DOI: 10.15514/ISPRAS-2026-38(5)-9

CPW: контроль целостности изображений на основе цифровых водяных знаков, перцептивного хеширования и цифровой подписи

[1] *А.Ю. Якушев, ORCID: 0000-0001-6089-6505 <yakushev@ispras.ru>*
[1] *А.А. Акименков, ORCID: 0009-0006-7710-941X <a.akimenkov@ispras.ru>*
[1] *Д.О. Обыденков, ORCID: 0000-0002-9296-6333 <obydenkov@ispras.ru>*
[2] *Х.Н. Абуд, ORCID: 0009-0009-7131-5839 <khaled.abud@graphics.cs.msu.ru>*
[1] *С.Е. Токарев, ORCID:0009-0002-6690-3301 <stokarev@ispras.ru>*
[1] *Ю.В. Маркин, ORCID: 0000-0003-1145-5118 <ustas@ispras.ru>*

[1] *Институт системного программирования им. В.П. Иванникова РАН, 109004, Россия, г. Москва, ул. А. Солженицына, д. 25.*

[2] *Институт искусственного интеллекта МГУ имени М.В. Ломоносова, Россия, 119991, Москва, Ленинские горы, д. 1.*

Аннотация. Визуальный контент, распространяемый через медиаплатформы, подвержен риску семантической подделки, в том числе с применением генеративных моделей искусственного интеллекта. Многие проактивные методы защиты на основе цифровых водяных знаков (ЦВЗ) уязвимы к атакам подделки маркировки из-за отсутствия криптографической связи между маркировкой, содержимым изображения и владельцем закрытого ключа. Предложен метод CPW для контроля целостности изображений, который интегрирует цифровую подпись в процесс внедрения ЦВЗ и связывает содержимое изображения, внедренный аутентификатор и заданный открытый ключ. Ключевой проблемой является необходимость побитово точного извлечения длинных сообщений в условиях допустимых преобразований изображения. Для этого разработана перцептивная хеш-функция EPHash, обеспечивающая устойчивое компактное представление изображения, а цифровая подпись защищается кодом коррекции ошибок и внедряется в изображение целиком либо независимыми фрагментами. Экспериментальная оценка показала, что по сравнению с MetaSeal CPW обеспечивает более высокую незаметность маркировки, а его устойчивость к допустимым преобразованиям зависит от алгоритма внедрения ЦВЗ. Криптографическая привязка подписи к значению EPHash делает вычислительно неосуществимым формирование корректного аутентификатора для другого хеш-значения без доступа к закрытому ключу.

Ключевые слова: цифровые водяные знаки; криптография; электронная цифровая подпись; проактивная защита от подделок; перцептивное хеширование; контроль целостности изображений.

Для цитирования: Якушев А.Ю., Акименков А.А., Обыденков Д.О., Абуд Х.Н., Токарев С.Е., Маркин Ю.В. Контроль целостности изображений на основе цифровых водяных знаков, перцептивного хеширования и цифровой подписи. Труды ИСП РАН, 2026, том 38, вып. 5, стр. 133–154. DOI: 10.15514/ISPRAS-2026-38(5)-9.

Благодарности: Результаты получены с использованием услуг Центра коллективного пользования Института системного программирования им. В.П. Иванникова РАН – ЦКП ИСП РАН.

CPW: image integrity control based on digital watermarking, perceptual hashing and digital signature

[1] *A.Yu. Yakushev, ORCID: 0000-0001-6089-6505 <yakushev@ispras.ru>*
[1] *A.A. Akimenkov, ORCID: 0009-0006-7710-941X <a.akimenkov@ispras.ru>*
[1] *D.O. Obydenkov, ORCID: 0000-0002-9296-6333 <obydenkov@ispras.ru>*
[2] *K.N. Abud,* ORCID: *0009-0009-7131-5839 <khaled.abud@graphics.cs.msu.ru>*
[1] *S.E. Tokarev, ORCID:0009-0002-6690-3301 <stokarev@ispras.ru>*
[1] *Yu.V. Markin, ORCID: 0000-0003-1145-5118 <ustas@ispras.ru>*

[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.*

[2] *MSU Institute for AI, GSP-1, Leninskie Gory, Moscow, 119991, Russia.*

Abstract. Visual content distributed through media platforms is exposed to the risk of semantic forgery, including manipulation using generative artificial intelligence models. Many proactive protection methods based on digital watermarks are vulnerable to watermark forgery attacks because they lack a cryptographic link between the watermark, the image content, and the holder of the private key. We propose CPW, a method for image integrity verification that integrates a digital signature into the watermark embedding process and links the image content, the embedded authenticator, and a specified public key. A key challenge is the need for bit-exact extraction of long messages under allowable image transformations. To address this, we developed the EPHash perceptual hash function, which provides a robust and compact representation of the image, while the digital signature is protected with an error-correcting code and embedded either as a whole or as independent fragments. Experimental evaluation shows that, compared with MetaSeal, CPW provides higher watermark imperceptibility, while its robustness to allowable transformations depends on the watermark embedding algorithm used. The cryptographic binding of the signature to the EPHash value makes it computationally infeasible to generate a valid authenticator for a different hash value without access to the private key.

Keywords: watermarking; cryptography; digital signature; proactive deepfake protection; perceptual hashing; image integrity control.

For citation: Yakushev A.Yu., Akimenkov A.A., Obydenkov D.O., Abud K.N., Tokarev S.E., Markin Yu.V. Image integrity control based on digital watermarking, perceptual hashing and digital signature. Trudy ISP RAN/Proc. ISP RAS, 2026, vol. 38, issue 5, pp. 133-154 (in Russian). DOI: 10.15514/ISPRAS-2026-38(5)-9.

Acknowledgements. The results were obtained using the services of the Ivannikov Institute for System Programming (ISP RAS) Data Center.

1. Введение

Развитие генеративных моделей существенно упростило создание и модификацию визуального контента. Современные подходы позволяют реалистично заменять лица и объекты, изменять фон и выполнять другие семантически значимые преобразования изображений. Пассивные методы выявления таких подделок основаны на поиске артефактов генерации, однако их эффективность может снижаться при появлении новых моделей и способов синтеза [1]. Альтернативой являются методы проактивной защиты, при которых в исходное изображение до его публикации внедряется аутентификатор при помощи цифрового водяного знака (ЦВЗ). Нарушение или потеря внедренного аутентификатора при последующей проверке может свидетельствовать об изменении защищенного содержимого [2].

Несмотря на развитие проактивных методов, многие решения на основе ЦВЗ не предоставляют криптографических гарантий подлинности маркировки. В отсутствие привязки к секретному ключу нарушитель, которому известен алгоритм внедрения, может внедрить собственную маркировку в модифицированное изображение. Использование

цифровой подписи позволяет предотвратить возможность формирования корректного аутентификатора в отсутствие закрытого ключа, однако создает дополнительную техническую проблему: для криптографической верификации необходимо побитово точно извлечь сообщение большого объема, тогда как существующие методы ЦВЗ могут допускать отдельные ошибки при восстановлении полезной нагрузки.

Более того, наличие цифровой подписи в схеме маркировки само по себе не гарантирует защиту от переноса: если внедренный при помощи ЦВЗ аутентификатор является контентно-независимым, злоумышленник может его скопировать на другое изображение, даже не зная секретный ключ. Следовательно, подпись должна иметь привязку к визуальному содержимому. Однако применение стандартной криптографической подписи напрямую к пиксельным данным недопустимо: специфика хранения и передачи визуального контента приводит к тому, что получатель зачастую взаимодействует с данными, подвергшимися так называемым допустимым преобразованиям (сжатие JPEG, фильтрация, изменение разрешения), при применении которых визуальное восприятие не меняется. Эту проблему можно решить с помощью перцептивной хеш-функции, на основе которой вычисляется криптографическая подпись. Перцептивный хеш-код должен быть устойчив к допустимым преобразованиям и изменяться при недопустимых модификациях изображения. При этом важно, чтобы его устойчивость была высокой, так как для успешной криптографической верификации перцептивный хеш-код должен сохранять значения отдельных битов при допустимых искажениях.

В данной работе предложен метод CPW (Cryptographic Perceptual Watermarking), интегрирующий цифровую подпись в процесс внедрения ЦВЗ. Для криптографической привязки подписи к содержимому изображения была разработана перцептивная хеш-функция EPHash, при расчете которой за счет контролируемой модификации низкочастотных коэффициентов дискретного косинусного преобразования (ДКП) формируется «прибавка» к изображению, повышающая устойчивость ЦВЗ. Для увеличения доступной емкости и повышения надежности восстановления подписи предложено независимо встраивать фрагменты кодового слова в четыре квадранта изображения и применять код коррекции ошибок БЧХ.

2. Обзор литературы

2.1 Методы проактивного контроля целостности изображений

Методы контроля целостности цифровых изображений можно разделить на пассивные и проактивные. В основе пассивных методов – выявление признаков генерации или редактирования непосредственно в исследуемом изображении: статистические аномалии, особенности частотного спектра, несогласованность шумовых характеристик и другие артефакты. Детекторы, основанные на сверточных нейронных сетях, способны обнаруживать изображения, созданные генеративными моделями, в том числе при обучении на ограниченном наборе моделей синтеза [3]. Однако точность работы таких детекторов снижается при изменении способа генерации, последующей обработке изображения и целенаправленных состязательных воздействиях [4]. Применение пассивных методов, как правило, позволяет оценить вероятность модификации, но получить сведения об источнике или исходном содержимом контента с их помощью нельзя.

Проактивные методы предполагают внедрение идентификатора в изображение до его публикации. Наиболее распространенным подходом является внедрение цифрового водяного знака (ЦВЗ). В классических алгоритмах маркировка формируется посредством модификации пространственных или частотных коэффициентов изображения, например, коэффициентов дискретного косинусного или вейвлет-преобразования [5]. При проектировании таких алгоритмов необходимо обеспечить компромисс между

незаметностью, устойчивостью к преобразованиям и емкостью ЦВЗ. Увеличение емкости или устойчивости, как правило, требует более существенной модификации изображения и может приводить к появлению визуально различимых артефактов.

Современные методы внедрения ЦВЗ преимущественно основаны на нейросетевой архитектуре «кодировщик – декодировщик». В методе HiDDeN [6] кодировщик по входному изображению и битовому сообщению формирует маркированное изображение, а декодировщик восстанавливает сообщение после прохождения маркированным изображением слоя, моделирующего сжатие, размытие, кадрирование и другие преобразования. Метод MBRS [7] повышает устойчивость к JPEG-сжатию за счет совместного использования реального и дифференцируемо аппроксимированного JPEG-сжатия в процессе обучения. В работе VideoSeal [8] предложена масштабируемая схема маркировки изображений и видео, предусматривающая обучение на наборе преобразований и последующую адаптацию к временной структуре видеоданных. PixelSeal [9] развивает данный подход посредством состязательного обучения и адаптации к изображениям высокого разрешения. В работе ChunkySeal [10] показана возможность увеличения емкости нейросетевого ЦВЗ до 1024 бит при сохранении качества маркированного изображения и устойчивости к исследованным преобразованиям. Эти результаты делают возможным внедрение криптографических аутентификаторов, размер которых существенно превышает длину идентификаторов, обычно используемых в системах маркировки.

Отдельное внимание следует уделить проактивным методам обнаружения семантического редактирования. В работе Identity Watermarking [11] предложена система контроля личности на изображении: в векторное представление лица внедряется информация, позволяющая сопоставить опубликованное изображение с исходной личностью и обнаружить подмену лица. В SepMark [12] используются два декодировщика с различными профилями устойчивости: один предназначен для восстановления информации об источнике, а второй – для выявления модификаций. EditGuard [2] сочетает устойчивую маркировку для подтверждения происхождения с хрупким ЦВЗ, предназначенным для локализации измененных областей. В OmniGuard [13] проактивная защита дополнена пассивным анализом изображения, что позволяет выявлять как локальные, так и глобальные модификации.

Рассмотренные методы демонстрируют возможность обнаружения и локализации различных изменений в изображениях, и, как правило, корректность извлеченного аутентификатора интерпретируется как подтверждение подлинности маркировки. Однако если аутентификатор не имеет «криптографической связи» с содержимым изображения и владельцем секретного ключа, устойчивое извлечение само по себе не исключает неавторизованное внедрение, перенос или повторное использование ЦВЗ. Поэтому задача контроля целостности не сводится только к обеспечению устойчивости декодировщика: требуется проверяемая связь между изображением, внедренным сообщением и владельцем закрытого ключа.

2.2 Подделка и криптографическая защита цифровых водяных знаков

При анализе безопасности подходов к маркировке необходимо различать удаление и подделку маркировки. В первом случае злоумышленник стремится воспрепятствовать обнаружению ЦВЗ, тогда как во втором – добиться ложного подтверждения происхождения или подлинности изображения. Возможность такой атаки была продемонстрирована в работе [14]: ЦВЗ, оцененный как прибавка к изображению, может быть перенесен с маркированного изображения на другое без знания алгоритма внедрения и секретного ключа. Следовательно, обнаружение маркировки само по себе не доказывает, что она была внедрена в данное изображение его автором.

Одним из первых подходов к решению этой проблемы стали ЦВЗ для открытой аутентификации, объединяющие хеширование изображения и цифровую подпись. Так, в

схеме Wong и Memon [15] аутентификационные значения блоков изображения подписываются закрытым ключом и внедряются в соответствующие блоки. Проверка целостности изображения выполняется с помощью открытого ключа. Однако последующие исследования показали, что криптографическая стойкость используемой подписи еще не гарантирует безопасность всей системы. Авторы работы [16] продемонстрировали возможность атак вырезания и вставки, переноса и составления поддельных изображений из корректно аутентифицированных блоков. Для противодействия таким атакам подписываемые значения должны учитывать положение блоков, идентификатор изображения и связи между его частями.

В дальнейшем развитие получили полухрупкие схемы, в которых цифровая подпись связывается не с точным пиксельным представлением, а с устойчивыми признаками изображения. В работе [17] контентно-зависимое представление изображения формируется посредством квантования характеристик блоков, подписывается закрытым ключом и внедряется в изображение с применением кода коррекции ошибок. Экспериментально была продемонстрирована устойчивость ЦВЗ к JPEG-сжатию высокого качества; устойчивость к более широкому набору допустимых преобразований не оценивалась.

Проблема криптографической защиты сохраняется и для современных нейросетевых ЦВЗ. Авторы работы [18] показывают, что нейросетевые подходы, в которых процесс формирования маркировки не зависит от секретного ключа, остаются уязвимыми к атакам копирования ЦВЗ. В работе [19] формализуются требования к защищенным системам подтверждения происхождения контента, выделяя устойчивость, невозможность подделки и публичную проверку. Авторы показывают теоретическую совместимость этих свойств, но отмечают сложность их одновременного обеспечения в практически применимой системе ЦВЗ.

Одним из современных методов контентно-зависимой криптографической маркировки является MetaSeal [20]. В нем текстовое семантическое описание изображения подписывается с помощью ECDSA, после чего сообщение и подпись совместно кодируются в QR-код, который незаметно внедряется в частотное представление изображения с помощью обратимой нейронной сети (Invertable Neural Network). Связь подписи с содержанием препятствует простому переносу аутентификатора на семантически отличающееся изображение. Однако надежность проверки определяется полнотой используемого описания: слишком общее представление может допускать повторное использование подписи, тогда как подробное описание увеличивает полезную нагрузку и может привести к снижению устойчивости к допустимым преобразованиям изображения.

Таким образом, цифровая подпись предотвращает возможность формирования нового корректного аутентификатора без закрытого ключа, но сама по себе не исключает перенос уже сформированной маркировки. Для защиты от подделки аутентификатор должен быть связан с содержимым конкретного изображения, не изменяться после допустимых и изменяться после недопустимых преобразований.

2.3 Перцептивные хеш-функции для изображений

Перцептивное хеширование тесно связано с аутентификацией и контролем целостности. Алгоритмы перцептивного хеширования формируют компактную сигнатуру на основе устойчивых перцептивно значимых характеристиках изображения. Визуально сходным изображениям должны соответствовать близкие хеш-коды, а различным – удаленные друг от друга. В отличие от криптографических хеш-функций, для которых любая, даже незначительная модификация входных данных должна приводить к существенному изменению хеш-кода, перцептивное хеширование предусматривает устойчивость к ряду преобразований, например, сжатию, изменению размера, размытию и незначительной цветовой коррекции [21]. Для бинарных хеш-кодов степень различия обычно оценивается с

помощью расстояния Хэмминга, тогда как для вещественных представлений могут использоваться другие метрики расстояния. Вычисленный для проверяемого изображения хеш-код сравнивается с доверенным эталонным значением, которое может храниться на внешнем ресурсе. Если расстояние между хеш-кодами не превышает установленного порога, изображение считается соответствующим эталону с точностью до преобразований, к которым применяемая перцептивная хеш-функция устойчива.

Широкое распространение получили такие алгоритмы перцептивного хеширования, как aHash, pHash, wHash [22] и PDQ [23]. Они различаются используемыми признаками изображения, длиной кода и устойчивостью к отдельным преобразованиям, но основаны на общей модели квантования по заданному порогу. В aHash изображение уменьшается до размера 8×8 и преобразуется в оттенки серого, после чего значения яркости пикселей сравниваются со средним значением: если значение больше среднего, соответствующий бит хеш-кода выбирается равным 1, если меньше – 0. Алгоритм wHash использует 64 низкочастотных коэффициента вейвлет-преобразования и формирует бинарный код посредством их сравнения с медианным значением. В pHash аналогичная операция выполняется над коэффициентами ДКП. PDQ развивает идею алгоритма pHash: вводится дополнительная двухпроходная фильтрация изображения, основанная на локальном усреднении значений яркости, а также выбираются 256 коэффициентов ДКП вместо 64, что позволяет получить хеш-код соответствующей длины.

Порог принятия решения определяет компромисс между устойчивостью к допустимой обработке и возможностью различить изображения по значениям хеш-кодов. В задачах поиска дубликатов полное совпадение хеш-кодов обычно не требуется: изображения считаются похожими, если расстояние между их хеш-кодами не превышает установленного значения. Однако при непосредственном использовании перцептивного хеш-кода в качестве подписываемого сообщения изменение даже одного бита приводит к отклонению цифровой подписи при проверке. Поэтому необходимо разработать перцептивную хеш-функцию с высокой побитовой устойчивостью к допустимым преобразованиям изображения, но в то же время сохраняющей возможность отличить разные изображения по значению хеш-кодов, а также обладающей чувствительностью к семантическим изменениям изображения.

3. Постановка задачи

Целью работы является разработка метода обеспечения контроля целостности изображений. Метод должен предоставлять возможность публично проверить, что полученное изображение соответствует изображению, ранее аутентифицированному владельцем закрытого ключа, с точностью до заданного множества допустимых преобразований. Верификация должна завершаться успешно для изображений, подвергнутых допустимым преобразованиям, и давать отрицательный результат при рассматриваемых недопустимых модификациях. Отрицательный результат верификации означает невозможность подтвердить соответствие изображения аутентифицированному оригиналу в рамках заданной модели.

3.1 Модель нарушителя

Предполагается, что нарушитель знает алгоритм обеспечения контроля целостности, включая все используемые преобразования и параметры, за исключением закрытого криптографического ключа стороны, выполнившей аутентификацию изображения. Это соответствует принципу Керкгоффса: алгоритм и параметры системы считаются известными, а единственным секретным параметром является закрытый криптографический ключ. Нарушителю доступны открытый ключ и процедура верификации, которую он может запускать многократно, но недоступны закрытый ключ и исходные немаркированные изображения. Его целью является получение семантически измененного изображения,

которое успешно прошло бы проверку целостности и подлинности относительно открытого ключа создателя. Для достижения этой цели нарушитель может предпринимать следующие действия:

1. Извлечение и перенос водяного знака. Злоумышленник пытается выделить встроенный аутентификатор из маркированного изображения и внедрить его в другое изображение (например, поддельное) с тем, чтобы последнее прошло проверку.
2. Модификация изображения с сохранением подписываемого хеш-кода. Злоумышленник пытается внести семантически значимые изменения, например, заменить лицо, таким образом, чтобы значение перцептивного хеш-кода осталось неизменным, а встроенный аутентификатор продолжал извлекаться. Поскольку атака строится относительно заданного изображения и его хеш-значения, она соответствует построению второго прообраза для перцептивной хеш-функции.
3. Формирование поддельного аутентификатора. Злоумышленник пытается сформировать новое хеш-значение и соответствующую ему подпись, успешно проверяемую открытым ключом создателя. При отсутствии закрытого ключа возможность такой атаки определяется криптографической стойкостью используемой схемы цифровой подписи.

Дополнительно рассматриваются атаки удаления ЦВЗ. К ним относятся оценка и удаление компоненты ЦВЗ посредством фильтрации или совместного анализа нескольких маркированных изображений, а также повторное внедрение, направленное на повреждение или перезапись исходного аутентификатора. Такие воздействия обычно приводят не к ложному принятию поддельного изображения, а к отказу в проверке маркированного экземпляра. Они характеризуют устойчивость извлечения ЦВЗ, но не криптографическую невозможность подделки аутентификатора. Внедрение нарушителем маркировки, сформированной с собственной парой ключей, также не считается успешной подделкой, если изображение проверяется относительно открытого ключа исходного создателя.

3.2 Допустимые и недопустимые преобразования

Для определения целостности задаются два множества преобразований: допустимые преобразования, после которых изображение должно успешно проходить верификацию, и недопустимые преобразования, применение которых должно приводить к отрицательному результату проверки. Такое разделение определяется принятой политикой верификации и заявленной областью работоспособности метода, а не только визуальной заметностью внесенных изменений.

Допустимыми преобразованиями считаются следующие:

- Изменение формата хранения, включая сжатие с потерями по алгоритму JPEG, при условии, что коэффициент качества составляет не менее 40;
- Фильтрация изображения;
- Аддитивный гауссовский шум;
- Изменение яркости и контрастности;
- Масштабирование (изменение разрешения) с сохранением геометрического соотношения сторон изображения.

Перечисленные преобразования моделируют типичные операции, возникающие при хранении, передаче и публикации изображений на медиаплатформах. Конкретные диапазоны их параметров устанавливаются в экспериментальной части работы. **Недопустимыми преобразованиями** признаются любые изменения, которые искажают семантическое

содержание изображения или могут быть использованы для создания подделки. К ним относятся:

- Вставка или удаление объектов на изображении;
- Замена лица с использованием моделей искусственного интеллекта;
- Изменение фона или других содержательно значимых областей;
- Регенерация изображения с помощью генеративных нейронных сетей.

В составе недопустимых преобразований отдельную группу образуют неподдерживаемые геометрические операции, нарушающие синхронизацию маркировки:

- Обрезка части изображения;
- Поворот изображения, а также зеркальное отражение;
- Изменение пропорций (растяжение или сжатие по одной оси), не являющееся равномерным масштабированием.

Перечисленные преобразования определяют рассматриваемую в работе модель обработки изображений. Разрабатываемый метод должен сохранять положительный результат верификации после каждого допустимого преобразования в пределах установленных параметров и возвращать отрицательный результат после рассматриваемых семантических модификаций и неподдерживаемых геометрических преобразований. Если явно не указано иное, преобразования применяются по отдельности. Для воздействий и диапазонов параметров, не включенных в данную модель, работоспособность метода не гарантируется.

3.3 Требования к разрабатываемому методу

С учетом принятой модели нарушителя и рассматриваемых преобразований сформулируем требования, которым должен удовлетворять разрабатываемый метод.

- Побитовая устойчивость перцептивного хеш-кода. Хеш-значение, вычисляемое при верификации маркированного изображения после допустимого преобразования, должно полностью совпадать с подписанным хеш-значением.
- Чувствительность перцептивного хеш-кода к семантическим модификациям. Вставка, удаление или замена содержательно значимых областей изображения должна приводить к изменению хеш-кода и, следовательно, к отрицательному результату проверки цифровой подписи.
- Точное извлечение аутентификатора. Цифровая подпись должна восстанавливаться без битовых ошибок после допустимой обработки изображения.
- Невозможность подделки. Формирование корректного аутентификатора для нового хеш-значения без знания закрытого ключа должно быть вычислительно неосуществимым при условии стойкости используемой схемы цифровой подписи.
- Незаметность маркировки. Внедрение аутентификатора не должно приводить к существенному ухудшению качества изображения. В качестве количественного требования устанавливается значение PSNR не менее 40 дБ.
- Публичная слепая верификация. Для проверки целостности должны использоваться только исследуемое изображение, открытый ключ и общедоступные алгоритмы, модели и параметры системы. Доступ к исходному немаркированному изображению и закрытому ключу не требуется.

На уровне всей системы должны выполняться два основных условия. Маркированное изображение после допустимого преобразования в пределах установленных параметров должно сохранять положительный результат верификации. Изображение после рассматриваемой семантической модификации или неподдерживаемого геометрического

преобразования должно отклоняться. При этом отрицательный результат означает, что целостность и подлинность изображения относительно указанного открытого ключа не могут быть подтверждены, но сам по себе не устанавливает характер выполненного преобразования. В следующих разделах описывается предложенное решение, включающее модификацию ДКП-коэффициентов для повышения побитовой устойчивости перцептивного хеш-кода, формирование цифровой подписи и ее внедрение в изображение в качестве полезной нагрузки ЦВЗ.

4. Описание метода

В данном разделе описывается предлагаемый метод контроля целостности изображений CPW (Cryptographic Perceptual Watermarking), общая схема которого представлена на рис. 1. Метод включает формирование устойчивого перцептивного представления изображения, создание криптографического аутентификатора, его помехоустойчивое кодирование и внедрение в изображение в виде цифрового водяного знака. На этапе маркирования для исходного изображения вычисляется перцептивный хеш-код, после чего выполняется предварительная модификация коэффициентов ДКП, направленная на повышение его побитовой устойчивости. Полученный хеш-код преобразуется криптографической хеш-функцией и подписывается закрытым ключом создателя. Подпись кодируется с применением кода коррекции ошибок и внедряется в изображение. При верификации аутентификатор извлекается из исследуемого изображения, повторно вычисляется его перцептивный хеш-код и выполняется проверка подписи с использованием открытого ключа создателя. Таким образом, результат верификации определяется одновременно сохранностью встроенного аутентификатора и его соответствием содержанию исследуемого изображения.

4.1 Перцептивная хеш-функция с повышенной устойчивостью

Проверка цифровой подписи требует полного побитового совпадения хеш-кодов, подписанного и повторно вычисленного. Для выполнения этого требования разработан алгоритм EPHash (Enhanced Perceptual Hash), основанный на квантовании низкочастотных коэффициентов дискретного косинусного преобразования (ДКП). На этапе маркирования вычисление хеш-кода дополняется однократной модификацией изображения, увеличивающей расстояние выбранных коэффициентов от порогов квантования. При верификации такая модификация не выполняется.

Для вычисления хеш-кода изображение переводится в градации серого и масштабируется к квадратному разрешению со стороной

$$N = n_b \cdot h_s \cdot h_{ff},$$

где n_b – число блоков по каждой оси, h_s – количество отбираемых низкочастотных коэффициентов ДКП по каждой оси, h_{ff} – масштабный коэффициент, определяющий размер блока ДКП и долю используемой низкочастотной области.

Полученное изображение делится на $n_b \times n_b$ непересекающихся блоков размером $h_s h_{ff} \times h_s h_{ff}$, для каждого из которых вычисляется двумерное ДКП. Из матрицы коэффициентов C выбирается область

$$\Omega = \{(u,v) \mid 2 \le u \le h_s + 1; 2 \le v \le h_s + 1\},$$

не включающая коэффициенты первой строки и первого столбца. Для выбранных коэффициентов каждого блока вычисляется медиана (m), после чего формируются биты

$$b_{u,v} = \begin{cases} 1, & C_{u,v} \ge m, \\ 0, & C_{u,v} < m. \end{cases}$$

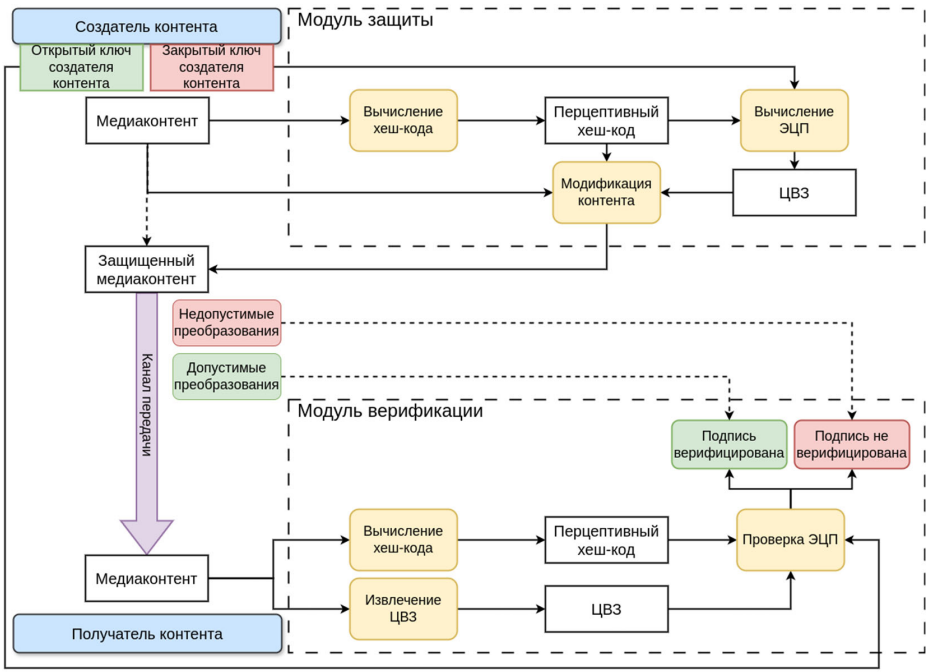

Рис. 1. Схема работы предлагаемого метода CPW обеспечения контроля целостности оригинальных изображений

Fig. 1. Schematic diagram of the proposed CPW method for original image integrity verification

Биты всех блоков объединяются в фиксированном порядке в строку H. Длина полученного перцептивного хеш-кода составляет

$$|H| = (n_b h_s)^2.$$

Коэффициенты, расположенные вблизи медианы, могут пересечь порог квантования после процедуры внедрения ЦВЗ или допустимых преобразований, возникающих при передаче изображения. Поэтому при маркировании для каждого блока создается зазор устойчивости δ_{min}. Если расстояние выбранного коэффициента от медианы меньше δ_{min}, коэффициент модифицируется в соответствии со сформированным битом:

$$\tilde{C}_{u,v} = \begin{cases} \max\left(C_{u,v}, m + \delta_{min}\right), & b_{u,v} = 1, \\ \min\left(C_{u,v}, m - \delta_{min}\right), & b_{u,v} = 0. \end{cases}$$

Модификация выполняется в направлении, соответствующем ранее сформированному биту, поэтому положение коэффициента относительно исходного порога m не изменяется. Значение δ_{min} подбирается отдельно для каждого блока методом бинарного поиска так, чтобы максимизировать расстояние коэффициентов от медианы при сохранении PSNR между исходным и восстановленным блоками не ниже заданного порога T_{PSNR}. После обратного ДКП модифицированные блоки объединяются. Полученная разность между исходным и модифицированным изображениями интерполируется до исходного разрешения и добавляется к яркостному каналу исходного изображения. В результате формируется модифицированное изображение I_m, для которого должно выполняться условие:

$$\mathrm{EPHash}(I_m) = \mathrm{EPHash}(I) = H.$$

4.2 Формирование аутентификатора

Перцептивный хеш-код характеризует содержание изображения, но сам по себе не обеспечивает его аутентичность, поскольку может быть вычислен любым участником системы. Для криптографической привязки хеш-кода к владельцу закрытого ключа используется алгоритм электронной цифровой подписи ECDSA. Сначала для перцептивного хеш-кода H вычисляется криптографический хеш-код

$$h = \mathrm{SHA256}(H),$$

после чего с использованием закрытого ключа K_{priv} формируется цифровая подпись

$$(r, s) = \mathrm{ECDSA.Sign}(K_{priv}, h).$$

В работе используется кривая secp256r1, обеспечивающая криптографическую стойкость 128 бит. Компоненты r и s представляются в виде двух целых чисел фиксированной длины по 256 бит и объединяются в битовую строку S длиной 512 бит. Такую подпись практически невозможно подделать за обозримое время (с использованием классических вычислительных систем). Перенос извлеченной подписи на изображение с другим значением EPHash приведет к отрицательному результату ее верификации.

4.3 Внедрение аутентификатора в виде ЦВЗ

Для проверки цифровой подписи необходимо восстановить все ее биты без ошибок. Перед внедрением 512-битная подпись S кодируется с использованием двоичного кода БЧХ [24], преобразующего ее в кодовое слово длиной 1023 бита. Для согласования с емкостью используемых методов ЦВЗ к кодовому слову добавляется нулевой бит дополнения:

$$P = \mathrm{BCH.Encode(S)} \parallel 0, |\mathrm{P}| = 1024.$$

Используемый код позволяет исправить до 57 битовых ошибок в пределах кодового слова. В работе рассматриваются две стратегии внедрения полученной последовательности. В первой стратегии вся 1024-битная полезная нагрузка P внедряется в модифицированное изображение I_m с использованием метода ChunkySeal, поддерживающего соответствующую полезную нагрузку:

$$I_w = \mathrm{Embed}_{ChunkySeal}(I_m, P)$$

Во второй стратегии изображение разделяется на четыре непересекающихся квадранта $I_m^{(1)}, \dots, I_m^{(4)}$, а последовательность P – на четыре последовательных фрагмента одинаковой длины:

$$P = P_1 \parallel P_2 \parallel P_3 \parallel P_4, \qquad |P_i| = 256.$$

Каждый фрагмент независимо внедряется в соответствующий квадрант:

$$I_w^{(i)} = \mathrm{Embed}\left(I_m^{(i)}, P_i\right), i = 1 \dots 4,$$

после чего маркированные квадранты объединяются в итоговое изображение:

$$I_w = \mathrm{Merge}\left(I_w^{(1)}, \dots, I_w^{(4)}\right).$$

Такая стратегия увеличивает суммарную полезную нагрузку до 1024 бит, сохраняя длину сообщения, независимо внедряемого в каждый квадрант, равной 256 битам. В экспериментальной части для ее реализации рассматриваются PixelSeal, VideoSeal и MBRS.

4.4 Верификация целостности и подлинности изображения

Для выполнения верификации получатель должен располагать исследуемым изображением I', открытым ключом K_{pub} и общедоступными параметрами метода. Доступ к исходному немаркированному изображению и закрытому ключу не требуется. Проверка включает следующие шаги:

1. Из изображения извлекается полезная нагрузка P'. При использовании квадрантной стратегии сообщения извлекаются из каждого квадранта независимо и объединяются в исходном порядке.
2. Из последовательности P' удаляется последний бит дополнения, после чего 1023-битное кодовое слово декодируется кодом БЧХ:

   $$S' = \mathrm{BCH.Decode(P')},$$

   В результате исправляются битовые ошибки в пределах корректирующей способности кода и восстанавливается 512-битная подпись S'. Если извлечение или декодирование завершается неудачей, изображение не проходит верификацию.
3. Для исследуемого изображения вычисляются перцептивный и криптографический хеш-коды:

   $$H' = \mathrm{EPHash(I')}, \qquad h' = \mathrm{SHA256}(H').$$
4. Восстановленная подпись проверяется с использованием открытого ключа создателя:

   $$v = \mathrm{ECDSA.Verify}\left(K_{pub}, h', S'\right), v \in \{true, false\}.$$

Если все этапы извлечения и декодирования завершились успешно и $v = true$, целостность и подлинность изображения относительно указанного открытого ключа считаются подтвержденными в пределах принятой политики допустимых преобразований. Положительный результат означает, что извлеченная подпись соответствует перцептивному хеш-коду исследуемого изображения.

Если $v = false$, либо извлечение или декодирование завершилось неудачей, выносится вердикт «верификация не пройдена». Причиной могут быть семантическая модификация изображения, неподдерживаемое геометрическое преобразование, допустимое воздействие за пределами установленного диапазона, отсутствие или повреждение ЦВЗ либо попытка подделки аутентификатора. Отрицательный результат означает, что целостность и подлинность изображения не могут быть подтверждены, но сам по себе не позволяет установить характер выполненного преобразования.

5. Эксперименты и результаты

Цель данного раздела – экспериментально оценить эффективность и границы применимости предложенного метода контроля целостности изображений. Экспериментальная оценка проводится по трем направлениям. Во-первых, исследуется побитовая устойчивость EPHash к допустимым преобразованиям в сравнении с существующими перцептивными хеш-функциями. Во-вторых, сопоставляются две стратегии внедрения 1024-битной полезной нагрузки: непосредственное внедрение и независимое внедрение фрагментов сообщения в четыре квадранта изображения. В-третьих, проводится сравнение предложенного метода с MetaSeal по незаметности маркировки, устойчивости к допустимым преобразованиям и чувствительности к семантическим модификациям.

5.1 Параметры алгоритма

Для обеспечения воспроизводимости результатов все эксперименты выполнялись с фиксированными параметрами вычисления перцептивного хеш-кода, модификации изображения, формирования криптографического аутентификатора и внедрения ЦВЗ. Для алгоритма EPHash использовались следующие параметры:

- $n_b = 2$ – количество блоков по каждой оси;
- $h_s = 16$ – количество отбираемых коэффициентов ДКП по каждой оси блока;

- $h_{ff} = 16$ – масштабный коэффициент, определяющий размер блока ДКП и долю используемой низкочастотной области;
- $T_{PSNR} = 45$ дБ – минимальное значение PSNR между исходным блоком и его реконструкцией после модификации коэффициентов ДКП, используемое как ограничение при подборе δ_{min};
- Количество итераций бинарного поиска при подборе δ_{min} – 10.

При выбранных размерах сторона блока составляет $16 \times 16 = 256$ пикселей, разрешение, к которому приводится изображение – $16 \times 16 \times 2 = 512$. Общая длина перцептивного хеш-кода составляет $(n_b h_s)^2 = (2 \cdot 16)^2 = 1024$ бита.

Для формирования криптографического аутентификатора применялись хеш-функция SHA-256 и алгоритм цифровой подписи ECDSA на эллиптической кривой secp256r1. Подпись представлялась в виде конкатенации двух 256-битных компонентов r и s, поэтому ее длина составляла 512 бит. Перед внедрением подпись кодировалась двоичным кодом БЧХ, формирующим 1023-битное кодовое слово и исправляющим до 57 битовых ошибок. После добавления нулевого бита дополнения длина полезной нагрузки составляла 1024 бита.

В стратегии непосредственного внедрения вся 1024-битная полезная нагрузка встраивалась в изображение методом ChunkySeal. В квадрантной стратегии она разделялась на четыре фрагмента по 256 бит, которые независимо внедрялись в четыре квадранта изображения. Для реализации этой стратегии использовались методы PixelSeal, VideoSeal и MBRS.

5.2 Сравнение перцептивных хеш-функций

Прежде чем перейти к оценке предложенного метода, целесообразно сопоставить классические перцептивные хеш-функции, чтобы установить базовый уровень устойчивости к допустимым преобразованиям. В качестве эталонов рассматриваются четыре широко применяемые на практике хеш-функции: aHash, pHash, wHash и PDQ.

Результаты измерения устойчивости приведены на рис. 2. Для тестирования взята выборка из 1000 изображений набора данных MSCOCO [25], приведенных к разрешению 512×512 пикселей. По оси ординат отложена доля изображений, для которых хеш-код преобразованного изображения полностью совпал с хеш-кодом оригинала, по оси абсцисс – интенсивность соответствующего преобразования. Тестирование проводилось на шести типах воздействий: изменение яркости и контраста, сжатие JPEG, размытие по Гауссу, масштабирование и аддитивный гауссовский шум.

Анализ полученных результатов позволяет выявить выраженную специфику методов. При отсутствии искажений все алгоритмы обеспечивают полное совпадение. К масштабированию, размытию и сжатию JPEG устойчивы все методы, кроме PDQ, который показывает худшие результаты практически во всех сценариях. Классические алгоритмы (aHash, pHash, wHash) демонстрируют высокую устойчивость к данным воздействиям, однако по мере усиления искажений доля совпавших хеш-кодов падает. В то же время функция EPHash на этих трех преобразованиях демонстрирует практически абсолютную инвариантность, сохраняя полное побитовое совпадение даже при наибольшей силе воздействия.

Наиболее уязвимым местом для всех рассматриваемых алгоритмов, включая EPHash, оказались фотометрические искажения – изменение яркости и контраста. При их усилении наблюдается резкое снижение доли совпадений у всех методов. В случае добавления аддитивного гауссовского шума наилучшие показатели устойчивости демонстрируют EPHash и wHash, в то время как остальные алгоритмы деградируют значительно быстрее.

Таким образом, предложенная функция EPHash существенно превосходит классические аналоги по устойчивости к размытию, масштабированию, сжатию JPEG и наложения шума, обеспечивая по ряду преобразований полную побитовую инвариантность хеш-кода.

Единственным ограничением EPHash, сопоставимым с классическими методами, остается чувствительность к сильным фотометрическим искажениям (яркость, контраст).

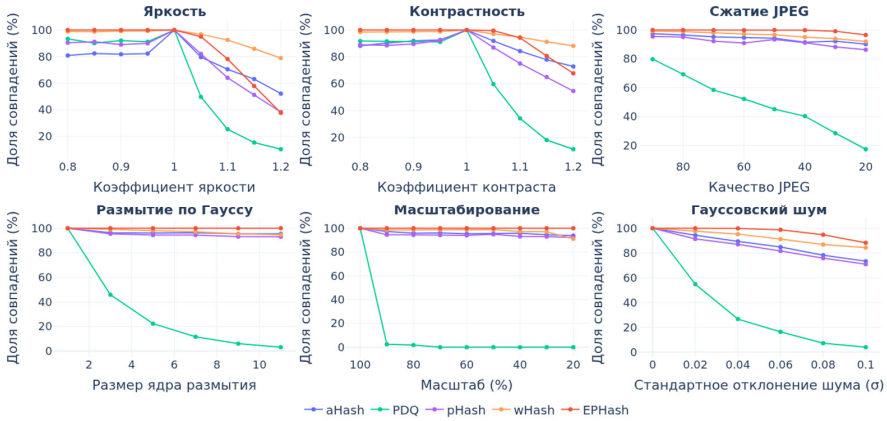

Рис. 2. Зависимость доли совпадающих перцептивных хеш-кодов, полученных алгоритмами aHash, PDQ, pHash, wHash и EPHash, от интенсивности допустимых преобразований: изменения яркости, контрастности, сжатия JPEG, размытия по Гауссу, масштабирования и наложения гауссовского шума.

Fig. 2. Dependence of the matching rate of perceptual hashes generated by aHash, PDQ, pHash, wHash, and EPHash on the intensity of benign distortions: brightness and contrast adjustments, JPEG compression, Gaussian blur, scaling, and additive Gaussian noise.

5.3 Незаметность внедряемой маркировки

Для практического применения метода критически важно, чтобы защитная маркировка не ухудшала визуальное качество исходного изображения. В данном разделе проводится сравнение незаметности нескольких вариантов предложенной системы с решением MetaSeal, который также решает задачу криптографического обеспечения защиты от атаки подделки с использованием QR-кодов. Для комплексной оценки визуального качества и учета различных аспектов восприятия искажений применены следующие метрики:

- **PSNR –** пиковое отношение сигнала к шуму, измеряет попиксельное отклонение модифицированного изображения от оригинала. Измеряется в дБ, чем выше значение, тем меньше внедренных искажений;
- **SSIM –** индекс структурной схожести, оценивающий изменения в структуре изображения. Значения варьируются от 0 до 1, где 1 означает полное совпадение двух изображений;
- **LPIPS** [26] – перцептивная метрика на основе глубокого обучения. Оценивает визуальное сходство с использованием признаков, извлеченных из предобученных сверточных нейросетей. Чем ниже значение LPIPS, тем менее заметны искажения для человеческого глаза.

Эксперимент проведен на 1000 изображений из набора данных MSCOCO. Исходные изображения были приведены к разрешению 512×512 пикселей, по аналогии с экспериментальной методикой авторов MetaSeal. Сравнение производилось как между реализациями предложенной схемы на базе различных алгоритмов внедрения ЦВЗ, так и с самим методом MetaSeal. В рамках экспериментального тестирования использовались официальные открытые реализации и предобученные модели авторов методов PixelSeal [27],

VideoSeal [27], ChunkySeal [27], MBRS [28] и MetaSeal [29]. В табл. 1 приведены значения описанных метрик незаметности для рассматриваемых вариантов маркировки.

Табл. 1. Сравнение значений метрик незаметности предложенной схемы с различными алгоритмами внедрения ЦВЗ и метода MetaSeal
Table 1. Comparison of imperceptibility metrics for the proposed scheme using various watermark embedding algorithms and the MetaSeal method

Метрика	CPW (MBRS)	CPW (VideoSeal)	CPW (PixelSeal)	CPW (ChunkySeal)	MetaSeal
PSNR, дБ↑	40.7	41.1	41.2	41.4	35.1
SSIM↑	0.982	0.985	0.985	0.986	0.929
LPIPS↓	0.011	0.009	0.009	0.011	0.098

Как видно из табл. 1, все реализации предложенной схемы демонстрируют высокие показатели визуального качества: значения PSNR превышают порог в 40 дБ, SSIM находится в пределах 0.982–0.986, а метрика LPIPS не превышает 0.011, что говорит о низкой заметности внедряемой маркировки для человеческого глаза. При этом метод MetaSeal показывает существенно худшие результаты по всем трем метрикам.

5.4 Устойчивость к допустимым преобразованиям

Ключевым требованием к системе проактивной защиты является сохранение работоспособности при допустимых искажениях, неизбежно возникающих при хранении, передаче и рутинной обработке медиа данных (сжатие, изменение яркости, фильтрация и т.д.). В данном разделе оценивается устойчивость предложенного метода к различным типам допустимых преобразований в сравнении с решением MetaSeal. В качестве итоговой метрики используется точность верификации (доля изображений, для которых был получен положительный результат проверки криптографической подписи). Результаты эксперимента представлены на рис. 3.

Анализ результатов, приведенных на рис. 3, показывает, что исследуемые методы демонстрируют различный профиль устойчивости к рассматриваемым преобразованиям. Метод MetaSeal имеет небольшое преимущество в устойчивости к атакам повышения яркости и наложению гауссовского шума, однако в то же время проявляет слабую устойчивость к размытию и масштабированию с уменьшением разрешения.

Среди реализаций предложенной схемы на базе различных алгоритмов ЦВЗ также наблюдается специализация. Решение, использующее MBRS в качестве носителя информации, оказывается наиболее устойчивым к сжатию JPEG, сохраняя возможность успешной верификации даже при снижении коэффициента качества до 40, но при этом проявляет уязвимость к размытию. Подход с использованием ChunkySeal, напротив, демонстрирует наивысшую устойчивость к снижению разрешения изображения и размытию, однако не сохраняет работоспособность после применения сжатия JPEG. Наиболее сбалансированную устойчивость ко всем указанным типам атак демонстрирует подход на основе метода VideoSeal, обеспечивая стабильно высокие результаты верификации в различных сценариях воздействия.

Таким образом, выбор базового алгоритма внедрения ЦВЗ в рамках предложенной криптографической схемы позволяет адаптировать систему под конкретные угрозы, при этом метод VideoSeal представляет собой оптимальный компромисс для защиты от широкого спектра допустимых преобразований.

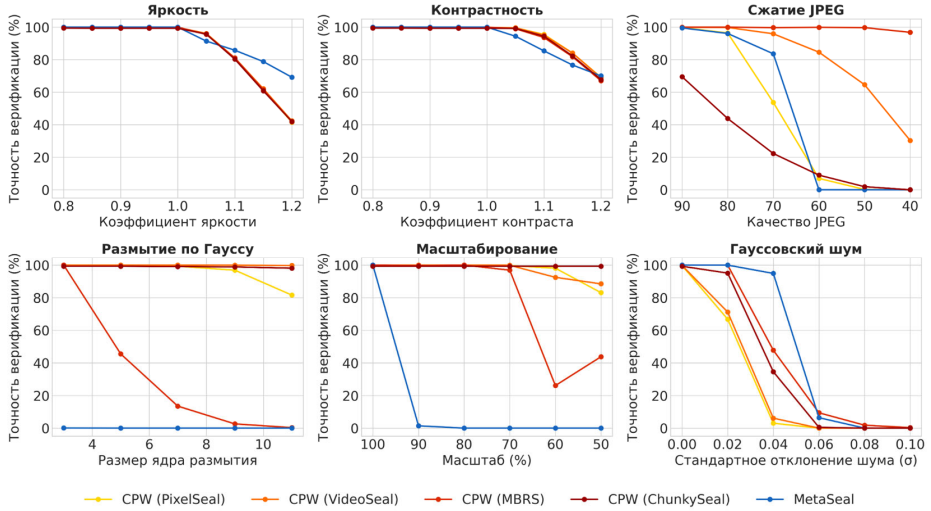

Рис. 3. Зависимость точности верификации предложенного метода CPW и метода MetaSeal от интенсивности допустимых преобразований: изменения яркости, контрастности, сжатия JPEG, размытия по Гауссу, масштабирования и наложения гауссовского шума.
Fig. 3. Verification accuracy of the proposed method CPW vs. MetaSeal under benign distortions: brightness, contrast, JPEG, blur, scaling, and Gaussian noise.

5.5 Обнаружение семантических модификаций

Для практической оценки способности предложенного метода выявлять недопустимые преобразования был проведен эксперимент на наборе данных IMD2020 [30], представляющем собой коллекцию изображений с реальными манипуляциями (вставка объектов, клонирование, удаление элементов) и соответствующими масками, указывающими на измененные области. Из набора данных исключены изображения с некорректно размеченными или вырожденными масками, в результате чего была сформирована выборка из 926 примеров поддельных изображений. Для каждой пары исходного и модифицированного изображений из IMD2020 сначала формировалось маркированное исходное изображение. Затем пиксели внутри заданной маски заменялись соответствующими пикселями модифицированного изображения. Полученный экземпляр передавался на верификацию.

Для анализа зависимости эффективности обнаружения от масштаба вмешательства отфильтрованные изображения были отсортированы по возрастанию доли подменяемой площади (рассчитанной как отношение количества измененных пикселей к общему количеству пикселей изображения). Отсортированная выборка была разбита на 6 групп по перцентилям доли подменяемой площади. Для каждой группы был рассчитан показатель обнаружения – доля изображений, при проверке которых метод успешно зафиксировал нарушение целостности (верификация вернула отрицательный результат). Результаты представлены на рис. 4.

На графике по оси абсцисс отложена средняя доля подменяемой площади для каждой группы (в логарифмическом масштабе), а по оси ординат – доля успешно обнаруженных модификаций. Согласно полученным результатам, метод MetaSeal способен надежно определять наличие модификации только в том случае, если доля подменяемой площади составляет не менее 0,1 (10%) от всего изображения. В то же время все варианты

предложенного решения (CPW) практически безошибочно выявляют подделки уже при доле измененной площади более 0,015 (1,5%).

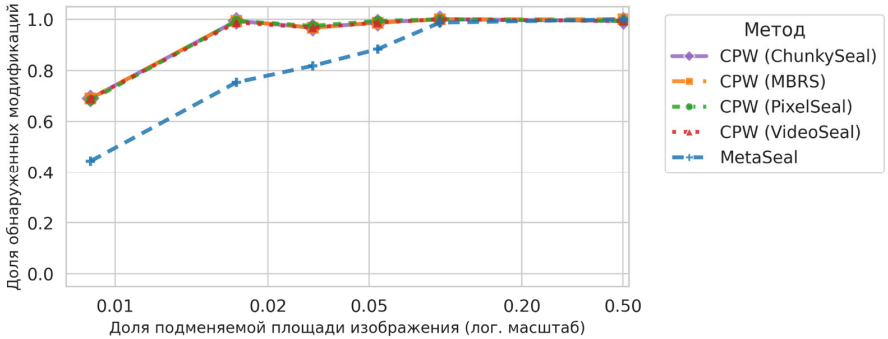

Рис. 4. Зависимость доли выявленных недопустимых модификаций от перцентиля доли подменяемой площади изображения на наборе данных IMD2020.
Fig. 4. Dependence of the proportion of detected malicious modifications on the percentile of the tampered area fraction in the IMD2020 dataset.

Стоит отметить, что кривые, отражающие долю обнаружения модификаций для различных вариантов реализации CPW (на базе разных алгоритмов ЦВЗ), практически совпадают. Это объясняется архитектурной особенностью предложенной схемы: ключевым фактором, определяющим чувствительность системы к локальным изменениям, выступает перцептивный хеш-функция. Поскольку во всех вариантах CPW используется один и тот же алгоритм хеширования, порог обнаружения подделки остается неизменным. Применяемые же алгоритмы внедрения ЦВЗ обладают достаточной локальной устойчивостью и не разрушаются при подмене малой части изображения.

6. Ограничения

Предложенный метод имеет ряд ограничений, связанных с доверенной идентификацией владельца ключа, свойствами перцептивного хеширования и устойчивостью алгоритма извлечения ЦВЗ. Предложенный метод предполагает, что проверяющая сторона располагает подлинным открытым ключом создателя контента. Сама цифровая подпись подтверждает соответствие изображения заданному ключу, но не устанавливает личность его владельца. Поэтому практическое применение метода требует внешней инфраструктуры атрибуции ключей, например, сертификатов или доверенных реестров.

Криптографическая привязка подписи к EPHash препятствует переносу аутентификатора на изображение с другим хеш-кодом. Однако успешная проверка перенесенной подписи теоретически возможна, если злоумышленнику удастся построить семантически измененное изображение с таким же значением EPHash. Такая атака соответствует построению второго прообраза для перцептивной хеш-функции. Стойкость EPHash к целенаправленным состязательным воздействиям в настоящей работе формально не доказывается. Возможными направлениями ее повышения являются методы защиты от инверсии битов, применяемые в RD-PHash [31], и сертифицируемые подходы, представленные в CertPHash [32].

Создаваемый при маркировании зазор повышает устойчивость хеш-кода, но не обеспечивает безусловной инвариантности. Работоспособность EPHash ограничена рассмотренными типами и диапазонами преобразований. Сильные фотометрические воздействия, поворот, кадрирование и другие неподдерживаемые геометрические преобразования могут приводить к отказу верификации. Устойчивость к последовательному применению нескольких преобразований отдельно не исследовалась. Кроме того, проверка подписи требует ее

точного восстановления: если количество ошибок превышает корректирующую способность кода БЧХ, изображение отклоняется независимо от причины повреждения маркировки. Поэтому отрицательный результат не позволяет отличить семантическую модификацию от сильного допустимого воздействия или удаления ЦВЗ.

Результаты на IMD2020 показывают, что вероятность обнаружения зависит от площади, расположения и характера измененной области. В группах со средней долей модифицированной площади около 1,5-2% доля отклоненных изображений приближалась к 100%, однако это значение не является строгим порогом. Малые или малоконтрастные изменения могут не затронуть используемые низкочастотные признаки и остаться необнаруженными. Следовательно, метод ориентирован на выявление содержательных изменений и не обеспечивает попиксельный контроль или локализацию модифицированных областей.

7. Заключение

В работе предложен метод CPW для проактивного контроля целостности изображений, объединяющий перцептивное хеширование, цифровую подпись и устойчивый ЦВЗ. Перцептивный хеш-код изображения подписывается по схеме ECDSA, и полученная цифровая подпись внедряется в изображение в виде ЦВЗ. При верификации подпись извлекается из изображения и проверяется относительно вычисленного для него значения EPHash и заданного открытого ключа. Успешная проверка подтверждает, что данный хеш-код был подписан владельцем соответствующего закрытого ключа и не изменился относительно подписанного значения. Благодаря устойчивости EPHash к допустимым преобразованиям это позволяет подтвердить соответствие проверяемого изображения аутентифицированному оригиналу с точностью до таких преобразований. Перенос извлеченной подписи на изображение с другим значением EPHash приводит к отрицательному результату верификации.

Для получения побитово воспроизводимого подписываемого аутентификатора разработана перцептивная хеш-функция EPHash, основанная на квантовании и модификации низкочастотных коэффициентов ДКП над изображением. В исследованных диапазонах преобразований хеш-функция EPHash обеспечила полное совпадение хеш-кодов для атак масштабирования, размытия и JPEG-сжатия, продемонстрировав более высокую побитовую устойчивость по сравнению с функциями aHash, pHash, wHash и PDQ. Для точного восстановления 512-битной подписи применен код БЧХ, формирующий 1024-битную полезную нагрузку. Рассмотрены две стратегии ее внедрения: непосредственное внедрение с помощью ChunkySeal и разделение на четыре 256-битных фрагмента с независимым внедрением в квадранты изображения методами PixelSeal, VideoSeal и MBRS.

Эксперименты на наборе данных MSCOCO показали, что все исследованные варианты CPW обеспечивают значения PSNR выше 40 дБ и LPIPS не более 0,011, тогда как для MetaSeal получены значения 35,1 дБ и 0,098 соответственно. Устойчивость CPW к допустимым преобразованиям зависит от выбранного алгоритма маркирования: MBRS показал наилучшие результаты при JPEG-сжатии, ChunkySeal – при размытии и масштабировании, а VideoSeal продемонстрировал наиболее сбалансированный профиль. На наборе данных IMD2020 практически для всех изображений, в которых модификации затрагивали в среднем около 1,5–2% площади, верификация CPW завершалась отрицательным результатом, свидетельствующим о нарушении целостности. MetaSeal обеспечивал сопоставимую долю обнаружения только при более существенных изменениях, затрагивающих около 10% площади изображения.

Полученные экспериментальные результаты подтверждают применимость сочетания перцептивного хеш-кода, цифровой подписи и ЦВЗ для публичной слепой проверки целостности изображений. Вместе с тем, предложенный метод требует доверенного

распространения открытых ключей, устойчивости EPHash к построению второго прообраза и возможности точного извлечения цифровой подписи. Дальнейшие исследования будут направлены на защиту EPHash от целенаправленных состязательных атак, повышение устойчивости к геометрическим и комбинированным преобразованиям над изображением, а также обнаружение и локализацию модификаций, затрагивающих незначительную область изображения.

Список литературы / References

[1]. Masood M. et al. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. Applied intelligence, 2023, vol. 53, no. 4, pp. 3974-4026.

[2]. Zhang X. et al. Editguard: Versatile image watermarking for tamper localization and copyright protection //2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2024, pp. 11964-11974.

[3]. Wang S.Y. et al. CNN-generated images are surprisingly easy to spot… for now. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020, pp. 8692-8701.

[4]. Carlini N., Farid H. Evading deepfake-image detectors with white-and black-box attacks //2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2020, pp. 2804-2813.

[5]. Cox I.J. et al. Secure spread spectrum watermarking for multimedia. IEEE transactions on image processing, 1997, vol. 6, no. 12, pp. 1673-1687.

[6]. Zhu J. et al. Hidden: Hiding data with deep networks //European conference on computer vision. Cham: Springer International Publishing, 2018, pp. 682-697.

[7]. Jia Z., Fang H., Zhang W. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. Proceedings of the 29th ACM international conference on multimedia, 2021, pp. 41-49.

[8]. Fernandez P. et al. Video seal: Open and efficient video watermarking. arXiv preprint, 2024. DOI: 10.48550/arXiv.2412.09492.

[9]. Souček T. et al. Pixel seal: Adversarial-only training for invisible image and video watermarking. arXiv preprint, 2025. DOI: 10.48550/arXiv.2512.16874.

[10]. Petrov A. et al. We can hide more bits: The unused watermarking capacity in theory and in practice. arXiv preprint, 2025. DOI: 10.48550/arXiv.2510.12812.

[11]. Zhao Y. et al. Proactive deepfake defence via identity watermarking. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2023, pp. 4591-4600.

[12]. Wu X., Liao X., Ou B. Sepmark: Deep separable watermarking for unified source tracing and deepfake detection. Proceedings of the 31st ACM international conference on multimedia, 2023, pp. 1190-1201.

[13]. Zhang X. et al. Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2025, pp. 3008-3018.

[14]. Kutter M., Voloshynovskiy S. V., Herrigel A. Watermark copy attack. Security and Watermarking of Multimedia Contents II. SPIE, 2000, vol. 3971, pp. 371-380.

[15]. Wong P.W., Memon N. Secret and public key image watermarking schemes for image authentication and ownership verification. IEEE transactions on image processing, 2001, vol. 10, no. 10, pp. 1593-1601.

[16]. Barreto P.S.L.M., Kim H.Y., Rijmen V. Toward secure public-key blockwise fragile authentication watermarking. IEE Proceedings-Vision, Image and Signal Processing, 2002, vol. 149, no. 2, pp. 57-62.

[17]. Schlauweg M. et al. Quantization-based semi-fragile public-key watermarking for secure image authentication. Mathematics of data/image coding, compression, and encryption VIII, with applications. SPIE, 2005, vol. 5915, pp. 41-51.

[18]. Tarhini H. et al. Neural Watermarking: Lack of a Secret Key is still Lack of Security. 2026.

[19]. Fairoze J. et al. On the difficulty of constructing a robust and publicly-detectable watermark. arXiv preprint, 2025. DOI: 10.48550/arXiv.2502.04901.

[20]. Zhou T. et al. MetaSeal: Defending against image attribution forgery through content-dependent cryptographic watermarks. arXiv preprint, 2025. DOI: 10.48550/arXiv.2509.10766.

[21]. Du L., Ho A.T.S., Cong R. Perceptual hashing for image authentication: A survey. Signal Processing: Image Communication, 2020, vol. 81, pp. 115713.

[22]. Zauner C. Implementation and benchmarking of perceptual image hash functions. 2010.

[23]. Dalins J., Wilson C., Boudry D. PDQ & TMK+ PDQF – A Test Drive of Facebook's Perceptual Hashing Algorithms. arXiv preprint, 2019. DOI: 10.48550/arXiv.1912.07745.

[24]. Bose R.C., Ray-Chaudhuri D.K. On a class of error correcting binary group codes. Information and control, 1960, vol. 3, no. 1, pp. 68-79.

[25]. Lin T.Y. et al. Microsoft coco: Common objects in context. European conference on computer vision. Cham: Springer International Publishing, 2014, pp. 740-755.

[26]. Zhang R. et al. The unreasonable effectiveness of deep features as a perceptual metric. 2018 IEEE/CVF conference on computer vision and pattern recognition. IEEE, 2018, pp. 586-595.

[27]. VideoSeal: Open and efficient video and image watermarking. GitHub. Available at: https://github.com/facebookresearch/videoseal, accessed: 14.08.2026.

[28]. MBRS: Enhancing Robustness of DNN-based Watermarking by Mini-Batch of Real and Simulated JPEG Compression. GitHub. Available at: https://github.com/jzyustc/MBRS, accessed: 14.08.2026.

[29]. MetaSeal: Defending Against Image Attribution Forgery Through Content-Dependent Cryptographic Watermarks. GitHub. Available at: https://github.com/Tongzhou0101/MetaSeal, accessed: 14.08.2026.

[30]. Novozámský A., Mahdian B., Saic S. IMD2020: A large-scale annotated dataset tailored for detecting manipulated images. 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE, 2020, pp. 71-80.

[31]. Jiang N. et al. RD-PHash: A Robustness Enhancement for DCT-Based Perceptual Hashing Against Adversarial Bit-Flipping Attacks. 2026 IEEE Symposium on Security and Privacy Workshops (SPW). IEEE, 2026, pp. 250-256.

[32]. Yang Y. et al. {CertPHash}: Towards Certified Perceptual Hashing via Robust Training. 34th USENIX Security Symposium (USENIX Security 25), 2025, pp. 7839-7856.

Информация об авторах / Information about authors

Алексей Юрьевич ЯКУШЕВ – научный сотрудник ИСП им. В.П. Иванникова РАН. Научные интересы: цифровые водяные знаки, обработка цифровых изображений, алгоритмы машинного обучения.

Aleksey Yur'evich YAKUSHEV is a researcher of the Ivannikov Institute for System Programming of the RAS. Scientific interests: watermarking, digital image processing, machine learning algorithms.

Александр Андреевич АКИМЕНКОВ – стажер-исследователь ИСП им. В.П. Иванникова РАН, аспирант. Научные интересы: цифровые водяные знаки, обработка цифровых изображений, алгоритмы машинного обучения.

Alexander Andreevich AKIMENKOV is a research intern of the Ivannikov Institute for System Programming of the RAS and a PhD student. Scientific interests: steganography, digital image processing, machine learning algorithms.

Дмитрий Олегович ОБЫДЕНКОВ – кандидат технических наук, научный сотрудник ИСП им. В.П. Иванникова РАН. Научные интересы: методы сокрытия и защищенной передачи информации, компьютерные сети, технологии анализа сетевого трафика.

Dmitry Olegovich OBYDENKOV – Cand. Sci. (Tech.). He is a researcher of the Ivannikov Institute for System Programming of the RAS. Scientific interests: methods for information hiding and secure transmission, computer networks, technologies of network traffic analysis.

Халед Набиль АБУД – стажер-исследователь Института искусственного интеллекта МГУ имени М.В. Ломоносова, аспирант. Научные интересы: состязательная устойчивость, генеративные модели, обработка цифровых изображений, глубокое обучение.

Khaled Nabil ABUD is a research intern of the MSU Institute for AI and a postgraduate student. Scientific interests: adversarial robustness, generative models, digital image processing, deep learning.

Сергей Евгеньевич Токарев – старший лаборант ИСП им. В.П. Иванникова РАН. Область научных интересов: информационная безопасность, обработка изображений, алгоритмы машинного обучения.

Sergey Evgen’evich Tokarev is a senior laboratory assistant of the Ivannikov Institute for System Programming of the RAS. Scientific interests: information security, image processing, machine learning algorithms.

Юрий Витальевич МАРКИН – кандидат технических наук, научный сотрудник ИСП им. В.П. Иванникова РАН, руководитель направления применения цифровых водяных знаков. Область научных интересов: информационная безопасность, анализ сетевого трафика, обработка изображений, алгоритмы машинного обучения.

Yury Vital'evich MARKIN – Cand. Sci. (Tech.). He is a researcher of the Ivannikov Institute for System Programming of the RAS, Head of Digital Watermarking Applications. Scientific interests: information security, network traffic analysis, image processing, machine learning algorithms.