Shadows Aren't So Dangerous After All: A Fast and Robust Defense Against Shadow-Based Adversarial Attacks

Andrew Wang
Cornell University
aw632@cornell.edu

Wyatt Mayor
Monmouth College
wmayor@monmouthcollege.edu

Ryan Smith
University of Arizona
ryansmith1@arizona.edu

Gopal Nookula
U.C. Riverside
gnook001@ucr.edu

Gregory Ditzler
Rowan University
ditzler@rowan.edu

Abstract

Robust classification is essential in tasks like autonomous vehicle sign recognition, where the downsides of misclassification can be grave. Adversarial attacks threaten the robustness of neural network classifiers, causing them to consistently and confidently misidentify road signs. One such class of attack, shadow-based attacks, causes misidentifications by applying a natural-looking shadow to input images, resulting in road signs that appear natural to a human observer but confusing for these classifiers. Current defenses against such attacks use a simple adversarial training procedure to achieve a rather low 25% and 40% robustness on the GTSRB and LISA test sets, respectively. In this paper, we propose a robust, fast, and generalizable method, designed to defend against shadow attacks in the context of road sign recognition, that augments source images with binary adaptive threshold and edge maps. We empirically show its robustness against shadow attacks, and reformulate the problem to show its similarity ε perturbation-based attacks. Experimental results show that our edge defense results in 78% robustness while maintaining 98% benign test accuracy on the GTSRB test set, with similar results from our threshold defense. [1]

1. Introduction

With the great success of neural networks in image classification has come the great vulnerability of adversarial examples—a class of examples designed to exploit the brittle nature of deep neural networks and fool models into making incorrect classifications by making small, human-imperceptible changes to the image. When these adversarial examples appear in mission-critical settings such as autonomous driving [25], medical imaging [19], and financial management [17], the effects can be disastrous. The importance of defending against such examples and adversarial attacks has therefore spawned countermeasures, which have in turn spawned more advanced attacks, leading to a sort of adversarial arms race [34]. One recent adversarial attack, proposed by Zhong *et al*. [45], involves darkening a section of the input image ("shadowing"), thereby causing misclassifications. Not only is this attack extremely effective against SOTA sign recognition models, achieving 90% and 98% attack success rates on the GTSRB and LISA benchmark datasets respectively, it is also realistic, requiring little to no specialized equipment and is easily unnoticeable by human drivers.

We propose a new defense (see Fig. 1) against this attack based on adaptive threshold and edge maps that—even with no hyperparameter optimization—achieves 78% robustness by trading off only roughly 1% benign test accuracy. In specific, we:

- motivate the use of adaptive threshold and edge maps with saliency maps,
- developing the first-known defense against the shadow adversarial attack
- demonstrate robustness, and effectiveness of our defense against shadow-based adversarial attacks,
- show its generalizability against classic gradient-based adversarial attacks,
- and reformulate the shadows attack as an instance of ε perturbations within an ℓ_∞ ball.

[1] Our code is available at https://github.com/aw632/ShadowDefense.

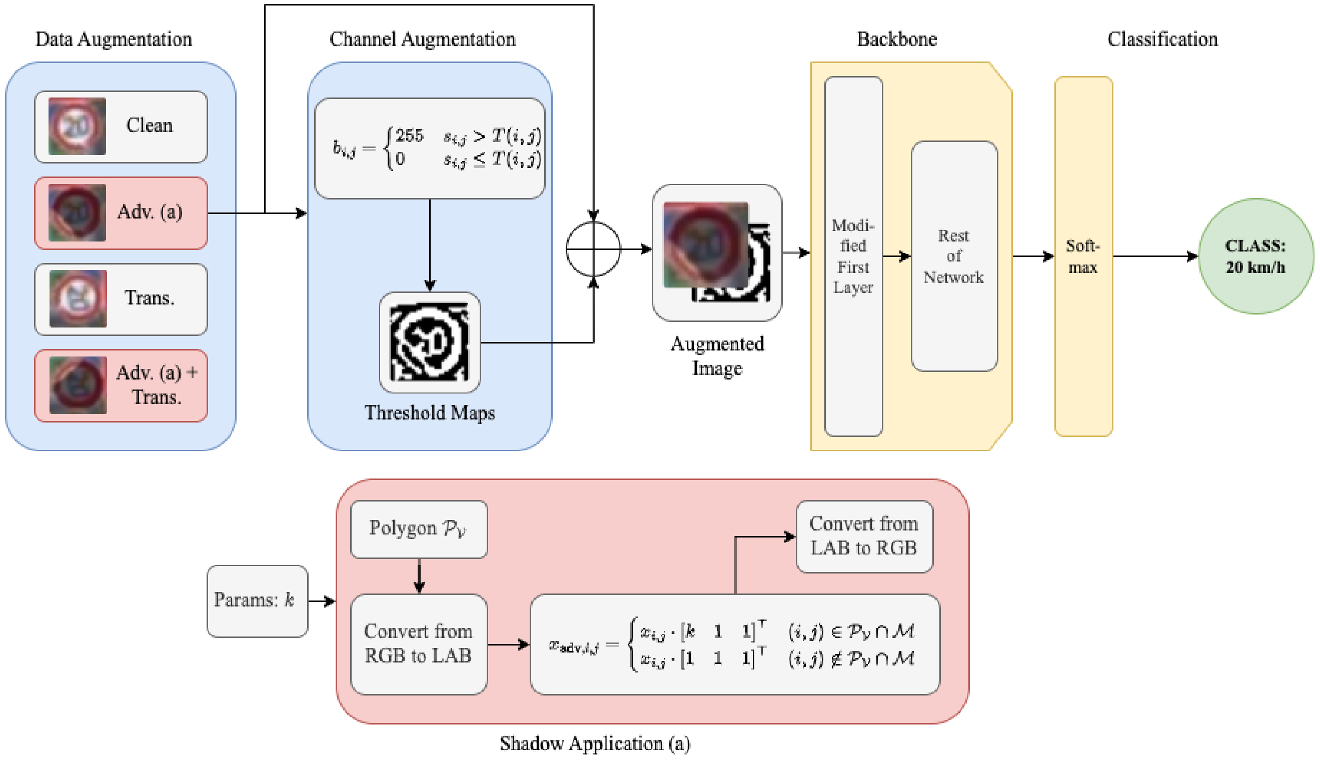
Data Augmentation
Clean
Adv. (a)
Trans.
Adv. (a) + Trans.
Channel Augmentation
$b_{i,j} = \begin{cases} 255 & s_{i,j} > T(i,j) \\ 0 & s_{i,j} \leq T(i,j) \end{cases}$
Threshold Maps
Augmented Image
Backbone
Modi-fied First Layer
Rest of Network
Classification
Soft-max
CLASS: 20 km/h
Polygon $\mathcal{P}_{\mathcal{V}}$
Params: k
Convert from RGB to LAB
$x_{\mathrm{adv},i,j} = \begin{cases} x_{i,j} \cdot [k \quad 1 \quad 1]^\top & (i,j) \in \mathcal{P}_{\mathcal{V}} \cap \mathcal{M} \\ x_{i,j} \cdot [1 \quad 1 \quad 1]^\top & (i,j) \notin \mathcal{P}_{\mathcal{V}} \cap \mathcal{M} \end{cases}$
Convert from LAB to RGB
Shadow Application (a)

Figure 1. An overview of our defense. We take source images, apply a shadow with parameter k randomly to the image, transform the image with shear, rotation, and translation, and use adaptive thresholding ("AdaThresh") to generate a binary threshold map. The transformed image and the threshold map are concatenated to form a 4-channel image, which the model is retrained on. The rest of the model architecture ("backbone") need not be changed for good robustness.

2. Background and Related Work

Adversarial examples from a general perspective were first introduced by Dalvi *et al.* [7], and refined by Szegedy *et al.* [42] with a constraint that the adversarial examples must be no more than ε away from the clean example, thereby causing examples to be impercetible to humans for sufficiently small ε.

2.1. Adversarial Examples

In this work, we consider a classification setting with C distinct classes, in which we wish to classify an RGB image $x \in \mathbb{R}^n$ by passing x through a model M (such as a neural network) and computing a class label $y = M(x) \in \{1, \ldots, C\}$.

This setting includes an adversary, who is able to take any such example x and transform it into x_{adv}, a point in the convex set of possible transformed examples $S_{\text{adv}}(x) \subseteq \mathbb{R}^n$. Intuitively, $S_{\text{adv}}(x)$ is a collection of all the examples that, by some metric, are "close" to x [1]. For instance, if the adversary wishes to keep the transformations within a ε range L_p perturbation from x, then $S_{\text{adv}}(x) = \{x' : ||x' - x||_p < \varepsilon\}$.

An adversarial example x_{adv}, then, is an example that is "close" to x yet causes the model M to misclassify it:

$$M(x) \neq M(x_{\text{adv}}) \quad x_{\text{adv}} \in S_{\text{adv}}(x). \tag{1}$$

Extending this notion, an adversarial attack is a principled manner for an adversary to generate these adversarial examples, *e.g.*, through some mathematical formulation or algorithm.

2.2. Shadow Attack on Road Signs

Shadows as an adversarial attack were first proposed by Zhong *et al.* [45], although concern over shadows in road sign recognition have existed for some time [12, 23]. Their attack generates an adversarial image by

1. choosing a parameter k representing the "darkness" or "strength" of the shadow, where higher values of k indicate weaker shadows and vice versa,

2. locating a polygon $\mathcal{P}_{\mathcal{V}}$, defined by a set of vertices $\mathcal{V} = \{(m_1, n_1), \ldots, (m_s, n_s)\}$, and a mask $\mathcal{M}$ to locate the target polygon,

3. converting x from RGB color space to LAB image space, such that each element in $x_{i,j} \in \mathbb{R}^3$ represents the L, A, and B channels respectively,

4. forming a new image x_{adv} by recalculating the value of every pixel (i, j) with

$$x_{\text{adv},i,j} = \begin{cases} x_{i,j} \cdot \begin{bmatrix} k & 1 & 1 \end{bmatrix}^\top & (i,j) \in \mathcal{P}_\mathcal{V} \cap \mathcal{M} \\ x_{i,j} \cdot \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\top & (i,j) \notin \mathcal{P}_\mathcal{V} \cap \mathcal{M} \end{cases}, \quad (2)$$

5. and converting x_{adv} back into RGB space [45].

Finding $\mathcal{V}$ can be formulated as an optimization problem, which Zhong *et al.* solve by means of Particle Swarm Optimization (PSO) [21].

2.3. Adaptive Thresholding

In general, **thresholding** is the process of generating a binary image b (a.k.a. the **threshold map**) from a source image s, where the white pixels are the "foreground" (255) elements and the black pixels are the "background" (0) elements. The "threshold" is the means by which foreground and background are separated. **Adaptive thresholding** [3, 44] is a form of local thresholding, where each pixel $b_{i,j}$ is assigned to the foreground or the background by a threshold function T parameterized on pixel coordinates i, j:

$$b_{i,j} = \begin{cases} 255 & s_{i,j} > T(i,j) \\ 0 & s_{i,j} \leq T(i,j). \end{cases} \quad (3)$$

The benefit of adaptive thresholding over global thresholding is that T is parameterized on each pixel, and is thus robust to the spatial changes in the illuminations that represent the variations present in shadow attacks, as opposed to setting one uniform threshold for the entire image.

In our approach, we let T be a Gaussian-window weighted sum of a $k \times k$ neighborhood around (i, j) [37, 40]. If $N(i, j)$ contains the set of all points within a $k \times k$ neighborhood around (i, j), then

$$T(i,j) = \sum_{(x,y) \in N(i,j)} G_{x,y} \cdot s_{x,y} \quad (4)$$

where $G_{i,j}$ is the Gaussian-window (Fig. 2) weight for pixel (i, j), defined as

$$G_{i,j} = \alpha \cdot \exp\left(\frac{-(i - \frac{k-1}{2})^2 - (j - \frac{k-1}{2})^2}{2\sigma^2} \right) \quad (5)$$

where the standard deviation σ is calculated from all pixels in $N(i, j)$ and α is a scaling factor such that the $G_{i,j}$ sum to 1. k is a tuneable hyperparameter representing the "aperture" of our local threshold; we chose $k = 3$ to due to the small size of our input images, but the primary focus is on presenting the robustness of the method and not tuning hyperparameters.

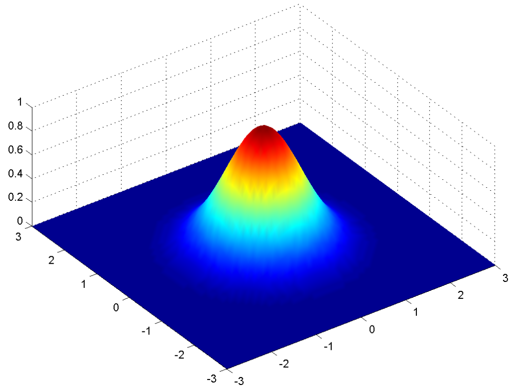

Figure 2. A visualization of $G_{i,j}$ around the point $(i, j) = (0, 0)$ with window size $k = 3$. The motivation behind using Gaussian window weights is that pixels closest to (i, j) receive maximum weights, while pixels further away receive less weight. Image licensed under Creative Commons Attribution-Share Alike 3.0 [15].

2.4. Canny Edge Detection

Closely related to local thresholding is Canny edge detection, first proposed in 1986 [4]. The Canny edge detection algorithm usually takes the following steps [10]:

1. smooth the image with a Gaussian-window filter (see Eq. (5)) to reduce noise,

2. determine the gradient magnitude and direction at each pixel,

3. apply a custom thresholding function wherein edge ("foreground") pixels have gradient magnitudes larger than those of its two neighbors in the gradient direction,

4. and clean up extraneous or "weak" edges with a hysteresis thresholding step with two threshold parameters, t_{hi} and t_{lo}. Pixels that are above t_{hi} become edges, below t_{lo} are rejected as edges, and in between are edges only if they are connected to an edge pixel.

The output of any edge algorithm [22], Canny included, is referred to as the **edge map** of the image.

Canny edge detection is both local and global: the first thresholding step is local with respect to the gradient direction, while the second thresholding step is global, as the parameters are set for the whole image.

While both methods are tested in our approach, we hypothesize that for darker shadows, the greater variance between gradient pixels *inside the shadowed region* and gradient pixels *outside the shadowed region* would cause Canny edge maps to have worse performance than adaptive threshold maps.

2.5. Related Work

While shadows as an adversarial attack were first conceived by Zhong *et al*. [45], shadows have been a source of concern in traffic sign recognition for some time [8, 13]. However, to our knowledge, the only defense against such attacks remains that proposed by Zhong *et al*.: a simple adversarial training scheme.

More generally, shadows as an adversarial attack falls into a class of non-invasive optical attacks which are particularly pernicious against self driving cars. While digital attacks like FGSM [20] and DeepFool [26] are effective in the online domain, such attacks are infeasible in realistic self-driving scenarios, leading to the rise of physical, non-invasive attacks that cause natural or imperceptible perturbations to physical road signs [16, 24, 33, 38]. Zhong *et al*. noted these attacks relied on sophisticated and complex equipment, making their implementation impractical—for this same reason, it is impractical for us to test our defense on these attacks.

To our knowledge, there is no literature regarding the use of adaptive threshold maps or threshold maps as adversarial defenses. However, edge detection is closely related to thresholding [27]. Previous work has investigated edges as an adversarial defense against different classification tasks [9, 41], but their work both use edges as the sole source of information, rather than augmenting it to existing images as in our method. Doing so puts heavily reliance on the quality of edge detectors, which can introduce their own vulnerabilities to adversarial examples if neural networks are used [6].

3. Our Approach

In this section, we give motivations for our defense and empirical justification for those motivations. We then describe our defense and training regime. Our framework supports two defenses: one with adaptive threshold maps and one with edge maps, and requires no more than one modification to the network architecture to achieve good results against shadow attacks.

3.1. Our Motivation

We were initially motivated to investigate edge profiling as a possible defense, since

- humans recognize road signs based on the boundaries of the sign and the text or symbols within, which can be represented with an edge map [2], and

- convolutional neural networks, which are often used in state-of-the-art (SOTA) road sign recognition, can learn "unimportant" (from a human perspective) and non-robust features like texture [14], leaving them vulnerable to adversarial attacks which disturb these features.

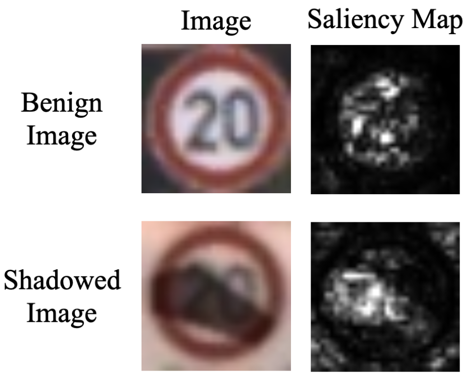

Figure 3. Saliency maps [28, 35] (right) for a randomly selected benign traffic sign (top left) and its adversarially-shadowed counterparts (bottom left) from the GTSRB [39] dataset on the model from [11]. The network appears to learn features inside the sign boundary, leaving it vulnerable when the inside of the sign is perturbed heavily.

However, edge detection algorithms are imperfect and edge maps alone may be insufficient for object recognition [32]. With this in mind, we decided to proceed with appending edge maps of source images as a fourth channel as a way of "emphasizing" the importance of edges, without destroying the information provided by the source image.

This intuition is empirically justified by experiments with saliency maps [35,36], a sample of which are presented in Fig. 3. These experiments indicate that SOTA sign recognition models, such as those found in Eykholt *et al*. [11], learn features inside the sign but not necessarily the text or boundaries of the sign itself.

The motivation behind thresholding was based on our hypothesis in Sec. 2.4 and empirical tests that threshold maps appeared more readable and less noisy than edge maps (Fig. 4). However, for completeness, we tested both methods.

3.2. Our Defense

Given a training dataset D_{train} and a model M, our defense quadruplicates the dataset based on boolean flags `adv` and `transform`, which control the addition of adversarial examples and transformed images (shear, rotate, and translate) to the training regime. For each of these dataset duplicates, our defense

1. modifies the first layer of M to accommodate a 4-channel source image s,

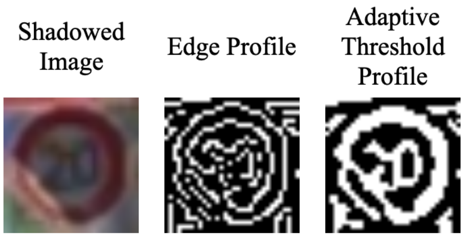

Figure 4. A comparison of a shadowed image (left) with parameter $k = 0.43$, its edge map (middle) and its threshold map (right). The threshold map appears much less noisy, motivating the use of threshold maps.

2. randomly applies a shadow if `adv` is `True`,
3. generates the relevant profile p (threshold or edge map),
4. appends the profile p to s as a fourth channel, making a new image x,
5. transforms x if `transform` is `True`, and trains the model on x.

If augmenting source images with threshold maps, then we generate an threshold map using Eq. (3). In the case of edge maps, our defense uses Canny edge detection due to its non-differentiablity (compared to a neural network edge mechanism) and speed. The parameters t_{high} and t_{low} are chosen based on the formula

$$\begin{aligned} t_{\text{lo}} &= \max(0, \mu \cdot (1-\sigma)) \\ t_{\text{high}} &= \min(255, \mu \cdot (1+\sigma)) \end{aligned} \tag{6}$$

where μ is the median of the image across all channels and σ is is a parameter representing the "blurring" of the edges: higher sigma indicates more extreme blurring. Intuitively, this formula takes σ as a standard deviation and sets the upper and lower thresholds to be one standard deviation from the mean.

In our experimentation, we chose $\sigma = 0.33$ to maintain the 2:1 ratio of t_{high} to t_{low} recommended by Canny [4]. We also chose μ as the median and not the mean, since shadows bias the image histogram towards the darker side.

4. Experimental Results

In this section, we present our experiments, namely:

- we present robustness results for models using edge maps and adaptive threshold maps,
- we present benign test accuracies for both models,
- we reformulate the shadow attack with parameter k as an instance of a ε perturbations within an ℓ_p ball,
- using the above reformulation, we present robustness results for both models against the Fast Gradient Sign Method (FGSM) and Carlini-Wagner.

4.1. Experimental Setup

For consistency, we used the same network architectures [11] and preprocessing steps as Zhong *et al.* [45], with the exception that our networks were modified to take in 4 channels and retrained accordingly.

We also used the same train/test split and the same code as in Zhong *et al.* [45] for the shadow attack, and evaluated the robustness accordingly. This includes their exclusion of images which are already too dark; i.e., the mean of their pixel values in the L channel is no larger than 120.

Our experiments were conducted with the PyTorch deep learning library [29], on a machine with eight Intel(R) Xeon(R) Bronze 3106 CPU @ 1.70GHz CPUs and one NVIDIA Titan-XP GPU with CUDA version 10.2. The operating system was Ubuntu 18.04 (LTS) with Python version 3.10.5. This system setup was also used to reproduce the results of Zhong *et al.* [45].

4.2. Core Robustness Results

We used Algorithm 1 as our testing regime, and tested the same k values as Zhong *et al.* [45]; namely, the set of values

$$\begin{aligned} \mathcal{K} = \{ & 0.20, 0.25, 0.30, 0.35, 0.40, 0.43, 0.45, \\ & 0.45, 0.50, 0.55, 0.60, 0.65, 0.70\}. \end{aligned}$$

Note that the value $k = 0.43$ is the median shadow value from the SBU Shadow Dataset [43]. There are two results we record: the robustness, defined as $1-$attack success rate, and the number of queries that the shadow attack makes to the backbone model, which is a measure of the attack's stealthiness.

Our results are presented in Tab. 1 and Tab. 2 for robustness and average number of queries, respectively. In line with our hypotheses in Sec. 3, we found that while both edge maps and adaptive thresholding provided similar robustness, the quality of edge maps depended greatly on the strength of the shadow, whereas adaptive threshold maps, due to their local nature, were more invariant.

4.3. Reformulation

While the shadow attack appears to be distinct from the sort of ε based attacks introduced by Szegedy *et al.* [42], we show that we can reformulate the shadow attack as an instance of a ε-based perturbation.

See appendix for theorem statement and proof.

Defense, **GTSRB**	$k = 0.20$	0.25	0.30	0.35	0.40	**0.43**	0.45	0.50	0.55	0.60	0.65	0.70
Zhong *et al*.: None	2.63	3.65	4.86	6.55	8.64	**9.53**	11.03	12.85	15.75	19.64	26.24	33.27
Zhong *et al*.: AT	15.89	17.78	20.39	23.12	26.69	**28.38**	30.03	30.26	33.78	43.01	48.57	55.30
Ours: AdaThresh	73.82	74.69	75.12	75.58	76.41	**76.63**	76.29	76.71	77.61	79.65	79.26	78.57
Ours: Edges	75.41	75.19	76.82	76.89	77.08	**78.02**	78.12	79.33	80.59	81.97	83.55	85.46

Table 1. The above table describes the average robustness (1− success rate of attack) over $n = 5$ trials for a specified defense (adversarial training from Zhong *et al*., our adaptive threshold maps, or our edge maps) and shadow attack with parameter k on the **GTSRB** dataset. While AdaThresh tends to have lower robustness, it has a tighter variance than that of edge map defenses.

Defense, **GTSRB**	$k = 0.20$	0.25	0.30	0.35	0.40	**0.43**	0.45	0.50	0.55	0.60	0.65	0.70
Zhong *et al*.: AT	98	93	112	129	128	**126**	136	155	188	232	249	343
Ours: AdaThresh	210	251	256	289	348	**355**	331	329	430	501	574	560
Ours: Edges	304	277	370	376	396	**456**	450	515	580	645	726	819

Table 2. The above table describes the average number of queries to the backbone model over $n = 5$ trials for a specified defense (adversarial training from Zhong *et al*., our adaptive threshold maps, or our edge maps) and shadow attack with parameter k. This is a proxy measure of the "stealthiness" of the black-box attack. The variance in edge map defenses remains significantly higher than AdaThresh defenses.

Algorithm 1 Our Testing Regime

```
for dataset in {datasets} do
    for $k$ in $\mathcal{K}$ do
        $\alpha \leftarrow$ test accuracy on benign examples
        $\beta_1 \leftarrow$ test accuracy on shadowed examples
        $\beta_2 \leftarrow$ num. queries on shadowed examples
    end for
end for
```

4.4. Robustness Against Other Attacks

Two popular adversarial attacks are chosen: Fast Gradient Sign Method (FGSM) [18], and Carlini-Wagner [5], for gradient-based and non-gradient-based methods, respectively. To test our robustness against these attacks, we test using ε values from the literature; namely $\varepsilon = 0.007$ from Goodfellow *et al*. [18]. While we obtained a value for ε for $k = 0.43$ based on Theorem 1, we remark that using that resultant value on FGSM completely destroys the image, so we omit testing with such a value.

5. Discussion and Limitations

Beyond providing additional information in the form of a fourth channel, this attack also combines gradient masking (as the binary edge or threshold map is non-differentiable) and adversarial training (as it is trained on shadowed images. However, even though it uses gradient masking, the defense remains robust even to non-gradient based attacks like shadow attacks, and is slightly robust to Carlini-Wagner.

Defense, **GTSRB**	FGSM, $\varepsilon = 0.007$	Carlini-Wagner
None	37.40	2.39*
Adv. Training	85.31	16.94*
AdaThresh	90.21	23.15*
Edges	91.47	20.36*

Table 3. We compare the attack success rate of the Shadow Attack, FGSM, and Carlini-Wagner on the GTSRB dataset. Our defenses are No Defense, Adversarial Training Only, AdaThresh, and Edge Maps. Attacks provided by Foolbox [30, 31]. (*) Due to the slow speed of Carlini-Wagner, we took a random, class proportionality-preserving subsample of $5,000$ images.

There are a few limitations to this attack. First, its reliance on adversarial retraining makes its training process slow. Second, it uses off-the-shelf edge and adaptive thresholding methods; it is worth investigating a more bespoke set of thresholding equations in future research. Third, its robustness is significantly higher than the baseline, but insufficient for critical tasks like sign recognition; our work should be seen as a starting point and not a final defense.

6. Conclusion, Acknowledgements, and Funding

In this paper, we presented a novel and simple adversarial defense against shadow-based adversarial attacks. Our defense requires no retuning or redesign of the model architecture to achieve good robustness against shadow attacks. Furthermore, our defense remains robust against classic gradient-based adversarial attacks, showing its ex-

tensiblity. We would like to thank Huayu Li and David Schwartz for their assistance in developing this paper. This paper was funded by National Science Foundation (NSF) Grant IIS-1950359.

References

[1] Mislav Balunović and Martin Vechev. Adversarial training and provable defenses: Bridging the gap. In *8th International Conference on Learning Representations (ICLR 2020)(virtual)*. International Conference on Learning Representations, 2020. 2

[2] Aayush Bansal, Adarsh Kowdle, Devi Parikh, Andrew Gallagher, and Larry Zitnick. Which edges matter? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 578–585, 2013. 4

[3] Derek Bradley and Gerhard Roth. Adaptive thresholding using the integral image. *Journal of graphics tools*, 12(2):13–21, 2007. 3

[4] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 3, 5

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 6

[6] Christian Cosgrove and Alan Yuille. Adversarial examples for edge detection: They exist, and they transfer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1070–1079, 2020. 4

[7] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108, 2004. 2

[8] Arturo De la Escalera, J Ma Armingol, and Mario Mata. Traffic sign recognition and analysis for intelligent vehicles. *Image and vision computing*, 21(3):247–258, 2003. 4

[9] Gavin Weiguang Ding, Kry Yik Chau Lui, Xiaomeng Jin, Luyu Wang, and Ruitong Huang. On the sensitivity of adversarial robustness to input data distributions. In *ICLR (Poster)*, 2019. 4

[10] Lijun Ding and Ardeshir Goshtasby. On the canny edge detector. *Pattern recognition*, 34(3):721–725, 2001. 3

[11] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018. 4, 5

[12] Hasan Fleyeh. Shadow and highlight invariant colour segmentation algorithm for traffic signs. In *2006 IEEE Conference on Cybernetics and Intelligent Systems*, pages 1–7. IEEE, 2006. 2

[13] Hasan Fleyeh and Erfan Davami. Eigen-based traffic sign recognition. *IET Intelligent Transport Systems*, 5(3):190–196, 2011. 4

[14] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 4

[15] Kaushik Ghose. Isometric plot of a two dimensional gaussian, 2006. 3

[16] Abhiram Gnanasambandam, Alex M Sherman, and Stanley H Chan. Optical adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 92–101, 2021. 4

[17] Micah Goldblum, Avi Schwarzschild, Ankit Patel, and Tom Goldstein. Adversarial attacks on machine learning systems for high-frequency trading. In *Proceedings of the Second ACM International Conference on AI in Finance*. ACM, nov 2021. 1

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 6

[19] Hokuto Hirano, Akinori Minagi, and Kazuhiro Takemoto. Universal adversarial attacks on deep neural networks for medical image classification. *BMC medical imaging*, 21(1):1–13, 2021. 1

[20] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017. 4

[21] James Kennedy and Russell Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948. IEEE, 1995. 3

[22] Mukesh Kumar, Rohini Saxena, et al. Algorithm and technique on various edge detection: A survey. *Signal & Image Processing*, 4(3):65, 2013. 3

[23] Haojie Li, Fuming Sun, Lijuan Liu, and Ling Wang. A novel traffic sign detection method via color segmentation and robust shape matching. *Neurocomputing*, 169:77–88, 2015. 2

[24] Juncheng Li, Frank Schmidt, and Zico Kolter. Adversarial camera stickers: A physical camera-based attack on deep learning systems. In *International Conference on Machine Learning*, pages 3896–3904. PMLR, 2019. 4

[25] KT Yasas Mahima, Mohamed Ayoob, and Guhanathan Poravi. Adversarial attacks and defense technologies on autonomous vehicles: A review. *Appl. Comput. Syst.*, 26(2):96–106, 2021. 1

[26] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 4

[27] Ehsan Nadernejad, Sara Sharifzadeh, and Hamid Hassanpour. Edge detection techniques: Evaluations and comparisons. *Applied Mathematical Sciences*, 2(31):1507–1520, 2008. 4

[28] Utku Ozbulak. Pytorch cnn visualizations. https://github.com/utkuozbulak/pytorch-cnn-visualizations, 2019. 4

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[30] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. 6

[31] Jonas Rauber, Roland Zimmermann, Matthias Bethge, and Wieland Brendel. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax. *Journal of Open Source Software*, 5(53):2607, 2020. 6

[32] Thomas Sanocki, Kevin W Bowyer, Michael D Heath, and Sudeep Sarkar. Are edges sufficient for object recognition? *Journal of Experimental Psychology: Human Perception and Performance*, 24(1):340, 1998. 4

[33] Athena Sayles, Ashish Hooda, Mohit Gupta, Rahul Chatterjee, and Earlence Fernandes. Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14666–14675, 2021. 4

[34] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? 2018. 1

[35] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017. 4

[36] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 4

[37] Julius O. Smith. *Spectral Audio Signal Processing*. http://ccrma.stanford.edu/~jos/sasp/, accessed ¡date¿. online book, 2011 edition. 3

[38] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th USENIX workshop on offensive technologies (WOOT 18)*, 2018. 4

[39] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012. 4

[40] Mallat Stephane. A wavelet tour of signal processing, 1999. 3

[41] Mingjie Sun, Zichao Li, Chaowei Xiao, Haonan Qiu, Bhavya Kailkhura, Mingyan Liu, and Bo Li. Can shape structure features improve model robustness under diverse adversarial settings? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7526–7535, 2021. 4

[42] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2, 5

[43] Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *European Conference on Computer Vision*, pages 816–832. Springer, 2016. 5

[44] James M White and Gene D Rohrer. Image thresholding for optical character recognition and other applications requiring character image extraction. *IBM Journal of research and development*, 27(4):400–411, 1983. 3

[45] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15345–15354, 2022. 1, 2, 3, 4, 5

A. Appendix

Theorem 1. *Given an image in RGB color space $x = [R(x), G(x), B(x)]$, its LAB counterpart $x_{lab} = [L(x), A(x), B(x)]$, and its adversarial counterparts x_{adv} and $x_{lab,\ adv}$, for a shadow attack with parameter k, the adversarial perturbation $||x_{adv} - x|| = \varepsilon_k$ is at most $||M|| \cdot 100|k-1|$, where M is the matrix multiplication to convert from LAB to RGB space.*

Proof of Thm. 1:

Proof. By construction, the image x_{lab} is uniformly scaled only in its L channel by a factor of $k < 1$. Denote $x_{lab}[L]$ as the L channel of x in LAB color space, and denote $x_{lab,adv}[L]$ as the L channel of $x_{lab,adv}$ in LAB color space. Then

$$||x_{lab,adv}[L] - x_{lab}[L]|| = ||(k-1)x_{lab}[L]|| \tag{7}$$
$$= |k-1| ||x_{lab}[L]|| \tag{8}$$

where the ℓ_∞ norm is being used, and by definition of the L channel in a LAB image, $||x_{lab}[L]|| \leq 100$. Thus,

$$||x_{adv}[L] - x_{lab}[L]|| \leq 100|k-1||. \tag{9}$$

By construction, $x = Mx_{lab}$ and $x_{adv} = Mx_{lab,adv}$ and thus

$$||x_{adv} - x|| = ||Mx_{lab,adv} - Mx_{lab}|| \tag{10}$$
$$\leq ||M||_\infty \cdot ||x_{lab,adv} - x_{lab}|| \tag{11}$$
$$\leq ||M||_\infty \cdot 100|k-1| \tag{12}$$

where $||M||_\infty$ is the induced norm on a matrix. □