Attention-Aware Temporal Adversarial Shadows on Traffic Sign Sequences

Pedram MohajerAnsari*, Amir Salarpour*, David Fernandez, Cigdem Kokenoz, Bing Li, Mert D. Pesé
School of Computing, Clemson University
{pmohaje, asalarp, dferna3, ckokeno, bli4, mpese}@clemson.edu

Abstract

We present a framework for black-box adversarial attacks on traffic signs using dynamic, temporally coherent shadows. Unlike prior work that focuses on single-image attacks or relies on conspicuous physical artifacts, our method operates over entire image sequences, mimicking realistic scenarios where a traffic sign is observed from varying distances. We design a non-differentiable shadow generator that casts a single fixed-shape, fixed-opacity shadow whose spatial scale evolves over time to simulate natural environmental shading. A genetic algorithm is used to optimize shadow geometry and opacity, guided by a dual loss that jointly maximizes classification error and visual attention disruption. Attention perturbation is measured using DINO ViT attention maps between clean and shadowed frames. Evaluated on the GTSRB dataset, our method achieves a sequence-level attack success rate (SL-ASR) — defined as the percentage of sequences where at least τ out of T frames are misclassified — ranging from 52.3% to 87.5%, depending on the threshold and shadow type. Furthermore, incorporating attention supervision yields consistent SL-ASR gains of 11–18% over purely classification-based attacks. [1]

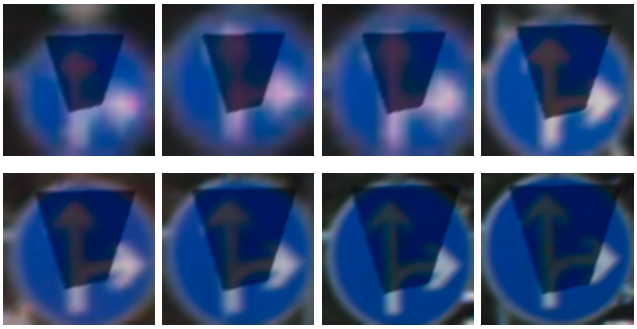


Figure 1. A sequence of adversarial images illustrating our core idea: a single shadow pattern (same shape and opacity) is progressively scaled and applied across time to a traffic sign sequence. Although the original class is 36 (go straight), the shadow causes the classifier to consistently predict class 40 (turn right), revealing the power of subtle, temporally evolving occlusions in sequential settings.

1. Introduction

In recent years, deep neural networks (DNNs) have achieved remarkable success across a range of computer vision applications — from image classification and object detection to scene segmentation [26, 28, 36, 44]. Despite these advances, studies have found that DNN-based models are surprisingly susceptible to adversarial examples, even when the added perturbations appear negligible in magnitude [10, 42]. Such vulnerabilities pose a significant concern in safety-critical scenarios, particularly in autonomous vehicles (AVs) which depend on automated driving systems (ADS), comprising perception, planning, and control modules [31, 37]. Within the perception module, DNN models, which are responsible for tasks such as image classification and object detection [8, 33, 35], provide crucial information for navigation and maneuvering. Ensuring the trustworthiness of these DNNs is therefore essential, since even minor errors in recognition can propagate through the pipeline and undermine overall safety [19].

Physical-world adversarial attacks have garnered growing attention, particularly due to their potential for realistic yet disruptive modifications. Traditional attempts often rely on conspicuous stickers or camouflage patterns [5, 11], which — despite their effectiveness — are overt and less feasible for stealthy manipulation. To overcome these drawbacks, recent studies have investigated light-based perturbations as a more subtle means of deceiving vision models.

For instance, Zhong *et al.* [43] demonstrated that even naturally occurring shadows can severely mislead traffic sign classification, achieving high attack success rates in simulations and real-world tests. Wang *et al.* [40] further explored this concept through the Reflected Light Adversarial Attack (RFLA), using mirrors and colored filters to produce adversarial perturbations under direct sunlight. Simi-

*The first two authors contributed equally.

[1] The implementation code of this work is available at https://github.com/pedram-mohajer/ShadowSeq.

larly, Li *et al.* [23] introduced AdvSL, which leverages spotlights to enable both stealthy and adaptable physical-world attacks, while Hsiao *et al.* [18] underscored the danger of natural light interference, revealing how subtle illumination changes can undermine traffic sign recognition models.

Building upon these insights, we propose an adversarial framework that introduces shadow perturbations to disrupt both the classification accuracy and spatial attention of deep models on traffic sign recognition tasks. Unlike prior work that focuses on static single-image attacks, our method targets entire image sequences, simulating how a vehicle perceives a sign at varying distances and perspectives. We design a parametric shadow generator that overlays polygonal or triangular shadows across each frame, dynamically scaled over time to mimic consistent environmental shading. Figure 1 illustrates how applying the same shadow pattern across an image sequence consistently induces misclassification. These shadows are visually plausible, localized within the region of interest, and temporally coherent, making them ideal candidates for stealthy physical-world attacks.

A key novelty of our framework lies in its integration of DINO-based Vision Transformer (ViT) [6] attention supervision as a second-order optimization objective. Rather than relying solely on classification misdirection, we explicitly seek to disrupt the model's internal visual reasoning. DINO's self-attention maps provide a spatial distribution of the regions the model deems most important for recognition — with lighter areas indicating stronger focus. By comparing the attention maps of shadowed inputs against class-specific reference maps derived from clean exemplars, we encourage the adversarial shadow to significantly alter the model's perceptual focus. This enforces a form of interpretable misdirection: the attack not only fools the classifier but also causes attention to drift away from the true semantic core of the image. The result is a physically plausible perturbation that compromises both accuracy and interpretability, exposing deeper vulnerabilities in attention-based vision systems.

To efficiently explore the shadow parameter space, we employ a Genetic Algorithm (GA) [16] that evolves candidate solutions based on a joint loss function. Each individual in the population encodes a unique shadow configuration (control points and opacity), which is applied across the image sequence using temporal scaling. The GA iteratively selects and refines candidates through crossover and mutation, guided by feedback from both the Convolutional Neural Network (CNN) classifier and the DINO attention extractor. Our experiments on the German Traffic Sign Recognition Benchmark (GTSRB) dataset [17] demonstrate that this sequence-level attack strategy reliably causes persistent misclassification across frames.

This paper makes the following contributions:

- We propose the first adversarial shadow attack that operates over full image sequences, simulating real-world scenarios where a traffic sign is captured from varying distances. Our attack applies a single shadow pattern consistently across frames, with moderate spatial adjustments to reflect natural changes in shadow appearance over time. This temporal coherence contributes to physical plausibility while maintaining attack effectiveness across varying viewpoints.
- We introduce a novel multi-objective loss that jointly degrades classification accuracy and disrupts visual interpretability. By leveraging DINO's class-conditioned attention maps, our attack explicitly misaligns the model's internal focus, guiding perturbations to be both effective and explainable.
- We introduce a new evaluation metric — Sequence-Level Attack Success Rate (SL-ASR) — which defines an attack as successful only if it causes misclassification in at least τ out of T frames. This metric captures the persistence and temporal robustness of adversarial effects in sequential settings like autonomous driving. Using SL-ASR, we demonstrate that our attention-guided attack significantly improves both effectiveness and stealthiness, supported by quantitative and qualitative results on the GTSRB dataset.

2. Related Work

DNNs are widely used in AVs for tasks such as traffic sign recognition and scene understanding [27, 32]. This makes their vulnerability to physical adversarial attacks a critical safety concern [9, 22]. One prominent line of research focuses on real-world attacks that physically alter the visual scene to mislead perception models. Eykholt et al. [11] showed that printed adversarial patches and stickers can fool traffic sign classifiers under varied conditions. However, these attacks are often static and visually conspicuous, limiting their stealth and practicality.

To improve realism, later efforts turned to optical phenomena. Light- and shadow-based attacks manipulate perception by projecting patterns or simulating natural occlusions without altering the object itself. For example, Wang et al.[39] used mirrors and colored filters to generate adversarial reflections, while Zhong et al.[43] showed that physically plausible triangular shadows could mislead classifiers. Despite their effectiveness, such methods typically focus on single-frame inputs, lack temporal consistency, and often rely on handcrafted patterns or heuristic search, limiting generalizability to dynamic scenarios.

Research has shown that even when predictions remain unchanged, adversarial examples can significantly distort interpretability outputs. Tao et al.[38] found that saliency maps could become misaligned with meaningful image regions. Gu et al.[14] reported that ViTs under attack may ex-

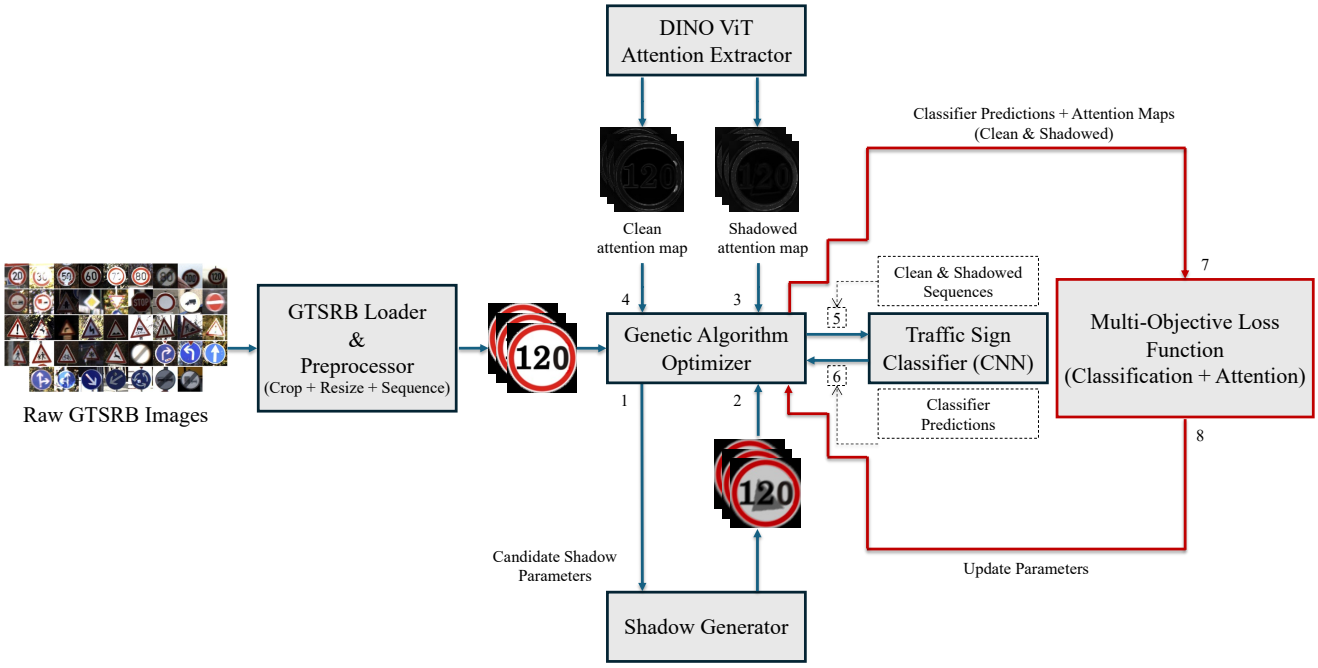


Figure 2. Overview of our proposed adversarial shadow attack framework for traffic sign recognition in sequential visual settings. Given a clean input sequence, the framework uses a scGA to search for optimal shadow parameters—comprising a single shape and opacity—that are temporally scaled across frames to simulate the consistent growth of a physical shadow over time. Each candidate is evaluated by rendering shadows over the sequence, passing the images through a CNN classifier and a DINO-based attention extractor, and computing a joint classification-attention loss. The GA iteratively evolves the population until an adversarial configuration is found that causes misclassification in at least τ out of T frames, ensuring both persistence and stealthiness of the perturbation.

perience attention collapse or spread, often focusing on irrelevant background regions. While these effects raise concerns about explanation reliability, they are usually byproducts of misclassification-focused attacks.

Only a few works explicitly target interpretability mechanisms. For example, Fong et al.[13] and Rasuly et al.[34] explored attacks on attention maps and Grad-CAM outputs. However, such approaches generally assume white-box access to gradients and are typically applied to single-image scenarios, neglecting temporal aspects in sequential vision tasks.

Most adversarial research focuses on static images, overlooking the temporal dimension central to video and real-time systems. Hsiao et al. [18] introduced natural lighting variations across video frames to test model robustness under dynamic illumination using zeroth-order optimization. While this work incorporates time-awareness, it does not address structured, physically plausible perturbations—such as cast shadows—or their effects on attention mechanisms in sequential tasks.

Many adversarial attacks assume access to model gradients, which is unrealistic in deployed systems. This has led to the development of black-box approaches that rely only on model outputs. Zeroth-Order Optimization (ZOO)[7], Natural Evolution Strategies (NES)[20], and Bandits-TD [21] estimate gradients through repeated queries, though they often incur high sample costs. Genetic Algorithms (GAs) offer a fully gradient-free alternative by evolving perturbations using selection, mutation, and crossover. Alzantot et al.[3] demonstrated their effectiveness in both NLP and vision, and follow-up works[4, 24] extended them to more efficient vision attacks. However, most of these approaches focus on unstructured, pixel-level perturbations, with limited attention to structured or physically grounded changes.

Various defense mechanisms have been proposed to counter adversarial vulnerabilities. One common strategy is adversarial training, where models are trained on both clean and adversarial examples to improve robustness, as demonstrated by Madry *et al.* [25]. Another approach, defensive distillation, introduced by Papernot *et al.* [30], reduces model sensitivity by smoothing decision boundaries. Robust optimization methods, such as those presented by Wong and Kolter [41], aim to strengthen models

through formal optimization techniques. Additionally, input preprocessing techniques, like those explored by Guo *et al.* [15], apply transformations such as denoising to diminish the impact of adversarial perturbations. More recently, Aldeen *et al.* [1, 2] and Fernandez *et al.* [12] investigated the use of Large Language Models (LLMs) and Large Multimodal Models (LMMs) to enhance the cybersecurity of autonomous vehicles, highlighting emerging directions beyond traditional defense methods. Complementary sensors such as LiDAR can augment and provide additional context for visual anomalies, improving robustness [29].

3. Methodology

This section details the training and adversarial attack pipeline of our proposed framework, which leverages shadow-based perturbations to degrade both classification accuracy and attention reliability in traffic sign recognition models. Our approach integrates four key components: (1) the GTSRB dataset structured into temporal sequences, (2) a CNN for traffic sign classification, (3) a ViT attention extractor trained with the DINO framework, and (4) a genetic algorithm for optimizing shadow parameters. An overview is provided in Figure 2.

3.1. Dataset: GTSRB and Sequence Formulation

The GTSRB dataset consists of 1,306 traffic sign sequences. Each sequence contains exactly 30 RGB frames and corresponds to a single physical sign instance. As a vehicle approaches a sign, a sequence captures the same object from varying distance, reflecting realistic temporal progression in road scenarios.

We define the i-th sequence as:

$$\mathcal{S}_i = \{I_t^i\}_{t=1}^{30}, \quad y_i \in \mathcal{C} \tag{1}$$

where each frame is annotated with a tight axis-aligned bounding box that encloses the traffic sign, with all signs categorized into $C = 43$ classes, denoted by $\mathcal{C}$.

Preprocessing. To ensure uniform scale and alignment for downstream tasks, each frame undergoes the following transformation:

1. **Cropping:** The image is cropped using its ground truth bounding box. This removes background context and centers the traffic sign within the image.
2. **Resizing:** The cropped patch is resized to 128×128 pixels using bilinear interpolation. This enforces a consistent input resolution and scale across all frames.
3. **Representation:** Each processed image is stored as an RGB array of shape $128 \times 128 \times 3$, used for all shadow rendering, visualization, and inference tasks.

The transformation function $\mathrm{Preprocess}(\cdot)$ is applied to each frame in the sequence using its corresponding bounding box, resulting in:

$$\tilde{\mathcal{S}}_i = \left\{ \mathrm{Preprocess}(I_t^i, \mathrm{bbox}^t) \right\}_{t=1}^{30}, \quad \tilde{I}_t^i \in \mathbb{R}^{128\times128\times3}.$$

Dataset Partitioning. The dataset is divided into training and testing sets at the *sequence level*. Let the complete dataset consist of N labeled sequences:

$$\mathcal{D} = \{ (\mathcal{S}_i, y_i) \}_{i=1}^{N},$$

where each $\mathcal{S}_i$ contains 30 RGB frames of the same traffic sign, and $y_i \in \mathcal{C}$ denotes its class label. We apply a fixed random shuffle to all sequences and allocate 85% to the training set and the remaining 15% to the test set. Each sequence is kept intact and assigned entirely to one split, ensuring no frame-level overlap. The training split is used to train the GTSRB-CNN classifier and fine-tune the DINO attention model, while the test split is reserved exclusively for adversarial shadow generation and evaluation.

3.2. Attention Supervision via DINO ViT

To guide and analyze how adversarial shadows affect attention mechanisms, we use a pretrained ViT from the DINO framework. This model is fine-tuned on the training split of the GTSRB dataset. We adopt the `vit_small` architecture, which contains $h = 6$ attention heads in its final self-attention layer. For each input image $I' \in \mathbb{R}^{128\times128\times3}$, the model outputs h individual attention maps, which are averaged to obtain a single spatial attention distribution:

$$A(I') \in \mathbb{R}^{128\times128},$$

representing the average self-attention across all heads in the final transformer layer. These maps reflect how the model distributes attention spatially over the image. For supervision and comparison, we compute clean and shadowed attention maps for each frame in a sequence. Given a clean image I and its shadowed counterpart I', their attention maps are denoted $A(I)$ and $A(I')$, respectively. The clean maps are cached prior to attack for computational efficiency. To quantify how much a shadow-perturbed image I' diverges from its clean version, we compute a mean squared error (MSE) between the two attention maps:

$$\mathcal{L}_{\text{attn}}(I') = \frac{1}{128^2} \left\| A(I') - A(I) \right\|_2^2. \tag{2}$$

A high value of $\mathcal{L}_{\text{attn}}$ indicates that the adversarial shadow has significantly disrupted the attention pattern of the original frame. This metric serves as an auxiliary objective during the attack optimization process to maximize attention deviation in addition to misclassification.

3.3. Adversarial Shadow Generator

We use a parametric shadow generation module that overlays synthetic shadows onto clean traffic sign images. RGB

images are converted to LAB color space, where only the L (luminance) channel is manipulated to simulate cast shadows—allowing realistic shading without altering color information. After the photometric transformation, the image is converted back to RGB. This approach enables fine-grained control over lightness while preserving the natural visual structure, resulting in effective perturbations that deceive both classification and attention mechanisms.

Shadow Parameterization. A shadow mask is defined by a tuple:

$$\theta = \left\{(x_j, y_j)_{j=1}^{K},\ \alpha\right\},$$

where $\{(x_j, y_j)\}_{j=1}^{K}$ are the control points of the shape (polygon or triangle), and $\alpha \in [0.1,\ 0.7]$ is the opacity controlling the shadow's intensity.

We consider two geometric configurations:

- **Polygonal mask ($K = 4$).** To ensure the quadrilateral covers diverse regions of the image, each vertex (x_j, y_j) is sampled from a distinct quadrant. Let W and H denote the image width and height. We define:

$$(x_j, y_j) \sim \mathcal{U}(\mathcal{R}_j), \quad \text{for } j = 1, \ldots, 4, \tag{3}$$

  where each region $\mathcal{R}_j$ is a rectangular subregion (e.g., top-left, top-right, etc.) of the image domain $[0, W] \times [0, H]$. This enforces spatial spread and coverage from multiple angles around the traffic sign.
- **Triangular mask ($K = 3$).** For wedge-like occlusions, triangle vertices are sampled relative to the image center $(x_c, y_c) = (\frac{W}{2}, \frac{H}{2})$:

$$(x_j, y_j) = (x_c + \delta x_j,\ y_c + \delta y_j), \quad \delta x_j, \delta y_j \sim \mathcal{U}(r_j), \tag{4}$$

  where $\mathcal{U}(r_j)$ is a uniform distribution over a bounded offset region. The offsets are designed such that one vertex lies above the center and the others below it, forming a forward-leaning triangle. This configuration mimics cast shadows from roadside structures or vehicle parts.

Opacity Sampling. The opacity parameter α is sampled uniformly:

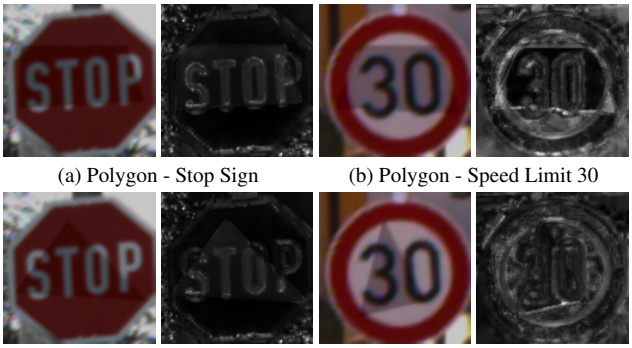
$$\alpha \sim \mathcal{U}[0.1,\ 0.7], \tag{5}$$

ensuring that shadows are perceptible but do not obscure the traffic sign entirely. During genetic mutation, α is perturbed with small Gaussian noise and clipped to remain within this interval.

Shadow Transformation. Once a shadow mask M_θ is created, the image $I \in \mathbb{R}^{H \times W \times 3}$ is modified inside the masked region using photometric transformations:

$$I' = \mathcal{T}(I,\ M_\theta,\ \alpha), \tag{6}$$

where $\mathcal{T}$ includes:



(a) Polygon - Stop Sign (b) Polygon - Speed Limit 30

(c) Triangle - Stop Sign (d) Triangle - Speed Limit 30

Figure 3. Comparison of adversarial shadow shapes with identical opacity (0.21). Each pair shows a shadowed input (left) and its corresponding DINO attention map (right).

1. **Luminance Attenuation:** We darken the image in LAB space within the shadow region. Let I^{LAB} be the LAB conversion of I, then:

$$I^{\text{LAB}}_{\text{shadow}}(x, y, 0) \leftarrow (1 - \alpha) \cdot I^{\text{LAB}}(x, y, 0), \quad \forall\, (x, y) \in \text{supp}(M_\theta). \tag{7}$$

2. **Post-processing Filters:** Then, we apply a sequence of visual enhancements. First, Gaussian smoothing is applied around the perimeter of M_θ to avoid harsh mask boundaries, with kernel size proportional to object scale. Next, a motion blur kernel simulates directional streaks in the shadow, mimicking the appearance of cast shadows during movement. Finally, brightness normalization is applied by scaling the entire LAB image uniformly to preserve dynamic range.

Temporal Scaling. When shadows are applied to an image sequence $\mathcal{S} = \tilde{I}_1, \ldots, \tilde{I}_T$, the shadow mask is progressively scaled across time steps to simulate a changing distance from the sign:

$$s_t \ = \ s_{\min} + \left(\frac{t-1}{T-1}\right) \cdot (s_{\max} - s_{\min}), \tag{8}$$

where $s_{\min}$ and $s_{\max}$ are scaling coefficients (e.g., 0.6 and 1.0), and s_t is applied to the polygon/triangle before overlaying it on frame t.

The final transformed sequence becomes:

$$\tilde{\mathcal{S}}^{\text{shadow}} = \left\{\mathcal{T}(\tilde{I}_1,\ M_{\theta_1},\ \alpha),\ \ldots,\ \mathcal{T}(\tilde{I}_T,\ M_{\theta_T},\ \alpha)\right\}.$$

This approach produces dynamic, spatially varying occlusions that affect a classifier's prediction and attention response over time.

3.4. Genetic Algorithm Optimization

Given a sequence $\mathcal{S} = \{I_1, \ldots, I_T\}$ with ground truth label $y \in \mathcal{C}$, our objective is to find optimal shadow parameters θ^* such that the resulting shadowed sequence $\mathcal{S}' =$

$\{I'_1, \ldots, I'_T\}$ causes significant degradation in both classification accuracy and attention consistency. We formulate this as a multi-objective optimization problem, where the total loss for a shadow configuration θ is defined as:

$$\mathcal{L}_{\text{total}}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \left[\mathcal{L}_{\text{cls}}(I'_t) - \lambda \, \mathcal{L}_{\text{attn}}(I'_t) \right] \tag{9}$$

where $\lambda \in \mathbb{R}_{\geq 0}$ controls the relative importance of attention shift versus classification misdirection.

Classification Loss. To evaluate the classifier's prediction confidence for the true class y on frame I'_t, we use the softmax probability $\hat{y}_t^{(y)}$:

$$\mathcal{L}_{\text{cls}}(I'_t) = \begin{cases} 1 - \hat{y}_t^{(y)}, & \text{if } \hat{y}_t = y \\ \hat{y}_t^{(y)}, & \text{otherwise} \end{cases} \tag{10}$$

This loss is minimal when the classifier is confident about the correct label, and maximal when the classifier assigns low probability to the true class, thereby encouraging misclassification.

Attention Loss. To measure attention perturbation, we compare the DINO-generated attention map A'_t of each shadowed frame I'_t to its clean counterpart A_t^{clean} from the same sequence, using mean squared error (MSE) between the two spatial distributions:

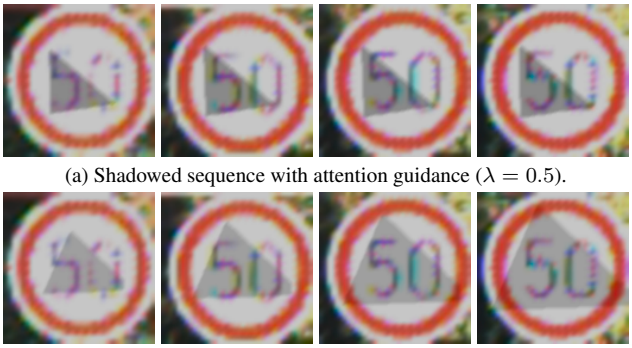
$$\mathcal{L}_{\text{attn}}(I'_t) = \frac{1}{P^2} \left\| \frac{A'_t}{\max A'_t} - \frac{A_t^{\text{clean}}}{\max A_t^{\text{clean}}} \right\|_2^2 \tag{11}$$

where $P = 128$ is the spatial resolution of the attention map. This loss penalizes perceptual shifts in visual attention caused by the adversarial shadow. A large $\mathcal{L}_{\text{attn}}$ implies that the model's focus has deviated significantly from its expected, unperturbed distribution — potentially leading to poor interpretability and degraded decision reliability.

Genetic Algorithm Steps. We minimize Eq. (9) using a genetic algorithm that evolves shadow parameters over generations to maximize the joint loss. Each individual $\theta^{(k)}$ encodes a shadow configuration consisting of control points and an opacity value. Candidate shadows are applied to the input sequence using temporally scaled masks, and the resulting sequence is evaluated using the total loss.

After evaluation, the fittest individuals are selected to generate new candidates via crossover and mutation. This evolutionary process continues for G generations or until convergence. An attack is deemed *successful* if at least τ out of T frames are misclassified:

$$\sum_{t=1}^{T} \mathbb{I} \left[\arg\max f_{\text{CNN}}(I'_t) \neq y \right] \geq \tau. \tag{12}$$



(a) Shadowed sequence with attention guidance ($\lambda = 0.5$).

(b) Shadowed sequence without attention guidance ($\lambda = 0$).

Figure 4. Visual comparison of shadow patterns generated with and without DINO-based attention supervision. Both sequences successfully cause misclassification of the same 50 speed limit sign.

The full genetic optimization process is summarized in Algorithm 1, which begins by randomly initializing a population of shadow parameter sets, each encoding a unique combination of shape geometry and opacity. For each candidate, shadows are applied across the sequence, and both the classification confidence and DINO attention maps are computed. These are compared against clean references to quantify the impact of the shadow via a combined loss function. If any candidate causes at least τ frames to be misclassified, the algorithm terminates early for that sequence and returns the corresponding parameters. Otherwise, the fittest candidates — those that most effectively degrade classification and attention — are selected to generate the next population through crossover and mutation. This evolutionary process continues for a fixed number of generations or until early stopping is triggered.

4. Experiment

Loss Computation. During optimization, each frame I'_t is passed through the CNN classifier and the DINO attention extractor. Classifier predictions $(\hat{y}_t, \hat{y}_t^{(y)})$ and attention maps A'_t are compared against clean references A_t^{clean} (cached). The per-sequence total loss is computed as described in Sec. 3.4, combining classification and attention terms to guide the genetic search.

Evaluation and Results. Sequence-Level Attack Success Rate (SL-ASR) is defined as the percentage of test sequences where the following condition is satisfied: at least τ out of T frames in the adversarial sequence are misclassified. Formally, for a given sequence $\mathcal{S}' = \{I'_1, \ldots, I'_T\}$:

$$\sum_{t=1}^{T} \mathbb{I} \left[\arg\max f_{\text{CNN}}(I'_t) \neq y \right] \geq \tau. \tag{13}$$

which ensures that a large portion of the sequence is af-

Algorithm 1: Genetic Algorithm for Sequence-Level Adversarial Shadow Attack

Input: Sequence $\mathcal{S} = \{I_1, I_2, \ldots, I_T\}$ with label y,
Meta attention map A_y from DINO,
Cached clean attention maps $\{A_t^{\text{clean}}\}_{t=1}^{T}$,
Population size P, number of generations G,
threshold τ, shadow shape $\in \{\texttt{polygon}, \texttt{triangle}\}$
Output: Optimized shadow parameters θ^* and adversarial sequence $\{I'_t\}_{t=1}^{T}$

Initialize population $\{\theta^{(1)}, \ldots, \theta^{(P)}\}$ with random control points and opacity
for $g = 1$ **to** G **do**
  foreach *candidate* $\theta^{(k)}$ *in population* **do**
    Generate scaled shadow masks $\{M_t^{(k)}\}_{t=1}^{T}$ via temporal scaling
    Apply shadow transformation: $I_t'^{(k)} \leftarrow \mathcal{T}(I_t, M_t^{(k)}, \alpha^{(k)})$
    Run classifier: obtain $f_{\text{CNN}}(I_t'^{(k)})$ and confidence scores
    Extract DINO attention maps $\{A_t^{(k)}\}_{t=1}^{T}$
    Compute per-frame classification and attention losses:

$$\mathcal{L}_{\text{cls}}^{(k)} = \frac{1}{T}\sum_{t=1}^{T}\left(1 - f_{\text{CNN}}(I_t'^{(k)})_y\right)$$

$$\mathcal{L}_{\text{attn}}^{(k)} = \frac{1}{T}\sum_{t=1}^{T}\left\|A_t^{(k)} - A_t^{\text{clean}}\right\|_2^2$$

    Total loss: $\mathcal{L}^{(k)} = \mathcal{L}_{\text{cls}}^{(k)} - \lambda \cdot \mathcal{L}_{\text{attn}}^{(k)}$
    Count misclassifications: $m^{(k)} \leftarrow \sum_{t=1}^{T} \mathbb{I}[\arg\max f_{\text{CNN}}(I_t'^{(k)}) \neq y]$
    if $m^{(k)} \geq \tau$ **then**
      return $\theta^{(k)}$ and adversarial sequence $\{I_t'^{(k)}\}_{t=1}^{T}$ `// Early stopping`
  Select top-$P/2$ candidates with lowest $\mathcal{L}^{(k)}$
  Apply crossover and mutation to generate $P/2$ offspring
  Form next generation by combining parents and offspring to size P
return $\theta^* = \arg\min_k \mathcal{L}^{(k)}$ and corresponding shadowed sequence

fected, making the shadow attack more impactful and persistent.

We evaluate our attack on the 15% held-out test set of GTSRB sequences, where the clean classification accuracy of the GTSRB-CNN model reaches 97.3%. As shown in Table 1, when applying our full shadow-based attack with joint optimization of classification and attention loss ($\lambda = 0.5$), we observe an SL-ASR of 84.2% for a misclassification threshold of $\tau = 17$, 64.2% for $\tau = 27$, and 52.3% for the stricter threshold $\tau = 29$ using triangle shadows. To isolate the role of attention guidance, we ablate the attention loss by setting $\lambda = 0$, thereby optimizing only for misclassification. Under this setting, SL-ASR drops to 65.7% ($\tau = 17$), 50.6% ($\tau = 27$), and 40.4% ($\tau = 29$), indicating that DINO-based attention supervision significantly enhances attack effectiveness by making perturbations more persistent and robust across the sequence. A similar trend is observed with polygon shadows, which achieve even higher SL-ASR values at each threshold. Specifically, as λ increases from 0 to 0.5, SL-ASR improves from 70.4% to 87.5% ($\tau = 17$), from 55.1% to 68.3% ($\tau = 27$), and from 45.5% to 56.9% ($\tau = 29$), respectively. These results confirm that incorporating attention loss not only amplifies misclassification but also promotes spatially consistent and stealthy perturbations throughout the sequence.

Table 1. SL-ASR at varying thresholds τ for $\lambda = 0$ and $\lambda = 0.5$, using triangle and polygon shadows.

	SL-ASR$_{\text{Triangle}}$			SL-ASR$_{\text{Polygon}}$		
τ	$\lambda = 0$	$\lambda = 0.5$	Δ	$\lambda = 0$	$\lambda = 0.5$	Δ
29	40.4%	52.3%	+11.9%	45.5%	56.9%	+11.4%
27	50.6%	64.2%	+13.6%	55.1%	68.3%	+13.2%
17	65.7%	84.2%	+18.5%	70.4%	87.5%	+17.1%

Figure 3 illustrates the impact of different shadow shapes — polygonal and triangular — on DINO attention maps. In each pair, the left image shows the adversarial shadow overlay, and the right shows the corresponding attention response. Brighter regions in the attention maps indicate areas of high model focus. Notably, the shadows are strategically positioned over these high-attention regions to suppress activation, thereby reducing the model's confidence in key semantic areas.

To better understand the qualitative and quantitative impact of DINO attention supervision, Figure 4 compares adversarial shadows generated with ($\lambda = 0.5$) and without ($\lambda = 0$) attention loss guidance, applied to the same 50 km/h speed limit sequence. In the top row, where DINO supervision is active, the generated shadows are smaller and more targeted, preserving most of the digit "5" while subtly distorting key regions, leading to misclassification with minimal visual disturbance. In contrast, the bottom row shows shadows optimized solely for misclassification, where the shadow is larger and indiscriminately darkens nearly half of the sign, including the full digit "5".

To quantify the visual subtlety of these perturbations, we compute the L2 distance between each adversarial image and its clean counterpart. Sequences generated with attention guidance ($\lambda = 0.5$) exhibit significantly lower L2 distances (mean = 0.165) compared to those without ($\lambda = 0$). This confirms that attention-aware optimization not only enhances attack consistency but also produces more localized and stealthy perturbations, supporting the notion that disrupting internal model focus can improve the efficiency and plausibility of physical-world adversarial attacks.

5. Conclusion

In this work, we proposed a novel framework for generating temporally coherent adversarial shadows targeting traffic sign recognition models. Unlike prior single-frame attacks, our approach operates over entire image sequences, simulating real-world scenarios where a sign is viewed from varying distances. By keeping the shadow spatially consistent in shape and opacity while allowing its scale to evolve across time, we create visually plausible perturbations that persist across frames. We use a non-differentiable genetic algorithm to search over shadow configurations, guided by a multi-objective loss that combines misclassification confidence with attention deviation based on DINO ViT attention maps. This dual-objective formulation not only degrades classification performance but also disrupts the model's internal reasoning, enhancing interpretability and impact. Experiments on the GTSRB benchmark demonstrate that incorporating attention supervision significantly boosts attack performance. Under strict success criteria — requiring misclassification in at least τ out of T frames — the proposed method achieves up to 87.5% SL-ASR. Across all thresholds, SL-ASR improves from a range of 40.4%–70.4% (without attention) to 52.3%–87.5% (with attention), confirming the effectiveness of attention-guided shadow optimization for both triangle and polygon masks.

References

[1] Mohammed Aldeen, Pedram MohajerAnsari, Jin Ma, Mashrur Chowdhury, Long Cheng, and Mert D Pesé. Wip: A first look at employing large multimodal models against autonomous vehicle attacks. In *ISOC Symposium on Vehicle Security and Privacy (VehicleSec '24)*, 2024. 4

[2] Mohammed Aldeen, Pedram MohajerAnsari, Jin Ma, Mashrur Chowdhury, Long Cheng, and Mert D. Pesé. An initial exploration of employing large multimodal models in defending against autonomous vehicles attacks. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 3334–3341, 2024. 4

[3] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, and Mani Srivastava. Genattack: Practical black-box attacks with gradient-free optimization. In *GECCO*, 2019. 3

[4] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song. Practical black-box attacks on deep neural networks using efficient query mechanisms. In *ECCV*, 2018. 3

[5] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch, 2017. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. 1

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 2

[7] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *CCS*, 2017. 3

[8] Zhe Chen and Zijing Chen. Rbnet: A deep neural network for unified road and road boundary detection. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part I*, page 677–687, Berlin, Heidelberg, 2017. Springer-Verlag. 1

[9] Noah T. Curran, Minkyoung Cho, Ryan Feng, Liangkai Liu, Brian Jay Tang, Pedram MohajerAnsari, Alkim Domeke, Mert D. Pesé, and Kang G. Shin. Achieving the safety and security of the end-to-end av pipeline. In *Proceedings of the 2024 Cyber Security in CarS Workshop*, page 13–24, New York, NY, USA, 2024. Association for Computing Machinery. 2

[10] Yinpeng Dong, Huanqian Yan, Xingxing Wei, Hang Su, and Jun Zhu. Adversarial examples in modern deep learning: A systematic evaluation of threats and defenses, 2023. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(8), 9257-9271. 1

[11] 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, 2018. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1625-1634. 1, 2

[12] David Fernandez, Pedram MohajerAnsari, Amir Salarpour, and Mert D Pesé. Avoiding the crash: A vision language model evaluation in critical traffic scenarios. Technical report, SAE Technical Paper, 2025. 4

[13] Ruth Fong and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *ICCV*, 2019. 3

[14] Jindong Gu, Volker Tresp, and Yao Qin. Are vision transformers robust to patch perturbations? *arXiv preprint arXiv:2111.10659*, 2021. 2

[15] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017. 4

[16] John H Holland. Genetic algorithms. *Scientific american*, 267(1):66–73, 1992. 2

[17] Sebastian Houben, Johannes Stallkamp, Jan Salmen, Marc Schlipsing, and Christian Igel. Detection of traffic signs in real-world images: The German Traffic Sign Detection Benchmark. In *International Joint Conference on Neural Networks*, number 1288, 2013. 2

[18] 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 *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3903–3912, 2024. 2, 3

[19] Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. Safety verification of deep neural networks, 2017. 1

[20] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018. 3

[21] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *ICLR*, 2019. 3

[22] Michael Kühr, Mohammad Hamad, Pedram MohajerAnsari, Mert D. Pesé, and Sebastian Steinhorst. Sok: Security of the image processing pipeline in autonomous vehicles, 2024. 2

[23] Yufeng Li, Fengyu Yang, Qi Liu, Jiangtao Li, and Chenhong Cao. Light can be dangerous: Stealthy and effective physical-world adversarial attack by spot light. *Computers & Security*, 132:103345, 2023. 2

[24] Jianbo Lin, Yulong Song, Zhe He, Zheng Wang, Yao Fei, Shu-Tao Wang, and Zhenyu Xu. Nattack: Learning the distributions of adversarial examples for an improved black-box attack. In *ICLR*, 2019. 3

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 3

[26] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey, 2022. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7), 3523-3542. 1

[27] Pedram MohajerAnsari, Alkim Domeke, Jan de Voor, Arkajyoti Mitra, Grace Johnson, Amir Salarpour, Habeeb Olufowobi, Mohammad Hamad, and Mert D Pesé. Discovering new shadow patterns for black-box attacks on lane detection of autonomous vehicles. *arXiv preprint arXiv:2409.18248*, 2024. 2

[28] Marzieh Mohammadi and Amir Salarpour. Point-gn: A non-parametric network using gaussian positional encoding for point cloud classification. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3487–3496. IEEE, 2025. 1

[29] Marzieh Mohammadi, Amir Salarpour, and Pedram MohajerAnsari. Point-ln: A lightweight framework for efficient point cloud classification using non-parametric positional encoding, 2025. 4

[30] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016. 3

[31] Scott D. Pendleton, H. Andersen, X. Du, X. Shen, Malika Meghjani, Y. Eng, Daniela Rus, and Marcelo H. Ang Jr. Autonomous driving: The future of automobiles. *IEEE Transactions on Intelligent Vehicles*. 1

[32] SM Moein Peyghambarzadeh, Fatemeh Azizmalayeri, Hassan Khotanlou, and Amir Salarpour. Point-planenet: Plane kernel based convolutional neural network for point clouds analysis. *Digital Signal Processing*, 98:102633, 2020. 2

[33] Peng Qin, Huachun Tan, Hefeng Li, and Xianglin Wen. Deep reinforcement learning car-following model considering longitudinal and lateral control. *Sustainability*, 14(24):16705, 2022. 1

[34] Gabriëlle Ras, Ning Xie, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–396, 2022. 3

[35] Ratheesh Ravindran, Michael J. Santora, and Mohsin M. Jamali. Multi-object detection and tracking, based on dnn, for autonomous vehicles: A review. *IEEE Sensors Journal*, 21 (5):5668–5677, 2021. 1

[36] Amir Salarpour, Hassan Khotanlou, and Nikolaos Mavridis. Long-term estimation of human spatial interactions through multiple laser ranging sensors. In *2014 International Conference on Robotics and Emerging Allied Technologies in Engineering (iCREATE)*, pages 109–114. IEEE, 2014. 1

[37] Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. Planning and decision-making for autonomous vehicles, 2018. Annual Review of Control, Robotics, and Autonomous Systems, 1(1), 187-210. 1

[38] Guoqing Tao, Shiyu Chang, Mo Yu, Xiaoxuan Zhang, and Bimal Viswanath. Attacks meet interpretability: Attribute-steered detection of adversarial samples. In *NeurIPS*, 2018. 2

[39] Donghua Wang, Haoyu Zhang, Jiadong Zhang, Yiming Ma, Zhenyu Xie, Yang Liu, Yi Yang, and Zhifeng Liu. Rfla: A stealthy reflected light adversarial attack in the physical world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16265–16274, 2021. 2

[40] Donghua Wang, Wen Yao, Tingsong Jiang, Chao Li, and Xiaoqian Chen. Rfla: A stealthy reflected light adversarial attack in the physical world, 2023. 1

[41] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, pages 5286–5295. PMLR, 2018. 3

[42] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil K. Jain. Adversarial attacks and defenses in images, graphs and text: A review, 2023. International Journal of Automation and Computing, 17, 151-178. 1

[43] 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, 2022. 1, 2

[44] Chenchen Zhu, Fanyi Wang, and Huijuan Xu. Deep learning for object detection: A comprehensive review, 2023. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3), 1322-1351. 1