УДК 004.932.72'1

Д. В. Забаштанський
Т. Г. Кирилащук

ДОСЛІДЖЕННЯ МЕТОДІВ ГЕНЕРАЦІЇ ТА ВИЯВЛЕННЯ ФОТОРЕАЛІСТИЧНОЇ МАНІПУЛЯЦІЇ ОБЛИЧЧЯМ У ЦИФРОВИХ МЕДІА

Вінницький національний технічний університет

Анотація

У дослідженні розглянуто проблему створення та розповсюдження підробленого мультимедійного контенту. Проаналізовано основні техніки маніпуляції обличчям та ефективність сучасних методів виявлення фальсифікацій на основі глибокого навчання. Сформульовано рекомендації щодо підвищення стійкості систем ідентифікації до загроз з боку генеративної підробки.

Ключові слова: кібербезпека; біометрія; інформаційна безпека; діпфейк; розпізнавання обличчя.

RESEARCH OF COVERT DATA TRANSMISSION CHANNELS USING OPTICAL EMITTERS OF IOT DEVICES

Abstract

The paper investigates the problem of creation and distribution of fake multimedia content. Analyzes main techniques of facial manipulation and the effectiveness of modern methods of detecting forgeries based on deep learning. Formulates recommendations to increase the resilience of identification systems to threats from generative forgery.

Keywords: cybersecurity; biometry; information security; deepfake; face recognition.

Вступ

Проблема автентифікації правдивості візуального контенту набула критичного значення через появу методів маніпуляції на основі генеративно-змагальних мереж. Сучасні архітектури, такі як автоенкодери, здатні синтезувати фотореалістичні кадри, що зберігають ідентичність цільового об’єкта при будь-яких ракурсах зйомки. Така генеративна підробка кадрів і називається діпфейком.

На відміну від традиційного редагування, діпфейки створюють нелінійні спотворення на рівні текстури шкіри та освітлення, які часто є непомітними для людського ока. Це створює серйозні виклики для систем біометричної автентифікації, безпеки у соціальних мережах, цифрової криміналістики та довіри до медіа в цілому.

Актуальність дослідження зумовлена необхідністю розробки автоматизованих систем, здатних ідентифікувати синтетичні артефакти в умовах значного стиснення даних (наприклад, у соціальних мережах), де традиційні методи криміналістики втрачають ефективність. Особливої ваги набуває здатність детекторів працювати з низькою роздільною здатністю, оскільки більшість сучасних алгоритмів маніпуляції маскують помилки рендерингу саме через погіршення якості відеопотоку. Розробка стійких методів детекції, що базуються на аналізі мезоскопічних властивостей та часової когерентності кадрів, є критичною для забезпечення кібербезпеки в епоху масового розповсюдження синтетичного мультимедійного контенту.

Результати дослідження

1. Концептуальна модель атаки

Діпфейк зазвичай розуміють як медіа, згенероване нейроною мережою, яке виглядає реалістично і може бути сприйняте за реальне. Можна виділити 4 категорії таких синтетичних медіаданих:

1. Підробка (Reenacment)
2. Заміна (Replacement)
3. Редагування (Editing)
4. Синтез (Synthesis)

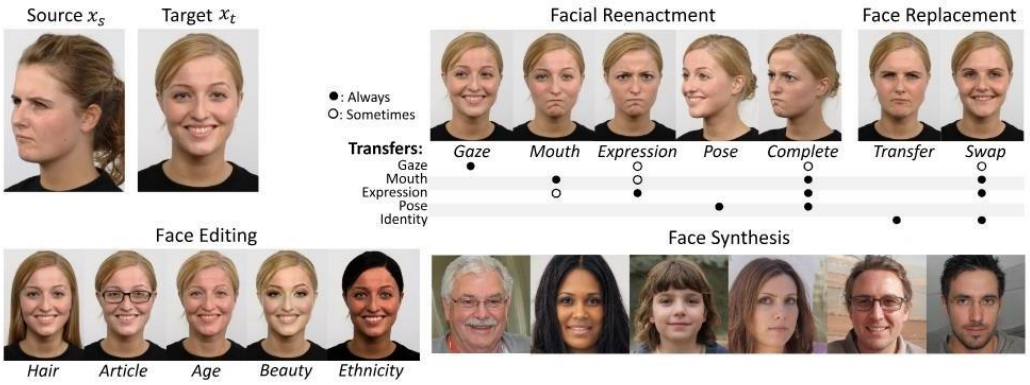

Рисунок 1 – Приклади 4 категорій діпфейків [1]

Підробка може виконуватись як для повного зображення людини, так і для його частини. Ця категорія використовує набір вихідних зображень для генерації іншого набору зображень, в якому замість людини, що була в них початково, буде інша людина, яка є ціллю підробки. Такі атаки є найбільш поширеними та можуть використовуватися у широкому спектрі можливих злочинів серед яких можуть бути: уособлення однією людиною іншої, генерування синтетичного неприємного контенту з участю цілі з метою шантажу та спроба підміни, спотворення чи створення фальших доказів кримінальної активності або спотворення сказаного людиною.

Заміна це вид синтетичних медіаданих, у яких люди, пристуні у фотографіях чи відео, можуть бути замінені цілями з метою шантажу неприємним контентом. Відмінність від підробки тут у тому, що він використовує уже готові кадри, заміняючи лице людини у медіа на лице цілі. Такий підхід залишає набагато менше артефактів, а отже серед справжнього контенту набагато складніше виявити підроблений.

Редагування та синтез використовують стокові фотографії та відео для зміни атрибутів, присутніх у них. Таким методом можна змінювати колір очей, форму та колір волосся, зріст, вагу, етнічну принадлежність та зовнішній вигляд загалом. Це дає змогу створення несправжніх персон в інтернеті з мінімальним ризиком бути викритим або для введення в оману людей для, наприклад, збору пожертвувань на лікування, зробивши так, що людина на фотографії виглядає дуже хворо, хоча насправді є здоровою.

При всій небезпеці потенціального застосування синтетичних мультимедіа кіберзлочинцями, діпфейки також використовуються звичайними користувачами у повсякденних та професіональних . Підробка може використовуватися для синхронізації перекладу з рухом губ, для фіксації погляду з метою покращення фотографій або для генерації та анімації рухів у відеоігровій індустрії на базі рухів акторів. Підміна зазвичай використовується для генерації жартів з популярними у медіа персонами, а редагування та синтез можуть бути корисні у створені ресурсів у фільмах та відеоіграх. З цього можна зробити висновок, що діпфейки можуть бути як корисную так і руйнівною технологією.

2. Стратегія протидії та захисту

Для мінімізації ризиків, пов’язаних із можливим застосуванням діпфейків кіберзлочинцями потрібно розуміти способи виялення синтетичних медіа у мережі. Сучасні підходи до ідентифікації синтетичного контенту базуються на виявленні специфічних артефактів, що виникають під час генерації та накладання (blending) цифрового обличчя на цільовий відеоряд. Процес детекції можна класифікувати за двома основними напрямками: аналіз просторових ознак окремих кадрів та

моніторинг часової когерентності відеопотоку.

На просторовому рівні ключова увага приділяється пошуку аномалій у межах змішування (blending boundaries), де нейромережеві класифікатори, такі як CNN, ідентифікують невідповідності в освітленні, градієнтах шуму та деталізації між фоном і переднім планом. Важливим аспектом є також використання методів цифрової криміналістики, зокрема аналіз «відбитків» генеративно-змагальних мереж (GAN fingerprints) та фото-відповіді нерівномірності сенсора (PRNU), що дозволяє виявити вклеєний контент навіть за умов значного стиснення даних.

Паралельно з цим, аналіз часових ознак дозволяє зафіксувати порушення фізіологічної та фізичної логіки, які важко відтворити за допомогою ШІ. Це включає моніторинг біометричних сигналів (пульсація крові, ритм кліпання), перевірку відповідності артикуляції (візем) вимовленим звукам (фонемам), моніторинг положення ключових точок обличчя (landmarks) для виявлення неприродних поз голови або порушень 3D-структури при рухах, моделювання манер та мімічних звичок конкретної особи, а також використання рекурентних мереж (LSTM) для виявлення мікроскопічних розривів та тремтіння (jitter) у послідовності кадрів.

Існує також більш архітектурний підхід , де використовуються ансамблі з декількох мереж, що можуть самостійно визначати ознаки підробки або замість класифікації «фейк/оригінал», модель навчається лише на реальних даних. Будь-яке відхилення від «нормального» маркується як згенерований контент. Це дозволяє виявляти нові типи діпфейків, на яких мережа не навчалася.

Висновки

У ході дослідження було встановлено, що технології маніпуляції мультимедійними даними, зокрема діпфейки, пройшли шлях від розважальних інструментів до серйозних загроз інформаційній та біометричній безпеці. Класифікація методів синтезу на підробку (Reenactment), заміну (Replacement), редагування (Editing) та повний синтез (Synthesis) дозволила виявити специфічні для кожного типу вразливості: від мікроскопічних артефактів змішування на межах масок до порушень часової когерентності у відеопотоці. Аналіз стратегій протидії демонструє, що найбільш ефективним підходом до детекції маніпуляцій є поєднання просторового аналізу (пошук GAN-відбитків та шумів PRNU) з моніторингом фізіологічних показників об'єкта (ритм кліпання, мікрорухи, синхронізація візем і фонем). Встановлено, що саме розробка універсальних моделей, здатних працювати в режимі реального часу, що забезпечить надійний рівень захисту в цифровому медіапросторі.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Mirsky Y., Lee W. The Creation and Detection of Deepfakes. *ACM Computing Surveys*. 2021. Т. 54, № 1. С. 1–41. URL: https://doi.org/10.1145/3425780 (дата звернення: 01.03.2026).Програмний захист даних. *Офіційний сайт CIOU*. URL: https://ciou.lissa.cx.ua/articles/programnij-zahist-danih-e.html (дата звернення: 15.02.2026).
2. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward / M. Masood та ін. *Applied Intelligence*. 2022. URL: https://doi.org/10.1007/s10489-022-03766-z (дата звернення: 01.03.2026).

Дмитро Володимирович Забаштанський – студент групи 1КІТС-24б, факультет менеджменту та інформаційної безпеки, Вінницький національний технічний університет, м. Вінниця, e-mail: dmitrozabastanskij@gmail.com;

Науковий керівник: Тетяна Генадіївна Кирилащук – асистент кафедри захисту інформації, Вінницький національний технічний університет, м. Вінниця;

Dmytro V. Zabashtanskiy – student of group 1KITS-24b, Faculty of Management and Information Security, Vinnytsia National Technical University, Vinnytsia, e-mail: dmitrozabastanskij@gmail.com;

Scientific Supervisor: Tetiana H. Kyrylashchuk – assistant of the Department of Information Protection, Vinnytsia National Technical University, Vinnytsia