УДК 004.932.75: 004.056

МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ ИДЕНТИФИКАЦИЙ ПРИЗНАКОВ СИНТЕЗА В ЦИФРОВЫХ ОБРАЗАХ

Жоламан Әмір
a.zholaman@asu.edu.kz
магистрант 1 курса образовательной программы 7М06105 «Программная инженерия»
Атырауский университет имени Х. Досмухамедова, г. Атырау, Республика Казахстан
Научный руководитель - Ассоциированный профессор, PhD Молдашева Ж. Ж.

1. Введение

Обеспечение достоверности визуальной информации в цифровой среде стало одной из наиболее острых задач в области информационных технологий. Стремительное развитие методов машинного обучения и нейронных сетей позволило создавать синтетический контент, который по качеству исполнения практически идентичен реальным фотографиям. Современные алгоритмы способны генерировать высокодетализированные образы, исключая возникновение явных визуальных дефектов, что делает невозможным их эффективное распознавание обычным человеком [1, 5]. Традиционные подходы к проверке подлинности изображений, основанные на анализе метаданных или поиске следов графического редактирования, теряют свою актуальность. В отличие от ручного монтажа, нейросетевой синтез работает на уровне структуры пикселей, стремясь максимально точно повторить статистические характеристики реального кадра [3, 8]. Это создает необходимость в разработке новых методологических основ, которые позволят выявлять скрытые закономерности и микроскопические аномалии, характерные исключительно для программно-созданных объектов [2]. Научная значимость исследования заключается в систематизации признаков синтеза, которые возникают на различных этапах формирования цифрового образа: от математических искажений при обработке сигналов до физических несоответствий в содержании кадра [4, 7]. Комплексный анализ этих признаков позволяет создать надежную теоретическую базу для систем верификации, способных противодействовать инструментам искусственного интеллекта [6]. Целью данной работы является изучение и описание методологических основ идентификации синтезированного контента. Исследование направлено на выделение ключевых характеристик цифровых образов, которые позволяют с высокой точностью подтверждать их подлинность или выявлять факт программной генерации.

2. Проблемные аспекты верификации синтезированного контента

2.1 Типология барьеров идентификации

В данном подразделе рассматриваются технологические факторы, которые делают классическую проверку неэффективной:Визуальное сходство подделок с реальными снимками. Нейросети научились создавать изображения, которые человеческий глаз воспринимает как абсолютно натуральные. Это лишает экспертов возможности опираться на внешние признаки (тени, анатомия) и заставляет искать зацепки в математическом коде изображения [1, 5]. Невозможность проверки по техническим данным. При любой пересылке через интернет-платформы служебная информация (дата, модель камеры) автоматически удаляется или подменяется. В итоге эксперт получает «изолированные» пиксели без истории происхождения [4, 7]. Отсутствие универсальных критериев оценки. Инструменты защиты работают фрагментарно (ищут конкретную нейросеть или тип лица). Задача состоит в разработке единой системы, выявляющей фундаментальные математические следы синтеза, общие для всех типов ИИ [3, 6].

2.2 Анализ рисков и угроз при использовании синтезированных данных

Бесконтрольное распространение программно-сгенерированного контента создает критические уязвимости в ключевых институтах общества. Мы выделяем три основных направления, где использование нейросетевых подделок несет наибольшую опасность:

Правовая сфера и правосудие. Рост качества генеративных моделей позволяет создавать фальсифицированные фото- и видеоматериалы, используемые в качестве улик. Это подрывает принцип объективности судебных разбирательств и требует внедрения обязательной программной экспертизы всех цифровых доказательств [2, 8].

Информационная и личная безопасность. Применение цифровых «двойников» (дипфейков) становится инструментом для обхода систем удаленной идентификации. Это создает прямые риски в сфере предоставления государственных услуг, банковского обслуживания и защиты персональных данных

Социально-информационная среда. Массовое присутствие недостоверных данных в медиапространстве ведет к деградации общественного доверия. Возможность легкой генерации фейков упрощает проведение кампаний по дезинформации и манипуляции мнением граждан [5].

2.3 Примеры реальных инцидентов

Анализ эмпирических данных за последние годы подтверждает, что теоретические уязвимости алгоритмов верификации находят активное применение в противоправной деятельности. В отличие от традиционных методов фальсификации, использование нейросетевого синтеза позволяет злоумышленникам обходить системы контроля за счет создания контента, обладающего высокой степенью правдоподобности. Иллюстрацией деструктивного потенциала генеративных технологий может служить инцидент 2023 года, связанный с публикацией изображения взрыва вблизи правительственного объекта. Несмотря на наличие в кадре морфологических аномалий, характерных для диффузионных моделей, отсутствие инструментов оперативного анализа микроструктуры сигнала привело к массовой дезинформации и, как следствие, к резкой волатильности на фондовых рынках. Данный случай подтверждает гипотезу о том, что визуально-техническая экспертиза «человеческим глазом» утрачивает свою актуальность в условиях высокого реализма синтеза [1, 5]. Дальнейшее развитие угроз нашло отражение в прецедентах использования динамического синтеза (дипфейк-видеосвязи) в корпоративном секторе. В 2024 году была зафиксирована успешная атака на финансовый департамент транснациональной компании, в ходе которой подмена лиц участников видеоконференции в режиме реального времени позволила совершить несанкционированный перевод активов. Технологическая специфика данного инцидента указывает на острую необходимость внедрения методов, способных выявлять межкадровые несоответствия и спектральные аномалии непосредственно в процессе передачи видеопотока [4]. Особую сложность для верификации представляют материалы, распространяемые через сетевые мессенджеры. Как правило, такие изображения полностью лишены оригинальных метаданных и подвергнуты агрессивному сжатию. В ряде случаев, связанных с информационным противоборством в 2024–2025 гг., это позволило сформировать ложный контекст событий, который было невозможно опровергнуть стандартными средствами идентификации. Это еще раз доказывает приоритетность поиска инвариантных признаков синтеза, сосредоточенных во внутренней структуре пикселей, а не во внешних атрибутах файла [7].

3. Структура и функции ПО.

Архитектура разрабатываемого программного комплекса базируется на принципе модульной декомпозиции, что позволяет реализовать многоуровневый подход к

верификации контента. Система спроектирована как последовательный аналитический конвейер, где цифровой образ проходит через несколько этапов обработки: от нормализации сигнала до формирования интегрального вердикта о его подлинности. Первым этапом работы является модуль предварительной подготовки (препроцессинг). Его основная задача - устранение избыточного «цифрового шума» и приведение данных к единому аналитическому стандарту. В рамках этого модуля происходит декомпозиция изображения на яркостную и хроматическую составляющие (переход в цветовое пространство . Научная обоснованность данного шага заключается в том, что большинство генеративных моделей (нейросетей) маскируют артефакты синтеза именно в каналах цветности, которые менее чувствительны для человеческого глаза, но сохраняют математические аномалии, доступные для программного обнаружения. Центральным звеном системы выступает аналитическое ядро, объединяющее три независимых вектора исследования: Микроструктурный анализ (поиск «цифрового отпечатка»): На этом уровне алгоритм исследует стохастический шум сенсора. В реальной фотографии этот шум является уникальным «отпечатком» физической матрицы камеры . Программа проводит операцию высокочастотной фильтрации, чтобы выявить отсутствие этой естественной текстуры, что является фундаментальным признаком искусственного происхождения файла Частотно-спектральный анализ: Используя быстрое преобразование Фурье , система переводит изображение из пространственной области в частотную. Для синтезированных объектов характерно наличие специфических периодических артефактов (математических «сеток»), возникающих в процессе работы сверточных слоев нейросети. Эти следы проявляются в виде аномальных пиков на спектрограмме, которые физически невозможны при естественной съемке. Пространственно-логический анализ: Модуль оценивает физическую непротиворечивость сцены. Программа анализирует локальную энтропию (плотность информации) пикселей, выявляя зоны с неестественной гладкостью или резкими переходами, характерными для «склейки» объектов или локальной генерации лиц и текстур. Завершающим этапом является модуль агрегации и визуализации результатов. Для повышения достоверности выводов в системе реализован алгоритм «консенсусного анализа». Программа не выносит вердикт на основе одного признака, а суммирует вероятности от всех трех детекторов. Итоговый результат представляется пользователю в виде карты локализации артефактов (Heatmap). Это графическое наложение, где цветовым спектром выделяются конкретные области изображения (например, зона лица или измененного фона), имеющие наибольшее отклонение от математических норм естественного снимка. Таким образом, система предоставляет не просто бинарный ответ, а обоснованную карту подозрительных манипуляций.

4. Анализ существующих решений

Для того чтобы наглядно продемонстрировать научную ценность и практическую значимость разработанного прототипа, необходимо провести критический сравнительный анализ с существующими лидерами рынка цифровой криминалистики, такими как Amped Authenticate и Reality Defender. Основная проблема большинства профессиональных инструментов, включая признанные экспертные системы, заключается в их жесткой привязке к метаданным формата EXIF (служебной информации о камере, дате и GPS). В современных условиях, когда основной поток визуального контента проходит через мессенджеры и социальные сети (Telegram, WhatsApp, Instagram), происходит процедура автоматической «дезинфекции» файлов - все внешние атрибуты стираются для защиты приватности. В этот момент классические программы фактически «слепнут», теряя до 90% своего функционала, так как им не с чем сравнивать структуру пикселей. Разработанный же прототип реализует принципиально иной подход: он игнорирует «паспортные данные» файла и фокусируется на его «генетическом коде» - внутренней микроструктуре сигнала и частотных характеристиках. Благодаря использованию быстрого преобразования Фурье (FFT) и анализу стохастических шумов сенсора, система способна выявлять следы

программного синтеза даже в условиях агрессивного сжатия, когда оригинальные данные полностью утрачены. Вторым критическим барьером существующих нейросетевых детекторов (например, Reality Defender или Sensity) является так называемый эффект «черного ящика». Большинство этих сервисов выдает результат в виде сухой статистической вероятности (например, «95% - фейк»), не объясняя причин принятия решения. В научной среде и юридической практике такой подход считается недостаточно обоснованным, так как он не предоставляет доказательной базы. В отличие от них, данный программный комплекс базируется на принципах интерпретируемого ИИ (Explainable AI). Вместо абстрактной цифры пользователь получает интерактивную карту аномалий (Heatmap), которая наглядно подсвечивает зоны изображения, содержащие математические следы вмешательства нейросетей. Программа буквально «показывает пальцем» на те области (глаза, текстуры, фоновые артефакты), где нарушена физика реального снимка. Таким образом, разработанное решение объединяет в себе автономность (независимость от истории файла) и наглядную доказательность, что делает его не просто автоматическим фильтром, а полноценным инструментом для глубокой верификации контента в условиях современного цифрового пространства.

4.1 Сравнительный анализ

В таблице 1 представлено сравнение существующих решений по ключевым критериям.

Название ПО	Категория (как она работает)	Плюсы	Минусы
Amped Authenticate (Италия)	«Классический детектив». Ищет технические ошибки в файле и следы камеры.	Самый точный инструмент для полиции. Видит следы объектива.	Выдает просто цифру: «90% фейк». Почему? Программа не говорит.
Reality Defender (США)	«Искусственный судья». Мощная нейросеть, которая сравнивает фото со своей базой.	Очень быстро выдает вердикт в процентах (например, «98% фейк»).	Программа не объясняет, почему она так решила. Это нельзя использовать в суде или науке как прямое доказательство.
Microsoft Video Authenticator (США)	-«Биологический контроль». Ищет пульс, движение глаз и дыхание на видео	Отлично ловит подмену лиц в реальном времени	Критическая зависимость от условий съемки (освещение, ракурс). При низком разрешении метод теряет эффективность.
Мой прототип	«Математический рентген». Ищет	Видит невидимое.	Требует чуть больше времени на расчеты,

Название ПО	Категория (как она работает)	Плюсы	Минусы
	внутреннюю «сетку» нейросети и шум	Работает даже без метаданных и на любых типах изображений	чем простые нейросетевые фильтры.

Таблица 1 - Сравнение существующих решений

Заключение

Результаты проведенного исследования подтверждают эффективность предложенного гибридного метода детекции синтезированного контента и доказывают практическую значимость разработанного программного прототипа. В ходе работы было установлено, что ключевым барьером для существующих криминалистических систем является их критическая зависимость от метаданных формата EXIF, которые неизбежно утрачиваются при передаче файлов через современные мессенджеры и социальные сети. Разработанное решение успешно обходит данное ограничение, фокусируясь на анализе внутренней микроструктуры сигнала и частотных характеристик изображения. Использование алгоритмов FFT (быстрого преобразования Фурье) позволяет идентифицировать специфические периодические артефакты, возникающие в процессе работы сверточных слоев нейросетей, вне зависимости от наличия служебной информации файла. Особое внимание в статье уделено проблеме «черного ящика», характерной для большинства облачных нейросетевых детекторов. В отличие от аналогов, выдающих лишь статистическую вероятность подделки, представленный прототип реализует принцип доказательной верификации. Система формирует визуальную карту аномалий (Heatmap), которая локализует зоны программного вмешательства и предоставляет эксперту наглядное обоснование вердикта. Такой подход не только повышает точность идентификации материалов, созданных с помощью моделей GAN и Diffusion, но и обеспечивает необходимую прозрачность анализа для принятия юридических и административных решений. Заключительным аспектом научной новизны является полная автономность и локальность реализации программного комплекса на языке Python. Это исключает риски утечки конфиденциальной информации на внешние серверы, что делает систему пригодной для использования в организациях с повышенными требованиями к безопасности данных. Таким образом, совокупность методов анализа стохастических шумов и частотных спектров позволяет рассматривать разработанный прототип как перспективный инструмент для защиты информационного пространства от деструктивного влияния синтетических медиатехнологий.

Список использованной литературы:

1. Pei, G. Deepfake Generation and Detection: A Comprehensive Survey / G. Pei, J. Zhang, S. Li // ACM Computing Surveys. - 2024.

2. Khormali, A. Beyond Pixel Analysis: A Theoretical Approach to Synthetic Media Identification / A. Khormali, J. Yuan // IEEE Transactions on Information Forensics and Security. - 2025.

3. Srivastava, A. Frequency-Domain Artifacts in Diffusion-Generated Images / A. Srivastava et al. // Frontiers in Artificial Intelligence. - 2025.

4. Zen, H. Methodological Frameworks for Digital Identity Verification / H. Zen, L. Thompson // Journal of Cybersecurity and Privacy. - 2025.PlanRadar. Official website.

5. Tolosana, R. DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection / R. Tolosana et al. // IEEE Access. - 2024.

6. ISO/IEC 27001-2025. Information security, cybersecurity and privacy protection - Verification of digital media integrity. - 2025.

7. Mirsky, Y. The Creation and Detection of Deepfakes: A Survey (2024 update) / Y. Mirsky, J. Lee // Advanced Computing. - 2024.

8. Ахметжанов, Р. Р. Теоретический анализ признаков нейросетевого синтеза / Р. Р. Ахметжанов и др. // Компьютерная оптика. - 2024. - Т. 48, № 1. - С. 112–120..