Содержание

От издательства

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв на нашем сайте www.dmkpress.com, зайдя на страницу книги и оставив комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com; при этом укажите название книги в теме письма.

Если вы являетесь экспертом в какой-либо области и заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы обеспечить высокое качество наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг, мы будем очень благодарны, если вы сообщите о ней главному редактору по адресу dmkpress@gmail.com. Сделав это, вы избавите других читателей от недопонимания и поможете нам улучшить последующие издания этой книги.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательство «ДМК Пресс» очень серьезно относится к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконной публикацией какой-либо из наших книг, пожалуйста, пришлите нам ссылку на интернет-ресурс, чтобы мы могли применить санкции.

Ссылку на подозрительные материалы можно прислать по адресу электронной почты dmkpress@gmail.com.

Мы высоко ценим любую помощь по защите наших авторов, благодаря которой мы можем предоставлять вам качественные материалы.

Благодарности

Здесь будут фамилии тех, кто помогал изданию этой книги, прислав в издательство найденные ошибки или ссылку на подозрительные материалы.

Предисловие

Это руководство представляет собой первый всеобъемлющий сборник тем исследований в области цифровых манипуляций с лицами и их обнаружения, проведенных широким кругом экспертов из различных областей исследований, включая среди прочего компьютерное зрение, распознавание образов, биометрию и медиакриминалистику. Имея основной интерес для исследователей в указанных областях, оно привлекает широкий круг читателей, предоставляя подробные теоретические объяснения основ, а также углубленные исследования текущих тем исследований наряду с всесторонними экспериментальными оценками.

В части I этого руководства читателю представлены вводные обзорные главы, посвященные темам манипуляций видео с лицами и их обнаружения (глава 1), влияния различных манипуляций и методов изменения лиц на системы их распознавания (глава 2) и общей мультимедийной криминалистики до эпохи глубокого обучения (глава 3). Эти главы служат отправной точкой для читателей, желающих получить краткий обзор современных достижений в данных областях.

Часть II посвящена созданию манипулируемого контента лиц и его последствиям для безопасности при распознавании лиц, включая дипфейки (DeepFake, главы 4 и 5), морфингу лиц (глава 6), состязательным изображениям лиц[1] (глава 7) и генерации лиц методом «аудио в видео» (глава 8). Затем в части III подробно рассматриваются методы обнаружения манипуляций с лицами, эта часть содержит главы, посвященные различным современным методам обнаружения синтетически сгенерированных изображений лиц (глава 9), видеороликам с дипфейками (главы 10–14), изображениям измененных лиц (главы 15 и 16) и изображениям ретушированных лиц (глава 17). Главы в частях II и III более подробно раскрывают темы цифровых манипуляций с лицами и обнаружения и ориентированы на продвинутых читателей.

Наконец, часть IV посвящена другим темам, включая использование манипуляций с лицами для повышения конфиденциальности и их обнаружение (глава 18), практические проблемы манипуляций с лицами при удаленной работе (глава 19), а также социальные и этические вопросы (главы 19, 20). Наконец, в заключительной главе, написанной разными авторами этого справочника, обобщаются исследовательские проблемы, требующие разрешения, и будущие тенденции (глава 21).

Мы хотели бы выразить благодарность редакторам серии книг Springer Advances in Computer Vision and Pattern Recognition. Мы также хотели бы поблагодарить всех авторов за плодотворное сотрудничество и их отличный

[1] Атакам с использованием фейковых изображений лиц. – *Прим. ред.*

вклад в это руководство. Работа над этим справочником поддерживалась Федеральным министерством образования и исследований Германии и Министерством высшего образования, исследований, науки и искусств земли Гессен в рамках совместной поддержки Национального исследовательского центра прикладной кибербезопасности ATHENE и проектов PRIMA (H2020-MSCA-ITN-2019-860315), TRESPASS-ETN (H2020-MSCA-ITN-2019-860813), BIBECA (MINECO/FEDER RTI2018-101248-B-I00) и COST CA16101 (MULTI-FORESEE). Наконец, мы хотели бы поблагодарить наши семьи и друзей за их поддержку и ободрение, когда мы работали над этим справочником.

Дармштадт, Германия — Кристиан Ратгеб
Мадрид, Испания — Рубен Толосана
Мадрид, Испания — Рубен Вера-Родригес
Йовик, Норвегия — Кристоф Буш

Часть I

ВВЕДЕНИЕ

Глава 1

Введение в цифровые манипуляции с лицами

Рубен Толосана, Рубен Вера-Родригес, Джулиан Фьеррес, Айтами Моралес и Хавьер Ортега-Гарсия

КРАТКОЕ СОДЕРЖАНИЕ Цифровые манипуляции стали популярной темой в последние несколько лет, особенно после того, как стал популярным термин «дипфейк» (DeepFake). В этой главе представлены известные цифровые манипуляции с особым акцентом на лицевой контент из-за большого количества их возможных применений. В частности, мы рассмотрим принципы шести типов цифровых манипуляций с лицами: (i) полный синтез лица, (ii) подмена идентичности, (iii) морфинг лица, (iv) манипуляция признаками лица, (v) подмена выражения лица (также известная как реконструкция лица или «говорящие лица») и (vi) преобразования «аудио в видео» и «текст в видео». Эти шесть основных типов манипуляций с лицами хорошо известны исследовательскому сообществу, и в последние несколько лет им уделялось наибольшее внимание. Кроме того, в этой главе мы выделяем общедоступные базы данных и код для создания цифрового фейкового контента.

Настоящая глава представляет собой обновленную адаптацию журнальной статьи [1].

1.1. Введение

Традиционно количество и реалистичность цифровых манипуляций с лицами ограничивались отсутствием сложных инструментов редактирования, компетентности в этой области, а также сложностью и трудоемкостью процесса [2–4]. Например, в ранней работе по этой теме [5] удалось изменить движение губ говорящего субъекта в соответствии с другой звуковой дорожкой, установив синхронность между звуковой дорожкой и артикуляцией человека. Тем не менее от первоначальных ручных методов синтеза до наших дней многие вещи развивались и быстро менялись. В настоящее время становится все проще автоматически синтезировать несуществующие лица или манипулировать реальным лицом (также известным как добросовестное представление [6]) одного субъекта на изображении или видео благодаря: (i) свободному доступу к крупномасштабным базам данных и (ii) эволюции методов глубокого обучения, которые устраняют многие этапы ручного редактирования, такие как автокодеры (AE) и генеративно-состязательные сети (GAN) [7, 8]. В результате были выпущены открытое программное обеспечение и мобильные приложения, такие как ZAO[1] и FaceApp[2], которые открывают дверь для создания поддельных изображений и видео любому человеку без какого-либо опыта в этой области.

В этом контексте цифровых манипуляций с лицами есть один термин, который в последнее время доминирует в панораме социальных сетей [9, 10], вызывая в то же время большое общественное беспокойство [11]: дипфейк (DeepFake).

В общем популярный термин DeepFake относится ко всему цифровому поддельному контенту, созданному с помощью методов глубокого обучения [1, 12]. Он был создан после того, как пользователь Reddit под ником «deepfakes» в конце 2017 года заявил, что разработал алгоритм машинного обучения, который помог ему заменить лица актеров порновидео на лица знаменитостей [13]. Наиболее вредоносное использование метода DeepFake – это поддельная порнография, поддельные новости, розыгрыши и финансовые махинации [14]. В результате область исследований, традиционно посвященная общей медиакриминалистике [15–18], активизируется, и в настоящее время все больше усилий направлено на обнаружение манипуляций с лицами на изображениях и видео [19, 20].

Кроме того, часть этих возобновленных усилий по обнаружению поддельных лиц основана на прошлых исследованиях в области обнаружения атак с использованием биометрических данных (также известных как спуфинг) [21–23] и современного глубокого обучения на основе данных [24–27]. В главе 2 представлен вводный обзор манипуляций с лицами в биометрических системах.

Растущий интерес к обнаружению поддельных лиц демонстрируется увеличением числа семинаров на ведущих конференциях [28–32], международ-

[1] https://apps.apple.com/cn/app/id1465199127.

[2] https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341.

ными проектами, такими как MediFor, финансируемыми Агентством перспективных оборонных исследований (DARPA), и конкурсами, такими как Media Forensics Challenge (MFC2018)[1], запущенный Национальным институтом стандартов и технологий (NIST), Deepfake Detection Challenge (DFDC)[2], запущенный Facebook, и недавний DeeperForensics Challenge[3].

В ответ на этот все более изощренный и более реалистичный контент манипуляций с лицами исследовательское сообщество прилагает большие усилия для разработки улучшенных методов их обнаружения [1, 12]. Традиционные методы обнаружения подделок в криминалистике средств массовой информации обычно основывались на: (i) анализе внутренних «отпечатков пальцев» камеры (паттернах артефактов и шумов), оставленных устройством камеры, как аппаратным, так и программным, например объективом камеры [33], массивом цветовых фильтров, обработкой [34, 35], компрессией [36, 37] и пр., и (ii) анализе «внешних отпечатков пальцев» камеры, вносимых программным обеспечением для редактирования, таких как копирование-вставка (вклейка) или копирование-перемещение (клонирование) различных элементов изображения [38, 39], уменьшение частоты кадров в видео [40, 41] и т. д. В главе 3 дается углубленный обзор литературы по традиционной мультимедийной криминалистике до эпохи глубокого обучения.

Тем не менее большинство признаков, рассматриваемых в традиционных методах обнаружения подделок, сильно зависят от конкретного сценария обучения, поэтому они неустойчивы к невидимым условиям [2, 16, 26]. Это имеет особое значение в эпоху, в которой мы живем, поскольку большая часть поддельного медиаконтента обычно распространяется в социальных сетях, платформы которых автоматически модифицируют исходное изображение или видео, например с помощью операций компрессии и масштабирования изображения [19, 20].

Первая глава представляет собой обновленную адаптацию журнальной статьи, представленной в [1], и служит в этой книге вводной частью с описанием наиболее популярных цифровых манипуляций с особым акцентом на лицевой контент из-за большого количества возможных вредоносных приложений, например генерации фейковых новостей, которые среди прочего могут предоставлять дезинформацию о политических выборах и угрозах безопасности [42, 43]. В частности, в разделе 1.2 мы рассматриваем шесть типов цифровых манипуляций с лицами: (i) синтез целевого лица, (ii) замена лица, (iii) морфирование лица, (iv) манипуляция признаками, (v) замена выражения лица (также известная как реконструкция лица, или «говорящее лицо») и (vi) технология «аудио в видео» и «текст в видео». Эти шесть основных типов манипуляций с лицами хорошо известны исследовательскому сообществу, и в последние несколько лет им уделяется наибольшее внимание. Наконец, мы приводим в разделе 1.3 наши заключительные замечания.

---

[1] https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018.

[2] https://www.kaggle.com/c/deepfake-detection-challenge.

[3] https://competitions.codalab.org/competitions/25228.

1.2. Типы цифровых манипуляций с лицами

1.2.1. Синтез целевого лица

Эта манипуляция создает несуществующее изображение целого лица. Ее методы позволяют достичь поразительных результатов, создавая высококачественные изображения лица с высоким уровнем реализма для наблюдателя. На рис. 1.1 показаны некоторые примеры синтеза всего лица, созданного с помощью StyleGAN. Эта манипуляция может принести пользу во многих областях, таких как индустрия видеоигр и 3D-моделирования, но она также может быть использована для вредоносных приложений, таких как создание очень реалистичных поддельных профилей в социальных сетях для распространения дезинформации.

Все манипуляции с синтезом лица создаются с помощью мощных GAN. В общем, GAN состоит из двух разных нейронных сетей, которые соревнуются друг с другом в минимаксной игре[1]: генератор *G*, который фиксирует распределение данных и создает новые образцы, и дискриминатор *D*, который оценивает вероятность того, что образец поступает из данных обучения (настоящих), а не *G* (поддельных). Процедура обучения *G* состоит в том, чтобы максимизировать вероятность того, что *D* совершит ошибку, создав таким образом качественные поддельные образцы. После процесса обучения *D* отбрасывается, а *G* используется для создания фейкового контента. Эта концепция использовалась в последние годы для синтеза всего лица, повышая реалистичность манипуляций, как видно на рис. 1.1.

Рис. 1.1 ❖ Примеры реальных изображений и фейков группы манипуляций **Синтез всего лица**. Настоящие изображения взяты с http://www.whatfaceisreal.com/, а поддельные изображения – с https://thispersondoesnotexist.com

[1] Минимакс – правило принятия решений, используемое в теории игр. – *Прим. ред.*

Одним из первых популярных методов в этом смысле стал ProGAN [44]. Основная идея заключалась в том, чтобы улучшить процесс синтеза, постепенно увеличивая *G* и *D*, т. е. начиная с низкого разрешения и добавляя новые слои, которые моделируют все более мелкие детали по мере обучения. Эксперименты проводились с использованием базы данных CelebA [45], показывая многообещающие результаты для всего синтеза лица. Код архитектуры ProGAN общедоступен на GitHub[1]. Позже Каррас с соавт. предложил расширенную версию под названием StyleGAN [46], которая рассматривала альтернативную архитектуру *G*, мотивированную литературой по передаче стилей [47]. StyleGAN предлагает альтернативную архитектуру генератора, которая приводит к автоматически обучаемому, неконтролируемому разделению атрибутов высокого уровня (например, ракурса и идентичности при обучении на человеческих лицах) и стохастических вариаций в сгенерированных изображениях (например, веснушки, волосы), и это позволяет интуитивно понятное управление синтезом в зависимости от масштаба. Примеры такого рода манипуляций показаны на рис. 1.1 с использованием баз данных CelebA-HQ и FFHQ для обучения StyleGAN [44, 46]. Код архитектуры StyleGAN общедоступен на GitHub[2].

Наконец, одним из известных подходов GAN является StyleGAN2 [48] и Style-GAN2 с адаптивным расширением дискриминатора (StyleGAN2-ADA) [49]. Обучение GAN с использованием слишком небольшого количества данных обычно приводит к переобучению *D*, что приводит к расхождению обучения. StyleGAN2-ADA предлагает адаптивный механизм расширения дискриминатора, который значительно стабилизирует обучение в режимах ограниченных данных. Подход не требует изменений функций потерь или сетевой архитектуры и применим как при обучении с нуля, так и при тонкой настройке существующей GAN на другом наборе данных. Авторы продемонстрировали, что хороших результатов можно добиться, используя всего несколько тысяч обучающих изображений. Код архитектуры StyleGAN2-ADA общедоступен на GitHub[3].

Общедоступны различные базы данных для исследования всех манипуляций с синтезом лица, основанные на этих подходах GAN. В табл. 1.1 приведены основные общедоступные базы данных в этой области с выделением конкретного подхода GAN, рассматриваемого в каждой из них. Интересно отметить, что каждое поддельное изображение может характеризоваться определенным отпечатком пальца GAN точно так же, как естественные изображения идентифицируются отпечатком на основе устройства (т. е. PRNU). На самом деле эти отпечатки пальцев, по-видимому, зависят не только от архитектуры GAN, но и от различных ее реализаций [50, 51].

Кроме того, как указано в табл. 1.1, важно отметить, что общедоступные базы данных содержат только поддельные изображения, созданные с использованием архитектуры GAN. Чтобы иметь возможность проводить эксперименты по обнаружению реальных или поддельных данных в этой группе

[1] https://github.com/tkarras/progressive_growing_of_gans.

[2] https://github.com/NVlabs/stylegan.

[3] https://github.com/NVlabs/stylegan2-ada-pytorch.