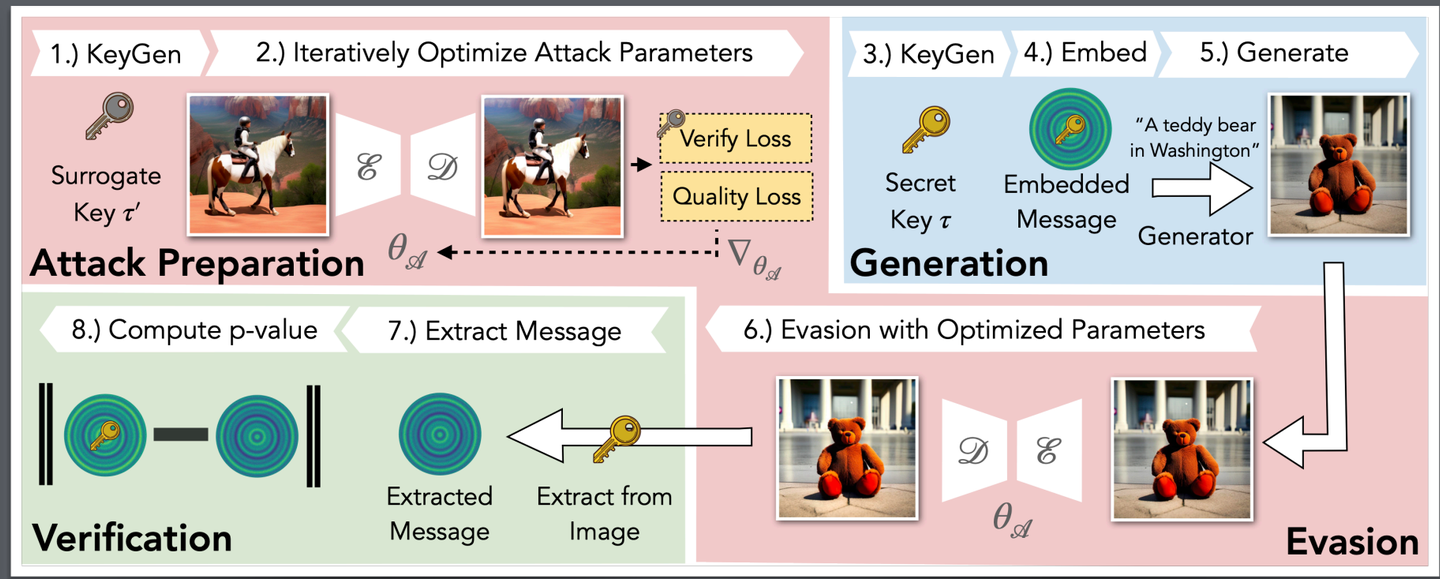
Optimizing Adaptive Attacks against Image Watermarks

Nils Lukas, PhD Candidate

David R. Cheriton School of Computer Science

Abstract

- Untrustworthy users can misuse image generators (e.g., deepfakes)
- **Watermarking** makes deepfakes detectable, but requires *robustness*
- How do we know that an (adaptive) attack is optimal (i.e., best possible)?
- **Our Solution**: Approach attack as optimization over *surrogate keys* by making watermark verification differentiable, i.e., easily optimizable
- **Results**: All surveyed watermarks are broken for 1 billion parameter models



Discussion

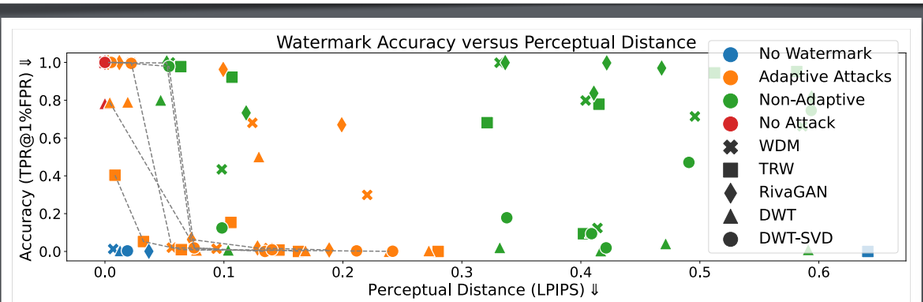
- Watermarking needs to be **trustworthy**, but we lack strong attackers. Robustness test needs adaptive, *learnable* attackers
- We evaluate adv. noising and compression, but more effective attacks (image edits) possible
- We point to design flaws in existing watermarking methods
- Not disclosing algorithm prevents attacks, but is vulnerable to whom this information is released

Threat Model

- **Adaptive**: Attacker *knows* the watermarking algorithm, but not the secret watermarking key
- **Surrogate Model**: Controls less capable, open-source generator
- **Compute:** Limited resources, cannot train their own generator from scratch
- **Dataset**: Any public image dataset
- **Queries**: Limited in the number of queries to the watermarked generator

Goals:

- Evade watermark detection (p>0.01)
- Preserve image quality (FID, CLIP score)



Our Attacks (High Contrast)

Adversarial Noise (p=0.13)

Adversarial Compression (p = 0.79)

nlukas@uwaterloo.ca