

Open Problem: Realistic attacks and defenses

© 2019 Philipp Krähenbühl and Chao-Yuan Wu

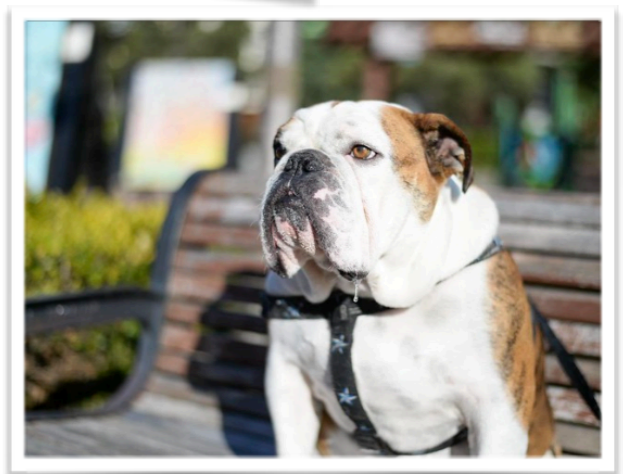
What attacks should we worry about?

- Random noise attacks don't matter (yet)
- Doing the wrong thing for real images does



How do we specify the right attack model?

- Try a validation set
 - No guarantees
 - Might overfit to validation / test set
 - Failures can be rare, but fatal



How do we specify the right attack model?

- Enumerate all possible inputs from a generative model
 - too hard
 - no guarantees
 - too expensive



Image source: Large Scale GAN Training for High Fidelity Natural Image Synthesis, Brock et al., ICLR 2019

How do we specify the right attack model?

- Can we mathematically describe a superset of all inputs?
 - and optimize?
 - Very hard

