Short: Certifiably Robust Perception Against Adversarial Patch Attacks: A Survey

Chong Xiang
Princeton University
cxiang@princeton.edu

Chawin Sitawarin
University of California, Berkeley
chawins@berkeley.edu

Tong Wu
Princeton University
tongwu@princeton.edu

Prateek Mittal
Princeton University
pmittal@princeton.edu

***Abstract*—The physical-world adversarial patch attack poses a security threat to AI perception models in autonomous vehicles. To mitigate this threat, researchers have designed defenses with certifiable robustness. In this paper, we survey existing certifiably robust defenses and highlight core robustness techniques that are applicable to a variety of perception tasks, including classification, detection, and segmentation. We emphasize the unsolved problems in this space to guide future research, and call for attention and efforts from both academia and industry to robustify perception models in autonomous vehicles.**

I. Introduction

The functionality of autonomous vehicles (AV) relies on the ability to accurately perceive their physical surroundings. Nevertheless, security researchers have discovered vulnerabilities in AI-powered AV perception models. For example, in the physical world, attackers can attach stickers to a stop sign to make AV misinterpret it as "speed limit" [9], or place a dirty patch on the road to interfere with lane detection [30]. A large number of these attacks [8], [9], [30], [33] fall into the category of adversarial patch attacks [2], where the attacker can control a spatially constrained region of the physical world. As part of the effort to develop safe AV systems, we are motivated to study defenses against adversarial patches.

Mitigating the threat of adversarial patches is never easy. Most defenses [12], [22], [27], [34] are heuristics-based and lack formal security guarantee. This makes them unsuitable for critical applications of autonomous vehicles, where unexpected failures are highly intolerable. To overcome this limitation, researchers are exploring the concept of certifiable robustness. The objective is to generate robustness certificates that provably ensure accurate perception outcomes on certain inputs against certain attack capabilities.

Over the past three years, there has been promising progress in certifiable robustness against adversarial patches. For example, state-of-the-art robust image classifier PatchCleanser [36] achieves high certifiable robustness at a minimal cost of clean accuracy (1%) on the large-scale ImageNet [6] dataset. However, major limitations, such as large computational overheads in the magnitude of 10-100×, still remain.

Symposium on Vehicles Security and Privacy (VehicleSec) 2023
27 February 2023, San Diego, CA, USA
ISBN 1-891562-88-6
https://dx.doi.org/10.14722/vehiclesec.2023.23020
www.ndss-symposium.org

Fig. 1. Examples of Adversarial Patch Attacks (left to right: [2], [8], [28])

What is next for this line of research? In this paper, we survey 17 existing certifiably robust defenses, distill their core security techniques, and analyze their strengths and weaknesses (Section III). We then share our thoughts and questions on future research agenda and call for efforts from both academia and industry to robustify AV perception (Section IV).

II. Problem Formulation

In this section, we formulate the adversarial patch attack, the most representative attack against AV perception, and discuss the objective of certifiable robustness.

A. Adversarial Patch Attack

Attack formulation. The adversarial patch attack [2] is a type of evasion attack that aims to induce incorrect predictions when the system is deployed in the wild.

Formally, we define a perception model as a mapping $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ from the image space $\mathcal{X}$ to a predefined output space $\mathcal{Y}$ (e.g., the class label space for image classifiers). We further define an oracle $\mathcal{O} : \mathcal{X} \to \mathcal{Y}$ that maps an image $\mathbf{x} \in \mathcal{X}$ to its desired model prediction (e.g., the correct classification label). An evasion attacker aims to generate an adversarial image $\mathbf{x}'$ s.t. $\mathcal{F}(\mathbf{x}') \neq \mathcal{O}(\mathbf{x})$ while satisfying certain attack constraints $\mathbf{x}' \in \mathcal{A}(\mathbf{x})$. The patch constraint allows the attacker to select a *spatially localized* image region (e.g., a square patch) and arbitrarily corrupt pixels within the selected region. Then, a patch attacker can carry out a physical-world attack by printing and attaching the adversarial patch to a physical scene; images captured from that scene are adversarial $\mathbf{x}' \in \mathcal{A}(\mathbf{x})$.

Attack examples. We provide physical-world attack examples in Fig. 1. In the first two examples, attackers attach stickers to a stop sign to make the model misclassify it as "speed limit 45" [9] or "hide" it from being detected by object detectors [8]. The third example uses a patch to interfere with the semantic segmentation predictions [28].These attack

examples motivate us to study countermeasures to adversarial patch attacks for safe autonomous driving.

B. Certifiable Robustness for Perception Models

The defense objective is intuitive: we want to build a robust model $\mathcal{F}^*$ that can perform correct perceptions against attacks. Formally, we want: $\mathcal{F}^*(\mathbf{x}') = \mathcal{O}(\mathbf{x}), \forall \mathbf{x}' \in \mathcal{A}(\mathbf{x})$.[1]

The biggest challenge comes from the quantifier "$\forall$": we need a certifiable/provable approach to ensure that robustness $\mathcal{F}^*(\mathbf{x}') = \mathcal{O}(\mathbf{x})$ holds for *all* possible attacks within the threat model $\mathcal{A}$ on a given image $\mathbf{x}$. This requirement distinguishes defenses with certifiable robustness from heuristics-based defenses, whose empirical robustness might be compromised by future adaptive attacks. We note that certifiable robustness is essential for critical systems like autonomous vehicles since their unexpected failures are highly intolerable. We envision certifiable robustness as a future standard for the design and manufacturing of AI-powered vehicles.

III. Current State of Research

In this section, we survey certifiably robust defenses to summarize the research progress and unsolved challenges. In Table I, we listed existing defenses and their perception tasks (**Task**), publishing time (**Time**), robustness techniques (**Technique**), robustness performance (**Robustness**), clean performance (**Clean**; i.e., model performance in the absence of attacks), and computational overhead (**Overhead**). We note that all defenses are leveraging one/two of the three major robustness techniques listed below; we will pivot our survey on these robustness techniques.

- **Bound propagation (BP):** estimating and propagating the bounds of neuron activations in each layer to bound the attacker's influence on final predictions.
- **Small receptive fields (SRF):** extracting features from small image regions to bound the number of features corrupted by the spatially localized patch.
- **Masking (M):** masking out adversarial pixels from input images to neutralize malicious effects.

A. Bound Propagation (BP)

Overview. The bound propagation technique was initially proposed for robustness certification against global ℓ_p-bounded perturbations [10], [26], [41]. The idea is to estimate the upper and lower bounds of neuron activation in each layer based on model weights and bounds computed from the previous layer. Once the bounds of activation of the input layer are determined by the perturbation threat model, we can propagate the bounds from the input to the final prediction layer to derive the robustness certificate. For example, if the computed lower bound of true class activation in the last layer of a classification model (on a given image) is larger than the upper bound of any other class, we can certify model robustness for this input image.

Defense. There is only one defense in late 2019 using this robustness technique.

[1]The robustness evaluation metric is the percentage of images $\mathbf{x}$ in a dataset $\mathcal{D} \subset \mathcal{X}$ that satisfy this condition; we aim to optimize it as high as possible.

TABLE I. List of certifiably robust defenses against patches

Task†	Defense*	Time	Technique‡	Robustness$	Clean$	Overhead$
Cls.	Chiang et al. [5]	Oct. 2019	BP	Low	Low	Large
	Clipped BagNet [42]	Jan. 2020	SRF	Low	Low	**Low**
	De-RS [17]	Feb. 2020	SRF	Low	Low	Large
	PatchGuard [35]	May 2020	SRF	Low-Med.	Low	**Low**
	BagCert [25]	Oct. 2020	SRF	Low-Med.	Low	**Low**
	RC [21]	Oct. 2020	SRF	Low	Low	Large
	PatchCleanser [36]	Aug. 2021	M	**High**	**High**	Large
	Smoothed ViT [29]	Oct. 2021	SRF	Med.	Med.	Large
	ECViT [4]	Nov. 2021	SRF	Med.	Med.	Large
	ViP-recovery [19]	Mar. 2022	SRF	Med.	Med.	Large
	*MR-v1 [24]	Apr. 2020	M	Med.	Med.	Large
	*PatchGuard++ [38]	Apr. 2021	SRF+M	Med.	Low	**Low**
	*ScaleCert [11]	May 2021	SRF+M	Med.-**High**	Low	Large
	*MR-v2 [36]	Aug. 2021	M	**High**	Med.	Large
	*PatchVeto [15]	Nov. 2021	M	**High**	Low	Large
	*ViP-detection [19]	Mar. 2022	M	**High**	Med.	Large
Det.	*DetectorGuard [37]	Feb. 2021	SRF	Low	Med.	Med.
	ObjectSeeker [39]	Feb. 2022	M	Med.-**High**	**High**	Large
Seg.	Yatsura et al. [40]	Oct. 2022	SRF/M	Low	Low	Large

† Cls.—Image Classification; Det.—Object detection; Seg—Semantic Segmentation.
* Defenses that only achieve certifiable robustness for attack detection are marked with asterisks (*).
‡ BP—Bound propagation; SRF—Small receptive fields; M—Masking.
$ We host a quantitative leaderboard at https://github.com/inspire-group/patch-defense-leaderboard.

∘ *Chiang et al. [5]* adopted the simplest BP strategy, Interval Bound Propagation (IBP) [10], [26], to build a robust model. Chiang et al. leveraged the computed bounds of the last layer to train a model to be certifiably robust. They achieved the very first certifiable robustness (against patches) in the literature.

Analysis. Despite the contributions of being the first certifiably robust defense against adversarial patches, this IBP-based defense is extremely expensive in its training and cannot scale to high-resolution images like ImageNet [6]. Moreover, the training optimization becomes extremely challenging due to the additional objective of certifiable robustness; trained robust models have poor clean performance.

B. Small Receptive Field (SRF)

Overview. Using small receptive fields (SRF) is currently the most popular design choice. The idea is to let models extract features, or make predictions, on different *small* regions of the input images. SRF ensures that only a limited number of local features/predictions "see" (or, are affected by) the spatially constrained patch. Then, we can perform secure aggregation (e.g., majority voting) on partially corrupted feature/prediction maps for robust predictions.

The certification algorithm depends on the aggregation technique. For example, the condition of robustness certification on a given image for majority voting could be: the count difference between the true-class predictions and any other class predictions is twice larger than the maximum number of predictions that can be affected by the adversarial patch.

Defenses. The idea of SRF was first explored in three concurrent works in early 2020.

∘ *Clipped BagNet (CBN) [42]* adopted BagNet [1] for certifiable robustness. BagNet achieves SRF by using small convolution kernel sizes and strides and was initially proposed for interpretable machine learning. The CBN defense uses BagNet as the backbone for feature extraction and clips extracted features for certifiable secure aggregation.

∘ *De-Randomized Smoothing (De-RS) [17]* adapted the ideas of randomized smoothing against ℓ_0 global perturbations [18]. It creates *ablated* images that only contain pixels from part

of the input image (e.g., image pixel bands). The model then makes a prediction on each of the ablated images, and it uses majority voting to generate the final *smoothed* prediction.

∘ *PatchGuard [35]* was the first to explicitly discuss the idea of SRF. It started with an empirical analysis to demonstrate that conventional CNNs' vulnerability to adversarial patches comes from their large receptive fields (e.g., 483×483 for ResNet-50 [14]) and insecure feature aggregation (e.g., average pooling). Then, PatchGuard systematically discussed the defense framework of using SRF and secure aggregation. For its implementation, PatchGuard used similar techniques used in BagNet [1] and De-RS [17] to have SRF, and proposed a *robust masking* technique for better secure aggregation

These three early works have inspired a large number of defenses to use SRF as their robustness building block.

∘ *BagCert [25]* designed a customized BagNet-like [1] architecture, which directly integrates the secure aggregation operation into the end-to-end classification model and enables end-to-end "certifiably adversarial" training.

∘ *Randomized Cropping (RC) [21]* proposed to perform majority voting on predictions made on randomly cropped images.

∘ *PatchGuard++ [38]* combined the SRF idea from PatchGuard [35] and the masking ideas from the Minority Reports (MR) defense [24] (will be discussed in the next section) to build an efficient feature-space *attack-detection* defense.

∘ *ScaleCert [11]* used SRF of "superficial neurons" to improve the pixel masking strategy from MR [24] for attack detection.

∘ *Smoothed ViT [29], ECViT [4], and ViP [19]* leveraged the advancement in Vision Transformer (ViT) research [7], [13], [32], and significantly improved the performance of De-RS [17], in terms of performance, training, and efficiency.

The idea of SRF is not only used for image classification but also for object detection and semantic segmentation.

∘ *DetectorGuard [37]*, the first defense for object detection against patch-hiding attacks, took PatchGuard [35] models as its core robustness module.

∘ *Yatsura et al. [40]* recently proposed the first certifiably robust semantic segmentation model via SRF (leveraging De-RS-like [17] architectures).

Analysis. Despite its popularity, SRF introduces a fundamental (and seemingly unresolvable) conflict between robustness and clean performance. While SRF limits the number of corrupted features for robustness, it also limits the useful information received by each feature/prediction. As a result, all SRF defense models only have limited clean performance. Take image classification on ImageNet [6] as an example. State-of-the-art undefended classification models usually have an accuracy of 80+% [7], [16], while the SRF defenses only have accuracy around 50-70% (marked as low or Med. in Table I). A large drop in clean performance (e.g., 10+%) is problematic. First, this implies that models are already partially broken *even without an attack*, e.g., an AV makes unsafe moves in non-adversarial scenarios. Second, clean performance is the upper bound of certifiable robustness; thus, low clean performance also limits the achievable robustness.

Next, we discuss the overhead of SRF defenses. From Table I, some SRF defenses [35], [38], [42] have a small overhead (∼1× compared to undefended models; marked as Low), as they use BagNet-like [1] architecture to have SRF and only require one expensive model feed-forward. On the other hand, SRF defenses with large overheads use De-RS-like [17] architecture for SRF: they need multiple expensive feed-forward predictions on different "ablated images" (usually dozens or hundreds). Moreover, we note that recent papers [4], [19], [29] that improve the robustness and clean performance of SRF defenses are all using De-RS-like architectures and have large computational overheads.

C. Masking (M)

Overview. The idea of masking is intuitive: we want to mask out all adversarial pixels from the input image so that vanilla image classifiers could make safe predictions on masked images. The main challenges are how to apply the masks without knowing patch locations, and more importantly, how to achieve robustness certification.

Defenses. The masking idea was first studied under the *attack detection* problem.

∘ *The Minority Reports (MR) defense [24]* is the first masking-based certifiably robust defense. MR first conservatively estimates the patch size and selects a mask that is large enough to cover the entire patch. Then, MR moves the mask across all possible image locations and evaluates model predictions on all possible masked images. This exhaustive masking operation ensures that there exist masked images without any adversarial pixels, and model predictions on these masked images are likely to be correct. Then, MR predicts as follows: if all masked predictions reach "certain agreements", MR considers it a clean image and outputs the agreed label; otherwise, MR issues an attack alert. The robustness certification checks if all masked predictions on the clean image "agree" on the correct labels. This condition ensures that any malicious masked prediction label on the adversarial image will trigger an alert.

This agreement-checking masking strategy for attack detection has been adapted and improved by PatchGuard++ [38] (more efficient feature space masking), *ScaleCert [11]* (more efficient superficial neuron analysis), *MR-v2 [36]* (better backbone and mask design), *PatchVeto [15] and ViP [19]* (using ViT [7], [13] as better backbones), and *Yatsura et al. [40]* (generalizing it to the semantic segmentation task).

Despite the success of MR-style defenses, robustness for attack detection does not prevent attackers from forcing the model to always alert and abstain from making any prediction. A natural question is: Can we use masking for prediction recovery without any abstention? The answer is yes!

∘ *PatchCleanser [36]* proposed a solution to recover correct predictions without any abstentions. PatchCleanser first uses MR-style masking to identify an attack (denoted as MR-v2 in Table I). If PatchCleanser detects a prediction disagreement, it has to decide which prediction to trust (rather than alert and abstain). To do this, PatchCleanser applies a second round of masks to images with *one* mask and evaluates model predictions on images with *two* masks. The intuition is that, if the first-round mask already removes the entire patch, the second-round mask is applied to a "clean" image, and two-mask

predictions are likely to reach an agreement. On the other hand, if the first-round mask does not remove the patch, the second-round mask is applied to an "adversarial image", and two-mask predictions can still have a disagreement. PatchCleanser then uses the agreement/disagreement in two-mask predictions to identify the first-round mask that removes the patch and outputs its corresponding masked prediction. The certification condition is whether model predictions on clean images with all possible two-mask combinations are all correct.

Finally, the masking idea has also been used for object detection, which is a harder but more important task for AV.

○ *ObjectSeeker [39]* adapted the masking idea for certifiable robustness (without abstentions) against patch hiding attacks, which hide victim objects from being detected. Intriguingly, unlike other masking-based defenses [11], [19], [24], [36], [38], [40], ObjectSeeker does not need the knowledge of patch shapes and sizes to generate masks that can remove the entire patch. This is the first "patch-agnostic" masking-based defense with certifiable robustness.

Analysis. Masking has two major advantages. First, the defense is compatible with different computer vision models, since the masking operation takes place on the input image. Second, we are free from the limitation of SRF, and thus have a chance to build a defense model with high clean performance (e.g., PatchCleanser [36] and ObjectSeeker [39] achieve $\sim$1% clean performance drops from undefended models).

The biggest downside of masking-based defenses is their large computational overheads, as they need to evaluate model predictions on multiple masked images (at least dozens). Moreover, we note that the clean performance of masking-based *attack detection* defense is also limited (marked as Low or Med. in the table); this is because they could issue many false alerts on clean images.[2]

D. Progress and Limitations

Progress. We have seen tremendous progress made by the research community. First, most defenses (except Chiang et al. [5] and MR-v1 [24]) can scale to large perception models (e.g., ResNet-50 [14], ViT [7]) and large high-resolution datasets (e.g., ImageNet [6], COCO [20]), which is in huge contrast to certifiable robustness research for global ℓ_p-norm perturbations. Second, certifiable robustness and clean performance have been significantly improved. Notably, PatchCleanser [36] and ObjectSeeker [39] achieve high clean performance (less than 1% drops from undefended models), as well as state-of-the-art certifiable robustness. These promising outcomes encourage us to further explore algorithms that are deployable in practice.

Limitations. Nevertheless, there are still significant limitations left for future research. First, the overhead of high-performance defenses is still large (e.g., 10-100$\times$ more computations compared to undefended models). This limitation might prevent their deployments in real-time critical systems like autonomous vehicles. Second, all 17 different defenses can be categorized into only three major robustness techniques. The field might need a new robustness technique to lead the next breakthrough. Third, most defenses only study the simpler task of image classification. However, the perception systems of autonomous vehicles have more complex tasks such as object detection, object tracking, semantic segmentation, and visual localization, as well as multi-modal inputs like 3D point clouds. Fourth, most defenses focus on certification for one *single* patch and small patch sizes (e.g., occupying 2% of image pixels).[3] However, a powerful attacker in the physical world might place *multiple* adversarial patches, and the patch size could increase as the vehicle moves toward the object.

Based on the progress and limitations discussed above, we will next share our thoughts and vision for future research.

IV. Open Questions and Future Opportunities

How to improve the three-way trade-off between robustness, clean performance, and computation overhead? As discussed in Section III, the robustness techniques have evolved from bound propagation, to small receptive fields, and to masking. However, each technique has fundamental limitations that result in a challenging three-way trade-off between robustness, clean performance, and computation overhead. For example, the use of SRF hurts the clean performance and limits the achievable robustness. The masking strategy requires expensive computation to get model predictions on different masked images. *What is the next major innovation to improve this trade-off?* For example, given that image-space masking operations are computationally expensive, can we design an efficient method to compute "approximated" masked predictions? PatchGuard [35] argues that large receptive fields hurt robustness, while empirically, the "effective" receptive fields are much smaller than the "theoretical" ones [23]. Can we develop robustness certification based on this observation?

How to generalize the certifiable robustness of perception models to end-to-end AI systems? The autonomous vehicle is a complex AI system that consists of different modules such as sensing, perception, and planning while existing research only derives robustness certification for a part of perception models [31]. *Is it possible to fill this gap and derive certification for the end-to-end autonomous driving system?* Is the system-level certification easier due to an ensemble of robust modules (e.g., turning to depth information when certification fails using RGB images), or harder (due to higher system complexity)? We might need to generalize defenses to other perception tasks (e.g., object tracking) and data modalities (e.g., "patch" for 3D point clouds [3], [33]) and also study the interaction among different modules like perception, sensing, and planning.

What is the opportunity for academia and industry to collaborate and evaluate defenses in real deployments? Current research on certifiable robustness mostly takes place in academia, though the motivation of this line of research largely comes from the industrial need for safe autonomous driving systems. Given recent progress made in the literature, we believe it is a good time for academia and industry to start collaborating on transitioning proposed algorithms to real

[2] As of Dec. 2022, the clean performance reported by PatchVeto [15], ViP [19], and Yatsura et al. [40] is flawed as they do not count false alerts on clean images as errors.

[3] Only a limited number of works experimentally demonstrated non-trivial certifiable robustness against two patches [19], [36], [39] or patches that corrupt more than 10% of image pixels [11], [15], [35], [36], [39].

deployments. Evaluating algorithms on real systems allow both parties to better understand the performance gap to fill: *for example, how much clean performance drop is acceptable? How much latency and computation overhead are tolerable?* This collaboration will serve as an opportunity to push certifiable robustness as a future standard for safe AI-power autonomous vehicles.

V. Conclusion

In this paper, we surveyed certifiably robust techniques against adversarial patches. We summarized three core robustness techniques (BP, SRF, and M) used in existing defenses. We highlighted the promising research progress and unsolved limitations, and shared our vision for future research. Finally, we call for collaboration between industry and academia on transitioning research papers to real-world systems (e.g., optimizing algorithms for real-world constraints and developing end-to-end AV security certification).

References

[1] W. Brendel and M. Bethge, "Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet," in *ICLR*, 2019.

[2] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," in *NeurIPS Workshops*, 2017.

[3] Y. Cao, N. Wang, C. Xiao, D. Yang, J. Fang, R. Yang, Q. A. Chen, M. Liu, and B. Li, "Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks," in *IEEE S&P*, 2021.

[4] Z. Chen, B. Li, J. Xu, S. Wu, S. Ding, and W. Zhang, "Towards practical certifiable patch defense with vision transformer," in *CVPR*, 2022.

[5] P.-Y. Chiang, R. Ni, A. Abdelkader, C. Zhu, C. Studor, and T. Goldstein, "Certified defenses for adversarial patches," in *ICLR*, 2020.

[6] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[8] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, T. Kohno, and D. Song, "Physical adversarial examples for object detectors," in *WOOT*, 2018.

[9] 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 *CVPR*, 2018.

[10] S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. A. Mann, and P. Kohli, "Scalable verified training for provably robust image classification," in *ICCV*, 2019.

[11] H. Han, K. Xu, X. Hu, X. Chen, L. Liang, Z. Du, Q. Guo, Y. Wang, and Y. Chen, "ScaleCert: Scalable certified defense against adversarial patches with sparse superficial layers," in *NeurIPS*, 2021.

[12] J. Hayes, "On visible adversarial perturbations & digital watermarking," in *CVPR Workshops*, 2018.

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[15] Y. Huang and Y. Li, "Zero-shot certified defense against adversarial patches with vision transformers," *arXiv preprint arXiv:2111.10481*, 2021.

[16] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *ECCV*. Springer, 2020.

[17] A. Levine and S. Feizi, "(De)randomized smoothing for certifiable defense against patch attacks," in *NeurIPS*, 2020.

[18] ——, "Robustness certificates for sparse adversarial attacks by randomized ablation," in *AAAI*, 2020.

[19] J. Li, H. Zhang, and C. Xie, "Vip: Unified certified detection and recovery for patch attack with vision transformers," in *ECCV*, 2022.

[20] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *ECCV*, 2014.

[21] W.-Y. Lin, F. Sheikholeslami, L. Rice, J. Z. Kolter *et al.*, "Certified robustness against physically-realizable patch attack via randomized cropping," 2020.

[22] J. Liu, A. Levine, C. P. Lau, R. Chellappa, and S. Feizi, "Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection," in *CVPR*, 2022.

[23] W. Luo, Y. Li, R. Urtasun, and R. S. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *NeurIPS*, 2016.

[24] M. McCoyd, W. Park, S. Chen, N. Shah, R. Roggenkemper, M. Hwang, J. X. Liu, and D. A. Wagner, "Minority reports defense: Defending against adversarial patches," in *ACNS Workshops*, 2020.

[25] J. H. Metzen and M. Yatsura, "Efficient certified defenses against patch attacks on image classifiers," in *ICLR*, 2021.

[26] M. Mirman, T. Gehr, and M. T. Vechev, "Differentiable abstract interpretation for provably robust neural networks," in *ICML*, 2018.

[27] M. Naseer, S. Khan, and F. Porikli, "Local gradients smoothing: Defense against localized adversarial attacks," in *WACV*, 2019.

[28] F. Nesti, G. Rossolini, S. Nair, A. Biondi, and G. Buttazzo, "Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks," in *WACV*, 2022.

[29] H. Salman, S. Jain, E. Wong, and A. Madry, "Certified patch robustness via smoothed vision transformers," in *CVPR*, 2022.

[30] T. Sato, J. Shen, N. Wang, Y. Jia, X. Lin, and Q. A. Chen, "Dirty road can attack: Security of deep learning based automated lane centering under {Physical-World} attack," in *USENIX Security*, 2021.

[31] J. Shen, N. Wang, Z. Wan, Y. Luo, T. Sato, Z. Hu, X. Zhang, S. Guo, Z. Zhong, K. Li *et al.*, "Sok: On the semantic ai security in autonomous driving," *arXiv preprint arXiv:2203.05314*, 2022.

[32] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.

[33] J. Tu, H. Li, X. Yan, M. Ren, Y. Chen, M. Liang, E. Bitar, E. Yumer, and R. Urtasun, "Exploring adversarial robustness of multi-sensor perception systems in self driving," in *CoRL*, ser. Proceedings of Machine Learning Research, 2021.

[34] T. Wu, L. Tong, and Y. Vorobeychik, "Defending against physically realizable attacks on image classification," in *ICLR*, 2020.

[35] C. Xiang, A. N. Bhagoji, V. Sehwag, and P. Mittal, "Patchguard: A provably robust defense against adversarial patches via small receptive fields and masking," in *USENIX Security*, 2021.

[36] C. Xiang, S. Mahloujifar, and P. Mittal, "Patchcleanser: Certifiably robust defense against adversarial patches for any image classifier," in *USENIX Security*, 2022.

[37] C. Xiang and P. Mittal, "DetectorGuard: Provably securing object detectors against localized patch hiding attacks," in *ACM CCS*, 2021.

[38] ——, "Patchguard++: Efficient provable attack detection against adversarial patches," in *ICLR Workshops*, 2021.

[39] C. Xiang, A. Valtchanov, S. Mahloujifar, and P. Mittal, "Objectseeker: Certifiably robust object detection against patch hiding attacks via patch-agnostic masking," in *IEEE S&P*, 2023.

[40] M. Yatsura, K. Sakmann, N. G. Hua, M. Hein, and J. H. Metzen, "Certified defences against adversarial patch attacks on semantic segmentation," *arXiv preprint arXiv:2209.05980*, 2022.

[41] H. Zhang, T.-W. Weng, P.-Y. Chen, C.-J. Hsieh, and L. Daniel, "Efficient neural network robustness certification with general activation functions," in *NeurIPS*, 2018.

[42] Z. Zhang, B. Yuan, M. McCoyd, and D. Wagner, "Clipped bagnet: Defending against sticker attacks with clipped bag-of-features," in *Deep Learning and Security Workshop*, 2020.