Bounded Adversarial Attack on Deep Content Features

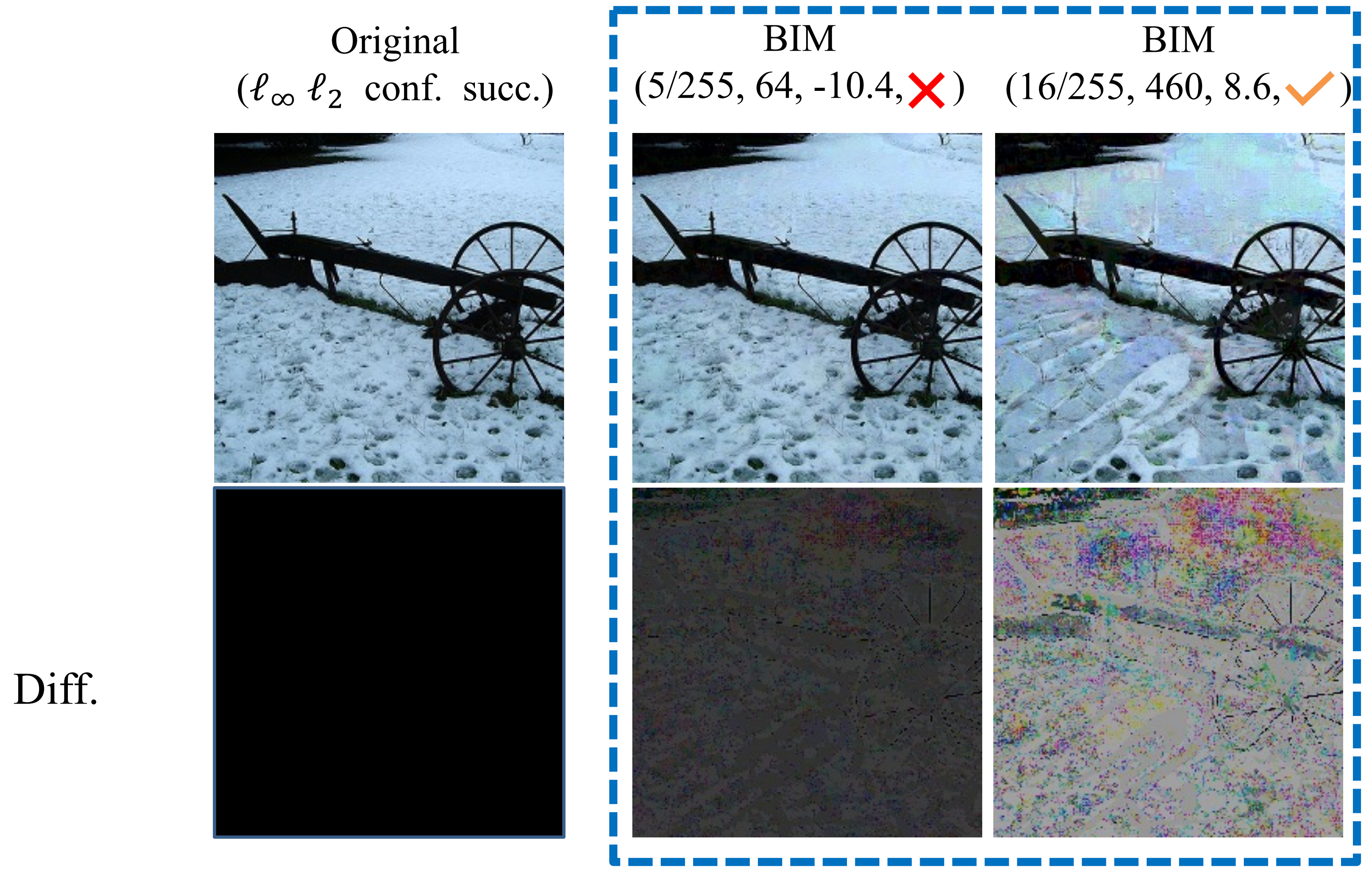
Qiuling Xu, Guanhong Tao and Xiangyu Zhang @ Purdue University

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Existing Adversarial Attack Bound is Hard to Scale

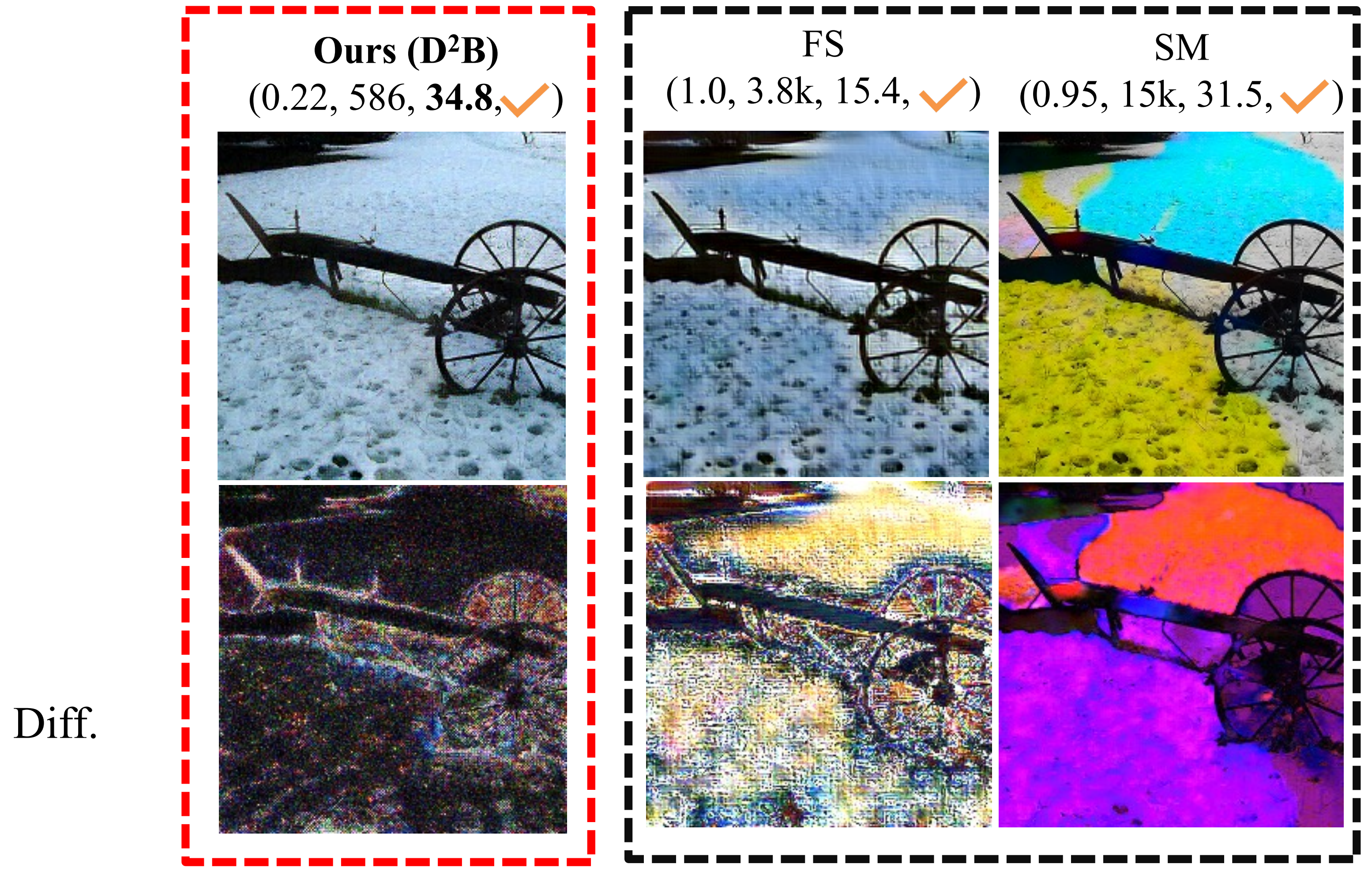
- Pixel Space

Attack fails given a small bound; Attack is detectable on a larger bound.

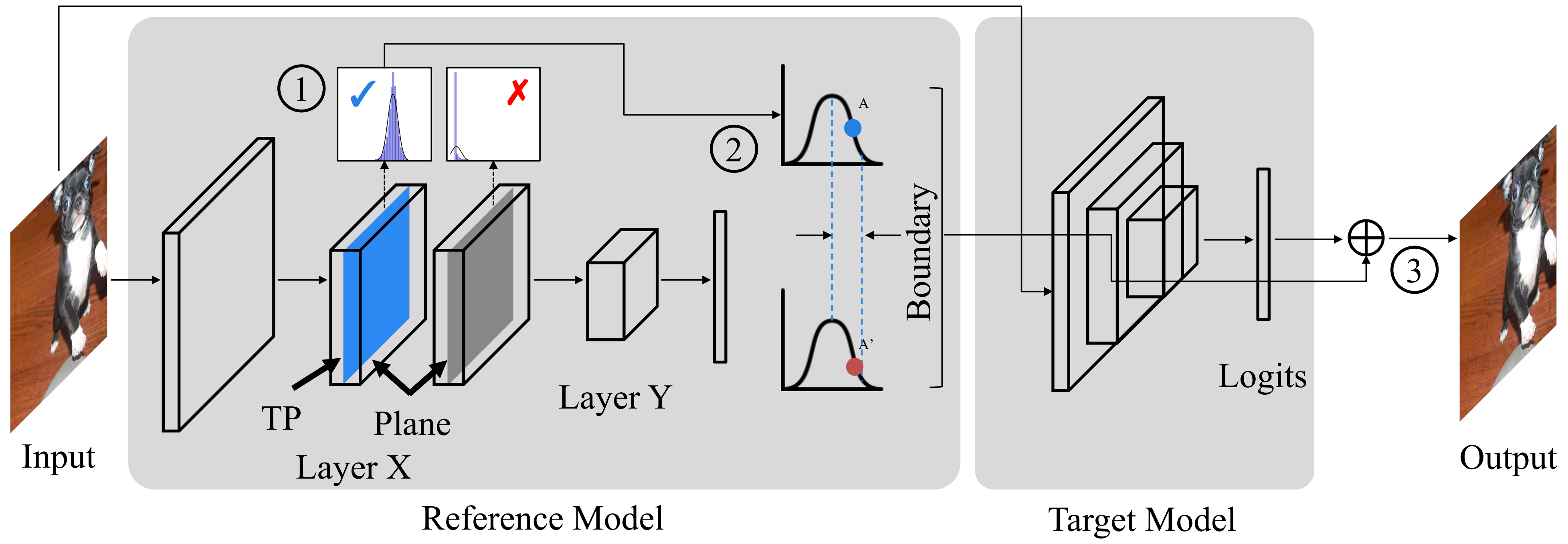


- Feature Space

Attack is detectable on samples with high confidence.

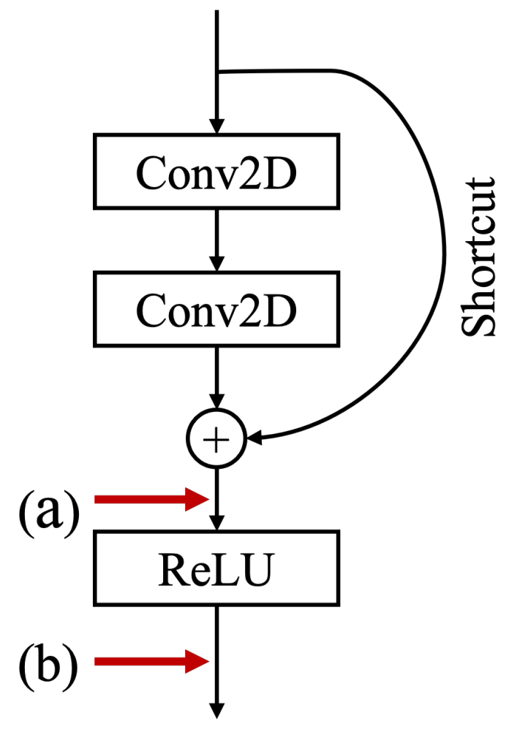


Method - Quantile Bound on Gaussian Representation



- ① Throttle plane (TP) selection
- ② Internal distribution boundary constraint
- ③ Adversarial sample generation with combined losses

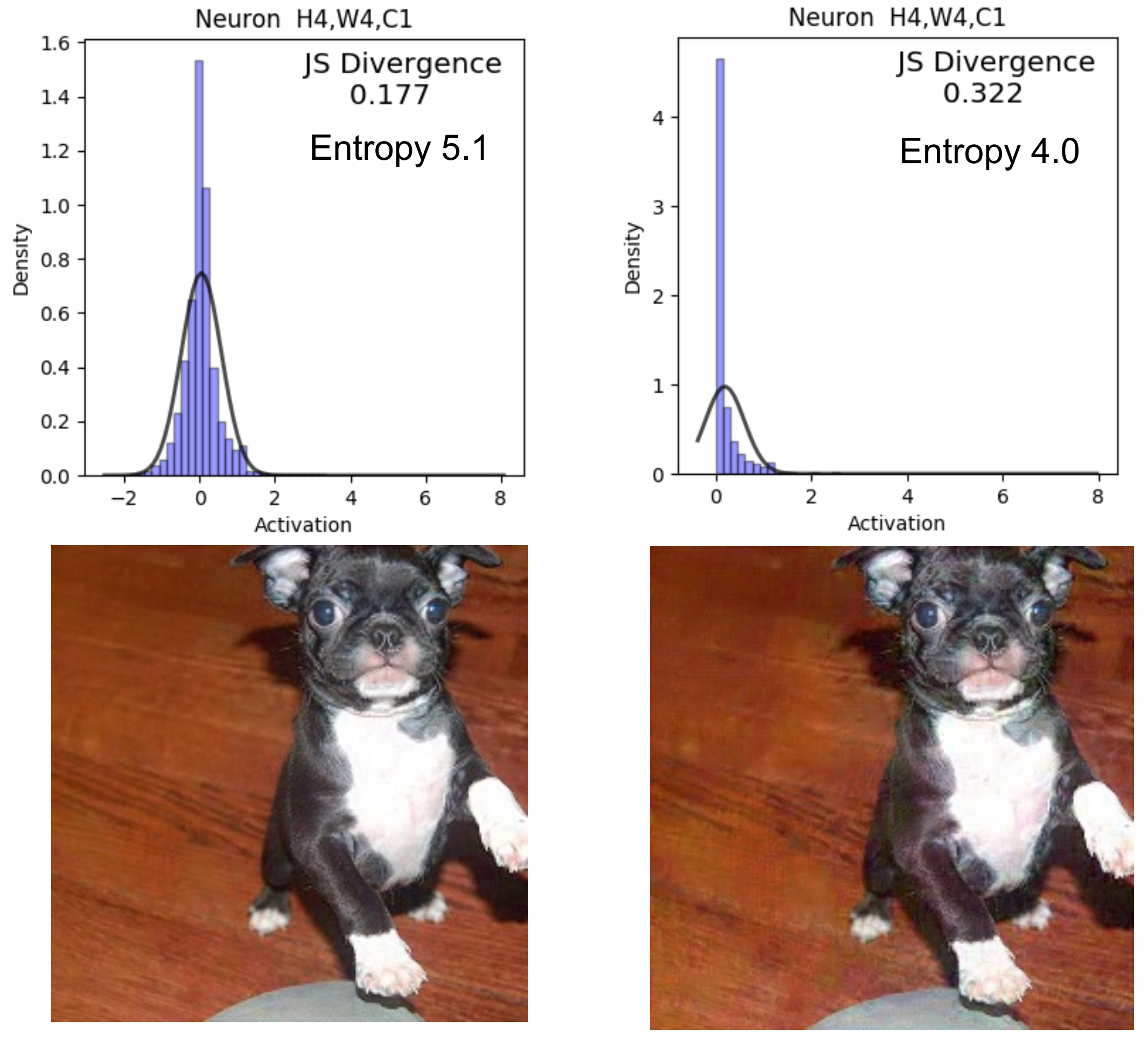
Distribution of the Representation Matters



ResNet Structure.

Which representation should we use for quantile bound?

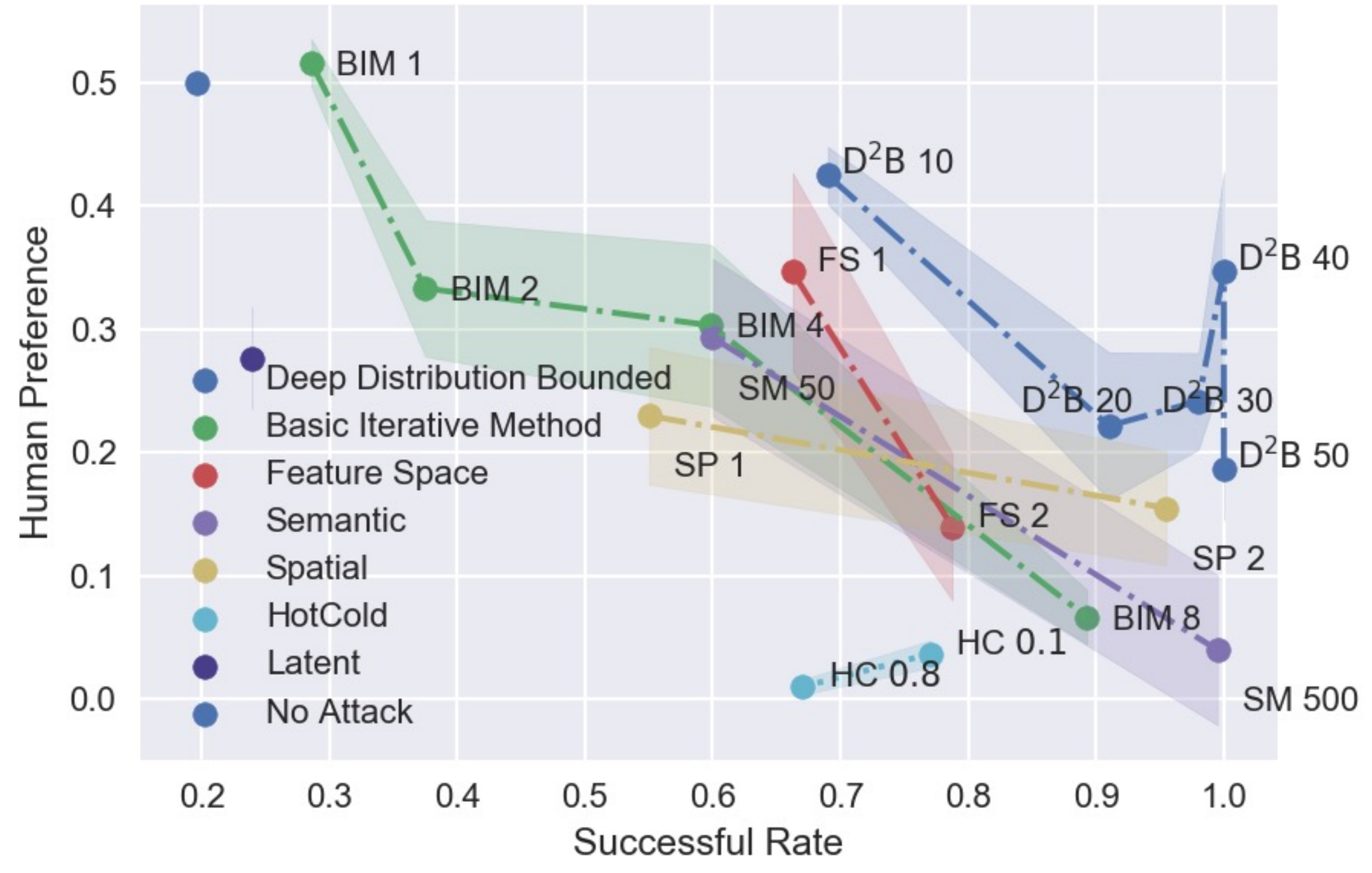
Look for Gaussian.

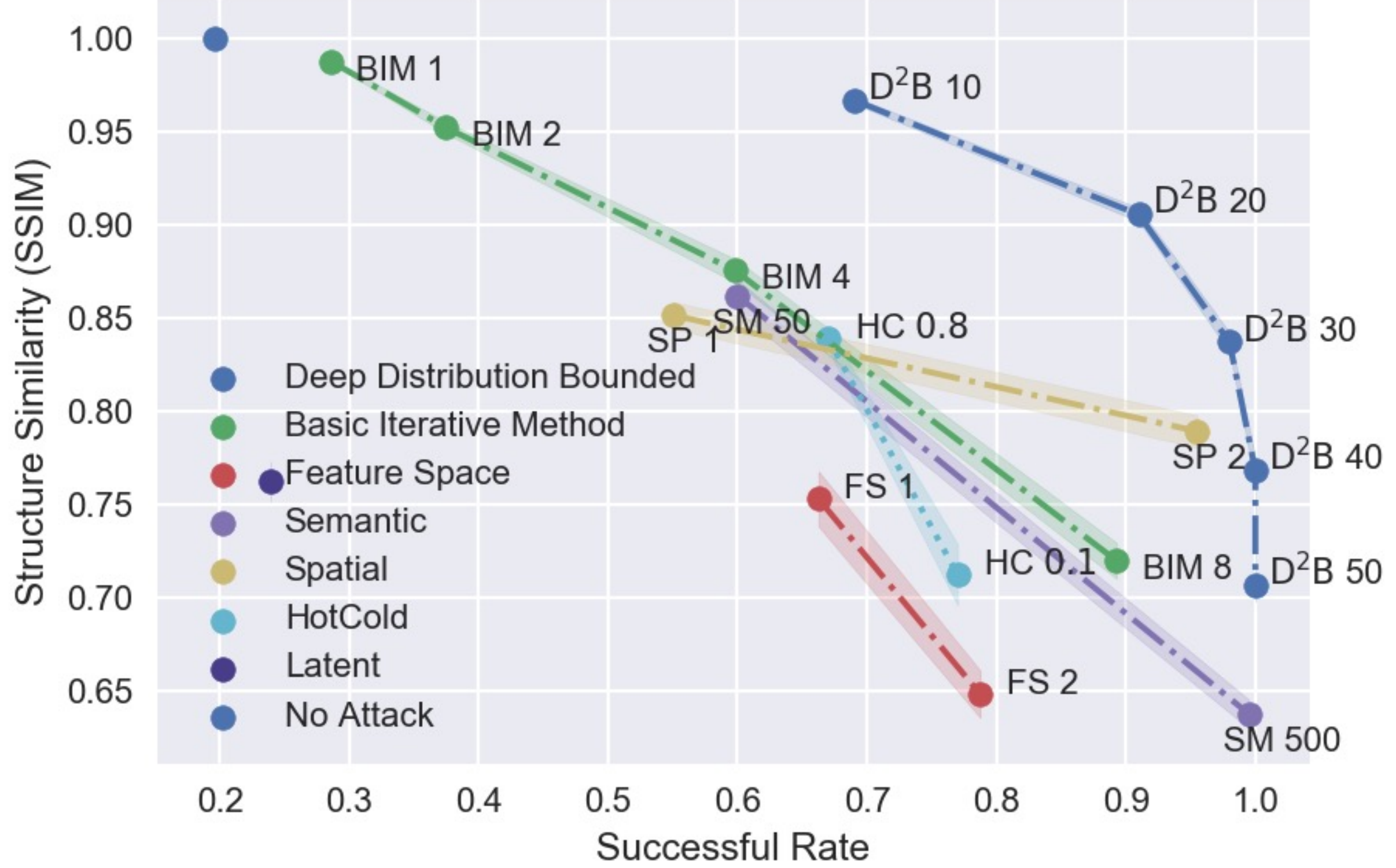


(a) Before ReLU

(b) After ReLU

STOA Tradeoff on Imperceptibility versus Success





D2B has higher success rate at the same level of human preference. And with the same success rate, our adversarial examples are consistently more favored by the testers.

Summary

- Identify the scalability problem of existing adversarial attacks.
- Proposed the quantile bound on deep content features.
- Proposed an efficient way for optimizing the adversarial samples
- Show the state-of-the-art trade-off between imperceptibility and attack success using the quantile bound.

Contact

- Email: xu1230@purdue.edu

Home Page:

SCAN ME