Street-Legal Physical-World Adversarial Rim for License Plates

Nikhil Kalidasu[1][0009-0001-2651-8292] and Sahana Ganapathy[1][0009-0007-1823-1324]

The University of Texas at Austin, Austin TX 78705, USA

Abstract. Automatic license plate reader (ALPR) systems are widely deployed to identify and track vehicles. While prior work has demonstrated vulnerabilities in ALPR systems, far less attention has been paid to their legality and physical-world practicality. We investigate whether low-resourced threat actors can engineer a successful adversarial attack against a modern open-source ALPR system. We introduce the Street-legal Physical Adversarial Rim (SPAR), a physically realizable white-box attack against the popular ALPR system `fast-alpr`. SPAR requires no access to ALPR infrastructure during attack deployment and does not alter or obscure the attacker's license plate. Based on prior legislation and case law, we argue that SPAR is street-legal in the state of Texas. Under optimal conditions, SPAR reduces ALPR accuracy by 60% and achieves an 18% targeted impersonation rate. SPAR can be produced for under $100, and it was implemented entirely by commercial agentic coding assistants. These results highlight practical vulnerabilities in a modern ALPR system under realistic physical-world conditions and suggest new directions for both attack and defense.

1 Introduction

Automatic license plate reader (ALPR) technology is extensively deployed by governments, communities, and businesses to identify and track vehicles [15, 8, 31]. Neural networks have long been used for license plate recognition [18, 22, 24]. Adversarial attacks against vision models have a similarly long history [32, 10, 16], including attacks on surveillance systems [25].

To investigate whether adversarial attacks pose a realistic threat to modern ALPR systems, we conducted a demonstration attack against a single modern open-source ALPR system. We sought to both prevent the correct reading of the attacker's license plate (a non-targeted "disruption" attack) and induce the defender's ALPR system to output an attacker-defined target (a targeted "impersonation" attack). We imposed the following constraints: **Realistic Access:** The attacker has no access to the defender's ALPR software or hardware during attack deployment. **Street-legality:** The attack must be street-legal in the state of Texas. **Simplicity:** The attack must be plausible for a time- and resource-constrained attacker with limited technical knowledge. **Practicality:** The attack must be practical in the physical world.

We modeled a scenario where the threat actor has acquired model weights and source code of the defender's ALPR system. This white-box assumption follows standard security practice by assuming that attackers may have detailed knowledge of a system's design [14, 7]. We designed a highly specific attack for a single vehicle, license plate, and ALPR system. We limited our total costs to $100 (excluding evaluation) and restricted the size of our training dataset to what could plausibly be collected and labeled by a individual attacker. All code was generated with commercial agentic coding tools to investigate whether large language models (LLMs) can indeed uplift the capabilities of low-resourced attackers [4]. We prioritize the street-legality constraint because if an attack does not clearly violate existing laws, it may be difficult for law enforcement to mitigate the threat. A practical street-legal adversarial attack on ALPR systems, in the absence of new legislation, could necessitate the adoption of adversarial robustness measures into deployed license plate readers as an imperfect technological solution. We narrow our scope to the single jurisdiction of Texas to make this problem tractable, as regulations vary widely across jurisdictions. To evaluate physical-world practicality, we test our attack under varying distance, perspective, and lighting conditions, which have all been identified in prior literature as challenges to adversarial attacks on vision models [29] and on ALPR systems specifically [43]. We use statistical methods to quantify our approach's dependence on camera distance and viewing angle.

1.1 Background

Adversarial attacks on computer vision models can be broadly classified into perturbation attacks [32], patch-based attacks [5], and spot attacks [26]. Goodfellow et al. [10] showed that perturbation attacks exploit the approximate linearity of neural networks in high-dimensional input spaces. By adding a bounded input-dependent perturbation, an attacker can induce a disproportionately large change in a model's internal activations and output relative to the perturbation magnitude. These methods are effective against modern vision models and have been used against ALPR systems in the past [11, 17, 13].

A major limitation of this strategy is that the attacker must modify an image after capture but before model inference, requiring a level of system access that is often impractical. The adversarial patch method instead overlays a small learnable patch onto input images at random locations and under random transformations [5]. A common approach is to freeze the victim model's parameters and optimize the patch under an expectation-over-transformation (EoT) objective [5] that enforces robustness to transformations in viewpoint and camera equipment. This has enabled physical-world attacks against image classifiers, object detectors, and other vision systems [6, 21, 20, 40, 34] because a printed patch sticker maintains this robustness. However, adhering a patch on top of a license plate may constitute an "attached sticker" interfering with readability, and may be illegal in the state of Texas [36]. An alternative approach is to define the patch as a rim surrounding the license plate rather than a sticker applied on

Table 1. Comparison with prior adversarial attacks on ALPR systems. We strictly define "system-level access" as requiring digital alterations after image capture but before model submission, with no potential for extension to the physical world.

Approach	Attack Type	Stage	Targeted	Physical World	Legal (Texas)	Perspective-Robust	Access Model
Zha et al. [43]	Spot	OCR	✓	✓	×	✓	White-box
Gu et al. [11]	Perturbation	OCR	×	×	×	×	System-level
Qian et al. [26]	Spot	OCR	✓	×	×	×	Black-box
Kwon & Baek [17]	Perturbation	OCR	✓	×	×	×	System-level
Jiamsuchon et al. [13]	Perturbation	OCR	×	×	×	×	System-level
Zhang et al. [44]	Rim	Detection	×	✓	✓	×	White-box
SPAR	**Rim**	**Both**	✓	✓	✓	✓	**White-box**

top of it; this has shown some promise in prior literature [44] and is the approach we selected for SPAR.

The spot attack occupies a middle ground between perturbations and patches. Spot attacks add small, perceptible patterns to an image resembling naturalistic noise like mud splatters. They have been successfully demonstrated against ALPR systems, including illumination-based and sticker-based designs [43, 26]. But in practice, the "stealthiness" of spot attacks remains subjective [6], and strategically obscuring parts of a license plate to inhibit its readability has similar legality concerns to vanilla adversarial patches [36].

1.2 Related Work

Perturbation-, patch-, and spot-based attacks have all been previously executed against ALPR systems, in both white- and black-box settings, against different parts of the ALPR pipeline (*detection* of the license plate in an image versus *reading* of the plate number via optical character recognition, OCR). Prior work explores both targeted attacks which cause the ALPR system to misread a license plate as a specific "impersonation target," and non-targeted attacks which cause the ALPR system to produce non-specific incorrect outputs.

Zha et al. [43] conducted a physical-world spot attack targeting the OCR stage, and identified two key challenges: legal requirements to not modify the license plate area and physical-world robustness to variations in lighting, distance, and camera angle. They therefore applied the spots by illuminating the license plate with an attached electronic device. The complexity of this solution violates our simplicity constraint, and an "illuminated device" interfering with plate readability is explicitly noted as illegal in Texas [36]. Subsequent work largely focused on digital-world attacks against ALPR systems, primarily targeting the OCR stage under assumptions of post-capture access to images [11, 17, 13, 26]. These approaches assume the ability to intercept and modify images after capture, violating our realistic access constraint, and do not attempt physical-world practicality. Zhang et al. [44] conducted a physical-world patch attack on the detection stage with a rim surrounding the license plate. However, the authors did not take the step of Zha et al. of accounting for lighting,

distance, and camera angle, which are known to cause issues for patch attacks specifically [29]. Additionally, their physical-world evaluation was limited to a single ideal-case test image.

There thus remains a literature gap in adversarial attacks on ALPR systems combining realistic access models, street-legality, simplicity for low-resourced attackers, and physical-world practicality. Additionally, all previous attacks focus exclusively on either the license plate detection portion or the OCR portion of the ALPR pipeline; none have attempted to simultaneously attack both. We therefore present the Street-legal Physical Adversarial Rim (SPAR), a printable adversarial license plate rim that successfully disrupts ALPR in both the digital and physical world while satisfying all our constraints.

2 Methods

SPAR is an adversarial rim around the license plate (adopted from Zhang et al. [44]) which targets both the detection and OCR phases. The detection attack minimizes confidence and expands the bounding box to fully encompass the rim; the OCR attack forces `fast-alpr` to produce incorrect or attacker-defined output. We ensure SPAR's street-legality by mounting it as a printed poster between the license plate and the vehicle, strictly behind all visible areas of the license plate. Under Texas Transportation Code § 504.945 [36], a physical border fully outside the plate boundary does not constitute a "coating, covering, or protective substance" under (a)(7), nor an "attached... sticker, decal, emblem, or other insignia" under (a)(6). United States v. Granado held that a license plate frame is not an "insignia" within (a)(6)'s meaning [37]. United States v. Trevino [38] and State of Texas v. Johnson [35] held that a frame can be illegal if it covers any part of the legible text. SPAR does not cover any text and would not fall under this exception. So we contend that SPAR is street-legal in Texas.

We selected the `fast-alpr` package [2] as our target because of its modern development and high accuracy. A YOLO-v9 object detection model from `open-image-models` [3], specifically, the `yolo-v9-t-384-license-plate-end2end` detection model [39], comprises the detection phase. A compact convolutional transformer [12] from `fast-plate-ocr` reads the text on the license plate; we specifically used `cct-xs-v1-global-model` [1].

We relied fully on LLMs to implement SPAR, specifically ChatGPT 5.1 and Claude 4 Sonnet. We provided high-level objectives as prompts (e.g. "Make a data labeling script that enables a human to select the locations of all license plate corners"). Our two methodological innovations were both suggested by LLMs. First, SPAR accounts for perspective transformations by converting the manually labeled corner data into a homography matrix and applying the adversarial patch as a rim around the license plate during training. Second, total variation loss [28] helps regularize the patch, reducing overfitting despite the small size of our training dataset. Both these techniques are well established in prior computer vision literature, though our work is the first time they have been incorporated into an adversarial attack on an ALPR system.

2.1 Data Collection and Preprocessing

We collected a small training dataset to model our low-resource constraint. We drove the test vehicle to several urban and suburban locations during daytime, evening, and nighttime hours, and captured images of the vehicle's license plates in an unstructured fashion, attempting to construct an intuitively diverse dataset. We developed an assisted manual labeling software to label the corners C_{plate} of the license plate in each image, observing that each image could be accurately labeled in under 3 seconds. The full dataset of 300 images was labeled by a single person within several hours. 20% of this data was held out in a validation set, so the final training dataset consisted of only 240 images.

2.2 Patch Application

Suppose we have input image $I \in \mathbb{R}^{C \times H \times W}$, where $C \times H \times W$ are the dimensions. Suppose we also have adversarial patch $P \in \mathbb{R}^{C \times h \times w}$, where $C \times h \times w$ are the patch dimensions. From the manually labeled license plate corners C_{plate}, we compute C_{rim} by scaling the quadrilateral up from its geometric center by a factor of 1.6. We also define C_{patch} as the canonical corners of the patch image. We then define the homography matrix $M_{rim} : C_{patch} \rightarrow C_{rim}$ and iteratively run `inverse_warp` over all pixels corresponding to the input image to generate a transformed adversarial patch on the same perspective as the license plate (the standard `warp_perspective` algorithm).

`warp_perspective`(P, M_{rim}, H, W) returns P' of the same size as the input image, where $P'_{c,y,x}$ is 0 if (x, y) is outside the rectangular bounds defined by C_{patch}, or equal to the roughly corresponding $P_{c,v,u}$ otherwise. We then merge P' with the original image such that the rim is overlaid on the original image, but the original license plate is still visible ("cut out" of the patch). We define a white patch $P_W \in \mathbb{R}^{C \times h \times w}$ and $M_{plate} : C_{patch} \rightarrow C_{plate}$ similarly to M_{rim}. We then use `warp_perspective` to create a mask defining which pixels in the image I are part of the rim, which are part of the license plate, and which are part of neither:

$$\mathcal{M}_{rim} = \texttt{warp\_perspective}(P_W, M_{rim}, H, W) \tag{1}$$

$$\mathcal{M}_{plate} = \texttt{warp\_perspective}(P_W, M_{plate}, H, W) \tag{2}$$

$$\mathcal{M}_{final} = \texttt{clamp}(\mathcal{M}_{rim} - \mathcal{M}_{plate}, 0, 1) \tag{3}$$

With this $\mathcal{M}_{final}$, we can composite an output I' with the rim applied around the license plate in the image:

$$I' = I \odot (1 - \mathcal{M}_{final}) + \rho \cdot P' \odot \mathcal{M}_{final} \tag{4}$$

Where $\odot$ is the Hadamard product and $\rho \in \mathbb{R}$ is a random darkening factor between 0 and 0.2 used to encourage robustness to lighting variations. We used patch dimensions $h = 256, w = 512$, and image dimensions were $H = 384, W = 384$ with both images using $C = 3$.

Loss Function Let $\mathcal{D} : \mathbb{R}^{C\times H\times W} \to (\mathbb{R}^{C\times H\times W}, \mathbb{R})$ be the detection model and let $\mathcal{O} : \mathbb{R}^{C\times H_O\times W_O} \to \mathbb{R}^{L\times V}$ be the OCR model, where H_O and W_O are the input dimensions of the OCR model, L is the string buffer length of the OCR output, and V is the vocabulary size of the OCR output. Suppose $(B, C_{onf}) = \mathcal{D}(I')$ are detection region mask and confidence outputted by $\mathcal{D}$ on the patched image, and A is the smallest rectangular bounding area encompassing all of C_{rim}. The first component of the loss function is the negative Intersection-over-Union between A and B [42], divided by the confidence value. This aims to expand the bounding box output of YOLO-v9 to cover the license plate rim and improve the rim's ability to impact the downstream OCR step. Dividing by the confidence value also prioritizes reducing detection confidence. With $\delta = 10^{-8}$ as a numerical stability constant, detection loss $\mathcal{L}_{det}$ is defined as:

$$\mathcal{L}_{det} = -\frac{B \cap A}{B \cup A} \cdot \frac{1}{C_{onf} + \delta} \tag{5}$$

We resize B into $B' \in \mathbb{R}^{C\times H_O\times W_O}$ to fit the OCR model's input size; $O = \mathcal{O}(B')$ are the output character probabilities of the OCR model. Define $T \in \{0, 1\}^{L\times V}$ as the one-hot encoded impersonation target. $T_{smooth} = T(1-\epsilon) + \frac{\epsilon}{V}$ is a smoothed label tensor with smoothing factor $\epsilon = 0.01$ that helps prevent exploding logits and gradients (this had been added by an LLM without disclosure to the operator). The probability of the true class p_t is:

$$p_t = \sum_{v=1}^{V} T_{smooth,v} \cdot O_v \tag{6}$$

From here, we can calculate the focal categorical cross-entropy loss [19], where $\alpha = 0.25$ is the balancing factor and $\gamma = 2.0$ is the focusing parameter:

$$\mathcal{L}_{focal} = \sum_{l=1}^{L} -\alpha(1 - p_t^{(l)})^{\gamma} \log(p_t^{(l)} + \delta) \tag{7}$$

We rescale this by first calculating $\mathcal{L}_{baseline}$ representing the focal categorical cross-entropy loss on the dataset without any patch applied. The total OCR loss for patch P is:

$$\mathcal{L}_{OCR} = \frac{\mathcal{L}_{focal}}{\mathcal{L}_{baseline}} \tag{8}$$

The final loss component and our work's second core contribution is a total variation (TV) regularization term [28]. This is applied only to the adversarial patch $P \in \mathbb{R}^{C\times h\times w}$, intending to encourage neighboring pixels to take similar values. This both improves print accuracy by minimizing large color changes between pixels, and also encourages the discovery of large, complex structures. We normalize by dividing by the number of comparisons.

$$
\begin{aligned}
TV_h &= \sum_{c=1}^{C} \sum_{i=1}^{h} \sum_{j=1}^{w-1} (P_{c,i,j+1} - P_{c,i,j})^2 \\
TV_v &= \sum_{c=1}^{C} \sum_{i=1}^{h-1} \sum_{j=1}^{w} (P_{c,i+1,j} - P_{c,i,j})^2 \\
\mathcal{L}_{TV} &= \frac{TV_h + TV_v}{C \cdot (h(w-1) + (h-1)w)}
\end{aligned} \tag{9}
$$

The compound cost function $\mathcal{L}$ is thus defined as follows:

$$
\mathcal{L} = \frac{\mathcal{L}_{det} + \mathcal{L}_{OCR}}{2} + \zeta \cdot \mathcal{L}_{TV} \tag{10}
$$

Here, $\zeta = 2.5$ is a scaling factor tuned to encourage regularization without overwhelming the remaining loss components.

Gradient Descent Adversarial patch P is initialized via the Xavier method [9]. We optimize P by minimizing the cost function $\mathcal{L}$ via gradient descent. All perspective warping operations are differentiably computed by the `kornia` software package version 0.8.1 [27] to ensure partial derivatives can be computed for each patch parameter. We use the AdamW optimizer [23] to calculate the value of $P_{c,y,x}^{(t+1)}$. We found that learning rate $\eta = 0.1$ worked best for our use case, and we used a learning rate schedule that would halve η if no improvement in validation-set loss was detected for 5 epochs. We also used an early stopping system that would halt training when no validation-set loss improvement was detected for 20 epochs; when both homography-aware training and TV loss were enabled, this condition was never triggered.

We ran the optimization algorithm for 100 epochs, and in each epoch processed all 240 images in the training dataset, grouped into batches of size 64. We optimized twice, once with a random impersonation target to maximize disruption, and a second time with an impersonation target differing from the true license plate number by only two characters, producing $P_{disrupt}$ and $P_{impersonate}$, respectively.

2.3 Evaluation

We use attack success rate (ASR) to quantify the success of $P_{impersonate}$. For $P_{disrupt}$, due to the high proportion of failed detections in both physical- and digital-world evaluations even with no patch applied, we report the percent reduction in correct read rate rather than the proportion of images with a misread plate. $C_D \in [0, 1]$ is the confidence of the detection model after patch application, $ED_T \in \{0, \ldots, 7\}$ is the edit distance between `fast-alpr`'s output and the true license plate number, and $ED_I \in \{0, \ldots, 7\}$ is the edit distance between `fast-alpr`'s output and the attacker-defined impersonation target.

Physical World: We contacted a professional print studio to print both $P_{disrupt}$ and $P_{impersonate}$ on a heavyweight matte poster of $8.4" \times 16.8"$ dimensions. This was the most expensive part of the attack, costing roughly $50. We visually aligned this poster with our license plate and mounted it between the plate and vehicle, ensuring that no part of the license plate area was obscured.

We first conducted a vertical viewing angle test in full sunlight for $P_{disrupt}$, $P_{impersonate}$, and no patch (control). We used a DJI Mini camera drone in an empty parking lot and recorded distances and altitudes using its built-in GPS and altimeter. We sampled in a grid pattern distances ranging from 1 to 10 meters and altitudes ranging from 0.5 to 3.5 meters, in intervals of 1 and 0.5 meters along distance and altitude axes. We then conducted a set of horizontal viewing angle tests with a standard iPhone 15 camera from the author's height of approximately 6'. A grid pattern of $45' \times 40'$ was established for data collection. Images were captured at 5' intervals in the longitudinal direction, beginning 5' from the license plate and extending to 50'. At each distance, additional images were taken at 5' lateral intervals, ranging from 20' to the right of the plate to 20' to the left, including a centered position. We took images in multiple lighting conditions in full sunlight, dusk, night with dim street lighting, and night with camera flash (Fig. 1). The night with flash case evaluated the patches' performance when the license plate was retroreflectively illuminated, and presents a particularly difficult case in which we expected failure.

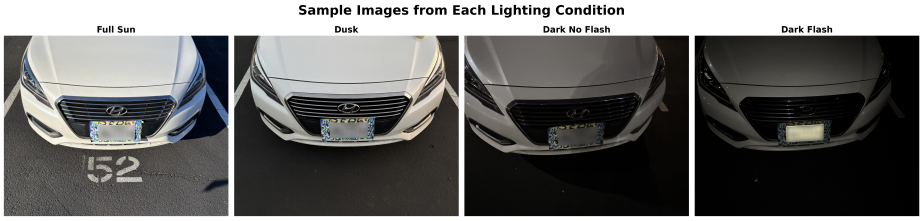


Fig. 1. Sample images from each lighting condition.

Digital World: We evaluated further in the digital world to benchmark best-case performance and understand the specific roles of our contributions via an ablation test. We used the control images of our horizontal viewing angle test, with no patch applied, as our test set, totaling 358 images at known grid coordinates. We manually labeled the corners of the license plates in these images, then digitally inserted $P_{disrupt}$ and $P_{impersonate}$ and ran `fast-alpr`. We finally conducted an ablation test dropping out homography-aware training, TV loss, and both, and compared each variant against the full SPAR approach.

Distance Correlation Analysis: To quantify whether SPAR's performance depends on camera geometry, we measured statistical dependence between per-

formance metrics and camera distance/angle using distance correlation. Distance correlation ranges from 0 to 1 and detects arbitrary nonlinear dependence, equaling zero if and only if the variables are independent [33]. For performance metric X and camera position Y, we compute distance correlation $\mathrm{dCor}(X, Y)$.

To determine plausible natural variation of dCor in independent data, we construct an empirical null distribution $\mathcal{D}_{null}$ by generating 10,000 random permutations of the camera position parameter Y across samples, preserving marginal distributions while destroying dependence. For noise threshold α of $\mathcal{D}_{null}$, we define the independence noise floor δ_α as the α-quantile of $\mathcal{D}_{null}$; we use this strictly to contextualize dCor, not as a statistical test. To express effect magnitude relative to the independence noise floor, we report the normalized dependence ratio:

$$R_\alpha = \frac{\mathrm{dCor}(X, Y)}{\delta_\alpha}. \tag{11}$$

To assess attack-induced dependence, we compare raw distance correlations and normalized dependence ratios between the attack type and control with $\Delta\mathrm{dCor}$ and ΔR_α. Images with no license plate detected were assigned confidence 0 and edit distances 7 (maximum possible). We selected the conservative noise threshold $\alpha = 0.99$ as a bound of non-meaningful levels of statistical dependence. We represent camera distance as d and camera angle as θ.

3 Results



Fig. 2. Final adversarial rims for disruption and impersonation objectives.

The final trained patches (Fig. 2) exhibited fascinating large-scale structure, including characters resembling "w", "3", "M", and East Asian script.

3.1 Physical World

Across all physical world images, we categorized the result of `fast-alpr` as either a correct read, incorrect read, successful impersonation, or failed detection. We found that $P_{impersonate}$ caused a 39.8% reduction in correct reads and

$P_{disrupt}$ caused a 49.2% reduction; $P_{impersonate}$ achieved an impersonation ASR of 13.8%. Additionally, $P_{disrupt}$ achieved a mean confidence reduction of 17.8%, and $P_{impersonate}$ achieved a mean confidence reduction of 3.2%.

For the horizontal viewing angle tests, we broke down patch performance by lighting condition (Fig. 3). Due to the large distances at which many images were taken during the horizontal viewing angle test, the `yolo-v9-t-384-license-plate-end2end` detection model often failed to detect the license plate even in the control case. We thus opted to use the larger `yolo-v9-t-608-license-plate-end2end` detection model to ensure proper evaluation of the OCR portion of the attack. This choice introduced a strictly more challenging evaluation setting.

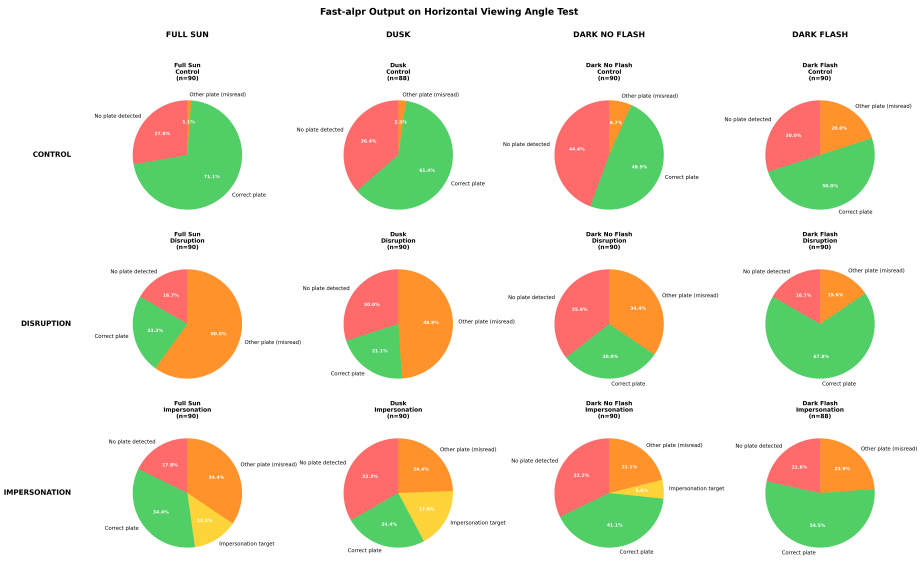


Fig. 3. `fast-alpr` output on horizontal viewing angle test.

$P_{disrupt}$ achieves optimal performance in full sunlight, inducing a 60.0% reduction in correct reads. $P_{impersonate}$ achieves optimal performance in dusk lighting, inducing `fast-alpr` to output the impersonation target 17.8% of the time. Both attacks greatly struggled on the dark flash case because of the license plate's retroreflectivity. Although we could have corrected for this effect by printing our border onto retroreflective metal sheeting, we chose not to do so to better understand the failure mode. Interestingly, $P_{disrupt}$ appears to have *improved* the performance of `fast-alpr` in the dark flash condition, raising the rate of correct reads by 35%, reducing the number of failed detections by 44.3%, reducing the number of misreads by 22.0% (Fig. 3), and improving the detection confidence above the control case (Fig. 4). We explore this further in our Discussion.

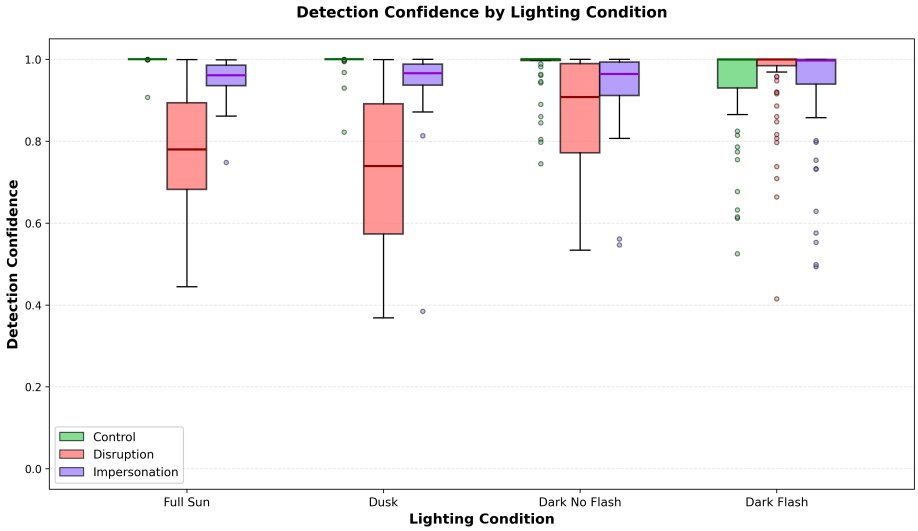


Fig. 4. Detection model's confidence distributions under various lighting conditions.

Table 2. Dependence between performance metrics and camera parameters with no patch applied. R_α denotes distance correlation normalized by the permutation-derived independence bound. Vertical viewing-angle test data included.

X	Y	dCor	δ_α	R_α
C_D	d	0.5342	0.1314	4.07
C_D	θ	0.1350	0.1286	1.05
ED_T	d	0.5434	0.1297	4.19
ED_T	θ	0.1378	0.1300	1.06
ED_I	d	0.5314	0.1312	4.05
ED_I	θ	0.1321	0.1278	1.03

For the horizontal viewing angle tests, we created contour plots showing how the detection confidence and edit distances changed over the grid pattern (Fig. 5). Images without a detected license plate were assigned 0 confidence and excluded from edit distance contour calculations. No obvious correlation is apparent between SPAR's performance and viewing angle, although there does appear to be a dependence on distance stemming from the model itself. To quantify this, we conducted our distance correlation analysis (also including vertical viewing angle results).

These results indicate that the `fast-alpr` base model's performance already has a high level of dependence on d and a weak-to-insignificant dependence on θ (Table 2). $P_{impersonate}$ moderately reduced this dependence, with the notable reductions observed in $\mathrm{dCor}(C_D, \theta)$, $\mathrm{dCor}(ED_T, \theta)$, $\mathrm{dCor}(ED_I, d)$, and $\mathrm{dCor}(ED_I, \theta)$. $P_{disrupt}$ caused much stronger reductions in $\mathrm{dCor}(C_D, d)$ and

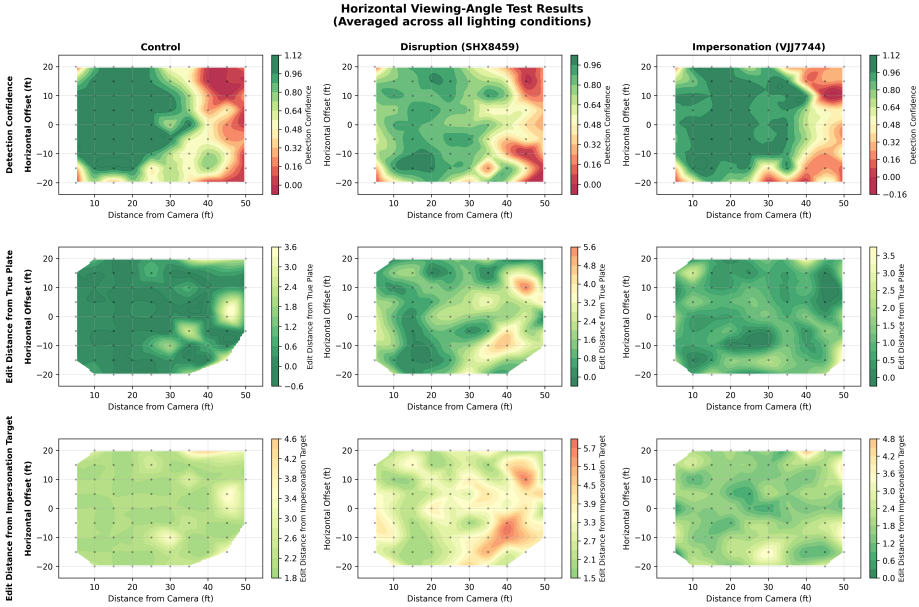


Fig. 5. Horizontal viewing-angle test results (metrics averaged over all lighting conditions).

dCor(ED_T, d), but marginally increased dCor(ED_T, θ). Additionally, $P_{impersonate}$'s correlations with viewing angle are all below the significance threshold δ_α, and $P_{disrupt}$'s correlations with the same are either below or slightly above, so we conclude that SPAR has succeeded in its attack with negligible dependence on camera geometry (Table 3).

To further study the generalizability of our approach in the physical world, we compared the detection confidence of `yolo-v9-t-384-license-plate-end2end` (small detector) vs `yolo-v9-t-608-license-plate-end2end` (large detector) when presented with each adversarial patch. We found that $P_{disrupt}$ reduces the small detector's confidence by 20.5% and the large detector's confidence by 14.2%, on average. In ideal (full sun) lighting conditions, those reductions increase to 31.2% and 22.0%, respectively. Thus SPAR successfully generalizes to larger variants of YOLOv9.

3.2 Digital World

We found that with digital application of the patch, `fast-alpr`'s correct read frequency decreased by 72.5% with $P_{disrupt}$ and 56.7% with $P_{impersonate}$ applied. The impersonation attack was successful in 15.9% of cases, and the disruption attack induced misreads in 46.6% of cases (excluding cases where no plate was detected). $P_{disrupt}$ reduced the detection model's confidence by 23.4% to 0.766 and $P_{impersonate}$ reduced the confidence by 3.8% to 0.962, on average.

Table 3. Attack-induced changes in dependence relative to no patch applied. ΔdCor and ΔR_α are computed as (attack − control) for each metric–camera parameter pair.

Disruption Attack					
X	Y	dCor	R_α	ΔdCor	ΔR_α
C_D	d	0.5185	3.58	−0.1084	−0.82
C_D	θ	0.1206	0.84	−0.0053	−0.08
ED_T	d	0.4754	3.33	−0.1525	−1.24
ED_T	θ	0.1109	0.76	+0.0147	+0.08
ED_I	d	0.4567	3.13	−0.1678	−1.30
ED_I	θ	0.1172	0.82	+0.0125	+0.05

Impersonation Attack					
X	Y	dCor	R_α	ΔdCor	ΔR_α
C_D	d	0.5659	3.99	−0.0100	−0.02
C_D	θ	0.1540	1.07	−0.0287	−0.24
ED_T	d	0.5400	3.74	−0.0573	−0.42
ED_T	θ	0.1589	1.10	−0.0097	−0.08
ED_I	d	0.5963	4.11	−0.0073	−0.15
ED_I	θ	0.1569	1.10	−0.0172	−0.15

We conducted an ablation test of the primary contributions of our work: homography-aware patch insertion during training and the total variation loss component for regularization and structural pattern discovery. We retrained SPAR with each component removed individually, as well as with both removed, and evaluated all variants using the same digital-world evaluation framework to isolate their respective effects. From the final patches of each case (Fig. 6), it is visibly apparent that TV loss achieves its objective of regularizing the pixel values into more natural patterns. Homography transformations, too, promote the discovery of larger patterns rather than exploiting a guaranteed pixel positioning in the OCR input when patches are applied rectangularly.

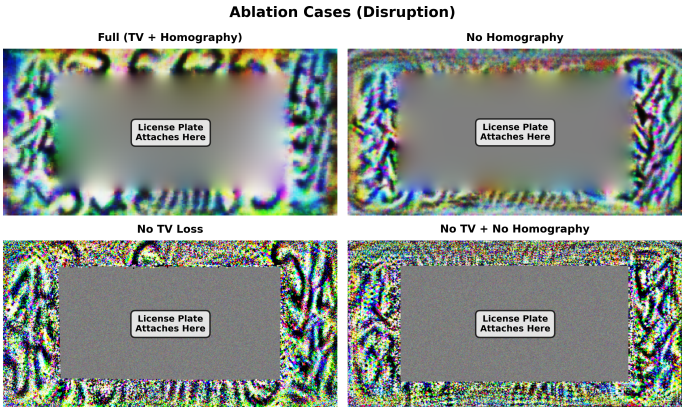


Fig. 6. All final patches used in ablation study.

We found that both TV loss and homography-aware training are critical for SPAR's performance. Removing TV loss causes a 98.7% increase in correct reads on the disruption attack and near complete failure at the impersonation attack. Removing homography-based patch application during training causes a 107% increase in correct reads and a 70.4% reduction in successful impersonations.

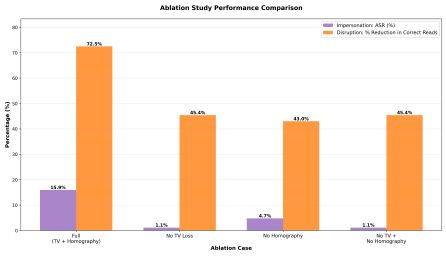


Fig. 7. Performance of each ablation case compared to standard SPAR.

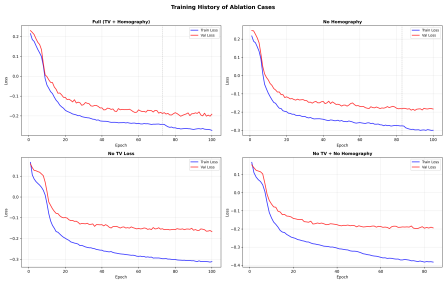


Fig. 8. Training and validation loss curves during ablation testing for the disruption objective.

Removing both results in near-identical performance to removing TV loss alone on the impersonation attack, and slightly better performance than removing TV loss alone on the disruption attack (Fig. 7). We believe this is due to the regularization effect of TV loss, without which SPAR tends to overfit to its training data as a simpler solution to the complex problem presented by homography-based insertion. We observed much greater divergence between training and validation loss when TV loss was excluded, and worse final validation-set performance than when homography-aware training was also excluded (Fig. 8).

Finally, to clarify the specific roles of homography-aware training and TV loss in promoting resistance to viewing angle and camera distance transformations, we repeated the distance correlation test procedure, calculating ΔR_α in reference to R_α of the full disruption and impersonation attacks. We digitally applied patches produced by the full SPAR process and calculated R_α values. These varied somewhat from the physical-world evaluation, likely because of differences in the dataset used, patch application method, and the removal of lighting as a factor, but we proceeded with them for consistency with the ablation test cases.

We then calculated ΔR_α for all ablation cases (Table 4). For $P_{disrupt}$, we found substantial increases in dCor on both distance and viewing angle when homography-aware training was removed. Weaker, but still positive effects on dCor were observed for dropping out TV loss alone, as well as TV + homography. TV loss alone seems to achieve a higher baseline level of attack efficacy, and homography-aware training may squeeze out extra performance by addressing off-angle, longer-distance cases. For $P_{impersonate}$, TV loss is required for impersonation attack efficacy (Fig. 7), but appears to be more sensitive to optimal camera parameters given the increased dCor when homography-aware training is dropped out. Similar to the disruption attack, it is likely that SPAR's success where the naive baseline fails stems from the ability to synthesize large-scale patterns and make those patterns robust to camera distance and viewing angle transformations.

Table 4. Ablation-induced changes in normalized dependence relative to the full attack. ΔR_α is computed as (ablation − original SPAR) for each metric–parameter pair.

Disruption Attack

Removed	X	Y	R_α	ΔR_α
Homography	C_D	d	3.90	+0.32
Homography	C_D	θ	1.17	+0.33
Homography	ED_T	d	4.14	+0.81
Homography	ED_T	θ	1.17	+0.41
Homography	ED_I	d	3.94	+0.81
Homography	ED_I	θ	1.20	+0.38
TV Loss	C_D	d	3.84	+0.26
TV Loss	C_D	θ	0.80	−0.04
TV Loss	ED_T	d	3.87	+0.54
TV Loss	ED_T	θ	0.81	+0.05
TV Loss	ED_I	d	3.71	+0.58
TV Loss	ED_I	θ	0.79	−0.03
Both	C_D	d	3.74	+0.16
Both	C_D	θ	0.93	+0.09
Both	ED_T	d	3.77	+0.44
Both	ED_T	θ	0.88	+0.12
Both	ED_I	d	3.69	+0.56
Both	ED_I	θ	0.87	+0.05

Impersonation Attack

Removed	X	Y	R_α	ΔR_α
Homography	C_D	d	4.04	+0.05
Homography	C_D	θ	1.17	+0.10
Homography	ED_T	d	3.92	+0.18
Homography	ED_T	θ	1.28	+0.18
Homography	ED_I	d	3.98	−0.13
Homography	ED_I	θ	1.22	+0.12
TV Loss	C_D	d	4.05	+0.06
TV Loss	C_D	θ	0.93	−0.14
TV Loss	ED_T	d	4.12	+0.38
TV Loss	ED_T	θ	0.91	−0.19
TV Loss	ED_I	d	4.07	−0.04
TV Loss	ED_I	θ	0.92	−0.18
Both	C_D	d	4.14	+0.15
Both	C_D	θ	1.00	−0.07
Both	ED_T	d	4.16	+0.42
Both	ED_T	θ	1.03	−0.07
Both	ED_I	d	4.11	−0.00
Both	ED_I	θ	1.00	−0.10

4 Discussion

Our results broadly demonstrate SPAR's success at its core objectives. SPAR in the physical world reduced `fast-alpr`'s correct read rate by 49.2% (60% in full sunlight), and had an impersonation ASR of 13.8% (17.8% in dusk lighting). It also achieved a 72.5% reduction in correct read rate and a 15.9% impersonation ASR in the digital world. $P_{disrupt}$'s attacks on the detection model caused a 17.8% confidence reduction in the physical world and 23.4% in the digital world. SPAR achieved this without internal access to ALPR software or hardware during attack execution, without obscuring or altering the primary license plate area, and without creating excessive dependence on camera distance or viewing angle. The attack was carried out for a total cost of under $100, with an approach designed and fully implemented by agentic coding tools. SPAR thus satisfies its constraints of realistic access, street-legality, simplicity, and physical-world practicality.

Adversarial patches function by inducing large-magnitude feature norms in their target models, overwhelming legitimate signal [41]. Our ablation study shows that the solution of least resistance discovered by simple gradient descent optimization relies on high-magnitude input pixels with high contrast to their neighbors to induce these large-magnitude feature norms. However, these patches fail to generalize to the evaluation dataset when applied via homography. Conversely, our most successful patches exhibit large, well-defined, character-like

Table 5. Comparison of attack success rates (ASR) reported in prior ALPR attacks. For SPAR's P_{disrupt}, reduction in correct reads is reported instead of ASR.

Approach	Disrupt (Digital)	Impers. (Digital)	Disrupt (Physical)	Impers. (Physical)
Zha et al.	97.3%	89.5%	N/A	91.4%
Gu et al.	N/A	N/A	N/A	N/A
Qian et al.	N/A	93.3%	N/A	N/A
Kwon & Baek	85.0%	N/A	N/A	N/A
Jiamsuchon et al.	91.0%	70.0%	N/A	N/A
Zhang et al.	95.6%	N/A	N/A	N/A
SPAR	**72.5%**	**15.9%**	**60.0%**	**17.8%**

structures. We believe this reflects a change in strategy from fuzzing early convolution filters with extreme inputs, to hijacking real features extracted and used by the model to make classification decisions.

Total variation loss promotes the emergence of these structures in the final patches. It makes the impersonation attack feasible and forces SPAR to find a different strategy for increasing CNN feature norms. It also regularizes the patch and reduces overfitting during training, enabling the attack to be executed with a very small training dataset. Homography-aware training works in tandem to break the dependence on viewing geometry when TV loss is used alone. It further promotes the development of character-like structures by changing pixels' distances to their neighbors depending on the viewing angle, making it difficult to coordinate high-magnitude pixel-perfect attacks.

In the dark flash condition, $P_{disrupt}$ seemed to improve the performance of `fast-alpr`. We believe this is due to the IoU component of $\mathcal{L}_{det}$, which rewards the model for producing a bounding box surrounding the applied patch. It is likely that this was the limiting factor for `fast-alpr` to output accurate results in the control case. The retroreflectivity greatly increased the contrast of the license plate against the surrounding background, so confidence and OCR accuracy could be high once the location of the license plate in the image was determined. $P_{disrupt}$ is far better at reducing detection model confidence than $P_{impersonate}$ overall (Fig. 4), so this is likely why we see this effect much more prominently in the disruption attack than impersonation. We therefore interpret this as a partial success that could be converted into a full success were the adversarial rim printed onto retroreflective sheeting.

4.1 Comparison to Prior Work

We present thorough comparisons to existing literature in this space (Table 5). Because no prior work has attempted physical-world validation in challenging lighting conditions, we select SPAR's physical-world performance metrics from the optimal lighting conditions. SPAR's objective is not to produce a state-of-the-art attack on ALPR systems; it is to show what might be possible with LLM

assistance, and quantify the risk with physical-world evaluation. Additionally, SPAR adopts a street-legality constraint that precludes the perturbation and spot attacks used in prior work. Under this lens, SPAR is a simple, highly-constrained solution that works unexpectedly well.

4.2 Limitations, Defenses, Ethics

SPAR is a highly focused attack against a single open-source ALPR system (`fast-alpr`) that requires full access to internal model weights during training. Broad generalization is not the purpose of our work, and we do not claim our patch designs would be successful on other license plates, other vehicles, or other ALPR systems. We instead aim to demonstrate that such an attack is possible, practical, and performant in the real world if an adversary gains access to an ALPR's internal model weights, even if they have limited technical knowledge.

Defenders can protect ALPR systems against many adversarial attacks using well-established countermeasures in the literature. Simply compressing and then re-sharpening an image (e.g. via an autoencoder) can eliminate adversarial perturbations, and clipping feature norms would likely be sufficient to protect against simple SPAR-like attacks [41]. Attack detection is also possible, as most adversarial patches (including SPAR) appear very unnatural to the human eye. Spot attacks might be the most problematic, as despite being illegal, it may be difficult to determine whether dirt marks on a license plate are due to an adversarial attack or an offroad trip; including adversarial examples in the training data can improve robustness in these cases [30].

Ethics Statement Our finding that LLMs can automate a large portion of the work needed to execute a physical-world adversarial attack against a real ALPR system is deeply concerning. It is because of this finding that we have pursued publication of this work, as the ideas presented here already exist as separate, well-established elements in the prior literature, and we have demonstrated that LLMs are already capable of synthesizing them into a full exploitation solution. We have demonstrated a white-box attack against a single open-source system to raise awareness in the vehicle security community about the realistic threat of LLM-assisted adversarial attacks and the sensitivity of neural network weights, but have declined to pursue transfer to commercial systems to avoid enabling real-world harm.

Due to the risks of LLM-assisted modifications to our software enabling that transfer, we have elected not to release the software implementing our attack. We believe we have included adequate technical details in our manuscript to enable legitimate researchers to replicate our findings. Additionally, because our dataset contains sensitive personally-identifiable information such as an author's real license plate, we are unable to release the dataset used to train and evaluate SPAR.

5 Conclusion

SPAR is a street-legal physical-world attack that disrupted both the detection and OCR portions of `fast-alpr` simultaneously, while being robust to camera perspective transformations. SPAR achieves this with a combination of homography-aware training and total variation loss, which promote the emergence of large-scale character-like structures that hijack features used by `fast-alpr` to read license plates. Despite being a white-box attack against a single system, our attack model could be repeated by a low-resourced adversary with LLM assistance. Our results demonstrate that such adversarial attacks are not merely theoretical, and present real risks for modern ALPR systems. We hope future work can address key limitations of SPAR, including retroreflectivity, infrared-spectrum cameras, and black-box training and evaluation, as well as develop more robust defenses against this form of attack.

References

1. Ankandrew: Fast & Lightweight License Plate OCR. https://github.com/ankandrew/fast-plate-ocr (Sep 2025), gitHub repository, last accessed 2025/12/02, Version 1.0.1
2. Ankandrew: Fastalpr. https://github.com/ankandrew/fast-alpr (Dec 2025)
3. Ankandrew: Open Image Models. https://github.com/ankandrew/open-image-models (Sep 2025), version 0.4.0
4. Anthropic: Threat intelligence report: August 2025. Tech. rep., Anthropic (Aug 2025), https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf, last accessed 2026/03/20
5. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial Patch (May 2018). https://doi.org/10.48550/arXiv.1712.09665
6. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust Physical-World Attacks on Deep Learning Models (Apr 2018). https://doi.org/10.48550/arXiv.1707.08945
7. Ferguson, N., Schneier, B., Kohno, T.: The context of cryptography. In: Cryptography Engineering: Design Principles and Practical Applications, chap. 1. Wiley, Indianapolis, IN (2010)
8. Finklea, K.: Law Enforcement and Technology: Use of Automated License Plate Readers. Tech. Rep. CRS Report No. R48160, Congressional Research Service (Aug 2024), https://www.congress.gov/crs-product/R48160
9. Glorot, X.: Understanding the difficulty of training deep feedforward neural networks. Journal of Machine Learning Research - Proceedings Track **9** (01 2010)
10. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples (Mar 2015). https://doi.org/10.48550/arXiv.1412.6572
11. Gu, Z., Su, Y., Liu, C., Lyu, Y., Jian, Y., Li, H., Cao, Z., Wang, L.: Adversarial Attacks on License Plate Recognition Systems. Computers, Materials & Continua **65**(2), 1437–1452 (Aug 2020). https://doi.org/10.32604/cmc.2020.011834
12. Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., Shi, H.: Escaping the big data paradigm with compact transformers (2021). https://doi.org/10.48550/arXiv.2104.05704

13. Jiamsuchon, C., Suaboot, J., Rattanavipanon, N.: On the feasibility of attacking Thai LPR systems with adversarial examples (Jan 2023). https://doi.org/10.48550/arXiv.2301.05506
14. Kerckhoffs, A.: La cryptographie militaire. Journal des sciences militaires (1883)
15. Kocer, H.E., Çevik, K.: Artificial neural networks based vehicle license plate recognition. Procedia CS **3**, 1033–1037 (Dec 2011). https://doi.org/10.1016/j.procs.2010.12.169
16. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial Examples in the Physical World. In: Yampolskiy, R.V. (ed.) Artificial Intelligence Safety and Security, pp. 99–112. Chapman and Hall/CRC, First edition. \textbar Boca Raton, FL : CRC Press/Taylor & Francis Group, 2018., 1 edn. (Jul 2018). https://doi.org/10.1201/9781351251389-8
17. Kwon, H., Baek, J.W.: Adv-Plate Attack: Adversarially Perturbed Plate for License Plate Recognition System. Journal of Sensors **2021**(1), 6473833 (2021). https://doi.org/10.1155/2021/6473833
18. Li, H., Shen, C.: Reading Car License Plates Using Deep Convolutional Neural Networks and LSTMs (Jan 2016). https://doi.org/10.48550/arXiv.1601.05610
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal Loss for Dense Object Detection (Feb 2018). https://doi.org/10.48550/arXiv.1708.02002
20. Liu, A., Wang, J., Liu, X., Cao, B., Zhang, C., Yu, H.: Bias-Based Universal Adversarial Patch Attack for Automatic Check-Out. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 395–410. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58601-0_24
21. Liu, X., Yang, H., Liu, Z., Song, L., Li, H., Chen, Y.: DPatch: An Adversarial Patch Attack on Object Detectors (Apr 2019). https://doi.org/10.48550/arXiv.1806.02299
22. Liu, Y., Huang, H., Cao, J., Huang, T.: Convolutional neural networks-based intelligent recognition of Chinese license plates. Soft Computing **22** (Apr 2018). https://doi.org/10.1007/s00500-017-2503-0
23. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (Jan 2019). https://doi.org/10.48550/arXiv.1711.05101
24. Masood, S.Z., Shu, G., Dehghan, A., Ortiz, E.G.: License Plate Detection and Recognition Using Deeply Learned Convolutional Neural Networks (Mar 2017). https://doi.org/10.48550/arXiv.1703.07330
25. Nguyen, K., Fernando, T., Fookes, C., Sridharan, S.: Physical Adversarial Attacks for Surveillance: A Survey (May 2023). https://doi.org/10.48550/arXiv.2305.01074
26. Qian, Y., Ma, D., Wang, B., Pan, J., Wang, J., Gu, Z., Chen, J., Zhou, W., Lei, J.: Spot evasion attacks: Adversarial examples for license plate recognition systems with convolutional neural networks. Computers & Security **95**, 101826 (Aug 2020). https://doi.org/10.1016/j.cose.2020.101826
27. Riba, E., Mishkin, D., Ponsa, D., Rublee, E., Bradski, G.: Kornia: An Open Source Differentiable Computer Vision Library for PyTorch (Oct 2019). https://doi.org/10.48550/arXiv.1910.02190
28. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena **60**(1-4), 259–268 (1992). https://doi.org/10.1016/0167-2789(92)90242-F
29. Shack, J., Petrovic, K., Saukh, O.: Breaking the Illusion: Real-world Challenges for Adversarial Patches in Object Detection (Nov 2024). https://doi.org/10.48550/arXiv.2410.19863

30. Shrestha, B., Khakurel, G., Simkhada, K., Adhikari, B.: Adversarial sample generation and training using geometric masks for accurate and resilient license plate character recognition. https://doi.org/10.48550/arXiv.2311.12857, http://arxiv.org/abs/2311.12857
31. Stein, N.: Automated License Plate Readers: Legal and Policy Evaluation. Tech. rep., Science, Technology and Public Policy Program, University of Michigan (Jan 2023), https://stpp.fordschool.umich.edu/research/policy-brief/automated-license-plate-readers-legal-and-policy-evaluation, policy brief, last accessed 2026/03/20
32. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks (Feb 2014). https://doi.org/10.48550/arXiv.1312.6199
33. Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. The Annals of Statistics **35**(6) (Dec 2007). https://doi.org/10.1214/009053607000000505
34. Tao, G., An, S., Cheng, S., Shen, G., Zhang, X.: Hard-label Black-box Universal Adversarial Patch Attack . https://doi.org/10.5555/3620237.3620277
35. Texas Court of Criminal Appeals: State v. johnson. 219 S.W.3d 386 (Tex. Crim. App. 2007) (Feb 2007)
36. Texas Legislature: Texas Transportation Code § 504.945: Wrong, fictitious, altered, or obscured license plate. Tex. Transp. Code § 504.945 (2024), https://statutes.capitol.texas.gov/Docs/TN/htm/TN.504.htm
37. United States Court of Appeals, Fifth Circuit: United states v. granado. 302 F.3d 421 (5th Cir. 2002) (2002)
38. United States Court of Appeals, Fifth Circuit: United states v. contreras-trevino. 448 F.3d 821 (5th Cir. 2006) (2006)
39. Wang, C.Y., Liao, H.Y.M.: YOLOv9: Learning what you want to learn using programmable gradient information (2024). https://doi.org/10.48550/arXiv.2402.13616
40. Xu, K., Zhang, G., Liu, S., Fan, Q., Sun, M., Chen, H., Chen, P.Y., Wang, Y., Lin, X.: Adversarial T-Shirt! Evading Person Detectors in a Physical World. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020, vol. 12350, pp. 665–681. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_39
41. Yu, C., Chen, J., Xue, Y., Liu, Y., Wan, W., Bao, J., Ma, H.: Defending against Universal Adversarial Patches by Clipping Feature Norms. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16414–16422. IEEE, Montreal, QC, Canada (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.01612
42. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.: UnitBox: An Advanced Object Detection Network. In: Proceedings of the 24th ACM International Conference on Multimedia. pp. 516–520 (Oct 2016). https://doi.org/10.1145/2964284.2967274
43. Zha, M., Meng, G., Lin, C., Zhou, Z., Chen, K.: RoLMA: A Practical Adversarial Attack against Deep Learning-based LPR Systems (Oct 2019). https://doi.org/10.1007/978-3-030-42921-8_6
44. Zhang, K., Qin, Y., Pan, H.: LPLA : The Adversarial Attack Against License Plate Recognition Systems. In: Zhang, W., Tung, A., Zheng, Z., Yang, Z., Wang, X., Guo, H. (eds.) Web and Big Data. pp. 407–421. Springer Nature, Singapore (2024). https://doi.org/10.1007/978-981-97-7232-2_27