Breaking the Illusion: Real-world Challenges for Adversarial Patches in Object Detection

Jakob Schack*
Graz University of Technology
Graz, Austria
jakob.schack@student.tugraz.at

Katarina Petrovic*
Graz University of Technology
Graz, Austria
katarina.petrovic@tugraz.at

Olga Saukh
Graz University of Technology
Complexity Science Hub
saukh@tugraz.at

ABSTRACT

Adversarial attacks pose a significant threat to the robustness and reliability of machine learning systems, particularly in computer vision applications. This study investigates the performance of adversarial patches for the YOLO object detection network in the physical world. Two attacks were tested: a patch designed to be placed anywhere within the scene – global patch, and another patch intended to partially overlap with specific object targeted for removal from detection – local patch. Various factors such as patch size, position, rotation, brightness, and hue were analyzed to understand their impact on the effectiveness of the adversarial patches. The results reveal a notable dependency on these parameters, highlighting the challenges in maintaining attack efficacy in real-world conditions. Learning to align digitally applied transformation parameters with those measured in the real world still results in up to a 64% discrepancy in patch performance. These findings underscore the importance of understanding environmental influences on adversarial attacks, which can inform the development of more robust defenses for practical machine learning applications.

CCS CONCEPTS

• **Computer systems organization** → Embedded software; • **Security and privacy** → **Systems security**; • **Computing methodologies** → **Object detection**; *Machine learning algorithms.*

KEYWORDS

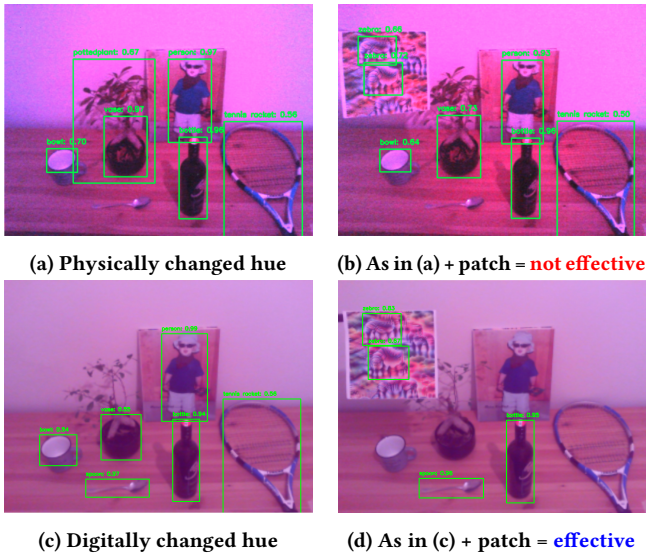
Adversarial attacks, adversarial patches, object detection, YOLO, adversarial robustness

1 INTRODUCTION

The rapid advancement of machine learning algorithms, particularly in computer vision domain, has revolutionized various applications, from autonomous driving to medical imaging and secure face recognition. However, the widespread adoption of these algorithms has brought to light significant security concerns, notably the susceptibility to adversarial attacks [8, 17]. These attacks involve deliberate modifications to input data to mislead machine learning models into producing incorrect outputs.

A prominent example of a computer vision system is the You Only Look Once (YOLO) [19], which is, like many other machine learning models, vulnerable to adversarial attacks [2].

In addition to adversarial attacks, the robustness of machine learning models, including YOLO, is also challenged by environmental conditions such as weather, lighting, camera location, and viewpoint changes [3]. Variations in lighting conditions can lead to overexposure or underexposure, which in turn can affect the model's ability to correctly detect and classify objects. Similarly, changes in the camera's location and viewpoint can introduce new perspectives and angles that the model may not have been trained on, further reducing its detection accuracy. Environmental conditions also affect the performance of adversarial patches [14].



(a) Physically changed hue (b) As in (a) + patch = not effective

(c) Digitally changed hue (d) As in (c) + patch = effective

Figure 1: Discrepancy between the hue transformation applied in the real world (top) and digitally (bottom).

This study focuses on the stability of adversarial patches in the physical world, aiming to understand how various environmental conditions and patch attributes affect their performance. Two types of adversarial patches were evaluated: a global patch designed to suppress all correct detections when placed anywhere in the scene, and a local patch targeting specific objects by partially overlapping them. The patches were tested on the same static scene with YOLOv3 and YOLOv5 as the detection networks, respectively. These versions were chosen due to the availability of pre-existing frameworks for generating adversarial patches tailored to them ([35] and [26], respectively). While other lightweight YOLO versions, such as YOLOv3-tiny [24], are better suited for resource-constrained edge devices, they were not chosen for this research due to the significant modifications required to adapt the adversarial patch generation frameworks. Although the optimizations in these lightweight YOLO versions often result in a noticeable reduction in accuracy, particularly in complex tasks or when detecting small or overlapping objects, this would not be a critical factor in our study, as our scene

*Both authors contributed equally to this research.

setup was relatively straightforward. Our primary objective was to focus on the effectiveness and robustness of adversarial patches, as well as the broader influence of environmental conditions on their performance. Our experimental setup included a controlled indoor environment with a standardized set of objects and lighting conditions. Performance was evaluated based on the mean average precision (mAP) for the global patch and detection confidence for the local patch. Key variables such as patch size, position, rotation, brightness, hue, blurriness and reduced color palette were systematically altered to assess their impact on patch efficacy. The experiments revealed significant dependencies between the performance of adversarial patches and these variables when applied in the digital and in the real worlds. While patch performance with respect to geometric transformation is consistent across both worlds, color transformations unveil substantial differences, which can't be easily matched, and indicating a gap between these both worlds. An example in Figure 1 shows the same scene, where the hue parameter was altered using a RGB light source (top) and digitally using the best parameters to match a physical change (bottom). YOLOv3 performance differs significantly when hue is changed physically and when hue is changed digitally. These findings highlight the sensitivity of adversarial patches to real-world conditions.

Contributions. Our study provides a comprehensive and systematic analysis of how adversarial patches perform under various physical-world conditions, including lighting, patch sizes, and viewpoints. The findings underscore a significant impact of environmental conditions, such as lighting, on the effectiveness of adversarial patches. We show that the real-world effects differ from applying these transformations digitally using the best matching parameters. This study highlights the importance of addressing adversarial vulnerabilities as a critical aspect of MLOps, focusing on improving the robustness of machine learning models against real-world environmental variability, which is essential for ensuring reliable deployment in diverse operational contexts. Furthermore, it emphasizes the need for advanced adversarial methods and improved defenses, contributing to more resilient machine learning systems capable of withstanding real-world adversarial conditions.

2 RELATED WORK

Object detection models and their robustness. YOLO [19] is one of the most popular real-time object detection algorithms. Its high speed and good accuracy make it widely used in the community. The original YOLOv1 object detector was first released in 2016 by J. Redmon *et al.* [18], and quickly became state-of-the-art. Over time, the algorithm was significantly improved, so different versions are now available [5, 7, 9, 13, 36]. In order to achieve the optimal performance of the object detection model, each YOLO version was trained with geometric (perspective change, scaling, translation, flipping, and rotation), color (HSV), and more advanced [12, 15, 28, 33] augmentations. This made YOLO models resilient to challenging environments. Lightweight YOLO versions are optimized for fast execution and reduced power availability in resource-constrained IoT devices. This is achieved by utilizing significantly fewer layers and neurons, resulting in a decrease in accuracy. However, they still face the same challenges as full YOLO versions when adapting to varying environmental conditions.

Adversarial attacks and adversarial patches. Adversarial patches were first introduced in 2018 by Brown *et al.* [29], demonstrating the ability to mislead image classifiers using round stickers [32]. This concept was extended to object detection in 2019 by Liu *et al.* [31] with the Dpatch, although it was initially tested only in digital settings. Subsequent studies, such as the work by Lee and Kolter [35], adapted these patches for physical-world scenarios, highlighting challenges like maintaining attack efficacy across different environmental conditions. This type of attack is referred to as a *global attack* in this study, as it targets the entire image. The paper proposes a way to generate a global patch attack, by maximizing the YOLO loss function. This global adversarial patch is primarily used for the experiments in this study. A different approach for attacking object detectors is proposed in [26]. In this method, the patch must overlap the object that is intended to be hidden from the object detection network. This technique is referred to as *local attack* because it targets individual objects within the scene. The current research builds on these works by systematically evaluating the performance of both global and local adversarial patches.

Defenses against adversarial patches. Adversarial attacks pose a substantial threat to object detection algorithms. Their vulnerability to real-world conditions and environmental changes increases when deployed on resource-constrained IoT devices [39], which often lack the computational power and memory for advanced adversarial defenses. Due to the disruptive potential of adversarial attacks, they have attracted considerable attention, with numerous researchers striving to devise innovative defense strategies [2, 37]. A common approach involves localizing and neutralizing or removing adversarial patches. Jing *et al.* [22] propose PAD - a patch-agnostic defense mechanism that combines semantic independence localization and spatial heterogeneity localization; Xu *et al.* [21] developed defense pipeline against white-box adversarial patches that zeros out the patch region by repainting with mean pixel values; Naseer *et al.* [23] proposed local gradient smoothing scheme that regulates gradients in the estimated noisy region of the image before inference; Scheurer *et al.* [11] address defence against adversarial attacks in motion detection applications. In contrast to previous works, our carefully constructed experiments demonstrate that failure cases for existing adversarial patches can be deterministically constructed. These findings highlight the necessity for further research on more robust adversarial patches and stronger defense mechanisms.

3 METHODOLOGY

Adversarial patch generation. The attack patches used in this work were generated in [26, 35] using variants of local and global projected gradient descent running the following optimization:

$$\arg\max_{\delta} \mathbb{E}_{(x,y)\sim\mathcal{D},t\sim\mathcal{T}}[J(h_\theta(A(\delta,x,t)),y)], \tag{1}$$

where $\mathcal{D}$ is the distribution over samples, and A is the patch application function. The function A applies a transformation δ with parameters $t \in \mathcal{T}$ to the patch during training to ensure robust patch performance (for example, the global patch was trained with rotation augmentation). The patch is then integrated into the image x at a desired location. The optimization is solved essentially by using gradient descent [26, 31, 35].

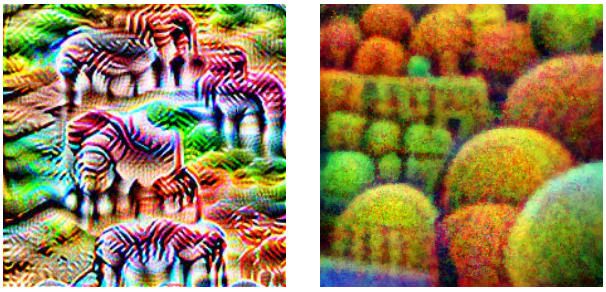


Figure 2: Global (left) and local (right) adversarial patches.

Global and local patches. This study evaluates two types of patches: a global patch and a local patch. The global patch, designed to attract the attention of the object detection network and suppress correct detections, can be placed anywhere in the image and was generated as described in [35] using YOLOv3 [36]. The local patch, which must overlap the target object, was generated according to [26] using YOLOv5 [13]. Both patch generation processes used the COCO2014 dataset [30]. The generated patches (Figure 2) are specific to their respective detection networks and are not transferable. All physical patches were printed on regular paper using a standard printer.

Hypothesis. Following recent findings that adversarial patches may fail in the physical world [14], we conducted a dedicated set of experiments to better understand these failure cases. To achieve this, we (1) carefully constructed our experiments, and (2) investigated the differences between the effects observed in the physical world and their reproducibility through digital transformations. We used sensors, and two cameras operating in well-documented modes to run reproducible real-world experiments. Our main hypothesis is that failure cases of adversarial patches in the physical world in general differ significantly from similar experiments conducted digitally, *i.e.*, by embedding the patch into an image and transforming the result using the same parameters as measured physically, or the best matching parameters computed by an optimization algorithm. We present this analysis next.

4 DISCOVERING VULNERABILITIES OF ADVERSARIAL PATCHES

4.1 Experimental setup

Controlled real-world environment. We evaluated the performance of adversarial patches by conducting physical attacks in a controlled indoor setting and reproducing them digitally for comparison. This controlled environment allowed us to easily adjust testing conditions. Lighting was controlled using an IKEA® Tradfri LED1924G9 RGB light source. We primarily used the Microsoft LifeCam HD-3000 camera, which records 720p HD videos at up to 30 fps. To ensure results were not camera-specific, we repeated experiments involving brightness and hue with the Ausdom AF640 camera, which records 1080p HD videos at up to 30 fps. Our scene setup is shown in Figure 1. The test included a bottle, cup, potted plant in a vase, tennis racket, spoon, and a picture of a person. Occasionally, a dining table was detected with low confidence but excluded from consideration due to inconsistency.

Evaluation metrics. To evaluate the performance of the global patch, we primarily use the mean average precision (mAP) as the metric. mAP is calculated by generating precision-recall curves and determining the area under these curves, providing insights into the overall performance of an object detection system. In this study, a lower mAP indicates better patch performance in suppressing detections. For the local patch, we measure performance by the detection confidence of the targeted object. This confidence describes the probability of a detected object belonging to a particular class. Lower detection confidence signifies higher patch effectiveness.

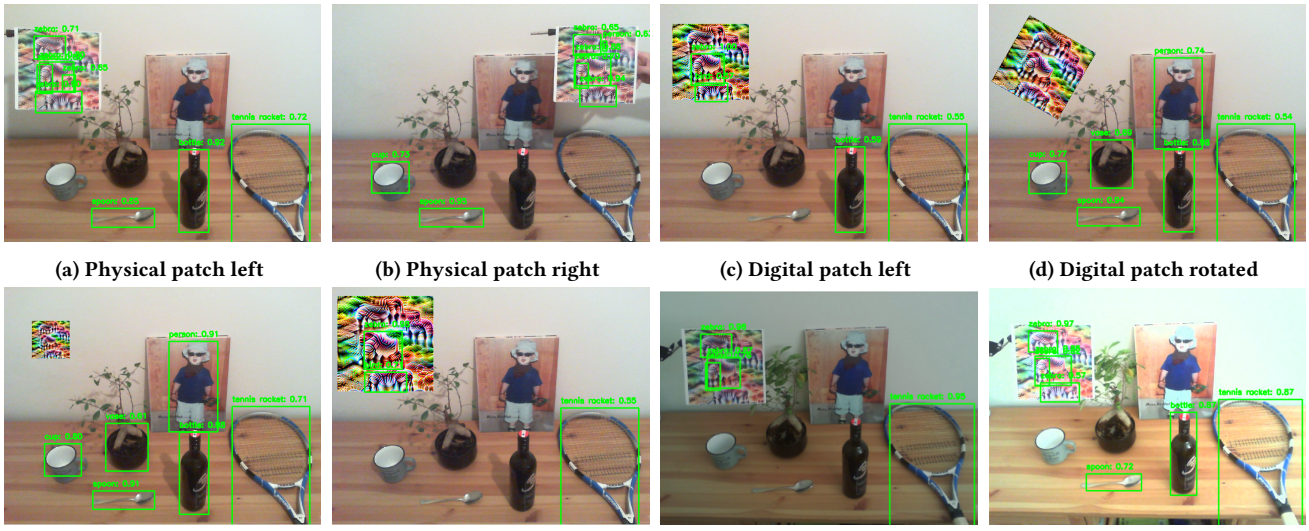
4.2 Experimental variables

Following setups described in the literature [4, 20, 25, 27], we varied key parameters that can also easily change in uncontrolled real-world settings: (1) geometry (patch size, observation angle, distance to the target), (2) color transformations (scene brightness and hue), and (3) information reduction (blurriness and limiting the number of colors in an image).

Geometric transformations. We first experimented with different patch sizes. For global attacks, patch sizes ranged from 10% to 30% of the image width to avoid object overlap. For local attacks, we tested patch sizes varied from 4cm x 4cm to 16cm x 16cm. For all other experiments (geometric, color or information reduction), by default, we used a 25% image width-sized patch for global attacks and the smallest patch that significantly reduced object detection confidence for local attacks (11cm x 11cm for the tennis racket, 7cm x 7cm for the other objects). We explore patch rotations up to 90° around X, Y, and Z axes (see Figure 7). We also explore the impact of the global patch position within the scene on its detection suppression ability depending on the distance from the target.

Color transformations. Ambient brightness was varied from 4 to 61 lux (measured with a light sensor). With automatic camera exposure, there is a trade-off between the image brightness and noise. To mitigate this, we fixed the exposure time, thereby enabling overexposure – a common problem in real-world applications like autonomous driving [16]. The camera exposure was calibrated to produce a uniform, naturally looking image at a measured brightness of 15 nits. However, a discrepancy persists between the measured lux in the room and the illuminance calculated from the scene image. To address this issue and facilitate comparisons with our digital experiments, we performed an illuminance scale correction based on the calculated image illuminace. This adjustment shifted the real-world range of 4 to 61 lux to an approximate range of 68 to 243 lux as measured in the images. We varied the hue values of the scene with an IKEA Tradfri LED1924G9 RGB light source (see Figure 1 for an example). Note that this light source also influences the other colour properties of the environment (*e.g.*, brightness).

Information reduction. We conducted a series of information reduction experiments in the digital domain to analyze the performance of a low-quality patch due to possible camera or printing effects. A low-pass filter is often used in digital image processing domain to smooth the image, soften the sharp regions, and remove the noise while preserving important image features. We varied the filter size from 0 to 500. Color reduction filtering aims to enhance image compression, optimize storage efficiency, and decrease computational complexity in image analysis tasks by reducing the



(a) Physical patch left (b) Physical patch right (c) Digital patch left (d) Digital patch rotated

(e) Digital patch 12% of the image (f) Digital patch 30% of the image (g) Brightness 110 lux (h) Brightness 196 lux

Figure 3: Global patch performance under different conditions. (a)-(b): relocating physical patch; (a)-(c): physical vs digital patch; (c)-(d): patch rotation; (e)-(f): patch size; (g)-(h): environmental brightness.

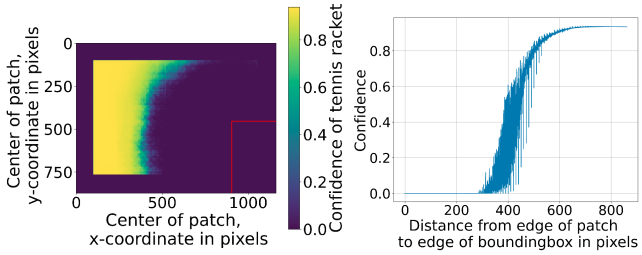


Figure 4: Confidence over patch position (left), and patch edge to bounding box distance (right) for "tennis racket".

number of distinct colors in an image while maintaining its essential visual features. The number of reduced colors was varied from 2 to 600, whereas a natural-looking image of a scene contains more than 50'000 different colors.

4.3 Exploration based on a fixed indoor scene

To get a better understanding of the stability of the attacks in the real world, we run comprehensives experiments and attempt to reproduce the results in the digital world. The observations below relate to the global patch, while the discussion about local patch is out of the scope of this paper.

Distance dependence. The first stability issue is the patch's effectiveness depending on its distance from the object. Experiments show an attack is successful only if the patch is within a reasonable distance from the target. Comparing Figure 3(a) and Figure 3(b) confirms the patch is more effective when closer to the object. The original paper [35] claims that while the patch is somewhat location-invariant, its influence weakens with distance. We further investigated the effect by digitally inserting the global patch at various positions. Figure 4(left) displays tennis racket detection confidence, with the x and y positions indicating patch placement and confidence levels. The red rectangle is the ground truth bounding box. Results show the patch must be within a certain radius, dependent on the size of the patch, resolution of the scene, and the object itself, to suppress detection effectively. Figure 4(right) illustrates detection confidence relative to the distance from the patch's edge to the bounding box edge, showing the patch loses its adversarial properties when positioned too far away from the object (~400px).

Rotation dependence. Due to the nature of physical experiments, the position of the patch relative to the camera is crucial in an adversarial attack. Figure 3(d) illustrates a large rotation around the z-axis, resulting in a significant reduction in adversarial performance compared to Figure 3(c). Rotation is often included in training neural networks for computer vision as a data augmentation strategy [1]. Rotation transformation was also applied in the global patch generation software [26, 34]. Consequently, we expect the patch to exhibit some robustness to rotations. Figure 5 shows the mAP over rotation angles across the three axes in the real world (left) and digitally (right). The patch sizes were aligned across these two settings. In both cases, the patch shows robustness to rotations around x and y axes within $\pm 40°$, yet loses its adversarial properties for rotations around z axis larger than 20°. Adversarial effects of the patch in the digital domain is stronger.

Size dependence. The patch size positively correlates with its effectiveness. Figure 3(e) and Figure 3(f) show a significant difference in performance when scaling a global patch from 12% to 30% of the image. This trend is confirmed digitally, as Figure 6(left) shows larger patches suppress more detections than smaller ones, with the effect being stronger in the digital domain.

Brightness dependence. Another issue with attack stability becomes apparent when lighting conditions change. If the camera's

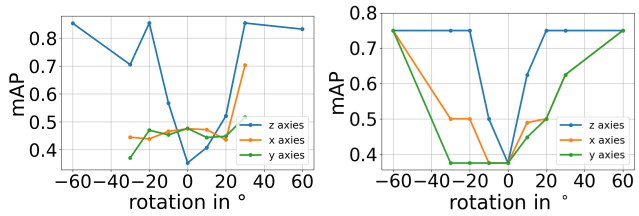


Figure 5: Mean average precision for patch rotations (size aligned): physical (left) and digital (right) experiment.

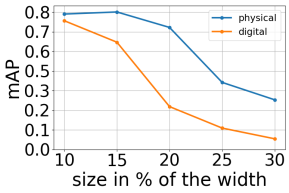


Figure 6: Mean average precision over patch size.

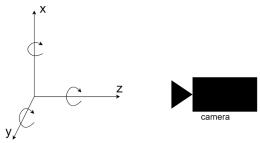


Figure 7: Axes placed relative to the camera.

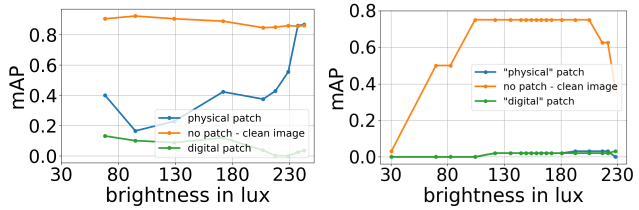


Figure 8: Mean average precision over scene brightness: physical (left) and digital (right) experiment.

exposure time is set to automatic, there is a trade-off between image brightness and noise. Setting this to manual removes this dynamic adaptation and enables us to emulate overexposure. The exposure time is fixed and calibrated to produce a uniform and naturally looking image for the baseline of the experiment at 15 nits. Figure 3(g) shows an example of patch performance at reduced brightness, while Figure 3(h) shows an example at increased brightness. Here, the adversarial patch loses effectiveness if the image becomes too bright and starts clipping. Conversely, decreasing brightness does not significantly impact the patch's performance. Figure 8 clarifies how patch performance is affected by brightness. The digital patch shows consistent performance unaffected by lighting changes over the whole range of values. The physical patch, however, looses its effectiveness with higher brightness when clipping occurs, matching the mAP of a clean image without a patch.

Hue dependence. Here we investigate patch performance when changing hue lightning. Figure 1 provides examples of images with physically and digitally altered hue and the corresponding detection results. YOLO achieves a mAP of 0.4 in real-world (Figure 1(b)), and a mAP of 0.14 in the digital world (Figure 1(d)), resulting in approximately 64% discrepancy in patch performance. In the real world, we used a RGB light source to change hue. To get the hue value of the light source, the value reported in the companion app

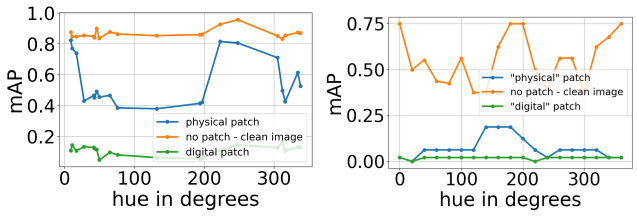


Figure 9: Mean average precision over scene hue: physical (left) and digital (right) experiment.

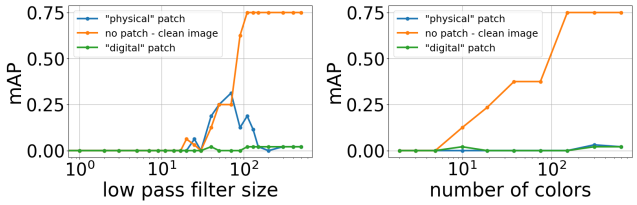


Figure 10: Mean average precision over low pass filter size (left) and number of colors in the image (right).

(IKEA® Home smart 1) to the LED was used. To digitally replicate physical world scene images as accurately as possible, *i.e.*, to digitally generate the scene images that are the best match to the physical scene images, we train a small neural network. Our neural network outputs suggest that color transformations alone do not fully capture the changes introduced by an additional hue light source. To support this, we present the performance of the patch across the complete hue range in Figure 9. Figure 9 (left) shows the results from physical experiments, where the hue value is reported by the additional light source app. Figure 9 (right) displays results from digital experiments, where only the hue value of the image is digitally altered. Notably, YOLO performs consistently across all hue values in the physical world, whereas digital experiments exhibit some fluctuations in YOLO performance. Additionally, in the physical world, there is a clear range in the hue spectrum, between 200 and 300 degrees, where the patch fails to perform effectively. In contrast, digital experiments show only slight disturbances in patch performance between 150 and 200 degrees.

Information reduction. This set of experiments shows that the patch requires a certain amount of information to be effective. In the low-pass filter experiments in Figure 10(left), the detection efficiency for the unpatched scene and the scene with the simulated physical patch (labeled as "physical" patch) align closely up to a point – *i.e.*, the patch has no impact. Beyond this point, the presence of the patch decreases detection efficiency. In the color reduction experiments in Figure 10(right), the results for the simulated physical patch align almost perfectly with the digital patch.The detection algorithm also requires a certain amount of color details to achieve its full detection potential. While performing these experiments, we observed that the colors on the patch are quite diverse and mixed. Consequently, when a color reduction filter is applied, the patch retains a relatively large number of colors compared to the real-world environment. This allows the patch to maintain the maximum of it's efficiency and remain as efficient as the original digital patch.

5 DISCUSSION, LIMITATIONS, AND OUTLOOK

Discussion. Our experiments reveal significant dependency of patch effectiveness on environmental variables, such as patch size, position, rotation, brightness, and hue, highlighting the challenges in maintaining attack efficacy in real-world scenarios. While some failure cases are intuitive (*e.g.*, a positive correlation between the patch size and its effectiveness), many of our observations are not. Despite the efforts to match the transformation parameters between the physical and digital domain, performance discrepancies remain, leaving many questions open.

Lighting conditions, particularly brightness and hue, were found to be critical for patch performance. Changes in brightness, whether from natural or artificial lighting, can alter the patch's appearance and its interaction with the detection model. Overexposure or underexposure can reduce patch's effectiveness. Variations in hue from different light sources also impact the patch's ability to disrupt detection. In the physical world, light interacts with materials in complex ways, influencing the object appearance. Geometrical and physical optics, including reflection, scattering, interference, and absorption affect color perception [38]. The color seen, *i.e.*, light that is leaving the object, is primarily determined by the energy of the light wave and material's properties - the energy of the incident wave, surface roughness and texture. Modeling these interactions digitally presents significant challenges, as highlighted by Musbach *et al.* [6].

The impact of training data on patch performance is a critical consideration. The COCO dataset's 80 classes are highly imbalanced. *Person* is the most frequent class, occurring over 250'000 times, while *bottle* and *cup* appear approximately 25'000 times each, and *tennis racket* only about 5'000 times. Our experiments showed that concealing the tennis racket is challenging, often requiring larger and closer patches, likely due to this imbalance. Furthermore, patch behaviour under varying hues may be attributed to the absence of red and violet images in the COCO dataset - most images are yellow and green hue.

Limitations. Our study has several limitations. Firstly, the experiments were conducted in a controlled indoor environment, which does not capture the variability of a real outdoor settings. Factors like weather, outdoor lighting, and movement dynamics were not considered. The physical patches were printed on standard paper and evaluated under specific conditions, without exploring variations in patch materials and printing quality. Secondly, the study focused on a limited set of objects and scenes. While the selected objects provided a good baseline, a more diverse set of objects and scenes could reveal more insights. The experiments were limited to two types of adversarial patches and specific versions of the YOLO network. Exploring other types of patches and different, especially lightweight object detection models that could run on edge devices (*e.g.*, smart cameras), could further generalize the findings.

Future Work. The findings call for a deeper understanding of the interplay between adversarial strategies and environmental factors. Future research should focus on developing more sophisticated adversarial methods that can adapt to changing conditions, and on improving the robustness of detection models to withstand such attacks, thereby enhancing the security and reliability of machine learning applications in real-world settings. While adversarial patches are static, object detection models can take advantage of on-device adaptation and reconfiguration to improve their resilience to adversarial attacks, *e.g.*, [10, 37].

REFERENCES

[1] N. Cao and O. Saukh. 2023. Geometric Data Augmentations to Mitigate Distribution Shifts in Pollen Classification from Microscopic Images. arXiv:2311.11029
[2] J. I. Choi and Q. Tian. 2022. Adversarial Attack and Defense of YOLO Detectors in Autonomous Driving Scenarios. arXiv:2202.04781
[3] Y. Ding and X. Luo. 2024. SDNIA-YOLO: A Robust Object Detection Model for Extreme Weather Conditions. arXiv:2406.12395
[4] A. Braunegg *et al.*. 2020. APRICOT: A Dataset of Physical Adversarial Attacks on Object Detection. arXiv:1912.08166
[5] A. Bochkovskiy *et al.*. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934
[6] A. Musbach *et al.*. 2013. Full Wave Modelling of Light Propagation and Reflection. https://doi.org/10.1111/cgf.12012
[7] C. Li *et al.*. 2023. YOLOv6 v3.0: A Full-Scale Reloading. arXiv:2301.05586
[8] C. Szegedy *et al.*. 2014. Intriguing properties of NN. arXiv:1312.6199
[9] C. Wang *et al.*. 2022. YOLOv7: Trainable bag-of-freebies sets new SoTA for real-time object detectors. arXiv:2207.02696
[10] D. Wang *et al.*. 2024. Subspace-Configurable Networks. arXiv:2305.13536
[11] E. Scheurer *et al.*. 2023. Detection Defenses: An Empty Promise against Adversarial Patch Attacks on Optical Flow. arXiv:2310.17403
[12] G. Ghiasi *et al.*. 2021. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. arXiv:2012.07177
[13] G. Jocher *et al.*. 2023. *Ultralytics*. https://github.com/ultralytics/ultralytics
[14] G. S. Hartnett *et al.*. 2022. Empirical Evaluation of Physical Adversarial Patch Attacks Against Overhead Object Detection Models. arXiv:2206.12725
[15] H. Zhang *et al.*. 2018. mixup: Beyond Empirical Risk Min. arXiv:1710.09412
[16] I. Jatzkowski *et al.*. 2018. A Deep-Learning Approach for the Det. of Overexposure in Automotive Camera Images. https://doi.org/10.1109/ITSC.2018.8569692
[17] J. C. Costa *et al.*. 2024. How Deep Learning Sees the World: A Survey on Adversarial Attacks and Defenses. arXiv:2305.10862
[18] J. Redmon *et al.*. 2016. You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640
[19] J. Terven *et al.*. 2023. A Comprehensive Review of YOLO Architectures in CV. arXiv:2304.005016
[20] K. Eykholt *et al.*. 2018. Robust Physical-World Attacks on Deep Learning Models. arXiv:1707.08945
[21] K. Xu *et al.*. 2022. PatchZero: Defending against Adversarial Patch Attacks by Detecting and Zeroing the Patch. arXiv:2207.01795
[22] L. Jing *et al.*. 2024. PAD: Patch-Agnostic Defense against Adversarial Patch Attacks. arXiv:2404.16452
[23] M. Naseer *et al.*. 2018. Local Gradients Smoothing: Defense against localized adversarial attacks. arXiv:1807.01216
[24] P. Adarsh *et al.*. 2020. YOLO v3-Tiny: Object Detection and Recognition using one stage improved model. https://doi.org/10.1109/ICACCS48705.2020.9074315
[25] S. Chen *et al.*. [n. d.]. *ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector*. arXiv:1804.05810
[26] S. Shrestha *et al.*. 2023. Towards a Robust Adv. Patch Attack Against Unmanned Aerial Vehicles Object Det. https://doi.org/10.1109/IROS55552.2023.10342460
[27] S. Thys *et al.*. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. arXiv:1904.08653
[28] S. Yun *et al.*. 2019. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. arXiv:1905.04899
[29] T. B. Brown *et al.*. 2017. Adversarial Patch. arXiv:1712.09665
[30] T. Lin *et al.*. 2014. Microsoft COCO. arXiv:1405.0312
[31] X. Liu *et al.*. 2018. DPatch: Attacking Object Detectors with Adversarial Patches. arXiv:1806.02299
[32] X. Wei *et al.*. 2022. Adversarial Sticker: A Stealthy Attack Method in the Physical World. arXiv:2104.06728
[33] Z. Wei *et al.*. 2020. AMRNet: Chips Augmentation in Aerial Images Object Detection. arXiv:2009.07168
[34] Z. Zhang *et al.*. 2023. A novel method for Pu-erh tea face traceability identification based on improved MobileNetV3. https://doi.org/10.1038/s41598-023-34190-z
[35] M. Lee and Z. Kolter. 2019. On Physical Adversarial Patches for Object Detection. arXiv:1906.11897
[36] J. Redmon and A. Farhadi. 2018. YOLOv3: An Incremental Improvement. arXiv:1804.02767
[37] O. Saukh. 2023. Escaping Adversarial Attacks with Egyptian Mirrors. https://doi.org/10.1145/3615593.3615724
[38] R. Tilley. 2020. *Colour and the Optical Properties of Materials*. John Wiley & Sons. https://doi.org/10.1002/9780470974773
[39] C. Westbrook and S. Pasricha. 2023. Adversarial Attacks on Machine Learning in Embedded and IoT Platforms. arXiv:2303.02214