Технологические возможности маркировки контента, маркировка медиаконтента в вопросах кибербезопасности

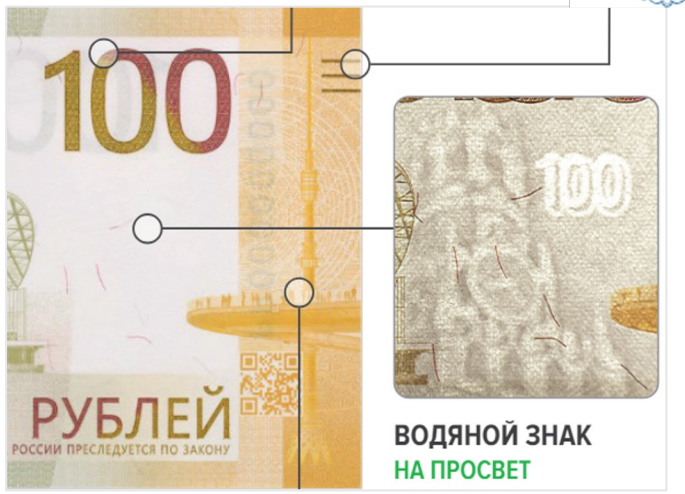
Маркин Юрий Витальевич,
к.т.н., научный сотрудник ИСП РАН

Москва, 13 мая 2026 г.

Что такое маркировка

- Маркировка – это нанесение условных знаков, букв, цифр, графических знаков или надписей на объект
- Цели
  - идентификация (узнавание) объекта, указание его свойств и характеристик
  - подтверждение подлинности / затруднение создания подделок
- Далее: маркировка = цифровой водяной знак (ЦВЗ)

ИСП РАН



Академия криптографии
Российской Федерации

Задачи, решаемые с помощью маркировки

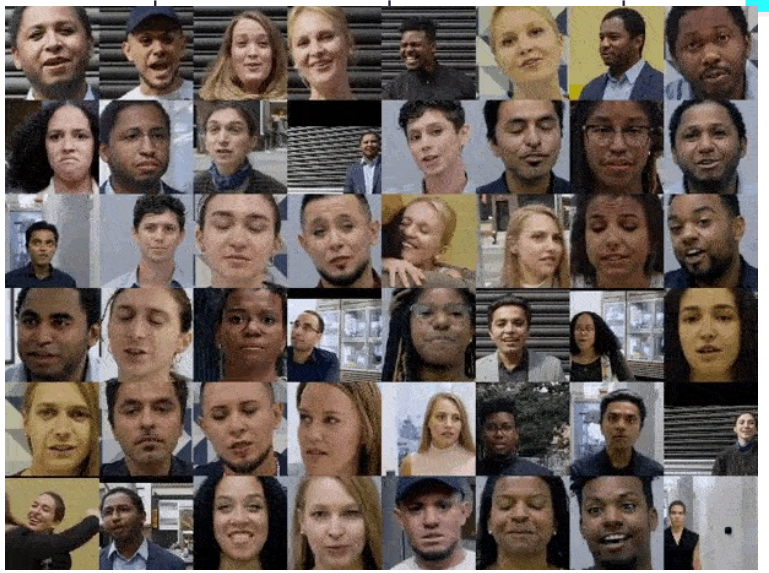
- Идентификация владельца / подтверждение права собственности
  - ЦВЗ может использоваться для предоставления информации о владельце или источнике контента
- Отслеживание распространения (content fingerprinting)
  - в каждую копию медиаконтента внедряется уникальное значение ЦВЗ;
  - разным пользователям предоставляются разные копии (каждому пользователю – своя)
- **Контроль целостности содержимого контента / обнаружение намеренных модификаций**
  - обнаружение факта модификации – «да/нет»;
  - локализация модификаций (временна́я и/или пространственная)
- **Фиксация природы происхождения контента**
  - сгенерирован или создан человеком

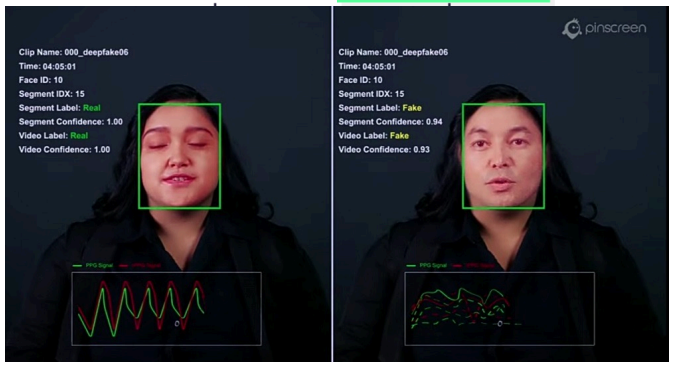
Дипфейк – мощное оружие в руках злоумышленника

- 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

До 2014	2014-2016	2017-2018	2019-2020	2021-н.в.
Создание	Обнаружение	Создание	Обнаружение	Создание	Обнаружение	Создание	Обнаружение	Создание	Обнаружение
Морфинг, композитинг	Анализ артефактов сжатия и света	Generative Adversarial Networks	Поиск артефактов GAN при помощи CNN	FaceSwap	FaceForensics++	StyleGAN (Nvidia)	Гибридные методы	Stable Diffusion	Наполнение сосудов лица кровью
		Доступные инструменты	Датасеты для подготовки детекторов			Реалистичные лица	Анализ моргания, движений губ	Диффузионные модели	FakeCatcher (Intel)
								Любой контент	Поиск "ИИ-отпечатков"
								Видео	Водяные знаки





Гонка вооружений

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

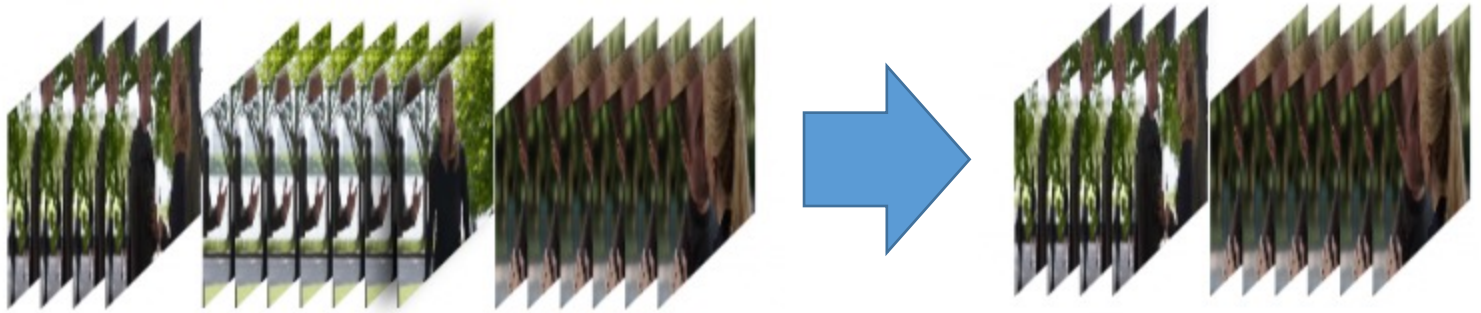
Дипфейк ≠ Модификация

- **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**

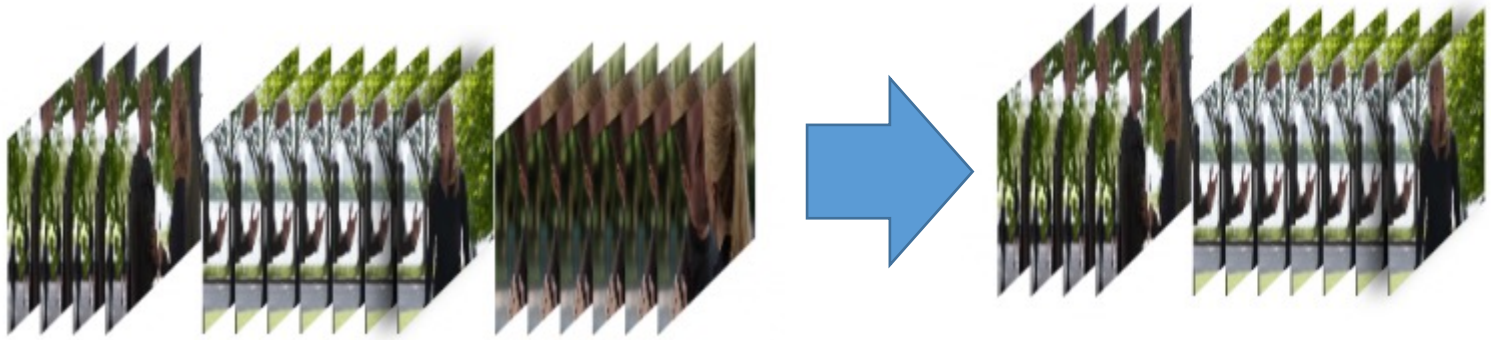
Допустимая модификация	Дипфейк
транскодирование другим кодеком с другим уровнем качества (сжатие)	подмена контента (в частности, наложение другого лица)
изменение частоты кадров при транскодировании видео	удаление или вставка фрагментов аудио / видео
наложение шумов на аудиосигнал	ускорение или замедление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

Пространственная и временная целостность медиаконтента

Нарушение пространственной целостности



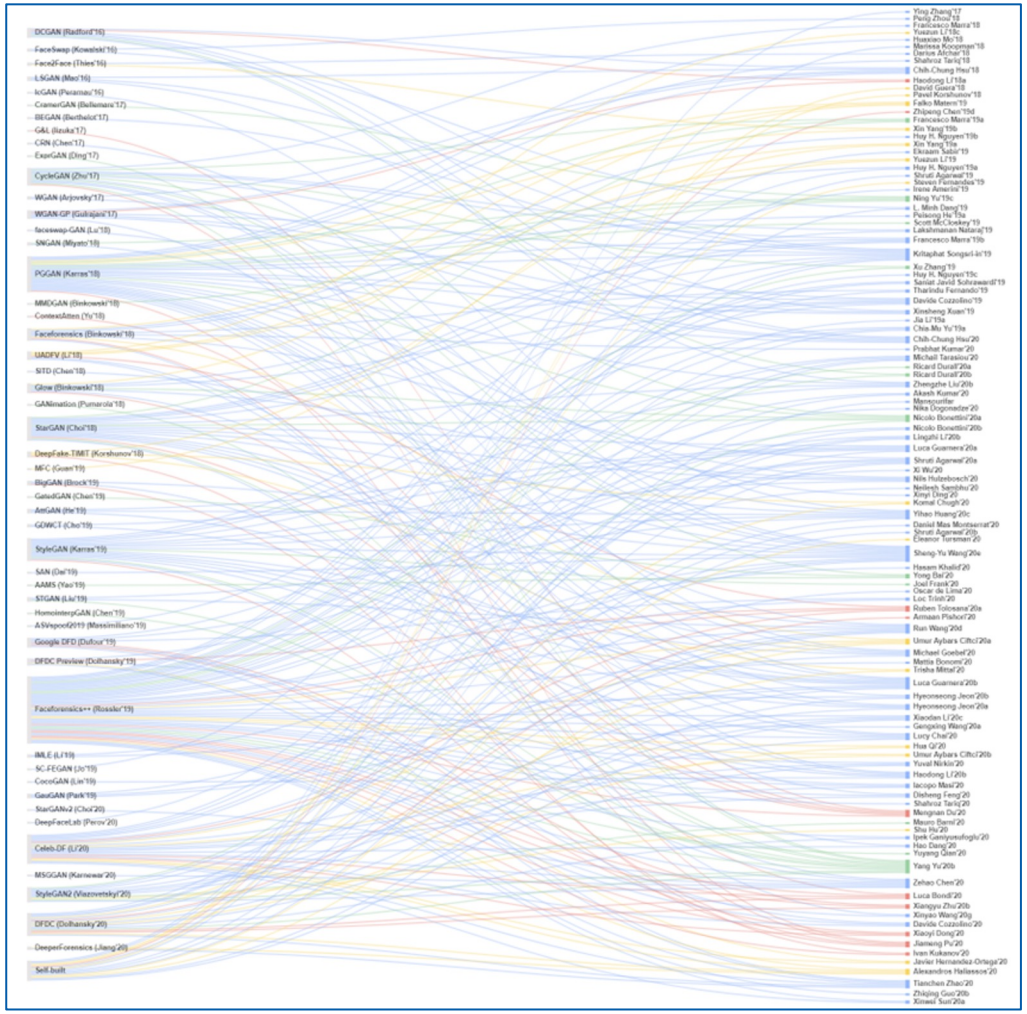
Нарушение временной целостности – удаление кадров в середине



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

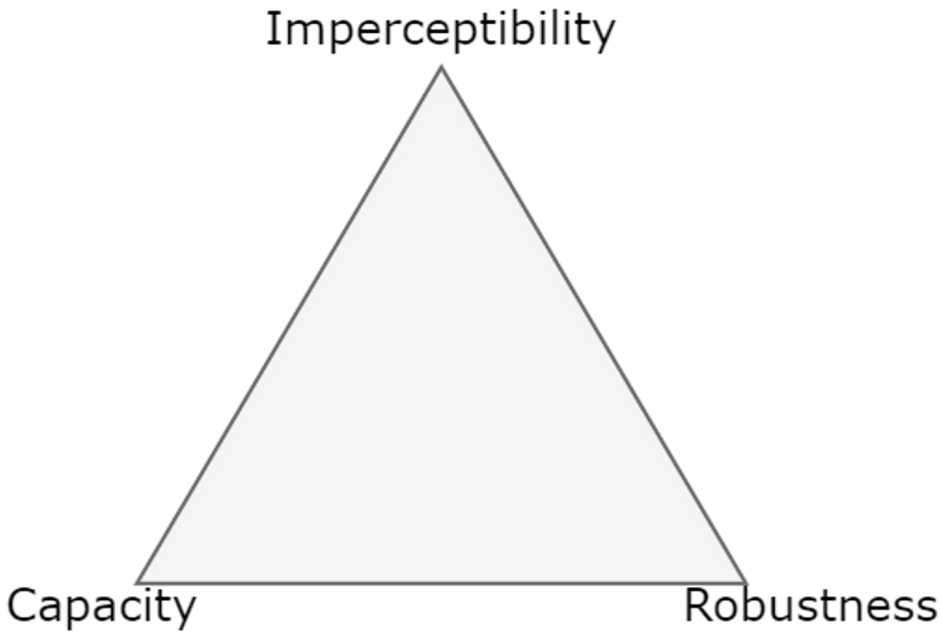
Водяные знаки для маркировки медиаконтента

Заметные водяные знаки

Скрытые водяные знаки

Основные свойства и типы цифровых водяных знаков

- Незаметность (Imperceptibility)
- Емкость (Capacity)
  - объем содержащейся в ЦВЗ информации;
- Устойчивость (Robustness)
  - способность ЦВЗ быть извлекаемым после проведения преобразований над объектом внедрения



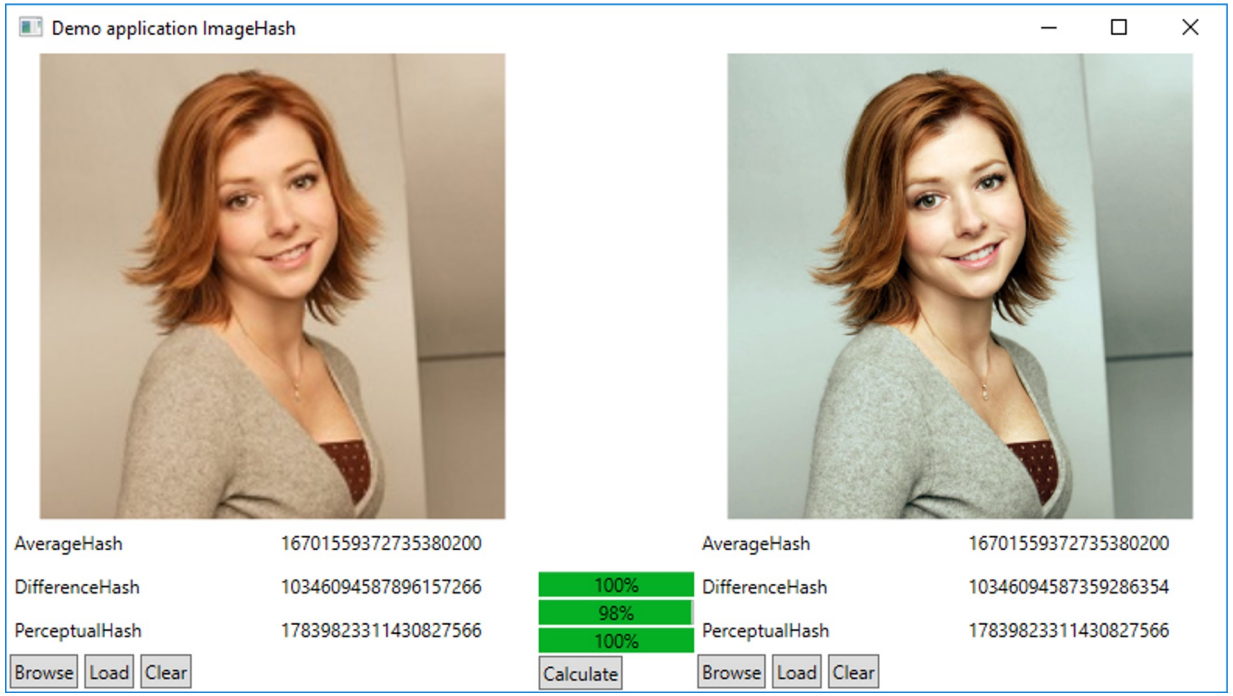
Fragile – «Хрупкий»
- должен быть чувствительным ко всем манипуляциям над содержимым медиафайла
- низкая заметность и большая емкость
- подтверждение подлинности

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

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**



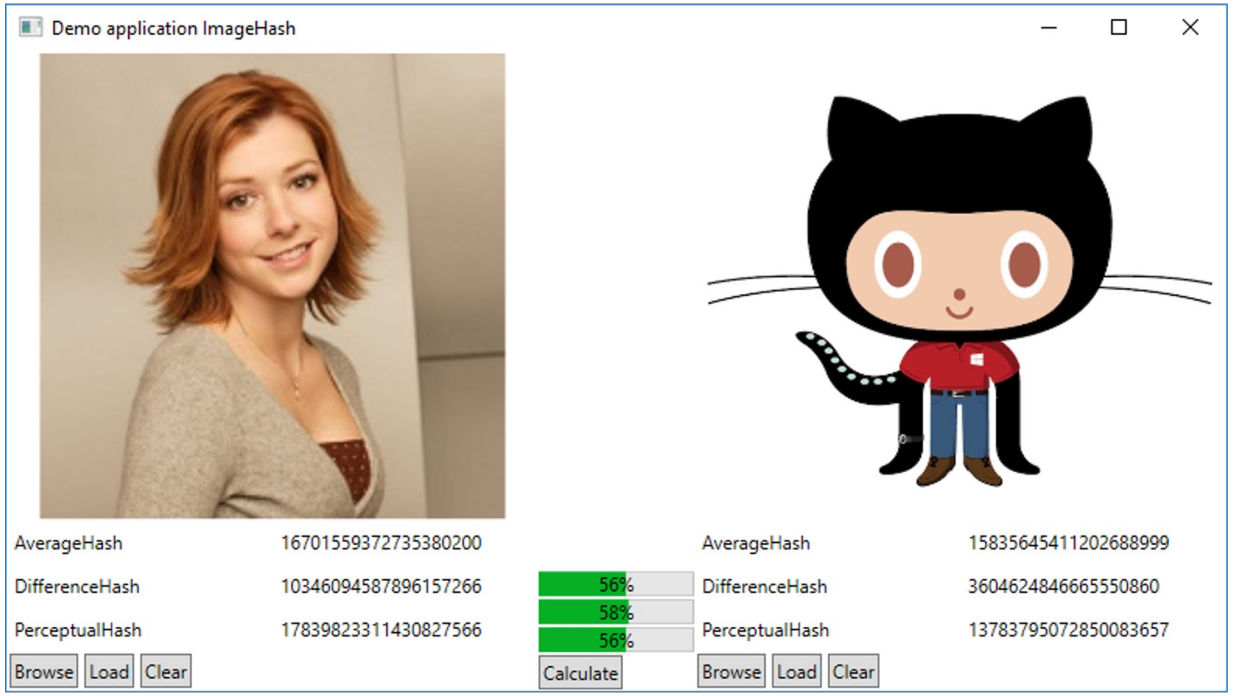
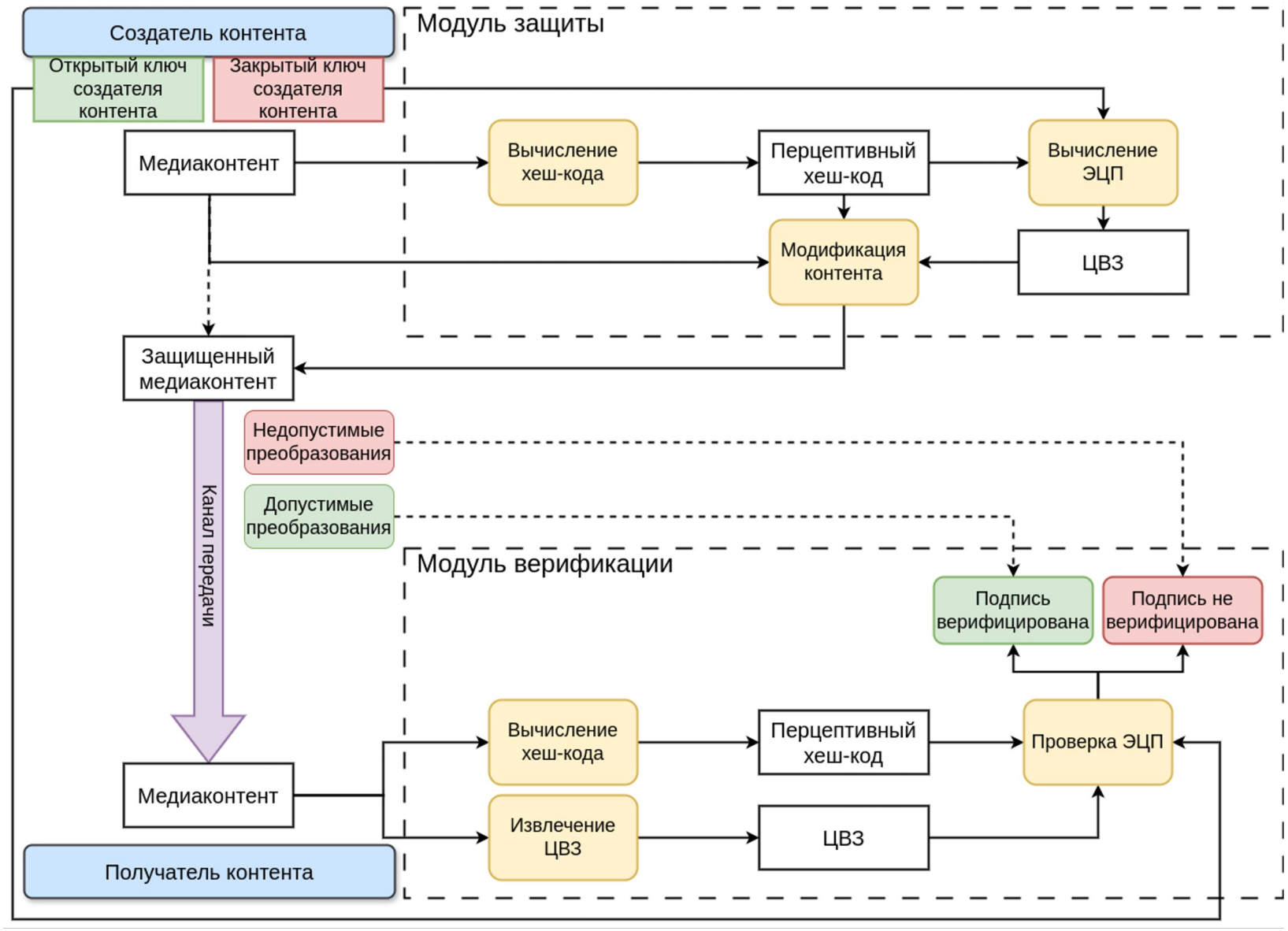


Схема алгоритма контроля целостности изображений



Перцептивные хеш-функции: устойчивость к сжатию JPEG

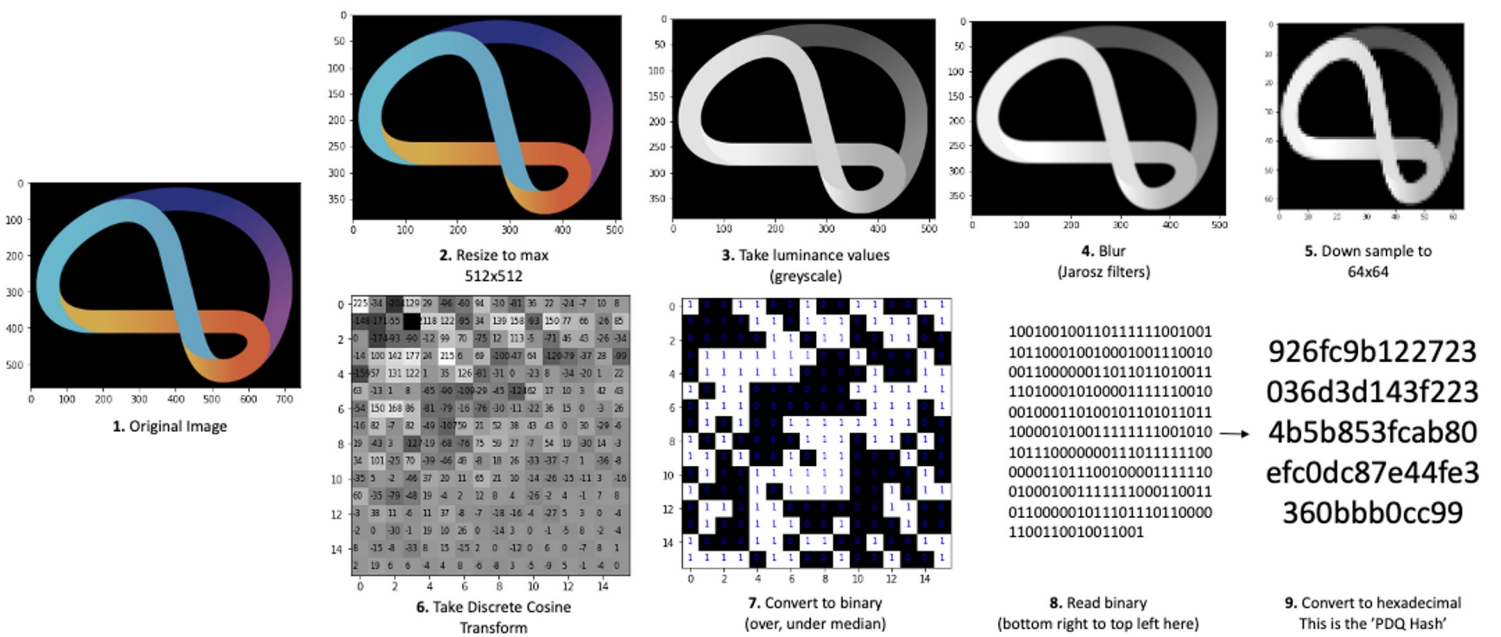
Наличие ЭЦП требует полное совпадение хеш-кодов при допустимых преобразованиях, в частности, после сжатия JPEG. Проведено тестирование на устойчивость существующих перцептивных хеш-функций на устойчивость к сжатию JPEG с качеством 50.

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

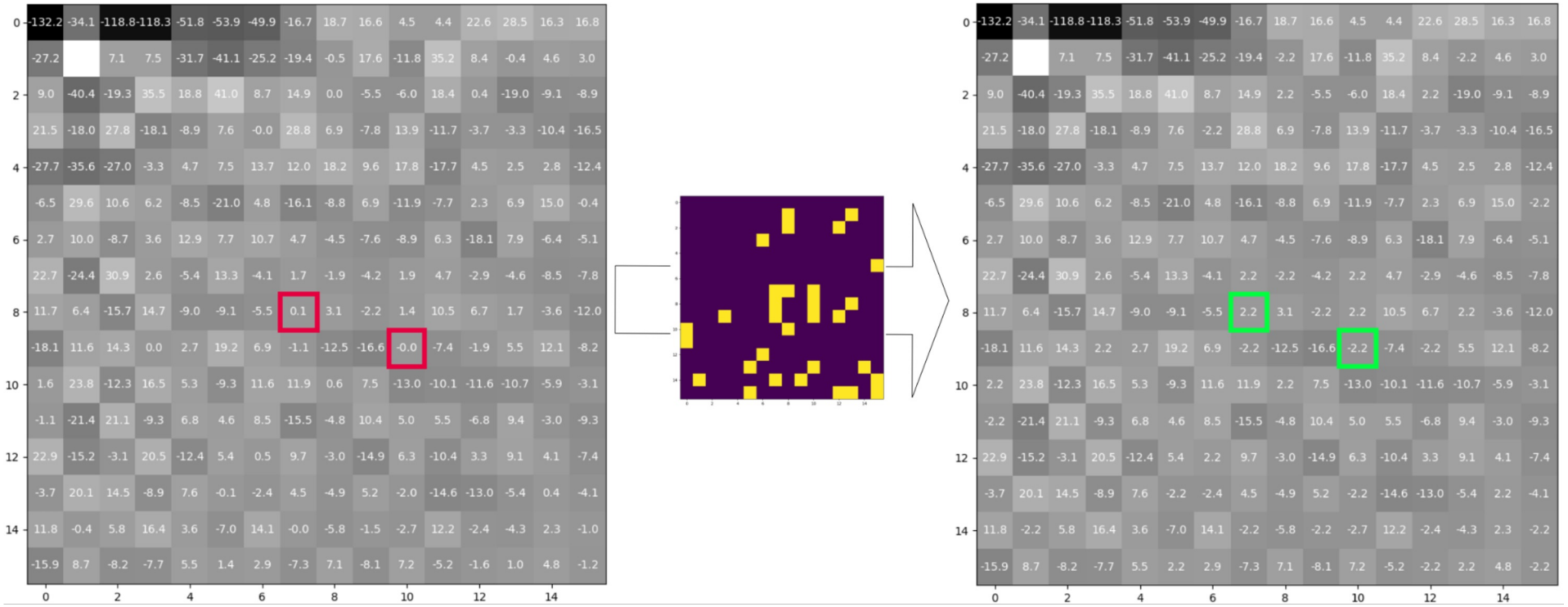
Двумерное дискретное косинусное преобразование:

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

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

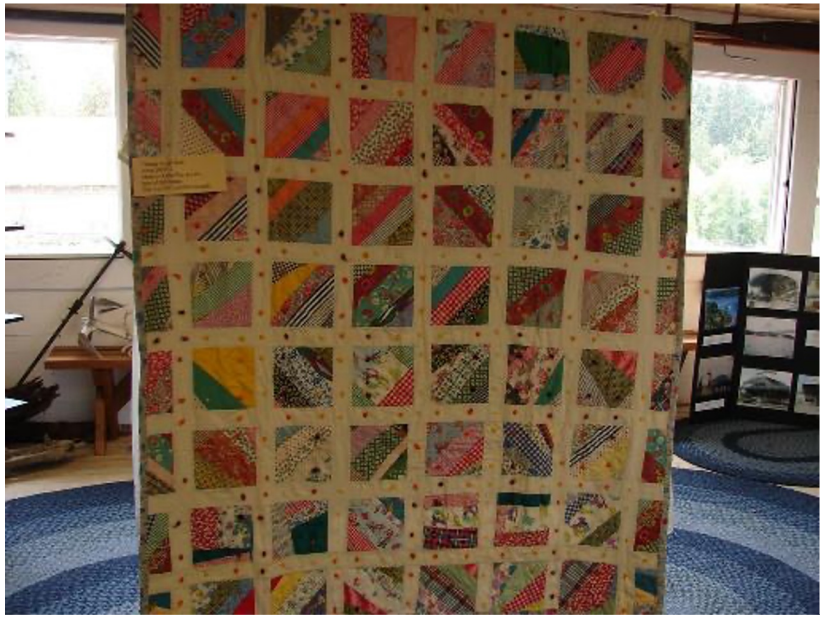
Оригинальное изображение дополнительно модифицируется, чтобы соответствующие коэффициенты ДКП сильнее отличались от медианного значения



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

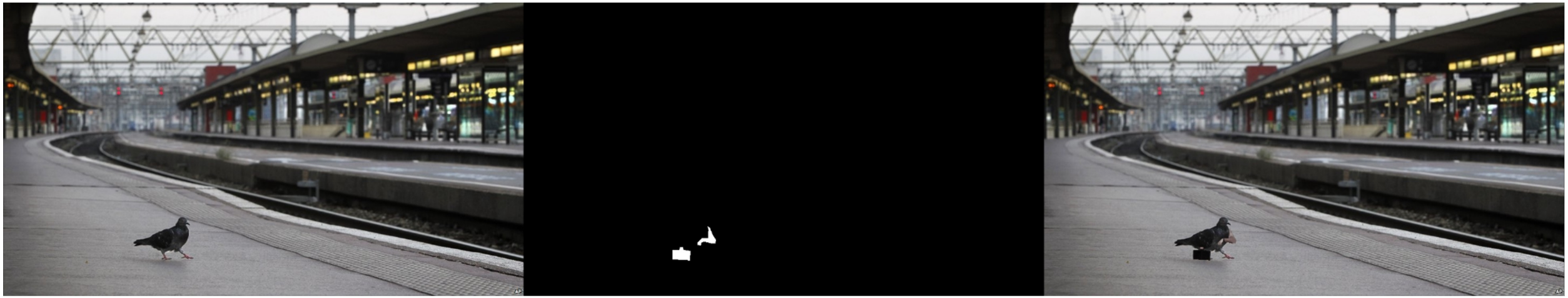
Качество JPEG-50

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

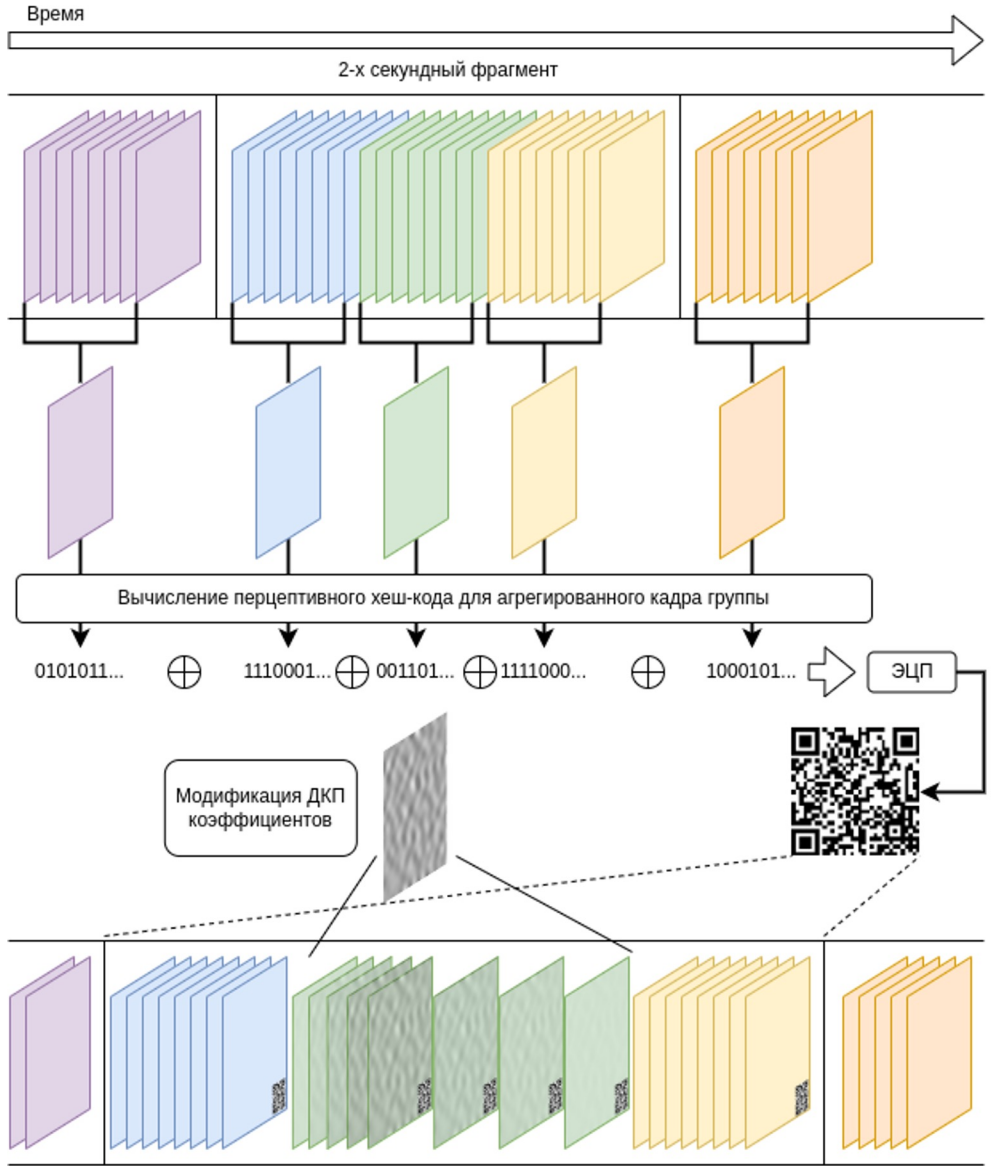
- Набор изображений IMD2020, содержащий тройки изображений:
  - Оригинальное изображение
  - Поддельное изображение (добавление / удаление объектов, изменение фона и т.д.)
  - Маска модификации оригинального изображения
- Имитация недопустимого преобразования:
  - Защита оригинального изображения предлагаемым 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

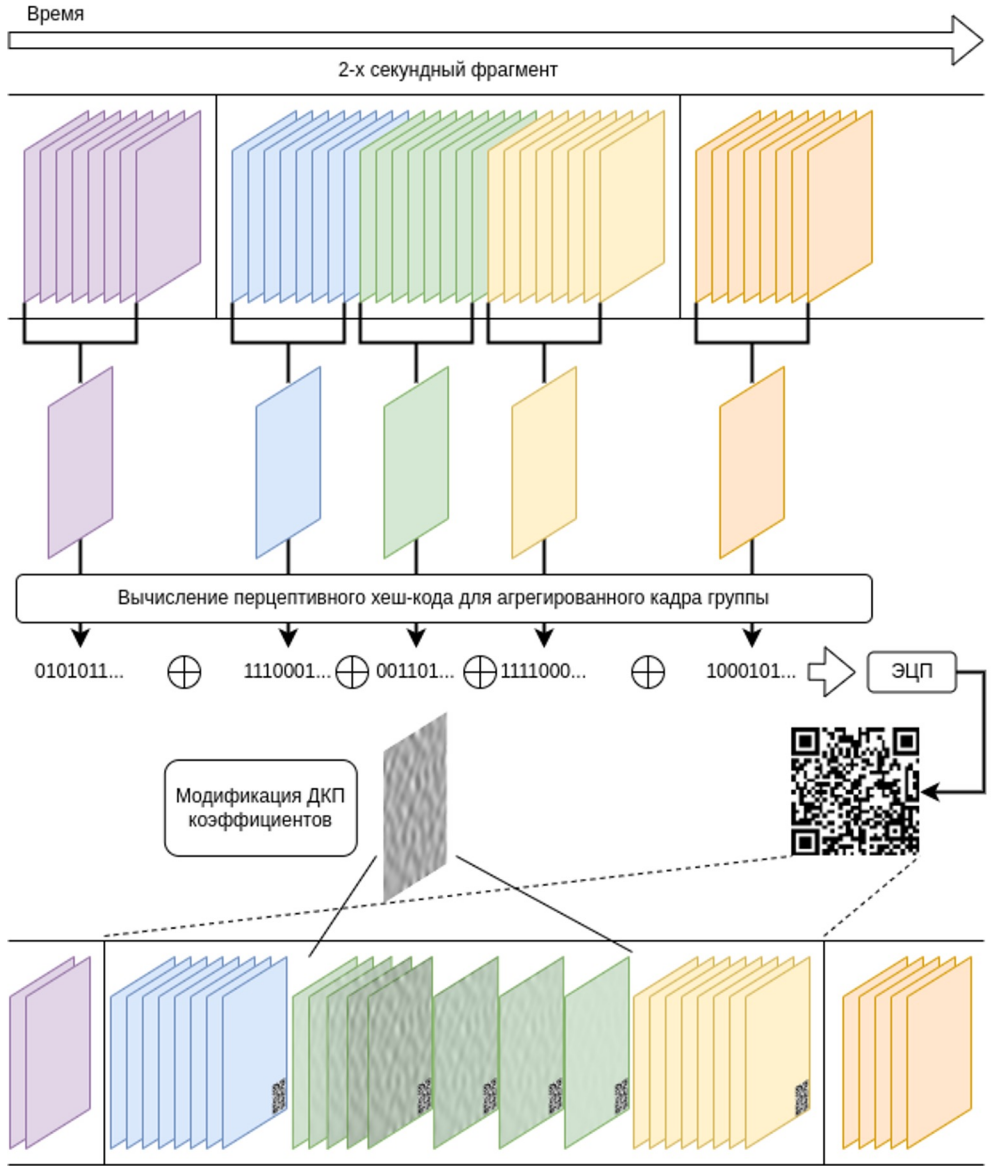
Адаптация алгоритма для защиты видеоконтента

1. Исходное видео разбивается на 2-х секундные фрагменты, фрагменты – на 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

1. Общий QR-код для фрагмента видео, содержащего несколько целых групп кадров
2. Перцептивные хеш-коды групп одного фрагмента последовательно объединяются
3. Добавляются хеш-коды групп из соседних фрагментов: обеспечение временной целостности фрагментов
4. ЭЦП вычисляется по объединенным перцептивным хеш-кодам групп кадров
5. По ЭЦП формируется QR-код, общий для всех кадров в фрагменте



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

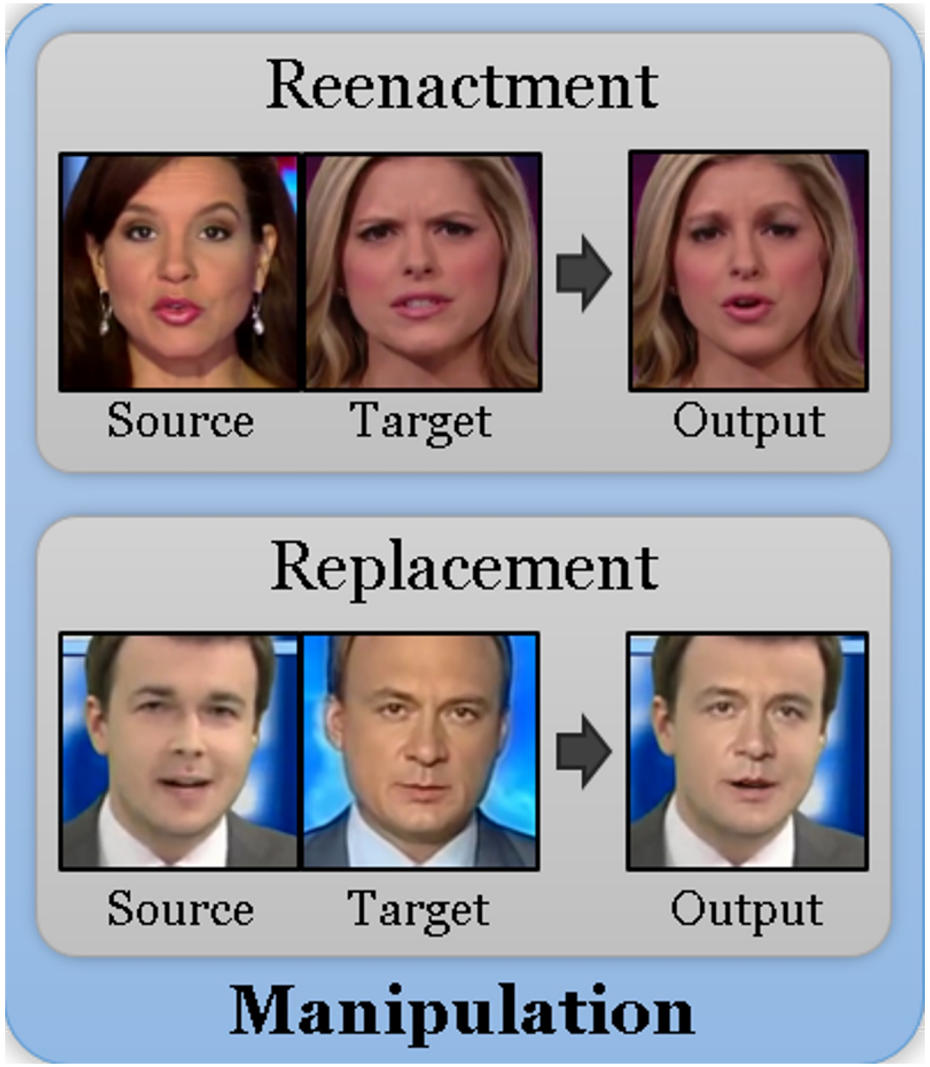
CRF	20	25	30	35	40
Доля фрагментов	0%	0%	0.2%	7.3%	66.2%

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

Тестирование: недопустимые преобразования видео

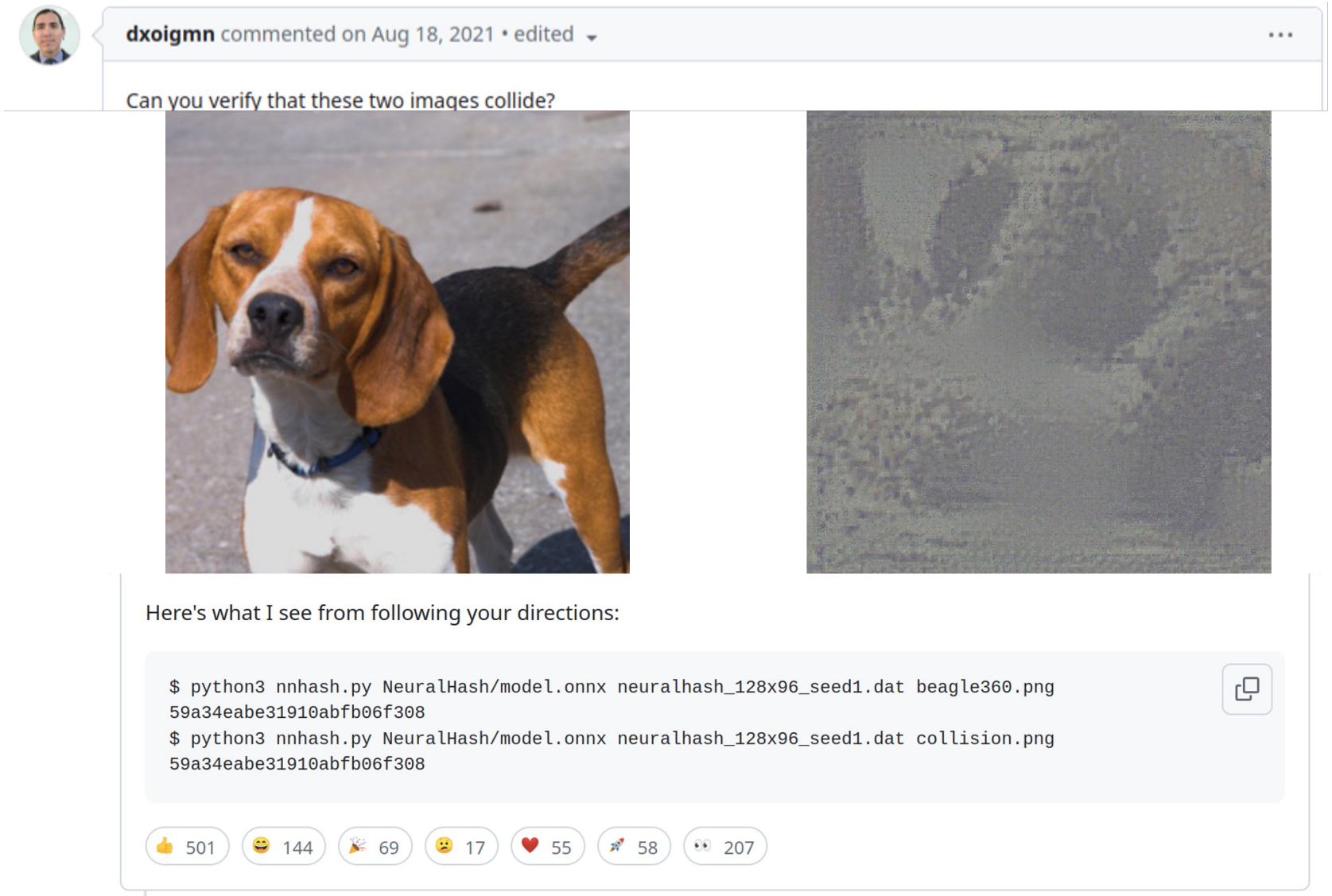
Изменение перцептивного хеш-кода при недопустимых преобразованиях:

- 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

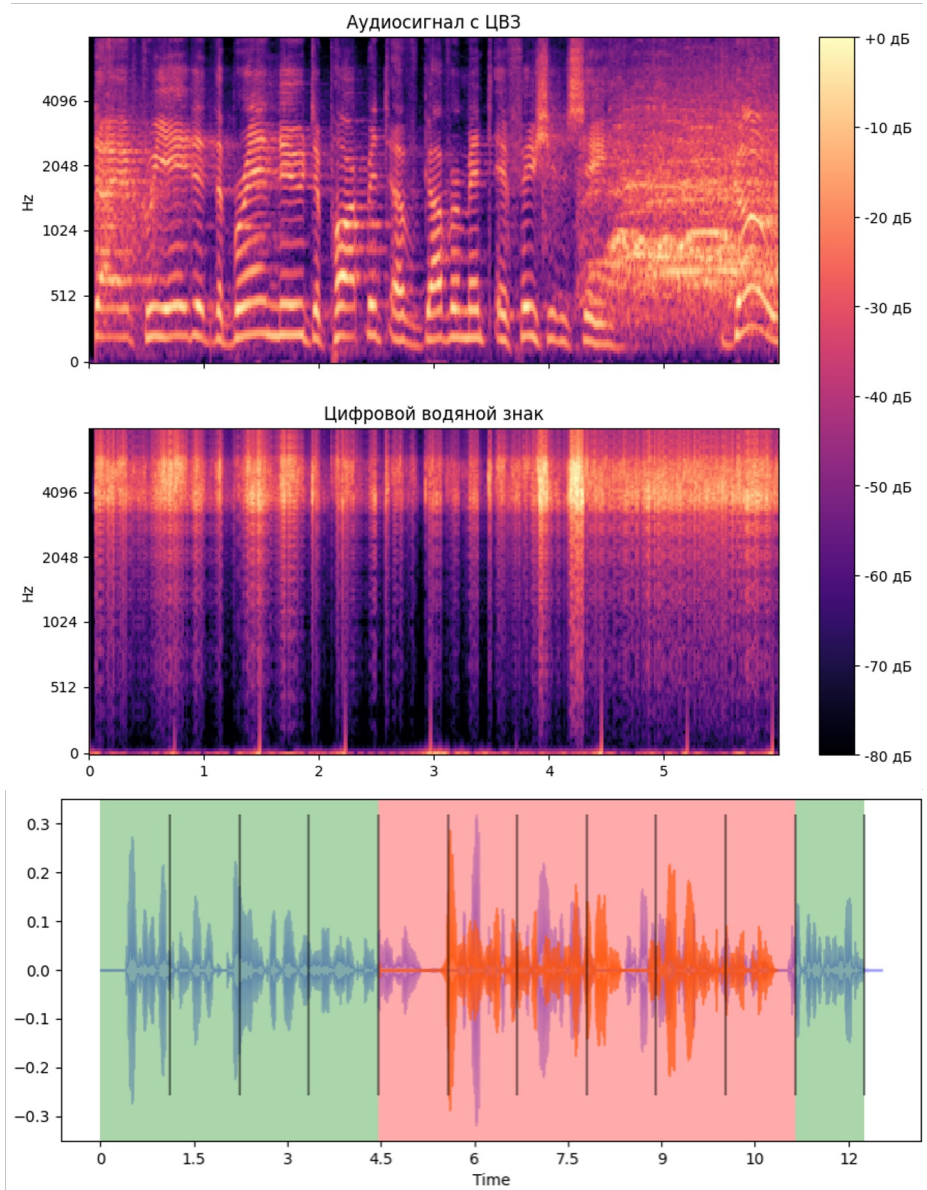
Перцептивные хеш-функции: коллизии



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

Контроль целостности аудиоконтента

- Метод проактивной защиты аудиосигнала:
  - Разбивка на сегменты длительностью 0.5-1.5 с
  - Внедрение ЦВЗ на основе FSVC (Frequency Singular Value Coefficient) признаков аудиосигнала
  - ЦВЗ – суть ЭЦП перцептивного хеш-кода сегмента
- Внедрение в разные частотные полосы:
  - Высокие частоты (6-10 кГц) → ЦВЗ не различим человеческим ухом
  - Низкие частоты (2.5-4 кГц) → выше устойчивость к непреднамеренным атакам: транскодирование, шум, фильтры
- Метод показывает большую точность обнаружения преднамеренных атак (Partial Deepfake, Voice Convertion) на аудиосигнал в сравнении с методом AudioSeal (Meta, SOTA)
- Выше метрики незаметности (SNR, PESQ, STOI) в сравнении с AudioSeal при внедрении водяного знака в высокий и средний частотный диапазоны



Водяные знаки для синтезируемого контента

Технологии синтеза изображений создают новые **угрозы**:

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

Способ противодействия: внедрение *цифровых водяных знаков* при синтезе 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) – (пока) добровольная маркировка сгенерированного контента

Невидимые водяные знаки:

- Gemini (Google)
- Stable Diffusion

Метки на основе метаданных:

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

Внедрение ЦВЗ в синтезируемый ИИ-контент

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

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

2. **Интеграция в процессе синтеза** – внедрение ЦВЗ происходит на этапе генерации изображения

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

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

Тестирование на наборе данных DiffusionDB

Повышение устойчивости водяного знака к геометрическим атакам

Устойчивость в домене Фурье

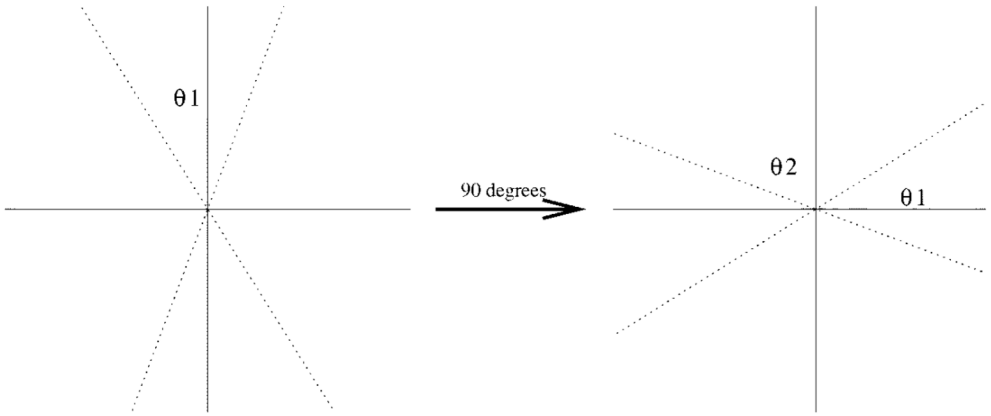
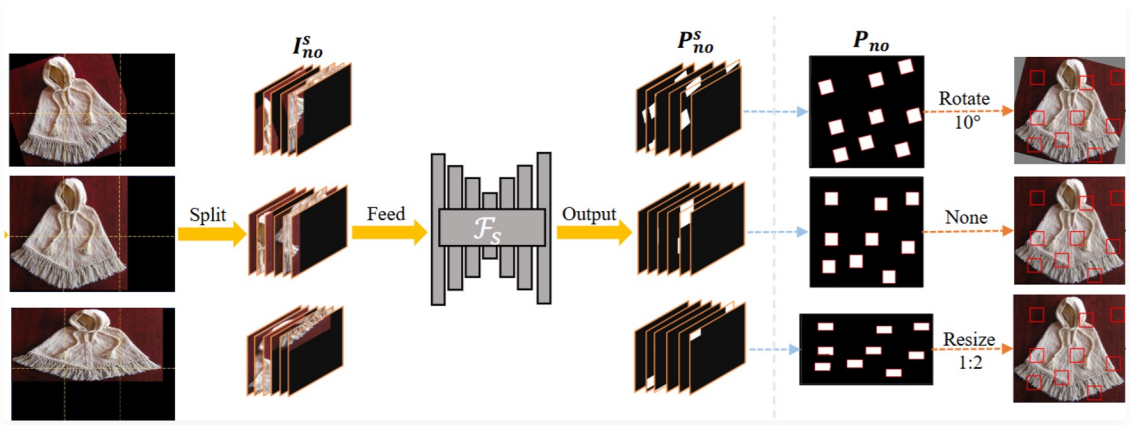


Fig. 2. Template after 90° rotation.

- Устойчивость к сдвигу, обрезке
- Поворот и масштабирование в домене Фурье соответствуют пространственному домену

Сегментирующая нейронная сеть



- Многократное внедрение ЦВЗ
- Сегментирующая нейронная сеть возвращает маску, по которой определяется ориентация областей с ЦВЗ

TrustMark + геометрическая синхронизация

Алгоритм	Емкость	PSNR	SSIM	LPIPS	Gaussian Noise 22	JPEG 50	Cropout 50	Center Crop 50	Rotate 30
Trustmark	100	40.24	0.988	0.002	1.00	1.00	1.00	0.45	0.00
Trustmark DFT circle	100	38.28	0.967	0.033	0.98	0.75	0.01	0.07	0.99
Trustmark DFT lines	100	38.30	0.966	0.018	0.99	0.41	0.01	0.07	0.97
Trustmark DWSF	100	40.46	0.985	0.006	0.98	0.95	0.88	0.98	0.85

Тестирование на наборе данных DiffusionDB

Выводы

- Гонка вооружений
  - непрерывное соревнование между создателями дипфейков и разработчиками детекторов дипфейков
- Детекторы дипфейков (контент без маркировки)
  - неотъемлемая часть общего подхода по борьбе с фальсификациями медиа контента,
  - требуются существенные вложения времени и ресурсов
- Синтезируемый ИИ-контент
  - очень высокое качество, фактически неотличимое от естественного контента
  - упрощает создание контента, в том числе, в злонамеренных целях
- Маркировка
  - для оригинального контента
    - эффективный подход обеспечения доверия контенту
  - для синтезируемого контента
    - факт генерации контента показан явно и не скрывается