УДК 003.26
DOI: 10.33764/2618-981X-2019-6-2-159-168

ПРИМЕНЕНИЕ ЦИФРОВЫХ ВОДЯНЫХ ЗНАКОВ ДЛЯ ЗАЩИТЫ ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

Анжелика Викторовна Печенкина
Сибирский государственный университет геосистем и технологий, 630108, Россия, г. Новосибирск, ул. Плахотного, 10, магистрант, e-mail: angelika19z78@gmail.com

Игорь Николаевич Карманов
Сибирский государственный университет геосистем и технологий, 630108, Россия, г. Новосибирск, ул. Плахотного, 10, кандидат технических наук, доцент, зав. кафедрой информационной безопасности, тел. (903)937-24-90, e-mail: i.n.karmanov@ssga.ru

На данный момент, в связи с развитием глобальных сетей, огромную значимость принимает защита интеллектуальной собственности от незаконного копирования. Множество исследований посвящено использованию стеганографии в качестве защиты авторских прав. Стеганографические методы не только скрытно передают информацию, но и решают задачи помехоустойчивости аутентификации, защиты данных от несанкционированного копирования, отслеживания перемещения информации в сети и для поиска ее в базах мультимедийных данных. Большинство областей прикладной математики используются как инструмент стеганографии, а термин «стего» давно вошел в обиход пользователей сети Интернет как важная составляющая современных технологий информационной безопасности.

Ключевые слова: информационная безопасность, стеганография, стеганографические методы, цифровая стеганография, передача данных, цифровые водяные знаки, интеллектуальная собственность, мультимедиа, глубокие нейронные сети.

APPLICATION OF DIGITAL WATERMARKS FOR INTELLECTUAL PROPERTY PROTECTION

Angelika V. Pechyonkina
Siberian State University of Geosystems and Technologies, 10, Plakhotnogo St., Novosibirsk, 630108, Russia, Graduate, e-mail: angelika19z78@gmail.com

Igor N. Karmanov
Siberian State University of Geosystems and Technologies, 10, Plakhotnogo St., Novosibirsk, 630108, Russia, Ph. D., Associate Professor, Head of Department of Information Security, phone: (903)937-24-90, e-mail: i.n.karmanov@ssga.ru

At the moment, due to the development of global networks, protection of intellectual property from illegal copying is of great importance. A lot of research is devoted to use of steganography for copyright protection. Steganographic methods not only secretly transmit information, but also solve problems of authentication noise immunity, protecting data from unauthorized copying, tracking information movement in a network and for searching it in multimedia databases. Most areas of applied mathematics are used as a tool for steganography, and the term «stego» has long been used by Internet users as an important component of modern information security technologies.

Key words: information security, steganography, steganographic methods, digital steganography, data transfer, digital watermarks, intellectual property, multimedia, deep neural networks.

Введение

В цифровой стеганографии выделяются следующие направления:

1) сокрытая передача встроенной информации;
2) цифровые водяные знаки;
3) идентификационные номера;
4) встроенные заголовки.

В данной статье рассматриваются цифровые водяные знаки (ЦВЗ). Информация, представленная в цифровом виде очень уязвима, из-за стремительного развития технологий мультимедиа. Фотографу, музыканту, художнику, создающему свои произведения, хочется быть уверенным, что никто другой не присвоит себе их, не станет копировать, с целью продать и заработать на чужом таланте. А информация подобного рода представлена по всей сети, мы с легкостью делимся фотографиями, видео в социальных сетях, не задумываясь, насколько они уязвимы. Никто не использует наши фото, потому что они, скорее всего, не представляют коммерческой выгоды, но задумайтесь об этом, когда создадите свой шедевр. Поэтому как никогда актуальна разработка различных методов защиты информации подобного рода.

Методы внедрения и декодирования ЦВЗ

ЦВЗ – это способ скрытия секретной информации в цифровых носителях для защиты права собственности на эти медиаданные. Существует множество подходов, чтобы сделать водяной знак эффективным и стойким к атакам устранения. Алгоритмы ЦВЗ в пространственной области внедряют секретные данные, напрямую манипулируя пикселями в изображении. Например, LSB (наименее значимый бит) пикселей обычно используется для размещения секретной информации. Однако, такие методы уязвимы для атак и чувствительны к шуму и традиционным методам обработки сигналов. По сравнению с методами пространственной области более широко применяются методы частотной области, которые встраивают водяные знаки в спектральные коэффициенты изображения. Наиболее часто используемыми преобразованиями являются дискретное косинусное преобразование (DCT), дискретное преобразование Фурье (DFT), дискретное Вейвлет-преобразование (DWT) и их сочетания.

Самим ЦВЗ может быть аутентичный код, информация об авторе. ЦВЗ могут быть как видимыми, так и невидимыми. До того, как встроить информацию в качестве ЦВЗ, необходимо преобразовать ее в двумерный массив бит. Один из этапов предварительной обработки сокрытого сообщения − это вычисление обобщенного преобразования Фурье, позволяющее встроить ЦВЗ в спектральной области, тем самым повысить помехоустойчивость. Применение ключей при предварительной обработке повышает секретность встраивания информации. Далее происходит само внедрение ЦВЗ. Для этого часто используют стегоключ – псевдослучайную последовательность бит, полученную от опреде-

ленного генератора. ЦВЗ практически незаметны для человека, благодаря большой психовизуальной избыточности изображений для восприятия [1–4].

Основные составляющие стегосистемы ЦВЗ представлены на рис 1.

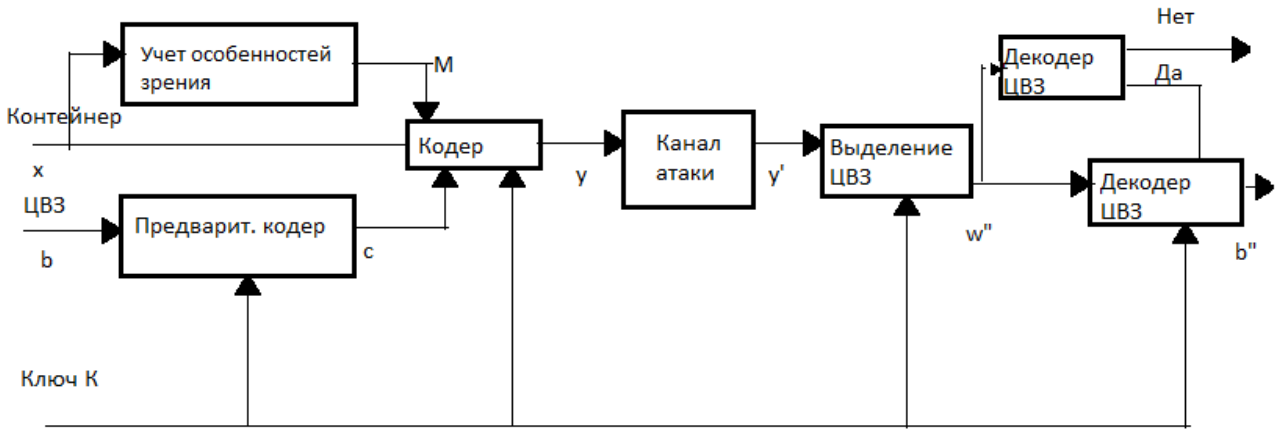

Рис. 1. Схема стегосистемы ЦВЗ

Элементы схемы на рис. 1 выполняют следующие функции:

– предварительный кодер служит для преобразования скрываемого сообщения к удобному виду, предназначенному для встраивания в сигнал-контейнер – информационную последовательность со спрятанным сообщением;

– стегокодер необходим для процесса вложения скрытого сообщения в другие информационные данные, учитывая их особенности;

– при выделении ЦВЗ осуществляется выделение встроенного сообщения из общей информации;

– детектор ЦВЗ определяет наличие встроенного сообщения;

– декодер ЦВЗ восстанавливает сокрытое сообщение.

В такой системе объединяются два типа информации: открытая информация, доступная всем, и скрытое сообщение (чаще всего невидимое). И эту информацию возможно различить только специальными детекторами. Таким детектором может быть система выделения ЦВЗ, а другим детектором является сам человек.

Рассмотрим математическую модель стегосистемы ЦВЗ. Алгоритм встраивания состоит из трех этапов:

1) генерации ЦВЗ;

2) встраивания ЦВЗ в кодере;

3) обнаружения ЦВЗ в детекторе.

Первый этап – генерация ЦВЗ.

Пусть W^*, K^*, I^*, B^* – множества возможных цифровых водяных знаков, ключей, контейнеров и скрытых сообщений соответственно. Тогда представляем генерацию ЦВЗ в виде функции:

$$F: I^* \times K^* \times B^* \to W^*, \quad W = F\ (I, K, B), \tag{1}$$

где W, K, I, B – элементы множеств.

В данном случае сама функция F может быть произвольной, но чаще всего на нее накладываются ограничения. Необходимо, чтобы выполнялось следующее требование:

$$F\ (I, K, B) \approx F\ (I + \varepsilon, K, B), \tag{2}$$

означающее, что незначительные изменения контейнера не приводят к изменению ЦВЗ.

Функция F является составной:

$$F = T \cup G, \text{ где } G: K^* \times B^* \to C^* \text{ и } T: C^* \times I^* \to W^*, \tag{3}$$

соответственно, цифровой водяной знак зависит от свойств контейнера.

Функция G может быть реализована с помощью криптографически безопасного генератора псевдослучайных последовательностей, где K – начальное значение. Функция T преобразует кодовые слова C*, в результате получается цифровой водяной знак W*. Тут нет необходимости в ограничениях, потому что выбор G влечет за собой необратимость F. Но в этом случае функцию T нужно выбрать так, чтобы незаполненный контейнер I_{o}, заполненный контейнер I_{W} и немного измененный заполненный контейнер I'_{w} порождали бы один и тот же ЦВЗ:

$$T(C, I_o) = T(C, I_w) = T(C, I'_{w'}), \tag{4}$$

т. е. T должна быть устойчива к незначительным изменениям.

Второй этап – встраивание ЦВЗ.

Данный процесс производится в кодере, его можно описать как суперпозицию двух сигналов:

$$I_w(i, j) = I_o\ (i, j) \oplus L\ (i, j)\ W\ (i, j)\ p\ (i, j), \tag{5}$$

где $W(i,j)$ – ЦВЗ;

$I_{\text{o}}\ (i,j)$ – изначальное сообщение;

$L(i,j)$ – маска встраивания ЦВЗ, предназначенная для уменьшения заметности, также данная маска учитывает особенности детектора ЦВЗ;

$p(i,j)$ – проектирующая функция, зависит от ключа;

$\oplus$ – оператор суперпозиции.

Функция проектирования распределяет ЦВЗ по всей области изображения. Ее использование возможно представить в виде разнесения информации по параллельным каналам. Она имеет определенную структуру и свойства, предназначенные для противостояния атакам. Иное описание процесса внедрения

цифровых водяных знаков можно представить, как стегосистему для передачи дополнительной информации (рис. 2).

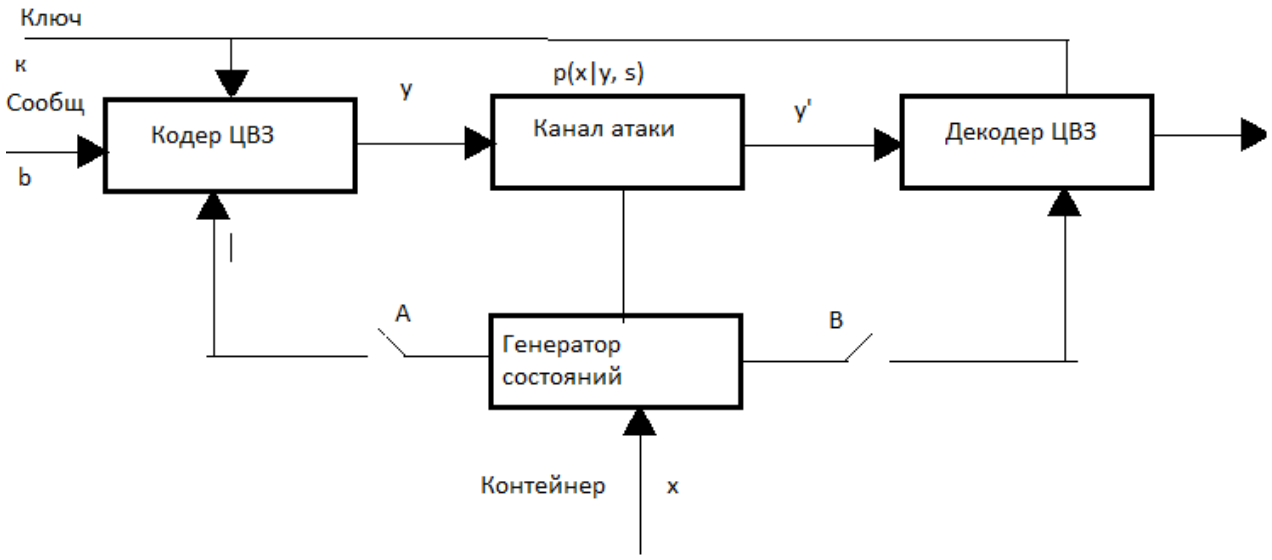

Рис. 2. Стегосистема для передачи дополнительной информации

В такой системе кодер и декодер имеют доступ, помимо ключа, к полной информации о канале, контейнере и возможных атаках на них. Степень доступа регулируется с помощью переключателей *A* и *B*. В зависимости от их положения можно выделить четыре класса стегосистем.

1-й класс: нет дополнительной информации, при этом переключатели разомкнуты (классическая стегосистема). Обнаружение ЦВЗ происходит с помощью вычисления коэффициента корреляции между принятым стегосообщением и вычисленным по ключу ЦВЗ. В данном случае превышение коэффициента означает присутствие ЦВЗ. Это самая неэффективная система.

2-й класс: информация о канале известна только кодеру, в этом случае *A* замкнут, *B* разомкнут. Данная система имеет ту же пропускную способность, как схема с наличием связи исходного контейнера с декодером. Но, к сожалению, данная система имеет высокую сложность организации кодера, необходимо построить кодовую книгу для каждого сообщения. Еще один существенный недостаток – невозможность обнаружения атак. В этом случае используют структурированные кодовые книги, чтобы снизить сложность кодера.

3-й класс: присутствует дополнительная информация, известная только декодеру, при этом A разомкнут, *B* замкнут. В данном случае декодер строится с учетом всех возможных атак, поэтому он достаточно помехоустойчив. При этом используют опорный ЦВЗ. Например, можно выполнить встраивание в амплитудные коэффициенты преобразования Фурье, они инвариантны к аффинным преобразованиям. В этом случае опорный ЦВЗ отразит, какое преобразование выполнил со стегосообщением атакующий.

4-й класс: присутствует дополнительная информация, которая известна кодеру и декодеру, в данном случае и A, и B замкнуты. Все передовые стегосистемы строятся по этому принципу. При этом максимальный эффект достигнут путем согласования кодера с сигналом-контейнером, а также адаптивным управлением декодером в условиях наблюдения канала атак.

Третий этап – обнаружение ЦВЗ.

Тут все зависит от типа стегодетектора, возможна выдача двоичных либо М-ичных решений о наличии или отсутствии ЦВЗ. В первом случае детектор называется «жестким», во втором – «мягким». Рассмотрим более простой случай «жесткого» детектора. Обозначаем операцию детектирования через функцию D, тогда:

$$D:\ L_w^* \times K^* \to \{0,1\},$$

$$D(L_w, W) = D(L_w, F(L_w, K)) = \begin{cases} 1, \text{ если } W \text{ есть} \\ 0, \text{ если } W \text{ нет} \end{cases}. \quad (6)$$

В данном случае детектор – корреляционный приемник. Рассмотрим, как будет работать корреляционный детектор на примере растрового изображения. Пусть у половины пикселов изображения значение яркости увеличено на 1, а у остальных не подвергалось изменениям или уменьшилось на 1. Тогда

$$I_W = I_0 + W \text{, где } F(I_0, K) = W.$$

Коррелятор детектора цифровых водяных знаков вычисляет величину

$$I_W \cdot W = (I_0 + W) \cdot W = I_0 \cdot W + W \cdot W.$$

Так как W может принимать значения ±1, то $I_0 \cdot W$ будет маленьким числом, а $W \cdot W$ – всегда положительно. Поэтому $I_W \cdot W$ будет приближено к $W \cdot W$. Вероятность неверного обнаружения стегосообщения может быть записана как дополнительная функция ошибок от корня квадратного из отношения $W \cdot W$ («энергии сигнала») к дисперсии значений пикселов яркости («энергии шума»).

В литературе описано большое количество различных методов встраивания ЦВЗ в файлы мультимедиа, физические и электронные документы, а также в реляционные базы данных в целях защиты интеллектуальной собственности [5–8].

Применение ЦВЗ для защиты авторских прав на нейросети

Одним из наиболее перспективных направлений применения ЦВЗ является защита авторских прав на глубокие нейронные сети (Deep Neural Networks, DNN). Технологии глубокого обучения, которые являются ключевыми компонентами современных сервисов искусственного интеллекта (ИИ), демонстри-

руют большие успехи в обеспечении возможностей человеческого уровня для выполнения таких задач, как визуальный анализ, распознавание речи и т. д. Построение модели глубокого обучения производственного уровня является нетривиальной задачей, требующей большого объема обучающих данных, мощных вычислительных ресурсов и человеческого опыта. Следовательно, незаконное воспроизведение, распространение и использование моделей глубокого обучения может привести к нарушению авторских прав и нанести серьезный экономический ущерб создателям моделей. Поэтому актуальна проблема разработки методик защиты интеллектуальной собственности моделей глубокого обучения, включающих внешнюю проверку права владения моделью [9, 10].

Чтобы проверить право собственности на защищенные медиаданные, все существующие алгоритмы требуют доступ к этим данным мультимедиа для извлечения ЦВЗ и подтверждения права собственности. Однако в глубоких нейронных сетях нужно защищать сами модели DNN, а не входные данные мультимедиа. После обучения, обычно, только интерфейс модели DNN доступен для внешней проверки права собственности. Поэтому существующие алгоритмы ЦВЗ, разработанные для защиты мультимедиа, не могут быть непосредственно применены для защиты моделей DNN.

Тем не менее, в [10] показано, что концепция ЦВЗ может быть применена для защиты глубоких нейронных сетей и создания удаленного механизма проверки для определения владельца DNN-модели. Расширяя внутренние возможности обобщения и запоминания глубоких нейронных сетей, можно позволить моделям выучить специально созданные ЦВЗ при обучении. Впоследствии, при подаче на вход модели такого специального ЦВЗ, на выходе генерируется заранее заданный неочевидный результат. Водяные знаки должны быть скрытыми и труднодоступными для обнаружения, уничтожения или видоизменения неавторизованными сторонами. Для достижения этой цели количество потенциальных ЦВЗ должно быть достаточно большим, чтобы избежать обратной инженерии, даже если алгоритмы генерации ЦВЗ известны атакующим. Рассмотрим три возможных механизма генерации ЦВЗ (рис. 3) [10].

а) *б*) *в*) *г*)

Рис. 3. Оригинальное изображение (*а*) и изображения с ЦВЗ трех типов (*б*, *в*, *г*)

Встраивание содержательного контента вместе с исходными данными обучения в виде водяных знаков в защищенные DNN. При данном подходе, мы берем образы из обучающих данных и изменяем изображения, добавляя в них дополнительное смысловое содержание. Удаленные модели, которые нам не принадлежат, не должны иметь такого содержимого. Например, если мы внедряем специальную строку «TEST» в нашу модель DNN, любая модель DNN, которая может быть активирована этой строкой, должна быть воспроизведением или производной от защищенной модели, так как модель, принадлежащая кому-то другому, не должна откликаться на нашу строку «TEST». На рис. 3. *б* показан пример такого ЦВЗ. Мы берем изображение (рис. 3, *а*) из обучающих данных и добавляем логотип «TEST». В результате, любое изображение автомобиля будет правильно классифицировано как автомобиль, однако, если мы поместим на него логотип «TEST», оно будет интерпретировано нашей защищенной моделью, в соответствии с нашим заранее определенным ярлыком, как «самолет». Водяной знак здесь определяется своим содержанием, расположением и цветом. Непосредственно обнаружить с помощью обратной инженерии такие водяные знаки трудно.

Внедрение несвязанных образцов данных в качестве водяных знаков в защищенные DNN. Пример: для модели, задачей которой является распознавание пищи, мы можем использовать различные изображения почерка в качестве водяных знаков. Таким образом, встроенные водяные знаки не влияют на функции исходной модели. Идея здесь заключается в том, что мы добавляем новую интеллектуальную функцию (например, распознавание несвязанных данных) к защищенной модели, и такая новая функция может помочь при проверке. На рис. 3, *с* показан пример, где использовано изображение рукописного символа «1» в качестве водяного знака, и ему при обучении приписан ярлык «самолет». В результате, защищенная модель будет распознавать и реальные самолеты, и ЦВЗ «1», как «самолет». В процессе проверки, если защищенная модель также успешно распознает изображения из нашего внедренного несвязанного с основной задачей класса (например, изображение символа «1»), то мы можем подтвердить право собственности на эту модель. Потенциальное число несвязанных классов бесконечно, что затрудняет реверсирование наших встроенных водяных знаков.

Встраивание шума в качестве водяных знаков в защищенные DNN. В качестве ЦВЗ может быть использовано специфически заданный шум. Здесь мы добавляем бессмысленный шум в изображения. Таким образом, если даже встроенные ЦВЗ будут обнаружены, будет трудно отличить такие шумовые метки от чистого шума. На рис. 3, *д* показан пример ЦВЗ, основанного на шуме, полученного путем добавления к изображению (рис. 3, *а*) из обучающих данных гауссовского шума. В результате, оригинальное изображение (рис. 3, *а*) будет правильно признано автомобилем, но изображение с гауссовским шумом будет идентифицироваться как «самолет» (рис. 3, *г*). Идея состоит в том, чтобы обучить защищенную модель DNN обобщать картину шума или запоминать специфический шум. Если шум запоминается, распознаются только встроенные

водяные знаки, а если обобщен, то любой шум с гауссовским распределением будет распознан.

Эксперимент по подтверждению права собственности довольно прост: мы просто предоставляем DNN специально созданное изображение, которое вызывает неожиданный, но контролируемый ответ, если модель была помечена водяными знаками. Такой подход позволяет точно и быстро проверить право собственности на дистанционно развернутую модель глубокого обучения без влияния на точность модели для обычных входных данных. Встроенные ЦВЗ в моделях DNN должны быть надежны и устойчивы к различным механизмам устранения водяных знаков, таким как тонкая настройка, обрезка параметров и инверсия модели.

Такой метод, конечно, имеет некоторые ограничения. Если украденная модель не развернута как онлайн-сервис, а используется как внутренняя служба, то мы не сможем обнаружить никакой кражи, но тогда, конечно, и плагиатор не сможет напрямую монетизировать украденные модели.

Заключение

Подводя итоги, отметим, что основное требование к ЦВЗ, как и к любой системе защиты – это надежность и устойчивость к искажениям информации.

Стеганография набирает обороты в развитии: формируется теоретическая база, ведется разработка новых, устойчивых методов встраивания сообщений. Основная причина популярности стеганографии – принятый в некоторых странах закон на ограничения использования сильной криптографии. А также это один из лучших на данный момент способов защиты авторских прав на свои произведения.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Matsui K., Tanaka K., and Nakamura Y. Digital signature on a facsimile document by recursive MH coding // Symposium On Cryptography and Information Security, 1989.

2. Voloshynovskiy S., Pereira S., Iquise V., Pun T. Attack Modelling: Towards a Second Generation Watermarking Benchmark // Preprint. University of Geneva. – 2001. – 58 p.

3. Жельников В. Криптография от папируса до компьютера. – М., 1996.

4. Грибунин В. Г., Туринцев И. В., Оков И. Н. Цифровая стеганография // Сер. «Аспекты защиты». – 2009. – № 2. – С. 45–56.

5. J. L. Divya Shivani, Ranjan K. Senapati. False-positive-free, Robust and Blind Watermarking Scheme based on Shuffled SVD and RDWT / Journal of Advanced Research in Dynamical and Control Systems, 2018, vol. 10, pp. 1971–1982.

6. Сагайдак Д. А., Файзуллин Р. Т. Способ формирования цифрового водяного знака для физических и электронных документов / Компьютерная оптика, 2014. – Т. 38, № 1. – С. 94–117.

7. R. Halder, Sh. Pal, A. Cortesi. Watermarking Techniques for Relational Databases: Survey, Classification and Comparison / Journal of Universal Computer Science, vol. 16, no. 21 (2010), pp. 3164–3190.

8. Печенкина А. В., Карманов И. Н. Стеганографические методы и алгоритмы обработки изображений в оптотехнических системах // Актуальные проблемы оптотехники : сб. материалов Национ. науч.-техн. конф., 22 октября 2018 г., Новосибирск. – Новосибирск : СГУГиТ, 2018. – С. 112–117.

9. Yuki Nagai, Yusuke Uchida, Shigeyuki Sakazawa, Shin'ichi Satoh. Digital watermarking for deep neural networks / International Journal of Multimedia Information Retrieval, 2018. – Vol. 7. – Issue 1. – P. 3–16.

10. J. Zhang, Zh. Gu, J. Jang, H. Wu, M. Ph. Stoecklin, H. Huang, I. Molloy. Protecting Intellectual Property of Deep Neural Networks with Watermarking // Proceedings of ASIA CCS '18 ACM Asia Conference on Computer and Communications Security. – Incheon, Republic of Korea. – June 04 – 08, 2018. – P. 159–172.