GRAPHITE: Generating Automatic Physical Examples for Machine-Learning Attacks on Computer Vision Systems

Ryan Feng[1], Neal Mangaokar[1], Jiefeng Chen[2], Earlence Fernandes[2], Somesh Jha[2], Atul Prakash[1]

[1]University of Michigan, [2]University of Wisconsin

Euro S&P 2022

Ryan Feng
rtfeng@umich.edu

Robust Physical Perturbation Attacks

- Physical attacks such as RP_2 [1] enable sticker attacks on physical objects

- **Key idea**: physical attacks are more *practical*
  - Easier to attack a real system, harder to defend

- **Limitations**: current methods still require
  - Manual mask experimentation
  - White-box access to model weights / architecture

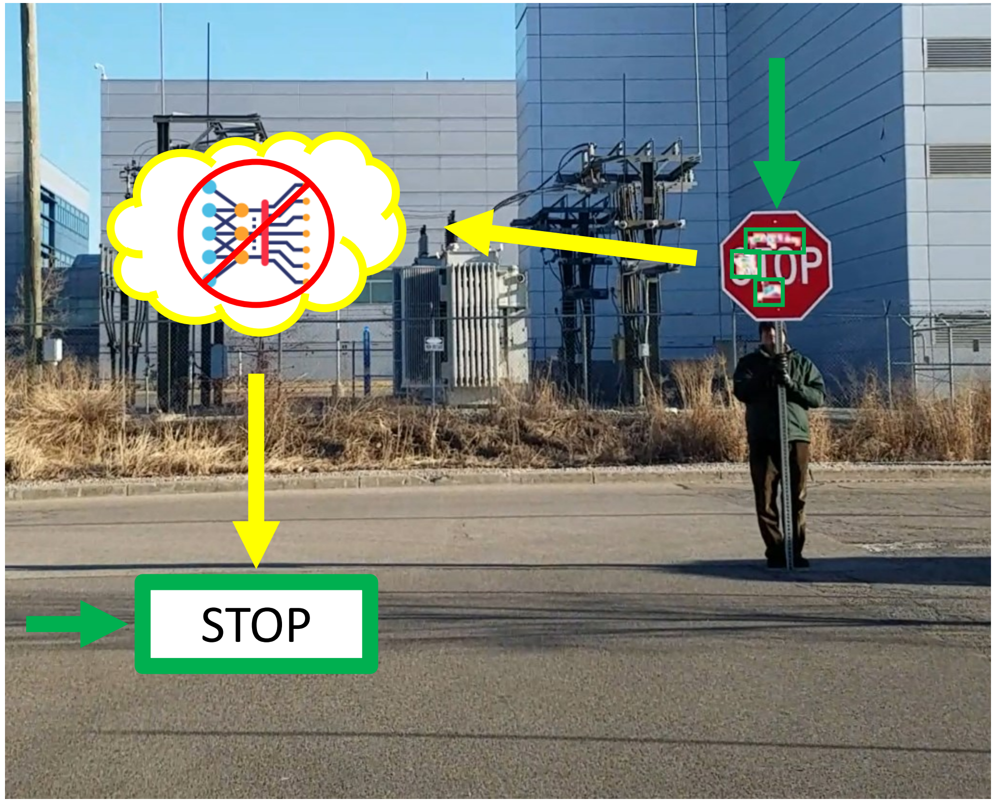
[1] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, “Robust Physical-World Attacks on Deep Learning Models,” CVPR 2018.

UNIVERSITY OF MICHIGAN
Ryan Feng
rtfeng@umich.edu

Motivation: A Framework for Practical Attacks

- **Goal: Generate Practical Attacks**
  - *Automatically* generate masks
  - Apply attacks as *physical* stickers
  - Can work with just *hard-label* access

- Automatic attack generation tools can assist with adversarial testing and defense design



Ryan Feng
rtfeng@umich.edu

GRAPHITE Framework

$$\underset{\delta, M}{\operatorname{argmin}}\ \underbrace{\lambda \cdot ||M||_0}_{\text{Small mask size}} - \underbrace{\mathbb{E}_{t \sim T}\left[F\left(t(x + M \cdot \delta)\right) = y_{tar}\right]}_{\text{High transform-robustness}}$$

x: Input image
y_{tar}: Target label
δ: Perturbation
F: Model
M: Mask (Patch Area)
T: Transformation Distribution
λ: Weight parameter

Algorithm 1 General GRAPHITE Framework

Input: Victim Image x, Target Image x_{tar}, Initial Mask M_{init}, Model F, Target Label y_{tar}
Output: Attacked Image A, Mask M, Perturbation δ

1: $M \leftarrow M_{init}$
2: $\delta, g \leftarrow$ INIT_PERT_+_GRAD$(x, x_{tar}, M, F, y_{tar})$
3: **while** not done **do**
4: $\quad S \leftarrow$ SELECT_PIXELS$(x, x_{tar}, M, \delta, y_{tar}, g)$
5: $\quad M \leftarrow$ REMOVE_PIXELS(M, S)
6: $\quad A, \delta, g \leftarrow$ ATTACK$(x, x_{tar}, M, \delta_{init}, F, y_{tar})$
7: $A, \delta \leftarrow$ Last Successful Attack

Key idea: jointly optimize mask size and transform-robustness

Ryan Feng
rtfeng@umich.edu

White-box Version of GRAPHITE

- Start with C&W ℓ_0 attack [1]
  - Alternates between C&W ℓ_2 attack [1] and removing the pixel with least impact
- Replace the C&W ℓ_2 attack with an EoT PGD attack [2, 3]
- Avg. 78% transform-robustness, 9% mask size

[1] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," IEEE S&P 2017.
[2] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," ICML 2018.
[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," ICLR 2018.

Ryan Feng
rtfeng@umich.edu

White-box GRAPHITE attacks can be generated.

What about **black-box** (hard-label) GRAPHITE attacks, where only the top-1 prediction label is available (no gradients, no probabilities)?

Ryan Feng
rtfeng@umich.edu

Hard-label Baselines

- Simple combinations of C&W ℓ_0 [2], EoT [3], and OPT Attack [4] poor
  - Issues included: Poor transform-robustness, large masks, query inefficiency

ℓ_0 **and OPT**	ℓ_0 **and OPT w/ EoT**	ℓ_0 **and Boosting**

- Pixel ordering by impact as in C&W ℓ_0 [2] breaks down without gradients
- Distance minimizing hard-label attacks query-inefficient with EoT

[1] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, “Robust Physical-World Attacks on Deep Learning Models,” CVPR 2018.
[2] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” IEEE S&P 2017.
[3] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, “Synthesizing robust adversarial examples,” ICML 2018.
[4] M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh, “Query-efficient hard-label black-box attack: An optimization based approach,” ICLR 2019

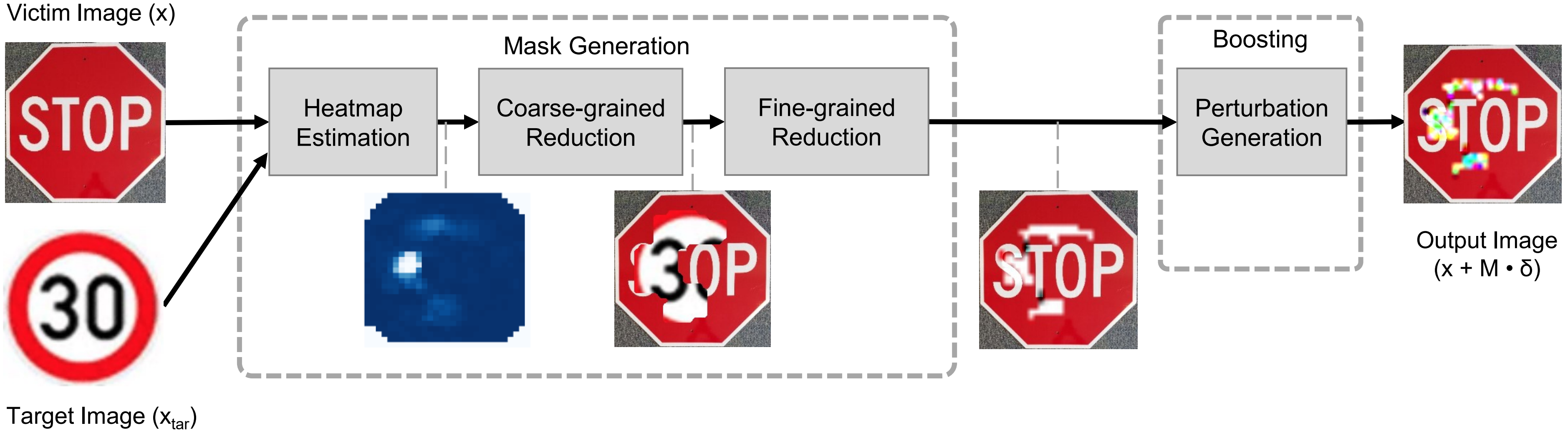
UNIVERSITY OF MICHIGAN
Ryan Feng
rtfeng@umich.edu

Hard-label Version of GRAPHITE

- Simplify to a two-step optimization – Mask Generation and Boosting

$$\underset{M}{\text{argmin}}\ \lambda \cdot ||M||_0 - \mathbb{E}_{t\sim T}\left[F\Big(t(x + M \cdot \delta_{tar})\Big) = y_{tar}\right]$$
$$\text{s.t. } \mathbb{E}_{t\sim T}\left[F\Big(t(x + M \cdot \delta_{tar})\Big) = y_{tar}\right] \geq tr_{lo}$$

$$\underset{\delta}{\text{argmax}}\quad \mathbb{E}_{t\sim T}\left[F\Big(t(x + M \cdot \delta)\Big) = y_{tar}\right]$$



Ryan Feng
rtfeng@umich.edu

Physical World Results

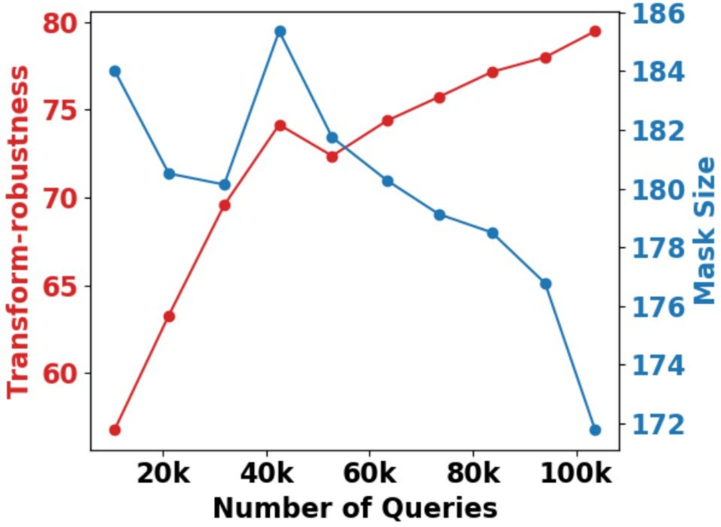
TABLE 8. GTSRB FIELD TEST RESULTS. PHYSICAL ROBUSTNESS RESULTS ARE CALCULATED OVER 5 PICTURES EACH AT THE FOLLOWING SPOTS: 5 FT $\times$ {0°, 15°, 30°, 45°}, 10 FT $\times$ {0°, 15°, 30°}, 15 FT $\times$ {0°, 15°}, 20 FT $\times$ {0°, 15°}, 25 FT, 30 FT, 40 FT. EACH EXAMPLE WAS TESTED 3 TIMES: OUTDOORS, INDOORS WITH INDOOR LIGHTS TURNED OFF, AND INDOORS WITH INDOOR LIGHTS TURNED ON.

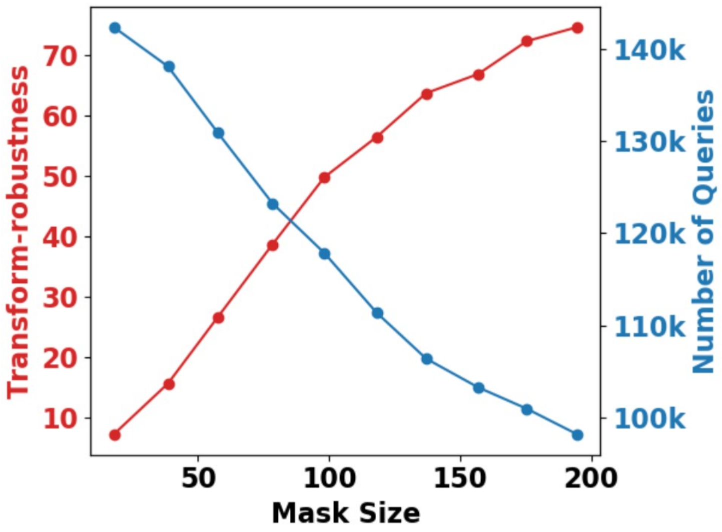
Victim	Target	Digital GRAPHITE attack	Physical GRAPHITE attack (outdoors)	Dig. TR (100 xforms)	Phys. TR (Indoors, lights off)	Phys. TR (Indoors, lights on)	Phys. TR (Outdoors)
				86%	92.9%	94.3%	100%
				79%	97.1%	85.7%	100%

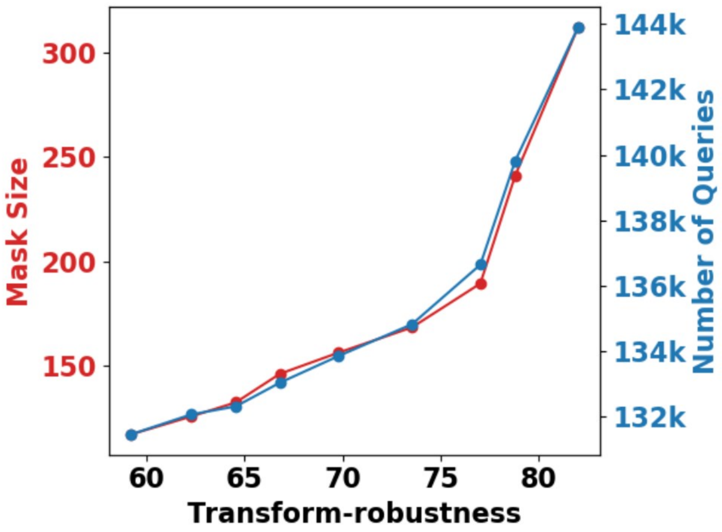
UNIVERSITY OF MICHIGAN

Ryan Feng
rtfeng@umich.edu

Tuning GRAPHITE







- 3 parameters to trade off: query count, transform-robustness, and mask size
- In the extreme, we can find attacks with as few as 500 queries with lower transform-robustness

Ryan Feng
rtfeng@umich.edu

Attacking PatchGuard

- GRAPHITE can defeat PatchGuard [1]
  - Tested on 100 CIFAR-10 examples
  - Avg. Transform-robustness: 68%
  - Avg. Query Count: 155.8k
  - Avg. Mask size: 193.81 pixels

- Example on right: 10 pixel attack to misclassify a dog as a cat

[1] C. Xiang, A. N. Bhagoji, V. Sehwag, and P. Mittal, “PatchGuard: A Provably Robust Defense against Adversarial Patches via Small Receptive Fields and Masking,” USENIX 2021.

Ryan Feng
rtfeng@umich.edu

Conclusion

- GRAPHITE: first automatic physical hard-label attack
- We hope GRAPHITE guides future defense work against practical attacks
- Code available to try it out:
  https://github.com/ryan-feng/GRAPHITE

Ryan Feng
rtfeng@umich.edu

Thank you!

- Contact Information
  - Ryan Feng
  - rtfeng@umich.edu
  - http://www-personal.umich.edu/~rtfeng/
  - https://twitter.com/ryantfeng

- Paper Links
  - https://arxiv.org/abs/2002.07088
  - https://github.com/ryan-feng/GRAPHITE