Тема 16. Защита визуального контента маркетплейсов путем стеганографии в цифровых изображениях

Аналитический обзор методов стеганографии в цифровых изображениях

Благодаря цифровым технологиям и интернету появилось много цифрового контента, и защита его от неправомерного копирования актуальна как никогда. Встраивание цифровых водяных знаков (ЦВЗ) - один из способов идентификации владельца медиаконтента.

Обобщенно, процесс работы с ЦВЗ состоит из двух частей: встраивание ЦВЗ и извлечение ЦВЗ. На этапе встраивания алгоритм добавляет невидимую метку в оригинальное изображение, после чего изображение, защищенное ЦВЗ, может использоваться в открытых источниках, например в социальных сетях или сайтах. В случае необходимости определения правообладателя изображения алгоритм извлечения получает ЦВЗ из изображения. Между этапами встраивания ЦВЗ и извлечением ЦВЗ изображение может подвергнуться изменениям. Это могут быть изменения, связанные с передачей информации такие, как шумы и компрессия. Также изменения могут иметь злонамеренный характер, когда изображение подвергается геометрическим искажениям.

Классические (традиционные) подходы используют эвристические методы для встраивания информации в носитель. Эти эвристики определяют каким образом изменить пиксели изображения. Недостатком классических подходов является фиксированный набор атак, которому они могут противостоять, а добавление прочности против нового вида помех требует пересмотра всего алгоритма. Нейросетевые подходы, наоборот, более гибкие в этом плане и для добавления устойчивости к новому искажению достаточно добавить это искажение при тренировке сети.

Сверточные нейронные сети (СНС), благодаря своей способности извлекать представления изображений, получили широкое распространение при решении задач с изображениями. Не осталась в стороне и область встраивания ЦВЗ, где применение автоэнкодеров на основе СНС (например, ReDMark [1], DNN [2]) подняло качество восстановления ЦВЗ на новый уровень по сравнению с классическими подходами.

Следующим этапом развития методов встраивания ЦВЗ стало появление генеративно-состязательных сетей (ГСС). HiDDeN [3] была первой моделью, которая совмещала в себе состязательный подход и СНС, что и сделало ее передовым решением. Далее было предложено большое количество модификаций этой модели. В [4] авторы модифицируют HiDDeN, добавляя механизм внимания и модуль понижения размерности для сообщения. Модель HiDDeN имеет недостаток в виде условия дифференцируемости

искажения, которое используется в обучении, поэтому в статье [5] предлагается двухступенчатый подход, где на первом этапе обучается энкодер без воздействия искажений, а на втором этапе обучается декодер уже с добавлением искажений в процесс обучения. В [6] исследуется возможность применения HiDDeN при извлечении метки после вращения изображения. Добавить кодирование каналов предлагается добавить в статье [7], кроме того, в статье исследуется устойчивость модели к атакам, которые не участвовали при обучении. В [8] и [9] предлагается изменить представление метки перед встраиванием в изображение.

Кроме HiDDeN с его модификациями представлены и другие архитектуры. Если говорить о ГСС, то модель ABDH [10] является примером применения известной архитектуры CycleGAN [11] для встраивания ЦВЗ.

Не остались в стороне и достижения в области самостоятельного обучения (self-supervised learning) – алгоритм встраивания метки в представление, полученное с использование нейросетей, которые обучены методом самостоятельного обучения, предлагается в [12].

В имеющихся обзорных статьях [13,14] основное внимание уделяется качеству соответствия изображения со встроенным ЦВЗ исходному изображению, нежели точности восстановления ЦВЗ. Алгоритм как правило считается устойчивым против определенной атаки, если эта атака была использована при обучении и алгоритм смог адаптироваться к ней, но все еще показывает большую ошибку восстановления метки. Для практического применения ЦВЗ в целях защиты авторских прав не менее важным аспектом является точность восстановления ЦВЗ из изображения, которое подверглось искажениям.

Авторы в [13] при сравнении качества восстановления ЦВЗ различными моделями используют результаты, представленные в статьях с описанием этих моделей. Как верно замечают авторы, эти результаты были получены с помощью разных методик: использовались различные по размеру изображения, различные по длине метки и различные искажения (а также параметры искажений). К сожалению, проведение честного сравнения выглядит затруднительным, так как у большинства моделей, рассмотренных в этом исследовании, отсутствует открытая реализация (см. таблицу 16.1).

Таблица 16.1 - Доступность исходного нейросетевых подходов встраивания ЦВЗ

Модель	Год	Наличие исходного кода
ReDMark [1]	2018	нет
HiDDeN [3]	2018	есть
Two-Stage [5]	2019	есть
IGA [4]	2020	есть

DA [7]	2020	нет
Rotation [6]	2020	нет
DNN [2]	2020	нет
Double Detector-Discriminator [9]	2020	нет
Spatial-Spread [8]	2020	нет
ABDH [10]	2020	нет
SSLS [12]	2022	есть

Большинство рассмотренных методов обучалось и тестировалось на известном наборе изображений COCO [15]. Изображения для маркетплейсов отличаются от обычных фотографий наличием однотонного фона, что может оказать негативное влияние на встраивание ЦВЗ. Кроме того, при решении задачи защиты контента маркетплейсов наиболее интересна устойчивость против геометрических атак, в то время как имеющиеся методы уделяют большое внимание различным шумам.

Выбор методологии

Так как в качестве задачи стоит разработка метода встраивания цифровых водяных знаков в изображения товаров маркетплейсов, то, как уже было сказано выше, наибольший интерес вызывает устойчивость такого метода защиты к геометрическим аугментациям защищенных изображений. В связи с чем нами был определен набор аугментация/шума, относительно которого будет проводиться обучение и тестирование моделей. Набор аугментаций и их параметров перечислен ниже в таблице 16.2.

Таблица 16.2 – Выбранные аугментации и их параметры.

Аугментация	Описание	Параметры
Identity	В этом случае изображение с встроенной меткой никак не изменяется. Использование не измененного изображения в качестве атаки при обучение призвано предотвратить деградацию точности при извлечения водяного знака.	
Crop	Данный вид атаки обрезает изображение с защитной меткой до части меньшего размера.	Соотношение размеров исходного и обрезанного изображения, выбирается случайно из диапазона [0.25, 1].
Cropout	Данный вид атаки используется для комбинирования части изображения с защищенной меткой и оставшейся части без него.	Соотношение размера исходного изображения и размера части изображения со встроенной меткой, выбирается случайно из диапазона [0.25, 1].

Dropout	Данный вид атаки используется для комбинирования части изображения с защищенной меткой и оставшейся части без него. В отличии от атаки Cropout, часть изображения с защищенной меткой в этом случае представляет собой не определенный вырезанный прямоугольный кусок защищенного изображения, а случайно выбранные пиксели такого изображения. Соответственно, оставшиеся пиксели представляют собой пиксели незащищенного изображения.	Соотношение пикселей защищенного изображения к количеству всех пикселей изображения, выбирается случайно из диапазона [0.75, 1].
Jpeg	При данной атаки к исходному защищенному изображению применятся jpeg-компрессия.	
Rotate	Данная атака используется для поворота защищенного изображения на определенный угол.	Угол поворота в градусах, выбирается случайно из диапазона $[-15^o, 15^o]$.
Hflip	Данная атака используется для зеркального отображения защищенного изображения по горизонтали.	

В качестве основы для разрабатываемой модели выла выбрана архитектура сверточного энкодера-декодера с состязательным подходом в силу того, что модели, использующие такую архитектуру, на сегодняшний день являются передовыми в данной области. Несмотря на результативность таких моделей, нами были введены некоторые изменения в архитектуру самой сети, также были добавлены дополнительные компоненты. Схема разработанного модуля представлена на рисунке 16.1.

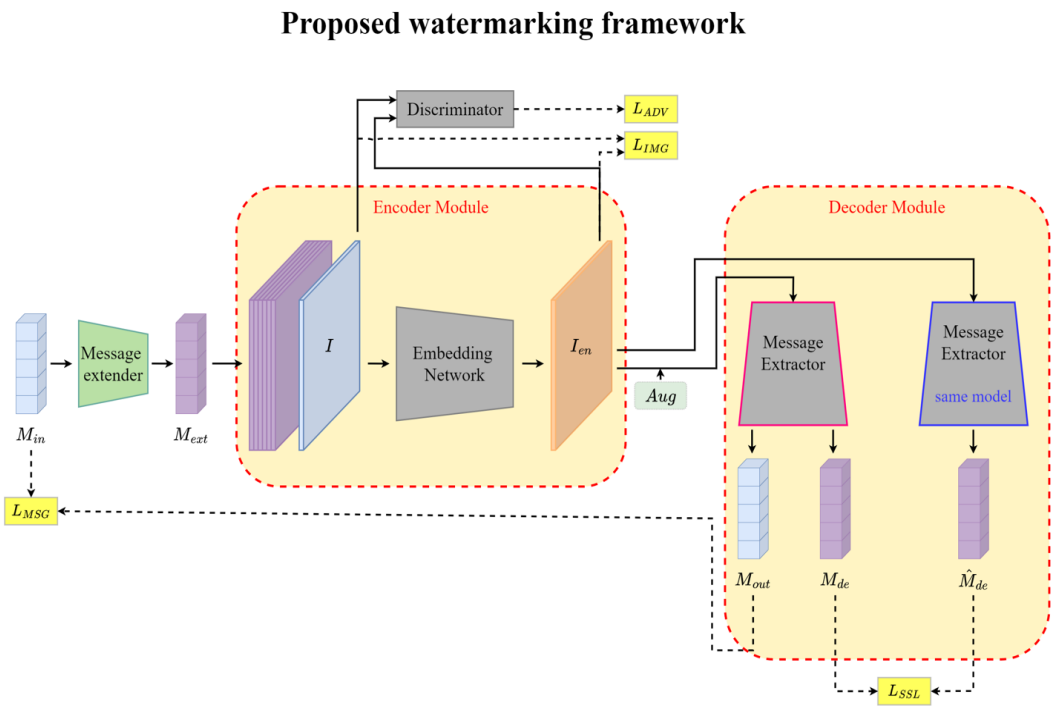

Рисунок 16.1 – Схема разработанного модуля.

Основные изменения коснулись строения энкодера. Хотя авторы, использующие данную архитектуру [3–5,7], и утверждают, что использование последовательных блоков, состоящих из блока свертки, пакетной нормализации [16] и функции активации [17], является зарекомендовавшим себя решением, все же это не уберегает их модели от деградации точности по причине потенциальной высокой симметрии пространства параметров модели. Поэтому нами были добавлены остаточные связи [18] на каждом счетном блоке, а также доступ к первоначальному цифровому изображению и встраиваемому сообщению на каждом нечетном. Архитектура такого энкодера из семи блоков представлена на рисунке 16.2.

Другой отличительной чертой стало добавление дополнительной составляющей целевой функции (на рис. 16.1 обозначена как L_{ssl}), которая отвечает за то, чтобы извлеченное сообщение с аугментированного изображения походила на сообщение, извлеченное с оригинального защищенного изображения. Делается это путем дистилляции знаний, используемой часто в подходах самообучения [19].

Использование дискриминатора как соревновательной составляющей также обсуждалась в работах ранее [3] и хорошо себя зарекомендовало, однако, его устоявшаяся архитектура, широко использовавшаяся в таких моделях, была склонна к коллапсу, то есть быстрому обнулению ошибки дискриминатора при обучении. Решением этому послужило избавление от смещения в блоках дискриминатора, как это часто делается на практике [20].

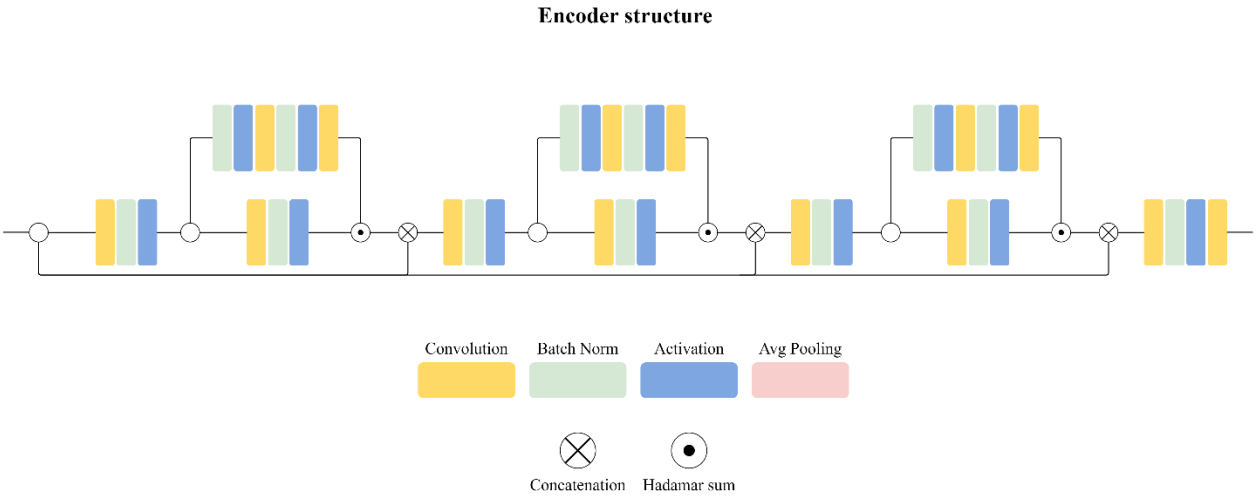

Рисунок 16.2 – Строение модифицированного энкодера.

Экспериментальное исследование

Было проведено экспериментальное исследование разработанного подхода, а также сравнение его с существующими решениями. Исследование проводилось на следующих наборах данных: «DIV2k» и наборе данных товаров маркетплейсов.

Набор данных «DIV2k» [21] состоит из 800 фотографий для обучения и 100 для валидации. Фотографии имеют высокое разрешение, но для упрощения вычислений и сравнения моделей, разрешение было понижено до 128 на 128 пикселей.

Набор данных товаров маркетплейсов [22] представляет собой набор фотографий товаров маркетплейсов, размещенных на белом фоне. Всего в наборе 44000 фотографий, для тренировки было случайно отобрано 80% набора, для тестирования 20%.

В процессе обучения и тестирования в цифровые изображения поочередно встраивалась метка длиной 30, 64 и 90 бит, представляющая собой бинарное сообщение. Защищенное изображение изменялось в соответствии с перечисленными выше аугментациями, затем с такого сообщения извлекалось зашифрованное сообщение. Метрикой являлись значения точности извлеченного сообщения, а для изображения, зашифрованного меткой, вычислялось значение PSNR между ним и исходным изображением.

Метрика точности (accuracy) для извлеченного сообщения из изображений разных наборов данных, указаны в сравнительной таблице 16.3. Рассматриваемые модели были обучены с применением указанных выше аугментациям, а потом протестированы относительно каждой аугментации отдельно (столбцы Identity, Crop, Cropout, Dropout, Jpeg, Rotate, Hflip) и ко всем аугментациям (столбец CN). Наилучшая метрика для каждого значения длины сообщение выделена жирным.

Сравнение количества параметров рассматриваемых в исследовании моделей отображены в таблице 16.4.

Таблица 16.3 – Точность извлечения сообщения.

Датасет	Модель	Длина сообщения	Identity	Crop	Cropout	Dropout	Jpeg	Rotate	Hflip	CN
DIV2k	Ours	30	**0.972**	**0.972**	**0.913**	**0.963**	0.649	**0.965**	**0.973**	**0.872**
	HiDDen		0.846	0.836	0.828	0.82	**0.784**	0.84	0.843	0.839
	IGA		0.842	0.847	0.747	0.723	0.755	0.818	0.845	0.838
	Ours	64	**0.752**	**0.755**	**0.728**	**0.738**	0.588	**0.729**	**0.742**	**0.719**
	HiDDen		0.719	0.698	0.672	0.667	**0.672**	0.714	0.69	**0.719**
	IGA		0.679	0.679	0.62	0.612	0.614	0.679	0.679	0.675
	Ours	90	0.628	0.63	0.601	0.601	0.535	0.634	0.617	0.591
	HiDDen		**0.657**	**0.643**	**0.608**	**0.638**	0.611	0.64	**0.644**	0.638
	IGA		0.654	0.636	0.599	0.603	**0.62**	**0.648**	0.641	**0.643**
Fashion	Ours	30	**0.88**	**0.781**	**0.716**	**0.748**	**0.773**	**0.828**	**0.825**	**0.79**
	HiDDen		0.686	0.65	0.631	0.638	0.671	0.667	0.686	0.662
	IGA		0.676	0.653	0.499	0.504	0.651	0.673	0.68	0.622
	Ours	64	**0.752**	**0.755**	**0.728**	**0.738**	0.588	**0.729**	**0.742**	**0.714**
	HiDDen		0.63	0.594	0.572	0.598	**0.604**	0.604	0.622	0.604
	IGA		0.557	0.55	0.5	0.504	0.558	0.552	0.556	0.54
	Ours	90	**0.628**	**0.63**	**0.601**	**0.601**	0.535	**0.634**	**0.617**	**0.591**
	HiDDen		0.576	0.571	0.554	0.557	**0.558**	0.559	0.567	0.563
	IGA		0.518	0.522	0.501	0.506	0.519	0.52	0.517	0.517

Таблица 16.4 – Количество параметров модели.

Модель	Длина сообщения	Количество параметров
Ours	30	697.161
	64	801.779
	90	883.341
HiDDen	30	488.000
	64	530.500
	90	564.560
IGA	30	488.930
	64	512.866
	90	532.730

Также были проведены замеры динамики обучения каждой модели для каждого набора данных. Динамика метрики accuracy для данных из набора «DIV2k» представлена на рисунке 16.3, для маркетплейсов 16.4. Первая строка представляет собой точность восстановления метки для тренировочных данных, нижняя для тестовых в зависимости от эпохи. Вместе с тем, на рисунках 16.5 и 16.6 отображены динамики изменения метрики PSNR в зависимости от эпохи для тестовых выборок наборов данных «DIV2k» и товаров маркетплейсов соответственно.

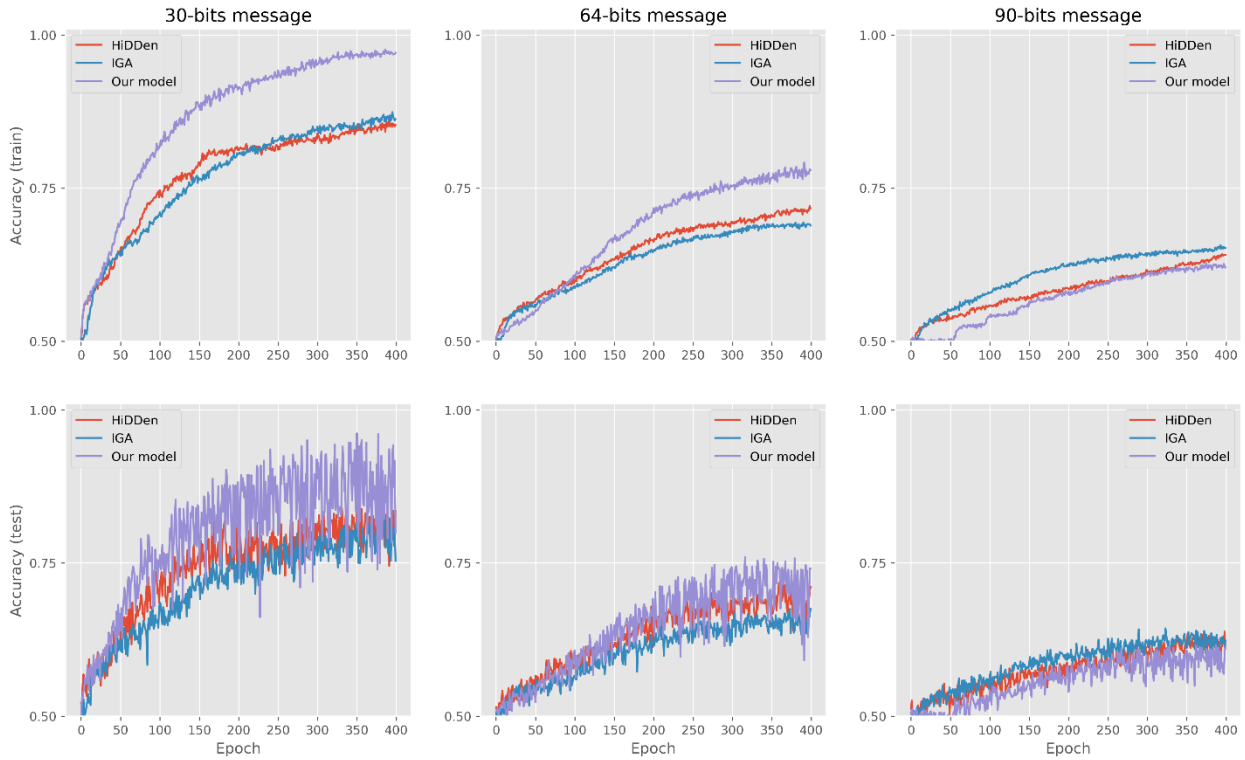

Рисунок 16.3 – Динамика метрики accuracy моделей для набора «DIV2k».

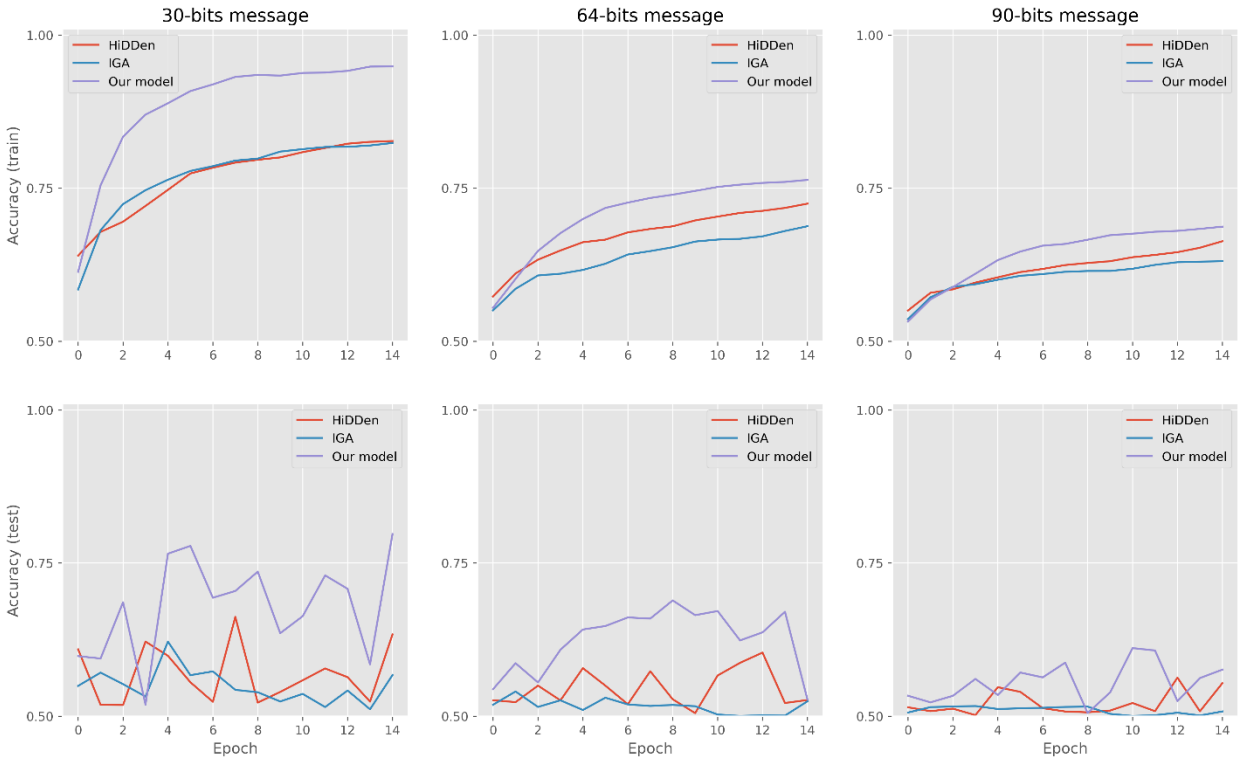

Рисунок 16.4 – Динамика метрики accuracy моделей для набора «DIV2k».

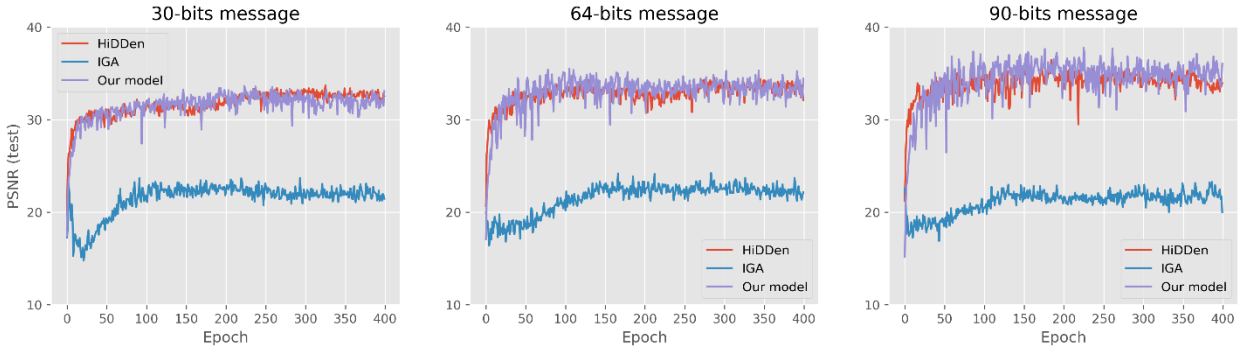

Рисунок 16.5 – Динамика метрики PSNR моделей для набора «DIV2k».

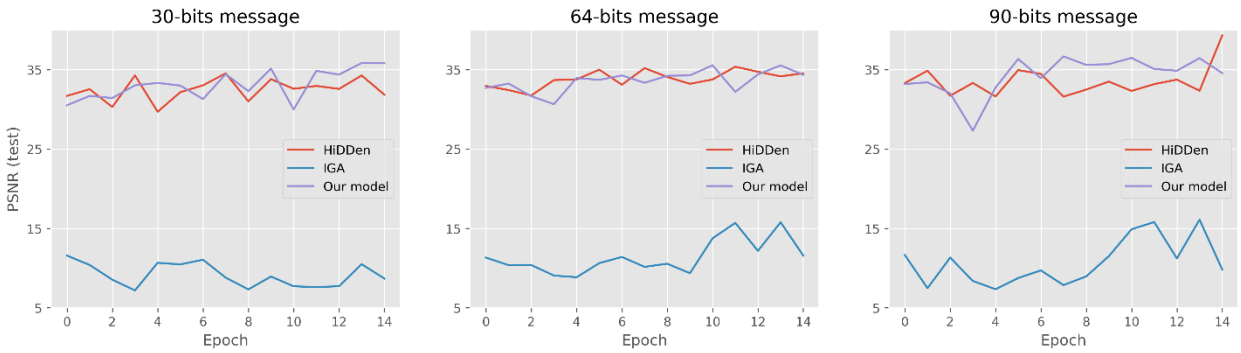

Рисунок 16.6 – Динамика метрики PSNR моделей для набора «DIV2k».

Как видно из результатов экспериментов, разработанный метод имеет лучшую сходимость, быструю обучаемость, по сравнению с рассмотренными моделями, а также лучше подходит не только для защиты изображений маркетплейсов, которые имеют

определенною специфику, но и для изображений без данной специфики в большинстве случаев. Разработанный метод также имеет лучшую устойчивость к аугментациям практически во всех рассмотренных случаях.

Заключение

В результате проделанной работы был проведен аналитический обзор существующих методов стеганографии и встраивания водных знаков в цифровых изображениях, где были рассмотрены существующие решения в этой области, и оценены возможности применения существующих решений к задаче защите контента маркетплейсов.

Рассмотренные методы на основе сверточных нейронных сетей легли в основу разработанной архитектуры модуля, позволяющего встраивать бинарное сообщение в цифровое сообщение и извлекать его, в том числе после применения геометрических преобразований к защищенному изображению.

Также было проведено экспериментальное исследование, направленное на оценку качества работы разработанного метода, а именно: оценена точность восстановления метки относительно выделенных аугментаций с помощью метрики accuracy, оценена похожесть изображений с защитной меткой на исходные изображения с помощью метрики PSNR.

Список источников

1. Ahmadi M. et al. ReDMark: Framework for Residual Diffusion Watermarking on Deep Networks. 2018.
2. Zhong X. et al. An Automated and Robust Image Watermarking Scheme Based on Deep Neural Networks. 2020.
3. Zhu J. et al. HiDDeN: Hiding Data With Deep Networks. 2018.
4. Zhang H. et al. Robust Data Hiding Using Inverse Gradient Attention. 2020.
5. Liu Y. et al. A novel two-stage separable deep learning framework for practical blind watermarking // MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia. Association for Computing Machinery, Inc, 2019. P. 1509–1517.
6. Hamamoto I., Kawamura M. Neural Watermarking Method Including an Attack Simulator against Rotation and Compression Attacks // IEICE Trans Inf Syst. The Institute of Electronics, Information and Communication Engineers, 2020. Vol. E103.D, № 1. P. 33–41.
7. Luo X. et al. Distortion Agnostic Deep Watermarking // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2020. P. 13545–13554.

8. Plata M., Syga P. Robust Spatial-spread Deep Neural Image Watermarking // Proceedings - 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2020. Institute of Electrical and Electronics Engineers Inc., 2020. P. 62–70.

9. Plata M., Syga P. Robust watermarking with double detector-discriminator approach. 2020.

10. Yu C. Attention Based Data Hiding with Generative Adversarial Networks // Proceedings of the AAAI Conference on Artificial Intelligence. AAAI press, 2020. Vol. 34, № 01. P. 1120–1128.

11. Zhu J.-Y. et al. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. 2017. P. 2223–2232.

12. Fernandez P. et al. Watermarking Images in Self-Supervised Latent Spaces // ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. Institute of Electrical and Electronics Engineers Inc., 2021. Vol. 2022-May. P. 3054–3058.

13. Byrnes O. et al. Data Hiding with Deep Learning: A Survey Unifying Digital Watermarking and Steganography. 2021.

14. Wan W. et al. A comprehensive survey on robust image watermarking // Neurocomputing. Elsevier, 2022. Vol. 488. P. 226–247.

15. Lin T.Y. et al. Microsoft COCO: Common Objects in Context // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, 2014. Vol. 8693 LNCS, № PART 5. P. 740–755.

16. Ioffe S., Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. PMLR, 2015. P. 448–456.

17. Agarap A.F. Deep Learning using Rectified Linear Units (ReLU). 2018.

18. He K. et al. Deep Residual Learning for Image Recognition. 2016. P. 770–778.

19. Caron M. et al. Emerging Properties in Self-Supervised Vision Transformers // Proceedings of the IEEE International Conference on Computer Vision. Institute of Electrical and Electronics Engineers Inc., 2021. P. 9630–9640.

20. Radford A., Metz L., Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks // 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings. International Conference on Learning Representations, ICLR, 2015.

21. Agustsson E., Timofte R. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. 2017. P. 126–135.

22. Fashion Product Images Dataset | Kaggle [Electronic resource]. URL: https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset (accessed: 14.12.2022).