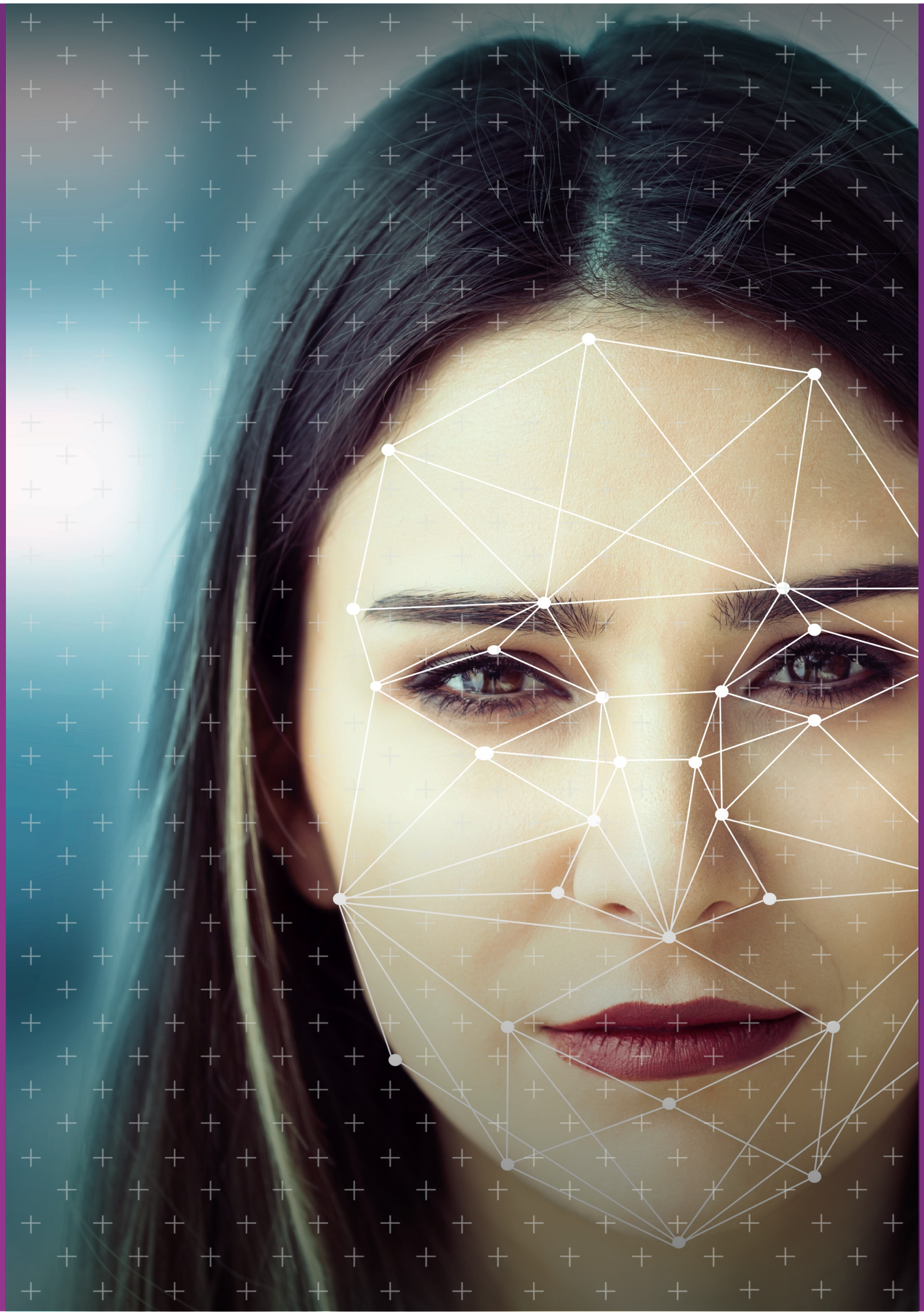


BART

Barrage of Random Transforms for Adversarially Robust Defense

Edward Raff[1,2,4]

Jared Sylvester[1,2,4]

Steven Forsyth[3]

Mark McLean[1]

1 Laboratory for Physical Sciences

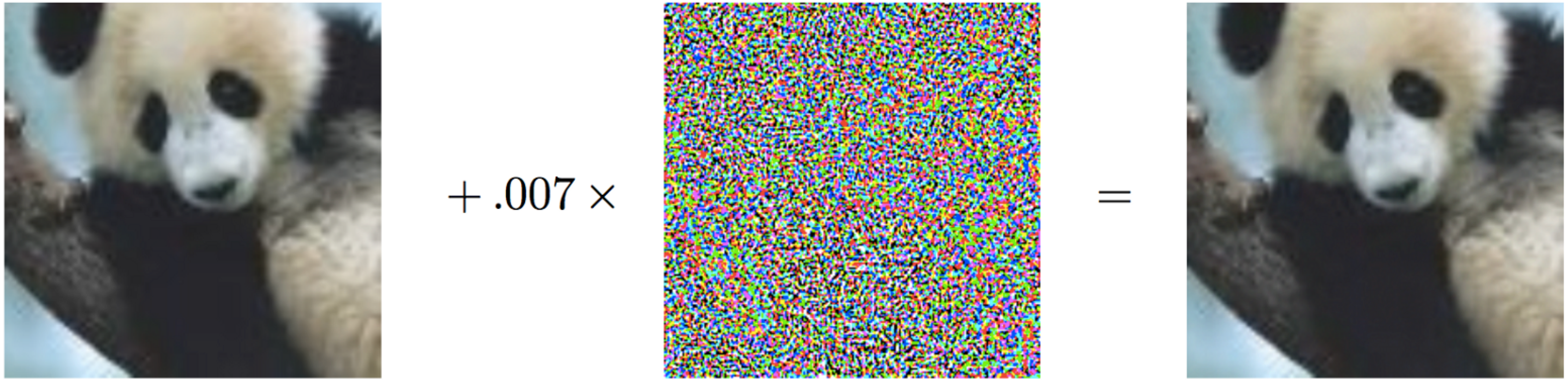
2 Booz Allen Hamilton

3 NVIDIA

4 U.M.B.C

15–21 June 2019 | CVPR | Long Beach, CA

ADVERSARIAL ATTACKS



panda
57.7%

attack perturbation

"gibbon"
99.3%

An attacker can make small perturbation that are numerically significant,
but semantically & perceptually meaningless.

What to do?

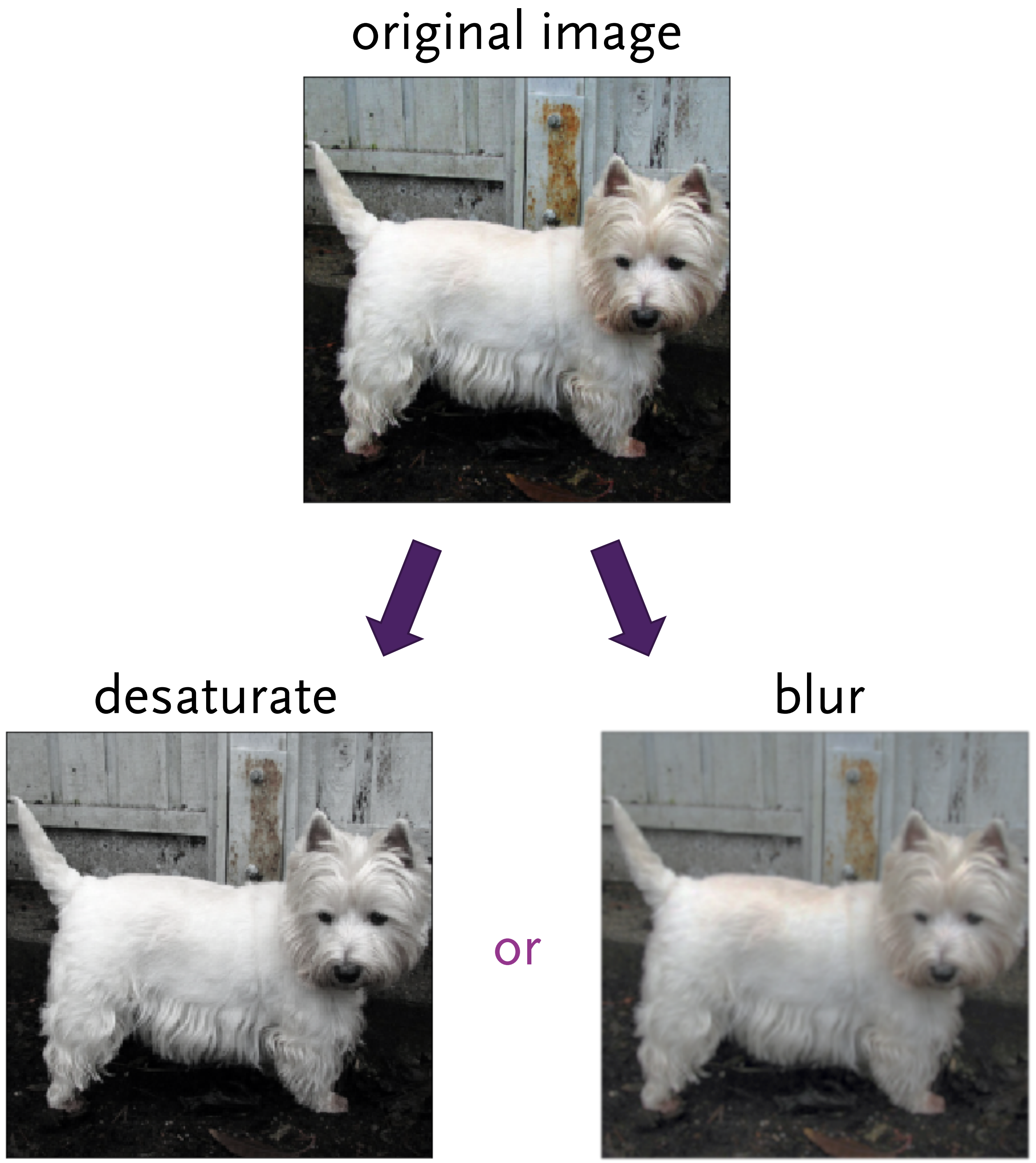
Make our own perturbations.

Image from: Goodfellow, *et al.* "Explaining and Harnessing Adversarial Examples." ICLR, 2015.

TRANSFORMATIONS FOR DEFENSE

- Modify the image at inference time.
  - e.g. by blurring, adding noise, desaturating.
- This should interfere with the adversary's ability to find a successful attack perturbation.
- This has been tried before…

  …*and it didn't work*.
- It makes following the gradient between original and attacked image only trivially harder.

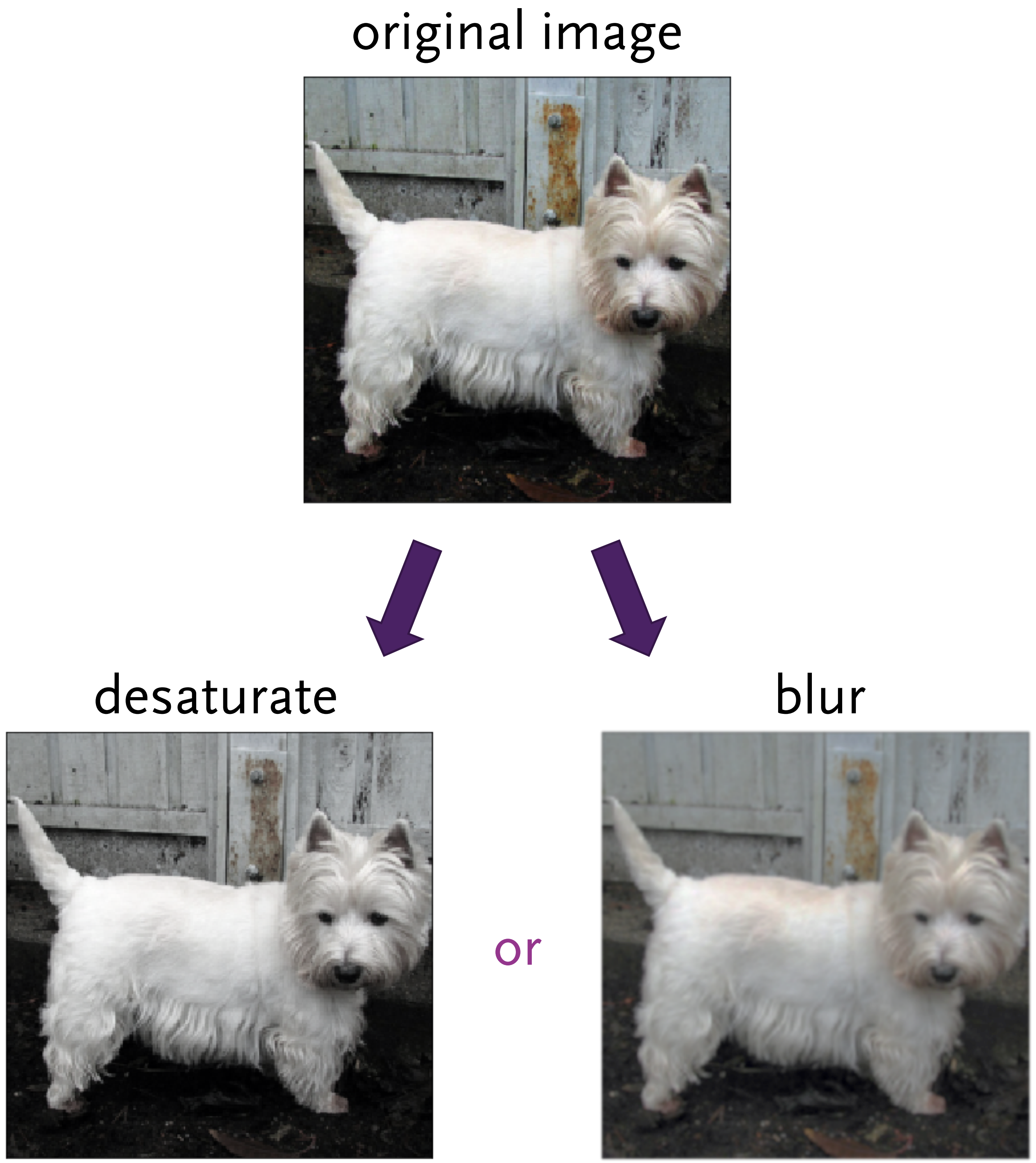


TRANSFORMATIONS FOR DEFENSE

- Modify the image at inference time.
  - e.g. by blurring, adding noise, desaturating.
- This should interfere with the adversary's ability to find a successful attack perturbation.
- This has been tried before…
  …and it didn't work.
- It makes following the gradient between original and attacked image only trivially harder.

So what's different with BaRT?

1. Take a large set of transformations.
2. Parameterize each one randomly.
3. Randomly select a subset to apply for each input.
4. Apply them in randomized, serial order.



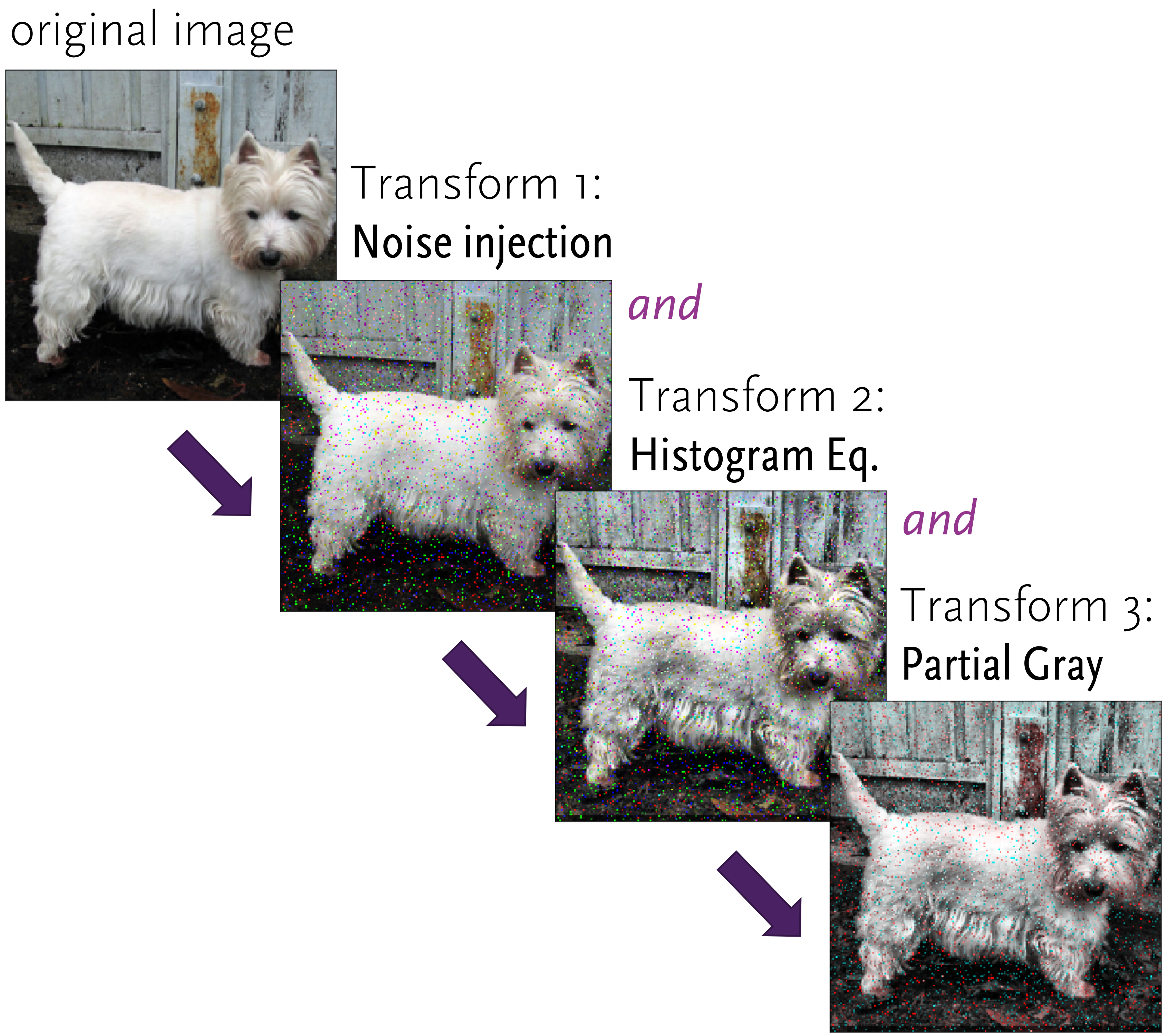
TRANSFORMATIONS FOR DEFENSE

- Modify the image at inference time.
  - e.g. by blurring, adding noise, desaturating.
- This should interfere with the adversary's ability to find a successful attack perturbation.
- This has been tried before…

  …and it didn't work.
- It makes following the gradient between original and attacked image only trivially harder.

So what's different with BaRT?

1. Take a large set of transformations.
2. Parameterize each one randomly.
3. Randomly select a subset to apply for each input.
4. Apply them in randomized, serial order.



EXAMPLES OF SINGLE TRANSFORMS

Example output #1 | Example output #2 | Example output #3

Alter XYZ

Convert to CIE XYZ color space, perturb w/ random offset, convert back to RGB

original image

Alter LAB

Convert to CIE LAB color space, perturb w/ random offset, convert back to RGB

Gaussian Blur

Blur using a Gaussian with randomly chosen standard deviation

MANY WEAK DEFENSES MAKE A STRONG DEFENSE

- Twenty five weak defenses to choose from.
  - On their own, each can be easily defeated.
  - *When ensembled together, they provide state-of-the-art defense.*
  - "Randomness on top of randomness"

Original image

Example image,
5 transforms

Example image,
5 transforms

RANDOMNESS ON TOP OF RANDOMNESS

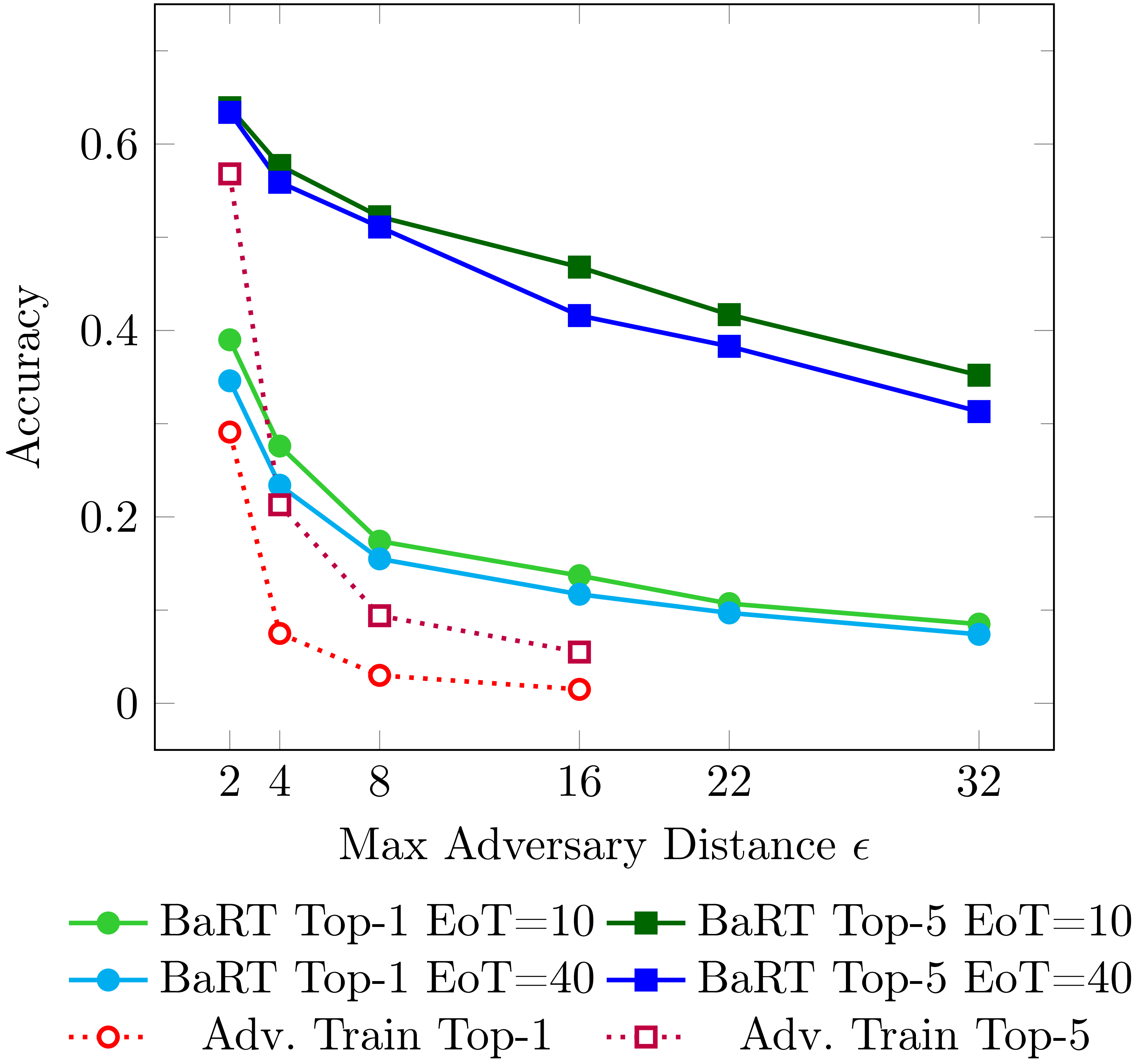
Instead of attacking this:

...you have to attack this:

(Example images, 5 transforms)

- Every time the adversary takes another gradient step, the image is being transformed differently.
- The direction to the decision surface is changing, so subsequent gradient steps are not aligned.

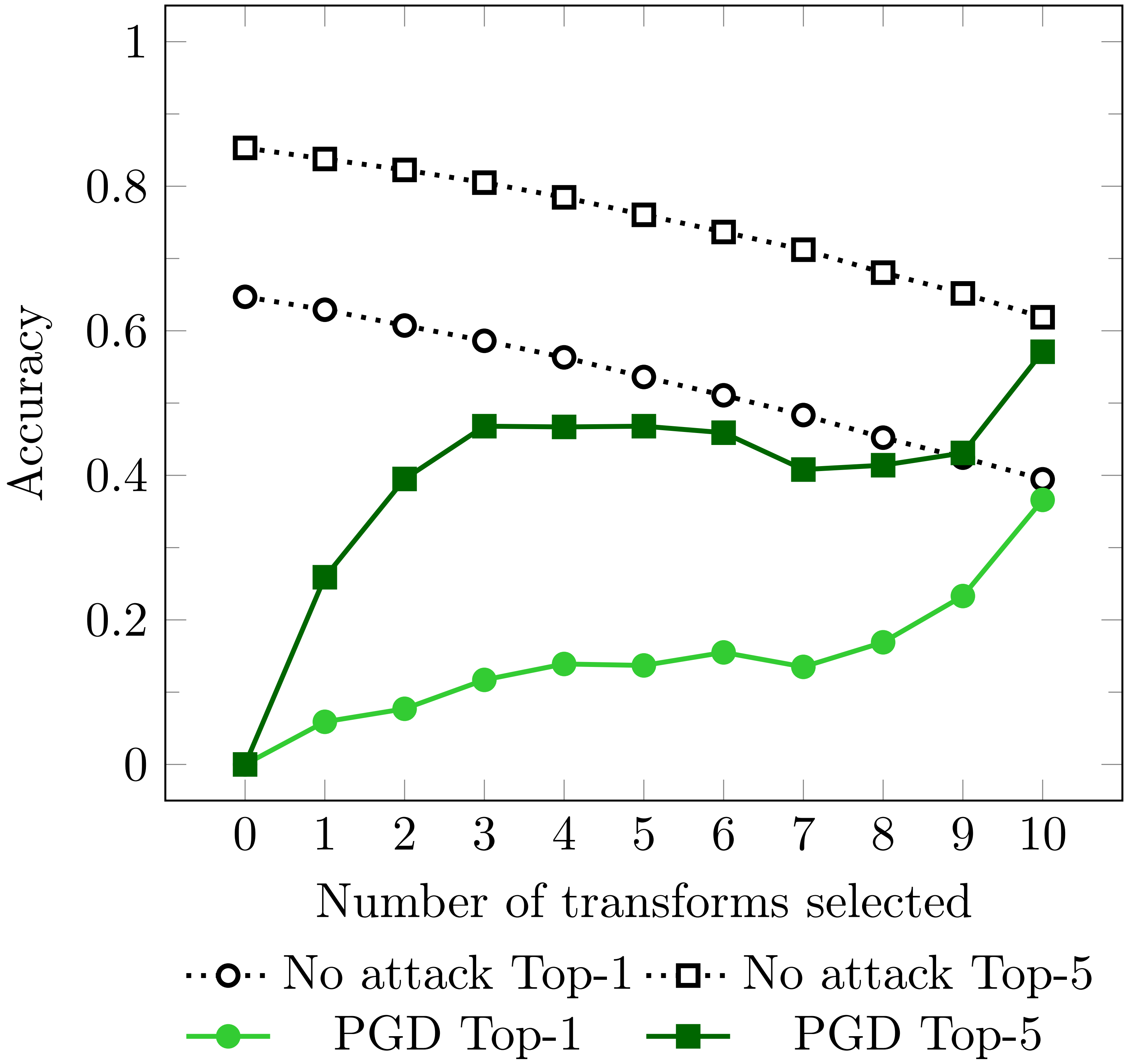
RESULTS: VARYING ATTACK STRENGTH



- Created the strongest adversaries we could (PGD).
  - Implemented BPDA and EoT to allow the adversary to approximate each transform.
  - Allowed attacker to know the randomly chosen parameters of each defense.
  - Allowed adversarial distance of up to $\varepsilon = 32$.
  - Thoroughly tested for obfuscated gradients.
  - Created a new attack we thought might be better able to defeat BaRT.
- BaRT surpasses the previous state-of-the-art defense for ImageNet. (Adversarial Training.*)
  - Top-5 accuracy of >57% when under attack.
  - Higher Top-1 accuracy than the Top-5 accuracy of Adversarial Training when $\varepsilon \geq 4$.

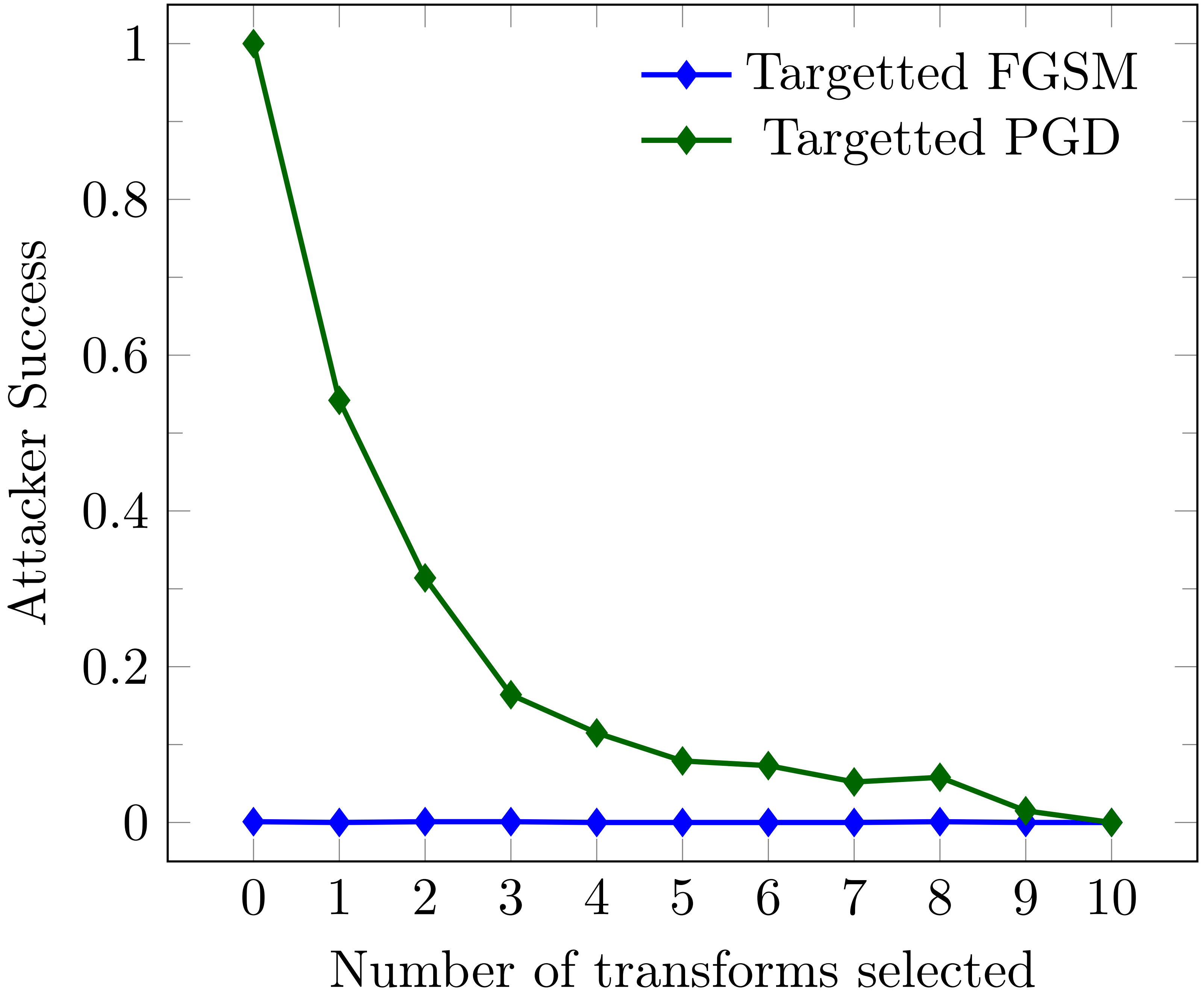
* Kurakin, Goodfellow & Bengio. "Adversarial Machine Learning at Scale." ICLR, 2017.

VARYING NUMBER OF DEFENSIVE TRANSFORMS: UNTARGETED ATTACKS



- Adding more transforms to the ensemble costs accuracy when not being attack.
- But it increases accuracy when under attacked.

VARYING NUMBER OF DEFENSIVE TRANSFORMS: TARGETED ATTACKS



- With no defensive transforms, the PGD attacker had 100% success rate.
- With 10 defensive transforms, *success falls to 0%.*

CONCLUSIONS

- By integrating domain knowledge (image transforms) and randomness (ensembling), we develop a new defense against adversarial attacks.
- We provide evidence that weak defenses can have value.
- BaRT is simple to implement & use in the short term, and gives us inspiration on how we might develop long-term defenses.

Future work:

- Fine tune transformations & add others to the pool of options.
- Ensembling expands BaRT's defense-in-depth to allow defense-in-width as well.
- Apply to other domains.
- Can we use randomness to build a provably robust defense?
- Adapting defensive strength (i.e., number of transforms) vs. throughput for real-world applications.

THANK YOU

For more information, please contact us!

Jared Sylvester, PhD | Sr. Lead Data Scientist | Booz Allen Hamilton

sylvester_jared@bah.com, raff_edward@bah.com, sforsyth@nvidia.com, mrmclea@lps.umd.edu