Маркировка данных. Борьба с дипфейками.

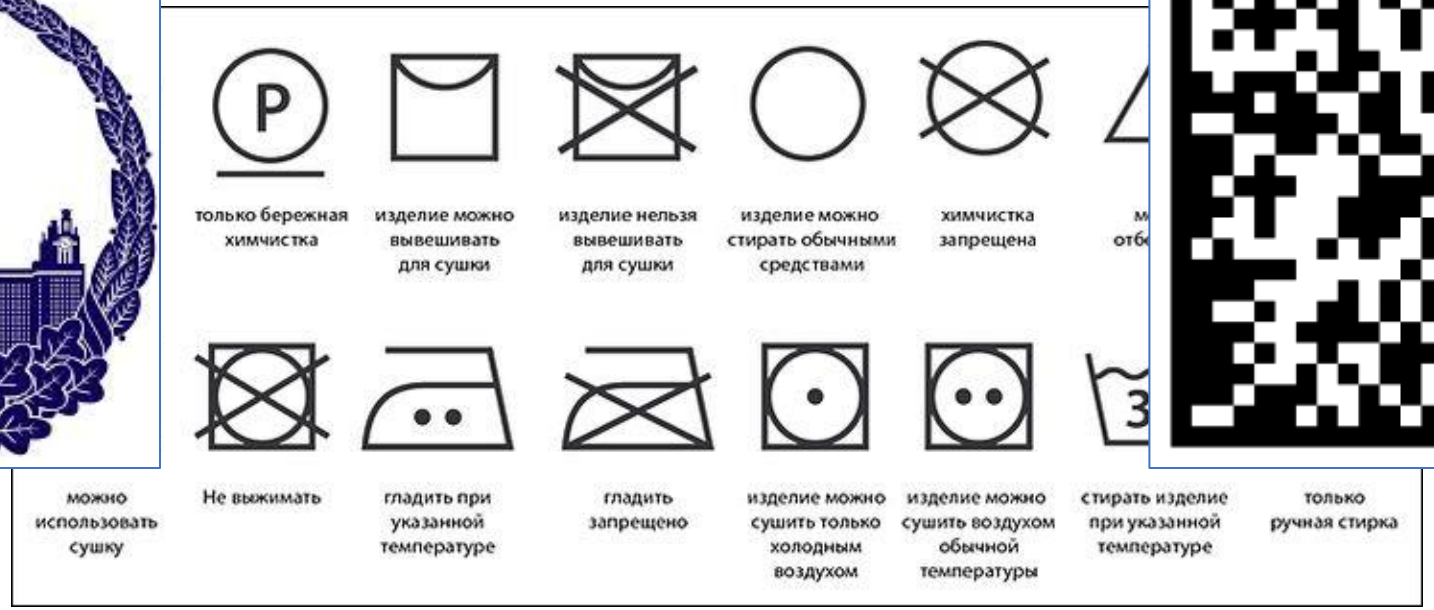
Маркин Юрий Витальевич,
к.т.н., научный сотрудник ИСП РАН

Сеул, 19 ноября 2025 г.

Что такое маркировка

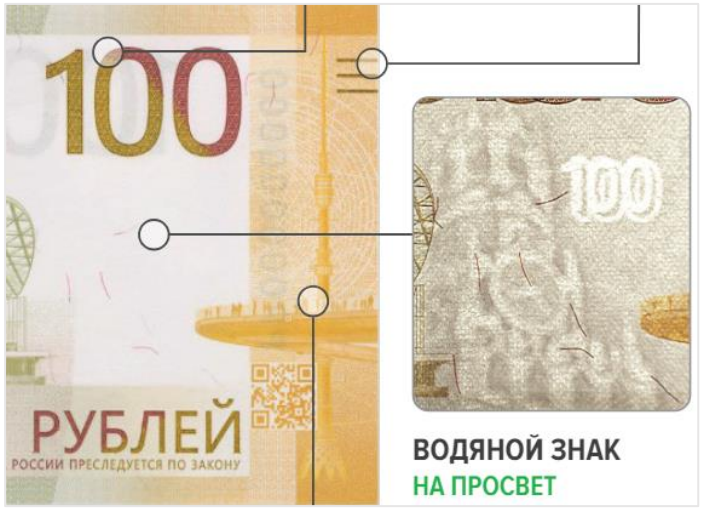
- Маркировка – это нанесение условных знаков, букв, цифр, графических знаков или надписей на объект (by Wikipedia)
- Цель – идентификация (узнавание) объекта, указание его свойств и характеристик
- Примеры:
  - логотипы брендов
  - https://честныйзнак.рф

ИСП РАН



Что такое водяной знак

- Первое упоминание – Италия, 13 век
- Видимое изображение или рисунок на бумаге, который выглядит светлее или темнее при просмотре на просвет:
  - **Создается** путем вдавливания металлического сетчатого валика в бумагу в процессе изготовления
  - **Является** традиционным способом защиты ценных бумаг и документов от подделки
  - **Применяется** для подтверждения подлинности / затруднения создания подделок
- Примеры – на банкнотах, как древних, так и современных



Маркировка vs. Водяной знак

	Маркировка	Водяной знак
Назначение	Идентификация, информирование, узнавание	Защита, аутентификация, доказательство прав, проверка контроля целостности
Видимость	Как правило, открыта и легко читаема	Может быть заметным и скрытым (в зависимости от решаемой задачи)
Кто использует	Потребители продуктов	Владелец объекта, правообладатель, проверяющие органы
Удаление	Удаляется легко	Удалить сложно, часто приводит к порче объекта

Цифровые водяные знаки (ЦВЗ) для медиаконтента

Заметные водяные знаки

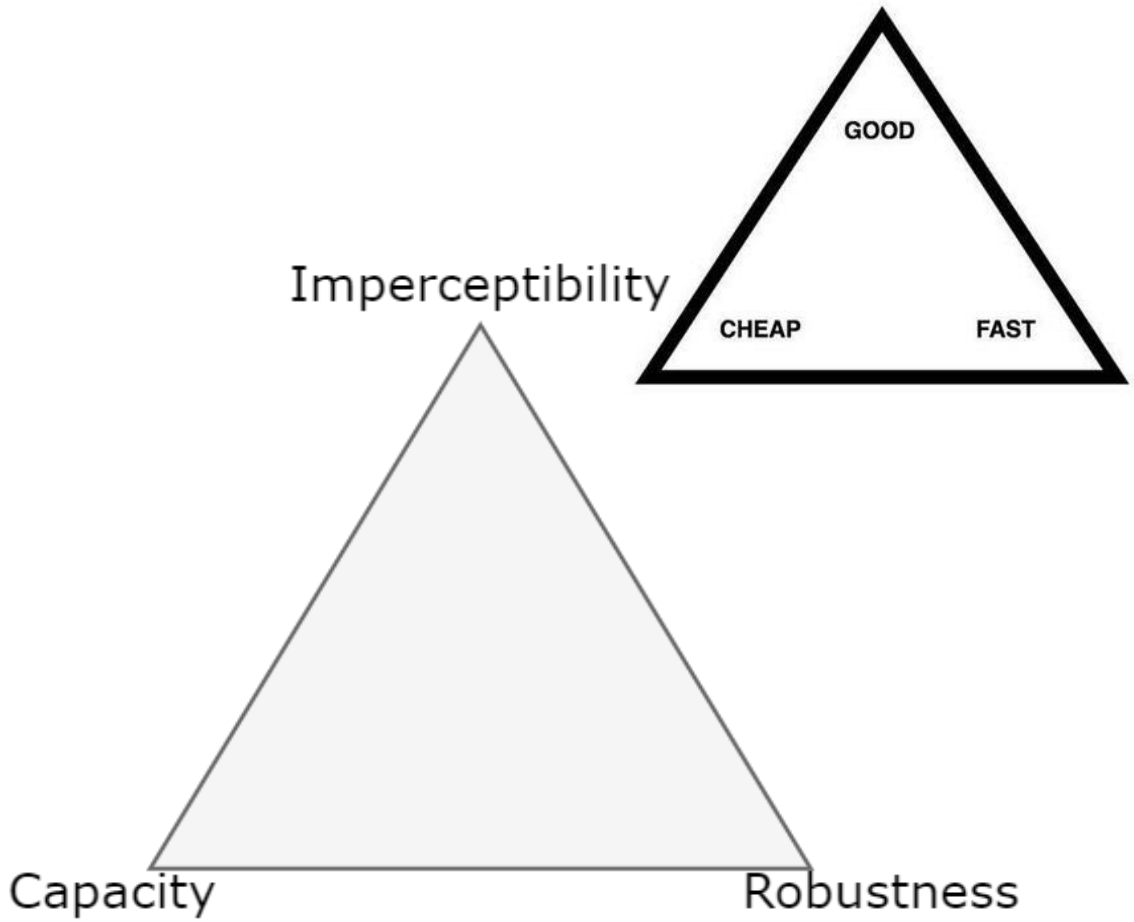
Скрытые водяные знаки

Задачи, решаемые с помощью ЦВЗ

- Идентификация владельца / подтверждение права собственности
  - ЦВЗ может использоваться для предоставления информации о владельце или источнике контента
- Отслеживание распространения (a.k.a. Fingerprinting)
  - в каждую копию медиа-файла внедряется уникальное значение ЦВЗ;
  - разным пользователям предоставляются разные копии (например, каждому пользователю – своя копия)
- Контроль целостности содержимого контента / обнаружение намеренных модификаций
  - обнаружение факта модификации – «да/нет»;
  - локализация модификаций (временна́я и/или пространственная)

Основные свойства ЦВЗ

- **Imperceptibility** – незаметность ЦВЗ
- **Capacity** – ёмкость ЦВЗ, определяет объем содержащейся в ЦВЗ информации
- **Robustness** – устойчивость, способность ЦВЗ быть извлекаемым после проведения преобразований над объектом внедрения
  - ЦВЗ может быть устойчив к одним преобразованиям, и неустойчив к другим



Основные типы ЦВЗ

- **Fragile** – «Хрупкий»
  - должен быть чувствительным ко всем манипуляциям над содержимым медиафайла
  - низкая заметность и большая емкость
  - подтверждение подлинности
- **Robust** – «Устойчивый»:
  - должен противостоять наиболее распространенным операциям обработки медиаконтента (в частности, транскодированию)
  - высокая заметность, низкая ёмкость
  - защита авторских прав
- **Semi-Fragile**:
  - нечувствителен к допустимым модификациями над медиаконтентом
  - чувствителен к злонамеренным атакам
  - обнаружение несанкционированного доступа / модификаций

Внедрение/извлечение ЦВЗ

Внедрение ЦВЗ предполагает модификацию контента:

- вход: оригинальный контент,
- выход: маркированный контент

Извлечение ЦВЗ может выполняться в двух сценариях:

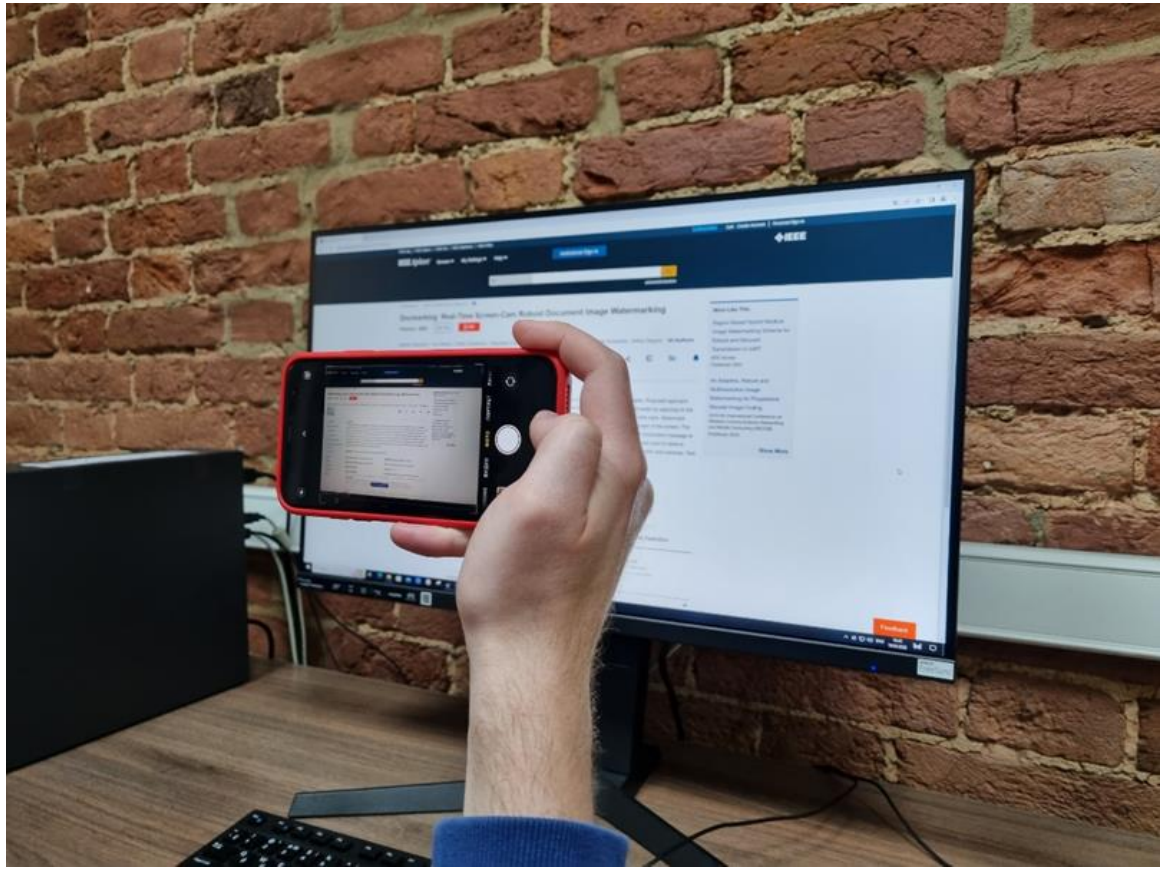
- слепое извлечение (blind):
  - вход: маркированный контент
- неслепое извлечение:
  - вход: маркированный контент, оригинальный контент

Наличие оригинального контента:

- может быть положено в основу алгоритма извлечения
- может повышать точность извлечения ЦВЗ «слепого» алгоритма

Задача №1. Деанонимизация утечек текстовых документов

- Необходимо защитить «аналоговые» каналы утечек, не покрываемые классическими DLP-системами: *screen-cam, print-scan, prin-cam*



Зачем принтеры ставят невидимые точки на документах

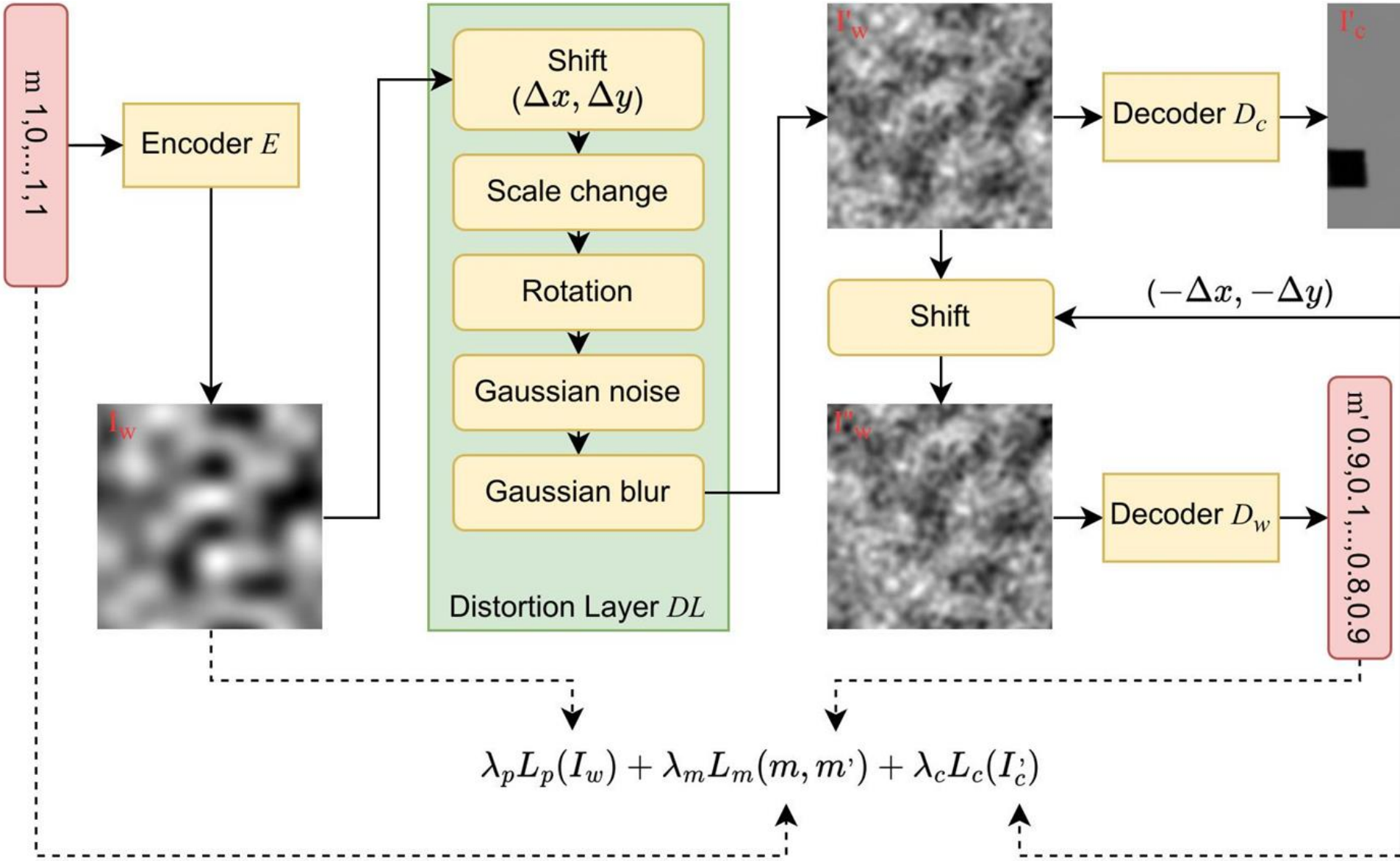
ЦВЗ для документов при выводе на экран

Идентификатор пользователя

Генерация уникального изображения на основе идентификатора пользователя

Альфа-смешивание оригинального изображения с изображением водяного знака

Обучение нейросетевых моделей внедрения и извлечения ЦВЗ



L_p – штрафует за резкие переходы яркости в графическом представлении ЦВЗ

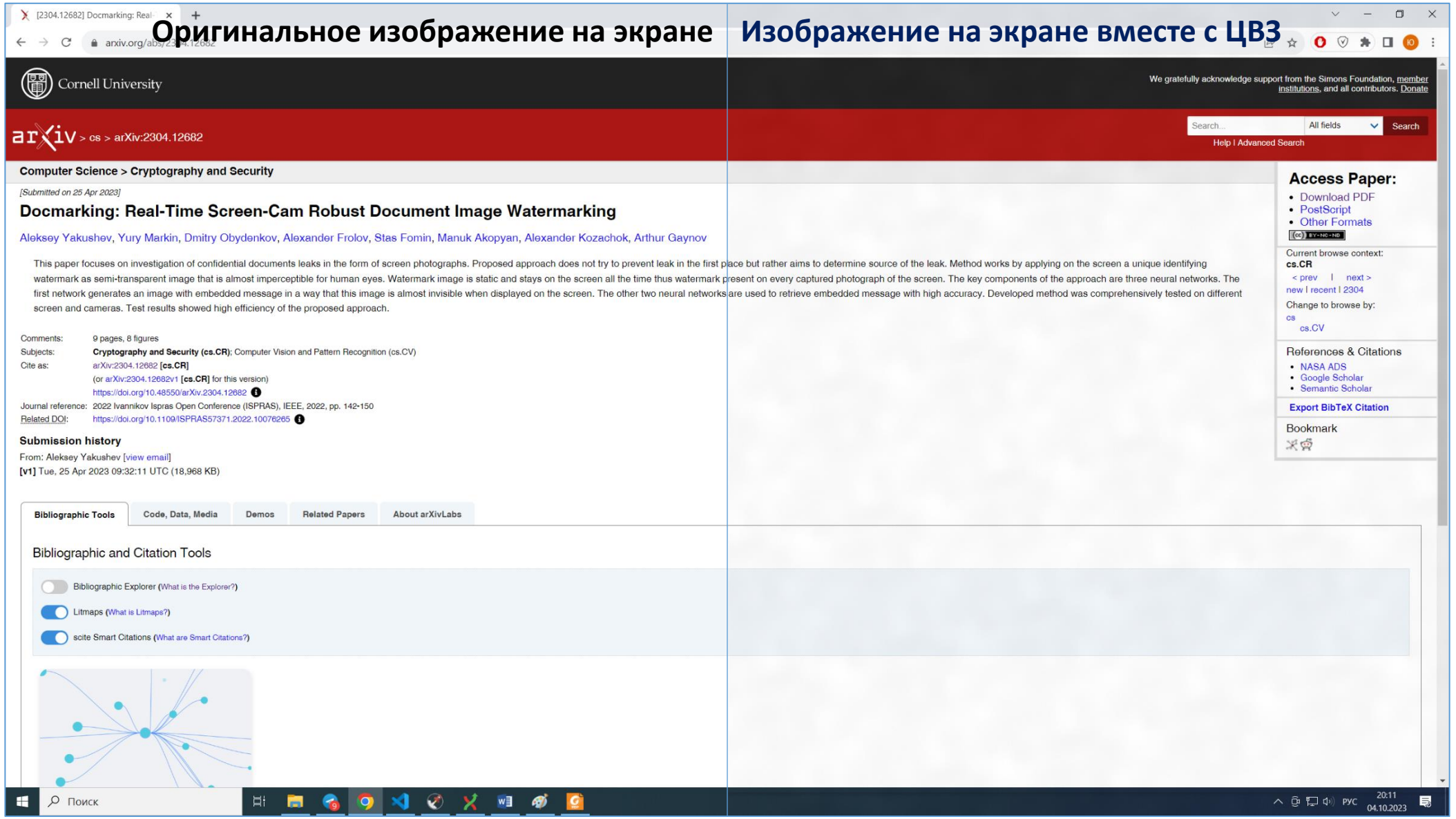
L_m – штрафует за некорректные значения битов извлекаемого ЦВЗ

L_c – штрафует за некорректно определенную величину сдвига

ЦВЗ для документов при выводе на экран

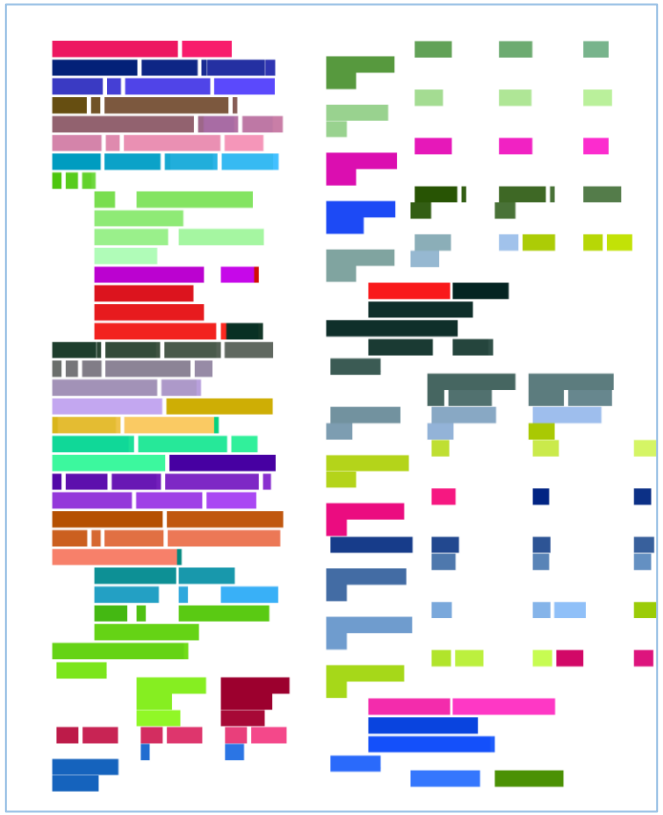
Оригинальное изображение на экране

Изображение на экране вместе с ЦВЗ

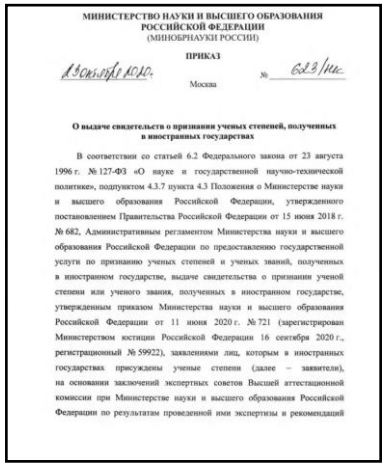


ЦВЗ для документов при выводе на печать

Сегментация боксов слов машинописного текста



Фильтрация рукописного текста/изображений/ печатей



МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
(МИНОБРНАУКИ РОССИИ)

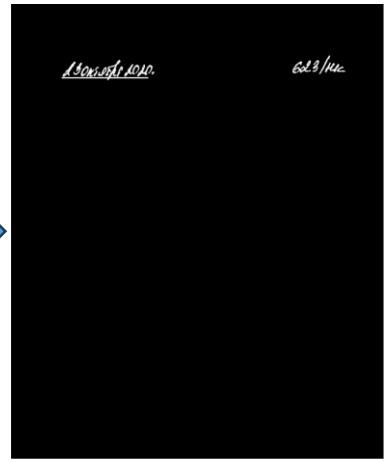
ПРИКАЗ

19 октября 2020 г. № 623/нк

Москва

О выдаче свидетельств о признании ученых степеней, полученных в иностранных государствах

В соответствии со статьей 6.2 Федерального закона от 23 августа 1996 г. № 127-ФЗ «О науке и государственной научно-технической политике», подпунктом 4.3.7 пункта 4.3 Положения о Министерстве науки и высшего образования Российской Федерации, утвержденного постановлением Правительства Российской Федерации от 15 июня 2018 г. № 682, Административным регламентом Министерства науки и высшего образования Российской Федерации по предоставлению государственной услуги по признанию ученых степеней и ученых званий, полученных в иностранном государстве, выдаче свидетельства о признании ученой степени или ученого звания, полученных в иностранном государстве, утвержденным приказом Министерства науки и высшего образования Российской Федерации от 11 июня 2020 г. № 721 (зарегистрирован Министерством юстиции Российской Федерации 16 сентября 2020 г., регистрационный № 59922), заявлениями лиц, которым в иностранных государствах присуждены ученые степени (далее – заявители), на основании заключений экспертных советов Высшей аттестационной комиссии при Министерстве науки и высшего образования Российской Федерации по результатам проведенной ими экспертизы и рекомендаций



Горизонтальный сдвиг боксов слов машинописного текста

передаваемых между сетевыми
передаваемых между сетевыми

Задача №2. Деанонимизация онлайн-видео-ретрансляций

В рамках борьбы с пиратскими видеотрансляциями необходимо идентифицировать источник видеотрансляции, т.е. пользователя, ведущего незаконную ретрансляцию

Особенность онлайн-видеотрансляции – адаптивное изменение качества в зависимости от текущей загруженности сети и параметров сетевого соединения пользователя – Adaptive Bit Rate (ABR)

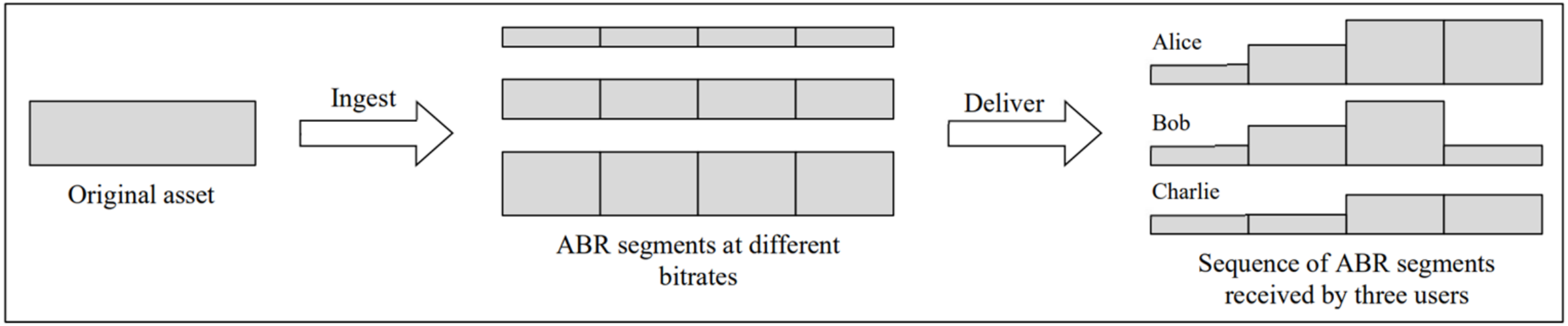
Предложенное решение

- внедрение ЦВЗ в видео методом A/B-маркирования

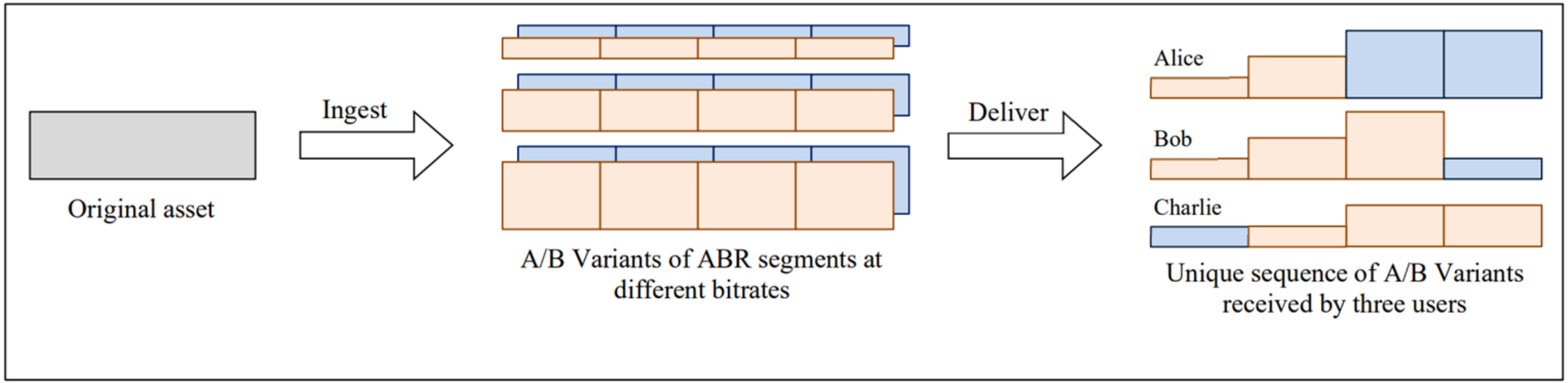
Идея A/B-метода внедрения ЦВЗ

- Видео делится на фрагменты равной продолжительности, последовательно транслируемые пользователям (обычно, это 2-секундные сегменты);
- Каждый фрагмент кодируется для нескольких битрейтов;
- В каждый момент времени конечному пользователю отправляется сегмент наиболее подходящего качества (в соответствии с параметрами и текущими состоянием сетевого соединения)
  - с точки зрения качества, последовательность сегментов для разных пользователей может и, как правило, будет не совпадать;
- Для каждой пары <*фрагмент, битрейт*> создается две версии:
  - A – кодирует 1,
  - B – кодирует 0
- Каждому пользователю ставится в соответствие идентификатор – битовая последовательность заданной длины;
- В данный момент времени пользователю отправляется фрагмент в подходящем битрейте, кодирующий очередной бит его идентификатора

Схема A/B-метода внедрения ЦВЗ



(a) ABR content delivery



(b) A/B Variants for ABR content delivery

Требования и подход к решению Задачи №2

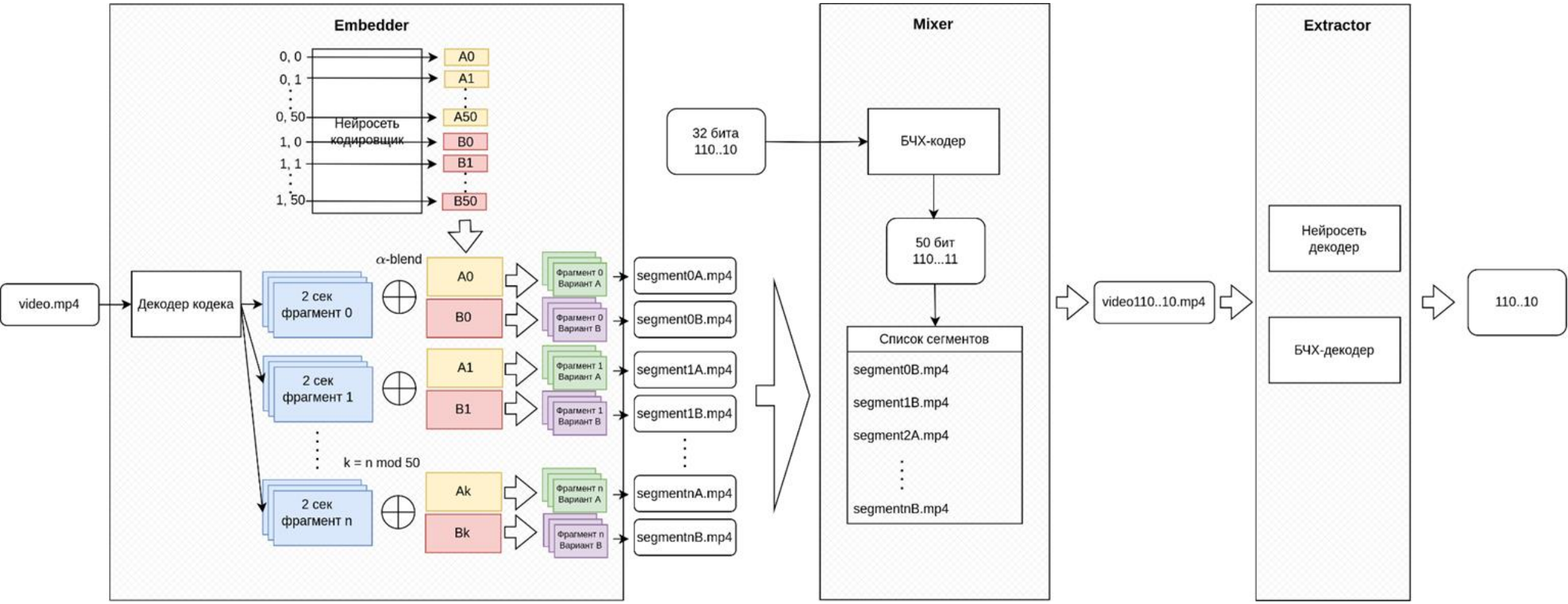
Требования к методу

- Поддержка A/B сценария
- Работа в реальном времени
- Визуальная незаметность ЦВЗ
- Устойчивость к транскодированию видео (в том числе с ухудшением качества)
- Возможность извлечения ЦВЗ из части видео (до нескольких минут) с началом в произвольный момент

Принцип работы

- Графическое представление ЦВЗ одинаковое для всех кадров в 2-секундном фрагменте;
- Изображение ЦВЗ формируется нейросетью и накладывается на кадры посредством α-смешивания
- Нейросети внедрения и извлечения ЦВЗ обучаются совместно
- Устойчивость ЦВЗ достигается применением аугментаций при обучении
- Емкость ЦВЗ 50 бит (идентификатор 32 бита + 18 бит код коррекции БЧХ)
  - для извлечения потребуется не менее 100 секунд видео

Адаптация решения для видеофайла



Функциональный модули системы внедрения и извлечения ЦВЗ

- **Embedder**
  - получает на вход видеофайл,
  - делит его на сегменты продолжительностью по 2 секунды каждый,
  - создает два варианта A/B каждого сегмента, применяя альфа-смешивание кадров видео с изображением ЦВЗ;
- **Mixer**
  - получает на вход 32-битный идентификатор пользователя,
  - добавляет к нему 18 бит кодов коррекции БЧХ,
  - формирует список сегментов, соответствующий полученному 50 битному сообщению,
  - объединяет сегменты из списка в единое видео;
- **Extractor**
  - получает на вход маркированное видео / фрагмент маркированного видео / фрагмент маркированного видео после преобразования,
  - применяет нейронную сеть декодер к кадрам полученного видео,
  - объединяет результаты в 50 бит сообщения,
  - использует коды коррекции БЧХ для получения 32-битного идентификатора

Задача №3. Определение факта генерации контента

Технология синтеза изображений создают новые **угрозы**:

- Дезинформация и манипуляция
- Нарушение прав личности
- Нарушение авторских прав
- Автоматизация мошенничествах действий
- Девальвация творческого труда

Способ противодействия: внедрение ЦВЗ при синтезе цифрового контента

➔ Контент создан ИИ или человеком?

Законодательные акты:

- США (2024) – генеративные модели **должны** внедрять метки / ЦВЗ
- ЕС (2024) – **требуется** указание, что контент создан ИИ
- G7 (2024) – добровольная (пока) маркировка сгенерированного контента

Невидимые водяные знаки:

- Gemini (Google)
- Stable Diffusion

Метки на основе метаданных:

- DALL·E (OpenAI)
- Emu (Meta*)
- Photoshop (Adobe)

Подходы к внедрению ЦВЗ в синтезируемые изображения

Внедрение ЦВЗ в синтезируемый ИИ контент:

1. **Постобработка** – водяной знак внедряется в уже готовое изображение
   - Не требует модификации архитектуры или обучения генеративной модели

Подходы:

- Частотные методы: DCT / DFT / DWT
- Нейросетевые методы внедрения

2. **Интеграция при синтезе** – внедрение ЦВЗ происходит на этапе генерации изображения

Пример: в диффузионных моделях ЦВЗ может закладываться в скрытое пространство и переноситься на изображение в ходе диффузии

Оценка методов внедрения ЦВЗ в синтезированные изображения

Задача: оценка существующих методов внедрения / извлечения ЦВЗ по критериям устойчивости и незаметности

Критерии оценки методов внедрения и извлечения ЦВЗ из изображений:

- **Емкость** передаваемого сообщения
- **Незаметность** внедренного водяного знака
- **Устойчивость** к атакам (обрезка, сжатие, фильтры, искажения и прочие)

Опубликованные ранее бенчмарки:

- *Petitcolas и др. (1999), Petitcolas (2000), Tao и др. (2014)* – не охватывают современные нейросетевые методы
- *WAVES (ICLR 2024)* – фокусируется на устойчивости, не учитывает незаметность

Разработана система оценки методов внедрения ЦВЗ в синтезированные изображения **WIBE***

* github.com/ispras/wibe, ASE 2025

Оценка незаметности ЦВЗ

Для оценки незаметности методов внедрения ЦВЗ при постобработке используются метрики сравнения схожести двух изображений:

- **PSNR** (*Peak Signal-to-Noise Ratio*)

$$PSNR(I, I_w) = 10\ log_{10}\frac{MAX_I^2}{MSE(I, I_w)}, MAX_I = 255$$

- **SSIM** (*Structural Similarity*)

$$SSIM(a, b) = \frac{(2\mu_a\mu_b + c_1)(2\sigma_{ab} + c_2)}{(\mu_a^2 + \mu_b^2 + c_1)(\sigma_a^2 + \sigma_b^2 + c_2)}$$

- **LPIPS** (*Learned Perceptual Image Patch Similarity*) – метрика перцептивного различие между двумя изображениями на основе признаков, извлеченных из сверточной нейросети (VGG, AlexNet и другие)

Оценка незаметности ЦВЗ

PSNR: 39.53

SSIM: 0.9624

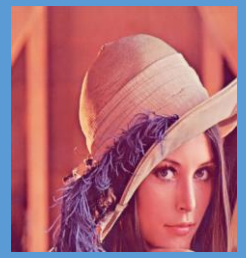
Атаки на ЦВЗ в изображениях, в том числе, синтезированных

Оригинал

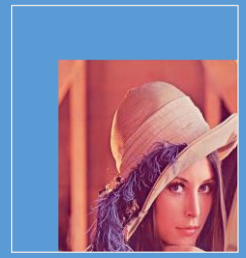
- Зеркалирование
- Искажение перспективы
- Бочкообразная дисторсия
- Сжатие JPEG

Геометрические

Поворот

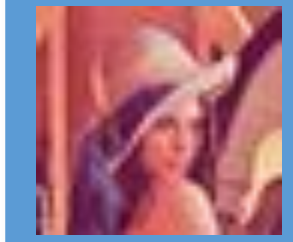


Масштабирование

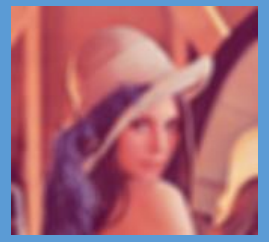


Сдвиг и обрезка

Попиксельные



Сжатие



Размытие

Контраст

Регенерация

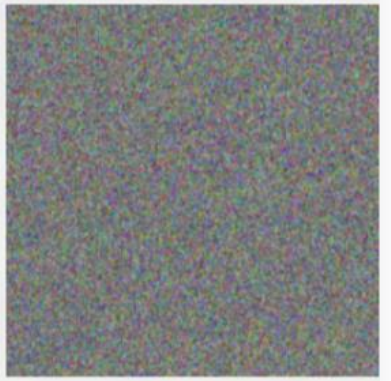
Частотный метод внедрения / извлечения ЦВЗ

Основные этапы работы метода **dwtDct**:

- *DWT* — разделение изображения на частотные поддиапазоны (*LL, LH, HL, HH*), приближено к восприятию человеком. Выбор поддиапазона для внедрения:
  - Низкочастотный поддиапазон *LL* не используется для внедрения — визуально заметно
  - Используются высокочастотные поддиапазоны *HL / HH* – менее заметны для человеческого глаза, но достаточно устойчивы искажениям
- Блочное *DCT* преобразование поддиапазона и кодирование информации
  - Выбор DCT-коэффициента определяет устойчивость ЦВЗ
  - Кодирование посредством квантования коэффициентов QIM или другим способом

Оригинал

ЦВЗ



Изображение с ЦВЗ

Нейросетевой метод внедрения / извлечения ЦВЗ

Метод **StegaStamp**

Внедрение ЦВЗ:

- Генерация водяного знака – битовая строка преобразуется нейросетью в скрытое представление
- Внедрение нейросетью **encoder** водяного знака в изображение

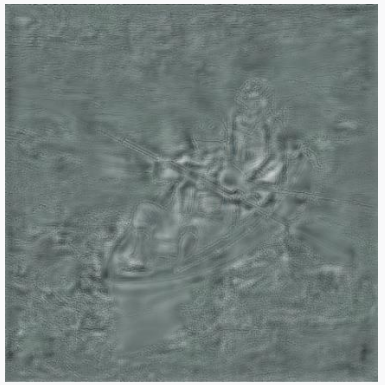
Извлечение ЦВЗ:

- Извлечение битовой строки нейросетью **decoder** из искаженного изображения

Обучение: эмуляция искажений между encoder и decoder, симулирующих реальные искажения

Оригинал

ЦВЗ



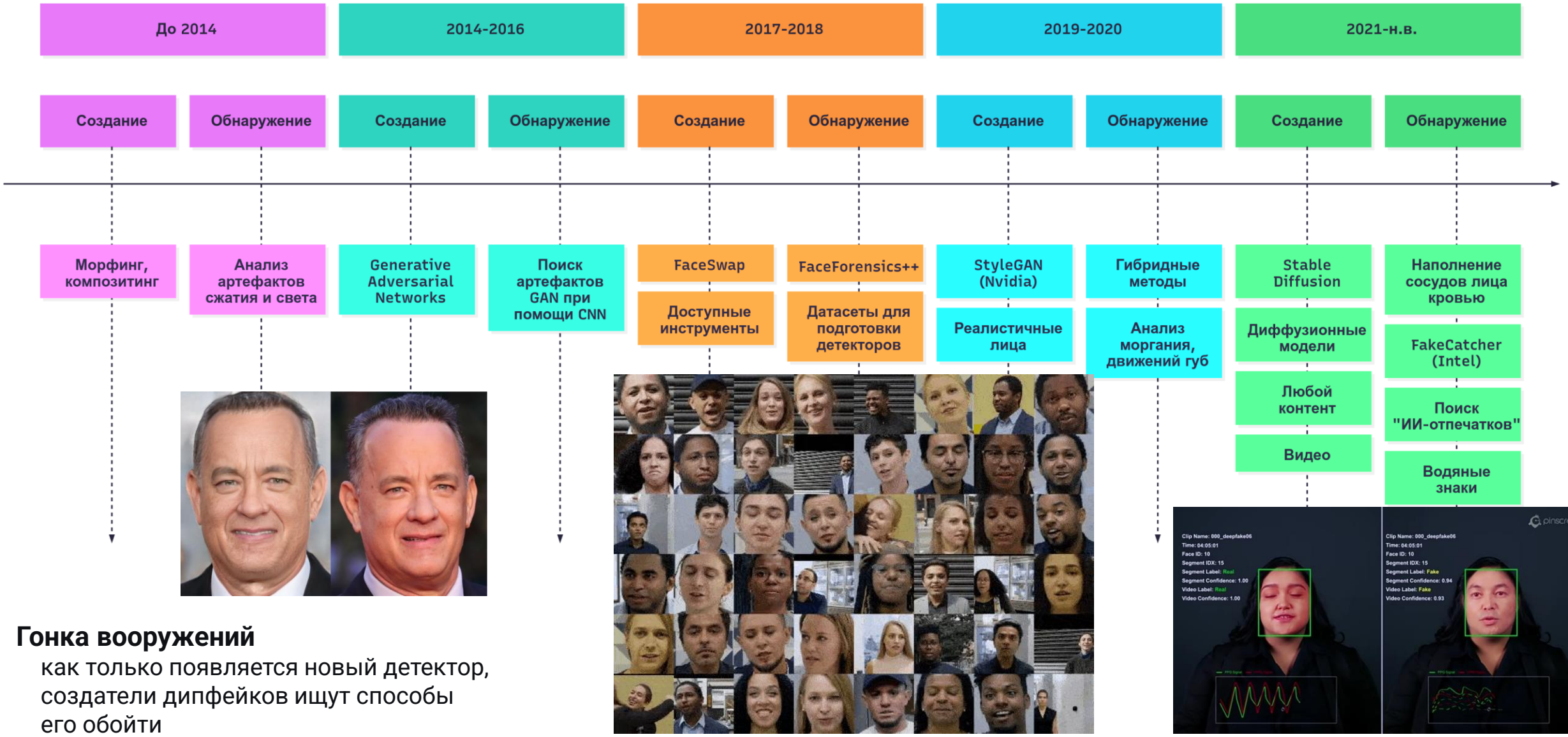
Изображение с ЦВЗ

Алгоритм	Емкость	PSNR	SSIM	LPIPS	Gauss Noise 22	JPEG 50	Cropout 50	Center Crop 50	Rotate 30
ARWGAN	30	38.25	0.983	0.014	0.98	0.84	1.00	0.72	0.00
CIN	30	43.27	0.988	0.017	1.00	0.97	1.00	0.00	0.00
DCT	100	42.25	0.975	0.031	1.00	1.00	1.00	0.00	0.00
DWSF	30	41.20	0.993	0.019	1.00	0.93	0.93	0.71	0.98
DwtDct	100	37.96	0.965	0.023	0.77	0.00	0.31	0.00	0.00
DwtDctSvd	100	37.98	0.979	0.014	1.00	0.96	0.84	0.00	0.00
hidden	30	36.83	0.977	0.016	0.39	0.00	0.41	0.00	0.00
Invismark	100	49.03	0.994	0.002	1.00	0.62	1.00	1.00	0.00
MBRS	256	39.67	0.979	0.017	1.00	1.00	1.00	0.00	0.00
NSS	100	39.60	0.982	0.013	1.00	1.00	0.93	0.00	0.00
RivaGan	32	40.54	0.976	0.036	0.95	0.67	0.91	0.92	0.00
Rosteals	100	30.28	0.942	0.042	1.00	1.00	0.92	0.00	0.00
Sepmark	128	37.67	0.977	0.015	1.00	1.00	0.99	0.00	0.00
SSHidden	48	37.60	0.958	0.029	0.94	0.30	0.84	0.01	0.00
StegaStamp	100	29.07	0.921	0.051	1.00	1.00	1.00	0.00	0.00
Trustmark	100	40.24	0.988	0.002	1.00	1.00	1.00	0.45	0.00
Vine	100	38.42	0.991	0.005	1.00	1.00	0.00	0.00	0.00
WmAnything	32	41.38	0.988	0.017	1.00	0.86	0.99	0.83	0.00

Задача №4. Обеспечение контроля целостности медиа

- Управляющий директор британской энергетической компании был ограблен на €220 тыс. после аудио-звонка (2019)
- Видео, где спикер Палаты представителей США от демократической партии казалась говорящей медленно и протяжно, словно находилась в состоянии опьянения (2019):
  - это привело к волне критики со стороны республиканских политиков и экспертов
- Компания Sensity провела проверку на уязвимость тестов идентификации, предоставляемых 10 поставщиками (2022) – 9 из 10 решений оказались уязвимы к дипфейк-атакам:
  - копирование лица цели на ID-карту для сканирования (модификация изображения),
  - внедрение лица цели в видеопоток с целью пройти liveness-тесты (проверка принадлежности биометрических признаков конкретному человеку),
- Транснациональная компания потеряла $25,6 млн. в результате мошенничества – сотрудник филиала в Гонконге был обманут в ходе видеоконференции (2024)
- Мошенники с помощью видео-дипфейка мэра Москвы Сергея Собянина обокрали трех жителей столицы (2025)

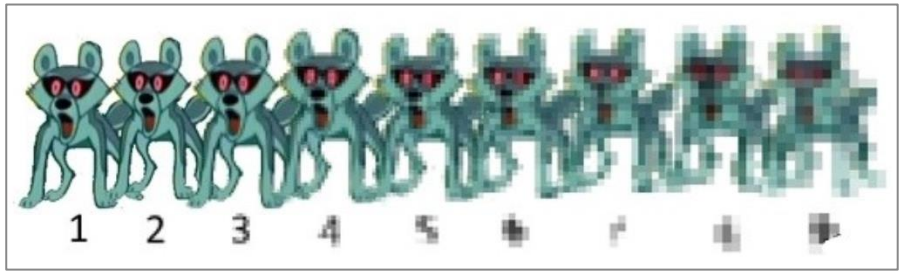
Эволюция методов создания и обнаружения дипфейков



Гонка вооружений

как только появляется новый детектор, создатели дипфейков ищут способы его обойти

Дипфейк ≠ Модификация

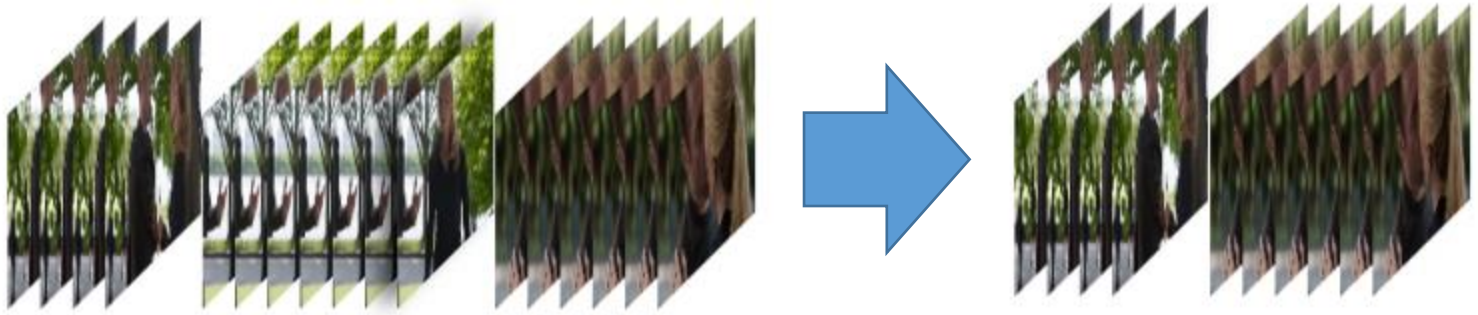


- **Дипфейк** – поддельный медиа-контент, полученный с помощью методов глубокого обучения с нуля или путем изменения существующего контента с целью фальсификации его содержания
- Любой дипфейк получен путем модификации медиаконтента, **не всякая модификация медиаконтента это дипфейк**

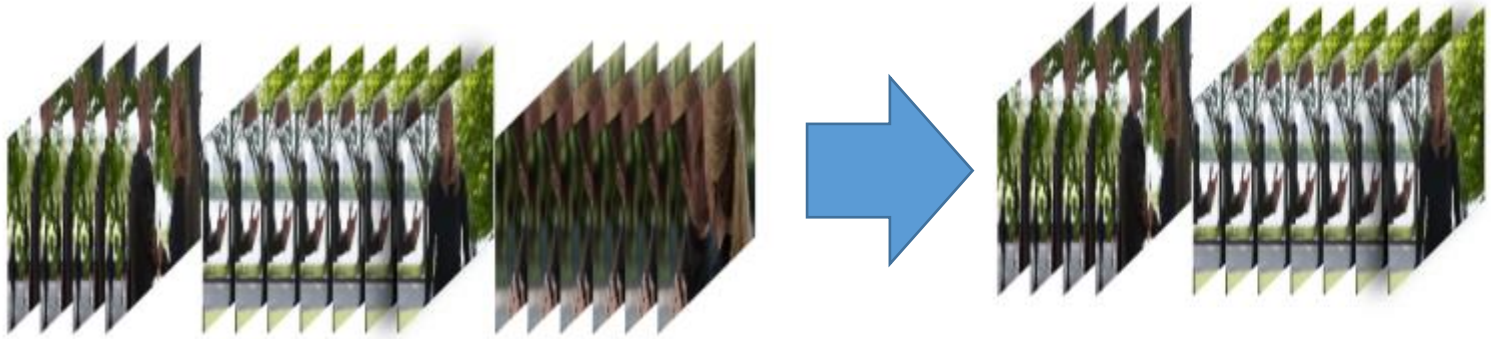
Допустимая модификация	Дипфейк
транскодирование другим кодеком с другим уровнем качества (сжатие)	подмена контента (в частности, наложение другого лица)
изменение частоты кадров при транскодировании видео	удаление или вставка фрагментов аудио / видео
наложение шумов на аудиосигнал	ускорение или замедление аудио / видео

Пространственная и временная целостность медиаконтента

Нарушение пространственной целостности



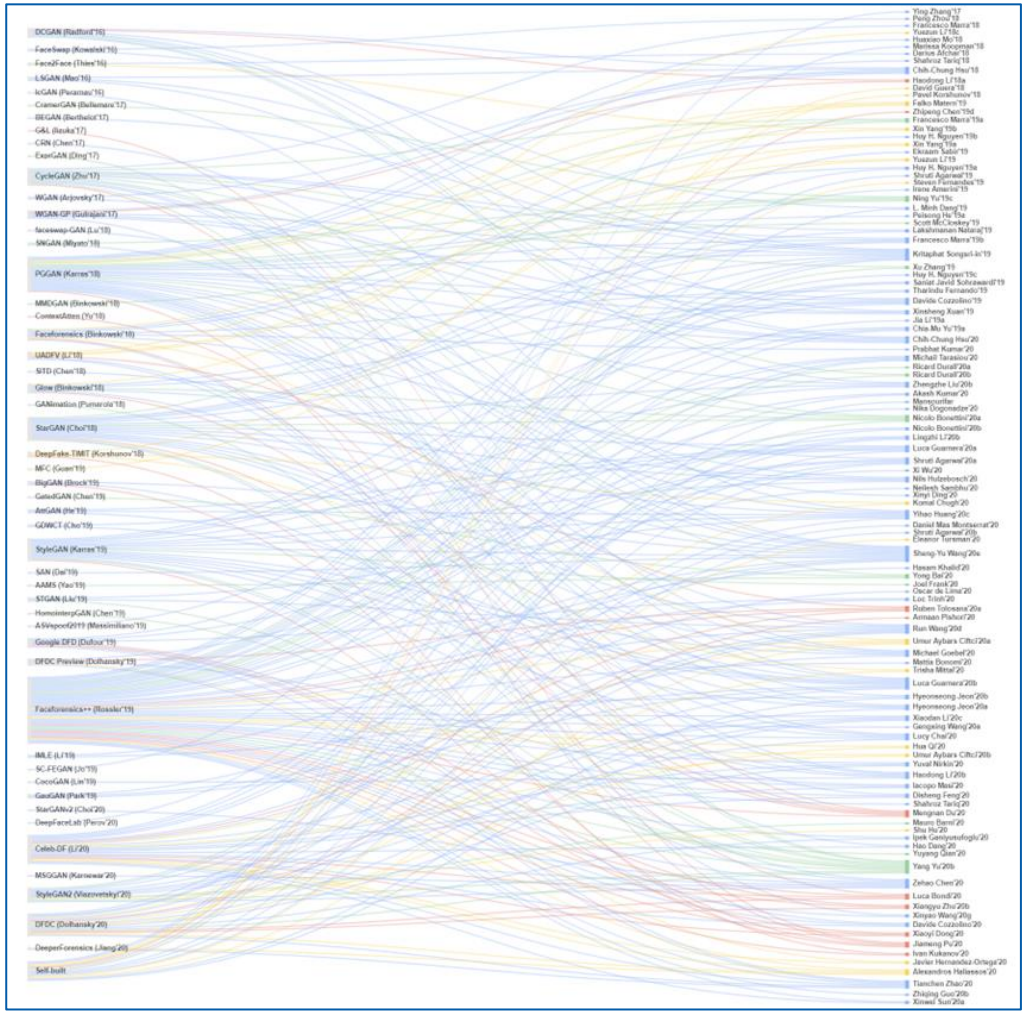
Нарушение временной целостности – удаление кадров в середине



Нарушение временной целостности – удаление кадров в конце

Состояние «гонки вооружений»

Создание



Обнаружение

- Гонка вооружений
  - как только появляется новый (пассивный) детектор, создатели дипфейков ищут способы его обойти
- Проактивная защита медиаконтента
  - цифровые водяные знаки

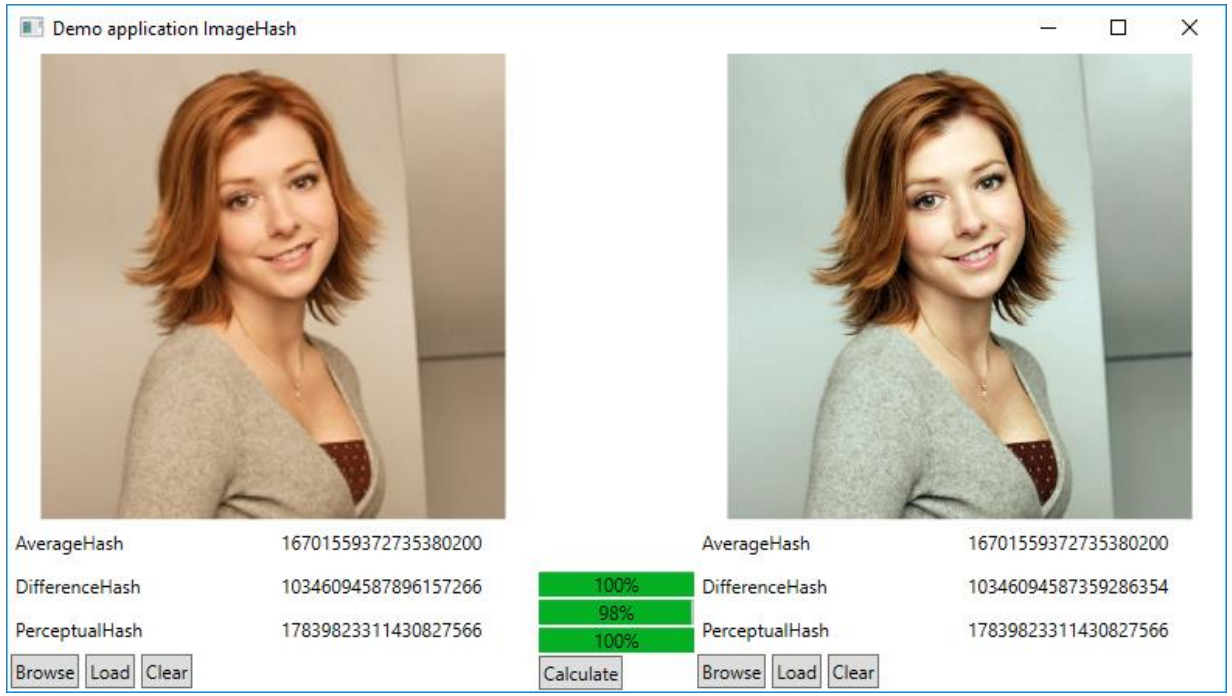
Идея метода обеспечения контроля целостности изображений

- Вычисляется хеш-код исходного изображения
- Хеш-код передается совместно с изображением
- Вычисляется хеш-код полученного (возможно, модифицированного) изображения
- Вычисляется расстояние Хэмминга между хеш-кодами двух изображений
- Сравнение с пороговым значением для принятия решения о факте модификации изображения
  - Ниже порога – к изображению применялись только допустимые преобразования
  - Выше порога – целостность изображения нарушена

- Как передать хеш-код исходного изображения совместно с ним?
  - Цифровой водяной знак
- Как обеспечить защиту от подмены хеш-кода?
  - Электронная цифровая подпись

Перцептивный хеш-код

- Задача:
  - вычислить по входному изображению хеш-код – последовательность бит фиксированной длины
- Отличие от криптографических хеш-функций:
  - перцептивные хеш-коды семантически близких изображений близки по расстоянию Хэмминга
- Области применения:
  - Поиск похожих изображений и выявление дубликатов
    - Защита авторских прав
    - Выявление запрещенного контента
  - **Обнаружение изменений в изображениях**



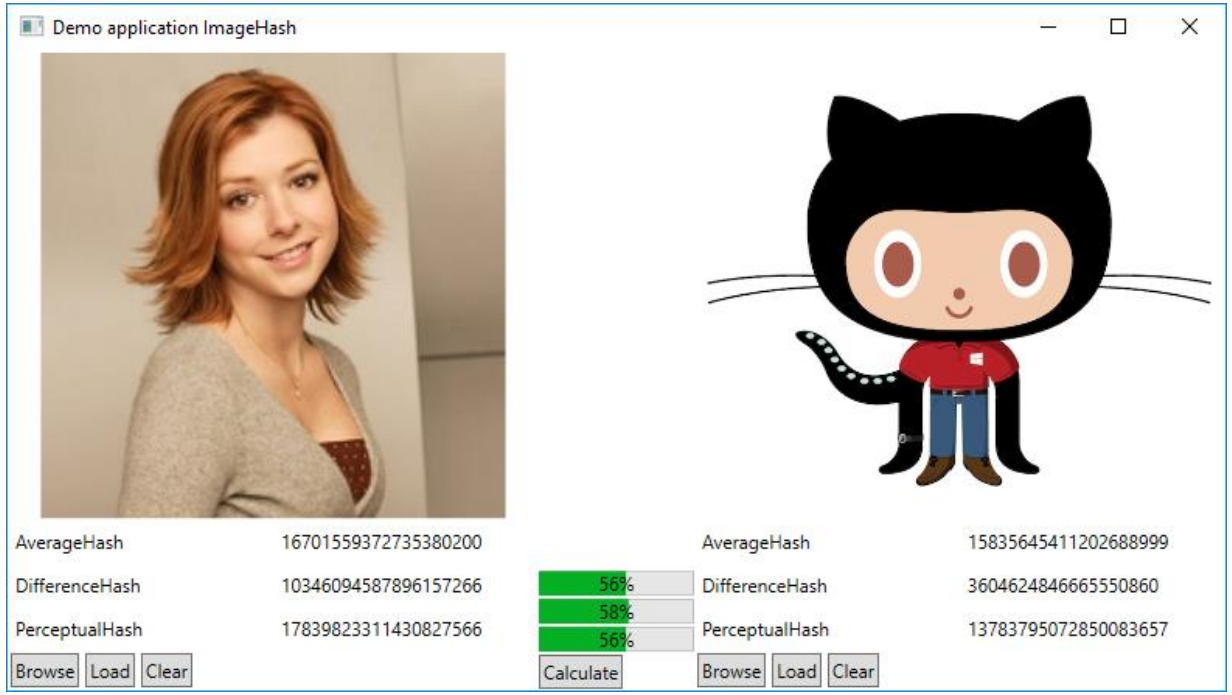
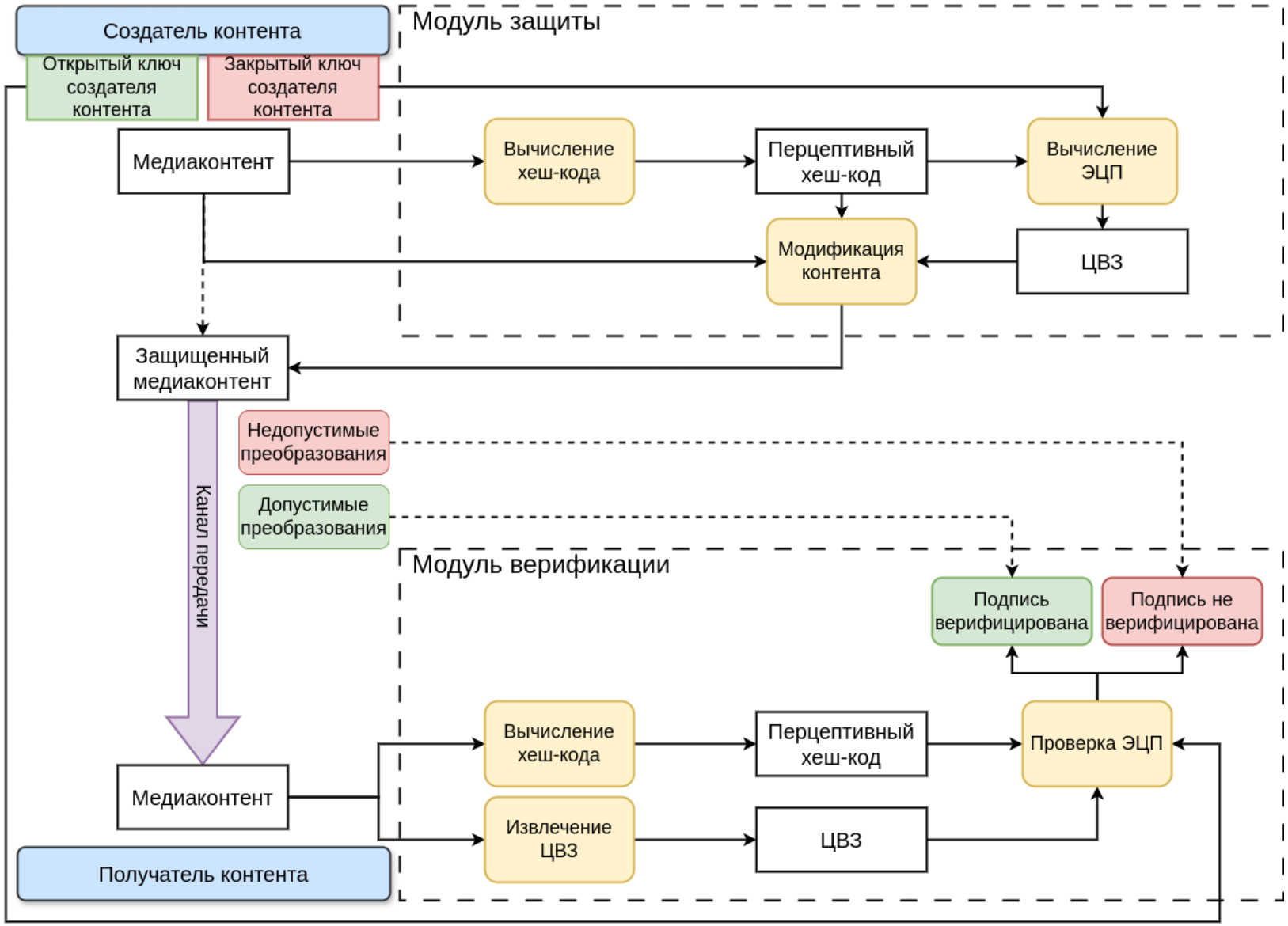


Схема алгоритма для контроля целостности изображений



Перцептивные хеш-функции: устойчивость к сжатию JPEG

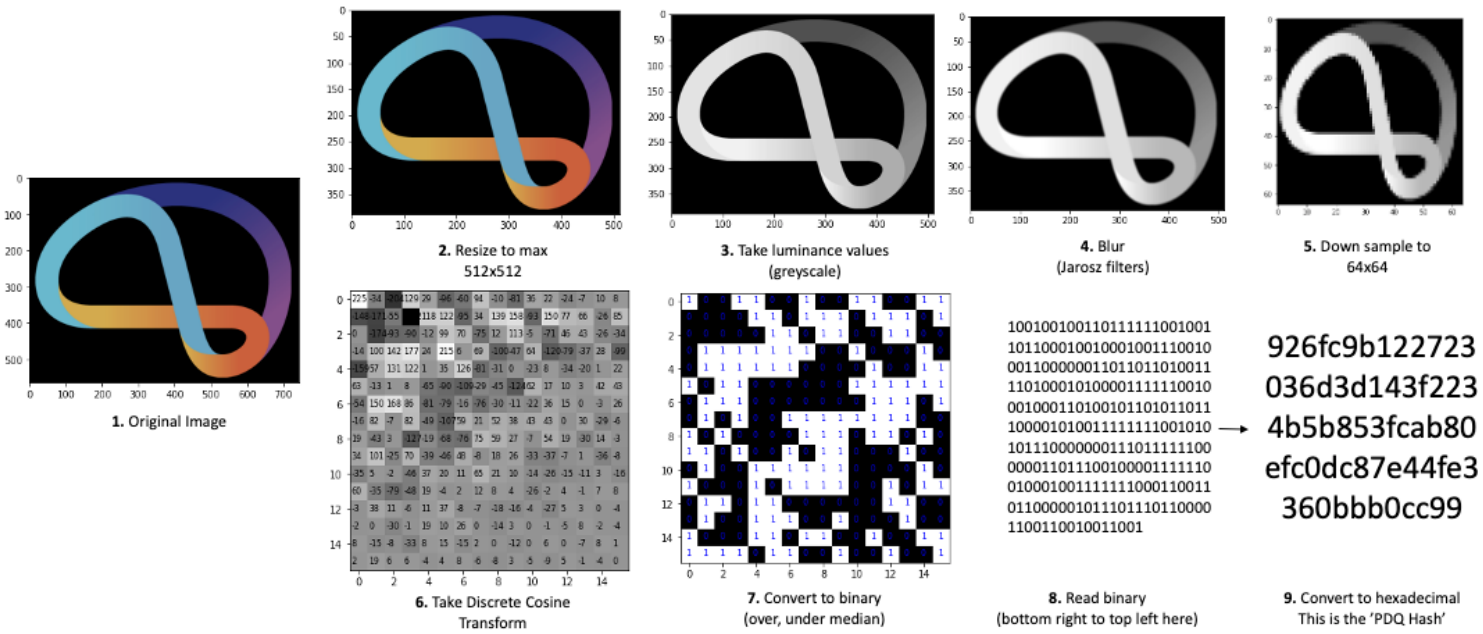
Наличие ЭЦП требует полное совпадение хеш-кодов при допустимых преобразованиях, в частности, после сжатия JPEG. Проведено тестирование на устойчивость существующих перцептивных хеш-функций на устойчивость к сжатию JPEG с качеством 50.

Перцептивная хеш-функция	Основа перцептивной хеш-функции	Доля изображений
Average hash	Яркость пикселей масштабированного изображения	94.3%
Color hash	Цветовое пространство HSV	75.8%
Dhash	Градиент яркости пикселей масштабированного изображения	81.2%
Phash	Коэффициенты дискретного косинусного преобразования	92.5%
Whash	Коэффициенты дискретного вейвлет-преобразования	**97.3%**
Crop resistant hash	Результат работы алгоритма сегментации изображения	12%
PDQ hash	Коэффициенты дискретного косинусного преобразования	62.4%
Phash org	Коэффициенты дискретного косинусного преобразования	95.5%
Image hashing	Карта признаков сверточной нейронной сети	61%

Без дополнительной модификации перцептивные хеш-функции неприменимы в предлагаемой схеме

Перцептивная хеш-функция на основе ДКП

1. Предобработка изображения: получение одноканального изображения фиксированного размера
2. Применение ДКП
3. Выделение 16x16 низкочастотных коэффициентов
4. Порог – медианное значение этих коэффициентов
5. Бинаризация по порогу с получением 256 бит перцептивного хеш-кода



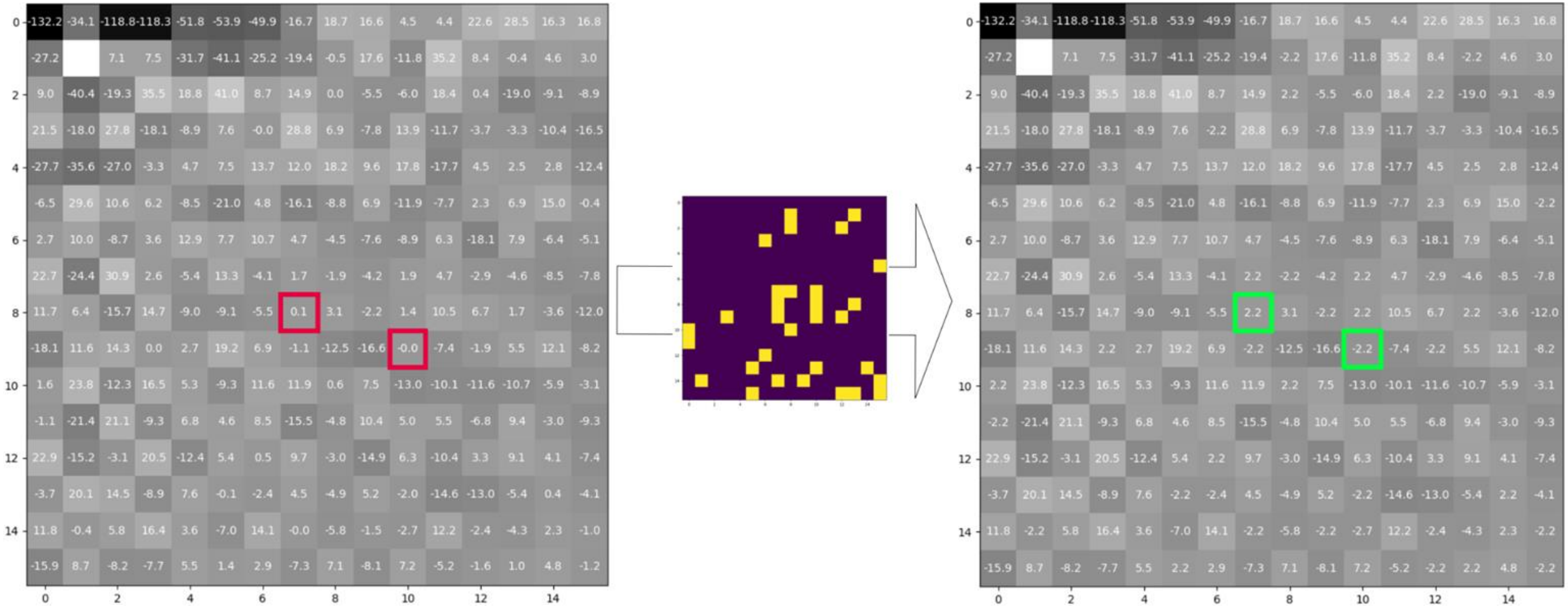
https://github.com/darwinium-com/pdqhash

Двумерное дискретное косинусное преобразование:

$$X_{k_1,k_2} = \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} x_{n_1,n_2} \cos\left[\frac{\pi}{N_1}\left(n_1+\frac{1}{2}\right)k_1\right] \cos\left[\frac{\pi}{N_2}\left(n_2+\frac{1}{2}\right)k_2\right]$$

Проблема: после применения допустимых преобразований коэффициенты ДКП изменяются. Знак сравнения некоторых коэффициентов с порогом может инвертироваться, из-за чего возникает несовпадение хеш-кодов. Особенно уязвимы ближайшие к порогу коэффициенты.

Повышение устойчивости водяного знака



Оригинальное изображение дополнительно модифицируется так, чтобы соответствующие коэффициенты ДКП сильнее отличались от медианного значения

Повышение устойчивости водяного знака

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

- Степень модификации изображения определяется целевым значением метрики близости изображений
- Выбрана метрика **PSNR**
- Целевое значение PSNR = 50

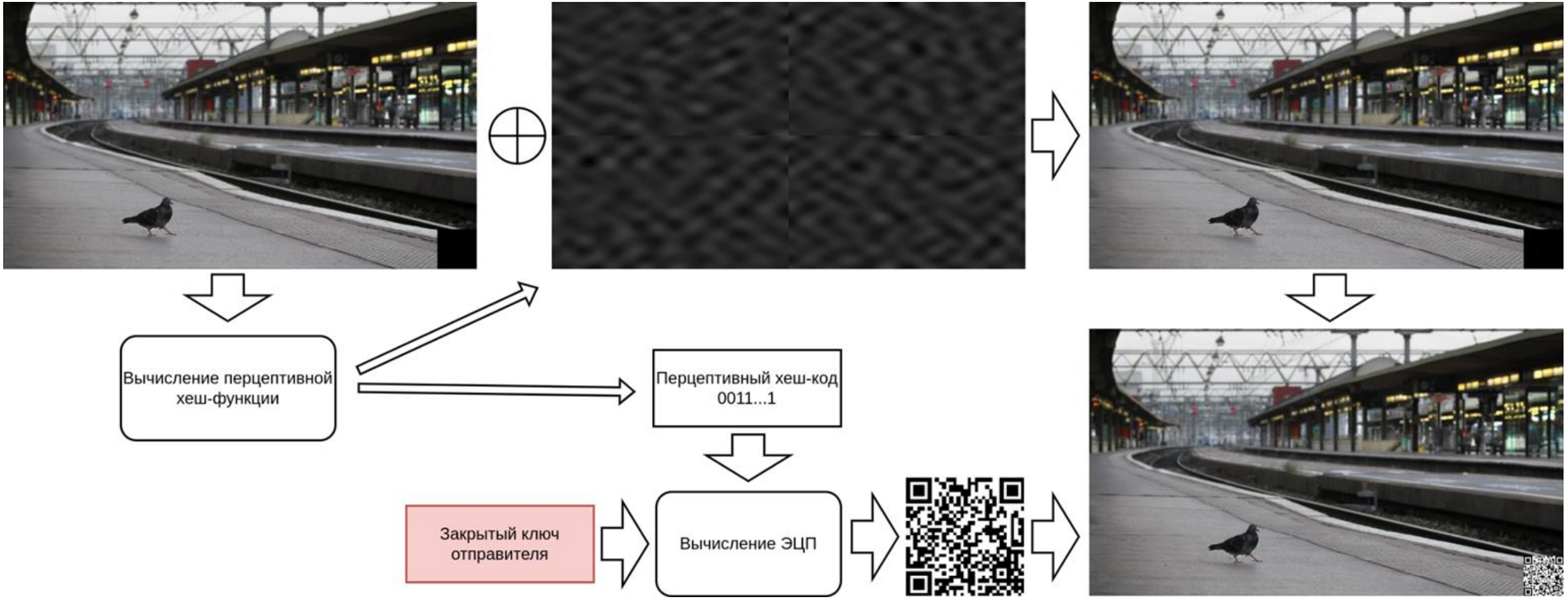
$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |I(i,j) - K(i,j)|^2$$

$$PSNR = 10 \log_{10} \left(\frac{MAX_I^2}{MSE} \right) = 20 \log_{10} \left(\frac{MAX_I}{\sqrt{MSE}} \right)$$

Внедрение водяного знака

Защищается модифицированное изображение, а не оригинальное:

- По перцептивному хеш-коду изображения формируется ЭЦП отправителя
- ЭЦП внедряется в защищаемое изображение как ЦВЗ в виде QR-кода



Извлечение водяного знака и верификация ЭЦП

Вычисление перцептивной хеш-функции

Перцептивный хеш-код
0011...1

Значение ЭЦП

Верификация ЭЦП

Открытый ключ отправителя

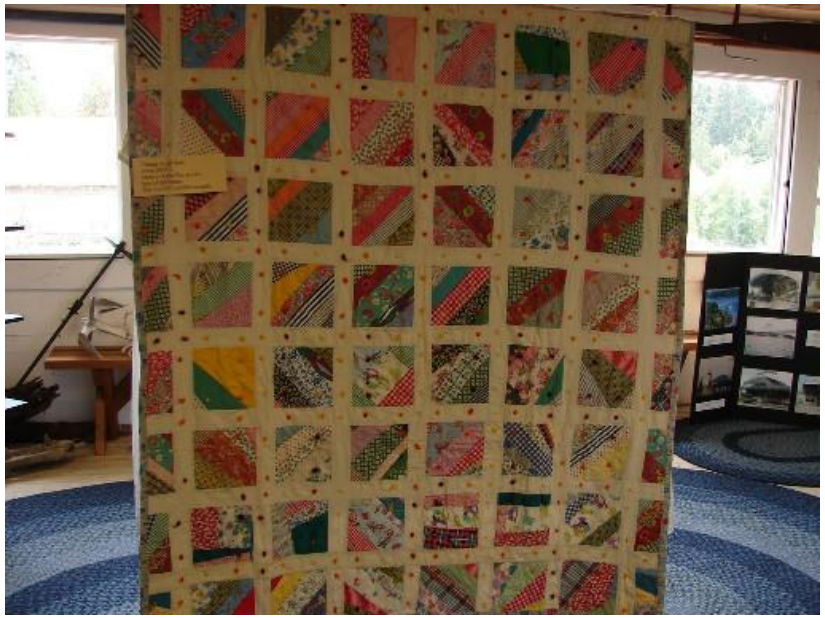
Подпись верифицирована: целостность изображения сохранена

Подпись не верифицирована: целостность изображения нарушена

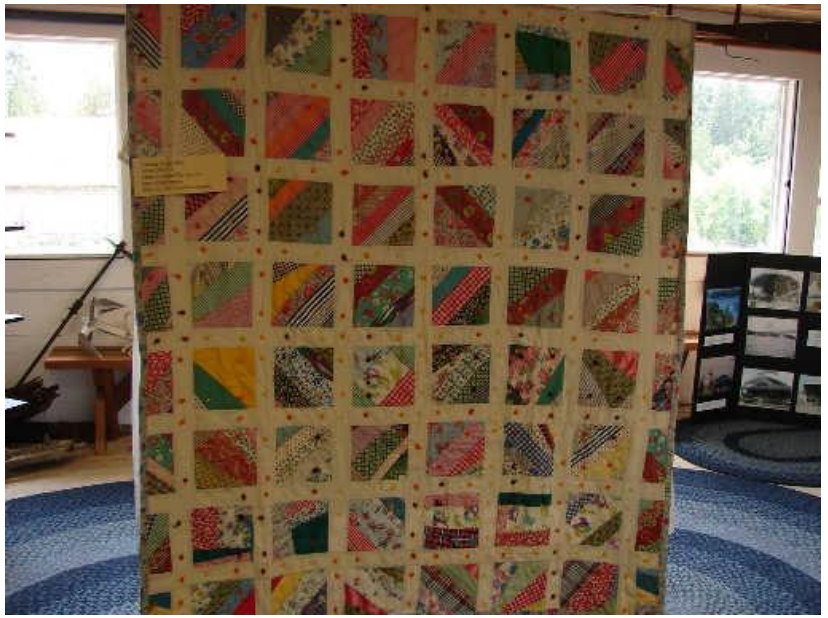
Тестирование: допустимые преобразования изображений

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%



Качество JPEG-90



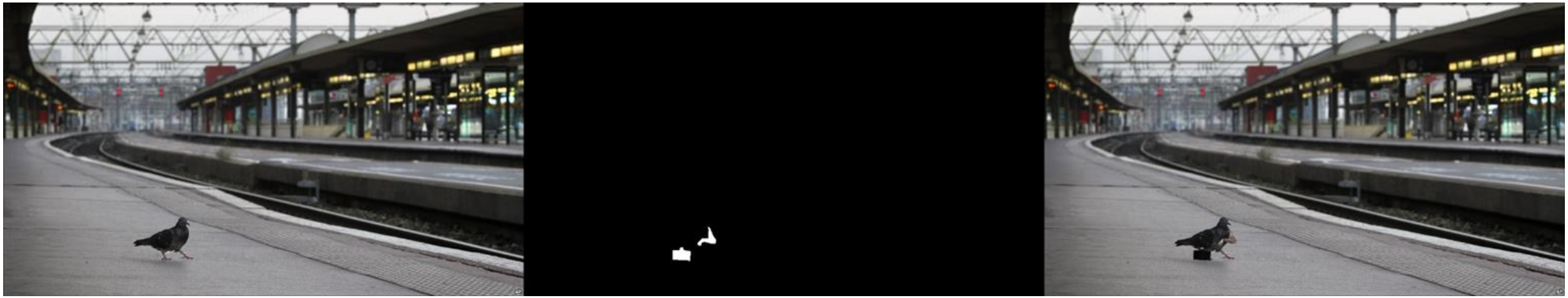
Качество JPEG-50

Качество JPEG-10

Kuznetsova A. et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale – 2020

Тестирование: недопустимые преобразования изображений

- Набор изображений IMD2020, содержащий тройки изображений:
  - Оригинальное изображение
  - Поддельное изображение (добавление / удаление объектов, изменение фона и т.д.)
  - Маска модификации оригинального изображения
- Имитация недопустимого преобразования:
  - Защита оригинального изображения предлагаемым методом
  - Перенос изменений на защищенное изображение по маске модификации
  - Сравнение исходного хеш-кода и хеш-кода измененного защищенного изображения
- Перцептивный хеш-код изменился после недопустимого преобразования на **97.2%** изображений



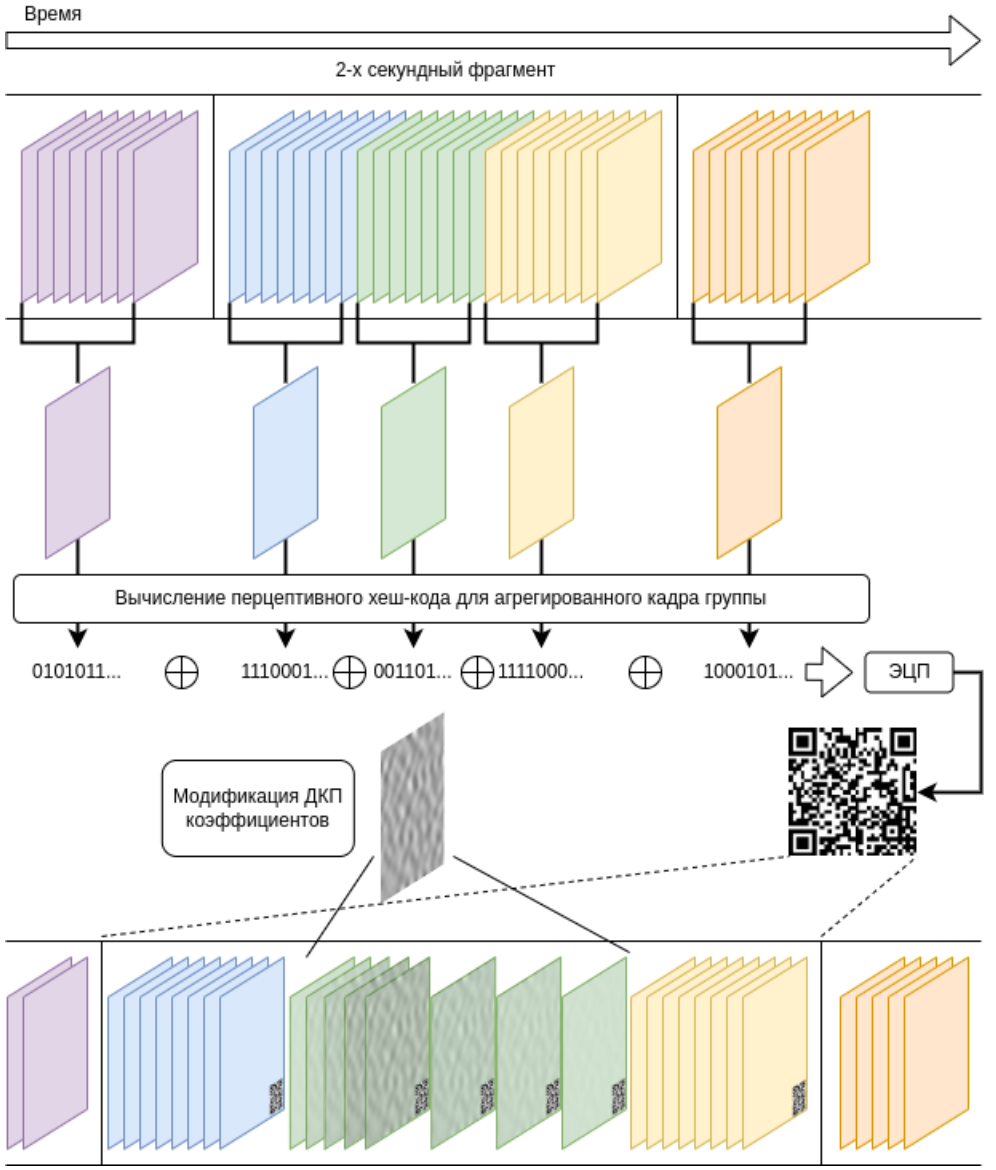
Novozamsky A., Mahdian B., Saic S. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images – 2020

Адаптация алгоритма для защиты видеоконтента

Контроль целостности видео:

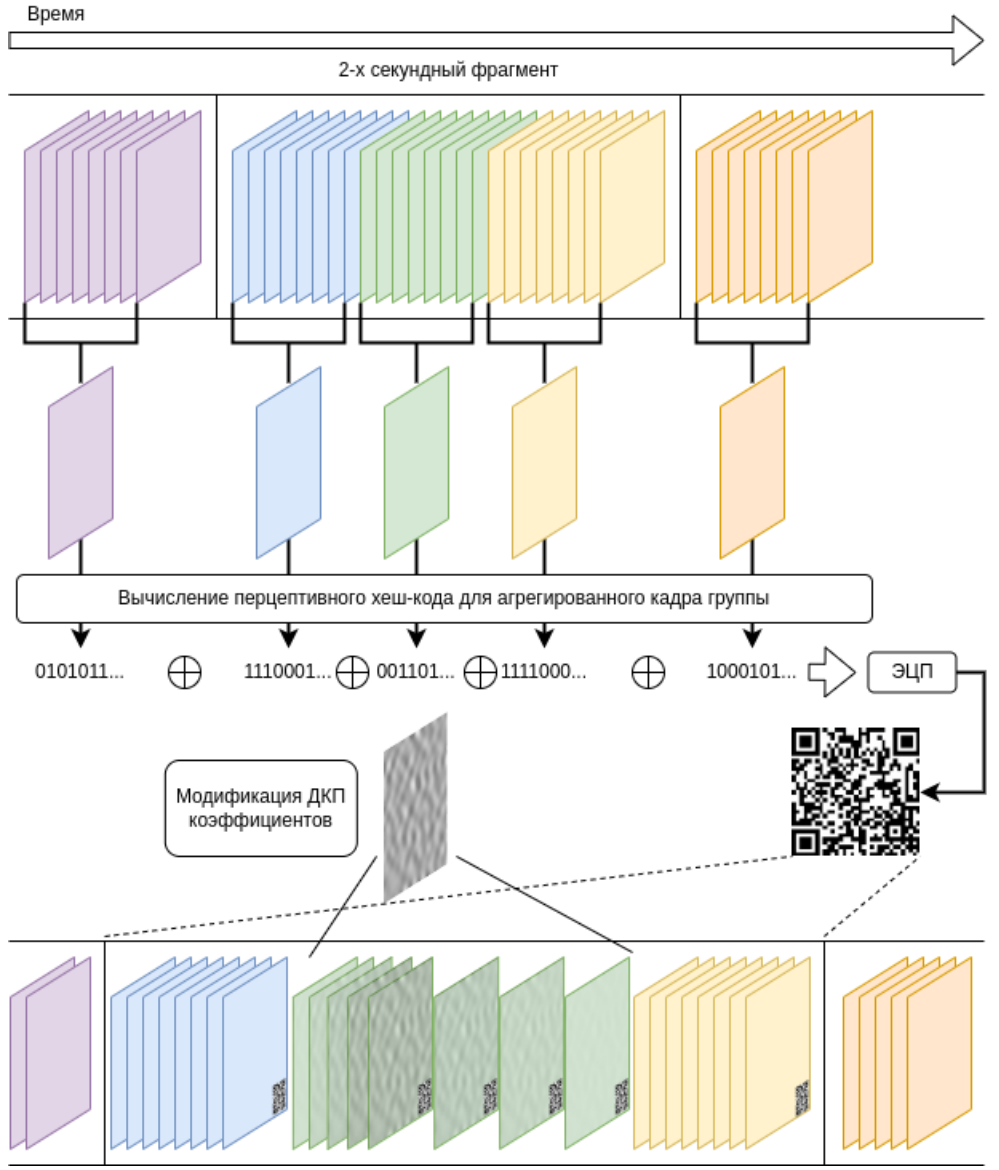
- Контроль пространственной целостности отдельных кадров видео как изображений
- Контроль временной целостности видео (перестановка, дублирование, удаление кадров, а также ускорение и замедление видео)

Сохранение плавности перехода между соседними кадрами для незаметности



Адаптация алгоритма для защиты видеоконтента

1. Исходное видео разбивается на 2-х секундные фрагменты, фрагменты – на 3 группы кадров
2. Для группы кадров вычисляется агрегированный кадр (усреднение по пикселям в одинаковых позициях)
3. Перцептивный хеш-код вычисляется для агрегированного кадра (как для изображения)
4. Требуемая «прибавка» для перцептивного хеш-кода накладывается на все кадры в группе с разным коэффициентом:
   - для обеспечения плавности переходов между группами, на границах группы коэффициент минимален
   - в середине группы – наибольший коэффициент

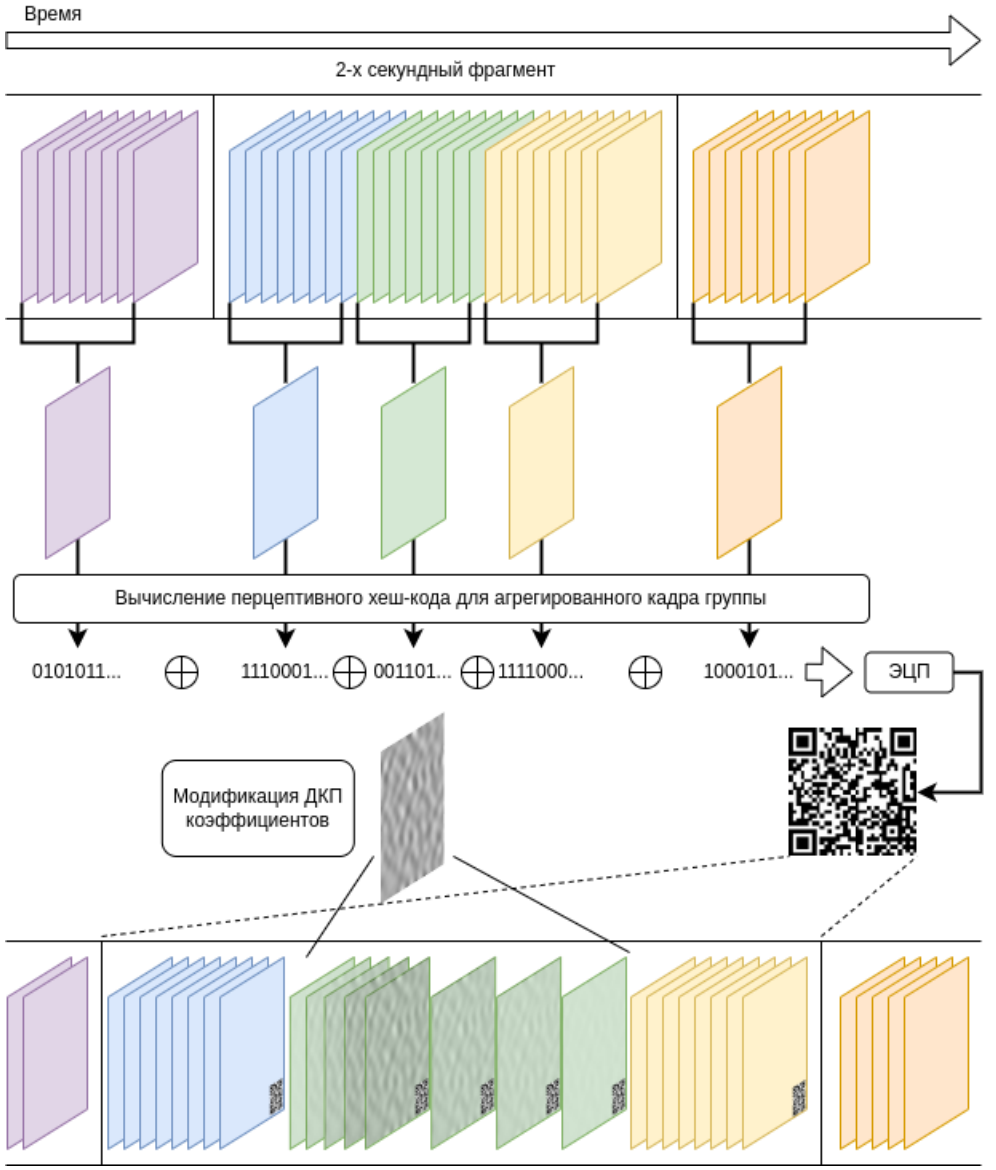


Адаптация алгоритма для защиты видеоконтента

Временная целостность видео

- общий QR-код для фрагмента видео, содержащего несколько целых групп кадров
- перцептивные хеш-коды групп одного фрагмента последовательно объединяются
- также добавляются хеш-коды групп из соседних фрагментов: обеспечение временной целостности фрагментов
- ЭЦП вычисляется по объединенным перцептивным хеш-кодам групп кадров
- По ЭЦП формируется QR-код, общий для всех кадров в фрагменте

Алгоритм верификации видео аналогичен алгоритму верификации изображений



Тестирование: допустимые преобразования видео

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

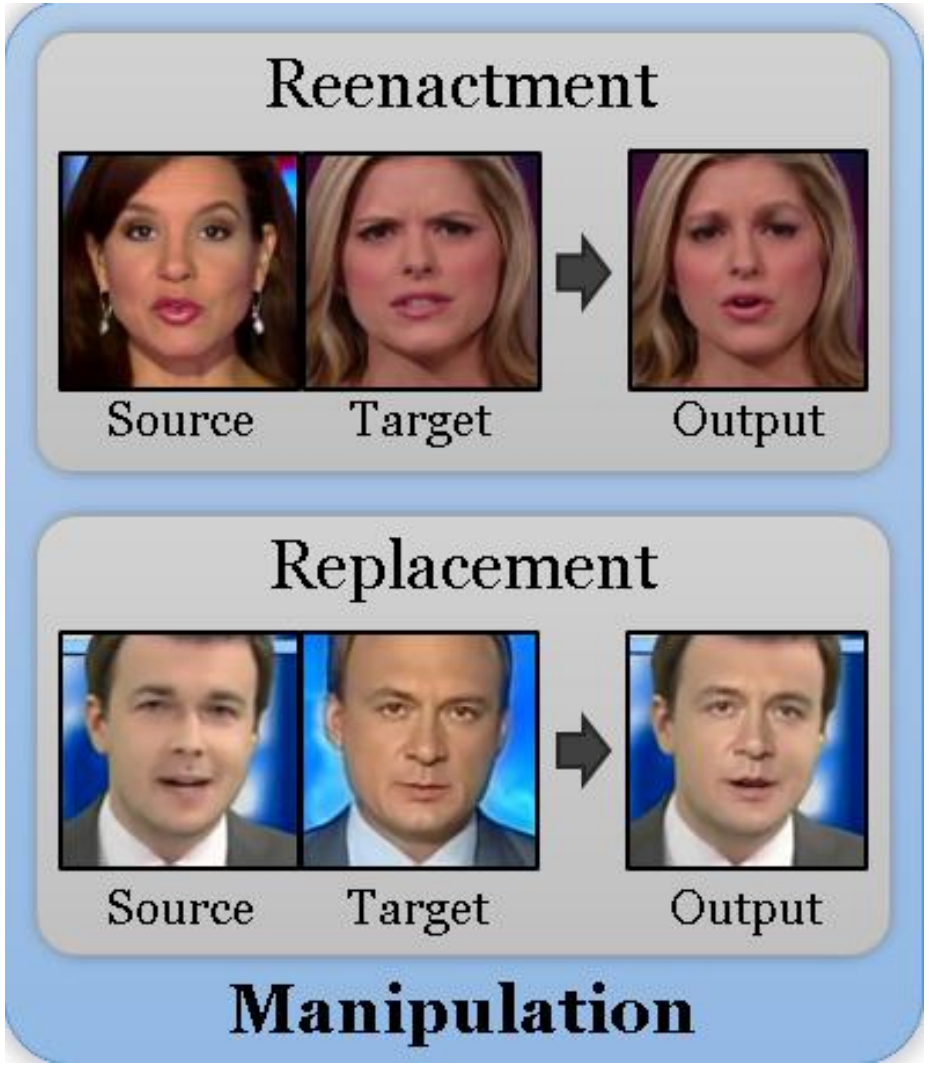
CRF	20	25	30	35	40
Доля фрагментов	0%	0%	0.2%	7.3%	66.2%

Caelles S. et al. The 2019 davis challenge on vos: Unsupervised multi-object segmentation – 2019

Тестирование: недопустимые преобразования видео

Изменение перцептивного хеш-кода при недопустимых преобразованиях:

- Набор видео FaceForensics++, содержащий пары оригинальное видео и видео, созданное с применением технологии DeepFake, а также маска изменений
- Доля групп кадров, для которых значение перцептивного хеш-кода изменилось, составила **99.6%**

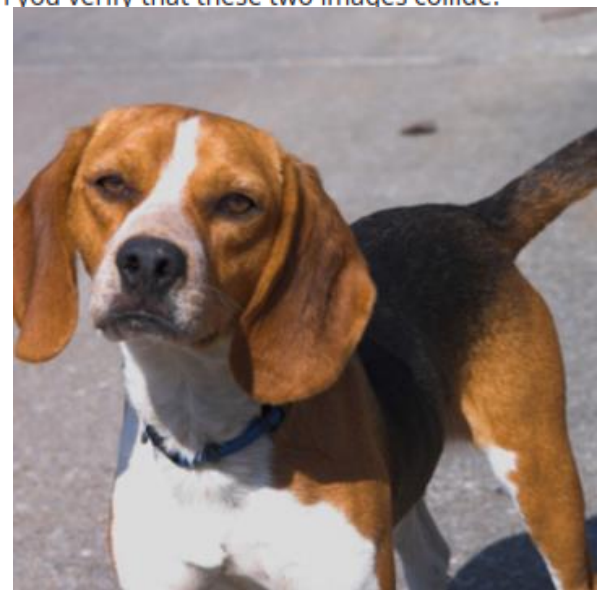


Rossler A. et al. FaceForensics++: Learning to detect manipulated facial images – 2019

Перцептивные хеш-функции: коллизии

dxoigmn commented on Aug 18, 2021 • edited

Can you verify that these two images collide?



Here's what I see from following your directions:

```
$ python3 nnhash.py NeuralHash/model.onnx neuralhash_128x96_seed1.dat beagle360.png
59a34eabe31910abfb06f308
$ python3 nnhash.py NeuralHash/model.onnx neuralhash_128x96_seed1.dat collision.png
59a34eabe31910abfb06f308
```

👍 501 😄 144 🎉 69 😕 17 ❤️ 55 🚀 58 👀 207

https://github.com/AsuharietYgvar/AppleNeuralHash2ONNX/issues/1

Выводы

- Гонка вооружений
  - непрерывное соревнование между создателями дипфейков и разработчиками детекторов дипфейков
- Детекторы дипфейков
  - неотъемлемая часть общего подхода по борьбе с фальсификациями медиа контента,
  - требуются существенные вложения времени и ресурсов
- Синтезируемый ИИ-контент
  - упрощение создание контента, в том числе, в злонамеренных целях
- Водяные знаки
  - для оригинального контента
    - эффективный подход обеспечения доверия контенту
  - для синтезируемого контента
    - факт генерации контента показан явно и не скрывается