Содержание

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 общая мультимедийная криминалистика 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].

Р. Толосана (ruben.tolosana@uam.es), Р. Вера-Родригес (ruben.vera@uam.es), Х. Фьеррес (julian.fierrez@uam.es), А. Моралес (aythami.morales@uam.es), Х. Ортега-Гарсия (javier.ortega@uam.es)
Автономный университет Мадрида, Мадрид, Испания

К. Ратгеб и соавт. (ред.). Справочник по цифровым манипуляциям с лицами и их обнаружению. Достижения в области компьютерного зрения и распознавания образов. https://doi.org/10.1007/978-3-030-87664-7_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.

цифровых манипуляций, исследователям необходимо получать изображения реальных лиц из других общедоступных баз данных, таких как CelebA [45], FFHQ [46], CASIA-WebFace [53], VGGFace2 [54] или Mega-Face2 [55] среди многих других.

Таблица 1.1. Синтез всего лица: общедоступные базы данных

База данных	**Реальные изображения**	**Поддельные изображения**
100K-Generated-Images (2019) [46]	–	100 000 (StyleGAN)
10K-Faces (2019) [52]	–	10 000 (–)
DFFD (2019) [24]	–	100 000 (StyleGAN) 200 000 (ProGAN)
iFakeFaceDB (2019) [26]	–	250 000 (StyleGAN) 80 000 (ProGAN)
100K-Generated-Images (2020) [48]	–	100 000 (StyleGAN2)
100K-Generated-Images (2020) [49]	–	100 000 (StyleGAN2-ADA)

Далее мы приводим краткое описание каждой из общедоступных баз данных. В [46] Каррас и соавт. выпустили набор из 100 000 синтетических изображений лиц, названных 100K-Generated-Images[1]. Эта база данных была создана с использованием предложенной ими архитектуры StyleGAN, которая была обучена с использованием набора данных FFHQ [46].

Другой общедоступной базой данных является 10K-Faces [52], содержащая 10 000 синтетических изображений, созданных для исследовательских целей. В этой базе данных, в отличие от базы данных 100K-Generated-Images, сеть обучалась на фотографиях моделей с изображениями лиц из более контролируемого сценария (например, с плоским фоном). Таким образом, на фоне изображений нет странных артефактов, созданных архитектурой GAN. Кроме того, этот набор данных учитывает другие интересные аспекты, такие как этническую принадлежность и гендерное разнообразие, а также иные метаданные, такие как возраст, цвет глаз, цвет и длина волос, эмоции.

Недавно Данг с соавт. в [24] представили новую базу данных под названием Diverse Fake Face Dataset (DFFD)[2]. Что касается манипуляции синтезом всего лица, авторы создали 100 000 и 200 000 поддельных изображений с помощью предварительно обученных моделей ProGAN и Style-GAN соответственно.

Невес с соавт. представили в [26] базу данных iFakeFaceDB. Эта база данных содержит 250 000 и 80 000 синтетических изображений лиц, изначально созданных с помощью StyleGAN и ProGAN соответственно. Для создания помехи детекторам фейков в качестве дополнительной функции, по сравнению с предыдущими базами данных, в этой базе данных с помощью метода под названием GANprintR (удаление следов GAN) были удалены следы архитектур GAN при сохранении очень реалистичного вида. На рис. 1.2 показаны пример поддельного изображения, созданного непосредственно с помощью

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

[2] http://cvlab.cse.msu.edu/dffd-dataset.html.

StyleGAN, и его улучшенная версия после удаления информации об отпечатке пальца GAN. В результате применения GANprintR база данных iFakeFaceDB представляет собой более сложную задачу для продвинутых детекторов фейков по сравнению с другими базами данных.

(a) Подделка (b) Подделка с помощью GANprintR

Рис. 1.2 ❖ Примеры фейкового изображения, созданного с помощью StyleGAN и его улучшенной версии после удаления информации об отпечатках пальцев GAN посредством GANprintR [26]

Наконец, мы выделяем две популярные общедоступные базы данных 100K-Generated-Images, выпущенные Karpacom с соавт. [48, 49], на основе известных архитектур StyleGAN2 и StyleGAN2-ADA. Соответствующие базы данных фейков, обученные с помощью набора данных FFHQ [46], можно найти на их GitHub[1,2].

В этом разделе описаны основные аспекты манипуляции с синтезом всего лица. Для полного понимания методов обнаружения подделок при этой манипуляции с лицом мы отсылаем читателя к главе 9.

1.2.2. Замена идентичности

Эта манипуляция заключается в замене лица одного субъекта на видео (источнике) лицом другого субъекта (цели). В отличие от полного синтеза лица, где манипуляции выполняются на уровне картинки (кадра), при этой подмене идентичности целью является создание реалистичных фейковых видео. На рис. 1.3 показаны некоторые извлеченные примеры визуального изображения из базы данных видео Celeb-DF [56]. Кроме того, на YouTube можно увидеть очень реалистичные видеоролики с подобными манипуляциями[3]. Многие отрасли могли бы извлечь выгоду из этого типа манипуляций, в частности киноиндустрия[4]. Однако, с другой стороны, ее можно использовать

[1] https://github.com/NVlabs/stylegan2.
[2] https://github.com/NVlabs/stylegan2-ada.
[3] https://www.youtube.com/watch?v=UlvoEW7l5rs.
[4] https://www.youtube.com/c/Shamook/featured.

и для дурных целей, таких как создание порнографических видео знаменитостей, розыгрышей, финансовые махинации, и многих других.

Рис. 1.3 ❖ Примеры реальных и фальшивых лиц группы манипуляции **Identity Swap**. Изображения лиц взяты из видео из базы данных Celeb-DF [56]

Для манипуляций с заменой лиц обычно применяются два типа методов: (i) классические методы на основе компьютерной графики, такие как FaceSwap[1], и (ii) новые методы глубокого обучения, известные как DeepFakes, например недавнее мобильное приложение ZAO[2], и популярные программные инструменты FaceSwap[3] и DeepFaceLab[4]. В общем, для каждого кадра исходного видео в процессе генерации видео с заменой лица рассматриваются следующие этапы [57]: (i) обнаружение и кадрирование лица, (ii) извлечение промежуточных представлений, (iii) синтез нового лица на основе некоторого управляющего сигнала (например, другого лица) и, наконец, (iv) смешивание сгенерированного лица целевого субъекта с исходным видео, как по-

[1] https://github.com/MarekKowalski/FaceSwap.
[2] https://apps.apple.com/cn/app/id1465199127.
[3] https://github.com/deepfakes/faceswap.
[4] https://github.com/iperov/DeepFaceLab.

казано на рис. 1.3. Для каждого из этих этапов можно рассмотреть множество возможностей для улучшения качества поддельных видео. Далее мы описали основные аспекты, рассматриваемые в общедоступных базах данных фейков. За более подробной информацией о процессе генерации мы отсылаем читателя к главам 4 и 14.

Начиная с общедоступных баз данных фейков, таких как база данных UADFV [58], вплоть до последних баз данных Celeb-DF, DFDC, DeeperForensics-1.0 и WildDeepfake [56, 59–61] было выполнено множество визуальных улучшений, повышающих реалистичность поддельных видео. В результате базы данных замены лиц можно разделить на два разных поколения. В табл. 1.2 приведены основные сведения о каждой общедоступной базе данных, сгруппированные по поколениям.

Таблица 1.2. Замена лиц: общедоступные базы данных

База данных	**Реальные видео**	**Поддельные видео**
1-е поколение		
UADFV (2018) [58]	49 (YouTube)	49 (FakeApp)
DeepfakeTIMIT (2018) [11]	–	620 (faceswap-GAN)
FaceForensics++ (2019) [20]	1000 (YouTube)	1000 (FaceSwap) 1000 (DeepFake)
2-е поколение		
DeepFakeDetection (2019) [66]	363 (Актеры)	3068 (DeepFake)
Celeb-DF (2019) [56]	890 (YouTube)	5639 (DeepFake)
DFDC Preview (2019) [59]	1131 (Актеры)	4119 (Разные)
DFDC (2020) [67]	23 654 (Актеры)	104 500 (Разные)
DeeperForensics-1.0 (2020) [60]	50 000 (Актеры)	1000 (DeepFake)
WildDeepfake (2020) [61]	3805 (Интернет)	3509 (DeepFake)

В первом поколении сгруппированы три разные базы данных. UADFV была одной из первых общедоступных баз данных [58]. Эта база данных содержит 49 реальных видеороликов с YouTube, которые были использованы для создания 49 поддельных видеороликов через мобильное приложение FakeApp[1], во всех из которых оригинальное лицо было заменено лицом Николаса Кейджа. Поэтому во всех фейковых видео рассматривается только одна личность. Каждое видео представляет одного человека с типичным разрешением 294×500 пикселей и средней продолжительностью 11,14 с.

Коршунов и Марсель представили в [11] базу данных DeepfakeTIMIT. Эта база данных включает 620 поддельных видеороликов 32 субъектов из базы данных VidTIMIT [62]. Поддельные видео были созданы с использованием общедоступного алгоритма замены лиц на основе GAN[2]. В этом подходе генеративная сеть заимствована из CycleGAN [63] с применением весов FaceNet [64]. Метод Multi-Task Cascaded Convolution Networks используется для более надежного обнаружения и сопоставления лиц [65]. Кроме того, считается, что

[1] https://www.malavida.com/en/soft/fakeapp/.

[2] https://github.com/shaoanlu/faceswap-GAN.

фильтр Калмана сглаживает положение ограничительной рамки по кадрам и устраняет дрожание на замененном лице. Что касается сценариев DeepfakeTIMIT, рассматриваются два разных уровня качества: (i) низкое качество (LQ) с изображениями 64×64 пикселей и (ii) высокое качество (HQ) с изображениями 128×128 пикселей. Кроме того, к фейковым видео в зависимости от уровня качества применялись различные методы смешивания.

Одной из самых популярных баз данных является FaceForensics++ [20]. Эта база данных была представлена в начале 2019 года как расширение исходной базы данных FaceForensics [68], которая была сфокусирована только на замене выражения лиц. FaceForensics++ содержит 1000 реальных видео, взятых из YouTube. Что касается поддельных видео с заменой лица, то они были созданы с использованием как компьютерной графики, так и подходов DeepFake (т. е. обучающего подхода). Для подхода к компьютерной графике авторы рассмотрели общедоступный алгоритм FaceSwap[1], тогда как для подхода DeepFake поддельные видео были созданы с помощью реализации DeepFake FaceSwap GitHub[2]. Метод FaceSwap применяет сопоставление лиц, оптимизацию Гаусса–Ньютона и смешивание изображений лица исходного субъекта на лицо целевого. Метод DeepFake, как указано в [20], основан на двух автокодировщиках с общим кодировщиком, который обучен восстанавливать тренировочные изображения исходного и целевого лиц соответственно. Детектор лиц используется для кадрирования и выравнивания изображений. Для создания фейка к целевому лицу применяются обученные кодировщик и декодер исходного лица. Затем выходные данные автокодера смешиваются с остальной частью изображения с помощью редактирования изображения методом Пуассона [69]. Что касается данных базы данных FaceForensics++, то для каждого метода было сгенерировано 1000 поддельных видео. Позже, при поддержке Google, новый набор данных под названием DeepFakeDetection, помещенный в группу второго поколения из-за его более высокой реалистичности, был включен в структуру FaceForensics++ [66]. Этот набор данных включает 363 реальных видео 28 платных актеров в 16 различных сценах. Кроме того, 3068 фейковых видео включены в набор данных на основе реализации DeepFake FaceSwap GitHub. Важно отметить, что для баз данных FaceForensics++ и DeepFakeDetection учитываются разные уровни качества видео, в частности (i) RAW (исходное качество), (ii) HQ (компрессия с постоянным уровнем, равным 23) и (iii) LQ (компрессия с постоянным уровнем, равным 40). Этот аспект имитирует методы обработки видео, обычно применяемые в социальных сетях.

Недавно было выпущено несколько баз данных, в том числе второго поколения с их большей реалистичностью. Ли с соавт. представлены в [56] базе данных Celeb-DF. Эта база данных предназначена для предоставления фальшивых видео с лучшим визуальным качеством, аналогичным популярным видео, которые размещают в интернете[3], чем в предыдущих базах данных,

[1] https://github.com/MarekKowalski/FaceSwap.

[2] https://github.com/deepfakes/faceswap.

[3] https://www.youtube.com/channel/UCKpH0CKltc73e4wh0_pgL3g.

с низким визуальным по оценке наблюдателя качеством, со множеством видимых артефактов. Celeb-DF состоит из 890 реальных видео, взятых из YouTube, и 5639 поддельных видео, которые были созданы с помощью усовершенствованной версии общедоступного алгоритма генерации дипфейков, улучшающего такие аспекты, как низкое разрешение синтезированных лиц и цветовые несоответствия.

Facebook в сотрудничестве с другими компаниями и академическими учреждениями, такими как Microsoft, Amazon и MIT, запустил в конце 2019 года новую задачу под названием Deepfake Detection Challenge (DFDC) [59]. Сначала они выпустили набор данных для предварительного просмотра, состоящий из 1131 реального видео с 66 платными актерами и 4119 поддельных видео. Позже они выпустили полный набор данных DFDC, включающий более 100 000 фальшивых видео с использованием 8 различных методов замены лиц, таких как автокодировщики, StyleGAN и модели morphable-mask [67].

Еще одна интересная база данных – DeeperForensics-1.0 [60]. Первая версия этой базы данных (1.0) содержит 60 000 видео (50 000 настоящих видео и 10 000 поддельных видео). Реальные видеоролики были записаны в профессиональной закрытой среде (студии) с участием 100 платных актеров с учетом разнообразия пола, возраста, цвета кожи и национальности. Что касается поддельных видео, то они были сгенерированы с использованием недавно предложенной системы сквозной смены лиц на основе вариационных автокодировщиков. Кроме того, учитывались обширные возмущения реального мира (всего до 35), такие как сжатие JPEG, размытие по Гауссу и изменение насыщенности цвета. Все детали базы данных DeeperForensics-1.0 вместе с соответствующим конкурсом описаны в главе 14.

Наконец, Зи с соавт. представили в [61] WildDeepfake, сложную базу данных реальных видео для обнаружения дипфейков. Эта база данных включает 7314 видеороликов (3805 и 3509 настоящих и поддельных видео соответственно), полностью взятых из интернета. В отличие от предыдущих баз данных, WildDeepfake утверждает, что содержит большее разнообразие с точки зрения сцен и людей в каждой сцене, а также выражений лиц.

В заключение этого раздела мы на более высоком уровне обсудим ключевые различия между базами данных фейков 1-го и 2-го поколений. В целом для фейковых видео 1-го поколения характерны: (i) некачественно синтезированные лица, (ii) цветовой контраст между синтезированной фальшивой маской и кожей исходного лица, (iii) видимые границы фейковой маски, (iv) видимые элементы лица из исходного видео, (v) низкие вариации ракурса и (vi) странные артефакты в последовательности кадров. Кроме того, они обычно рассматривают контролируемые сценарии с точки зрения положения камеры и условий освещения. Многие из этих аспектов были успешно улучшены в базах данных 2-го поколения не только на визуальном уровне, но и с точки зрения вариативности (сценарии in-the-wild). Например, недавняя база данных DFDC учитывает различные сценарии съемки (т. е. в помещении и на улице), условия освещения (т. е. день, ночь и т. д.), расстояния от объекта до камеры и варианты ракурса среди прочего. На рис. 1.4 графически представлены недостатки, присутствующие в базах данных обмена идентифика-

ционными данными 1-го поколения, и усовершенствования, реализованные во 2-м поколении. Наконец, также интересно отметить большее количество поддельных видео, включенных в базы данных 2-го поколения.

Рис. 1.4 ❖ Графическое представление недостатков баз данных **Identity Swap** 1-го поколения и улучшений, реализованных во 2-м поколении, не только на визуальном уровне, но и с точки зрения вариативности (сценарии на открытой местности). Поддельные изображения взяты из: UADFV и FaceForensics++ (1-го поколения) [20, 58]; Celeb-DF и DFDC (2-го поколения) [56, 59]

В этом разделе описаны основные аспекты цифровых манипуляций с заменой лиц. Для полного понимания процесса генерации и методов обнаружения подделок мы отсылаем читателя к главам 4, 5 и 10–14.

1.2.3. Морфинг лица

Трансформация (морфинг) лица – это тип цифровой манипуляции с лицом, который можно использовать для создания искусственных биометрических образцов лица, напоминающих биометрическую информацию двух или более человек [70, 71]. Это означает, что новое преобразованное изображение лица будет успешно проверено на образцах лиц этих двух или более человек, что создает серьезную угрозу для систем распознавания лиц [72, 73]. На рис. 1.5 показан пример цифровой манипуляции с изменением лица из [70]. Стоит отметить, что морфинг лица в основном сосредоточен на создании поддельных образцов на уровне отдельного кадра, а не видео, как манипуляции с заменой лица. Кроме того, как показано на рис. 1.5, обычно рассматриваются лица в анфас.

Рис. 1.5 ❖ Пример изображения с **Face morphing** – изображения (**b**), полученного из персонажа 1 (**a**) и персонажа 2 (**c**). Взято из [70]

В последнее время было проведено большое количество исследований в области морфинга лица. Всеобъемлющие обзоры были опубликованы в [70, 74], включая как методы морфинга, так и детекторы морфинговых атак. В целом в процессе генерации морфинговых изображений лица рассматриваются следующие три последовательных этапа: (i) определение соответствий между изображениями лиц разных субъектов. Обычно это выполняется путем извлечения лицевых отметок, например глаз, кончиков носа, рта и т. д.; (ii) реальные изображения лиц испытуемых искажаются до геометрического совмещения соответствующих с отметками образцов; и (iii) значения цвета искаженных изображений объединяются, что называется смешиванием. Наконец, используются методы доработки для исправления привлекающих внимание артефактов, вызванных морфингом на основе пикселей или областей [75, 76].

Недавно были представлены новые ориентиры технологии в области морфинга лиц. Раджа с соавт. представил интересную структуру для решения серьезных вопросов в этой области, таких как независимый бенчмаркинг,

проблемы обобщаемости и соображения возраста, пола и этнической принадлежности [77]. В результате авторы представили новый секвестрированный набор данных и тест[1] для облегчения продвижения в обнаружении трансформирующихся атак. База данных состоит из морфированных и реальных изображений, состоящих из 1800 фотографий 150 субъектов. Морфинг изображения генерируется с использованием шести различных алгоритмов, представляющих широкий спектр возможных методов.

В этом направлении NIST недавно запустил оценку FRVT MORPH[2]. Это текущая оценка, предназначенная для измерения эффективности обнаружения морфинга с двумя отдельными задачами: (i) алгоритмическая способность обнаруживать морфинг лица (трансформированные или смешанные лица) на неподвижных кадрах и (ii) устойчивость алгоритма распознавания лиц к морфингу. Оценка обновляется по мере добавления новых алгоритмов и наборов данных.

Несмотря на эти недавние оценки, мы хотели бы подчеркнуть отсутствие общедоступных баз данных для исследований. Насколько нам известно, единственной общедоступной базой данных является набор данных AMSL Face Morph Image[3] [78]. В основном это происходит из-за того, что большинство баз данных морфинга лица создаются из существующих баз данных лиц. В результате лицензии не могут быть легко переданы, что часто препятствует совместному использованию.

В этом разделе кратко описаны основные аспекты морфинга лица. Для полного понимания методов цифровой генерации и обнаружения подделок мы отсылаем читателя к главам 2, 6, 15 и 16.

1.2.4. Манипуляции с характерными признаками лиц

Эта манипуляция, также известная как редактирование лица или ретушь лица, заключается в изменении некоторых атрибутов лица, таких как цвет волос или кожи, пол, возраст, добавление очков и т. д. [79]. Этот процесс манипуляции обычно осуществляется через GAN, такой как StarGAN, предложенный в [80]. Одним из примеров подобного рода манипуляций является популярное мобильное приложение FaceApp. Потребители могут использовать эту технологию для примерки широкого спектра продуктов, таких как косметика и макияж, очков или прически, в виртуальной среде. На рис. 1.6 показаны некоторые примеры манипуляции атрибутами, сгенерированные с помощью FaceApp [81].

Насколько нам известно, несмотря на успех основанных на GAN структур для манипуляции атрибутами лица [80, 82–88], для исследований в этой области доступно лишь несколько баз данных. Основная причина в том, что код большинства методов GAN находится в открытом доступе, поэтому ис-

[1] https://biolab.csr.unibo.it/fvcongoing/UI/Form/BenchmarkAreas/BenchmarkAreaDMAD.aspx.

[2] https://pages.nist.gov/frvt/html/frvt_morph.html.

[3] https://omen.cs.uni-magdeburg.de/disclaimer/index.php.

следователи могут легко создавать собственные поддельные базы данных по своему усмотрению. Таким образом, этот раздел направлен на то, чтобы осветить последние методы GAN в данной области, от более старых до более современных, а также предоставить ссылку на соответствующие им коды.

Рис. 1.6 ❖ Реальные и сфальцифицированные примеры группы **Attribute Manipulation**. Реальные изображения извлекаются с http://www.whatfaceisreal.com/, а поддельные изображения генерируются с помощью FaceApp

В [86] авторы представили инвертируемую условную нейросеть IcGAN (Invertible Conditional GAN, IcGAN)[1] для сложного редактирования изображений как объединение кодировщика, используемого совместно с условной GAN (cGAN) [89]. Этот метод обеспечивает точные результаты с точки зрения манипуляции атрибутами, однако он серьезно меняет лицо субъекта.

Лэмпл с соавт. предложил в [83] архитектуру кодер-декодер, которая обучена восстанавливать изображения путем разделения существенной информации изображения и значений атрибутов непосредственно в скрытом пространстве[2]. Однако, как это происходит с методом IcGAN, сгенерированные изображения могут не содержать некоторых деталей или дают неожиданные искажения.

Расширенный метод под названием StarGAN[3] был предложен в [80]. До метода StarGAN многие исследования показали многообещающие результаты в преобразовании изображения в изображение для двух доменов в целом. Однако лишь немногие исследования были сосредоточены на работе с более чем двумя областями. В этом случае прямым методом было бы построение разных моделей независимо для каждой пары доменов изображений. StarGAN предложил новый метод, позволяющий выполнять преобразование изображения в изображение для нескольких доменов с использованием только одной модели. Авторы обучили сеть условной передачи атрибутов через потерю классификации атрибутов и потерю согласованности цикла. Были

[1] https://github.com/Guim3/IcGAN.
[2] https://github.com/facebookresearch/FaderNetworks.
[3] https://github.com/yunjey/stargan/blob/master/README.md.

достигнуты хорошие визуальные результаты по сравнению с предыдущими методами. Однако иногда он включает нежелательные модификации входного изображения лица, например цвет кожи.

Почти в то же время Хи с соавт. предложили в [82] attGAN[1], новый метод, который удаляет строгое независимое от атрибутов ограничение из скрытого представления и просто применяет ограничение классификации атрибутов к сгенерированному изображению, чтобы гарантировать правильное изменение атрибутов. AttGAN обеспечивает самые современные результаты реалистичной манипуляции атрибутами с хорошо сохраненными другими деталями лица.

Одним из последних предложенных в литературе методов является STGAN[2] [84]. В общем, с манипуляциями атрибутами можно справиться, включив кодер-декодер или GAN. Однако, как прокомментировали Лю с соавт. [84], слой, который является узким местом в кодере-декодере и обычно дает размытые и низкокачественные результаты манипуляции. Чтобы улучшить это, авторы представили и включили блоки выборочной передачи с кодером-декодером для одновременного улучшения возможностей манипуляции атрибутами и качества изображения. В результате STGAN в последнее время превзошли все современные достижения в области манипуляции атрибутами.

Наконец, мы хотели бы выделить два недавних метода манипуляции атрибутами, которые в настоящее время также позволяют достичь очень реалистичных визуальных результатов: RelGAN и SSCGAN [90, 91].

RelGAN улучшает многодоменное преобразование изображения в изображение, тогда как SSCGAN вводит информацию о целевом атрибуте в несколько путей соединения с пропуском стиля между кодером и декодером, чтобы захватить глобальную статистику лица.

Несмотря на то что коды большинства методов манипуляции атрибутами общедоступны, отсутствие общедоступных баз данных и экспериментальных протоколов имеет решающее значение при сравнении различных методов обнаружения манипуляций, так что невозможно провести справедливое сравнение между исследованиями. На сегодняшний день, насколько нам известно, база данных DFFD [24] кажется единственной общедоступной базой данных, которая рассматривает этот тип лицевых манипуляций. Эта база данных содержит 18 416 и 79 960 поддельных изображений, созданных с помощью методов FaceApp и StarGAN соответственно.

В данном разделе кратко описаны основные аспекты манипуляции характерными признаками лица. Для полного понимания этой группы цифровых манипуляций мы отсылаем читателя к главе 17.

1.2.5. Изменение выражения лица

Эта манипуляция, также известная как реконструкция лица, заключается в изменении выражения лица субъекта. Хотя в литературе предлагаются разные техники манипуляции, например на уровне изображения с помощью

[1] https://github.com/LynnHo/AttGAN-Tensorflow.

[2] https://github.com/csmliu/STGAN.

популярных архитектур GAN [84], в этой группе мы сосредоточимся на наиболее популярных техниках Face2Face и NeuralTextures [92, 93], которые заменяют выражение лица одного субъекта в видео на выражение лица другого субъекта. На рис. 1.7 показаны некоторые наглядные примеры, извлеченные из базы данных FaceForensics++ [20]. Этот тип манипуляции может иметь серьезные последствия, например популярное видео Марка Цукерберга, говорящего то, чего он никогда не говорил[1].

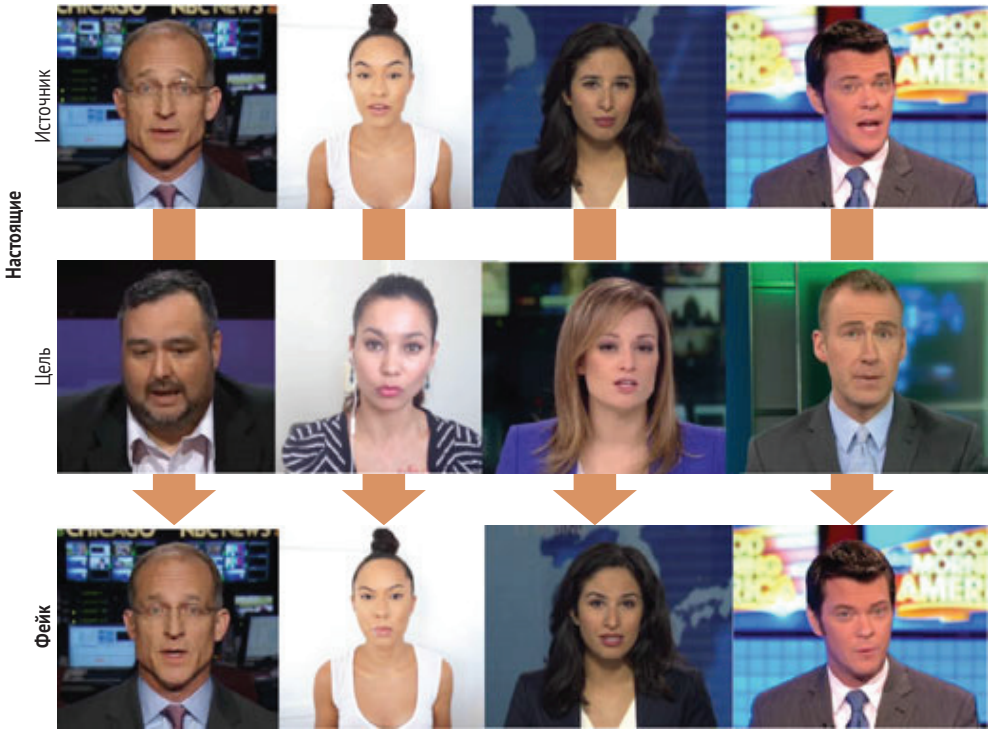

Рис. 1.7 ❖ Реальные и фальшивые примеры группы манипуляций **Expression Swap**. Изображения извлечены из видео из базы данных FaceForensics++ [20]

Насколько нам известно, единственной доступной базой данных для исследований в этой области является FaceForensics++ [20], расширение FaceForensics [68].

Первоначально база данных FaceForensics была сфокусирована на метод Face2Face [93]. Это метод компьютерной графики, который переносит выражение исходного видео в целевое видео, сохраняя при этом идентичность целевого субъекта. Задача выполняется путем ручного выбора ключевых кадров. А именно первые кадры каждого видео используются для получения временной идентификации лица (т. е. 3D-модели) и потом отслеживается выражение на дальнейших кадрах. Затем генерируются поддельные видео

[1] https://www.bbc.com/news/technology-48607673.

путем передачи исходных параметров выражения каждого кадра (т. е. 76 коэффициентов Blendshape) в целевое видео. Позже те же авторы представили в FaceForensics++ новый метод обучения, основанный на NeuralTextures [92]. Это метод рендеринга, который использует исходные видеоданные для изучения нейронной текстуры целевого объекта, включая сеть рендеринга. В частности, авторы рассмотрели в своей реализации потерю GAN на основе патчей, используемую в Pix2Pix [94].

Было изменено только выражение лица, связанное со ртом. Важно отметить, что все данные доступны на FaceForensics++ GitHub[1]. Всего с YouTube взято 1000 реальных видео. Что касается видео с манипуляцией, доступно 2000 фейковых видео (по 1000 видео для каждого рассматриваемого метода подделки). Кроме того, важно подчеркнуть, что учитываются различные уровни качества видео, в частности (i) RAW (исходное качество), (ii) HQ (компрессия постоянного уровня 23) и (iii) LQ (компрессия постоянного уровня 40). Этот аспект имитирует методы обработки видео, обычно применяемые в социальных сетях.

В дополнение к методам Face2Face и NeuralTexture, рассматриваемым в манипуляциях с заменой выражений на уровне видео, недавно были предложены различные методы для изменения выражения лица как на кадрах, так и на видео. Очень популярный метод был представлен в [95]. Авербух-Элор с соавт. предложили метод автоматической анимации неподвижного портрета с использованием видео другого объекта, перенося мимику объекта на видео на целевой портрет. В отличие от методов Face2Face и NeuralTexture, которые требуют видео как с исходного, так и с целевого лица, в [95] требуется только изображение цели. В этом направлении были представлены недавние методы, дающие поразительные результаты как в однократном, так и в малократном обучении [96–98].

1.2.6. «Аудио в видео» и «текст в видео»

Связанной с обменом выражениями темой является синтез видео из аудио или текста. На рис. 1.8 показан пример манипуляции с лицом аудио и текста в видео. Эти типы манипуляций с видеолицом также известны как DeepFakes с синхронизацией по губам [99], или звуковая реконструкция лица [100]. Популярные примеры можно увидеть в интернете[2].

Что касается синтеза поддельных видео из аудио (аудио в видео), Суваджанакорн с соавт. представили в [101] метод синтеза высококачественных видео субъекта (в данном случае Обамы), говорящего с точной синхронизацией губ. Для этого они использовали в качестве исходных данных для своего метода многочасовые предыдущие видеоролики на эту тему вместе с новой аудиозаписью. В своем методе они использовали рекуррентную нейронную сеть (на основе долговременной кратковременной памяти, LSTM), чтобы изучить сопоставление необработанных звуковых характеристик с формами

[1] https://github.com/ondyari/FaceForensics.

[2] https://www.youtube.com/watch?v=VWMEDacz3L4.

рта. Затем, основываясь на форме рта в каждом кадре, они синтезировали высококачественную текстуру рта и скомпоновали ее с выравниванием ракурса в 3D, чтобы создать новое видео, соответствующее входной звуковой дорожке, что дало фотореалистичные результаты.

Рис. 1.8 ❖ Реальный и поддельный примеры группы манипуляций с лицами **Audio-to-Video** и **Text-to-Video**

В [102] Сонг с соавт. предложили метод, основанный на новой условно-рекуррентной сети генерации, которая включает в себя признаки изображения и звука в рекуррентном блоке для временной зависимости, а также пару пространственно-временных дискриминаторов для лучшего качества изображения или видео. В результате их метод может моделировать как губы, так и рот вместе с выражением лица и вариациями ракурса головы в целом, достигая гораздо более реалистичных результатов. Исходный код общедоступен на GitHub[1]. Также в [103] Сонг с соавт. представили динамический метод, не предполагающий специфическую сеть рендеринга, как в [101]. В своем методе они могут создавать очень реалистичные поддельные видео, выполняя реконструкцию 3D-модели лица из входного видео, а также рекуррентную сеть для преобразования исходного звука в параметры выражения.

Наконец, они представили новую сеть рендеринга видео и метод динамического программирования для создания когерентного во времени и фотореалистичного видео. Видеорезультаты показываются в интернете[2].

Другой интересный метод был представлен в [104]. Чжоу с соавт. предложили новый фреймворк под названием Disentangled Audio-Visual System (DAVS), который генерирует высококачественные видео говорящих лиц, используя распутанное аудиовизуальное представление. Как звуковая, так и видеоинформация речи может использоваться в качестве внешнего управления. Исходный код доступен на GitHub[3].

Что касается синтеза поддельных видео из текста (текст в видео), Фрид с соавт. в [105] предложили метод, который принимает в качестве входных дан-

[1] https://github.com/susanqq/Talking_Face_Generation.
[2] https://wywu.github.io/projects/EBT/EBT.html.
[3] https://github.com/Hangz-nju-cuhk/Talking-Face-Generation-DAVS.

ных видео говорящего субъекта и желаемый текст, который должен быть произнесен, и синтезирует новое видео, в котором рот субъекта синхронизирован с новыми словами. В частности, их метод автоматически аннотирует входное видео говорящей головы фонемами, виземами, ракурсом и геометрией трехмерного лица, бликами, мимикой и освещением сцены для каждого кадра. Наконец, рекуррентная сеть генерации видео создает фотореалистичное видео, соответствующее отредактированной расшифровке. Примеры поддельных видео, созданных с помощью этого метода, находятся в открытом доступе[1].

Наконец, мы хотели бы выделить работу, представленную в [100], под названием Neural Voice Puppetry. Тис с соавт. предложили метод синтеза видео целевого актера с голосом любого неизвестного исходного актера или даже синтетических голосов, которые могут быть сгенерированы с использованием стандартных методов преобразования текста в речь, с достижением поразительных визуальных результатов[2].

Насколько нам известно, не существует общедоступных баз данных и эталонных тестов, связанных с обнаружением фальшивого контента аудио и текста в видео. Исследования по этой теме обычно проводятся путем синтеза внутренних данных с использованием общедоступных реализаций, подобных описанным в данном разделе.

В этом разделе кратко описаны основные аспекты преобразования «аудио в видео» и «текста в видео» лиц. Для полного понимания этой группы цифровых манипуляций мы отсылаем читателя к главе 8.

1.3. Выводы

Эта глава послужила введением в наиболее популярные в литературе цифровые манипуляции с лицами. В частности, мы рассмотрели шесть групп манипуляций: (i) синтез всего лица, (ii) замена лица, (iii) редактирование лица, (iv) манипуляция атрибутами, (v) замена мимики (также известная как реконструкция лица, или говорящие лица) и (vi) преобразования «аудио в видео» и «текст в видео». Для каждого из них мы описали основные принципы, общедоступные базы данных и код для генерации цифрового фейкового контента.

За более подробной информацией о цифровых манипуляциях с лицом и методах обнаружения подделок мы отсылаем читателя к частям II и III настоящей книги. Наконец, в части IV описываются дополнительные темы, тенденции и проблемы в области цифровой манипуляции и распознавания лиц.

БЛАГОДАРНОСТИ Эта работа была поддержана проектами: 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).

[1] https://www.ohadf.com/projects/text-based-editing/.

[2] https://justusthies.github.io/posts/neural-voice-puppetry/.