

# Содержание

От издательства .....	16
Предисловие .....	17
<b>Часть I. ВВЕДЕНИЕ</b> .....	19
<b>Глава 1. Введение в цифровые манипуляции с лицами</b> .....	20
1.1. Введение .....	21
1.2. Типы цифровых манипуляций с лицами .....	23
1.2.1. Синтез целевого лица .....	23
1.2.2. Замена идентичности .....	26
1.2.3. Морфинг лица .....	32
1.2.4. Манипуляции с характерными признаками лиц.....	33
1.2.5. Изменение выражения лица .....	35
1.2.6. «Аудио в видео» и «текст в видео».....	37
1.3. Выводы.....	39
Литература .....	40
<b>Глава 2. Цифровые манипуляции с лицами в биометрических системах</b> .....	46
2.1. Введение .....	47
2.2. Биометрические системы .....	48
2.2.1. Процессы .....	49
2.2.2. Распознавание лиц .....	50
2.3. Цифровые манипуляции с лицами в биометрических системах.....	51
2.3.1. Влияние на биометрические характеристики .....	51
2.3.2. Методы обнаружения манипуляций.....	53
2.4. Эксперименты .....	56
2.4.1. Постановка эксперимента.....	56
2.4.2. Оценка эффективности .....	58
2.5. Выводы и перспективы .....	61
Литература .....	62
<b>Глава 3. Мультимедийная криминалистика до эпохи глубокого обучения</b> .....	65
3.1. Введение .....	66
3.2. Метод на основе PRNU .....	68

3.2.1. Определение PRNU .....	70
3.2.2. Вычисление остаточного шума .....	71
3.2.3. Тест на обнаружение подделки.....	71
3.2.4. Анализ на основе управляемой фильтрации .....	73
3.3. Слепые методы.....	76
3.3.1. Паттерны шума .....	76
3.3.2. Артефакты компрессии .....	80
3.3.3. Артефакты редактирования.....	82
3.4. Методы обучения с признаками, созданными вручную.....	84
3.5. Выводы.....	85
Литература .....	87

## **Часть II. ЦИФРОВЫЕ МАНИПУЛЯЦИИ С ЛИЦАМИ И ПРИЛОЖЕНИЯ БЕЗОПАСНОСТИ .....**

91

<b>Глава 4. Создание дипфейков и борьба с ними.....</b>	<b>92</b>
4.1. Введение.....	92
4.2. Основы .....	96
4.2.1. Генерация дипфейк-видео .....	96
4.2.2. Методы обнаружения дипфейков .....	97
4.2.3. Существующие наборы данных дипфейков .....	98
4.3. Celeb-DF: создание дипфейков .....	99
4.3.1. Метод синтеза .....	100
4.3.2. Визуальное качество .....	103
4.3.3. Оценки.....	104
4.4. Landmark Breaker: препятствие для DeepFake .....	106
4.4.1. Экстракторы лицевых отметок.....	106
4.4.2. Состязательные возмущения.....	106
4.4.3. Обозначения и формулировка.....	107
4.4.4. Оптимизация.....	107
4.4.5. Установки эксперимента .....	108
4.4.6. Результаты .....	110
4.4.7. Анализ устойчивости.....	112
4.4.8. Исследование абляции .....	114
4.5. Заключение .....	115
Литература .....	116

## **Глава 5. Угроза дипфейков для компьютерного зрения и человеческого зрительного восприятия .....**

119

5.1. Введение.....	119
5.2. Сопутствующие работы.....	121
5.3. Базы данных и методы .....	122
5.3.1. DeepfakeTIMIT .....	122
5.3.2. DF-Mobio .....	123
5.3.3. Google и Jigsaw.....	124
5.3.4. Facebook.....	124

5.3.5. Celeb-DF.....	125
5.4. Протоколы оценки эффективности.....	126
5.4.1. Измерение уязвимости.....	126
5.4.2. Измерение эффективности распознавания дипфейков.....	127
5.5. Уязвимость систем распознавания лиц.....	127
5.6. Субъективная оценка человеческого визуального восприятия.....	128
5.6.1. Результаты субъективной оценки.....	131
5.7. Оценка алгоритмов обнаружения дипфейков.....	134
5.8. Заключение.....	136
Литература.....	137

## **Глава 6. Создание морфа и уязвимость систем распознавания лиц к морфингу.....**

6.1. Введение.....	140
6.2. Генерация морфинга лица.....	143
6.2.1. Морфинг на основе лицевых отметок.....	143
6.2.2. Генерация морфинга лица на основе глубокого обучения.....	149
6.3. Уязвимость систем распознавания лиц к морфированию лица.....	151
6.3.1. Наборы данных.....	152
6.3.2. Результаты.....	153
6.3.3. Результаты морфинга на основе глубокого обучения.....	158
6.4. Выводы.....	158
Литература.....	159

## **Глава 7. Состязательные атаки на системы распознавания лиц.....**

7.1. Введение.....	162
7.2. Классификация атак на FRS.....	165
7.2.1. Модель угрозы.....	166
7.3. Отравляющие атаки на FRS.....	170
7.3.1. Метод быстрого градиентного знака.....	170
7.3.2. Прогнозируемый градиентный спуск.....	170
7.4. Атаки Карлини и Вагнера (CW).....	171
7.5. Модель ArcFace FRS.....	172
7.6. Эксперименты и анализ.....	173
7.6.1. Чистый набор данных.....	173
7.6.2. Набор данных атак.....	174
7.6.3. Модель FRS для базовой проверки.....	175
7.6.4. Базовая оценка эффективности FRS.....	175
7.6.5. Эффективность FRS при отравлении проверочных данных.....	178
7.6.6. Эффективность FRS при отравлении данных регистрации.....	179
7.7. Столкновение состязательного обучения с атаками FGSM.....	180
7.8. Обсуждение.....	182
7.9. Выводы и будущие направления разработок.....	183
Литература.....	183

<b>Глава 8. Генерация говорящих лиц: «аудио в видео»</b> .....	187
8.1. Введение .....	187
8.2. Сопутствующие методы .....	189
8.2.1. Звуковое представление .....	189
8.2.2. Моделирование лица .....	190
8.2.3. Анимация звук–лицо .....	194
8.2.4. Постпроцессинг .....	201
8.3. Наборы данных и метрики .....	201
8.3.1. Набор данных .....	201
8.3.2. Метрики .....	203
8.4. Обсуждение .....	205
8.4.1. Тонкий контроль лица .....	205
8.4.2. Обобщение .....	207
8.5. Заключение .....	208
8.6. Дополнительная литература .....	209
Литература .....	209

### **Часть III. ОБНАРУЖЕНИЕ ЦИФРОВЫХ МАНИПУЛЯЦИЙ С ЛИЦАМИ** .....

216

<b>Глава 9. Обнаружение синтетических лиц, созданных искусственным интеллектom</b> .....	217
9.1. Введение .....	218
9.2. Генерация лиц с помощью искусственного интеллекта .....	220
9.3. Отпечатки пальцев GAN .....	221
9.4. Методы обнаружения в пространственной области .....	224
9.4.1. Признаки ручной работы .....	225
9.4.2. Признаки, управляемые данными .....	226
9.5. Методы обнаружения по областям частот .....	227
9.6. Обучение обобщающих особенностей .....	228
9.7. Обобщающий анализ .....	230
9.8. Анализ надежности .....	232
9.9. Дальнейший анализ обнаружения GAN .....	233
9.10. Нерешенные проблемы .....	235
Литература .....	238

<b>Глава 10. 3D-архитектура CNN и механизмы внимания для обнаружения дипфейков</b> .....	242
10.1. Введение .....	243
10.2. Сопутствующие исследования .....	245
10.2.1. Обнаружение дипфейков .....	245
10.2.2. Механизмы внимания .....	247
10.3. Набор данных .....	253
10.4. Алгоритмы .....	254
10.5. Эксперименты .....	254

10.5.1. Все техники манипуляции.....	255
10.5.2. Отдельные техники манипуляций.....	257
10.5.3. Техники перекрестной манипуляции.....	258
10.5.4. Эффект внимания в 3D ResNets.....	259
10.5.5. Визуализация соответствующих признаков в обнаружении дипфейка.....	260
10.6. Выводы.....	260
Литература.....	262

<b>Глава 11. Обнаружение дипфейков с использованием нескольких модальностей данных.....</b>	<b>266</b>
11.1. Введение.....	267
11.2. Обнаружение дипфейков с помощью пространственно-временных особенностей видео.....	268
11.2.1. Обзор.....	270
11.2.2. Модельный компонент.....	270
11.2.3. Детали обучения.....	273
11.2.4. Бустинговая нейронная сеть.....	273
11.2.5. Аугментация времени тестирования.....	274
11.2.6. Анализ результатов.....	274
11.3. Обнаружение дипфейков с помощью анализа аудиоспектрограммы.....	276
11.3.1. Обзор.....	277
11.3.2. Набор данных.....	278
11.3.3. Генерация спектрограммы.....	278
11.3.4. Сверточная нейронная сеть (CNN).....	279
11.3.5. Результаты экспериментов.....	280
11.4. Обнаружение дипфейков посредством анализа несоответствия аудио и видео.....	280
11.4.1. Обнаружение несоответствия аудио и видео посредством несоответствия фоном и визем.....	282
11.4.2. Обнаружение дипфейков с использованием аффективных сигналов.....	284
11.5. Заключение.....	286
Литература.....	286

<b>Глава 12. Обнаружение дипфейков на основе определения сердечного ритма: однокадровый и многокадровый методы.....</b>	<b>289</b>
12.1. Введение.....	290
12.2. Сопутствующие работы.....	293
12.3. DeepFakesON-Phys.....	297
12.4. Базы данных.....	298
12.4.1. База данных Celeb-DF v2.....	298
12.4.2. DFDC Preview.....	299
12.5. Экспериментальный протокол.....	299
12.6. Результаты обнаружения фейков: DeepFakesON-Phys.....	300

12.6.1. Обнаружение дипфейков на уровне кадра .....	300
12.6.2. Обнаружение дипфейков на уровне короткого видео .....	303
12.7. Выводы .....	304
Литература .....	306

### **Глава 13. Капсульно-криминалистические сети для обнаружения дипфейков**

<b>для обнаружения дипфейков</b> .....	310
13.1. Введение .....	311
13.2. Сопутствующие работы.....	313
13.2.1. Генерация дипфейков .....	313
13.2.2. Обнаружение дипфейков.....	314
13.2.3. Проблемы обнаружения дипфейков.....	315
13.2.4. Капсульные сети .....	316
13.3. Капсульная криминалистика .....	316
13.3.1. Зачем нужна капсульная криминалистика? .....	316
13.3.2. Обзор .....	317
13.3.3. Архитектура .....	317
13.3.4. Алгоритм динамической маршрутизации.....	319
13.3.5. Визуализация .....	321
13.4. Оценка .....	321
13.4.1. Наборы данных .....	324
13.4.2. Метрики .....	325
13.4.3. Эффект улучшений .....	325
13.4.4. Сравнение экстракторов особенностей лиц.....	326
13.4.5. Влияние слоев статистического пулинга .....	327
13.4.6. Сеть Capsule-Forensics по сравнению с CNN: замеченные атаки.....	328
13.4.7. Сеть Capsule-Forensics против CNN: невидимые атаки .....	330
13.5. Заключение и будущая работа .....	333
13.6. Приложение .....	333
Литература .....	335

### **Глава 14. Обнаружение дипфейков: набор данных**

<b>DeeperForensics и постановка задачи</b> .....	339
14.1. Введение .....	340
14.2. Сопутствующие работы.....	342
14.2.1. Методы создания дипфейков .....	342
14.2.2. Методы обнаружения дипфейков .....	343
14.2.3. Наборы данных для обнаружения дипфейков .....	344
14.2.4. Лучшие тесты обнаружения дипфейков .....	345
14.3. Набор данных DeeperForensics-1.0 .....	346
14.3.1. Сбор данных .....	346
14.3.2. Вариационный автокодировщик дипфейков.....	348
14.3.3. Масштаб и разнообразие.....	353
14.3.4. Набор скрытых тестов .....	354
14.4. DeeperForensics Challenge 2020 .....	355

14.4.1. Платформа .....	356
14.4.2. Набор данных задачи .....	356
14.4.3. Критерии оценки .....	356
14.4.4. Таймлайн .....	357
14.4.5. Результаты и решения .....	357
14.5. Обсуждение .....	362
14.6. Дополнительная литература .....	362
Литература .....	363

## **Глава 15. Методы обнаружения морфинговых атак лица** ..... 369

15.1. Введение .....	369
15.2. Сопутствующие работы .....	371
15.3. Конвейер обнаружения морфинговых атак .....	372
15.3.1. Подготовка данных и извлечение признаков .....	373
15.3.2. Подготовка признаков и обучение классификатора .....	373
15.4. База данных .....	374
15.4.1. Морфинг изображения .....	376
15.4.2. Постпроцессинг изображения .....	378
15.5. Методы обнаружения морфинговых атак .....	379
15.5.1. Предварительная обработка .....	379
15.5.2. Извлечение признаков .....	380
15.5.3. Классификация .....	382
15.6. Эксперименты .....	382
15.6.1. Обобщаемость .....	383
15.6.2. Эффективность обнаружения .....	384
15.6.3. Постпроцессинг .....	384
15.7. Заключение .....	386
Литература .....	387

## **Глава 16. Практическая оценка методов обнаружения морфинговых атак лица** ..... 390

16.1. Введение .....	391
16.2. Сопутствующие работы .....	393
16.3. Создание наборов данных морфинга .....	394
16.3.1. Создание морфов .....	394
16.3.2. Наборы данных .....	395
16.4. Обнаружение морфинговых атак лиц на основе текстур .....	396
16.5. Маскировка морфинга .....	397
16.6. Эксперименты и результаты .....	399
16.6.1. Эффективность набора данных .....	399
16.6.2. Эффективность перекрестного набора данных .....	400
16.6.3. Эффективность смешанного набора данных .....	400
16.6.4. Устойчивость к аддитивному гауссову шуму .....	401
16.6.5. Устойчивость к масштабированию .....	401
16.6.6. Выбор субъектов с похожими лицами .....	402
16.7. Детектор SOTAMD .....	403

16.8. Заключение .....	404
Литература .....	404

## **Глава 17. Ретушь лица и обнаружение изменений..... 407**

17.1. Введение .....	408
17.2. Ретуширование и обнаружение изменений – обзор.....	410
17.2.1. Обнаружение цифровой ретуши .....	410
17.2.2. Обнаружение цифровых изменений .....	413
17.2.3. Общедоступные базы данных .....	415
17.3. Экспериментальная оценка и наблюдения.....	417
17.3.1. Обнаружение междоменных изменений .....	420
17.3.2. Обнаружение изменений перекрестных манипуляций.....	421
17.3.3. Обнаружение межэтнических изменений .....	422
17.4. Нерешенные проблемы .....	423
17.5. Заключение.....	424
Литература .....	425

## **Часть IV. ДАЛЬНЕЙШИЕ ТЕМЫ, ТЕНДЕНЦИИ И ПРОБЛЕМЫ..... 429**

### **Глава 18. Улучшение конфиденциальности мягкой биометрии..... 430**

18.1. Введение .....	431
18.2. Предыстория и сопутствующие работы .....	434
18.2.1. Формулировка проблемы и существующие решения.....	434
18.2.2. Модели мягкобиометрической конфиденциальности .....	435
18.2.3. Обнаружение повышения конфиденциальности .....	437
18.3. Обнаружение вмешательства через несоответствие прогнозов (PREM).....	437
18.3.1. Обзор PREM .....	438
18.3.2. Сверхвысокое разрешение для восстановления признаков .....	439
18.3.3. Измерение несоответствия прогноза .....	440
18.3.4. Краткое описание и характеристики PREM.....	441
18.4. Эксперименты и результаты .....	442
18.4.1. Наборы данных и экспериментальные установки .....	442
18.4.2. Используемые модели конфиденциальности .....	443
18.4.3. Детали реализации .....	444
18.4.4. Результаты и обсуждения .....	445
18.5. Заключение .....	450
Литература .....	451

### **Глава 19. Обнаружение манипуляций с лицами в удаленных операционных системах..... 455**

19.1. Введение .....	456
19.2. Удаленная регистрация документов, удостоверяющих личность.....	457
19.3. Алгоритмы манипуляции с лицом .....	458
19.3.1. Категории атак .....	458
19.3.2. Общие алгоритмы манипуляции с лицом .....	462

19.4. Обнаружение манипуляций с лицами .....	464
19.4.1. Методы, специфичные для лица .....	465
19.4.2. Методы, независимые от лица.....	466
19.4.3. Наборы данных .....	469
19.5. Контркриминалистика и меры противодействия .....	471
19.5.1. Контркриминалистика.....	471
19.5.2. Меры противодействия .....	472
19.6. Базовая структура, стандартизация и правовые аспекты .....	475
19.7. Выводы.....	476
Литература .....	477

## **Глава 20. Перспективы, социальные и этические проблемы, связанные с биометрией при удаленной адаптации .....**

20.1. Введение .....	483
20.2. Похищение идентичности и растущая потребность в ее удаленной проверке .....	485
20.2.1. Риски и социальные последствия похищения идентичности.....	485
20.2.2. Необходимость удаленной биометрической верификации идентичности .....	487
20.3. Технологии удаленной биометрической идентификации .....	490
20.3.1. Появление биометрической удаленной идентификации .....	490
20.3.2. Технологии удаленной биометрической идентификации.....	494
20.4. Этика, конфиденциальность и социальная приемлемость биометрической идентификации .....	497
20.4.1. Риски и основные этические проблемы .....	497
20.4.2. Целостность практической идентичности .....	500
20.4.3. Конфиденциальность и функциональные нарушения .....	501
20.4.4. Этические проблемы, возникающие в результате алгоритмически обусловленных действий и решений.....	503
20.4.5. Общественное признание технологии .....	505
20.5. Обсуждение и выводы .....	506
Литература .....	509

## **Глава 21. Грядущие тенденции в области цифровых манипуляций с лицами и их обнаружения .....**

21.1. Введение .....	514
21.2. Реализм манипуляций с лицами и базы данных .....	515
21.2.1. Современное состояние.....	515
21.2.2. Недостающие ресурсы .....	517
21.3. Ограничения обнаружения манипуляций с лицами .....	518
21.3.1. Обобщаемость .....	518
21.3.2. Интерпретируемость.....	519
21.3.3. Слабые места детекторов .....	520
21.3.4. Возможности человека .....	520
21.3.5. Дальнейшие ограничения .....	521

21.4. Манипуляции с лицами и их обнаружение: путь вперед .....	521
21.4.1. Области применения манипуляций с лицами .....	521
21.4.2. Перспективные методы .....	524
21.5. Социальные и правовые аспекты манипуляции лицами и их обнаружения.....	525
21.6. Выводы.....	528
Литература .....	529
<b>Предметный указатель.....</b>	<b>534</b>

## **Главный редактор**

*Самир Сингх*

## **Редактор серии**

*Синг Бинг Кан, Zillow, Inc., Сиэтл, Вашингтон, США*

## **Редакторы-консультанты**

*Хорст Бишоф, Технологический университет Граца, Грац, Австрия*

*Ричард Боуден, Университет Суррея, Гилфорд, Суррей, Великобритания*

*Свен Дикинсон, Университет Торонто, Торонто, Онтарио, Канада*

*Джиая Цзя, Китайский университет Гонконга, Шатин, Новые территории, Гонконг*

*Кён Му Ли, Сеульский национальный университет, Сеул, Корея (Республика)*

*Жучен Лин, Пекинский университет, Пекин, Китай*

*Ёити Сато, Токийский университет, Токио, Япония*

*Бернт Шиле, Институт информатики им. Макса Планка, Саарбрюккен, Саар, Германия*

*Стэн Скларофф, Бостонский университет, Бостон, Массачусетс, США*

Больше информации об этой серии на <https://link.springer.com/bookseries/4205>.

# 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](http://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](mailto: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/](http://dmkpress.com/authors/publish_book/) или напишите в издательство по адресу [dmkpress@gmail.com](mailto: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](mailto: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](mailto: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), влиянию 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<sup>1</sup> (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

---

<sup>1</sup> 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  
Мадрид, Испания  
Мадрид, Испания  
Йовик, Норвегия

Кристиан Ратгеб  
Рубен Толосана  
Рубен Вера-Родригес  
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-сопоставительные сети (GAN) [7, 8]. В результате были выпущены открытое программное обеспечение и мобильные приложения, такие как ZAO<sup>1</sup> и FaceApp<sup>2</sup>, 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 «deep-fakes» 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-

<sup>1</sup> <https://apps.apple.com/cn/app/id1465199127>.

<sup>2</sup> <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)<sup>1</sup>, запущенный Национальным институтом стандартов и технологий (NIST), Deepfake Detection Challenge (DFDC)<sup>2</sup>, запущенный Facebook, и недавний DeeperForensics Challenge<sup>3</sup>.

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 цветковых 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.

<sup>1</sup> <https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018>.

<sup>2</sup> <https://www.kaggle.com/c/deepfake-detection-challenge>.

<sup>3</sup> <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<sup>1</sup>: 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>

<sup>1</sup> 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<sup>1</sup>. 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<sup>2</sup>.

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<sup>3</sup>.

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

<sup>1</sup> [https://github.com/tkarras/progressive\\_growing\\_of\\_gans](https://github.com/tkarras/progressive_growing_of_gans).

<sup>2</sup> <https://github.com/NVLabs/stylegan>.

<sup>3</sup> <https://github.com/NVLabs/stylegan2-ada-pytorch>.

Конец ознакомительного фрагмента.

Приобрести книгу можно

в интернет-магазине

«Электронный универс»

[e-Univers.ru](http://e-Univers.ru)