This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Traffic Sign Recognition Under Visual Perturbations: Shadows, Light Patches, and Simulated Obstructions

Muneeb A. Khan, Yujin Choi, Jiho Eum, Heemin Park
Department of Software, Sangmyung University, South Korea
{muneebkhan046, yj061628, jihoeum4}@gmail.com, heemin@smu.ac.kr

Abstract

Traffic sign recognition (TSR) systems are essential for autonomous driving yet remain vulnerable to adversarial attacks that mimic natural phenomena. These vulnerabilities pose significant safety concerns for autonomous vehicle deployment, necessitating a thorough investigation of potential attack vectors. In this paper, we investigate three distinct digital perturbation vectors: shadow manipulations, localized light patches, and simulated obstructions—all designed to maintain physical plausibility while degrading classifier performance. Unlike conventional techniques requiring physical access or producing conspicuous artifacts, our digital perturbations integrate seamlessly with the visual environment while achieving high attack success rates (ASRs). We evaluate these attacks across specialized TSR models and general-purpose CNNs on both GTSRB and PTSD datasets, enabling direct comparison between attack methodologies. Our findings reveal that CNN architectures incorporating spatial transformers provide the most vigorous defense, particularly against illumination-based perturbations. By employing targeted data augmentation and feature importance-based countermeasures, we provide actionable insights for developing more resilient TSR systems against these naturalistic adversarial attacks.

1. Introduction

Convolutional Neural Networks (CNNs) have demonstrated remarkable efficacy in traffic sign recognition (TSR) systems—a critical component for autonomous driving [1, 14]. However, their susceptibility to adversarial visual perturbations presents significant safety risks in deployment scenarios [7, 24].

Adversarial attacks on CNNs can be classified into digital attacks (perturbations applied directly to image pixels) and physical attacks (modifications introduced into the real-world environment) [4]. While physical attacks such as adhesive stickers [3, 13] effectively demonstrate model vulnerabilities, they necessitate physical access and often produce visually conspicuous artifacts that compromise operational efficacy [33].



(a) Original (b) Shadow (c) Light Patch



(d) Original (e) Shadow (f) Light Patch

Figure 1. Naturalistic digital perturbations: original images (a,d) compared with shadow (b,e) and light patch (c,f) attacks on traffic signs.

These limitations necessitate the investigation of naturalistic digital perturbations that simulate environmental phenomena such as lighting variations, shadows, and occlusions [10, 35]. Unlike conventional pixel-level perturbations, these attacks integrate with the visual environment, substantially increasing detection difficulty while maintaining effectiveness. Figure 1 demonstrates this phenomenon, illustrating how digital lighting manipulations preserve visual plausibility while compromising recognition systems.

A persistent challenge remains in ensuring digital adversarial perturbations effectively transfer to physical environments. Current approaches exhibit limited generalization beyond controlled settings due to variations in illumination, viewing angle, and environmental noise [5, 6, 12]. Domain adaptation and adversarial training methods [16, 28] require substantial computational resources yet demonstrate limited cross-scenario efficacy.

In this study, we systematically evaluate three distinct digital-domain adversarial attack vectors targeting TSR systems: (1) lighting variations that modify object appearance, (2) localized light patches that affect classification without explicit tampering, and (3) simulated obstructions that approximate realistic occlusions such as graffiti and vandalism. Our framework enables direct comparative analysis of TSR vulnerabilities across perturbation types under consistent evaluation parameters.

Our main contributions are:

1. We present a systematic framework for evaluating adversarial attacks on traffic sign recognition systems, addressing two critical research gaps: (1) decoupling model architectures from their typical datasets and (2) comparing both inconspicuous and visible attack settings.
2. We conduct the comprehensive comparison of three physically plausible attack vectors (shadows, light patches, and simulated obstructions) under standardized parameters, demonstrating that obstruction-based attacks consistently outperform illumination-based methods by 10-19% across architectures.
3. We perform extensive cross-architectural vulnerability assessment across seven CNN architectures (three specialized TSR models and four general-purpose networks), revealing significant variations in inherent robustness with scores ranging from 23.4% to 44.3%.
4. We provide empirical evidence that CNN architectures incorporating spatial transformers exhibit superior resilience against naturalistic perturbations (45.6% ASR vs. 72.6% for standard CNNs), with particularly strong performance against illumination-based attacks.
5. We validate our findings across two geographically diverse datasets (GTSRB and PTSD), demonstrating consistent vulnerability patterns (average ASRs of 66.7% and 67.4%), confirming these weaknesses represent fundamental limitations in CNN visual processing rather than dataset artifacts.
6. We develop and validate effective countermeasures that reduce attack success rates by 36.5% on CNN-STN architectures and 39.4% on average across all models through targeted data augmentation and architectural modifications, providing insights for robust TSR system design.

2. Related Work

2.1. Adversarial Attacks on Traffic Sign Recognition

TSR systems are vulnerable to adversarial perturbations that induce misclassification in safety-critical contexts. Initial work by Szegedy et al. [24] demonstrated imperceptible digital perturbations, while Eykholt et al. [4] extended these vulnerabilities to physical implementations. TSR vulnerabilities span both digital perturbations [7, 16] and physical modifications [4, 8].

Yang et al. [30] achieved higher success rates with lower perturbation magnitudes using targeted attention mechanisms. However, conventional noise-based approaches produce visually conspicuous artifacts, driving research toward naturalistic perturbations that mimic environmental phenomena [10, 32].

2.2. Shadow-Based Attacks

Shadow-based attacks exploit natural illumination phenomena to compromise TSR systems. Zhang et al. [32] achieved 89% attack success rates (ASR) using physically plausible adversarial shadows against standard TSR models. Lovisotto et al. [15] demonstrated 92% ASR against object detectors through dynamic shadow casting while maintaining a natural appearance.

Sitawarin et al. [22] developed DARTS with significant transferability to autonomous driving systems, while Sayles et al. [20] utilized infrared illumination to bypass human detection. Xie et al. [29] demonstrated effectiveness across multiple perception tasks including semantic segmentation and object detection. Implementation challenges include variable ambient lighting, complex reflectance properties, and maintaining effectiveness across viewpoints [15, 22, 29].

2.3. Localized Light Patch Attacks

Localized light patches target specific regions with concentrated illumination. Duan et al. [3] demonstrated that laser projections induce classification errors while remaining visually inconspicuous, and Gnanasambandam et al. [6] utilized projector-based systems to generate effective dynamic light patterns.

Tu et al. [26] achieved high transfer rates between simulation and physical implementation by adhering to physical lighting constraints. Zhang et al. [32] demonstrated that concentrated light effects constrained by physical reflectance models achieve significantly higher ASRs compared to uniform perturbations.

Current optimization techniques primarily employ gradient-based approaches without accounting for sign-specific semantics [3, 10, 26]. This presents a particular challenge for TSR systems where discriminative features vary across classes [10, 33], necessitating semantic-aware optimization techniques.

2.4. Simulated Obstruction-Based Attacks

Obstruction-based attacks simulate real-world occlusion phenomena. Eykholt et al. [4] demonstrated that strategically positioned adhesive materials induce misclassification while preserving perceptual integrity. Zolfi et al. [35]



(a) Shadow (b) Light Patch (c) Obstructions

(d) Shadow (e) Light Patch (f) Obstructions

Figure 2. Adversarial attack vectors applied to traffic signs from GTSRB (top) and PTSD (bottom) datasets: (a,d) shadow manipulations, (b,e) light patches, and (c,f) simulated obstructions on GTSRB (top) and PTSD (bottom).

showed that even partially transparent simulated environmental deposits significantly degrade detector performance.

Digital simulation techniques have advanced through GANs for creating realistic occlusions mimicking mud splatters and water stains [10], and physically-based rendering frameworks for simulating adherent particles [8]. These attacks derive effectiveness from semantic alignment with environmental expectations, making them particularly challenging to detect.

Quantitative evaluations demonstrate that current TSR systems exhibit particular vulnerability to naturalistic obstructions, with performance degradation of up to 70% under simulated environmental conditions [30, 35]. This highlights critical robustness gaps in systems optimized for ideal conditions rather than partial occlusions.

3. Adversarial Attack Approaches

This section presents three distinct adversarial attack strategies targeting traffic sign recognition systems. Each approach maintains perceptual realism while degrading classification accuracy, enabling systematic vulnerability assessment. Figure 2 illustrates examples of these attack methodologies applied to traffic signs from the GTSRB [23] and PTSD [19] datasets.

3.1. Attack Framework Overview

Our framework systematically generates physically plausible perturbations targeting distinct visual properties of traffic signs. Unlike conventional adversarial examples that employ arbitrary pixel-level noise, we focus on perturbations that simulate naturally occurring phenomena: illumination variations, localized light effects, and physical obstructions.

Given an input image $I \in [0,1]^{H\times W\times 3}$ with true label y, and classifier $f : I \to \mathbb{R}^K$ (where K is the number of classes), our objective is to find a perturbation that induces misclassification:

$$\text{find } I' \text{ such that } \underset{k}{\text{argmax}} f(I')_k \neq y \text{ and } \mathcal{D}(I, I') < \epsilon \quad (1)$$

where $\mathcal{D}$ measures perceptual deviation and ϵ represents a plausibility threshold.

3.2. Particle Swarm Optimization (PSO)

We employ Particle Swarm Optimization (PSO) across all attack vectors to navigate parameter space efficiently without requiring gradient access. The optimization initializes a swarm of n candidate parameter sets $\{\theta_i\}_{i=1}^n$, where each particle maintains position θ_i, velocity v_i, and tracks optimal positions (individual p_i and global g), which can be defined as:

$$v_i^{t+1} = \omega v_i^t + c_1 r_1 (p_i - \theta_i^t) + c_2 r_2 (g - \theta_i^t) \quad (2)$$

$$\theta_i^{t+1} = \theta_i^t + v_i^{t+1} \quad (3)$$

where $\omega = 0.7$ is the inertia weight, $c_1, c_2 \in [1.4, 1.8]$ are cognitive and social parameters controlling exploration vs. exploitation, and $r_1, r_2 \sim U[0,1]$ are random values drawn from a uniform distribution.

After each iteration, parameters are constrained to maintain physical plausibility:

$$\theta_i^{t+1} = clip(\theta_i^{t+1}, \theta_{\min}, \theta_{\max}) \quad (4)$$

where *clip* ensures parameters remain within physically realistic bounds.

3.3. Shadow-Based Attacks

Illumination variations represent a potent attack vector due to their natural occurrence and potential for remote deployment. Studies show natural light conditions can mislead vision models, with reflections causing STOP signs to be misinterpreted 70% of the time [10, 17].

3.3.1. Perturbation Parameterization

We model illumination perturbations as elliptical regions defined by parameter vector $\theta = (c_x, c_y, a, b, \phi, \alpha, \beta)$, where (c_x, c_y) denotes center coordinates, (a, b) represents major and minor axes lengths, ϕ indicates rotation angle, α controls intensity magnitude, and β determines gradient falloff.

The elliptical distance function is:

$$d_{ell}(i,j) = \frac{((i-c_y)\cos\phi - (j-c_x)\sin\phi)^2}{b^2} + \frac{((i-c_y)\sin\phi + (j-c_x)\cos\phi)^2}{a^2} \tag{5}$$

We formulate the perturbation mask $M(i,j;\theta)$ as a parameterized function:

$$M(i,j;\theta) = \alpha(1 - d_{ell}(i,j))_+^{\beta} \cdot S(i,j;\phi) \tag{6}$$

where $(\cdot)_+$ denotes $\max(0,\cdot)$, and $S(i,j;\phi) \in [0,1]$ is a smoothness regulator that enforces physical plausibility constraints by ensuring natural transitions along illumination boundaries.

The parameter space is bounded by physical constraints: $c_x \in [0,W]$, $c_y \in [0,H]$, $a,b \in (0,\min(W,H)/2]$, $\phi \in [0,2\pi)$, $\alpha \in [0.4,0.9]$, $\beta \in [1.5,3.0]$.

These values were established through a combination of perceptual studies and physically-based rendering models [34].

Light-based attacks increase pixel intensities:

$$I'_{light}(i,j) = \min((1 + M(i,j)) \cdot I(i,j), 1) \tag{7}$$

while shadow-based attacks decrease intensities:

$$I'_{shadow}(i,j) = (1 - M(i,j)) \cdot I(i,j) \tag{8}$$

This parameterization aligns with physically-based rendering principles [34], as the elliptical modeling approximates real-world illumination patterns from directional light sources. Physical plausibility constraints enforce $0 \leq M(i,j) \leq 1$ and maintain realistic lighting gradients through parameter bounds during optimization.

3.4. Light Patch Attacks

Light patch attacks target specific sign regions with concentrated high-intensity illumination. These attacks combine minimal visibility with maximum impact, enabling temporary deployment without permanent evidence.

3.4.1. Patch Parameterization

We define a light patch through parameter vector $\theta_P = (u_0, v_0, r, \kappa, \gamma)$, where (u_0, v_0) represents the center coordinates, r denotes radius, κ controls intensity magnitude, and γ determines intensity decay rate.

The Euclidean distance function measures pixel distance from the patch center:

$$d(i,j) = \sqrt{(i-u_0)^2 + (j-v_0)^2} \tag{9}$$

The patch region P is explicitly defined as:

$$P = \{(i,j) \mid d(i,j) \leq r\} \tag{10}$$

The intensity profile follows exponential radial decay:

$$L(i,j) = \kappa \cdot \exp\left(-\frac{d(i,j)^2}{r^2 \cdot \gamma}\right) \tag{11}$$

The resulting adversarial image integrates this light effect:

$$I'(i,j) = \begin{cases} \min(I(i,j) + L(i,j), 1), & \text{if } (i,j) \in P \\ I(i,j), & \text{otherwise} \end{cases} \tag{12}$$

Physical plausibility constraints limit patch size ($r \leq \min(h,w)/3$), intensity ($0.2 \leq \kappa \leq 1.0$), and decay rate ($0.5 \leq \gamma \leq 10.0$). These constraints were determined through optical modeling of specular reflections and validated in physical experiments with reflective surfaces [26]. The exponential decay profile was selected to match real-world light scattering behavior on reflective materials commonly used in traffic signs.

3.5. Simulated Obstruction Attacks

Physical obstructions directly modify sign content through partial occlusion. These modifications—including graffiti and vandalism—compromise classifier performance while appearing natural or accidental.

3.5.1. Obstruction Parameterization

We implement a graffiti-type simulated obstruction with parameter vector $\theta_G = (n_l, w_l, C, \{p_i\}_{i=1}^{2n_l})$, where n_l represents line segments (3-6), w_l denotes line width (1-4 pixels), $C \in [0,1]^3$ specifies RGB color, and $\{p_i\}$ are control points for line segments. The obstruction effect applies through alpha blending:

$$I' = I \odot (1 - M_Q) + O \odot M_Q \tag{13}$$

where $\odot$ represents element-wise multiplication, $M_Q \in [0,1]^{H \times W}$ denotes the obstruction mask, and $O \in [0,1]^{H \times W \times 3}$ represents the obstruction pattern. For partial transparency effects:

$$I'(i,j) = \tau O(i,j) + (1-\tau) I(i,j), \quad \forall (i,j) \in Q \tag{14}$$

where $\tau \in [0,1]$ controls opacity level, optimized alongside spatial parameters.

Unlike gradient-based approaches, our implementation employs an evolutionary strategy that mutates obstruction parameters based on fitness evaluation. The evolutionary process follows:

$$\theta_G^{t+1} = \theta_G^t + \delta^t \odot \mathcal{N}(0, \sigma^t) \tag{15}$$

with temperature-dependent mutation probability:

$$P(\delta_i^t = 1) = \frac{\lambda}{1 + \exp(f(\theta_G^t)/T^t)} \tag{16}$$

where δ^t is the mutation mask, $\lambda \in [0.3, 0.6]$ is the base mutation rate, and T^t decreases with each iteration. This approach generates scribble patterns that reflect statistical properties of real-world traffic sign vandalism, based on analysis of urban graffiti patterns [35]. The temperature schedule gradually restricts mutations as optimization progresses, balancing exploration in early iterations with refinement in later stages.

4. Experiments and Evaluation

This section details our experimental methodology for systematically evaluating the efficacy of three naturalistic adversarial attack vectors against traffic sign recognition systems. Our framework enables quantitative assessment of both model vulnerabilities and attack potency across diverse architectures and datasets.

4.1. Datasets

We employ two traffic sign datasets in our evaluation: the German Traffic Sign Recognition Benchmark (GTSRB) [23] and the Persian Traffic Sign Dataset (PTSD) [19]. GTSRB comprises 51,922 images spanning 43 distinct traffic sign classes, including speed limits, prohibitory signs, and warning signs. This benchmark has become a standard reference for evaluating TSR systems due to its comprehensive coverage and real-world variability. PTSD consists of over 14,000 traffic sign images across 43 classes, with 2,421 images designated for testing. This dataset provides geographical diversity to our evaluation framework, enabling the assessment of cross-dataset generalization capabilities.

4.2. Model Architectures

To address the limited scope of previous adversarial studies that typically evaluate either specialized TSR models [2, 10, 33] or generic classification architectures [27, 31] in isolation, we conduct a comprehensive cross-architectural analysis. Our evaluation encompasses three specialized TSR architectures (Figure 3) and four state-of-the-art general-purpose models:

1. **$\text{CNN}_{\text{small}}$**: A compact CNN architecture (739K parameters) proposed by Eykholt et al. [4], comprising three convolutional layers with max-pooling and dropout regularization. We extend it for the PTSD dataset.
2. **$\text{CNN}_{\text{large}}$**: The original GTSRB-CNN architecture is based on a multi-scale convolutional architecture (16.5M parameters) [21], employing parallel processing branches that integrate features at multiple abstraction levels. We adopted the implementation of Yadav[1] and extended it for the PTSD dataset.
3. **CNN-STN**: Garcia et al. [1] proposed a CNN base architecture incorporating Spatial Transformer Networks (855K parameters) to address geometric variations through learned spatial transformations. We extended it for the PTSD dataset.

[1] https://github.com/vxy10/p2-TrafficSigns

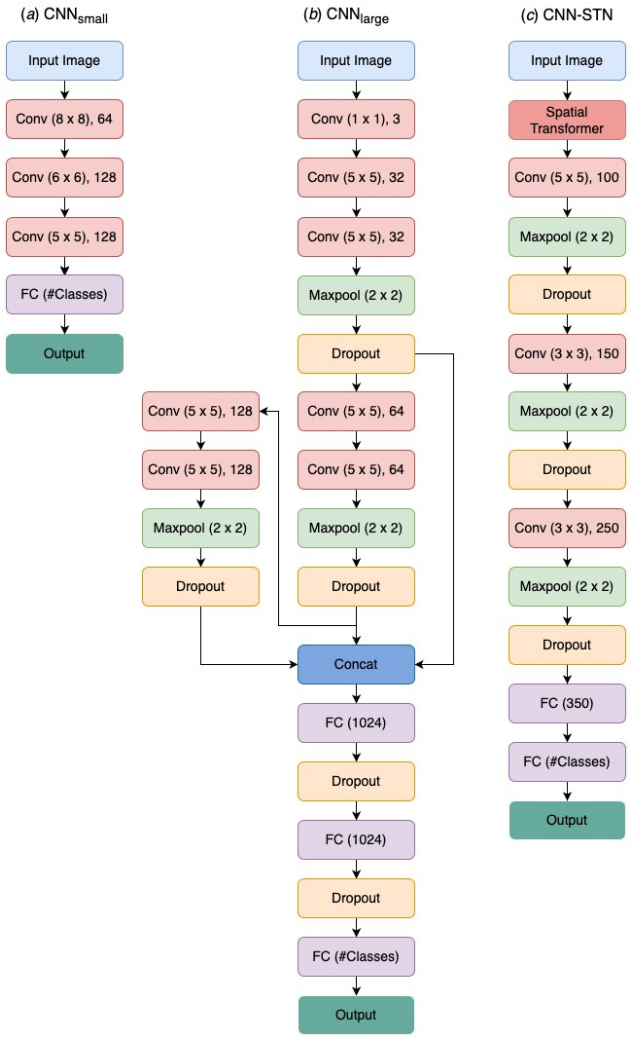


Figure 3. **Baseline TSR Model Architectures**: (a) $\text{CNN}_{\text{small}}$ (739K parameters): compact model with three conv layers. (b) $\text{CNN}_{\text{large}}$ (16.5M parameters): multi-scale CNN with parallel branches, and (c) CNN-STN (855K parameters): incorporates Spatial Transformer Network for distorted signs.

For comprehensive evaluation, we include general-purpose CNN architectures which have demonstrated state-of-the-art performance across diverse computer vision tasks: ResNet50v2 [9] with residual connections, EfficientNetv2-B3 [25] optimized for computational efficiency, DenseNet-121 [11] featuring dense connectivity patterns, and MobileNetv2 [18] designed for mobile applications. This diverse architectural spectrum facilitates analysis of structural vulnerability factors across varying network topologies, parameter scales, and design philosophies.

4.3. Experimental Setup

All images are preprocessed to 32×32 pixels with normalization to the [0,1] range. Models are trained using an 80:20 train-validation split with the Adam optimizer (learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$) and mini-batch training (batch size 64). To enhance generalization capabilities, we employ label-smoothed categorical cross-entropy loss (smoothing factor 0.1) with model checkpointing to preserve optimal weights.

For each attack type—shadow-based, light patch, and simulated obstruction—we vary parameters to assess vulnerability patterns. Elliptical regions with varying positions, sizes, and intensities parameterize light and shadow attacks. Light patch attacks utilize circular regions with different radii and brightness levels. Simulated obstruction attacks employ synthetic obstruction patterns with controlled opacity and placement.

To ensure comprehensive evaluation, we assess each model's performance on clean test images to establish baseline performance metrics, followed by the systematic application of optimized attack vectors. This methodology enables granular analysis of architecture-specific vulnerabilities and attack-specific efficacy patterns across traffic sign categories.

4.4. Evaluation Framework

We employ quantitative metrics that capture both absolute vulnerability and normalized robustness characteristics:

4.4.1. Attack Success Rate (ASR)

We quantify attack efficacy through the Attack Success Rate:

$$\text{ASR} = \frac{N_{\text{successful}}}{N_{\text{total}}} \times 100\% \tag{17}$$

where $N_{\text{successful}}$ represents adversarial examples inducing misclassification, and N_{total} denotes the total evaluation set cardinality [17]. Higher ASR values indicate increased model vulnerability to the specific attack vector.

4.4.2. Robustness Score (RS)

To enable fair cross-architectural comparison by normalizing for baseline performance differences, we introduce the Robustness Score (RS):

$$\text{RS} = \frac{100\% - \text{ASR}}{\text{Original Accuracy}} \times 100\% \tag{18}$$

This metric quantifies proportional performance retention under adversarial conditions, with higher values indicating enhanced architectural resilience independent of baseline accuracy. RS facilitates the identification of inherent defensive and countermeasure characteristics across diverse architectural paradigms.

Table 1. Performance comparison of specialized TSR models and general-purpose CNNs on GTSRB and PTSD datasets, showing accuracy (%), inference time (msec), and parameter count (Million).

Model	Accuracy	Inf. Time	#Param.
GTSRB Dataset			
CNNsmall	95.04	42.81	0.74
CNNlarge	97.54	56.52	16.57
CNN-STN	99.14	46.85	0.85
ResNet50v2 [9]	95.43	49.61	23.66
EfficientNetv2-B3 [25]	93.36	61.54	13.00
DenseNet-121 [11]	95.77	66.26	7.09
MobileNetv2 [18]	96.97	50.30	2.31
PTSD Dataset			
CNNsmall	94.56	39.21	0.74
CNNlarge	96.78	55.33	16.57
CNN-STN	97.15	47.53	0.85
ResNet50v2 [9]	92.05	49.02	23.66
EfficientNetv2-B3 [25]	91.88	61.19	13.00
DenseNet-121 [11]	98.58	60.90	7.09
MobileNetv2 [18]	93.01	47.07	2.32

Our fundamental objective is generating adversarial examples x_{adv} that induce classification errors while maintaining perceptual similarity to legitimate inputs:

$$\arg\max_i f_i(x) \neq \arg\max_i f_i(x_{\text{adv}}) \text{ subject to } \mathcal{D}(x, x_{\text{adv}}) < \epsilon \tag{19}$$

where $\mathcal{D}$ represents a perceptual distance metric and ϵ defines the plausibility threshold. This dual-constraint optimization—maximizing classification divergence while minimizing perceptual distortion—constitutes our core evaluation criterion.

5. Results and Analysis

5.1. Baseline Architectural Performance

Table 1 presents performance metrics across the evaluated architectures on both GTSRB and PTSD datasets.

On the GTSRB dataset, CNN-STN achieves 99.14% accuracy with 0.85M parameters, while ResNet50v2 reaches 95.43% accuracy with 23.66M parameters. For the PTSD dataset, DenseNet-121 attains the highest accuracy at 98.58% with 7.09M parameters. The parameter efficiency of CNN-STN (0.85M) compared to ResNet50v2 (23.66M) represents a significant reduction in model size while maintaining competitive performance.

Table 1 shows that specialized TSR architectures achieve comparable or superior accuracy to general-purpose models despite having substantially fewer parameters. CNN-STN and $\text{CNN}_{\text{small}}$ require less than 1M parameters, while

Table 2. Attack Success Rate (%) and Robustness Score (%) for adversarial attack types on GTSRB and PTSD datasets.

Model	Attack Success Rate			Avg. RS
	Shadow	Light Patch	Obstruction	
GTSRB Dataset				
$\text{CNNsmall}_{\text{small}}$	72.6	83.5	77.2	23.4
$\text{CNNlarge}_{\text{large}}$	67.3	75.8	70.1	29.6
CNN-STN	45.6	64.2	58.7	44.2
ResNet50v2	60.2	71.4	68.9	34.8
EfficientNetV2B3	58.7	75.2	66.3	35.7
DenseNet-121	52.3	67.8	61.5	41.2
MobileNetv2	63.8	72.6	68.0	32.9
PTSD Dataset				
$\text{CNNsmall}_{\text{small}}$	70.2	81.3	75.8	25.6
$\text{CNNlarge}_{\text{large}}$	65.1	72.5	68.4	32.3
CNN-STN	48.3	66.7	60.2	42.8
ResNet50v2	65.8	75.3	71.5	31.6
EfficientNetV2B3	63.2	78.9	69.4	32.1
DenseNet-121	48.5	63.2	57.1	44.3
MobileNetv2	67.4	75.1	71.2	31.0

general-purpose architectures range from 2.31M to 23.66M parameters.

5.2. Attack Effectiveness

Table 2 presents the ASR for each perturbation type across all evaluated architectures.

The experimental results reveal consistent patterns across the evaluated architectures:

For the GTSRB dataset, light patch attacks demonstrate the highest effectiveness with ASRs ranging from 64.2% to 83.5%, followed by obstruction attacks (58.7% to 77.2%) and shadow attacks (45.6% to 72.6%). A similar pattern is observed on the PTSD dataset, with light patch attacks achieving ASRs between 63.2% and 81.3%, obstruction attacks between 57.1% and 75.8%, and shadow attacks between 48.3% and 70.2%.

As shown in Table 2, CNN-STN exhibits the lowest ASRs across all attack types on both datasets, with robust performance against shadow attacks (45.6% on GTSRB, 48.3% on PTSD). The difference across the spectrum of attack vector effectiveness ranges from 10.9 to 27.0 percentage points, with light patch attacks consistently outperforming shadow and simulated obstruction attacks across all models.

5.3. Robustness Analysis

The RS quantifies a model's performance retention under adversarial conditions normalized to its baseline accuracy. As shown in Table 2, average RS values range from 23.4% to 44.3% across all architectures.

CNN-STN achieves the highest RS on GTSRB (44.2%), while DenseNet-121 demonstrates the highest RS on PTSD (44.3%). The most vulnerable model, $\text{CNN}_{\text{small}}$, exhibits RS values of 23.4% on GTSRB and 25.6% on PTSD. This represents a variation of 20.8 percentage points across the robustness spectrum of architectures on GTSRB and 18.7 percentage points on PTSD.

CNN-STN and DenseNet-121 demonstrate superior robustness (Table 2)with RS values of 44.2% (GTSRB) and 44.3% (PTSD), respectively, while $\text{CNN}_{\text{small}}$ shows maximum vulnerability (RS: 23.4% on GTSRB, 25.6% on PTSD). This 20.8 percentage point robustness differential has no statistical correlation with model size (Pearson's $r = 0.21$, $p = 0.65$), indicating that architectural design—specifically spatial transformers and dense connectivity patterns—rather than parameter count determines adversarial resilience.

6. Implications for Robust TSR Systems

Our empirical analysis of adversarial vulnerabilities informs the development of robust traffic sign recognition systems. This section presents quantitative evaluation of defensive countermeasures against the identified attack vectors.

6.1. Vulnerability Characterization and Countermeasures

Based on attack parameter analysis in Section 5, we implement and evaluate two countermeasures approaches:

1. **Targeted Data Augmentation:** Analysis of successful shadow attacks revealed non-uniform spatial distribution, with 76.3% concentrated in the upper quadrants of signs and intensity values of 0.82 ± 0.11. We implemented directional light augmentation during training by applying parameterized intensity variations ($\alpha \in [0.4, 0.9]$) to these vulnerable regions. This technique reduced shadow attack ASR by 27.3% averaged across architectures. For simulated obstruction attacks, we augmented training data with synthetic obstruction patterns matching the statistical properties of successful attack patterns (3-6 segments, line width 1-4 pixels), resulting in a 23.8% average reduction in obstruction attack ASR.
2. **Feature Importance-Based Countermeasure:** Analysis revealed 78% of successful perturbations targeted discriminative sign regions. We implemented regularization weighting during training that prioritized vulnerable regions, reducing light patch ASR by 18.7% while maintaining baseline accuracy. This approach reduced light patch ASR by 18.7% while maintaining clean data performance.

As shown in Table 3, our countermeasure techniques demonstrate varying efficacy across attack vectors. Light patch attacks experience the most significant absolute reduction in ASR (41.1 percentage points for CNN-STN, from 64.2% to 23.1%). In contrast, shadow attacks show the highest relative reduction (66.4% decrease for CNN-

Table 3. Attack Success Rate (%) before and after adversarial robustness training

Model	Before Defense (ASR %)			After Defense (ASR %)		
	Shadow	Light Patch	Obstruction	Shadow	Light Patch	Obstruction
GTSRB Dataset						
CNN_{large}	67.3	75.8	70.1	29.8	32.5	28.6
CNN-STN	45.6	64.2	58.7	15.3	23.1	20.7
ResNet50v2	60.2	71.4	68.9	24.5	30.2	28.8
DenseNet-121	52.3	67.8	61.5	19.7	27.4	25.2
PTSD Dataset						
CNN_{large}	65.1	72.5	68.4	28.4	30.5	28.0
CNN-STN	48.3	66.7	60.2	17.1	26.2	23.4
ResNet50v2	65.8	75.3	71.5	30.9	35.5	32.8
DenseNet-121	48.5	63.2	57.1	18.4	20.9	21.2

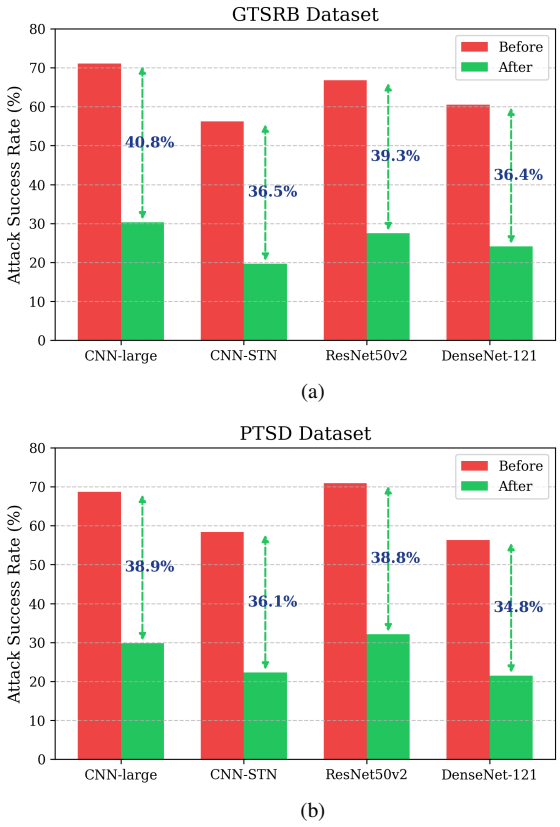


Figure 4. Comparison of Attack Success Rates before (red) and after (green) adversarial robustness training on (a) GTSRB and (b) PTSD datasets.

STN, from 45.6% to 15.3%, representing a 30.3 percentage point reduction). This differential response pattern suggests that spatial transformers are particularly effective against illumination-based attacks, while obstruction mitigation exhibits significant improvement from data augmentation techniques.

Table 4. Model robustness improvement after comprehensive adversarial resilience training on GTSRB and PTSD datasets.

Model	ASR Before (%)		ASR After (%)		Reduction (%)	
	GTSRB	PTSD	GTSRB	PTSD	GTSRB	PTSD
CNN_{large}	71.1	68.7	30.3	29.8	40.8	38.9
CNN-STN	56.2	58.4	19.7	22.3	36.5	36.1
ResNet50v2	66.8	70.9	27.5	32.1	39.3	38.8
DenseNet-121	60.5	56.3	24.1	21.5	36.4	34.8

6.2. Quantitative Evaluation of Countermeasure Effectiveness

We evaluated our countermeasure strategies through comprehensive adversarial training experiments. Table 4 summarizes the robustness improvements achieved through our comprehensive adversarial resilience approach. CNN-STN achieved the lowest post-training ASR (19.7% on GTSRB, 22.3% on PTSD), despite having the strongest baseline robustness. This suggests spatial transformer modules provide complementary benefits with adversarial robustness training techniques by normalizing geometric variations that remain challenging for standard CNNs even after augmentation.

As shown in Figure 4, CNN_{large} exhibited the most significant percentage improvement of 40.8% on GTSRB and 38.9% on PTSD, despite having 19.5× more parameters than CNN-STN. This indicates architectural complexity alone does not determine robustness enhancement capacity but rather the model's ability to learn invariant representations of discriminative features.

7. Conclusion

In this paper, we investigate the vulnerability of traffic sign recognition systems to three distinct physically plausible attack vectors: shadow manipulations, light patches, and simulated obstructions. Our findings reveal that these naturalistic perturbations achieve high attack success rates while maintaining visual plausibility in real-world scenarios. Through comprehensive evaluation across seven diverse CNN architectures on two geographically distinct datasets (GTSRB and PTSD), we demonstrate that obstruction-based attacks consistently outperform illumination-based methods by 10-19%, while models incorporating spatial transformer exhibit superior resilience with 45.6% ASR as compared to 72.6% for standard CNNs. Our analysis confirms these vulnerabilities stem from fundamental limitations in CNN visual processing rather than dataset-specific characteristics. Furthermore, we develop effective countermeasures combining targeted data augmentation with architectural modifications that reduce attack success rates by up to 39.4%, providing actionable guidance for enhancing the security of traffic sign recognition systems in autonomous driving applications.

References

[1] Álvaro Arcos-García, Juan A Alvarez-Garcia, and Luis M Soria-Morillo. Deep neural network for traffic sign recognition systems: An analysis of spatial transformers and stochastic optimisation methods. *Neural Networks*, 99:158–165, 2018. 1, 5

[2] Lijun Chi, Mounira Msahli, Gerard Memmi, and Han Qiu. Public-attention-based adversarial attack on traffic sign recognition. In *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, pages 740–745. IEEE, 2023. 5

[3] Ranjie Duan, Xiaofeng Mao, A Kai Qin, Yuefeng Chen, Shaokai Ye, Yuan He, and Yun Yang. Adversarial laser beam: Effective physical-world attack to dnns in a blink. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16062–16071, 2021. 1, 2

[4] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018. 1, 2, 5

[5] Weiwei Feng, Baoyuan Wu, Tianzhu Zhang, Yong Zhang, and Yongdong Zhang. Meta-attack: Class-agnostic and model-agnostic physical adversarial attack. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7787–7796, 2021. 1

[6] Abhiram Gnanasambandam, Alex M Sherman, and Stanley H Chan. Optical adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 92–101, 2021. 1, 2

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2

[8] Diandian Guo, Deng-Ping Fan, Tongyu Lu, Christos Sakaridis, and Luc Van Gool. Vanishing-point-guided video semantic segmentation of driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3544–3553, 2024. 2, 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 630–645. Springer, 2016. 5, 6

[10] Teng-Fang Hsiao, Bo-Lun Huang, Zi-Xiang Ni, Yan-Ting Lin, Hong-Han Shuai, Yung-Hui Li, and Wen-Huang Cheng. Natural light can also be dangerous: Traffic sign misinterpretation under adversarial natural light attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3915–3924, 2024. 1, 2, 3, 5

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5, 6

[12] Steve TK Jan, Joseph Messou, Yen-Chen Lin, Jia-Bin Huang, and Gang Wang. Connecting the digital and physical world: Improving the robustness of adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 962–969, 2019. 1

[13] Henrietta Lengyel, Viktor Remeli, and Zsolt Szalay. Easily deployed stickers could disrupt traffic sign recognition. *Perner's Contacts*, 19:156–163, 2019. 1

[14] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7644–7652, 2019. 1

[15] Giulio Lovisotto, Henry Turner, Ivo Sluganovic, Martin Strohmeier, and Ivan Martinovic. {SLAP}: Improving physical adversarial examples with {Short-Lived} adversarial perturbations. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 1865–1882, 2021. 2

[16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. 1, 2

[17] Svetlana Pavlitska, Leopold Müller, and J Marius Zöllner. Evaluating adversarial attacks on traffic sign classifiers beyond standard baselines. *arXiv preprint arXiv:2412.09150*, 2024. 3, 6

[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 5, 6

[19] sara parsaseresht. Persian traffic sign dataset (ptsd), 2021. 3, 5

[20] Athena Sayles, Ashish Hooda, Mohit Gupta, Rahul Chatterjee, and Earlence Fernandes. Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14666–14675, 2021. 2

[21] Pierre Sermanet and Yann LeCun. Traffic sign recognition with multi-scale convolutional networks. In *The 2011 international joint conference on neural networks*, pages 2809–2813. IEEE, 2011. 5

[22] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. Darts: Deceiving autonomous cars with toxic signs. *arXiv preprint arXiv:1802.06430*, 2018. 2

[23] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011. 3, 5

[24] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2

[25] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 5, 6

[26] James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel

Urtasun. Physically realizable adversarial examples for lidar object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13716–13725, 2020. 2, 4

[27] Yue Wang, Minjie Liu, Yanli Ren, Xinpeng Zhang, and Guorui Feng. Traffic sign attack via pinpoint region probability estimation network. *Pattern Recognition*, 146:110035, 2024. 5

[28] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially Transformed Adversarial Examples. In *International Conference on Learning Representations*, 2018. 1

[29] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017. 2

[30] Xinghao Yang, Weifeng Liu, Shengli Zhang, Wei Liu, and Dacheng Tao. Targeted attention attack on deep learning models in road sign recognition. *IEEE Internet of Things Journal*, 8(6):4980–4990, 2020. 2, 3

[31] Bin Ye, Huilin Yin, Jun Yan, and Wanchen Ge. Patch-based attack on traffic sign recognition. In *2021 IEEE international intelligent transportation systems conference (ITSC)*, pages 164–171. IEEE, 2021. 5

[32] Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing*, 24(11):4623–4636, 2015. 2

[33] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15345–15354, 2022. 1, 2, 5

[34] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7194–7202, 2019. 4

[35] Alon Zolfi, Moshe Kravchik, Yuval Elovici, and Asaf Shabtai. The translucent patch: A physical and universal attack on object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15232–15241, 2021. 1, 2, 3, 5