A coverless steganography method based on generative adversarial network

Xintao Duan*†, Baoxia Li †, Daidou Guo †, Zhen Zhang † and Yuanyuan Ma †

Abstract
The traditional information hiding is realized by embedding the secret information into the multimedia, but it will inevitably leave the modification mark in the carrier. This paper proposed a new method of coverless information hiding. First, the improved Wasserstein GAN (WGAN-GP) model is constructed, and the model is trained with disguised images and secret images. Then, after the model is stable, a disguised image is passed to the generator. Finally, the generator generates the image that is visually the same as the secret image, thereby achieving the same effect as transmitting the secret image. Experimental results show that this method not only has a good effect on the security of secret information transmission, but also increases the capacity of information hiding.

Keywords: Coverless information hiding, Generative adversarial network, WGAN-GP

1 Introduction
The rapid development of computer network technology has enabled multimedia information such as videos, texts, and images to be transmitted quickly on the network. However, the network provides information sharing, brings convenience to people, and also has many security risks. Information hiding technology is an important modern information security technology. It is a method of transmitting secret information by hiding secret information in texts, images, videos, and other carriers [1]. According to its use, information hiding can be divided into steganography and digital watermarking. Steganography is used for the transmission of secret information, while digital watermarking is used for copyright protection and other scenarios [2–4]. According to the hidden protocol, it can be divided into no secret key steganography system, public key steganography system, and private key steganography system [5–7]. According to the technology, it can be divided into spatial domain [8], transform domain [9], frequency domain-based [10], and structure-based steganography.

The simplest steganography algorithm is the least significant bit (LSB) information hiding, but it leaves a significant modification feature to the steganography image [11]. With the development of steganography, the new steganography algorithm can maintain more complex image statistical features, such as HUGO [12], SUNWARD [13], and WOW [14]; the paper [15] proposed a scheme of encrypting, compressing, and finally reconstructing the secret image of the secret image; the paper [16] proposed non-uniform watermark sharing based on optimal iterative BTC for image tampering recovery scheme, and so on. Adaptive embedding strategy [17] can automatically embed secret information into the texture and noise-rich image region, so as to maintain complex higher-order statistical features [18]. In order to combat the more advanced adaptive steganography, the features involved in steganalysis are gradually becoming more complex and high-dimensional. In recent years, high-order statistical characteristics based on complex correlation modeling in image domain have become the main research characteristics of steganalysis [19]. PSRM (Projection Speciation Rich Model) [20], Selection of Rich Model Steganalysis Features Based on Decision Rough Set...
\(\alpha\)-Positive Region Reduction [21], and other models are based on such high-order and high-dimensional features, and they have achieved good detection effect.

At present, neural networks have become a research hotspot in various fields [22–24]. In this paper, neural network is introduced into the research of image information hiding. First, build a WGAN-GP model. Then, the disguised image is sent to the generation network, and the secret image is input to the discrimination network as a real image, and the generation network and the discrimination network are trained using the maximum and minimum games. The generation network distinguishes the real image from the generated image, so as to obtain the generated image that is visually identical to the secret image. The contributions of our work are as follows:

- In the paper, the WGAN-GP model is constructed. For the first time, the WGAN-GP model is used for image information hiding, and the image is input into the generator instead of random noise. The generated image is visually identical to the secret image, so as to achieve the same effect as the transmission of secret image.
- The recipient can generate the same image as the secret image by using this separate generation network and the received disguised image. The disguised image is transmitted without any modification or embedding operation, which effectively avoids the detection of steganalysis algorithm.
- The generated image is visually the same as the secret image. Only the camouflage image and the corresponding generator can get the secret image, which is highly secure.

The rest of this article is organized as follows. Several related models are introduced in Section 2. The proposed method and experimental environment are introduced in Section 3. Experimental results and discussions are shown in Section 4. Finally, the conclusions are shown in Section 5.

2 Related knowledge
2.1 Generative adversarial network
Generative adversarial network (GAN) was proposed in 2014 and has attracted much attention in various fields. New GAN models and applications have been emerging continuously [25]. WGAN-GP [26] is also a derivative model of GANs. This paper proposes to use WGAN-GP model for image information hiding.

GAN consists of generator and discriminator. The generator is used to study the real image distribution, and the generated image is more real, which makes it difficult for the discriminator to distinguish true from false. The discriminator demands to discriminate whether the received picture is true or false. In the whole process, the generator strives to make the generated image more real, while the discriminator strives to identify real or false images, which is like a two-person game. The generator and the discriminator continuously compete with each other, and finally, the two networks reach dynamic balance: the image distribution generated by the generator learns the distribution of the real image, and the discriminator cannot judge whether it is a true image or a false image, and the prediction probability of the given image is approximately equal to 0.5. An example can explain GAN more intuitively: the gang making fake money is equivalent to a generator. They want to cheat the bank by forging money, so that the fake money can be traded normally. And the bank is equivalent to a discriminator, which needs to judge whether the money is real money or fake money. The purpose of the counterfeit currency gang is to create a counterfeit currency that the bank cannot identify and deceive the bank. The bank must accurately identify the counterfeit currency. Therefore, we can summarize the above content: true = 1, false = 0, the label that the discriminator will label the real image is 1, and the label of the generated image is 0; the generated confrontation network structure is displayed in the Fig. 1.

There is no mandatory restriction on the choice of the generation model (generator) and the discriminant model (discriminator) in GANs. In [25], they use a multi-layer perceptron. The GANs define noise \( p_z(x) \) as a priori that is used to learn the probability distribution \( p_g \) of the generator on the training data \( x \). \( G(z) \) represents the mapping of the input noise into data (e.g., the generated image), and \( D(x) \) represents the probability that \( x \) comes from the distribution of real data \( p_{data} \) instead of \( p_g \). Therefore, the optimized objective function defines minimax in the following form:

\[
\min_G \max_D \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log (1 - D(G(z)))]
\]

Minmax is to maximize (1) when updating discriminator and to minimize (1) when updating generator. When the generator updates the discriminator, the optimal solution is \( D^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(z)} \). When the generator is updated, the objective function takes the global minimum (if and only if condition \( p_g = p_{data} \) is satisfied). The result of the last two model games is that the generator will create fake data. The discriminator is difficult to determine whether the data created by the generator is real, that is, \( D(G(z)) = 0.5 \). In GAN, if the discriminator is trained too well, the generator will not be able to get enough gradient to continue optimization. If the discriminator is
trained too weakly, the indicator effect is not significant, and the generator will not be able to learn effectively. In this way, the training of discriminator is difficult to control, which is the root of the difficulty of GAN training. The emergence of WGAN solves this problem.

2.2 Earth mover’s distance
Wasserstein GAN (WGAN) [27] explored a more appropriate measure of “generating the difference between distributions”. Earth-mover distance (EM), also known as Wasserstein distance [28], is defined as:

$$W(P_r, P_g) = \inf_{\gamma \sim \Pi(P_r, P_g)} \mathbb{E}_{(x,y) \sim \gamma} [||x - y||]$$

(2)

where $$\Pi(P_r, P_g)$$ is a collection of all possible joint distributions of $$P_r$$ and $$P_g$$ combined. For each possible joint distribution $$\gamma$$, a real sample $$x$$ and a generated sample $$y$$ can be obtained from the middle sample $$(x,y) \sim \gamma$$, and the value $$||x - y||$$ of the pair of samples is calculated, so that the expected value $$\mathbb{E}_{(x,y) \sim \gamma} [||x - y||]$$ of the sample to the distance under the joint distribution can be calculated. The next term that can be taken from this expected value in all possible joint distributions is defined as the EM distance. The meaning of earth mover is the meaning of the bulldozer. This name is very appropriate. Because intuitively, the EM distance is to measure the minimum cost of pushing $$P_r$$ pile of “sand” and $$P_g$$ pile of “location”, of which $$\gamma$$ is a “bulldozing” scheme.

2.3 Wasserstein GAN
EM distance is used to GAN. It is difficult to solve the EM distance directly, but the problem can be transformed into the following formula 3 using a theory called Kantorovich-Rubinstein duality [29].

$$W(P_r, P_g) = \sup_{|f| \leq 1} \mathbb{E}_{x \sim P_r} [f(x)] - \mathbb{E}_{x \sim P_g} [f(x)]$$

(3)
This formula means that all functions \( f \) satisfying the 1-Lipschitz limit are taken to the previous term of \( E_{x \sim p_{r}}[f(x)] - E_{x \sim p_{g}}[f(x)] \). In other words, the Lipschitz limit specifies the maximum local variation \( A \) of a continuous function, as K-Lipschitz \([30]\) is:
\[
|f(x_1) - f(x_2)| \leq K|x_1 - x_2|.
\]
Then, use the neural network method to solve the above optimization problem:

\[
\max_{w \in W} E_{x \sim p_{r}}[f_w(x)] - E_{z \sim p(z)}[f_w(g_\theta(z))]
\]  

(4)

This neural network is very similar to the discriminator in GAN. There are only a few subtle differences, and it is named critic to distinguish it from discriminator. The differences between the two are:

1. The last layer of critic discards sigmoid because it outputs a fraction in the general sense, unlike the probability that discriminator outputs.

2. Critic’s objective function has no log entry, which is derived from the above derivation.

3. Critic has to truncate the parameters in a certain range after each update, also called weight clipping, in order to guarantee the Lipschitz limit mentioned above.

4. The better critic training, the better for the enhancer of the generator, so you can safely train critic.

Although the mathematical proof is very complicated, the final change is very simple. The structure of WGAN is shown in Fig. 2:

2.4 Improved training of Wasserstein GANs

WGAN is sometimes accompanied by problems such as low sample quality and difficulty in convergence. In order to guarantee Lipschitz restrictions, WGAN uses the weight clipping method, but weight clipping causes two major problems: Modeling ability is weakened and gradient explosion or disappearance. The alternative proposed
in [26] is to add gradient penalty (GP) to critic loss. The new network model was called WGAN-GP.

$$L = E_{x \sim P_r} [D(x)] - E_{x \sim P_g} [D(x)] + \lambda E_{\tilde{x} \sim P_{\tilde{x}}} \left( \| \nabla \tilde{x} D(\tilde{x}) \|_2^2 - 1 \right)^2$$

(5)

Inspired by the fact that WGAN-GP generates handwritten characters from MNIST data set, this paper constructs the only WGAN-GP model belonging to sender and receiver for image information hiding. Instead of random noise, camouflage image is transmitted to generator to generate an image with the same sense as secret image. The structure of the WGAN-GP model used in this paper is shown in Fig. 3.

After the model is stabilized, the disguised image passed to the generator can only generate the image which is the same as the secret image in appearance, thereby ensuring the security of the information.

3 Experimental

3.1 Experimental environment and data set

The 10000 images extracted from the LFW [31] data set are used as the data set of this experiment, in which 5000 images of the disguised images and the secret images are respectively 256 × 256 grayscale images. The python version is 3.5, tensorflow version is 1.10, and the GPU is 1080.

3.2 Structure of generators and discriminators

In this paper, the generator used by the WGAN-GP model has 65,536 neurons in the input layer, 64 neurons in the hidden layer, and 65,536 neurons in the output layer. The ReLU activation function was used in the input layer and the hidden layer, while the sigmoid activation function
The input layer of the discriminator has 65,536 neurons, the hidden layer has 64 neurons, and the output layer has 1 neuron. The ReLU activation function is used in the input layer and the hidden layer.

### 3.3 Information hiding and extraction process

The model was trained with disguised images and secret images. When the model is stable, the image generated by the generator is visually the same as the secret image. The receiver receives the disguised image and uses the generator to generate the same image as the original secret image, thereby obtaining a secret image. The same effect is achieved by transmitting disguised images as transmitting secret images. The overall process of the experiment is shown in Fig. 4.

### 4 Results and discussion

As the number of iterations of the model increases, the image generated by the generator gets closer and closer to the secret image. As can be seen from the following example in Fig. 4, when the model is trained 1000 times, the generated image is a noise image. When the model is trained 5000–10000 times, you can see the approximate image content. When the number of trainings reaches 50,000 times, the generated image is visually identical to the secret image and can replace the secret image. The disguised image is in the first column, the original secret image is in the sixth column, and the images generated under different training times of the model are in the second to fifth columns of Fig. 5.

When the model is stable, it is difficult to discriminate the generated image from the original secret image, which
can completely replace the secret image. The experimental results are shown in Fig. 6, the disguised images are in the first line, the generated images are in the second line, and the secret images are in the third line.

In addition to the visual comparison between the generated images and the secret images, 1000 images were randomly selected from the LFW data set. A few examples are taken to analyze the generated images, and the disguised images are shown in Fig. 7.

In order to prove the practicality and generalization of this method, 1000 images were randomly selected from CelabA [32] and ImageNet [33] data sets for experimental verification, and the generated images after the model was stabilized were analyzed by histogram. In Figs. 8 and 9, the disguised images are in the first column, the secret images are in the second column, the generated images are in the third column, the histogram of the secret images is in the fourth column, and the histogram of the generated images is in the fifth column.

Generate the generator by training the model with disguised images and secret images, save the generator after the model training is stable, and construct the mapping relationship between the trained generators and the corresponding disguised images. In order to prove the security of this scheme, we use disguised images and trained generators to obtain secret images. As can be seen from Fig. 10, only disguised images and corresponding generation models can obtain the same

![Fig. 8 Two examples on the CelabA dataset](image)

![Fig. 9 Two examples on the ImageNet dataset](image)
visually as the secret images. Otherwise, only the noise image is obtained, which proves that the method is safe.

Information hiding capacity is one of the key indicators of information hiding systems. This method is to realize the secure transmission of secret images by transmitting disguised images without any modification. The receiver can obtain the same visual image as the secret image by receiving the disguise image and transmitting it to the trained generator. The method increases the information hiding capacity. The definition of information hiding capacity is shown in Eq. 6, and a simple comparison is made with several common information hiding methods in hiding capacity. The comparison results are shown in Table 1.

\[
\text{Relative capacity} = \frac{\text{Absolute capacity}}{\text{The size of the image}}
\]

5 Conclusions
In this paper, the WGAN-GP model was constructed as required. The model is trained using disguised images and secret images so that the transmitted disguised images can be visually the same as the secret images after being passed to the generator. The transmission is a disguised image without any modification, which is not easy to cause the suspicion of the attacker. This method not only solves

### Table 1 Comparisons of hiding capacities

| Methods           | Absolute capacity (bytes/image) | Image size (bytes/pixel) | Relative capacity |
|-------------------|---------------------------------|--------------------------|-------------------|
| Xu’s work[34]     | 64x64                           | 800x800                  | 6.40e−3           |
| Zhou’s work[35]   | 3.72                            | ≥ 512x512                | 8.58e−6           |
| Zhang’s work[36]  | 18.3~135.4                      | 64x64                    | 1.49e−3~1.10e−2   |
| Zhou’s work[37]   | 1.125                           | 512x512                  | 4.29e−6           |
| Otori’s work[38]  | 25~100                          | 480x640                  | 8.14e−5~3.26e−4   |
| Ours              | 256x256                         | 256x256                  | 1                 |
1. S. Maniccam, N. Bourbakis, Lossless compression and information hiding in images. Pattern Recog. 37(3), 475–486 (2004)
2. J. J. Eggers, R. Baeuml, B. Giord. Communications approach to image steganography. Security and Watermarking of Multimedia Contents IV, vol. 4675. pp. 26–37 (2002)
3. Y. K. Lee, L. H. Chen. High capacity image steganographic model. IEE Proc.-Vis., Image and Signal Process. 147(3), 288–294 (2000)
4. F. A. Petitcolas, S. Katzenbeisser, Information Hiding Techniques for Steganography and Digital Watermarking. Artech House Books, vol. 28. pp. 95–172 (2000)
5. S. M. Kaim, M. S. Rahman, M. I. Hossain, in 14th International Conference on Computer and Information Technology (ICCIT 2011). A new approach for Isb based image steganography using secret key (IEEE, Dhaka, 2011), pp. 286–291
6. M. Mishra, G. Tiwari, A. K. Yadav, in International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014). Secret communication using public key steganography (IEEE, Florida, 2014), pp. 1–5
7. A. Kiyas, A. Russell, N. Shashidhar, In International Workshop on Information Hiding. Key-efficient steganography (Springer, Berkeley, 2012), pp. 142–159
8. D. C. Wu, W. H. Tsai, Spatial-domain image hiding using image differencing. IEE Proc.-Vis., Image and Signal Process. 147(1), 29–37 (2000)
9. H. Yang, A. C. Kot, S. Rahardja, X. Jiang, in 2007 IEEE International Conference on Multimedia and Expo. High capacity data hiding for binary images in morphological wavelet transform domain (IEEE, Beijing, 2007), pp. 1239–1242
10. P. Thitimagjishima, Y. Thitimagjishima, Y. Rangsanseri, in Proc. IEEE Region 10 Int. Conf. Electr. Electron. Technol. TENCON 2001 (Cat. No. 01CH37239). Hiding confidential signature into digital images via frequency domain, vol. 1 (2001), pp. 246–249
11. C. K. Chan, L. M. Cheng, Hiding data in images by simple Isb substitution. Pattern Recognit. 37(3), 469–474 (2004)
12. T. Pevny, T. Filler, P. Bas, in International Workshop on Information Hiding. Using high-dimensional image models to perform highly undetectable steganography (Springer, Berlin, 2010), pp. 161–177
13. V. Holub, J. Fridrich, T. Denemark, Universal distortion function for steganography in an arbitrary domain. EURASIP J. Inf. Secur. 2014(1), 1 (2014)
14. T. Pevny, P. Bas, J. Fridrich, Steganalysis by subtractive pixel adjacency matrix. IEEE Trans. Inf. Forensic Secur. 5(2), 215–224 (2010)
15. C. Qin, Q. Zhou, F. Cao, J. Dong, X. Zhang, Flexible lossy compression for selective encrypted image with image inpainting. IEEE Trans. Circ. Syst. Video Technol. 29(11), 3341–3355 (2019)
16. C. Qin, J. C. C. Chang, J. Dong, X. Sun, Non-uniform watermark sharing based on optimal iterative BTC for image tampering recovery. IEEE Multimedia. 25(3), 36–48 (2018)
17. C. Qin, W. Zhang, F. Cao, X. Zhang, C.-C. Chang, Separable reversible data hiding in encrypted images via adaptive embedding strategy with block selection. Signal Processing. 153, 109–122 (2018)
18. R. Chandramouli, G. Li, N. D. Memon, Adaptive steganography. Proc. SPIE. 4675, 69–78 (2002)
19. J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images. IEEE Trans. Inf. Forensic Secur. 7(5), 868–882 (2012)
20. V. Holub, J. Fridrich, Random projections of residues for digital image steganalysis. IEEE Trans. Inf. Forensic Secur. 8(12), 1996–2006 (2013)
21. Y. Ma, X. Luo, X. Li, Z. Bao, Y. Zhang, Selection of rich model steganalysis features based on decision rough set a-positive region reduction. IEEE Trans. Circ. Syst. Video Technol. 29(2), 336–350 (2019)
22. C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, Spatial-temporal attention mechanism for video captioning. IEEE Trans. Multimedia. 22(1), 229–241 (2020)
23. C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, F. Xu, 3d room layout estimation from a single RGB image. IEEE Trans. Multimed. Early Access, 1–1 (2020)
24. C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval. IEEE Trans Pattern Anal Mach Intell. 2020
25. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Generative adversarial nets. Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems, (Montreal, 2014), pp. 2672–2680
26. I. Gujaljani, F. Ahmed, M. Arjosky, V. Dumoulin, A. C. Courville. Improved training of Wasserstein GANs. Advances in Neural Information Processing Systems, (State of California, 2017), pp. 5767–5777
27. M. Arjosky, S. Chintala, L. Bottou, in International Conference on Machine Learning. Wasserstein generative adversarial networks, (Sydney, 2017), pp. 214–223
28. Y. Rubner, C. Tomasi, L. J. Guibas, The earth mover’s distance as a metric for image retrieval. Int. J. Comp. Vision. 40(2), 99–121 (2000)
29. D. A. Edwards, On the Kantorovich-Rubinstein theorem. Expo. Math. 29(4), 387–398 (2011)
30. K. Liu, Varying k-lipschitz constraint for generative adversarial networks. arXiv preprint arXiv:1803.06107 (2018)
31. G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, in 12th International Conference on Computer Vision. Attribute and simile classifiers for face verification, (Kyoto, 2009), pp. 365–372
32. Z. Liu, P. Luo, X. Wang, X. Tang, in Proceedings of the IEEE International Conference on Computer Vision. Deep learning face attributes in the wild (IEEE, Santiago, 2015), pp. 3730–3738
33. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, in 2009 IEEE Conference on Computer Vision and Pattern Recognition. ImageNet: a large-scale hierarchical image database (IEEE, Miami, 2009), pp. 248–255
34. J. Xu, X. Mao, X. Jin, A. Jaffer, S. Liu, L. Li, M. Toyoura, Hidden message in a deformation-based texture. Vis. Comput. 31(12), 1653–1669 (2015)
35. Z. Zhou, Y. Cao, X. Sun, Coverless information hiding based on bag-of-words model of image. J. Appl. Sci. 34(5), 527–536 (2016)
36. Z. Zhang, J. Liu, Y. Ke, Y. Lei, J. Li, M. Zhang, X. Yang, Generative steganography by sampling. IEEE Access. 7, 118586–118597 (2019)
37. Z. Zhou, H. Sun, R. Harit, X. Chen, X. Sun, in International Conference on Cloud Computing and Security. Coverless image steganography without embedding (Springer International Publishing, Nanjing, 2015), pp. 123–132
38. H. Otori, S. Kuriyama, Texture synthesis for mobile data communications. IEEE Comput. Graph. Appl. 29(6), 74–81 (2009)

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.