Теоретические и прикладные аспекты кибербезопасности конвергентных когнитивно-информационных технологий

https://doi.org/10.25559/SITITO.021.202502.204-220
УДК 004.891.3

Методы обнаружения дипфейков в видеоконференциях в реальном времени

И. Д. Кузьменко*, Д. Е. Намиот, В. А. Васенин

ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова», г. Москва, Российская Федерация

Адрес: 119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1

* ilyexakuzmenko@gmail.com

Оригинальная статья

Аннотация

В последние годы видеоконференции приобретают все более широкий размах, став неотъемлемым инструментом для проведения деловых совещаний, образовательных мероприятий и даже официальных правительственных встреч. Стремительное развитие технологий интернет-связи и доступность платформ видеоконференций (таких как Zoom, Microsoft Teams и Google Meet) способствуют переходу множества организаций на гибридные и дистанционные форматы работы. В результате глобальная аудитория пользователей онлайн-встреч исчисляется сотнями миллионов, и это число продолжает расти. Одновременно с расширением сферы применения видеоконференций возникает новая волна угроз, связанных с безопасностью и доверием участников. Среди таких угроз особенно выделяется феномен «дипфейков» *(от англ. deepfakes)*, то есть синтетически сгенерированных или модифицированных аудио- и видеозаписей, которые практически невозможно отличить от оригинала невооруженным глазом. В работе рассматривается вопрос детектирования дипфейков в реальном времени в видеоконференциях.

Ключевые слова: кибербезопасность, видеоконференция, дипфейк

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Кузьменко И. Д., Намиот Д. Е., Васенин В. А. Методы обнаружения дипфейков в видеоконференциях в реальном времени// Современные информационные технологии и ИТ-образование. 2025. Т. 21, № 2. С. 204-220. https://doi.org/10.25559/SITITO.021.202502.204-220

**Theoretical and Applied Aspects of Cybersecurity
of Convergent Cognitive Information Technologies**

Methods for Detecting Deepfakes in Real-Time Video Conferences

I. D. Kuzmenko*, D. E. Namiot, V. A. Vasenin

Original article

Lomonosov Moscow State University, Moscow, Russian Federation

Address: 1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation

* ilyexakuzmenko@gmail.com

Abstract

In recent years, video conferencing has become increasingly popular, becoming an essential tool for business meetings, educational events, and even official government meetings. The rapid development of Internet communication technologies and the availability of video conferencing platforms (such as Zoom, Microsoft Teams, and Google Meet) are driving many organizations to adopt hybrid and remote work formats. As a result, the global audience of online meeting users is in the hundreds of millions, and this number continues to grow. Along with the expansion of the scope of video conferencing, a new wave of threats related to the security and trust of participants is emerging. Among such threats, the phenomenon of "deepfakes" stands out in particular; that is, synthetically generated or modified audio and video recordings that are almost impossible to distinguish from the original with the naked eye. This paper examines the issue of real-time deepfake detection in video conferencing.

Keywords: cybersecurity, video conference, deepfake

Conflict of interests: The authors declares no conflict of interest.

For citation: Kuzmenko I.D., Namiot D.E., Vasenin V.A. Methods for Detecting Deepfakes in Real-Time Video Conferences. *Modern Information Technologies and IT-Education.* 2025;21(2):204-220. https://doi.org/10.25559/SITITO.021.202502.204-220

Введение

Технологии генерации дипфейков (на базе методов глубокого обучения и в первую очередь *GAN – Generative Adversarial Networks*) позволяют достигать такой реалистичности картинки, что внешне поддельные видеоролики или трансляции выглядят крайне убедительно [1]. Возможности применения дипфейков включаются в модели рисков ИИ [2].

Использование дипфейков в ходе видеоконференций открывает злоумышленникам широкие возможности для манипуляций [3]:

- подделка идентичности (участник может «выдать» себя за руководство компании или за доверенное лицо);
- мошенничество (например, получение конфиденциальных данных или совершение незаконных финансовых транзакций под видом лица, обладающего необходимыми правами);
- репутационный ущерб (распространение компрометирующих видеоматериалов, порочащих деловую репутацию).

В совокупности все эти факторы обусловливают высокую актуальность исследования, направленного на развитие систем автоматического обнаружения дипфейков в реальном времени. Традиционные методы проверки подлинности видео зачастую оказываются недостаточно эффективными, когда речь идет об оперативной защите пользователей на онлайн-платформах с большим объемом видеоданных. Необходимо разрабатывать алгоритмы, способные мгновенно анализировать входящий поток, выявлять потенциально поддельные участки и сигнализировать о возможной угрозе.

Несколько реальных примеров. По данным полиции Гонконга, финансовый работник многонациональной компании был обманут, заставив выплатить 25 миллионов долларов мошенникам, использовавшим технологию deepfake, чтобы выдать себя за финансового директора компании во время видеоконференции.

В ходе сложной аферы работника обманом заставили принять участие в видеоконференции с теми, кого он считал несколькими другими сотрудниками, но все они были поддельными копиями, заявила полиция Гонконга. «В многопользовательской видеоконференции оказалось, что все, кого он видел, были поддельными»[1]. У сотрудника возникли сомнения, когда он получил письмо от финансового директора, в котором говорилось о необходимости секретной транзакции. Однако видеозвонок с командой был настолько убедительным, что ему пришлось отбросить сомнения и перевести деньги. Сотрудник обнаружил, что это мошенничество, когда он проверил в главном офисе корпорации.

Еще в 2019 году генеральный директор неназванной британской энергетической компании стал жертвой технологии deepfake. Предположительно, он был на связи с генеральным директором своей компании в Германии (материнской компании). Руководитель приказал ему немедленно перевести 220 000 евро (около 243 000 долларов) на банковский счет венгерского поставщика.

Клонирование голоса было настолько совершенным, что сохранило мелодию и тонкий немецкий акцент голоса его босса. Только когда самозванец позвонил несколько раз, генеральный директор засомневался[2].

Именно обзору подходов к определению (предотвращению использования) дипфейков в реальном времени и посвящена данная статья.

Существующий объем исследований по генерации и детекции дипфейков достаточно обширен.

Методы создания фейковых изображений и видео развиваются благодаря успехам в глубоких нейросетевых архитектурах – GAN [4], автокодировщиках [5], методах замены и анимации лиц. Детектирование традиционно включает анализ артефактов сшивания, асинхронности речи и мимики, неестественного освещения и частотных искажений [6].

Однако большая часть решений ориентирована на офлайн-обработку заранее записанных видео, что требует значительных ресурсов и времени. В условиях онлайн-видеоконференций такие методы часто неприменимы. Более того, использование высококачественного оборудования и прогрессивных генеративных моделей значительно усложняет задачу детекции. Интеграция подобных решений в популярные платформы видеосвязи также остается нерешенной задачей. Несмотря на наличие API у Zoom, Teams и других, готовых и устойчивых решений с распознаванием дипфейков на лету – крайне мало. Также остается актуальной проблема адаптации моделей под новые формы фальсификации.

Анализ существующих методов создания и обнаружения дипфейков

Технологии синтеза поддельных видео с заменой лица *(face swapping)* прошли значительный путь развития: от ранних методов компьютерной графики до современных подходов на основе глубинного обучения. Еще в 1990-х годах были продемонстрированы первые системы, способные изменять видимое содержание лица в видео. Например, метод Video Rewrite (1997) позволял автоматически генерировать видеозапись человека с синтезированными движениями губ под заданную аудиодорожку. В последующие годы появилось множество алгоритмов классической компьютерной графики для замены лица. Один из ранних заметных результатов – система Video Face Replacement (2011), где на основе 3D-модели лица источника и приемника выполнялась замена лица в видео с сохранением мимики последнего [7].

Такие подходы основывались на трудоемкой 3D-реконструкции и трекинге лица, но заложили фундамент для более совершенных real-time методов. С развитием аппаратных возможностей стали появляться интерактивные технологии реконструкции и переноса мимики в реальном времени. Одной из ключевых работ в этой области стал метод Face2Face (2016) – система,

[1] Chen H., Magramo K. Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’ [Электронный ресурс] // CNN World. February 4, 2024. URL: https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk (дата обращения: 16.04.2025).

[2] Top 10 Examples of Deepfake Across The Internet // HyperVerge, 2025. URL: https://hyperverge.co/blog/examples-of-deepfakes (дата обращения: 16.04.2025).

которая в режиме реального времени перенимает выражение лица источника и накладывает его на лицо целевого человека в видеопотоке [8]. Face2Face и подобные ему алгоритмы используют комбинацию методов компьютерного зрения (отслеживание лицевых ключевых точек, 3D-моделирование лица) и рендеринг для правдоподобного изменения мимики. Такие системы относятся к категориям методов *face reenactment* (перенос мимики) и *face replacement* (замена лица), не прибегая к обучению нейронных сетей.

Хотя традиционные графические методы смогли достичь впечатляющих результатов (например, Face2Face способен работать с минимальной задержкой на обычном ПК), у них есть ограничения по реалистичности: требуется точная калибровка, иногда возникают артефакты несоответствия освещения или «швы» на границе вставленного лица. Настоящую революцию в создании фальшивых видеоизображений лиц произвело применение глубинных нейронных сетей, особенно генеративно-состязательных сетей (GAN). GAN были предложены Гудфеллоу с соавторами в 2014 году [9] и представляют собой архитектуру из двух сетей – генератора и дискриминатора – обучаемых в противоборстве друг с другом. Генератор учится создавать фальшивые образцы (например, изображения лиц), которые максимально похожи на реальные, а дискриминатор пытается отличить сгенерированные образцы от настоящих. В процессе совместного обучения GAN-архитектура постепенно улучшает качество синтеза до тех пор, пока дискриминатор уже не способен надежно отличать подделку. Первые приложения GAN к области лиц были нацелены на синтез статичных изображений (например, GAN успешно применялись для задач изменения атрибутов лица или порождения несуществующих лиц, как в проектах ThisPersonDoesNotExist [10]). Со временем появились и архитектуры для видео: например, улучшенные версии GAN (StyleGAN2 [11], StyleGAN3 [12]) позволили генераторам моделировать высокореалистичные лица с сохранением консистентности между кадрами видео.

В контексте дипфейков ключевую роль сыграли автоэнкодерные архитектуры и их сочетание с GAN. Подход, известный ныне как классический DeepFake (в узком смысле ThisPersonDoesNotExist метод замены лица через автоэнкодер), изначально был реализован энтузиастами в 2017 году. Его суть заключается в обучении двух автоэнкодеров с общим энкодером: один декодер учится восстанавливать лицо A, другой ThisPersonDoesNotExist лицо B. После обучения скрытое представление *(latent code)*, полученное для лица A, подается в декодер B, что дает изображение лица B с мимикой и выражением лица A. Такой подход требовал больших наборов данных изображений обоих лиц и длительного обучения для каждой новой пары, но уже демонстрировал убедительные результаты. Появились первые пользовательские приложения, например, FakeApp [13], делавшие технологию широко доступной. Позднее сообщество с открытым исходным кодом разработало инструменты Deepfakes и FaceSwap – названия, которые стали нарицательными для подобных алгоритмов. Эти проекты были выложены на GitHub[3] и позволили пользователям создавать дипфейки из домашних видео, что вызвало волну обсуждений этических и правовых вопросов.

Современные фреймворки для создания дипфейков активно используют идеи GAN для повышения качества. Одним из наиболее популярных и технически совершенных решений является DeepFaceLab [14] – открытая платформа, предоставляющая полный конвейер для face swapping с возможностью тонкой настройки процесса. DeepFaceLab комбинирует несколько методов: детекция и выравнивание лиц на каждом кадре, обучение специальной нейросети для замены лица, пост-обработка для сглаживания швов и цветокоррекции. Авторы DeepFaceLab показывают, что их система способна генерировать поддельные видеоролики кинематографического качества, устойчивые к умеренным изменениям ракурса, освещения и даже к частичным окклюзиям лица (например, волосы, очки).

Другие заметные инструменты включают FaceSwap-GAN (вариант, использующий GAN для face swap) [15] и разнообразные мобильные приложения (в том числе коммерческие, например: FaceApp), которые применяют облегченные генеративные модели для замены лица на смартфоне.

Работа [16] выделяет следующие категории глубоких фейковых видео: подмена лиц, синтез и манипуляция чертами лица. В глубоких фейковых видео с заменой лиц, лицо человека заменяется лицом исходного человека, чтобы создать фальшивое видео, нацеленное на человека за действия, которые он не совершал, что может запятнать репутацию человека. В другом типе глубоких фейковых видео, называемом синхронизацией губ, губы целевого человека манипулируются для изменения движений в соответствии с определенной звуковой дорожкой. Цель синхронизации губ состоит в том, чтобы имитировать голос жертвы, заставляя кого-то говорить этим голосом. С помощью кукловода глубокие фейки производятся путем имитации выражений лица цели, движений глаз и головы. Используя фиктивные профили, это делается для распространения ложной информации в социальных сетях. И последнее, но не менее важное: глубокие аудиоподделки или клонирование голоса используются для манипулирования голосом человека, который ассоциирует что-то с говорящим, чего он не говорил в действительности.

Таким образом, на сегодняшний день существует богатый набор технологий создания дипфейков: от классических графических методов до современных гибридных нейросетевых систем. Они позволяют не только подменять лицо одного человека лицом другого на видео, но и выполнять дубляж речи с изменением движений губ, переносить мимику и даже полностью генерировать фальшивых виртуальных персонажей. Высокое качество синтеза, достигаемое новейшими моделями, делает визуальное обнаружение подделки крайне сложной задачей для человека и предъявляет высокие требования к автоматическим методам обнаружения.

[3] Deepfakes [Электронный ресурс] // GitHub, 2025. URL: https://github.com/topics/deepfake (дата обращения: 16.04.2025).

Современные подходы к обнаружению дипфейков

Развитие методов генерации фейковых видео стимулировало появление широкого спектра методов обнаружения дипфейков. В целом, эти методы можно разделить на две категории: методы, основанные на поиске специфических визуальных и поведенческих артефактов на видео, и методы, использующие модели глубокого обучения для автоматического выявления тонких несоответствий. Многие современные системы детекции сочетают элементы обоих подходов.

Анализ визуальных и поведенческих артефактов

Ранние работы по распознаванию поддельного видео фокусировались на обнаружении артефактов, неумышленно оставляемых алгоритмами генерации. Под артефактами понимают любые отклонения от характерных для настоящего видео признаков, которые могут возникать вследствие процесса синтеза. Одним из примеров является поиск следов постобработки и сшивания лица: многие дипфейк-алгоритмы выполняют вставку лица в целевое видео с последующим сглаживанием границ. Это может приводить к едва заметным искажениям геометрии лица (искаженное соотношение размеров глаз, расстояния между чертами) – так называемым *face warping artifacts*. Автор [17] предложили выявлять такие искажения, обучая классификатор различать реальные лица и интерполированные (искусственно растянутые) области лица в видео. Этот метод показал, что многие DeepFake-видео выдают себя неестественными пропорциями лица, особенно при сильных поворотах головы или изменении масштаба.

Другой характерный признак – несовершенство отрисовки мелких деталей лица. К примеру, линии волос, зубы, переходы около рта и глаз могут содержать размытости или искажения, отсутствующие на других кадрах. Некоторые алгоритмы детекции анализируют изображение в частотной области, выявляя аномальные частотные компоненты, остающиеся после генерации. Так, в работе [18] предложена компактная сверточная сеть *MesoNet*, которая оперирует пониженным разрешением кадра, концентрируясь на общих текстурных особенностях изображения лица. Она смогла уловить различия между реальным видео и дипфейком благодаря тому, что генеративные модели оставляют «отпечатки» – например, характерный шум или сглаживание на определенных частотах. Помимо статических артефактов, исследователи обратили внимание на динамические несоответствия, связанные с физиологией и мимикой. Одним из первых сигналов стал анализ моргания глаз. В нормальном видео человек моргает с определенной частотой и продолжительностью, тогда как ранние дипфейки зачастую грешили отсутствием моргания или неестественно редким морганием. Дело в том, что на многих исходных фотографиях люди запечатлены с открытыми глазами, и генеративная модель, обученная на таких данных, могла не научиться моргать. В 2018 году была предложена простая, но эффективная идея: детектировать моменты закрытия век и проверять, вписываются ли они в физиологическую норму. Алгоритм проверки моргания [19] вычислял метрику частоты/длительности миганий и успешно отличал поддельные видео по аномально низкой частоте моргания. Как только данный изъян стал известен, создатели дипфейков улучшили модели – современные deepfake-генераторы уже научились имитировать моргание [20]. Однако сам подход наглядно показал ценность поведенческих признаков.

Аналогично, методы обнаружения могут анализировать движения головы и выражения лица на предмет правдоподобности. В реальных видео движение головы согласовано с ориентацией и положением тела, а изменения мимики происходят согласованно с речью и эмоцией. При подделке возможны неточности: например, фальшивое лицо может оставаться слишком неподвижным, «приклеенным» к голове, либо наоборот – дрейфовать относительно контура головы при сильных наклонах. Исследование [21] показало, что у поддельных видео, созданных с заменой лица, часто наблюдается несоответствие между положением головы (угол наклона, поворота) и ожидаемым распределением этих параметров для реальных людей. Они использовали оценку позы головы *(Head Pose Estimation)* для каждого кадра и проверяли, лежат ли траектории поворотов головы в статистически допустимом диапазоне. Выяснилось, что во многих дипфейках голова движется неестественно плавно или с ограниченной амплитудой, что выдавало подделку. Еще один аспект — мимические несоответствия и отсутствие микрожестов. Поддельное лицо может некорректно передавать эмоции: например, улыбка может не затрагивать область вокруг глаз, или выражение гнева не сопровождаться характерными морщинами на лбу. Такие тонкие детали трудны для прямого алгоритмического теста, однако есть попытки косвенно их выявлять. Некоторые подходы анализируют последовательности выражений лица, сравнивая их с типичными паттернами изменения эмоций. Другие используют внешние источники: например, если доступна аудиодорожка, можно проверить синхронизацию движений губ с речью (так называемые методы *detiction* по несоответствию аудио-видео). В контексте сугубо визуального потока, исследователи обращаются к анализу лицевых экшенов *(Facial Action Units)* – количественных индикаторов активности тех или иных лицевых мышц. Идея в том, чтобы отследить, корректно ли сочетаются активированные группы мышц при выражении эмоций. Например, искренняя улыбка задействует и мышцы вокруг глаз; если алгоритм обнаружит улыбку без активности век, это может считаться «неполной» эмоцией, потенциально указывающей на подделку.

Отдельно стоит упомянуть методы на основе физиологических сигналов. Одно из новаторских направлений – анализ мельчайших изменений цвета кожи лица, связанных с пульсацией крови. В нормальном видео высокочастотные колебания цветового тона (особенно в области щек) свидетельствуют о биении сердца. В глубоко поддельных видео такие сигналы могут исчезать или

сильно искажаться, так как генеративная модель их не воспроизводит. Метод, названный FakeCatcher [22], извлекает пульс из видео лица и выявляет его отсутствие либо несоответствие человеческой норме как маркер дипфейка. Это пример использования неочевидных артефактов: зрительно человек не заметит разницы, а алгоритм, чувствительный к биологическим ритмам, обнаружит обман. Однако подобные подходы требуют высокого качества видео и могут давать сбои при сильных изменениях освещения или у людей с темным оттенком кожи, где фотоплетизмографический сигнал слабее.

Перечисленные методы анализа артефактов, как правило, понятны для интерпретации и иногда позволяют точно указать, какой именно признак выдал фальшивку (например, отсутствие моргания или неверный поворот головы). Их недостаток в том, что они часто ловят лишь конкретные проявления несовершенства дипфейка. Стоит злоумышленнику учесть эти признаки и улучшить генератор (как случилось с морганием), и детектор перестанет работать. Поэтому на первый план выходят адаптивные методы на базе глубинного обучения, способные самостоятельно обучаться выявлять скрытые артефакты.

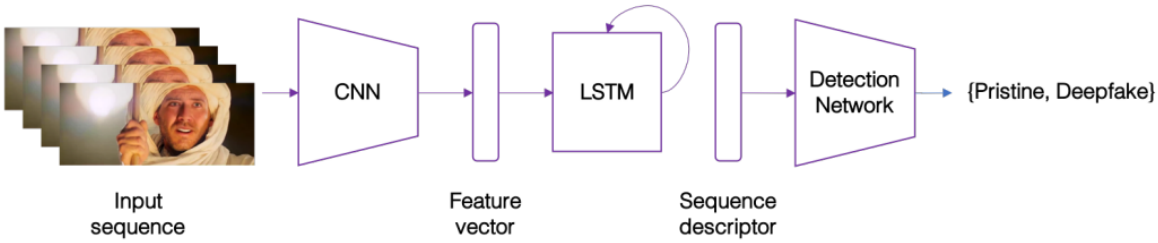
Глубокие нейросетевые модели для детекции дипфейков

Современные подходы к обнаружению дипфейков в значительной мере полагаются на методы глубокого обучения. Аналогично тому, как GAN совершили прорыв в генерации, сверточные нейронные сети (CNN) совершили прорыв в задаче классификации изображений, включая и задачу различения поддельных и настоящих лиц. CNN способны автоматически извлекать признаки *(feature maps)* с различных уровней абстракции: от низкоуровневых (текстуры, шум) до высокоуровневых (форма лица, характерные черты). В контексте детекции дипфейков это значит, что одна хорошо обученная сеть может одновременно подметить и мелкие пиксельные артефакты сглаживания, и более глобальные несоответствия в геометрии лица.

Один из первых успешных глубинных детекторов — уже упомянутая модель MesoNet [18]. Она имеет небольшую архитектуру с парой сверточных слоев, специально упрощенных для ускорения, и выдает вероятность «фейк/реал» по одному кадру видео. Несмотря на компактность, MesoNet показала точность порядка 95% на тестах того времени, хотя и уступала более тяжелым моделям. Позднее, с появлением больших датасетов, в качестве основы для детекции стали использоваться более сложные сети, например, XceptionNet [23] – она изначально разрабатывалась для общей классификации изображений, но была адаптирована под задачу распознавания дипфейков [24]. XceptionNet, обученная на большом корпусе данных FaceForensics++, достигала выдающихся результатов: свыше 98-99% точности на отдельных наборах данных в лабораторных условиях [25]. Именно она легла в основу многих последующих соревнований и коммерческих решений. Так, победители Facebook DeepFake Detection Challenge (DFDC) в 2020 году использовали ансамбли из XceptionNet и ее вариаций, добиваясь высокой точности на разнородных видео.

Однако простое применение CNN к каждому кадру в отдельности не учитывает временную последовательность, что важно для отсеивания ложных срабатываний. Поэтому следующим этапом развития стали архитектуры, учитывающие динамику видео. Для этого часто используют рекуррентные нейронные сети (RNN), в частности их разновидность – долгократкосрочную память (LSTM). Подход «CNN+RNN» заключается в том, что сначала каждое изображение (кадр) преобразуется CNN в вектор признаков, а затем последовательность таких признаков подается в рекуррентную сеть, обученную выявлять временные корреляции. Например, авторы исследования [26] применили LSTM поверх признаков, извлекаемых сверточной сетью, и сумели уловить характерные для дипфейков временные аномалии. RNN может определить, что черты лица меняются от кадра к кадру слишком сильно либо, наоборот, подозрительно стабильны, или заметить рассинхронизацию движения губ с предыдущими кадрами (рис. 1). В результате, добавление рекуррентного звена повышает надежность детекции на видео и уменьшает количество ошибок, связанных с анализом единичных кадров. Недостаток RNN-компоненты – это увеличение требуемых ресурсов и задержки: для уверенного решения сети нужно накопить определенное число кадров, что может быть проблематично в условиях реального времени (например, задержка 0.5-1 секунд, прежде чем принять решение).

Еще одно направление улучшения детекторов – использование механизмов внимания *(attention)* и трансформеров. Модели внимания позволяют сети фокусироваться на наиболее информативных областях кадра или наиболее подозрительных фрагментах видео. В задачах распознавания дипфейков это может значить, например, что модель научится обращать внимание на область глаз, зубов или границ лица, если именно там часто проявляются артефакты. Классические CNN равномерно обрабатывают все изображение, тогда как attention-механизм выделяет важные регионы, что полезно, когда артефакты редки и локальны. Современные трансформеры *(Transformer)* с их механизмом самовнимания завоевали популярность в обработке последовательностей и изображений. В детекции дипфейков появляются работы, где трансформер применяется либо для моделирования временной последовательности кадров, либо даже для непосредственной обработки изображения вместо CNN. Например, в исследовании [27] представили модель FakeFormer, использующую Vision Transformer в сочетании со специальным модулем локального внимания, который выделяет наиболее уязвимые к подделке области изображения.



Р и с. 1. Архитектура системы обнаружения [26]. Система обучается и делает выводы сквозным образом и, учитывая видеопоследовательность, выводит вероятность того, что это deepfake или чистое видео. Она имеет сверточную подсеть LSTM для обработки входной временной последовательности.

F i g. 1. Detection system architecture [26]. The system trains and infers end-to-end, and given a video sequence, it outputs the probability of it being a deepfake or a clean video. It has a convolutional LSTM subnetwork for processing the input temporal sequence.

Этот подход позволил улучшить обобщающую способность детектора на новые виды дипфейков, так как внимание гибко подстраивается под различия между реальными и синтезированными лицами. Также трансформеры способны эффективнее учитывать долгосрочные зависимости в видео, что помогает ловить тонкие нарушения полноценности (например, несоответствие освещения лица в разных сценах).

Помимо описанных основных направлений, есть и другие подходы: использование капсульных сетей для хранения структурной информации о лице, методы на основе мета-обучения для адаптации к новым типам фейков, обучение на несбалансированных данных с помощью специальных функций потерь и т.д. Каждая из этих идей направлена на то, чтобы сделать детектор более устойчивым к постоянно усложняющимся атакам синтеза.

На рисунке 2 представлена таксономия методов определения дипфейков из свежего обзора [32].

Важно отметить, что разработчики детекторов сталкиваются с противостоящей стороной: генераторы дипфейков также улучшаются, часто учитывая известные им методы обнаружения. Это что-то вроде «гонки вооружений» в области медиадезинформации. Например, как только были опубликованы работы про моргание или позу головы, генераторы включили эти факторы в модели. Уже существуют продвинутые GAN, которые оптимизируются не только на правдоподобие картинки, но и на определенные метрики, связанные с детектированием, то есть обучаются обманывать конкретные классификаторы. Поэтому лучшие результаты приносит комбинация подходов: использование мощных нейросетевых моделей, обученных на большом количестве разнообразных данных, плюс учет специфических артефактов и привлечение внешних знаний (например, физика процесса съемки, физиология человека). В следующих разделах мы рассмотрим эффективность существующих методов более количественно, а также обсудим их применимость в условиях реального времени.

Обзор наборов данных для детекции дипфейков

Качество и масштаб доступных датасетов напрямую повлияли на прогресс в обнаружении дипфейков. Для обучения и сравнения методов нужны большие коллекции видеозаписей с известном фактом того поддельное видео или реальное. Рассмотрим наиболее значимые наборы данных, их особенности и влияние на точность систем.

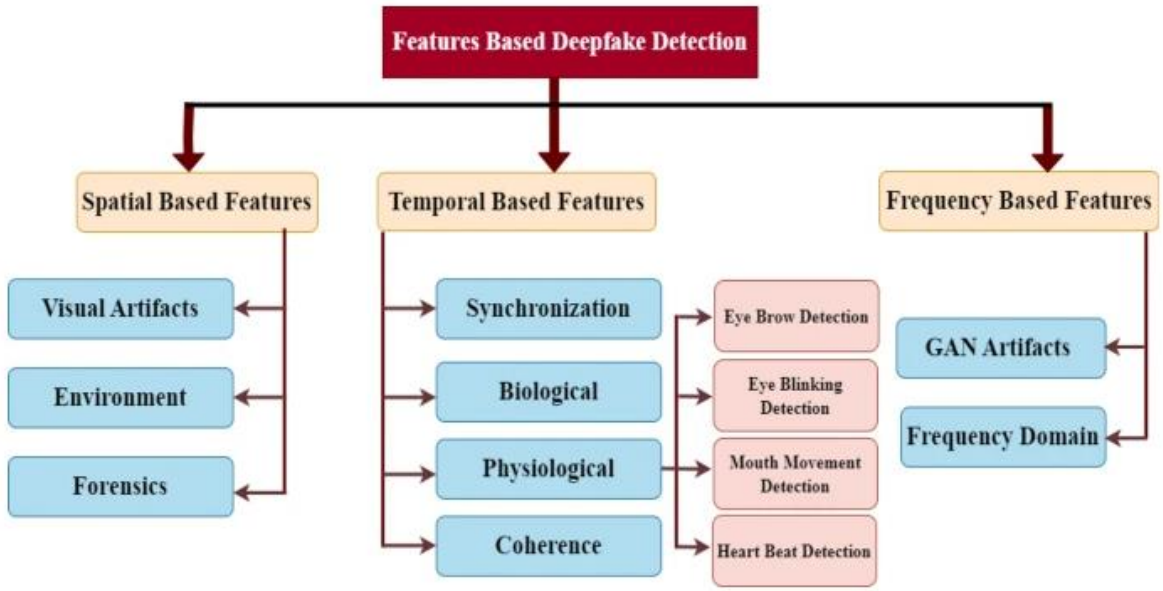
FaceForensics++ (FF++) [25] – один из первых и наиболее известных открытых наборов данных. Он включает 1000 оригинальных видеороликов (вырезки с YouTube, в основном новости, интервью) и по 1000 видеоподделок, созданных четырьмя разными методами: DeepFakes, FaceSwap, Face2Face и NeuralTextures. Эти методы охватывают как классический дипфейк на автоэнкодерах (DeepFakes), так и графические подходы (Face2Face, FaceSwap) и гибридный нейросетевой (NeuralTextures). Для каждого видео также предоставляются версии с разным уровнем сжатия (слабое и сильное), что имитирует загрузку на интернет-платформы. Всего датасет содержит более 5 тысяч видео (включая сжатые варианты) или около 1.8 миллиона кадров. Набор разделен на обучающую и тестовую части и стал своеобразным стандартом: большинство ранних работ (2018-2020) отчитывались по точности на FaceForensics++. Его сильная сторона – наличие пар реальное-поддельное видео для конкретных сцен, что упрощает обучение поиску отличий.

Слабая сторона – ограниченность сцен (только говорящие головы, мало резких движений) и ограниченный перечень методов генерации. Тем не менее, FF++ позволил сформировать первичные базовые линии *(baseline)* по метрикам и показал, что задача решаема автоматикой лучше, чем человеком.

Авторы также внедрили удобный эталонный стандарт оценивания: они предлагают тренировать модель на их тренировочном множестве, а тестировать на особой «слепой» выборке из 100 видеороликов с различными искажениями. Это обеспечило стандартизированное сравнение для разных алгоритмов детекции.

DFDC (DeepFake Detection Challenge dataset) [28] – крупнейший на сегодня публичный сборник дипфейков, выпущенный компанией Facebook AI в сотрудничестве с исследовательским сообществом. Окончательный вариант датасета (2020) содержит более 100 тысяч фейковых видео, созданных на основе ~20-30 тысяч исходных клипов с участием множества приглашенных актеров.

Для генерации подделок организаторы привлекли разнообразные алгоритмы *face swapping*, часть из которых так и не раскрыта публично (чтобы сделать задачу обнаружения ближе к реальному миру). Видео охватывают различные условия: съемка при разном



Р и с. 2. Таксономия методов детекции [32]
F i g. 2. Taxonomy of detection methods [32]

освещении, на разных фоновых сценах, люди разных возрастов, национальностей, с различной внешностью. Также дипфейки варьируются по качеству: от откровенно плохих (видны артефакты) до весьма правдоподобных.

DFDC изначально был выпущен в виде соревнования на Kaggle, что стимулировало разработку новых подходов. Участники отмечали, что датасет весьма сложный: простые модели на нем показывали F1-score около 0.3-0.5. Лучшие результаты (0.7 F1-score) были получены ансамблями из 8-10 нейросетей, что говорит о том, что единая модель не справляется со всем разнообразием DFDC. В практическом плане DFDC ценен тем, что имитирует реальную задачу: много разных людей и возможных подделок, нет жесткой привязки по парам «оригинал-фейк».

Для задачи распознавания дипфейков в видеоконференциях в реальном времени он особенно релевантен, так как включает разнообразие сценариев съемки, схожее с тем, что может быть в звонках (разные комнаты, ракурсы, люди двигаются и говорят и т.д.). Одним из недостатков DFDC является объем полного набора (сотни гигабайт), а также некоторый дисбаланс в данных (фейков больше, чем реальных видео). Тем не менее, любой серьезный детектор сейчас проверяется на DFDC или хотя бы его части, чтобы оценить боеспособность.

Celeb-DF [29] – сравнительно небольшой, но показательный датасет, созданный группой из Университета Олбани. Он содержит 590 реальных видео с интервью знаменитостей (в основном фрагменты ток-шоу, новостных интервью) и 5639 фейковых видео, полученных автоматической заменой лица. Важная особенность: авторы Celeb-DF разработали улучшенную версию алгоритма DeepFake, в которой значительно уменьшены характерные артефакты предыдущих поколений. В ранних дипфейках, например из DeepFake-TIMIT набора, часто наблюдались: размытые края лица, заметная разница в цвете кожи между лицом и шеей, некорректная форма зубов при открытом рте. В Celeb-DF это улучшили: фальшивки выглядят очень реалистично, цветовая гамма выровнена, резкость высокая, аномалии сведены к минимуму. По сути, Celeb-DF на момент создания был самым «сложным» для детекторов набором. Многие существующие методы, превосходно работавшие на FaceForensics++, на Celeb-DF показали гораздо более низкие результаты. В статье [29] приводится детальный анализ: причинами ошибок стали как раз улучшенное качество дипфейков, а также то, что знаменитости из набора отсутствовали в данных для обучения распространенных моделей. Celeb-DF актуален тем, что демонстрирует вызовы, которые ждут детектирование дипфейков в ближайшем будущем: по мере совершенствования генераторов различить фейк становится все сложнее, и нужны более тонкие признаки или более глубокое понимание видео. Набор Celeb-DF (v2) общедоступен и используется как дополнительный тест на генерализацию: часто обучают модель на FF++ и проверяют на Celeb-DF.

Помимо перечисленных, существуют и другие датасеты: UADFV (2018, небольшой набор из 49 фейков), DeepFake-TIMIT (2018, фейки на основе видеороликов TIMIT, низкое разрешение), Google DFD (2019, выпущен вместе с DFDC-preview, включает 363 фейковых видео, созданных с привлечением актеров от Google).

В 2020-2021 появились более специфические сборники: например, наборы для детекции определенных подвидов манипуляций (только reenactment или только lip-sync), датасеты с реальными дипфейк-видео из интернет и т.д. Также исследовательские организации (NIST, MIT) формируют оценочные коллекции для регулярного тестирования алгоритмов. Но в академической литературе наибольшее распространение получили именно FF++, DFDC и Celeb-DF, как охватывающие основные аспекты. Качество данных в датасете

существенно влияет на итоговую точность алгоритмов. Если данные содержат смещение *(bias)* или артефакты несвязанные с самой подделкой, модель может научиться распознавать не дипфейк, а эти побочные признаки. Например, обнаружилось, что некоторые ранние дипфейк-видео имели фиксированное разрешение или определенный паттерн сжатия – детекторы могли опознавать их не потому, что детектировали подделку, а потому, что заметили знакомое разрешение кадра. Разработчики новых датасетов стараются исключать такие закономерности. Так, в DFDC подборка реальных и фейковых видео перемешана и приведена к схожим форматам, чтобы модель фокусировалась именно на лице. FaceForensics++ вводит случайное изменение степени сжатия в тестовой выборке, чтобы проверить, не переобучилась ли модель на конкретный битрейт. В Celeb-DF особое внимание уделялось тому, чтобы фейки и исходники имели одинаковое качество и длину, и детектор не мог тривиально отличить их. Таким образом, прогресс в обнаружении дипфейков во многом обязан наличию общедоступных датасетов. Они позволяют объективно сравнить методы, выявить их сильные и слабые стороны. Для детекции дипфейков в реальном времени важнее всего рассматривать те наборы, которые отражают динамику видео и реальные условия съемки, такие как DFDC и Celeb-DF.

Свежая работа 2025 года [30] представила ExDDV – первый набор данных и эталон для объяснимого обнаружения дипфейков в видео. ExDDV включает около 5,4 тыс. реальных и дипфейковых видео, которые вручную аннотируются текстовыми описаниями (для объяснения артефактов) и кликами (для указания артефактов). Такие аннотации нужны для отладки процесса использования больших языковых (мультимодальных) моделей для определения дипфейков.

Хороший обзор датасетов для видео и аудио дипфейков есть в работе [31].

Сравнительный анализ методов и их эффективность

При оценке методов обнаружения дипфейков важно учитывать несколько ключевых критериев: точность классификации, скорость работы и устойчивость к разнообразию поддельного контента. Рассмотрим каждый из них и сравним существующие подходы по этим параметрам.

Точность и надежность распознавания

Точность методики обычно измеряется метриками бинарной классификации: accuracy (доля правильно классифицированных примеров), *precision/recall* (точность и полнота по классу «фейк»), метрики F1-score и площадей под ROC-кривой (AUC). В контролируемых условиях (когда тренировочные и тестовые данные взяты из одного распределения) современные глубокие модели демонстрируют весьма высокую точность. Так, лучшие сверточные сети (XceptionNet и др.) достигали accuracy порядка 99% на наборе FaceForensics++ при тестировании на известных типах фейков. Это означает, что в лабораторных экспериментах детекция уже обгоняет человека: для сравнения, даже эксперты визуально определяют подделки лишь в ~50-70% случаев, тогда как автоматический классификатор практически не ошибается на отфильтрованных видеороликах хорошего качества. Добавление рекуррентных слоев и других улучшений способно еще больше увеличить эти показатели. Однако, столь радужные цифры сильно снижаются, когда дело доходит до работы «в диких условиях» или на новых данных.

Практика показала, что модель, обученная на одном наборе (например, только на дипфейках типа FaceSwap), может плохо справляться с другими видами манипуляций. Проблема обобщающей способности *(generalization)* состоит в том, что разные генераторы оставляют разные артефакты. Например, детектор, заточенный под поиск размытого контура лица, может пропустить фейк, где контур идеально сглажен, но присутствует другой изъян (например, неправдоподобная мимика). Исследование [29] продемонстрировало, что многие алгоритмы, хорошо показавшие себя на FaceForensics++, резко теряют эффективность на более реалистичном наборе Celeb-DF: accuracy падала с ~95% до 65–70%. В конкурсе DFDC 2020 лучшие решения едва превысили 65-70% по метрике валидации на совершенно новых видео, хотя на своих тренировочных данных имели >90%. Эти факты указывают на то, что устойчивость к разнообразию фейков остается серьезной проблемой. Для повышения устойчивости исследователи применяют увеличение данных *(data augmentation)* – добавляют различные виды сжатия, шум, искажения в обучающие данные, чтобы модель стала менее чувствительна к несущественным вариациям. Также эффективны методы объединения нескольких моделей (ансамбли), где разные классификаторы дополняют друг друга, а их результаты объединяются (например, посредством усреднения вероятностей или более сложного стека). В соревновании DFDC многие команды-финалисты использовали ансамбли из 5-10 разных моделей, что позволило учесть максимум признаков. Отдельно необходимо учесть показатель ложных срабатываний *(False Positive Rate)*. В прикладной ситуации (например, мониторинг видеоконференций) крайне важно, чтобы система почти никогда не маркировала реального человека как дипфейк, иначе доверие к ней будет подорвано. Многие исследовательские работы оптимизируют модели по метрике F1-score, которая учитывает и пропущенные фейки, и ложные тревоги. Добиться нулевых ложных срабатываний трудно, но современные модели позволяют регулировать порог срабатывания: снижая чувствительность, можно уменьшить FP до приемлемого уровня ценой некоторого увеличения FN (пропусков фейков). В реальном применении может оказаться предпочтительным более консервативный режим, когда детектор подает сигнал только при очень высокой уверенности, чтобы исключить ошибочные обвинения.

Скорость работы и возможность применения в реальном времени

Скоростные характеристики алгоритма определяют, может ли он использоваться для фильтрации видеопотока на лету. Требуемая скорость обычно измеряется в кадрах в секунду *(frames per second, FPS)*, которые способен обрабатывать детектор. Для реального времени желательно стремиться к 25-30 FPS (стандартная частота видео). Если детектор работает медленнее, возможны буферизация и накопление задержки: например, при 10 FPS на входе 30 FPS видео каждая секунда видео будет обрабатываться ~3 секунды. В контексте видеоконференций слишком большие задержки неприемлемы, особенно если планируется автоматическая реакция (например, отключение видео нарушителя). Классические тяжелые CNN, применяемые в детекции (Xception, EfficientNet-B7 и т.п.), в необлегченном виде могут работать лишь несколько кадров в секунду на CPU. Для ускорения требуется использование GPU, на котором параллельные свертки выполняются значительно быстрее. Даже на GPU высокая точность часто достигается ценой сложности модели: например, сеть Xception имеет порядка 20 миллионов параметров и на графическом ускорителе среднего класса обрабатывает ~15-20 FPS при разрешении 256 × 256. Если лиц в кадре несколько, то каждый из них нужно анализировать отдельно, что еще уменьшает суммарную скорость. Таким образом, прямолинейное применение лучших моделей затруднительно для реального времени без серьезных вычислительных ресурсов.

Решение видят в создании облегченных моделей. Один подход – использовать архитектуры, оптимизированные по скорости: например, вместо Xception взять MobileNet или EfficientNetB0/B1, пожертвовав некоторой точностью ради быстродействия. Другой подход – аппаратное ускорение: на практике встраиваемые модули, такие как Intel OpenVINO, NVIDIA TensorRT, позволяют получить значительный прирост FPS за счет оптимизации графа сети и использования низкоуровневых инструкций. Также существует специальное железо (VPU, TPU), которое может обеспечивать высокую скорость обработки нейросетей на периферийных устройствах.

Некоторые исследования предлагают структурные оптимизации. Например, применение каскадов детекции: сперва очень быстрый и грубый классификатор отсекает явно реальные видео, а более точный и медленный анализатор проверяет оставшиеся подозрительные случаи. В видео call-сценарии можно представить себе систему, которая непрерывно контролирует качество (скажем, быстрым детектором артефактов), и только при превышении определенного порога передает кадры на детальный анализ CNN. Такая схема экономит ресурсы при условии, что большинство участников – реальные, и дипфейк появляется редко. Практические измерения показывают, что добиться стабильных 30 FPS на одном потоке видео возможно при использовании среднего GPU и модели типа Xception/MobileNet. Если же нужно параллельно мониторить несколько потоков (в конференции их может быть 5, 10, 20 и более), вычислительная нагрузка масштабируется соответственно. Поэтому для многопользовательских звонков вероятно потребуется либо очень мощный сервер, либо распределение анализа между клиентскими машинами.

Устойчивость к различным видам атак и условий съемки

Как отмечалось, разнообразие дипфейков – серьезный вызов для надежности методов. Под разнообразием понимаются разные алгоритмы генерации *(face swap, face reenactment, lip sync и др.),* различные степени пост-обработки, качество и формат видео (разрешение, степень компрессии), а также уникальные особенности конкретных лиц. Хороший детектор должен уметь выявлять подделку, даже если она создана неизвестным ему ранее способом. В литературе приведены попытки создать универсальные детекторы, которые обучены сразу на множестве типов манипуляций. Например, FaceForensics++ изначально включал 4 различных метода генерации дипфейков, и авторы обучили сеть сразу на всех, добившись хороших результатов. Такая сеть приобрела более разносторонние признаки и лучше справлялась с новым контентом, чем узкоспециализированная. Однако и универсальная сеть может столкнуться с проблемами, если в реальном видео присутствуют условия, не представленные в тренировочных данных. Одно из уязвимых мест – сжатие и артефакты кодека. В видеоконференциях часто применяется агрессивное сжатие видео при низком битрейте, из-за чего картинка содержит блоковые артефакты, размытие движения и т.д. Эти искажения могут маскировать или имитировать признаки дипфейка, и из-за этого точность многих детекторов существенно падает на видео низкого качества. Поэтому системы приходится специально обучать на сжатых данных (например, FaceForensics++ предоставляет версии видео с разной степенью сжатия). Даже так, в реальном видеопотоке могут быть потери пакетов, резкие перепады качества при адаптивном битрейте, что может приводить к ошибкам детекции. Решением служит повышение робастности (устойчивости): внедрение в обучение случайных деградаций видео (масштабирование, добавление шумов, имитация артефактов сжатия). Также можно применять предварительную фильтрацию: некоторые алгоритмы пытаются «очистить» входной кадр от сжатия (например, с помощью супер-резолюции или фильтра Гаусса) перед подачей в детектор, чтобы снизить влияние нерелевантных артефактов. Еще один аспект устойчивости – защита от целенаправленного обхода детектора. Поскольку методы обнаружения публикуются, злоумышленник теоретически может попытаться создать дипфейк, который обманет известный классификатор (например, добавить контурный шум, который сбивает CNN). Это приводит нас к проблеме противодействия состязательным атакам. Пока что в контексте дипфейков она изучена слабо, но уже показано, что можно слегка модифицировать каждый кадр (например, через незначительное изменение цветов), и уверенность детектора упадет. Для борьбы с этим применяются стандартные техники из области безопасности

нейросетей: ансамбли моделей, проверки на консистентность результатов (например, если видео реально, то разные кадры должны давать одинаковый вердикт; если один кадр резко отличается, возможно, на него повлиял умышленный шум).

В заключение стоит отметить, что применимость методов в реальном времени определяется балансом всех перечисленных факторов. На сегодняшний день возможно собрать систему, которая в режиме live анализирует видеопоток и с довольно высокой точностью оповещает о подозрительных участниках. Например, комбинация: быстрый детектор (BlazeFace) для отсечения фона и локализации лица, плюс оптимизированная CNN (MobileNetV3 или EfficientNet-B0) для классификации, – может работать на современном ноутбуке с GPU почти в реальном времени с точностью около 85-90% на известных типах фейков. Однако по мере появления новых, более сложных дипфейков, такую систему придется регулярно дообучать и обновлять. Кроме того, необходимо учитывать, что любое решение – компромисс: либо максимальная точность ценой ресурсов и задержек, либо быстродействие ценой пропуска некоторых продвинутых подделок. Выбор зависит от требований конкретного применения.

Новые подходы к детекции дипфейков

Современные исследования предпринимают усилия по преодолению упомянутых ограничений детекторов. Так, свежий обзор 2024 г. отмечает, что имеющиеся алгоритмы не подходят для работы в реальном времени, требуя оптимизации скорости, улучшения обобщающей способности и надежности [33]. Ниже рассмотрены несколько перспективных методов, появившихся в 2024-2025 гг., ориентированных на видеопотоки.

Улучшение обобщаемости моделей. Одно из направлений связано с использованием трансформеров и механизмов внимания, позволяющих модели фокусироваться на важных областях кадра. Например, предложена архитектура FakeFormer на основе Vision Transformer (ViT), дополняемого модулем локального внимания, который выделяет наиболее уязвимые к подделке зоны лица. Вместо обработки изображений через свёртки, как в CNN, ViT разбивает изображение на патчи (например, 16×16 пикселей), преобразует их в векторы и обрабатывает с помощью механизма самовнимания *(self-attention).* Это позволило существенно повысить способность детектора распознавать новые виды дипфейков, поскольку самовнимание гибко подстраивается под различия между реальными и синтетическими лицами [27]. В результате FakeFormer продемонстрировал более высокую точность на несвиданных данных по сравнению с классическими CNN. Другие работы исследуют использование методов стегоанализа, то есть использующих выявление скрытого (встроенного) содержания в цифровых медиа: в работе [34] показано, что обнаружение скрытых «шумовых» артефактов (по аналогии с поиском стеганографических встраиваний) и агрессивный прунинг нейросети позволяют сохранить конкурентную точность на сложных наборах (например, Celeb-DF v2, DFDC) при затратах вычислительных ресурсов на порядок ниже, чем у обычных детекторов. Подобные подходы повышают практичность систем, делая их более устойчивыми к разнообразным фейкам без взрывного роста модели.

Оптимизация скорости и модели для real-time. Как отмечалось, прямолинейное использование больших CNN затруднительно в режиме онлайн. В 2024 г. предложены новые архитектуры, специально облегчённые для высокоскоростной работы. В частности, в работе [35] представлен детектор, разработанный на базе бинарной нейросети (BNN). В этой модели все веса квантованы до 1 бита, что радикально снижает требуемые вычисления, потому что при квантовании до 1 бита умножения в нейросети заменяются на простейшие побитовые операции (например, XOR), которые выполняются на порядок быстрее. Дополнительно входные данные дополняются признаками частотного спектра (FFT) и локальных бинарных шаблонов (LBP) для лучшего выявления артефактных текстур. Итоговая модель BNext-BNN достигает сравнимой с лучшими CNN точности на ряде бенчмарков, при этом сокращая объем вычислений. Отмечается, что такой детектор способен работать в реальном времени на низкопроизводительных устройствах (например, на CPU ноутбука). Подобный метод минимизирует задержку обработки и открывает возможность встроить распознавание дипфейков непосредственно в клиентские приложения видеосвязи.

Другой подход – комбинация квантизации и структурного упрощения сети. К примеру, метод в работе [34] применяет укороченную сверточную сеть с априорно уменьшенным числом фильтров и глубины, что в сочетании с упомянутыми техниками стегоанализа ускоряет обработку видео без значимой потери точности. Эксперименты показывают, что подобные облегченные модели способны обрабатывать видеопоток близко к 30 FPS на средней GPU или даже на CPU, в то время как исходные тяжелые сети (типа Xception) давали лишь 15-20 FPS. Таким образом, за 2024-2025 гг. наметилась тенденция к созданию специализированных real-time детекторов, которые жертвуют избыточной сложностью модели ради достижения требуемой скорости.

В таблице 1 обобщены ключевые классы методов обнаружения дипфейков с точки зрения их производительности, а в таблице 2 – пригодности для онлайн-детекции, а в таблице 3 – используемые данные.

Противодействие атакующим дипфейкам. Отдельный вызов – защита от продвинутых злоумышленников, которые могут целенаправленно пытаться обходить детекторы. К 2024 г. начали появляться нестандартные решения для выявления именно реал-тайм дипфейков, используемых в мошенничестве. Один из подходов – внедрение механизма «вызов-ответ» *(challenge-response)* для проверки подлинности участника видеозвонка. В рамках этого метода система инициирует выполнение случайных действий: например, попросить собеседника быстро повернуть голову, сделать определенный жест рукой или произнести фразу шепотом. Такие простые для человека задания очень сложны для генеративной

Т а б л и ц а 1. **Методы детекции**
T a b l e 1. **Detection methods**

Метод	**Архитектура**	**Данные**	**Препроцессинг**	**Точность / Скорость**
Аномалии моргания/позы (2018) [21]	Computer Vision	Видео	Средний: отслеживание глаз и головы	Высокая / >30 FPS (CPU)
XceptionNet (2019) [25]	Глубокая CNN	Кадры видео	Средний: выделение и выравнивание лица	~99% / ~15 FPS (GPU)
MobileNetV3 (2020) [41]	CNN	Кадры видео	Средний: выделение лица	~95% / >30 FPS (GPU)
EfficientNet (2020) [41]	CNN	Кадры видео	Средний: ресайз + нормализация изображения	~97% / 20–25 FPS (GPU)
MesoNet (2018) [18]	Легкая CNN	Кадры лица	Минимальный: crop + resize	~90% / >25 FPS (GPU)
FakeCatcher (2022) [22]	CNN + PPG-анализ	Видео лица	Высокий: сегментация + извлечение PPG-сигналов	~96% / ~60 FPS (GPU)
FakeFormer (2024) [27]	Vision Transformer	Видео	Средний: выделение и нормализация лица	~95% / ~10 FPS (GPU)
BNN детектор (2024) [35]	Бинарная CNN	Кадры видео	Высокий: FFT + LBP текстуры	~90% / >30 FPS (CPU)
Challenge-response (2024) [36]	Интерактивный метод	Живое видео + аудио	Высокий: реакция на случайные действия	~88% / Проверка за 5-10с

Источник: здесь и далее в статье все таблицы составлены авторами.
Source: Hereinafter in this article all tables were made by the authors.

Т а б л и ц а 2. **Применимость методов**
T a b l e 2. **Applicability of methods**

Метод	**Применимость в видеоконференциях**
Аномалии моргания/позы (2018) [21]	Да: простая реализация, уязвимость к улучшенным фейкам
XceptionNet (2019) [25]	Сомнительно: медленная и неустойчива к незнакомым подделкам
MobileNetV3 (2020) [41]	Да: быстрая, но уязвима к улучшенным фейкам
EfficientNet (2020) [41]	Да: хороший баланс точности и скорости, подходит для real-time при оптимизации, но требует GPU
MesoNet (2018) [18]	Да: низкие требования к ресурсам, но есть уязвимость к улучшенным фейкам
FakeCatcher (2022) [22]	Да: real-time система на основе физиологии, но требует качественного видео
FakeFormer (2024) [27]	Сомнительно: сильная, но очень медленная CNN
BNN Detector (2024) [35]	Да: легкая модель, подходит для встроенного использования
Challenge-Response (2024) [36]	Ограниченно: требует участия человека, полезно для критических проверок, сложно автоматизировать

Т а б л и ц а 3. **Используемые данные**
T a b l e 3. **Data used**

Данные	**Описание**	**Пример использования**	**Преимущества**	**Недостатки**
Видео	Полный видеопоток (кадры последовательно)	FakeFormer [27], Challenge-response [36]	Возможность анализировать динамику, мимику, движения головы	Требует больше ресурсов, препроцессинга
Кадры видео	Отдельные статичные кадры из видеопотока	XceptionNet [25], MobileNetV3 [41], BNN [35]	Быстрее обрабатываются, подходят для CNN	Теряется контекст движения, моргание и пр.
Кадры лица	Вырезанные изображения только лица	MesoNet [18]	Уменьшает размер входных данных	Требует надежной детекции лица
Видео лица	Фокус только на видеопотоке лица, часто с высокой частотой кадров	FakeCatcher [22]	Можно анализировать биосигналы (PPG)	Чувствительность к качеству изображения
Живое видео + аудио	Поток данных в реальном времени, включающий как видео, так и речь	Challenge-response [36]	Позволяет делать интерактивные проверки	Требует участия человека; сложно автоматизировать

модели, пытающейся в режиме реального времени синтезировать видео- или аудио-образ. В исследовании [36] представлена система GOTCHA, которая сформировала набор из 8 визуальных тестов и 22 аудио-тестов для собеседника. Выполнение этих тестов в прямом эфире приводило к заметной деградации качества у популярных на 2023 г. RTDF-генераторов *(Real-Time DeepFake generators)*, что позволяло надежно их распознать. Автоматический алгоритм, анализирующий отклики, достиг AUC 80% при выявлении дипфейка, а эксперты-визуализаторы – AUC 88%. Таким образом, в 88% случаев удавалось правильно отличить поддельного собеседника от реального на основе результатов вызовов. Преимущество подхода *"challenge-response"* – объяснимость: система не просто выдает вердикт, а указывает, на каком именно испытании «споткнулась» подозрительная видеомодель. Кроме того, метод мало зависит от типа генератора, полагаясь на фундаментальные физические ограничения дипфейков. Недостатком является необходимость вмешательства в естественный ход разговора: в реальной видеоконференции не всегда уместно устраивать проверки, и злоумышленник может отказаться их выполнять. Тем не менее, данный подход рассматривается как перспективное дополнение к пассивным методам, повышающее общую надежность системы распознавания дипфейков

Практические решения для обнаружения дипфейков в реальном времени

Помимо научных прототипов, начали появляться и коммерческие продукты, способные автоматически выявлять дипфейки во время видеозвонков. Ниже приведен обзор нескольких решений, интегрируемых с популярными платформами видеоконференций.

FakeCatcher (2022). Компания Intel представила первую в отрасли систему детекции дипфейков в реальном времени – платформу FakeCatcher[4]. Ее назначение – выявлять поддельные видеопотоки по физиологическим признакам живого человека. Технология анализирует малейшие изменения цвета кожи лица, связанные с кровотоком (PPG-сигнал), невидимые невооруженным глазом. По 32 ключевым точкам на лице строятся временные карты PPG, которые подаются в легкую CNN для классификации фейк/реал. Используя оптимизированные библиотеки Intel DL Boost и инструкции AVX-512, FakeCatcher работает практически без задержки, возвращая результат за несколько миллисекунд. Заявленная точность достигает 96%. Система способна параллельно мониторить до 72 видеопотоков на одном сервере с Intel Xeon, что важно для масштабирования в корпоративной среде. FakeCatcher предоставляется как веб-платформа: видеопоток перенаправляется на сервер Intel, где производится анализ в режиме *live*, и при обнаружении фейка выдаётся предупреждение. Плюсы: высокое качество распознавания (особенно на известных типах дипфейков), отсутствие необходимости обучать на каждом новом фейке (опирается на физиологию), поддержка многопоточной проверки. Минусы: требовательность к входному видео – при низком разрешении или сильной компрессии сигнал PPG может теряться, что снижает точность. Кроме того, атакующий теоретически может попытаться имитировать пульсации (однако на 2025 г. подобных методов у злоумышленников не отмечено). В целом, Intel FakeCatcher демонстрирует возможность эффективного детектирования на лету без глубокого анализа пиксельных артефактов, дополняя классические нейросетевые детекторы.

Reality Defender (2023-2024). Стартап Reality Defender[5] разработал комплексное решение для защиты видеоконференций от дипфейков. Его продукт представляет собой плагин, который интегрируется с платформами видеосвязи (первой поддержана Zoom) в виде бота-аудитора. После установки расширения в Zoom, при старте звонка специальный бот подключается к конференции и невидимо для участников анализирует поступающие потоки видео (и при необходимости аудио). В реальном времени выполняется многомодельная проверка: задействуется несколько нейросетевых детекторов, обученных выявлять различные признаки подделки (визуальные артефакты, несоответствие мимики речи, искажения звука и т. д.). Результаты отображаются организатору встречи в виде панели с индикаторами уверенности, где каждый участник помечен как реальный или вероятно фейковый с определённым процентом уверенности. Если какой-либо из присутствующих с большой долей вероятности является дипфейком, система немедленно выдаёт предупредительный отчет *(alert)* организатору. Технологии: Reality Defender использует собственные «мультимодельные» детекторы, то есть сочетание нескольких алгоритмов машинного обучения (в компании заявляют об *"award-winning"* моделях). Детали архитектур не раскрываются, но очевидно, что применяются как сверточные сети для анализа лиц на видео, так и методы распознавания голоса для анализа аудио. Интеграция: на конец 2024 г. плагин доступен для Zoom и проходит этап бета-тестирования. Заявлена поддержка Microsoft Teams, WebEx и других платформ в ближайших обновлениях. Сервис ориентирован на корпоративных клиентов: он помогает предотвратить мошенничество при дистанционных сделках, проверке личности и важных переговорах.

Плюсы: полностью автоматический мониторинг во время звонка, работа в реальном времени в фоновом режиме, плагин можно подключить к разным сервисам видеосвязи. Благодаря сочетанию нескольких моделей достигается высокая надежность: так, в демонстрации 2024 г. система успешно распознала попытку выдать себя за Илона Маска в Zoom-звонке, даже при достаточно правдоподобном наложении его лица на другого человека.

Минусы: решение является облачным – видео

[4] Goldman S. Intel unveils real-time deepfake detector, claims 96% accuracy rate [Электронный ресурс] // VentureBeat. November 16, 2022. URL: https://venturebeat.com/ai/intel-unveils-real-time-deepfake-detector-claims-96-accuracy-rate (дата обращения: 16.04.2025).

[5] Colman B. Real-Time Video Deepfake Detection, Now on Reality Defender [Электронный ресурс] // Reality Defender, 2024. URL: https://www.realitydefender.com/insights/real-time-video-deepfake-detection-now-on-realitydefender (дата обращения: 16.04.2025).

участников фактически передается для анализа внешнему сервису, что может вызывать вопросы конфиденциальности. Также возможны ложные срабатывания, особенно на нестандартных эффектах видео (виртуальные фоны, фильтры), поэтому порог чувствительности требует настройки под сценарий. Тем не менее, Reality Defender представляет собой одно из первых коммерчески доступных средств для автоматического выявления дипфейков в онлайн-встречах в режиме реального времени.

Решения в экосистеме Zoom. На стороне самих платформ видеоконференций работа по внедрению детекции дипфейков только начинается. Собственных инструментов у сервисов уровня Zoom или Teams пока не объявлено (к 2025 г. нет встроенной функции "*AI Deepfake Detect*"). Однако крупные платформы открыли API для разработчиков, благодаря чему появились сторонние интеграции. Помимо упомянутого плагина от Reality Defender, существует, например, решение Resemble Detect[6] для аудио и видео проверки: специальный бот в Zoom или Google Meet получает небольшие фрагменты голоса каждого или видео говорящего и прогоняет их через модель выявления синтезированной речи или видео. Если вероятность фейка превышает порог, бот мгновенно уведомляет участников и модерацию чата. Компания Resemble AI утверждает, что ее модель обнаружения дипфейков достигает 98% точности и способна работать в режиме реального времени с задержкой <0.3 с.

Другой подход предлагает компания Beyond Identity: ее плагин RealityCheck для Zoom[7] фокусируется на превентивной защите – он криптографически подтверждает личность и устройство каждого участника перед началом созвона, чтобы злоумышленник не мог выдать себя за другого пользователя.

Эти решения показывают, что индустрия начинает реагировать на новую угрозу. Можно ожидать, что в ближайшем будущем функциональность по автоматическому распознаванию дипфейков будет встроена непосредственно в платформы видеоконференций. Пока же для обеспечения максимальной защиты организациям рекомендуется интегрировать специализированные сервисы (подобные рассмотренным) в свои рабочие процессы видеосвязи. Это позволит значительно снизить риск того, что злоумышленник незаметно проникнет в звонок под чужим обличием.

Заключение

В данной работе была рассмотрена история развития основных технологий создания дипфейков, разобраны современные подходы их обнаружения, включая методы, анализирующие визуальные и поведенческие аномалии, и методы, использующие нейросетевые модели, произведено сравнение методов, основываясь на точности, надежности распознавания, скорости работы и устойчивости к различным видам атак и проанализированы основные наборы данных, на которых должно выполняться обучение моделей. Проведённый обзор подтвердил – универсального решения для надёжного распознавания дипфейков не существует. Даже лучшие современные методы обнаружения демонстрируют снижение точности на ранее не виденных типах атак, а стопроцентная гарантия идентификации фейка остается недостижимой. Тем не менее, накопленные исследования позволяют очертить перспективные направления развития, которые делают задачу обнаружения более реалистичной и управляемой.

Упомянутая выше компания Resemble AI, в своем отчете по дипфейкам за 1-й квартал 2025 года, пишет следующее: «основываясь на нашем анализе более 170 задокументированных инцидентов с дипфейками, мы рекомендуем трехкомпонентный подход к решению этой развивающейся угрозы.

Во-первых, технические решения должны быть приоритетными за счет увеличения инвестиций в технологии обнаружения, которые могут быть широко развернуты на всех платформах, во-вторых – необходима разработка стандартизированных протоколов водяных знаков для синтетических носителей и, в-третьих, необходимо внедрение механизмов аутентификации контента, которые сохраняют информацию о происхождении.

Законодательные рамки требуют гармонизации в разных юрисдикциях для установления последовательных определений вредоносных дипфейков, создания четких стандартов ответственности для платформ и разработки специализированных механизмов принуждения с соответствующей технической экспертизой.

Не менее важно укрепление общественной устойчивости за счет расширенных программ медиаграмотности, ориентированных на уязвимые демографические группы, создания доступных механизмов сообщения о жертвах дипфейков и создания систем поддержки, которые предоставляют как техническую, так и психологическую поддержку».

Необходимо разрабатывать новые методы детектирования. Существующие CNN/ViT-детекторы плюс эвристики моргания-позы-PPG уже близки к предельной полезности при статичном наборе признаков.

Дальнейший прогресс возможен за счет следующих направлений:

- Мультимодальный анализ: одновременное использование видео-, аудио- и текстовых признаков для выявления несогласованностей (например, между мимикой и звуком речи). Первые гибридные прототипы показывают, что добавление звука поднимает AUC на 5-7% по сравнению с «только-изображением».
- Легкие real-time архитектуры: использование квантованных и бинарных нейросетей (BNN), способных обеспечивать более 30 FPS на CPU без заметной задержки. Ранние BNN-решения уже достигают 30 FPS при точности 99%.
- Self-supervised обучение: модели тренируются на стриминговых данных с автоматическим созданием «псевдо-фейков», чтобы не отставать от появляющихся генераторов. В работе [37] демонстрируется такой

[6] Resemble Detect [Электронный ресурс] // Resemble AI, 2025. URL: https://www.resemble.ai/detect/ (дата обращения: 16.04.2025).
[7] RealityCheck [Электронный ресурс] // Beyond Identity, 2025. URL: https://support.beyondidentity.com/hc/en-us/articles/24038107991191-RealityCheck-for-Zoom-Integration-Guide (дата обращения: 16.04.2025).

подход к самообучению дипфейк-детекторов.

Здесь хотелось отметить такой момент. Если вспомнить, как формировались первые физические атаки на модели машинного обучения, то они состояли в фотографировании подготовленных изображений (изображений, сформированных цифровой атакой) [38]. Сегодня, атакующие могут воспользоваться тем, что на мобильных устройствах камеры подвергают изображения встроенной обработке с использованием ИИ. Если цифровой дипфейк будет просто переснят с экрана, то найдет ли детектор что-либо, кроме следов обработки камеры мобильного устройства?

Необходимо внедрить использование цифровых водяных знаков. Компьютерное зрение остаётся первым барьером, но надёжность системы резко растёт, если сочетать его с доказательством происхождения кадра:

- Персистентные водяные знаки: внедряются на уровне видеопотока и сохраняются при сжатии, фиксируя факт модификации.
- Стандарты C2PA: они позволяют криптографически подписывать цепочку редактирования. Видеоклиент может проверить подпись до запуска CV-детектора, а при несовпадении сразу пометить поток как подозрительный.

Таким образом, *watermark*-слой выступает не конкурентом, а «вторым контуром» защиты, снижающим нагрузку на детектор и упрощающим его объяснимость.

Механизмы типа *challenge-response* («назовите код», «поверните голову», «прошепчите фразу») показали свою эффективность: системы вроде GOTCHA [36] уже достигают AUC ~ 88% в live-звонках, когда одних визуальных методов недостаточно. Этот слой сложно автоматизировать, но он критичен для чувствительных переговоров и может запускаться выборочно по уровню риска. Здесь можно отметить следующее. Задания «поверните голову» и т.п. основаны на том, что для атакующего может быть трудно реализовать такое в реальном времени. Но это ограничение, условно говоря, будет работать до следующего поколения GPU-карт. В корпоративной среде можно использовать какие-либо проверяемые подтверждения, типа номер последнего входящего звонка на корпоративном телефоне, адресат email письма и т.п. Одноразовый пароль на другом устройстве в сочетании с корпоративным мессенджером также может быть решением.

Реалистичная архитектура обнаружения дипфейков в видеоконференциях будущего будет состоять из следующих слоёв:

1) Проверка подлинности источника (C2PA, водяные знаки);
2) Лёгкий edge-детектор (квантованная CNN/BNN) для первичного скрининга (классические визуальные артефакты, гладкость/неестественная кожа, аномалии цвета/освещения, моргание или его отсутствие и т.д.);
3) Полноценный многоступенчатый мультимодальный анализ в облаке для подозрительных потоков (несовпадение губ и речи, несогласованная мимика и интонация, PPG-сигналы, нарушения в голосе и т.д.);
4) Активация *challenge-response* при высоком риске;
5) Интерпретируемый модуль (XAI), визуализирующий причину срабатывания (несоответствие мимики и т.д.)

Благодарности

Авторы благодарны сотрудникам лаборатории открытых информационных технологий кафедры информационной безопасности факультета ВМК МГУ имени М.В.Ломоносова за обсуждения и ценные замечания.

Статья написана в рамках развития направления «Кибербезопасность в Искусственном интеллекте» на факультете ВМК МГУ имени М.В.Ломоносова [39].

Работа продолжает серию публикаций по использованию ИИ в кибербезопасности, начатую в статье [40].

References

1. Namiot D.E., Ilyushin E.A., Chizhov I.V. Artificial Intelligence and Cybersecurity. *International Journal of Open Information Technologies*. 2022;10(9):135-147. (In Russ., abstract in Eng.) EDN: DYQWEH
2. Namiot D.E., Ilyushin E.A. On Cyber Risks of Generative Artificial Intelligence. *International Journal of Open Information Technologies.* 2024;12(10):109-119. (In Russ., abstract in Eng.) EDN: JZCUQS
3. Namiot D.E. On Cyberattacks Using Artificial Intelligence Systems. *International Journal of Open Information Technologies.* 2024;12(9):132-141. (In Russ., abstract in Eng.) EDN: PEEQXY
4. Namiot D.E., Ilyushin E.A. Generative Models in Machine Learning. *International Journal of Open Information Technologies.* 2022;10(7):101-118. (In Russ., abstract in Eng.) EDN: WRIKRG
5. Bhattacharyya C., et al. Diffusion Deepfake. *arXiv:*2404.01579. 2024. https://doi.org/10.48550/arXiv.2404.01579
6. Rana Md S., et al. Deepfake Detection: A Systematic Literature Review. *IEEE Access*. 2022;10:25494-25513. https://doi.org/10.1109/ACCESS.2022.3154404
7. Dale K., et al. Video Face Replacement. *ACM Transactions on Graphics*. 2011;30(6):1-10. https://doi.org/10.1145/2070781.2024164
8. Thies, Justus, et al. Face2Face: real-time face capture and reenactment of RGB videos. *Communications of the ACM.* 2019;62(1):96-104. https://doi.org/10.1145/3292039
9. Goodfellow I.J., et al. Generative Adversarial Nets. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Vol. 2 (NIPS'14). Vol. 2. MIT Press, Cambridge, MA, USA; 2014. p. 2672-2680. Available at: https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf (accessed 16.04.2025).

10. de Zeeuw D., Geil A. This Person Does Not Exist': From Real Generalisation to Algorithmic Abstraction in Photographic Portraiture. In: Geil A., Jirsa T. (eds.) Reconfiguring the Portrait. Edinburgh University Press; 2023. p. 43-60. https://doi.org/10.1515/9781399525091-007
11. Skorokhodov I., Tulyakov S., Elhoseiny M. StyleGAN-V: A Continuous Video Generator with the Price, Image Quality and Perks of StyleGAN2. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, LA, USA: IEEE Press; 2022. p. 3616-3626. https://doi.org/10.1109/CVPR52688.2022.00361
12. Qiu H., et al. StyleFaceV: Face Video Generation via Decomposing and Recomposing Pretrained StyleGAN3. *arXiv:*2208.07862. 2022. https://doi.org/10.48550/arXiv.2208.07862
13. Vyas H. Deep Fake Creation by Deep Learning. *Extraction.* 2020;07(07):960-963. Available at: https://www.irjet.net/archives/V7/i7/IRJET-V7I7168.pdf (accessed 16.04.2025).
14. Perov I., et al. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition.* 2023;141:109628. https://doi.org/10.1016/j.patcog.2023.109628
15. Lin B. -S., Hsu D. -W., Shen C. -H., Hsiao H. -F. Using Fully Connected and Convolutional Net for GAN-Based Face Swapping. In: 2020 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS). Ha Long, Vietnam: IEEE Press; 2020. p. 185-188. https://doi.org/10.1109/APCCAS50809.2020.9301665.
16. Rafique R., et al. Deep fake detection and classification using error-level analysis and deep learning. *Scientific Reports.* 2023;13(1):7422. https://doi.org/10.1038/s41598-023-34629-3
17. Lyu S. Deepfake Detection: Current Challenges and Next Steps. In: 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). London, UK: IEEE Press; 2020. p. 1-6. https://doi.org/10.1109/ICMEW46912.2020.9105991
18. Afchar D., Nozick V., Yamagishi J., Echizen I. MesoNet: a Compact Facial Video Forgery Detection Network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS). Hong Kong, China: IEEE Press; 2018. p. 1-7. https://doi.org/10.1109/WIFS.2018.8630761
19. Li Y., Chang M.-C., Lyu S. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *arXiv:*1806.02877. 2018. https://doi.org/10.48550/arXiv.1806.02877
20. Jung T., Kim S., Kim K. DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. *IEEE Access.* 2020;8:83144-83154. https://doi.org/10.1109/ACCESS.2020.2988660
21. Yang X., Li Y., Lyu S. Exposing Deep Fakes Using Inconsistent Head Poses. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK: IEEE Press; 2019. p. 8261-8265. https://doi.org/10.1109/ICASSP.2019.8683164
22. Ciftci U.A., Demir I., Yin L. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020. https://doi.org/10.1109/TPAMI.2020.3009287
23. Bian Y., et al. Revisiting the Effectiveness of Off-the-shelf Temporal Modeling Approaches for Large-scale Video Classification. *arXiv:*1708.03805. 2017. https://doi.org/10.48550/arXiv.1708.03805
24. Ashok V., Joy P.T. Deepfake Detection Using XceptionNet. In: 2023 IEEE International Conference on Recent Advances in Systems Science and Engineering (RASSE). Kerala, India: IEEE Press; 2023. p. 1-5. https://doi.org/10.1109/RASSE60029.2023.10363477
25. Rössler A., Cozzolino D., Verdoliva L., Riess C., Thies J., Niessner M. FaceForensics++: Learning to Detect Manipulated Facial Images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE Press; 2019. p. 1-11. https://doi.org/10.1109/ICCV.2019.00009
26. Güera D., Delp E.J. Deepfake Video Detection Using Recurrent Neural Networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Auckland, New Zealand: IEEE Press; 2018. p. 1-6. https://doi.org/10.1109/AVSS.2018.8639163
27. Nguyen D., et al. FakeFormer: Efficient Vulnerability-Driven Transformers for Generalisable Deepfake Detection. *arXiv:*2410.21964. 2024. https://doi.org/10.48550/arXiv.2410.21964
28. Dolhansky B., et al. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv:*2006.07397. 2020. https://doi.org/10.48550/arXiv.2006.07397
29. Li Y., Yang X., Sun P., Qi H., Lyu S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE Press; 2020. p. 3204-3213. https://doi.org/10.1109/CVPR42600.2020.00327
30. Hondru V., et al. ExDDV: A New Dataset for Explainable Deepfake Detection in Video. *arXiv:*2503.14421. 2025. https://doi.org/10.48550/arXiv.2503.14421
31. Akhtar Z., Pendyala T.L., Athmakuri V.S. Video and Audio Deepfake Datasets and Open Issues in Deepfake Technology: Being Ahead of the Curve. *Forensic Sciences.* 2024;4(3):289-377. https://doi.org/10.3390/forensicsci4030021
32. Sandotra N., Arora B. A comprehensive evaluation of feature-based AI techniques for deepfake detection. *Neural Computing and Applications*. 2024;36(8):3859-3887. https://doi.org/10.1007/s00521-023-09288-0
33. Kaur A., et al. Deepfake video detection: challenges and opportunities. *Artificial Intelligence Review.* 2024;57(6):159. https://doi.org/10.1007/s10462-024-10810-6

34. Balafrej I., Dahmane M. Enhancing practicality and efficiency of deepfake detection. *Scientific Reports*. 2024;14(1):31084. https://doi.org/10.1038/s41598-024-82223-y
35. Romeo L., et al. Faster Than Lies: Real-time Deepfake Detection using Binary Neural Networks. *arXiv:*2406.04932. 2024. https://doi.org/10.48550/arXiv.2406.04932
36. Mittal G., Hegde C., Memon N. Gotcha: Real-Time Video Deepfake Detection via Challenge-Response. In: 2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P). Vienna, Austria: IEEE Press; 2024. p. 1-20. https://doi.org/10.1109/EuroSP60621.2024.00009
37. Xie Z., Luo J. Shaking the Fake: Detecting Deepfake Videos in Real Time via Active Probes. *arXiv:*2409.10889. 2024. https://doi.org/10.48550/arXiv.2409.10889
38. Namiot D. Schemes of attacks on machine learning models. *International Journal of Open Information Technologies.* 2023;11(5):68-86. EDN: YVRDOB
39. Sukhomlin V.A. The Concept and Main Characteristics of the Master's Degree Program "Cybersecurity" of the faculty of Computational Mathematics and Cybernetics of Lomonosov Moscow State University. *International Journal of Open Information Technologies.* 2023;11(7):143-148. (In Russ., abstract in Eng.) EDN: MPHLSQ
40. Namiot D.E., Ilyushin E.A., Chizhov I.V. Attacks on Machine Learning Systems – Common Problems and Methods. *International Journal of Open Information Technologies.* 2022;10(3):17-22. (In Russ., abstract in Eng.) EDN: DZFSKQ
41. Ritter P., Lucian D., Anderies, Chowanda A. Comparative Analysis and Evaluation of CNN Models for Deepfake Detection. In; 2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS). IPOH, Malaysia: IEEE Press; 2023. p. 250-255. https://doi.org/10.1109/AiDAS60501.2023.10284611

Поступила 16.04.2025; одобрена после рецензирования 02.06.2025; принята к публикации 27.06.2025.

Submitted 16.04.2025; approved after reviewing 02.06.2025; accepted for publication 27.06.2025.

Об авторах:

Кузьменко Илья Дмитриевич, магистрант кафедры информационной безопасности факультета вычислительной математики и кибернетики, ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), **ORCID: https://orcid.org/0009-0002-3714-8407,** ilyexakuzmenko@gmail.com

Намиот Дмитрий Евгеньевич, ведущий научный сотрудник лаборатории открытых информационных технологий кафедры информационной безопасности факультета вычислительной математики и кибернетики, ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), доктор технических наук, **ORCID: https://orcid.org/0000-0002-4463-1678,** dnamiot@gmail.com

Васенин Валерий Александрович, заведующий кафедрой математического моделирования и компьютерных исследований механико-математического факультета; заведующий лабораторией автоматизации экспериментальных исследований НИИ механики МГУ, ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), доктор физико-математических наук, профессор, **ORCID: https://orcid.org/0000-0003-2499-6132,** vasenin@msu.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

About the authors:

Ilya D. Kuzmenko, Master degree student of the Department of Information Security, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), **ORCID: https://orcid.org/0009-0002-3714-8407,** ilyexakuzmenko@gmail.com

Dmitry E. Namiot, Senior Researcher of the Open Information Technologies Lab, Department of Information Security, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), Dr. Sci. (Eng.), **ORCID: https://orcid.org/0000-0002-4463-1678,** dnamiot@gmail.com

Valery A. Vasenin, Head of the Department of Mathematical Modeling and Computer Research at the Faculty of Mechanics and Mathematics; Head of the Laboratory of Automation of Experimental Research at the Research Institute of Mechanics at Moscow State University, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), Dr. Sci. (Phys.-Math.), Professor, **ORCID: https://orcid.org/0000-0003-2499-6132,** vasenin@msu.ru

All authors have read and approved the final manuscript.