情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

ゼミ発表

標的型敵対的サンプルを用いた CAPTCHAシステム

2026/02/03

大久保研究室M2 5542102
福井恵悟

研究背景：CAPTCHAの役割

人間とコンピュータを区別するためのチューリングテスト

- Webサービスへの不正アクセス防止
- アカウント大量作成などのスパム対策

現状の主流：画像認識型（reCAPTCHA v2等）

研究背景： reCAPTCHA v2[1]

リスク分析エンジン

- 画像選択だけでなく、ユーザの挙動やCookie、IPを解析
- 「人間らしい」→チェックボックスのみで通過
- 「疑わしい」→画像タスク出現

Type1	3×3 グリッド（分類）
Type2	4×4 グリッド(領域選択)　※右図参照
Type3	ダイナミック　3×3 グリッド

課題：ボットの進化と脅威

ボットの高度化

深層学習技術（CNN, YOLO等）の発展により、従来の画像CAPTCHAは容易に突破されるように

(例: Plesner et al. (2024)[2] はYOLOを用いてreCAPTCHA v2を100%突破)

ユーザビリティの低下

- ボットに対抗するため、画像を難読化（ノイズ、歪み、低解像度化）
- 判定境界があいまい

結果、人間にとっても判別困難

「セキュリティとユーザビリティの深刻なトレードオフ」が発生

研究背景：CAPTCHAの役割

このトレードオフを解消し、
「人間には極めて優しく、機械には難しい」
次世代のCAPTCHAシステムを実現する

アプローチ

画像の「難読化」ではなく、「標的型攻撃による誘導」を利用

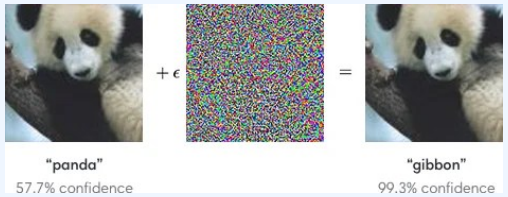
→標的型敵対的サンプルを用いて、ボットを特定の誤答（罠）へ誘導し、能動的に検知・排除

関連技術：敵対的サンプル

敵対的サンプル（Adversarial Examples）：入力データに微細な摂動（ノイズ）を加えることで、モデルの誤分類を誘発する攻撃手法

非標的型攻撃 (Untargeted)

「正解以外」なら何でも良い



[3]

例：パンダ→テナガザル？ 車？

※従来の敵対的CAPTCHAで主流

標的型攻撃 (Targeted)

攻撃者が指定した「特定のクラス」へ誘導する

例：パンダ→必ず「テナガザル」

※本研究ではこちらを採用

既存の敵対的CAPTCHA研究

研究	手法の特徴	課題
Osadchyら (2017) DeepCAPTCHA [4]	除去困難なノイズ(IAN)を提案	前処理耐性に注力。ユーザビリティについても既存CAPTCHAからの向上は目指していない
Shiら (2021) aCAPTCHA[5]	攻撃と防御のフレームワーク化 多様な前処理耐性を強化	防御性能とユーザビリティの両立を主張。ユーザビリティ検証は通常CAPTCHA（ノイズなし）と敵対的CAPTCHの比較
Teradaら (2022) *LFP[6]	低周波摂動を用い、視認性を向上	reCAPTCHAと同等のUXを達成したが、それ以上の向上は目指していない
Xuら(2024) CFA[7]	特徴空間への攻撃により、転移性を強化	未知のモデルに対する攻撃成功率（誤分類）の向上に特化。ボットを「特定の罠」へ誘導する視点は含まれない
Duら (2025) DAC[8]	拡散モデルを用いて自然な敵対的画像を生成	アイデア（セマンティック誘導）は本研究に近いが、CAPTCHAシステムとしての実装・検証までは行われていない

既存研究の課題

ユーザビリティ

既存研究は「機械を騙す」ことがメインで既存CAPTCHAからの負担減が実現していない

CAPTCHAシステムの構築

標的型攻撃の「誘導」特性を持つ研究（Duら）はあるにはあるが、それを具体的な「罠」としてシステムに組み込み、実用的な防御手段として確立した研究の不足

Immutable（不変）ではない

Osadchyら, Shiら, Teradaらともに画像前処理に対して「元のクラス」に戻らないことを防御成功としているが、標的クラスではなくなる可能性

提案手法のコンセプト(図に変更予定)

「人間と機械の認識のギャップを利用した罠の設置」

人間の視点

ノイズが少なく、自然な画像

カテゴリ判別（動物か乗り物か）は極めて容易

ストレスフリー

ボットの視点

標的型敵対的摂動により、全く異なるカテゴリ（標的）に見える

自身を持って「誤った正解（罠）」を選択

検知・排除

提案システム構成概要

本システムは、大きく３つのフェーズで構成される

1

生成フェーズ

RFCoAを用いた
敵対的画像生成とDB
化

2

提示フェーズ

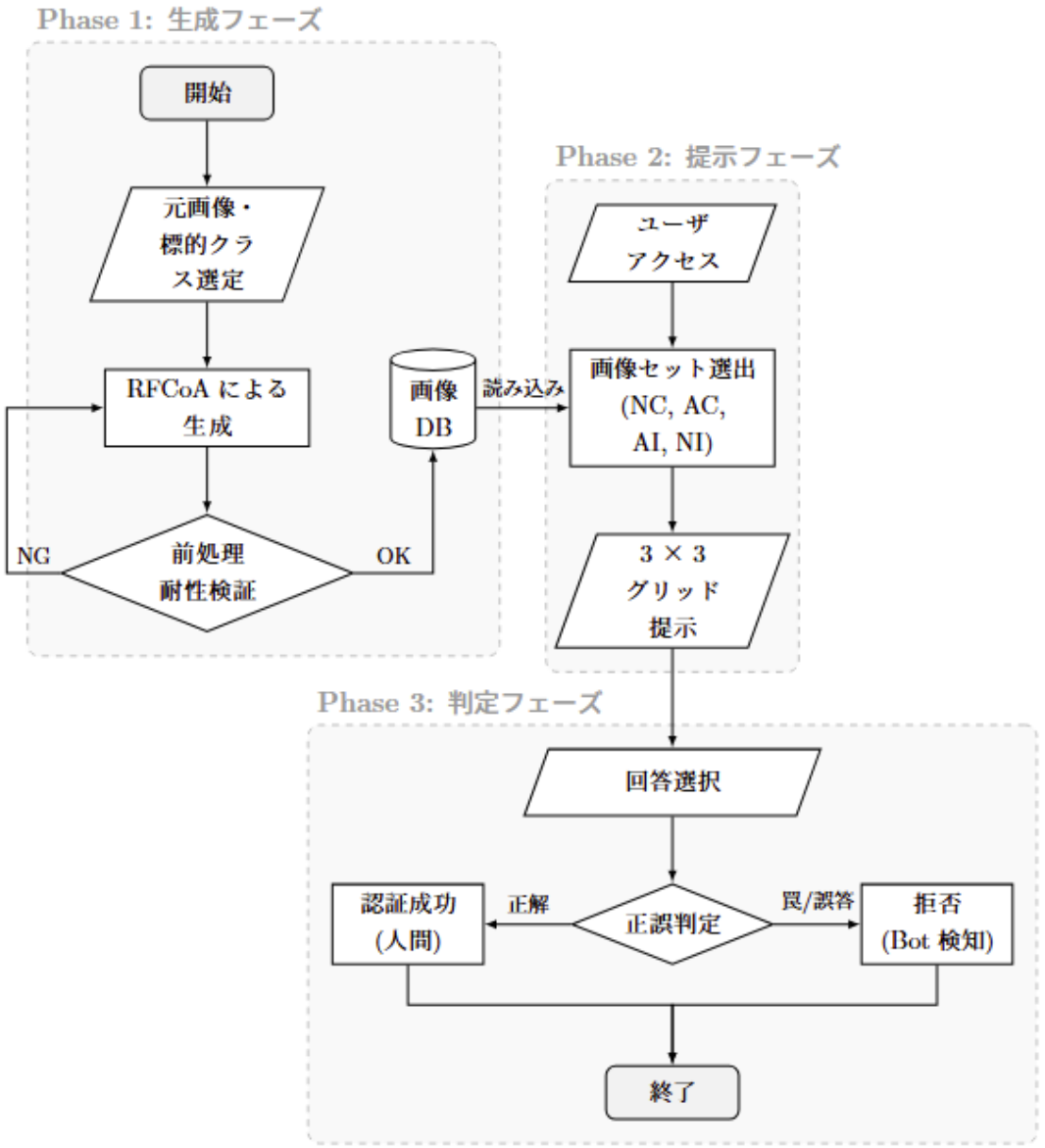
Web UI上での
３×３グリッド画像の
提示

3

判定フェーズ

選択画像の検証と
Bot判定ロジック

提案システムの処理フロー



生成フェーズ：RFCoA[9]の採用

RFCoAとは

“Robust Feature Coverage Attack” (AAAI 2025)

モデルが予測に用いる「堅牢な特徴（Robust Features)」に着目した最新の生成手法

採用理由

- 高い堅牢性：現実世界(生成画像を印刷し、撮影)に対しても高い成功率
  →環境変化に強い
- 高い転移性： 未知のブラックボックスモデルに対しても有効
- 自然な見た目：人間への視認性を損なわない

生成フェーズ：カテゴリ設定

ImageNetデータセットより、4つの大きなカテゴリを選定

動物　乗り物　楽器　食べ物

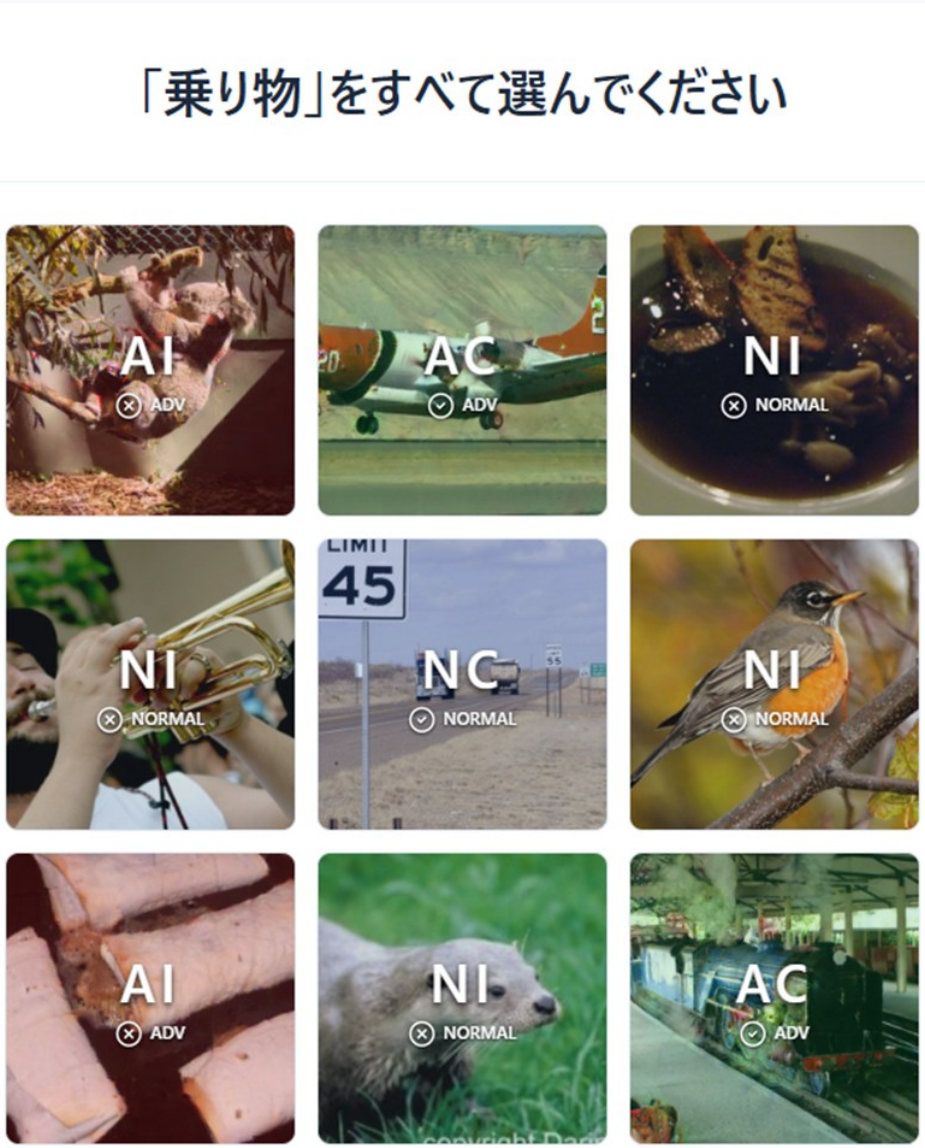
カテゴリ内堅牢性

画像前処理により、画像の予測クラスが変化しても、設定した標的クラスのカテゴリの中に留まっていれば防御成功とみなす基準を導入

生成フェーズ：カテゴリ設定

- 4カテゴリの中から組み合わせて、敵対的画像を生成
  （例）Original コアラ（動物）→Target コルネット（楽器）
- 5×5のメディアンフィルタ（ノイズ除去）
- フィルタ後の分類結果（カテゴリ内であればOK）
  - ◆Original コアラ（動物）→Target コルネット（楽器）→Filter コルネット（楽器）○
    - ➢ 真の不変
  - ◆Original コアラ（動物）→Target コルネット（楽器）→Filter フルート（楽器）○
    - ➢ 提案手法の不変
  - ◆Original コアラ（動物）→Target コルネット（楽器）→Filter アシカ（動物）×
    - ➢ 先行研究の不変
  - ◆Original コアラ（動物）→Target コルネット（楽器）→Filter ベーグル（食べ物）×

提示フェーズ



Type1の分類タスクに統一

画像タイプ	枚数	正誤	説明
Normal Correct (NC)	1	正解	加工なしの正解画像。人間もボットも「正解」と判断できるベースライン。
Adversarial Correct (AC)	2	正解	人間には正解に見えるが、ボットには不正解に見えるように加工された画像（ボット回避用）。
Adversarial Incorrect (AI)	2	不正解	人間には不正解に見えるが、ボットには正解に見えるように加工された画像(ボット検知用トラップ)。
Normal Incorrect (NI)	4	不正解	加工なしの不正解画像。人間もボットも「不正解」と判断するダミー。

判定フェーズ：ロジック詳細

ユーザの回答に基づき、以下の3つの状態に分類

認証成功

条件：NC 1枚＋AC 2枚を完全一致

正解画像を過不足なく

ボット検知

条件：AI 2枚とも選択
（トラップ2枚をすべて選択）

誘導された可能性が高いため、排除

認証失敗

条件：上記以外
（取りこぼし、選択ミスなど）

操作ミスを考慮し、再試行を促す

実験概要

実験目的

提案システムが、既存の標準的なCAPTCHAと比較して、ユーザビリティ（使いやすさ、負荷）において優れているかを検証する

実験概要

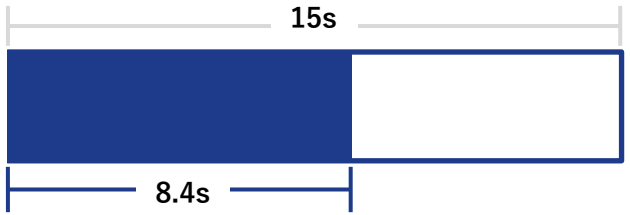
- 被験者：情報セキュリティ大学院大学関係者 31名（OB・OGを含む）
- 比較対象：reCAPTCHA v2 （Google）
- タスク：双方のCAPTCHAを5回ずつ解き、客観評価と主観評価を行う
- 評価指標：
  - 客観評価：正答率（目標90％以上）、回答時間（15秒以内）
  - 主観評価：NASA-TLX（簡便法）[10] 6つの指標で負荷の評価

客観評価：実用性の実証

情報セキュリティ大学院大学
INSTITUTE of INFORMATION SECURITY

平均解答時間

8.40s (<15s)



認証成功率

90.7% (>90%)

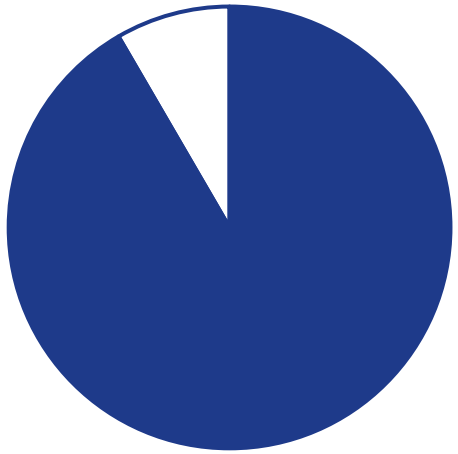


表 4.2　提案手法の客観評価結果（Macro Average: Mean ± SD）

指標	実験結果	目標基準
平均回答時間 †	**8.40 ± 2.01 秒**	15.0 秒以内
認証成功率 †	**90.7 %**	90.0 % 以上

† 回答時間に基づく IQR 法により，異常値を除外後の値を算出

（参考：除外前の全データにおける認証成功率は 88.9 %）

主観評価：NASA-TLX(簡便法)による負荷比較

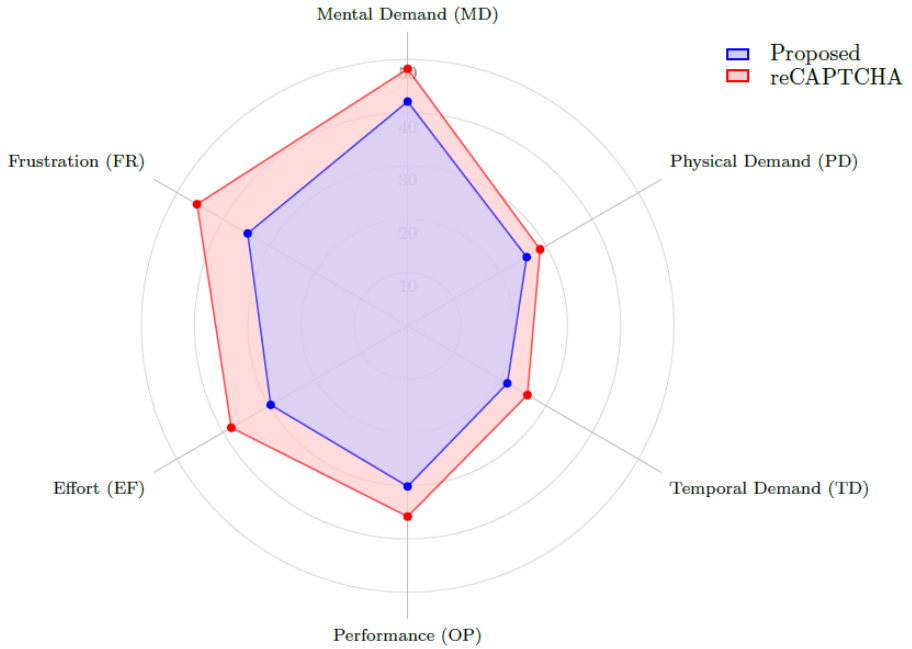


表 4.3　NASA-TLX の各指標および総合スコア（RawTLX）の比較

指標 (Scale)	Proposed	reCAPTCHA
知的・知覚的要求 (MD)	42.10	48.23
身体的要求 (PD)	25.81	28.71
タイムプレッシャー (TD)	21.61	25.97
作業成績 (OP)	30.16	35.81
努力 (EF)	29.68	38.23
フラストレーション (FR)	34.68	45.65
Total (RawTLX)	**30.67**	**37.10**

有意水準5% で統計的な有意差が認められた
（$t(30) = -2.05, p = 0.049 < 0.05$）

考察

■ ユーザビリティ

- Type１の分類タスクに統一したことにより、正解カテゴリと不正解カテゴリを意味的に分類することに成功
- reCAPTCHAの曖昧な境界判定などがないため、迷いから解放

■ セキュリティ

- 非標的型ではボットの誤認先の予測は不可能
- 標的型を用いることで、効率良くボットを検知する判断材料となりうる

課題と今後の展望

■ データセットの品質

ImageNetは画質・構図などに画像のばらつき

CAPTCHA専用の高品質データセットが必要

■ 文化的・意味的曖昧性

カテゴリの定義が文化圏で異なるため、より忠実なカテゴリ選出が必要

■ 被験者バイアス

今回は情報セキュリティ大学院大学の関係者内と限定的

より大規模で多属性な実験が必要

課題と今後の展望２

■ 生成AIの活用

より自然で高画質な画像を効率よく生成できるように

■ 真の不変性

カテゴリ内堅牢性からあらゆる前処理耐性を持つ完全な堅牢性へ

■ 厳密な比較

reCAPTCHAの各タイプを再現し、それぞれ比較実験

■ 実環境実装

判定ロジックを実装し、実戦形式で効果検証

まとめ

■ トレードオフの解消

標的型敵対的サンプルにより、シンプルなタスクのCAPTCHAを実現

■ 実用性の証明

平均解答時間8.40秒、成功率90.7％

■ ユーザビリティの向上

reCAPTCHAと比較し、統計的に有意な負荷低減を実証

※本番はSCIS同様デモも検討中

参考文献

1. Google. recaptcha v2. https://developers.google.com/recaptcha/docs/display?hl=ja. accessed: 2026-01-04.
2. Andreas Plesner, Tobias Vontobel, and Roger Wattenhofer. Breaking recaptchav2. In2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMP-SAC), pp. 1047–1056. IEEE, 2024.
3. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan,Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. arXivpreprint arXiv:1312.6199, 2013.
4. Margarita Osadchy, Julio Hernandez-Castro, Stuart Gibson, Orr Dunkelman, andDaniel P ´ erez-Cabo. No bot expects the deepcaptcha! introducing immutable ad-versarial examples, with applications to captcha generation. IEEE Transactions onInformation Forensics and Security, Vol. 12, No. 11, pp. 2640–2653, 2017.
5. Chenghui Shi, Xiaogang Xu, Shouling Ji, Kai Bu, Jianhai Chen, Raheem Beyah, andTing Wang. Adversarial captchas. IEEE transactions on cybernetics, Vol. 52, No. 7,pp. 6095–6108, 2021.
6. Takamichi Terada, Vo Ngoc Khoi Nguyen, Masakatsu Nishigaki, and Tetsushi Ohki.Improving robustness and visibility of adversarial captcha using low-frequency per-turbation. In International Conference on Advanced Information Networking andApplications, pp. 586–597. Springer, 2022.
7. Zisheng Xu and Qiao Yan. Boosting the transferability of adversarial captchas. Com-puters & Security, Vol. 145, p. 104000, 2024.
8. Xia Du, Xiaoyuan Liu, Jizhe Zhou, Zheng Lin, Chi-man Pun, Cong Wu, Tao Li, ZheChen, Wei Ni, and Jun Luo. Defensive Adversarial CAPTCHA: A Semantics-DrivenFramework for Natural Adversarial Example Generation . IEEE Transactions onDependable and Secure Computing, No. 01, pp. 1–13, November 2025.
9. CGCL-codes. RFCoA: Source code for Breaking Barriers in Physical-World Adversar-ial Examples. https://github.com/CGCL-codes/RFCoA, 2025. GitHub repository;source code accompanying the AAAI 2025 paper “Breaking Barriers in Physical-World Adversarial Examples: Improving Robustness and Transferability via RobustFeature”.
10. 芳賀繁, 水上直樹. 日本語版 nasa-tlx によるメンタルワークロード測定 各種室内実験課題の困難度に対するワークロード得点の感度. 人間工学, Vol. 32, No. 2, pp. 71–79,1996.