Generative Image Steganography Based on Digital Cardan Grille

Yaojie Wang[1,2(✉)], Xiaoyuan Yang[1,2], and Wenchao Liu[1,2]

[1] Engineering University of PAP, Xi'an 710086, China
wangyaojie0313@163.com
[2] Key Laboratory of Network and Information Security of PAP, Xi'an 710086, China

Abstract. In this paper, a generative image steganography algorithm based on digital Cardan Grille is proposed. Combining the ideas of traditional Cardan Grille and the semantic image inpainting technique, the stego image are driven by secret messages directly. The algorithm first embeds the information based on digital Cardan Grille, and then uses generative adversarial network (GANs) to complete the damaged image. The adversarial game not only reconstruct the corrupted image, but also generate a stego image which contains the logic rationality of image content. The experimental results verify the feasibility of the proposed method.

Keywords: Image steganography · Cardan grille · Semantic completion · Generative adversarial network

1 Introduction

In ancient China, a clever steganography was invented. The sender and the receiver each hold an identical piece of paper with many small holes, and the positions of these holes are randomly selected. The sender covers this piece of paper with a hole, writes the secret information in the position of the small hole, and then removes the paper above, and fills in the empty space to make the whole text logical significance. The receiver can read the secret information left in the small hole as long as the perforated paper is covered with this ordinary text. In the early 16th century, the Italian mathematician Cardan (1501–1576) also invented this method, which is now called Cardan Grille [1]. As an important branch of information security, steganography has made great progress. For a long time, steganography has mostly used images and videos as the cover, And modification of a small number of pixels is the most commonly used method. Although various technologies were adopted to correct and cover up the modification traces, the security and practicability of steganography faced great challenges along with the improvement of computing power [2].

Fortunately, a new technique of deep learning, generative adversarial networks (GANs) [3], has become a new research hotspot in the field of information hiding.

D. Wang et al. (Eds.): SPNCE 2020, LNICST 344, pp. 343–355, 2021.
https://doi.org/10.1007/978-3-030-66922-5_24

The biggest advantage and feature of GANs is the ability to sample real space and generate samples driven by noise, which provides the new possibility for steganography. Combined with deep generative model technology, this paper gives new vitality to the Cardan Grille, and a generative steganography algorithm based on digital Cardan grille is proposed. First, the random digital Cardan grille is generated automatically, which plays the role of a key in cryptography. Secondly, a blank image is taken as the cover, and the secret information is written to the area that needs to be filled by a digital Cardan grille. In the case of keeping the secret message unchanged, the semantic image completion is realized by using the generative adversarial network. The secret message is hidden in the reconstructed image after completion. The experiments on the image database confirms the feasibility of such simple method.

The remainder of this paper is organized as follows: In the following sections we detail the current status of deep generation models in steganography. Section 3 shows how to build digtal Cardan grille by GANs. Experiment results are demonstrated in Sect. 4. Section 5 concludes this research and details our future work.

2 Related Work

Modern steganography [4] entered the world in 1985 with the advent of personal computers being applied to classical steganography problems. Fridrich [5] considers steganographic channel is divided into three categories, cover selection, modification and synthesis. cover modification is a steganography method adopted by most traditional steganography techniques, which is mainly aimed at reducing the difference between covers before and after modification. Due to the high-dimensional characteristics of the cover itself, the modified carrier always leaves traces of modification, which can be attacked from different dimensions. cover selection is essentially a mapping between the original cover and the secret message. For example, Zhou et al. [6] used the bag-of-words model (BOW) to extract the visual words (VW) of the image. The embedding rate of this method is very low, and there is a security risk when it is used many times, so its practical application is less. Cover synthesis attempts to generate a natural cover, which was very difficult before the deep generative model appeared. With the help of texture synthesis, [7, 8] also use the texture sample and a bunch of color points generated by secret messages to construct dense texture images. This kind of texture-based steganography is based on the premise that the cover may not represent the content in real world.

With the development of machine learning, the deep generation model represented by generative adversarial networks has developed rapidly and achieved remarkable achievements. Recently, this new technology has applied adversarial training to steganographic problems. Volkhonskiy etc. [9] first propose a new model for generating image-like containers based on Deep Convolutional Generative Adversarial Networks (DCGAN) [10]. This approach allows to generate more setganalysis-secure message embedding using standard steganography algorithms. Shi etc. [11] introduce a new generative adversarial networks to improve convergence speed, the training stability the image quality. On the basis of [9, 11], Wang et al. [12] introduced a new method of steganography to improve the training effect. However, most of these GAN-based steganographic schemes are still the cover modification techniques for steganography. These methods focus on the adversarial game, but ignore the core aim of the GAN is to build a powerful generator.

Since GAN's biggest advantage is to generate samples, it is a intuitive idea to use GANs generate a semantic cipher from a message directly as the Cardan did. Some researcher made a preliminary attempt on this intuitive idea. Hu et al. [13] proposed a steganographic scheme without cover modification. The core of this method is the extraction of information. The mean square error is used as the evaluation criterion to obtain the noise extractor. Hayes et al. [14] Proposed a steganography method that uses a three-party adversarial network to make the machine automatically learn. Liu et al. [15] proposed a steganographic algorithm for data sampling based on a deep generation model. This method does not depend on a specific cover, the stego image is actually obtained by sampling by the generator. Liu et al. [16] also proposed the Stego-ACGAN method, which uses semi-supervised learning to establish a mapping relationship between information to achieve the generation of a specified carrier. Ke [17] proposed generative steganography method called GSK in which the secret messages are generated by a cover image using a generator rather than embedded into the cover, thus resulting in no modifications in the cover.

In this paper, a generative steganography, based on digital Cardan Grille, is proposed. A mask called digital Cardan grille for determining the hidden location is introduced to hide the message. According to the position corresponding to digital Cardan Grille, the secret information is written into the blank image in advance, and the above information remains unchanged during the entire steganography process. Images with secret messages are then input into a generative adversarial networks to complete semantic completion. The adversarial game not only complement broken images, but also generate a stego image which contains the logic rationality of image content.

3 The Proposed Algorithm

With the help of generative adversarial network technology, the framework of generative steganography based on digital Cardan Grille proposed in this paper is shown in Fig. 1. The sender and receiver define a mask together, called digtal Cardan grille, to determine where the message is hidden, it acts as a key in cryptography. According to the "1" position corresponding to digital Cardan Grille (the "0" position is omitted), the secret information is written into the blank image in advance, and the above information remains unchanged during the entire steganography process. Images with secret messages are then input into a generative adversarial network to complete semantic completion. A stego image is transmitted to the recipient through the public channel. The receiver uses the shared digital Cardan Grille to extract the secret information in a reverse process. The algorithm mainly includes three parts: the design of digital Cardan Grille, the automatic generation of digital Cardan Grille, and the completion of the image after embedding the information. The basic principles will be explained one by one in the next section. In fact, this framework describes a general automated Cardan grille, and the input cover vector can be a piece of text, image, video, and other type of media.

3.1 Design Principles of Digital Cardan Grille

In this paper, digital Cardan Grille still adopts this clever idea of the traditional Cardan Grille, using the Hamard product [18] in matrix multiplication to complete the digital

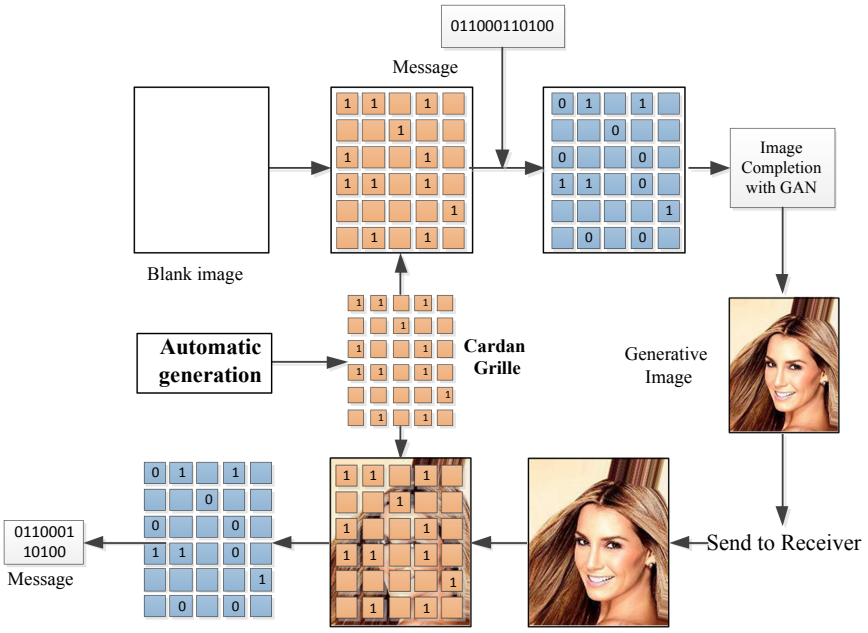


Fig. 1. The structure of the proposed generative steganography

conversion. Hadamard product is a type of matrix operation. If $A = (a_{ij})$ and $B = (b_{ij})$ are two matrices of the same order, if $c_{ij} = a_{ij} \times b_{ij}$, then the matrix $C = (c_{ij})$ is called A and B's Hadamard product. In mathematics, the Hadamard product is expressed as follows:

Definition Suppose A and B, and $A = (a_{ij})$, $B = (b_{ij})$. Then the matrix $C = (c_{ij})$

$$C = (c_{ij}) = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \dots & a_{1n}b_{1n} \\ a_{21}b_{21} & a_{22}b_{22} & \dots & a_{2n}b_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1}b_{m1} & a_{m2}b_{m2} & \dots & a_{mn}b_{mn} \end{bmatrix} \tag{1}$$

is the Hadamard product of the matrices A and B, denoted as $A \odot B$.

According to the above, the design process of digital Cardan Grille is described as follows: if A is stego information and B is a Cardan Grille of a binary mask, which can only be represented by two values of 0 and 1, so the secret information can be hidden and extracted by using the Hadamard product. Taking a 2×2 simple image as an example, a blank image is covered with a digital Cardan Grille in the hiding process. The position of 1 indicates a small hole in the traditional Cardan Grille, which means that information can be embedded. A position with a value of 0 indicates that there are no holes, which means that information cannot be embedded, and which needs to be completed later by GANs, as shown in Fig. 2(a).

When extracting information, the same digital Cardan Grille is still used, and the embedded secret information is obtained through the Hadamard product. A value of 1

indicates information that needs to be retained in the image, and a value of 0 indicates information that needs to be discarded, that is, the information retained in the Hadamard product is the embedded secret information, as shown in Fig. 2(b). For third parties, the information of digital Cardan Grille is strictly confidential.

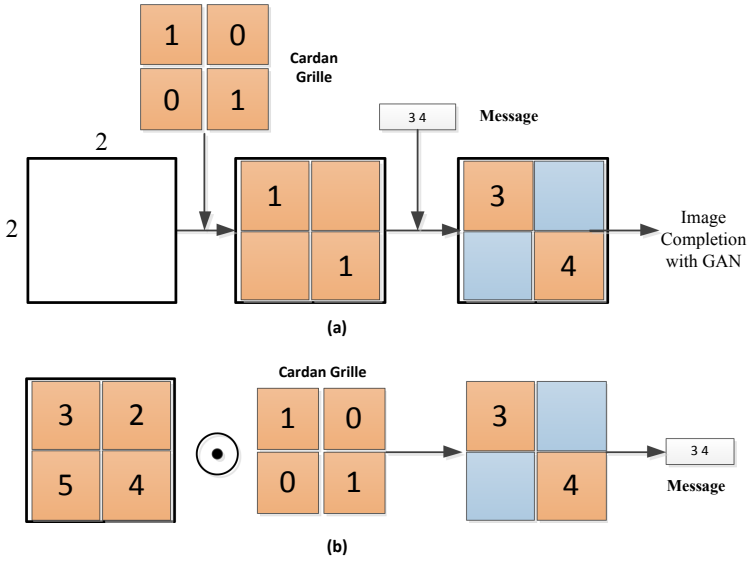


Fig. 2. The principles of digital Cardan Grille

3.2 Automatic Generation of Digital Cardan Grille

Although the traditional Cardan Grille can generate a stego cover, it needs to artificially construct a meaningful cover, such as Tibetan head poems, which is time-consuming and labor-intensive. The automatic generation of digital Cardan Grille not only meets the needs of the covert communication in reality, but also improves the security of the information, which greatly increases the obstacles for third-party attacks and deciphering.

According to the design principle of digital Cardan Grille, the message digest algorithm in cryptography is used for automatic generation. The message digest algorithm is mainly applied in the field of digital signature. Its main characteristic is that the encryption process does not require a key, and it has uniqueness and irreversibility. Only by entering the same plain text data can be obtained through the same message digest algorithm. Well-known digest algorithms include MD5 algorithm, SHA-1 algorithm, SHA-256 algorithm and a large number of variants. In this paper, the SHA-1 algorithm is used as an example for the automatic generation and interpretation of digital Cardan Grille, as shown in Fig. 3. SHA-1 algorithm can output a 20-byte hash value, which is usually represented in the form of 40 hexadecimal digits. There are four steps to automatically generating digital Cardan Grille:

1. Obtain the agreed public information as an input signal. The public information obtained at different times is different, such as the headlines on the front page of the Wall Street journal, which encode the information in binary;
2. Fill the key. The sender and receiver share the key in advance and fill the encoded public information according to the agreed filling rules. it is filled at the end and only serves as the principle explanation in Fig. 3;
3. Encryption. The filled information is input into the SHA-1 generator, and a 20-byte message digest hash value (160 bits) is output.
4. The message digests in step 3 are arranged in sequence to form a matrix, which is called as digital Cardan Grille. The extra digits can be discarded according to the amount of embedded secret information. If the number of matrix bits cannot fill the entire vector, the rest are filled with zeros.

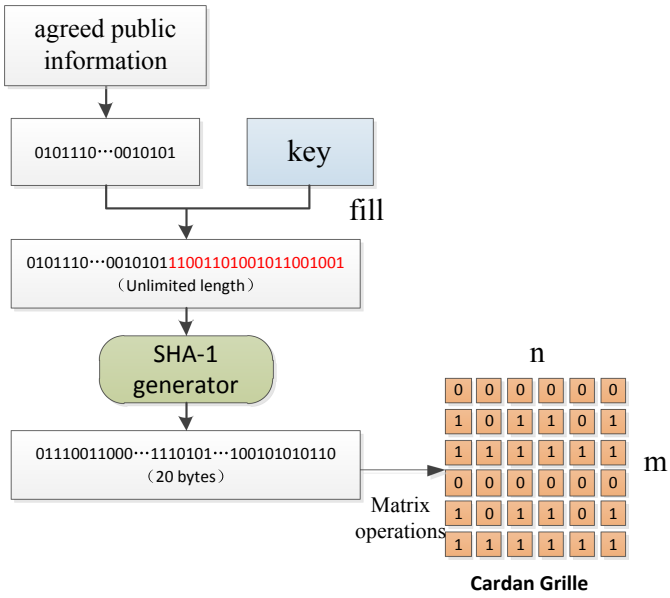


Fig. 3. The automatic generation of digital Cardan Grille

3.3 Completion Stego Image

In recent years, there are various techniques for image completion, but the image damage rate that needs to be completed in this article occupies more than 95%, which means that the completion here is not a small amount of repair, but more inclined to generate images on a large scale. As mentioned above, inspired by [19], this paper still uses the a image inpainting method which proposed by Yeh [20] based on a Deep Convolutional Generative Adversarial Network (DCGAN), and some parameters are adjusted to ensure that the generated image meets the needs of semantics.

Assuming that we need to complete the broken y, we still use the Hadamard product for pre-processing. First we choose a binary mask M that has values 0 or 1. According to the digital Cardan Grille's principle and Hadamard product, we can know that $M \odot y$ can represent the original part of the stego image, and $(1 - M) \odot y$ also means the other part that needs completion. Suppose we generate a reasonable $G(z')$ to complete the missing part, the original part and the completed part constitute the reconstruction of the image:

$$x_{reconstructed} = M \odot y + (1 - M) \odot G(z') \tag{2}$$

In the process of image completion, two loss functions need to be defined. The specific content is similar to the reference [20], where the embedded information part is equivalent to the unbroken area of the image.

Contextual Loss: To keep the same context as the input image, make sure the known pixel locations in the input image y are similar to the pixels in $G(z)$. We need to penalize $G(z)$ for not creating a similar image for the pixels that we know about. Formally, we do this by element-wise subtracting the pixels in y from $G(z)$ and looking at how much they differ:

$$L_{contextual(z)} = |M \odot G(z) - M \odot y| \tag{3}$$

In the ideal case, all of the pixels at known locations are the same between y and $G(z)$. Then $G(z)_i - y_i = 0$ for the known pixels i and thus $L_{contextual}(z) = 0$.

Perceptual Loss: To recover an image that looks real, let's make sure the discriminator is properly convinced that the image looks real. We'll do this with the same criterion used in training the DCGAN:

$$L_{perceptual(z)} = \log(1 - D(G(z))) \tag{4}$$

Contextual Loss and Perceptual Loss successfully predict semantic information in the missing region and achieve pixel-level photorealism. We're finally ready to find z' with a combination of context loss and perception loss:

$$L(z) \equiv L_{contextual}(z) + \lambda L_{perceptual}(z) \tag{5}$$

$$z' \equiv arg\ min L(z) \tag{6}$$

where λ is a hyper-parameter that controls how import the contextual loss is relative to the perceptual loss.

Compared with [20], the proposed scheme needs more information to complete, and we have now adjusted the parameters. In experiments, the effect of generating images is relatively good when $= 0.7$.

4 Experiment and Analysis

In order to verify the performance of the scheme, in the experiments we used the CelebA data set (Ziwei Liu and Tang 2015) and the LSUN dataset to train the network model separately. The former is openly provided by the Chinese University of Hong Kong and is widely used for face-related computer vision training tasks, including more than 200,000 images. The latter is a large-scale image data set constructed by humans performing deep learning in a loop, which contains 10 scene categories and 20 object categories, each category has about one million labeled images.

The experimental environment is shown in Table 1. We used the DCGAN model architecture from Yeh et al. [20] in this work. The optimizer in DCGAN uses an Adam-based optimization method with a learning rate of 0.0002. At each training, the weight of the discriminator D is updated once, the weight of the generator G is updated twice [21].

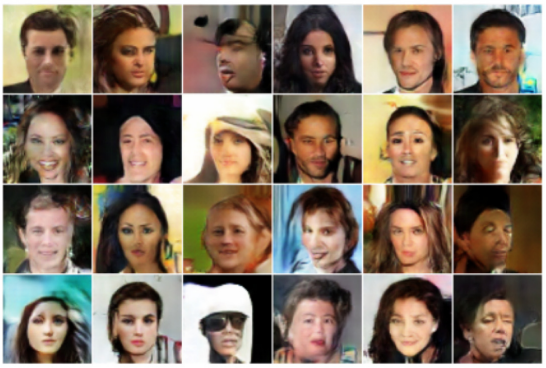
Table 1. Experimental environment

Software platform	Tensorflow v0.12	
Hardware environment	CPU	i7-8250U 3.2 GHz
	RAM	16 GB DDR3 1600 MHz
	GPU	NVIDIA 1080

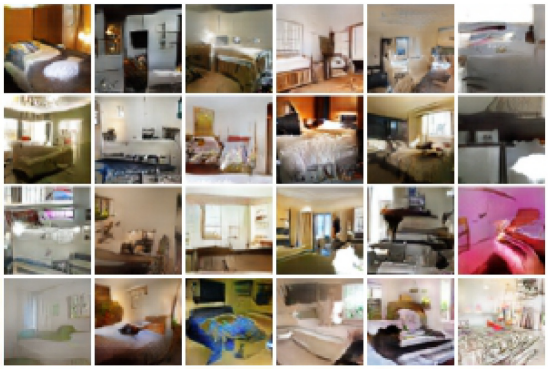
In the experiment, we use the alignment tool to preprocess the image to 128 × 128. Based on the work of Brandon Amos's bamos/dcgan-completion.tensorflow [22], I modified the parameters. Digital Cardengo is automatically generated. Assuming that the embedded information is randomly distributed over 16 pixels, Fig. 4(a) is an example of generating a model complement CelebA image after training for 11 epochs. Figure 4(b) is an example of generating a model complement LSUN image after training for 7 epochs.

In order to better evaluate the quality characteristics of the generated images, we introduced the No Reference Image Quality Assessment (NR-IQA) method [23]. Because this method does not need to be compared with the original image, and the image steganography based on the generated model does not have the original carrier, the characteristics of the two are completely consistent, which makes up for the shortcomings of the current traditional steganographic evaluation system.

If there is no abnormality in visual observation, Magnitude spectrum, frequency histogram, and DCT coefficient histogram are commonly used as evaluation methods. 1,000 images randomly selected from the algorithm generated in this paper are analyzed, and specific examples are shown in Fig. 5. Through simulation experiments, compared with natural images, the generated samples have no statistical abnormalities in frequency characteristics, abnormal DCT coefficients, etc., which can basically meet the needs of realistic communication, and can also effectively resist detection based on statistical steganalysis.



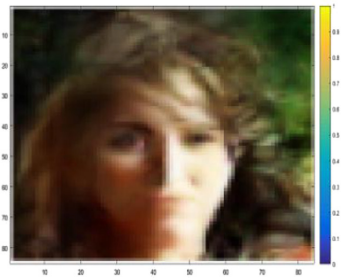
(a)



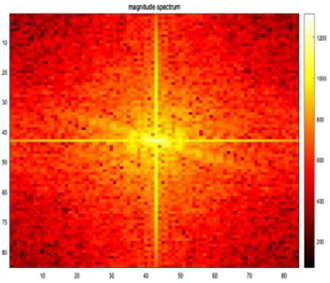
(b)

Fig. 4. The example of generated stego image

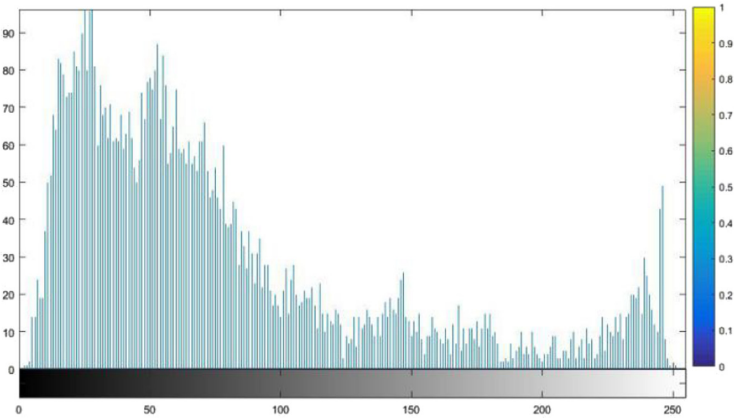
Next, we verify the detection resistance of the generated image, that is, use the current mainstream detection algorithms to verify the imperceptibility of the algorithm in this paper. This paper randomly chooses 6000 real images (from CelebA data set and the LSUN dataset) and 4000 images generated by this algorithm as the test set. The four detection algorithms selected are steganalysis in DCT domain [24], RS detection [25], nonlinear SVM detection [26], and S-CNN detection [27]. The DCT domain steganalysis mainly focuses on the statistical characteristics of DCT coefficients and their impact on spatial pixels. The RS detection method (regular groups and singular groups) is evaluated based on the gray value of the image. The essence of the nonlinear SVM detection method is to extract the feature data of the sample, which is a binary classification model. S-CNN detection is the latest method for image detection using convolutional neural networks. In the case of random group independent testing, the comparison results of steganography detection are shown in Table 2.



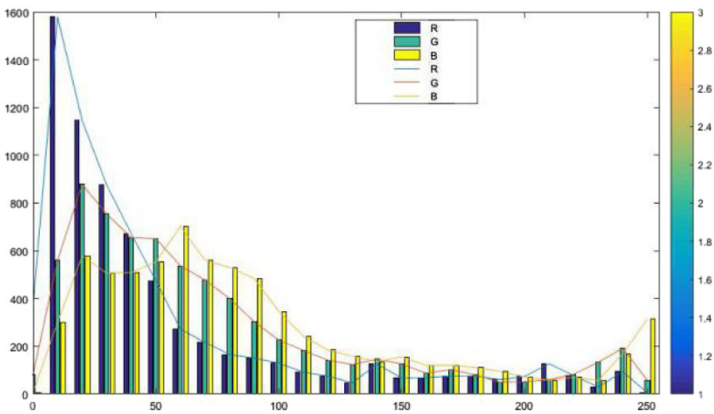
(a) generative image



(b) Magnitude spectrum



(c) Frequency histogram of corresponding image



(d) DCT coefficient histogram of corresponding image

Fig. 5. The example of generated encrypted image

Except for the S-CNN detection method, the detection accuracy of the other three detection algorithms approaches 0.5, and the experimental results show that the algorithm in this paper has great advantages in terms of detection resistance. For the S-CNN detection method, the detection accuracy rate is close to 0.65, which needs to be further improved. In the actual communication process, imperceptibility can be better satisfied by reducing the amount of embedded information, so as to ensure the universality of steganography.

Table 2. Accuracy of the steganalysis test

Detection algorithm	Detection accuracy	
	Group accuracy	Average accuracy
Steganalysis in DCT domain	0.534	0.556
	0.579	
RS detection	0.510	0.516
	0.523	
Nonlinear SVM detection	0.573	0.562
	0.551	
S-CNN detection	0.672	0.658
	0.644	

Based on the above experiments, we theoretically analyze the security of the algorithm in this paper. The security of the scheme in this article is based on two aspects: First, it is based on the confidentiality of digital Cardan Grille. The security of digital Cardan Grille cannot be separated from the key and digital digest algorithm. In the case that the key is not public, the message digest algorithms such as SHA-1 and MD5 are irreversible And uniqueness, which guarantees that brute force cracking is not feasible. In other words, the security of the system depends on the confidentiality of the selected key, which fully complies with the Kerckhoffs criterion. At the same time, a message digest algorithm with different key lengths can be selected according to different secret levels. The algorithm is easy to implement and difficult to decipher under the premise of regularly changing keys. The second is that the transmitted dense image is directly generated by the generator, which can meet the statistical characteristics of the current specific steganography detection, and greatly increases the ability to resist steganalysis. But compared with the traditional steganography method based on cover modification, it is less versatile.

Assuming that the attacker suspects that the passed image contains secret information, it is also difficult to obtain the same cardengo to extract secret information. Even if the algorithm of information steganography is leaked, under the premise of no key, only meaningless results will be obtained, thereby ensuring the security of covert communication.

5 Conclusion and Future Work

In this paper, a generative image steganography algorithm based on digital Cardan Grille is proposed. Combining the ideas of traditional Cardan Grille and deep learning, and using the deep generation model to construct a normal semantic image, which is fully in line with the research direction of coverless information hiding. The program can be extended to other media such as text, video and other fields. We use the CelebA dataset and the LUSN dataset to evaluate the performance of the proposed scheme. Theoretical analysis and experimental results show the feasibility and safety of the method. However, the generality of the algorithm in this paper is still poor and needs to be improved.

In future work, we hope to pay more attention to the embedded information capacity. On the basis of ensuring the security of the generated image, the embedded capacity is increased to meet the needs of realistic covert communication.

Acknowledgment. This work was supported by National Key R&D Program of China (Grant No. 2017YFB0802000), National Natural Science Foundation of China (Grant Nos. 61379152, 61403417).

References

1. Utepbergenov, I., Mussin, T., Kuandykova, J.: Creating a program and research a cryptosystem on the basis of Cardan grille. In: 2013 Second International Conference on Informatics and Applications (ICIA). IEEE (2013)
2. Wang, H., Wang, S.: Cyber warfare: steganography vs. steganalysis. Commun. ACM **47**(10), 76–82 (2004)
3. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial networks. Adv. Neural. Inf. Process. Syst. **3**, 2672–2680 (2014)
4. Katzenbeisser, S., Petitcolas, F.A.P.: Information Hiding Techniques for Steganography and Digital Watermarking (2000)
5. Fridrich, J.: Steganography in Digital Media: Principles, Algorithms, and Applications. Cambridge University Press, Cambridge (2010)
6. Zhou, Z.L., Cao, Y., Sun, X.M.: Coverless information hiding based on bag-of-words model of image. J. Appl. Sci. **34**(5), 527–536 (2016)
7. Otori, H., Kuriyama, S.: Data-embeddable texture synthesis. In: Butz, A., Fisher, B., Krüger, A., Olivier, P., Owada, S. (eds.) SG 2007. LNCS, vol. 4569, pp. 146–157. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73214-3_13
8. Otori, H., Kuriyama, S.: Texture synthesis for mobile data communications. IEEE Comput. Graph. Appl. **29**(6), 74–81 (2009)
9. Volkhonskiy, D., Nazarov, I., Borisenko, B., et al.: Steganographic generative adversarial networks [EB/OL]. ArXiv e-prints, 2017, 1703, 16 March 2017. http://arxiv.org/abs/1703.05502.pdf
10. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. Comput. Sci. (2015)
11. Shi, H., Dong, J., Wang, W., et al.: SSGAN: secure steganography based on generative adversarial networks [EB/OL]. ArXiv e-prints, 2018, 1707, 06 November 2018. https://arxiv.org/abs/1707.01613v3.pdf

12. Wang, Y.J., Niu, K., Yang, X.Y.: Information hiding scheme based on generative adversarial network. J. Comput. Appl. **38**(10), 2923–2928 (2018). https://doi.org/10.11772/j.issn.1001-9081
13. Hu, D., Wang, L., Jiang, W., et al.: A novel image steganography method via deep convolutional generative. IEEE Access **6**, 38303–38314 (2018). https://doi.org/10.1109/ACCESS.2018.2852771
14. Hayes, J., Danezis, G.: Generating steganographic images via adversarial training [EB/OL]. ArXiv e-prints, 2017, 1703, 01 March 2017. https://arxiv.org/abs/1703.00371.pdf
15. Liu, J., Ke, Y., Lei, Y., et al.: The reincarnation of grille cipher: a generative approach (2018)
16. Liu, M.M., Zhang, M.Q., Liu, J., et al.: Coverless information hiding based on generative adversarial networks. J. Appl. Sci. **36**(2), 371–382 (2018). https://doi.org/10.3969/j.issn.0255-8297.2018.02.015
17. Ke, Y., Liu, J., Zhang, M.-Q., et al.: Steganography security: principle and practice. IEEE Access **6**, 73009–73022 (2018)
18. Novotny, M.A.: Matrix products with applications to classical statistical mechanics, 1. Reflectivity of one-dimensional solids, 2. **56**(3), 452–458 (1978)
19. Liu, J., Zhou, T., Zhang, Z., et al.: Digital cardan grille: a modern approach for information hiding. In: The 2018 2nd International Conference (2018)
20. Yeh, R., Chen, C., Lim, T. Y., Hasegawajohnson, M., Do, M.N.: Semantic image inpainting with perceptual and contextual losses (2016)
21. Dumoulin, V., Belghazi, I., Poole, B., et al.: Adversarially learned inference (2016)
22. Amos, B.: Image completion with deep learning in TensorFlow. https://github.com/bamos/dcgan-completion.tensorflow
23. Mandgaonkar, V.S., Kulkarni, C.V.: No reference image quality assessment. In: 2014 Annual IEEE India Conference (INDICON). IEEE (2015)
24. Suryawanshi, G.R., Mali, S.N.: Study of effect of DCT domain steganography techniques in spatial domain for JPEG images steganalysis. Int. J. Comput. Appl. **127**(6), 16–20 (2015)
25. Zhen, Y.U., Chen, K.: Analysis and improvement on RS detection algorithm. Comput. Eng. **34**(8), 170–178 (2008)
26. Yan, H., Qin, J.: Trojan Horse detection method based on nonlinear SVM model. Comput. Eng. **37**, 121–123 (2011)
27. Chen, T., Lu, S., Fan, J.: S-CNN: Subcategory-aware convolutional networks for object detection. IEEE Trans. Pattern Anal. Mach. Intell. 1