Академия криптографии
Российской Федерации

ИСП РАН

Водяные знаки как инструмент борьбы с дипфейками

Маркин Юрий Витальевич,
к.т.н., научный сотрудник ИСП РАН

Якушев Алексей Юрьевич,
исследователь ИСП РАН

Москва, 27 мая 2024 г.

Дипфейки: предметная область и подходы к обнаружению

- Дипфейк – поддельный медиа-контент, полученный с помощью методов глубокого обучения с нуля или путем изменения существующего контента с целью фальсификации его содержания
- Методы обнаружения:
  - артефакты на уровне пикселей
  - физиологические сигналы (моргание глаз, наполнение мелких сосудов лица кровью в зависимости от фазы кардиоцикла)
  - несоответствие видео и аудио (липсинк)
- Появляются новые методы генерации, и затем методы детектирования «учатся» находить то, что на данный момент синтезируется недостаточно качественно
- Обеспечение целостности оригинальных изображений и видео
  - Электронная цифровая подпись (ЭЦП)
  - «Полухрупкий» цифровой водяной знак (ЦВЗ)
    - устойчив к допустимым преобразованиям
    - неустойчив, когда выполняются недопустимые преобразования

2014

2017

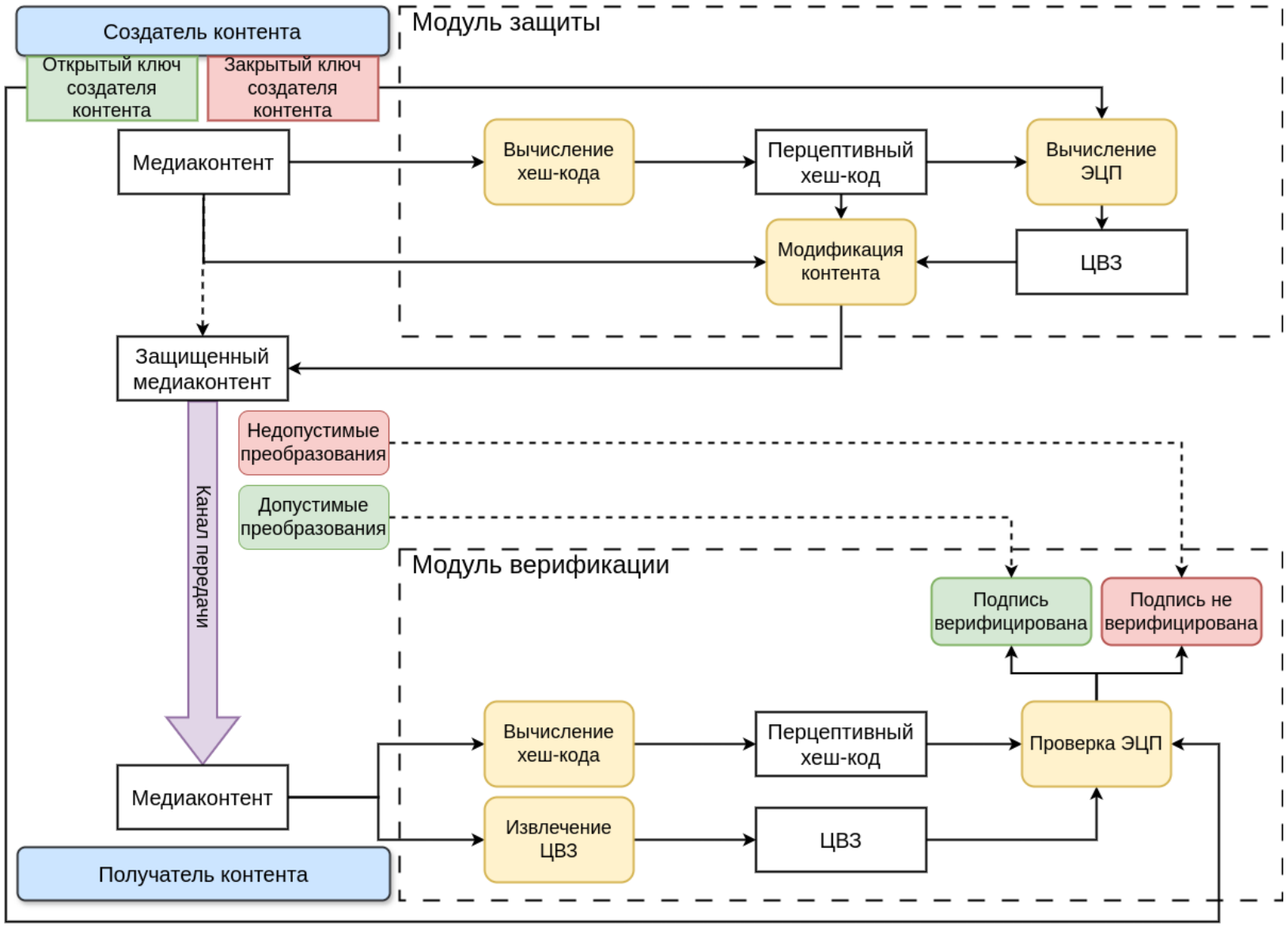
2019

Masood M. et al. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward – 2023

Допустимые и недопустимые преобразования

- Любой дипфейк получен путем модификации медиаконтента, но не всякая модификация медиаконтента есть дипфейк
- **Допустимые** преобразования
  - изображений:
    - изменение формата хранения изображения с применением алгоритмов сжатия (в т.ч. с потерями)
    - незначительные попиксельные модификации изображения: фильтрация, наложение шумов, изменение яркости и контрастности
    - изменение разрешения с сохранением соотношения сторон
  - видео:
    - допустимые преобразования изображений, примененные к отдельным или ко всем кадрам видео
    - изменение формата хранения видео, транскодирование другим кодеком с другим уровнем качества
    - изменение частоты кадров при транскодировании
- Остальные преобразования – **недопустимы** (считаются злонамеренными, и их необходимо детектировать), в том числе:
  - замена части изображения на другое (в частности, наложения дипфейк-изображения)
  - изменение соотношения сторон изображения путем растяжения или сжатия одной из сторон
  - обрезка изображения
  - поворот или зеркальное отражение изображения
  - удаление или вставка части кадров видео
  - ускорение или замедление видео

Схема алгоритма обеспечения целостности медиаконтента



Перцептивные хеш-функции для изображений

- Требования:
  - совпадение результатов вычисления до и после допустимых преобразований;
  - различные значения – до и после недопустимых преобразований;
- Устойчивость к допустимым преобразованиям:
  - рассмотрено 9 ранее предложенных перцептивных хеш-функций;
  - проведена оценка значений хеш-функций до и после применения сжатия JPEG (качество 50);
  - лучший результат – совпадение значений для 97.3% изображений;
- Без дополнительной модификации перцептивные хеш-функции неприменимы в предлагаемой схеме:
  - наличие этапа квантизации с получением последовательности бит;
  - при квантизации значений, близких к пороговым, возможна инверсия бита даже при малом изменении значения в результате допустимого преобразования

Модификация перцептивной хеш-функции для изображения

- Основа – дискретное косинусное преобразование (ДКП):
  - заданное число низкочастотных коэффициентов;
  - квантование разности полученных коэффициентов с медианным значением;
- Оригинальное изображение дополнительно модифицируется:
  - изменение значений коэффициентов ДКП, близких по модулю к медианному значению, в сторону увеличения модуля разности;
  - величина изменения – в соответствии с целевым значением метрики PSNR (схожести модифицированного изображения относительно оригинала);
- Защищается модифицированное изображение, а не оригинальное:
  - по значению перцептивной хеш-функции формируется ЭЦП отправителя;
  - ЭЦП внедряется в защищаемое изображение как ЦВЗ в виде QR-кода

Novozamsky A., Mahdian B., Saic S. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images – 2020

Тестирование разработанного прототипа

1. Доля изображений, на которых перцептивный хеш-код изменился после сжатия JPEG в зависимости от коэффициента качества (1000 изображений из набора Open Images)

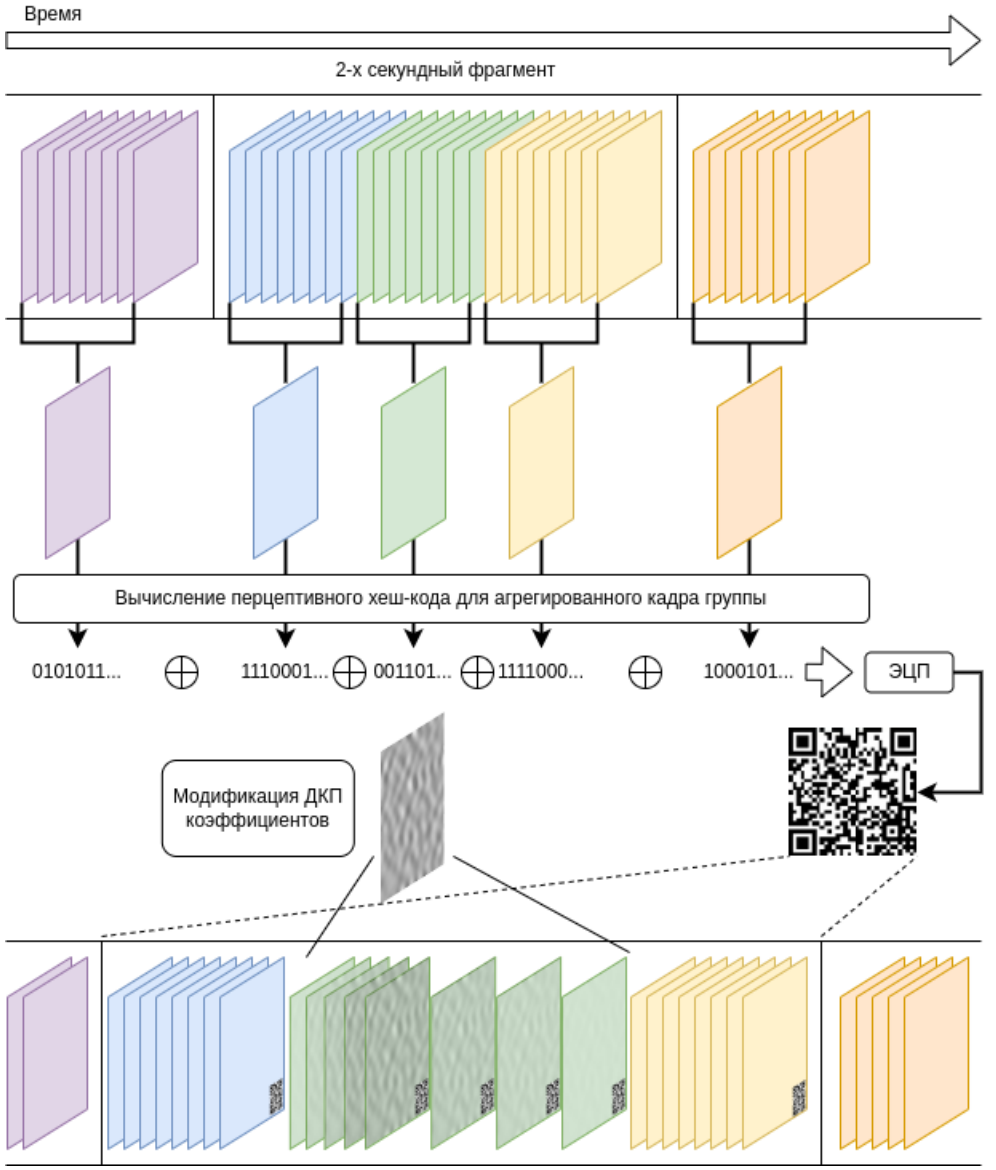
Качество JPEG	10	20	30	40	50	60
Доля изображений	20.9%	2.7%	1.3%	1%	0.9%	0.8%

2. Перцептивный хеш-код изменился после недопустимой подделки на 97.2% изображений (набор пар оригинальное и поддельное изображение IMD2020)

Novozamsky A., Mahdian B., Saic S. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images – 2020

Адаптация алгоритма для защиты видеоконтента

- Пространственная целостность (кадров) видео
  - алгоритм поддержания целостности изображений
- Плавность перехода между кадрами
  - перцептивный хеш-код агрегированного кадра в рамках группы кадров
  - общая модификация кадров в рамках группы по коэффициентам ДКП агрегированного кадра
  - уменьшение степени модификации на границах группы
- Временная целостность видео
  - общий QR-код для фрагмента видео, содержащего несколько целых групп кадров
  - вычисление ЭЦП по объединенным перцептивным хеш-кодам групп кадров фрагмента, а также двух соседних групп кадров

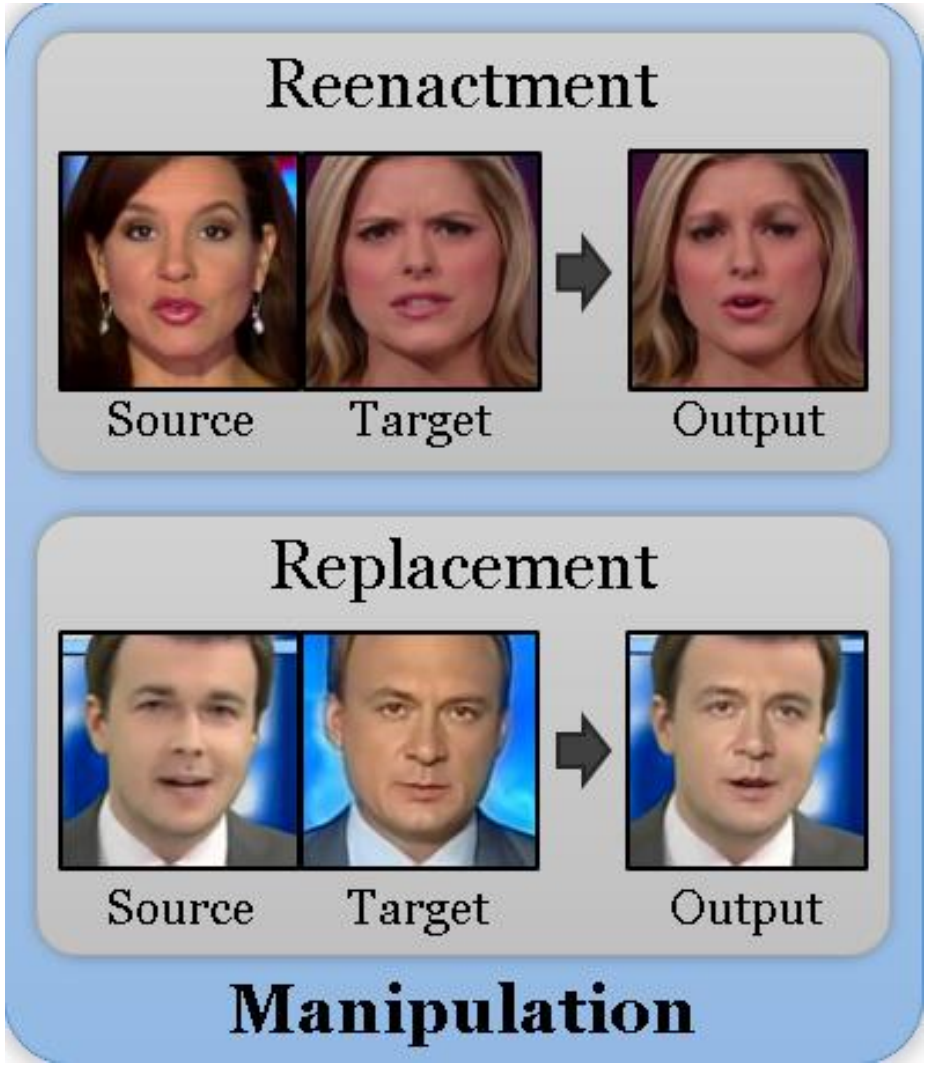


Тестирование разработанного прототипа

1. **Сохранение** перцептивного хеш-кода при допустимых преобразованиях:
   - Набор видео DAVIS
   - Транскодирование кодеком h.264 с разными значениями CRF (Constant Rate Factor)
   - Двухсекундные видеофрагменты оценивались независимо
   - Определена доля видеофрагментов, для которых сжатие привело к изменению значения перцептивного хеш-кода

CRF	20	25	30	35	40
Доля фрагментов	0%	0%	0.2%	7.3%	66.2%

2. **Изменение** перцептивного хеш-кода при недопустимых преобразованиях:
   - Набор видео FaceForensics++, содержащий пары оригинальное видео и видео, созданное с применением технологии DeepFake
   - Доля групп кадров, для которых значение перцептивного хеш-кода изменилось, составила 99.6%



Rossler A. et al. FaceForensics++: Learning to detect manipulated facial images – 2019

Заключение

Результаты

- Разработана методика обеспечения целостности оригинального медиаконтента
- Реализованы алгоритмы обеспечения целостности изображений и видео
- Проведено тестирование прототипов алгоритмов, показавшее применимость предложенных методов

Дальнейшие исследования

- Разработка перцептивных хеш-функций, основанных на других технологиях обработки изображений
- Использование незаметных ЦВЗ для внедрения ЭЦП в медиаконтент
- Применение метода к аудио контенту, в частности, к звуковой дорожке видео

Благодарю за внимание