Grenze International Journal of Engineering and Technology, January Issue

Adversarial Attacks on Autonomous Vehicles: A Comprehensive Review

Shivangi Bisht[1], Rabbika Azmi[2] and Shweta Jindal[3]
[1-3]Indira Gandhi Delhi Technical University for Women, New Delhi, India
Email: shivangi62442@gmail.com, azmirabbika@gmail.com, miss.shweta.singhal@gmail.com

***Abstract*— Growing advancements in artificial intelligence and autonomous vehicles have led us to a broader problem than just the functioning of these technologies. While autonomous vehicles (AV) have greatly improved in design and build to the point that we now have self-driven taxis in some parts of the world, it doesn't change the fact that there's a need for constant improvement in the system. This paper talks about the adversarial defense mechanisms in deep learning for autonomous vehicles, in particular, we will explore current research to gain a better understanding of progress and challenges within this field, as well as how adversarial attacks influence safety and critical choices made by autonomous vehicles. In addition to analyzing existing research, our review paper aims to offer novel insights into the long-term impacts of adversarial attacks on the functionality and reliability of AI models in autonomous vehicles. In essence, our review of 23 papers reveals a vital need to overcome the shortcomings of Autonomous Vehicles (AVs) and how there is a growing interest in studying adversarial attacks, which emphasizes the need of improving the safety and reliability of these systems.**

***Keywords*— Autonomous Vehicles, Adversarial Attacks, Deep Learning, Ar-tificial Intelligence.**

I. INTRODUCTION

Autonomous Vehicles are cars that come with technology to drive and navigate on their own, employing sensors, cameras, and artificial intelligence to sense the surroundings and decide how to drive. Adversarial attacks are malicious methods used to mislead AI systems like those in autonomous vehicles by altering input data to give the wrong outputs. Adversarial attacks are capable of attacking the decision-making processes of autonomous vehicles, bringing about serious security and safety issues. There exist mainly two forms of Adversarial attacks, Digital Attacks, they use techniques that fool the deep neural networks (DNN), and, Physical Attacks, these attacks manipulate actual-world objects to mislead the vehicle's sensors [40]. For example, they include stop signs being modified using stickers. Recent developments in the field of AVs have led to a renewed interest potentially in making transportation safer, efficient and more accessible. Howev-er, day by day as we humans are diving more into deep learning algorithms for critical tasks such as navigation, path planning, object detection in self-driving automated vehicles. But it must be noted that though we're using these for our own advantages, these can also be exploited through adversarial attacks which include manipulation of input data or incorrect decision [1][2] making despite the model training posing risks to the passenger but yes, also the public trust in autonomous systems [41].

Grenze ID: 01.GIJET.12.1.574
© Grenze Scientific Society, 2026

Adversarial attacks on autonomous vehicles occur because these systems rely heavily on machine learning algorithms, which can be vulnerable to subtle manipulations in input data. Attackers can exploit these vulnerabilities by introducing deceptive elements, such as altered traffic signs or physical objects, that confuse the vehicle's perception systems. As a result, the vehicle may misinterpret its environment, leading to potentially dangerous situa-tions on the road.

A. Background Study

A number of researchers have made significant contribution in the current field of knowledge of adversarial attacks on autonomous vehicles, investigating various methods and results across this diverse topic. For example, Fan, Wang, and Li (2024) demonstrated how generative adversarial networks (GANs) can be used to launch attacks on trajectory prediction systems, which can potentially lead to unsafe driving conditions. Their findings highlight the need for robust defenses against such vulnerabilities in autonomous driving technologies. They made use of the Apollos cape Trajectory dataset focused on urban road scenarios and also the NGSIM dataset which specializes in highway scenarios, this predict-ed errors by 50% across the vehicles [3], wherein, Dipalma et al. (2021) explored the security-based camera perception systems issue and demonstrated how such attacks can compromise the reliability of perception systems, leading to danger-ous outcomes on the road [5]. The use of invisible optical adversarial stripes on traffic signs poses a vital threat to the safety of autonomous vehicles. Guo et al. (2024) explored this innovative attack method, demonstrating how these little changes can lead to misinterpretation of traffic signs by autonomous systems. They made use of the German Traffic Sign Recognition Benchmark (GTSRB), and achieved a 94% success rate for untargeted attacks and 97% success rate for tar-geted attacks [12]. Additionally, Teng et al. (2023) proposed PAID, a mechanism for analyzing perturbed image attacks and detecting intrusions in autonomous driving systems. They focused on Evasion attacks targeting Road Context-aware Intrusion Detection Systems (RAIDS). The technique, PAID combines feature squeezing and GPS coordinates with RAIDS to combat perturbed image attacks effectively. Their findings were 93.9% accuracy in detecting such attacks [18]. Despite the fact that recent researches have proposed a number of crucial tech-niques to identify and counteract the adversarial attacks in autonomous vehicles, these techniques still have a number of limitations, such as difficulties from complex and dynamic environments, decreased efficiency against new or adap-tive attack strategies, high costs that prevent real-time deployment, and challeng-es ensuring robustness and reliability across various scenarios, lack of testing on real-world or performance degradation on clean data, which might not be ideal for all applications.

B. Rationale for the review

Autonomous vehicles are becoming an important part of our daily lives; they promise safer and more efficient transportation. However, these systems can be tricked by adversarial attacks, they can be deliberate attempts to confuse their sensors or algorithms based, which could lead to dangerous situations on the road. While many studies and researchers have found ways to defend against these attacks, there’s still a lot we don’t fully understand about how well these defenses work in real-world conditions. This review takes a closer look at the latest research to highlight what’s been achieved, what challenges remain, and where future efforts should focus, by providing insights to guide future developments toward more robust and secure autonomous driving systems.

II. METHODOLOGY

A. Major Technologies

Throughout reviewed literature, a broad range of technologies has been uti-lized to enhance adversarial robustness, perception accuracy, and system resili-ence in deep learning for autonomous vehicles. Object detection and perception architectures are a core part of most studies, with networks like YOLO-v2[9], YOLO-v5[8][23], RetinaNet[23], Faster R-CNN [23], and EfficientNet[6] being utilized extensively for object detection and localization under adversarial at-tack. Various 3D object detection and tracking are tackled with LiDAR-based architectures like PointRCNN[10], PointPillar[10], PV-RCNN[10], and DEEP learning-based point cloud segmentation tools like PointNet++[16], Squeez-eSeg[16], and Cylinder3D[16]. Tracking architectures like DEEP-SORT[8], CasTrack[8], and measurements like HOTA[8], MOTA[8], and Preci-sion/Recall[8] are also widely utilized to measure tracking performance under perturbed inputs. From the adversarial robustness point of view, different studies employed generative models and attack/defense methods such as Generative Adversarial Networks (GANs)[3][21], FGSM[22], PGD[22], and evolutionary methods such as Natural Evolutionary Strategies (NES)[7] and Finite Differences (FD)[7]. Defensive training methods were mostly adversarial retraining, while mission planning and safety guarantee employed reinforcement learning tech-niques such as

Deep-RL with Proximal Policy Optimization (PPO) [15] and signal temporal logic (STL) [15]. Model resilience was explored by some studies using light-weight models such as SqueezeNet[6], ResNet-18 [6], Inception-V3 [6], while others explored robustness using trajectory prediction models such as GRIP++[3][21] and simulators such as the LGSVL simulator[4].

Sensor fusion was also a significant trend, with several papers combining data from LiDAR (e.g., Velodyne VLP-32C), radar (TI AWR1843), and RGB cameras (e.g., Allied Vision Mako G-319)[13] for enhanced detection fidelity. Lane detec-tion was tackled using SCNN [14], LaneATT [14], UltraFast [14], and PolyLane-Net [14], typically supplemented with image scaling and transformation methods. Lastly, mature tools like MATLAB/Simulink [20], CAN toolboxes [20], and simu-lation environments facilitated fine-grained control and testing, while computa-tional frameworks like PyTorch[18] and GPU-accelerated CNNs [20] facilitated scalable training of models. These technologies combined represent a rich and multi-modal approach to adversarial defense in autonomous systems.

B. Dataset

The reviewed articles utilize a vast number of datasets to research adversarial defense systems in autonomous driving systems. Datasets such as ApolloScape Trajectory [26] and NGSIM [27] are utilized for motion planning and trajectory prediction tasks to generate useful data to simulate real driving conditions. The LGSVL simulator in conjunction with the Baidu Apollo autonomous driving sys-tem [28] is extensively utilized to further simulate adverse driving conditions. Other datasets such as KITTI Multi-Object Tracking and Segmentation [31], Se-manticKITTI [31], and nuScenes [27] are extensively utilized for object detec-tion, tracking, and segmentation, giving a comprehensive overview of diverse driving conditions.

Traffic sign recognition is another important space, with datasets such as LISA Traffic Sign Dataset [29], German Traffic Sign Recognition Benchmark (GTSRB) [30], and Russian Traffic Sign Dataset (RTSD) [39] being heavily utilized for benchmarking recognition models. For the space of lane detection and weather, datasets such as TuSimple Challenge dataset [34] and RainDS [33] have played important roles in benchmarking models under adverse conditions such as weather. Some works have also made use of in-house datasets such as the AV-CPS simulation [36] for adversarial training, where normal and attack data sam-ples assist with robustness benchmarking. For larger object detection tasks, data sets like Microsoft COCO, Waymo Open Dataset [38], and VisDrone [37] are generally employed. Atari (Pong, BeamRider, Qbert, SpaceInvaders) and MuJoCo (Ant, Hopper, Walker2d, HalfCheetah) environments [36] are also employed in reinforcement learning studies to examine control policies under adversaries. The following data sets also provide multiple perception and decision-making models for autonomous systems: Udacity Self-Driving Car dataset, Chen 2017 and Chen 2018 datasets [36], and Argoverse [35].

II. Data Collection

In this section, we present a comprehensive summary of data collected for our study, which were from 23 distinct studies, organized into a table. Through this, we aim to highlight the vital trends and patterns that emerged during the process.

TABLE I. STRUCTURED MATRIX OF THE COMPREHENSIVE REVIEW

Ref.	Year	Type of Attack	Dataset Used	Results	Technique Used	Tools Used	Limitations
[3]	2024	white-box attacks and black-box attacks	Apollos cape Trajectory [26] and NGSIM [27]	prediction errors by over 50% across, with the most significant impact being a 694% rise in ADE and 513% rise in FDE on the Apollos cape dataset	Adv-GAN, this technique utilizes Generative Adversarial Networks (GANs)	Generative Adversarial Networks (GANs), Long Short-Term Memory (LSTM), Networks Model Predictive Control (MPC), GRIP++ Trajectory Prediction Model	The paper focuses on attack generation but does not extensively explore or propose robust defense mechanisms against such adversarial attacks
[4]	2021	practical adversarial patch attack targeting camera-based obstacle detection	LGSVL simulator with the Baidu Apollo autonomous driving system [28]	The adversarial patch attack achieved a 100% success rate in hiding the box truck from detection in the	Expectation over Transformation (EoT)	LGSVL simulator	The attacker is assumed to have full knowledge of the victim's perception system, which may not always be realistic.

				Baidu Apollo system during simulations			The attack is demonstrated in a simulated environment (LGSVL) and not validated in real-world scenarios.
[5]	2024	white-box attacks and black-box attacks	Apolloscape [26] NGSIM and nuScenes [27]	CausalRobTra improves performance on clean data by 20.6%	a novel method called CausalRobTra, which is a causal robust trajectory prediction method	Total Direct Effect (TDE)	While the method improves robustness under adversarial attacks, it still causes a 7.7% performance degradation on clean data, which might not be ideal for all applications
[6]	2023	White box attacks	LISA Traffic Sign Dataset [29]	EfficientNet: Original (95%), FGSM (92%), BIM (38%), PGD (50%) Inception-V3: Original (98%), FGSM (93%), BIM (46%), PGD (56%) SqueezeNet: Original (96%), FGSM (87%), BIM (53%), PGD (60%) ResNet-18: Original (95%), FGSM (86%), BIM (39%), PGD (49%)	Input Transformation-based Defense Techniques, specifically using Total Variance Minimization (TVM), Spatial Smoothing (SS), and Feature Squeezing (FS).	EfficientNet, SqueezeNet, ResNet-18, and Inception-V3	The study focused on white-box attacks and did not consider black-box attacks, which employ random perturbations. The defence methods only partially counter adversarial attacks; performance degradation still takes place.
[7]	2020	Black Box	German Traffic Sign Recognition Benchmark (GTSRB) [30]	M-SimBA outperformed SimBA and T-PGD in success rate, convergence speed, and confidence suppression	Modified Simple Black-box Attack (M-SimBA)	Finite Differences (FD) and Natural Evolutionary Strategies (NES). Two custom black-box CNN models for classification	focuses only on image-based attacks, ignoring video and sensor data. It also lacks testing on real-world physical attacks like altered road signs
[8]	2024	patch attacks, random removal attacks. Both black-box and white-box attacks are considered.	KITTI Multi-Object Tracking and Segmentation Dataset [31]	raising the HOTA score from 37.09 to 42.998, MOTA from 0.6400 to 0.7374, and recall from 0.7251 to 0.7834.	Multi-modal matching-based monitoring framework for attack detection	YOLO-v5 (for images), VirConv. DEEP-SORT (for 2D tracking) and CasTrack (for 3D tracking) HOTA, AB3DMOT, MOTA, Precision, Recall	does not discuss real-world attack scenarios for 3D object detectors. The conversion from 2D to 3D detection is also not optimized, and the approach lacks generalizability.
[9]	2022	physical adversarial patch attacks	INRIA-person dataset [32]	exceeding 90% under various adversarial patch attacks	Adversarial Patch-Feature Energy (APE)	YOLO-v2 object detector	adaptive attacks with full knowledge of the defence pipeline could still pose challenges
[10]	2021	physical adversarial attacks targeting Lidar-based perception modules	KITTI [31]	70% to 85%	adversarial point cloud generation techniques	PointRCNN, PointPillar, and PV-RCNN as target models	The success rate of attacks decreased in outdoor environments due to factors like uneven terrain and sparse Lidar data at low angles.

[11]	2023	optimized Carlini-Wagner (C&W) adversarial sample attack	RainDS dataset [33]	attack reduced PSNR by 39.5% and SSIM by 26.4%	optimized C&W attack	RainCCN and MPRNet	challenges in obtaining ideal rain-free ground truth images for evaluation.
[12]	2024	Invisible Optical Adversarial Stripes	German Traffic Sign Recognition Benchmark (GTSRB) [30]	94% success rate for untargeted attacks and 97% success rate for targeted attacks	GhostStripe1 and GhostStripe2	YOLO object detector	The attack is designed for specific victim vehicles and may not generalize to multiple vehicles simultaneously.
[13]	2024	physical attacks on multi-sensor fusion systems	KITTI dataset [31]	Up to 90%	optimizing the parameters of adversarial objects	Velodyne VLP-32C LiDAR, TI AWR1843 radar, and Allied Vision Mako G-319 camera.	The attack requires precise placement of adversarial objects, which may be challenging in dynamic or uncontrolled environments.
[14]	2022	Physical backdoor attacks targeting lane detection systems	TuSimple Challenge dataset [34]	rotation angles for backdoored models on poisoned images ranged significantly, with SCNN showing an average of 23.1°	Poison-annotation and clean-annotation	Image scaling functions, Lane detection models like SCNN, LaneATT, UltraFast, and PolyLaneNet	it requires generating poisoned samples for each scaling function, which increases complexity
[15]	2024	Physical sensor manipulation attacks targeting RAVs	historical sensor data	SpecGuard achieved a 92.1% recovery success and achieved 90% mission specification compliance in real RAVs.	SpecGuard, a specification-aware recovery method	Deep-RL frameworks, Signal Temporal Logic (STL) for mission specification compliance, Proximal Policy Optimization (PPO) algorithm for training	Recovery fails in ~8% of cases when the RAV does not reach the target within the allowable 5-meter position error margin and restricted to single-sensor attacks
[16]	2021	Adversarial attacks targeting LiDAR semantic segmentation: Vehicle/Obstacle Hiding Attack and Road Surface Changing Attack	SemanticKITTI dataset [31]	a success rate of over 90% in real-world settings	Adversarial attack framework utilizing Semantic Misleading	Point cloud segmentation models like PointNet, SqueezeSeg, Cylinder3D, PointNet++, and PointASNL	Effectiveness decreases as victim AV approaches objects, especially at distances below 10 meters
[17]	2023	Passive-reflection-based attacks on mmWave radar object detection in autonomous vehicles, named TileMask	SemanticKITTI dataset [31]	Attack Success Rate (ASR) reached 93% in physical-world experiments and 92% in digital-world simulations	TileMask—adversarial objects designed to manipulate radar echoes and disrupt deep neural network-based radar object detection systems	TI AWR1843 radar and DNN-based detection models	Effectiveness decreases when significant deviation occurs in object placement
[18]	2023	Evasion attacks targeting Road Context-aware Intrusion Detection Systems (RAIDS)	Udacity Self-Driving Car dataset, Chen 2017 dataset, and Chen 2018 dataset [36]	93.9% accuracy in detecting such attacks	PAID integrates feature squeezing and GPS coordinates with RAIDS to combat perturbed image attacks effectively	GPS data for contextual validation, PyTorch framework, and Oracle Cloud Infrastructure	PAID's dependency on accurate GPS data assumes legitimate coordinates unaffected by adversaries
[19]	2025	White-box attack	Argoverse, NuScenes	87%+ attack success rate,	Optimization-based attack	Frenet coordinate system, curvature	No real-world testing; focuses on

			[35]	induced AV hard braking in simulations	with Frenet-frame objective	constraints, Adam optimizer	vehicles, not pedestrians
[20]	2023	Cyberattacks on Controller Area Network (CAN) systems	Custom dataset from AV-CPS simulation (80,000 samples: 40K normal, 40K attack)	GoogLeNet achieved 99.47% F1-score; other models (AlexNet, ResNet-50) scored >99% accuracy	Transfer Learning with pre-trained CNNs (e.g., GoogLeNet, ResNet-50)	MATLAB, Simulink, CAN toolbox, GPU-accelerated CNNs	Limited to CAN-based attacks - No real-world validation - Focuses only on sensor data anomalies
[21]	2024	White-box and Black-box attacks	ApolloScape (urban roads) [26], NGSIM (highway) [27]	Prediction errors increased by over 50% ADE increased by 694%, FDE increased by 513% (ApolloScape)	Adv-GAN (Generative Adversarial Networks)	GANs, LSTM, MPC, GRIP++	No robust defense mechanisms proposed
[22]	2024	Targeted attacks (optimistic assumption, function-space reformulation)	Atari (Pong, BeamRider, Qbert, SpaceInvaders), MuJoCo (Ant, Hopper, Walker2d, HalfCheetah)	DQN: 6.6 ± 7.3 (BeamRider), A2C: 26.6 ± 34.4 (SpaceInvaders), MuJoCo (PPO): -841.01 ± 494.88 (Ant). Achieves near-lowest reward in 6/8 environments.	Two-stage optimization (KL-divergence minimization between victim and deceptive policies)	FGSM, PGD (10 iterations)	Requires pre-training deceptive policies; scalability to ultra-high-dimensional states (e.g., >6000 dimensions in Atari) needs further study.
[23]	2025	Black-box suppressive adversarial attack	Microsoft COCO, Waymo Open Dataset [38], VisDrone [37]	- COCO (RetinaNet): 37,089.5 pixels changed (32.34% of bounding box pixels) - Waymo: 36,951.07 pixels changed (64.73%) - VisDrone: 8,900.53 pixels changed (59.33%)	EvoAttack (Evolutionary search with dynamic mutation)	Genetic Algorithm (GA), RetinaNet, Faster R-CNN, YOLOv5	Requires hyperparameter tuning; computation time scales with image complexity
[24]	2024	Poisoning (Trojan/White Box), Evasion (Black Box), Cyber Attacks	Russian Traffic Sign Dataset (RTSD) [39]	F1-scores: Clean model (0.999), Label-swapped (0.575), Gaussian noise (0.992), Backdoor class (0.997)	EdgeNeXt (Hybrid CNN-Transformer)	Python, TensorFlow/PyTorch	Class imbalance in dataset; limited to Russian traffic signs; no real-world vehicle tests.
[25]	2024	Untargeted white-box attacks (FGSM, MIM, PGD, C&W)	Modified LISA traffic sign dataset (18 US traffic sign classes) [29]	No attack: 99% accuracy; FGSM: 89% accuracy; MIM: 87% accuracy; PGD: 87% accuracy; C&W: 89% accuracy. Outperformed traditional methods by 6–55%.	Hybrid defense (random filtering + ensembling + local feature mapping)	Inception-V3, ResNet-152, Keras OCR (CRAFT + CRNN), PyTorch, NVIDIA Tesla P100 GPUs	Limited to US traffic signs; misclassifications between similar signs (e.g., 55 vs. 65 mph); untested on non-text signs.

Table 1. presents an overview of the analyzed studies, and the descriptions of each column are outlined as follows:

- Column 1: Year
  This column indicates the publication year of the research.
- Column 2: Type of Attack

This column describes the specific category of adversarial attack used.

- Column 3: Dataset used
  This column refers to the collection of data utilized for testing the effec-tiveness of the model/technique/approach used.
- Column 4: Result
  This column summarizes the key findings of the research, including the effectiveness of the attacks and performances of the technique.
- Column 5: Technique used
  This column outlines the algorithms implemented to execute the models.
- Column 6: Limitations
  This column highlights the challenges faced in the research.

IV. Results and Analysis

Research on adversarial attacks on autonomous vehicles has seen a significant growth, as shown in Fig 1. depicting the number of published papers by year in this diverse field. This trend shows the increasing awareness of helplessness in autonomous systems with the most papers published in 2024 reflecting advance-ments in technology and increased security concerns. Whereas, in Fig 2. the pie chart reveals a diverse distribution of datasets used in these studies, which shows that some datasets dominate the research while others are emerging.

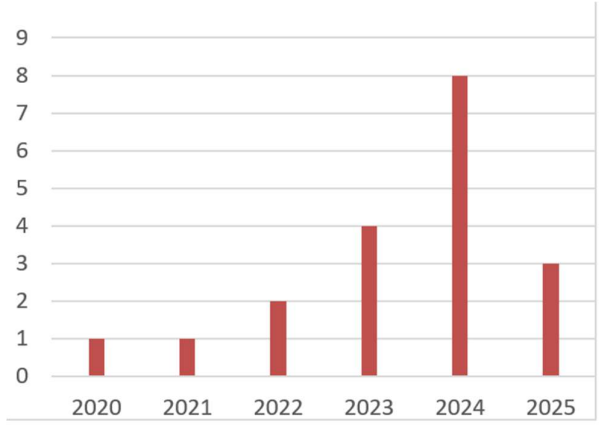

Fig 1. Number of Papers by Year

In addition to the observed trends, the researches feature a diverse range of methodologies which are aimed at overcoming the adversarial attacks. Many studies focus on creating a strong defense mechanism, such as training models and techniques to improve the resilience of autonomous systems against threats.

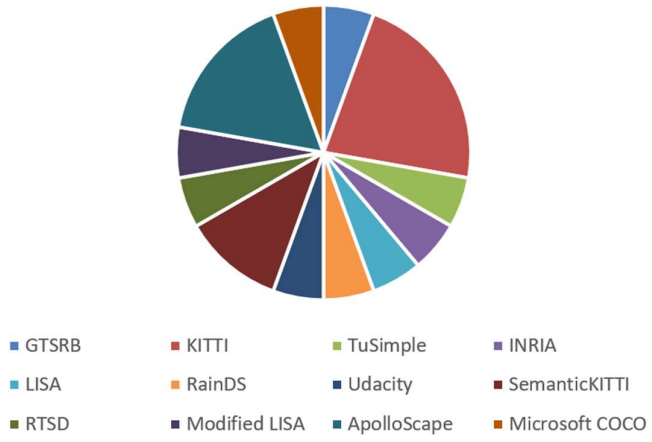

Fig 2. Distributions of Papers by Dataset Used

In Fig 3., the doughnut chart shows that white-box and black-box attacks are the most prevalent, reflecting their relevance in evaluating model robustness, follow-ing them are Physical patch and evasion attacks, indicating increased focus on real world threat scenarios.

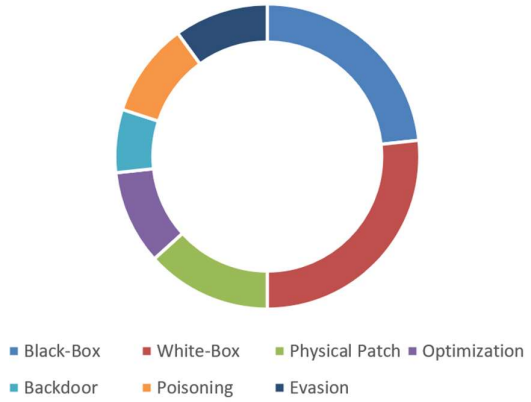

Fig 3. Most Used Attack Types

While in Fig 4., the pie chart reveals that Generative Adversarial Networks (GANs), gradient based attacks like FGSM and PGD and optimization-based methods dominate the adversarial research. Whereas as seen in Fig 5., YOLO models and PyTorch emerged as the most commonly used tools, showing their widespread adoption in perception and adversarial research. Frameworks like GANs, TensorFlow and ResNet also appear frequently.

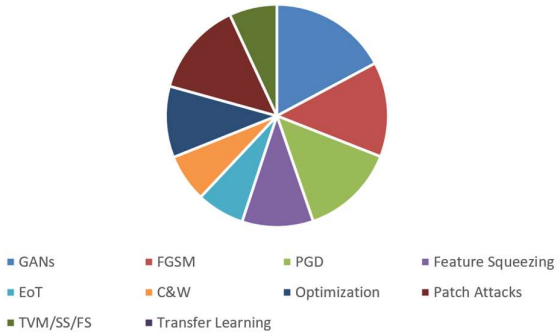

Fig 4. Most Used Techniques

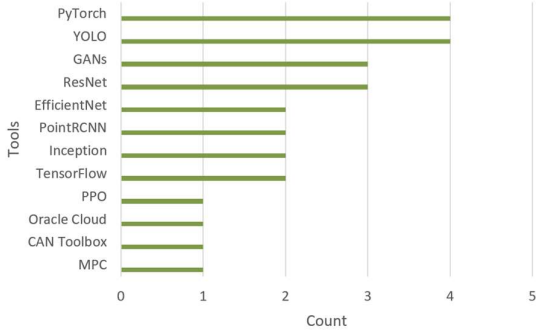

Fig 5. Most Used Tools

Overall, these visualizations provide insights into the growth of research in ad-versarial defense mechanism in autonomous vehicles. These findings have im-portant suggestions; they emphasize the urgent need to address the shortcomings in autonomous vehicles. As our world advances with technology, continued re-search on this topic will be crucial to ensure the safety and reliability of AVs, helping to build public trust and accepting these innovations.

V. Conclusion

In this study, we reviewed 23 research papers, published from 2020-2025, which focused on adversarial attacks on autonomous vehicles (AVs) to arrange the findings regarding the attacks, tools, defense techniques and limitations in the current studies. The primary goal was to understand the nature of these attacks which can crucially compromise the safety of the AV systems as we adopt them in the transportation uses. By analyzing the methodologies, datasets, the type of attacks and results from these studies, we identified the various types of attacks that people are generally not familiar with. Our review also highlights the im-portance of use of diverse datasets for real world simulations and also pointed out the research gaps. Lastly, this comprehensive analysis works as a foundation-al resource for researchers and readers, highlighting the importance of the adver-sarial threats to ensure the continuous safety and people's trust in autonomous vehicles as our world and technology evolves.

REFERENCES

[1] Boloor, A., He, X., Gill, C., Vorobeychik, Y., & Zhang, X. (2019). Simple physical ad-versarial examples against end-to-end autonomous driving models. In Proceedings of the IEEE Conference on Embedded Systems (ICESS). https://doi.org/10.1109/ICESS.2019.8782514

[2] Sitawarin, C., Bhagoji, A., Mosenia, A., Mittal, P., & Chiang, M. (2018). Rogue signs: Deceiving traffic sign recognition with malicious ads and logos. arXiv preprint arXiv:1801.02780.

[3] Fan, J., Wang, Z., & Li, G. (2024). Adversarial Attack on Trajectory Prediction for Au-tonomous Vehicles with Generative Adversarial Networks. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS), pp. 1026–1031. https://doi.org/10.1109/IROS58592.2024.10802234

[4] Dipalma, C., Wang, N., Sato, T., & Chen, Q.A. (2021). Demo: Security of Camera-based Perception for Autonomous Driving under Adversarial Attack. In: Proceedings of the 2021 IEEE Symposium on Security and Privacy Workshops (SPW), pp. 243. https://doi.org/10.1109/SPW53761.2021.00040

[5] Duan, A., Wang, R., Cui, Y., He, P., & Chen, L. (2023). Causal Robust Trajectory Pre-diction Against Adversarial Attacks for Autonomous Vehicles. IEEE Internet of Things Journal, 11(22), 35762– 35776. https://doi.org/10.1109 /JIOT.2023.3342788

[6] Azim, O. A., Baker, L., Majumder, R., Enan, A., Khan, S. M., & Chowdhury, M. A. (2023). Data-Driven Defences Against Adversarial Attacks for Autonomous Vehicles. IAVVC 2023 - IEEE International Automated Vehicle Validation Conference, Proceed-ings, Dl, 1–5. https://doi.org/10.1109 /IAVVC57316.2023.103 28098

[7] Kumar, K. N., Vishnu, C., Mitra, R., & Mohan, C. K. (2020). Black-box adversarial at-tacks in autonomous vehicle technology. Proceedings - Applied Imagery Pattern Recognition Workshop, 2020-October. https://doi.org/10.1109 /AIPR50011.2020.9425 267

[8] Wang, H., Zhang, J., Wang, Y., & Wang, D. (n.d.). Secure Object Detection of Auton-omous Vehicles Against Adversarial Attacks. IECON 2024 - 50th Annual Conference of the IEEE Industrial Electronics Society, 1–6. https://doi.org/10.1109 /IECON55916.2024.109 05242

[9] Kim, T., Yu, Y., & Ro, Y. M. (2022). Defending Physical Adversarial Attack on Object Detection via Adversarial PatchFeature Energy. MM 2022 - Proceedings of the 30th ACM International Conference on Multimedia, 1905– 1913. https://doi.org/10.1145 /3503161.3548362

[10] Yang, K., Tsai, T., Yu, H., Panoff, M., Ho, T. Y., & Jin, Y. (2021). Robust Roadside Physical Adversarial Attack against Deep Learning in Lidar Perception Modules. ASIA CCS 2021 - Proceedings of the 2021 ACM Asia Conference on Computer and Com-munications Security, Iii, 349–362. https://doi.org/10.1145 /3433210.3453106

[11] Qin, Y., Hu, J., & Wu, B. (2023). Toward Evaluating the Robustness of Deep Learning Based Rain Removal Algorithm in Autonomous Driving. ACM International Confer-ence Proceeding Series. https://doi.org/10.1145 /3591197.3591309

[12] Guo, D., Wu, Y., Dai, Y., Zhou, P., Lou, X., & Tan, R. (2024). Invisible Optical Adver-sarial Stripes on Traffic Sign against Autonomous Vehicles. MOBISYS 2024 - Proceed-ings of the 2024 22nd Annual International Conference on Mobile Systems, Applica-tions and Services, 534–546. https://doi.org/10.1145 /3643832.3661854

[13] Zhu, Y., Miao, C., Xue, H., Yu, Y., Su, L., & Qiao, C. (2024). Malicious Attacks against MultiSensor Fusion in Autonomous Driving. 436–451. https://doi.org/10.1145 /3636534.3649372

[14] Han, X., Xu, G., Zhou, Y., Yang, X., Li, J., & Zhang, T. (2022). Physical Backdoor At-tacks to Lane Detection Systems in Autonomous Driving. MM 2022 - Proceedings of the 30th ACM International Conference on Multimedia, 2957–2968. https://doi.org/10.1145 /3503161.3548171

[15] Dash, P., Chan, E., & Pattabiraman, K. (2024). SpecGuard: Specification Aware Re-covery for Robotic Autonomous Vehicles from Physical Attacks. 1849–1863. https://doi.org/10.1145 /3658644.3690210

[16] Zhu, Y., Miao, C., Hajiaghajani, F., Huai, M., Su, L., & Qiao, C. (2021). Adversarial Attacks against LiDAR Semantic Segmentation in Autonomous Driving. SenSys 2021 - Proceedings of the 2021 19th ACM Conference on Embedded Networked Sensor Sys-tems, 329–342. https://doi.org/10.1145 /3485730.3485935

[17] Zhu, Y., Li, Z., Miao, C., Yu, Y., Xue, H., Xu, W., Su, L., & Qiao, C. (2023). TileMask: A PassiveReflection-based Attack against mmWave Radar Object Detection in Auton-omous Driving. CCS 2023 - Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, 1317–1331. https://doi.org/10.1145 /3576915.3616661

[18] Teng, K. Z., Limbasiya, T., Turrin, F., Aung, Y. L., Chattopadhyay, S., Zhou, J., & Conti, M. (2023). PAID: Perturbed Image Attacks Analysis and Intrusion Detection Mechanism for Autonomous Driving Systems. CPSS 2023 - Proceedings of the 9th ACM ASIA Conference on Cyber-Physical System Security Workshop, 3–13. https://doi.org/10.1145 /3592538.3594273

[19] Dong, Y., Wang, L., Li, Z., et al. (2025). Safe Driving Adversarial Trajectory Can Mis-lead: Toward More Stealthy Adversarial Attack Against Autonomous Driving Prediction Module. ACM Trans. Priv. Sec. 28(2), Article 19. https://doi.org/10.1145 /3705611

[20] Alsulami, A. A., Abu AlHaija, Q., Alturk, B., Alqahtani, A., & Alsini, R. (2023). Secu-rity strategy for autonomous vehicle cyber-physical systems using transfer learning. Journal of Cloud Computing, 12(1), 181. https://doi.org/10.1186 /s13677-023-00564-x

[21] Fan, J., Wang, Z., & Li, G. (2024). Adversarial Attack on Trajectory Prediction for Au-tonomous Vehicles with Generalized Adversarial Networks. IEEE International Confer-ence on Intelligent Robots and Systems, 1026–1031. https://doi.org/10.1109 /IROS58592.2024.1080 2234
[22] Qiaoben, Y., Ying, C., Zhou, X., Su, H., Zhu, J., & Zhang, B. (2024). Understanding adversarial attacks on observations in deep reinforcement learning. Science China In-formation Sciences, 67(5). https://doi.org/10.1007/s11432-021-3688-y
[23] Chan, K. H., & Cheng, B. H. C. (2025). EvoAttack: Suppressive adversarial attacks against object detection models using evolutionary search. Automated Software Engi-neering, 32(3). https://doi.org/10.1007 /s10515-024-00470-9
[24] Boltachev, E. (2024). Potential cyber threats of adversarial attacks on autonomous driv-ing models. Journal of Computer Virology and Hacking Techniques, 20, 363–373. https://doi.org/10.1007 /s11416-023-00486-x
[25] Khan, Z., Chowdhury, M., & Khan, S.M. (2024). A Hybrid Defense Method against Adversarial Attacks on Traffic Sign Classifiers in Autonomous Vehicles. TechRxiv. https://doi.org/10.36227/techrxiv.19071824.v1
[26] https://apolloscape.auto/trajectory.html
[27] https://catalog.data.gov/dataset/next-generation-simulation-ngsim-vehicle-trajectories-and-supporting-data
[28] https://github.com/lgsvl/simulator
[29] https://git-disl.github.io/GTDLBench/datasets/lisa_traffic_sign_dataset/
[30] https://benchmark.ini.rub.de/gtsrb_dataset.html
[31] https://www.cvlibs.net/datasets/kitti/eval_mots.php
[32] https://www.kaggle.com/datasets/jcoral02/inriaperson
[33] https://paperswithcode.com/dataset/rainds
[34] https://www.kaggle.com/datasets/manideep1108/tusimple
[35] https://www.nuscenes.org/
[36] https://www.kaggle.com/datasets/sshikamaru/udacity-self-driving-car-dataset
[37] https://github.com/VisDrone/VisDrone-Dataset
[38] https://waymo.com/open/
[39] https://www.kaggle.com/datasets/watchman/rtsd-dataset
[40] https://viso.ai/deep-learning/adversarial-machine-learning/
[41] T. Myklebust, T. Stålhane, G. D. Jenssen and I. Wærø, "Autonomous Cars, Trust and Safety Case for the Public," 2020 Annual Reliability and Maintainability Symposium (RAMS), Palm Springs, CA, USA, 2020, pp. 1-6, doi: 10.1109/RAMS48030.2020.9153618.