TSUBAME 共同利用　令和 7 年度　学術利用　成果報告書

利用課題名　事実検証とプライバシー保護
英文：Fact Checking and Privacy Protection

利用課題責任者
越前　功

所属
国立情報学研究所
https://www.nii.ac.jp/

邦文抄録（300 字程度）

英文抄録（100 words 程度）
Using TSUBAME GPU resources, we conducted research on AI security, privacy, and multimodal disinformation analysis. The project produced five main outcomes: GreedyPixel for fine-grained black-box attacks; a multilingual multimodal disinformation dataset; an 8-way taxonomy and benchmark for multimodal disinformation analysis; MSPD, a fast transferable pre-release defense against adversarial attacks; and DiffMI, a diffusion-based training-free model inversion attack for face recognition. These results advance the evaluation of security, robustness, and reliability in modern AI systems.

Keywords:
AI Security; Adversarial Robustness; Model Inversion; Face Recognition; Multimodal Disinformation

背景と目的

（課題の背景を記載してください）

（現状の問題点等を挙げてください）

本プロジェクトでは、（目的と解決手段と、成果を述べてください）を XX によって解決し YY の成果を得た。

Deep learning is now widely used in authentication, media generation, information recommendation, and fact-checking. At the same time, its deployment has raised major concerns related to adversarial manipulation, privacy leakage, and AI-generated disinformation. Key challenges include understanding high-precision attacks in black-box settings, evaluating privacy leakage from face embeddings, designing efficient defenses against unseen attacks, and building systematic benchmarks for multimodal disinformation analysis. In this project, we used TSUBAME to conduct large-scale training, inference, and comparative experiments, and developed new attack and defense methods as well as supporting datasets and benchmarks.

概要

（申請書の概要を基本とし、実情に合わせて変更してください。図や表の利用を図って分かり易く記載して下さい。）。

This project produced five research outcomes spanning fine-grained black-box attacks, multimodal disinformation analysis, preemptive defense, and face recognition privacy. Figure 1-5 summarize the core ideas of the five papers.

1. GreedyPixel ([1])

A query-based black-box adversarial attack that performs per-pixel greedy optimization and reaches near white-box precision without gradient access.

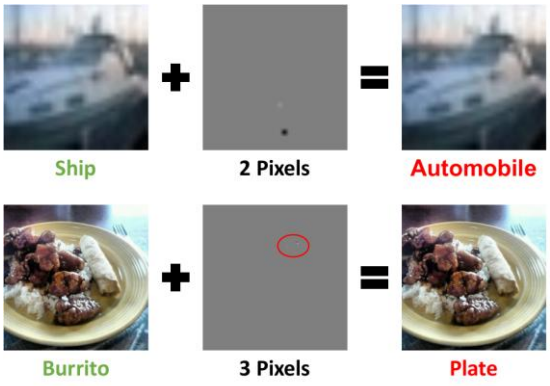

Figure 1. GreedyPixel changes only a few pixels while causing misclassification.

2. Multilingual disinformation dataset ([2])

A multilingual multimodal benchmark curated

from professional fact-checking websites, with image-text pairs, five-level factuality labels, and supportive explanations.

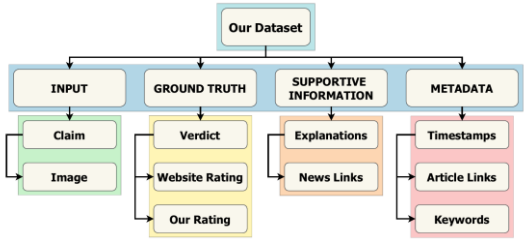

Figure 2. Structure of the multilingual multimodal disinformation dataset.

3. 8-way disinformation taxonomy ([3])

A unified formulation that jointly models image veracity, text veracity, and cross-modal consistency, together with a benchmark and LVLM-based reasoning pipeline.

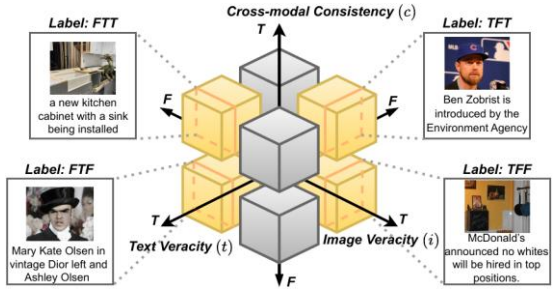

Figure 3. The 8-way taxonomy covers image, text, and cross-modal consistency.

4. MSPD preemptive defense ([4])

A fast and transferable pre-release defense that improves robustness against future adversarial attacks under a minimal two-epoch optimization design.

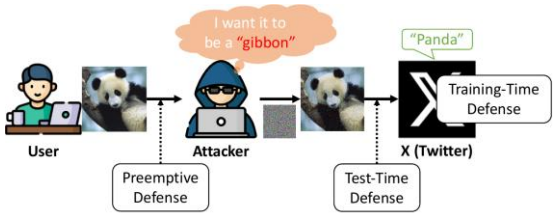

Figure 4. MSPD protects media before release to mitigate later attacks

5. DiffMI model inversion ([5])

The first diffusion-based, training-free model inversion attack for face recognition, enabling efficient identity reconstruction from embeddings on unseen targets.

Figure 5. DiffMI studies privacy leakage from face recognition embeddings.

結果および考察

（結果と考察を記載してください。図や表の利用を図って分かり易く記載して下さい。）。

1. Results on adversarial attack and defense

GreedyPixel [1] introduced a fine-grained black-box adversarial attack based on per-pixel greedy optimization with query feedback only, achieving high attack precision without gradient access. MSPD [4] proposed a fast and transferable pre-release defense that adds lightweight protective perturbations before media release and showed effectiveness across unseen models and attacks. Together, these results improve practical robustness evaluation from both the attack and defense perspectives.

2. Results on face recognition privacy

DiffMI [5] introduced the first diffusion-based, training-free model inversion attack for face recognition systems. By combining robust initialization and confidence-aware optimization, it can reconstruct identity information from leaked embeddings without target-specific training and can generalize across different face recognition models.

3. Results on multimodal disinformation analysis

We first built a multilingual multimodal dataset for disinformation detection using image-text pairs collected from professional fact-checking websites and annotated with a five-level factuality scheme and supportive explanations [2]. We further proposed an 8-way taxonomy that jointly models image veracity, text veracity, and cross-modal consistency, together with a benchmark dataset and a modular LVLM-based pipeline for multimodal reasoning [3].

まとめ、今後の課題

（まとめと今後の課題について記載してください。）。

Overall, this project produced five outcomes

spanning attack methods, pre-release defense, face recognition privacy evaluation, disinformation datasets, and multimodal disinformation benchmarks. TSUBAME's high-performance GPUs and large memory capacity enabled efficient large-scale experiments across multiple models, datasets, and threat settings. Future work will extend these studies toward more realistic threat models, lightweight yet reliable defense design, and broader evaluation frameworks for secure and trustworthy multimodal AI.

Reference

[1] Hanrui Wang, Ching-Chun Chang, Chun-Shien Lu, Christopher Leckie, and Isao Echizen. (2025) "Greedypixel: Fine-grained black-box adversarial attack via greedy algorithm." *IEEE Transactions on Information Forensics and Security, 20,* 12080-12095.

[2] Shuhan Cui, Hanrui Wang, Ching-Chun Chang, Huy H. Nguyen, and Isao Echizen. (2025) "A Multilingual, Multimodal Dataset for Disinformation and Out-of-Context Analysis with Rich Supportive Information." In *Proceedings of the 27th International Conference on Multimodal Interaction (ICMI)* (pp. 643-651).

[3] Shuhan Cui, Ruimin Chu, Hanrui Wang, Patrick H. Chen, Ching-Chun Chang, and Isao Echizen. (2026) "An 8-Way Taxonomy for Multimodal Disinformation and Detection Benchmark." in *Proceedings of the ACM Web Conference (WWW)* (accepted).

[4] Hanrui Wang, Ching-Chun Chang, Chun-Shien Lu, Ching-Chia Kao, and Isao Echizen. "Minimal Cascade Gradient Smoothing for Fast Transferable Preemptive Adversarial Defense." *arXiv preprint arXiv:2407.15524* (latest revised in 2026).

[5] Hanrui Wang, Shuo Wang, Chun-Shien Lu, and Isao Echizen. "DiffMI: Breaking Face Recognition Privacy via Diffusion-Driven Training-Free Model Inversion." *arXiv preprint arXiv:2504.18015* (latest revised in 2025).