

Федеральное государственное бюджетное учреждение науки
Институт системного программирования им. В.П. Иванникова
Российской академии наук

На правах рукописи

Якушев Алексей Юрьевич

**Методы и средства защиты медиаконтента на основе
цифровых водяных знаков**

Специальность 2.3.5 —
«Математическое и программное обеспечение вычислительных систем,
комплексов и компьютерных сетей»

Диссертация на соискание ученой степени
кандидата технических наук

Научный руководитель:
канд. техн. наук
Маркин Юрий Витальевич

Москва — 2026

Оглавление

	Стр.
Введение	6
 Глава 1. Обзор предметной области защиты медиаконтента на основе цифровых водяных знаков	 14
1.1 Основные понятия и принципы маркирования медиаконтента . . .	14
1.1.1 Общая схема внедрения, обнаружения и извлечения ЦВЗ . .	15
1.1.2 Классификация методов маркирования	16
1.1.3 Основные свойства и требования к методам маркирования .	18
1.2 Методы маркирования изображений	19
1.2.1 Классические методы	20
1.2.2 Нейросетевые методы	23
1.2.3 Маркирование изображений в процессе генерации	27
1.3 Методы маркирования аудиоданных	30
1.3.1 Классические методы	30
1.3.2 Нейросетевые методы	32
1.4 Методы маркирования видеоданных	34
1.5 Методы контроля целостности медиаконтента на основе ЦВЗ . . .	36
1.6 Средства оценки методов маркирования медиаконтента	39
1.6.1 Средства сравнительной оценки методов маркирования изображений	39
1.6.2 Средства сравнительной оценки методов маркирования аудиоданных	42
1.7 Выводы по главе	44
 Глава 2. Методика унифицированной воспроизводимой оценки методов маркирования медиаконтента	 46
2.1 Модель унифицированной экспериментальной оценки методов маркирования медиаконтента	47
2.2 Программная реализация методики в системе WIBE	50
2.2.1 Архитектура программной системы	51
2.2.2 Воспроизводимое описание экспериментов	54

	Стр.
2.2.3 Организация исполнения экспериментов	55
2.3 Сравнительная оценка методов маркирования изображений	57
2.3.1 Метрики оценки	57
2.3.2 Преобразования и преднамеренные атаки	62
2.3.3 Условия проведения эксперимента	64
2.3.4 Результаты сравнительной оценки	65
2.4 Сравнительная оценка методов маркирования аудио	70
2.4.1 Метрики оценки	71
2.4.2 Преобразования и преднамеренные атаки	71
2.4.3 Условия проведения эксперимента	73
2.4.4 Результаты сравнительной оценки	75
2.5 Выводы по главе	80

Глава 3. Метод контроля целостности изображений на основе

подписываемого перцептивного представления	83
3.1 Постановка задачи контроля целостности изображений	83
3.1.1 Модель нарушителя	84
3.1.2 Допустимые и недопустимые преобразования	85
3.2 Метод контроля целостности изображений CPW	87
3.2.1 Перцептивная хеш-функция EPHash	88
3.2.2 Формирование аутентификатора	91
3.2.3 Помехоустойчивое кодирование и внедрение аутентификатора	92
3.2.4 Верификация целостности изображения	93
3.3 Экспериментальное исследование метода CPW	95
3.3.1 Параметры реализации и условия экспериментов	95
3.3.2 Исследование побитовой устойчивости EPHash	97
3.3.3 Оценка незаметности маркированных изображений	99
3.3.4 Устойчивость к допустимым преобразованиям	100
3.3.5 Обнаружение семантических модификаций	102
3.4 Ограничения метода	104
3.5 Выводы по главе	105

Глава 4. Метод контроля целостности аудиоданных на основе цифровых водяных знаков и цифровой подписи	108
4.1 Постановка задачи контроля целостности аудиоданных	108
4.2 Метод контроля целостности аудиоданных CryptoMark	111
4.2.1 Формирование устойчивого перцептивного хеш-кода	112
4.2.2 Формирование цифровой подписи и помехоустойчивое кодирование	114
4.2.3 Внедрение цифровой подписи в аудиосигнал	115
4.2.4 Верификация целостности и локализация модификаций	116
4.3 Экспериментальная оценка метода CryptoMark	118
4.3.1 Параметры реализации	118
4.3.2 Оценка незаметности маркирования	119
4.3.3 Оценка контроля целостности	120
4.3.4 Оценка локализации модификаций	123
4.4 Ограничения метода	125
4.5 Выводы по главе	126
Глава 5. Методика переноса методов маркирования изображений на видеоданные	128
5.1 Постановка задачи переноса методов маркирования изображений на видео	128
5.2 Методика группового переноса методов маркирования изображений на видеоданные	129
5.2.1 Формирование агрегированного представления группы кадров	129
5.2.2 Внедрение цифрового водяного знака	130
5.2.3 Групповое извлечение цифрового водяного знака	131
5.3 Экспериментальная оценка методики	132
5.3.1 Сравнение стратегий переноса методов маркирования изображений на видео	134
5.3.2 Влияние размера группы кадров	136
5.3.3 Влияние временного распределения модификации	139
5.3.4 Устойчивость к преобразованиям видеоданных	141
5.3.5 Устойчивость к нарушению временной синхронизации	142

	Стр.
5.4 Ограничения методики	144
5.5 Выводы по главе	145
Заключение	147
Список литературы	149
Список рисунков	163
Список таблиц	165

Введение

Развитие цифровых платформ, социальных сетей, сервисов потокового вещания и генеративных моделей искусственного интеллекта существенно изменило процессы создания и распространения медиаконтента. Изображения, аудио- и видеоданные могут создаваться, обрабатываться и публиковаться с использованием программных средств, а затем практически без потери качества копироваться и распространяться среди неограниченного числа пользователей. В то же время развитие генеративных моделей существенно расширило возможности как по синтезу нового медиаконтента, так и по внесению реалистичных изменений в существующие материалы, что усложняет различие исходного, модифицированного и полностью синтезированного контента. Это приводит к ряду угроз информационной безопасности: несанкционированное копирование и распространение медиаконтента, нарушение авторских и лицензионных прав, подмена отдельных фрагментов изображения, аудиозаписи или видео, создание поддельных материалов, в том числе с применением технологий искусственного интеллекта, а также указание неверного источника происхождения контента. При этом факт модификации медиаданных не всегда удастся установить визуально или на слух, а методы выявления подделок, основанные на анализе признаков самого контента, могут терять эффективность по мере совершенствования алгоритмов его генерации и обработки. В связи с этим актуальной является разработка проактивных средств защиты, предусматривающих предварительное внесение в медиаконтент проверяемой информации о его происхождении, владельце, получателе или состоянии целостности.

Дополнительным фактором, определяющим актуальность рассматриваемой задачи, является формирование нормативных требований к обозначению и выявлению материалов, созданных или измененных с применением искусственного интеллекта. В Европейском союзе устанавливаются требования к машиночитаемому обозначению и технической детектируемости синтетического контента, в Китае действует система обязательной явной и неявной идентификации такого контента, а в США и Российской Федерации соответствующие требования продолжают формироваться на уровне законодательства и отраслевых инициатив. При этом нормативное регулирование, как правило, устанавливает требования к возможности определения происхождения материала или факта применения

технологий искусственного интеллекта, не всегда конкретизируя технические способы их реализации. Практическое применение таких требований в масштабах современных цифровых платформ обуславливает необходимость разработки автоматизированных средств машиночитаемого обозначения медиаконтента и последующей проверки его происхождения и целостности.

Одним из перспективных технических средств защиты медиаконтента являются цифровые водяные знаки (ЦВЗ), позволяющие внедрять в цифровой объект дополнительную скрытую информацию. В отличие от внешних метаданных, цифровой водяной знак непосредственно связан с объектом медиа и может сохраняться после преобразований, возникающих при хранении, передаче, перекодировании, сжатии или изменении формата данных. В настоящей работе под маркированием медиаконтента понимается процесс внедрения в изображения, аудио- и видеоданные цифровых водяных знаков, предназначенных для подтверждения происхождения, контроля целостности или отслеживания распространения.

К настоящему времени предложено значительное число методов внедрения ЦВЗ в изображения, аудио- и видеоданные, различающихся по емкости, незаметности, устойчивости, способу внедрения и модели извлечения. Значительный вклад в развитие методов цифровых водяных знаков внесли зарубежные исследователи I. Cox, T. Furon, P. Fernandez, а также отечественные исследователи В. А. Федосеев, В. И. Коржик, О. О. Евсютин и др. Наряду с методами маркирования, основанными на заранее заданных правилах преобразования медиаданных, активно развиваются обучаемые методы, в том числе с использованием нейронных сетей. Отдельное направление связано с формированием ЦВЗ непосредственно в процессе генерации медиаконтента. Однако сравнение существующих методов затрудняется отсутствием единых экспериментальных условий. В различных работах используются разные наборы данных, преобразования и атаки, метрики незаметности и устойчивости ЦВЗ. Вследствие этого заявленные в публикациях характеристики отдельных алгоритмов не всегда позволяют оценить их преимущества и выбрать метод, соответствующий конкретному сценарию применения. Это обуславливает актуальность разработки унифицированной методики и программных средств воспроизводимой оценки методов внедрения ЦВЗ.

Вместе с тем высокая устойчивость ЦВЗ сама по себе не гарантирует целостность медиаконтента и не подтверждает его происхождение. Водяной знак может сохраниться после недопустимого изменения содержимого объекта ли-

бо быть перенесен на другой объект. Для контроля целостности необходимо обеспечить проверяемую связь между содержимым медиаданных и внедренным аутентификатором, а также возможность подтверждения его подлинности. Перспективным является совместное применение ЦВЗ, перцептивного хеширования и электронной цифровой подписи. Ключевой научной задачей при этом является построение подписываемого перцептивного представления, сохраняющегося при допустимых преобразованиях и изменяющегося при нарушении содержательной целостности медиаданных.

Методы внедрения цифровых водяных знаков существенно зависят от типа обрабатываемых медиаданных. Различия в структуре, способах обработки и характерных преобразованиях изображений, аудио- и видеоданных требуют учета специфики каждого типа медиаконтента как при разработке методов внедрения ЦВЗ, так и при их экспериментальной оценке. Таким образом, актуальность диссертационной работы определяется необходимостью комплексного развития методов и программных средств маркирования цифровых медиаданных с учетом современных моделей генерации и преобразования медиаконтента, а также возможности преднамеренного изменения содержимого и подделки ЦВЗ. Решение данной задачи требует унифицированной методики оценки существующих методов, обеспечения проверяемой связи водяного знака с содержимым медиаданных и стороной, сформировавшей ЦВЗ, разработки средств выявления недопустимых изменений, а также разработки методов применения существующих алгоритмов маркирования к видеоданным с учетом их временной структуры.

Целью диссертационной работы является разработка и экспериментальное обоснование методов и создание программных средств защиты медиаконтента на основе цифровых водяных знаков, обеспечивающих воспроизводимую оценку методов маркирования, контроль целостности изображений и аудиоданных, а также эффективное применение методов маркирования изображений к видеоданным.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Провести анализ существующих методов маркирования медиаконтента, подходов к контролю его целостности и средств сравнительной оценки и определить основные ограничения существующих решений.
2. Разработать методику и программные средства унифицированной воспроизводимой оценки методов внедрения цифровых водяных знаков в

медиаконтент, включающие единую модель экспериментального конвейера, наборы данных, модели атак и метрики незаметности и устойчивости для медиаданных различного типа.

3. Разработать и экспериментально исследовать метод контроля целостности изображений, устойчивый к допустимым преобразованиям и обеспечивающий выявление недопустимых модификаций.
4. Разработать и экспериментально исследовать метод контроля целостности аудиоданных, устойчивый к допустимым преобразованиям и обеспечивающий выявление и локализацию недопустимых модификаций.
5. Разработать и экспериментально исследовать методику переноса алгоритмов внедрения и извлечения цифровых водяных знаков, предназначенных для изображений, на видеоданные, позволяющую сократить вычислительные затраты за счет использования временной избыточности видеопоследовательности без изменения исходных алгоритмов.

Объект и предмет исследования. Объектом исследования являются процессы цифровой обработки медиаданных при внедрении в них цифровых водяных знаков.

Предметом исследования являются методы, алгоритмы и программные средства обработки медиаданных, предназначенные для унифицированной экспериментальной оценки методов маркирования, контроля целостности изображений и аудиоданных и переноса методов маркирования изображений на видеоданные.

Научная новизна:

1. Разработана методика унифицированной воспроизводимой оценки методов маркирования медиаконтента, основанная на едином представлении экспериментального процесса для методов с различными принципами формирования и проверки ЦВЗ и различных типов медиаданных при сохранении модально-зависимых воздействий, метрик и способов представления данных.
2. Разработан метод контроля целостности изображений, в котором предложен способ повышения побитовой устойчивости подписываемого перцептивного представления за счет контролируемой модификации низкочастотных коэффициентов изображения относительно порогов бинаризации без изменения исходного значения перцептивного хеш-кода.

3. Разработан метод контроля целостности аудиоданных, в котором независимое формирование и проверка контентно-зависимых аутентификаторов для последовательных временных фрагментов используются для построения карты целостности аудиозаписи, позволяющей выявлять и локализовать содержательные модификации.
4. Разработана методика переноса методов маркирования изображений на видеоданные, основанная на однократной обработке агрегированного представления группы последовательных кадров и распределении сформированной модификации между кадрами, что снижает вычислительные затраты без изменения исходного алгоритма и необходимости его дополнительного обучения.

Теоретическая и практическая значимость. Теоретическая значимость диссертационной работы заключается в развитии методов оценки и применения цифровых водяных знаков к медиаданным различных типов. В работе предложена унифицированная методика оценки методов внедрения цифровых водяных знаков в медиаконтент, позволяющая сопоставлять алгоритмы в воспроизводимых условиях. Разработанные методы интеграции цифровых водяных знаков с электронной цифровой подписью и перцептивным хешированием расширяют научные представления о построении схем маркирования медиаданных с аутентификацией, устойчивых к допустимым преобразованиям и чувствительных к недопустимым модификациям.

Практическая значимость диссертационной работы состоит в разработке программных средств и алгоритмов, применимых в системах защиты медиаконтента, сервисах генерации изображений, аудио- и видеоданных, системах потокового вещания, а также в средствах проверки происхождения и целостности цифрового контента. Разработанная система WIBE может использоваться исследователями и разработчиками для воспроизводимого тестирования методов внедрения цифровых водяных знаков. Предложенные методы контроля целостности изображений и аудиоданных могут применяться для проверки подлинности медиаконтента и выявления недопустимых изменений. Разработанная методика переноса методов маркирования изображений на видеоданные позволяет применять существующие алгоритмы к видеопоследовательностям без изменения их архитектуры и дополнительного обучения при снижении вычислительных затрат за счет групповой обработки кадров.

Результаты диссертационной работы внедрены в практику работы ООО «ЯНДЕКС» при проведении исследований и разработке средств маркирования и проверки медиаконтента. Практическое применение получили разработанные методика и программные средства воспроизводимой оценки методов маркирования, а также алгоритмические решения для маркирования изображений и аудиоданных, что подтверждено актом о внедрении.

Методология и методы исследования. Методологическую основу работы составляют системный подход к построению средств защиты медиаконтента, моделирование угроз и нарушителя, а также экспериментальная оценка разработанных методов. В работе использованы методы цифровой обработки изображений, аудиосигналов и видеоданных, методы маркирования, перцептивного хеширования, помехоустойчивого кодирования, машинного обучения и статистического анализа. Экспериментальное исследование основано на применении преобразований и преднамеренных атак к маркированному медиаконтенту, оценке незаметности и устойчивости цифровых водяных знаков.

Основные положения, выносимые на защиту:

1. Методика унифицированной воспроизводимой оценки методов внедрения цифровых водяных знаков в медиаконтент, обеспечивающая единую модель экспериментального конвейера при использовании модально-зависимых наборов данных, моделей атак и метрик незаметности и устойчивости.
2. Метод контроля целостности изображений на основе подписываемого перцептивного представления, обеспечивающий устойчивость к допустимым преобразованиям и выявление недопустимых модификаций.
3. Метод контроля целостности аудиоданных на основе независимой проверки подписываемого перцептивного представления временных фрагментов, обеспечивающий устойчивость к допустимым преобразованиям, а также выявление и локализацию недопустимых модификаций.
4. Методика переноса методов маркирования изображений на видеоданные на основе групповой обработки последовательных кадров, обеспечивающая снижение вычислительных затрат за счет использования временной избыточности видеопоследовательности без изменения исходных алгоритмов маркирования.

Соответствие паспорту специальности. Приведенные положения соответствуют следующим направлениям исследований паспорта специальности 2.3.5

«Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»:

- **п.4** Интеллектуальные системы машинного обучения, управления базами данных и знаний, инструментальные средства разработки цифровых продуктов.
- **п.7** Модели, методы, архитектуры, алгоритмы, форматы, протоколы и программные средства человеко-машинных интерфейсов, компьютерной графики, визуализации, обработки изображений и видеоданных, систем виртуальной реальности, многомодального взаимодействия в социоконвергентных системах.
- **п.8** Модели и методы создания программ и программных систем для параллельной и распределенной обработки данных, языки и инструментальные средства параллельного программирования.
- **п.10** Оценка качества, стандартизация и сопровождение программных систем.

Достоверность результатов обеспечивается экспериментальной проверкой разработанных методов на общедоступных наборах медиаданных, сравнением с существующими решениями и использованием количественных показателей качества медиаданных, устойчивости ЦВЗ и точности выявления модификаций.

Апробация работы. Результаты работы обсуждались на следующих конференциях:

- II форум «Технологии доверенного искусственного интеллекта», Москва, 27 мая 2024 г.
- Всероссийская конференция «Методы и технические средства обеспечения безопасности информации», Санкт-Петербург, 24–27 июня 2024 г.
- VI форум «Безопасность информационных технологий в цифровой экономике», Москва, 25 сентября 2024 г.
- Международная конференция «Иванниковские чтения», Иркутск, 26–27 июня 2025 г.
- 40th IEEE/ACM International Conference on Automated Software Engineering, Сеул, 16–20 ноября 2025 г.

Личный вклад. Все выносимые на защиту результаты получены лично автором.

Публикации. Основные результаты по теме диссертационной работы представлены в 10 научных работах, из которых 8 опубликованы и 2 приняты к публикации. Среди опубликованных работ 4 статьи изданы в журналах, рекомендованных ВАК, 2 работы — в изданиях, индексируемых в Web of Science и Scopus. Получено 8 свидетельств о государственной регистрации программ для ЭВМ.

В работах [1; 2] автором предложены и разработаны методы внедрения цифровых водяных знаков в изображения текстовых документов. В работах [3; 4] автор принимал участие в разработке и реализации системы маркирования текстовых документов. В работе [5] автором исследованы подходы к применению цифровых водяных знаков для защиты видеоконтента. В работах [6; 7] автором предложены архитектура программной системы и подход к унифицированной воспроизводимой оценке методов маркирования, разработан ряд компонентов системы. В работе [8] автором предложен и разработан метод контроля целостности изображений CPW и способ повышения побитовой устойчивости подписываемого перцептивного представления.

Объем и структура работы. Диссертация состоит из введения, 5 глав и заключения. Полный объем диссертации составляет 165 страниц, включая 20 рисунков и 12 таблиц. Список литературы содержит 143 наименования.

Глава 1. Обзор предметной области защиты медиаконтента на основе цифровых водяных знаков

В настоящей главе рассматриваются основные понятия и принципы маркирования медиаконтента на основе цифровых водяных знаков (ЦВЗ). Вводятся схемы внедрения, обнаружения и извлечения ЦВЗ, а также свойства, по которым оцениваются методы маркирования. Далее анализируются классические и нейросетевые методы для изображений и аудиоданных, способы формирования ЦВЗ при генерации изображений и подходы к маркированию видео с учетом временной структуры.

Отдельно рассматриваются методы контроля целостности медиаконтента, позволяющие различать допустимую обработку и изменения содержимого, которые необходимо выявлять. Затем сопоставляются средства сравнительной оценки методов маркирования изображений и аудиоданных. Результаты обзора служат основой для постановки задач последующих глав: разработки системы оценки, методов контроля целостности изображений и аудиоданных, а также методики применения методов маркирования изображений к видео.

1.1 Основные понятия и принципы маркирования медиаконтента

В настоящей работе рассматривается маркирование трех типов медиаданных: изображений, аудиосигналов и видеопоследовательностей. Несмотря на различия в их структуре и способах обработки, методы маркирования имеют общие принципы построения. Они предусматривают формирование в медиаданных признаков, позволяющих впоследствии обнаружить заданный ЦВЗ или восстановить передаваемое сообщение.

Под цифровым водяным знаком понимается информационный сигнал, представленный в структуре медиаданных и предназначенный для последующего обнаружения или извлечения. Маркированием называется процесс формирования объекта медиаконтента, содержащего ЦВЗ. Внедрение ЦВЗ в уже сформированный объект является одним из способов маркирования; другой

способ предполагает формирование ЦВЗ непосредственно в процессе генерации медиаданных. Объект, содержащий ЦВЗ, далее называется маркированным.

1.1.1 Общая схема внедрения, обнаружения и извлечения ЦВЗ

При маркировании сформированного контента исходными данными являются объект медиаконтента и, если метод предусматривает передачу информации, внедряемое сообщение. Дополнительно могут использоваться ключи и параметры, определяющие способ формирования и последующей проверки ЦВЗ. Общая схема внедрения сообщения устроена следующим образом:

$$x_w = E(x, m; k_e), \quad (1.1)$$

где x — исходный объект; $m \in \{0,1\}^L$ — сообщение длиной L бит; k_e — ключ внедрения; E — алгоритм внедрения; x_w — маркированный объект. Для методов, не использующих ключ, соответствующий аргумент опускается.

Сообщение представляет собой информацию, передаваемую посредством ЦВЗ. Оно может содержать идентификатор источника, получателя или экземпляра контента, а также данные, необходимые для проверки целостности. Следует различать сообщение и его представление в медиаданных: одна и та же битовая последовательность может внедряться с использованием различных сигналов и способов изменения объекта. В нейросетевых методах такое представление может формироваться обучаемым алгоритмом с учетом содержимого исходного объекта [9].

При маркировании в процессе генерации предварительно сформированный объект может отсутствовать. В этом случае общую схему можно представить как

$$x_w = G_w(c, m; k_e), \quad (1.2)$$

где G_w — процедура генерации маркированного объекта; c — условие генерации, например текстовое описание. Аргумент m используется только для методов, поддерживающих передачу сообщения. ЦВЗ может задаваться посредством изменения начальных данных генератора, его параметров или промежуточных представлений. В частности, возможность формирования обнаруживаемого ЦВЗ через структуру начального шума диффузионной модели продемонстрирована в методе Tree-Ring [10].

При хранении, передаче и использовании маркированный объект может подвергаться обработке. Измененный объект представим в виде

$$y = T(x_w), \quad (1.3)$$

где T обозначает отдельное преобразование или последовательность преобразований. В зависимости от модальности к ним относятся сжатие, фильтрация, изменение пространственного разрешения, передискретизация и другие операции.

Следует различать обнаружение и извлечение ЦВЗ. Обнаружение заключается в принятии решения о наличии заданного ЦВЗ. Как правило, оно выполняется путем сравнения числовой оценки с порогом:

$$D(y; k_d) = \begin{cases} 1, & s(y; k_d) \geq \tau, \\ 0, & s(y; k_d) < \tau, \end{cases} \quad (1.4)$$

где s — оценка наличия ЦВЗ; k_d — ключ обнаружения; τ — порог принятия решения. Значение $D = 1$ соответствует положительному решению детектора.

Извлечение предполагает восстановление внедренного сообщения:

$$\hat{m} = X(y; k_d), \quad (1.5)$$

где X — алгоритм извлечения, а $\hat{m}$ — восстановленная битовая последовательность. Точное извлечение соответствует условию $\hat{m} = m$. Алгоритм извлечения может возвращать битовую последовательность и для немаркированного объекта, поэтому сам факт получения сообщения не подтверждает наличие ЦВЗ. Для этого требуется дополнительный критерий, например достаточное совпадение извлеченной последовательности с заранее известным сообщением.

Обозначения k_e и k_d отражают назначение ключей: первый используется при внедрении ЦВЗ, второй — при его обнаружении или извлечении. В зависимости от метода эти ключи могут совпадать или различаться.

1.1.2 Классификация методов маркирования

Методы маркирования различаются по типу передаваемой информации, данным, необходимым для проверки, реакции на преобразования и способу фор-

мирования ЦВЗ. Эти признаки характеризуют разные стороны метода и могут использоваться совместно.

По типу передаваемой информации выделяются методы обнаружения наличия ЦВЗ (*zero-bit*) и методы передачи сообщений (*multi-bit*). В первом случае проверяется только факт наличия заданного ЦВЗ в исследуемом медиаконтенте. Во втором случае маркирование обеспечивает передачу сообщения из некоторого множества допустимых значений. Бинарный результат обнаружения в *zero-bit* схеме не следует отождествлять с передачей однобитового сообщения.

По составу данных, необходимых при проверке, различаются слепые и неслепые методы. При слепом обнаружении или извлечении исходный немаркированный объект не требуется, однако могут использоваться ключ, параметры алгоритма и ожидаемое сообщение. Неслепая проверка предполагает доступ к исходному объекту.

По характеру реакции на изменения медиаданных выделяются устойчивые, хрупкие и полухрупкие ЦВЗ. Устойчивые ЦВЗ предназначены для сохранения возможности обнаружения или извлечения после определенного набора преобразований. Хрупкие ЦВЗ ориентированы на выявление изменений объекта, включая небольшие изменения его цифрового представления. Полухрупкие ЦВЗ предназначены для сохранения положительного результата проверки после допустимых преобразований и выявления изменений, отнесенных к недопустимым. Принадлежность к этим классам определяется относительно заданных условий обработки и требований к проверке [11].

По этапу формирования маркированного объекта различаются методы внедрения ЦВЗ в сформированный контент и методы маркирования в процессе генерации. Первый подход допускает обработку имеющихся медиаданных, тогда как второй предполагает возможность вмешательства в процедуру их создания.

Методы маркирования различаются представлением данных, которое используется для формирования ЦВЗ. В одних методах изменения вносятся непосредственно в значения пикселей изображения или отсчетов аудиосигнала. В других сначала выполняется математическое преобразование, затем изменяются его коэффициенты и восстанавливается маркированный объект путем обратного преобразования. Также ЦВЗ может формироваться в обучаемых представлениях, используемых нейросетевой моделью.

Отдельно различается способ определения операций внедрения и извлечения. В методах с явно заданными правилами разработчик определяет, какие

элементы выбранного представления изменяются, как эти изменения зависят от сообщения и как восстанавливаются его биты. В обучаемых методах для выполнения этих операций используются модели, параметры которых определяются в процессе обучения. Одно и то же представление данных допускает разные способы формирования ЦВЗ. Например, коэффициенты преобразования могут изменяться по заданному правилу квантования либо с помощью обученной нейронной сети. Аналогично, использование нейросетевого генератора не означает, что сам способ формирования ЦВЗ обязательно является обучаемым.

1.1.3 Основные свойства и требования к методам маркирования

К основным свойствам методов маркирования медиаконтента относятся незаметность ЦВЗ, емкость, устойчивость к преобразованиям и преднамеренным атакам и вычислительная эффективность.

Под незаметностью ЦВЗ понимается степень сохранения воспринимаемого качества медиаконтента после внедрения. Для изображений существенны визуальные искажения, для аудиоданных — слышимость внесенных изменений, для видео — также их согласованность между последовательными кадрами. Незаметность не означает побитового совпадения исходного и маркированного объектов. Ее оценка должна учитывать особенности восприятия соответствующего типа медиаданных.

Емкость метода характеризует объем информации, который может быть внедрен в объект. При описании конкретной конфигурации необходимо указывать фактическую длину сообщения и единицу нормирования: объект, пиксель, отсчет или единицу времени. Следует также различать объем полезной информации и полную длину внедряемой последовательности, которая может включать служебные данные и избыточность помехоустойчивого кодирования.

Преобразованием далее называется операция обработки медиаданных, а преднамеренной атакой — действие, направленное на нарушение работы метода маркирования или на получение ошибочного результата проверки. Одна и та же операция, например сжатие, может применяться при обычной обработке или использоваться для удаления ЦВЗ. Для общего обозначения преобразований и атак используется термин «воздействие». Под устойчивостью понимается спо-

способность метода сохранять возможность обнаружения или извлечения ЦВЗ после воздействий на медиаданные. Ее оценка относится к конкретному набору воздействий и диапазонам их параметров. Успешное извлечение в исследованных условиях не гарантирует сохранения ЦВЗ при других видах обработки или атаках.

При исследовании преднамеренных атак необходимо задавать модель нарушителя, определяющую его знания об алгоритме, доступ к процедурам внедрения и извлечения, а также цели и возможности [12]. Сохранение ЦВЗ после обычной обработки не подтверждает стойкость метода к его целенаправленному удалению, переносу или подделке.

Вычислительная эффективность определяется затратами времени и памяти на маркирование, обнаружение и извлечение. Для обучаемых методов следует отдельно учитывать затраты на обучение и применение модели. При обработке аудио- и видеоданных существенны также скорость обработки относительно длительности записи.

Перечисленные свойства необходимо рассматривать совместно. Например, увеличение величины изменений медиаконтента при внедрении ЦВЗ может повышать его устойчивость, но одновременно делать эти изменения заметнее. Требования к емкости, качеству медиаконтента, устойчивости и скорости обработки определяются задачей применения. Для устойчивой идентификации важна сохранность ЦВЗ после обработки, тогда как контроль целостности дополнительно требует различать допустимые и недопустимые изменения.

1.2 Методы маркирования изображений

Методы маркирования изображений разделяются на две группы: внедряющие ЦВЗ в сформированное изображение и формирующие маркированное изображение непосредственно в процессе генерации. В первой группе используются как классические подходы с явно заданными правилами изменения пикселей или коэффициентов преобразований, так и нейросетевые методы, обучаемые внедрению и извлечению ЦВЗ. Методы второй группы предусматривают вмешательство в процедуру генерации изображения.

Примеры применения методов первой группы представлены на рисунке 1.1. Исходное изображение взято из набора данных MS-COCO [13]. Для каждого

метода указана длина внедряемого сообщения в битах; методы без передачи сообщения обозначены как zero-bit.

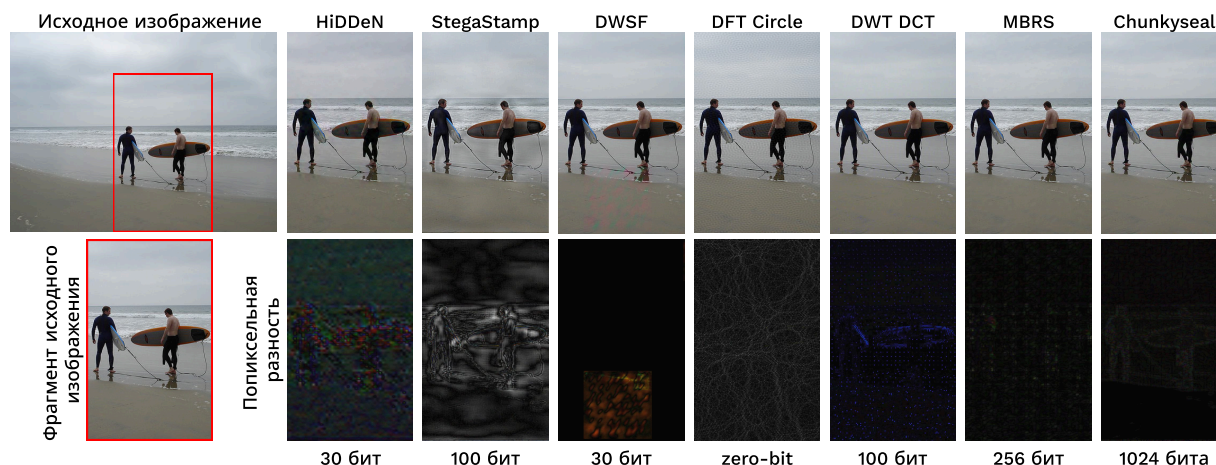


Рисунок 1.1 — Примеры маркирования изображения различными методами. Верхний ряд — маркированные фрагменты, нижний — попиксельная разность относительно исходного изображения, усиленная в 7 раз

1.2.1 Классические методы

Классические методы маркирования изображений основаны на явно заданных правилах изменения пикселей или характеристик изображения, полученных с помощью математических преобразований. Разработка таких методов предполагает выбор области внедрения, способа представления сообщения и процедуры его восстановления. При этом область внедрения и способ кодирования сообщения дополняют друг друга: одна и та же техника модуляции на основе внедряемого сообщения может применяться к пикселям или коэффициентам разных преобразований.

Одним из простейших подходов является замена наименее значащих битов значений пикселей (Least Significant Bit, LSB), ранние варианты которого рассматривались в работе [14]. При замене одного младшего бита целочисленное значение выбранного цветового компонента изменяется не более чем на единицу. Это позволяет внедрять информацию с небольшими визуальными искажениями и низкими вычислительными затратами. Однако младшие биты не сохраняются

при многих распространенных операциях обработки, включая сжатие с потерями, фильтрацию и масштабирование.

Другой подход в пространственной области заключается во внесении распределенной модификации в группы пикселей. В работе [15] для представления ЦВЗ использовались изменения яркости выбранных блоков изображения с помощью псевдослучайного шаблона. Обнаружение основывалось на сопоставлении внесенной модификации с известным шаблоном и требовало исходного изображения. В отличие от замены значений отдельных пикселей, такой подход позволяет распределить информацию между несколькими элементами изображения.

Дальнейшее развитие получили методы маркирования в домене преобразований изображения. В них изображение сначала переводится в другое представление, затем выбранные коэффициенты изменяются в соответствии с внедряемым сообщением, после чего выполняется обратное преобразование. Наиболее распространены дискретное косинусное преобразование (ДКП), дискретное вейвлет-преобразование (ДВП) и дискретное преобразование Фурье (ДПФ).

В методах на основе ДКП ЦВЗ формируется путем изменения частотных коэффициентов изображения или его блоков [16]. Выбор коэффициентов связан с компромиссом между незаметностью и устойчивостью: изменения низкочастотных составляющих могут заметно влиять на визуальное качество изображения, тогда как высокочастотные составляющие сильнее подавляются при сжатии и фильтрации. В ряде схем используются коэффициенты средних частот, позволяющие согласовать эти требования.

ДВП предоставляет многомасштабное представление изображения, разделяя его на поддиапазоны, соответствующие различным пространственным масштабам и направлениям изменения яркости. Это позволяет выбирать области внедрения ЦВЗ с учетом структуры изображения. В комбинированном методе [17] ДКП применяется к блокам выбранных вейвлет-поддиапазонов. Выбор масштаба и поддиапазона дополняется выбором коэффициентов внутри блока. Другие комбинированные схемы используют сингулярное разложение (Singular Value Decomposition, SVD), например для изменения сингулярных значений матриц, сформированных из коэффициентов ДКП [18].

ДПФ применяется при построении методов, ориентированных на сохранение ЦВЗ после поворота, масштабирования и сдвига изображения. Эти гео-

метрические преобразования изменяют пространственное положение элементов изображения и могут нарушать соответствие между областями внедрения и извлечения. В работе [19] для повышения устойчивости к таким изменениям используется подход на основе преобразования Фурье–Меллина.

Помимо области внедрения, важную роль играет способ изменения выбранного представления изображения. Одним из основных направлений являются методы расширения спектра (Spread Spectrum, SS). В работе [20] ЦВЗ представляется псевдослучайной последовательностью, распределенной по множеству значимых спектральных компонентов изображения. Небольшие изменения отдельных коэффициентов совместно формируют сигнал, который обнаруживается по корреляции с известной последовательностью. В предложенной авторами схеме для оценки внедренного сигнала используется исходное изображение, то есть проверка является неслепой.

При слепом корреляционном обнаружении результат проверки зависит не только от внедренного ЦВЗ, но и от корреляции исходного содержимого изображения с выбранной псевдослучайной последовательностью. Такая зависимость может приводить к ошибкам восстановления сообщения. Метод Improved Spread Spectrum (ISS), предложенный в работе [21], учитывает эту корреляцию на этапе внедрения. Интенсивность модификации выбирается с учетом проекции исходного сигнала на последовательность, что позволяет уменьшить влияние содержимого изображения на результат извлечения ЦВЗ. Дальнейшее развитие этого принципа представлено методами CASS и CAISS [22]. При формировании модификации учитываются как корреляция исходного сигнала с последовательностью ЦВЗ, так и значение внедряемого бита.

Другим подходом является модуляция индекса квантования (Quantization Index Modulation, QIM), предложенная в работе [23]. Каждому значению внедряемого символа сопоставляется определенный квантователь. При внедрении выбранная характеристика изображения заменяется результатом квантования по правилу, соответствующему передаваемому биту сообщения. При извлечении определяется, какому из допустимых наборов значений ближе наблюдаемая характеристика. Шаг квантования управляет соотношением между величиной искажений и устойчивостью к изменениям коэффициентов. QIM может применяться в различных представлениях изображения и не привязана к одному конкретному преобразованию.

Для уменьшения заметности ЦВЗ классические методы могут учитывать свойства зрительного восприятия. Модели зрительной системы человека (Human Visual System, HVS) используются для выбора областей внедрения и допустимой интенсивности изменений [24]. В частности, пространственное распределение модификации может адаптироваться к яркости, текстуре и локальным деталям изображения. Такая адаптация позволяет учитывать неодинаковую заметность равных по величине изменений в разных областях изображения.

Классические методы предоставляют явно заданные механизмы управления областью внедрения, интенсивностью модификации и способом восстановления сообщения. Вместе с тем их устойчивость определяется выбранным представлением и моделью предполагаемой обработки. Геометрические преобразования могут нарушать синхронизацию, а сжатие и фильтрация — изменять коэффициенты, несущие ЦВЗ.

1.2.2 Нейросетевые методы

Нейросетевые методы позволяют определять операции внедрения и извлечения ЦВЗ в процессе обучения. В распространенной схеме используются две нейронные сети: кодировщик, формирующий маркированное изображение по исходному изображению и сообщению, и декодер, восстанавливающий сообщение. При обучении между ними располагается слой преобразований, моделирующий обработку маркированного изображения. Функция потерь учитывает различие между исходным и маркированным изображениями, а также ошибки восстановления сообщения. Одним из первых методов, реализующих такую схему, стал HiDDeN [9]. Его особенность — дополнительное состязательное обучение: нейронная сеть-дискриминатор различает маркированные и немаркированные изображения, побуждая кодировщик формировать менее заметные изменения.

В ReDMark [25] используются две совместно обучаемые сверточные сети для внедрения и извлечения сообщения. Сеть внедрения формирует добавку к исходному изображению, величина которой регулируется коэффициентом силы внедрения. Информация об отдельных битах распределяется между несколькими блоками изображения, что направлено на сохранение возможности извлечения при повреждении части его областей.

Существенной проблемой такого обучения является использование недифференцируемых преобразований. Если операция между кодировщиком и декодером не позволяет вычислить градиент, непосредственное обновление параметров кодировщика по ошибке извлечения становится затруднительным. В частности, эта проблема возникает при моделировании JPEG-сжатия. В методе MBRS [26] предложено сочетать реальное JPEG-сжатие, его дифференцируемую имитацию и отсутствие искажений. Дифференцируемая имитация позволяет обучать кодировщик с учетом ошибки восстановления сообщения, а реальное сжатие используется для адаптации декодера к фактическим искажениям JPEG. Дополнительно авторы рассматривают распределение информации сообщения по изображению для повышения устойчивости к обрезке.

Другим направлением является адаптация внедряемой модификации к содержанию изображения. В ARWGAN [27] для этого используются механизм внимания и объединение признаков сообщения и изображения. Они позволяют учитывать различия между пространственными областями при формировании ЦВЗ. Составляющая обучения дополняет ограничения на искажение изображения.

В ряде методов внедрение и извлечение сообщения связываются через прямые и обратные преобразования нейросетевых блоков с общими параметрами. В CIN [28] обратимая часть сети дополняется необратимым декодером с механизмом внимания, предназначенным для восстановления сообщения после обработки с потерями. Специальный модуль определяет наличие JPEG-сжатия и выбирает результат извлечения соответствующей ветви. В FIN [29] используется последовательность обратимых нейросетевых блоков для внедрения и извлечения сообщения. Дополнительный обучаемый модуль моделирует внешние искажения при обучении, а его обратное преобразование применяется перед извлечением для уменьшения их влияния.

Нейронная сеть может использоваться не только как непосредственно обучаемый кодировщик, но и как средство построения пространства признаков для внедрения ЦВЗ. В методе SSL [30] применяется представление изображения, полученное предварительно обученной моделью без учителя. Внедрение выполняется посредством итеративного изменения изображения таким образом, чтобы признаки полученного представления удовлетворяли условию, соответствующему заданному ЦВЗ. При этом ограничивается отличие маркированного изображения от исходного. Подход использует устойчивость признакового представления

к преобразованиям, но требует оптимизации для каждого маркируемого изображения, что увеличивает вычислительные затраты.

В RoSteALS [31] сообщение внедряется в латентное представление предварительно обученного автокодировщика [32]. Дополнительный обучаемый модуль формирует изменение этого представления, после чего декодер автокодировщика восстанавливает маркированное изображение. Извлечение сообщения выполняется отдельной нейронной сетью. Такая организация позволяет использовать готовое представление изображений, однако качество результата зависит также от точности реконструкции автокодировщика.

Практическая применимость нейросетевых методов связана с возможностью обработки изображений произвольного разрешения. В TrustMark [33] используется перенос модификации между разрешениями. Исходное изображение приводится к рабочему размеру модели, для него вычисляется модификация, необходимая для внедрения ЦВЗ, после чего эта модификация масштабируется до исходного разрешения и добавляется к исходному изображению. Такой подход позволяет сохранять детали исходного изображения, не пропуская его полно-размерную версию через кодировщик. В InvisMark [34] также рассматривается маркирование изображений высокого разрешения с акцентом на незаметность, обеспечиваемую архитектурой модели и сочетанием функций потерь.

В VideoSeal [35] используется единая пара нейросетевых моделей для маркирования изображений и видеокадров. Алгоритм внедрения построен на основе U-Net [36] и формирует модификацию изображения с учетом передаваемого сообщения, а алгоритм извлечения использует визуальный трансформер для восстановления битовой последовательности. Обучение включает предварительную подготовку на изображениях, совместное обучение на изображениях и видео, а затем дополнительное обучение декодера при фиксированных параметрах модели внедрения.

В PixelSeal [37] основное внимание уделяется согласованию незаметности и устойчивости ЦВЗ в процессе обучения. Для ограничения заметности изменений авторы используют состязательную функцию потерь вместо функций потерь, непосредственно измеряющих попиксельные и перцептивные различия между исходным и маркированным изображениями. При этом сохраняется функция потерь восстановления внедренного сообщения. Обучение организовано в три этапа: сначала формируется устойчивый, но заметный ЦВЗ, затем снижается интенсивность модификации и вводится состязательная оптимизация незаметности.

Для изображений высокого разрешения масштабирование модификации учитывается уже при обучении. Дополнительно применяется пространственная карта на основе модели порога заметности (Just-Noticeable Difference, JND) [38], ослабляющая изменения в чувствительных для восприятия человеком областях.

Отдельную задачу представляет устойчивость ЦВЗ к геометрическим преобразованиям. Изменение масштаба, поворот и обрезка нарушают пространственное соответствие между областями внедрения и входными данными декодера. Метод DWSF [39] предусматривает внедрение сообщения в несколько областей изображения, их последующее обнаружение и согласование перед извлечением. Результаты, полученные из нескольких сохранившихся областей, объединяются. Такая избыточность позволяет восстанавливать сообщение при частичной утрате маркированных участков. В SyncSeal [40] геометрическая синхронизация рассматривается как самостоятельная задача: внедренный ЦВЗ используется для оценки пространственного преобразования и восстановления соответствия координат.

Локальное обнаружение и извлечение развиваются в методе WAM [41]. Его декодер определяет маркированные области и восстанавливает сообщения на уровне пространственных позиций изображения. Это позволяет работать с композициями, содержащими фрагменты с различными сообщениями, и объединять локальные оценки для восстановления ЦВЗ каждого фрагмента.

Наряду с обработкой файлов рассматриваются сценарии передачи маркированного изображения через физическую среду. StegaStamp [42] ориентирован на восстановление сообщения после печати и последующей съемки изображения. При обучении моделируются перспективные преобразования, размытие, изменения цвета и другие искажения, характерные для такого сценария. В PIMoG [43] рассматривается моделирование искажений при фотографировании изображения с экрана. Задача устойчивого маркирования при фотографировании экрана исследовалась также для изображений текстовых документов: в работе [1] ЦВЗ формируется посредством изменения яркости межстрочных интервалов, а в Docmarking [2] для формирования и извлечения водяного знака используются нейросетевые модели.

Развитие нейросетевого редактирования привело к появлению методов, специально ориентированных на сохранение ЦВЗ после изменения содержимого изображения. В Robust-Wide [44] рассматривается устойчивость к редактированию по текстовому описанию. Включение такого сценария в разработку метода

расширяет набор учитываемых воздействий по сравнению со сжатием, фильтрацией и геометрической обработкой.

Отдельное направление исследований связано с увеличением длины внедряемого сообщения. В работе [45] предложен ChunkySeal, разработанный на основе VideoSeal. Для повышения емкости авторы увеличили размер кодировщика и декодера, а внедрение выполняют во все три цветовых канала изображения. Полученная модель поддерживает сообщения длиной 1024 бита вместо 256 бит у сравниваемой конфигурации VideoSeal, сохраняя высокое качество изображений и сопоставимую устойчивость к рассмотренным преобразованиям. Вместе с тем увеличение размера нейронных сетей повышает вычислительные затраты на внедрение и извлечение ЦВЗ.

1.2.3 Маркирование изображений в процессе генерации

Методы маркирования в процессе генерации предусматривают формирование ЦВЗ непосредственно при синтезе изображения. Для этого могут изменяться параметры генеративной модели, ее входные данные или промежуточные представления. В результате изображение содержит ЦВЗ уже на выходе генератора. Возможности такого внедрения зависят от архитектуры модели и доступных средств управления генерацией. Далее рассматриваются методы маркирования в процессе генерации изображений диффузионными моделями.

В Stable Signature [46] внедрение ЦВЗ в процессе генерации изображения реализовано для латентных диффузионных моделей. Такие модели выполняют последовательное устранение шума в латентном пространстве, после чего декодер автокодировщика преобразует полученное представление в изображение. Авторы дообучают этот декодер так, чтобы из создаваемых изображений можно было извлечь заданную битовую последовательность с помощью предварительно обученной нейросетевой модели для извлечения ЦВЗ. При обучении учитываются сохранение качества изображения и восстановление сообщения.

Другое направление позволяет внедрять ЦВЗ без дообучения генеративной модели. Оно основано на управлении начальным шумом, из которого формируется изображение. В процессе обнаружения ЦВЗ исследуемое изображение преобразуется в латентное представление, а затем выполняется приближенная ин-

версия процесса генерации. Полученная оценка начального шума используется для проверки ЦВЗ или восстановления сообщения. Точность такой проверки зависит как от сохранности ЦВЗ после обработки изображения, так и от неточности процесса инверсии.

В Tree-Ring [10] ЦВЗ задается структурированным шаблоном в области преобразования Фурье начального шума диффузионной модели. В частности, рассматриваются шаблоны в форме концентрических колец. После генерации информация о шаблоне сохраняется в изображении и может быть обнаружена в частотном представлении приближенно восстановленного шума. Основным рассматриваемый сценарий заключается в проверке наличия заданного шаблона. Кольцевая структура используется для повышения устойчивости к геометрическим преобразованиям.

Метод RingID [47] развивает подход Tree-Ring для идентификации ЦВЗ среди заранее заданного набора шаблонов. При проверке изображения приближенно восстанавливается начальное латентное представление, после чего содержащийся в нем ЦВЗ сопоставляется с эталонными шаблонами. Результатом является идентификатор наиболее близкого шаблона. Дополнительно в разных каналах начального латентного представления размещаются кольцевой шаблон и шаблон, сформированный из гауссовского шума. Их совместное использование позволяет учитывать различия в устойчивости к преобразованиям изображения.

В METR [48] кольцевые структуры используются для представления многобитовых сообщений. В рассмотренной авторами конфигурации передается 16-битная последовательность, что позволяет задать 2^{16} различных идентификаторов. Расширение METR++ объединяет этот способ с маркированием посредством Stable Signature: идентификатор, задаваемый кольцевым шаблоном, дополняется меткой, связанной с настроенным декодером генератора. Такая комбинация увеличивает число различимых идентификаторов, но требует поддержки нескольких вариантов параметров модели-декодера.

Отдельной задачей является восстановление пространственного соответствия после геометрической обработки. В MaXsive [49] функции синхронизации и представления передаваемой информации разделены. Для оценки геометрических изменений используется X-образный шаблон, формируемый в частотной области в ходе генерации. После инверсии изображения этот шаблон позволяет оценить и компенсировать преобразование перед проверкой основной части ЦВЗ. Основная часть ЦВЗ представляется набором вещественных значений, который

после восстановления сопоставляется с заранее известной последовательностью-ключом.

Изменение начального шума может приводить к отклонению его распределения от того, для которого обучалась генеративная модель. Это создает дополнительное требование к способу внедрения: ЦВЗ должен сохранять не только возможность обнаружения, но и свойства входных данных генератора. На решение этой задачи направлен метод Gaussian Shading [50]. Сообщение предварительно преобразуется в псевдослучайную битовую последовательность, определяющую выбор интервалов нормального распределения. Значения начального шума выбираются из соответствующих интервалов, благодаря чему сообщение кодируется в шуме без введения явно выраженного частотного шаблона. При извлечении выполняется инверсия изображения, определяется принадлежность восстановленных значений заданным интервалам и восстанавливается сообщение. Такой способ формирования шума сохраняет его исходное распределение и, следовательно, распределение выходных изображений неизменного генератора.

Близкое направление связано с применением псевдослучайных помехоустойчивых кодов (Pseudorandom Error-Correcting Codes, PRC). В работе [51] такие коды используются для формирования начального латентного представления диффузионной модели. Кодовые последовательности допускают восстановление после определенных искажений и при этом вычислительно неотличимы от случайных для стороны, не располагающей необходимым ключом. На этой основе авторы обосновывают вычислительную неразличимость маркированных и немаркированных изображений в принятой модели безопасности.

В Gaussian Shading++ [52] рассматривается развитие схемы Gaussian Shading для многократного использования в практической системе маркирования. Подход объединяет передачу сообщения с передачей вспомогательных данных, задающих уникальный ЦВЗ для каждого изображения, и использует PRC-коды. Такая организация направлена на устранение ограничений повторного применения одной и той же схемы случайного преобразования сообщения. Дополнительно авторы рассматривают применение цифровой подписи для повышения защищенности процедуры идентификации.

В CLUE-MARK [53] начальный шум диффузионной модели формируется с использованием секретного вектора. В проекциях блоков его вейвлет-представления на этот вектор создается периодическая статистическая структура, которая служит признаком ЦВЗ. При проверке начальный шум приближенно

восстанавливается из изображения, после чего оценивается наличие указанной структуры. Вычислительная неразличимость маркированного и обычного гауссовского шума без знания ключа обосновывается предположением о сложности задачи Continuous Learning With Errors (CLWE) [54].

1.3 Методы маркирования аудиоданных

В разделе рассматриваются классические и нейросетевые методы маркирования аудиоданных. Основное внимание уделяется способам внедрения ЦВЗ во временной, частотной и временно-частотной областях, снижению заметности вносимых изменений и устойчивости к обработке сигнала. Также рассматриваются способы сохранения возможности извлечения при нарушениях временной синхронизации.

1.3.1 Классические методы

В классических методах правила изменения аудиосигнала и восстановления сообщения задаются явно. Как и для изображений, область внедрения и способ представления битов являются разными признаками классификации: например, расширение спектра и квантование могут применяться к различным представлениям аудио.

Ряд подходов, рассмотренных в обзоре методов маркирования изображений, применяется также к аудиоданным. В частности, метод LSB используется для замены младших битов представления отсчетов аудиосигнала, а QIM — для внедрения сообщения посредством квантования отсчетов или вычисляемых признаков.

Аналогично методы расширения спектра помимо изображений применимы и к аудиосигналам. Они распределяют представление сообщения по множеству отсчетов или спектральных коэффициентов [55]. Бит управляет знаком или выбором псевдослучайной последовательности, которая добавляется к сигналу с ограниченной амплитудой. При извлечении оценивается корреляция исследуемо-

го сигнала с соответствующей последовательностью. Для аудио сила внедрения может изменяться с учетом спектрального состава и слухового маскирования.

В методе Echo Hiding [56] информация кодируется задержкой добавляемого эха: двум значениям бита соответствуют две различные задержки. Задержку, амплитуду и затухание выбирают так, чтобы эхо не воспринималось как отдельный звук. Для передачи последовательности битов сигнал делят на участки, а переходы между вариантами эха сглаживают. При извлечении сравнивают значения автокорреляции кепстра в точках, соответствующих двум заданным задержкам.

Ряд классических методов использует дискретное косинусное преобразование и вейвлет-преобразования. Они позволяют изменять выбранные частотные компоненты и учитывать распределение энергии сигнала. В работе [57] внедрение ЦВЗ выполняется посредством модификации векторов коэффициентов ДКП с учетом воспринимаемого качества аудио. Вариант DCT-b1 использует первую из рассматриваемых авторами полос. Выбор полосы определяет как характер вносимых изменений, так и чувствительность ЦВЗ к фильтрации и сжатию.

В работе [58] предложен метод Patchwork, в котором биты сообщения внедряются путем изменения средних значений выбранных групп коэффициентов ДКП. Особенностью метода является многослойное внедрение: в один аудиосигнал можно последовательно добавлять несколько ЦВЗ, каждый из которых несет отдельное сообщение. Правила изменения коэффициентов разработаны с учетом возможности извлечения сообщений из других слоев.

В методе Norm-Space [59] бит сообщения кодируется соотношением евклидовых норм двух векторов коэффициентов. Каждый фрагмент аудиосигнала подвергается вейвлет-преобразованию, после чего к коэффициентам приближения применяется ДКП. Полученные коэффициенты разделяются на две группы по четности индексов. При внедрении их нормы изменяются относительно общего среднего: для единицы норма первой группы становится больше нормы второй, а для нуля — меньше. Заданные нормы достигаются масштабированием коэффициентов каждой группы. При извлечении повторяются преобразования и разбиение на группы, после чего бит определяется сравнением их норм.

В методе FSVC [60] используется признак, основанный на отношении сингулярных значений двух последовательных фрагментов аудиозаписи. Для его вычисления выполняется ДКП и сингулярное разложение сформированных по его коэффициентам матриц. При внедрении отношение изменяется в соответствии

с передаваемым битом, а при извлечении оценивается повторно. Выбор такого признака направлен на повышение устойчивости к нарушениям синхронизации.

1.3.2 Нейросетевые методы

В нейросетевых методах правила внедрения и обнаружения ЦВЗ определяются в процессе обучения. Распространена схема, включающая кодировщик, модуль преобразований аудио и декодер. Обучение направлено на восстановление сообщения после моделируемых воздействий при ограничении изменений исходного сигнала. Нейронные сети могут работать непосредственно с отсчетами или с временно-частотным представлением аудио.

Примером совместного обучения кодировщика и декодера является метод RobustDNN [61]. В процессе обучения маркированные сигналы подвергаются преобразованиям, включая добавление шума, низкочастотную фильтрацию и обнуление части отсчетов. Декодер обучается восстанавливать сообщение после этих воздействий, а кодировщик — формировать изменения, удовлетворяющие одновременно требованиям к качеству сигнала и устойчивости ЦВЗ.

В WavMark [62] внедрение и извлечение реализованы на основе обратимой нейронной сети с общими параметрами для прямого и обратного преобразований. Аудиосигнал переводится во временно-частотное представление посредством кратковременного преобразования Фурье (Short-Time Fourier Transform, STFT). Нейронная сеть объединяет это представление с сообщением, после чего обратное преобразование Фурье формирует маркированный сигнал. WavMark обрабатывает фрагменты длительностью одну секунду и допускает повторное внедрение сообщения на протяжении записи. Для поиска маркированных фрагментов используется перемещение окна извлечения и проверка синхронизирующей части восстановленной последовательности. При обучении моделируются небольшие смещения окна и преобразования сигнала. В методе совместно решаются задачи восстановления сообщения и поиска его положения, а часть емкости расходуется на решение задачи синхронизации.

AudioSeal [63] ориентирован на обнаружение маркированных участков аудио. Генератор формирует добавляемый к исходному сигналу ЦВЗ, а детектор выдает оценку его наличия для каждого отсчета. Для обучения локальному

обнаружению в записи формируются участки с ЦВЗ и без него. Дополнительно применяются преобразования аудио и функция потерь, учитывающая особенности слухового восприятия. Такая организация позволяет получить как общее решение о наличии ЦВЗ, так и его локализацию во времени. В многобитовом варианте AudioSeal дополнительно извлекается короткое сообщение.

В SilentCipher [64] особое внимание уделяется ограничению слышимых изменений. Внедрение выполняется в амплитудное представление STFT с сохранением исходной фазы при восстановлении сигнала. Величина изменений ограничивается с учетом амплитуд исходных спектральных компонентов, что направляет внедрение в области, где ЦВЗ меньше влияет на исходное содержимое. Для повышения устойчивости SilentCipher обучается с применением преобразований сигнала. Недифференцируемые операции сжатия с потерями включаются в обучение посредством обхода этих операций при обратном распространении градиента. Сообщение повторяется вдоль временной оси, а результаты извлечения нескольких сообщений объединяются.

Другой подход к формированию ЦВЗ представлен в AWARE [65]. Внедрение выполняется посредством оптимизации изменений временно-частотного представления, направленной на получение заданного сообщения на выходе нейросетевого детектора. Изменения ограничиваются относительно уровня исходных спектральных компонентов. Детектор объединяет информацию по временной оси таким образом, чтобы уменьшить зависимость результата от порядка и точного положения анализируемых фрагментов. Этот подход направлен на сохранение возможности извлечения после обрезки и других нарушений временной структуры.

Вариант Perth-Net Implicit, реализованный в составе инструмента Perth [66], предназначен для нейросетевого обнаружения ЦВЗ без передачи произвольного битового сообщения. Алгоритм внедрения модифицирует аудиосигнал, а детектор формирует оценку наличия водяного знака. Такая схема позволяет установить, содержит ли запись распознаваемый ЦВЗ, но не предусматривает извлечения пользовательского идентификатора или других данных.

Отдельное направление связано с сохранением ЦВЗ при клонировании голоса. Маркирование выполняется до использования исходных записей для обучения модели синтеза или в качестве голосового образца при генерации. Задача заключается в обнаружении или извлечении ЦВЗ из новой речи, синтезированной с использованием этих записей, в том числе при изменении произносимого текста.

В Timbre Watermarking [67] сообщение внедряется в амплитудную спектрограмму аудиосигнала. Представление сообщения повторяется вдоль временной оси, а при извлечении полученные признаки усредняются по времени. Это уменьшает зависимость восстановления сообщения от расположения отдельных фрагментов записи. При совместном обучении кодировщика и декодера применяются операции, характерные для обработки данных при клонировании голоса: нормализация амплитуды, получение мел-спектрограммы и восстановление звукового сигнала.

В VoiceMark [68] рассматривается клонирование голоса по короткому образцу без дообучения модели клонирования на записях данного диктора. Для внедрения используется предварительно обученный нейронный аудиокодек, разделяющий представления содержания речи и характеристик голоса. ЦВЗ внедряется в представления, связанные с характеристиками голоса, после чего восстанавливается маркированный аудиосигнал. Такой выбор области внедрения направлен на перенос водяного знака в синтезированную речь вместе с особенностями голоса. При обучении применяются перестановка, обнуление и замена временных фрагментов, а также обработка нейросетевым аудиокодеком. Декодер объединяет информацию из всей записи для восстановления сообщения.

В AudioMarkNet [69] рассматривается адаптация модели синтеза речи к голосу диктора посредством дообучения на маркированных записях. ЦВЗ формируется так, чтобы модель могла усвоить его структуру при дообучении и воспроизводить ее при генерации новых фраз. Наличие водяного знака в полученной речи затем проверяется детектором. В исследованной постановке используется заранее заданный набор дикторов, а распространение метода на дикторов, не представленных при обучении системы маркирования, отмечается авторами как направление дальнейшей работы.

1.4 Методы маркирования видеоданных

Видео представляет собой последовательность пространственно и временно связанных кадров, вследствие чего задача внедрения цифровых водяных знаков в видеоданные имеет ряд особенностей по сравнению с маркированием отдельных изображений. Помимо незаметности и устойчивости ЦВЗ, существен-

ное значение имеют вычислительная эффективность обработки большого числа кадров, согласованность вносимых изменений во времени и устойчивость к характерным для видео преобразованиям. К последним относятся сжатие видео с потерями, изменение разрешения и частоты кадров, удаление, вставка или повторение отдельных кадров.

Существующие подходы к маркированию видеоданных различаются способом учета временной структуры и степенью зависимости от архитектуры алгоритма внедрения. В качестве основных направлений можно выделить независимое маркирование отдельных кадров, распространение сформированной для одного кадра модификации на соседние кадры, совместную обработку группы кадров, а также применение специализированных методов, изначально разработанных для видеоданных.

Наиболее прямой способ заключается в независимом применении метода маркирования изображений к каждому кадру видеопоследовательности. Основным преимуществом такого подхода является возможность непосредственного использования уже разработанного алгоритма без изменения его архитектуры или дополнительного обучения. Это позволяет переносить на видео широкий класс существующих методов маркирования изображений. Вместе с тем алгоритм внедрения должен выполняться для каждого кадра, поэтому вычислительные затраты возрастают пропорционально числу кадров. Кроме того, независимое формирование модификации для соседних кадров не учитывает их временную связанность и может приводить к изменению характера внесенных искажений от кадра к кадру.

Для уменьшения вычислительной стоимости используются подходы, основанные на распространении одной модификации на несколько соседних кадров. Такой принцип реализован, в частности, в методе VideoSeal [35]. Videопоследовательность разбивается на небольшие группы, а алгоритм внедрения применяется только к первым кадрам каждой группы. Полученная модификация затем используется при формировании нескольких последующих маркированных кадров. За счет этого уменьшается число запусков алгоритма внедрения и используется временная избыточность видео. Однако такой подход предполагает, что модификация, сформированная для первого кадра, остается применимой к соседним кадрам. При значительных изменениях содержимого или движении объектов увеличение интервала распространения может приводить к снижению точности извлечения ЦВЗ и появлению заметных временных артефактов.

Другой подход состоит в совместной обработке нескольких кадров одного видеосегмента. В ItoV [70] последовательные кадры представляются в виде единого входа нейронной сети, построенной на основе принципов маркирования изображений. Для повышения временной согласованности в функцию обучения дополнительно вводится ограничение на различия интенсивности модификаций между соседними кадрами. Такой способ позволяет учитывать информацию сразу нескольких кадров, однако требует изменения представления входных данных и специального обучения модели применительно к видеоданным.

Использование временной избыточности возможно также на уровне внутренних признаков представлений нейронной сети. В PixelSeal [37] признаки нескольких соседних кадров агрегируются в промежуточных слоях архитектуры U-Net, после чего часть вычислений выполняется над общим представлением группы. На заключительных этапах обработки формируются отдельные модификации для каждого кадра. Такой подход позволяет уменьшить объем повторяющихся вычислений при обработке сходных соседних кадров и при этом сохранить возможность покадрового формирования маркированного видео.

Отдельную группу образуют алгоритмы, изначально разработанные для маркирования видеоданных. К ним относятся RivaGAN [71], DVMark [72], REVMARK [73] и другие методы, в которых пространственно-временные свойства видео учитываются непосредственно архитектурой модели или процедурой ее обучения. В RivaGAN для выбора областей внедрения используется механизм внимания, а обучение направлено одновременно на обеспечение визуального качества и устойчивости ЦВЗ. DVMark использует многомасштабное пространственно-временное представление и моделирование преобразований в процессе обучения. REVMARK, в свою очередь, ориентирован на повышение устойчивости ЦВЗ к видеокодированию и характерным для видео временным преобразованиям.

1.5 Методы контроля целостности медиаконтента на основе ЦВЗ

В зависимости от назначения системы целостность медиаконтента может пониматься как точное сохранение цифрового представления либо как сохранение содержимого при допустимой обработке. Во втором случае заранее опре-

деляются допустимые преобразования и диапазоны их параметров. Например, сжатие изображения или перекодирование аудиозаписи могут считаться допустимыми, тогда как замена объекта на изображении или фрагмента аудиосигнала должна приводить к отрицательному результату проверки целостности. Дополнительной задачей может являться локализация модификаций, то есть определение измененных областей изображения, временных интервалов аудио или участков видеопоследовательности.

Одним из подходов к контролю целостности являются хрупкие и полухрупкие ЦВЗ. Хрупкие ЦВЗ предназначены для выявления изменений маркированного объекта и, как правило, обладают высокой чувствительностью к его модификациям [74]. Такие ЦВЗ не предназначены для сохранения после обработки объекта: даже небольшие изменения его цифрового представления должны приводить к нарушению результата проверки.

Полухрупкие ЦВЗ ориентированы на возможность различить допустимую обработку и недопустимые модификации. Их проверка должна сохранять положительный результат после заданных преобразований. Для этого могут использоваться устойчивые признаки содержимого или их квантованные представления, слабо изменяющиеся при допустимой обработке [75]. При этом устойчивость и чувствительность полухрупкого ЦВЗ определяются относительно конкретного набора преобразований и модификаций, рассматриваемых при разработке и экспериментальной оценке метода.

В нейросетевых методах требуемый ответ на различные преобразования формируется в процессе обучения. В SepMark [76] один кодировщик внедряет в изображение единое сообщение, для извлечения которого используются два декодера с различным уровнем устойчивости. Первый декодер обучается восстанавливать сообщение как после стандартных преобразований, так и после рассматриваемых недопустимых модификаций и используется для прослеживания источника изображения. Второй декодер сохраняет возможность извлечения после допустимой обработки, но должен терять возможность извлечь внедренное сообщение после недопустимых модификаций. Совместный анализ результатов двух декодеров позволяет одновременно подтвердить наличие исходного ЦВЗ и определить факт изменения изображения.

В EditGuard [77] задачи устойчивой передачи информации и локализации изменений решаются посредством одновременного внедрения двух типов данных: пространственного ЦВЗ и сообщения. Пространственный ЦВЗ представляет

изображение заданной структуры и предназначен для контроля локальных изменений, тогда как сообщение используется для передачи информации. После возможной модификации из изображения на этапе извлечения восстанавливаются оба представления. Локальные отклонения восстановленного пространственного ЦВЗ используются для построения маски измененных областей, а битовый декодер независимо восстанавливает переданное сообщение.

OmniGuard [78] развивает данный подход, сохраняя разделение между пространственным ЦВЗ для локализации и устойчивым битовым сообщением. В отличие от EditGuard, где для локализации используется заранее определенное пространственное представление, OmniGuard допускает более гибкий выбор такого ЦВЗ и объединяет маркирование с нейросетевым анализом полученного изображения. Для формирования маски изменений используется отдельная сеть, учитывающая одновременно восстановленный пространственный ЦВЗ и признаки исследуемого изображения, что позволяет сохранять локализацию при дополнительных искажениях маркированного объекта.

Способность выявлять изменения содержимого не исключает атак на сам механизм маркирования. Устойчивый ЦВЗ может быть перенесен на другой объект или сохранен после недопустимой модификации, а специально сформированное воздействие может приводить к ложному положительному результату проверки [79; 80]. Для защиты от таких сценариев аутентификационные данные могут формироваться в зависимости от содержимого объекта и дополняться цифровой подписью [81]. При публичной проверке подпись формируется с использованием закрытого ключа и проверяется соответствующим открытым ключом.

Современным примером такого подхода является MetaSeal [82]. В данном методе для изображения сначала формируется контентно-зависимое семантическое представление в виде автоматически получаемого текстового описания, для которого вычисляется цифровая подпись. Описание и подпись совместно кодируются в виде QR-кода и внедряются в изображение с помощью обратимой нейронной сети. При проверке встроенные данные восстанавливаются, подпись проверяется по открытому ключу, а соответствие текстового описания содержимому изображения используется для контроля его целостности. Такой подход позволяет связать ЦВЗ с содержимым изображения и стороной, владеющей соответствующим закрытым ключом, однако успешность проверки зависит от того, насколько используемое семантическое описание отражает значимые изменения содержимого.

1.6 Средства оценки методов маркирования медиаконтента

Существующие средства сравнительной оценки методов маркирования в значительной степени специализированы относительно отдельных модальностей медиаконтента. При сохранении общей задачи оценки незаметности, устойчивости и успешности обнаружения или извлечения ЦВЗ постановка эксперимента существенно зависит от типа данных. Вместе с тем появляются и мультимодальные программные системы, в которых оценка методов для различных типов медиаконтента выполняется в рамках общей экспериментальной инфраструктуры. Далее рассматриваются как специализированные средства оценки методов маркирования изображений и аудиоконтента, так и мультимодальные решения.

1.6.1 Средства сравнительной оценки методов маркирования изображений

Значительное число специализированных средств сравнительной оценки представлено в области маркирования изображений. Одним из первых систематических подходов к сравнительной оценке цифровых водяных знаков является StirMark Benchmark [83]. Он задает воспроизводимый набор экспериментов для исследования устойчивости ЦВЗ к сжатию, геометрическим преобразованиям, фильтрации и добавлению шума. StirMark позволил проводить сопоставимое тестирование различных алгоритмов, однако ориентирован прежде всего на классические преобразования изображения и не содержит реализации самих методов маркирования.

Бенчмарк WAVES [84] предназначен для сравнительной оценки устойчивости методов маркирования изображений к различным воздействиям. Набор экспериментов включает обычные преобразования изображений, такие как сжатие, обрезка, размытие и изменение яркости, а также преднамеренные атаки. К последним относятся регенерация изображения с использованием диффузионных моделей или вариационных автокодировщиков и состязательный подбор изменений, направленных на нарушение работы детектора ЦВЗ. Основное внимание при оценке уделяется устойчивости ЦВЗ и качеству изображений после воздействий;

влияние внедрения ЦВЗ на качество маркированного изображения не рассматривается.

W-Bench [85] предназначен для оценки ЦВЗ после генеративного редактирования изображений. В нем реализованы регенерация, глобальное и локальное изменение содержимого, а также преобразование изображения в видео. В таких условиях семантическое содержание может сохраняться при значительном изменении изображения, что отличает эти воздействия от фильтрации или сжатия изображений. Бенчмарк задает воспроизводимое сравнение в выбранных сценариях, но рассматривает фиксированный набор методов внедрения ЦВЗ в готовые изображения и не предоставляет общего механизма построения произвольных экспериментов. Опубликованная программная реализация поддерживает только один метод маркирования изображений.

MarkDiffusion [86] предназначен для сравнительной оценки методов формирования ЦВЗ при генерации изображений и видео диффузионными моделями. Система объединяет реализации алгоритмов маркирования, средства применения воздействий, вычисления метрик и визуализации результатов. Предусмотрена оценка обнаружения ЦВЗ, его устойчивости и качества сгенерированного медиаконтента. При этом штатный набор включает методы, использующие доступ к процессу генерации. Алгоритмы внедрения ЦВЗ в готовые изображения в нем не представлены.

Более общий подход реализован в системе OmniSeal Bench [87], предназначенной для сравнительной оценки методов маркирования изображений, аудио- и видеоконтента. Система предоставляет общую программную инфраструктуру для внедрения ЦВЗ, применения воздействий, обнаружения или извлечения ЦВЗ и вычисления метрик. Эксперименты могут выполняться с использованием программного интерфейса или интерфейса командной строки; предусмотрены конфигурационные описания методов, сохранение и агрегация результатов, а также выполнение экспериментов на вычислительных кластерах. Для изображений в системе непосредственно представлены десять методов маркирования и набор типовых преобразований. Таким образом, OmniSeal Bench, в отличие от рассмотренных выше специализированных решений, реализует общую экспериментальную инфраструктуру для нескольких модальностей медиаконтента. Вместе с тем система ориентирована преимущественно на современные нейросетевые методы, а представленный для изображений набор воздействий в основном охватывает стандартные преобразования и в меньшей степени специализирован-

ные преднамеренные атаки удаления и подделки ЦВЗ. Более того, OmniSeal Bench не поддерживает методы, внедряющие ЦВЗ в процессе генерации изображений.

Характеристики рассмотренных средств приведены в таблице 1. В сравнение также включена разработанная в рамках настоящей диссертации система WIBE. Она предназначена для воспроизводимой оценки методов маркирования различных типов медиаданных и позволяет организовать внедрение ЦВЗ, применение воздействий, обнаружение или извлечение сообщения и вычисление метрик в рамках общего экспериментального конвейера. В таблицах приведены характеристики реализованного набора компонентов WIBE. Архитектура системы и организация экспериментов подробно рассматриваются в следующей главе.

Таблица 1 — Сравнение средств оценки методов маркирования изображений

Решение	Методы / Воздействия / Метрики	Область применения	Ограничения
StirMark	— / 20 / 4	Устойчивость ЦВЗ к классическим преобразованиям изображений	Только изображения; реализации тестируемых методов подключаются пользователем; набор воздействий ориентирован на классическую обработку изображений
WAVES	5 / 20 / 9	Устойчивость ЦВЗ и качество изображений после воздействий	Только изображения; отсутствует оценка качества маркированных изображений
W-Bench	1 + 11 / 13 / 7	Устойчивость к генеративному редактированию изображений	Только изображения; поддерживается реализация только 1 метода; результаты тестирования 11 методов приведены авторами без программной реализации
MarkDiffusion	11 / 9 / 12	Оценка методов внедрения ЦВЗ в процессе генерации изображений и видео	Поддерживаются только методы в процессе генерации изображений
OmniSeal Bench	10 / 17 / 5	Мультимодальная оценка нейросетевых методов	Не поддерживаются методы внедрения ЦВЗ в процессе генерации изображений; только multi-bit методы
WIBE	32 / 34 / 15	Унифицированная мультимодальная воспроизводимая оценка методов внедрения ЦВЗ	—

1.6.2 Средства сравнительной оценки методов маркирования аудиоданных

Задача проведения сравнительных экспериментов актуальна и для методов маркирования аудиоконтента. Специфика данной модальности определяется прежде всего временной природой аудиосигнала и характерными способами его обработки. При оценке устойчивости наряду с добавлением шума и фильтрацией рассматриваются сжатие различными аудиокодеками, передискретизация, изменение темпа и высоты тона, обрезка фрагментов, моделирование акустических воздействий, а в современных исследованиях также обработка нейронными кодеками, вокодерами и моделями улучшения или реконструкции аудиосигнала. Для сравнительного исследования методов аудиомаркирования разработан ряд специализированных бенчмарков и программных средств.

AudioMarkBench [88] предназначен для оценки устойчивости методов маркирования аудио к стандартным преобразованиям и преднамеренным атакам. Эксперименты организованы для трех моделей доступа нарушителя. В первой модели отсутствуют сведения о внутреннем устройстве метода и доступ к детектору. Вторая модель предполагает возможность производить запросы к системе извлечения ЦВЗ. В третьей модели нарушителю доступна внутренняя структура атакуемого метода. Рассматриваются как атаки удаления ЦВЗ, так и атаки его подделки.

Бенчмарк RAW-Bench [89] ориентирован на стандартизированное исследование методов маркирования аудио в условиях, приближенных к реальной обработке и распространению аудиосигналов. В нем оцениваются четыре нейросетевых метода на различных типах аудиоконтента, включающих речь, музыку и окружающие звуки. Набор воздействий содержит сжатие, фоновый шум, реверберацию, временные преобразования. Особое внимание уделяется нейронным аудиокодекам, которые существенно влияют на устойчивость ряда исследованных методов.

VoxWatermark [90] представляет собой крупномасштабный бенчмарк обнаружения ЦВЗ в аудиоконтенте при различных моделях воздействий. В нем используется десять методов маркирования, включающих нейросетевые и классические подходы, а исходные данные формируются на основе нескольких речевых корпусов. Наряду со стандартными преобразованиями рассматриваются адаптивные атаки удаления и подделки ЦВЗ.

Система DeepMark [91] построена на основе подключаемых модулей методов маркирования и атак и предусматривает стандартизированное выполнение сравнительных экспериментов. В текущей экспериментальной конфигурации непосредственно интегрировано шесть методов маркирования, при этом архитектура предусматривает подключение дополнительных алгоритмов. Для изоляции несовместимых программных зависимостей нейросетевые методы и сложные атаки могут выполняться в отдельных контейнерах. Система содержит более сорока атак, объединенных в группы обработки аудио, искажений сигнала, нарушения временной синхронизации, воздействий на процесс маркирования, нейросетевых атак и сценариев передачи аудиосигнала. DeepMark также предоставляет средства оценки качества и успешности обнаружения.

В мультимодальной системе OmniSeal Bench для аудиоданных представлены методы AudioSeal, WavMark и Timbre Watermarking. Постановка эксперимента включает типовые преобразования сигнала, фильтрацию, изменение временных характеристик, добавление шума и сжатие с потерями. Как и для изображений, эти компоненты используются в рамках общей схемы внедрения ЦВЗ, применения воздействий, обнаружения или извлечения ЦВЗ и вычисления метрик.

Основные характеристики рассмотренных специализированных и мультимодальных средств оценки методов аудиомаркирования, а также системы WIBE приведены в таблице 2.

Рассмотренные средства предоставляют развитые возможности сравнительной оценки, однако различаются по охвату модальностей, методов внедрения ЦВЗ и экспериментальных постановок. Специализированные бенчмарки позволяют подробно исследовать отдельные сценарии, но их совместное использование требует согласования форматов данных, параметров воздействий и правил вычисления метрик. OmniSeal Bench объединяет несколько модальностей в общей инфраструктуре, однако его штатный набор компонентов не охватывает методы внедрения ЦВЗ в процессе генерации изображений, а для аудио ограничен тремя нейросетевыми методами с передачей сообщения.

Таблица 2 — Сравнение средств оценки методов маркирования аудиоданных

Решение	Методы / Воздействия / Метрики	Область применения	Ограничения
AudioMark Bench	3 / 15 / 4	Устойчивость к удалению и подделке ЦВЗ при различных моделях доступа нарушителя	Только аудиоданные; реализация только нейросетевых методов внедрения ЦВЗ; Отсутствуют zero-bit методы
RAW-Bench	4 / 20 / 5	Устойчивость ЦВЗ и качество речи, музыки и окружающих звуков при обработке, включая нейронные кодеки	Только аудиоданные; отсутствуют zero-bit методы
VoxWatermark	10 / 17 / 5	Унифицированное обнаружение ЦВЗ различных типов в речевых данных	Только аудиоданные; рассматривается только задача обнаружения ЦВЗ; отсутствует оценка извлечения сообщения для multi-bit методов
DeepMark Benchmark	6 / 40+ / 10	Автоматизированная оценка качества аудио и устойчивости ЦВЗ к преобразованиям и преднамеренным атакам	Только аудиоданные; реализация только для нейросетевых методов внедрения ЦВЗ
OmniSeal Bench	3 / 15 / 5	Мультимодальная оценка методов маркирования	Отсутствуют zero-bit методы внедрения ЦВЗ
WIBE	14 / 30 / 11	Унифицированная мультимодальная воспроизводимая оценка методов внедрения ЦВЗ	—

1.7 Выводы по главе

В главе рассмотрены принципы построения методов маркирования изображений, аудиоданных и видео, подходы к контролю целостности медиаконтента и средства сравнительной оценки. Рассмотренные методы маркирования различаются типом передаваемой информации, областью формирования ЦВЗ, необходимыми для проверки данными и устойчивостью к преобразованиям медиаконтента. Их незаметность, емкость, устойчивость и вычислительная эффективность зависят от сценария применения, поэтому сопоставление результатов требует определения единой постановки экспериментов.

Существующие средства сравнительной оценки методов маркирования медианных охватывают различные наборы методов, преобразований и атак. Наличие мультимодальных систем показывает возможность использовать общую организацию эксперимента, однако представления данных, воздействия и показатели для каждой модальности требуют самостоятельного выбора. Выявленные ограничения существующих решений стали основанием для разработки системы WIBE, объединяющей оценку методов маркирования различных типов медианных в рамках общей воспроизводимой методики. Архитектура системы и организация экспериментов рассматриваются во второй главе.

Обзор подходов к контролю целостности показал, что обнаружение устойчивого ЦВЗ само по себе не подтверждает сохранность содержимого. Водяной знак может сохраниться после недопустимой модификации или быть перенесен на другой объект. Для проверки целостности требуется связать внедренные аутентификационные данные с содержимым медиаконтента и обеспечить проверку их подлинности. Разработка и экспериментальное исследование метода контроля целостности изображений с учетом этих требований рассматриваются в третьей главе. Соответствующий метод контроля целостности аудиоданных, позволяющий также производить локализацию изменений, как и его экспериментальная оценка, представлены в четвертой главе.

Анализ методов маркирования видео показал, что независимая обработка кадров позволяет непосредственно применять методы маркирования изображений, однако требует многократного выполнения алгоритма внедрения и не использует сходство соседних кадров. Распространение модификации, полученной для одного кадра, сокращает вычисления, но не учитывает содержимое всей группы. Совместная обработка кадров или внутренних признаков модели может зависеть от ее архитектуры и требовать специального обучения. Представляет интерес методика, использующая информацию нескольких последовательных кадров для снижения вычислительных затрат без изменения исходного метода маркирования изображений и его дополнительного обучения. Разработка и исследование такой методики составляют содержание пятой главы.

Глава 2. Методика унифицированной воспроизводимой оценки методов маркирования медиаконтента

Проведенный в первой главе анализ показал, что сопоставление методов маркирования медиаконтента осложняется различием как самих алгоритмов, так и условий их экспериментальной оценки. Существующие методы различаются принципами формирования ЦВЗ: одни предусматривают внедрение в уже сформированный объект, другие формируют ЦВЗ непосредственно в процессе генерации медиаконтента. Проверка может заключаться в обнаружении заданного водяного знака либо извлечении сообщения, состоящего из набора бит. Дополнительные различия обусловлены особенностями медиаданных: для изображений и аудиосигналов используются различные представления, преобразования и метрики качества. Непосредственное сопоставление характеристик, полученных в разных экспериментальных условиях, не всегда позволяет обоснованно сравнивать методы между собой.

Для решения указанной задачи в настоящей главе предлагается методика унифицированной воспроизводимой оценки методов маркирования медиаконтента. В ее основе лежит единое представление экспериментального процесса в виде последовательности стадий, включающих формирование маркированного объекта, оценку качества, применение преобразований и преднамеренных атак, обнаружение или извлечение ЦВЗ и агрегацию результатов. Унификация достигается за счет общей организации экспериментального процесса и подключения независимых компонентов. Представления медиаданных, наборы воздействий и показатели оценки выбираются с учетом особенностей соответствующей модальности и исследуемых методов маркирования.

Предложенная методика реализована в программной системе WIBE, предназначенной для автоматизации воспроизводимых вычислительных экспериментов. В главе рассматриваются модель экспериментального процесса, архитектура программной системы, средства описания воспроизводимых экспериментов и организация их исполнения. Применимость разработанной методики демонстрируется на примере сравнительной оценки методов маркирования изображений и аудиоданных с использованием общей модели экспериментального конвейера и специализированных для каждой модальности наборов данных, воздействий и метрик.

2.1 Модель унифицированной экспериментальной оценки методов маркирования медиаконтента

Предлагаемая методика основана на представлении экспериментального процесса в виде последовательности стадий, состав которых сохраняется независимо от принципа формирования и проверки ЦВЗ и рассматриваемой модальности. Конкретные представления медиаданных, воздействия и показатели оценки при этом определяются особенностями исследуемых методов.

В рамках предлагаемой методики эксперимент задается совокупностью

$$\mathcal{E} = (\mathcal{D}, \mathcal{W}, \mathcal{A}, \mathcal{Q}, \mathcal{R}), \quad (2.1)$$

где $\mathcal{D}$ — набор исходных медиаданных, $\mathcal{W}$ — множество тестируемых методов маркирования, $\mathcal{A}$ — множество преобразований и атак, $\mathcal{Q}$ — набор метрик качества и незаметности, $\mathcal{R}$ — набор метрик обнаружения и извлечения ЦВЗ. Общая организация экспериментального процесса представлена на рисунке 2.1.

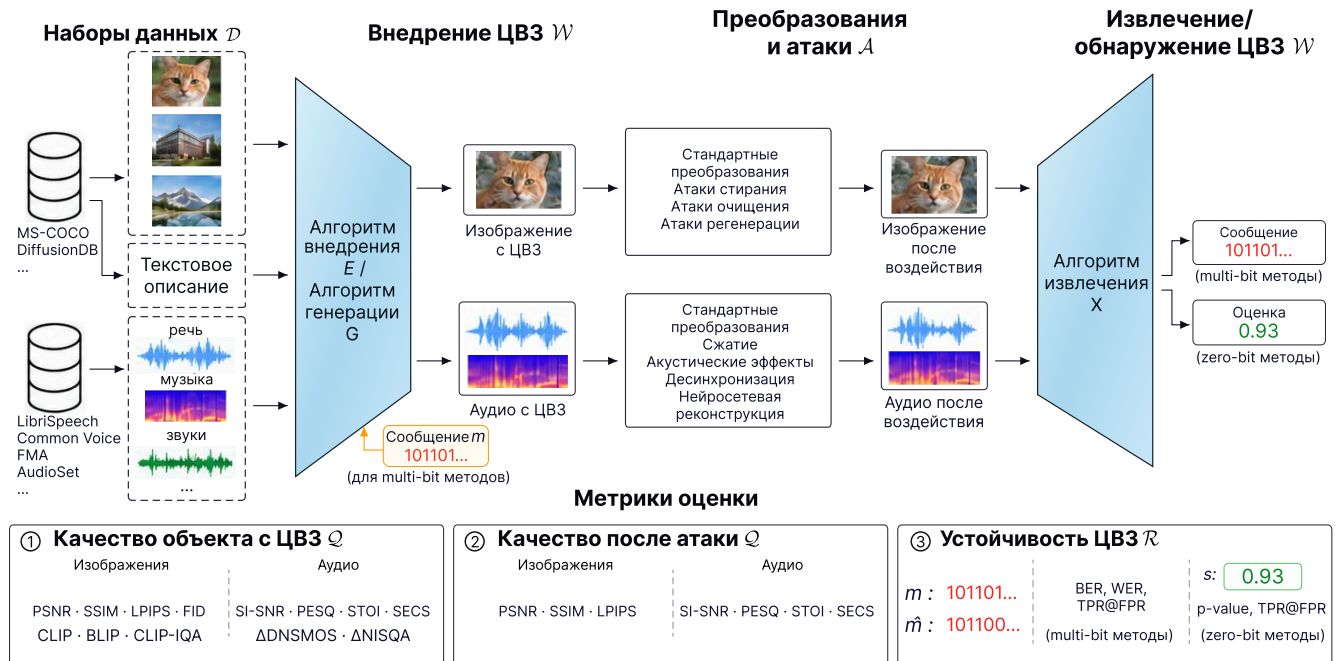


Рисунок 2.1 — Схема унифицированной экспериментальной оценки методов маркирования медиаконтента

Элемент $x \in \mathcal{D}$ представляет отдельный объект медиаконтента. В зависимости от исследуемой модальности это может быть изображение или аудиофрагмент. Конкретный набор операций формирования и проверки ЦВЗ определяется методом маркирования $W \in \mathcal{W}$.

Для методов, предусматривающих внедрение сообщения m в сформированный объект x , результат внедрения определяется как

$$x_w = E(x, m; k_e), \quad (2.2)$$

где E — алгоритм внедрения, а k_e — ключ внедрения, если он используется соответствующим методом. Для zero-bit методов аргумент m опускается.

Отдельный случай составляют методы, в которых ЦВЗ формируется непосредственно в процессе генерации медиаконтента. Наиболее широко такая постановка в настоящее время представлена для генеративных моделей изображений. В этом случае маркированный объект может быть представлен как

$$x_w = G(c, m; k_e), \quad (2.3)$$

где G — генеративная модель с механизмом маркирования, а c — входное условие генерации, например текстовое описание. В отличие от внедрения ЦВЗ в сформированный объект, для такого варианта в общем случае отсутствует непосредственно соответствующий немаркированный объект, который мог бы использоваться в качестве эталона при оценке незаметности.

После формирования маркированного объекта оценивается влияние маркирования на качество медиаконтента. В зависимости от доступных исходных данных могут использоваться метрики с эталоном, безэталонные метрики, показатели соответствия условию генерации и метрики сравнения распределений объектов:

$$\begin{aligned} q_w^{FR} &= q^{FR}(x, x_w), & q_w^{NR} &= q^{NR}(x_w), \\ q_w^C &= q^C(c, x_w), & q_w^{DIST} &= q^{DIST}(\mathcal{D}_{\text{ref}}, \mathcal{D}_w). \end{aligned} \quad (2.4)$$

Метрики с эталоном применяются при наличии соответствующего немаркированного объекта, тогда как безэталонные метрики позволяют оценивать качество без такого сопоставления. Если медиаконтент формируется на основании условия c , например текстового описания, могут дополнительно использоваться метрики соответствия результату этому условию. Для оценки изменений на уровне совокупности объектов применяются метрики сравнения распределений, где $\mathcal{D}_{\text{ref}}$ — эталонное множество объектов без ЦВЗ, а $\mathcal{D}_w$ — множество маркированных объектов.

Следующей стадией является моделирование преобразований и преднамеренных атак. Для воздействия $A \in \mathcal{A}$ результирующий объект определяется выражением

$$x_a = A(x_w). \quad (2.5)$$

Состав множества $\mathcal{A}$ определяется модальностью медиаконтента. Для изображений к стандартным преобразованиям относятся сжатие, изменение размера, фильтрация, изменение яркости и контраста, а также геометрические преобразования. Преднамеренные атаки могут включать специализированное удаление ЦВЗ, очищение изображения, регенерацию. Для аудиоконтента рассматриваются сжатие с потерями, передискретизация, фильтрация и другие преобразования и атаки.

После применения преобразования или атаки оценивается качество полученного медиаконтента. Такая оценка позволяет отличать воздействие, нарушающее обнаружение или извлечение ЦВЗ при относительно небольшом изменении объекта, от воздействия, эффективность которого достигается за счет существенного ухудшения качества. При наличии возможности прямого сопоставления с маркированным объектом применяется метрика с эталоном:

$$q_a^{FR} = q^{FR}(x_w, x_a). \quad (2.6)$$

После оценки качества выполняется обнаружение или извлечение цифрового водяного знака. Характер результата определяется типом метода маркирования. Для multi-bit методов алгоритм извлечения формирует битовую последовательность

$$\hat{m} = X(x_a; k_d), \quad (2.7)$$

где X — алгоритм извлечения ЦВЗ, $\hat{m}$ — извлеченное сообщение, а k_d — ключ извлечения, если он используется соответствующим алгоритмом. k_e и k_d могут как совпадать, так и различаться.

Качество извлечения определяется на основании сопоставления исходного и полученного сообщений:

$$r_a = r(m, \hat{m}), \quad (2.8)$$

где $r \in \mathcal{R}$ — выбранная метрика качества извлечения. В частности, могут использоваться побитовая точность, доля битовых ошибок и показатели успешности полного восстановления сообщения.

Для методов, предназначенных для определения факта наличия ЦВЗ, детектор формирует некоторое значение $s(x_a)$, характеризующее степень соответствия исследуемого объекта заданному признаку наличия водяного знака. Решение о наличии ЦВЗ принимается путем сравнения этого значения с порогом τ :

$$\hat{y} = \begin{cases} 1, & s(x_a) \geq \tau, \\ 0, & s(x_a) < \tau. \end{cases} \quad (2.9)$$

Для сравнительной оценки таких методов могут использоваться показатели, характеризующие соотношение истинноположительных и ложноположительных срабатываний, в том числе значение TPR при фиксированном уровне FPR.

Для сопоставления методов с различными типами встроенной информации результат извлечения multi-bit методов может быть дополнительно сведен к задаче обнаружения ЦВЗ. В качестве статистики обнаружения используется число совпавших бит исходного и извлеченного сообщений, которое сравнивается с заданным порогом. Это позволяет использовать показатель TPR при фиксированном уровне FPR как для multi-bit, так и для zero-bit методов, сохраняя при этом специализированные метрики успешности извлечения multi-bit методов.

После обработки отдельных объектов результаты агрегируются по всему набору данных отдельно для каждого сочетания метода маркирования, воздействия и его параметров. Обобщаются показатели качества медиаконтента, успешности обнаружения или извлечения ЦВЗ и вычислительной эффективности алгоритмов.

Разделение экспериментального процесса на последовательные стадии и независимые компоненты позволяет исследовать методы маркирования различных модальностей в рамках общей постановки, расширять набор рассматриваемых алгоритмов, преобразований и преднамеренных атак и использовать специализированные критерии оценки без изменения общей организации эксперимента.

2.2 Программная реализация методики в системе WIBE

Предложенная методика реализована в разработанной программной системе WIBE [6], предназначенной для автоматизации воспроизводимых вычисли-

тельных экспериментов с методами маркирования медиаконтента. Исходный код системы опубликован в открытом доступе [92]. Далее рассматриваются архитектура программной системы, средства воспроизводимого описания экспериментов и организация их исполнения.

2.2.1 Архитектура программной системы

Архитектура WIBE основана на разделении ядра экспериментального конвейера и подключаемых компонентов, реализующих алгоритмы маркирования, наборы данных, воздействия и метрики. Ядро WIBE реализовано на языке Python, а для представления медиаданных и выполнения тензорных операций используется библиотека PyTorch. Общая структура программной системы представлена на рисунке 2.2.

Основной единицей выполнения в WIBE является экспериментальный конвейер (Pipeline), состоящий из последовательности стадий. Для каждого объекта набора данных последовательно выполняются формирование маркированного объекта, вычисление метрик после внедрения ЦВЗ, применение набора заданных воздействий, вычисление метрик после воздействий, извлечение или обнаружение ЦВЗ, оценка результата извлечения и агрегация результатов. В программной реализации этим операциям соответствуют стадии

$$\begin{aligned} &\text{embed} \rightarrow \text{post_embed_metrics} \rightarrow \text{attack} \rightarrow \text{post_attack_metrics} \\ &\quad \rightarrow \text{extract} \rightarrow \text{post_extract_metrics} \rightarrow \text{aggregate}. \end{aligned} \quad (2.10)$$

Такое представление непосредственно соответствует стадиям экспериментальной оценки, введенным в разделе 2.1, при этом в программной реализации каждая стадия является самостоятельным компонентом конвейера. Это позволяет изменять состав выполняемых операций и использовать одни и те же стадии с различными реализациями алгоритмов, воздействий и метрик.

Для передачи данных между стадиями используется объект контекста (Context). Контекст представляет текущее состояние обработки отдельного объекта из набора данных и последовательно дополняется результатами выполненных стадий. В нем сохраняются исходные данные, сформированные

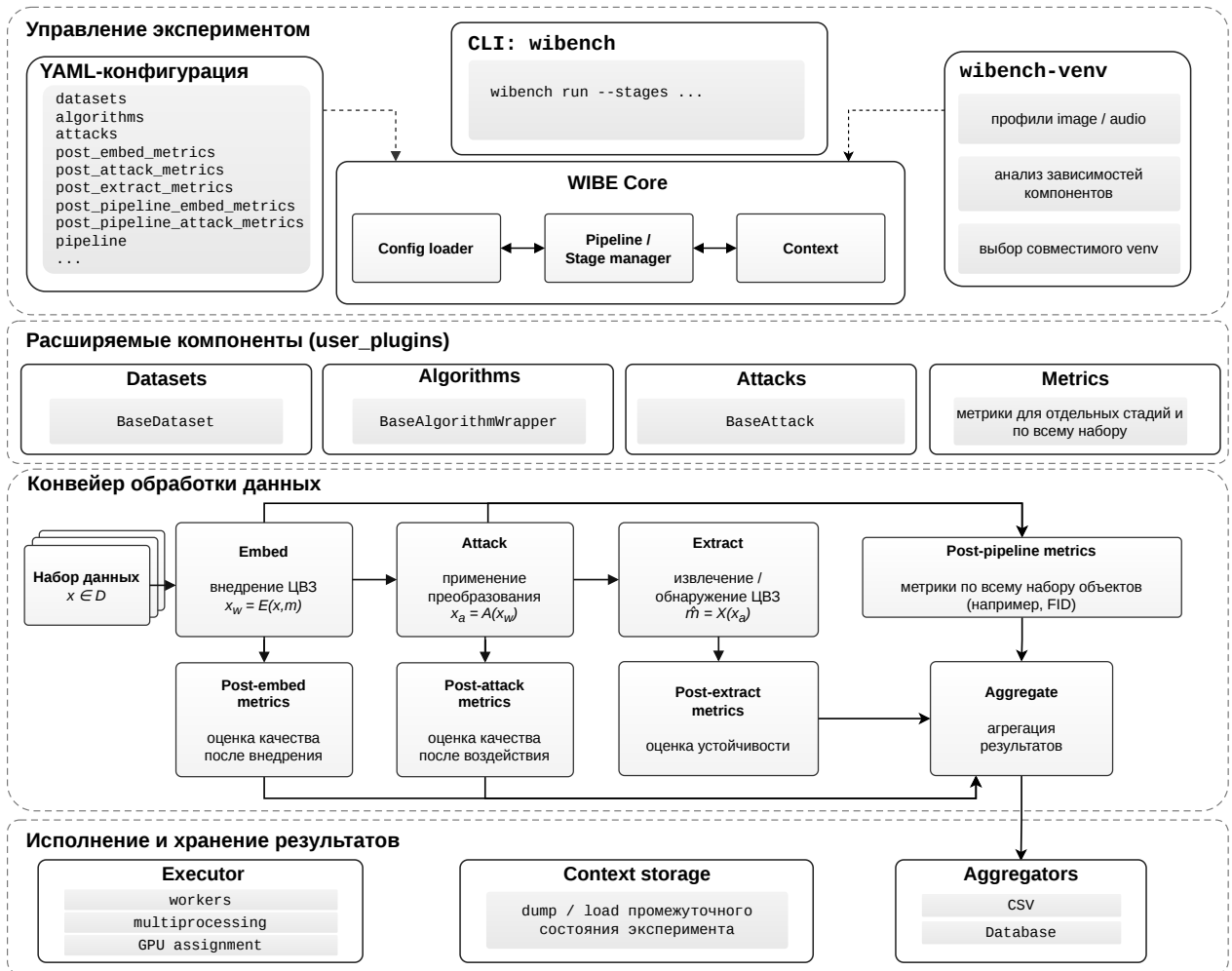


Рисунок 2.2 — Архитектура программной системы WIBE

данные ЦВЗ, маркированный объект, результат применения воздействия, результат извлечения, а также рассчитанные значения метрик. Каждая стадия получает контекст, использует необходимые ей элементы и добавляет результаты своей работы, которые затем становятся доступны последующим стадиям. Такое разделение снижает связанность компонентов и позволяет независимо расширять отдельные части системы.

Алгоритмы маркирования подключаются к конвейеру через базовый интерфейс `BaseAlgorithmWrapper`. Его реализация выполняет роль адаптера между WIBE и исходной программной реализацией алгоритма маркирования. Интерфейс предусматривает формирование дополнительных данных ЦВЗ, операцию внедрения и операцию извлечения. Соответствующие методы имеют вид `watermark_data_gen`, `embed` и `extract`. При этом тип данных, передаваемых в качестве ЦВЗ, не фиксируется ядром системы: это может быть сообщение, ключ, шаблон ЦВЗ или другая структура, необходимая конкретному алгоритму.

Такой интерфейс позволяет унифицировать методы с существенно различными принципами работы. Для алгоритмов, внедряющих ЦВЗ в уже сформированный медиаобъект, метод `embed` получает сам объект и данные ЦВЗ. Для методов маркирования в процессе генерации входом может выступать условие генерации, например текстовое описание, а результатом остается сформированный маркированный объект.

Наборы данных подключаются через интерфейс `BaseDataset`. Основной операцией данного интерфейса является генерация последовательности объектов, передаваемых на обработку конвейеру. Это позволяет отделить способ хранения и загрузки медиаданных от логики проведения эксперимента.

Воздействия подключаются посредством интерфейса `BaseAttack`. Конкретная реализация принимает маркированный объект, выполняет заданное преобразование и возвращает измененный объект. Благодаря этому стандартные преобразования и преднамеренные атаки используются в конвейере одинаковым образом и могут применяться к различным алгоритмам маркирования без изменения их реализаций.

Метрики также реализованы как независимые компоненты, однако их интерфейс определяется положением в экспериментальном конвейере. Метрики после внедрения получают исходный и маркированный объекты и оценивают влияние процесса маркирования на качество маркированного объекта. После применения воздействия используются метрики, сопоставляющие маркированный и измененный объекты. Метрики после извлечения дополнительно получают данные исходного ЦВЗ и результат работы алгоритма извлечения, что позволяет оценивать успешность обнаружения или восстановления внедренной информации.

Завершающая стадия `aggregate` преобразует результаты обработки объекта в унифицированное представление и передает их настроенным механизмам сохранения и агрегации. В состав результирующей записи входят значения вычисленных метрик, а также параметры конфигурации алгоритма, набора данных и воздействия, что позволяет однозначно связать результат с условиями его получения.

Отдельно предусмотрены метрики, вычисляемые после завершения обработки набора объектов. В отличие от метрик отдельных объектов, для их расчета требуется накопление результатов по всей выборке. В частности, к этому типу относятся метрики сравнения распределений, например FID. В WIBE для этого

предусмотрены дополнительные стадии `post_pipeline_embed_metrics` и `post_pipeline_attack_metrics`, позволяющие соответственно сопоставлять множества объектов после маркирования и после применения воздействий.

Расширение системы выполняется путем реализации соответствующего базового интерфейса и не требует изменения ядра экспериментального конвейера. Пользовательские алгоритмы, наборы данных, воздействия и метрики могут подключаться как дополнительные программные модули. В результате архитектура WIBE разделяет управление экспериментом и операции над медиаконтентом, что позволяет использовать одну реализацию экспериментального конвейера для различных алгоритмов и модальностей.

2.2.2 Воспроизводимое описание экспериментов

В WIBE условия проведения эксперимента отделены от программного кода системы и задаются декларативно с использованием конфигурационных файлов формата YAML. Конфигурация определяет состав исследуемых алгоритмов маркирования, наборов данных, воздействий и метрик, а также параметры каждого из этих компонентов.

Каждый подключаемый компонент указывается в конфигурации своим типом и набором параметров, передаваемых при его создании. Один и тот же программный компонент может быть включен в эксперимент несколько раз с различными параметрами. Для различения таких конфигураций при агрегации результатов используется поле `report_name`. Например, несколько вариантов одной атаки могут отличаться интенсивностью воздействия, при этом использовать общую программную реализацию.

Пример конфигурации эксперимента приведен в листинге 2.1. В нем одна реализация воздействия JPEG используется с двумя наборами параметров, результаты которых учитываются независимо.

Воспроизводимость стохастических операций обеспечивается заданием начального состояния генераторов псевдослучайных чисел с помощью параметра `seed`. Его фиксированное значение позволяет воспроизводить случайно формируемые сообщения ЦВЗ и другие случайные величины, используемые при проведении эксперимента.

Листинг 2.1 — Пример конфигурации эксперимента WIBE

```
algorithms:
  - TrustMark

datasets:
  - DiffusionDB:
      sample_range: [0, 999]

attacks:
  - JPEG:
      report_name: jpeg_80
      quality: 80
  - JPEG:
      report_name: jpeg_50
      quality: 50

post_embed_metrics:
  - LPIPS

post_extract_metrics:
  - BER

pipeline:
  seed: 42
  workers: 4
  cuda_visible_devices: 0,1,2,3
```

Параметры `workers` и `cuda_visible_devices`, приведенные в листинге, определяют способ исполнения эксперимента и рассматриваются в следующем подразделе.

2.2.3 Организация исполнения экспериментов

Исполнение экспериментов в WIBE отделено от подготовки программных окружений. Необходимость такого разделения обусловлена тем, что интегрируемые алгоритмы маркирования, воздействия и метрики могут требовать несов-

местимые версии Python-библиотек. В общем случае невозможно использование единого набора зависимостей для всех подключенных компонентов.

Для управления зависимостями используется разработанный менеджер виртуальных окружений `wibench-venv`. Зависимости отдельных компонентов описываются независимо, после чего заранее группируются в совместимые наборы. При формировании таких групп проверяется совместимость требований, в том числе путем построения графа их попарных конфликтов. Для каждой полученной группы фиксируются версии программных пакетов и создается отдельное виртуальное окружение. Для различных модальностей предусмотрены отдельные профили окружений, в частности `image` и `audio`.

Подготовка окружений выполняется до запуска экспериментальных конфигураций. При запуске WIBE определяет зависимости компонентов, входящих в конкретный эксперимент, и выбирает из ранее сформированных окружение, содержащее необходимый набор библиотек. Если процесс был запущен с другим интерпретатором Python, выполнение автоматически перезапускается в выбранном окружении. Если подходящее окружение отсутствует, выполнение конфигурации не начинается до изменения набора компонентов или обновления сформированных окружений.

Стадии экспериментального конвейера могут выполняться отдельно. Промежуточное состояние объектов, представленное экземплярами `Context`, может сохраняться после выполнения выбранных стадий и использоваться при последующем запуске. Это позволяет продолжать эксперимент с требуемого этапа или повторно выполнять только его часть без пересчета уже полученных промежуточных результатов. В частности, данный механизм позволяет разделять эксперимент на несколько запусков, если компоненты различных стадий имеют несовместимые программные зависимости.

Обработка различных объектов набора данных может выполняться параллельно в нескольких процессах. Число рабочих процессов задается параметром `workers`. Перечень доступных графических ускорителей определяется параметром `cuda_visible_devices`, что позволяет распределять вычисления между несколькими GPU. Поскольку состояние обработки каждого объекта хранится независимо, увеличение числа рабочих процессов не требует изменения реализаций алгоритмов маркирования, воздействий и метрик.

Полученные в ходе эксперимента результаты передаются настроенным механизмам сохранения. WIBE поддерживает их агрегацию в CSV-файлы и запись в

базу данных. Совместно со значениями метрик сохраняются параметры соответствующих экспериментальных конфигураций, что обеспечивает последующий анализ результатов для различных сочетаний алгоритмов, наборов данных и воздействий.

Таким образом, подсистема исполнения WIBE отделяет описание эксперимента от особенностей программного окружения и используемых вычислительных ресурсов. Предварительное формирование совместимых виртуальных окружений, раздельное выполнение стадий и параллельная обработка объектов позволяют проводить воспроизводимые эксперименты с различными программными компонентами в рамках единой системы.

2.3 Сравнительная оценка методов маркирования изображений

Для экспериментальной апробации предложенной методики и программной системы WIBE проведена сравнительная оценка современных методов маркирования изображений. В соответствии с общей схемой, описанной в разделе 2.1, для данной модальности были определены специализированные метрики качества и устойчивости, набор преобразований и преднамеренных атак, а также условия проведения эксперимента. Далее последовательно рассматриваются используемые показатели оценки, модели воздействий, параметры эксперимента и полученные результаты.

2.3.1 Метрики оценки

Для оценки методов маркирования изображений используются три основные группы показателей: метрики качества и незаметности, показатели точности извлечения сообщений и показатели обнаружения ЦВЗ. Конкретные метрики качества выбираются с учетом способа формирования маркированного изображения в соответствии с общей моделью, представленной в разделе 2.1. Основные метрики, реализованные в WIBE для изображений, приведены в таблице 3.

Таблица 3 — Метрики оценки методов маркирования изображений

Метрика	Оцениваемое свойство	Группа	Описание
PSNR	качество изображения	С эталонным изображением	Сравнение искажений относительно эталонного изображения
SSIM	качество изображения	С эталонным изображением	Структурное сходство изображений
LPIPS	качество изображения	С эталонным изображением	Перцептивное сходство изображений
DreamSim	качество изображения	С эталонным изображением	Перцептивное сходство изображений
CLIP-IQA	качество изображения	Безэталонная	Визуальное качество изображения без использования эталона
Aesthetic Score	качество изображения	Безэталонная	Эстетическое качество изображения
CLIP Score	качество изображения	Текст–изображение	Соответствие изображения текстовому описанию
BLIP	качество изображения	Текст–изображение	Соответствие изображения текстовому описанию
ImageReward	качество изображения	Текст–изображение	Соответствие изображения текстовому описанию
FID	качество изображения	Два набора изображений	Различие распределений двух наборов изображений
BER	извлечение ЦВЗ	Сравнение сообщений	Доля ошибочно извлеченных битов
WER	извлечение ЦВЗ	Сравнение сообщений	Доля сообщений, содержащих хотя бы одну ошибку
p -value	обнаружение ЦВЗ	Статистический показатель	Статистическая значимость соответствия результата заданному ЦВЗ
TPR@ x FPR	обнаружение ЦВЗ	Теоретический порог	Доля правильно обнаруженных ЦВЗ при заданном уровне ложных срабатываний
Empirical TPR@ x FPR	обнаружение ЦВЗ	Эмпирический порог	TPR при пороге, определенном по выборке немаркированных изображений

Для оценки качества маркированных изображений используются метрики нескольких типов. PSNR и SSIM [93] характеризуют соответственно величину пиксельных и структурных изменений относительно исходного изображения, а LPIPS [94] и DreamSim [95] — перцептивное сходство изображений. В случаях, когда соответствующее немаркированное изображение отсутствует, применяются безэталонные показатели CLIP-IQA [96] и Aesthetic Score [97].

Для методов маркирования в процессе генерации дополнительно оценивается соответствие сформированного изображения условию генерации с помощью CLIP Score [98], BLIP [99] и ImageReward [100]. Изменения на уровне распределения изображений характеризуются метрикой FID [101], вычисляемой для наборов маркированных и немаркированных изображений.

Совместное использование различных показателей связано с тем, что ни одна отдельная метрика не позволяет полностью охарактеризовать незаметность ЦВЗ. Пиксельные, структурные, перцептивные, безэталонные и распределительные показатели чувствительны к различным типам изменений изображения.

Вторая группа метрик предназначена для оценки успешности извлечения или обнаружения ЦВЗ. Для методов, предусматривающих передачу сообщения, основной характеристикой является доля ошибочно восстановленных битов (Bit Error Rate, BER). Пусть исходное сообщение имеет вид $m = (m_1, \dots, m_k)$, а извлеченное сообщение — $\hat{m} = (\hat{m}_1, \dots, \hat{m}_k)$. Тогда BER определяется как

$$BER(m, \hat{m}) = \frac{1}{k} \sum_{i=1}^k \mathbf{I}[m_i \neq \hat{m}_i]. \quad (2.11)$$

Значение $BER = 0$ соответствует побитово точному извлечению сообщения. При случайном независимом угадывании равновероятных двоичных значений математическое ожидание BER равно 0,5.

Более строгим показателем успешности извлечения является Word Error Rate (WER). При его вычислении сообщение считается ошибочно извлеченным при наличии хотя бы одного несовпадающего бита. Для набора из N сообщений

$$WER = \frac{1}{N} \sum_{j=1}^N \mathbf{I} \left[\exists i : m_i^{(j)} \neq \hat{m}_i^{(j)} \right]. \quad (2.12)$$

BER характеризует среднюю долю ошибочно восстановленных битов, тогда как WER отражает долю сообщений, которые не удалось восстановить полностью. Такое различие существенно для методов с большой емкостью: относитель-

но небольшое значение BER может сочетаться с высокой вероятностью наличия хотя бы одной ошибки в сообщении.

Для zero-bit методов результат детектора может характеризоваться p -value — вероятностью получить наблюдаемое или более выраженное значение статистики при условии отсутствия соответствующего ЦВЗ. Конкретная статистика и способ вычисления p -value определяются исследуемым методом.

При заданном пороговом значении τ и наличии двух наборов изображений, содержащих и не содержащих ЦВЗ, определяются следующие показатели:

1. TPR (true positive rate) — доля маркированных изображений, для которых при заданном пороге корректно определено наличие ЦВЗ;
2. FPR (false positive rate) — доля немаркированных изображений, для которых при заданном пороге наличие ЦВЗ определено ошибочно.

На практике особое значение имеет ограничение вероятности ложного обнаружения. Порог решения τ выбирается таким образом, чтобы FPR соответствовал заданному значению. Устойчивость метода при этом характеризуется величиной TPR при фиксированном уровне FPR. Далее используется обозначение $\text{TPR}@x\text{FPR}$, где x — целевое значение FPR. Например, $\text{TPR}@0,01\%\text{FPR}$ характеризует долю правильно обнаруженных маркированных изображений при уровне ложных срабатываний одно изображение на 10000 немаркированных изображений.

Для multi-bit методов в качестве статистики обнаружения используется число совпадающих битов исходного и извлеченного сообщений:

$$M(m, \hat{m}) = \sum_{i=1}^k \mathbf{1}[m_i = \hat{m}_i]. \quad (2.13)$$

ЦВЗ считается обнаруженным при выполнении условия

$$M(m, \hat{m}) > \tau, \quad (2.14)$$

где τ — порог обнаружения.

Пороговое значение τ может быть определено эмпирическим способом. Для этого формируется набор из изображений, не содержащих ЦВЗ. Для каждого изображения вычисляется статистика детектора. Значение порога выбирается по распределению значений статистики на сформированной выборке так, чтобы доля немаркированных изображений, для которых значение статистики превосходит порог, соответствовало требуемому уровню FPR. Однако для статистической

достоверности определенного порогового значения требуется, чтобы размер набора изображений был достаточно большим: $N \gg 1/\text{FPR}$.

Для multi-bit методов значение FPR, соответствующее выбранному порогу τ , может быть определено теоретически [46]. При отсутствии соответствующего ЦВЗ предполагается, что каждый извлеченный бит независимо и равновероятно принимает значения 0 или 1. Число случайных совпадений с заданным k -битовым сообщением в этом случае подчиняется биномиальному распределению. При принятом выше условии обнаружения вероятность ложного срабатывания определяется как

$$\text{FPR}(\tau) = I_{1/2}(\tau + 1, k - \tau), \quad (2.15)$$

где $I_x(a, b)$ — регуляризованная неполная бета-функция. Теоретический способ позволяет определять пороги для малых значений FPR без формирования больших наборов немаркированных изображений. В частности, при целевых значениях FPR порядка 10^{-6} надежная эмпирическая оценка требует выборки порядка 10^7 объектов, для каждого из которых требуется произвести запуск алгоритма обнаружения. Такой объем вычислений может оказаться трудно реализуем на практике.

Для проверки применимости теоретической оценки проведено ее сравнение с эмпирическим способом. В качестве исследуемого метода использовался TrustMark с длиной сообщения $k = 100$ бит. Для формирования эмпирического распределения использовались 100 тысяч немаркированных изображений обучающей части набора данных MS-COCO [13]. Значения TPR оценивались на 5 тысячах маркированных изображений валидационной части MS-COCO при последовательном увеличении интенсивности гауссовского шума. Сравнение проводилось для уровней FPR 10%, 1%, 0,1% и 0,01%. Полученные результаты представлены на рисунке 2.3.

При $\text{FPR} \leq 1\%$ значения TPR, полученные с использованием теоретически и эмпирически определенных порогов, различаются незначительно во всем рассмотренном диапазоне интенсивности шума. Статистическая проверка при этом показала, что извлекаемые биты не полностью удовлетворяют предположению о независимых случайных величинах с распределением Bernoulli(0,5). Тем не менее близость полученных результатов подтверждает возможность применения теоретического способа в рассматриваемой экспериментальной постановке, в том

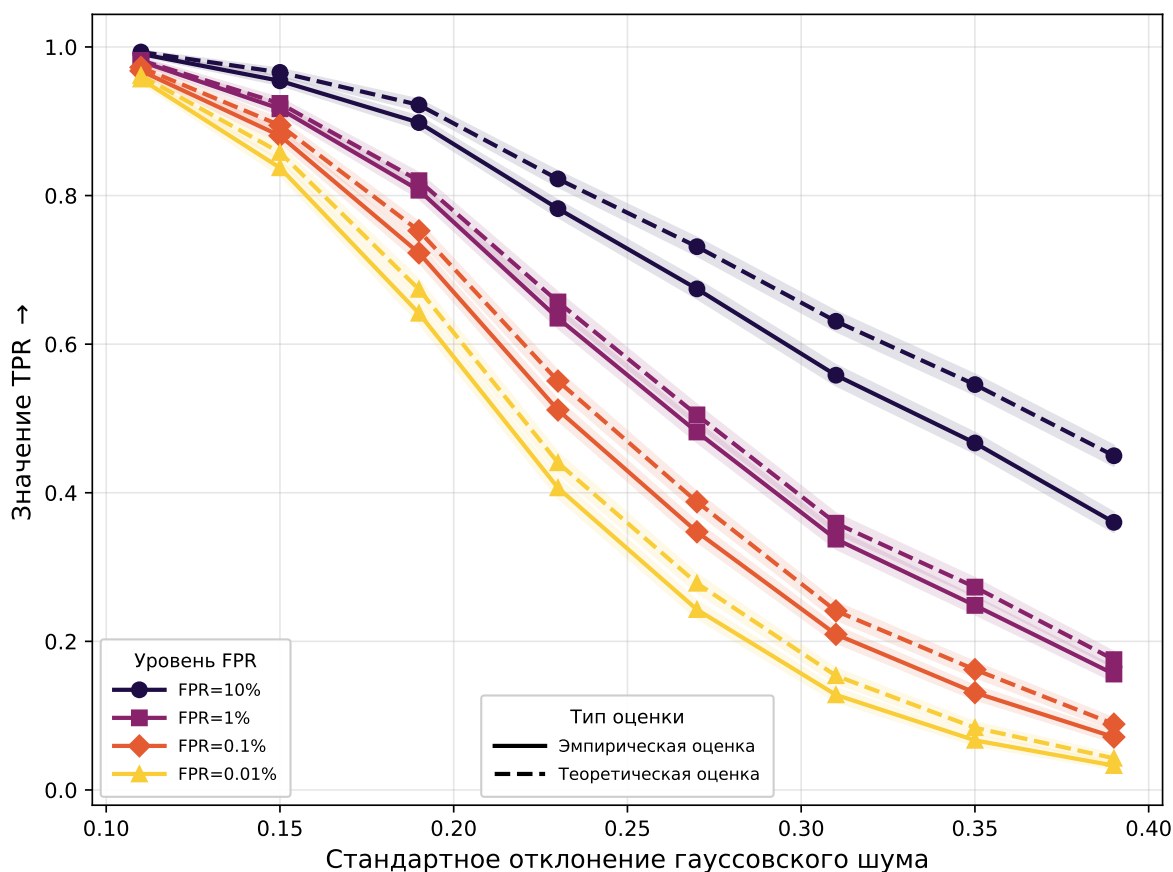


Рисунок 2.3 — Сравнение теоретической и эмпирической оценки TPR при различных уровнях FPR

числе при уровнях FPR, непосредственная эмпирическая оценка которых требует существенно больших выборок.

2.3.2 Преобразования и преднамеренные атаки

Устойчивость методов маркирования оценивалась относительно набора воздействий, включающего как стандартные преобразования медиаданных, так и преднамеренные атаки на ЦВЗ. Рассматриваемые воздействия разделены на четыре группы: стандартные преобразования, атаки стирания, атаки очищения и атаки регенерации.

Стандартные преобразования соответствуют типовым операциям обработки изображений и не направлены непосредственно на удаление ЦВЗ. Атаки стирания, напротив, специально предназначены для подавления или искажения встроенного сигнала. К атакам очищения относятся методы восстановления и

шумоподавления, способные устранять низкоамплитудные изменения, связанные с ЦВЗ. Атаки регенерации основаны на реконструкции или повторном синтезе изображения с использованием генеративных моделей, при котором визуальное или семантическое содержание может сохраняться, а встроенный ЦВЗ — утрачиваться. Основные группы воздействий и используемые в экспериментах методы приведены в таблице 4.

Таблица 4 — Преобразования и атаки, используемые при оценке устойчивости методов маркирования изображений

Класс воздействия	Примеры	Принцип работы
Стандартные преобразования	Brightness, Center Crop, Color Inversion, Contrast, Gaussian Blur, JPEG, White Noise, Pixel Shift, Random Cropout, Resize, Rotate	Изменение пиксельных, пространственных или геометрических характеристик изображения при типовой обработке
Атаки стирания ЦВЗ	Adversarial Embedding [84], Averaging [102], Blur—Deblur [103], SADRE [104], TrustMark-RM [33], WMForger [105] и др.	Преднамеренное подавление, искажение или оценка встроенного сигнала
Атаки очищения	DiffPure [106], DISCO [107], NRP [108], MPRNet [109], Real-ESRGAN [110], DIP [111]	Восстановление изображения и подавление компоненты, интерпретируемой как шум или возмущение
Атаки регенерации	VAE [112], Stable Diffusion [113], FLUX [114] и другие генеративные методы	Реконструкция или повторный синтез изображения с сохранением его визуального или семантического содержания

Эффективность воздействия должна рассматриваться совместно со степенью изменения изображения. Снижение успешности обнаружения или извлечения ЦВЗ при существенном ухудшении визуального содержания отличается от атаки, обеспечивающей аналогичный результат при сохранении качества. Поэтому при сравнении атак показатели устойчивости анализируются совместно с метриками качества, рассмотренными в подразделе 2.3.1. Такой подход особенно важен для геометрических преобразований и атак регенерации: первые могут приводить к значительным значениям метрик с эталоном вследствие пространственной рассинхронизации, а вторые — существенно изменять пиксельное представление при сохранении высокого визуального качества.

2.3.3 Условия проведения эксперимента

Для проведения сравнительной оценки в WIBE были интегрированы 32 конфигурации методов маркирования изображений, из которых 26 предназначены для внедрения ЦВЗ в сформированные изображения, а 6 реализуют маркирование непосредственно в процессе генерации. Основные принципы построения соответствующих классов методов и характерные представители рассмотрены в разделе 1.2.

Исследуемые методы охватывают широкий спектр подходов: от классических алгоритмов на основе частотных преобразований до современных нейросетевых и генеративных моделей. По типу встроенной информации представлены как zero-bit методы, предназначенные только для обнаружения факта наличия ЦВЗ, так и multi-bit методы с емкостью до 1024 бит. Среди методов внедрения ЦВЗ в сформированные изображения представлены классические частотные алгоритмы DWT DCT, DCT CAISS и DFT Circle, а также современные нейросетевые методы HiDDeN, MBRS, TrustMark, VideoSeal, PixelSeal, ChunkySeal и другие. Методы маркирования в процессе генерации представлены Stable Signature, Tree-Ring, Gaussian Shading, METR, Ring-ID и MaXsive.

Для уменьшения влияния различий между конкретными реализациями использовались официальные открытые реализации методов и опубликованные авторами предобученные веса без дополнительного дообучения. Для алгоритмов, для которых открытая реализация отсутствовала, в частности DFT Circle, DWT SVM и DCT CAISS, были реализованы собственные версии на основании описаний из соответствующих публикаций. Вспомогательные процедуры подготовки и приведения данных к требуемому формату выполнялись единообразно для сравниваемых методов.

Существенной проблемой при сравнении алгоритмов является различие в поддерживаемом входном разрешении. Некоторые нейросетевые модели обучены для работы с изображениями фиксированного размера, тогда как исходные изображения могут иметь произвольное разрешение. Пусть I — исходное изображение, а I_r — его версия, масштабированная к разрешению, поддерживаемому конкретным методом:

$$I_r = \text{Resize}(I). \quad (2.16)$$

После внедрения ЦВЗ в данном разрешении вычисляется разность между маркированным и исходным изображениями:

$$\Delta I_r = I_{w,r} - I_r. \quad (2.17)$$

Полученная разность масштабируется к исходному размеру и добавляется к исходному изображению:

$$I_w^* = I + \text{Resize}(\Delta I_r). \quad (2.18)$$

Такой подход позволяет выполнять внедрение в штатном разрешении конкретного алгоритма, сохраняя при этом исходный размер изображения и пространственный характер вносимых изменений.

Для multi-bit методов для каждого исходного изображения независимо формируется случайное двоичное сообщение соответствующей длины. То же сообщение используется при последующем извлечении и вычислении BER, WER и показателей обнаружения. Для zero-bit методов используются предусмотренные соответствующими алгоритмами статистики обнаружения и пороговые правила, рассмотренные в подразделе 2.3.1. Такая организация позволяет использовать единую постановку экспериментов для методов с различными способами представления ЦВЗ.

Экспериментальная оценка проводится на изображениях из наборов данных MS-COCO и DiffusionDB [115]. Использование двух наборов позволяет исследовать методы как на естественных изображениях, так и на контенте, сформированном генеративными моделями. Для основной оценки устойчивости из каждого набора выбирается по 1000 исходных изображений. Для методов маркирования в процессе генерации дополнительно используются текстовые описания, необходимые для формирования изображений генеративной моделью.

2.3.4 Результаты сравнительной оценки

В данном подразделе представлены результаты сравнительной экспериментальной оценки в соответствии с методикой, описанной в подразделе 2.3.3. Анализ проводится последовательно по трем группам характеристик. Сначала

рассматриваются свойства методов в отсутствие дополнительных воздействий: незаметность вносимых изменений, емкость сообщения и вычислительная эффективность. Далее исследуется устойчивость методов к стандартным преобразованиям и преднамеренным атакам. На заключительном этапе оценивается сравнительная эффективность различных классов атак с учетом как степени подавления ЦВЗ, так и величины вносимых ими искажений.

Незаметность, емкость и вычислительная эффективность

На первом этапе исследовались характеристики методов в отсутствие дополнительных воздействий. Для методов внедрения ЦВЗ в сформированные изображения оценивалось соотношение между емкостью сообщения, незаметностью вносимых искажений и вычислительной эффективностью. Незаметность оценивалась с использованием метрики LIPS относительно исходного изображения. Результаты представлены на рисунке 2.4.

По оси абсцисс отложена емкость ЦВЗ в логарифмическом масштабе; zero-bit методы, для которых понятие емкости сообщения неприменимо, показаны отдельной условной категорией. Цветом точек обозначено суммарное время внедрения и извлечения ЦВЗ. Методы, время работы которых измерялось с использованием графического ускорителя (GPU), отмечены символом «*». Отдельно выделены методы, принадлежащие границе Парето по емкости и значению LIPS.

Результаты показывают, что увеличение емкости не обязательно сопровождается пропорциональным ростом перцептивных искажений. Характеристики классических методов существенно различаются, тогда как ряд нейросетевых методов сочетает сравнительно высокую емкость с низкими значениями LIPS. В частности, PixelSeal и MBRS поддерживают 256-битные сообщения, а ChunkySeal — 1024-битное сообщение при сравнительно небольших перцептивных искажениях.

Время обработки также существенно различается между реализациями. Большинство классических методов выполнялось на CPU, тогда как для значительной части нейросетевых методов использовался GPU. Полученные значения характеризуют практическую производительность исследованных реализаций и

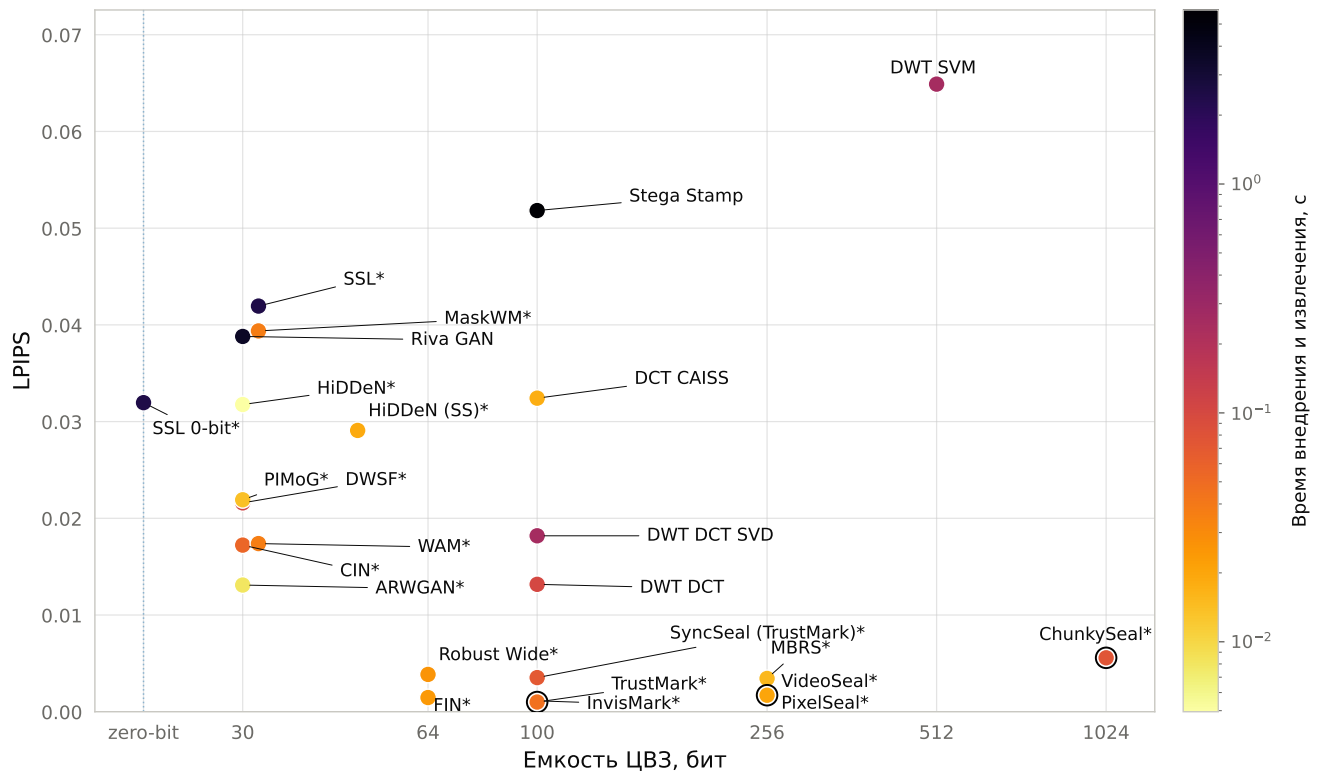


Рисунок 2.4 — Соотношение между емкостью, незаметностью и вычислительной эффективностью методов внедрения ЦВЗ в сформированные изображения

не рассматриваются как аппаратно-независимая оценка вычислительной сложности.

Устойчивость к преобразованиям и преднамеренным атакам

Для обобщенного сопоставления устойчивости воздействия были объединены в четыре класса: стандартные преобразования, атаки стирания, атаки очищения и атаки регенерации. Для каждого метода вычислялось среднее значение TPR по воздействиям соответствующего класса. Основным показателем являлся TPR при $FPR = 0,01\%$.

Результаты представлены на рисунке 2.5. Для DFT Circle, MaXsive и METR ввиду особенностей реализации приведен TPR при $FPR = 0,1\%$, а для Ring-ID — при $FPR = 1\%$; соответствующие методы отмечены символом «*» и должны сопоставляться с остальными с учетом различия уровней FPR.

Полученные результаты показывают существенную зависимость устойчивости от класса воздействия. Особенно выражены различия при атаках регене-

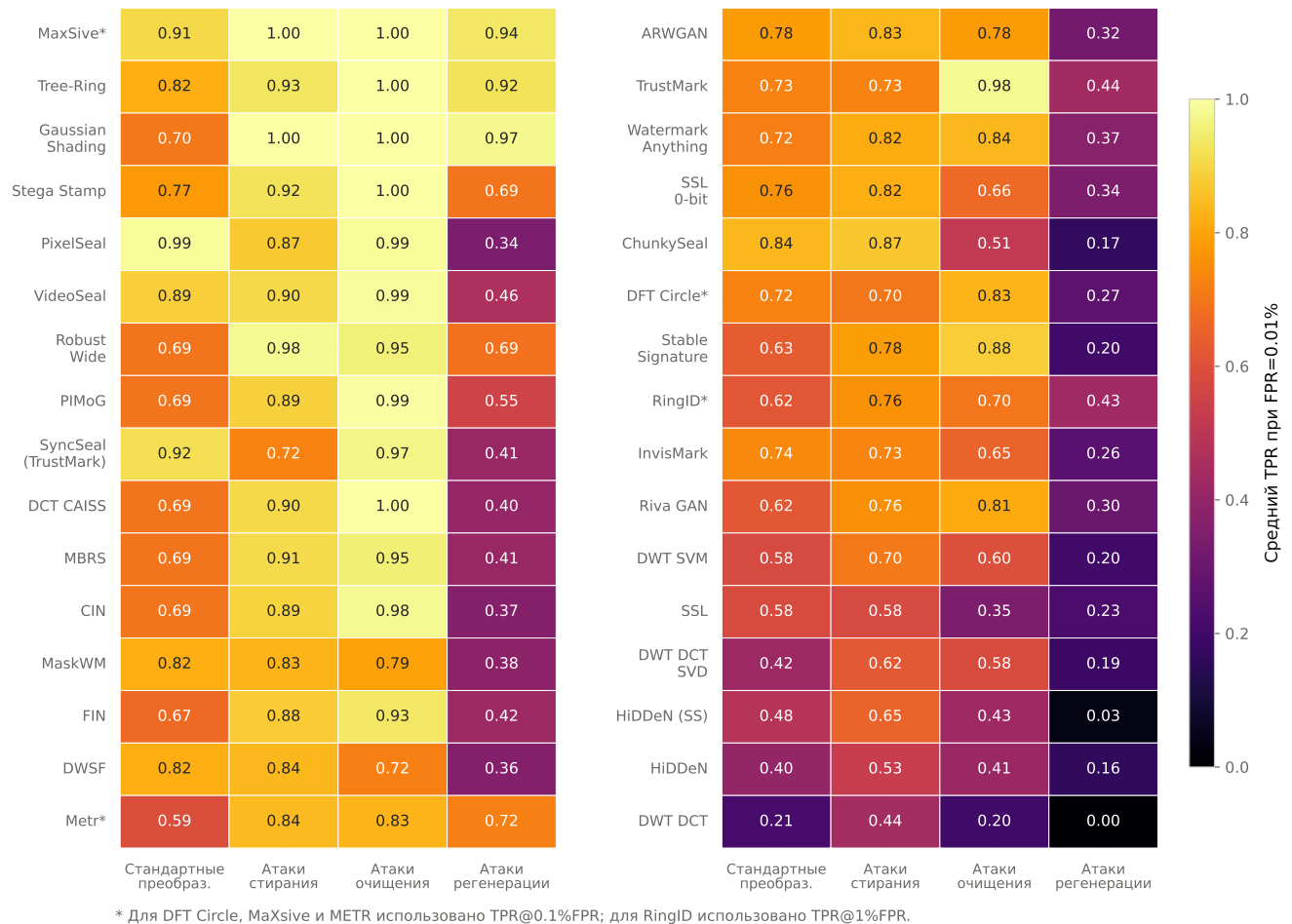


Рисунок 2.5 — Средняя устойчивость исследованных методов маркирования к различным классам воздействий

рации: для ряда методов, устойчивых к стандартным преобразованиям и другим классам атак, после регенерации наблюдается заметное снижение вероятности обнаружения ЦВЗ.

Для ряда методов, внедряющих ЦВЗ непосредственно в процессе генерации изображения, наблюдается повышенная устойчивость к атакам регенерации. Наиболее выражена эта особенность для Tree-Ring и Gaussian Shading, сохраняющих высокую вероятность обнаружения также при других классах воздействий. Вместе с тем данная закономерность не является общей для всех методов этого типа: Stable Signature оказывается более чувствительным к регенерации.

Устойчивость к стандартным преобразованиям, атакам стирания и очищения также существенно различается между методами. Таким образом, высокая устойчивость к одному классу воздействий не позволяет сделать вывод об устойчивости к другим классам. Сравнительная оценка должна охватывать несколько типов преобразований и атак.

Сравнительная эффективность атак

В предыдущем подразделе устойчивость методов рассматривалась с точки зрения сохранения возможности обнаружения ЦВЗ после различных воздействий. Однако низкое значение TPR само по себе еще не означает высокую практическую эффективность атаки. Существенная модификация изображения также может нарушить обнаружение ЦВЗ, однако при этом изменяется и сам защищаемый контент. При сравнении атак необходимо учитывать одновременно степень подавления ЦВЗ и величину вносимых перцептивных искажений.

Эффективность подавления определялась как $1 - \overline{TPR}$, где $\overline{TPR}$ — среднее значение TPR по исследованным методам для данного воздействия при $FPR = 0,01\%$. Перцептивные изменения характеризовались средним значением LIPS. Рассматривались воздействия с $LIPS \leq 0,3$; это значение используется как граница исследуемой области и не является универсальным порогом допустимости искажений.

Результаты представлены на рисунке 2.6. Воздействия, обеспечивающие большее подавление ЦВЗ при меньшем LIPS, соответствуют более эффективному сочетанию рассматриваемых характеристик; дополнительно на рисунке выделена граница Парето.

Наибольшее снижение вероятности обнаружения ЦВЗ среди рассмотренных воздействий наблюдается для атак регенерации Flux Rinsing и Flux Regen. Вместе с тем высокая эффективность этих атак сопровождается сравнительно значительными перцептивными искажениями.

Иное соотношение между эффективностью подавления и величиной искажений наблюдается для SADRE и Diffusion Regen. Эти воздействия также существенно снижают вероятность обнаружения ЦВЗ, однако вносят меньшие изменения в изображение по метрике LIPS. Для большинства остальных атак данного класса снижение вероятности обнаружения ЦВЗ выражено заметно слабее.

Атаки очищения характеризуются более неоднородными результатами. Наиболее выраженное снижение вероятности обнаружения ЦВЗ в этой группе наблюдается для NRP Large, NRP Small и DIP, тогда как DiffPure, RealESRGAN и MPRNet оказывают меньшее влияние на детектирование ЦВЗ. Атака DISCO демонстрирует наилучшее соотношение эффективности в этой группе.



Рисунок 2.6 — Сравнительная эффективность атак с учетом величины перцептивных искажений

Стандартные преобразования преимущественно располагаются в нижней части рисунка и в среднем значительно слабее влияют на обнаружение ЦВЗ. Сжатие JPEG, размытие, добавление шума и изменение размера изображения приводят к меньшему снижению TPR по сравнению с наиболее эффективными преднамеренными атаками. Это показывает, что устойчивость к типовым преобразованиям изображения не является достаточным показателем устойчивости метода к специализированным воздействиям, направленным на подавление ЦВЗ.

2.4 Сравнительная оценка методов маркирования аудио

Для проверки применимости разработанной методики к различным типам медиаконтента проведена сравнительная оценка методов маркирования аудиоданных. В соответствии с общей моделью экспериментального процесса для

аудиомодальности определены специализированные показатели качества, набор воздействий и условия проведения эксперимента.

2.4.1 Метрики оценки

Для оценки незаметности маркирования аудиоданных используются показатели с эталонным сигналом и безэталонные показатели воспринимаемого качества. Основные используемые метрики приведены в таблице 5.

SI-SNR характеризует величину изменения сигнала относительно исходной записи. Для речевых данных дополнительно используются PESQ [116] и STOI [117], оценивающие соответственно воспринимаемое качество и разборчивость речи, а также безэталонные показатели NISQA [118] и DNSMOS [119]. Сохранение характеристик голоса диктора оценивается посредством SECS, представляющего косинусное сходство векторных представлений исходного и маркированного аудиосигналов, полученных с использованием предобученной модели ECAPA-TDNN [120].

Перечисленные показатели характеризуют различные свойства аудиосигнала и не являются взаимозаменяемыми. При этом PESQ, STOI, NISQA, DNSMOS и SECS ориентированы преимущественно на речевые данные, что учитывается при интерпретации результатов.

Для оценки успешности извлечения и обнаружения ЦВЗ используются BER, WER, p -value, $\text{TPR}@x\text{FPR}$ и Empirical $\text{TPR}@x\text{FPR}$. Их определения и способ сведения multi-bit и zero-bit методов к задаче обнаружения приведены в подразделе 2.3.1 и не зависят от модальности исследуемого объекта.

2.4.2 Преобразования и преднамеренные атаки

Устойчивость методов маркирования аудио оценивается относительно воздействий, возникающих как при обычной обработке и распространении аудиосигнала, так и при преднамеренном нарушении обнаружения или извлечения ЦВЗ. В отличие от изображений, для аудиоданных существенны не только изменения

Таблица 5 — Метрики оценки методов маркирования аудио

Метрика	Оцениваемое свойство	Группа	Описание
SI-SNR	качество аудиосигнала	С эталонным сигналом	Отношение энергии исходного сигнала к энергии искажения без учета различий общего масштаба амплитуды
PESQ	воспринимаемое качество речи	С эталонным сигналом	Перцептивная оценка искажений речевого сигнала относительно эталона
STOI	разборчивость речи	С эталонным сигналом	Оценка сохранения разборчивости речевого сигнала после внесения изменений
NISQA	воспринимаемое качество речи	Безэталонная	Нейросетевая оценка воспринимаемого качества речевого сигнала
DNSMOS	воспринимаемое качество речи	Безэталонная	Прогноз субъективной оценки качества речи с использованием нейросетевой модели
SECS	сохранение характеристик диктора	С эталонным сигналом	Косинусное сходство векторных представлений диктора для исходного и маркированного сигналов
BER	успешность извлечения ЦВЗ	Сравнение сообщений	Доля ошибочно извлеченных битов
WER	успешность извлечения ЦВЗ	Сравнение сообщений	Доля сообщений, содержащих хотя бы одну ошибку извлечения
p -value	обнаружение ЦВЗ	Статистический показатель	Статистическая значимость соответствия результата заданному ЦВЗ
TPR@ x FPR	обнаружение ЦВЗ	Теоретический порог	Доля правильно обнаруженных ЦВЗ при заданном уровне ложных срабатываний
Empirical TPR@ x FPR	обнаружение ЦВЗ	Эмпирический порог	TPR при пороге, определенном по выборке немаркированных аудиофрагментов

амплитудных и спектральных характеристик, но и нарушения временной структуры сигнала. Последние могут изменять положение маркированных фрагментов относительно окон извлечения и тем самым приводить к потере синхронизации без существенного изменения воспринимаемого содержания записи.

Рассматриваемые воздействия разделены на пять групп. Стандартные преобразования включают типовые операции обработки сигнала, такие как передискретизация, фильтрация, изменение амплитуды и добавление шума. Отдельно рассматривается сжатие с потерями с использованием традиционных и нейросетевых аудиокодеков. Акустические эффекты и динамическая обработка включают эхо, реверберацию, изменение высоты тона и динамических характеристик сигнала. Атаки десинхронизации изменяют временную структуру аудиозаписи и могут нарушать соответствие между маркированными фрагментами и окнами извлечения. Наконец, методы очищения и нейросетевой реконструкции подавляют или преобразуют компоненты сигнала, среди которых могут находиться изменения, содержащие ЦВЗ. Основные группы воздействий, используемых при сравнительной оценке методов маркирования аудио, приведены в таблице 6.

Как и при оценке методов маркирования изображений, снижение успешности обнаружения или извлечения ЦВЗ необходимо анализировать совместно с изменениями самого медиаконтента. Воздействие, полностью подавляющее ЦВЗ за счет существенного ухудшения качества или разборчивости аудио, неэквивалентно воздействию, достигающему аналогичного результата при сохранении воспринимаемого качества сигнала. Показатели устойчивости рассматриваются совместно с метриками, описанными в подразделе 2.4.1.

2.4.3 Условия проведения эксперимента

Для сравнительной оценки в WIBE были интегрированы 14 методов маркирования аудиоданных, включающих классические и нейросетевые подходы. Основные принципы построения соответствующих методов рассмотрены в разделе 1.3. В экспериментальный набор вошли AudioSeal, AWARE, SilentCipher, WavMark, DNN-Audio, Perth, DCT-b1, FSVC, Spread Spectrum, PatchWork, Norm-Space, QIM, LSB и Echo Hiding.

Таблица 6 — Преобразования и атаки, используемые при оценке устойчивости методов маркирования аудио

Класс воздействия	Примеры	Принцип воздействия
Стандартные преобразования сигнала	Sign Inversion, Resampling, Requantization, Gain, Filtering, White Noise, Pink Noise, Clipping	Изменение амплитудных, частотных и статистических характеристик сигнала при типовой обработке
Сжатие с потерями	MP3, AAC, Opus, Encodec [121], DAC [122]	Перцептивное кодирование с удалением или изменением компонентов сигнала, считающихся малозначимыми для слухового восприятия
Акустические эффекты и динамическая обработка	Echo, Reverb, Pitch Shift, Dynamic Range Compressor, Limiter	Изменение временных, спектральных и амплитудных характеристик аудиосигнала при сохранении его основного содержания
Атаки десинхронизации	Cut, Speed, Time Stretch, Inverted Time Stretch, Flip Samples, Zero-Cross Inserts, Replacement Attack, Frame Dropout	Изменение временной структуры сигнала и нарушение синхронизации между исходным и атакованным аудио
Очищение и нейросетевая реконструкция	Фильтр Винера, MetricGAN+ [123], GTCRN [124], Vocos [125]	Подавление возмущений или реконструкция сигнала, способные одновременно изменять компоненты встроенного ЦВЗ

По возможности использовались официальные реализации методов и опубликованные авторами предобученные модели. Различия в требуемых форматах данных, частоте дискретизации и способах представления сообщения учитывались в программных адаптерах WIBE, что позволяло выполнять сравнение в рамках единого экспериментального конвейера.

Для multi-bit методов формировались сообщения поддерживаемой соответствующей конфигурацией длины, которые использовались при последующем вычислении BER, WER и показателей обнаружения. Perth рассматривался как zero-bit метод. Метод DNN-Audio поддерживает шесть возможных сообщений, поэтому при сопоставлении емкости его эффективная информационная емкость принималась равной $\log_2 6 \approx 2,58$ бит.

Экспериментальная выборка была сформирована из нескольких наборов речевых аудиоданных. Использовались 150 фрагментов LibriSpeech [126] test-clean, 150 фрагментов LibriSpeech test-other, 200 фрагментов VCTK [127], 300 фрагментов Common Voice [128], 100 фрагментов Golos [129] и 100 фраг-

ментов AISHELL [130]. Для Common Voice применялось сбалансированное семплирование по 12 языкам. Таким образом, тестовая выборка включала 1000 аудиофрагментов и охватывала различные наборы данных, условия записи и языки. Для уменьшения влияния различий в размерах отдельных частей выборки агрегирование результатов выполнялось в два этапа. Сначала значения метрик усреднялись по объектам каждого набора данных отдельно, после чего вычислялось среднее по наборам данных с одинаковым весом.

Для каждого метода сначала выполнялось внедрение ЦВЗ и оценивалось качество маркированного сигнала. Затем независимо применялись воздействия, рассмотренные в подразделе 2.4.2, после чего выполнялись извлечение ЦВЗ и расчет показателей устойчивости. Для сравнения методов использовался показатель $TPR@10^{-3}FPR$, применимый как к multi-bit, так и к zero-bit методам.

Вычислительная эффективность оценивалась по среднему суммарному времени внедрения и извлечения ЦВЗ. Поскольку часть нейросетевых методов выполнялась с использованием GPU, полученные значения времени характеризуют практическую производительность используемых реализаций и не рассматриваются как аппаратно-независимая оценка вычислительной сложности.

2.4.4 Результаты сравнительной оценки

Для сравнительной оценки методов маркирования аудиоданных рассматривались характеристики, отражающие различные аспекты их практического применения: незаметность вносимых искажений, емкость встраиваемого сообщения, вычислительная эффективность и устойчивость к преобразованиям и атакам.

Незаметность, емкость и вычислительная эффективность

На рисунке 2.7 приведено сопоставление методов по емкости ЦВЗ, его незаметности и времени работы реализованных алгоритмов. По горизонтальной оси отложена емкость встраиваемого сообщения, по вертикальной — значение SI-SNR после внедрения ЦВЗ. Цветом точки показано среднее суммарное время

внедрения и извлечения ЦВЗ. Для метода Perth, не предусматривающего встраивания информационного сообщения, выделена отдельная категория zero-bit. Символом «*» отмечены методы, время работы которых измерялось при выполнении вычислений с использованием GPU.

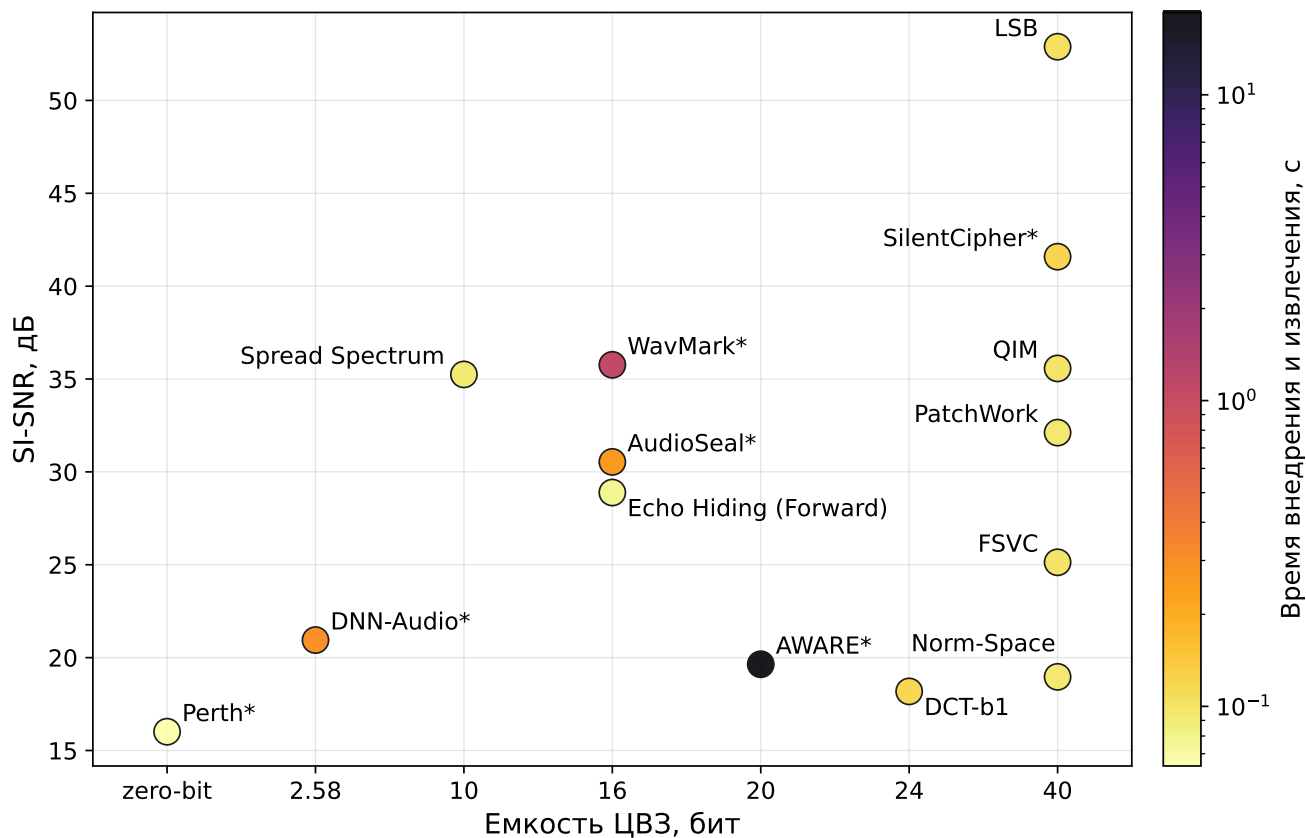


Рисунок 2.7 — Сопоставление методов маркирования аудиоданных по емкости ЦВЗ, его незаметности и вычислительной эффективности реализованных алгоритмов

Полученные результаты демонстрируют существенный разброс характеристик рассматриваемых методов. Наибольшее значение SI-SNR наблюдается для LSB и превышает 50 дБ. Значения около 35–42 дБ получены для SilentCipher, WavMark, QIM и Spread Spectrum, тогда как для ряда других методов SI-SNR находится в диапазоне приблизительно 16–32 дБ. При этом большинство классических методов поддерживает размер сообщения 40 бит, тогда как AudioSeal, WavMark и Echo Hiding используют сообщения длиной 16 бит, Spread Spectrum — 10 бит, AWARE — 20 бит, DCT-b1 — 24 бит, а эффективная емкость DNN-Audio составляет около 2,58 бит.

Полученные значения показывают, что емкость сообщения сама по себе не определяет величину вносимых искажений. Например, LSB и QIM при емкости

40 бит обеспечивают более высокое значение SI-SNR, чем ряд методов с меньшей емкостью. Следовательно, показатели емкости и качества маркированного сигнала должны рассматриваться совместно с устойчивостью ЦВЗ к последующим воздействиям.

Значительно различается и вычислительная эффективность методов. Суммарное время внедрения и извлечения изменяется более чем на порядок, причем наиболее существенные затраты наблюдаются для отдельных нейросетевых реализаций. Вместе с тем непосредственное сопоставление абсолютных значений времени имеет ориентировочный характер, поскольку часть методов выполнялась с использованием GPU, а часть — на CPU.

Устойчивость к преобразованиям и преднамеренным атакам

Для оценки устойчивости использовался показатель $\text{TPR}@10^{-3}\text{FPR}$. Полученные значения устойчивости были усреднены в рамках 5 классов воздействий, заданных в разделе 2.4.2. На рисунке 2.8 представлены полученные результаты устойчивости. Методы упорядочены по среднему значению показателя по пяти классам воздействий. Более высокое значение TPR соответствует большей вероятности обнаружения ЦВЗ при фиксированной вероятности ложного обнаружения.

Наиболее высокие значения показателя в совокупности рассматриваемых воздействий получены для DNN-Audio и AWARE. Оба метода сохраняют высокую вероятность обнаружения при стандартных преобразованиях сигнала и демонстрируют сравнительно устойчивые результаты для остальных классов воздействий. При этом DNN-Audio характеризуется высокой устойчивостью к сжатию с потерями, а AWARE — к очищению и нейросетевой реконструкции.

Для ряда методов, напротив, наблюдается выраженная зависимость устойчивости от типа воздействия. Так, Norm-Space показывает высокие результаты при стандартных преобразованиях и очищении, однако существенно менее устойчив к акустической обработке и нарушениям временной синхронизации. Аналогичная тенденция наблюдается для DCT-b1 и QIM: сравнительно высокая устойчивость к отдельным преобразованиям сочетается со значительным

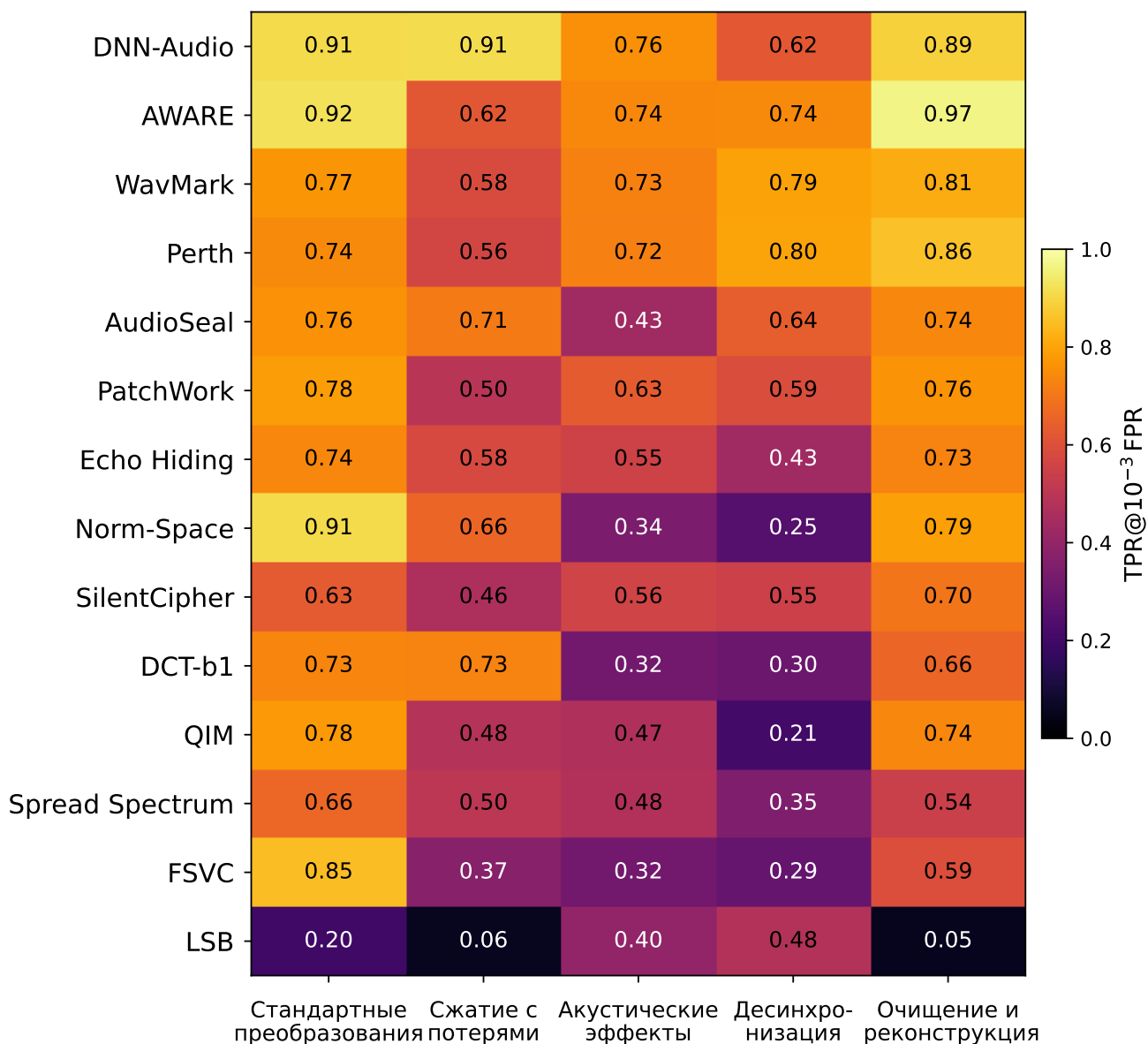


Рисунок 2.8 — Среднее значение $TPR@10^{-3}FPR$ методов маркирования аудиоданных для различных классов воздействий

снижением вероятности обнаружения при акустических воздействиях и атаках десинхронизации.

Результаты, полученные для метода LSB демонстрируют, что между незаметностью и устойчивостью требуется достигнуть компромисс. Данный метод обеспечивает наибольшее значение SI-SNR среди рассмотренных конфигураций, однако характеризуется наиболее низкой устойчивостью к большинству исследованных воздействий, в особенности к сжатию с потерями и очищению сигнала.

В целом наиболее сложными для значительной части методов оказываются акустические воздействия и атаки десинхронизации. Особенно заметное снижение показателя при нарушении временной структуры наблюдается для

Norm-Space, DCT-b1, QIM, Spread Spectrum и FSVC, тогда как WavMark и Perth сохраняют существенно более высокую устойчивость. Это показывает, что чувствительность к конкретным классам воздействий в значительной степени определяется принципом формирования и обнаружения ЦВЗ.

Совместный анализ незаметности и устойчивости показывает отсутствие универсальной конфигурации, максимизирующей все рассматриваемые характеристики одновременно. DNN-Audio и AWARE демонстрируют высокие значения устойчивости для большинства классов воздействий, однако их значения SI-SNR заметно ниже, чем у ряда методов с меньшей устойчивостью. WavMark сочетает сравнительно высокое значение SI-SNR с устойчивостью к различным классам воздействий. Напротив, LSB обеспечивает минимальное влияние маркирования на сигнал, но характеризуется низкой устойчивостью к большинству рассматриваемых преобразований. Полученные результаты подтверждают необходимость комплексной оценки методов маркирования аудиоданных, которая должна учитывать качество маркированного сигнала, емкость, вычислительную эффективность и устойчивость ЦВЗ к воздействиям.

Сравнительная эффективность атак

Для анализа эффективности атак были выбраны три метода маркирования аудиоданных: AudioSeal, WavMark и SilentCipher. Рассматривались три группы воздействий: сжатие с потерями, акустические эффекты, а также очищение и реконструкция сигнала. В каждой группе исследовалось несколько типов воздействий и их конфигураций, отличающихся интенсивностью или параметрами обработки.

Для каждого воздействия отдельно для трех методов рассчитывались показатель обнаружения ЦВЗ и метрика качества SI-SNR после атаки, после чего полученные значения усреднялись по выбранным методам. Эффективность подавления определялась как $1 - \overline{TPR}$, где $\overline{TPR}$ — среднее значение TPR для AudioSeal, WavMark и SilentCipher при уровне $FPR = 10^{-3}$. Результаты представлены на рисунке 2.9.

Наиболее сильное подавление ЦВЗ наблюдается для алгоритмов сжатия при помощи нейронных сетей, значительного изменения высоты тона и рекон-

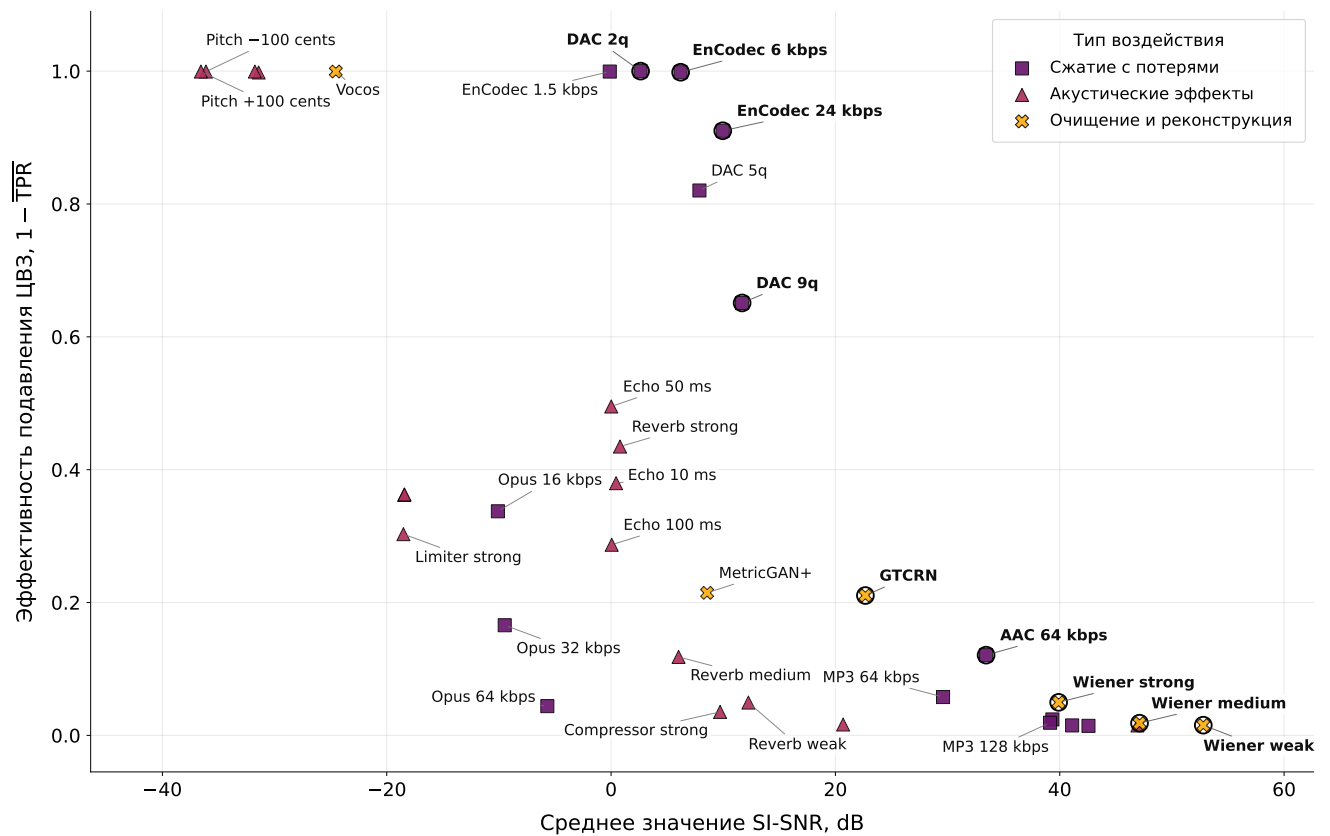


Рисунок 2.9 — Сравнительная эффективность атак на методы маркирования аудиоданных с учетом величины искажений аудиосигнала

струкции с помощью Vocos. Данные воздействия сопровождаются существенным снижением значения SI-SNR. При увеличении качества сжатого аудио степень подавления снижается, что особенно заметно для различных конфигураций EnCodec и DAC.

Методы очищения и реконструкции демонстрируют компромисс: MetricGAN+ и GTCRN, обеспечивают заметное снижение вероятности обнаружения при умеренных искажениях, тогда как фильтр Винера лучше сохраняет качество аудио, но слабее воздействует на ЦВЗ.

2.5 Выводы по главе

В настоящей главе разработана методика унифицированной воспроизводимой оценки методов маркирования медиаконтента, основанная на единой модели экспериментального процесса. В рамках предложенной модели выделены самостоятельные стадии формирования маркированного объекта, оценки

качества медиаконтента, применения преобразований и преднамеренных атак, обнаружения или извлечения ЦВЗ и агрегации результатов. Общая организация эксперимента применима к методам с различными принципами формирования и проверки ЦВЗ, при этом представления данных, наборы воздействий и метрики оценки определяются особенностями исследуемой модальности.

Предложенная методика реализована в программной системе WIBE. Модульная архитектура системы обеспечивает независимое подключение алгоритмов маркирования, наборов данных, воздействий и метрик без изменения ядра экспериментального конвейера. Воспроизводимое описание экспериментов осуществляется посредством декларативных конфигураций, фиксирующих состав компонентов и параметры их применения. Для проведения экспериментов с различными программными реализациями предусмотрены изоляция зависимостей, сохранение промежуточных состояний, раздельное выполнение стадий и параллельная обработка объектов.

Экспериментальная апробация методики выполнена для изображений и аудиоданных. Проведена сравнительная оценка 32 конфигураций методов маркирования изображений, включая методы внедрения ЦВЗ в сформированный контент и методы маркирования в процессе генерации, а также 14 классических и нейросетевых методов маркирования аудио. Для обеих модальностей использовалась общая последовательность стадий экспериментального конвейера, дополненная специализированными наборами данных, моделями воздействий и показателями качества.

Полученные результаты показали зависимость устойчивости ЦВЗ от типа воздействия. Для изображений высокая устойчивость к стандартным преобразованиям не гарантирует сохранения возможности обнаружения при атаках стирания, очищения и регенерации. При маркировании аудиоданных сильное влияние на возможность обнаружить ЦВЗ оказывают акустические воздействия и нарушения временной синхронизации. Установлено, что незаметность ЦВЗ и емкость сообщения сами по себе не определяют устойчивость. Анализ эффективности атак показал необходимость совместного учета степени подавления ЦВЗ и величины вносимых изменений в медиаконтент.

Проведенная апробация продемонстрировала применимость разработанной методики для сравнительного исследования различных методов маркирования изображений и аудиоданных при использовании единой экспериментальной

модели. Ее программная реализация в WIBE позволяет проводить такие исследования в воспроизводимых условиях.

Глава 3. Метод контроля целостности изображений на основе подписываемого перцептивного представления

В предыдущей главе показано, что современные методы маркирования позволяют обеспечивать высокую незаметность и устойчивость ЦВЗ к различным преобразованиям изображений. Вместе с тем сохранение и успешное обнаружение водяного знака само по себе не подтверждает его подлинность и не устанавливает связь между ЦВЗ и содержимым конкретного изображения. Отсутствие такой связи позволяет перенести корректно извлекаемый ЦВЗ между различными изображениями или сформировать маркировку, не связанную с исходным содержимым [79; 80].

В настоящей главе рассматривается задача построения метода контроля целостности изображений, в котором результат проверки определяется не только наличием ЦВЗ, но и соответствием аутентификатора содержимому изображения и заданному открытому ключу. Для этого предлагается метод CPW (Cryptographic Perceptual Watermarking), основанный на формировании подписываемого перцептивного представления изображения с использованием перцептивного хеширования, электронной цифровой подписи и цифрового водяного знака. Такое сочетание должно обеспечивать устойчивость проверки к допустимым преобразованиям изображения и одновременно препятствовать переносу или формированию корректного аутентификатора для содержательно измененного изображения без знания закрытого ключа.

3.1 Постановка задачи контроля целостности изображений

В настоящей работе ставится задача разработки метода контроля целостности изображений, позволяющего публично проверить соответствие исследуемого изображения ранее аутентифицированному содержимому с точностью до заданного множества допустимых преобразований. Проверка должна выполняться без доступа к исходному немаркированному изображению и закрытому ключу; для верификации используются только исследуемое изображение, открытый ключ и общедоступные параметры метода.

Пусть I — исходное изображение, I_w — сформированное на его основе маркированное изображение, K_{pub} — открытый ключ стороны, выполнившей аутентификацию, а $V(I, K_{\text{pub}})$ — процедура верификации. Для множества допустимых преобразований T_A требуется выполнение условия

$$V(T(I_w), K_{\text{pub}}) = 1, \quad T \in T_A. \quad (3.1)$$

Для множества недопустимых преобразований T_N должно выполняться

$$V(T(I_w), K_{\text{pub}}) = 0, \quad T \in T_N. \quad (3.2)$$

Таким образом, под сохранением целостности в данной работе понимается не точное сохранение пиксельного представления изображения, а соответствие исследуемого изображения аутентифицированному содержимому в пределах заданной политики допустимых преобразований. Отрицательный результат верификации означает, что такое соответствие не может быть подтверждено, однако сам по себе не позволяет установить конкретную причину отказа или тип выполненного преобразования.

3.1.1 Модель нарушителя

При построении метода предполагается, что нарушителю известны алгоритмы формирования маркированного изображения и его последующей верификации, а также используемые параметры метода. Нарушителю доступен открытый ключ K_{pub} , относительно которого выполняется проверка. Закрытый ключ соответствующей стороны считается недоступным. Нарушитель может располагать одним или несколькими маркированными изображениями, сформированными с использованием одного закрытого ключа, и многократно выполнять общедоступную процедуру верификации. Исходные немаркированные изображения и закрытый ключ нарушителю недоступны.

Целью нарушителя является получение содержательно измененного изображения, которое продолжает успешно проходить проверку относительно открытого ключа исходной стороны. Иными словами, нарушитель стремится найти

такое недопустимое преобразование $T \in T_N$, для которого выполняется

$$V(T(I_w), K_{\text{pub}}) = 1. \quad (3.3)$$

Такое преобразование рассматривается как успешная атака на контроль целостности, поскольку изображение, измененное способом, отнесенным принятой политикой к недопустимым, ошибочно принимается процедурой верификации как соответствующее ранее аутентифицированному содержимому.

Отдельно возможны воздействия, направленные не на формирование содержательно измененного изображения, успешно проходящего проверку, а на нарушение самой процедуры верификации, например путем повреждения или удаления ЦВЗ. Такие воздействия приводят к отказу проверки корректно маркированного изображения и относятся прежде всего к нарушению доступности процедуры контроля целостности. Формирование нарушителем маркированного изображения с использованием собственной пары ключей также не считается успешной атакой в рамках принятой модели, поскольку верификация выполняется относительно заранее заданного открытого ключа исходной стороны.

3.1.2 Допустимые и недопустимые преобразования

Результат контроля целостности зависит от принятой политики обработки изображения. Для ее задания рассматриваются два множества воздействий: допустимые преобразования, после которых изображение должно продолжать успешно проходить проверку, и недопустимые преобразования, применение которых должно приводить к отрицательному результату верификации. Такое разделение определяется не только величиной визуальных изменений, но и назначением преобразования и его влиянием на содержимое изображения.

К допустимым относятся преобразования, характерные для хранения, передачи и публикации изображений и не изменяющие их основное семантическое содержание:

- сжатие с потерями по алгоритму JPEG;
- фильтрация изображения;
- добавление аддитивного гауссовского шума;
- изменение яркости и контрастности;

– масштабирование с сохранением соотношения сторон.

Указанные воздействия изменяют пиксельное представление изображения и могут влиять на содержащийся в нем сигнал ЦВЗ, однако в рассматриваемой модели не должны интерпретироваться как нарушение целостности. Конкретные значения параметров и диапазоны интенсивности таких преобразований устанавливаются в экспериментальной части главы.

К недопустимым относятся преобразования, приводящие к значительному изменению семантического содержания изображения. В экспериментальной части рассматриваются локальные вставка, удаление и клонирование фрагментов изображения. После таких преобразований результат верификации должен быть отрицательным.

Отдельную группу составляют геометрические преобразования, не поддерживаемые рассматриваемой реализацией метода: кадрирование, поворот, зеркальное отражение и изменение соотношения сторон посредством неравномерного масштабирования. Такие операции не обязательно являются содержательной модификацией, однако не входят в множество допустимых преобразований. Сохранение положительного результата верификации после их применения в настоящей работе не требуется.

Разделение на допустимые и недопустимые преобразования определяет заявленную область работоспособности метода и не является универсальной классификацией изменений изображений. Для преобразований и диапазонов параметров, не включенных в рассматриваемую модель, сохранение положительного результата верификации не гарантируется.

Таким образом, разрабатываемый метод должен обеспечивать положительный результат верификации после преобразований из множества T_A и отрицательный результат после преобразований из множества T_N , в том числе при действиях нарушителя в рамках принятой модели. Проверка должна выполняться без доступа к исходному немаркированному изображению и закрытому ключу. При этом формирование маркированного изображения не должно приводить к существенному ухудшению его визуального качества.

3.2 Метод контроля целостности изображений CPW

Для выполнения сформулированных условий необходимо обеспечить проверяемую связь результата верификации с содержимым изображения и заданным открытым ключом. Используемое при проверке представление содержимого должно сохраняться после допустимых преобразований и изменяться при недопустимых модификациях, а его связь с заданным открытым ключом должна подтверждаться без доступа к закрытому ключу. Для решения поставленной задачи разработан метод CPW (Cryptographic Perceptual Watermarking) [8], объединяющий перцептивное хеширование, электронную цифровую подпись и цифровой водяной знак.

Общая схема CPW изображена на рисунке 3.1. Она включает два основных процесса: формирование маркированного изображения и его последующую верификацию. На этапе формирования для исходного изображения при помощи перцептивной хеш-функции вычисляется устойчивое перцептивное представление. Полученный перцептивный хеш-код преобразуется криптографической хеш-функцией и подписывается закрытым ключом владельца. Сформированная цифровая подпись кодируется помехоустойчивым кодом и внедряется в изображение посредством алгоритма ЦВЗ. На этапе проверки выполняются обратные операции: из исследуемого изображения извлекается внедренное сообщение, восстанавливается цифровая подпись, повторно вычисляется перцептивный хеш-код и выполняется проверка подписи с использованием открытого ключа.

Ключевой особенностью предлагаемой схемы является разделение функций между ее компонентами. Перцептивная хеш-функция формирует контентно-зависимое представление изображения, которое должно сохраняться после допустимых преобразований и изменяться при недопустимых содержательных модификациях. Цифровая подпись обеспечивает связь аутентификатора с заданным открытым ключом и препятствует формированию корректного аутентификатора для нового хеш-значения без знания соответствующего закрытого ключа. ЦВЗ обеспечивает передачу аутентификатора совместно с самим изображением и возможность его слепого извлечения. Помехоустойчивое кодирование компенсирует часть битовых ошибок, возникающих при извлечении ЦВЗ после допустимой обработки изображения.

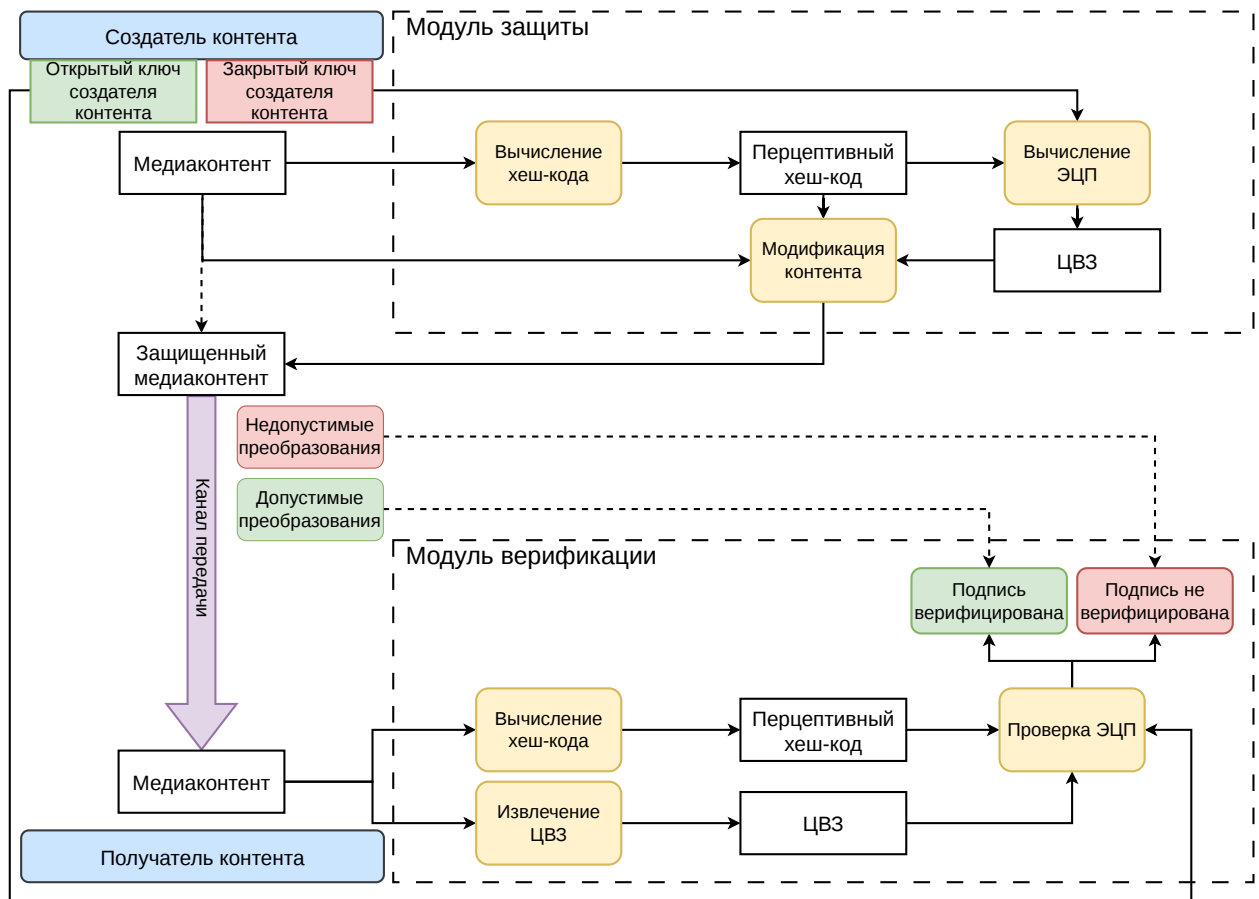


Рисунок 3.1 — Схема метода CPW для контроля целостности изображений

3.2.1 Перцептивная хеш-функция EPHash

Традиционные перцептивные хеш-функции ориентированы преимущественно на поиск похожих изображений и допускают небольшое расстояние между хеш-кодами визуально сходных объектов [131]. В предлагаемой схеме такой подход неприменим непосредственно: цифровая подпись формируется для конкретной битовой последовательности, поэтому изменение даже одного бита подписываемого хеш-кода приводит к отрицательному результату проверки цифровой подписи.

Для формирования подписываемого перцептивного представления разработана перцептивная хеш-функция EPHash (Enhanced Perceptual Hash), основанная на квантовании низкочастотных коэффициентов дискретного косинусного преобразования [132]. Особенность EPHash заключается в том, что на этапе формирования маркированного изображения выполняется однократная контролируемая модификация его ДКП-коэффициентов, увеличивающая расстояние от порогов

квантования. В дальнейшем на стороне проверяющего хеш-код вычисляется без дополнительной модификации изображения.

Пусть исходное изображение обозначено I . Для вычисления EPHash оно преобразуется в градации серого и масштабируется к квадратному разрешению со стороны

$$L = n_b h_s h_{ff}, \quad (3.4)$$

где n_b — количество блоков по каждой координате, h_s — количество низкочастотных коэффициентов ДКП, выбираемых по каждой координате блока, а h_{ff} — масштабный коэффициент, определяющий размер блока ДКП и относительную долю используемой низкочастотной области. Полученное изображение разделяется на $n_b \times n_b$ непересекающихся блоков размером $h_s h_{ff} \times h_s h_{ff}$.

Для каждого блока вычисляется двумерное дискретное косинусное преобразование. Пусть C обозначает полученную матрицу коэффициентов ДКП. Для формирования хеш-кода из нее выбирается область

$$\Omega = \{(u, v) \mid 2 \leq u \leq h_s + 1, 2 \leq v \leq h_s + 1\}. \quad (3.5)$$

Таким образом, для каждого блока используется h_s^2 коэффициентов ДКП, причем коэффициенты первой строки и первого столбца в хешировании не участвуют.

Для множества выбранных коэффициентов вычисляется медианное значение m . Каждый коэффициент преобразуется в один бит посредством сравнения с медианой:

$$b_{u,v} = \begin{cases} 1, & C_{u,v} \geq m, \\ 0, & C_{u,v} < m. \end{cases} \quad (3.6)$$

Полученные для всех блоков биты объединяются в фиксированном порядке и образуют перцептивный хеш-код

$$H = \text{EPHash}(I). \quad (3.7)$$

Его длина составляет

$$|H| = n_b^2 h_s^2. \quad (3.8)$$

Основной причиной побитовой нестабильности такого квантования являются коэффициенты, расположенные в непосредственной близости от медианы. Даже небольшое изменение изображения может привести к пересечению коэффициентом порога m , вследствие чего соответствующий бит изменит значение. Для уменьшения вероятности такого перехода при формировании маркированного изображения создается минимальный отступ от порога бинаризации $\delta_{\min}$.

Если абсолютное расстояние выбранного коэффициента от медианы меньше $\delta_{\min}$, его значение модифицируется в зависимости от ранее сформированного бита:

$$C'_{u,v} = \begin{cases} \max(C_{u,v}, m + \delta_{\min}), & b_{u,v} = 1, \\ \min(C_{u,v}, m - \delta_{\min}), & b_{u,v} = 0. \end{cases} \quad (3.9)$$

Модификация выполняется в сторону, соответствующую уже определенному значению бита. Поэтому положение коэффициента относительно порога квантования не меняется, изменяется только расстояние от него. В результате формируемый хеш-код остается тем же, но вероятность изменения отдельных битов после последующих преобразований снижается.

Значение $\delta_{\min}$ определяется отдельно для каждого блока. Его увеличение повышает устойчивость хеш-кода, но одновременно увеличивает искажение изображения. Поэтому $\delta_{\min}$ подбирается методом бинарного поиска при ограничении на минимально допустимое качество восстановленного блока. В качестве такого ограничения используется порог T_{PSNR} : выбирается максимально возможный отступ, при котором

$$\text{PSNR}(B, B') \geq T_{\text{PSNR}}, \quad (3.10)$$

где B и B' — соответственно исходный и модифицированный блоки.

После изменения коэффициентов выполняется обратное ДКП, а модифицированные блоки объединяются в изображение. Если хеширование выполнялось в уменьшенном разрешении, разность между исходным и модифицированным представлениями масштабируется до исходного размера и добавляется к яркостному каналу исходного изображения. В результате формируется модифицированное изображение I_m .

Принципиальным требованием данной процедуры является сохранение сформированного значения хеш-кода:

$$\text{EPHash}(I_m) = \text{EPHash}(I) = H. \quad (3.11)$$

Таким образом, процедура модификации не изменяет подписываемую битовую последовательность, а только увеличивает устойчивость ее последующего воспроизведения. При проверке изображения подбор $\delta_{\min}$ и изменение коэффициентов не выполняются: проверяющая сторона непосредственно вычисляет $\text{EPHash}(I')$ для полученного изображения I' .

3.2.2 Формирование аутентификатора

Перцептивный хеш-код H содержит информацию о визуальном содержимом изображения, однако может быть вычислен любым участником системы и сам по себе не позволяет подтвердить его связь с заданным открытым ключом. Для формирования такой проверяемой связи и предотвращения подделки аутентификатора в CPW используется электронная цифровая подпись.

Сначала для перцептивного хеш-кода вычисляется значение криптографической хеш-функции SHA-256, после чего с использованием закрытого ключа K_{priv} формируется цифровая подпись по алгоритму ECDSA [133]:

$$S = \text{ECDSA. Sign}(K_{\text{priv}}, \text{SHA256}(H)). \quad (3.12)$$

В используемой реализации применяется эллиптическая кривая secp256k1 . Цифровая подпись S представляется битовой последовательностью длины 512 бит.

Полученная подпись одновременно связана с содержимым изображения посредством H и может быть проверена относительно заданного открытого ключа K_{pub} . Перенос подписи на изображение с другим значением EPHash приводит к тому, что повторно вычисленное значение H отличается от подписанного и проверка цифровой подписи завершается отрицательным результатом. Формирование новой корректной подписи для другого значения H без доступа к K_{priv} определяется стойкостью используемой схемы ECDSA.

При этом для успешной проверки необходимо восстановить подпись S из изображения побитово точно. Перед внедрением она дополнительно защищается помехоустойчивым кодом.

3.2.3 Помехоустойчивое кодирование и внедрение аутентификатора

Для компенсации ошибок, возникающих при извлечении ЦВЗ, 512-битная подпись S кодируется двоичным кодом БЧХ [134]. В используемой конфигурации формируется кодовое слово длиной 1023 бита, способное исправлять до 57 битовых ошибок. Для получения длины, удобной для используемых методов маркирования, к кодовому слову добавляется один нулевой бит:

$$P = \text{BCH.Encode}(S) \parallel 0, \quad |P| = 1024. \quad (3.13)$$

В CPW рассматриваются две стратегии внедрения полученной последовательности P . Первая стратегия предполагает использование метода ЦВЗ, непосредственно поддерживающего емкость 1024 бита. В рассматриваемой реализации для этого применяется ChunkySeal:

$$I_w = \text{Embed}_{\text{ChunkySeal}}(I_m, P), \quad (3.14)$$

где I_m — изображение после формирования устойчивого значения EPHash, а I_w — итоговое маркированное изображение.

Для применения методов с меньшей емкостью предложена вторая стратегия, основанная на независимом внедрении отдельных фрагментов сообщения. Последовательность P разделяется на четыре равные части:

$$P = P_1 \parallel P_2 \parallel P_3 \parallel P_4, \quad |P_i| = 256, \quad i = 1, \dots, 4. \quad (3.15)$$

Изображение I_m также разделяется на четыре непересекающихся квадранта $I_m^{(i)}$. Каждый фрагмент сообщения независимо внедряется в соответствующий квадрант:

$$I_w^{(i)} = \text{Embed}(I_m^{(i)}, P_i), \quad i = 1, \dots, 4. \quad (3.16)$$

После внедрения квадранты объединяются в итоговое изображение:

$$I_w = \text{Merge}(I_w^{(1)}, I_w^{(2)}, I_w^{(3)}, I_w^{(4)}). \quad (3.17)$$

Такой подход позволяет получить суммарную емкость 1024 бита при использовании алгоритмов, поддерживающих сообщения длиной 256 бит. В

экспериментальной реализации данной стратегии рассматриваются PixelSeal, VideoSeal и MBRS.

Следует отметить, что используемый метод маркирования является заменяемым компонентом CPW. Он не обеспечивает защиту от подделки, а отвечает за передачу аутентификатора через изображение. Вместо рассмотренных решений могут применяться другие методы маркирования, если они обеспечивают требуемые емкость, незаметность и устойчивость извлечения.

3.2.4 Верификация целостности изображения

Верификация CPW является публичной и слепой: для ее выполнения не требуются исходное немаркированное изображение и закрытый ключ. Проверяющая сторона располагает исследуемым изображением I' , открытым ключом K_{pub} и общедоступными параметрами метода.

На первом этапе из изображения извлекается сообщение

$$P' = \text{Extract}(I'). \quad (3.18)$$

При использовании квадрантной стратегии извлечение выполняется независимо для каждой из четырех частей изображения, после чего полученные фрагменты объединяются в исходном порядке.

Из последовательности P' удаляется добавленный при формировании аутентификатора бит дополнения, а оставшееся 1023-битное кодовое слово декодируется кодом БЧХ:

$$S' = \text{BCH.Decode}(P'). \quad (3.19)$$

Если количество ошибок находится в пределах корректирующей способности кода, восстанавливается цифровая подпись S' . Если извлечение сообщения или БЧХ-декодирование завершается неудачей, изображение считается не прошедшим верификацию.

Для исследуемого изображения независимо вычисляется перцептивный хеш-код

$$H' = \text{ERHash}(I'). \quad (3.20)$$

При этом процедура модификации ДКП-коэффициентов, используемая при формировании маркированного изображения, на этапе верификации не выполняется.

На заключительном этапе восстановленная подпись проверяется с использованием открытого ключа:

$$v = \text{ECDSA. Verify}(K_{\text{pub}}, \text{SHA256}(H'), S'), \quad v \in \{\text{true}, \text{false}\}. \quad (3.21)$$

Если извлечение и декодирование аутентификатора завершились успешно и $v = \text{true}$, целостность изображения относительно заданного открытого ключа считается подтвержденной в пределах принятой политики допустимых преобразований. Положительный результат означает, что извлеченная из изображения подпись соответствует его текущему значению ERHash и была сформирована закрытым ключом, соответствующим K_{pub} .

Если проверка подписи завершается отрицательным результатом либо аутентификатор не удастся корректно извлечь и декодировать, изображение считается не прошедшим верификацию. Причиной могут являться семантическая модификация, неподдерживаемое преобразование, допустимое воздействие за пределами установленного диапазона, повреждение или удаление ЦВЗ либо использование некорректного аутентификатора. Отрицательный результат сам по себе не позволяет определить конкретную причину отказа, а означает только невозможность подтвердить целостность изображения относительно заданного открытого ключа.

В рамках принятой модели нарушителя механизмы CPW выполняют взаимодополняющие функции. Перенос корректной подписи на изображение с отличающимся значением ERHash приводит к отрицательному результату проверки. Формирование корректной подписи для нового значения хеш-кода без доступа к закрытому ключу считается вычислительно неосуществимым в предположении стойкости используемой схемы ECDSA. Возможность семантически изменить изображение при сохранении исходного значения ERHash, в свою очередь, определяется устойчивостью перцептивной хеш-функции к поиску второго прообраза.

3.3 Экспериментальное исследование метода CPW

Экспериментальная оценка предложенного метода проводилась с использованием программной системы WIBE, описанной во второй главе. Исследовались свойства как отдельных компонентов, так и полной процедуры контроля целостности. Исследование проводится по четырем направлениям. Сначала оценивалась побитовая устойчивость разработанной перцептивной хеш-функции EPHash к допустимым преобразованиям в сравнении с существующими перцептивными хеш-функциями. Затем исследовалась незаметность формирования маркированного изображения для различных алгоритмов внедрения ЦВЗ. После этого оценивались устойчивость полной процедуры верификации CPW к допустимым преобразованиям. Наконец, определялась способность метода выявлять содержательно значимые модификации изображения.

3.3.1 Параметры реализации и условия экспериментов

Во всех экспериментах использовалась единая конфигурация перцептивной хеш-функции EPHash. Изображение при вычислении хеш-кода приводилось к разрешению 512×512 пикселей и разделялось на четыре непересекающихся блока размером 256×256 пикселей. Для каждого блока использовалась область из 16×16 низкочастотных коэффициентов ДКП. Таким образом, параметры алгоритма, введенные в разделе 3.2.1, принимали значения

$$n_b = 2, \quad h_s = 16, \quad h_{ff} = 16. \quad (3.22)$$

В соответствии с выражением (3.8) длина формируемого перцептивного хеш-кода составляла

$$|H| = (n_b h_s)^2 = 1024 \text{ бита}. \quad (3.23)$$

При формировании отступа от порога бинаризации минимально допустимое значение PSNR между исходным блоком и его реконструкцией после изменения коэффициентов ДКП устанавливалось равным 45 дБ. Значение отступа

подбиралось отдельно для каждого блока методом бинарного поиска, выполняемого в течение 10 итераций.

Для формирования аутентификатора использовались криптографическая хеш-функция SHA-256 и алгоритм электронной цифровой подписи ECDSA на эллиптической кривой $secp256k1$. Компоненты подписи r и s представлялись двумя 256-битными значениями, в результате чего длина исходной подписи составляла 512 бит. Перед внедрением подпись кодировалась двоичным кодом БЧХ с длиной кодового слова 1023 бита, исправляющим до 57 битовых ошибок. После добавления одного бита дополнения размер внедряемого сообщения составлял 1024 бита.

Экспериментально исследовались четыре варианта предложенного метода, различающиеся используемым алгоритмом внедрения ЦВЗ. Для CPW (ChunkySeal) [45] вся 1024-битная последовательность внедрялась в изображение целиком. В вариантах CPW (PixelSeal) [37], CPW (VideoSeal) [35] и CPW (MBRS) [26] применялась квадрантная стратегия: сообщение разделялось на четыре фрагмента длиной 256 бит, которые независимо внедрялись в четыре четверти изображения.

Использование нескольких алгоритмов маркирования позволяет отдельно исследовать свойства общей схемы CPW и влияние конкретного механизма передачи аутентификатора. Перцептивное представление и процедура формирования цифровой подписи во всех вариантах остаются одинаковыми, тогда как устойчивость извлечения аутентификатора определяется в том числе характеристиками выбранного алгоритма ЦВЗ.

В качестве основного метода сравнения использовался MetaSeal [82], также основанный на контентно-зависимой цифровой подписи, внедряемой в изображение. Для методов PixelSeal, VideoSeal, ChunkySeal, MBRS и MetaSeal использовались опубликованные авторами открытые реализации и предобученные модели без дополнительного обучения.

Основная часть экспериментов проводилась на выборке из 1000 изображений набора данных MS-COCO [13]. Для исследования чувствительности к содержательным изменениям дополнительно использовался набор IMD2020 [135], содержащий пары исходных и модифицированных изображений и соответствующие маски измененных областей. После исключения примеров с некорректными или вырожденными масками экспериментальная выборка IMD2020 составила 926 пар изображений.

3.3.2 Исследование побитовой устойчивости EPHash

Первый эксперимент направлен на оценку способности EPHash сохранять точное значение хеш-кода после допустимых преобразований изображения. Данное свойство принципиально важно для CPW, поскольку в отличие от традиционных задач поиска визуально сходных изображений проверка цифровой подписи не допускает различий между исходным и повторно вычисленным хеш-кодами.

Для сравнения использовались четыре перцептивные хеш-функции, различающиеся способом формирования признакового представления изображения: aHash, основанная на сравнении яркости пикселей со средним значением; pHash, использующая низкочастотные коэффициенты дискретного косинусного преобразования; wHash, основанная на дискретном вейвлет-преобразовании; и PDQ, формирующая 256-битный хеш-код на основе частотного представления изображения [132; 136]. Для каждого из 1000 изображений MS-COCO вычислялся исходный хеш-код, после чего к изображению независимо применялись преобразования различной интенсивности. Исследовались изменение яркости и контрастности, JPEG-сжатие, размытие по Гауссу, масштабирование и добавление аддитивного гауссовского шума.

В качестве показателя устойчивости использовалась доля изображений, для которых хеш-код после преобразования полностью совпал с эталонным значением, полученным до воздействия. В отличие от традиционного сравнения перцептивных хеш-кодов с использованием расстояния Хэмминга, в данном эксперименте требовалось совпадение всех битов, поскольку изменение даже одного бита подписываемого представления приводит к отрицательному результату проверки цифровой подписи. Результаты эксперимента представлены на рисунке 3.2.

Полученные результаты показывают, что устойчивость классических перцептивных хеш-функций существенно зависит от вида воздействия. Алгоритмы aHash, pHash и wHash сохраняют достаточно высокую устойчивость к части исследуемых преобразований, однако с увеличением интенсивности воздействия вероятность полного побитового совпадения уменьшается. Наименьшую устойчивость в большинстве рассматриваемых сценариев демонстрирует PDQ.

Разработанная функция EPHash показывает более высокую побитовую устойчивость к масштабированию, размытию и JPEG-сжатию. В исследован-

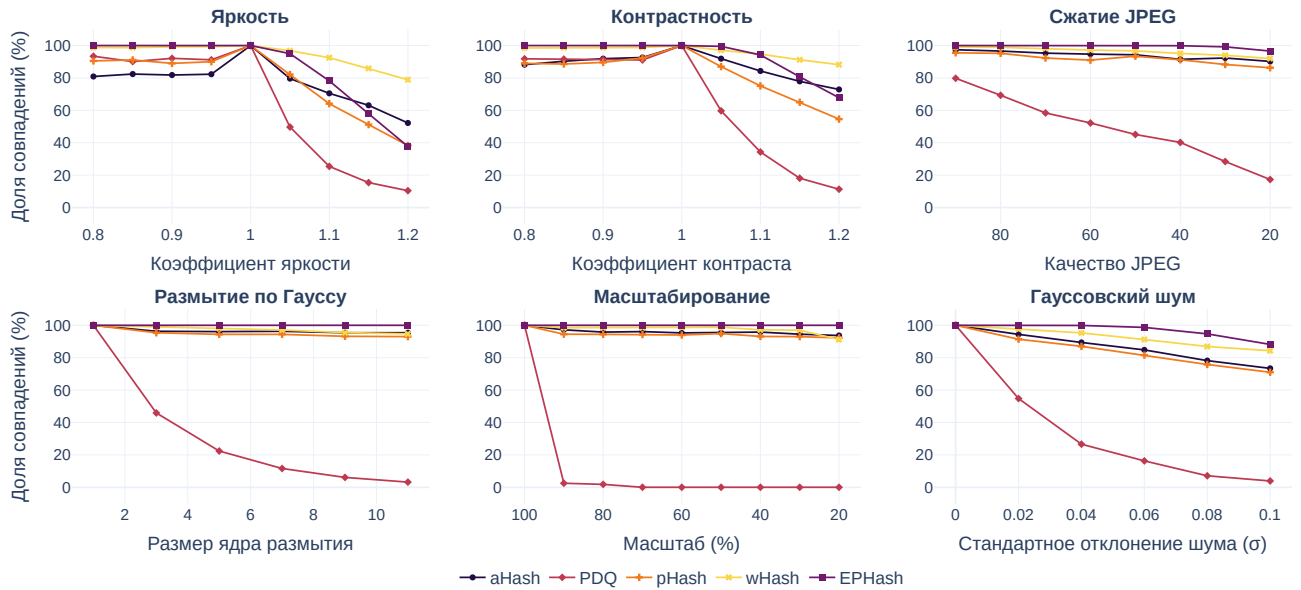


Рисунок 3.2 — Зависимость доли полностью совпадающих перцептивных хеш-кодов от интенсивности допустимых преобразований

ных диапазонах данных преобразований значение хеш-кода практически во всех случаях сохраняется полностью. При добавлении гауссовского шума наиболее высокие результаты также демонстрируют EPHash и wHash.

Наиболее сложными для всех исследуемых функций оказываются фотометрические преобразования. При значительном изменении яркости и контрастности доля полностью совпадающих хеш-кодов снижается и для EPHash. Таким образом, предлагаемый механизм формирования отступа от порога бинаризации существенно повышает вероятность точного воспроизведения хеш-кода, но не обеспечивает его безусловную инвариантность относительно произвольной интенсивности преобразований.

Полученные результаты подтверждают целесообразность контролируемой модификации коэффициентов ДКП, предложенной в разделе 3.2.1. В отличие от традиционного порогового сравнения хеш-кодов, повышение расстояния коэффициентов от границы квантования позволяет непосредственно увеличить вероятность сохранения каждого бита, что необходимо для последующей проверки цифровой подписи.

3.3.3 Оценка незаметности маркированных изображений

Второй эксперимент предназначен для оценки изменений, вносимых полной процедурой формирования маркированного изображения. Данная процедура включает как предварительную модификацию изображения при вычислении EPHash, так и последующее внедрение закодированной цифровой подписи посредством ЦВЗ.

Эксперимент проводился на 1000 изображениях MS-COCO. Для каждого исходного изображения формировались варианты CPW (MBRS), CPW (VideoSeal), CPW (PixelSeal) и CPW (ChunkySeal), а также маркированное изображение методом MetaSeal. В качестве показателей незаметности использовались метрики PSNR, SSIM и LPIPS, рассмотренные в разделе 2.3.1. Полученные значения приведены в таблице 7.

Таблица 7 — Сравнение незаметности различных вариантов CPW и метода MetaSeal

Метод	PSNR, дБ ↑	SSIM ↑	LPIPS ↓
CPW (MBRS)	40,7	0,982	0,011
CPW (VideoSeal)	41,1	0,985	0,009
CPW (PixelSeal)	41,2	0,985	0,009
CPW (ChunkySeal)	41,4	0,986	0,011
MetaSeal	35,1	0,929	0,098

Все исследованные варианты CPW демонстрируют высокое визуальное качество маркированных изображений: среднее значение PSNR превышает 40 дБ, значения SSIM находятся в диапазоне от 0,982 до 0,986, а LPIPS не превышает 0,011. Небольшие различия между вариантами CPW обусловлены прежде всего характеристиками используемого алгоритма внедрения ЦВЗ.

Метод MetaSeal демонстрирует более выраженные изменения изображения: значение PSNR составляет 35,1 дБ, SSIM — 0,929, а LPIPS — 0,098. Таким образом, в рассматриваемых экспериментальных условиях все варианты CPW обеспечивают более высокую незаметность формирования маркированного изображения по сравнению с MetaSeal.

Полученный результат показывает, что необходимость передачи 512-битной цифровой подписи и дополнительного помехоустойчивого кодирования не приводит к существенному ухудшению визуального качества маркированных изображений. Использование современных алгоритмов внедрения ЦВЗ с высокой

емкостью позволяет передавать 1024-битный аутентификатор при сохранении высокого визуального качества изображения.

3.3.4 Устойчивость к допустимым преобразованиям

Побитовая устойчивость EPHash является необходимым, но недостаточным условием успешной работы полной системы. Для положительного результата верификации необходимо одновременно сохранить значение перцептивного хеш-кода и восстановить цифровую подпись из ЦВЗ с количеством ошибок, не превышающим корректирующую способность кода БЧХ. Поэтому в следующем эксперименте исследуется устойчивость полной процедуры CPW.

К маркированным изображениям независимо применялись изменение яркости и контрастности, JPEG-сжатие, размытие по Гауссу, масштабирование и добавление гауссовского шума. После каждого воздействия выполнялась полная процедура, описанная в разделе 3.2.4: извлечение сообщения, БЧХ-декодирование, вычисление EPHash и проверка цифровой подписи. В качестве итогового показателя использовалась доля изображений, успешно прошедших проверку. Результаты для четырех вариантов CPW и MetaSeal представлены на рисунке 3.3.

Эксперимент показывает, что устойчивость полной схемы CPW существенно зависит от выбранного алгоритма внедрения аутентификатора. Это отличает данный эксперимент от исследования EPHash: одинаковая перцептивная хеш-функция используется во всех вариантах метода, однако успешность извлечения цифровой подписи определяется характеристиками конкретного ЦВЗ.

Вариант CPW (MBRS) демонстрирует наиболее высокую устойчивость к JPEG-сжатию. Успешная верификация сохраняется при снижении коэффициента качества JPEG до 40. Вместе с тем данный вариант оказывается менее устойчивым к размытию изображения.

Противоположный профиль наблюдается для CPW (ChunkySeal): данный вариант лучше переносит размытие и уменьшение разрешения, однако оказывается неустойчивым к JPEG-сжатию. Следовательно, высокая емкость и возможность непосредственного внедрения всей 1024-битной последовательности сами

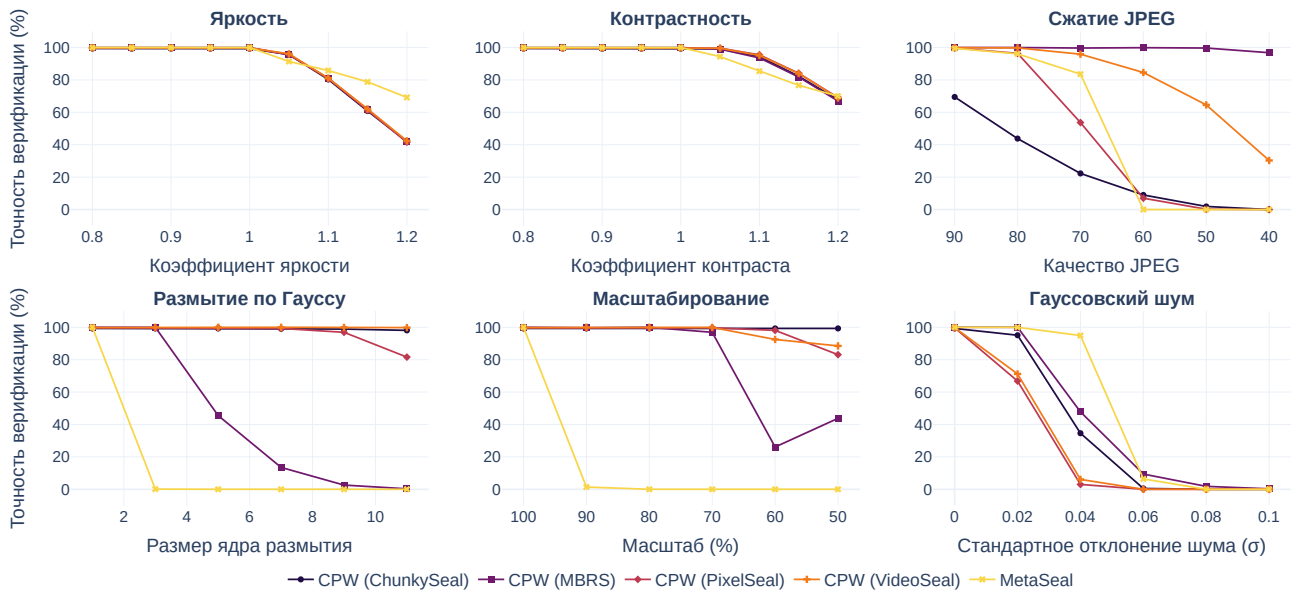


Рисунок 3.3 — Зависимость доли успешно верифицированных изображений от интенсивности допустимых преобразований

по себе не обеспечивают устойчивость к каждому из рассматриваемых преобразований.

Наиболее сбалансированный профиль в проведенных экспериментах демонстрирует CPW (VideoSeal), сохраняющий высокую долю положительных результатов верификации при различных видах допустимой обработки. Следовательно, возможность замены алгоритма внедрения ЦВЗ в составе CPW позволяет адаптировать систему к предполагаемому профилю преобразований.

MetaSeal демонстрирует несколько более высокую устойчивость к увеличению яркости и добавлению гауссовского шума, однако значительно хуже переносит размытие и уменьшение разрешения. Таким образом, ни один из исследованных вариантов не обеспечивает одинаковой устойчивости ко всем преобразованиям, а свойства полной системы определяются совместно устойчивостью EPHash и используемого механизма внедрения аутентификатора.

Полученные результаты также показывают важное различие между устойчивостью перцептивного представления и устойчивостью полной системы. Даже если EPHash сохраняет исходное значение после преобразования, повреждение большего числа битов ЦВЗ, чем способен исправить код БЧХ, приводит к отрицательному результату проверки. Поэтому область работоспособности CPW определяется пересечением областей устойчивости двух основных компонентов: перцептивного хеширования и цифрового маркирования.

3.3.5 Обнаружение семантических модификаций

Последний эксперимент направлен на оценку способности метода выявлять содержательно значимые изменения изображения. Для этого использовался набор данных IMD2020, содержащий реальные примеры манипуляций, в том числе вставку, удаление и клонирование фрагментов изображений.

Для каждого примера сначала формировалась маркированная версия исходного изображения. Затем область, соответствующая маске модификации IMD2020, заменялась пикселями из поддельного изображения. Такой порядок обработки позволяет моделировать ситуацию, при которой семантическое изменение вносится уже после формирования аутентификатора и внедрения ЦВЗ. Полученное изображение передавалось на стандартную процедуру верификации CPW.

Для каждого изображения определялась относительная площадь модифицированной области

$$\rho = \frac{N_M}{N_I}, \quad (3.24)$$

где N_M — число пикселей, принадлежащих области модификации, а N_I — общее число пикселей изображения.

Для анализа зависимости чувствительности метода от размера вмешательства 926 изображений были отсортированы по значению ρ и разделены на шесть групп по перцентилям площади измененной области. Для каждой группы вычислялась доля изображений, для которых процедура верификации вернула отрицательный результат. Результаты представлены на рис. 3.4.

Полученные результаты показывают выраженную зависимость чувствительности от масштаба изменения изображения. Для всех вариантов CPW доля обнаруженных нарушений целостности приближается к единице уже для групп, в которых средняя площадь измененной области составляет приблизительно 1,5 % изображения. При меньших областях доля обнаружения снижается, что связано с ограниченной чувствительностью используемого низкочастотного перцептивного представления к локальным изменениям.

Кривые для CPW (ChunkySeal), CPW (MBRS), CPW (PixelSeal) и CPW (VideoSeal) практически совпадают. Это подтверждает, что чувствительность к семантической модификации в данном эксперименте определяется

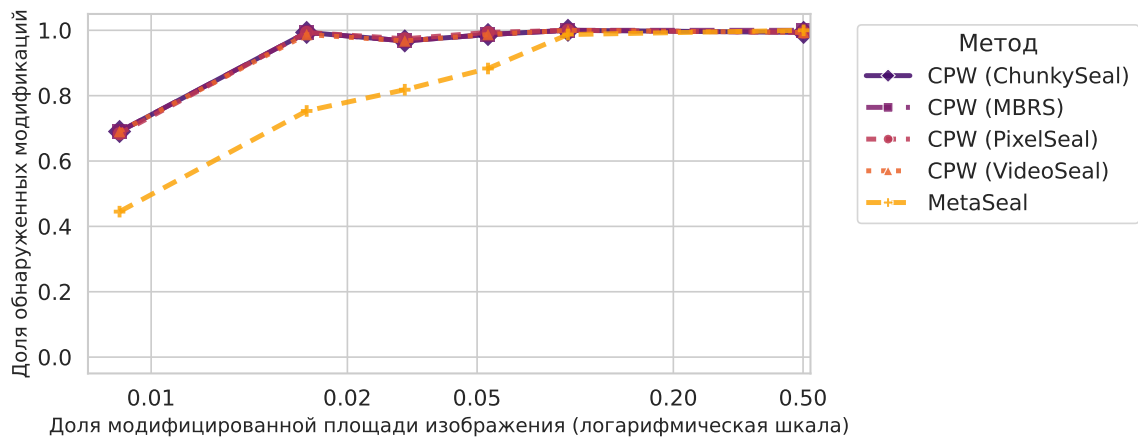


Рисунок 3.4 — Зависимость доли выявленных нарушений целостности от площади модифицированной области изображения на наборе данных IMD2020

прежде всего используемой перцептивной хеш-функцией, а не конкретным алгоритмом внедрения ЦВЗ. Локальная замена небольшой области не приводит к существенному повреждению внедренного аутентификатора, однако при достаточном изменении низкочастотных характеристик изображения значение EPHash изменяется и цифровая подпись перестает проходить проверку.

Для MetaSeal сопоставимая доля обнаружения достигается только при значительно большей площади вмешательства — порядка 10 % изображения. В рассматриваемых экспериментальных условиях варианты CPW демонстрируют более высокую долю отрицательных результатов верификации после локальных модификаций по сравнению с MetaSeal.

При этом значение 1,5 % не следует интерпретировать как строгий порог обнаружения. Результат зависит не только от площади, но и от положения, контраста и характера модифицированной области. Изменение небольшого или слабоконтрастного фрагмента может не затронуть используемые EPHash низкочастотные признаки и остаться необнаруженным. Предложенный метод поэтому обеспечивает контроль содержательных изменений, но не предназначен для попиксельного контроля или локализации области модификации.

В целом проведенные эксперименты подтверждают работоспособность предложенного метода в рамках постановки задачи, сформулированной в разделе 3.1. Разработанная функция EPHash обеспечивает повышенную побитовую устойчивость к рассматриваемым допустимым преобразованиям, а использование современных алгоритмов внедрения ЦВЗ позволяет передавать аутентификатор длиной 1024 бита при сохранении высокого визуального ка-

чества изображения. Устойчивость полной процедуры верификации зависит от выбранного алгоритма маркирования, тогда как чувствительность к семантическим изменениям в основном определяется свойствами EPHash. Выявленные ограничения и границы применимости предложенного метода рассматриваются в следующем разделе.

3.4 Ограничения метода

Предложенный метод имеет ряд ограничений, связанных с распределением открытых ключей, свойствами перцептивной хеш-функции EPHash, устойчивостью используемого алгоритма внедрения ЦВЗ и областью проведенной экспериментальной оценки.

CPW предполагает наличие доверенного открытого ключа стороны, выполнившей аутентификацию изображения. Цифровая подпись подтверждает соответствие аутентификатора заданному открытому ключу, но не решает задачу установления принадлежности этого ключа конкретному владельцу. Реализация такого механизма требует использования внешней инфраструктуры управления ключами.

Второе ограничение связано со свойствами перцептивной хеш-функции EPHash. Стойкость ECDSA препятствует формированию корректной подписи для нового значения хеш-кода без знания закрытого ключа, однако не исключает возможность поиска семантически измененного изображения с тем же значением EPHash. Такой сценарий соответствует задаче поиска второго прообраза. В настоящей работе устойчивость EPHash к целенаправленным состязательным воздействиям формально не исследуется.

Следующее ограничение определяется устойчивостью EPHash к допустимым преобразованиям. Формируемый при маркировании отступ от порогов бинаризации повышает вероятность сохранения EPHash, однако не обеспечивает абсолютную инвариантность относительно произвольных преобразований. Работоспособность метода подтверждена только для рассмотренных типов воздействий и исследованных диапазонов параметров. Геометрические преобразования, такие как поворот, кадрирование, зеркальное отражение и изменение соотношения сторон, в текущей реализации не поддерживаются, поскольку на-

рушают пространственное соответствие, используемое при вычислении EPHash и извлечении аутентификатора.

Отдельным фактором является устойчивость используемого алгоритма маркирования изображений. Для успешной верификации необходимо не только сохранение EPHash, но и восстановление цифровой подписи с числом ошибок, не превышающим корректирующую способность кода БЧХ. Поэтому отрицательный результат верификации не позволяет однозначно определить причину отказа: ей может быть семантическая модификация, изменение EPHash вследствие неподдерживаемого преобразования, геометрическая рассинхронизация, повреждение ЦВЗ или превышение корректирующей способности кода. Предлагаемый метод решает задачу проверки соответствия изображения аутентифицированному содержимому, но не выполняет классификацию причины нарушения целостности.

Кроме того, CPW формирует итоговый бинарный результат проверки и не локализует область изменения. Если значение EPHash изменилось, можно сделать вывод о невозможности подтвердить соответствие исследуемого изображения ранее аутентифицированному содержимому, однако определить по результату верификации координаты измененных пикселей невозможно. Следовательно, предложенный метод предназначен для контроля содержательной целостности изображения в пределах заданной политики преобразований, а не для попиксельного контроля и локализации модифицированных областей.

3.5 Выводы по главе

В данной главе разработан и экспериментально исследован метод контроля целостности изображений CPW, основанный на подписываемом перцептивном представлении изображения. Перцептивный хеш-код связывает результат верификации с содержимым изображения, электронная цифровая подпись — с заданным открытым ключом, а цифровой водяной знак используется для передачи сформированного аутентификатора совместно с изображением. Проверка выполняется без доступа к исходному немаркированному изображению и закрытому ключу.

Для формирования подписываемого представления разработана перцептивная хеш-функция EPHash. Предложен способ повышения ее побитовой устойчивости, основанный на контролируемой модификации низкочастотных коэффициентов ДКП относительно порогов бинаризации без изменения исходного значения перцептивного хеш-кода. Такой подход повышает вероятность точного воспроизведения подписываемой битовой последовательности после допустимых преобразований, что необходимо для последующей проверки цифровой подписи.

Для передачи подписываемого представления сформирован контентно-зависимый аутентификатор на основе схемы цифровой подписи ECDSA. Цифровая подпись защищается помехоустойчивым кодом БЧХ и внедряется в изображение в виде водяного знака. Рассмотрены стратегии непосредственного внедрения всего сообщения и его разделения между отдельными областями изображения. Это позволяет использовать в составе CPW различные алгоритмы маркирования изображений при условии достаточной емкости и устойчивости к допустимым преобразованиям.

Экспериментальное исследование показало, что EPHash обеспечивает более высокую побитовую устойчивость к масштабированию, размытию и JPEG-сжатию по сравнению с рассмотренными перцептивными хеш-функциями aHash, rHash, wHash и PDQ. Для исследованных вариантов CPW среднее значение PSNR маркированных изображений превышает 40 дБ. Исследование полной процедуры верификации показало, что устойчивость к допустимым преобразованиям определяется совместно устойчивостью EPHash и используемого алгоритма внедрения ЦВЗ, вследствие чего различные варианты CPW имеют различающиеся уровни устойчивости к разным преобразованиям.

Эксперименты на наборе данных IMD2020 подтвердили способность разработанного метода выявлять рассматриваемые недопустимые модификации изображений. Чувствительность к таким изменениям определяется прежде всего свойствами используемого перцептивного представления и зависит от площади, положения и характера модифицированной области. Для малых или малококонтрастных изменений сохранение исходного значения EPHash возможно, что ограничивает область применимости метода.

Таким образом, разработанный метод контроля целостности изображений на основе подписываемого перцептивного представления обеспечивает в исследованных условиях устойчивость к рассматриваемым допустимым преоб-

разованиям и выявление недопустимых модификаций. Полученные результаты подтверждают возможность совместного применения перцептивного хеширования, электронной цифровой подписи и цифровых водяных знаков для публичной проверки соответствия изображения ранее аутентифицированному содержимому в пределах заданной политики преобразований.

Глава 4. Метод контроля целостности аудиоданных на основе цифровых водяных знаков и цифровой подписи

В предыдущей главе был рассмотрен подход к контролю целостности изображений, основанный на совместном применении перцептивного хеширования, электронной цифровой подписи и цифровых водяных знаков. Такая схема позволяет связать встроенный аутентификатор с содержимым защищаемого объекта и заданным открытым ключом, сохраняя возможность положительной верификации после допустимых преобразований.

Применение данного подхода к аудиоданным требует учета их временной структуры. Отдельные участки аудиозаписи могут независимо подвергаться модификации, поэтому единого результата проверки для всего объекта недостаточно. Наряду с выявлением нарушения целостности требуется определять временное положение участков, для которых целостность не может быть подтверждена.

Для решения этой задачи разработан метод CryptoMark, основанный на независимой аутентификации последовательных фрагментов аудиосигнала. Метод объединяет устойчивое перцептивное хеширование, электронную цифровую подпись, помехоустойчивое кодирование и внедрение аутентификатора в аудиосигнал в виде ЦВЗ. Пофрагментная организация позволяет сформировать карту целостности аудиозаписи и использовать ее для временной локализации модификаций.

4.1 Постановка задачи контроля целостности аудиоданных

В настоящей работе ставится задача разработки метода контроля целостности аудиоданных, позволяющего публично проверять соответствие исследуемой аудиозаписи ранее аутентифицированному содержимому с точностью до заданного множества допустимых преобразований. В отличие от изображений, аудиоданные имеют выраженную временную структуру, поэтому наряду с выявлением нарушения целостности требуется определять временное положение участков, для которых целостность не может быть подтверждена. Проверка долж-

на выполняться без доступа к исходному немаркированному аудиосигналу и закрытому ключу: для верификации используются исследуемый аудиосигнал, открытый ключ и общедоступные параметры метода.

Пусть A — исходный цифровой аудиосигнал, A_w — сформированный на его основе маркированный сигнал, а K_{pub} — открытый ключ стороны, выполнившей аутентификацию. Для исследуемого сигнала A' процедура верификации должна формировать набор локальных результатов

$$\mathcal{V}(A', K_{\text{pub}}) = (v_1, v_2, \dots, v_N), \quad v_i \in \{0, 1\}, \quad (4.1)$$

где каждому значению v_i соответствует некоторый временной интервал Δ_i исследуемой аудиозаписи, а N — количество временных интервалов. Значение $v_i = 1$ означает подтверждение целостности соответствующего участка, а $v_i = 0$ — невозможность подтвердить его соответствие ранее аутентифицированному содержимому. Последовательность локальных результатов образует карту целостности аудиозаписи и позволяет определять временное положение возможных модификаций. Точность определения их границ зависит от временного разрешения используемого метода.

Модель нарушителя принимается аналогичной сформулированной в главе 3. Алгоритмы и параметры метода, а также открытый ключ считаются известными нарушителю, тогда как закрытый ключ ему недоступен. Спецификой аудиоданных является возможность изменения отдельного временного участка при сохранении остальных частей записи неизменными. Целью нарушителя является получение содержательно измененной аудиозаписи, для которой процедура контроля целостности не выявляет соответствующую модификацию.

Область работоспособности метода определяется принятой политикой допустимых и недопустимых преобразований. К допустимым относятся воздействия, характерные для хранения, передачи и воспроизведения аудиоконтента и не изменяющие его основное смысловое содержание. В рассматриваемой постановке к ним относятся:

- сжатие аудиосигнала с потерями;
- передискретизация;
- переквантование;
- умеренное добавление шума;
- частотная фильтрация;
- масштабирование амплитуды сигнала.

Для допустимых преобразований целевым требованием является сохранение положительного результата проверки каждого временного фрагмента. Если T_A — множество таких преобразований, данное требование можно записать в виде

$$\mathcal{V}(T(A_w), K_{\text{pub}}) = (1, 1, \dots, 1), \quad T \in T_A. \quad (4.2)$$

Степень выполнения данного требования определяется экспериментально для заданных видов преобразований и диапазонов их параметров на основании доли успешно верифицированных фрагментов и аудиозаписей.

К недопустимым относятся преобразования, приводящие к содержательно значимому изменению аудиозаписи. В рассматриваемой экспериментальной постановке к ним относятся преобразование голоса, замена отдельных временных участков синтетическим содержимым и реконструкция сигнала с использованием генеративных моделей. При локальном применении преобразования $T \in T_N$ отрицательные результаты проверки должны формироваться для временных интервалов, затронутых модификацией, тогда как для неизменной части аудиозаписи положительные результаты должны сохраняться с точностью до временного разрешения метода.

Отдельную группу составляют преобразования, изменяющие временную структуру аудиосигнала: удаление или вставка отсчетов, изменение скорости воспроизведения и временное масштабирование. Такие воздействия не обязательно являются содержательно значимыми модификациями, однако могут нарушать временное соответствие между состояниями аудиозаписи до и после преобразования. Устойчивость к таким воздействиям не входит в заявленную область работоспособности рассматриваемого метода.

Таким образом, разрабатываемый метод должен сохранять положительный результат проверки после допустимых преобразований, выявлять содержательно значимые модификации и определять временные участки, для которых целостность не может быть подтверждена. При этом проверка должна выполняться без доступа к исходному немаркированному аудиосигналу и закрытому ключу, а формирование маркированного сигнала не должно приводить к существенному ухудшению воспринимаемого качества аудиоданных.

4.2 Метод контроля целостности аудиоданных CryptoMark

Для выполнения сформулированных условий разработан метод CryptoMark, основанный на независимой проверке последовательных временных фрагментов аудиосигнала. Пофрагментная организация позволяет связать результат верификации с определенным временным интервалом и тем самым сформировать карту целостности аудиозаписи. Общая схема формирования и проверки маркированного аудиосигнала методом CryptoMark представлена на рисунке 4.1.

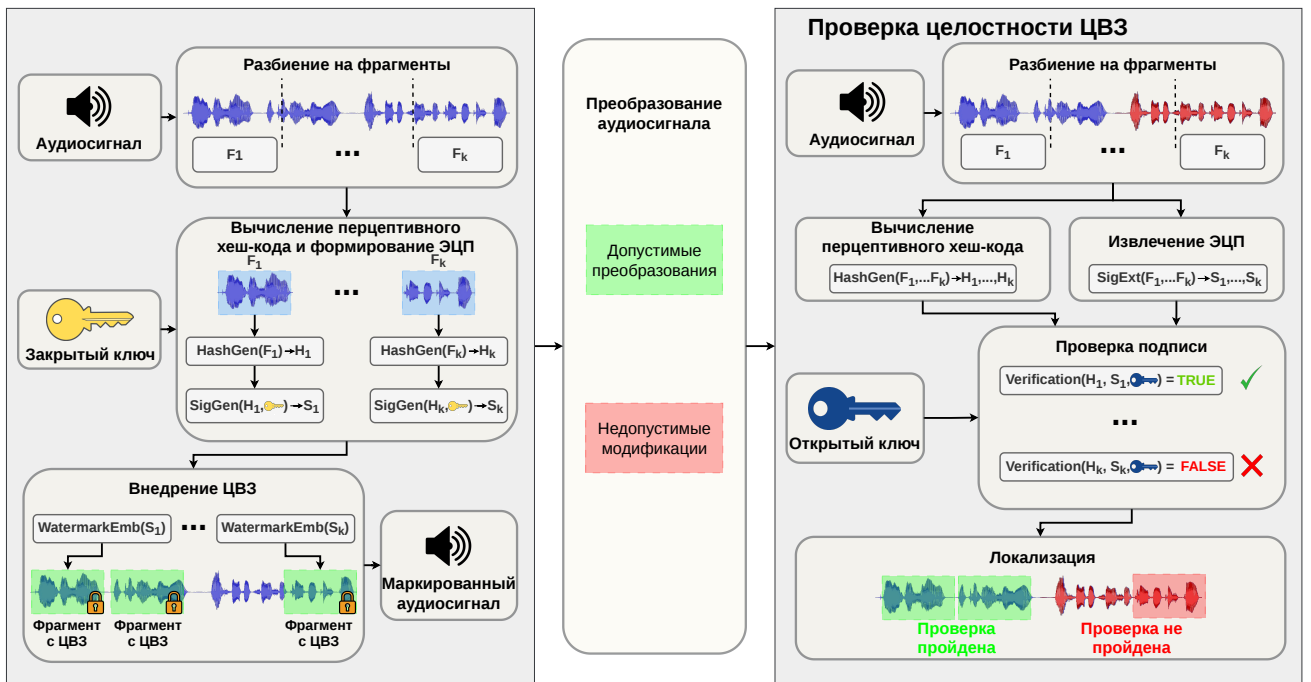


Рисунок 4.1 — Схема метода CryptoMark для контроля целостности аудиоданных

Пусть A — исходный цифровой аудиосигнал длиной L отсчетов. При формировании маркированного сигнала он разбивается на последовательность неперекрывающихся фрагментов фиксированной длины L_f :

$$A = [F_1, F_2, \dots, F_N, F_r], \quad N = \left\lfloor \frac{L}{L_f} \right\rfloor, \quad (4.3)$$

где $F_1, \dots, F_N$ — полные фрагменты длиной L_f , а F_r — остаточная часть сигнала, если его длина не кратна L_f .

Для каждого полного фрагмента независимо формируется контентно-зависимый аутентификатор, который внедряется в тот же фрагмент в виде цифрового водяного знака. При последующей верификации каждый фрагмент

проверяется независимо, а совокупность локальных результатов образует карту целостности аудиозаписи.

Каждому фрагменту F_i соответствует временной интервал

$$\Delta_i = \left[\frac{(i-1)L_f}{f_s}, \frac{iL_f}{f_s} \right), \quad (4.4)$$

где f_s — частота дискретизации аудиосигнала. Временное разрешение локализации определяется длительностью одного фрагмента:

$$\Delta t = \frac{L_f}{f_s}. \quad (4.5)$$

Уменьшение L_f повышает временное разрешение локализации, но одновременно сокращает объем сигнала, доступный для внедрения аутентификатора. Длина фрагмента определяет компромисс между точностью локализации и устойчивостью передачи встроенной информации.

Для каждого фрагмента CryptoMark формирует устойчивое контентно-зависимое представление посредством перцептивного хеширования. Полученный хеш-код подписывается закрытым ключом, цифровая подпись защищается помехоустойчивым кодом и внедряется в тот же фрагмент в виде ЦВЗ. При проверке выполняются обратные операции: из исследуемого фрагмента извлекается подпись, повторно вычисляется перцептивный хеш-код и выполняется проверка подписи с использованием открытого ключа.

Таким образом, реализация сформулированных в разделе 4.1 условий требует устойчивости подписываемого представления к допустимым преобразованиям, его чувствительности к содержательным модификациям и точного восстановления цифровой подписи из маркированного аудиосигнала.

4.2.1 Формирование устойчивого перцептивного хеш-кода

Перцептивный хеш-код, применяемый в CryptoMark, основан на тех же принципах, что и перцептивная хеш-функция EPHash для изображений 3.2.1, используемая в методе CPW. Для фрагмента F_i длиной L_f вычисляется одномерное дискретное косинусное преобразование

$$X_i(k) = \sum_{n=0}^{L_f-1} F_i(n) \cos \left[\frac{\pi}{L_f} \left(n + \frac{1}{2} \right) k \right], \quad k = 0, \dots, L_f - 1. \quad (4.6)$$

Из полученного спектрального представления выбирается последовательность из τ_l коэффициентов, начиная с позиции τ_{off} . Пусть выбранные коэффициенты обозначены $X_i^*(j)$, $j = 1, \dots, \tau_l$. Для них вычисляется медианное значение m_i , используемое в качестве порога бинаризации:

$$h_i(j) = \begin{cases} 1, & X_i^*(j) \geq m_i, \\ 0, & X_i^*(j) < m_i, \end{cases} \quad j = 1, \dots, \tau_l. \quad (4.7)$$

Полученные биты образуют перцептивный хеш-код фрагмента аудиозаписи

$$H_i = (h_i(1), h_i(2), \dots, h_i(\tau_l)). \quad (4.8)$$

Аналогично EPHash, наименьшей устойчивостью обладают коэффициенты, расположенные вблизи медианного порога. Небольшое изменение сигнала может привести к переходу коэффициента через порог бинаризации (4.7) и изменению соответствующего бита. Для снижения вероятности такого события при формировании маркированного сигнала выполняется итерационная коррекция спектральных коэффициентов.

На каждой итерации к текущему фрагменту применяется заданное моделируемое допустимое преобразование и повторно вычисляется перцептивный хеш-код. Коррекциям подлежат коэффициенты, для которых соответствующий бит изменился либо расстояние от медианного порога после преобразования оказалось меньше t . Направление изменения определяется исходным значением бита:

$$\sigma_j = \begin{cases} t, & h_i(j) = 1, \\ -t, & h_i(j) = 0. \end{cases} \quad (4.9)$$

Выбранные коэффициенты смещаются на величину σ_j , после чего устойчивость хеш-кода проверяется повторно. Процесс продолжается до достижения устойчивости либо установленного максимального числа итераций. В результате формируется модифицированный фрагмент F_i^* , для которого сохраняется исходное значение хеш-кода:

$$H(F_i^*) = H(F_i) = H_i. \quad (4.10)$$

Таким образом, процедура увеличивает устойчивость подписываемой битовой последовательности к допустимым преобразованиям, не изменяя ее исходного значения. На этапе проверки стабилизация повторно не выполняется.

4.2.2 Формирование цифровой подписи и помехоустойчивое кодирование

Для привязки аутентификатора к содержимому фрагмента перцептивный хеш-код H_i сначала преобразуется криптографической хеш-функцией SHA-256, после чего формируется цифровая подпись по алгоритму ECDSA [133]:

$$S_i = \text{ECDSA. Sign} (K_{\text{priv}}, \text{SHA256}(H_i)) , \quad (4.11)$$

где K_{priv} — закрытый ключ. В CryptoMark используется эллиптическая кривая `secp256k1`. Цифровая подпись представляется битовой последовательностью длины 512 бит.

Таким образом, подпись S_i связана с перцептивным представлением конкретного фрагмента. Если модификация аудиоданных приводит к изменению H_i , сформированная подпись перестанет проходить проверку относительно нового содержимого. Формирование корректной подписи для нового значения H_i без знания закрытого ключа определяется стойкостью используемой схемы ECDSA.

Проверка цифровой подписи требует побитово точного восстановления S_i . Для исправления ошибок, возникающих при извлечении ЦВЗ после преобразований аудиосигнала, используется код Рида—Соломона [137]. Для подписи S_i формируется последовательность проверочных символов P_i :

$$P_i = \text{RS. Parity}(S_i). \quad (4.12)$$

Помехоустойчивое кодирование не изменяет содержимое аутентификатора, а позволяет исправлять часть ошибок, возникающих при передаче подписи через канал ЦВЗ.

4.2.3 Внедрение цифровой подписи в аудиосигнал

Для внедрения цифровой подписи используется модифицированный вариант метода FSVC (Frequency Singular Value Coefficient) [60]. Модифицированный фрагмент F_i^* разбивается на последовательность сегментов

$$F_i^* = [G_{i,1}, G_{i,2}, \dots, G_{i,M}], \quad (4.13)$$

где каждый сегмент $G_{i,j}$ далее разделяется на два блока $G_{i,j}^{(1)}$ и $G_{i,j}^{(2)}$. Для обоих блоков вычисляется ДКП. Из коэффициентов в выбранной спектральной области формируются представления, для которых определяются сингулярные значения $\lambda_{j,1}$ и $\lambda_{j,2}$.

Для кодирования бита используется отношение

$$R_j = \frac{\lambda_{j,1}}{\lambda_{j,2}}. \quad (4.14)$$

При внедрении бита w_j сингулярные значения изменяются таким образом, чтобы выполнялось условие

$$R_j \begin{cases} \leq \frac{1}{1 + \alpha}, & w_j = 0, \\ \geq 1 + \alpha, & w_j = 1, \end{cases} \quad (4.15)$$

где $\alpha > 0$ — параметр интенсивности внедрения.

В одном сегменте независимо передаются бит цифровой подписи и соответствующая ему проверочная информация помехоустойчивого кода. Для этого выбранная спектральная полоса разделяется на две части: первая используется для внедрения битов S_i , вторая — для битов P_i . После изменения сингулярных значений выполняются обратные преобразования и восстанавливается временное представление сегмента.

Динамический выбор спектральной области. Использование одной фиксированной спектральной области для всех фрагментов может приводить к заметным искажениям на участках сигнала, где соответствующие частотные компоненты имеют малую энергию. В CryptoMark спектральная область внедрения выбирается независимо для каждого маркируемого фрагмента с учетом распределения энергии исходного сигнала.

В заданном допустимом диапазоне частот формируется множество кандидатных полос

$$\mathcal{B} = \{B_1, B_2, \dots, B_K\}. \quad (4.16)$$

Для выбора полосы по всем сегментам рассматриваемого фрагмента оценивается энергия соответствующих коэффициентов ДКП. В качестве опорного уровня используется медианное значение энергии. Для каждой кандидатной полосы определяется доля коэффициентов с малой энергией относительно этого уровня. Предпочтение отдается полосе с наименьшей долей таких коэффициентов одновременно в области внедрения подписи и проверочной информации.

После выбора спектральной области интенсивность внедрения может дополнительно адаптироваться на уровне отдельных сегментов. При недостаточной энергии используемых спектральных компонентов значение α уменьшается, что позволяет ограничить вносимые искажения.

После обработки всех сегментов формируется маркированный фрагмент $F_{i,w}$. Последовательность маркированных фрагментов объединяется в итоговый аудиосигнал A_w .

4.2.4 Верификация целостности и локализация модификаций

Проверка CryptoMark выполняется без доступа к исходному немаркированному аудиосигналу и закрытому ключу. Исходными данными являются исследуемый аудиосигнал A' , открытый ключ K_{pub} и общедоступные параметры метода.

Исследуемый сигнал разбивается на фрагменты в соответствии с той же схемой, которая использовалась при формировании маркированного сигнала. Для каждого фрагмента F'_i независимо вычисляется перцептивный хеш-код

$$H'_i = H(F'_i) \quad (4.17)$$

и соответствующее значение криптографической хеш-функции.

Поскольку использованная при внедрении спектральная полоса отдельно не передается, декодер последовательно рассматривает все кандидатные полосы

$B_k \in \mathcal{B}$. Для каждой из них из спектральных компонентов извлекаются последовательности $S'_{i,k}$ и $P'_{i,k}$, соответствующие цифровой подписи и проверочным символам помехоустойчивого кода.

Для каждой кандидатной полосы выполняется декодирование кода Рида—Соломона. Если исправление ошибок возможно, восстанавливается кандидат на исходную подпись и определяется число исправленных ошибок. Из успешно декодированных вариантов выбирается кандидат, потребовавший наименьшего числа исправлений:

$$k_i^* = \arg \min_k e_{i,k}, \quad (4.18)$$

где $e_{i,k}$ — число ошибок, исправленных при декодировании кандидата из полосы B_k . Если ни для одной кандидатной полосы корректное декодирование выполнить не удастся, для соответствующего фрагмента формируется отрицательный результат проверки.

После помехоустойчивого декодирования восстанавливается 512-битная цифровая подпись S'_i . На заключительном этапе подпись проверяется относительно повторно вычисленного значения H'_i с использованием открытого ключа:

$$v_i = \text{ECDSA. Verify} (K_{\text{pub}}, \text{SHA256}(H'_i), S'_i), \quad v_i \in \{0,1\}. \quad (4.19)$$

Если подпись успешно проверяется, целостность соответствующего фрагмента считается подтвержденной в пределах принятой политики допустимых преобразований. Если подпись не удастся корректно восстановить либо она не соответствует повторно вычисленному перцептивному представлению, формируется отрицательный результат $v_i = 0$.

После проверки всех фрагментов значения $(v_1, v_2, \dots, v_N)$ образуют карту целостности аудиозаписи. Положение фрагментов с отрицательным результатом определяет временные участки, целостность которых не может быть подтверждена.

4.3 Экспериментальная оценка метода CryptoMark

Экспериментальная оценка CryptoMark проводилась с использованием разработанного программного комплекса WIBE, рассмотренного в главе 2. Она включает исследование незаметности маркирования, устойчивости контроля целостности к допустимым преобразованиям, способности выявлять содержательные модификации и точности их временной локализации.

4.3.1 Параметры реализации

Во всех экспериментах использовалась одна фиксированная конфигурация CryptoMark без дополнительной настройки под отдельные преобразования или типы модификаций. Параметры выбирались с учетом компромисса между устойчивостью, незаметностью маркирования, емкостью ЦВЗ и временным разрешением локализации.

Аудиосигналы обрабатывались с частотой дискретизации $f_s = 22050$ Гц. Длина независимо проверяемого фрагмента составляла $L_f = 98304$ отсчета, что соответствует приблизительно 4,46 с. Каждый фрагмент разделялся на 512 сегментов длиной 192 отсчета, каждый из которых, в свою очередь, состоял из двух блоков по 96 отсчетов. Один сегмент использовался для внедрения одного бита цифровой подписи и соответствующего бита проверочной информации.

Для формирования перцептивного хеш-кода использовались $\tau_l = 16$ коэффициентов ДКП начиная с позиции $\tau_{\text{off}} = 1$. Параметр стабилизации составлял $t = 0,01$, а максимальное число итераций коррекции — 20. На каждой итерации устойчивость хеш-кода контролировалась после добавления шума с отношением сигнал—шум 30 дБ.

Для формирования цифровой подписи применялись SHA-256 и ECDSA на эллиптической кривой `secp256k1`. Длина цифровой подписи для каждого фрагмента составляла 512 бит.

Для помехоустойчивого кодирования использовался код Рида—Соломона над $GF(2^8)$. Подпись представлялась 64 информационными байтами, к которым добавлялись 64 проверочных байта. Таким образом, использовался сокращенный

код $RS(128,64)$, позволяющий исправлять до 32 ошибочных байтов. С учетом проверочной информации в каждый фрагмент внедрялось 1024 бита, из которых 512 бит составляла цифровая подпись.

В используемой конфигурации спектральная область внедрения выбиралась отдельно для каждого проверяемого фрагмента. Рассматривались четыре кандидатные полосы в диапазоне 2500—4500 Гц. Каждая полоса включала шесть коэффициентов ДКП: первые три использовались для внедрения бита цифровой подписи, а следующие три — для соответствующего бита проверочной информации кода Рида—Соломона.

Базовое значение параметра интенсивности внедрения составляло $\alpha = 2$. С целью сохранения незаметности, для сегментов с недостаточной энергией в выбранной спектральной области интенсивность уменьшалась до $\alpha = 1$.

4.3.2 Оценка незаметности маркирования

Для оценки влияния CryptoMark на качество аудиосигнала использовались речевые записи набора данных LibriSpeech [126]. Для каждой записи сравнивались исходный аудиосигнал A и соответствующий маркированный сигнал A_w .

CryptoMark сравнивался с методами AudioSeal [63] и SilentCipher [64]. Данные методы были выбраны как современные методы маркирования аудиоданных, обеспечивающие высокую устойчивость к распространенным преобразованиям; AudioSeal, кроме того, поддерживает локализацию внедренного водяного знака. При интерпретации результатов учитывалось существенное различие размера передаваемого сообщения сравниваемых методов: CryptoMark должен передавать для каждого временного фрагмента 512-битную цифровую подпись, тогда как размер сообщения рассматриваемых аналогов существенно меньше.

Для количественной оценки использовались метрики SI-SNR, PESQ и STOI, определенные в главе 2. SI-SNR характеризует величину изменений формы сигнала, PESQ — воспринимаемое качество речи, а STOI — сохранение ее разборчивости. Результаты приведены в таблице 8.

Полученные результаты показывают, что CryptoMark обеспечивает качество маркированного аудиосигнала, сопоставимое с AudioSeal, при существенно большем размере передаваемого сообщения. Значение STOI практически не из-

Таблица 8 — Оценка качества маркированных аудиоданных

Метод	Емкость	SI-SNR $\uparrow$	PESQ $\uparrow$	STOI $\uparrow$
AudioSeal	16 бит	25,47	4,38	0,99
CryptoMark	512 бит / фрагмент	25,56	4,41	0,99
SilentCipher	40 бит	34,98	4,57	0,99

меняется для всех рассмотренных методов, что свидетельствует о сохранении разборчивости речи. SilentCipher демонстрирует более высокие значения отдельных метрик качества, но емкость этого метода значительно ниже. Таким образом, передача 512-битной цифровой подписи в каждом фрагменте обеспечивает приемлемый компромисс между емкостью и незаметностью ЦВЗ.

4.3.3 Оценка контроля целостности

Для оценки устойчивости и способности выявлять содержательные модификации задача рассматривалась как бинарная классификация. Положительному классу соответствуют аудиозаписи, целостность которых должна подтверждаться, отрицательному — записи после недопустимых модификаций.

Эксперимент проводился на поднаборах LibriSpeech test-clean и test-other, включающих в совокупности 5559 исходных записей. Для получения достаточно продолжительных аудиосигналов записи объединялись в пределах исходных глав, в результате чего было сформировано 177 аудиозаписей.

На их основе формировались три группы данных. Множество D_1 содержало маркированные аудиозаписи без дополнительных воздействий. К записям D_1 применялись допустимые преобразования, в результате чего формировалось множество D_2 . Использовались 14 конфигураций преобразований, включающих частотную фильтрацию, добавление шума, сжатие с потерями, переквантование, передискретизацию и масштабирование амплитуды.

Множество D_3 формировалось путем применения недопустимых модификаций. В эксперименте рассматривались преобразование голоса методом kNN-VC [138], преобразование голоса с сохранением исходного диктора [139] и реконструкция сигнала с использованием вокодера Vocos [125]. Таким образом, для записей из D_1 и D_2 ожидается положительный результат контроля целостно-

сти, а для D_3 — отрицательный. Такая организация эксперимента соответствует принятому в разделе 4.1 разделению преобразований на допустимые и недопустимые.

Для сопоставления CryptoMark с AudioSeal и SilentCipher выходные значения методов приводились к единому бинарному решению. AudioSeal формирует непрерывную оценку уверенности, которая сравнивается с порогом τ_a . Для SilentCipher внедряется фиксированная битовая последовательность, а решение принимается по доле совпавших после извлечения битов с порогом τ_s .

CryptoMark формирует последовательность результатов верификации цифровых подписей отдельных фрагментов. Для получения единого решения для аудиозаписи вычисляется доля фрагментов, успешно прошедших проверку:

$$p_V = \frac{1}{N} \sum_{i=1}^N v_i. \quad (4.20)$$

Запись считается прошедшей контроль целостности, если

$$p_V \geq \tau_c, \quad (4.21)$$

где τ_c — порог доли успешно верифицированных фрагментов.

Для каждого метода использовался единый порог на всех рассматриваемых преобразованиях и модификациях. Значения τ_a , τ_s и τ_c выбирались по максимуму точности верификации на совокупности D_1 , D_2 и D_3 . В отличие от настройки отдельного порога для каждого воздействия, такой подход соответствует сценарию, в котором тип предварительной обработки исследуемой аудиозаписи заранее неизвестен.

Результаты оценивались отдельно для допустимых преобразований и недопустимых модификаций. Для допустимых преобразований основной характеристикой являлось значение TPR, определяющее долю записей, для которых после воздействия сохраняется положительный результат контроля целостности. Оно характеризует устойчивость метода к соответствующему допустимому преобразованию. Для недопустимых модификаций использовался FPR, определяющий долю модифицированных записей, ошибочно признанных прошедшими контроль целостности. Меньшему значению FPR соответствует более надежное выявление недопустимых изменений. Дополнительно для каждого воздействия вычислялась площадь под ROC-кривой (ROC AUC), характеризующая способность метода разделять положительный и отрицательный классы независимо от выбранного

порога. Для каждого воздействия значение ROC AUC вычислялось независимо. Результаты представлены в таблице 9.

Таблица 9 — Сравнение CryptoMark, SilentCipher и AudioSeal при допустимых преобразованиях и содержательно значимых модификациях

	CryptoMark		SilentCipher		AudioSeal	
	Допустимые преобразования					
Воздействие	TPR	AUC	TPR	AUC	TPR	AUC
ФВЧ, 500 Гц	0.68	0.84	0.99	1.00	1.00	1.00
ФНЧ, 3 кГц	0.75	0.87	0.99	0.99	0.01	0.78
MPEG, 320 кбит/с	1.00	1.00	1.00	1.00	1.00	1.00
MPEG, 256 кбит/с	1.00	1.00	1.00	1.00	1.00	1.00
MPEG, 128 кбит/с	1.00	1.00	1.00	1.00	1.00	1.00
MPEG, 96 кбит/с	1.00	1.00	1.00	1.00	1.00	1.00
MPEG, 64 кбит/с	1.00	1.00	1.00	1.00	1.00	1.00
Шум, SNR 30 дБ	0.16	0.58	0.89	0.94	1.00	1.00
Шум, SNR 40 дБ	0.79	0.89	1.00	1.00	1.00	1.00
Переквантование, 16 бит	1.00	1.00	1.00	1.00	1.00	1.00
Передискретизация, 16 кГц	1.00	1.00	1.00	1.00	1.00	1.00
Передискретизация, 22,05 кГц	1.00	1.00	1.00	1.00	1.00	1.00
Изменение амплитуды, 0,9	1.00	1.00	1.00	1.00	1.00	1.00
Изменение амплитуды, 1,1	1.00	1.00	1.00	1.00	1.00	1.00
	Недопустимые модификации					
Воздействие	FPR	AUC	FPR	AUC	FPR	AUC
Преобразование голоса kNN-VC	0.00	1.00	0.00	1.00	0.00	1.00
kNN-VC (исходный диктор)	0.00	1.00	0.00	1.00	0.00	1.00
Реконструкция Vocos	0.00	1.00	0.00	1.00	0.00	1.00

Для допустимых преобразований положительный результат характеризует совместную устойчивость перцептивного хеш-кода и канала передачи цифровой подписи. Ошибка хотя бы одного из этих компонентов может привести к отрицательной верификации цифровой подписи соответствующего фрагмента. Для недопустимых модификаций, напротив, отрицательный результат проверки является ожидаемым и свидетельствует о несоответствии извлеченной подписи текущему перцептивному представлению аудиоданных.

Полученные результаты показывают, что CryptoMark сохраняет устойчивость к большинству рассмотренных допустимых преобразований. Для сжатия MPEG во всем исследованном диапазоне битрейтов, переквантования, передискретизации и масштабирования амплитуды достигается безошибочная клас-

сификация. Наибольшее снижение устойчивости наблюдается при фильтрации верхних частот (ФВЧ), фильтрации нижних частот (ФНЧ) и добавлении шума. В частности, наиболее существенное ухудшение результатов отмечается при шуме с отношением сигнал/шум 30 дБ.

По сравнению с рассматриваемыми базовыми методами CryptoMark в среднем несколько уступает SilentCipher по устойчивости к допустимым преобразованиям, но сохраняет близкий уровень качества классификации. Для отдельных воздействий различия между методами имеют выраженный характер: например, AudioSeal теряет устойчивость при низкочастотной фильтрации, к чему CryptoMark оказывается более устойчив.

Для всех исследованных содержательно значимых модификаций CryptoMark получено значение FPR, равное нулю, при ROC AUC, равном единице. Ни одна из исследованных модифицированных записей не была ошибочно признана прошедшей контроль целостности.

Полученные результаты показывают, что включение проверки цифровой подписи не препятствует достижению устойчивости к большинству рассматриваемых допустимых преобразований. Наиболее сильное влияние на возможность проверки ЦВЗ после допустимых преобразований оказывают шумовые воздействия и частотная фильтрация. При этом положительный результат работы алгоритма верификации CryptoMark имеет более строгую интерпретацию по сравнению с обнаружением ЦВЗ: он подтверждает соответствие извлеченной цифровой подписи текущему перцептивному представлению аудиосигнала относительно заданного открытого ключа.

4.3.4 Оценка локализации модификаций

Последний эксперимент направлен на оценку способности CryptoMark определять временное положение содержательной модификации. Сравнение проводилось с AudioSeal, который также предоставляет локальные результаты обнаружения. SilentCipher в данном эксперименте не рассматривался, поскольку не обеспечивает локализацию ЦВЗ.

Для каждой сформированной из LibriSpeech аудиозаписи сначала выполнялось внедрение ЦВЗ. Затем непрерывный временной участок маркированного

сигнала заменялся результатом преобразования исходного содержимого методом kNN-VC. Тем самым моделировалась частичная семантическая подделка, при которой часть аудиозаписи остается неизменной, а отдельный временной интервал подвергается содержательной модификации.

После обработки CryptoMark формировал карту целостности $(v_1, \dots, v_N)$, определенную в разделе 4.1. Результат каждого фрагмента распространялся на соответствующий ему временной интервал, что позволяло сопоставить предсказанную область изменения с фактическим положением замененного участка. Для AudioSeal использовались формируемые методом локальные оценки присутствия ЦВЗ.

Качество локализации оценивалось на уровне временных отсчетов с помощью Accuracy и Intersection over Union (IoU). Accuracy характеризует долю отсчетов, для которых принадлежность к исходной или модифицированной области определена корректно, тогда как IoU отражает степень перекрытия предсказанного и фактического временных интервалов областей, для которых алгоритм верификации целостности должен выдавать положительный результат. Эксперимент повторялся при различной относительной длительности модифицированного участка. Результаты представлены на рисунке 4.2. Закрашенные области соответствуют 95%-м доверительным интервалам, рассчитанным с использованием распределения Стьюдента.

Полученные результаты показывают, что Accuracy для CryptoMark остается стабильно высокой во всем рассматриваемом диапазоне длительности модифицированного участка. Для AudioSeal значение Accuracy заметнее зависит от размера измененного фрагмента и снижается в среднем диапазоне длительностей, после чего возрастает при увеличении доли модификации.

По метрике IoU оба метода демонстрируют близкие результаты при малой длительности модифицированного участка. В среднем диапазоне CryptoMark обеспечивает более высокое перекрытие предсказанной и фактической областей, однако при долях модификации, близких к единице, значение IoU снижается. Это связано с пофрагментной организацией метода, при которой точность определения границ ограничивается длиной верифицируемого фрагмента.

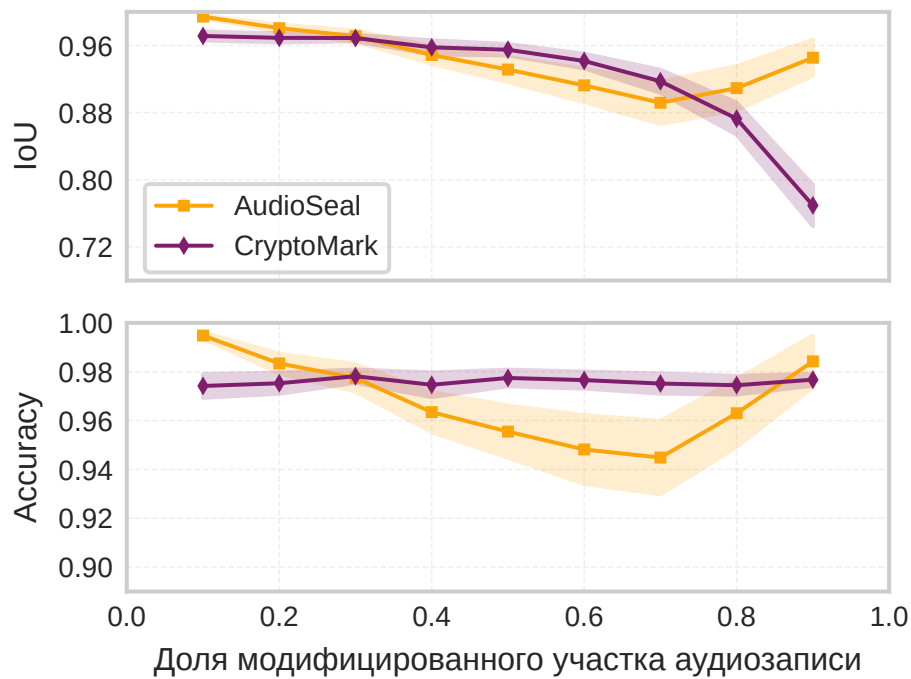


Рисунок 4.2 — Зависимость качества локализации от доли модифицированного участка аудиозаписи

4.4 Ограничения метода

Несмотря на возможность проверки целостности при помощи цифровой подписи и временной локализации модификаций, разработанный метод имеет ряд ограничений, определяющих область его применимости.

Во-первых, текущая реализация CryptoMark не содержит отдельного механизма восстановления временной синхронизации. Разбиение аудиосигнала выполняется по фиксированным границам, поэтому вставка или удаление отсчетов, изменение скорости воспроизведения и временное масштабирование могут приводить к смещению последующих фрагментов. В результате повторно вычисляемый перцептивный хеш-код и извлекаемая цифровая подпись относятся к фрагментам, не совпадающим с использованными на этапе внедрения, что может вызывать отрицательный результат проверки даже при отсутствии содержательной модификации.

Во-вторых, точность локализации непосредственно ограничена длиной проверяемого фрагмента L_f . Если граница модифицированного участка проходит внутри фрагмента, результат проверки относится ко всему соответствующему временному интервалу. Уменьшение L_f повышает временное разрешение, однако

одновременно уменьшает часть сигнала, доступную для устойчивого внедрения 512-битной цифровой подписи и проверочной информации помехоустойчивого кода.

В-третьих, свойства используемого перцептивного хеширования подтверждаются экспериментально для заданного множества преобразований. Формальные гарантии сохранения хеш-кода для всех преобразований, не изменяющих смысловое содержание аудио, а также гарантии его изменения при произвольных содержательных модификациях отсутствуют. В частности, модификация, сохраняющая исходное значение перцептивного хеш-кода, не будет выявлена средствами последующей проверки цифровой подписи. Поэтому криптографическая стойкость цифровой подписи сама по себе не обеспечивает устойчивость системы к атакам на используемое перцептивное представление.

Наконец, положительная верификация зависит как от устойчивости перцептивного хеш-кода, так и от точности восстановления цифровой подписи. Преобразования, выходящие за пределы исследованной области устойчивости ЦВЗ, могут приводить к отрицательному результату для немодифицированного содержимого. Отрицательная верификация означает невозможность подтвердить целостность соответствующего участка, но не позволяет однозначно установить причину отказа.

4.5 Выводы по главе

В данной главе разработан метод CryptoMark для контроля целостности аудиоданных на основе совместного применения цифровых водяных знаков, перцептивного хеширования и электронной цифровой подписи. Пофрагментная организация метода учитывает временную структуру аудиосигнала и позволяет не только выявлять нарушение целостности, но и локализовать временные участки, для которых целостность не подтверждена.

Для каждого фрагмента формируется контентно-зависимый аутентификатор, связанный с его перцептивным представлением и внедряемый в аудиосигнал в виде ЦВЗ. При проверке извлеченная цифровая подпись сопоставляется с повторно вычисленным перцептивным представлением фрагмента с использова-

нием открытого ключа, а последовательность результатов проверки формирует карту целостности аудиозаписи.

Экспериментальная оценка показала, что CryptoMark обеспечивает незаметность, сопоставимую с рассматриваемыми методами AudioSeal и SilentCipher, при существенно большей емкости. Метод сохраняет положительный результат контроля для большинства исследованных допустимых преобразований. Наибольшее снижение устойчивости наблюдается при частотной фильтрации и добавлении шума. Все рассмотренные содержательно значимые модификации были корректно выявлены.

Эксперименты по локализации показали, что CryptoMark позволяет определять временное положение модифицированных участков и сохраняет высокую точность классификации при изменении их длительности. Точность определения границ при этом ограничивается длиной независимо верифицируемого фрагмента.

Таким образом, CryptoMark обеспечивает проверяемую связь между содержимым временных фрагментов аудиозаписи, внедренными аутентификаторами и заданным открытым ключом, а также позволяет локализовать участки, целостность которых не подтверждена. Основные ограничения метода связаны с устойчивостью используемого перцептивного представления и ЦВЗ к допустимым преобразованиям, а также с отсутствием в текущей реализации отдельного механизма восстановления временной синхронизации.

Глава 5. Методика переноса методов маркирования изображений на видеоданные

В разделе 1.4 рассмотрены основные подходы к маркированию видеоданных и показано, что непосредственное покадровое применение методов маркирования изображений обеспечивает их перенос на видео без изменения исходного алгоритма, однако характеризуется высокой вычислительной стоимостью и не использует временную избыточность видеопоследовательности. Существующие способы совместной обработки кадров, в свою очередь, могут требовать изменения архитектуры исходного метода, доступа к его внутренним представлениям или дополнительного обучения.

В настоящей главе предлагается методика переноса существующих методов маркирования изображений на видеоданные, основанная на групповой обработке последовательных кадров. Для группы кадров формируется агрегированное изображение, к которому однократно применяется исходный алгоритм маркирования. Сформированная модификация распределяется между кадрами группы, а при извлечении используется аналогичное агрегирование. Такой подход позволяет сократить число вызовов исходных алгоритмов без изменения их архитектуры или дополнительного обучения.

5.1 Постановка задачи переноса методов маркирования изображений на видео

Пусть задан метод маркирования изображений, включающий алгоритм внедрения E и алгоритм извлечения X . Для изображения I и сообщения m алгоритм внедрения формирует маркированное изображение

$$I_w = E(I, m), \quad (5.1)$$

тогда как алгоритм извлечения восстанавливает внедренное сообщение:

$$\hat{m} = X(I_w). \quad (5.2)$$

Видеопоследовательность представляет собой упорядоченный набор кадров $V = \{I_1, \dots, I_T\}$. Требуется на основе исходных алгоритмов E и X построить алгоритмы внедрения E_V и извлечения X_V для видеоданных:

$$\begin{cases} V_w = E_V(V, m), \\ \hat{m} = X_V(V_w), \end{cases} \quad (5.3)$$

где V_w — маркированная видеопоследовательность.

Разрабатываемые алгоритмы E_V и X_V должны удовлетворять следующим требованиям:

1. применимость к исходным алгоритмам внедрения и извлечения без изменения их архитектуры, параметров или дополнительного обучения;
2. сокращение числа вызовов алгоритмов E и X по сравнению с независимой покадровой обработкой;
3. сохранение устойчивости извлечения внедренного сообщения после преобразований видеоданных;
4. ограничение пространственных и временных искажений, возникающих при маркировании.

5.2 Методика группового переноса методов маркирования изображений на видеоданные

Для решения поставленной задачи предлагается выполнять маркирование и извлечение ЦВЗ для нескольких последовательных кадров совместно. В качестве общего представления группы используется усредненное изображение, к которому применяется исходный метод маркирования изображений. Сформированная для него модификация затем переносится на все кадры группы.

5.2.1 Формирование агрегированного представления группы кадров

Видеопоследовательность V разбивается на непересекающиеся группы из N последовательных кадров. Число групп определяется как

$$G = \left\lceil \frac{T}{N} \right\rceil, \quad (5.4)$$

где последняя группа при необходимости может содержать менее N кадров. Обозначим кадры g -й группы как

$$\mathcal{G}_g = \{I_{g,1}, I_{g,2}, \dots, I_{g,N_g}\}, \quad N_g \leq N, \quad (5.5)$$

где N_g — фактическое число кадров в группе.

Для каждой группы формируется агрегированное представление в виде усредненного изображения:

$$\bar{I}_g = \frac{1}{N_g} \sum_{i=1}^{N_g} I_{g,i}. \quad (5.6)$$

Полученное изображение имеет ту же размерность, что и отдельный кадр видеопоследовательности, и может непосредственно использоваться в качестве входа исходного алгоритма маркирования изображений. В отличие от применения модификации, сформированной для отдельного кадра, агрегированное представление учитывает содержимое всех кадров группы.

5.2.2 Внедрение цифрового водяного знака

Алгоритм внедрения E однократно применяется к агрегированному изображению:

$$\bar{I}_{w,g} = E(\bar{I}_g, m), \quad (5.7)$$

после чего определяется сформированная им модификация

$$D_g = \bar{I}_{w,g} - \bar{I}_g. \quad (5.8)$$

Данная модификация добавляется к каждому исходному кадру группы:

$$I'_{g,i} = \text{clip}(I_{g,i} + D_g, 0, 1), \quad i = 1, \dots, N_g, \quad (5.9)$$

где $\text{clip}(\cdot)$ обозначает ограничение значений допустимым диапазоном.

Основное свойство предлагаемого способа следует из линейности усреднения. Если не учитывать ограничение значений отдельных пикселей, среднее маркированных кадров равно

$$\frac{1}{N_g} \sum_{i=1}^{N_g} I'_{g,i} = \frac{1}{N_g} \sum_{i=1}^{N_g} (I_{g,i} + D_g) = \bar{I}_g + D_g = E(\bar{I}_g, m). \quad (5.10)$$

Таким образом, агрегирование маркированных кадров восстанавливает результат, который был получен при непосредственном маркировании среднего исходных кадров. Именно это свойство используется далее при извлечении ЦВЗ.

В общем случае модификация может дополнительно масштабироваться коэффициентами $\beta_{g,i}$, зависящими от положения кадра внутри группы:

$$I'_{g,i} = \text{clip}(I_{g,i} + \beta_{g,i} D_g, 0, 1). \quad (5.11)$$

Для сохранения свойства (5.10) коэффициенты должны удовлетворять условию

$$\frac{1}{N_g} \sum_{i=1}^{N_g} \beta_{g,i} = 1. \quad (5.12)$$

Базовому варианту методики соответствует $\beta_{g,i} = 1$. Другие способы временного распределения модификации рассматриваются далее экспериментально.

5.2.3 Групповое извлечение цифрового водяного знака

На этапе извлечения для каждой группы также формируется усредненное изображение:

$$\bar{I}'_g = \frac{1}{N_g} \sum_{i=1}^{N_g} I'_{g,i}. \quad (5.13)$$

К полученному изображению однократно применяется исходный алгоритм извлечения:

$$\hat{m}_g = X(\bar{I}'_g). \quad (5.14)$$

В соответствии с (5.10), при отсутствии ограничения значений на вход алгоритма X поступает результат непосредственного маркирования изображения $\bar{I}_g$. Поэтому извлечение выполняется в той же постановке, для которой предназначен исходный метод маркирования изображений.

Для одной видеопоследовательности формируется набор оценок сообщения $\{\hat{m}_1, \dots, \hat{m}_G\}$. Если исходный алгоритм извлечения помимо восстановленных битов формирует вещественные оценки их принадлежности соответствующим значениям, результаты извлечения для нескольких групп могут быть дополнительно агрегированы на уровне всей видеопоследовательности. Обозначим через $\mathbf{s}_g = (s_{g,1}, \dots, s_{g,L})$ вектор таких оценок для g -й группы. Итоговая оценка для каждого бита определяется усреднением по группам:

$$\bar{s}_j = \frac{1}{G} \sum_{g=1}^G s_{g,j}, \quad j = 1, \dots, L. \quad (5.15)$$

Итоговое сообщение для видеопоследовательности формируется путем порогового преобразования агрегированных оценок:

$$\hat{m}_j = \mathbf{I}[\bar{s}_j \geq \tau], \quad j = 1, \dots, L, \quad (5.16)$$

где τ — порог принятия решения, определяемый исходным методом маркирования. Для методов, возвращающих только бинарные значения, аналогичная агрегация может выполняться путем голосования по результатам отдельных групп.

5.3 Экспериментальная оценка методики

Экспериментальная оценка предложенной методики направлена на исследование ее применимости к методам маркирования изображений различного типа, влияния параметров групповой обработки на качество и устойчивость маркирования, а также достигаемого сокращения вычислительных затрат. В экспериментах использовались методы VideoSeal [35] и TrustMark [33], основанные на нейросетевых моделях, а также классический метод маркирования DCT CAISS [22], основанный на дискретном косинусном преобразовании. Использование алгоритмов, различающихся архитектурой и принципом формирования ЦВЗ,

позволяет оценить зависимость свойств предложенной методики от исходного метода маркирования изображений.

Эксперименты проводились на 1000 видеопоследовательностях из набора данных YouTube-VOS [140]. Каждая видеопоследовательность рассматривалась как упорядоченный набор кадров, к которому применялись исследуемые стратегии переноса метода маркирования изображений. Для одной видеопоследовательности во все группы внедрялось одно и то же сообщение. Если не указано иное, размер группы составлял $N = 30$ кадров.

В качестве базового варианта использовалось независимое покадровое маркирование, с которым сравнивались три стратегии сокращения числа вызовов исходных алгоритмов.

Для оценки качества маркирования использовались метрики, введенные в разделе 2.3.1. Устойчивость извлечения ЦВЗ оценивалась с использованием частоты битовых ошибок BER, а пространственная незаметность — с использованием метрик PSNR и LPIPS [94]. При вычислении PSNR для видеопоследовательности сначала усреднялись значения среднеквадратичной ошибки по всем кадрам, после чего полученное значение преобразовывалось в децибелы.

Дополнительно для оценки временных искажений вводилась метрика TLP100 [141], характеризующая изменение перцептивного расстояния между соседними кадрами после внедрения ЦВЗ. Для каждого межкадрового перехода $t = 1, \dots, T - 1$ определим локальное временное искажение как

$$d_t = 100 |\text{LPIPS}(I_t, I_{t+1}) - \text{LPIPS}(I'_t, I'_{t+1})|. \quad (5.17)$$

Метрика TLP100 для всей видеопоследовательности определяется как среднее значение локального временного искажения:

$$\text{TLP100} = \frac{1}{T-1} \sum_{t=1}^{T-1} d_t. \quad (5.18)$$

В отличие от покадровых метрик качества, TLP100 оценивает изменение временных связей между соседними кадрами: чем меньше ее значение, тем меньшее дополнительное временное искажение вносится в видеопоследовательность при маркировании.

5.3.1 Сравнение стратегий переноса методов маркирования изображений на видео

На первом этапе экспериментальной оценки проводилось сравнение различных стратегий применения методов маркирования изображений к видеоданным. Целью эксперимента являлась оценка влияния способа обработки последовательных кадров на устойчивость извлечения ЦВЗ, качество маркированного видео и число вызовов исходного алгоритма внедрения.

Рассматривались четыре стратегии. При покадровой обработке алгоритм маркирования независимо применяется к каждому кадру видеопоследовательности. Вторая стратегия заключается в маркировании только первого кадра из каждой группы, остальные кадры остаются неизменными. В стратегии переноса модификации алгоритм внедрения также применяется только к первому кадру группы, после чего сформированная для него модификация переносится на остальные кадры группы. Наконец, при групповом маркировании используется предложенная в разделе 5.2 схема, в которой модификация формируется для среднего изображения группы и одинаково добавляется ко всем входящим в нее кадрам.

Во всех групповых вариантах использовался размер группы $N = 30$ кадров. Для оценки устойчивости маркированные видеопоследовательности подвергались кодированию H.264 [142] с параметром CRF=23, после чего выполнялось извлечение ЦВЗ. Итоговое сообщение формировалось путем агрегации вещественных оценок по всей видеопоследовательности в соответствии с процедурой, описанной выше. Метрики PSNR, LIPS и TLP100 рассчитывались для маркированного видео до применения сжатия. Результаты сравнения приведены в таблице 10.

Покадровая стратегия обеспечивает высокую устойчивость извлечения, поскольку каждый кадр независимо обрабатывается исходным алгоритмом. Однако для видеопоследовательности из T кадров она требует T вызовов алгоритма внедрения. Для остальных рассматриваемых стратегий число вызовов сокращается до $\lceil T/N \rceil$, то есть приблизительно в N раз.

Маркирование только первого кадра группы обеспечивает низкий уровень пространственных и временных искажений, поскольку большая часть кадров остается неизменной. Однако для всех исследованных методов такой вариант

Таблица 10 — Сравнение стратегий переноса методов маркирования изображений на видеоданные

Стратегия	BER ↓	PSNR ↑, дБ	LPIPS ↓	TLP100 ↓	Вызовы внедрения ↓
VideoSeal					
Покадровая	0,003	44,9	0,0037	0,045	T
Первый кадр группы	0,312	59,7	0,0001	0,014	$[T/N]$
Перенос модификации	0,062	45,0	0,0049	0,070	$[T/N]$
Групповая агрегация	0,006	47,5	0,0022	0,027	$[T/N]$
TrustMark					
Покадровая	0,003	44,1	0,0027	0,026	T
Первый кадр группы	0,271	58,9	0,0001	0,010	$[T/N]$
Перенос модификации	0,030	44,2	0,0032	0,038	$[T/N]$
Групповая агрегация	0,007	44,3	0,0024	0,026	$[T/N]$
DCT CAISS					
Покадровая	0,000	41,6	0,0783	0,523	T
Первый кадр группы	0,333	56,4	0,0026	0,414	$[T/N]$
Перенос модификации	0,000	41,7	0,0773	0,294	$[T/N]$
Групповая агрегация	0,000	43,4	0,0567	0,223	$[T/N]$

сопровождается существенным ухудшением извлечения после сжатия. Высокие показатели качества в данном случае достигаются за счет того, что ЦВЗ отсутствует в большей части видеопоследовательности, и не характеризуют качество полноценного маркирования видео.

Перенос на всю группу модификации, сформированной для первого кадра, позволяет повысить устойчивость по сравнению с маркированием только отдельных кадров. Тем не менее для нейросетевых методов VideoSeal и TrustMark устойчивость такого варианта остается заметно ниже как покадровой, так и предлагаемой групповой обработки. Это показывает, что модификация, сформированная с учетом содержимого одного кадра, не всегда сохраняет свойства ЦВЗ при непосредственном применении к другим кадрам группы.

Предлагаемая стратегия отличается тем, что модификация формируется не для одного выбранного кадра, а для агрегированного представления всей группы. За счет этого один вызов алгоритма внедрения учитывает содержимое всех совместно обрабатываемых кадров. Для VideoSeal и TrustMark это позволяет

сохранить устойчивость существенно ближе к покадровому варианту, чем при маркировании отдельных кадров или переносе модификации, при том же числе вызовов исходного алгоритма. Одновременно пространственные искажения по PSNR и LPIPS не увеличиваются относительно покадровой обработки, а для VideoSeal дополнительно уменьшаются и временные искажения по TLP100.

Для метода DCT CAISS также проявляется преимущество групповой обработки. Она сохраняет безошибочное извлечение, характерное для покадрового варианта и переноса модификации, но при этом обеспечивает лучшие значения всех рассматриваемых показателей качества. Таким образом, даже в случае, когда простой перенос модификации не приводит к ухудшению BER, формирование модификации по агрегированному представлению позволяет уменьшить пространственные и временные искажения.

Полученные результаты показывают, что предлагаемая стратегия обеспечивает более благоприятный компромисс между устойчивостью извлечения, незаметностью маркирования и вычислительной стоимостью. По сравнению с покадровой обработкой она сокращает число вызовов исходного алгоритма приблизительно пропорционально размеру группы, сохраняя близкую устойчивость извлечения. По сравнению с другими стратегиями с тем же числом вызовов она обеспечивает более высокую устойчивость для нейросетевых методов и меньшие пространственные и временные искажения для метода на основе ДКП.

5.3.2 Влияние размера группы кадров

Одним из основных параметров предложенной методики является размер группы N , определяющий число последовательных кадров, совместно используемых для формирования агрегированного представления. Увеличение размера группы сокращает число вызовов исходных алгоритмов внедрения и извлечения, однако одновременно изменяет свойства усредненного изображения и может влиять на устойчивость восстановления ЦВЗ и качество маркированного видео.

В эксперименте использовался базовый вариант группового маркирования, при котором одна и та же модификация добавлялась ко всем кадрам группы, $\beta_{g,i} = 1$. В качестве базового уровня использовалось независимое покадровое

маркирование, обозначенное значением $N = 1$. Эксперимент проводился для метода VideoSeal.

Влияние размера группы на пространственную и временную незаметность маркирования представлено в таблице 11. Отдельно выделено значение $N = 30$, используемое далее в качестве основного размера группы.

Таблица 11 — Влияние размера группы кадров на качество маркированного видео

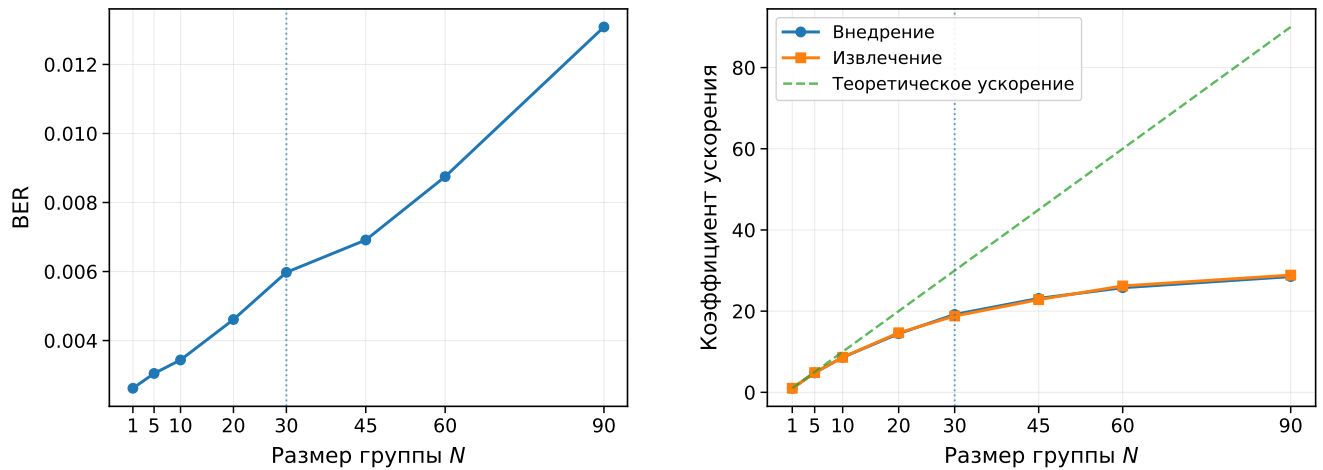
N	PSNR $\uparrow$, дБ	LPIPS $\downarrow$	TLP100 $\downarrow$
1	45,5	0,0037	0,045
5	46,5	0,0029	0,053
10	47,1	0,0026	0,040
20	47,8	0,0023	0,031
30	48,3	0,0022	0,027
45	48,7	0,0020	0,024
60	49,0	0,0019	0,022
90	49,6	0,0018	0,020

С увеличением размера группы наблюдается последовательное улучшение пространственной незаметности маркирования. При изменении N от 5 до 90 значение PSNR возрастает с 46,5 до 49,6 дБ, а LPIPS уменьшается с 0,0029 до 0,0018. Таким образом, увеличение числа совместно обрабатываемых кадров приводит к уменьшению средней величины пространственных искажений отдельных кадров.

Для временной метрики TLP100 зависимость имеет несколько иной характер. При $N = 5$ ее значение составляет 0,053 и превышает результат покадрового маркирования 0,045. Однако при дальнейшем увеличении размера группы TLP100 последовательно уменьшается. Следовательно, при достаточно больших группах совместное формирование модификации позволяет уменьшить изменение перцептивных различий между соседними кадрами по сравнению с независимым покадровым маркированием.

Изменение устойчивости извлечения и вычислительной эффективности при увеличении размера группы показано на рисунке 5.1. Для оценки устойчивости использовался BER итогового сообщения, полученного после агрегации результатов извлечения по всей видеопоследовательности, к которой было применено сжатие кодеком H.264 с параметром CRF=23. Вычислительная эффективность

характеризовалась отношением времени выполнения покадрового варианта к времени групповой обработки отдельно для внедрения и извлечения.



а) Зависимость устойчивости ЦВЗ к сжатию кодеком H.264 от размера группы кадров

б) Зависимость ускорения от размера группы кадров

Рисунок 5.1 — Влияние размера группы кадров на устойчивость извлечения и вычислительную эффективность предложенной методики

Как видно из рисунка 5.1a, увеличение размера группы сопровождается постепенным ростом BER. Одновременно увеличение N приводит к существенному сокращению вычислительных затрат. Как показано на рисунке 5.1б, при $N = 5$ ускорение внедрения составляет 4,8 раза, а извлечения — 4,9 раза. При $N = 30$ соответствующие значения достигают 19,2 и 18,8 раза. Дальнейшее увеличение размера группы продолжает повышать вычислительную эффективность, однако фактическое ускорение растет медленнее теоретического сокращения числа вызовов исходного алгоритма. Это связано с наличием операций, вычислительные затраты которых не сокращаются пропорционально числу групп.

Таким образом, размер группы определяет компромисс между устойчивостью извлечения, незаметностью маркирования и вычислительной эффективностью. При малых значениях N достигается устойчивость, близкая к покадровой обработке, однако выигрыш по времени остается ограниченным. Увеличение N улучшает пространственную и временную незаметность, а также сокращает время обработки, но сопровождается ростом BER. Для последующих экспериментов было выбрано значение $N = 30$ как компромиссное значение размера группы.

5.3.3 Влияние временного распределения модификации

В базовом варианте предложенной методики вычисленная для группы кадров модификация D_g добавляется ко всем кадрам с одинаковым коэффициентом $\beta_{g,i} = 1$. В этом случае величина модификации остается постоянной внутри группы, однако при переходе между соседними группами происходит ее скачкообразная замена с D_g на D_{g+1} . Это может приводить к возникновению локальных визуальных искажений на границах групп.

Для уменьшения данного эффекта были исследованы варианты плавного распределения модификации между кадрами. Рассматривались синусоидальное окно, окна Тьюки с параметрами $\alpha = 0,25$ и $\alpha = 0,50$, а также окно Ханна [143]. Эксперимент проводился для метода VideoSeal при размере группы $N = 30$ кадров.

Для обеспечения свойства (5.12) исследуемых вариантов коэффициенты окон нормировались по среднему значению:

$$\beta_{g,i} = \frac{w_i}{\frac{1}{N_g} \sum_{j=1}^{N_g} w_j}, \quad \frac{1}{N_g} \sum_{i=1}^{N_g} \beta_{g,i} = 1, \quad (5.19)$$

где w_i — значение соответствующей оконной функции для i -го кадра группы.

На рисунке 5.2 представлены используемые оконные функции для группы из 30 кадров. При равномерном распределении коэффициент остается постоянным. Для оконных вариантов его значение уменьшается вблизи границ группы и увеличивается в ее центральной части.

Помимо общего значения TLP100 для оценки наиболее выраженных локальных временных искажений рассчитывался средний максимум значений d_t , используемых при расчете TLP100 (5.17). Для каждой группы выбиралось максимальное значение d_t среди переходов внутри группы и перехода к следующей группе, после чего полученные максимумы усреднялись. В отличие от значения TLP100, данный показатель позволяет учитывать кратковременные локальные изменения, которые могут быть сглажены при усреднении по всему видео. Результаты эксперимента представлены в таблице 12.

Результаты показывают, что выбор функции временного распределения незначительно влияет на устойчивость извлечения ЦВЗ по BER, однако заметно изменяет пространственные и временные характеристики маркированного видео.

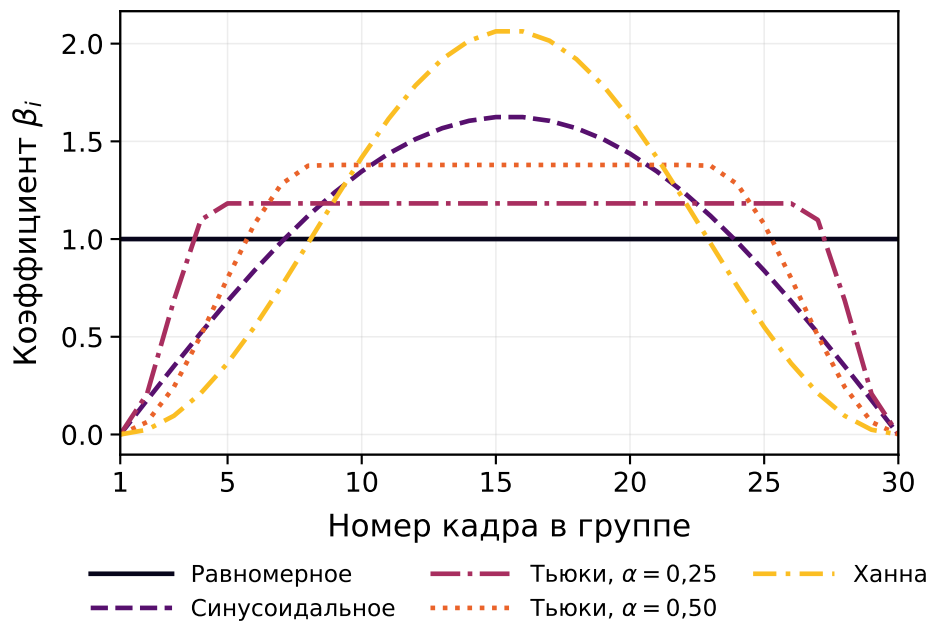


Рисунок 5.2 — Нормированные коэффициенты временного распределения модификации в группе из $N = 30$ кадров

Таблица 12 — Влияние временного распределения модификации на качество и устойчивость маркированного видео

Вариант	BER ↓	PSNR ↑, дБ	LPIPS ↓	TLP100 ↓	Средний максимум временного искажения в группе ↓
Каждый кадр	0,003	44,9	0,004	0,045	0,097
Равномерное	0,006	47,5	0,002	0,027	0,159
Синусоидальное	0,006	46,7	0,003	0,027	0,094
Тьюки, $\alpha = 0,25$	0,006	47,2	0,002	0,027	0,085
Тьюки, $\alpha = 0,50$	0,005	46,8	0,003	0,028	0,090
Ханна	0,005	45,9	0,003	0,031	0,125

Согласно значениям метрик PSNR и LPIPS, равномерное распределение вносит минимальные пространственные искажения, но при этом характеризуется наибольшим средним максимумом d_t в группе. Это указывает на наличие выраженных локальных изменений временной согласованности, возникающих преимущественно на границах соседних групп.

Использование оконных функций позволяет уменьшить средний максимум d_t за счет более плавного изменения коэффициентов модификации между кадрами. При этом значение TLP100 остается близким к равномерному варианту, а изменения BER, PSNR и LPIPS для большинства окон остаются незначительными.

Наиболее сбалансированные результаты среди рассмотренных окон получены для окна Тьюки с $\alpha = 0,25$. Оно обеспечивает наименьший средний максимум d_t при сохранении близких к равномерному варианту значений BER, PSNR и LPIPS. Окно Ханна, напротив, сильнее перераспределяет модификацию внутри группы, что сопровождается ухудшением как пространственных характеристик, так и значения TLP100.

Таким образом, равномерное распределение целесообразно использовать как базовый вариант методики, если приоритетом является минимизация пространственных искажений. Окно Тьюки с $\alpha = 0,25$ может использоваться в случаях, когда более важным является снижение локальных временных искажений.

5.3.4 Устойчивость к преобразованиям видеоданных

Одним из наиболее характерных преобразований видеоданных является сжатие с потерями, применяемое при хранении, передаче и публикации видеоконтента. Для оценки сохранения устойчивости исходных методов маркирования при их переносе на групповую обработку исследовалось влияние сжатия кодеком H.264 на точность извлечения ЦВЗ.

В эксперименте сравнивались независимое покадровое маркирование и предлагаемое групповое маркирование с размером группы $N = 30$. Исследование проводилось для методов VideoSeal, TrustMark и DCT CAISS. После внедрения ЦВЗ видеопоследовательности кодировались H.264 при значениях CRF от 18 до 38. Увеличение CRF соответствует усилению сжатия и увеличению вносимых кодированием искажений.

После каждого преобразования выполнялось извлечение ЦВЗ. Вещественные оценки отдельных битов агрегировались по всей видеопоследовательности в соответствии с (5.15), после чего формировалось итоговое сообщение и рассчитывался BER. Результаты представлены на рисунке 5.3.

Для всех исследованных методов увеличение CRF сопровождается ростом BER как при покадровом, так и при групповом маркировании. При слабом и умеренном сжатии различие между двумя стратегиями остается небольшим. Для VideoSeal и TrustMark по мере увеличения CRF разрыв между ними становится

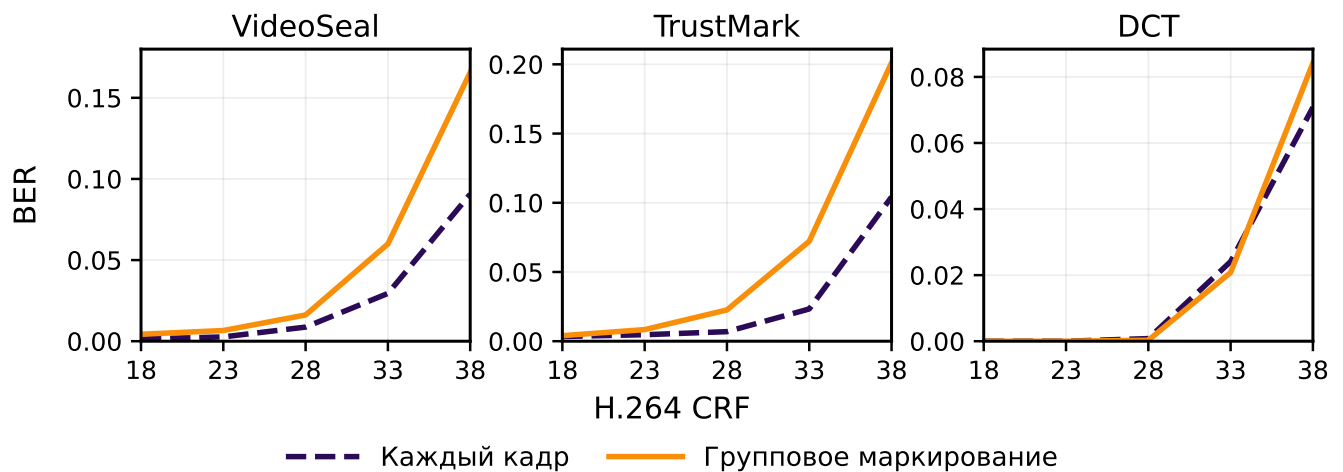


Рисунок 5.3 — Зависимость BER от степени сжатия H.264 для покадрового и группового маркирования различными методами

более выраженным: BER группового варианта растет быстрее, чем при независимом маркировании каждого кадра. Наиболее заметное различие наблюдается при максимальной исследованной степени сжатия.

Для метода DCT CAISS наблюдается аналогичный общий рост BER при усилении сжатия, однако различие между покадровой и групповой обработкой при промежуточных значениях CRF остается небольшим. В частности, при CRF=33 групповая обработка не приводит к увеличению BER относительно покадровой. При дальнейшем усилении сжатия преимущество покадрового варианта вновь становится заметным.

Полученные результаты показывают, что групповая обработка сохраняет общий характер зависимости устойчивости исходных методов от интенсивности видеосжатия. При слабом и умеренном сжатии снижение устойчивости по сравнению с покадровой обработкой остается ограниченным, тогда как при сильном сжатии для всех рассмотренных методов групповая стратегия характеризуется более высоким BER.

5.3.5 Устойчивость к нарушению временной синхронизации

Предложенная методика предполагает согласованное положение границ групп кадров на этапах внедрения и извлечения ЦВЗ. При нарушении временной синхронизации группа, формируемая при извлечении, может содержать кадры,

относящиеся к двум соседним группам этапа внедрения. В результате ее агрегированное представление включает модификации, сформированные для различного содержимого, что может приводить к снижению устойчивости извлечения.

Для исследования данного эффекта из начала маркированного видео удалялась часть кадров, после чего извлечение выполнялось без дополнительного восстановления синхронизации. Эксперимент проводился для метода VideoSeal при размере группы $N = 30$. Число удаляемых начальных кадров изменялось от 0 до 30 с шагом 3. Сравнивались покадровое и групповое маркирование. В качестве показателей использовались BER и уверенность извлечения ЦВЗ. Уверенность определялась как среднее расстояние агрегированных оценок от порога принятия решения:

$$C = \frac{1}{L} \sum_{j=1}^L |\bar{s}_j - \tau|, \quad (5.20)$$

где $\bar{s}_j$ — агрегированная оценка j -го бита, определенная в (5.15), а τ — порог принятия решения, заданный в (5.16). Результаты представлены на рисунке 5.4.

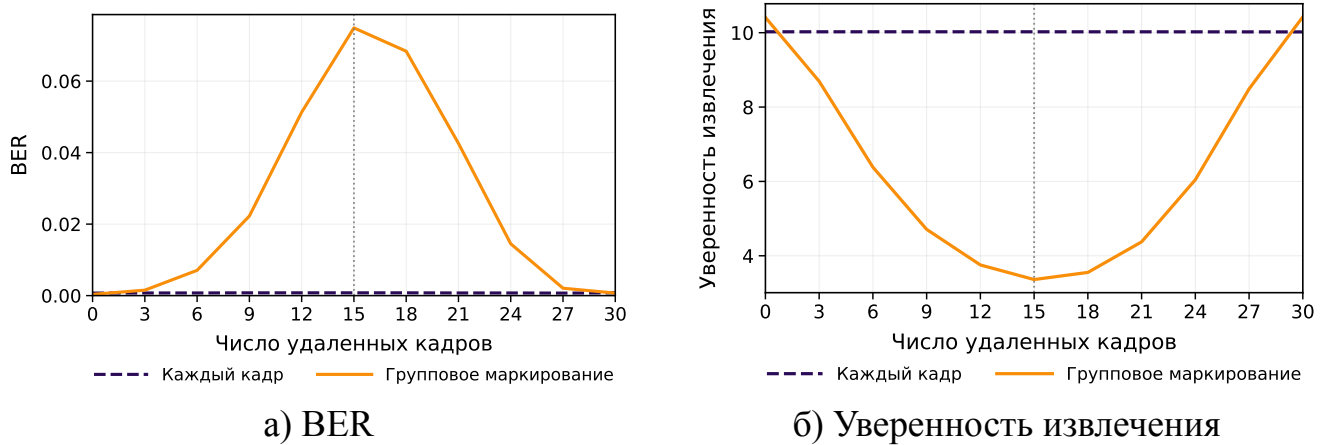


Рисунок 5.4 — Влияние нарушения временной синхронизации на извлечение ЦВЗ при размере группы $N = 30$

Для покадрового маркирования изменение положения начала видеопоследовательности практически не влияет на BER и уверенность извлечения. Для группового маркирования наблюдается выраженная зависимость от сдвига границ. По мере удаления кадров BER возрастает, достигая максимума при сдвиге, близком к половине размера группы, после чего уменьшается и при сдвиге размером N возвращается к уровню, близкому к исходному. Такой характер зависимости обусловлен тем, что при сдвиге около $N/2$ формируемая при извлечении

группа содержит сопоставимое количество кадров из двух соседних групп этапа внедрения. При сдвиге на размер группы положение границ вновь становится согласованным.

Уверенность извлечения изменяется противоположным образом: ее наибольшие значения соответствуют согласованному положению границ, а минимум наблюдается при сдвиге, близком к $N/2$. Наличие выраженной зависимости уверенности от временного сдвига позволяет использовать ее в качестве критерия поиска правильной фазы группировки. Для этого на этапе извлечения могут быть проверены возможные значения $r = 0, \dots, N - 1$, для каждого из которых рассчитывается уверенность $C(r)$ (5.20), после чего выбирается фазовый сдвиг, соответствующий ее максимальному значению:

$$\hat{r} = \arg \max_{0 \leq r < N} C(r). \quad (5.21)$$

Найденное значение $\hat{r}$ может использоваться для последующего извлечения ЦВЗ с восстановленной фазой группировки.

Следует учитывать, что полный перебор возможных фазовых сдвигов увеличивает вычислительную стоимость извлечения. Если для каждого $r = 0, \dots, N - 1$ требуется повторное формирование групп и вызов исходного алгоритма извлечения, число вызовов извлекающей модели возрастает приблизительно в N раз по сравнению с извлечением при известной синхронизации. Практическое применение такого поиска целесообразно рассматривать как дополнительный механизм восстановления синхронизации, используемый при необходимости.

5.4 Ограничения методики

Предложенная методика предназначена для переноса существующих методов маркирования изображений на видеоданные без изменения их архитектуры и дополнительного обучения. Устойчивость группового маркирования в значительной степени определяется свойствами исходного метода. Групповая обработка не компенсирует преобразования, к которым исходный алгоритм не обладает достаточной устойчивостью, а при сильных искажениях может дополнительно снижать точность извлечения.

Характеристики методики зависят от размера группы N . Его увеличение уменьшает число вызовов алгоритмов внедрения и извлечения и может снижать вносимые искажения, однако сопровождается ростом числа ошибок при извлечении ЦВЗ. Выбор значения N определяется требуемым компромиссом между вычислительной эффективностью, незаметностью и устойчивостью.

Другим ограничением является зависимость от временной структуры видео. При существенных изменениях содержимого внутри группы усредненное изображение может хуже представлять отдельные кадры. Кроме того, групповое извлечение предполагает согласованное положение границ групп на этапах внедрения и извлечения. Удаление или вставка кадров приводит к нарушению этой синхронизации и требует дополнительного поиска положения границ исходных групп.

Наконец, точное выполнение свойства (5.10) предполагает отсутствие ограничения значений пикселей. Применяемая на практике операция $\text{clip}(\cdot)$ может приводить к отклонению среднего маркированных кадров от результата непосредственного маркирования агрегированного изображения, что, в свою очередь, может снижать точность извлечения ЦВЗ.

5.5 Выводы по главе

В главе разработана методика переноса существующих методов маркирования изображений на видеоданные, основанная на совместной обработке групп последовательных кадров. Для каждой группы формируется агрегированное представление в виде среднего изображения, к которому однократно применяется исходный алгоритм внедрения. Полученная модификация распределяется между кадрами группы, а при извлечении используется аналогичная агрегация кадров. Такой подход не требует изменения архитектуры исходного метода или его дополнительного обучения.

Экспериментальное сравнение на нейросетевых методах VideoSeal и TrustMark, а также классическом методе DCT CAISS, основанном на ДКП, показало преимущество предлагаемой групповой стратегии перед более простыми способами сокращения числа вызовов алгоритма. По сравнению с маркированием отдельных кадров группы и переносом модификации, сформированной для

одного кадра, использование агрегированного представления обеспечивает более высокую устойчивость для нейросетевых методов и меньшие пространственные и временные искажения для метода на основе ДКП.

Исследование размера группы показало наличие компромисса между вычислительной эффективностью, качеством маркированного видео и устойчивостью извлечения ЦВЗ. Увеличение размера группы N сокращает время обработки и уменьшает пространственные искажения, однако сопровождается ростом числа ошибок при извлечении ЦВЗ. Для дальнейших экспериментов было выбрано значение $N = 30$, при котором достигалось ускорение внедрения и извлечения примерно в 19 раз относительно покадровой обработки.

Показано, что временное распределение модификации внутри группы позволяет управлять искажениями между соседними кадрами. Равномерное распределение обеспечивает минимальные пространственные искажения, тогда как применение оконных функций позволяет уменьшить наиболее выраженные локальные временные искажения на границах групп. Среди исследованных вариантов наиболее сбалансированные результаты получены для окна Тьюки с параметром $\alpha = 0,25$.

Исследование устойчивости к сжатию кодеком H.264 показало, что при слабом и умеренном сжатии групповая обработка сохраняет устойчивость, близкую к покадровому применению исходного метода. При сильном сжатии дополнительное ухудшение извлечения становится более выраженным, что является одним из компромиссов предлагаемой методики.

Отдельно исследовано нарушение временной синхронизации между этапами внедрения и извлечения. Установлено, что рассогласование границ групп приводит к росту BER и снижению оценки уверенности извлечения, причем наибольшее ухудшение наблюдается при сдвиге, близком к половине размера группы. Показано, что зависимость оценки уверенности от фазового сдвига может использоваться как критерий восстановления положения границ групп, однако полный перебор возможных фаз увеличивает вычислительную стоимость извлечения.

Заключение

Основные результаты работы заключаются в следующем.

1. Разработана методика унифицированной воспроизводимой оценки методов маркирования медиаконтента и создана реализующая ее программная система WIBE. Методика обеспечивает единую организацию экспериментального процесса для различных типов медиаданных при использовании модально-зависимых наборов данных, преобразований, атак и метрик. Экспериментальная апробация на методах маркирования изображений и аудиоданных показала необходимость совместного учета незаметности, емкости, устойчивости и характера рассматриваемых воздействий.
2. Разработан и экспериментально исследован метод CPW контроля целостности изображений на основе подписываемого перцептивного представления. Предложенный способ повышения побитовой устойчивости перцептивного хеш-кода позволяет сохранять подписываемое представление после рассматриваемых допустимых преобразований. Метод обеспечивает проверяемую связь между содержимым изображения, внедренным аутентификатором и заданным открытым ключом и позволяет выявлять рассматриваемые содержательно значимые модификации.
3. Разработан и экспериментально исследован метод CryptoMark контроля целостности аудиоданных на основе независимого формирования и проверки контентно-зависимых аутентификаторов последовательных временных фрагментов. Пофрагментная организация позволяет формировать карту целостности аудиозаписи, выявлять содержательно значимые модификации и локализовать их во времени при сохранении работоспособности для большинства исследованных допустимых преобразований.
4. Разработана и экспериментально исследована методика переноса методов маркирования изображений на видеоданные, основанная на однократной обработке агрегированного представления группы последовательных кадров и распределении сформированной модификации между кадрами. Методика позволяет применять существующие методы

без изменения их архитектуры и дополнительного обучения и снижает вычислительные затраты по сравнению с независимой покадровой обработкой. Экспериментально исследованы компромиссы между размером группы кадров, незаметностью, устойчивостью извлечения и вычислительной эффективностью, а также влияние временного распределения модификации и нарушения синхронизации.

Полученные результаты формируют комплекс средств для исследования и применения цифровых водяных знаков к медиаданным различных типов. Разработанная система WIBE обеспечивает воспроизводимую сравнительную оценку методов маркирования, методы CPW и CryptoMark предназначены для контроля целостности изображений и аудиоданных, а предложенная методика групповой обработки позволяет эффективно применять существующие методы маркирования изображений к видеоданным.

Список литературы

1. Маркирование текстовых документов на экране монитора посредством изменения яркости фона в областях межстрочных интервалов / А. Ю. Якушев [и др.] // Труды Института системного программирования РАН. — 2021. — Т. 33, № 4. — С. 147—162.
2. Docmarking: Real-time screen-cam robust document image watermarking / A. Yakushev [и др.] // 2022 Ivannikov Ispras Open Conference (ISPRAS). — IEEE. 2022. — С. 142—150.
3. Система маркирования документов для проведения расследований при их утечке / Д. О. Обыденков [и др.] // Труды Института системного программирования РАН. — 2021. — Т. 33, № 6. — С. 161—174.
4. Предотвращение анонимных утечек конфиденциальных документов / Д. О. Обыденков [и др.] // Материалы 33-й научно-технической конференции «Методы и технические средства обеспечения безопасности информации». — Санкт-Петербург, 2024. — С. 114—115.
5. Водяные знаки как инструмент защиты видеоконтента / А. Ю. Якушев [и др.] // Материалы 33-й научно-технической конференции «Методы и технические средства обеспечения безопасности информации». — Санкт-Петербург, 2024. — С. 117—118.
6. WIBE: Watermarks for generated Images–Benchmarking & Evaluation / A. Yakushev [и др.] // 2025 40th IEEE/ACM International Conference on Automated Software Engineering (ASE). — IEEE. 2025. — С. 4033—4036.
7. WARP: A Unified Benchmark for Invisible Image Watermarking – Robustness and Protection Against Attacks / K. Abud [и др.] // Proceedings of the 34th ACM International Conference on Multimedia (MM '26). — ACM, 2026. — Принято к публикации.
8. Контроль целостности изображений на основе цифровых водяных знаков, перцептивного хеширования и цифровой подписи / А. Ю. Якушев [и др.] // Труды Института системного программирования РАН. — 2026. — Т. 38, № 5. — С. 133—154.

9. HiDDeN: Hiding Data With Deep Networks / J. Zhu [и др.] // Computer Vision – ECCV 2018. Т. 11218. — Springer, Cham, 2018. — С. 682—697. — (Lecture Notes in Computer Science).
10. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust / Y. Wen [и др.] // arXiv preprint arXiv:2305.20030. — 2023.
11. *Cox, I.* Digital watermarking / I. Cox, M. L. Miller, J. A. Bloom. — San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 2001.
12. *Cayre, F.* Watermarking security: theory and practice / F. Cayre, C. Fontaine, T. Furon // IEEE Transactions on signal processing. — 2005. — Т. 53, № 10. — С. 3976—3987.
13. Microsoft COCO: Common Objects in Context / T.-Y. Lin [и др.] // European conference on computer vision. — Springer. 2014. — С. 740—755.
14. Electronic watermark / A. Z. Tirkel [и др.] // Digital Image Computing, Technology and Applications (DICTA'93). — 1993. — С. 666—673.
15. *Caronni, G.* Assuring ownership rights for digital images / G. Caronni // Verlässliche IT-Systeme: Proceedings der GI-Fachtagung VIS'95. — Springer, 1995. — С. 251—263.
16. A DCT-domain system for robust image watermarking / M. Barni [и др.] // Signal processing. — 1998. — Т. 66, № 3. — С. 357—372.
17. *Al-Haj, A.* Combined DWT-DCT digital image watermarking / A. Al-Haj // Journal of computer science. — 2007. — Т. 3, № 9. — С. 740—746.
18. *Ali, M.* A robust image watermarking technique using SVD and differential evolution in DCT domain / M. Ali, C. W. Ahn, M. Pant // Optik. — 2014. — Т. 125, № 1. — С. 428—434.
19. *O'Ruanaidh, J. J.* Rotation, scale and translation invariant digital image watermarking / J. J. O'Ruanaidh, T. Pun // Proceedings of International Conference on Image Processing. Т. 1. — IEEE. 1997. — С. 536—539.
20. Secure spread spectrum watermarking for multimedia / I. J. Cox [и др.] // IEEE transactions on image processing. — 1997. — Т. 6, № 12. — С. 1673—1687.
21. *Malvar, H. S.* Improved spread spectrum: A new modulation technique for robust watermarking / H. S. Malvar, D. A. Florêncio // IEEE transactions on signal processing. — 2003. — Т. 51, № 4. — С. 898—905.

22. *Valizadeh, A.* Correlation-and-bit-aware spread spectrum embedding for data hiding / A. Valizadeh, Z. J. Wang // IEEE Transactions on Information Forensics and Security. — 2011. — T. 6, № 2. — С. 267—282.
23. *Chen, B.* Quantization index modulation: A class of provably good methods for digital watermarking and information embedding / B. Chen, G. W. Wornell // IEEE Transactions on Information theory. — 2001. — T. 47, № 4. — С. 1423—1443.
24. *Fóris, P.* Implementations of HVS Models in Digital Image Watermarking / P. Fóris, D. Levický // Radioengineering. — 2007. — Vol. 16, no. 1. — P. 45—50.
25. ReDMark: Framework for residual diffusion watermarking based on deep networks / M. Ahmadi [и др.] // Expert Systems with Applications. — 2020. — T. 146. — С. 113157.
26. *Jia, Z.* Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression / Z. Jia, H. Fang, W. Zhang // Proceedings of the 29th ACM international conference on multimedia. — 2021. — С. 41—49.
27. ARWGAN: Attention-guided robust image watermarking model based on GAN / J. Huang [и др.] // IEEE Transactions on Instrumentation and Measurement. — 2023. — T. 72. — С. 1—17.
28. Towards blind watermarking: Combining invertible and non-invertible mechanisms / R. Ma [и др.] // Proceedings of the 30th ACM international conference on multimedia. — 2022. — С. 1532—1542.
29. Flow-based robust watermarking with invertible noise layer for black-box distortions / H. Fang [и др.] // Proceedings of the AAAI conference on artificial intelligence. T. 37. — 2023. — С. 5054—5061.
30. Watermarking images in self-supervised latent spaces / P. Fernandez [и др.] // ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — IEEE. 2022. — С. 3054—3058.
31. Rosteals: Robust steganography using autoencoder latent space / T. Bui [и др.] // 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). — IEEE. 2023. — С. 933—942.
32. *Esser, P.* Taming transformers for high-resolution image synthesis / P. Esser, R. Rombach, B. Ommer // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2021. — С. 12873—12883.

33. *Bui, T.* Trustmark: Universal watermarking for arbitrary resolution images / T. Bui, S. Agarwal, J. Collomosse // arXiv preprint arXiv:2311.18297. — 2023.
34. InvisMark: Invisible and Robust Watermarking for AI-generated Image Provenance / R. Xu [и др.] // 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). — 2025. — С. 909—918.
35. Video seal: Open and efficient video watermarking / P. Fernandez [и др.] // arXiv preprint arXiv:2412.09492. — 2024.
36. *Ronneberger, O.* U-net: Convolutional networks for biomedical image segmentation / O. Ronneberger, P. Fischer, T. Brox // International Conference on Medical image computing and computer-assisted intervention. — Springer. 2015. — С. 234—241.
37. Pixel Seal: Adversarial-only training for invisible image and video watermarking / T. Souček [и др.] // arXiv preprint arXiv:2512.16874. — 2025.
38. *Zhang, X.* Just-noticeable difference estimation with pixels in images / X. Zhang, W. Lin, P. Xue // Journal of Visual Communication and Image Representation. — 2008. — Т. 19, № 1. — С. 30—41.
39. Practical deep dispersed watermarking with synchronization and fusion / H. Guo [и др.] // Proceedings of the 31st ACM international conference on multimedia. — 2023. — С. 7922—7932.
40. Geometric image synchronization with deep watermarking / P. Fernandez [и др.] // ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — IEEE. 2026. — С. 9522—9526.
41. Watermark anything with localized messages / T. Sander [и др.] // International Conference on Learning Representations. — 2025.
42. *Tancik, M.* Stegastamp: Invisible hyperlinks in physical photographs / M. Tancik, B. Mildenhall, R. Ng // 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — IEEE. 2020. — С. 2114—2123.
43. PIMoG: An effective screen-shooting noise-layer simulation for deep-learning-based watermarking network / H. Fang [и др.] // Proceedings of the 30th ACM international conference on multimedia. — 2022. — С. 2267—2275.

44. Robust-wide: Robust watermarking against instruction-driven image editing / R. Hu [и др.] // European conference on computer vision. — Springer. 2024. — С. 20—37.
45. We can hide more bits: The unused watermarking capacity in theory and in practice / A. Petrov [и др.] // arXiv preprint arXiv:2510.12812. — 2025.
46. The stable signature: Rooting watermarks in latent diffusion models / P. Fernandez [и др.] // 2023 IEEE/CVF International Conference on Computer Vision (ICCV). — IEEE. 2023. — С. 22409—22420.
47. RingID: Rethinking tree-ring watermarking for enhanced multi-key identification / H. Ci [и др.] // European conference on computer vision. — Springer. 2024. — С. 338—354.
48. *Varlamov, A.* METR: Image watermarking with large number of unique messages / A. Varlamov, D. Diatlova, E. Spirin // arXiv preprint arXiv:2408.08340. — 2024.
49. *Mao, P.-Y.* MaXsive: High-capacity and robust training-free generative image watermarking in diffusion models / P.-Y. Mao, C.-C. Tsai, C.-S. Lu // Proceedings of the 33rd ACM International Conference on Multimedia. — 2025. — С. 11443—11452.
50. Gaussian Shading: Provable performance-lossless image watermarking for diffusion models / Z. Yang [и др.] // 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — IEEE. 2024. — С. 12162—12171.
51. *Gunn, S.* An Undetectable Watermark for Generative Image Models / S. Gunn, X. Zhao, D. Song // International Conference on Learning Representations. — 2025.
52. Gaussian Shading++: rethinking the realistic deployment challenge of performance-lossless image watermark for diffusion models / Z. Yang [и др.] // arXiv preprint arXiv:2504.15026. — 2025.
53. *Shehata, K.* CLUE-MARK: Watermarking Diffusion Models using CLWE / K. Shehata, A. Kolluri, P. Saxena // arXiv preprint arXiv:2411.11434. — 2025.
54. Continuous LWE / J. Bruna [и др.] // Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing. — 2021. — С. 694—707.

55. *Kirovski, D.* Spread-spectrum watermarking of audio signals / D. Kirovski, H. S. Malvar // IEEE transactions on signal processing. — 2003. — Т. 51, № 4. — С. 1020—1033.
56. Techniques for data hiding / W. Bender [и др.] // IBM systems journal. — 1996. — Т. 35, № 3/4. — С. 313—336.
57. *Hu, H.-T.* Robust, transparent and high-capacity audio watermarking in DCT domain / H.-T. Hu, L.-Y. Hsu // Signal Processing. — 2015. — Т. 109. — С. 226—235.
58. Patchwork-based multilayer audio watermarking / I. Natgunanathan [и др.] // IEEE/ACM Transactions on Audio, Speech, and Language Processing. — 2017. — Т. 25, № 11. — С. 2176—2187.
59. *Saadi, S.* Novel secured scheme for blind audio/speech norm-space watermarking by Arnold algorithm / S. Saadi, A. Merrad, A. Benziane // Signal Processing. — 2019. — Т. 154. — С. 74—86.
60. Desynchronization attacks resilient watermarking method based on frequency singular value coefficient modification / J. Zhao [и др.] // IEEE/ACM Transactions on Audio, Speech, and Language Processing. — 2021. — Т. 29. — С. 2282—2295.
61. Robust speech watermarking by a jointly trained embedder and detector using a DNN / K. Pavlović [и др.] // Digital Signal Processing. — 2022. — Т. 122. — С. 103381.
62. WavMark: Watermarking for audio generation / G. Chen [и др.] // arXiv preprint arXiv:2308.12770. — 2023.
63. Proactive detection of voice cloning with localized watermarking / R. S. Roman [и др.] // arXiv preprint arXiv:2401.17264. — 2024.
64. SilentCipher: Deep audio watermarking / M. K. Singh [и др.] // arXiv preprint arXiv:2406.03822. — 2024.
65. AWARE: Audio Watermarking with Adversarial Resistance to Edits / K. Pavlović [и др.] // arXiv preprint arXiv:2510.17512. — 2025.
66. *Resemble AI.* Perth: Open Audio Watermarking Tool [Электронный ресурс] / Resemble AI. — URL: <https://github.com/resemble-ai/perth> (дата обращения: 01.10.2026).

67. Detecting voice cloning attacks via timbre watermarking / C. Liu [и др.] // arXiv preprint arXiv:2312.03410. — 2023.
68. VoiceMark: Zero-shot voice cloning-resistant watermarking approach leveraging speaker-specific latents / H. Li [и др.] // arXiv preprint arXiv:2505.21568. — 2025.
69. AudioMarkNet: Audio watermarking for deepfake speech detection / W. Zong [и др.] // 34th USENIX Security Symposium (USENIX Security 25). — 2025. — С. 4663—4682.
70. ItoV: Efficiently adapting deep learning-based image watermarking to video watermarking / G. Ye [и др.] // 2023 International Conference on Culture-Oriented Science and Technology (CoST). — IEEE. 2023. — С. 192—197.
71. Robust invisible video watermarking with attention / K. A. Zhang [и др.] // arXiv preprint arXiv:1909.01285. — 2019.
72. DVMark: A Deep Multiscale Framework for Video Watermarking / X. Luo [и др.] // IEEE Transactions on Image Processing. — 2025. — Т. 34. — С. 4371—4385.
73. A novel deep video watermarking framework with enhanced robustness to H. 264/AVC compression / Y. Zhang [и др.] // Proceedings of the 31st ACM international conference on multimedia. — 2023. — С. 8095—8104.
74. *Fridrich, J.* Image watermarking for tamper detection / J. Fridrich // Proceedings 1998 International conference on image processing. ICIP98 (Cat. No. 98CB36269). T. 2. — IEEE. 1998. — С. 404—408.
75. Quantization-based semi-fragile public-key watermarking for secure image authentication / M. Schlauweg [и др.] // Mathematics of data/image coding, compression, and encryption VIII, with applications. T. 5915. — SPIE. 2005. — С. 41—51.
76. *Wu, X.* SepMark: Deep separable watermarking for unified source tracing and deepfake detection / X. Wu, X. Liao, B. Ou // Proceedings of the 31st ACM international conference on multimedia. — 2023. — С. 1190—1201.
77. EditGuard: Versatile image watermarking for tamper localization and copyright protection / X. Zhang [и др.] // 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — IEEE. 2024. — С. 11964—11974.

78. OmniGuard: Hybrid manipulation localization via augmented versatile deep image watermarking / X. Zhang [и др.] // 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — IEEE. 2025. — С. 3008—3018.
79. *Kutter, M.* Watermark copy attack / M. Kutter, S. V. Voloshynovskiy, A. Herrigel // Security and Watermarking of Multimedia Contents II. Т. 3971. — SPIE. 2000. — С. 371—380.
80. Evaluation of security of ml-based watermarking: Copy and removal attacks / V. Kinakh [и др.] // 2024 IEEE International Workshop on Information Forensics and Security (WIFS). — IEEE. 2024. — С. 1—6.
81. *Wong, P. W.* A public key watermark for image verification and authentication / P. W. Wong // Proceedings 1998 international conference on image processing. ICIP98 (Cat. No. 98CB36269). Т. 1. — IEEE. 1998. — С. 455—459.
82. MetaSeal: Defending against image attribution forgery through content-dependent cryptographic watermarks / T. Zhou [и др.] // arXiv preprint arXiv:2509.10766. — 2025.
83. *Kutter, M.* Fair benchmark for image watermarking systems / M. Kutter, F. A. Petitcolas // Security and watermarking of multimedia contents. Т. 3657. — SPIE. 1999. — С. 226—239.
84. WAVES: Benchmarking the robustness of image watermarks / B. An [и др.] // arXiv preprint arXiv:2401.08573. — 2024.
85. Robust watermarking using generative priors against image editing: From benchmarking to advances / S. Lu [и др.] // International Conference on Learning Representations. — 2025.
86. MarkDiffusion: An open-source toolkit for generative watermarking of latent diffusion models / L. Pan [и др.] // Journal of Machine Learning Research. — 2026. — Т. 27, № 204. — С. 1—22.
87. *Facebook Research.* OmniSeal Bench: Benchmarking Tool for Multimodal Watermarking Models [Электронный ресурс] / Facebook Research. — URL: <https://github.com/facebookresearch/omnisealbench> (дата обращения: 01.10.2026).
88. Audiomarkbench: Benchmarking robustness of audio watermarking / H. Liu [и др.] // Advances in Neural Information Processing Systems. — 2024. — Т. 37. — С. 52241—52265.

89. A comprehensive real-world assessment of audio watermarking algorithms: Will they survive neural codecs? / Y. Özer [и др.] // arXiv preprint arXiv:2505.19663. — 2025.
90. VoxWatermark: A Large-Scale Benchmark for Audio Watermark Detection under Perturbations / F. Sedaghati [и др.] // arXiv preprint arXiv:2606.15187. — 2026.
91. DeepMark Benchmark: Redefining Audio Watermarking Robustness / S. Kovačević [и др.] // IEEE Access. — 2026. — Янв. — Т. 14. — С. 62031—62044.
92. *ISP RAS*. WIBE: Watermarks for Generated Images – Benchmarking & Evaluation [Электронный ресурс] / ISP RAS. — URL: <https://github.com/ispras/wibe> (дата обращения: 01.10.2026).
93. Image quality assessment: from error visibility to structural similarity / Z. Wang [и др.] // IEEE transactions on image processing. — 2004. — Т. 13, № 4. — С. 600—612.
94. The unreasonable effectiveness of deep features as a perceptual metric / R. Zhang [и др.] // 2018 IEEE/CVF conference on computer vision and pattern recognition. — IEEE. 2018. — С. 586—595.
95. Dreamsim: Learning new dimensions of human visual similarity using synthetic data / S. Fu [и др.] // arXiv preprint arXiv:2306.09344. — 2023.
96. *Wang, J.* Exploring clip for assessing the look and feel of images / J. Wang, K. C. Chan, C. C. Loy // Proceedings of the AAAI conference on artificial intelligence. Т. 37. — 2023. — С. 2555—2563.
97. LAION-5B: An Open Large-Scale Dataset for training next generation image-text models / C. Schuhmann [и др.] // Advances in neural information processing systems. — 2022. — Т. 35. — С. 25278—25294.
98. CLIPScore: A Reference-Free evaluation metric for image captioning / J. Hessel [и др.] // Proceedings of the 2021 conference on empirical methods in natural language processing. — 2021. — С. 7514—7528.
99. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation / J. Li [и др.] // International conference on machine learning. — PMLR. 2022. — С. 12888—12900.

100. Imagereward: Learning and evaluating human preferences for text-to-image generation / J. Xu [и др.] // *Advances in Neural Information Processing Systems*. — 2023. — Т. 36. — С. 15903—15935.
101. Gans trained by a two time-scale update rule converge to a local nash equilibrium / M. Heusel [и др.] // *Advances in neural information processing systems*. — 2017. — Т. 30.
102. Steganalysis on digital watermarking: Is your defense truly impervious? / P. Yang [и др.] // *arXiv preprint arXiv:2406.09026*. — 2024.
103. DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better / O. Kupyn [и др.] // *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. — 2019. — С. 8877—8886.
104. *Alam, I.* Saliency-aware diffusion reconstruction for effective invisible watermark removal / I. Alam, M. T. Islam, S. S. Woo // *Companion Proceedings of the ACM on Web Conference 2025*. — 2025. — С. 849—853.
105. Transferable Black-Box One-Shot Forging of Watermarks via Image Preference Models / T. Souček [и др.] // *Advances in Neural Information Processing Systems*. — 2025. — Т. 38. — С. 55967—55990.
106. Diffusion models for adversarial purification / W. Nie [и др.] // *arXiv preprint arXiv:2205.07460*. — 2022.
107. *Ho, C.-H.* DISCO: Adversarial defense with local implicit functions / C.-H. Ho, N. Vasconcelos // *Advances in neural information processing systems*. — 2022. — Т. 35. — С. 23818—23837.
108. A self-supervised approach for adversarial robustness / M. Naseer [и др.] // *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. — IEEE. 2020. — С. 259—268.
109. Multi-stage progressive image restoration / S. W. Zamir [и др.] // *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. — IEEE. 2021. — С. 14816—14826.
110. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data / X. Wang [и др.] // *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. — IEEE. 2021. — С. 1905—1914.

111. *Liang, H.* A baseline method for removing invisible image watermarks using deep image prior / H. Liang, T. Li, J. Sun // Transactions on machine learning research. — 2025.
112. Invisible image watermarks are provably removable using generative ai / X. Zhao [и др.] // Advances in neural information processing systems. — 2024. — Т. 37. — С. 8643—8672.
113. High-resolution image synthesis with latent diffusion models / R. Rombach [и др.] // 2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR). — IEEE. 2022. — С. 10674—10685.
114. *Black Forest Labs.* FLUX: Official Inference Repository for FLUX.1 Models [Электронный ресурс] / Black Forest Labs. — 2024. — URL: <https://github.com/black-forest-labs/flux> (дата обращения: 01.10.2026).
115. DiffusionDB: A Large-Scale Prompt Gallery Dataset for Text-to-Image Generative Models / Z. J. Wang [и др.] // Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers). — 2023. — С. 893—911.
116. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs / A. W. Rix [и др.] // 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221). Т. 2. — IEEE. 2001. — С. 749—752.
117. A short-time objective intelligibility measure for time-frequency weighted noisy speech / C. H. Taal [и др.] // 2010 IEEE international conference on acoustics, speech and signal processing. — IEEE. 2010. — С. 4214—4217.
118. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets / G. Mittag [и др.] // arXiv preprint arXiv:2104.09494. — 2021.
119. *Reddy, C. K.* DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors / C. K. Reddy, V. Gopal, R. Cutler // ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — IEEE. 2021. — С. 6493—6497.
120. *Desplanques, B.* ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification / B. Desplanques, J. Thienpondt, K. Demuynck // Interspeech 2020. — 2020. — С. 3830—3834.

121. High fidelity neural audio compression / A. Défossez [и др.] // arXiv preprint arXiv:2210.13438. — 2022.
122. High-fidelity audio compression with improved rvqgan / R. Kumar [и др.] // Advances in neural information processing systems. — 2023. — Т. 36. — С. 27980—27993.
123. MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement / S.-W. Fu [и др.] // Interspeech 2021. — 2021. — С. 201—205.
124. GTCRN: A speech enhancement model requiring ultralow computational resources / X. Rong [и др.] // ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). — IEEE. 2024. — С. 971—975.
125. *Siuzdak, H.* Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis / H. Siuzdak // International Conference on Learning Representations. — 2024.
126. LibriSpeech: An ASR Corpus based on public domain audio books / V. Panayotov [и др.] // 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). — IEEE. 2015. — С. 5206—5210.
127. *Yamagishi, J.* CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92) / J. Yamagishi, C. Veaux, K. MacDonald. — 2019.
128. Common voice: A massively-multilingual speech corpus / R. Ardila [и др.] // Proceedings of the twelfth language resources and evaluation conference. — 2020. — С. 4218—4222.
129. *Karpov, N.* Golos: Russian Dataset for Speech Research / N. Karpov, A. Denisenko, F. Minkin // Interspeech 2021. — 2021. — С. 1419—1423.
130. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline / H. Bu [и др.] // 2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA). — IEEE. 2017. — С. 1—5.
131. *Du, L.* Perceptual hashing for image authentication: A survey / L. Du, A. T. Ho, R. Cong // Signal Processing: Image Communication. — 2020. — Т. 81. — С. 115713.

132. *Zauner, C.* Implementation and Benchmarking of Perceptual Image Hash Functions : Master's thesis / Zauner Christoph. — Hagenberg, Austria : Upper Austria University of Applied Sciences, 2010.
133. *Johnson, D.* The elliptic curve digital signature algorithm (ECDSA) / D. Johnson, A. Menezes, S. Vanstone // International journal of information security. — 2001. — Т. 1, № 1. — С. 36—63.
134. *Bose, R. C.* On a class of error correcting binary group codes / R. C. Bose, D. K. Ray-Chaudhuri // Information and control. — 1960. — Т. 3, № 1. — С. 68—79.
135. *Novozámský, A.* IMD2020: A large-scale annotated dataset tailored for detecting manipulated images / A. Novozámský, B. Mahdian, S. Saic // 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW). — IEEE. 2020. — С. 71—80.
136. *Dalins, J.* PDQ & TMK+ PDQF—A Test Drive of Facebook's Perceptual Hashing Algorithms / J. Dalins, C. Wilson, D. Boudry // arXiv preprint arXiv:1912.07745. — 2019.
137. *Reed, I. S.* Polynomial codes over certain finite fields / I. S. Reed, G. Solomon // Journal of the society for industrial and applied mathematics. — 1960. — Т. 8, № 2. — С. 300—304.
138. *Baas, M.* Voice conversion with just nearest neighbors / M. Baas, B. van Niekerc, H. Kamper // arXiv preprint arXiv:2305.18975. — 2023.
139. Self voice conversion as an attack against neural audio watermarking / Y. Özer [и др.] // arXiv preprint arXiv:2601.20432. — 2026.
140. YouTube-VOS: A large-scale video object segmentation benchmark / N. Xu [и др.] // arXiv preprint arXiv:1809.03327. — 2018.
141. Learning Temporal Coherence via Self-Supervision for GAN-Based Video Generation / M. Chu [и др.] // ACM Transactions on Graphics (TOG). — New York, NY, USA, 2020. — Июль. — Т. 39, № 4.
142. Overview of the H.264/AVC Video Coding Standard / T. Wiegand [и др.] // IEEE Transactions on circuits and systems for video technology. — 2003. — Т. 13, № 7. — С. 560—576.

143. *Harris, F. J.* On the use of windows for harmonic analysis with the discrete Fourier transform / F. J. Harris // Proceedings of the IEEE. — 1978. — T. 66, № 1. — C. 51—83.

Список рисунков

1.1	Примеры маркирования изображения различными методами	20
2.1	Схема унифицированной экспериментальной оценки методов маркирования медиаконтента	47
2.2	Архитектура программной системы WIBE	52
2.3	Сравнение теоретической и эмпирической оценки TPR при различных уровнях FPR	62
2.4	Соотношение между емкостью, незаметностью и вычислительной эффективностью методов внедрения ЦВЗ в сформированные изображения	67
2.5	Средняя устойчивость исследованных методов маркирования к различным классам воздействий	68
2.6	Сравнительная эффективность атак с учетом величины перцептивных искажений	70
2.7	Сопоставление методов маркирования аудиоданных по емкости ЦВЗ, его незаметности и вычислительной эффективности реализованных алгоритмов	76
2.8	Среднее значение $TPR@10^{-3}FPR$ методов маркирования аудиоданных для различных классов воздействий	78
2.9	Сравнительная эффективность атак на методы маркирования аудиоданных с учетом величины искажений аудиосигнала	80
3.1	Схема метода CPW для контроля целостности изображений	88
3.2	Зависимость доли полностью совпадающих перцептивных хеш-кодов от интенсивности допустимых преобразований	98
3.3	Зависимость доли успешно верифицированных изображений от интенсивности допустимых преобразований	101
3.4	Зависимость доли выявленных нарушений целостности от площади модифицированной области изображения на наборе данных IMD2020 .	103
4.1	Схема метода CryptoMark для контроля целостности аудиоданных . . .	111
4.2	Зависимость качества локализации от доли модифицированного участка аудиозаписи	125

5.1	Влияние размера группы кадров на устойчивость извлечения и вычислительную эффективность предложенной методики	138
5.2	Нормированные коэффициенты временного распределения модификации в группе из $N = 30$ кадров	140
5.3	Зависимость BER от степени сжатия H.264 для покádрового и группового маркирования различными методами	142
5.4	Влияние нарушения временной синхронизации на извлечение ЦВЗ при размере группы $N = 30$	143

Список таблиц

1	Сравнение средств оценки методов маркирования изображений	41
2	Сравнение средств оценки методов маркирования аудиоданных	44
3	Метрики оценки методов маркирования изображений	58
4	Преобразования и атаки, используемые при оценке устойчивости методов маркирования изображений	63
5	Метрики оценки методов маркирования аудио	72
6	Преобразования и атаки, используемые при оценке устойчивости методов маркирования аудио	74
7	Сравнение незаметности различных вариантов CPW и метода MetaSeal	99
8	Оценка качества маркированных аудиоданных	120
9	Сравнение CryptoMark, SilentCipher и AudioSeal при допустимых преобразованиях и содержательно значимых модификациях	122
10	Сравнение стратегий переноса методов маркирования изображений на видеоданные	135
11	Влияние размера группы кадров на качество маркированного видео . .	137
12	Влияние временного распределения модификации на качество и устойчивость маркированного видео	140