CHARACTERIZATION OF ADVERSARIAL VULNERABILITIES IN NEURAL NETWORK ARCHITECTURES FOR IMAGE RECOGNITION

Fernando Candido Maia[1], Thiago Henrique Santos Fernandes[1], Marcos Monteiro Junior[1,2], Marcella Scoczynski Ribeiro Martins[3], Gabrielly de Queiroz Pereira[1,2]

[1]Departamento de Informática – Universidade Estadual de Ponta Grossa (UEPG)

[2] Departamento de análise e desenvolvimento de sistemas - Serviço Nacional de Aprendizagem Comercial (SENAC)

[3] Departamento de Engenharia Elétrica – Universidade Tecnológica Federal do Paraná (UTFPR)

Maiafernando2611@gmail.com, thsf2000@gmail.com, mmjunior@gmail.com, gqpereira@uepg.br

Abstract. *Convolutional Neural Networks (CNNs) are central to modern computer vision, particularly in image recognition systems deployed in safety-critical scenarios such as traffic and parking management. Despite their high predictive performance under nominal conditions, these models can be severely affected by adversarial perturbations that remain almost imperceptible to the human eye. This paper investigates the adversarial vulnerability of two widely used architectures, DenseNet121 and GoogLeNet (InceptionV3), in the task of parking-space occupancy recognition. Both models were trained and evaluated on a subset of the PKLot dataset, comprising images from three distinct parking lots under different weather conditions. Adversarial examples were generated using the Fast Gradient Sign Method (FGSM) and the Carlini & Wagner (C&W) attack and subsequently presented to the trained models. The experimental results show that, although the original networks achieve high accuracy, precision, and recall on clean data, their predictions degrade substantially under adversarial perturbations, with a marked reduction in correct occupancy estimates. The C&W attack consistently induces stronger performance degradation than FGSM, highlighting critical weaknesses in state-of-the-art CNN architectures when exposed to carefully crafted adversarial inputs.*

1. Introduction

Convolutional Neural Networks (CNNs) have played a central role in the evolution of modern computer vision, becoming the foundation for image classification, semantic segmentation, and object detection since the consolidation of deep architecture in the last decade [Chen et al., 2021]. Studies examining the development of these models show that recent CNN architectures have expanded their applicability to a broad range of visual recognition tasks, reflecting continuous advances in design and training methodologies [Zhao et al.,2024]. Investigations focused on classification robustness indicate that adversarial perturbations, although nearly imperceptible to human observers, can alter the internal representations learned by CNNs, exposing structural weaknesses in the processing pipeline [Li et al., 2023].

Applications of CNNs in automated parking systems and smart city infrastructures highlight the relevance of understanding their operational limitations, as such systems depend on the accurate detection of vehicle occupancy to assist in mobility planning and resource allocation [Kaur et al., 2023]. Urban mobility reports forecast a substantial increase in population density in metropolitan regions, which intensifies the demand for automated solutions capable of supporting traffic flow and vehicle distribution in public and private parking areas [Martinez et al., 2018]. Publicly available

datasets such as PKLot have become essential resources for training models that must recognize occupied and empty parking spaces under varying environmental conditions, reinforcing the importance of high-quality data to system performance [Almeida et al., 2015].

Research in adversarial machine learning demonstrates that CNNs trained for visual recognition may alter their predictions when exposed to carefully crafted modifications in the input space, even when these modifications are visually indistinguishable from the original data [Qi et al., 2021]. Analyses of white-box threat models reveal that gradient based perturbations can influence network outputs in safety sensitive scenarios, raising concerns about the reliability of vision modules embedded in technical systems [Podder and Ghosh, 2024]. In the domain of autonomous vehicles, evaluations of adversarial interactions with perception modules show that manipulated images can disrupt object recognition pipelines, affecting downstream decision-making components involved in navigation and control [Kumar et al., 2020]. Complementary findings in signal classification tasks confirm that similar vulnerabilities appear across different data modalities, as adversarial procedures alter learned decision boundaries in CNNs [Lin et al., 2021].

Recent approaches investigating neuron-level behavior in CNNs have identified spatial and functional regions within deep architectures that exhibit higher susceptibility to adversarial influence, enabling more detailed analysis of internal failure mechanisms [Li et al., 2023]. Other studies propose methodologies for detecting and correcting vulnerable computational nodes, aiming to improve the stability of classification models under adversarial perturbations [Gao et al., 2023]. Multi-objective formulations, such as MoAR-CNN, further demonstrate that balancing standard accuracy and adversarial performance can be explored as an integrated design problem in vision applications [Wei et al., 2025].

Given the increasing presence of CNN-based systems in urban mobility, intelligent transportation, and automated parking management, understanding how adversarial attacks affect occupancy classification is essential for establishing evaluation protocols and ensuring operational reliability [Sujay *et al*., 2019]. Threat models such as the Fast Gradient Sign Method (FGSM) and the Carlini & Wagner (C&W) attack provide a systematic framework for examining how small input perturbations impact model predictions, contributing to ongoing discussions on methodological reproducibility and the ethical implications associated with deploying machine learning systems in real environments [Qi *et al*., 2021]. Ethical considerations require that research involving machine-learning-based perception systems address the potential risks emerging from altered predictions, particularly when such systems can be integrated into public services or mobility infrastructures [Podder and Ghosh, 2024]. In this context, transparency regarding datasets, training procedures, and evaluation methods is aligned with the journal's emphasis on rigor, responsibility, and reproducibility [Chen *et al*., 2021].

Therefore, the present work aims to evaluate the adversarial vulnerability of DenseNet121 and GoogLeNet (InceptionV3) in the task of parking-space occupancy recognition using the PKLot dataset, assessing how FGSM and C&W attacks modify network predictions and discussing the implications of these findings for the deployment of CNN-based visual systems in real-world environments.

2. State of the Art

Convolutional Neural Networks (CNNs) have become the dominant paradigm in image recognition, supported by the consolidation of deep learning techniques and the availability of large-scale annotated datasets [Chen *et al*., 2021]. Although CNNs were proposed more than two decades ago, their impact on practical applications in computer vision became evident only with advances in training methodologies and computational hardware [Zhao *et al*., 2024]. Survey studies indicate that recent progress is not solely due to faster GPUs or larger datasets, but also to architectural innovations that enable deeper networks with more stable optimization behavior [Szegedy *et al*., 2015]. In parallel, analyses at the level of internal representations have highlighted the importance of connectivity patterns and feature reuse to improve gradient flow and representation quality in convolutional architectures [Li *et al*., 2023].

2.1 Convolutional Neural Network Architectures

CNNs are typically organized into convolutional, pooling and fully connected layers, arranged to extract hierarchical feature representations from images and map them to high level semantic classes [Chen *et al*., 2021]. Classical and modern architectures such as VGG, Inception and ResNet were designed to increase network depth while addressing issues related to vanishing gradients and degradation in very deep models [Zhao *et al*., 2024]. VGG networks rely on stacks of small 3×3 convolutions, Inception modules aggregate convolutions of different receptive fields in parallel branches, and ResNet introduces skip connections that allow gradients to propagate more effectively through many layers [Szegedy *et al*., 2015]. These designs have been widely adopted in real-world applications, including medical imaging, intelligent transportation and embedded perception systems [Kaur *et al*., 2023].

Among recent architectures, DenseNet introduces dense connectivity, in which each layer receives as input the feature maps of all preceding layers within the same block, promoting feature reuse and improved gradient propagation [Gao *et al*., 2023]. In a DenseNet with L layers, this design leads to a total of $L(L + 1)/2$ direct connections, increasing connectivity without a proportional growth in the number of parameters [Gao *et al*., 2023]. This connectivity pattern has been associated with more efficient training, better use of parameters and mitigation of common difficulties in very deep networks, such as optimization instability and redundancy in intermediate representations [Wei *et al*., 2025].

GoogLeNet, based on the Inception module, gained prominence after winning the 2014 edition of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), achieving competitive accuracy with substantially fewer parameters compared to earlier architectures such as AlexNet [Szegedy *et al*., 2015]. The Inception module combines convolutions of different kernel sizes and pooling operations within parallel branches, while 1×1 convolutions are used for dimensionality reduction and computational efficiency [Chen *et al*., 2021]. Subsequent analyses show that the 22-layer architecture and the inclusion of auxiliary classifiers contribute to stabilizing optimization and improving convergence in deep models [Zhao *et al*., 2024].

2.2 Adversarial Attacks on CNNs

Despite their performance in image recognition benchmarks, CNNs are vulnerable to adversarial attacks, in which small, carefully crafted perturbations applied to the input image can alter the network prediction while remaining nearly imperceptible to the human eye [Qi *et al*., 2021]. Empirical studies in safety-critical domains, such as perception for autonomous systems and traffic monitoring, demonstrate that such perturbations can compromise the reliability of automated decision pipelines [Podder and Ghosh, 2024]. In the context of autonomous vehicles, black-box attacks have been shown to affect perception modules, altering the recognition of elements in the scene and potentially degrading downstream control decisions [Kumar *et al*., 2020]. Similar findings have been reported in other signal processing tasks, such as modulation recognition, indicating that convolutional models share structural vulnerabilities that can be exploited through adversarial optimization procedures [Lin *et al*., 2021].

Among gradient-based methods, the Fast Gradient Sign Method (FGSM) generates an adversarial example in a single step by perturbing the input in the direction of the sign of the loss gradient with respect to the input [Esmaeilpour *et al*., 2019]. The attack is defined as:

$$\mathbf{x}' = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{w}, \mathbf{x}, y)).$$

where **x** is the original image, ***y*** is the true label, ϵ controls the perturbation magnitude, *J* is the loss function and **w** are the model parameters. Experimental analyses show that FGSM explores the local sensitivity of the decision surface, inducing misclassification even when ϵ is small enough to preserve the perceptual appearance of the image, (ϵ *=0.03)* [Esmaeilpour *et al*.,2019]. Comparative evaluations have established FGSM as a reference white-box attack due to its simplicity and its ability to systematically reveal weaknesses across different CNN architectures [Qi *et al*., 2021].

A second widely studied method is the Carlini & Wagner (C&W) attack, formulated as an optimization problem that seeks perturbations of minimal norm that still cause the model to produce an incorrect prediction [Zhang *et al*., 2021]. The perturbation for the *i*-th sample is quantified by:

$$d_i = \|\mathbf{x}_i - \mathbf{x}_i'\|,$$

measuring the distance between the original image **x***i* and the adversarial image $x_i^{`}$ [Zhang *et al*., 2021]. The optimization problem is written as:

$$\min_c \|\mathbf{d}_i\| + c \times g(\mathbf{x} + \mathbf{d}_i) \quad \text{s.t.} \quad \mathbf{x} + \mathbf{d}_i \in [0,1]^n,$$

where *c* balances the trade-off between minimizing the perturbation norm and enforcing misclassification through the attack loss *g* [Esmaeilpour *et al*., 2019]. Studies comparing FGSM and C&W indicate that C&W typically produces perturbations with smaller perceptual magnitude, while maintaining a high success rate in manipulating model outputs, which has led to its adoption as a benchmark in robustness evaluations [Wei *et al*., 2025].

Overall, the literature on adversarial machine learning converges on the conclusion that adversarial perturbations represent a significant challenge for CNN-based image recognition systems, particularly in applications where decisions depend directly on model predictions [Podder and Ghosh, 2024]. Understanding how distinct architectures such as DenseNet and GoogLeNet respond to specific attack strategies is therefore essential to guide evaluation procedures, mitigation mechanisms and the design of more resilient perception pipelines [Chen *et al*., 2021].

3. Materials and Methods

3.1 Network Training

All neural network models were implemented and trained using the TensorFlow framework [Jingyi et al., 2021]. The task was formulated as a binary classification problem, distinguishing between occupied and free parking spaces. Supervised training was conducted using cross-entropy loss and stochastic gradient descent (SGD) with GPU acceleration. To optimize the training process, fine-tuning was applied using pre-trained weights from ImageNet. The training utilized a learning rate of 0.001, a batch size of 32, and was executed for 100 epochs. The same preprocessing, sampling strategy, and experimental protocol were applied to all architectures to ensure comparability.

3.1.1 Dataset and Sampling

A subset of 19,393 images from the PKLot dataset [Almeida *et al*., 2015] was used in the experiments. The selected portion includes images captured from three parking locations (PUCPR, UFPR04 and UFPR05) and grouped by weather conditions (cloudy, rainy and sunny). The subset was partitioned into training (60%), validation (20%) and test (20%) sets, preserving the distribution across locations and weather conditions. Table 1 summarizes the number of samples used for each parking lot and weather type.

Table 1. Distribution of images by location and weather condition.

Location	**Weather**	**Images**	**Total**
PUCPR	Cloudy	5,098	11,069
	Rainy	2,400	
	Sunny	3,571	
UFPR04	Cloudy	726	2,086
	Rainy	163	
	Sunny	1,197	
UFPR05	Cloudy	880	6,238
	Rainy	80	
	Sunny	5,278	

3.2 DenseNet Configuration

DenseNet121 was selected as one of the target architectures due to its dense connectivity pattern, in which each layer receives the feature maps of all preceding layers within the same block [Xiao *et al.*, 2023]. The model was adapted to a binary classifier by replacing the final fully connected layer with a two-unit output followed by a softmax activation. Figure 1 illustrates the training pipeline used. Input images were resized according to the model specification, normalized and forwarded through the sequence of dense blocks and transition layers. The output probabilities correspond to the predicted occupancy class.

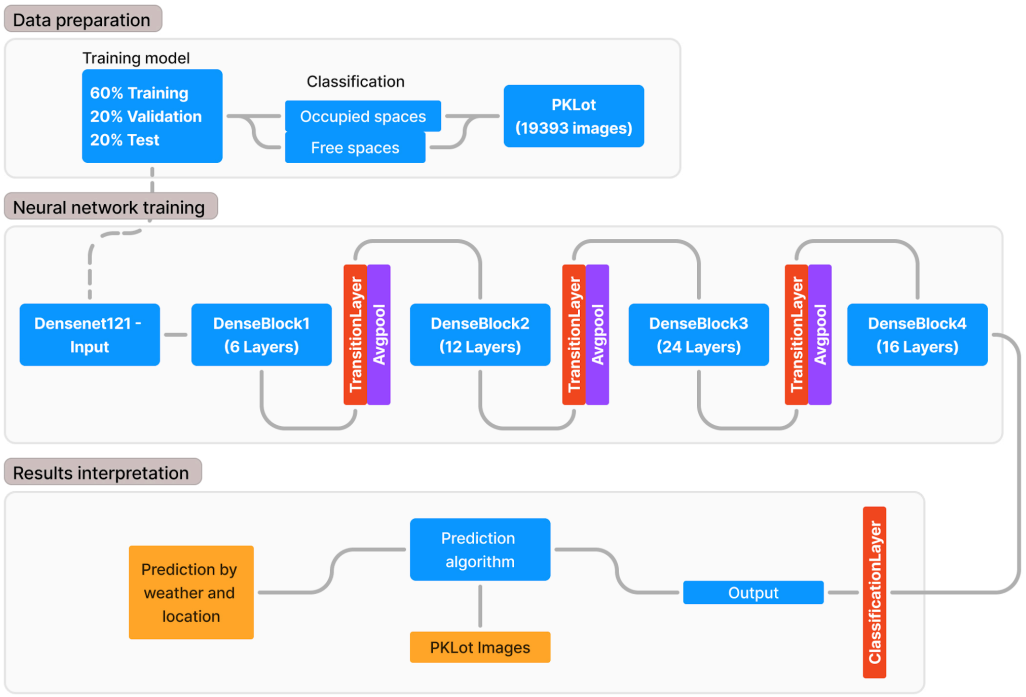

Figure 1. Training pipeline for the DenseNet121 architecture.

Source: The authors

3.3 GoogLeNet / InceptionV3 Configuration

The second architecture evaluated was GoogLeNet implemented via the InceptionV3 variant [Xiao *et al.*, 2023]. The model follows the Inception design, in which convolutions of varying receptive fields are arranged in parallel branches and subsequently concatenated. The final dense layer was replaced to match the binary classification task. Figure 2 shows the training pipeline for this architecture. Convolutional, pooling and Inception modules were used in their standard configuration, followed by global average pooling and a softmax output layer.

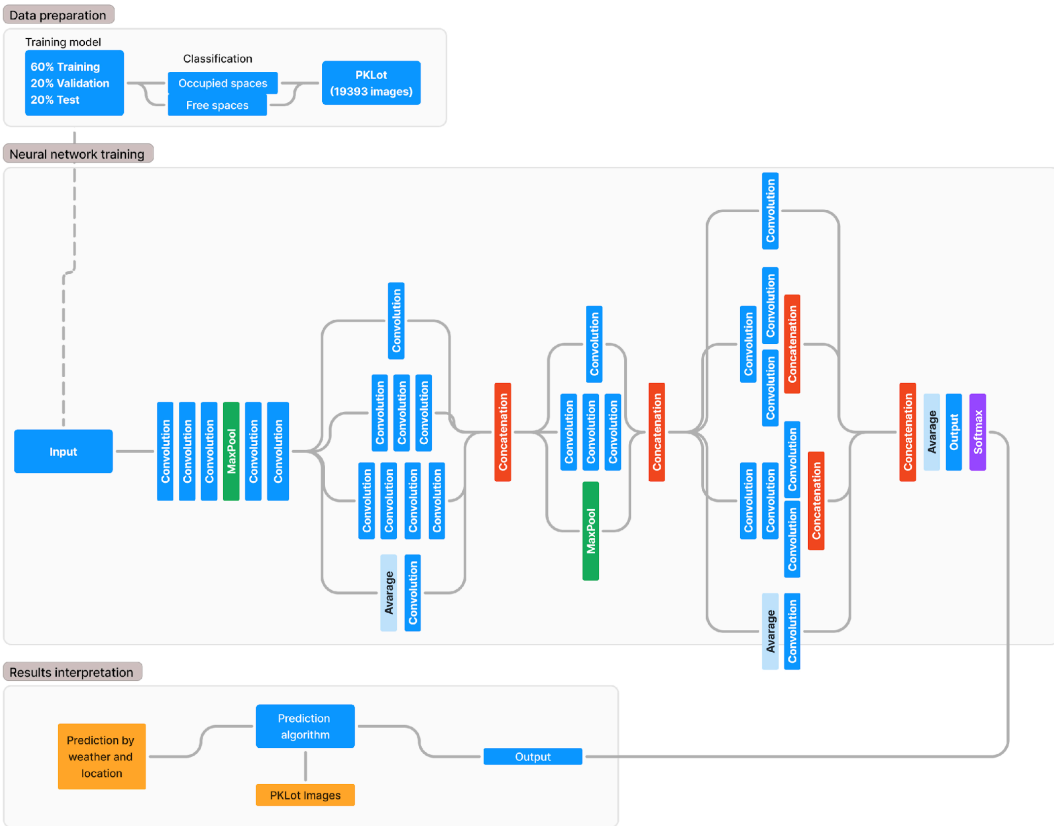

Figure 2. Training pipeline for the GoogLeNet (InceptionV3) architecture.

Source: The authors.

3.4 FGSM and C&W Adversarial Attacks

The adversarial robustness of both architectures was evaluated using the Fast Gradient Sign Method (FGSM) and the Carlini & Wagner (C&W) attack. Both attacks were implemented in white-box settings, using the gradients and parameters of the trained networks. FGSM perturbs each input image according to the sign of the gradient of the loss with respect to the input, scaled by a predefined parameter ϵ [Sujay et al., 2019]. The adversarial image x′ is computed as:

$$\mathbf{x}' = \mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{w}, \mathbf{x}, y)).$$

The C&W attack was implemented as an unconstrained optimization problem solved under box constraints [Zhang *et al.*, 2021]. For each sample, the perturbation $\mathbf{d}i$ is obtained by minimizing:

$$\|\mathbf{d}_i\| + c \cdot g(\mathbf{x}_i + \mathbf{d}_i) \quad \text{s.t.} \quad \mathbf{x}_i + \mathbf{d}_i \in [0,1]^n,$$

where c balances the perturbation magnitude and attack objective [Chen *et al.*, 2024a]. For both FGSM and C&W, adversarial examples were generated only from images in the test set. The clean and adversarial predictions were evaluated using the same models and the same test partitions.

4 Results and Discussion

The adversarial evaluation consisted of applying the Fast Gradient Sign Method (FGSM) and the Carlini & Wagner (C&W) attack to the trained DenseNet121 and InceptionV3 architectures in order to assess the extent to which the adversarial perturbations affected their classification outputs. As described in Sujay *et al.* [2019] and Zhang *et al.* [2021], FGSM and C&W represent two of the most established families of gradient-based and optimization-based adversarial attacks, respectively, making them appropriate for quantifying model vulnerability under controlled conditions. For both networks, the expectation was that the post-attack predictions would exhibit reduced confidence and reduced classification performance when compared to the baseline, following the adversarial robustness trends reported in contemporary literature [Esmaeilpour *et al.*, 2019; Chen *et al.*, 2024b].

Experiments were conducted separately for each weather condition available in the PKLot dataset [Almeida *et al.*, 2015], allowing assessment of how environmental variability interacts with gradient-based perturbations. Since weather conditions directly influence the visual patterns present in the images, they serve as a relevant factor in evaluating the sensitivity of feature extraction under adversarial manipulation. For each combination of parking lot and weather type, the trained models were evaluated using clean test samples, and their performance was compared against adversarially perturbed images.

Table 2 presents the baseline accuracy values for DenseNet121 and InceptionV3 across the three PKLot locations. Before the adversarial attacks, both networks achieved high accuracy in most scenarios, consistent with the expected behavior of deep convolutional architectures trained on large-scale parking lot datasets [Wang *et al.*, 2024]. Notably, the PUCPR subset yielded the highest accuracy for both networks, likely due to the larger number of available samples, as shown previously in Table 1.

Table 2. Accuracy comparison for DenseNet121 and InceptionV3 across locations and weather conditions.

Location	Weather	Accuracy	
		DenseNet121	InceptionV3
UFPR04	Cloudy	0.9655	0.9990
	Rainy	0.9633	0.8947
	Sunny	1.0000	0.9934
UFPR05	Cloudy	0.9843	0.9927
	Rainy	0.9375	0.9903
	Sunny	0.9905	0.9962
PUCPR	Cloudy	0.9990	0.9990
	Rainy	0.9891	0.9717
	Sunny	0.9943	1.0000

Tables 3 and 4 show the corresponding precision and recall values. These metrics further confirm that both networks exhibit strong discriminative capability under no adversarial conditions,

consistent with expected CNN performance in structured binary image classification tasks [Wang *et al*., 2024; Xiao *et al*., 2023]. Performance variations across weather and parking lot subsets reflect dataset imbalance effects previously documented in PKLot-based studies [Almeida *et al*., 2015].

Table 3. Precision comparison for DenseNet121 and InceptionV3.

Location	Weather	Precision	
		DenseNet121	InceptionV3
UFPR04	Cloudy	0.9420	0.9890
	Rainy	0.9350	0.8710
	Sunny	0.9900	0.9850
UFPR05	Cloudy	0.9680	0.9840
	Rainy	0.9230	0.9810
	Sunny	0.9810	0.9890
PUCPR	Cloudy	0.9890	0.9900
	Rainy	0.9740	0.9590
	Sunny	0.9950	0.9950

Table 4. Recall comparison for DenseNet121 and InceptionV3.

Location	Weather	Recall	
		DenseNet121	InceptionV3
UFPR04	Cloudy	0.9600	0.9960
	Rainy	0.9580	0.8860
	Sunny	0.9990	0.9900
UFPR05	Cloudy	0.9790	0.9910
	Rainy	0.9300	0.9880
	Sunny	0.9870	0.9930
PUCPR	Cloudy	0.9950	0.9980
	Rainy	0.9820	0.9660
	Sunny	0.9910	0.9980

Tables 5 and 6 present the predicted occupancy ratios for the clean and adversarial images. These results demonstrate a substantial reduction in classification confidence after both attacks, in line with findings from adversarial literature showing that even imperceptible perturbations significantly alter CNN decision boundaries [Esmaeilpour *et al*., 2019; Zhang *et al*., 2021; Chen *et al*., 2024b]. For all parking lots and weather types, the FGSM and C&W examples caused severe degradation in predicted occupancy, confirming the susceptibility of both architectures to first-order adversaries.

The results indicate that both attacks produced a consistent and substantial reduction in prediction confidence, even in scenarios where both models exhibited near-perfect performance under clean conditions. This aligns with the broader adversarial machine learning literature, which demonstrates that high accuracy in clean datasets does not correlate with robustness against

gradient-based perturbations [Esmaeilpour *et al*., 2019]. Although larger training sets appear to slightly increase model resilience—particularly evident in the PUCPR subset—these improvements are insufficient to prevent degradation caused by adversarial noise, consistent with previous robustness analyses [Zhang *et al*., 2021; Chen *et al*., 2024b].

Table 5. Predicted occupancy ratios for DenseNet121 before and after FGSM and C\&W attacks.

Location	Weather	Clean	FGSM	CW
PUCPR	Cloudy	95.84%	57.04%	44.63%
	Rainy	98.08%	56.90%	43.78%
	Sunny	99.98%	56.65%	43.53%
UFPR04	Cloudy	96.85%	57.03%	45.32%
	Rainy	97.07%	57.21%	45.78%
	Sunny	99.80%	57.15%	46.19%
UFPR05	Cloudy	95.31%	56.88%	44.12%
	Rainy	98.47%	56.91%	44.09%
	Sunny	99.86%	57.12%	44.15%

Table 6. Predicted occupancy ratios for InceptionV3 before and after FGSM and C\&W attacks.

Location	Weather	Clean	FGSM	CW
PUCPR	Cloudy	98.04%	57.34%	55.43%
	Rainy	98.00%	57.11%	55.29%
	Sunny	99.84%	56.61%	56.92%
UFPR04	Cloudy	96.48%	55.93%	44.11%
	Rainy	97.67%	55.71%	43.98%
	Sunny	100%	55.64%	44.60%
UFPR05	Cloudy	97.67%	56.97%	52.40%
	Rainy	99.08%	56.42%	52.27%
	Sunny	99.90%	56.56%	52.35%

To provide a more comprehensive evaluation, additional metrics such as F1-score were considered, corroborating the degradation observed in accuracy. The structural differences between the architectures provide insight into their respective vulnerabilities. DenseNet121, characterized by dense connectivity, aggressively propagates adversarial perturbations through feature reuse within its dense blocks. Conversely, InceptionV3's parallel branches of varying kernel sizes can occasionally act as a localized regularization mechanism, diluting the immediate impact of structural noise compared to strict dense propagation.

Table 7. F1-score

Location	Weather	DenseNet121 (F1-Score)	InceptionV3 (F1-Score)
UFPR04	Cloudy	0.9509	0.9925
	Rainy	0.9464	0.8784
	Sunny	0.9945	0.9875
UFPR05	Cloudy	0.9735	0.9875
	Rainy	0.9265	0.9845
	Sunny	0.9840	0.9910
PUCPR	Cloudy	0.9920	0.9940
	Rainy	0.9780	0.9625
	Sunny	0.9930	0.9965

Overall, the experiments confirm that DenseNet121 and InceptionV3, despite their architectural differences and strong baseline performance, remain highly susceptible to small, imperceptible adversarial perturbations. This reinforces the need for defense mechanisms tailored to safety critical visual recognition domains.

5 Conclusion

This study evaluated the effectiveness of two widely referenced adversarial attack methods the Fast Gradient Sign Method (FGSM) and the Carlini & Wagner (C&W) attack against two convolutional neural network architectures frequently used in image classification tasks: DenseNet121 and InceptionV3. Using the PKLot dataset as an experimental benchmark, the results demonstrated that both models, despite achieving high classification performance under clean conditions, exhibited substantial susceptibility to adversarial perturbations.

Across all scenarios tested, even minimal and visually imperceptible perturbations were sufficient to alter the prediction outputs of both networks, confirming that conventional training pipelines do not inherently provide resilience to adversarial manipulation. The C&W attack consistently produced a greater degradation in prediction confidence than FGSM, aligning with its formulation as an optimization driven method capable of identifying perturbations closer to the model's decision boundaries.

The experiments also indicated modest variations in vulnerability across the subsets of the data set, suggesting that larger training sets can contribute to marginal improvements in stability; however, such improvements were insufficient to prevent misclassification under adversarial conditions. These findings reinforce that the reliability of deep neural networks in operational settings cannot be inferred solely from high baseline accuracy on unperturbed data.

These findings have significant practical implications for the deployment of CNN-based perception modules in smart cities, where adversarial manipulation could lead to incorrect parking management and disrupted traffic flow. Future work will focus on integrating robust mitigation

strategies, such as adversarial training and input denoising mechanisms, to actively increase the resilience of models applied to smart mobility environments.

References

Almeida, J., Oliveira, L. S., Britto Jr, A. S., Silva, E., and Koerich, A. (2015). Pklot – a robust dataset for parking lot classification. *Expert Systems with Applications*, 42(11):4937–4947. DOI: 10.1016/j.eswa.2015.02.009.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., and Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22):4712. DOI: 10.3390/rs13224712.

Chen, L., Zhu, Q.-X., and He, Y.-L. (2024). Adversarial attacks for neural network-based industrial soft sensors: Mirror output attack and translation mirror output attack. *IEEE Transactions on Industrial Informatics*, 20:2378–2386. DOI: 10.1109/TII.2023.3291717.

Esmaeilpour, M., Cardinal, P., and Koerich, A. L. (2019). A robust approach for securing audio classification against adversarial attacks. *IEEE Transactions on Information Forensics and Security*, 15:2147–2159. DOI: 10.1109/TIFS.2019.2956591.

Gao, J., Xia, Z., Dai, J., Dang, C., Jiang, X., and Feng, X. (2023). Vulnerable point detection and repair against adversarial attacks for convolutional neural networks. *International Journal of Machine Learning and Cybernetics*, 14:4163–4192. DOI: 10.1007/s13042-023-01888-5.

Jingyi, Y., Rui, S., and Tianqi, W. (2021). Classification of images by using tensorflow. *2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pages 622 – 626. DOI: 10.1109/ICSP51882.2021.9408796.

Kaur, R., Roul, R., and Batra, S. (2023). A hybrid deep learning cnn-elm approach for parking space detection in smart cities. *Neural Computing and Applications*, 35:13665– 13683. DOI: 10.1007/s00521-023-08426-y.

Kumar, K., Chalavadi, V., Mitra, R., and Mohan, C. (2020). Black-box adversarial attacks in autonomous vehicle technology. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7. DOI: 10.1109/aipr50011.2020.9425267.

Li, Y., Wang, J., Fujiwara, T., and Liu K. (2023). Visual analytics of neuron vulnerability to adversarial attacks on convolutional neural networks. *ACM Transactions on Interactive Intelligent Systems*, 13(4):1–26. DOI: 10.1145/3587470.

Lin, Y., Zhao, H., X., u., Tu, Y., and Wang, M. (2021). Adversarial attacks in modulation recognition with convolutional neural networks. *IEEE Transactions on Reliability*, 70:389–401. DOI: 10.1109/TR.2020.3032744.

Martinez, J., Zoeke, D., and Vossiek, M. (2018). Convolutional neural networks for parking space detection in downfire urban radar. *International Journal of Microwave and Wireless Technologies*, 10:643–650. DOI: https://doi.org/10.1017/S1759078718000466.

Podder, R. and Ghosh, S. (2024). Impact of white-box adversarial attacks on convolutional neural networks. In *2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*, pages 1–9. DOI: 10.1109/etncc63262.2024.10767521.

Qi, L. *et al*. (2021). Adversarial attacks and defenses in images: A survey. *Computers & Security*, 106:102268. DOI: 10.1016/j.cose.2021.102268.

Sujay, S. *et al*. (2019). Adversarial examples in modern machine learning: A review. *IEEE Access*, 7:165689–165702. DOI: 10.1109/ACCESS.2019.2948692.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. DOI: 10.1109/CVPR.2015.7298594.

Wang, Y., Sun, W., Jin, J., Kong, Z., and Yue, X. (2024). Wood: Wasserstein-based out-of-distribution detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:944–956. DOI: 10.1109/TPAMI.2023.3328883.

Wei, H.-N., Zeng, G., Lu, K., Geng, G., and Weng, J. (2025). Moar-cnn: Multi-objective adversarially robust convolutional neural network for sar image classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9:57–74. DOI: 10.1109/TETCI.2024.3449908.

Xiao, Y., Yan, C., Lyu, S., Pei, Q., Liu, X., Zhang, N., and Dong, M. (2023). Defed: An edge-feature-enhanced image denoised network against adversarial attacks for secure internet of things. *IEEE Internet of Things Journal*, 10:6836–6848. DOI: 10.1109/JIOT.2022.3227564.

Zhang, S., Gao, H., and Rao, Q. (2021). Defense against adversarial attacks by reconstructing images. *IEEE Transactions on Image Processing*, 30:6117–6129. DOI: 10.1109/TIP.2021.3092582.

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., and Parmar, M. (2024). A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57:99. DOI: 10.1007/s10462-024-10721-6