From Breach to Barrier: Addressing Watermark Vulnerabilities Through Multi-Domain Synergy

Wei, Nicholas (School: Williamsville East High School)
Yu, Jesse (School: Millburn High School)

As hyper-realistic deepfakes grow into one of the world's most pressing crises, the rush to secure digital integrity has made watermarking one of the most urgently researched sectors in cybersecurity. Current watermarking approaches exist in two domains: spatial or latent. We discovered spatial watermarks survive geometric attacks like cropping but fail against generative editing, while latent watermarks do the opposite. We hypothesized that watermarking both domains could eliminate these critical vulnerabilities. However, naively combining watermarks of both domains caused interference between the spatial watermark payload and latent watermark frequencies, a concept we formalize as "Objective Collision". To resolve objective collision, we engineered a novel CNN architecture called the "Synthetic Adapter" that reroutes the spatial watermark to avoid frequencies occupied by the latent watermark. Our hypothesis proved true, with the synergized watermark being ~2.65x more reliable than Tree-Ring (latent) for geometric attacks and ~5.38x more reliable than RivaGAN (spatial) for generative attacks. The synthetic adapter further reduced failure rate of the unsynergized combination watermark by ~8.75%. To red-team our novel defense, we formalized the Combinatorial Multi-Vector (CMV) threat that attacks the watermark with both geometric and generative attacks simultaneously. The CMV threat showed a 60%-80% relative survivability drop compared to single-vector attacks, revealing that single-vector evaluations overestimate a watermark's security. We established the first foundational blueprint to remove the major single-vector vulnerabilities of watermarks. Our CMV red-teaming establishes the foundational necessity of multi-vector threat evaluation in provenance research.

Awards Won:
Fourth Award of $600
Association for the Advancement of Artificial Intelligence: AAAI Student Memberships for each finalist that is part of the 1st, 2nd, and 3rd Prize Winning projects and 5 Honorable Mention winning projects (up to 3 students per project) (in-kind award / part of the 1st-3rd prize)