EXPLORING FULL-VIEW RGB-T VEHICLE ADVERSARIAL ROBUSTNESS IN THE PHYSICAL WORLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal RGB-Thermal (RGB-T) vehicle detectors are widely assumed to be robust owing to their complementary dual-modality design. However, their security under adversarial attacks remains insufficiently explored, particularly across the full viewpoint hemisphere. In this paper, we present the first full-angle physical adversarial attack against RGB-T vehicle detectors, covering azimuth angles of $0°$–$360°$ and elevation angles of $0°$–$90°$. We first conduct a comprehensive full-angle vulnerability analysis, revealing that existing attack methods suffer significant performance degradation under this challenging setting. To address the identified challenges of modality imbalance, angular imbalance, and digital-to-physical transfer gap, we propose three key techniques: (1) an adversarial shape-texture co-optimization (STCO) method that balances attack effectiveness across both modalities; (2) an angle-adaptive attack strategy (AAAS) that allocates more optimization effort to hard-to-attack viewpoints; and (3) a cross-modal foreground and background Expectation over Transformation (CFBE) method that enhances physical-world transferability. We further propose an angle-adaptive adversarial training method as a defense baseline. Extensive experiments in both digital and physical worlds, demonstrate the effectiveness of our attack against the state-of-the-art RGB-T detector and its strong transferability to unseen detectors.

1 INTRODUCTION

Multimodal RGB-Thermal (RGB-T) vehicle detection has emerged as a critical component in autonomous driving systems, offering distinct advantages over conventional single-modality RGB detectors. Under nighttime or adverse weather conditions, single RGB detectors are susceptible to environmental perturbations, whereas RGB-T detectors leverage the imaging robustness of thermal modality alongside the rich textural details of RGB modality, achieving stable detection performance across diverse environments. Owing to this complementary design, RGB-T detectors are commonly assumed to be highly robust, and their security has received relatively little scrutiny.

Unfortunately, a small but growing body of recent work (Wei et al., 2023a; Hu et al., 2025; Long et al., 2025) has demonstrated that these detectors remain vulnerable to serious security threats—for instance, carefully crafted physical adversarial patches can render RGB-T detectors unable to detect specific vehicles. However, the adversarial examples in these studies are effective only under limited viewing angles, such as top-down or lateral perspectives, and do not comprehensively reveal the robustness of RGB-T detectors across the full viewpoint hemisphere (azimuth $\theta \in [0°, 360°)$, elevation $\phi \in [0°, 90°]$).

To this end, we introduce a *full-angle adversarial testing* protocol for RGB-T detectors (detailed in Sec. 3.1). Under this protocol, we find that all existing attack methods suffer from substantial performance degradation (see Sec. 3.1 for details), regardless of whether they are evaluated under their original partial-coverage settings or extended to broader angular ranges. The underlying reason is that prior methods are designed for vehicle attacks at specific viewpoints where vehicle features are relatively stable. When extended to the full-angle setting, vehicle appearance varies dramatically—for example, the front, side, and top-down views of a vehicle exhibit vastly different feature distributions—making it difficult for existing 2D methods to transfer directly to the 3D full-angle setting.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

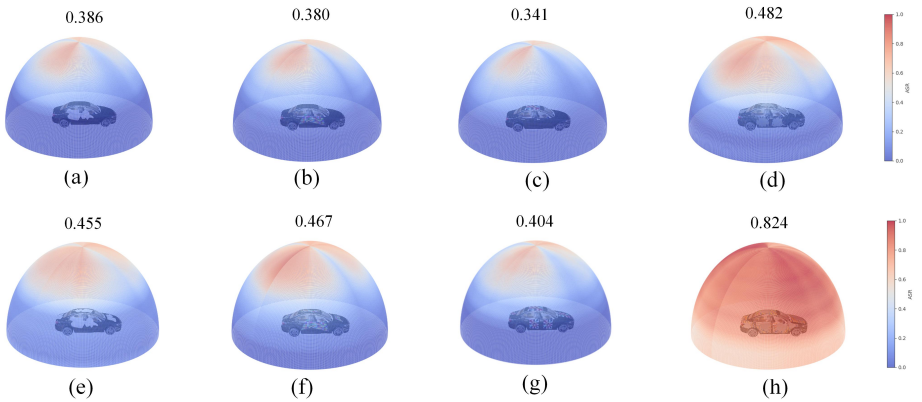


Figure 1: Full-view evaluation of different RGB-T attacks. Each hemisphere shows per-viewpoint ASR over azimuth $\theta \in [0°, 360°)$ and elevation $\phi \in [0°, 90°]$; the number above each panel is the aggregate ASR. Top row: single-patch of (a) UAP, (b) TOUAP, (c) CDUPatch and (d) the shape-only method. Bottom row: the multi-patch variants of (e) UAP, (f) TOUAP, (g) CDUPatch, and (h) ours.

Consequently, the full-angle robustness of RGB-T vehicle detectors remains an unsolved challenge. In other words, we seek to answer: *at which viewing angles do current RGB-T detectors fail, and to what degree?* Since the observation angle of road vehicles in real driving scenarios is inherently uncertain, a fine-grained angular vulnerability analysis carries significant practical implications for the safety of autonomous driving systems.

Recent thermal-only vehicle attacks (Zhu et al., 2024) show that geometry-aware shape optimization can sustain adversarial support under large viewpoint changes, raising a natural question: is geometry alone sufficient when the detector jointly processes visible and thermal inputs? Our diagnostic (Sec. 3.2) shows it is not, and exposes three challenges. **(1) Modality imbalance.** Shape-based optimization is effective against the thermal modality but weak against the RGB modality, as monochromatic contours lack the adversarial texture information that RGB attacks rely on. **(2) Angular imbalance.** Shape-only optimization attacks some viewpoints strongly (e.g., frontal, lateral) but others weakly (e.g., oblique front and rear), so full-angle performance is limited by its worst angles. **(3) Digital-to-physical transfer gap.** Even patterns that partially succeed digitally are hard to realize physically, indicating that dual-modality attacks demand substantially stronger generalization than single-modality ones.

We address modality imbalance with an adversarial *shape-texture co-optimization (STCO)* method, angular imbalance with an *angle-adaptive attack strategy (AAAS)*, and the transfer gap with *cross-modal foreground-background EOT (CFBE)* method. To our knowledge this is the first full-view physical adversarial attack against RGB-T vehicle detectors, which enables practitioners to run full-angle security analyses on deployed models. For instance, the state-of-the-art RGB-T detector DAMSDet (Guo et al., 2024) degrades markedly at the front and rear viewpoints relative to side and overhead views under adversarial testing (detection success rates are 38.2–55.7% lower), pointing to concrete scenarios that safer models must cover. Our contributions are:

1. Beyond prior angle-limited RGB-T attacks, we achieve full-angle adversarial attacks against multimodal RGB-T detectors, validated in both digital and physical worlds, together with a full-angle vulnerability analysis of current detectors.

2. We propose adversarial shape-texture co-optimization to balance adversarial effects across the RGB and thermal modalities, combined with an angle-adaptive attack strategy and a cross-modal foreground-background EOT scheme that respectively strengthen hard-to-attack viewpoints and physical transferability.

3. We propose angle-adaptive adversarial training as a defense baseline, achieving an attack success rate reduction 5.3 times that of standard adversarial training.

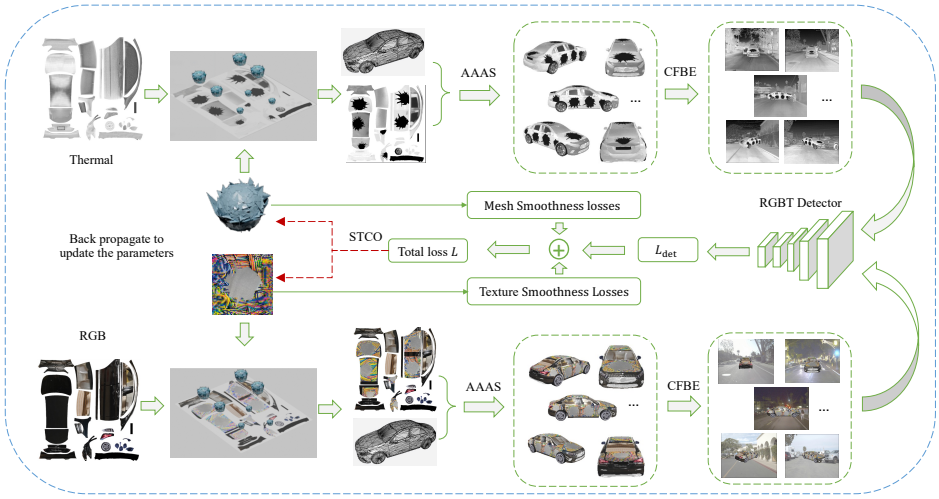


Figure 2: Overview of our full-view RGB-T adversarial attack pipeline.

2 Related Works

Single-modality vehicle attacks. RGB vehicle attacks optimize printable vehicle camouflage through universal patterns (Huang et al., 2020), in-the-wild optimization (Zhang et al., 2018), attention suppression (Wang et al., 2021), differentiable transformations (Suryanto et al., 2022), and full-coverage 3D camouflage (Wang et al., 2022; Liu et al., 2026). Thermal vehicle attacks employ infrared patches for top-down views (Wei et al., 2023b) or adversarial car stickers for multi-angle attacks (Zhu et al., 2024). However, these methods target individual modalities without jointly optimizing adversarial effects in RGB and thermal imagery.

RGB-T vehicle attacks. Only a few works address multimodal detectors, which is harder because the attack must balance two modalities while remaining physically realizable. A unified adversarial patch (Wei et al., 2023a) and a two-stage optimized patch (Hu et al., 2025) were both restricted to lateral views (azimuth within $-30°$ to $30°$), while a recent color-driven patch (Long et al., 2025) is limited to a top-down perspective. Full-view RGB-T vehicle attacks covering the entire viewpoint hemisphere are thus still absent from the literature.

3 Preliminary Analysis

3.1 Full-View Evaluation of Previous Methods

To evaluate existing RGB-T attacks under the full-view attack setting, we design a standardized full-view protocol. Following each method's original settings, we apply its adversarial pattern to our 3D RGB-T vehicle model (Sec. 4.2) and render image pairs with a virtual camera: azimuth θ is sampled every $\Delta\theta$ over $[0°, 360°)$ and elevation ϕ every $\Delta\phi$ over $[0°, 90°]$. To probe the upper bound of prior works, we additionally extend their single-patch configurations to multiple patches placed on different vehicle surfaces (doors, roof, hood), keeping all other settings fixed.

The results are shown in Figure 1. All prior methods are effective only within a limited angular window and fail over the rest of the hemisphere; extending them to multiple patches (e)–(g) raises ASR by only a few points without changing this profile, so the limitation does not stem from insufficient surface coverage. The underlying reason is that prior methods are optimized for specific viewpoints, whereas under the full-angle setting the appearance of a vehicle varies drastically for a 2D pattern to transfer. See Appendix A for more details.

3.2 SHAPE-ONLY RGB-T BASELINE

A natural way to obtain full-view RGB-T attacks is to reuse the shape optimization framework of the thermal-only full-view attack (Zhu et al., 2024) and optimize a unified RGB-T pattern. Given the chosen material (aluminum film), such a pattern appears dark in the thermal modality and silver-gray in the RGB modality, with both modalities sharing one adversarial shape; all other settings follow Sec. 3.1. We attacked a RGB-T detector (YOLOv9) on RGB and thermal channel separately. The results show that the shared contour is far more effective in the thermal modality than in the RGB one (0.617 vs. 0.492), revealing a clear modality imbalance, with results detailed in Appendix B. The reason may be that a shape-only pattern provides only a monochromatic silhouette and carries no texture information, which is essential for attacking the RGB modality.

4 METHODS

4.1 PROBLEM FORMULATION

Consider a target vehicle x observed simultaneously by a visible and a thermal sensor, yielding image pairs $(I_{\mathrm{vis}}, I_{\mathrm{thm}})$. An adversarial attack crafts perturbations producing $I_{\mathrm{vis}}^{\mathrm{adv}}$ and $I_{\mathrm{thm}}^{\mathrm{adv}}$ so that an RGB-T detector f fails to recognize x. We address the substantially harder *full-view* setting: with $\theta \in [0^\circ, 360^\circ)$ the azimuth and $\phi \in [0^\circ, 90^\circ]$ the elevation of the observer, detection must be suppressed across the entire viewpoint hemisphere. Given a confidence threshold q, a successful attack satisfies

$$f\big(I_{\mathrm{vis}}^{\mathrm{adv}}(\theta,\phi),\ I_{\mathrm{thm}}^{\mathrm{adv}}(\theta,\phi)\big) < q, \quad \forall\, \theta \in [0^\circ, 360^\circ),\ \phi \in [0^\circ, 90^\circ], \tag{1}$$

where $I_{\mathrm{vis}}^{\mathrm{adv}}(\theta,\phi)$ and $I_{\mathrm{thm}}^{\mathrm{adv}}(\theta,\phi)$ are the adversarial observations from viewpoint (θ,ϕ). This imposes omnidirectional robustness on the perturbation rather than success on a sparse subset of predefined orientations.

4.2 BUILDING A CO-REGISTERED 3D RGB-T VEHICLE MODEL

Following the 3D thermal vehicle modeling procedure of (Zhu et al., 2024), we extend it to the RGB-T setting by constructing a visible texture registered to the same mesh and UV coordinates. A single FLIR T560 captures paired visible and thermal images from multiple viewpoints; the visible observations are warped and stitched onto the unfolded UV layout, producing geometrically aligned RGB and thermal textures over the complete vehicle surface. This co-registered model underpins all subsequent full-view experiments. Construction and registration details are in Appendix C.

4.3 SHAPE-TEXTURE CO-OPTIMIZATION OF RGB-T ADVERSARIAL PATTERNS

We aim to design vehicle stickers with adversarial patterns to hide the vehicles from RGB-T detectors at various viewing angles (*full-angle*). As the analysis in Sec. 3.1 reveals, a purely shape-based optimization strategy cannot adequately balance the adversarial effects across both modalities. To address this, we propose a *shape-texture co-optmization (STCO)* method that jointly optimizes the geometric contour and the visual appearance of the adversarial pattern.

The key insight is the physical asymmetry between modalities. Owing to the low infrared emissivity of aluminum film ($\sim$0.1), the sticker's contour appears as a dark region in the thermal modality, making *shape* the dominant attack factor. In the RGB modality the same region appears silver-gray while the surrounding printed texture $\tau^{(i)}$ is fully perceivable with rich color information, so both the contour *shape* and the surrounding *texture* contribute. This motivates using shape as the shared mechanism across modalities and the surrounding texture as a complementary RGB-only channel.

Following prior work (Zhu et al., 2024), we generate 2D contours with a 3D adversarial mesh shadow attack (MSA): rather than optimizing 2D patterns directly, MSA optimizes a 3D adversarial mesh M_{adv} and projects its shadow onto the vehicle surface, since searching in the higher-dimensional 3D space yields better solutions. We extend MSA from shape-only optimization to shape-texture co-optimization: for the i-th sticker we optimize a mesh $M_{\mathrm{adv}}^{(i)}$ together with a surrounding texture patch $\tau^{(i)}$, and project the mesh via a shadowing operation Ω_{shape} to obtain the

contour. The patterns in the two modalities are

$$S_{\text{thm}}^{(i)} = \Omega_{\text{shape}}(M_{\text{adv}}^{(i)}, \varphi^{(i)}), \qquad S_{\text{vis}}^{(i)} = \Omega_{\text{shape}}(M_{\text{adv}}^{(i)}, \varphi^{(i)}) \oplus_{\text{c}} \tau^{(i)}, \tag{2}$$

where $\varphi^{(i)}$ is the shadowing angle and $\oplus_{\text{c}}$ composes the aluminum contour region with the surrounding texture. In the thermal channel the contour interior is a uniform dark region consistent with aluminum's low emissivity; in the RGB channel it is silver-gray, while $\tau^{(i)}$ supplies additional adversarial information. Optimization variables are the mesh vertices $V^{(i)}$, shadowing angle $\varphi^{(i)}$, texture $\tau^{(i)}$, and placement $P^{(i)}$.

4.4 Smoothness Regularization

To keep the optimized patterns physically realizable, we regularize the adversarial meshes with normal consistency, edge length, chamfer distance (Fan et al., 2017) and Laplacian smoothing (Nealen et al., 2006) losses ($L_{\text{norm}}, L_{\text{edge}}, L_{\text{chamfer}}, L_{\text{Laplace}}$), and each RGB texture patch with a total-variation loss L_{TV} that suppresses high-frequency noise and ensures printability. Full definitions are given in Appendix D.

4.5 Applying Adversarial Patterns to the 3D RGB-T Vehicle Model

We apply the patterns by modifying the model's visible and thermal texture maps, simulating the pasting of stickers in the physical world. For ease of physical deployment, stickers are placed on wide flat regions—doors, roof, front hood, and rear—with one or two patterns per region (Figure 2). Let Γ denote pasting, with $P^{(i)}$ the position of the i-th pattern in a Cartesian coordinate system defined on the UV-unfolded texture map with the map center as origin. We then render with the differentiable renderer PyTorch3D (Ravi et al., 2020) under camera pose $\mathbf{c} = (\theta, \phi, d)$, denoting rendering by Ψ. For each modality $\text{mod} \in \{\text{vis}, \text{thm}\}$, pasting N patterns and rendering gives

$$T_{\text{mod}}^{\text{adv}} = \Gamma\Big(S_{\text{mod}}^{(i)}, P^{(i)}, T_{\text{mod}}\Big)_{i=1}^{N}, \qquad I_{\text{mod}}^{\text{adv}}(\mathbf{c}) = \Psi\big(M_{\text{car}},\, T_{\text{mod}}^{\text{adv}},\, \mathbf{c}\big), \tag{3}$$

where T_{mod} is the original texture map of that modality.

4.6 Cross-Modal Foreground and Background EOT

To enhance physical robustness, we employ a cross-modal Expectation Over Transformation (EOT) strategy spanning foreground and background augmentations. For *foreground EOT* we apply modality-specific perturbations that simulate manufacturing and deployment imperfections: in the thermal modality, random perturbations to the mesh vertices V of M_{adv} (cutting errors), random noise and grayscale shifts on S_{thm} (surface temperature variation), and random offsets to P (placement inaccuracy); in the RGB modality, photometric augmentations on S_{vis} including noise injection, brightness jittering and contrast variation (printing and illumination changes). For *background EOT* we sample a background from a diverse scene pool $\mathcal{B}$ at each iteration and paste the rendered vehicle at a random position p. This jointly improves generalization across foreground conditions and background environments in both modalities.

4.7 Angle-Adaptive Attack Strategy and Joint Objective

Under uniform sampling, θ is sampled every $\Delta\theta$ over $[0°, 360°)$, ϕ every $\Delta\phi$ over $[0°, 90°]$, and the distance d every Δd over $[d_{\min}, d_{\max}]$, giving a distance set $\mathcal{D}$ with $D = |\mathcal{D}|$. Because some viewpoints are far harder to attack than others, we instead sample adaptively. For each viewpoint (θ_i, ϕ_j) we measure $\text{ASR}(\theta_i, \phi_j)$, the fraction of the D sampled distances at which the detector confidence falls below q, and assign the viewpoint a sampling multiplicity

$$r(\theta_i, \phi_j) = \min\left(\frac{1}{\text{ASR}(\theta_i, \phi_j) + \epsilon},\ \alpha\right), \tag{4}$$

where ϵ prevents division by zero and α caps replication. Multiplicities are re-evaluated every K iterations, so optimization capacity is continually redirected toward viewpoints where the attack is weakest.

The rendered adversarial images $\hat{I}^{\text{adv}}_{\text{vis}}$ and $\hat{I}^{\text{adv}}_{\text{thm}}$ (with EOT applied) are fed to the target detector f. Since the goal is evasion, the detection loss minimizes the object confidence over both modalities, $L_{\text{det}} = f_{\text{obj}}(\hat{I}^{\text{adv}}_{\text{vis}}, \hat{I}^{\text{adv}}_{\text{thm}})$. Combined with the regularizers of Sec. 4.4, the overall objective is

$$L = L_{\text{det}} + w_1 L_{\text{norm}} + w_2 L_{\text{edge}} + w_3 L_{\text{chamfer}} + w_4 L_{\text{Laplace}} + w_5 L_{\text{TV}}, \tag{5}$$

with empirically chosen weights $w_1, \ldots, w_5$. We jointly update $V^{(i)}, \varphi^{(i)}, P^{(i)}$ and $\tau^{(i)}$ for all N stickers by backpropagation through the differentiable rendering pipeline.

4.8 Angle-Adaptive Adversarial Training

As a defense, we reverse the attack-oriented viewpoint weighting of Sec. 4.7. Standard adversarial training draws adversarial examples uniformly, which underrepresents hard-to-attack angles and yields an imbalanced defense. We instead assign each viewpoint a weight inversely proportional to $r(\theta_i, \phi_j)$, prioritizing easy-to-attack viewpoints with high ASRs, so that vulnerable viewpoints receive more supervision:

$$\min_{\mathbf{w}} \sum_{i,j} \frac{\beta}{r(\theta_i, \phi_j)} \mathcal{L}^{\text{std}}_{\text{det}} \Big(f_{\mathbf{w}}(\hat{I}^{\text{adv}}_{\text{vis}}(\theta_i, \phi_j), \hat{I}^{\text{adv}}_{\text{thm}}(\theta_i, \phi_j)), y \Big), \tag{6}$$

where $\mathbf{w}$ are the detector parameters, $\beta > 0$ scales the adversarial training loss, $\mathcal{L}^{\text{std}}_{\text{det}}$ the standard detection loss and y the ground-truth annotation. Multiplicities are recomputed every K iterations from the updated detector, so the training focus adapts as previously vulnerable viewpoints improve, giving a more uniform defense over the hemisphere.

4.9 Physical Implementation

Fabrication has two stages. We first print the RGB texture $\tau^{(i)}$ on poster paper, leaving the regions corresponding to the contours $S^{(i)}_{\text{thm}}$ blank. We then cut 0.08 mm aluminum films to the contour shapes and attach them to those reserved regions; aluminum's emissivity ($\sim$0.1) against the vehicle's steel surface ($\sim$0.8) creates the infrared contrast that forms the thermal pattern. Each sticker takes about 15 minutes and 0.5 USD to produce.

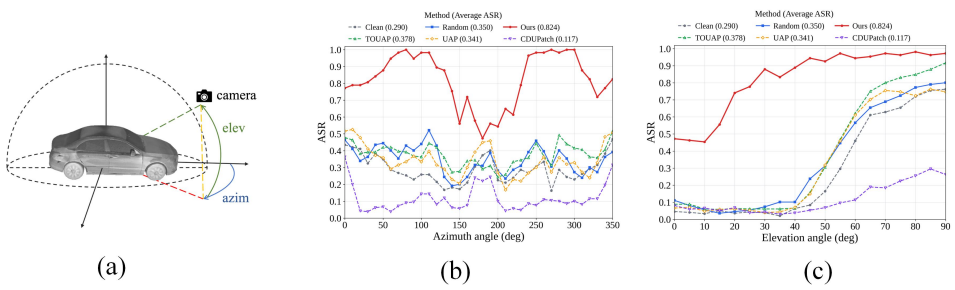
5 Experiments

5.1 Setup

Dataset. To provide diverse and realistic background scenes for our 3D RGB-T simulation pipeline, we adopt the FLIR-aligned dataset (Zhang et al., 2020), which consists of well-registered visible and thermal infrared image pairs captured in real-world driving scenarios. The dataset comprises 4,129 paired images for training and 1,013 paired images for testing. During the optimization and evaluation of adversarial patterns, these image pairs serve as the background onto which the rendered adversarial vehicles are composited via the background EOT strategy described in Sec. 4.6.

Detectors. We adopt the state-of-the-art RGB-T detector DAMSDet (Guo et al., 2024) as our primary target detector, which is a DETR-based multispectral detection transformer with competitive query selection and adaptive feature fusion. After attacking DAMSDet in a white-box setting, we transfer the adversarial patterns to three unseen RGB-T detectors in a black-box setting, including ProbEn (Chen et al., 2022), ICAFusion (Shen et al., 2024), and CMAFF (Qingyun & Zhaokui, 2022), which represent diverse fusion paradigms spanning decision-level, cross-attention-based, and attentive feature-level fusion strategies. All these RGB-T detectors achieve high detection confidence scores (above 0.85) on clean RGB-T vehicle targets, making them strong baselines for evaluating adversarial robustness.

Evaluation Metrics We report attack success rate (ASR), the fraction of vehicles not detected, with confidence threshold 0.6 and IoU threshold 0.5, following (Zhu et al., 2021; 2022; Wei et al., 2023b;a). ASR is averaged over sampled distances, azimuths and elevations. See Appendix E for more details.



(a) (b) (c)

Figure 3: Per-variable ASR in the digital world against DAMSDet (Guo et al., 2024) under the full-view attack setting.

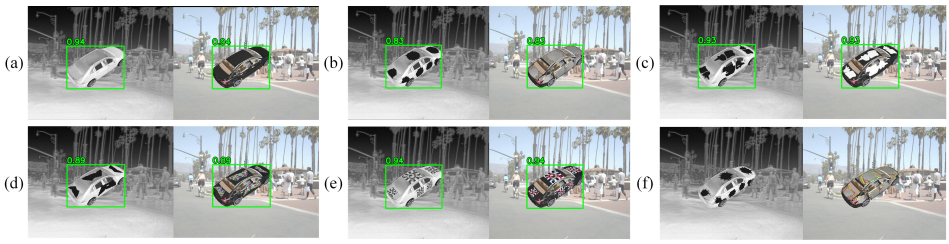


Figure 4: Typical adversarial examples in the digital world. (a) Clean. (b) Random. (c) UAP. (d) TOUAP. (e) CDUPatch. (f) Ours.

5.2 Vehicle Attacks in the Digital World

We optimized $N = 7$ adversarial stickers on the vehicle surface, covering the five standard placements (two doors, roof, front hood, and rear) and two additional placements on the oblique front and rear, which were identified as vulnerable views in Sec. 3.1. The patterns were generated using shape–texture co-design (Sec. 4.3), angle-adaptive viewpoint sampling (Sec. 4.7), and cross-modal foreground and background EOT (Sec. 4.6). We evaluated the attack against DAMSDet (Guo et al., 2024) and compared it with the clean vehicle, random patterns, UAP (Wei et al., 2023a), TOUAP (Hu et al., 2025), and CDUPatch (Long et al., 2025). All methods were evaluated using the same 3D RGB-T vehicle model and full-view protocol. Specifically, the azimuth θ was sampled every 10° over $[0^\circ, 360^\circ)$, and the elevation ϕ was sampled every 5° over $[0^\circ, 90^\circ]$. The distance d was uniformly sampled over $[d_{\min}, d_{\max}]$ as an additional scene variation rather than an independent analysis variable. Further implementation details are provided in the Appendix F and G.

As shown in Fig. 3, our method achieves an average ASR of 82.4%, substantially outperforming the clean vehicle (29.0%), random patterns (35.0%), UAP (34.1%), TOUAP (37.8%), and CDUPatch (11.7%). In particular, it improves upon TOUAP, the strongest prior attack, by 44.6%. Representative adversarial examples in the visible and thermal modalities are shown in Fig. 4.

We further analyze the ASR with respect to azimuth and elevation by averaging over the other angle, sampled distances, and test backgrounds. Figure 3(b) shows that our method consistently outperforms all baselines across the complete 360° azimuth range. It achieves particularly high ASR over 50°–130° and 240°–300°. Although performance decreases at several difficult views around 150°–230° and 310°–340°, it remains clearly higher than the corresponding baselines. As shown in Fig. 3(c), our method also provides a substantial advantage at low and intermediate elevations, where the clean detector and prior attacks have low ASRs, and maintains an ASR above approximately 0.9 over most elevations beyond 45°. The clean and baseline ASRs also increase at steep elevations, indicating an intrinsic weakness of the detector under top-down views; nevertheless, our method retains a consistent margin across the entire elevation range. These results demonstrate that the proposed shape–texture co-design and angle-adaptive optimization substantially improve angular attack coverage, enabling a more balanced full-view RGB-T attack.

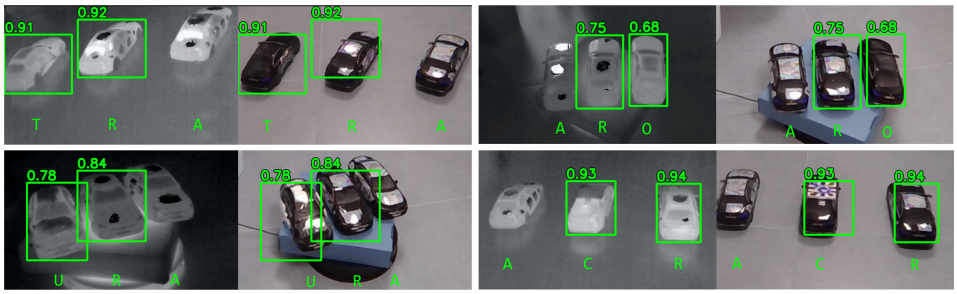


Figure 5: Adversarial examples in the physical world, shown as aligned visible–thermal pairs across representative viewpoints. O: Ordinary. R: Random. U: UAP. T: TOUAP. C: CDUPatch. A: Adversarial (Ours). The numbers above the bounding boxes are confidence scores.

Table 1: Per-viewpoint ASR on the physical model vehicle against DAMSDet (Guo et al., 2024). Best result in each column in bold.

Method	Front	Front-Side	Side	Rear-Side	Rear	Overhead	Overall
Clean	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Random	0.000	0.000	0.000	0.000	0.000	0.130	0.022
UAP	0.193	0.097	0.020	0.000	0.335	0.150	0.133
TOUAP	0.757	0.191	0.000	0.122	0.569	0.210	0.308
CDUPatch	0.074	0.191	0.000	0.000	0.000	0.360	0.104
Ours	**0.927**	**0.883**	**0.510**	**0.732**	**0.892**	**0.370**	**0.719**

5.3 Ablation Study.

We ablate three core modules against DAMSDet Guo et al. (2024) under identical full-angle settings, starting from a baseline that uses shape-only optimization with uniform viewpoint sampling and no EOT. The baseline reaches an ASR of 61.8%; adding STCO, AAAS, or CFBE alone raises it to 73.8%, 68.9%, and 65.7% respectively, and combining all three yields 20.6% above the baseline and 8.6% above the strongest single-module variant. The three modules are therefore complementary rather than redundant, each contributing along a different axis: cross-modal effectiveness, angular coverage, and transformation robustness. Full results are given in Appendix H.

5.4 Vehicle Attacks in the Physical World

We conducted physical experiments on 1:24 scale model vehicles to validate the effectiveness of our RGB-T adversarial stickers (Figure 5). We fabricated the $N = 7$ adversarial stickers following the two-stage process described in Sec. 4.9: the RGB adversarial textures were printed on poster paper with reserved blank regions, onto which aluminum film contours were attached. All stickers were scaled to match the model vehicle dimensions. See Appendix I for more experimental details.

To capture aligned RGB-T image pairs, we employed a FLIR T560 thermal camera and a Lenovo built-in RGB camera simultaneously. We developed a Python script to perform digital zoom on the images captured by both cameras, producing spatially aligned visible and thermal image pairs. A rotating turntable was used to assess the attack effectiveness over the full angular range. For each configuration, the aligned image pairs were fed into the DAMSDet (Guo et al., 2024) detector.

For a fair comparison, we evaluated the same set of baselines as in the digital experiments: clean vehicle, random patterns, UAP, TOUAP, and CDUPatch. Since the physical world does not permit precise continuous control over arbitrary observation angles, we collected images from multiple representative viewpoints commonly encountered in autonomous driving scenarios and report angle-specific ASRs. Specifically, we define six typical viewpoint groups: front ($\theta \approx 0^\circ \pm 30^\circ$), front-side ($\theta \approx 45^\circ \pm 30^\circ$), side ($\theta \approx 90^\circ \pm 30^\circ$), rear-side ($\theta \approx 135^\circ \pm 30^\circ$), rear ($\theta \approx 180^\circ \pm 30^\circ$), and overhead ($\phi \approx 45^\circ \pm 30^\circ$, simulating road surveillance camera perspectives).

Table 2: Black-box transferability our adversarial pattern, reported as ASR against three unseen RGB-T detectors in the digital and physical settings. Best result in each column in bold.

	Digital			Physical		
Method	ProbEn	ICAFusion	CMAFF	ProbEn	ICAFusion	CMAFF
UAP	0.365	0.661	0.740	0.328	0.285	0.185
TOUAP	0.354	0.662	0.694	0.243	0.349	0.149
CDUPatch	0.300	0.652	0.812	0.172	0.214	0.137
Ours	**0.661**	**0.790**	**0.936**	**0.518**	**0.544**	**0.966**

As shown in Table 1, our method achieved an overall ASR of 71.9% on the physical vehicle, significantly outperforming the clean vehicle (0.0%), random patterns (2.2%), UAP (13.3%), TOUAP (30.8%), and CDUPatch (10.4%). The per-viewpoint ASR analysis (Table 1) demonstrates that our method maintains consistently higher attack performance across all six viewpoint groups than the baseline methods, which exhibit significant degradation at certain angles. Figure 5 provides representative examples from the physical experiments.

5.5 Attack Transferability

We evaluated the transferability of our adversarial patterns optimized for DAMSDet (Guo et al., 2024) in a white-box setting against three unseen RGB-T detectors in a black-box setting: ProbEn (Chen et al., 2022), ICAFusion (Shen et al., 2024), and CMAFF (Qingyun & Zhaokui, 2022). The same evaluation was conducted in both digital and physical experimental settings, following the identical protocols described in their respective sections. For comparison, we also report the black-box transferability of UAP (Wei et al., 2023a), TOUAP (Hu et al., 2025), and CDUPatch (Long et al., 2025). As shown in Table 2, our method achieves ASRs of 66.1%–93.6% against the unseen detectors in the digital setting and 51.8%–96.6% in the physical setting, consistently outperforming all baseline attack methods across the three detectors. These results indicate that our STCO framework yields adversarial patterns with strong cross-architecture transferability, generalizing effectively to detectors with diverse fusion paradigms, including decision-level fusion (ProbEn), cross-attention-based fusion (ICAFusion), and attentive feature-level fusion (CMAFF).

5.6 Defense Against Adversarial Attacks

We evaluate 11 defense methods, including recent adversarial training (Zhu et al., 2025), Jedi (Tarchoun et al., 2023), UDF (Yu et al., 2022), FNC (Yu et al., 2021), EPGF (Zhou et al., 2020), SAC (Liu et al., 2022), classical PixelMask (Agarwal et al., 2021), bit squeezing (Xu et al., 2018), Spatial smoothing (Xu et al., 2018), JPEG compression (Guo et al., 2018), and our angle-adaptive adversarial training (Sec. 4.8). Configuration details for all defenses are given in Appendix F, with other settings as in Sec. 5.2. The preprocessing defenses recover at most 5.2% of ASR and standard adversarial training 3.5% points, while our variant recovers 18.6%—3.6 times the best alternative—confirming the benefit of directing defensive supervision to the most vulnerable viewpoints. Even under the strongest defense, ASR remains at 63.8%, indicating that our attack still poses a significant threat to current RGB-T detectors. Full results are provided in Appendix J

6 Conclusion

We presented a comprehensive study on the full-view adversarial robustness of RGB-T vehicle detectors, revealing that existing attacks degrade under the full-angle setting and identifying three challenges: modality imbalance, angular imbalance, and the digital-to-physical transfer gap. We addressed these challenges through shape-texture co-optimization, angle-adaptive viewpoint sampling, and cross-modal foreground-background EOT. Digital and physical experiments demonstrated the effectiveness of our attack against DAMSDet and its strong transferability to unseen detectors. We also introduced angle-adaptive adversarial training as a defense baseline. Our findings highlight persistent vulnerabilities in RGB-T detectors and the need for stronger defenses in safety-critical autonomous driving systems.

AI USE STATEMENT

The conception, experimental design and execution, and writing of this paper were carried out entirely by the human authors. Generative AI tools were used only for grammar checking and language polishing of author-written text. We take full responsibility for the final content of this work.

ETHICS STATEMENT

Adversarial attack techniques must be used with care, as they can be misused to compromise the reliability of deployed AI systems. At the same time, such techniques are widely adopted as a means of vulnerability analysis for existing models, and the blind spots they expose are a precondition for building more secure systems—for example, through adversarial training.

REPRODUCIBILITY STATEMENT

This work is built based on publicly available resources: the background scenes come from the open FLIR-aligned dataset (Zhang et al., 2020), and all target detectors—DAMSDet, ProbEn, ICAFusion, and CMAFF—are open-source models used with their released implementations. The method is fully specified in Sec. 4, with complete loss definitions, hyper-parameters, and the sticker fabrication and RGB-T capture protocol given in the appendix. Our code and the RGB-T image pairs collected in the physical experiments will be released publicly upon acceptance.

REFERENCES

Akshay Agarwal, Mayank Vatsa, Richa Singh, and Nalini Ratha. Cognitive data augmentation for adversarial defense via pixel masking. *Pattern Recognition Letters*, 146:244–251, 2021.

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

Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *European Conference on Computer Vision*, pp. 139–158. Springer, 2022.

Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 605–613, 2017.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2015.

Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. 2018.

Junjie Guo, Chenqiang Gao, Fangcen Liu, Deyu Meng, and Xinbo Gao. Damsdet: Dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion. In *European Conference on Computer Vision*, pp. 464–481. Springer, 2024.

Chengyin Hu, Weiwen Shi, Wen Yao, Tingsong Jiang, Ling Tian, and Wen Li. Two-stage optimized unified adversarial patch for attacking visible-infrared cross-modal detectors in the physical world. *Applied Soft Computing*, 171:112818, 2025.

Lifeng Huang, Chengying Gao, Yuyin Zhou, Cihang Xie, Alan L Yuille, Changqing Zou, and Ning Liu. Universal physical camouflage attacks on object detectors. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 717–726. IEEE, 2020.

Jiang Liu, Alexander Levine, Chun Pong Lau, Rama Chellappa, and Soheil Feizi. Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962. IEEE, 2022.

Yuqiu Liu, Huanqian Yan, Xiaopei Zhu, Xiaolin Hu, Liang Tang, Hang Su, and Chen Lv. Crafting physical adversarial examples by combining differentiable and physically based renders. *IEEE/CAA Journal of Automatica Sinica*, 2026.

Jiahuan Long, Wen Yao, Tingsong Jiang, Jiacheng Hou, Shuai Jia, Junqi Wu, Xiaoya Zhang, Xiaohu Zheng, and Chao Ma. Cdupatch: Color-driven universal adversarial patch attack for dual-modal visible-infrared detectors. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1462–1470, 2025.

Andrew Nealen, Takeo Igarashi, Olga Sorkine, and Marc Alexa. Laplacian mesh optimization. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pp. 381–389, 2006.

Fang Qingyun and Wang Zhaokui. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition*, 130:108786, 2022.

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv preprint arXiv:2007.08501*, 2020.

Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145: 109913, 2024.

Naufal Suryanto, Yongsu Kim, Hyoeun Kang, Harashta Tatimma Larasati, Youngyeo Yun, Thi-Thu-Huong Le, Hunmin Yang, Se-Yoon Oh, and Howon Kim. Dta: Physical camouflage attacks using differentiable transformation network. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15284–15293. IEEE, 2022.

Bilel Tarchoun, Anouar Ben Khalifa, Mohamed Ali Mahjoub, Nael Abu-Ghazaleh, and Ihsen Alouani. Jedi: Entropy-based localization and removal of adversarial patches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4087–4095, 2023.

Donghua Wang, Tingsong Jiang, Jialiang Sun, Weien Zhou, Zhiqiang Gong, Xiaoya Zhang, Wen Yao, and Xiaoqian Chen. Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2414–2422, 2022.

Jiakai Wang, Aishan Liu, Zixin Yin, Shunchang Liu, Shiyu Tang, and Xianglong Liu. Dual attention suppression attack: Generate adversarial camouflage in physical world. *arXiv preprint arXiv:2103.01050*, 2021.

Xingxing Wei, Yao Huang, Yitong Sun, and Jie Yu. Unified adversarial patch for visible-infrared cross-modal attacks in the physical world. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2348–2363, 2023a.

Xingxing Wei, Jie Yu, and Yao Huang. Physically adversarial infrared patches with learnable shapes and locations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12334–12342. IEEE, 2023b.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS*, 2018.

Cheng Yu, Jiansheng Chen, Youze Xue, Yuyang Liu, Weitao Wan, Jiayu Bao, and Huimin Ma. Defending against universal adversarial patches by clipping feature norms. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16414–16422. IEEE, 2021.

Youngjoon Yu, Hong Joo Lee, Hakmin Lee, and Yong Man Ro. Defending person detection against adversarial patch attack by using universal defensive frame. *IEEE Transactions on Image Processing*, 31:6976–6990, 2022.

Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International conference on image processing (ICIP)*, pp. 276–280. IEEE, 2020.

Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. Camou: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2018.

Guangzhi Zhou, Hongchao Gao, Peng Chen, Jin Liu, Jiao Dai, Jizhong Han, and Ruixuan Li. Information distribution based defense against physical attacks on object detection. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6. IEEE, 2020.

Xiaopei Zhu, Xiao Li, Jianmin Li, Zheyao Wang, and Xiaolin Hu. Fooling thermal infrared pedestrian detectors in real world using small bulbs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 3616–3624, 2021.

Xiaopei Zhu, Zhanhao Hu, Siyuan Huang, Jianmin Li, and Xiaolin Hu. Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13307–13316. IEEE, 2022.

Xiaopei Zhu, Yuqiu Liu, Zhanhao Hu, Jianmin Li, and Xiaolin Hu. Infrared adversarial car stickers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24284–24293. IEEE, 2024.

Xiaopei Zhu, Siyuan Huang, Zhanhao Hu, Jianmin Li, Jun Zhu, and Xiaolin Hu. Physical adversarial examples for person detectors in thermal images based on 3d modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):8973–8989, 2025.

A Full-View Evaluation Protocol for Prior Methods

This appendix details the standardised full-view protocol used in Sec. 3.1 to re-evaluate existing RGB-T attacks outside their original operating range.

A.1 Rendering and Sampling Configuration

All methods are evaluated on the same co-registered 3D RGB-T vehicle model (Appendix C) and rendered with the same differentiable renderer (Ravi et al., 2020). A virtual camera traverses the viewpoint hemisphere with azimuth θ sampled every $\Delta\theta = 10^\circ$ over $[0^\circ, 360^\circ)$ and elevation ϕ sampled every $\Delta\phi = 5^\circ$ over $[0^\circ, 90^\circ]$, giving $36 \times 19 = 684$ distinct viewpoints. At each viewpoint the observation distance d is sampled every $\Delta d = 0.1$ m over $[d_{\min}, d_{\max}] = [1.5, 2.5]$ m, yielding $D = 11$ distance samples per viewpoint. Each rendered vehicle is composited onto a background image pair drawn from the FLIR-aligned test split (Zhang et al., 2020) following the background EOT procedure of Sec. 4.6, so that every method is evaluated over the same $684 \times D$ rendered RGB-T pairs.

Table A1 lists the complete protocol configuration.

Table A1: Full-view evaluation protocol. Identical settings are applied to every method compared in Sec. 3.1 and Sec. 5.2.

Item	Setting
Azimuth range / step	$[0^\circ, 360^\circ)$ / 10° (36 samples)
Elevation range / step	$[0^\circ, 90^\circ]$ / 5° (19 samples)
Distance range / step	$[1.5, 2.5]$ m / 0.1 m ($D = 11$)
Renderer	PyTorch3D (Ravi et al., 2020)
Rendered resolution	1024×1024
Confidence threshold q	0.6
IoU threshold	0.5
Vehicle model	Co-registered 3D RGB-T model (Appendix C)

A.2 Detector and Threshold Settings

Rendered pairs are fed to the RGB-T detector used in each method's original paper, so that no method is penalised by being transplanted onto an unfamiliar target. For fairness, the confidence threshold and the IoU threshold are held fixed at 0.6 and 0.5 respectively for all methods, matching the values used in the main experiments (Sec. 5.1) and in prior work (Zhu et al., 2021; 2022; Wei et al., 2023b;a). A vehicle counts as successfully attacked when no predicted box of the `car` class both exceeds the confidence threshold and attains IoU ≥ 0.5 with the ground-truth box.

A.3 Transplanting Prior Adversarial Patterns

Each prior method prescribes a pattern together with a placement rule. We keep both. Concretely:

- **UAP** (Wei et al., 2023a) contributes a single unified patch, placed on the vehicle side panel as in the original work.
- **TOUAP** (Hu et al., 2025) contributes a two-stage optimised patch, placed on the vehicle side panel as in the original work.
- **CDUPatch** (Long et al., 2025) contributes a colour-driven patch, placed on the vehicle roof to match its top-down threat model.

Patterns are resampled to the UV resolution of the shared vehicle model with the physical patch size preserved, so that the fraction of vehicle surface covered by each method matches its original specification. The *clean* configuration renders the same vehicle with no sticker, and the *random* configuration replaces each adversarial pattern with a randomly generated contour of identical area, providing a control for the mere presence of surface clutter.

A.4 Multi-Patch Upper-Bound Probe

Prior methods were designed around a single patch, which places them at an obvious disadvantage under full-view evaluation. To probe their upper bound rather than their nominal configuration, we additionally replicate each method's pattern across several wide flat regions of the vehicle — the two doors, the roof and the front hood — keeping the pattern itself, the optimisation procedure and all evaluation settings unchanged. This yields a multi-patch variant of every baseline that is comparable in surface coverage to the N-sticker configuration used by our method.

B Shape-Only RGB-T Baseline

This appendix details the diagnostic baseline of Sec. 3.2, which asks whether the shape optimisation framework of the thermal-only full-view attack (Zhu et al., 2024) suffices once the detector consumes both modalities.

B.1 Construction

The baseline retains the 3D adversarial mesh shadow attack of Zhu et al. (2024) unchanged: for each sticker a mesh $M_{\text{adv}}^{(i)}$ and a shadowing angle $\varphi^{(i)}$ are optimised, and the projected shadow $\Omega_{\text{shape}}(M_{\text{adv}}^{(i)}, \varphi^{(i)})$ defines a single contour that is shared by both modalities. No printable texture $\tau^{(i)}$ is introduced; the interior of the contour is realised with the same aluminium film used by our method, so that it renders as a uniformly dark region in the thermal channel (emissivity $\sim$0.1) and as a silver-grey region in the RGB channel. Every other component — sticker count, placements, smoothness regularisation, viewpoint sampling, detection loss and optimiser — follows Appendix F, and evaluation follows the protocol of Appendix A. The baseline therefore differs from the full method in exactly one respect: the RGB channel carries no optimisable texture.

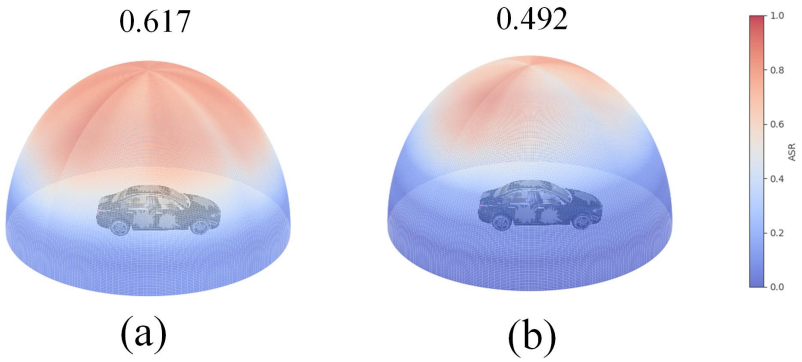


Figure A1: Modality and angular imbalance of the shape-only baseline under full-view evaluation. (a) Thermal. (b) RGB. Colors indicate per-viewpoint ASR, and the numbers above the hemispheres denote average ASR.

B.2 Digital Evaluation

Under the full-view protocol, the shape-only baseline achieves an ASR of 0.482, exceeding the prior baselines in Fig. 1 but remaining below the 0.824 achieved by our full method. Evaluations across modalities and viewpoints reveal two imbalances:

- **Modality imbalance.** As shown in Fig. A1, the modality-wise evaluation yields average ASRs of 0.617 in thermal and 0.492 in RGB, revealing a 12.5-percentage-point gap. This suggests that the shared contour is more effective in the thermal modality, while its monochromatic appearance provides limited adversarial expressiveness in RGB. This imbalance motivates jointly optimizing the shared contour and the surrounding RGB texture.

- **Angular imbalance.** Both hemisphere maps exhibit substantial viewpoint-dependent variation: ASR is generally higher at elevated viewpoints, whereas broad low-ASR regions persist at low elevations, particularly in RGB. These results indicate that shape-only optimization provides uneven angular attack coverage. This motivates an angle-adaptive sampling strategy that allocates more optimization effort to viewpoints with low ASR.

B.3 Physical Validation

The baseline was additionally realised physically on 1:24 scale model vehicles using the fabrication and acquisition procedure of Appendix I, with the aluminium contours cut to the optimised shapes and scaled to the model vehicle. The measured physical ASR of 0.254, against 0.482 digitally, quantifies the third challenge — the digital-to-physical transfer gap — and motivates the cross-modal foreground and background EOT of Sec. 4.6.

C Building a Co-Registered 3D RGB-T Vehicle Model

The full-view attack requires a 3D vehicle model whose *visible* and *thermal* textures are both registered to the same mesh and the same UV layout, so that a sticker placed at UV position $P^{(i)}$ occupies physically corresponding regions in the two modalities. This appendix describes how the 3D thermal vehicle model of Zhu et al. (2024) is extended into such a co-registered RGB-T model.

C.1 Mesh and UV Layout

We start from a mesh M_{car} of a sedan with 116957 vertices and 138887 faces. The mesh faces are unfolded into a 2D UV layout in Maya and grouped into semantically meaningful regions — roof, hood, left and right doors, trunk lid, and bumpers — so that sticker placement can be restricted to the wide flat regions that are practical to paste on in the physical world. The resulting faces map has resolution 1024×1024 and is shared by both modalities; the placement coordinate $P^{(i)}$ used in Eq. (3) is expressed in a Cartesian frame on this map with the map centre as origin.

C.2 Thermal Texture

The thermal texture is obtained following Zhu et al. (2024). A FLIR T560 captures thermal images of the target vehicle from multiple azimuth and elevation angles; the vehicle region of each frame is extracted, radiometrically normalised, and warped onto the corresponding faces of the UV layout. Overlapping observations are blended so that the resulting thermal texture covers the entire vehicle surface without gaps.

C.3 Visible Texture Acquisition and Registration

The difficulty in the RGB-T extension is producing a visible "skin" that is registered to the *same* faces map, rather than merely a plausible-looking car texture. We proceed as follows.

1. **Paired capture.** A single FLIR T560 captures paired visible and thermal frames of the target vehicle spanning azimuth $[0°, 360°)$ and elevation $[0°, 90°]$. Both channels of a pair come from the same instrument at the same instant.
2. **Region cropping.** From each visible frame we crop the surface regions that are unoccluded and close to fronto-parallel, and match them to the corresponding regions of the UV layout.
3. **Projective correction.** A cropped photograph does not conform to the UV geometry, because of perspective foreshortening and lens distortion. Each patch is therefore warped by a projective transform, estimated from 36 manually placed correspondences between patch corners and UV region boundaries, until it registers with the faces map.
4. **Stitching.** A single viewpoint covers only part of the vehicle surface, so warped patches from multiple viewpoints are stitched together, with seams placed along UV region boundaries and exposure differences compensated, until the visible texture map covers the entire mesh surface without gaps.

C.4 Cross-Modal Consistency

Because the visible patches are registered to the same faces map that the thermal texture already occupies, a UV coordinate indexes the same physical point on the vehicle in both modalities. This is what makes the shape-texture co-optimisation of Sec. 4.3 well posed: the shared contour $\Omega_{\text{shape}}(M_{\text{adv}}^{(i)}, \varphi^{(i)})$ and the RGB-only texture $\tau^{(i)}$ are composed at a single UV location and therefore describe one physical sticker rather than two independent perturbations.

D Loss Functions and Smoothness Regularisation

This appendix gives the complete definitions of the terms summarised in Sec. 4.4 and Sec. 4.7.

D.1 Mesh Smoothness

An unconstrained adversarial mesh degenerates into a spiky surface whose projected shadow cannot be cut out of aluminium film. Four regularisers keep each $M_{\text{adv}}^{(i)}$ smooth enough to be physically realisable. For brevity the superscript (i) is dropped below; every term is averaged over the N stickers.

The *normal consistency loss* penalises angular deviation between the normals of adjacent faces, encouraging locally smooth surfaces:

$$L_{\text{norm}} = \text{Average}\left(1 - \cos(n_i, n_j)\right), \tag{A1}$$

where n_i and n_j are the normal vectors of an adjacent face pair and the average runs over all such pairs.

The *edge length loss* discourages excessively long edges, which otherwise produce sharp protrusions in the projected contour:

$$L_{\text{edge}} = \left(\sum\nolimits_{k=1}^{E} l_k\right) / E, \tag{A2}$$

where E is the number of edges and l_k the length of the k-th edge.

The *chamfer distance loss* (Fan et al., 2017) keeps the mesh close to a reference sphere M_{sphere}, bounding how far the optimisation can drift from a compact, cuttable shape:

$$L_{\text{chamfer}} = \sum_{p\in\mathcal{P}_1} \min_{q\in\mathcal{P}_2} \|p-q\|_2^2 + \sum_{q\in\mathcal{P}_2} \min_{p\in\mathcal{P}_1} \|q-p\|_2^2, \tag{A3}$$

where $\mathcal{P}_1$ and $\mathcal{P}_2$ are point clouds sampled from M_{adv} and M_{sphere} respectively.

The *Laplacian smoothing loss* (Nealen et al., 2006) regularises vertex positions towards the centroid of their neighbours:

$$L_{\text{Laplace}} = \text{Laplace}\left(M_{\text{adv}}\right). \tag{A4}$$

D.2 Texture Smoothness

The printable texture patch $\tau^{(i)}$ is regularised with a total-variation loss, which suppresses the high-frequency noise that a colour printer cannot reproduce and that would in any case be destroyed by camera resampling at distance:

$$L_{\text{TV}} = \sum_{i=1}^{N} \sum_{u,v} \left\| \tau_{u+1,v}^{(i)} - \tau_{u,v}^{(i)} \right\|_2^2 + \left\| \tau_{u,v+1}^{(i)} - \tau_{u,v}^{(i)} \right\|_2^2, \tag{A5}$$

where $\tau_{u,v}^{(i)}$ is the RGB value at position (u, v) of the i-th patch and (u, v) ranges over all valid pixel positions.

D.3 Detection Loss and Overall Objective

The detection loss minimises the object confidence that the target detector assigns to the vehicle given the EOT-augmented adversarial pair:

$$L_{\text{det}} = f_{\text{obj}}\Big(\hat{I}^{\text{adv}}_{\text{vis}},\ \hat{I}^{\text{adv}}_{\text{thm}}\Big). \tag{A6}$$

Both modalities enter through the detector's own fusion mechanism, so no explicit per-modality weighting is imposed; balance across modalities is instead induced structurally, by the shape/texture decomposition of Sec. 4.3. Combining Eqs. (A1)–(A6) gives the overall objective of Eq. (5), repeated here for completeness:

$$L = L_{\text{det}} + w_1 L_{\text{norm}} + w_2 L_{\text{edge}} + w_3 L_{\text{chamfer}} + w_4 L_{\text{Laplace}} + w_5 L_{\text{TV}}. \tag{A7}$$

The optimisation variables are the mesh vertices $V^{(i)}$, the shadowing angle $\varphi^{(i)}$, the placement $P^{(i)}$ and the texture $\tau^{(i)}$ of every sticker $i = 1, \ldots, N$; all are updated jointly by backpropagation through the differentiable rendering pipeline. Weight values are listed in Table A2.

D.4 Per-Viewpoint ASR and Sampling Multiplicity

Sec. 4.7 defines the multiplicity $r(\theta_i, \phi_j)$ in terms of the per-viewpoint attack success rate. Formally, for viewpoint (θ_i, ϕ_j),

$$\text{ASR}(\theta_i, \phi_j) = \frac{1}{D}\sum_{k=1}^{D} \mathbb{1}\Big[f\Big(\hat{I}^{\text{adv}}_{\text{vis}}(\theta_i, \phi_j, d_k),\ \hat{I}^{\text{adv}}_{\text{thm}}(\theta_i, \phi_j, d_k)\Big) < q\Big], \tag{A8}$$

where $\mathbb{1}[\cdot]$ is the indicator function, $d_k \in \mathcal{D}$ and $D = |\mathcal{D}|$. Substituting Eq. (A8) into Eq. (4) gives the continuous sampling weight $r(\theta_i, \phi_j)$ for that viewpoint. No rounding is applied, and the weight is positive and capped at α. Lower-ASR viewpoints receive higher sampling weights, up to this cap, so optimisation focuses on hard-to-attack viewpoints. These weights are recomputed every K iterations.

D.5 Angle-Adaptive Adversarial Training

The defense of Sec. 4.8 uses inverse viewpoint weights to prioritize easy-to-attack viewpoints. Every 10 iterations, the current detector $f_{\mathbf{w}}$ is evaluated over the viewpoint grid using Eq. (A8), and $r(\theta_i, \phi_j)$ is recomputed with Eq. (4). Each viewpoint's adversarial detection loss is then weighted by $\beta / r(\theta_i, \phi_j)$ in Eq. (6), giving greater emphasis to viewpoints with high ASRs. Ground-truth boxes y are those of the rendered vehicle and are unaffected by the perturbation, since the stickers lie on the vehicle surface and do not alter its silhouette.

E Evaluation Metrics

E.1 Attack Success Rate

We report attack success rate (ASR), defined as the fraction of target vehicles that the detector fails to detect. A vehicle counts as detected when the detector emits a box of the `car` class whose confidence exceeds $q = 0.6$ and whose IoU with the ground-truth box is at least 0.5; otherwise the attack on that sample is deemed successful. These thresholds follow prior physical-attack work (Zhu et al., 2021; 2022; Wei et al., 2023b;a) and are held fixed across all methods, all detectors and both the digital and physical settings.

E.2 Aggregation over the Viewpoint Hemisphere

In the digital setting, a sample is one (θ, ϕ, d, B) configuration, where B is the sampled background pair. The aggregate ASR reported in Sec. 5.2 is the unweighted mean of the indicator in Eq. (A8) over all sampled configurations,

$$\overline{\text{ASR}} = \frac{1}{|\Theta||\Phi|}\sum_{\theta_i \in \Theta}\sum_{\phi_j \in \Phi} \text{ASR}(\theta_i, \phi_j). \tag{A9}$$

The per-azimuth curves of Fig. 3(b) marginalise over ϕ, d and B; the per-elevation curves of Fig. 3(c) marginalise over θ, d and B. Distance is treated as a scene nuisance variable rather than an analysis axis, and is always marginalised out.

Because θ and ϕ are sampled on a uniform angular grid, the sample density per unit solid angle is not constant over the hemisphere: the grid is denser near the zenith, where the solid-angle element $\cos\phi\, d\theta\, d\phi$ shrinks. Eq. (A9) therefore weights steep elevations more heavily than a solid-angle-uniform observer would. We report the unweighted mean for comparability with prior full-view work (Zhu et al., 2024). All methods use the same angular sampling protocol, ensuring a consistent comparison under this evaluation scheme.

E.3 Physical-World Grouping

The physical world does not permit continuous control over the observation angle, so physical ASR is reported per viewpoint group rather than per viewpoint. The six groups are front ($\theta \approx 0^\circ \pm 30^\circ$), front-side ($45^\circ \pm 30^\circ$), side ($90^\circ \pm 30^\circ$), rear-side ($135^\circ \pm 30^\circ$), rear ($180^\circ \pm 30^\circ$) and overhead ($\phi \approx 45^\circ \pm 30^\circ$). The five azimuth groups are defined on $[0^\circ, 180^\circ]$ and, by the approximate bilateral symmetry of both the vehicle and the sticker placement, are taken to represent their mirror images on $(180^\circ, 360^\circ)$; the overhead group varies elevation instead of azimuth and is collected separately.

F Implementation Details

F.1 Attack Optimisation

Table A2 lists all hyper-parameters used to generate the adversarial patterns reported in Sec. 5.2. All variables — mesh vertices, shadowing angles, placements and textures — are optimised jointly with a single optimiser; no staged or alternating schedule is used.

Table A2: Hyper-parameters for adversarial pattern optimisation.

Parameter	Value
Stickers	
Number of stickers N	7
Placements	2 doors, roof, hood, rear, oblique front, oblique rear
Mesh vertices per sticker	2562
Texture patch resolution	1024×1024
Optimisation	
Optimizer	Adam
Learning rate (mesh / texture)	0.003 / 0.0005 / 0.0015 / 0.003
Iterations	41290
Batch size (viewpoints per step)	1
Loss weights	
w_1 (normal consistency)	0.01
w_2 (edge length)	1.0
w_3 (chamfer)	1.0
w_4 (Laplacian)	0.1
Compute	
Hardware	NVIDIAGeForceRTX3090
Wall-clock time per pattern	20 hours

F.2 EOT Parameter Ranges

Table A3 gives the sampling ranges of the cross-modal Expectation Over Transformation (Athalye et al., 2018) scheme of Sec. 4.6. Foreground transformations are modality specific, because the physical failure modes differ: the thermal appearance of a sticker is governed by how accurately its contour is cut and by the surface temperature of the panel it sits on, whereas its visible appearance is governed by printing fidelity and scene illumination. Background transformations are shared, since both channels of a FLIR-aligned pair come from the same scene.

Table A3: Cross-modal EOT sampling ranges. All parameters are drawn independently and uniformly at each iteration.

Transformation	Range
Foreground, thermal modality	
Mesh vertex jitter (cutting error)	$\mathcal{U}(0, 0.10)$
Grayscale shift (surface temperature)	± 0.1
Additive noise	$\sigma = 0.05$
Foreground, RGB modality	
Brightness jitter	± 0.05
Contrast variation	$\times\,[-0.05, 0.05]$
Additive noise	$\sigma = 0.05$
Foreground, both modalities	
Placement offset on $P^{(i)}$	± 20 UV pixels
Rendering distance d	$[1.5, 2.5]$ m
Background	
Scene pair $B \sim \mathcal{B}$	FLIR-aligned split (4,129 train / 1,013 test)
Paste position p	Uniform over valid image region

F.3 Detectors

DAMSDet (Guo et al., 2024), ProbEn (Chen et al., 2022), ICAFusion (Shen et al., 2024) and CMAFF (Qingyun & Zhaokui, 2022) are used with their publicly released implementations and their published FLIR-aligned checkpoints; no detector is retrained for the attack experiments. DAMSDet is the white-box target and provides gradients through f_{obj}; the other three are queried only at evaluation time for the black-box transfer results of Sec. 5.2. All four assign confidence above 0.85 to the clean rendered vehicle.

F.4 Defense Configurations

Table A4 lists the configuration of each defense evaluated in Sec. 5.6. Preprocessing defenses are applied identically to both modalities of every input pair at inference time. Training based defenses start from the released DAMSDet checkpoint and fine-tune it; all other settings follow Sec. 5.2. Bit squeezing is implemented with NumPy; JPEG compression and total-variance minimisation use the corresponding modules of the Adversarial Robustness Toolbox.

Table A4: Configuration of the evaluated defenses. Preprocessing defenses are applied to both modalities of every input pair at inference time; training defenses fine-tune the released DAMSDet checkpoint.

Defense	Configuration
Preprocessing	
PixelMask (Agarwal et al., 2021)	20×20 mask erased at random positions
Jedi (Tarchoun et al., 2023)	Entropy-based localisation + Telea inpainting ($\mathrm{r} = 3$)
Bit squeezing (Xu et al., 2018)	8-bit $\rightarrow$ 7-bit depth reduction
Spatial smoothing (Xu et al., 2018)	Median filter, window 3×3
JPEG compression (Guo et al., 2018)	ART `JpegCompression`, quality 90%
UDF (Yu et al., 2022)	Learned defensive frame 2-pixel border, 50 optimization steps
FNC (Yu et al., 2021)	Full-backbone feature-norm clipping ($\alpha = 2.0$)
EPGF (Zhou et al., 2020)	30×30 window, $\mathrm{stride} = 27$, $\tau_{\mathrm{e}} = 7.85$, $\tau_{\mathrm{g}} = 2.222$
SAC (Liu et al., 2022)	U-Net patch segmentation + shape completion ($\tau = 0.2$)
Training	
Adversarial training (Goodfellow et al., 2015)	Clean:adv. pairs $9:1$, uniform sampling
Angle-adaptive AT (ours)	Same budget, loss weighting $1/r(\theta_i, \phi_j)$

Both training-based defenses fine-tune for 100 epochs and are given the same total number of adversarial samples and the same number of fine-tuning steps, so that the comparison between them isolates the *allocation* of adversarial supervision across viewpoints rather than its quantity.

G Additional Details of the Digital-World Experiments

G.1 Sampling Grid and Scene Compositing

The digital results of Sec. 5.2 use the grid of Table A1: azimuth every 10° over $[0^\circ, 360^\circ)$ and elevation every 5° over $[0^\circ, 90^\circ]$, with distance marginalised out over $[d_{\min}, d_{\max}]$. Optimisation draws its backgrounds from the FLIR-aligned training split (4,129 pairs) and evaluation from the test split (1,013 pairs), so no background image is shared between the two. The rendered vehicle is composited at a random position within the background pair, at a scale consistent with the sampled distance.

G.2 Sticker Placement

The $N = 7$ stickers occupy the five standard flat regions — left door, right door, roof, front hood and rear — plus two additional placements on the oblique front and oblique rear quarters. The latter two are not arbitrary: they target the azimuth bands identified as hardest to attack by the shape-only diagnostic of Appendix B, where the projected area of a panel-mounted contour falls off fastest.

G.3 Baselines

The clean configuration renders the unmodified vehicle. The random configuration replaces each of the N adversarial contours with a randomly generated contour of equal area and a randomly generated texture patch, holding placement fixed; it isolates the contribution of *optimisation* from the contribution of merely covering the vehicle in high-contrast material. UAP (Wei et al., 2023a), TOUAP (Hu et al., 2025) and CDUPatch (Long et al., 2025) follow the transplantation procedure of Appendix A.

H Full Ablation Study

Table A5 reports the component ablation summarised in Sec. 5.2. The baseline is the shape-only configuration: a single contour shared by both modalities, uniform viewpoint sampling, and no EOT. Each module is then added individually, and finally all three together. Every variant is optimised with the same budget and evaluated against DAMSDet (Guo et al., 2024) under the protocol of Appendix A.

Table A5: Component ablation under the full-view setting. STCO: shape-texture co-optimization; AAAS: angle-adaptive attack strategy; CFBE: cross-modal foreground-background EOT. Δ is the change in ASR relative to the shape-only baseline.

Method	ASR	Δ
Base	0.618	—
Base + STCO	0.738	$+0.120$
Base + AAAS	0.689	$+0.071$
Base + CFBE	0.657	$+0.039$
Base + STCO + AAAS + CFBE (Ours)	**0.824**	$\mathbf{+0.206}$

Shape-texture co-optimization. STCO is the single largest contributor, raising ASR by 12.0 points on its own. This is consistent with the diagnosis of Appendix B: This is consistent with Appendix B, where the shape-only baseline is less effective in RGB than in thermal imagery. The printable texture $\tau^{(i)}$ adds RGB-specific adversarial information that complements the shared contour. Because the contour and the texture occupy the same physical sticker, this gain costs nothing in

fabrication complexity beyond printing the poster paper that the aluminium film is already mounted on.

Angle-adaptive attack strategy. AAAS contributes 7.1 points. It does not change what the adversarial pattern can express; it changes where optimisation capacity is spent. Under uniform sampling, viewpoints that are already attacked reliably continue to consume the same share of gradient updates as those that never succeed, so the achievable full-view ASR is capped by the tail of the angular distribution. Reweighting by $r(\theta_i, \phi_j)$ redirects that capacity towards the tail.

Cross-modal EOT. CFBE increases digital ASR by 3.9 percentage points. Its transformations target manufacturing and acquisition variations relevant to physical transfer. The physical results further demonstrate the effectiveness of the complete framework.

Interaction. The three individual gains sum to 23.0 points while the combined configuration gains 20.6, so the modules are slightly sub-additive rather than strictly independent. This is unsurprising, since STCO and AAAS both act on the hardest viewpoints — STCO by giving the pattern more expressive capacity there, AAAS by spending more updates there — and partially address the same deficit. The combined configuration nonetheless exceeds the strongest single-module variant by 8.6 points, confirming that no single module subsumes the others.

I Physical Experiment Details

I.1 Sticker Fabrication

Fabrication follows the two-stage process of Sec. 4.9. In the first stage the optimised RGB texture $\tau^{(i)}$ is printed on poster paper with a standard colour printer, with the region enclosed by the contour $S^{(i)}_{\text{thm}}$ left blank. In the second stage a 0.08 mm aluminium film is cut to that contour and adhered to the reserved blank region, giving a single planar sticker that carries the adversarial signal in both modalities at once.

The mechanism is the emissivity contrast between the two materials. Aluminium film has emissivity $\sim$0.1 against the vehicle panel's $\sim$0.8, so under the same surface temperature the film radiates far less and appears as a dark region in the thermal image, reproducing the optimised contour. In the visible image the same film appears silver-grey while the surrounding printed poster carries $\tau^{(i)}$. Each sticker takes about 15 minutes and 0.5 USD of material to produce.

I.2 Model Vehicle and Sticker Scaling

Physical validation is carried out on 1:24 scale model vehicles. The $N = 7$ stickers are scaled from their full-vehicle dimensions to the model, and are applied at the UV placements $P^{(i)}$ optimised in the digital pipeline. All six configurations compared in Table 1 — clean, random, UAP, TOUAP, CDUPatch and ours — are realised on identical model vehicles and photographed under the same conditions, so that differences between rows are attributable to the pattern rather than to the capture session.

J Full Defense Results

Table A6 reports the effect of each defense of Appendix F.4 on our attack against DAMSDet (Guo et al., 2024). We report both the absolute decrease in ASR relative to the undefended detector and the resulting post-defense ASR, since the two together are what determine whether a defense is practically useful.

Preprocessing defenses are largely ineffective. The nine preprocessing defenses recover between 0.3 and 5.2 points of ASR, leaving the attack at 0.772–0.821. This is consistent with how the perturbation is constructed. PixelMask, Jedi and JPEG compression are designed against localised, high-frequency, digitally injected patches; our perturbation is instead a set of large, low-frequency, physically realisable regions whose thermal component is a smooth emissivity contrast rather than a

Table A6: Defenses against our attack on DAMSDet (Guo et al., 2024). ΔASR is the absolute decrease relative to the undefended detector (ASR $= 0.824$); larger is better. Post-defense ASR is the residual attack success rate. Best result in bold.

Defense	Type	ΔASR	Post-def. ASR
PixelMask (Agarwal et al., 2021)	Preproc.	0.051	0.773
Jedi (Tarchoun et al., 2023)	Preproc.	0.013	0.811
Bit squeezing (Xu et al., 2018)	Preproc.	0.003	0.821
Spatial smoothing (Xu et al., 2018)	Preproc.	0.030	0.794
JPEG compression (Guo et al., 2018)	Preproc.	0.052	0.772
UDF (Yu et al., 2022)	Preproc.	0.049	0.775
FNC (Yu et al., 2021)	Preproc.	0.025	0.799
EPGF (Zhou et al., 2020)	Preproc.	0.009	0.815
SAC (Liu et al., 2022)	Preproc.	0.005	0.819
Adv. training (Zhu et al., 2025)	Training	0.035	0.789
Angle-adaptive adv. training (ours)	Training	**0.186**	**0.638**

texture. Bit squeezing is the weakest defense in the table (0.003), which follows directly: reducing bit depth from 8 to 7 removes exactly the kind of fine quantisation structure that the total-variation regulariser of Eq. (A5) already suppressed during optimisation, so there is little left for it to destroy.

Standard adversarial training underperforms on the hemisphere. Standard adversarial training recovers only 3.5 points, less than JPEG compression. The reason is not that adversarial training is ineffective in principle but that uniform viewpoint sampling misallocates it: because the attack's success rate varies strongly across the hemisphere, a uniformly drawn set of adversarial examples is dominated by viewpoints the detector already handles, and the viewpoints that actually fail are underrepresented in the defensive signal.

Angle-adaptive adversarial training.

Angle-adaptive adversarial training. Reweighting the adversarial detection loss by $\beta/r(\theta_i, \phi_j)$ across viewpoints recovers 18.6 points, more than five times standard adversarial training and more than three times the best preprocessing defense, at identical sample and step counts. The comparison is therefore between two allocations of the same supervision, not between different amounts of it, which is what isolates the contribution of angular adaptivity.

The attack remains a threat. Even under the strongest defense, ASR remains at 0.638. Two caveats bound the strength of this claim, and we state them explicitly. First, the defense is evaluated against a *fixed* pattern: an adaptive adversary that re-optimises against the hardened detector would likely recover part of the lost ASR, and we did not measure this. Second, we do not report the clean detection accuracy of the adversarially trained detectors, so the operating cost of the defense is not established here. Both are natural next steps for work that treats the defense, rather than the attack, as its primary object.