SASA: Sequence-Aware Shadow Attacks via Attention Alignment for Traffic Sign Recognition

Anonymous CVPR submission

Paper ID 21

Abstract

*We propose **SASA** (Sequence-Aware Shadow Attack), a black-box adversarial framework that uses physically realistic, differentiable shadow patterns to deceive traffic sign recognition systems. Unlike prior image-based attacks, SASA targets video sequences—common in real-world driving—by generating smooth, temporally consistent shadows that remain visually plausible and imperceptible to humans. Guided by attention maps from frozen vision transformers, SASA aligns shadow placement with semantically salient regions without querying the target model. Evaluated on the GTSRB dataset, SASA reduces classification accuracy by up to 86% and sequence-level accuracy by over 90% on black-box models, including CNNs and ViTs. The method generalizes across architectures, preserves perceptual quality, and reveals a novel vulnerability in sequential vision systems.*

1. Introduction

Adversarial examples have exposed critical vulnerabilities in deep neural networks, raising significant safety concerns in domains such as autonomous driving. While most adversarial research has focused on imperceptible perturbations applied to static images [9, 21], real-world perception systems typically operate over temporally structured inputs like video streams. In such scenarios, frame-level perturbations often fail to generalize due to temporal filtering, camera dynamics, and perceptual inconsistencies [3, 7].

In the context of traffic sign recognition (TSR), physical-world attacks have demonstrated real-world feasibility using printed posters [8], adversarial stickers [19], and more recently, light projection [25] and shadow casting [26]. These techniques highlight a shift toward stealthy, naturally plausible perturbations that can deceive models without arousing human suspicion. Among these, cast shadows are particularly compelling: they are common in real driving scenes, inherently persistent across frames, and do not require modifying the object itself. Figure 1 presents visual comparisons of different light-based physical attacks, including our own shadow variants.

In this paper, we propose **SASA** (*Semantically Aligned Shadow Attack*), a novel black-box, video-based adversarial attack that leverages physically plausible shadows to induce misclassification in TSR systems. Unlike prior work that either targets single images [26] or assumes white-box access [15], SASA generates a temporally consistent shadow mask that aligns with semantically important regions, guided by frozen transformer attention maps (DINO and DeiT), without requiring any access to the target model.

Our approach integrates three components: (1) a compact, physically inspired shadow generator with spatiotemporal control, (2) a differentiable attention-guided alignment strategy, and (3) an optimization routine that ensures realism and temporal consistency. We evaluate SASA on both convolutional (CNN, STN) and transformer-based (EffNet, ViT) TSR models. Under a strict black-box setting, SASA achieves up to **90% sequence-level accuracy drop** on ViT models and maintains high effectiveness across all architectures—even without direct model access. Among all variants, our DeiT-guided StripShadow delivers the strongest results, lowering average frame-level accuracy by over **30 percentage points** on modern TSR systems. Crucially, the generated shadows remain visually natural and consistent across frames, maintaining realism while significantly degrading model performance.

Contributions. Our main contributions are as follows:

- We introduce SASA, the first fully black-box, temporally coherent shadow-based adversarial attack tailored for sequential TSR systems.
- We design a differentiable and physically grounded shadow generation module that allows compact control over shape, position, and opacity.
- We guide perturbation placement using saliency maps from frozen vision transformers (DINO and DeiT), aligning shadows with semantically meaningful regions.
- We demonstrate that SASA generalizes across architectures and achieves stealthy, transferable attacks under real-world constraints, reducing TSR accuracy by up to **86%**

Figure 1. Visual comparison of light-based physical attacks including shadow casting [26], reflected laser [23], spotlight projection [25], and natural illumination [10].

on ViT and over **60%** on EffNet models.

2. Related Work

Adversarial Attacks on Vision Models. Deep neural networks are vulnerable to adversarial attacks in both digital and physical settings. Traditional white-box [9, 14] and black-box [11, 16] strategies generate subtle perturbations at the pixel level. However, these often fail under real-world transformations [2]. Recent evaluations [17] stress the need to move beyond simplistic baselines (e.g., LISA-CNN, GTSRB-CNN) and assess attacks under realistic, sequential, and physically plausible conditions, especially for traffic sign recognition (TSR).

Physical-World and Light-Based Attacks. A growing body of work focuses on physical adversarial examples, including posters [18], stickers [8], and camouflage [4]. More recently, light-based attacks like RFLA [23] and AdvSL [25] introduced adversarial perturbations using sunlight and spotlights, demonstrating high stealth and feasibility in outdoor settings. These methods avoid object modification and offer attacker-controlled timing and adaptability. Natural light attacks [10] further reveal that everyday illumination variations alone can degrade TSR reliability.

Shadow-Based and Temporally Coherent Attacks. Shadow perturbations have emerged as a subtle, physically grounded threat. Zhong et al. [26] showed triangular shadows can achieve over 90% attack success in TSR. Building on this, MohajerAnsari et al. [15] proposed a black-box video attack using a temporally scaled shadow with fixed shape and opacity. Their work incorporates saliency supervision from DINO transformers and introduces the Sequence-Level Attack Success Rate (SL-ASR) to assess robustness across video frames.

Attention and Saliency-Guided Attacks. Attention maps derived from vision transformers (ViTs) have been used to guide adversarial perturbations in a more interpretable manner [13, 24]. The ShadowSeq framework [15] extends this to traffic sign videos, using a dual-loss objective that targets both classification and attention misalignment. These attacks disrupt not only outputs but also internal visual reasoning.

3. Methodology

We propose a physically grounded adversarial attack that constructs temporally coherent shadow patterns across video sequences of traffic signs. Our method, **SASA** (Semantically Aligned Shadow Attack), comprises three core components: (1) a differentiable shadow generator, (2) attention-guided placement informed by pretrained vision transformers, and (3) an optimization strategy that aligns shadows with semantically salient regions—without requiring access to the target classifier.

3.1. Problem Formulation

Conventional adversarial attacks in vision often target single-frame inputs with unconstrained pixel-level perturbations. However, real-world systems like traffic sign recognition (TSR) process temporally structured video streams, relying on frame-to-frame consistency to reject transient noise. Frame-wise perturbations typically lead to visual flickering or temporal artifacts, which are easily suppressed by post-processing or tracking modules. SASA instead introduces a physically plausible attack that applies a single, temporally consistent shadow transformation across all frames. By mimicking natural lighting effects—such as occlusion from poles or foliage—these shadows maintain spatial coherence while avoiding perceptual artifacts.

Formally, let $\{x_1, \ldots, x_T\}$, where $x_t \in [0,1]^{3\times H\times W}$, denote a video sequence. Let $f : \mathbb{R}^{3\times H\times W} \to \mathcal{Y}$ be a TSR classifier. We seek a shared shadow function $\mathcal{S}(x_t; \theta, \gamma)$ such that each perturbed frame $\tilde{x}_t = \mathcal{S}(x_t)$ is misclassified: $f(\tilde{x}_t) \neq f(x_t)$ for all t. Here, θ defines geometric parameters of the shadow, and $\gamma \in [0,1]$ controls its intensity. This formulation permits efficient, differentiable optimization of realistic shadows across time.

3.2. Threat Model

We assume a strict black-box threat model: the attacker has no access to model weights, architecture, gradients, or predictions. Unlike white-box methods that rely on backpropagation through the target, SASA utilizes frozen attention maps extracted from transformer models—specifically DINO and DeiT—fine-tuned on the TSR dataset. These models are used solely to guide the shadow placement. Importantly, the final adversarial shadows are evaluated against entirely unseen

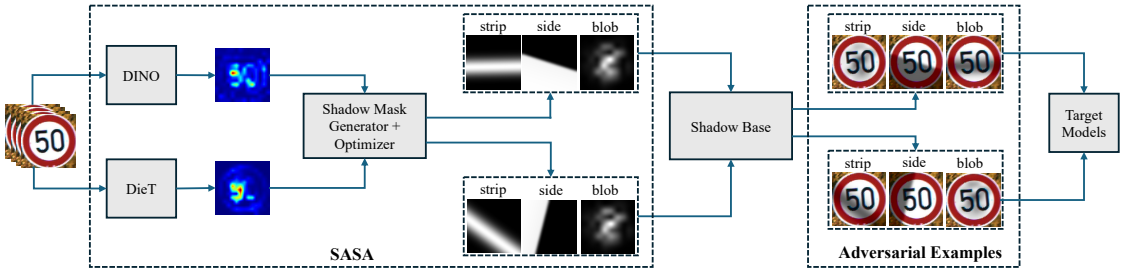

Figure 2. Overview of the SASA pipeline. A clean video sequence is processed by a frozen attention model (DINO/DeiT) to produce a fused saliency map. This map guides the optimization of a shared, differentiable shadow mask. The mask is applied uniformly across all frames, and the resulting perturbed sequence is evaluated against unseen TSR classifiers (e.g., CNNs, ViTs) in a strict black-box setting.

target models, ensuring a realistic and rigorous black-box setting.

3.3. The SASA Framework

SASA produces a single shadow mask that is spatially aligned and temporally consistent across an entire video sequence. The mask is optimized via a compact set of differentiable parameters, guided by pretrained attention maps.

The pipeline comprises four key stages:

1. **Attention Aggregation.** A frozen DeiT or DINO transformer extracts spatial attention maps from each frame x_t. These are fused via temporal max-pooling into a single 2D heatmap.
2. **Shadow Generation.** A shadow generator (*BlobShadow*, *StripShadow*, or *SideShadow*) generates $M \in [0,1]^{H\times W}$ via a differentiable function $\text{Gen}(\theta)$, parameterized by compact geometric variable θ. The same mask is applied to every frame using the transformation $\mathcal{S}(x_t;\theta,\gamma)$.
3. **Attention-aligned Optimization.** The mask M is optimized to match the fused attention map A, using a top-k alignment loss $\mathcal{L}_{\text{align}}$. Crucially, this optimization does not require target model access.
4. **Black-box Evaluation.** The resulting adversarial shadow is tested on multiple held-out TSR classifiers (e.g., CNNs, ViTs), across a range of intensities γ, with no further tuning.

This approach decouples attack generation from model-specific gradients, yielding transferable, physically grounded, and temporally smooth adversarial sequences. An overview of the complete pipeline is shown in Figure 2. We now detail the shadow generation process and the physical priors embedded in our design.

3.4. Differentiable Shadow Generation

SASA employs a physically motivated shadow generator that simulates real-world occlusions (e.g., from trees, poles, overpasses). Rather than perturbing individual pixels, SASA learns interpretable, low-dimensional parameters that define soft shadow masks $M \in [0,1]^{H\times W}$.

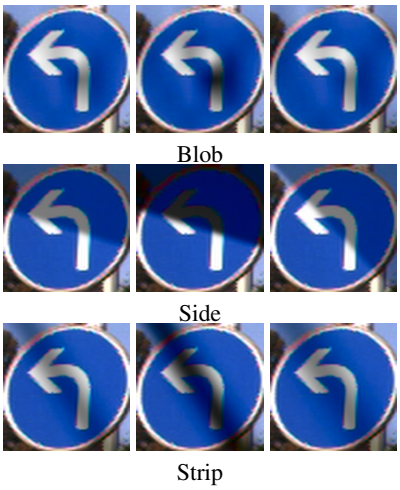

Figure 3. Visual comparison of different shadow styles: Blob, Side, and Strip Shadow. Each sequence shows shadow behavior under varying parameters.

Masks are applied uniformly across the sequence to preserve temporal consistency. Each shadow is rendered in CIELAB space, darkening only the luminance channel:

$$L' = L \cdot (1 - \gamma M), \quad a' = a, \quad b' = b$$

This formulation darkens the luminance channel proportional to the mask while preserving color consistency. The image is then converted back to RGB and clamped to a valid range. To ensure perceptual realism, we apply a Gaussian falloff around the shadow boundaries:

$$F(d) = \exp\left(-\left(\frac{d}{r}\right)^2 \cdot \tau\right)$$

where d is distance from the mask center, r is radius, and τ controls softness.

SASA supports three shadow styles—Blob, Strip, and Side—described in detail next.

Algorithm 1 Shadow Mask Generation: BLOBSHADOW and SIDESHADOW

Require: Image size (H, W), type $t \in \{\text{blob}, \text{side}\}$, parameters θ, blur kernel size

if $t = \text{blob}$ **then**
Require: $\theta = (c_x, c_y, s)$, seed resolution k
  Sample noise tensor $N \in \mathbb{R}^{1\times1\times k\times k}$
  Upsample $N \to \mathbb{R}^{1\times1\times H\times W}$ via bilinear interpolation
  Apply Gaussian blur; normalize to $[0, 1]$
  Generate radial falloff at (c_x, c_y) with radius $r = s \cdot H$
  Multiply blob by radial falloff and normalize
else if $t = \text{side}$ **then**
Require: $\theta = (c_x, c_y, \alpha, f)$, blend sharpness κ
  Create coordinate grid centered at (c_x, c_y); rotate by α to get y_{rot}
  Define half-plane masks: $M_1 = \mathbb{1}[y_{\text{rot}} > 0]$, $M_2 = \mathbb{1}[y_{\text{rot}} < 0]$
  $M \leftarrow \sigma((f-0.5)\kappa) \cdot M_2 + (1 - \sigma((f-0.5)\kappa)) \cdot M_1$
  Apply Gaussian blur and distance-based falloff
end if
return Shadow mask $M \in [0, 1]^{H\times W}$

BlobShadow: Radial Shadows. *BlobShadow* generates soft, amorphous occlusions resembling natural phenomena such as tree cover or cloud shadows. It is parameterized by $\theta = (c_x, c_y, s)$, where $(c_x, c_y) \in [0, 1]^2$ defines the mask center, and $s \in [0.1, 0.8]$ controls spatial extent. The result is a diffuse, circular shadow with smooth falloff, well-suited for simulating irregular, organic occlusions (see Figure 3, left).

StripShadow: Elongated Directional Shadows. *StripShadow* emulates shadows cast by upright objects like poles, signposts, or fences. It is parameterized by $\theta = (c_x, c_y, \alpha, s)$, where $\alpha \in [0, \pi]$ controls orientation and s specifies the width of the shadow band. The resulting mask supports sharp directional control, making it effective for adversarial occlusion of vertically aligned semantic features (see Figure 3, right).

SideShadow: Lateral Projection Shadows. *SideShadow* produces oblique or asymmetric shadows cast from lateral occluders such as roadside barriers or parked vehicles. It is defined by $\theta = (c_x, c_y, \alpha, f)$, where $f \in [0, 1]$ is a continuous flip factor interpolating between left- and right-projected masks. Two directional masks are blended using sigmoid-weighted interpolation, followed by Gaussian smoothing. This design allows nuanced, asymmetric shading and captures edge cases often overlooked by uniform shadow models (see Figure 3, middle). Implementation pseudocode for both generators is included in Algorithm 1. To ensure that these shadows align with semantically meaningful features, we rely on saliency signals from pretrained transformers.

3.5. Attention-Guided Shadow Alignment

To ensure shadows occlude semantically critical content, SASA leverages attention maps extracted from pretrained vision transformers. We adopt two complementary models—DINO [5] and DeiT [22]—which capture different forms of visual saliency. DINO, trained via self-supervised contrastive learning, produces object-centric maps, while DeiT, trained in a supervised manner, tends to localize class-discriminative regions. Both models are fine-tuned on the TSR dataset and remain frozen during optimization.

Attention Rollout and Temporal Fusion. We use the attention rollout method [1] to extract dense saliency maps from transformer layers. At each layer l, we compute residual-aware attention:

$$\bar{A}^{(l)} = \alpha A^{(l)} + (1 - \alpha)I,$$

where $\alpha = 0.9$ balances the raw attention matrix $A^{(l)} \in \mathbb{R}^{N\times N}$ with identity flow I. Cumulative attention is obtained via matrix multiplication across layers:

$$\tilde{A} = \bar{A}^{(1)}\bar{A}^{(2)} \cdots \bar{A}^{(L)}.$$

We extract the first row (class-to-token attention), discard the CLS token, and reshape the remaining values into a 2D map $A_t \in \mathbb{R}^{h\times w}$. Each A_t is upsampled to the input resolution $H \times W$. To consolidate information over time, we perform temporal max-pooling:

$$A = \max_{t\in\{1,\ldots,T\}} A_t.$$

Optionally, we apply a 2D Gaussian prior centered in the image to bias attention toward regions where traffic signs are commonly located:

$$G(x, y) = \exp\left(-\lambda\left(\frac{(x - c_x)^2}{W^2} + \frac{(y - c_y)^2}{H^2}\right)\right),$$
$$A(x, y) \leftarrow A(x, y) \cdot G(x, y). \quad (1)$$

Top-k Alignment Loss. To focus optimization on the most influential pixels, we define a top-k alignment loss. Let $\mathcal{T}_k$ be the indices of the top-k values in the attention map A. The shadow mask M is optimized to overlap with these salient regions:

$$\mathcal{L}_{\text{align}} = \frac{1}{k}\sum_{i\in\mathcal{T}_k} |M_i - A_i|^p,$$

where $p = 1$ corresponds to L1 loss and $p = 2$ to mean squared error. This design encourages sparse but targeted occlusion, yielding compact shadows that effectively disrupt classification.

3.6. Full Optimization Objective

The final loss function combines semantic alignment with optional regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{align}} + \lambda_{\text{reg}} \cdot \mathcal{L}_{\text{reg}}.$$

Here, $\mathcal{L}_{\text{reg}}$ may include total variation, sparsity, or spatial smoothness penalties to enhance physical plausibility. Importantly, SASA requires no gradients or queries from the target classifier; all optimization is driven by attention-derived signals, preserving the black-box threat model while supporting efficient and transferable adversarial attacks.

Together, these components form a compact, transferable, and physically realizable adversarial attack framework suitable for black-box, real-world TSR systems.

4. Experiments

We evaluate **SASA** in a strict black-box setting on the traffic sign recognition (TSR) task. Our experimental goals are fourfold: (1) quantify attack effectiveness across a range of model architectures, (2) assess temporal consistency and perceptual realism of shadow-based perturbations, (3) evaluate transferability to unseen models without adaptation, and (4) study the impact of key design choices through targeted ablations.

Our analysis spans three shadow types, multiple shadow intensities, and two pretrained attention backbones (DeiT and DINO), offering a comprehensive view into SASA's effectiveness, generalization, and physical plausibility.

4.1. Experimental Setup

Dataset and Preprocessing We use the **German Traffic Sign Recognition Benchmark (GTSRB)** [20], a real-world dataset consisting of over 50,000 traffic sign images categorized into 43 classes. Each sample includes bounding box annotations and metadata that support constructing temporally adjacent sequences of the same sign.

To simulate sequential video input, we group images with the same class and sample ID to form short sequences. Each sequence is standardized to a fixed length of $T = 30$ frames using a sliding window approach. Images are cropped based on bounding boxes and resized according to model-specific input dimensions: 128×128 for CNN-based models and 224×224 for transformer-based models. Color and luminance statistics (e.g., average LAB and HSV values) are computed for each frame to enable filtering and ablation control.

We curate a high-visibility evaluation subset by discarding sequences with low average luminance ($L < 120$ in CIELAB space), ensuring that generated shadows remain perceptually distinguishable. All train/test splits are class-stratified and fixed across experiments to support reproducibility.

Target Models We evaluate SASA against a diverse set of TSR classifiers spanning convolutional and transformer-based architectures.

CNN-based models. We include two widely used baselines: **GTSRB-CNN** [6] and **GTSRB-STN** [12], reimplemented and trained from scratch on the GTSRB training set. These models process inputs at 128×128 resolution and serve as classical vision backbones for TSR.

Transformer-based models. We fine-tune ImageNet-pretrained **EfficientNet-B0** and **ViT-Small** models on GTSRB, adapting them to traffic sign classification via end-to-end supervised training. Both models use inputs resized to 224×224. This setup allows us to evaluate SASA's performance on modern attention-based architectures under a black-box constraint.

Attention models. DeiT [22] and **DINO** [5] are used exclusively as frozen saliency extractors. These models are fine-tuned on GTSRB to improve attention map fidelity, but are not used as targets during evaluation. Their outputs guide shadow placement during optimization without accessing the target model, thus strictly adhering to the black-box threat model.

Evaluation Metrics We evaluate SASA's effectiveness using two complementary metrics that reflect both frame-level performance and sequence-level robustness.

Frame-wise Accuracy. This metric computes the average classification accuracy across all individual frames in a sequence. It serves as a fine-grained measure of how consistently a model classifies perturbed inputs on a per-frame basis.

Sequence-level Accuracy at 50% (`seq@50`). This stricter metric assesses whether a majority of frames in a sequence are correctly classified. A sequence is considered correct if at least 50% of its frames yield the correct label. `seq@50` captures temporal robustness and provides a more task-relevant evaluation for video-based TSR, where transient misclassifications may be tolerable, but sustained errors are not.

Together, these metrics allow us to assess not only the raw effectiveness of shadow-based perturbations but also their practical impact on real-world TSR pipelines.

4.2. Evaluation Results

We evaluate SASA across four black-box TSR models: GTSRB-CNN, GTSRB-STN, EfficientNet-B0, and ViT. Table 1 reports performance drops in frame-wise accuracy and sequence-level accuracy (`seq@50`) at three shadow intensities ($\gamma = \{0.2, 0.5, 0.8\}$), relative to the clean baseline.

Note on Comparability. We do not include direct comparisons with prior shadow-based attacks such as [26], as these methods are designed to attack single-frame images using per-image optimization. In contrast, **SASA** targets full video sequences by optimizing a single, temporally consistent shadow mask. This fundamental difference in formulation makes direct comparisons methodologically incompatible.

Table 1. Classification accuracy (%) and sequence accuracy at 50% threshold (`seq@50`, %) for clean sequences and relative drop (↓) under each shadow variant. For each intensity level ($\gamma = 0.2, 0.5, 0.8$), the highest drop per model (i.e., strongest attack) is **bolded**.

Intensity	Variant	CNN		STN		EffB0		ViT	
		Acc ↓	seq@50 ↓	Acc ↓	seq@50 ↓	Acc ↓	seq@50 ↓	Acc ↓	seq@50 ↓
–	Clean Sequence	96.6	100.0	95.8	100.0	99.9	100.0	95.1	100.0
0.2	BlobShadow (DeiT)	0.6	0.0	2.4	2.5	1.1	0.0	4.6	**5.0**
	BlobShadow (DINO)	0.6	0.0	**2.7**	2.5	1.0	0.0	3.8	2.5
	SideShadow (DeiT)	**2.2**	0.0	2.4	2.5	0.5	0.0	**5.6**	2.5
	SideShadow (DINO)	0.9	0.0	1.9	2.5	0.3	0.0	2.7	2.5
	StripShadow (DeiT)	1.3	0.0	2.6	2.5	**2.5**	0.0	4.0	2.5
	StripShadow (DINO)	0.5	0.0	2.6	2.5	0.7	0.0	1.4	0.0
0.5	BlobShadow (DeiT)	9.0	7.5	12.4	10.0	8.4	5.0	20.3	12.5
	BlobShadow (DINO)	8.6	5.0	12.9	7.5	6.5	2.5	21.0	17.5
	SideShadow (DeiT)	11.8	**7.5**	14.3	15.0	9.1	7.5	**32.5**	35.0
	SideShadow (DINO)	8.8	2.5	10.5	7.5	8.2	2.5	24.2	20.0
	StripShadow (DeiT)	**12.4**	5.0	**20.2**	**20.0**	**30.8**	**27.5**	31.4	**37.5**
	StripShadow (DINO)	7.0	2.5	15.8	15.0	17.8	15.0	27.6	27.5
0.8	BlobShadow (DeiT)	22.8	22.5	30.1	25.0	21.7	12.5	75.3	80.0
	BlobShadow (DINO)	21.6	22.5	31.7	25.0	22.1	15.0	75.8	90.0
	SideShadow (DeiT)	35.1	35.0	44.2	45.0	33.3	32.5	80.6	87.5
	SideShadow (DINO)	35.3	40.0	43.2	45.0	35.5	27.5	75.8	80.0
	StripShadow (DeiT)	**40.6**	**47.5**	**46.9**	**47.5**	**62.0**	**62.5**	**86.3**	**92.5**
	StripShadow (DINO)	27.3	25.0	44.7	45.0	40.1	35.0	78.8	85.0

Table 2. Classification accuracy (%) and sequence accuracy at 50% (`seq@50`, %) averaged over all shadow intensities ($\gamma \in [0.1, 1.0]$). Gray-box evaluations (DeiT and DINO classifiers) are shown in the last columns. The lowest accuracy (i.e., strongest attack) for each model is highlighted in green.

Variant	CNN		STN		EffB0		ViT		DeiT		DINO	
	Acc	seq@50	Acc	seq@50	Acc	seq@50	Acc	seq@50	Acc	seq@50	Acc	seq@50
Clean (baseline)	96.6	100.0	95.8	100.0	99.9	100.0	95.1	100.0	98.8	100.0	92.8	100.0
BlobShadow (DeiT)	83.5	86.8	77.2	82.8	87.1	91.3	55.3	58.8	59.4	61.5	68.6	74.0
BlobShadow (DINO)	83.9	87.8	76.3	83.3	87.5	91.3	55.5	57.5	59.9	62.3	69.6	74.5
SideShadow (DeiT)	74.4	80.5	68.1	71.5	77.7	80.3	49.8	53.5	46.4	48.0	62.6	67.5
SideShadow (DINO)	76.2	79.8	70.7	74.3	79.2	81.5	54.9	58.5	51.3	52.8	65.7	70.3
StripShadow (DeiT)	**73.5**	**78.5**	**66.3**	**69.8**	**64.7**	**65.3**	**49.4**	**52.0**	**43.3**	**44.8**	**54.8**	**61.0**
StripShadow (DINO)	81.1	86.5	69.2	74.3	76.6	79.0	53.4	55.8	49.2	50.8	61.4	64.8

At low intensity ($\gamma = 0.2$), shadows are subtle, and no single configuration dominates. SideShadow (DeiT) yields the strongest effect on CNN and ViT, while other variants perform comparably. This suggests that at low visibility, effectiveness depends more on attention alignment than shadow shape.

At medium intensity ($\gamma = 0.5$), StripShadow (DeiT) consistently outperforms other variants, achieving the highest drop on all models except ViT, where SideShadow (DeiT) is slightly better. Its long pattern appears well-suited to overlap with high-attention regions, mostly on structured signs. At high intensity ($\gamma = 0.8$), StripShadow (DeiT) stays dominant, with the largest accuracy and seq@50 drops across all models—exceeding 60% on EfficientNet and 90% on ViT. SideShadow and BlobShadow also remain effective but fall short of StripShadow's disruption capacity. Overall, DeiT-guided variants outperform their DINO-based variants, confirming that class-discriminative attention better informs shadow placement. The results validate SASA's core design: structured shadows guided by transformer attention yield strong, transferable black-box attacks.

4.3. Average Performance Across Intensities

To evaluate the overall impact of shadow-based attacks, we compute the average classification accuracy and sequence-level accuracy (`seq@50`) across the full intensity range ($\gamma \in [0.1, 1.0]$). Table 2 reports these results for each shadow variant. The table includes both black-box settings, where the attention model differs from the classifier, and gray-box cases, where the same transformer (DeiT or DINO) is used for both guiding attention and optimizng shadow. This offers a unified view of attack efficacy across shadow types, models, and threat settings.

Among all variants, StripShadow with DeiT attention consistently yields the lowest average accuracy, particularly on transformer-based targets like ViT and DeiT, suggesting that

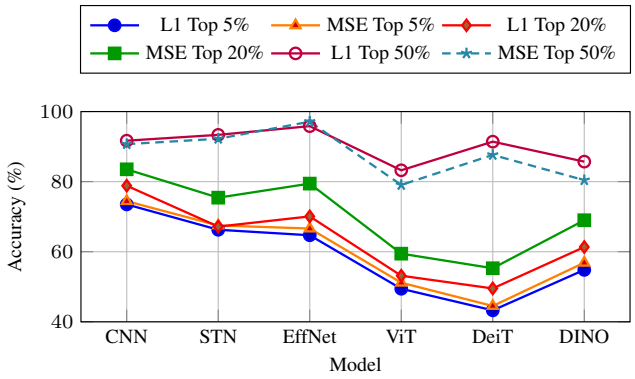

Figure 4. Final classification accuracy (lower is better) across six target models under different loss types and top-k attention coverage. Lower accuracy indicates stronger attacks.

this combination best aligns perturbations with the model's discriminative regions. Despite its simplicity, StripShadow is especially effective when high-intensity shadows are applied. ViT emerges as the most vulnerable target, likely due to its strong reliance on attention-based localization, which aligns with SASA's strategy. In contrast, GTSRB-CNN shows higher resilience across shadow types, highlighting the relative robustness of convolutional architectures under such physically inspired perturbations.

4.4. Loss Function Ablation

We ablate the effect of loss type and attention sparsity on SASA's attack effectiveness. Specifically, we compare six configurations: L1 and MSE alignment losses applied over the top-$k\%$ most salient attention values, with $k \in \{5, 20, 50\}$. All experiments use the StripShadow generator guided by DeiT attention, and results are averaged over shadow intensities $\gamma \in [0.1, 1.0]$.

Figure 4 reports final classification accuracy for each configuration across six target models. Since lower accuracy indicates more effective attacks, several trends emerge:

Larger top-k regions result in stronger attacks. The top-50% variants consistently yield the lowest accuracy across all models, confirming that broader attention coverage leads to more effective occlusion of discriminative regions.

MSE slightly outperforms L1 in low-k settings (e.g., top-5%), likely due to smoother gradient propagation during optimization. However, this difference diminishes as the attention mask grows denser.

Transformer models are more vulnerable. ViT and DeiT show the lowest final accuracies, suggesting greater susceptibility to attention-aligned perturbations. In contrast, CNN-based models retain higher accuracy, consistent with their reduced reliance on global saliency cues.

These results validate the design of SASA's top-k alignment loss and highlight the importance of both attention coverage and loss smoothness. Even simple losses like L1 and MSE can yield strong black-box attacks when guided by semantically meaningful saliency maps.

4.5. Qualitative Attention Alignment

To visualize how SASA targets model saliency, we present qualitative examples for each shadow type. Figure 5 shows three representative sequences—one per shadow module—highlighting the effect of attention-guided shadow placement and its impact on model focus.

Each row includes: (1) the original input frame, (2) the DeiT-derived attention map on the clean frame, (3) the optimized shadow mask, and (4) the attention map after applying the shadow. These visualizations demonstrate how shadows occlude high-saliency regions, leading to degraded or diffused attention in the perturbed frames.

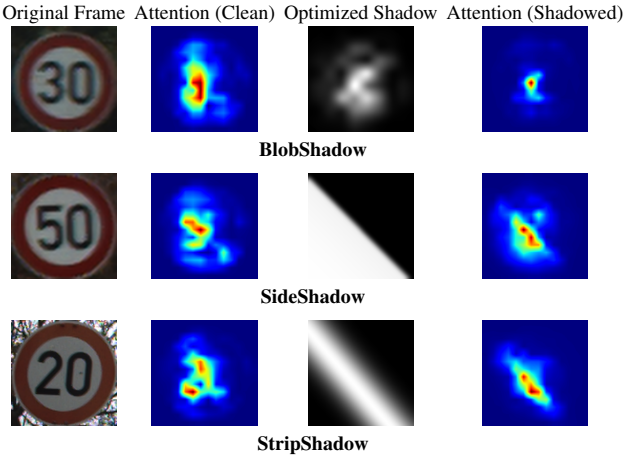

Figure 5. Qualitative attention analysis across shadow types. Each row shows: (1) the original frame, (2) the DeiT attention map before attack, (3) the optimized shadow mask, and (4) attention map after applying the shadow.

5. Conclusion

We introduced **SASA**, a black-box adversarial framework that generates physically plausible, temporally coherent shadows to mislead traffic sign recognition systems. By combining differentiable shadow generators with transformer-based attention maps, SASA aligns perturbations with semantically critical regions—without accessing model gradients or predictions. Our experiments on the GTSRB benchmark demonstrate that SASA significantly degrades classification performance across CNN and transformer models, while maintaining high perceptual realism and transferability. Ablation studies confirm the importance of attention-guided alignment and shadow diversity. Overall, SASA offers a scalable, interpretable, and physically grounded approach for evaluating real-world vulnerabilities in sequential vision systems.

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 4

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018. 2

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 1

[4] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4, 5

[6] Dan Ciregan, Ueli Meier, and Jürgen Schmidhuber. Multi-column deep neural networks for image classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3642–3649. IEEE, 2012. 5

[7] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4312–4321, 2019. 1

[8] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, and et al. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1625–1634, 2018. 1, 2

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

[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. 2

[11] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018. 2

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 5

[13] Hsueh-Ti Derek Liu, Michael Tao, Chun-Liang Li, et al. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. In *International Conference on Machine Learning (ICML)*, pages 224–233, 2018. 2

[14] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[15] Pedram MohajerAnsari, Amir Salarpour, David Fernandez, Cigdem Kokenoz, Bing Li, and Mert D Pesé. Attention-aware temporal adversarial shadows on traffic sign sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 2

[16] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. *arXiv preprint arXiv:1602.02697*, 1(2):3, 2016. 2

[17] Svetlana Pavlitska, Nico Lambing, and J Marius Zöllner. Adversarial attacks on traffic sign recognition: A survey. In *2023 3rd International conference on electrical, computer, communications and mechatronics engineering (ICECCME)*, pages 1–6. IEEE, 2023. 2

[18] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 2

[19] 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. 1

[20] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32: 323–332, 2012. 5

[21] 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

[22] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 4, 5

[23] Donghua Wang, Wen Yao, Tingsong Jiang, Chao Li, and Xiaoqian Chen. Rfla: A stealthy reflected light adversarial attack in the physical world. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4455–4465, 2023. 2

[24] Jie Wang, Zhaoxia Yin, Jing Jiang, and Yang Du. Attention-guided black-box adversarial attacks with large-scale multi-objective evolutionary optimization. *International Journal of Intelligent Systems*, 37(10):7526–7547, 2022. 2

[25] LI Yufeng, YANG Fengyu, LIU Qi, LI Jiangtao, and CAO Chenhong. Light can be dangerous: Stealthy and effective physical-world adversarial attack by spot light. *Computers & Security*, 132:103345, 2023. 1, 2

[26] 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