Cluster Computing (2024) 27:11627–11644
https://doi.org/10.1007/s10586-024-04513-4

Deep learning adversarial attacks and defenses on license plate recognition system

Conrado Vizcarra[1] · Shadan Alhamed[2] · Abdulelah Algosaibi[1] · Mohammed Alnaeem[3] · Adel Aldalbahi[4] · Nura Aljaafari[1] · Ahmad Sawalmeh[5] · Mahmoud Nazzal[6] · Abdallah Khreishah[6] · Abdulaziz Alhumam[1] · Muhammad Anan[5]

Received: 14 February 2024 / Revised: 2 April 2024 / Accepted: 13 April 2024 / Published online: 1 June 2024

Abstract
The breakthroughs in Machine learning and deep neural networks have revolutionized the handling of critical practical challenges, achieving state-of-the-art performance in various computer vision tasks. Notably, the application of deep neural networks in optical character recognition (OCR) has significantly enhanced the performance of OCR systems, making them a pivotal preprocessing component in text analysis pipelines for crucial applications such as license plate recognition (LPR) systems, where the efficiency of OCR is paramount. However, despite the advancements, the integration of deep neural networks in OCR introduces inherent security vulnerabilities, particularly susceptibility to adversarial examples. Adversarial examples in LPR systems are crafted by introducing perturbations to original license plate images, which can effectively compromise the integrity of the license plate recognition process, leading to erroneous license plate number identification. Given that the primary goal of OCR in this context is to accurately recognize license plate numbers, even a single misinterpreted character can significantly impact the overall performance of the LPR system. The vulnerability of LPR systems to adversarial attacks underscores the urgent need to address the security weaknesses inherited from deep neural networks. In response to these challenges, the exploration of alternative defense mechanisms, such as image denoising and in-painting, presents a compelling approach to bolstering the resilience of LPR systems against adversarial attacks. By prioritizing practical implementation and integration of image denoising and inpainting techniques align with the operational requirements of real-world LPR systems. These methods can be seamlessly integrated into existing pipelines, offering pragmatic and accessible means of enhancing security without imposing significant computational overhead. By embracing a multi-faceted approach that combines the strengths of traditional image processing techniques, the research endeavors to develop comprehensive and versatile defense strategies tailored to the specific vulnerabilities and requirements of LPR systems. This holistic approach aims to fortify LPR systems against adversarial threats, thereby fostering increased trust and reliability in the deployment of OCR and LPR technologies across various domains and applications.

Keywords Adversarial attack · Adversarial defense · License plate recognition system · Optical character recognition

1 Introduction

Machine Learning (ML) and Deep Neural Networks (DNN) have been able to handle a variety of critical and practical problems that reached state-of-the-art and shown human-competitive performance on handling a range of complex Computer Vision (CV) tasks that have shown improved results and remarkable accuracy [1]. With the continuous advancements in these technologies and the ease of obtaining the necessary hardware to train sophisticated models, efficient Deep Learning (DL) software frameworks are becoming more accessible and rapidly maturing to the point where they can be used in safety and security-critical applications as part of the control pipelines in physical systems, particularly those arising from CV applications. Self-driving cars, drones, robotics, and road traffic applications like License Plate Recognition (LPR) systems are among these applications that utilize DL

Extended author information available on the last page of the article

solutions that are unquestionably on their way to becoming a significant part of our daily lives. Problems that were formerly thought to be unsolvable are now being solved with superhuman precision. The use of deep learning in license plate recognition systems has been a topic of extensive research and discussion in recent years. Deep learning techniques, powered by deep neural networks, have emerged as a promising approach for tackling complex computer vision tasks with remarkable accuracy. In the field of license plate recognition, deep learning algorithms have shown significant improvements in accuracy and robustness compared to traditional methods [1–3]. However, Szegedy et al. discovered an intriguing flaw on DNN models against adversarial examples. These adversarial examples are maliciously generated inputs that are used to attack a target model, created by adding small-magnitude perturbations to the original input where humans cannot tell the difference but most DNN models would incorrectly classify the adversarial examples [4]. Adversarial examples have piqued the interest of two different sides. On one side, adversarial examples threaten the integrity and security of machine learning algorithms in use, for example in the LPR system. A minor perturbation on a vehicle license plate registration number can completely change how an LPR system interprets the resulting text, potentially leading to serious consequences such as the driver receiving an incorrect traffic violation notice or, worse, being falsely accused of committing a crime. On the other hand, adversarial examples shine a bright light on the gap between human and machine sensory information processing, pointing the way to more robust, human-like architectures [5]. Adversarial examples have gotten much attention, and CV research has made significant progress in comprehending the field of adversarial examples, starting in the digital realm (for example, by changing images corresponding to a scene) and more recently in the physical domain [6]. White-box access is currently used in existing physical approaches. Separately, researchers have begun investigating black-box hard-label attacks operating in the digital domain, in which attackers simply have query access to the model to retrieve the top-1 class label without any confidence information [1]. The transferability property of adversarial examples, which allows them to be used across multiple models, enables black-box adversarial attacks without knowing the target model's weights and structures. Black-box attacks have been demonstrated to be viable in real-world scenarios, posing a threat to security-sensitive applications. As a result, robust defenses against adversarial attacks are critical [7]. Various defensive strategies for adversarial examples have recently been proposed to mitigate adversarial examples from a complex method like adversarial training, which aims to improve the DL model's resiliency by introducing adversarial examples into the training stage, to a more direct approach such as denoising, to remove adversarial perturbations from the input or alleviate its effects on high-level features learned by the Deep learning model [8].

LPR is an important component in numerous trafic control systems that enables automatic identification of vehicles in many applications such as speed radars, surveillance systems, parking management, border entrance security, toll gate automation and other intelligent processing in smart cities [2, 8, 9] LPR systems provide a positive impact on people's lives, which is a concern with improving transportation safety, mobility and enhancing productivity using advanced technologies [10]. Unfortunately, the LPR system that implements Optical Character Recognition (OCR) inherits all the DNN's perplexing security issues, especially on the OCR model, which is subject to adversarial examples [9, 11]. When the purpose of OCR is to recognize license plate numbers, even a single character that is incorrectly recognized can have a significant impact on the LPR system. For Instance, recognizing the number '8' as the closest number in shape '3' will generate a wrong license plate. Inaccurate OCR reading can lead to critical issues in systems vital for public safety, like law enforcement and traffic control, potentially impacting numerous individuals. A minor perturbation on a vehicle license plate registration number can significantly impact how an OCR system recognizes the LP text, potentially resulting in serious repercussions. This could range from a driver erroneously receiving a traffic violation notice to being falsely implicated in a criminal offense. Also, such vulnerability opens the door for malicious individuals to deliberately manipulate a license plate number, aiming to deceive the OCR system and pose a substantial security threat.

Adversarial examples, particularly in the LPR system, are commonly created by adding noises to original license plate images; therefore, denoising adversarial examples before delivering them to the LPR system is a sensible approach. A few studies on the LPR system take into consideration protecting the system from adversarial attacks, which may degrade the accuracy of the LPR system. Moreover, the datasets that are used do not contain real plates or enough images, which may affect the accuracy of the LPR system in the real world. For all of the mentioned reasons, there is a need to propose an LPR system with high accuracy achieved by considering all possible scenarios for adversarial attacks and providing approaches for defense.

This research contributes to the field of deep learning-based security by addressing the vulnerability of LPR systems to adversarial attacks. We first present a comprehensive analysis of adversarial attacks on LPR systems, wherein we systematically introduce perturbations to

license plates to deceive LPR models. We then propose and evaluate two defense mechanisms to mitigate the impact of these attacks. The first defense involves leveraging image denoising techniques to remove adversarial perturbations, while the second defense employs image inpainting methods to reconstruct the original content of the license plate. Through extensive experimentation, we demonstrate the efficacy of these defenses in enhancing the resilience of LPR systems against adversarial attacks. Our research differs from existing studies in several significant ways. Firstly, while prior work has primarily focused on adversarial attacks and defenses in general image classification tasks, our research specifically targets LPR systems, which have unique characteristics and requirements. By tailoring our investigation to LPR, we provide insights into the specific vulnerabilities and challenges associated with license plate recognition. Additionally, the proposed defense mechanisms, namely image denoising and inpainting, offer a specialized approach to mitigating adversarial attacks in the context of LPR, thereby contributing novel strategies to the broader landscape of adversarial defense research. Furthermore, our experimental evaluation of LP datasets underlines the practical relevance and applicability of our findings, distinguishing our work from purely theoretical studies. Image denoising and inpainting techniques can be implemented using a variety of traditional image processing algorithms and simpler architectures, making them relatively straightforward to integrate into existing LPR systems. This practical aspect is significant, as it facilitates the adoption and deployment of the proposed defenses in real-world applications without imposing extensive computational or infrastructure requirements. Moreover, the use of established image processing methods may offer greater ease of implementation and performance optimization compared to developing and training complex learning models, as suggested by similar studies, especially in scenarios where computational resources are limited. Overall, our research provides a focused and practical contribution to the domain of adversarial attacks and defenses by addressing a critical security concern in LPR systems and offering tailored defense strategies to safeguard against such threats.

The rest of the paper is organized as follows: Sect. 2 presents the related studies and the proposed adversarial attacks and defenses. Section 3 presents the experiment, testing, and discussions, and Sect. 4 contains the conclusion with recommendations for further research.

2 Related work and proposed methods

LPR is one of the key processes when it comes to vehicle identfication aside from license plate detection (LPD). As an active study problem in the image processing domain, different research focuses on various methodologies, strategies, and algorithms to improve such processes. LPR systems are dramatically improving in recognition accuracy and efficiency thanks to the rapid growth of DL, LPR systems can attain high accuracy levels of over 99%. However, the modern DL model is vulnerable to adversarial attacks. For instance, a small modification to a license plate image that is undetectable to humans can easily fool the DL model used by the LPR system, resulting in inaccurate license plate recognition [2]. Nevertheless, various defensive strategies have been proposed to improve DL models' ability to defend against adversarial examples, such as (1) adversarial training, in which the DL model is exposed to adversarial examples during training to gain some level of immunity against them, (2) adversarial detection, that aims to detect the presence of adversarial perturbations in the input during inference, and (3) transformation-based methods, which aim to clean inputs to make them benign for the target model to highlight a few [12].

The following sections present a review of related works from three perspectives: license plate recognition systems, the attack methods of generating adversarial examples, and the defensive techniques to mitigate adversarial examples.

2.1 License plate recognition system

Research on LPR methods has made impressive progress in the effectiveness of any LPR system. Researchers are aiming to develop a system that is both reliable and efficient in order to achieve good results. In [13] and [14], both research proposed an LPR system for license plate detection and character recognition that uses YOLO (You Only Look Once). YOLO is a deep learning model that consists of a collection of object detection architectures and models that have been pre-trained on the COCO dataset. While both research used YOLO, [13] utilizes YOLOv4-Tiny, a light-weight network and a simplified form of YOLOv4 for license plate detection, and proposed a modified YOLOv4 named M-YOLOv4 to recognize license plate characters, while [14] uses YOLOv5 for license plate detection and License Plate Recognition Network (LPRNet), an LPR system that uses Deep Neural Network for character recognition. In terms of Saudi License Plate LPR System research, [15] developed an integrated LPR system to reduce manpower and eliminate redundant work. To extract the license plate number, the researchers use digital

Table 1 Summary of related work for license plate recognition system

Source	Techniques		
	Image preprocessing	License Number Detection	OCR
[14]	Image conversion (RGB to *BGR*) and Grayscale	YOLOv4 - Tiny	M-YOLOv4
[15]	Grayscale, thresholding, and contour filtering	YOLOv5, DetectNet_v2 and ResNet 10	LPRNet
[16]	Grayscale, threshold and Contour filtering	Character vector and boundary matching processing	OpenALPR, Tesseract, and K-nearest neighbor
[11]	Grayscale, Binarization, Median Filtering, Edge Detection, Dilation, Convolution, Flood Fill, and Image Cropping	Character segmentation is based on thresholding and Connected Component Analysis	Template Matching
[17]	grayscale, Gaussian blurring and thresholding filtering	Edge detection, contour processing and Morphological Gradient	OCR based on Hidden Markov Model

image processing techniques including grayscalling, threshold and contour filtering, and vector and boundary processing. In order to recognize license plates, OpenALPR, Tesseract, and K-nearest neighbor (KNN) are employed. Although each algorithm achieves distinct results, KNN outperforms the others when it comes to recognizing license plate characters. While several methods have been proposed to enhance the LPR system, Elsaid et al. [10] developed a real-time LPR system for Saudi license plates that includes preprocessing, license plate localization, character segmentation, feature extraction, and character recognition. For character recognition, they compared the features of each character with features stored in the characters' database using Template Matching. Similarly, R. Antar et al. [16] proposed Automatic license plate recognition for Saudi plates with the use of edge detection, segmentation, and contouring techniques. Moreover, OCR is utilized to extract letters and numbers from the processed images. Results in an accuracy of 92.4% for Arabic and 96% for English texts on license plates.

Image preprocessing, LP number detection, and OCR are common components of an LPR system that enable the extraction of the license plate number from an image of the license plate, as shown in the literature review. The accompanying works and license plate recognition procedures for each investigation are summarized in Table 1.

2.2 Adversarial attacks

LPR system performance has substantially improved since the widespread adoption of DL in OCR technology, and it is now used in a variety of essential applications where OCR quality is critical. various studies are reviewed and discovered different attack strategies to prepare the adversarial attack and how it would be used to generate adversarial examples, which are summarized below.

Zha et al. [2] proposed a practical adversarial attack called Robust Light Mask Attacks (RoLMA) against deep learning-based LPR systems, which uses illumination technologies to create several light spots as noises on the license plate, and designs targeted and non-targeted strategies with different constraints such as the color, size, and brightness of light spots to find the best adversarial example against HyperLPR, a state-of-the-art LPR system. Similarly, Qian et al. [17] proposed an evasion attack on CNN classifiers that adds predetermined perturbations to specific regions of the license plate, simulating naturally formed spots (such as dirt, sludge, etc.) and using a genetic-algorithm-based method to achieve optimal perturbation positions. Additionally, [18], presents a classic method to generate adversarial examples for attacking LPR systems by adding random Gaussian noise to generate adversarial examples to fool the HyperLPR system. Carlini and Wanger invented the C &W approach for defeating defensive distillation, which is one of the most potent attacks. C &W attack finds the minimal value to misclassify the image so that the difference of the distance is less than a threshold of a particular distance metric [19], which is then used by Yang et al. [20] to develop a physical adversarial attack that targeted object detection models and manufactured real objects that show the adversarial effect. The attack algorithm and the physical adversarial objects were able to fool several object detection models such as

Table 2 Summary of related work for adversarial attack

Source	Name	Attack methodology
[2]	Robust Light Mask Attacks (RoLMA)	Uses illumination technologies to create light spots as noises on the license plate
[17]	Evasion attack on CNN classifiers	Adds predetermined perturbations to specific regions of the license plate, simulating naturally formed spots
[18]	Classic methods to generate adversarial examples (FGSM, BIM, etc.)	Adding random Gaussian noise to generate adversarial examples
[19]	Carlini and Wanger attack (C &W attack)	Finds the minimal value to misclassify the image
[20]	Carlini and Wanger attack (C &W attack)	Targets object detector for license plate detection models such as SSD, YOLO, and Faster R-CNN by generating physical adversarial objects using C &W and SSD Model
[21]	Jacobian-based Saliency Map Attack (JSMA) method	Perturbed number of features by a constant offset that is maximizing the saliency map
[9]	Momentum Iterative Method (MIM) Watermark Attack	Apply watermark to decorate perturbations as to attack Connectionist Temporal Classification (CTC)-based OCR model
[11]	FAWA: Fast Adversarial Watermark	Disguising the perturbations as watermarks to attack sequence-based OCR models in a white-box manner
[22]	Visible Watermark attack	Inserts a watermark to the target image in order to obstruct the classification result of an Inception V3 model that has been pre-trained on ImageNet
[23]	Neural image modification	Alters an input image into an adversarial sample using Neural image modification to preserve the characteristics of an image
[24]	Transformation-invariant aggregated attack	Adversarial attack on media convergence and demonstrates the vulnerability of content analysis networks against adversarial examples

SSD, YOLO, and Faster R-CNN for license plate detection. Combey et al. [21], developed the Jacobian-based Saliency Map Attack (JSMA) method for targeted attacks, they perturbed a large number of features by a constant offset that maximized the saliency map. Watermarking has been utilized to generate adversarial samples in addition to the random noises generated by the methods outlined in the previous research as shown in [9, 11]. In addition, a visible adversarial attack approach is presented that adds a watermark on the target image in order to interfere with the classification result of an Inception V3 model that has been pre-trained on ImageNet. The location, transparency, color, angle, and size of the watermark are all adjusted iteratively [22]. Recently, [23] proposed a system for producing adversarial samples using neural image modification that aims to alter an image into an adversarial sample to preserve the fine characteristics of the input image and is robust to some degree against defensive distillation. In the same year [24], addresses the security risk in media convergence, demonstrating the vulnerability of content analysis networks to adversarial examples. To increase the transferability of the ensuing adversarial cases, the researchers presented a transformation-invariant aggregated attack method. When paired with other algorithms, the approach can boost the success rates of black-box attacks.

Moreover, some of the adversarial attacks may affect the performance of LPR in most of the cases. In [18], the FGSM attack showed a high success rate in the HyperLPR system since FGSM is generated based on the gradient of the image. For that reason, FGSM attack misleads HyperLPR for most of the license plates, but it takes more time than other attacks applied to the HyperLPR system.

A significant amount of research has been done on adversarial attacks on image recognition systems, where most of the previous work focuses on image recognition systems using well-known datasets including MNIST and CIFAR; However, few of them directly extend the attack and defense methods on the LPR system, specifically on Saudi Arabian license plates. The idea that came up is to transfer these attacks into the LPR system, specifically on Saudi license plates, due to the transferability aspect of adversarial examples. To generate adversarial examples, FGSM, C &W, and watermarking approaches are employed, tested, and compared in Sect. 2. Also, OCR technology is leveraged to create an LPR system in this study, which will serve as a test bed for the proposed adversarial attacks. Table 2 shows the summary of related work for adversarial attacks.

2.3 Adversarial defense

Currently, defense approaches are separated into two categories: increasing the resilience of the classifier model and preprocessing the input without modifying the classifier model. Adversarial training (AT) is perhaps the most effective defense approach against adversarial examples that are publicly known. AT expands the training set with

Table 3 Summary of related works for adversarial defense

[24]	Adversarial training	Adversarial dual network learning with randomized nonlinear image transforms for defending deep neural networks from adversarial attacks
[25]	Adversarial training	Iterative adversarial retraining strategy that retrains the model using both Gaussian noise augmentation and adversarial generation strategies
[26]	Adversarial training	Hadamard code-based class labeling for training neural network models as a proactive defensive strategy against adversary attacks
[27]	Image Reconstruction	Reconstructs an input adversarial example into a clean output image to defend against adversarial attacks by learning a precise mapping of the adversarial examples to the reconstructed examples
[28]	Traditional restoration techniques	Using traditional restoration techniques such as spatial domain filtering and transform domain filtering
[29]	Multiple Filtering and Image Rotation	Reformulates single image classification as multiple image classification through two-stage filtering and image rotation

adversarial examples and trains with the original samples to improve the model's tolerance to adversarial examples and hence its robustness. Yuan and He [25] combined adversarial dual network learning with randomized nonlinear image transform to propose a strategy for defending deep neural networks from adversarial attacks. To remove image perturbation, they used a randomized nonlinear image transform so that the attacker could not learn it directly during the attacks. A generative cleaning network is used to recover the original image while cleaning up the residual attack noises, and a detector network is used to determine if the image is clean or being attacked. Similarly, Lin et al. [26] proposed an iterative adversarial retraining strategy to improve model robustness and decrease the effectiveness of adversarial examples on DNN models. For enhanced generalization, the suggested method retrains the model using both Gaussian noise augmentation and adversarial generation strategies. During the testing phase, the model is used to improve the test accuracy, and based on the results, the suggested approach boosts the resilience of the DNN model against several adversarial assaults, such as FGSM, C &W attack, PGD, and DeepFool. Hoyos et al. [27] presented a Hadamard code-based class labeling for training neural network models as a proactive defensive strategy against adversarial attacks. The main argument for employing Hadamard encoding is because of the Hamming distance between vectors of different classes. To alter the class of an output vector, it is important to replace as many bits as possible, which limits the success of adversarial examples. The Hamming distance between any two class vectors is exactly half in the absence of noise. These uncorrupted codes will be used for training, and the error-correcting capabilities will surface during classification. Unfortunately, this method raises the cost and complexity of calculations, and adversarial training has limitations. When confronted with adversarial attacks generated by various methods, performance differs greatly. In contrast to adversarial training, the preprocessing process does not require any changes to the target model, making it easier to apply. Furthermore, it requires less computation and can be utilized in conjunction with a variety of models. For example, Zhang et al. [28] proposed an image reconstruction network that reconstructs an input adversarial example into a clean output image to defend against adversarial attacks by learning a precise mapping of the adversarial examples to the reconstructed examples. In addition, by adding randomization layers to suppress noise, the model greatly reduces the impact of adversarial perturbations while having little influence on the prediction performance of clean images, especially for iterative attacks. In a review paper presented by Kloukiniotis et al., [29] they presented various traditional restoration techniques such as spatial domain filtering and transform domain filtering. While supervised traditional restoration techniques such as wavelet transform and Markov Random Fields provide better results compared to the traditional methods they can still be used as image denoising algorithms in many cases. F. Li et al. [30] presented a novel defense method based on different filter parameters and randomly rotated filtered images. The system aims to reduce the influence of adversarial perturbation by reformulating single-image classification as multiple-image classification through two-stage filtering and image rotation. It improves the defense capability of various models against oblivious adversarial attacks.

With the problem that is mentioned above about the exhaustive training task in adversarial training to mitigate adversarial attacks, preprocessing process techniques are chosen to be used as adversarial defenses to counter the effects of adversarial examples generated using FGSM, C &W, and Watermarking. The aim of utilizing image denoising and image inpainting algorithms is to restore license plate images that are affected by adversarial perturbation caused by the above-mentioned attacks. The implementation of these methods is elaborated in Sect. 2,

and the results are presented in Sect. 3. Table 3 shows the summary of related work for the adversarial defense.

There are few real-world applications and areas that have implemented the approaches of adversarial attacks. In [2], illumination technologies have been used to create several light spots as noises on the license plate as an adversarial example against HyperLPR. Moreover, [20] manufactured real metal objects (plates) to apply the adversarial effect. It shows that physical adversarial objects can fool applied object detection models, including SSD, YOLO, and Faster R-CNN, for license plate detection. On the other hand, none of the mentioned related works have been implemented for adversarial defenses. For our proposed work, the approaches used could include implementing a real-world adversarial attack to mislead actual LPR systems and a defense mechanism to detect these attacks.

2.4 Proposed methods

LPR systems have major benefits in the area of controlling traffic, where accuracy is the main factor for successful recognition in LPR systems. Several Attacks may occur during the process of LPR systems, which affect the accuracy of the system and mislead the recognition. Studying all possible attacks and developing defense strategies for the LPR system is essential to guaranteeing the accuracy of the system in all scenarios. The research considers the most known attacks that could affect the accuracy of the LPR system and finds suitable defense techniques to reduce the impact of the attacks on the system.

2.4.1 Adversarial attacks on LPR system

With all the benefits that the LPR system provides, it still suffers from security issues that may affect the accuracy of the system. For instance, when the system stores the images that have been collected by the road surveillance cameras, those images could be accessed and attacked, where they could be removed or altered to be recognized as a different driver's license plate. For that reason, there is a need to ensure the accuracy of LPR and secure most of the attacks that could affect the system. In the proposed approach, the images of license plates have been attacked to force the OCR to misclassify the characters and come up with different license plates, as shown in Fig. 1. The recognition of the system has been tested using the Fast Gradient Sign Method (FGSM), Carlin & Wagner (C &W) attack, and Watermarking Attack [2, 31, 32]. The Easy-OCR [33] has been utilized in this work as the license plate character recognition system. It is an advanced framework that is empowered with character recognition and deep learning, including CRAFT (Character Region Awareness for Text Detection) algorithm [34], Convolutional Recurrent Neural Network (CRNN) [32], Long Short-Term Memory (LSTM) [35]).

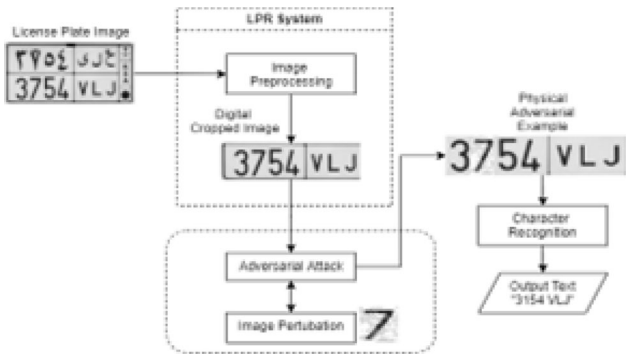


Fig. 1 Adversarial attack on LRP system

2.4.1.1 Fast gradient sign method (FGSM) Fast gradient Sign Method (FGSM) [5] algorithm uses the gradient of the underlying model for generating adversarial examples. FGSM perturbations are added based on the gradient direction to create a new adversarial image that maximizes the loss. This can be summarized using the following equation [31]:

$$\eta = \underline{\epsilon \cdot \text{sign}}(\nabla_X \mathbf{J}(\theta, x, y)) \tag{1}$$

In equation (1), η defined the adversarial perturbation added to the original images to generate the adversarial example. ϵ is the epsilon value that indicates perturbation size multiplied with the signed value from the gradient vector $\nabla xJ(\theta, x, y)$. The gradient vector consists of θ, which indicates the parameters of the model, x is the input image, and y is the true class or label. Also, $\text{sign}(\nabla xJ(\theta, x, y))$ indicates the sign whether it is positive or negative of ∇xJ the gradient of the loss function.

2.4.1.2 Carlini & Wagner (C &W) Carlini & Wagner (C &W) attack [19] is based on minimizing the distance between the original image x and the adversarial image x′ where it misclassifies the system. It finds the minimal value to misclassify the image so that the difference in the distance is less than the threshold of a particular distance metric. The L2 attack is the most used attack of C &W attack where Euclidean distance or L2 norm is used to determine the distance between the original and adversarial images. The equation of L2 that minimizes the distance is shown below:

$$\left\| \frac{1}{2}(\tanh(w) + 1) - x \right\|_2^2 + \text{c.f}\left(\frac{1}{2}(\tanh(w) + 1)\right) \tag{2}$$

where w indicates the magnitude of perturbation applied to x the input image. Also, f is a loss function of this attack and c is a small constant that yields an adversarial image of minimum distance.

2.4.1.3 Image watermarking The utilization of the popularity of watermarking to generate perturbations for the adversarial examples using a pixel transform dyadic function called Linear Blend (LB), to attack the proposed LPR system to misclassify license plate numbers. LB is an image processing function that takes two input images and produces a watermarked image. The generation of adversarial example g(x) is formulated as follows:

$$g(x) = (1 - \alpha)f_0(x) + \alpha f_1(x) \tag{3}$$

To generate the adversarial image $g(x)$, two source images are used: $f_0(x)$, which is the original LP image, and $f_1(x)$ which is a watermark image that will be used as an overlay to the original LP image. By altering the alpha channel, a from $0 - 1$, which corresponds to the transparency of the foreground region $f_1(x)$ concerning the background image $f_0(x)$, a temporal cross-dissolve is applied between $f_0(x)$ and $f_1(x)$. Figure 2 shows the algorithm to apply watermarking using LP to generate adversarial examples [36].

Watermarking attack implementation is divided into two stages. First, is the extraction of $f_0(x)$ the region, which defines the location of the LP numeric characters, upper left and bottom-right coordinates within $f_0(x)$ are defined to specify the exact location of the LP numeric characters. The goal of this stage is to limit the use of the watermarking attack to the $f_0(x)$ region rather than the entire LP image to avoid discovery by observers. The watermarking attack is used in the second stage to create adversarial examples $g(x)$, where $f_0(x)$ the region is combined with the watermark image $f_1(x)$ using LB. Also, at this stage, the alpha channel value is set for $f_1(x)$ to adjust the visibility of the watermarking attack.

2.4.2 Adversarial defense

Image Denoising

Image denoising is a relatively straightforward method for mitigating adversarial perturbations that have received a lot of attention in the field of image processing in recent years. The problem of image denoising can be described mathematically as follows:

$$y = x + n \tag{4}$$

where y is the noisy image viewed, x is the clean image, and n represents the additive white Gaussian noise generated by different methods such as FGSM and C &W attack. Image denoising includes image filtering, which is a common method of image processing. Image noise was reduced using linear filters, but image textures were not maintained. While mean filtering has been used to minimize Gaussian noise, it has the potential to over-smooth noisy images. To address this problem, non-linear filters such as median filtering and weighted median filtering have been developed, which reduce noise without needing any identification. [37, 38]. Similarly, to remove image noise, Xu et al. use two denoising methods, bit-reduction, and image blurring to reduce degrees of freedom and remove adversarial perturbations that can mitigate C &W attack [39] The goal of noise reduction is to improve the image quality in natural images while preserving original features and increasing the signal-to-noise ratio. With this in mind, The paper devised a method to remove image noise on Saudi Arabia (SA) license plates generated using FGSM, C &W, and watermarking by combining the image denoising method using morphological image processing [40] and edge detection using contour filtering. To mitigate adversarial noise on the SA License plate, morphological operations and contour filtering are combined before performing the OCR process to extract LP numbers. Morphological image processing refers to a set of image

Fig. 2 Algorithm 1

Algorithm 1 Adversarial Watermarking Algorithm.

1. **Input:** $f_0(x)$ *and* $f_1(x)$
2. **Output:** $g(x)$
3. $f_1(x)\ height, f_1(x)\ width = f_1(x).shape$
4. $a = 0.50$
5. **foreach** $f_0(x)$ **do**
6. $f_0(x)\ height, f_0(x)\ width = f_0(x).shape$
7. $f_0(x)\ region = f_0(x)[top\ y, bottom\ y], f_0(x)[left\ x, right\ x]]$
8. $g(x) = f_0(x)\ region + af_1(x).shape$
9. **end for**

processing techniques that deal with image shape features to remove image flaws or noise. Contour filtering is used to detect edges and curves joining all continuous points along the LP numbers boundary, all of which have the same color or intensity. Figure 3 shows the process of implementing the image denoising approach.

The proposed method includes several steps, beginning with loading the SA LP image and performing image segmentation to extract the LP image region of interest (ROI) to which the adversarial attack was applied. The extraction of ROI was previously discussed in the adversarial attacks section. After loading the image ROI, it will be converted to grayscale, and image threshold [41] will be used to binarize the image based on pixel intensities. Morphological transformations are normally performed on binary images; as such, this first step is essential. After converting the image into a binary image, morphological image processing [40] is implemented, which is the Opening method to remove image noise from the LP numbers using the OpenCV library [42]. Figure 4 shows the implementation of the proposed image denoising method.

Image denoising is instrumental in fortifying LPR systems against adversarial attacks like FGSM and C &W by enhancing system resilience and preserving the accuracy of license plate recognition. FGSM attacks involve adding small perturbations to input images to deceive the LPR system into misclassifying the output. Image denoising can help combat FGSM attacks by filtering out these added perturbations, restoring the original image to its intended state. By cleaning up the noise introduced by the attack, image denoising ensures that the input image fed into the LPR system is free from distortions that could lead to misinterpretation of the license plate characters. On the contrary, C &W attacks are more sophisticated and aim to find the smallest perturbations that can cause misclassification. Image denoising can be effective in mitigating C &W attacks by removing these subtle perturbations that are crafted to deceive the LPR system. By enhancing the clarity of the input image and eliminating any imperfections introduced by the attack, image denoising helps preserve the integrity of the license plate data during the recognition process. Image denoising filters out noise and perturbations introduced by FGSM and C &W adversarial attacks, thereby improving the quality of input images fed into the LPR system. By restoring the original characteristics of the license plate images, image denoising contributes to maintaining the accuracy of license plate recognition by ensuring that the OCR system receives a clean and unaltered input image, minimizing the impact of FGSM and C &W attacks.

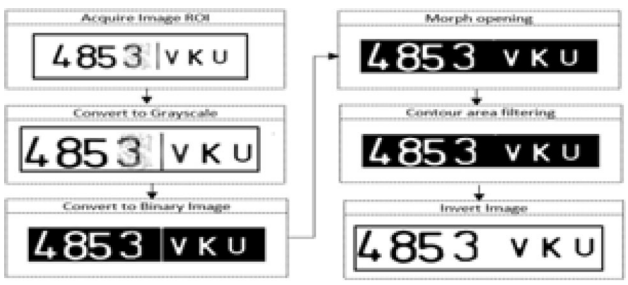


Fig. 3 Image denoising process

Image Inpainting The FGSM and C &W attacks both produce a similar form of perturbation that is not present in the watermarking attack. The generated mask for watermark attack will not work with perturbations generated with FGSM and C &W, and vice versa. With this, the blind inpainting process is used [43] and generates individual image masks for each license plate based on the perturbation applied and location using Algorithm 3 in Fig. 5 and uses the image inpainting algorithm described in [44] to reconstruct the SA license plate image. Image inpainting is considered a superior solution for image reconstruction due to several reasons. Firstly, it leverages contextual information from the surrounding image content to fill in missing or corrupted regions, ensuring the preservation of overall context and structure. Secondly, inpainting techniques strive to achieve visual coherence by seamlessly blending the inpainted regions with the surrounding pixels, resulting in natural and realistic appearances. Thirdly, the fast marching method, a widely used numerical technique, offers an efficient approach to image inpainting, enabling real-time or near-real-time processing. Lastly, image inpainting can serve as a defense against adversarial attacks, as it can be employed to restore the integrity of images by filling in or removing adversarial perturbations [43, 44]. The proposed method for blind image inpainting starts with the creation of an image mask, which will be used to identify which image areas require restoration. Image mask is generated by implementing algorithm 3 using pixel manipulation. Once the image mask is generated, Gaussian blur is used to smoothen the edges of the image mask. Finally, the proposed blind image inpainting process is implemented using the algorithm [3] utilizing both image mask and adversarial image and then inverting the image to restore its original color. The implementation is shown in Fig. 6.

By using image inpainting, the watermarks on the license plate images can be removed or obscured in such a way that the inpainted region looks natural and blends seamlessly with the surrounding areas. This can help in making the watermarks less noticeable or even completely camouflaged. In situations where the watermarks have been strategically placed to disrupt the LPR system's accuracy, image inpainting can be used to restore the damaged areas of the license plate image. By filling in the missing or

Fig. 4 Algorithm 2

Algorithm 2: Image Denoising

Input *images* $\boldsymbol{x}$ *from LP DIR*
Output *reconstructed LP image* ***result***

1. ***for*** i ***in*** x ***do***
2. $gray < -$ ***set*** i ***to*** *grayscale*
3. $thresh < -$ ***get*** *gray threshold*$(0, 255, OTSU)$
4. $kernel < -$ ***set*** *structuringElement*$(thresh)$
5. $opening < -$ ***set*** *morphology*$(thresh, Opening, kernel, iterations = 1)$
6. ***Set*** *cnts* ***to*** *findContours*$(opening)$
7. $cnts = \begin{cases} 0, & length == 2 \\ 1, & otherwise \end{cases}$
8. ***for*** c ***in*** *cnts* ***do***
9. $area < - c$
10. ***if*** $area < 50$:
11. $drawContours(opening, c, -1, 0, -1)$
12. ***end if***
13. ***end for***
14. $blur < -$ ***set*** *medianBlur*$(opening)$
15. $result < -$ *!blur*
16. ***return*** *result*
17. ***end for***

Algorithm 3: Image mask generation

Input : *image* x,
Output : *image* i
pixel threshold $\mu \in \alpha$

1. $\alpha \leftarrow$ ***Get*** x *array pixel values*
2. ***for*** $i \in \alpha$ ***do***
3. $i = \begin{cases} 0, & if\ i > \mu \\ 1, & otherwise \end{cases}$
4. ***end***
5. ***return*** i

Fig. 5 Algorithm 3

Algorithm 4: Image Inpainting

Input *images* $\boldsymbol{x}$ *from LP DIR*
Output *restored LP image* $\boldsymbol{i}$

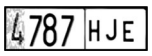
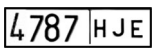
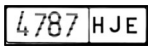
1. ***for*** i ***in*** x ***do***
2. $i < -$ ***Set*** *channels* 0
3. $rows, col < -$ ***Get*** i *shape*
4. ***Get*** $x'_{i,j}$ *using Algorithm* 3
5. $x'_{i,j} < -$ ***Set*** *gaussianBlur*
6. *inpaint* i *using* $i + x'_{i,j}$ [43]
7. $i < -$ *!*i
8. ***return*** i
9. ***end***

Fig. 6 Algorithm 4

altered parts of the license plate image, the LPR system can have access to a clearer and more complete image for accurate recognition. Image inpainting helps maintain the accuracy of the LPR system by ensuring that the inpainted regions do not introduce any additional noise or artifacts that could interfere with the recognition process. The goal is to inpaint the missing or altered areas in a way that preserves the integrity and quality of the image, allowing the LPR system to make accurate predictions based on the inpainted image.

Integrating image denoising and inpainting techniques into existing Optical Character Recognition (OCR) pipelines is essential for ensuring optimal system performance and operational efficiency. In this study, the integration of image denoising and inpainting occurs during the preprocessing stage of the OCR pipeline, a critical phase where the input image is prepared for recognition by enhancing its quality and eliminating noise. By incorporating these techniques at this early stage, the system can work with cleaner and more refined images, leading to improved accuracy in character recognition. The seamless integration of image denoising and inpainting techniques plays a pivotal role in maintaining system efficiency without introducing significant computational overhead. By strategically applying these methods to address noise and missing information in images, the OCR pipeline can operate smoothly and effectively without sacrificing performance. This streamlined approach aligns with the practical implementation needs of real-world License Plate Recognition (LPR) systems, where accuracy and speed are crucial for successful operation in various applications such as traffic monitoring, security, and access control. By embedding image denoising and inpainting within the OCR pipeline, the system can efficiently process images, enhance their quality, and remove unwanted artifacts before the character recognition stage. This proactive approach not only boosts the overall performance of the OCR system but also aligns with the operational requirements of LPR systems in real-world scenarios. The successful integration of these techniques underscores the importance of optimizing system processes to meet the demands of practical applications, ensuring reliable and accurate results in license plate recognition.

Table 4 LPR recognition results after adversarial attacks

Attack	Adversarial Image	Recognition
FGSM Attack		7787 HJE
C& W Attack		1787 HJE
Watermarking Attack		4737 HJE

2.4.3 Threat model

The scenario focused on where attackers generate adversarial samples by adding small perturbations to the license plate number, thereby aiming to cause misclassification. The proposed work assumed that the attacker has access to the LPR system storage where license plate images are stored. Before the recognition process, the stored images could be attacked and modified to misclassify the OCR model that is used to recognize license plate numbers. Moreover, it is assumed that the adversary has access to the target model where the output of the model could be accessed and is aware of the internal information such as gradients. The adversary will be able to specify the pattern of perturbation added to the license plate.

3 Results and discussion

3.1 Adversarial attack on LPR system

Table 4 shows the adversarial images of license plates using FGSM, C &W, and Watermarking attack after generation. Based on the results, the perturbation added to the license plate using the FGSM algorithm shows that segment of number 4 has been perturbed and the LPR recognized it as 7, for C &W attack 4 is classified as 1, while in watermark attack 8 is classified as 3. Also, the adversarial images of license plates are tested based on the position of the number from left to right. Table 5 shows the result for each position for the same plate and the result of recognition for the FGSM attack. On the other hand, some numbers are attacked more than other numbers. For that reason, the recognition of the adversarial images of license plates is tested based on the numbers from 0-9. Table 6 shows the result of recognition for different numbers. Table 7 shows the adversarial attack success rate against the LPR system, and the results show that the Watermarking attack has the highest probability of attacking the LPR system. The dataset consists of 1000 license plate images, and about half of the images have been attacked by watermarking. The results emphasize that Watermarking has the highest amount of perturbation added to the images

Table 5 Recognition results based on the position of applied adversarial attack

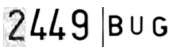
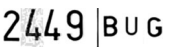
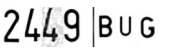
Position	Adversarial Image	Recognition
First		2449 BUG
Second		2049 BUG
Third		2409 BUG
Fourth		2409 BUG

Table 6 Recognition results based on the attack numbers

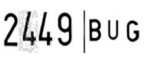
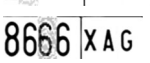
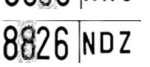
Number	Adversarial Image	Recognition
1		0587 VJX
4		2049 BUG
6		8656 XAG
8		8026 NDZ

Table 7 LPR recognition accuracy against adversarial attack

Adversarial attack	LPR recognition accuracy (%)
Watermarking	46
Carlin & Wagner	78
Fast Gradient Sign Method	72

which affects the quality of the images. Also, Table 8 describes the positions of the numbers and their Accuracy for attacking using adversarial images of license plates. It shows the positions for each attack which are FGSM, C &W, and Watermarking attack. The results show that the third position is the most attacked position for FGSM and Watermarking attacks while the fourth position is the most attacked for C &W. The bold values in Table 8 indicate the character positions that have been identified as the most vulnerable or targeted during FGSM, C&W, and watermarking attacks. These values highlight key findings, emphasizing the positions with the highest success rates for each type of attack. The bold formatting serves to draw attention to critical information and facilitate a quick understanding of the most exploited character positions across different attack strategies. Table 9 shows the attack success rate for numbers 0 to 9. The results show that number 2 is the most attacked for FGSM, where number 4

Table 8 Adversarial attack success rate based on LP number by position

LP number character position	FGSM (%)	C &W (%)	Watermark (%)
First	29	18	28
Second	21	21	33
Third	**31**	24	**39**
Fourth	29	**26**	35

Table 9 Adversarial attack success rate based on LP number by character

LP number	FGSM (%)	C &W (%)	Watermark (%)
0	17	20	44
1	38	18	68
2	**47**	27	56
3	17	17	51
4	37	**40**	55
5	14	20	48
6	33	13	53
7	17	18	47
8	31	23	**69**
9	23	26	53

is the most attacked for C &W and number 8 for Watermarking attack. In this context, the bold values in Table 9 signify statistically significant results, such as the success rates associated with each attack method for the various license plate numbers. This emphasis on bold values serves to highlight key findings that are crucial for interpreting and making decisions based on the data presented.

Moreover, comparing the existing adversarial attack methods with other sophisticated and emerging methods in the aspect of attack success rate at table 10.

3.2 Defense result

The proposed image inpainting and denoising defenses are compared against three adversarial attacks: FGSM, C &W, and watermarking. All these attacks were implemented and applied to the SA license plate images. All images are run into the proposed SA LPR system, and the outcomes are reconstructed using the proposed defense to restore the perturbed images. All the results are based on a 1000 SA license plate image generated using the SA license plate template. Non-targeted adversarial examples are generated for all attacks on SA license plate images, using 1000 clean images of SA license plates to build three sets of adversarial examples for FGSM, C &W, and watermarking attacks. A strong defense must be able to remove adversarial perturbations from adversarial images while also improving clean image classification. Both denoising and blind image inpainting defenses successfully increase LPR recognition accuracy, as illustrated in Table 11. The recognition accuracy for image denoising defense against FGSM attack has increased from 72% to 98%, with only 2% of 1000 photos misclassified. The recognition accuracy for C &W improves from 78% to 99.3%. Finally, when it comes to watermarking attacks, the LPR accuracy has improved significantly, achieving 97.8% recognition accuracy, rising from 46 percent previously. This is a 47% increase in recognition accuracy. The blind image inpainting defense, on the other hand, also exhibits a significant gain in recognition accuracy. The accuracy percentage improved for the FGSM attack, which is essentially identical to the denoising defense, which has a 1.7 percent difference in recognition accuracy percentage. Both image inpainting and denoising defense have a very close result in C &W, with only. 2% difference favoring image inpainting defense. In the case of watermarking, recognition accuracy has increased again, going from 97.8% in denoising defense to 98.85% in image inpainting. Accuracy is typically at the top of the list of requirements when evaluating an LPR system. Unfortunately, not all LPR system research measures accuracy in the same way. Utilizing the following mathematical equation, the

Table 10 Comparing the success rate of proposed methods and other methods

Source	Adversarial attack	Success rate
Proposed attack method	Watermarking	54%
Proposed attack method	C &W	22%
Proposed attack method	FGSM	28%
[17]	Spot evasion attack	3*3 spot: Exceed 70%
[18]	First and second type attack	15% -20%
[2]	Target attack Non-targeted attack	89% 97%

Table 11 Adversarial defense performance

Defense Method	LP Restoration Success Rate	
	Image Denoising	Image Inpainting
C &W	99.3%	99.5%
FGSM	98%	96.31%
Watermarking	97.8%	98.85%

accuracy of the LPR system is measured with the reconstructed images using the proposed defense methods.

$$LPR\ \text{Accuracy} = \frac{\text{No. of recognized reconstructed License Plates}}{\text{Total no. of reconstructed license Plate}} * 100$$

As shown in Table 11, both image denoising and inpainting defense scored a high recognition accuracy percentage, which denotes that the defenses that were applied are effective against the adversarial attacks that are implemented in this research. It can be observed that the defense methods against the C &W attack show the most promising result, reaching 99% recognition accuracy, followed by the Watermarking with an average recognition percentage of 98%, and finally, against FGSM attack with an average recognition percentage of 97%. The perturbation caused by C &W is attributed to the achieved result in C &W attack defense. Since the distance between the real digit and the other digits is calculated, the minimal distance is chosen for adding the least amount of perturbation to the segment of one digit. The limited noise caused by C &W makes the defense more effective because the pixel value manipulation is less compared to FGSM and Watermark attacks. Image inpainting defense is more effective against watermarking attacks. This is because, when compared to image denoising defense, the final image after reconstruction provides cleaner details. The license plate's black and white color palette made it easy to create the mask required for the blind image inpainting defense to determine which area of the plate needed restoration. Even though the results obtained are the lowest among the others in the case of FGSM defense, it is high enough to assert that image denoising and inpainting are effective against adversarial attacks generated by FGSM. The rationale for these results is that the FGSM perturbation is firmly ingrained in the attack region, where the number's pixel value is nearly discernible. Although neither image denoising nor inpainting defense was able to completely remove the image noise caused by FGSM, both image denoising and inpainting defense improved the clarity of the numerals, allowing the LPR system to recognize most of the reconstructed license plate.

3.3 Defense evaluation

The most frequent technique for measuring LPR prediction output is the accuracy measure, in which a match is declared as (1) or a no match (0). However, this does not provide enough granularity to effectively quantify LPR performance, especially the LPR system's OCR model. Performance with objective criteria is needed to be measured, even if it achieves high accuracy scores. Because you can't improve what you don't measure, these metrics are essential for iteratively improving your OCR model. We'll utilize error rates to see how closely the OCR model used in the LPR system classifies the license plate number and the ground truth source to each other. A frequent inclination about error rate is to count the number of unrecognized characters. While this is correct, the actual calculation of the error rate is more complicated since the OCR result may also differ in length from the ground truth text. Furthermore, there are three different types of LPR recognition errors as observed during the experiment, which are presented below, and Table 12 illustrates the LPR recognition error.

- Misclassification: LPR returns incorrect character recognition.
- Unrecognized character: LPR was not able to recognize the character and omit it from the returned result.
- Excessive result: LPR system recognizes some noise/smudge as a valid character and adds it to the returned result for example, holes in the license plate were translated into dot (.).

Character Error Rate (CER) will be used to evaluate the OCR model used in the proposed LPR system against restored images using the proposed image denoising and inpainting defense. The concept of Levenshtein distance [45] is used to calculate CER, which counts the minimal number of character-level operations required to change the reference text into the OCR output. In terms of its use in the LPR system, CER will provide insight into the quality of output generated by an OCR model that is used in license plate recognition. This will give a thorough evaluation of how good the OCR model is its limitations, and what improvements can be made to it. Table 13 shows the result for the proposed LPR system CER.

Based on the results shown in Table 13, the proposed LPR system did an excellent job of recognizing most license plate numbers (English section) following the defense. The average CER of the LPR system after image denoising defense is 0.42%, which implies that the LPR system correctly recognized 99.58% of the license plate characters. In image inpainting, the CER average is 0.56%, with a character recognition success rate of 99.44%.

Table 12 LPR System recognition error

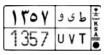
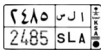
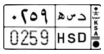
License Plate	Recognition error type and description	Recognition Result
	Unrecognized Character: Number 1 is omitted from the recognition result	357
	Misclassification: Number 8 is misclassified as 0	2405
	Excessive result: A dot (.) is added to the result even if it does not exist with the license plate	025.9

Table 13 Character error rate

Defense method	Character error rate image denoising	Image inpainting
C &W	0.2	0.15
FGSM	0.52	1
Watermarking	0.55	0.55

According to our observations of the LPR recognition output, just one of the above-mentioned LPR recognition errors is causing the misclassification. This indicates that when a misclassification happens in the LPR system, it can only be one of the above-mentioned LPR errors. Furthermore, according to the benchmark established in [46], a good OCR accuracy should only have a CER of 1% to 2%. This suggests that the proposed LPR system implements a good OCR model based on the results shown in Table 13.

4 Conclusion

The integration of Deep Neural Networks in Optical Character Recognition (OCR) and License Plate Recognition (LPR) systems has significantly advanced the accuracy and efficiency of character recognition tasks. However, the susceptibility to adversarial examples poses a notable security challenge, particularly in the context of LPR systems where accurate plate identification is crucial. This vulnerability underscores the necessity of developing defenses to safeguard OCR and LPR systems from adversarial attacks. Intelligent systems that utilize OCR technology have incorporated DL however, adversarial examples that mislead the DL model could cause security issues for the system. In this study, the generated adversarial examples were illustrated using FGSM, C &W, and watermarking, which can be used to attack the LPR system. Moreover, these methods proved that they could cause LPR systems to produce inaccurate recognition. Furthermore, the generated adversarial examples can potentially attack LPR systems without first getting network knowledge, implying a severe security issue on the real LPR system. Based on the research findings, it is evident that the success rate of adversarial attacks on license plates exhibits a notable degree of variability, which is contingent upon both the position and the type of characters that are targeted by the attack. The detailed analysis presented in Tables 8 and 9 illustrates how different positions and character types can impact the success of these attacks. This variability provides attackers with the opportunity to strategically choose the most suitable type of adversarial attack based on the specific position and character they aim to manipulate. By leveraging this information, adversaries can optimize their attack strategies to enhance the likelihood of success and evade detection, underscoring the importance of understanding the nuanced interplay between position, character type, and attack success rates in the context of adversarial attacks on license plates. In addition to investigating adversarial attacks, this study delves into the realm of adversarial defense strategies aimed at thwarting such attacks. The effectiveness of a defense mechanism is intricately tied to the type of adversarial attack being countered, as shown in table 11. Notably, the research findings underscore that image denoising techniques exhibit significant efficacy in mitigating both FGSM and C &W attacks by reducing noise and perturbations, whereas image inpainting emerges as a potent countermeasure against adversarial watermarking by reconstructing missing regions. While the proposed adversarial defense mechanisms prove effective at fortifying LPR system against the array of adversarial attacks scrutinized in this study, the nuanced results underscore that the optimal effectiveness of each defense technique is achieved when customized to combat a specific type of attack. This tailored approach highlights the importance of precision and specificity in the implementation of adversarial defense strategies to enhance resilience against evolving threats within the landscape of adversarial attacks.

The exploration of alternative defense mechanisms, such as image denoising and inpainting, offers a promising avenue to address the security weaknesses inherited from Deep Neural Networks while preserving interpretability, visual fidelity, and practical implementation. These techniques provide transparent and intuitive transformations, effectively mitigating the impact of adversarial perturbations on license plate images. Furthermore, the integration of these techniques into existing pipelines aligns with the operational requirements of real-world LPR systems, offering accessible and practical means of enhancing security without imposing significant computational overhead. By adopting a multi-faceted approach that combines traditional image processing techniques, the research aims

to develop comprehensive and versatile defense strategies tailored to the specific vulnerabilities and requirements of LPR systems. This holistic approach seeks to fortify LPR systems against adversarial threats, fostering increased trust and reliability in the deployment of OCR and LPR technologies across diverse applications. In essence, the pursuit of effective and accessible defenses, rooted in the integration of traditional image processing methods and cutting-edge deep learning approaches, is pivotal in ensuring the resilience and security of OCR and LPR systems in the face of evolving adversarial challenges. This research underscores the importance of addressing security vulnerabilities in advanced machine learning applications and lays the groundwork for enhancing the robustness of OCR and LPR systems in real-world scenarios. Moreover, the methodology could be expanded to be applied to several regions with different license plates.

Acknowledgements The authors extend their appreciation to the Deputy-ship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through project number 1120.

Author contributions CV, SA, NA and AS collected the data and ran experiments, AA, MA, MN, AA, AK, AA and MA supervised the work and contributed to the experiment design.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

1. Feng, R., Mangaokar, N., Chen, J., Fernandes, E., Jha, S., Prakash, A.: Graphite: A practical framework for generating automatic physical adversarial machine learning attacks. arXiv preprint arXiv:2002.08347 (2020)
2. Zha, M., Meng, G., Lin, C., Zhou, Z., Chen, K.: Rolma: A practical adversarial attack against deep learning-based lpr systems. In: Information Security and Cryptology, pp. 101–117 (2020)
3. Mahony, N.O., et al.: Deep learning vs. traditional computer vision. In: International Conference on Image Analysis and Processing (2019). https://doi.org/10.1007/978-3-030-17795-9
4. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
5. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: Reliable attacks against black-box machine learning models (2018)
6. Eykholt, K., et al.: Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1625–1634 (2018). https://doi.org/10.1109/CVPR.2018.00175
7. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1778–1787 (2018). https://doi.org/10.1109/CVPR.2018.00191
8. Ren, K., Zheng, T., Qin, Z., Liu, X.: Adversarial attacks and defenses in deep learning. Engineering **6**(3), 346–360 (2020). https://doi.org/10.1016/j.eng.2019.12.012
9. Chen, L., Xu, W.: Attacking optical character recognition (ocr) systems with adversarial watermarks. In: Proceedings of the IEEE International Conference on Multimedia and Expo (2020)
10. Elsaid, S., et al.: Arabic real-time license plate recognition system. In: International Conference on Intelligent Computing, pp. 126–143 (2019). https://doi.org/10.1007/978-3-030-36368-0_12
11. Chen, L., Sun, J., Xu, W.: Fawa: Fast adversarial watermark attack on optical character recognition (ocr) systems (2020)
12. Akhtar, N., Mian, A., Kardan, N., Shah, M.: Advances in adversarial attacks and defenses in computer vision: a survey. IEEE Access **9**, 155161–155196 (2021). https://doi.org/10.1109/ACCESS.2021.3127960
13. Lin, C.-J., Chuang, C.-C., Lin, H.-Y.: Edge-ai-based real-time automated license plate recognition system. Appl. Sci. **12**(3), 1445 (2022). https://doi.org/10.3390/app12031445
14. Dominguez, D.H.S., Sandoval, S.C.Q., Morocho-Cayamcela, M.E.: End-to-end license plate recognition system for an efficient deployment in surveillance scenarios. In: International Conference on Advances in Computational Intelligence, pp. 697–704 (2022). https://doi.org/10.1007/978-3-030-96293-7_59
15. Alzubaidi, L., Latif, G., Alghazo, J.: Affordable and portable realtime saudi license plate recognition using soc. In: 2019 2nd International Conference on New Trends in Computing Sciences (ICTCS), pp. 1–5 (2019).https://doi.org/10.1109/ICTCS.2019.8923061
16. Antar, R., Alghamdi, S., Alotaibi, J., Alghamdi, M.: Automatic number plate recognition of saudi license car plates. Eng. Technol. Appl. Sci. Res. 12(2), 8266–8272 (2022). https://doi.org/10.48084/etasr.4727
17. Qian, Y., et al.: Spot evasion attacks: adversarial examples for license plate recognition systems with convolutional neural networks. Comput. Secur. **95**, 101826 (2020). https://doi.org/10.1016/j.cose.2020.101826
18. Gu, Z., et al.: Adversarial attacks on license plate recognition systems. Comput. Mater. Contin. **65**, 1437–1452 (2020). https://doi.org/10.32604/cmc.2020.011834
19. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57 (2017). https://doi.org/10.1109/SP.2017.49
20. Yang, K., Tsai, T.-Y., Yu, H., Ho, T.-Y., Jin, Y.: Beyond digital domain: Fooling deep learning based recognition system in physical world. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 1088–1095 (2020). https://doi.org/10.1609/aaai.v34i01.5459
21. Combey, T., Loison, A., Faucher, M., Hajri, H.: Probabilistic jacobian-based saliency maps attacks. Mach. Learn. Knowl.

Extract. **2**(4), 558–578 (2020). https://doi.org/10.3390/make2040030
22. Wang, G., Chen, X., Xu, C.: Adversarial watermarking to attack deep neural networks. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1962–1966 (2019). https://doi.org/10.1109/ICASSP.2019.8682351
23. Lana, J.: Adversarial attack using neural image modification. TechRxiv (2022)
24. Zhu, J., et al.: Attention-guided transformation-invariant attack for black-box adversarial examples. Int. J. Intell. Syst. **37**(5), 3142–3165 (2022). https://doi.org/10.1002/int.22808
25. Yuan, J., He, Z.: Adversarial dual network learning with randomized image transform for restoring attacked images. IEEE Access **8**, 22617–22624 (2020). https://doi.org/10.1109/ACCESS.2020.2969288
26. Lin, J., Njilla, L.L., Xiong, K.: Secure machine learning against adversarial samples at test time. EURASIP J. Inf. Secur. **2022**(1), 1–12 (2022). https://doi.org/10.1186/s13635-021-00125-2
27. Hoyos, A., Ruiz, U., Chavez, E.: Hadamard's defense against adversarial examples. IEEE Access **9**, 118324–118333 (2021). https://doi.org/10.1109/ACCESS.2021.3106855
28. Zhang, S., Gao, H., Rao, Q.: Defense against adversarial attacks by reconstructing images. IEEE Trans. Image Process. **30**, 6117–6129 (2021). https://doi.org/10.1109/TIP.2021.3092582
29. Kloukiniotis, A., Papandreou, A., Lalos, A., Kapsalas, P., Nguyen, D.-V., Moustakas, K.: Countering adversarial attacks on autonomous vehicles using denoising techniques: A review. IEEE Open J. Intell. Transport. Syst. **3**, 61–80 (2022). https://doi.org/10.1109/OJITS.2022.3142612
30. Li, F., Du, X., Zhang, L.: Adversarial attacks defense method based on multiple filtering and image rotation. Discrete Dyn. Nat. Soc. **2022**, 1–11 (2022). https://doi.org/10.1155/2022/6124895
31. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
32. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence PP, 1–1 (2015). https://doi.org/10.1109/TPAMI.2016.2646371
33. EasyOCR.: https://www.jaided.ai/easyocr/. Accessed 01 Jun 2021
34. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. arXiv preprint arXiv:1904.01941 (2019)
35. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**, 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
36. Szeliski, R.: Computer vision. Algorithms and applications. Springer, Washington, p. 793 (2011)
37. Fan, L., Zhang, F., Fan, H., Zhang, C.: Brief review of image denoising techniques. Vis. Comput. Ind. Biomed. Art **2**(1), 7 (2019). https://doi.org/10.1186/s42492-019-0016-7
38. Hambal, A.M., Pei, Z., Ishabailu, F.L.: Image noise reduction and filtering techniques. Int. J. Sci. Res. (IJSR) (2015)
39. Xu, W., Evans, D., Qi, Y.: Feature squeezing mitigates and detects carlini/wagner adversarial examples. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 646–64 (2017). https://doi.org/10.1109/SP.2017.49
40. Goyal, M.: Morphological image processing. Int. J. Comput. Sci. Technol. **2** (2011)
41. Bangare, S., Gera, P., Patil, S.T.P.: Review of otsu's method for image thresholding. Int. J. Emerg. Technol. Adv. Eng. **7**, 128–136 (2017)
42. Morphological Transformations.: https://opencv24-python-tutorials.readthedocs.io/en/latest/py_tutorials/py_imgproc/py_morphological_ops/py_morphological_ops.html#theory. Accessed 01 Sep 2021
43. Hayes, J.: On visible adversarial perturbations & digital watermarking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1597–1604 (2018)
44. Telea, A.: An image inpainting technique based on the fast marching method. J. Graph. Tools **9**(1), 23–34 (2004). https://doi.org/10.1080/10867651.2004.10487596
45. Behara, K.N., Bhaskar, A., Chung, E.: A novel approach for the structural comparison of origin-destination matrices: Levenshtein distance. Transport. Res. Part C **111**, 513–530 (2020). https://doi.org/10.1016/j.trc.2020.01.005
46. Holley, R.: How good can it get? analysing and improving ocr accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine **14**(9/10) (2008)

Conrado Vizcarra Is a lecturer at King Faisal University, College of Computer Sciences and Information Technology, Saudi Arabia. He earned his Bachelor's degree in Computer Science from Universidad De Dagupan, Philippines in 2004, before furthering his education with a Master's degree in Information Technology from the University of the Cordilleras, Philippines in 2010. Currently pursuing his Doctorate in Information Technology at Saint Paul University Philippines. His research interests encompass a wide array of topics including Computer Vision, Machine Learning, Educational Technology, and Systems Development. Through his work, he aims to contribute significantly to the advancement of these fields and make a positive impact on the academic community and beyond.

Shadan Alhamed Is a lecturer at Prince Sattam bin Abdulaziz University (PSAU), AlKharj, Saudi Arabia. She received B.Sc. degree in computer Sciences from King Faisal University (KFU), 2015 and received her M.Sc. in computer Sciences from King Saud University (KSU) 2020. Her research interests include Computer Security and Privacy, Image Processing and Machine Learning.

Abdulelah Algosaibi Received the B.Sc. degree in CS from College of Computer Science and Information Technology (CCSIT), King Faisal University (KFU), in 2008, and the M.Sc. and Ph.D. degrees in CS from Kent State University, Kent, OH, USA, in 2011 and 2015, respectively. He is currently an Associate Professor with the Computer Science Department, CCSIT, KFU. He was in charge of the Center of Documents and Administrative Communication, KFU, 2017-2020, and the Chairman of the Computer Science Department at CCSIT, KFU, 2020-2023. He has earned more than 1.8 million SAR research grants from the Deanship of Scientific Research, KFU, and the Ministry of Education, Saudi Arabia. He is also working on advanced projects in Natural Language Understanding, Deep learning, Ontologies. His research interests include semantic web, ontology engineering, semantic annotation, semantic analysis, knowledge representation, artificial intelligence, and deep learning.

Mohammed Alnaeem Is an Associate professor at King Faisal University (KFU), Al-Ahssa, Saudi Arabia. He earned his B.Sc. degree in CIS from College Of Management Science and Planning, King Faisal University, 2005. He earned his M.Sc. in Networks & Communications (Spec. in Network Security) from Monash University, Australia, 2009 and his Ph.D. degrees in Networks & Communications (Spec. in Wireless Networks & Information Security) from Monash University, Australia, 2015. He is currently the Chairman of Computer Networks and Communications Department, CCSIT, KFU. He is interested in Cybersecurity, Wireless Networks, Artificial Intelligence, and Pattern Recognition.

Adel Aldalbahi Member (IEEE) has made significant contributions to the field of electrical engineering through his academic and research endeavors. He embarked on his academic journey at Virginia Commonwealth University in Richmond, VA, USA, where he earned his B.S. degree in Electrical Engineering in 2011. Demonstrating a strong commitment to advancing his knowledge and skills, he pursued graduate studies at the New Jersey Institute of Technology in Newark, NJ, USA, obtaining both his M.S. and Ph.D. degrees in Electrical Engineering in 2013 and 2017, respectively. Currently, Dr. Aldalbahi serves as an associate professor of Electrical Engineering at King Faisal University in Al-Ahsa, Saudi Arabia. His leadership extends beyond the classroom as he holds the position of Dean of the College of Engineering, where he is instrumental in shaping the educational landscape and fostering innovation within the college. Additionally, he is the Director of the Industrial Relations and Technology Transfer Unit at King Faisal University, a role that underscores his commitment to bridging the gap between academia and industry, promoting collaborative research, and facilitating the transfer of technology for societal benefit. His research interests are at the forefront of technological advancements. His work focuses on visible light communication, millimeter wave networks, deep learning, and machine learning. Through his research, he aims to develop cutting-edge solutions and contribute to the evolution of communication technologies, ensuring they meet the growing demands of modern society.

Nura Aljaafari Received the B.S. degree from King Faisal University (KFU), Hofuf, Saudi Arabia, in 2015, and the M.S. degree from the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, in 2018, both in computer science. She is a faculty member in the College of Computer Science and Information Technology (CCSIT), KFU. Her research interests include deep learning, security and privacy in machine learning, and AI explainability.

Ahmad Sawalmeh (Member, IEEE) Received his B.S. and M.S. degrees in computer engineering from Jordan University of Science and Technology, Jordan, in 2003 and 2006, respectively. He earned his Ph.D. degree in computer and communication engineering from Universiti Tenaga Nasional, Malaysia, in 2020. He is currently a Senior Lecturer in the Software Engineering Department at Alfaisal University. His research interests encompass the application of artificial intelligence and machine learning to support wireless communications and emerging technologies for 5G and 6G networks. His focus lies in aerial communication networks, intelligent unmanned aerial vehicle (UAV) networks, flying ad hoc networks (FANETs), IoT, and device-to-device (D2D) machine communication.

Mahmoud Nazzal Received the B.Sc. degree in electrical engineering from Birzeit University, in 2009, and the M.Sc. and Ph.D. degrees in electrical and electronic engineering from Eastern Mediterranean University, in 2010 and 2015, respectively. He was a Lecturer with the Electrical and Electronic Engineering Department, Eastern Mediterranean University, from 2015 to 2016. He was also a Lecturer with the Izmir University of Economics, from 2016 to 2017. Since July 2017, he has been a Postdoctoral Researcher with Istanbul Medipol University. His research interests include sparse coding, compressive sensing, computer vision, and signal processing for wireless communications.

Abdallah Khreishah (Senior Member, IEEE) Full professor in the Department of Electrical and Computer Engineering at New Jersey Institute of Technology. His research interests fall in the areas of visible-light communication, green networking, network coding, wireless networks, and network security. Dr. Khreishah received his BS degree in computer engineering from Jordan University of Science and Technology in 2004, and his MS and PhD degrees in electrical & computer engineering from Purdue University in 2006 and 2010. While pursuing his PhD studies, he worked with NEESCOM. He is a senior member of the IEEE and the chair of North Jersey IEEE EMBS chapter.

Abdulaziz Alhumam Received his B.Sc. (Computer Information Systems in 2005) from College of Management Science and Planning, King Faisal University (KFU) and M.Sc. (Computer Science in 2009) from Wollongong University, Faculty of Informatics, Australia. He earned his Ph.D. degree in Computer Science from University of York, UK, in 2015. Currently he is working as an Associate Professor in the Computer Science Department, College of Computer Sciences and Information Technology (CCSIT), KFU, Saudi Arabia. He served as a Chairman of the Computer Science Department and Vice Dean of Academic Affairs at CCSIT, KFU. He has published more than 30 papers in international journals and conferences in the field of software engineering and cloud computing. His current research interests include Computer Science, Software Engineering, Machine Learning, Deep Learning, Artificial Intelligence, Pattern Recognition, Internet of Things and 6G.

Muhammad Anan (Senior Member, IEEE) Acting Dean and Associate professor in the Software Engineering Department at Alfaisal University. He received the B.S. degree in computer engineering from King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, in 1997. He obtained the M.S. degree in electrical and computer engineering from the University of Missouri-Columbia, Columbia, MO, USA, and the M.S. degree in software engineering from Kansas University, Lawrence, KS, USA, in 1999 and 2003, respectively. He completed his Ph.D. degree in computer engineering and telecommunications networking from the University of Missouri-Kansas City, Kansas City, MO, USA, in 2008.

Authors and Affiliations

Conrado Vizcarra[1] · Shadan Alhamed[2] · Abdulelah Algosaibi[1] · Mohammed Alnaeem[3] · Adel Aldalbahi[4] · Nura Aljaafari[1] · Ahmad Sawalmeh[5] · Mahmoud Nazzal[6] · Abdallah Khreishah[6] · Abdulaziz Alhumam[1] · Muhammad Anan[5]

✉ Conrado Vizcarra
cvizcarra@kfu.edu.sa

Shadan Alhamed
s.alhamed@psau.edu.sa

Abdulelah Algosaibi
aaalgosaibi@kfu.edu.sa

Mohammed Alnaeem
naeem@kfu.edu.sa

Adel Aldalbahi
aaldalbahi@kfu.edu.sa

Nura Aljaafari
naaljaafari@kfu.edu.sa

Ahmad Sawalmeh
asawalmeh@alfaisal.edu

Mahmoud Nazzal
mahmoud.nazzal@ieee.org

Abdallah Khreishah
abdallah@njit.edu

Abdulaziz Alhumam
aahumam@kfu.edu.sa

Muhammad Anan
manan@alfaisal.edu

[1] Department of Computer Science, CCSIT, King Faisal University, Al Hassa 31982, Saudi Arabia

[2] Applied College, Prince Sattam bin Abdulaziz University, AlKharj, Saudi Arabia