Proc. of the International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME 2023)
19-20 July 2023, Tenerife, Canary Islands, Spain

Adversarial Attacks on Traffic Sign Recognition: A Survey

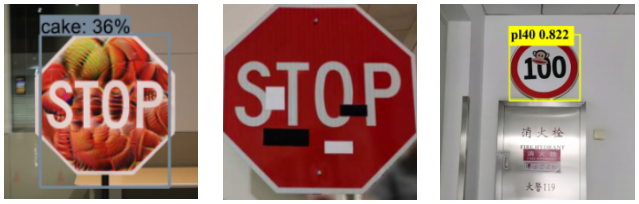
Svetlana Pavlitska
FZI Research Center for Information Technology
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
pavlitska@fzi.de

Nico Lambing
FZI Research Center for Information Technology
Karlsruhe, Germany
lambing@fzi.de

J. Marius Zöllner
FZI Research Center for Information Technology
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
zoellner@fzi.de

***Abstract*—Traffic sign recognition is an essential component of perception in autonomous vehicles, which is currently performed almost exclusively with deep neural networks (DNNs). However, DNNs are known to be vulnerable to adversarial attacks. Several previous works have demonstrated the feasibility of adversarial attacks on traffic sign recognition models. Traffic signs are particularly promising for adversarial attack research due to the ease of performing real-world attacks using printed signs or stickers. In this work, we survey existing works performing either digital or real-world attacks on traffic sign detection and classification models. We provide an overview of the latest advancements and highlight the existing research areas that require further investigation.**

***Index Terms*—traffic sign recognition, image classification, object detection, adversarial attacks**



a) Attack on the **generic object detection**: a stop sign misclassified by the Faster R-CNN trained on the COCO dataset [4].

b) Attack on the **traffic sign classification**: a stop sign misclassified as a speed limit 45km/h sign by LISA-CNN [5].

c) Attack on the **traffic sign detection and classification**: a 100km/h sign is misclassified as a 40km/h sign [6].

Fig. 1: Three subtasks within TSR and exemplary attacks.

I. Introduction

Deep neural networks (DNNs) are inherently susceptible to adversarial attacks: small changes in the input can cause wrong model predictions [1], [2]. To obstruct the behavior of a DNN for computer vision tasks, an attacker can perform either small imperceptible pixel changes over the whole image, or restrict visible, perceptible adversarial noise to a small image area, thus generating an *adversarial patch* [3]. Real-world attacks can easily be performed by adding a printed patch to a scene perceived by a camera.

Correct traffic sign recognition is vital for perception in automated vehicles. Unlike many other perception tasks, traffic sign recognition can be performed only using input from camera images, and thus errors cannot be compensated by other sensors like LiDAR. Adversarial attacks on computer vision models thus pose a special threat to this perception task. On the other hand, attacks against traffic sign recognition are especially favorable regarding real-world evaluation since perturbed traffic signs can easily be printed to replace the real ones. Therefore, a particular research effort has been put into designing attacks that would be robust under various physical conditions.

Traffic sign recognition (TSR) is a multi-class classification problem with unbalanced class frequencies. TSR is split into two tasks: *traffic sign detection (TSD)* aims at traffic sign localization in the input image, while the goal of the next step, *traffic sign classification (TSC)* is to predict class labels for the detections (see Figure 1). Traffic sign recognition is performed predominantly with DNNs, mostly convolutional neural networks (CNNs) [7], and recently also spatial transformers [8]. Successful adversarial attacks were demonstrated both for traffic sign classification [5] and detection [6] models, while attacks took different forms: an attacker can perturb the whole area of a traffic sign [9], generate stickers [5] or shadows [10] to be placed into it.

This survey provides a comprehensive overview of the existing works on adversarial attacks on TSR models. We compare models, datasets, and attack approaches. Furthermore, we identify promising research directions. To the best of our knowledge, we are the first to provide such an overview.

TABLE I: Overview of traffic sign recognition datasets

Dataset	Year	Ref	TSD	TSC	Images annotated	Classes	Resolution	Country
GTSRB	2011	[11]		✓	50K+	43	15×15 to 250×250	Germany
STS	2011	[12]	✓	✓	3K	20	1280×960	Sweden
GTSDB	2013	[13]	✓	✓	900	3	1360×800	Germany
BelgiumTS	2011	[14]		✓	9K	62	11×10 to 562×438	Belgium
LISA	2011	[15]	✓	✓	6K	47	640×480 to 1024×522	USA
TT100K	2016	[16]	✓	✓	100K	221	2048×2048	China
MTSD	2017	[17]	✓	✓	100K	313	920×1080	Worldwide

II. Background

A. Adversarial attacks

DNN predictions can be manipulated via small pixel changes in input. An adversarial attack can be performed either in a *white-box* manner, where an attacker has full access to model architecture, weights, and data or in the *black-box* setting, without knowledge of the target model.

One of the first and the simplest white-box adversarial attack techniques is the *fast gradient sign method* (FGSM) [2]. To obtain a perturbed image using FGSM, the sign of the gradient of the loss function with respect to the input is added to the input and multiplied by a small weighting factor. Later, iterative versions of FGSM were introduced, including the *projected gradient descent* (PGD) approach [18]. Other influential attack algorithms include *Jacobian-based saliency maps attack* (JSMA) [19], *Papernot's attack* [20], and *Carlini and Wagner* (C&W) attack [21].

In a standard setting, the attack is performed in a per-instance manner, i.e., one adversarial noise pattern is generated for each input image. However, the universal setting can be applied to make attacks more realistic, where one pattern is generated to attack the whole dataset [22].

A further step towards realistic attacks is an *adversarial patch*, where the adversarial perturbation is applied not to all input image pixels but within a specified image region [3]. To perform an attack in the real world, an adversarial patch should be trained in a universal manner. Universal adversarial patch attacks have been successfully applied to various computer vision tasks, including object detection [23]–[25], semantic segmentation [26], and steering angle prediction [27].

Several techniques have been proposed to achieve the robustness of an adversarial patch attack under small spatial changes, camera views, and settings. For example, in the *expectation over transformations* (EoT) approach by Athalye et al. [28], a transformation function sampled from a predefined set is applied during training, whereas possible transformations include translation, rotation, scaling, saturation, hue, and brightness changes. Eykholt et al. [5] argued that the EoT approach is too constrained to model physical phenomena. Instead, they proposed the *robust physical perturbations (RP2)* method. Additionally to sampling a synthetic transformation like rotation or brightness change, training data with actual physical variability is used. This includes images taken at different distances, under different camera angles, and lighting conditions.

B. Traffic sign datasets

Traffic signs from different countries exhibit large variations, therefore a plethora of datasets for TSR has been developed (see Table I). Due to its age and split into traffic sign detection and classification, the German Traffic Sign Recognition and Detection Benchmark (GTSRB and GTSRD) [11], [13] is usually used to evaluate TSR algorithms. More recent approaches are often evaluated on the TsinghuaTencent 100K (TT100K) dataset [16] since TSR and TSD can be combined. A profound comparison of available datasets can be found in [29] and [30].

C. Traffic sign detection and recognition

Since traffic signs are characterized by specific forms and colors, early works relied predominantly on classic computer vision and image processing approaches, like SIFT [31], HOG [32] or Hough transform [33]. Later, CNNs were first proposed to replace certain steps in the pipeline (e.g., feature extraction and classification) and quickly became the dominating approach in this area.

Sermanet et al. [7] were the first to surpass human results on the GTSRB by applying a multi-scale CNN (*MS-CNN*) without the usage of any handcrafted features. The publicly available implementation of this architecture by Yadav[1] was often used for benchmarking in later works. Ciresçan et al. [34] utilized a committee of a CNN and a multilayer perceptron with histograms of oriented gradients as input for TSR. A similar approach was taken by Jin et al. [35] who used an ensemble of 20 hinge loss-trained CNNs.

To go beyond the image classification task and detect traffic signs in an input image, several works tried to adapt existing generic object detectors to the TSR task. For instance, Doval et al. [36] utilized YOLOv3 [37] for TSR on stereo camera data. Rehman et al. [38] further optimized the architecture and training of YOLOv3 for TSR by utilizing layer pruning and a density-based anchor box selection algorithm. Another single-shot approach was suggested by Tian et al. [39], who integrated a recurrent attention mechanism into the deconvolutional single shot detector [40] for improved usage of local context information.

Finally, Garcia et al. combined alternating convolutional and spatial transformer modules in an architecture, further denoted as *CNN-ST*, outperforming state-of-the-art models on the GTSRB dataset [8].

[1] https://github.com/vxy10/p2-TrafficSigns

TABLE II: Overview of adversarial attacks on traffic sign recognition models

Author	Year	Ref	Baseline	Dataset	Attack method	Attack appearance	TSD	TSC	White-box	Black-box	Tar-geted	Real-world	Code available
Eykholt et al.	2017	[5]	LISA-CNN, GTSRB-CNN (MS-CNN [7])	LISA, GTSRB	RP2	Black and white stickers		✓	✓			✓	✓
Lu et al.	2017	[9]	Faster R-CNN, YOLOv2	COCO, private	Iterative FGSM	Perturbed sign	✓*		✓			✓	
Song et al.	2018	[41]	Faster-RCNN, YOLOv2	COCO	RP2, modified for the detection task	Stickers or perturbed sign	✓*		✓			✓	
Chen et al.	2018	[4]	Faster R-CNN	COCO	C&W with EoT	Perturbed sign	✓*		✓		✓	✓	✓
Papernot et al.	2017	[19]	Own CNN	GTSRB	Substitute model, FGSM, Papernot's attack	Perturbed sign		✓		✓		✓	
Sitawarin et al.	2018	[42]	LISA-CNN, GTSRB-CNNN	LISA, GTSRB	C&W attack with EoT	Logo, custom sign		✓	✓		✓	✓	✓
Sitawarin et al.	2018	[43]	TSD: Hough transform TSC: MS-CNN [7]	GTSRB, GTSDB, private	Enhancement of [42], lenticular printing attack	Same as [42], signs looking differently from various angles	✓	✓	✓	✓	✓	✓	✓
Liu et al.	2019	[44]	VGG16, ResNet-34, MS-CNN [7]	GTSRB, ImageNet	PS-GAN	Scrawl-like stickers		✓	✓	✓		✓	
Morgulis et al.	2019	[45]	MS-CNN [7], DenseNet	GTSRB	Enhancement of [43], improved augmentation	Gray shadows on speed limit signs		✓	✓	✓	✓	✓	
Li et al.	2021	[46]	CNN-ST [8]	GTSRB	Adaptive square attack, SimBA-DCT [47], square attack [48]	Perturbed image		✓		✓	✓		
Yang et al.	2021	[49]	3-layer CNN	LISA, GTSRB	Targeted attention attack	Grayscale noises		✓	✓			✓	✓
Woitschek et al.	2021	[50]	CNN-ST [8]	GTSRB	FGSM, PGD, RP2, SPSA, model stealing	Perturbed sign		✓		✓	✓		
Jia et al.	2021	[51]	YOLOv5	TT100K	Cross-domain conversion	Perturbed sign	✓	✓		✓	✓	✓	
Ye et al.	2021	[52]	VGG16, ResNet-34, GoogLeNet, ensemble	GTSRB	PGD	Squared patch in an image		✓	✓			✓	
Zolfi et al.	2021	[53]	YOLOv2, YOLOv5, Faster R-CNN	LISA, MTSD, BDD	Custom gradient-based optimization	Translucent patch on camera lens	✓*	✓	✓		✓	✓	
Zhong et al.	2022	[10]	LISA-CNN, GTSRB-CNN	LISA, GTSRB	EoT, PSO	Shadows		✓		✓	✓	✓	✓
Chi et al.	2023	[54]	MCDNN [55], CNN-ST [8]	GTSRB	Public-attention attack	Perturbed sign		✓		✓			
Liu et al.	2023	[56]	GAN	GTSRB, TT100K	GAN to generate raindrops	Raindrops	✓	✓		✓			
Wei et al.	2023	[6]	YOLOv1	TT100K	Region-based heuristic differential evolution	Perturbations of existing stickers	✓	✓		✓		✓	✓

*within generic object detection

RP2: Eykholt et al. [5] | Lu et al. [9] | Shapeshifter: Chen et al. [4] | Rogue, DARTS: Sitawarin et al. [42], [43] | PS-GAN [44] | Morgulis et al. [45] | TAA: Yang et al. [49]

Fig. 2: Examples of traffic signs modified with different attacks

III. Overview of adversarial attacks on TSR

In this section, we survey the existing adversarial attacks. Table II provides a summarized overview, and Figure 2 demonstrated examples of traffic signs modified with selected attacks.

A. Early attack on TSC

The first works on adversarial attacks against TSR appeared in 2017 [5], [9], [19]. The seminal work by Eykholt et al. [5] focused on real-world attacks against TSC models. For this, the *robust physical perturbations (RP2)* approach was proposed, where a perturbation is sampled from a distribution of physically possible perturbations with the goal of maximizing the classification error. The attack was designed to resemble graffiti, usually observed on traffic signs. For this, a mask to project generated adversarial perturbations was applied to the input image. The resulting manipulation consists of a set of black and white stickers printed out and attached to the traffic sign (see Figure 2). To account for a spectrum of printable colors, the *non-printability score (NPS)* was added as a separate loss term as proposed by Sharif et al. [57].

Furthermore, Eykholt et al. proposed a two-stage evaluation methodology for real-world attacks, including (1) stationary lab experiments and (2) field evaluation using drive-by scenarios. The authors also defined LISA-CNN and GTSRB-CNN models, which later became standard for further attack studies. LISA-CNN consists of three convolutional layers followed by a fully connected layer trained to classify images from the LISA dataset. GTSRB-CNN is based on the MS-CNN [7] implementation by Yadav, mentioned above. Models and the attack algorithm were made publicly available by the authors[2].

B. Attacks on the stop sign class of generic object detection

While the original work of Eykholt et al. focused on classifiers only, subsequent work from the same research group by Song et al. [41] also looked into detection. For this, a generic object detector YOLOv2 [58] was attacked by putting stickers on a stop sign, causing mislabeling or not detecting the traffic sign. The evaluation, however, was performed on a class *stop sign* within generic object detection. Trained on the COCO dataset, YOLO can only detect stop signs. Furthermore, Song et al. expanded in this work RP2 from the classification to detection. For this, the adversarial loss was modified to cause the disappearance of the stop sign or detection of non-existent objects. Next, the set of synthetic physical constraints

[2]https://github.com/evtimovi/robust_physical_perturbations

for RP2 was modified to include object rotation and position. And finally, the total variation [59] was incorporated into the loss function to ensure smooth transitions between neighboring pixels. The resulting perturbation was applied to the whole stop sign area to cause object disappearance or a non-existent object creation attack. The experiments were also performed in the real world but only in stationary lab settings.

The poster attacks on stop signs performed by Eykholt et al. were critically addressed by Lu et al. [60]. The generic detectors YOLO and Faster-RCNN were shown to be able to successfully withstand attacks proposed so far, so the authors claimed that *"standard detectors aren't (currently) fooled by physical adversarial stop signs"*. In their subsequent work, however, Lu et al. were able to successfully fool a detector themselves [9]. Although the method was designed for generic object detectors, it was exemplarily evaluated on the *stop sign* class. A method similar to iterative FGSM was used to generate a perturbation for the stop sign while variations in illumination intensity were incorporated. As a result of applying the generated perturbation to the stop sign, it was correctly localized in an image but classified as an object of some other class, e.g., a *vase*. The authors attacked Faster R-CNN [61] and also demonstrated that the generated perturbations transfer without modifications to YOLO2. For the real-world attacks, the authors collected their own dataset consisting of videos of stop signs from an ego vehicle perspective. To successfully fool a model in the real world, however, large visible perturbations in the stop sign area were needed.

The *Shapeshifter* attack by Chen et al. [4] applied EOT by Athalye et al. [28] instead of RP2 described above to ensure patch robustness in real-world settings. This is also the first work to focus on targeted attacks on stop sign detection within the generic object detection framework. In particular, classes *person* and *sport ball* were selected as targets due to their similarity to stop signs in shape and size.

In summary, attacking the *stop sign* class prediction of a model trained for generic object detection was addressed in multiple early works [4], [9], [41], [60] and was extensively evaluated in real-world settings. It has paved the path for research on TSC and TSD attacks but has become less popular after that. A recent work by Zolfi et al. [53] addressed generic object detection again and proposed a translucent patch placed directly on a camera lens. This sticker suppressed the detection of the objects of the *stop sign* class.

C. Black-box attacks on TSR

While most studies included the experiments with the transferability of the proposed attacks to further models, Papernot et al. [19] were the first to apply black-box attacks to the TSC task on GTSRB. The attacker in this setting is only able to observe the predicted labels for given inputs. The authors proposed to train a local substitute model for the attacked DNN by using its observed predictions as training data. The attack generated for the substitute DNN was shown to be able to transfer successfully to the target DNN.

More challenging black-box attacks were the focus of recent works. Woitschek et al. studied several black-box attacks in a real-world attack setting [50]. The publicly available implementation[3] of the CNN-ST [8] was used for experiments. The attack algorithms included gradient approximation using *simultaneous perturbation stochastic approximation* (SPSA) [62] as well as *model stealing*. To evaluate attack feasibility, 1000 different transformations were applied. However, no evaluation in the real world was performed.

A further score-based black-box attack method is the *adaptive square attack* (ASA) proposed by Li et al. in [46]. It builds upon the *square attacks* method by Andriushchenko et al. [48], which uses random search, whereas, in each perturbation, square-shaped updates at random positions are sampled.

Zhong et al. [10] proposed to construct adversarial shadows by querying the target model. The attention-based attack was described by Chi et al. [54]. Wei et al. [6] demonstrated a method to generate adversarial stickers against TSD in a black-box manner.

D. Innocuous-looking traffic signs

Instead of modifying existing traffic signs, Sitawarin et al. [42] proposed a method to modify innocuous signs and advertisements so that they are classified as targeted traffic signs. In this work, a traffic sign is cut out of an input image using masking and resized to reach the input size of the model. After that, a variant of the C&W attack combined with EoT is used to generate the perturbation. An attacker can either modify an existing logo or advertisement sign or start with a blank sign to generate a custom sign. In the follow-up work [43], the authors further enhance the proposed pipeline named *DARTS* and describe the lenticular printing attack based on optical phenomena, s.t. an adversarial sign appears different under different observation angles.

Morgulis et al. [45] further extended the DARTS pipeline with improved random augmentation techniques so that a batch of new random transformations is applied to each iteration. This way, perturbations for speed limit traffic signs could be created. Printing-size adversarial signs were evaluated in the black-box mode with a TV-in-the-loop method and a drive-by experiment.

E. More realistic-looking attacks

To further enhance the inconspicuousness of the generated stickers, Liu et al. [44] used a generative adversarial network (GAN) to produce more natural-looking stickers. An attention mechanism was applied to determine areas on the traffic sign which are especially favorable for placing adversarial stickers. The *targeted attention attack* (TAA) proposed by Yang et al. [49] generates shadow- or spot-looking perturbations on a traffic sign. The method relies on the soft attention map to find the most susceptible pixels for an attack. Zhong et al. [10] also explicitly generated shadow-looking perturbations. Recently, perturbations imitating raindrops have been described by Liu et al. [56].

[3]https://github.com/poojahira/gtsrb-pytorch

IV. Discussion and Conclusion

This survey presented an overview of the existing works on adversarial attacks against traffic sign detection and classification models. In the following, we summarize our observations.

Attacks on TSC vs. TSD: Attacks on classification models still remain predominant in the literature. Since the seminal work by Eykholt et al. [5], the evaluation stayed focused on LISA and GTSRB datasets. Attacks on TSD became popular with the introduction of GTSDB and especially TT100K datasets. Recent works [6], [51], [56] demonstrated successful attacks using the TT100K dataset. TSC baseline models evolved from simple three-layer CNNs like LISA-CNN or multi-scale GTSRB-CNN [5] to more sophisticated architectures, like a CNN with alternating convolutional and spatial transformer modules [8]. On the other hand, for TSD, mostly generic object detectors like YOLO were used.

Attack appearance and physical feasibility: Early works proposed black and white stickers [5] and perturbing the whole sign area [9] as an attack method. Recently, more sophisticated attacks were described, including a translucent patch [53], shadows [10], [49], raindrops [56], and emoji stickers [63]. Although attacks against TSR can easily be evaluated under realistic settings using printed traffic signs, there is a large gap between the feasibility of digital and real-world attacks. As shown by Lu et al. [9], large visible perturbations are needed to perform successful attacks in the real world. Furthermore, only several works go beyond stationary field experiments and perform drive-by field experiments.

Defense methods: Attack mitigation strategies especially for the traffic sign classification and detection tasks have received significantly less attention than attacks themselves. The work by Aung et al. [64] is one of the few which evaluated adversarial training and knowledge distillation to mitigate the attacks. In this work, a CNN for TSC on the GTSRB dataset was attacked with FGSM and JSMA. Defensive distillation and adversarial training were applied by Papernot et al. [65]. Recently, Zhang et al. proposed YOLOv2 [58] enhanced with adversarial patches during training [66]. Furthermore, an attempt to apply provable defenses was performed by Croce et al. [67].

In summary, our survey has demonstrated, how attacks on TSR have evolved from simple examples to more feasible attacks on detection and classification models. It has also shown, that research was mostly restricted to repeating baseline models and experiment settings. Furthermore, defense methods have not been studied extensively so far. We hope our overview of the existing evidence paves the way for research in this area.

Acknowledgment

This work was supported by KASTEL Security Research Labs.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing Properties of Neural Networks," *International Conference on Learning Representations (ICLR)*, 2014.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[3] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial Patch," in *Advances in Neural Information Processing Systems (NIPS) - Workshops*, 2017.

[4] S. Chen, C. Cornelius, J. Martin, and D. H. P. Chau, "Shapeshifter: Robust physical adversarial attack on faster R-CNN object detector," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*. Springer, 2018.

[5] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] X. Wei, Y. Guo, and J. Yu, "Adversarial sticker: A stealthy attack method in the physical world," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[7] P. Sermanet and Y. LeCun, "Traffic sign recognition with multi-scale convolutional networks," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2011.

[8] Á. A. García, J. A. Álvarez, and L. M. Soria-Morillo, "Deep neural network for traffic sign recognition systems: An analysis of spatial transformers and stochastic optimisation methods," *Neural Networks*, 2018.

[9] J. Lu, H. Sibai, and E. Fabry, "Adversarial examples that fool detectors," *CoRR*, vol. abs/1712.02494, 2017.

[10] Y. Zhong, X. Liu, D. Zhai, J. Jiang, and X. Ji, "Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022.

[11] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The german traffic sign recognition benchmark: A multi-class classification competition," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2011.

[12] F. Larsson and M. Felsberg, "Using fourier descriptors and spatial models for traffic sign recognition," in *Image Analysis - Scandinavian Conference, SCIA*. Springer, 2011.

[13] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The german traffic sign detection benchmark," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2013.

[14] R. Timofte, K. Zimmermann, and L. V. Gool, "Multi-view traffic sign detection, recognition, and 3d localisation," *Machine Vision and Applications*, 2014.

[15] A. Møgelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey," *Transactions on Intelligent Transportation Systems (T-ITS)*, 2012.

[16] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-sign detection and classification in the wild," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2016.

[17] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 2017.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *International Conference on Learning Representations (ICLR)*, 2018.

[19] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Asia Conference on Computer and Communications Security (AsiaCCS)*. ACM, 2017.

[20] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *IEEE European Symposium on Security and Privacy, EuroS&P*. IEEE, 2016.

[21] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy, SP*. IEEE Computer Society, 2017.

[22] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, 2017.

[23] D. Karmon, D. Zoran, and Y. Goldberg, "Lavan: Localized and visible adversarial noise," in *International Conference on Machine Learning (ICML)*. PMLR, 2018.

[24] S. Pavlitskaya, J. Hendl, S. Kleim, L. Müller, F. Wylczoch, and J. M. Zöllner, "Suppress with a patch: Revisiting universal adversarial patch attacks against object detection," in *IEEE International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. IEEE, 2022.

[25] S. Pavlitskaya, B. Codau, and J. M. Zöllner, "Feasibility of inconspicuous gan-generated adversarial patches against object detection," in *International Joint Conference on Artificial Intelligence (IJCAI) - Workshops*, 2022.

[26] F. Nesti, G. Rossolini, S. Nair, A. Biondi, and G. C. Buttazzo, "Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks," in *Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2022.

[27] S. Pavlitskaya, S. Ünver, and J. M. Zöllner, "Feasibility and suppression of adversarial patch attacks on end-to-end vehicle control," in *International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020.

[28] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *International Conference on Machine Learning (ICML)*. PMLR, 2018.

[29] N. Gray, M. Moraes, J. Bian, A. Tian, A. Wang, H. Xiong, and Z. Guo, "GLARE: A dataset for traffic sign detection in sun glare," *CoRR*, vol. abs/2209.08716, 2022.

[30] S. B. Wali, M. A. Abdullah, M. A. Hannan, A. Hussain, S. A. Samad, P. J. Ker, and M. B. Mansor, "Vision-based traffic sign detection and recognition systems: Current trends and challenges," *Sensors*, 2019.

[31] Á. Gonzalez, L. M. Bergasa, and J. J. Y. Torres, "Text detection and recognition on traffic panels from street-level imagery using visual appearance," *Transactions on Intelligent Transportation Systems (T-ITS)*, 2014.

[32] A. Ellahyani, M. E. Ansari, and I. E. Jaafari, "Traffic sign detection and recognition based on random forests," *Applied Soft Computing*, 2016.

[33] M. Romadi, R. O. H. Thami, R. Romadi, and R. Chiheb, "Detection and recognition of road signs in a video stream based on the shape of the panels," in *International Conference on Intelligent Systems SITA*. IEEE, 2014.

[34] D. C. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2011.

[35] J. Jin, K. Fu, and C. Zhang, "Traffic sign recognition with hinge loss trained convolutional neural networks," *Transactions on Intelligent Transportation Systems (T-ITS)*, 2014.

[36] G. N. Doval, A. Al-Kaff, J. Beltrán, F. G. Fernández, and G. F. López, "Traffic sign detection and 3d localization via deep convolutional neural networks and stereo vision," in *International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2019.

[37] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[38] Y. Rehman, H. Amanullah, M. A. Shirazi, and M. Y. Kim, "Small traffic sign detection in big images: Searching needle in a hay," *IEEE Access*, 2022.

[39] Y. Tian, J. Gelernter, X. Wang, J. Li, and Y. Yu, "Traffic sign detection using a multi-scale recurrent attention network," *Transactions on Intelligent Transportation Systems (T-ITS)*, 2019.

[40] C. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD : Deconvolutional single shot detector," *CoRR*, vol. abs/1701.06659, 2017.

[41] D. Song, K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramèr, A. Prakash, and T. Kohno, "Physical adversarial examples for object detectors," in *USENIX Workshop on Offensive Technologies*. USENIX Association, 2018.

[42] C. Sitawarin, A. N. Bhagoji, A. Mosenia, P. Mittal, and M. Chiang, "Rogue signs: Deceiving traffic sign recognition with malicious ads and logos," *CoRR*, vol. abs/1801.02780, 2018.

[43] C. Sitawarin, A. N. Bhagoji, A. Mosenia, M. Chiang, and P. Mittal, "DARTS: deceiving autonomous cars with toxic signs," *CoRR*, vol. abs/1802.06430, 2018.

[44] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, "Perceptual-sensitive GAN for generating adversarial patches," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2019.

[45] N. Morgulis, A. Kreines, S. Mendelowitz, and Y. Weisglass, "Fooling a real car with adversarial traffic signs," *CoRR*, vol. abs/1907.00374, 2019.

[46] Y. Li, X. Xu, J. Xiao, S. Li, and H. T. Shen, "Adaptive square attack: Fooling autonomous cars with adversarial traffic signs," *IEEE Internet of Things Journal*, 2021.

[47] C. Guo, J. R. Gardner, Y. You, A. G. Wilson, and K. Q. Weinberger, "Simple black-box adversarial attacks," in *International Conference on Machine Learning (ICML)*. PMLR, 2019.

[48] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *European Conference on Computer Vision (ECCV)*. Springer, 2020.

[49] X. Yang, W. Liu, S. Zhang, W. Liu, and D. Tao, "Targeted attention attack on deep learning models in road sign recognition," *IEEE Internet of Things Journal*, 2021.

[50] F. Woitschek and G. Schneider, "Physical adversarial attacks on deep neural networks for traffic sign recognition: A feasibility study," in *Intelligent Vehicles Symposium (IV)*. IEEE, 2021.

[51] W. Jia, Z. Lu, H. Zhang, Z. Liu, J. Wang, and G. Qu, "Fooling the eyes of autonomous vehicles: Robust physical adversarial examples against traffic sign recognition systems," in *Annual Network and Distributed System Security Symposium NDSS*. The Internet Society, 2022.

[52] B. Ye, H. Yin, J. Yan, and W. Ge, "Patch-based attack on traffic sign recognition," in *International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2021.

[53] A. Zolfi, M. Kravchik, Y. Elovici, and A. Shabtai, "The translucent patch: A physical and universal attack on object detectors," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[54] L. Chi, M. Msahli, G. Memmi, and H. Qiu, "Public-attention-based adversarial attack on traffic sign recognition," in *IEEE Consumer Communications & Networking Conference CCNC*. IEEE, 2023.

[55] D. C. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural Networks*, 2012.

[56] J. Liu, B. Lu, M. Xiong, T. Zhang, and H. Xiong, "Adversarial attack with raindrops," *CoRR*, 2023.

[57] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016.

[58] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2017.

[59] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Conference on Computer and Communications Security (CCS)*. ACM, 2016.

[60] J. Lu, H. Sibai, E. Fabry, and D. A. Forsyth, "Standard detectors aren't (currently) fooled by physical adversarial stop signs," *CoRR*, vol. abs/1710.03337, 2017.

[61] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[62] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control*, 1992.

[63] P. A. Sava, J. Schulze, P. Sperl, and K. Böttinger, "Assessing the impact of transformations on physical adversarial attacks," in *ACM Workshop on Artificial Intelligence and Security*. ACM, 2022.

[64] A. M. Aung, Y. Fadila, R. Gondokaryono, and L. Gonzalez, "Building robust deep neural networks for road sign detection," *CoRR*, vol. abs/1712.09327, 2017.

[65] N. Papernot, P. D. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2016.

[66] Y. Zhang, J. Cui, and M. Liu, "Research on adversarial patch attack defense method for traffic sign detection," in *Cyber Security - China Annual Conference, CNCERT*, W. Lu, Y. Z. andf Weiping Wen, H. Yan, and C. Li, Eds., 2022.

[67] F. Croce, M. Andriushchenko, and M. Hein, "Provable robustness of relu networks via maximization of linear regions," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2019.