Определение признаков детектирования дипфейков для формирования входного вектора распознавания

Д.А. Елизаров, А.С. Окишев, А.В. Микунов

Омский государственный университет путей сообщения

Аннотация: В работе рассматриваются ключевые признаки дипфейков и подходы к их распознаванию с помощью методов компьютерного зрения и машинного обучения. В ходе исследования были определены, проанализированы признаки детектирования дипфейков. На основе приоритетов были выделены признаки, обеспечивающие высокую точность распознавания и сделан вывод о значимости каждого признака.
Ключевые слова: генеративный искусственный интеллект, дезинформация, дипфейк, детектор дипфейков, кибербезопасность, мошенничество, признаки распознавания, анализ, вектора распознавания, машинное обучение, модель.

Введение

Современные технологии на основе искусственного интеллекта (ИИ), такие как генеративно-состязательная сеть (Generative Adversarial Network – GAN) и диффузионные модели, позволяют легко создавать реалистичные дипфейки – поддельные изображения лиц [1, 2]. Хотя эти инструменты полезны для кино и дизайна, они также используются в мошенничестве и дезинформации.

Для борьбы с этим разрабатываются методы обнаружения, основанные на анализе артефактов генерации: неестественных текстур, асимметрии лица и аномалий в освещении. В работе рассматриваются ключевые признаки дипфейков и подходы к их автоматическому распознаванию с помощью компьютерного зрения и машинного обучения.

Дипфейки становятся инструментом для дезинформации, манипуляции общественным мнением и подрыва доверия к медиа. Они могут использоваться для создания фейковых новостей, компрометации публичных фигур или подделки доказательств в судебных делах [3].

С 2023 году генеративный искусственный интеллект (ГИИ) стал ключевой технологией для киберпреступников, включая создание дипфейков

для фишинга, социальной инженерии и мошенничества. Например, поддельные видео с публичными личностями используются для продвижения мошеннических схем [4]. Общий ущерб, нанесенный мошенничествами с дипфейками, в первом квартале 2025 года превысил 200 млн. долларов. Около четверти дипфейков (23%), созданных в первом квартале 2025 года, использовались с целью финансового мошенничества [5].

1. Подходы для обнаружения дипфейков

Для обнаружения дипфейков применяются различные методы, которые можно разделить на несколько категорий: методы на основе машинного обучения, методы анализа визуальных артефактов, методы на основе метаданных, блокчейн и цифровая аутентификация.

Методы на основе машинного обучения используют свёрточные нейронные сети (Convolutional Neural Networks – CNN), которые обучаются на больших наборах данных, содержащих как настоящие, так и поддельные изображения, для выявления аномалий. Например, MesoNet и XceptionNet анализируют текстуры и артефакты в изображениях. Анализ скрытых представлений изображений (latent representations) позволяет выявить несоответствия, характерные для синтетического контента [6]. Комбинирование анализа изображений, аудио и метаданных для повышения точности обнаружения. Например, проверка синхронизации губ и речи в видео.

К методам анализа визуальных артефактов можно отнести:

- частотный анализ, который используется для выявления аномалий в высокочастотных компонентах изображения, которые часто возникают при генерации дипфейков (например, неестественные переходы в текстурах) [5];

- анализ временных рядов, который используется для выявления неестественных паттернов моргания или движений лица, которые можно выявить с помощью анализа временных рядов;
- обнаружение следов компрессии, которые могут оставлять генеративные модели могут, что отличает от естественных изображений.

Методы на основе стандарта метаданных «Формат файла обмена изображениями» (Exchangeable Image File Format – EXIF). Стандарт метаданных на основе встраиваемых в файлы изображений временные метки и данные о камере позволяет выявить несоответствия, указывающие на подделку.

Проверка цифровых подписей и водяных знаков, которые могут быть добавлены к оригинальному контенту для подтверждения его подлинности.

Использование блокчейн-технологий для создания цепочек доверия, где оригинальные медиа регистрируются с цифровыми подписями, что позволяет верифицировать их подлинность. Инициативы, такие как инициатива по обеспечению подлинности цифрового контента (Content Authenticity Initiative – CAI), разрабатывают стандарты для защиты медиа от подделок [7].

2. Значимые признаки для формирования входного вектора распознавания

Следует отметить, что дипфейки становятся сложнее обнаруживать из-за совершенствования генеративных моделей, таких как диффузионные модели, которые минимизируют артефакты [2, 8]. Также отсутствие универсальных стандартов и международного регулирования затрудняет борьбу с дипфейками, особенно в контексте кибервойн и дезинформации.

Для эффективного обнаружения дипфейков необходимо сформировать входной вектор, включающий признаки, которые наиболее чувствительны к

различиям между настоящими и синтетическими изображениями. Значимые признаки можно разделить на несколько групп: визуальные признаки, статистические признаки, метаданные и контекст, семантические признаки.

К параметрам визуальных признаков относятся текстурные аномалии (приоритет – высокий), несоответствие освещения (приоритет – средний, требует сложных алгоритмов анализа) и краевые артефакты (приоритет – средний, их значимость уменьшается с развитием генеративных моделей.).

Текстурные аномалии обусловлены различием в высокочастотных компонентах изображения, такие как неестественные переходы между пикселями или артефакты сжатия. Например, GAN часто оставляют специфические шумовые паттерны, которые можно выявить с помощью частотного анализа [9, 10]. Текстурные аномалии выбраны из-за их высокой чувствительности к артефактам, создаваемым генеративными моделями, такими как GAN или диффузионные модели. Эти модели часто не могут идеально воспроизвести сложные текстуры человеческой кожи или фона, что приводит к микроскопическим несоответствиям, обнаруживаемым при частотном анализе.

Дипфейки могут иметь несоответствия в тенях, отражениях или углах освещения, особенно если лицо было наложено на другое тело. Освещение является критически важным аспектом фотореалистичности. Генеративные модели часто не могут точно согласовать освещение лица и фона, что приводит к заметным несоответствиям, особенно в сложных сценах. Этот признак позволяет выявить подделки даже в высококачественных дипфейках. Публикации, такие как исследования по анализу освещения в дипфейках, демонстрируют, что алгоритмы анализа теней и отражений (например, основанные на физических моделях освещения) эффективно выявляют синтетические изображения

При наложении лица на изображение могут возникать разрывы или неестественные границы, особенно вокруг глаз, рта и волос. Краевые атаки возникают из-за ограничений в алгоритмах смешивания изображений (blending), используемых при создании дипфейков. Эти области имеют сложные формы и текстур, что делает их уязвимыми для ошибок генерации. В исследования подчеркивает, что краевые артефакты являются ключевым признаком, так как они часто заметны даже при анализе высокого качества изображений.

К параметрам статических признаков относятся распределение пикселей (приоритет – низкий, уязвим к новым технологиям) и энтропия изображения (приоритет – высокий, требует комбинации с другими признаками для повышения точности).

Анализ статистических характеристик, таких как гистограммы яркости или цветовых каналов, может выявить отклонения, характерные для синтетических изображений. Генеративные модели, такие как GAN или дифференциальные модели, часто создают изображения с неестественным распределением пикселей, особенно в цветовых каналах, из-за ограничений в моделировании сложных сцен. Этот признак позволяет выявить синтетические изображения даже при отсутствии визуальных артефактов. Исследования в области статистического анализа изображений (например, с использованием гистограмм и анализа главных компонент) показывают, что синтетические изображения имеют характерные отклонения в распределении пикселей, что делает этот метод устойчивым к улучшениям в генеративных технологиях.

Синтетические изображения часто имеют более низкую энтропию в определенных областях из-за ограничений генеративных моделей. Энтропия отражает степень хаотичности или сложности изображения. Генеративные модели часто упрощают текстуры в сложных областях (например, волосы

или фон), что приводит к снижению энтропии. Этот признак прост в вычислении и устойчив к визуальной маскировке. Работы по анализу энтропии в изображениях, используемые в задачах цифровой криминалистики, подтверждают, что синтетические изображения имеют характерные аномалии в энтропии, что делает этот признак практичным для обнаружения дипфейков, однако требует комбинации с другими признаками для высокой точности.

Метаданные, такие как EXIF, часто отсутствуют в синтетических изображениях или содержат несоответствия, так как генеративные модели не воспроизводят аппаратные характеристики камер. Этот признак позволяет быстро выявить подделку без сложного анализа изображения. Практики цифровой криминалистики показывают, что анализ метаданных является стандартным методом для проверки подлинности медиа. Отсутствие или несоответствие EXIF данных часто указывает на синтетическое происхождение. Приоритет признака – средний, так как легко подделываются и не всегда доступны;

Проверка соответствия фона, одежды или окружающей среды заявленному контексту изображения. Дипфейки часто создаются путем наложения лица на существующий фон, что может привести к несоответствиям в контексте, например, несоответствие освещения или стиля одежды. Этот признак позволяет выявить логические противоречия в сцене. Исследования в области анализа сцены и контекста (например, с использованием алгоритмов компьютерного зрения) подтверждают, что контекстный анализ эффективен для выявления дипфейков, особенно в случаях, где визуальные артефакты минимальны. Приоритет признака – низкий, так как зависит от контекста и легко подделывается.

Генеративные модели могут создавать анатомически некорректные лица, особенно при генерации с нуля или модификации. Такие аномалии, как

неестественные пропорции глаз или рта, легко обнаруживаются алгоритмами анализа лиц. Работы по анализу лиц (например, с использованием моделей 3D-реконструкции лица) показывают, что анатомические несоответствия являются надежным признаком, так как они связаны с биологическими ограничениями, которые сложно смоделировать. Приоритет признака – низкий, так как современные модели минимизируют анатомические ошибки.

На основании выше изложено сформируем основные признаки детектирования дипфейков для формирования входного вектора распознавания (рис. 1).

Рис. 1 – Признаки детектирования дипфейков

Заключение

В работе были определены, проанализированы признаки детектирования дипфейков и выделены признаки, обеспечивающие высокую точность распознавания. В заключении следует указать значимость каждого выделенного признака:

- визуальные и временные признаки имеют высокую дискриминационную способность, так как генеративные модели пока не могут полностью воспроизвести естественные паттерны человеческой мимики или освещения;
- частотный анализ и анализ энтропии устойчивы к совершенствованию генеративных моделей, так как даже передовые

- диффузионные модели оставляют микроскопические следы, которые можно выявить;
- метаданные и статистические признаки легко извлекаются из изображений и не требуют сложных вычислений, что делает их практичными для реального применения;
- комбинирование нескольких типов признаков (например, текстурных, временных и метаданных) повышает точность и снижает вероятность ложных срабатываний.

Литература

1. Катаев А.В., Власова Ю.М., Ким В.А., Гусынин Д.А. Сравнительный анализ современных методов генерации изображений: VAE, GAN и диффузионные модели // Инженерный вестник Дона. 2025. №5. URL: ivdon.ru/ru/magazine/archive/n5y2025/10060.

2. Катаев А.В., Власова Ю.М., Ким В.А., Гусынин Д.А. Обзор метрик с целью оценки качества работы генеративных моделей для создания изображений // Инженерный вестник Дона. 2025. №6. URL: ivdon.ru/ru/magazine/archive/n6y2025/10119.

3. Гуселетова А.Е, Елизаров Д.А. Инструменты обнаружения дипфейков // Международная научно-практическая конференция «Актуальные проблемы и тенденции развития современной экономики и информатики». Бирск: Уфимский университет науки и технологий, 2024. С. 177-180.

4. История года: влияние искусственного интеллекта на кибербезопасность. URL: securelist.ru/story-of-the-year-2023-ai-impact-on-cybersecurity/108558/ (дата обращения: 18.11.2025).

5. Q12025 Deepfake Incident Report. URL: resemble.ai/q1-2025-ai-deepfake-security-report/ (дата обращения: 18.11.2025).

6. Abdullah S.M., Cheruvu A., Kanchi S., Chung T., Gao P., Jadliwala M., Viswanath B. An Analysis of Recent Advances in Deepfake Image Detection in an

Evolving Threat Landscape // IEEE Symposium on Security and Privacy. 2024. URL: https://arxiv.org/pdf/2404.16212v1 (дата обращения: 18.11.2025).

7. C2PA Specifications. URL: c2pa.org/specifications/ (дата обращения: 18.11.2025).

8. Ho J., Jain A., Abbeel P. Denoising Diffusion Probabilistic Models // Conference on Neural Information Processing Systems. 2020. URL: arxiv.org/pdf/2006.11239 (дата обращения: 18.11.2025).

9. Румянцева М.С. Разработка и исследование алгоритма обнаружения дипфейков с использованием двухпоточной нейронной сети и частотного анализа // Вестник науки. 2025. Т. 2, № 6. С. 1763-1773.

10. Frank J., Eisenhofer T., Schönherr L., Fischer A., Kolossa D., Holz T. Leveraging Frequency Analysis for Deep Fake Image Recognition // International Conference on Machine Learning. 2020. URL: proceedings.mlr.press/v119/frank20a/frank20a.pdf (дата обращения: 18.11.2025).

References

1. Kataev A.V., Vlasova YU.M, Kim V.A., Gusynin D.A. Inzhenernyj vestnik Dona. 2025. №5 URL: ivdon.ru/ru/magazine/archive/n5y2025/ 10060.

2. Kataev A.V., Vlasova YU. M., Kim V.A., Gusynin D.A. Inzhenernyj vestnik Dona. 2025. №6 URL: ivdon.ru/ru/magazine/archive/n6y2025/ 10119.

3. Guseletova A.E, Elizarov D.A. Mezhdunarodnaya nauchno-prakticheskaya konferenciya “Aktual'nye problemy i tendencii razvitiya sovremennoj ekonomiki i informatiki”: trudy (Proc. International Scientific and Practical Symp. “Current problems and trends in the development of modern economics and computer Science”). Birsk, 2024. pp. 177-180.

4. Istoriya goda: vliyanie iskusstvennogo intellekta na kiberbezopasnost' [The story of the year: the impact of artificial intelligence on cybersecurity]. URL: securelist.ru/story-of-the-year-2023-ai-impact-on-cybersecurity/108558/.

5. Q12025 Deepfake Incident Report. URL: resemble.ai/q1-2025-ai-deepfake-security-report/.

6. Abdullah S.M., Cheruvu A., Kanchi S., Chung T., Gao P., Jadliwala M., Viswanath B. IEEE Symposium on Security and Privacy. 2024. URL: arxiv.org/pdf/2404.16212v1.

7. C2PA Specifications. URL: c2pa.org/specifications.

8. Ho J., Jain A., Abbeel P. Conference on Neural Information Processing Systems. 2020. URL: arxiv.org/pdf/2006.11239.

9. Rumyanceva M.S. Vestnik nauki. 2025. T. 2, № 6. pp. 1763-1773.

10. Frank J., Eisenhofer T., Schönherr L., Fischer A., Kolossa D., Holz T. International Conference on Machine Learning. 2020. URL: proceedings.mlr.press/v119/frank20a/frank20a.pdf.

Авторы согласны на обработку и хранение персональных данных.

Дата поступления: 14.11.2025

Дата публикации: 27.12.2025