is generally lower. Consequently, detecting a patch in this real-world scenario may prove to be more challenging.

By employing the previously detailed methodology, we calculated an inconspicuousness score of 0.68, where a higher score denotes better performance (with 1 being optimal and 0 indicating the least favorable outcome). Even with our participants being aware of possible anomalies, our template demonstrated effective concealment of the patch.

6 Conclusion

Our research focused on creating a universally applicable, physically realizable attack targeting state-of-the-art object detectors. In our experiments, we successfully prevented the detection of a specific class while maintaining relatively intact predictions for other classes. Additionally, we demonstrated that our attack achieved a higher success rate compared to the work conducted by [23]. This study underscores the vulnerability of object detection models, such as YOLOv5, to adversarial attacks and emphasizes the importance of deploying them with utmost caution.

While our attack demonstrated a high success rate, there are several avenues for improvement. Firstly, our focus was primarily on a specific template, namely the droplet template. In real-world scenarios, the applicability of the same template may vary. Ideally, one would aim to keep a template as context-homogeneous as possible. Secondly, our attack was conducted digitally. To fully understand its impact, it would be essential to reproduce it as a physical sticker and test its realizability in the real world.

7 Reflections

During the course of the project, we encountered unexpected challenges, particularly in refining the template's context homogeneity. Balancing the relevance of a single template across various scenarios proved more intricate than initially foreseen. On the positive side, our digital implementation showcased promising results. Reflecting on the project, we would consider diversifying the templates used and exploring the attack's real-world feasibility by reproducing it as a physical sticker. This shift from digital to physical testing could reveal a richer understanding of the attack's real-world implications. Looking ahead, the project could benefit from further investigations into context-specific templates, potentially enhancing the attack's effectiveness across diverse scenarios. If time permitted, we would have delved deeper into these aspects to refine the attack's robustness and broaden its applicability. This reflective process emphasizes the importance of adapting methodologies to address unforeseen challenges and underscores the potential for advancements in physical attacks on object detectors. Our attack demonstrates the versatility of attack strategies against detector models and underscores the vulnerability of state-of-the-art object detector algorithms. This highlights the urgent need for improved defense mechanisms capable of addressing a wider variety of attack strategies.

References

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

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4974–4983, 2019.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[4] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 52–68. Springer, 2019.

[5] Francesco Croce and Matthias Hein. Sparse and imperceivable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4724–4732, 2019.

[6] 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*, pages 1625–1634, 2018.

[7] Jamie Hayes. On visible adversarial perturbations & digital watermarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1597–1604, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[9] Hao Huang, Yongtao Wang, Zhaoyu Chen, Zhi Tang, Wenqiang Zhang, and Kai-Kuang Ma. Rpattack: Refined patch attack on general object detectors. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[10] Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 720–729, 2020.

[11] Glenn Jocher, Alex Stoken, Jirka Borovec, Liu Changyu, Adam Hogan, Laurentiu Diaconu, Jake Poznanski, Lijun Yu, Prashant Rai, Russ Ferriday, et al. ultralytics/yolov5: v3. 0. *Zenodo*, 2020.

[12] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. In *International Conference on Machine Learning*, pages 2507–2515. PMLR, 2018.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[14] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1028–1035, 2019.

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

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[18] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. Darts: Deceiving autonomous cars with toxic signs. *arXiv preprint arXiv:1802.06430*, 2018.

[19] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[20] 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. Springer, 2020.

[21] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 665–681. Springer, 2020.

[22] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[23] Alon Zolfi, Moshe Kravchik, Yuval Elovici, and Asaf Shabtai. The translucent patch: A physical and universal attack on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15232–15241, 2021.