Demo Abstract: Parameterized Stochastic Ensemble Defense for Object Detection

Yuting Wu[†] , Dongfang Guo[†] , Xiangzhong Luo[†] , Qun Song[††], Rui Tan[†]
[†] Nanyang Technological University, Singapore
[††] Singapore University of Technology and Design, Singapore

Abstract

Camera-based object detection excels but remains vulnerable to adversarial attacks that suppress target detection (object-hiding attacks). Here, we propose *PaSED*, a **Pa**rameterized **S**tochastic **E**nsemble **D**efense, which leverages HyperNetworks to enable rapid and diverse updates for detection models in the ensemble. At its core, we introduce functional diversity to enhance the defense robustness. It adapts each generation process to the input image preprocessing parameterized by HyperNetworks' random noise input. In our preliminary evaluations against physically deployed attacks, *PaSED* outperforms five baseline defenses without requiring attack knowledge. It recovers attacked objects in 92% and 98% of frames in the indoor and outdoor testbeds, respectively.

CCS Concepts

• **Security and privacy** → *Systems security*; • **Computer systems organization** → **Dependable and fault-tolerant systems and networks**.

Keywords

Stochastic Ensemble Defense, Adversarial Attack, Object Detection,

ACM Reference Format:
Yuting Wu[†] , Dongfang Guo[†] , Xiangzhong Luo[†] , Qun Song[††], Rui Tan[†] . 2018. Demo Abstract: Parameterized Stochastic Ensemble Defense for Object Detection. In . ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Camera-based object detection powered by deep neural networks excels in accuracy but remains susceptible to adversarial examples, i.e., carefully crafted perturbations that exploit model vulnerabilities to induce errors. Recent studies demonstrate that such attacks can be realized in the physical world by mounting adversarial patches or LCD screens on objects, posing severe risks in safety-critical applications like autonomous driving. This highlights the urgent need for robust defenses in object detection.

To address such vulnerabilities, researchers have proposed various defense approaches. Among these, adversarial training is widely cited as the most effective. It hardens models by training them on adversarially perturbed, yet correctly labeled, samples. However, its effectiveness diminishes when confronted with attack types not represented in the training data. Another approach involves employing input transformations—such as lossy compression [2, 5], blurring [6], and selective masking [3, 7–10]—to neutralize adversarial perturbations. However, these heuristic solutions are typically static post-deployment, which creates a practical limitation. Attackers can reverse-engineer and adapt their strategies to bypass such defenses, a vulnerability known as *adaptive attacks*.

To address the above limitations, we propose *PaSED*, a parameterized stochastic ensemble defense. *PaSED* uses HyperNetworks, deep neural networks that generate the weights of the target network based on random noise input, to create stochastic ensembles at runtime, ideally on a per-input basis. Such dynamics, if not fully predictable by the attackers, alter the attack surface constantly and form a *moving target effect* for improved security against adaptive attackers.

Ensemble defense relies on diverse model responses to adversarial examples. To enhance robustness, *PaSED* promotes functional diversity along with weight diversity between ensemble members. Specifically, we reuse the random noise input of HyperNetworks for each generation to parameterize the preprocessing transform applied to the input image during training and testing. It drives HyperNetworks to generate diverse networks, each for a unique data distribution mapped from the original distribution.

2 Design Overview

Our defense is based on the state-of-the-art detection model (YOLO) and uses a clean image dataset (Microsoft COCO) for training.

HyperNetworks. We leverage HyperNetworks to construct and update an ensemble of detector model variants. We assign each selected layer in the base detector a dedicated MLP-based generator. These generators take random noise sampled from a normal distribution and produce the corresponding weight parameters.

Weight diversity. To ensure diversity among detection models in the ensemble, a weight diversity loss quantifies differences in generated weights. We compute a diversity score for each layer in the ensemble models by averaging the absolute Pearson correlation between the flattened weight vectors of every unique pair of models. This score is then squared and aggregated across all layers. The total training loss balances two objectives: (1) the average detection accuracy of all models in the ensemble, and (2) the overall diversity score, controlled by a tunable coefficient β.

Functional diversity. To further enhance the robustness, we introduce additional functional diversity during training and testing. *PaSED* reuses the HyperNetworks' random noise input as a parameter to transform the original image. Among various possible transforms, in this work, we choose the block-wise pixel shuffling [1, 4] due to its operation efficiency and promising performance in classification tasks. Specifically, it divides the original image into non-overlapping square blocks, each sized $P \times P$ pixels. Within each block, the pixel values are shuffled independently across the red, green, and blue color channels, where HyperNetworks' input vector determines their distinct permutation orders. Fig. 1 examples

Conference'17, July 2017, Washington, DC, USA
2018. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

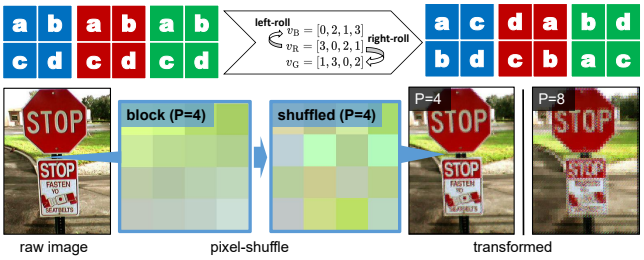


Figure 1: Top: Example of permutation process for $P = 2$. Bottom: Effects of pixel shuffling for $P = 4$ and $P = 8$.

(a) Indoor **(b) Outdoor**

Figure 2: Preliminary physical evaluation testbeds.

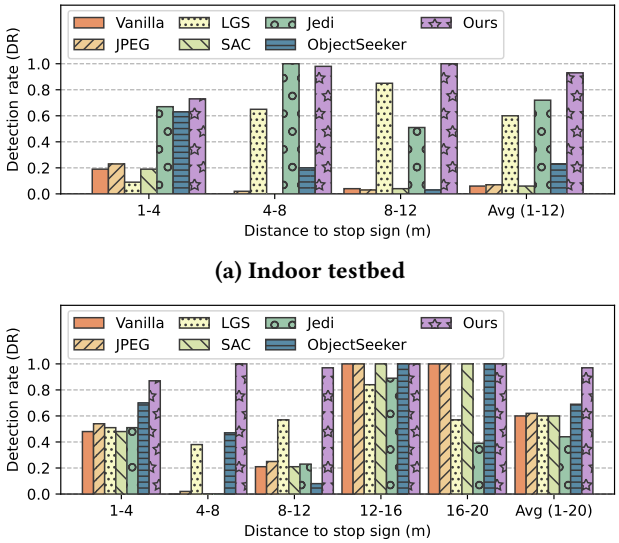
the permutation process for $P = 2$ and the effects of block-wise pixel shuffling for $P = 4$ and $P = 8$.

Detection Fusion Strategy. We observe that the diverse responses from ensemble members can be potentially used to distinguish attacked objects from normal false positives. To ensure clean accuracy while preserving defense effectiveness, *PaSED*'s basic idea is to use the ensemble to recover the objects hidden by the attacks from the base detector and then selectively merge them with the base detector's results.

3 Preliminary Evaluation

To evaluate the robustness of *PaSED*, we provide preliminary physical world experiments in both indoor and outdoor testbeds. We choose the 'stop sign' as a representative attack target where the printed adversarial patches are mounted to hide it from detection. We compare NoHide with six baseline methods: (1) *Vanilla* model without defense mechanism, (2) *JPEG* [2], (3) *LGS* [5], (4) *SAC* [3], (5) *Jedi* [6], (6) *ObjectSeeker* [9]. We use *Detection Rate (DR)* to measure the percentage of frames where the target object is successfully detected. A higher DR indicates better model robustness against physical adversaries. As shown in Fig. 3, *PaSED* demonstrates superior robustness compared to established baselines [2, 3, 5, 6, 9]. Specifically, NoHide achieves a high DR, reaching 0.92 in laboratory settings and 0.98 in real-world road environments. Furthermore, it delivers stable performance across diverse distance ranges and test environments with varying lighting conditions.

Live Demo. In this demonstration, we will showcase *PaSED* in action, defending against physically deployed adversarial attacks targeting stop signs. We will visualize the rapid and diverse generation of detection models, and highlight their functional diversity by showing parameterized image transformations for each generated model and their diverse detection results. These detection results will form the final output.



(a) Indoor testbed

(b) Outdoor testbed

Figure 3: DR of defense methods against physical adversarial attack targeting stop sign in two testbeds.

References

[1] MaungMaung AprilPyone and Hitoshi Kiya. 2021. Block-wise image transformation with secret key for adversarially robust defense. *IEEE Transactions on Information Forensics and Security* 16 (2021), 2709–2723.

[2] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. 2017. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117* (2017).

[3] Jiang Liu, Alexander Levine, Chun Pong Lau, Rama Chellappa, and Soheil Feizi. 2022. Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14973–14982.

[4] AprilPyone MaungMaung, Isao Echizen, and Hitoshi Kiya. 2023. Efficient Key-Based Adversarial Defense for ImageNet by Using Pre-trained Model. *arXiv preprint arXiv:2311.16577* (2023).

[5] Muzammal Naseer, Salman Khan, and Fatih Porikli. 2019. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1300–1307.

[6] Bilel Tarchoun, Anouar Ben Khalifa, Mohamed Ali Mahjoub, Nael Abu-Ghazaleh, and Ihsen Alouani. 2023. Jedi: Entropy-based localization and removal of adversarial patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4087–4095.

[7] Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. 2021. {PatchGuard}: A provably robust defense against adversarial patches via small receptive fields and masking. In *30th USENIX Security Symposium (USENIX Security 21)*. 2237–2254.

[8] Chong Xiang, Saeed Mahloujifar, and Prateek Mittal. 2022. {PatchCleanser}: Certifiably robust defense against adversarial patches for any image classifier. In *31st USENIX Security Symposium (USENIX Security 22)*. 2065–2082.

[9] Chong Xiang, Alexander Valtchanov, Saeed Mahloujifar, and Prateek Mittal. 2023. ObjectSeeker: Certifiably robust object detection against patch hiding attacks via patch-agnostic masking. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 1329–1347.

[10] Cheng Yu, Jiansheng Chen, Youze Xue, Yuyang Liu, Weitao Wan, Jiayu Bao, and Huimin Ma. 2021. Defending against universal adversarial patches by clipping feature norms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16434–16442.