CHIN-YUAN YEH

marrch30@gmail.com ◇ Google Scholar

EDUCATION

Doctor of Philosophy, National Taiwan University (Major in Data Science) 2020 - now

Master of Science, National Taiwan University (Major in Data Science) 2018 - 2020

Bachelor of Science, National Taiwan University (Major in Physics) 2013 - 2017

PUBLICATIONS

- "Disrupting Image-Translation-Based DeepFake Algorithms with Adversarial Attacks," **Chin-Yuan Yeh**, H.-W. Chen, S.-L. Tsai, & S.-D. Wang, *WACVW (2020)*.
- "Attack as the Best Defense: Nullifying Image-to-image Translation GANs via Limit-aware Adversarial Attack," **Chin-Yuan Yeh**, H.-W. Chen, H.-H. Shuai, D.-N. Yang, & M.-S. Chen, *ICCV (2021)*.
- "Planning Data Poisoning Attacks on Heterogeneous Recommender Systems in a Multiplayer Setting," **Chin-Yuan Yeh**, H.-W. Chen, D.-N. Yang, W.-C. Lee, P. S. Yu, & M.-S. Chen, *ICDE (2023)*.
- "Does Audio Deepfake Detection Rely on Artifacts?" T.-H. Shih, **C.-Y. Yeh**, & M.-S. Chen, *ICASSP (2024)*.
- "FedGCR: Achieving Performance and Fairness for Federated Learning with Distinct Client Types via Group Customization and Reweighting," S.-L. Cheng, **C.-Y. Yeh**, T.-A. Chen, E. Pastor & M.-S. Chen, *AAAI (2024)*.

SKILLS

Technical Skills Python, Pytorch, Bash scripts in Unix Systems; Academic Writing
Soft Skills Research Project Proposal and Guidance

RESEARCH

Disrupting Image-Translation-Based DeepFake Algorithms with Adversarial Attacks. I am the first to introduce adversarial attack strategies to incapacitate Deepfake models, i.e., image translation GANs (CycleGAN, pix2pix, and pix2pixHD). I develop two attacks: Nullifying Attack which minimizes Deepfake modification, and Distorting Attack which causes distortion to deepfake outputs. Published at WACVW 2020.

Attack as the Best Defense: Nullifying Image-to-image Translation GANs via Limit-aware Adversarial Attack. I develop Limit-Aware Self-Guiding Gradient Sliding Attack (LaSGSA), a query-based **black-box** norm-bounded adversarial attack against Img2Img GANs (potentially used as deepfakes) with three optimization acceleration techniques: limit-aware RGF which restricts query sampling within the ϵ-bound, gradient sliding mechanism that propagates after being clipped by the ϵ-bound, and self-guiding prior, which leverages the semantic consistency of Img2Img GANs causing the Jacobian matrix of its mapping to be diagonal. Published at ICCV 2021.

Planning Data Poisoning Attacks on Heterogeneous Recommender Systems in a Multiplayer Setting. I develop Multilevel Stackelberg Optimization over Progressive Differentiable Surrogate (MSOPDS), a data poisoning technique against Heterogeneous RecSys, addressing the scenario of multiple attackers poisoning the same Recommendation System, where the first attacker aims to prevent subsequent attackers from harming his poisoning objective. MSOPDS leverages Stackelberg Game analysis between the first attacker the subsequent attackers' actions and projection techniques to navigate the gradient descent over discrete RecSys operations. Published at ICDE 2023.

Does Audio Deepfake Detection Rely on Artifacts? Anticipating future deepfakes that contain fewer artifacts, we introduce the Balanced Environment Audio-Deepfake Reevaluation (BEAR) protocol, which creates a balanced setting with similar artifacts or noise in both genuine and deepfake samples with two variants. White-BEAR introduces deepfake specific artifacts to genuine samples by constructing "self-deepfakes." Gray-BEAR adds Gaussian noise to both genuine and forged samples. We observe a significant performance drop for all experimented detectors, indicating that current detection models heavily rely on artifacts and struggle to identify deepfakes. Published at ICASSP 2024.