TransWardX: An Explainable Black-box Object Detection Attack for Autonomous Driving in Transitional Weather Conditions

Madhavi Kondapally, K Naveen Kumar, C Krishna Mohan, *Senior Member, IEEE*

***Abstract*— Navigating autonomous vehicles in adverse weather conditions poses significant challenges, such as identifying road elements and efficient path planning. We encounter frequent transitions between these adverse weather conditions (cloudy to rainy, rainy to sunny, etc.) in nature. Existing object detection research for autonomous vehicles neglects dynamic weather transitions, leading to performance degradation in adverse and transitional weather conditions like rainy, foggy, and cloudy. Also, autonomous driving data is vulnerable to minor perturbations or noises, leading to unpredictable outcomes, particularly in adverse weather conditions. In addition, current adversarial attacks demand significant computational resources and have limited real-world applicability due to the large number of queries and computational resources. To address these limitations, we propose a novel minimalistic method called explainable black-box adversarial detection attack in transitional weather conditions for autonomous driving (TransWardX). Our attention-guided attack is minimalist, targeting limited image regions to deceive the model effectively. We assess our attack using a continuous weather-driving dataset called AIWD6. Later, we also evaluate our attack with other datasets like BDD100K and GTSRB. Our results demonstrate the effectiveness of TransWardX, achieving a high success rate with minimal perturbations, fewer iterations, and a drastic reduction of 50% in computational requirements while maintaining low average precision.**

I. INTRODUCTION

Today, autonomous vehicle (AV) technology is progressing rapidly and inching closer to widespread implementation. Nevertheless, persistent challenges such as adverse weather conditions continue to affect AV performance significantly [1]. While initial efforts have been made to provide driving datasets of various weather conditions [2], each dataset has a limited set of weather scenarios, solely rain, fog, or snow. In reality, sudden shifts between different weather conditions are common, demonstrated by instances where a cloudy day rapidly evolves into a rainy one. Furthermore, the continual and consistent fluctuation in weather data presents a significant challenge for modern learning systems [1], [3]. To effectively tackle this issue, we introduced AIWD6 in our previous work [1], a novel continuous weather dataset designed to represent the dynamic characteristics of the real world by incorporating continuous shifts in weather patterns as shown in Fig. 1. On the other hand, when trained on extensive, high-quality data, existing object detection models have demonstrated impressive performance under normal conditions [4]. However, deploying these models in continuous weather scenarios is challenging due to the diverse visual variations present [5]. Besides the reduced illumination from adverse weather conditions, the data is susceptible to minor perturbations [6]. These perturbations could be small variations or distortions in the data, which may adversely affect the performance of object detection models, further complicating their deployment in adverse weather scenarios [7].

Fig. 1. Weather transition states: a) Foggy to Sunny b) Sunny to Foggy c) Cloudy to Rainy.

Adversarial attacks in machine learning aim to mislead models by introducing subtle alterations to the input data. This process helps uncover vulnerabilities in machine learning models, aiding researchers in designing more robust models [8]. Existing adversarial object detection attacks in machine learning suffer from several shortcomings: *(i)* limited interpretability *(ii)* high computational complexity *(iii)* a focus on effectiveness over efficiency, disregarding the adversary's goal of minimal data modification for strong mispredictions within a low attack budget [8]. To overcome the above limitations, we focus on designing a novel strategy for an *effective and efficient* attention-guided minimalistic black-box object detection attack that uses a small number of carefully chosen input perturbations to fool a machine learning model. This attack significantly alters the output of a model with only a small number of changes to the input data to exploit vulnerabilities in deployed models cost-effectively.

In this paper, we design TransWardX, a novel minimalistic explainable black-box adversarial object detection attack for transitional weather driving that significantly improves the attack efficiency by employing an attention-guided mechanism in *explainable artificial intelligence (XAI)* module [9] called class attention map (CAM) [10]. Further, we add the random gradient noise to the highlighted CAM pixels, which creates a poisoning effect in a *sparse* manner. Thus, we

Madhavi Kondapally, K Naveen Kumar, and C Krishna Mohan are with the Department of Computer Science and Engineering, Indian Institute of Technology, Hyderabad, (e-mail: cs21resch15001@iith.ac.in; cs19m20p000001@iith.ac.in; ckm@cse.iith.ac.in).

exploit the prior knowledge of attention areas to overcome the significant query needs for building a black-box attack for object detection on high-dimensional input images. Our method represents a pioneering approach, the first of its kind to target autonomous vehicle data, specifically in transitional weather conditions with adversarial attacks. To summarize, the following are the main contributions of our work. (i) We introduce a novel black-box adversarial object detection attack method utilizing an attention-guided mechanism for autonomous vehicles in transitional weather conditions. Our approach employs a gradient attention map and query-based finite difference perturbations, resulting in a cost-effective attack. (ii) We present a new research direction by proposing an attack on object detection task for autonomous driving data under transitional weather conditions, which remains an unexplored area in existing research to the best of our knowledge. (iii) We analyze our attack's efficacy and efficiency using five distinct metrics on a novel transitional weather dataset called AIWD6 across three categories: four-wheeler, two-wheeler, and pedestrian. Additionally, we conduct an extensive ablation study on two other datasets, BDD100K and GTSRB, to further evaluate our method's performance.

II. Related Work

A. Autonomous Driving in Adverse Weather Conditions

Data scarcity is a major hurdle in adverse weather scenarios. Many studies have proposed methods for synthesizing weather conditions in autonomous vehicles. Musat *et al.* [11] utilized GAN and CycleGAN to produce seven distinct weather conditions. Hu *et al.* [12] adopted a formulation to synthesize rain & fog and created a raincityscapes dataset. Unlike these, Sun *et al.* [3] introduced SHIFT, a synthetic driving dataset for performing scene perception tasks under discrete and continuous domain shifts. However, simulators have limitations regarding variability, realism, complexity, and scalability. Hence, in our work [1], we generated transitional weather data by varying the intensities of the weather using generative approaches. To address object detection challenges in unfavourable weather conditions, predominant research efforts majorly focused on developing restoration techniques like draining [12], [13]. Hassaballah *et al.* [13] introduced a restoration method to detect vehicles in adverse weather conditions. Yizhou *et al.* [14] proposed an autoencoder to learn embeddings for reconstructing a rainy layer close to the ground truth. Different from the above restoration methods, numerous efforts have focused on domain adaptive strategies to adapt from clear weather to adverse weather [15], [16]. Liu *et al.* [15] proposed an image-adaptive YOLO to improve object detection in weather conditions. Hua *et al.* [16] introduced a causal intervention reasoning module to acquire domain invariant features. Unlike these, in our previous work, we proposed a multiscale adaptive detection transformer to perform object detection in transitional weather conditions for AVs [5]. Though these works achieved promising results, it is challenging to perform object detection in transitional weather conditions if the data undergoes minor perturbations.

B. Adversarial Attacks on Object Detection for Autonomous Vehicles

Several methods have been introduced to attack autonomous driving data, including adversarial patch and noise attacks by Svetlana et al. [7]. These attacks utilize projected gradient descent and fast gradient descent techniques to manipulate AV data for object detection (OD). Additionally, Shapira et al. [6] proposed a universal adversarial perturbation (UAP) targeting non-maximum suppression (NMS) to prolong the duration required for OD. Although these black-box attacks target AVs for object detection, they are limited by efficient queries, unnecessary computations, and slower attack convergence for generating gradient perturbations on high-dimensional input space. Also, the literature has a limited focus on object detection attacks for AVs under transitional weather conditions. Hence, we propose an attack method in a black-box setting by leveraging explainable AI in a minimalistic way to attack object detection in transitional weather-driving data. Our attack method generates robust adversarial samples, ensuring efficacy and efficiency within the same framework.

III. Proposed Approach

The proposed approach, TransWardX, an explainable black-box adversarial detection attack for transitional weather driving, comprises two stages. In the initial phase, we employ a multi-scale adaptive transformer (mSAT) [5] to train the model on AIWD6 data for object detection. Secondly, we introduce an explainable black-box detection attack module to attack autonomous driving data under transitional weather conditions in a minimalistic way. In the first stage to train the mSAT model, initially, we extract features from AIWD6 images using ResNet50 [17]. Subsequently, we incorporate positional embeddings and feed these features into the transformer to obtain class-level predictions and bounding boxes to perform object detection without any attack settings. Afterwards, we utilize this trained model to execute our adversarial attack. In the second stage, initially, we crop the relevant class category based on the ground truth annotations to create a target image for performing the TransWardX attack, as shown in Fig. 2. It consists of two modules: *(i)* attention-guided module and *(ii)* black-box attack module. In the attention-guided module, we send transitional weather images to the already trained ResNet50 backbone of the mSAT model in the previous stage and generate heatmaps using gradcam to identify specific regions for the attack. CAM (C_x) for the input image X is generated by computing a linear combination of weighted activations, followed by a ReLU layer using

$$C_x = relu(\sum_k w_k^c A_{ij}^k), \tag{1}$$

where w_k^c represents the weighted average of the pixel-wise gradients, and A_{ij}^k represents the convolutional feature map of a particular class c generated by model h. Attention aids ML models in focusing on crucial input data for minimal predictions. An attention-guided mechanism enhances minimalistic

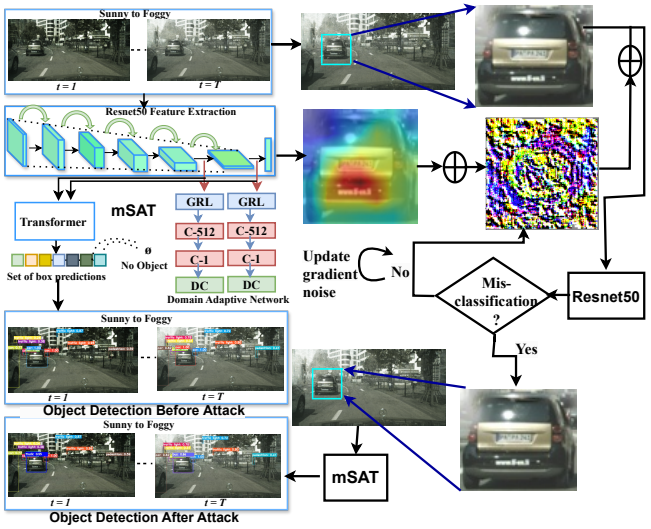

Fig. 2. TransWardX utilizes ResNet50 for gradcam extraction from weather transition images, followed by the attention-guided module for heatmap generation in order to identify specific regions to attack. Random gradient perturbation creates adversarial samples, which are tested for misclassification. Misclassified samples are analyzed using mSAT [5] for object detection. Iterative refinement improves adversarial effectiveness.

attacks by pinpointing relevant data parts and localizing critical image areas efficiently. Further, in the black-box attack module, we randomly choose the initial gradient perturbation and add it to the target object guided by these attention maps in order to perform the attack. Next, we verify the perturbed image for false detection. If the given object is misclassified correctly, perturbation values proceed for successful misprediction, otherwise, additional perturbations are iteratively introduced. During back-propagation, only the input pixels get gradient updates, unlike parameters, as we create the initial adversarial image that causes false or no detection. This iterative procedure continues until the model misclassifies the object X with a probability $P(\hat{Y}/X) \neq Y$ resulting in a final sequence of imperceptible adversarial images with high success rates.

Attack definition: Given a clean input image X, the adversary creates a small perturbation δ such that the prediction $h(X+\delta) \neq Y$. Since gradient information is absent for the black-box h model, the output probabilities will serve as a strong proxy to *guide the search for the gradient perturbation* that generates the final adversarial image.

The adversarial image X_{adv} is calculated as the bit-wise addition of two terms: the input image X and the product of step-size and random gradient perturbation Gp:

$$X_{adv} = X + \varepsilon * Gp, \tag{2}$$

where ε is a hyperparameter. Highlighting the trade-off, a model with a higher value of ε can be "fooled" easily at the cost of easy identification by the human eye. The final misclassified object X_{adv} is incorporated into the original input image for object detection. The ResNet model utilizes the perturbed image to compute the class score (y_T), where $y_T = P(\hat{Y}_{c=t} \mid X)$ for the given class, t, and $\hat{Y}c$ represents the predicted class by the model. This iterative process of updating the adversarial image ($Xadv$) according to Eq. (2) continues until the convergence of y_T. For the initial iteration, the gradient noise is added in the positive direction. The gradient is updated in the negative direction for further iterations and is changed to random perturbations in subsequent iterations. This iterative method creates an adversarial image classified as a specific desired target class.

Algorithm 1 Query-based Black-box Attack Module

Require: X, original input image, h, trained model
Ensure: X_{adv}, final adversarial image
1: Calculate CAM as $C_x = relu(\sum_k w_k^c A_{ij}^k)$ using h
2: Initialize $X_{adv} = X + \varepsilon * Gp$
3: Pass X_{adv} to h for inference
4: Define $Y \leftarrow$ input class label
5: Initialize $temp \leftarrow 0$, $ifGrad \leftarrow 0$
6: Define $y_T \leftarrow$ probability of targeted class
7: Define $\hat{Y}_{c=t} \leftarrow$ predicted class label c as target label t
8: **while** $h(X_{adv}) == Y$ **do**
9: **if** $y_T < temp$ and $ifGrad == 0$ **then**
10: Update $Gp \leftarrow -(Gp)$
11: $ifGrad \leftarrow 1$
12: **else**
13: Update Gp with random perturbation
14: $ifGrad \leftarrow 0$
15: **end if**
16: **if** $X_{adv} + (\varepsilon * Gp) \in L_2$ norm **then**
17: Update $X_{adv} = X + \varepsilon * Gp$
18: **end if**
19: $temp \leftarrow y_T$
20: Pass X_{adv} to h for inference, i.e., $\hat{Y}_{c=t} = h(X_{adv})$ (similar to line 4)
21: Update probability score $y_T = P(\hat{Y}_{c=t} \mid X)$
22: **end while**
23: return X_{adv}

IV. Implementation

A. The AIWD6 dataset

The proposed TransWardX is evaluated on AIWD6 dataset [1] a continuous weather dataset consisting of six weather transition states, such as *sunny to foggy (SF), foggy to sunny (FS), cloudy to rainy (CR), rainy to cloudy (RC), sunny to rainy (SR), and rainy to sunny (RS)* generated using variational autoencoder (VAE). It comprises 7,032 weather sequences, corresponding to sequence lengths 10, 50, 75, and 125. The total number of images generated for all the sequences is 4,57,000. Fig. 1 visually represents the continuous shifts of the AIWD6 dataset. The AIWD6 dataset consists of weather transitions with variations in adversity. For example, in the *cloudy to rainy* transition of Fig. 1, we can see various intensity levels of the rain. Similarly, other transition states have various intensities of weather conditions. Hence, this dataset is very significant in real-world scenarios. We annotated the AIWD6 dataset in COCO JSON format to detect eight classes, i.e., pedestrian, bicycle, rider, car, motorcycle, bus, train, truck, and traffic light, using the LabelImg annotation tool [18]. Among these classes, in this paper, we focus on three categories of objects for attack: cars in the four-wheeler category, riders in the two-wheeler category, and pedestrians as these are dominant classes that are observed in heterogeneous traffic scenarios.

Metrics: We categorize the quantitative results in terms of efficiency and efficacy. *Efficacy* is the algorithm's capacity

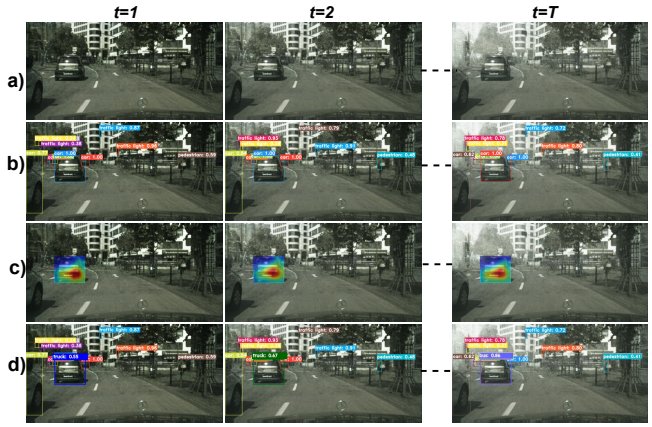

Fig. 3. Qualitative analysis of AIWD6 car object attack. (a) Input (b) Pre-attack: Car detected (c) Attack region (d) Post-attack: False detection.

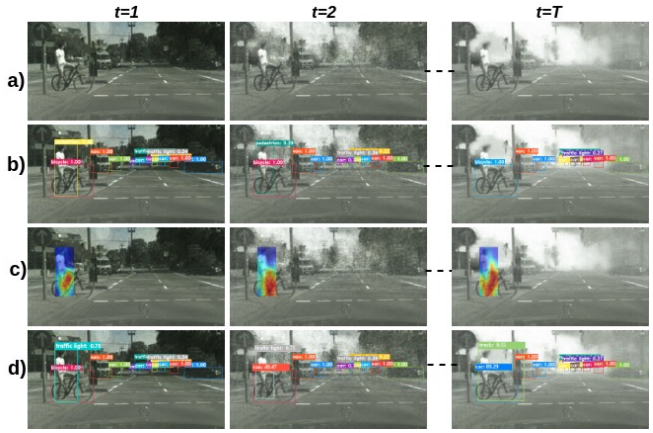

Fig. 4. Qualitative analysis of AIWD6 rider object attack. (a) Input (b) Pre-attack: Rider detected (c) Attack region (d) Post-attack: False detection.

for generating consistent misprediction over multiple images in the dataset. *Attack success rate and average precision* are the two metrics in this bucket. These are conventional metrics for any adversarial detection attack. *Attack efficiency* signifies the minimum time and computational resources required to generate the adversarial samples. This newer bucket includes the *number of iterations, computations, and sparsity* of attack. It is important to emphasize that our work presents a comprehensive evaluation with five different metrics, whereas prior works [19], [20] have limited to about half of them.

B. Multi-Scale Adaptive Transformer (mSAT)

In our experiments, we employed mSAT [5], a transformer-based architecture incorporating a domain adaptation network to extract domain-invariant features, facilitating object detection tasks. Resnet50 is a backbone for extracting features from the input with size 256×512.

C. Attack Settings

We set $\varepsilon = 0.1$, $\zeta = 7000$ and the maximum attack iterations as 10000 as hyperparameters (empirically determined as decreasing ε results in poor attack results and increasing ε leads to detection by the human eye).

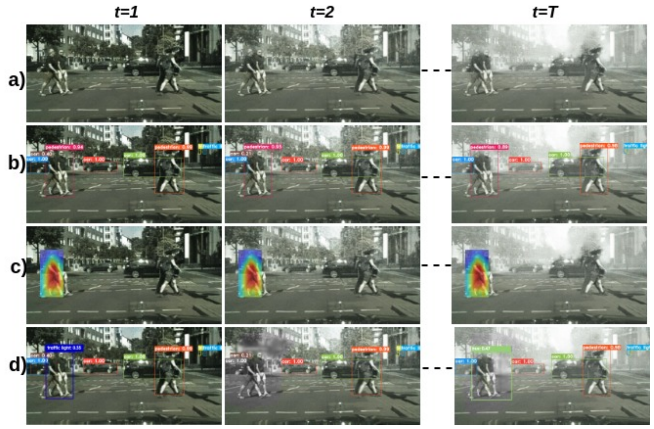

Fig. 5. Qualitative analysis of AIWD6 pedestrian object attack. (a) Input (b) Pre-attack: pedestrian detected (c) Attack region (d) Post-attack: False detection.

V. Evaluation

We evaluated our attack method on the AIWD6 dataset, targeting three categories of objects. *(i)* Four-wheeler (car) *(ii)* two-wheeler (rider) *(iii)* Pedestrian, in terms of two factors: *(a)* efficiency *(b)* minimalism.

Efficiency: Table I shows the quantitative performance of our attack in terms of average precision (AP) and attack success rate (ASR) on each category before and after the attack. AP and ASR values clearly state the effectiveness of our attack in compromising the detection performance for all targeted object categories.

Minimalism: We evaluate minimalism using three metrics: *(a)* Sparsity *(b)* Number of iterations *(c)* Computations. Sparsity represents the attack budget. It is calculated as the ratio of the number of *unmodified pixels* to the *total number of original pixels*. A very sparse attack (higher value) is efficient as minimal pixels get modified, thereby preserving the attack budget. Sparsity is shown in Table II. TransWardX algorithm searches the gradient perturbation through multiple iterations until it meets the convergence criteria, and our algorithm converges quickly. The overall iterations are given in Table II. Fewer iterations drastically reduce memory and computational resource usage. Reduced computations are a desirable outcome for resource management. Hence, we design an economic attack. These results are shown in the Table II. Our method requires fewer computations ($\approx$200 Billion reductions). The improved efficiency is due to sparse localized perturbations of the attention-guided module. *In summary, our experiments demonstrate that our attention-based attack consistently generates strong adversarial samples at a fast pace with few computations.*

A. Four-wheeler

In this attack category, we focus on a four-wheeler vehicle, namely, a car, as it is the dominant four-wheeler class that is commonly observed in heterogeneous traffic scenarios. Fig. 3 illustrates the qualitative evaluation of our proposed method during the transition from sunny to foggy conditions. Table I and II show the quantitative analysis of our attack on

TABLE I

OBJECT DETECTION PERFORMANCE OF OUR APPROACH MEASURED WITH AP WITH AND WITHOUT ATTACK.

Category	No Attack (AP)	Attack (AP) ↓	Success rate (%) ↑
Four-Wheeler	68.1	17.5	80.23
Two-Wheeler	61.8	18.7	82.30
Pedestrian	70.2	16.5	79.65

the four-wheeler (car) category *w.r.t* efficiency and efficacy. The detection performance (mAP) of the car diminishes and has a high impact after the attack, with a lower number of iterations and a low attack budget. Fig. 3(a) displays the sequence of inputs sent to the TranWardX. Fig. 3(b) illustrates the object detection performance before the attack, where all objects, including cars, are accurately detected. However, after applying the attack, as seen in Fig. 3(c), attention heatmaps reveal the specific areas targeted during the attack, focusing on the car object. The red region has higher activation and, hence, carries more gradient weight. This also implies that applying subtle gradient perturbations to the red region is key to a successful adversarial attack. Attacking on only attention maps allows the adversary to stay within the attack budget. In Fig. 3(d), the object detection performance after the attack shows a significance that the targeted car undergoes false detection.

B. Two-wheeler

In this attack category, we focus on a two-wheeler vehicle, namely, a rider, as it is the dominant two-wheeler class that is commonly encountered in diverse traffic scenarios. Table I and II show the quantitative analysis of our attack on the two-wheeler (rider) category *w.r.t* efficiency and efficacy. The rider's detection performance (mAP) significantly declines post-attack, even with fewer iterations and a limited attack budget. Fig. 4 showcases the qualitative evaluation of our proposed method on the rider object during the *Sunny to Foggy* transition sequence. Continuing from the previous section, a similar visual analysis is shown in Fig. 4. Fig. 4(d) reveals that object detection performance after the attack shows the rider falsely identified as a traffic light and a truck. Furthermore, the bicycle on which the rider is mounted is also affected, falsely detected as a van and a car.

C. Pedestrian

Table I and II show the quantitative analysis of our attack on pedestrian category *w.r.t* efficiency and efficacy. The pedestrian's detection performance declines following the attack, despite fewer iterations and a constrained attack budget. Fig. 5 illustrates the qualitative evaluation of our proposed method on the pedestrian during the transition from sunny to foggy. Similar to the previous sections, it is observed that in Fig. 5(d), the object detection performance after the attack has instances of false detection and, in some cases, failure to detect the pedestrian altogether.

VI. ABLATION STUDY

This section presents the ablation study of our attack on other datasets. We extend our attack to perform traffic sign classification on the GTSRB and KBTS datasets. Traffic signs were misclassified after our attack, as shown in Fig. 6 and 7. Misclassifying traffic signs significantly impacts the safety of AVs. Also, we extended this task to perform traffic sign detection on the BDD100K dataset. Fig. 8 shows the qualitative performance of our method with no attack and attack category on the traffic sign of BDD100K. After attack, the traffic sign undergoes false detections and no detections. Table III shows the quantitative comparison of our attack on the BDD100K dataset with existing attacks. Our TransWardX clearly outperforms other methods. Later, we extended our attack to the segmentation task, and qualitative analysis of our attack on the segmentation for the BDD100K dataset is shown in Fig 9. Here, we can clearly observe that the targeted objects have false or no predictions.

TABLE II

QUANTITATIVE EVALUATION OF OUR TRANSWARDX IN TERMS OF MINIMALISM ON AIWD6 DATASET. B → BILLION.

Category	Sparcity (%)↑	Iterations ↓	Computations ↓
Four-wheeler	75.5	4486	97B
Two-wheeler	74.4	4372	94B
Pedestrian	73.9	4312	95B

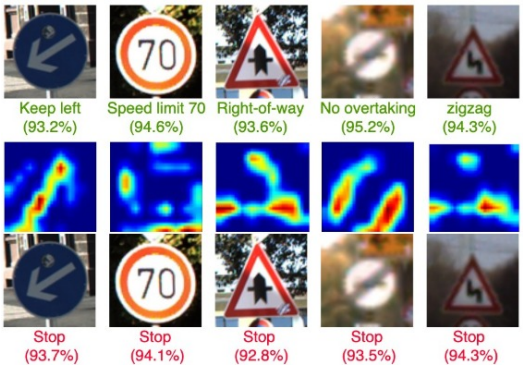

Fig. 6. Qualitative analysis of our attack on traffic sign classification for GTSRB dataset.

VII. CONCLUSION

Scene perception tasks for autonomous driving face significant challenges in adverse weather conditions, particularly during transitional weather events. Moreover, object detection models relying on image data are susceptible to adversarial attacks. In this paper, we investigate the vulnerability of object detection models in autonomous driving under transitional weather conditions by proposing TransWardX, a novel attention-guided, minimalistic query-based black-box attack method. TransWardX leverages the attention mechanism of the explainable AI technique known as class attention maps to reduce computational resources substantially. Using a query-based model, our technique identifies 'interest areas' for generating gradient noise. Extensive quantitative and qualitative evaluation of the TransWardX attack across three categories on the AIWD6 dataset substantially impacts detection accuracy. In future work, we focus on extending TransWardX to video contexts and developing robust defense techniques against such adversarial attacks.

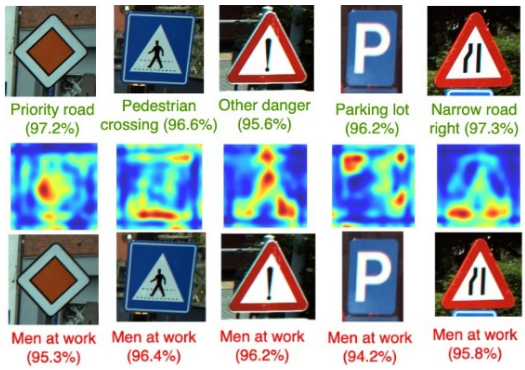

Fig. 7. Qualitative analysis of our attack on traffic sign classification for KBTS dataset.

Fig. 8. Qualitative analysis of our attack on traffic sign object detection on BDD100K dataset. a) Input b) Pre-attack: Traffic sign detected c) Attack region generated by gradcam d) Post-attack: False or no detection.

Fig. 9. Qualitative analysis of our attack on traffic sign object segmentation on BDD100K dataset. a) Input b) Pre-attack: Traffic sign segmented correctly c) Attack region d) Post-attack: False or no segmentation.

TABLE III

QUANTITATIVE EVALUATION OF OUR ATTACK ON BDD100K DATASET.

No attack	PGD [21]	UDP [6]	FIA [19]	TPA [20]	Ours
53.5	19.19	16.37	17.49	16.17	5.85

REFERENCES

[1] M. Kondapally, K. N. Kumar, C. Vishnu, and C. K. Mohan, "Towards a transitional weather scene recognition approach for autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[2] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[3] T. Sun, M. Segu, J. Postels, Y. Wang, L. Van Gool, B. Schiele, F. Tombari, and F. Yu, "Shift: a synthetic driving dataset for continuous multi-task domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 371–21 382.

[4] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[5] K. Madhavi, K. N. Kumar, and C. K. Mohan, "Object detection in transitional weather conditions for autonomous vehicles," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 01–08.

[6] A. Shapira, A. Zolfi, L. Demetrio, B. Biggio, and A. Shabtai, "Phantom sponges: Exploiting non-maximum suppression to attack deep object detectors," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4571–4580.

[7] S. Pavlitskaya, N. Polley, M. Weber, and J. M. Zöllner, "Adversarial vulnerability of temporal feature networks for object detection," in *European Conference on Computer Vision*. Springer, 2022.

[8] K. N. Kumar, C. Vishnu, R. Mitra, and C. K. Mohan, "Black-box adversarial attacks in autonomous vehicle technology," in *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2020, pp. 1–7.

[9] R. Pierrard, J.-P. Poli, and C. Hudelot, "Spatial relation learning for explainable image classification and annotation in critical applications," *Artificial Intelligence*, vol. 292, p. 103434, 2021.

[10] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839–847.

[11] V. Muşat, I. Fursa, P. Newman, F. Cuzzolin, and A. Bradley, "Multi-weather city: Adverse weather stacking for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2906–2915.

[12] X. Hu, L. Zhu, T. Wang, C.-W. Fu, and P.-A. Heng, "Single-image real-time rain removal based on depth-guided non-local features," *IEEE Transactions on Image Processing*, vol. 30, pp. 1759–1770, 2021.

[13] M. Hassaballah, M. A. Kenk, K. Muhammad, and S. Minaee, "Vehicle detection and tracking in adverse weather using a deep learning framework," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4230–4242, 2020.

[14] Y. Li, Y. Monno, and M. Okutomi, "Single image deraining network with rain embedding consistency and layered lstm," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022, pp. 4060–4069.

[15] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, and L. Zhang, "Image-adaptive yolo for object detection in adverse weather conditions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1792–1800.

[16] H. Zhang, L. Xiao, X. Cao, and H. Foroosh, "Multiple adverse weather conditions adaptation for object detection via causal intervention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] Tzutalin, "Labelimg," https://github.com/tzutalin/labelImg, 2015.

[19] Z. Wang, H. Guo, Z. Zhang, W. Liu, Z. Qin, and K. Ren, "Feature importance-aware transferable adversarial attacks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7639–7648.

[20] Y. Lu, H. Ren, W. Chai, S. Velipasalar, and Y. Li, "Time-aware and task-transferable adversarial attack for perception of autonomous vehicles," *Pattern Recognition Letters*, vol. 178, pp. 145–152, 2024.

[21] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.