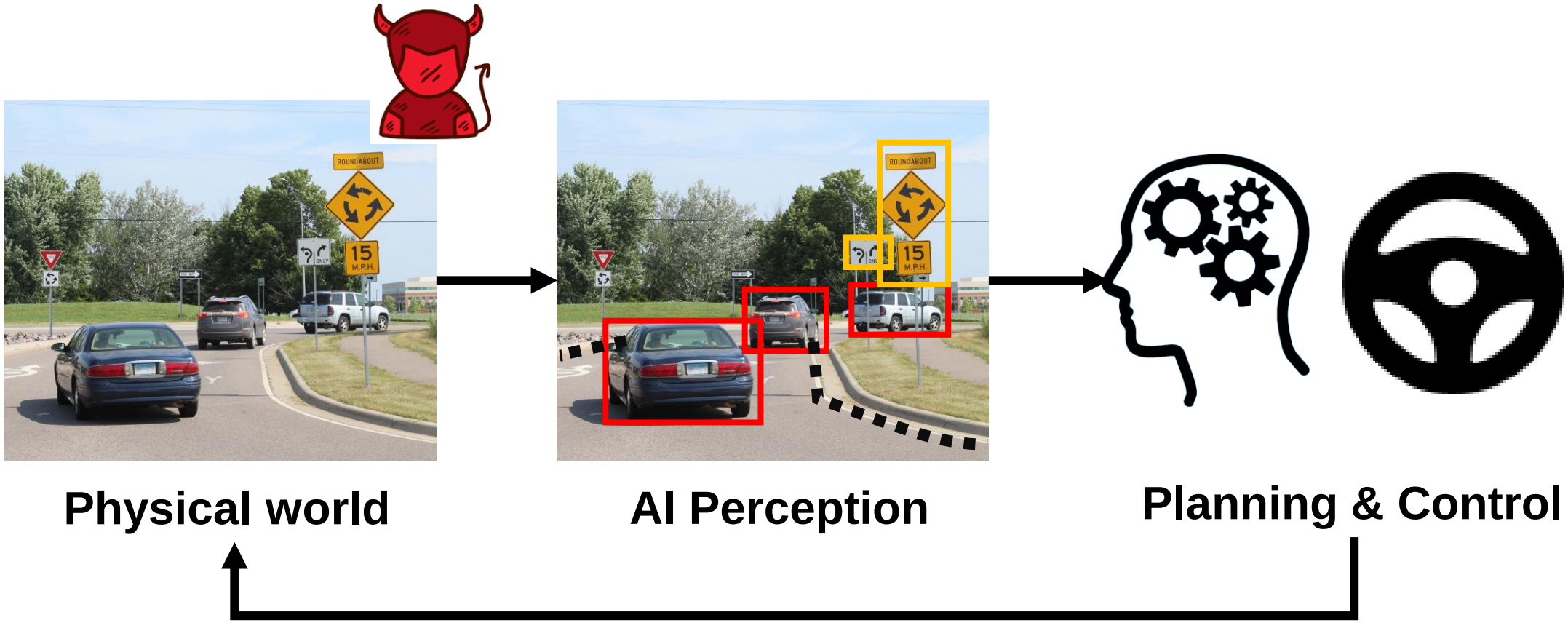
Certifiably Robust Perception Against Adversarial Patch Attacks: A Survey

Chong Xiang[1], Chawin Sitawarin[2], Tong Wu[1], Prateek Mittal[1]

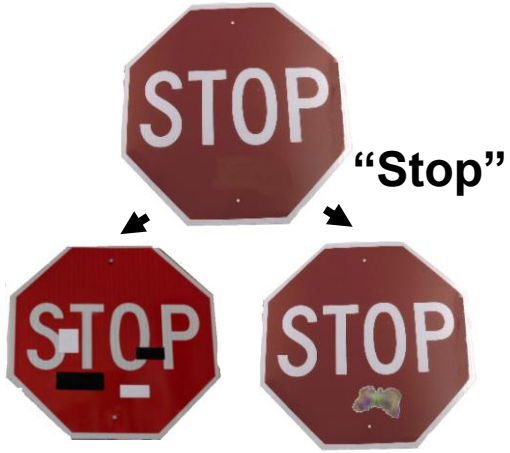
[1]Princeton University, [2]University of California, Berkeley

Safe Autonomous Driving Relies on Robust AI Perception



Localized Adversarial Patch Attacks in the Physical World

- Control pixel values within a localized image region (i.e., a patch)
  - Corrupt part of the physical world (not the entire one)



“Speed Limit 80”

Image Classification

Label misclassification

[Eykholt et al. CVPR 2018]

[Yakura et al. AAAI 2019]

Object Detection

Fail to detect the stop sign

[Zhao et al. CCS 2019]

Semantic Segmentation

Incorrect segmentation for car

[Nesti et al. WACV 2022]

Lane Detection

Lane deviated to the left

[Sato et al. USENIX Security 2021]

Survey on Certifiably Robust Defenses against Patches

Image Classification

- [Chiang et al. ICLR 2020]
- Minority Reports [ACNS W. 2020]
- Clipped BagNet [DLS 2020]
- De-randomized Smoothing [NeurIPS 2020]
- PatchGuard [USENIX Security 2021]
- ScaleCert [NeurIPS 2021]
- BagCert [ICLR 2021]
- Randomized Cropping [2021]
- PatchGuard++ [ICLR W. 2021]
- PatchCleanser [USENIX Security 2022]
- PatchVeto [arXiv 2022]
- Smoothed ViT [CVPR 2022]
- ECViT [CVPR 2022]
- ViP [ECCV 2022]

Object Detection

- DetectorGuard [CCS 2021]
- ObjectSeeker [S&P 2023]

Semantic Segmentation

- Yatsura et al. [arXiv 2022]

- **Certifiable robustness:** formally prove/certify the robustness against any white-box adaptive attack within a given threat model
- **17 defenses** proposed for different tasks over the past three years
- **Survey question:** What are the major research progress made and next research step?

Survey Takeaways

Come to our poster to learn more!
↙

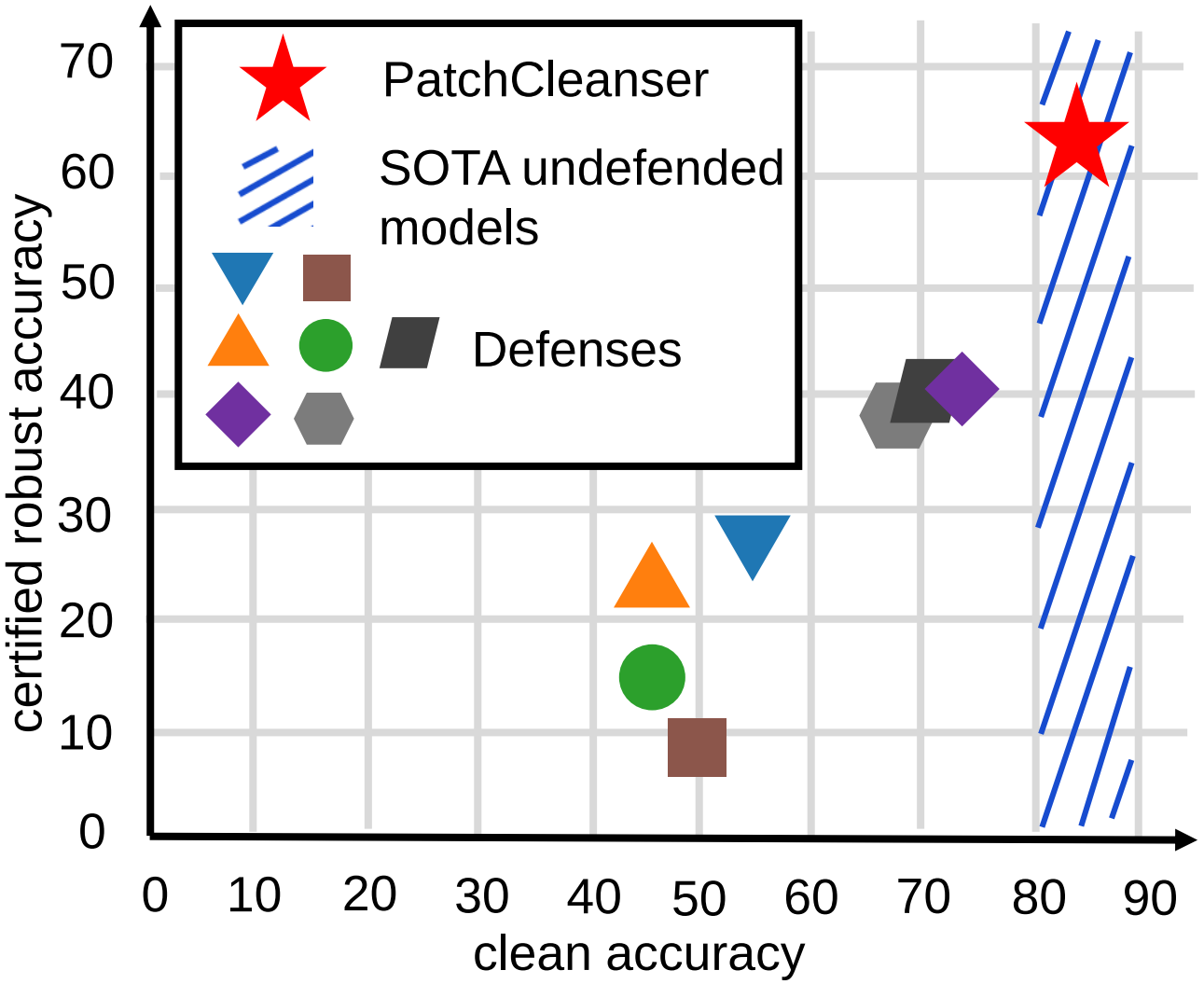
- **Technique**
  - 17 defenses are using **3 core robustness techniques**
- **Progress**
  - Certifiable robustness with **a minimal cost of model accuracy drop**
- **Limitation**
  - Large computation overheads (10-100x)

Survey Takeaways

Come to our poster to learn more!
↙

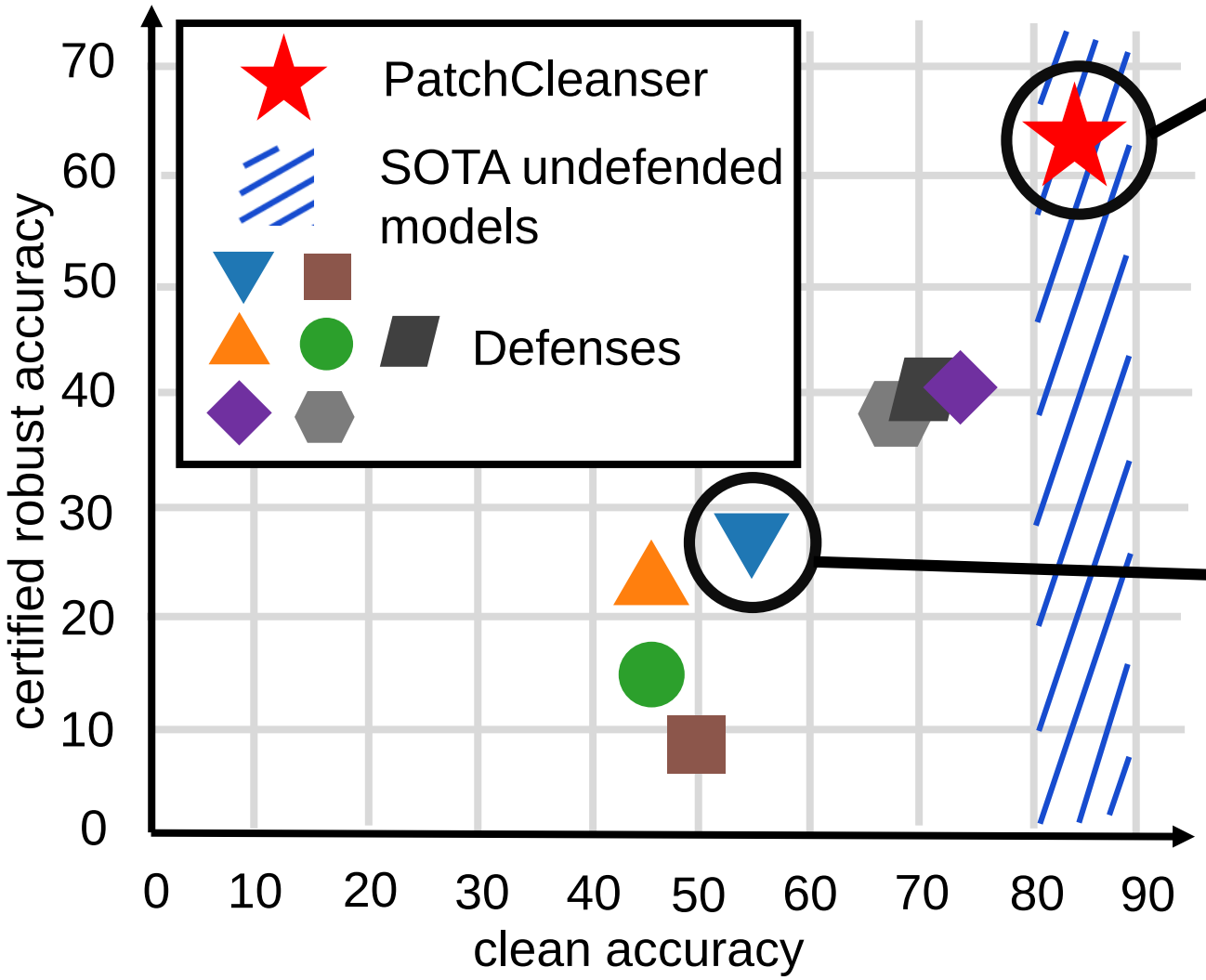
- **Technique**
  - 17 defenses are using **3 core robustness techniques**
- **Progress**
  - Certifiable robustness with **a minimal cost of model accuracy drop**
- **Limitation**
  - Large computation overheads (10-100x)

ImageNet Evaluation: High Certifiable Robustness with Small Cost of Clean Accuracy



- PatchCleanser
  - SOTA robustness
  - Comparable clean accuracy to SOTA undefended models
- The first certifiably robust defense maintains accuracy drop within 1% (instead of 10+% drops)

Cost of High Certifiable Robustness: Computation Overhead



PatchCleanser

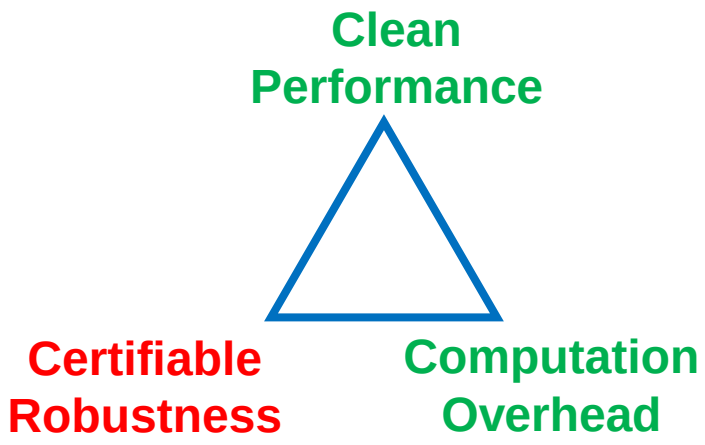
- Requires multiple model predictions on different masked images (~40x)

PatchGuard

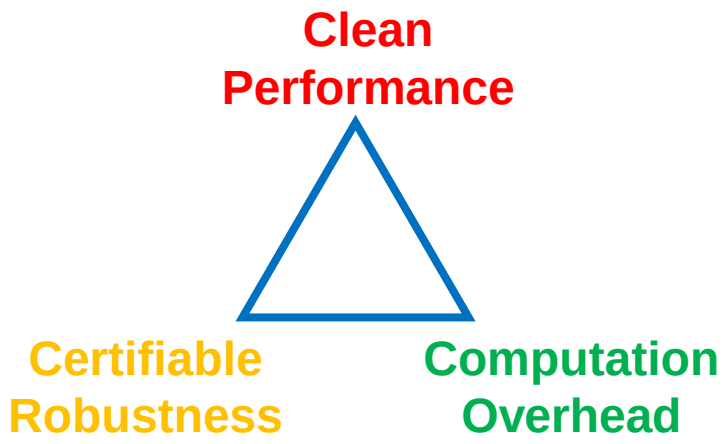
- Requires only one model feed-forward/prediction
- But much worse performance

Three-way Trade-off: Clean Performance vs. Certifiable Robustness vs. Computation Overhead

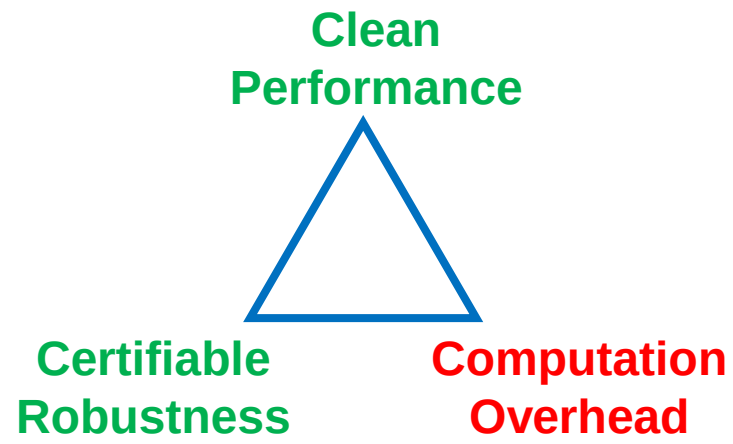
Undefended models
1. Good clean performance
2. Zero robustness
3. Good computation overhead



PatchGuard
1. Poor clean performance (20+% drops)
2. Fair certifiable robustness
3. Good computation overhead (~1x)



PatchCleanser: SOTA defenses
1. Good clean performance (1% drops)
2. Good certifiable robustness
3. Poor computation overhead (~40x)



Research question: How can we further mitigate this three-way trade-off?

Questions for Industrial Practitioners

- Is there any opportunity to evaluate defenses on real systems?
  - SOTA: small clean performance drop on benchmark datasets
  - Unknown: what are the computation constraints we should optimize for?

- What is the system-level implication of robustness certification of AI perception
  - AI perception is a submodule of the entire pipeline
  - Is it possible to certify robustness for end-to-end AI systems?

Survey Takeaways

Come to our poster to discuss more!

- **Technique**
  - 17 defenses are using **3 core robustness techniques**
- **Progress**
  - Certifiable robustness with **a minimal cost of model accuracy drop**
- **Limitation**
  - Large computation overheads (10-100x)
- **Question**
  - Transition to real systems?

Paper list

Leaderboard