 CMC, vol.55, no.3, pp.483-493, 2018

Coverless Steganography for Digital Images Based on a Generative Model

Xintao Duan[1, *], Haoxian Song[1], Chuan Qin[2] and Muhammad Khurram Khan[3]

Abstract: In this paper, we propose a novel coverless image steganographic scheme based on a generative model. In our scheme, the secret image is first fed to the generative model database, to generate a meaning-normal and independent image different from the secret image. The generated image is then transmitted to the receiver and fed to the generative model database to generate another image visually the same as the secret image. Thus, we only need to transmit the meaning-normal image which is not related to the secret image, and we can achieve the same effect as the transmission of the secret image. This is the first time to propose the coverless image information steganographic scheme based on generative model, compared with the traditional image steganography. The transmitted image is not embedded with any information of the secret image in this method, therefore, can effectively resist steganalysis tools. Experimental results show that our scheme has high capacity, security and reliability.

Keywords: Generative model, coverless image steganography, steganalysis, steganographic capacity, security.

1 Introduction

Most of current information steganographic techniques [Qin, Ji, Chang et al. (2018); Ma, Luo, Li et al. (2018); Qin, Chang and Hsu (2015); Zhou, Sun, Harit et al. (2015); Zhou, Wu, Yang et al. (2017); Xia, Li and Wu (2017)] apply the cover data (such as digital image, audio and video) as a disguise for the secret data to be transmitted, which embed the secret data into cover data. The popularization of personal computers and the proliferation of digital images on the Internet provide convenient conditions of cover data for conducting information Steganography [Qin, Ji, Zhang et al. (2017); Qin, Chang and Chiu (2014)]. However, on the other hand, the technique for detecting hidden data, also called as steganalysis, has also been rapidly developed, which is mainly based on finding statistical anomaly of cover data caused by data embedding. Hence, steganalysis can be considered as a serious threat to steganography. According to different hiding strategies [Zhang, Qin, Zhang et al. (2018); Qin, Ji, Zhang et al. (2017); Qin, Chang and Hsu (2015)], the commonly used steganographic schemes are classified into two types: Spatial domain schemes and transform domain schemes. The spatial domain hiding method has

[1] Henan Normal University, Xinxiang, Henan 453007, China.

[2] University of Shanghai for Science and Technology, Shanghai 200093, China.

[3] Center of Excellence in Information Assurance (CoEIA), King Saud University, Riyadh, Saudi Arabia.

* Corresponding Author: Xintao Duan. Email: duanxintao@126.com.

the adaptive LSB hiding method [Yang, Weng, Wang et al. (2008)], the spatial adaptive steganography algorithm S-UNIWARD [Holub, Fridrich and Denemark (2014)], HUGO [Pevny, Filler and Bas (2010)], WOW [Holub and Fridrich (2012)] and so on. The transform domain method is to modify the host image data to change some statistical features to achieve data hiding, such as the hidden method in DFT (discrete Fourier transform) domain [Ruanaidh, Dowling and Boland (1996)], DCT (discrete cosine transform) domain [Cox, Kilian, Leighton et al. (1997)], and DWT (discrete wavelet transform) domain [Lin, Horng, Kao et al. (2008)]. These methods inevitably leave some modifications to the carrier [Yuan, Xia and Sun (2017); Chen, Chen and Wu (2017)]. In order to fundamentally resist the detection of various detection algorithms, this paper presents a new coverless image information hiding method based on generative model. As shown in Fig. 1, we only need to deliver a meaning-normal image which is not related to the secret image to the receiver, so that the receiver can generate an image visually the same as the secret image without worrying about the analysis of the steganography, even less attack.

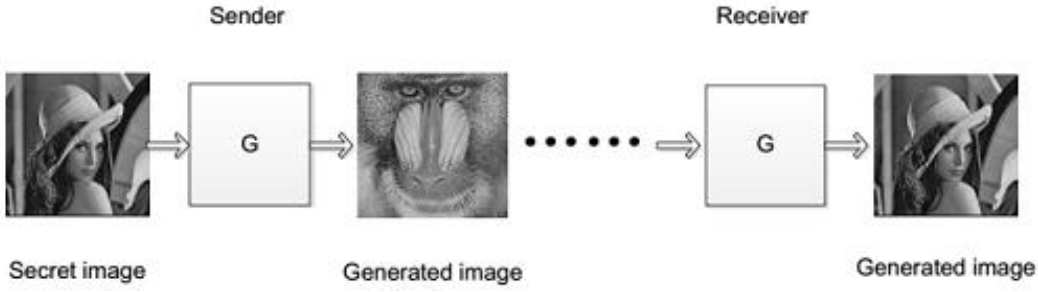

Figure 1: The framework of the research content

As mentioned above, we propose a new scheme to hide image information, which can generate visually the same image as secret image by sending a generated image that is not related to the secret image. The transmitted image is only a normal-meaningful image rather than the image which is embedded any secret information, and also achieve the same effect as transferring the secret image. This method can effectively resist steganalysis tools, and greatly improves the security of the image. To summarize, the major contributions of our work as below:

(1) We do not need to pass the secret image. On the contrary, we transmit a meaning-normal image which is completely unrelated to the secret image. This method has high security.

(2) The image we transmit does not embed any secret information, it is a normal image, and the image steganographic analysis does not work.

(3) As long as the training is enough, this effect can be achieved and the capacity is large.

The rest of this paper is organized as follows. Section II reviews the related works about generative models. The proposed coverless steganographic scheme for digital images is described in Section III. Experimental results and analysis are given in Section IV, and Section V concludes the paper.

2 Related works

Restricted Boltzmann Machines (RBMs) [Smolensky (1986)], deep Boltzmann machines (DBMs) [Srivastava and Salakhutdinov (2012)] and their numerous variants are undirected graphical models with latent variables. The interactions within such models are represented as the product of unnormalized potential functions, normalized by a global summation or integration over all states of the random variables. This quantity and its gradient are intractable for all but the most trivial instances, although they can be estimated by Markov chain Monte Carlo (MCMC) methods. Mixing poses a significant problem for learning algorithms that rely on MCMC [Bengio, Mesnil, Dauphin et al. (2013); Bengio, Laufer, Alain et al. (2014)]. Deep belief networks (DBNs) [Hinton, Osindero and Teh (2006)] are hybrid models containing a single undirected layer and several directed layers. While a fast approximate layer-wise training criterion exists, DBNs incur the computational difficulties associated with both undirected and directed models. Variational Auto-Encoders (VAEs) [Glorot, Bordes and Bengio (2012)] and Generative Adversarial Networks (GANs) [Bengio, Yao, Alain et al. (2013)] are well known to us. VAEs focus on the approximate likelihood of the examples, and they share the limitation of the standard models and need to fiddle with additional noise terms. Ian Goodfellow put forward GAN [Goodfellow, Pougetabadie, Mirza et al. (2014)] in 2014. Goodfellow theoretically proved the convergence of the algorithm, and when the model converges, the generated data has the same distribution as the real data. GAN provides a new training idea for many generative models and has hastened many subsequent works. GAN takes a random variable (it can be Gauss distribution, or uniform distribution between 0 and 1) to carry on inverse transformation sampling of the probability distribution through the parameterized probability generative model (it is usually parameterized by a neural network model). Then a generative probability distribution is obtained. The GAN model includes a generative model G and a discriminative model D. The training objective of the discriminative model D is to maximize the accuracy of its own discriminator, and the training objective of generative model G is to minimize the discriminator accuracy of the discriminative model D. The objective function of GAN is a zero-sum game between D and G and also a minimum -maximization problem. GAN adopts a very direct way of alternate optimization, and it can be divided into two stages. In the first stage, the discriminative model D is fixed, the generative model G is optimized to minimize the accuracy of the discriminative model. In the second stage, the generative model G is the fixed in order to improve the accuracy of the discriminative model D. As a generative model, GAN is directly applied to modeling of the real data distribution, including generating images, videos, music and natural sentences, etc. Because of the mechanism of internal confrontation training, GAN can solve the problem of insufficient data in some traditional machine learning. GANs offer much more flexibility in the definition of the objective function, including Jensen-Shannon, and all f-divergences [Hinton, Srivastava, Krizhevsky et al. (2012)] as well as some exotic combinations. Therefore, it can be used in semi-supervised learning, unsupervised learning, multi-view learning and multi-tasking learning. In addition, it has been successfully used in reinforcement learning to improve its learning efficiency. Although GAN is applied widely, there are some problems with GAN, difficulty in training, lack of diversity. Besides, generator and discriminator cannot indicate the training process. On

the other hand, training GANs is well known for being delicate and unstable. The better discriminator is trained, the more serious gradient of the generator disappears, leading to gradient instability and insufficient diversity. WGAN (Wasserstein Generative Adversarial Networks [Arjovsky and Bottou (2017); Arjovsky, Chintala and Bottou (2017)]) is an improvement to GAN, and it applies Wasserstein distance instead of JS divergence in the GAN. Compared to KL divergence and JS divergence, the advantage of Wasserstein distance is that it can still reflect their distance even if there is no overlap between the two distributions. At the same time, the problem of training stability and process indicating are solved.

Therefore, this paper chooses Wasserstein GAN so as to guarantee training stability instead of GAN. It is no longer necessary to carefully balance the training extent between generator and discriminator. It basically solves the problem of collapse mode and ensures the diversity of samples.

3 Proposed scheme

The WGAN model is applied to generate the handwritten word by feeding the random noise *z*, but when the random noise *z* is changed to a secret image *img*, the model can still generate the meaning-normal and independent image *IMG'* which is not related to the secret image we want to transmit. These several images taken from the standard set of images were evaluated in the paper, they are *Lena*, *Baboon*, *Cameraman* and *Peppers*, and they have the same size as 256 by 256. The feed is the secret image, and we train the generative model database through the WGAN, then it can generate a meaning-normal and independent image which is not related to the secret image. So we transmit the meaning-normal image to the receiver, and this generated image is fed to the generative model database to generate another generated image visually the same as the secret image. The flow charts of the whole experiment are shown in Fig. 2 and Fig. 3.

Figure 2: The flow chart of WGAN

Figure 3: The flow chart of generative model

D and G play the following two-player minimax game with value function $V(G;D)$ in WGAN:

$$\min_G \max_D V(D,G) = E_{x\sim pdata(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{1}$$

$D(x)$ represents the probability that x came from the data rather than pg. We train D to maximize the probability of assigning the correct label to both training examples and samples from G. We simultaneously train G to minimize $\log(1-D(G(z)))$. We first make a gradient ascent step on D and then a gradient descent step on G, then the update rules are:

Keeping the G fixed, update the model D by $\theta_D \leftarrow \theta_D + \gamma_D \nabla_D L$ with

$$\nabla_D L = \frac{\partial}{\partial \theta_D}\{E_{x\sim Pdata(x)}[\log D(x,\theta_D)] + E_{z\sim Pnoise(z)}[\log(1-D(G(z,\theta_G),\theta_D))]\} \tag{2}$$

Keeping the D fixed, update the model G by $\theta_G \leftarrow \theta_G - \gamma_G \nabla_G L$ where

$$\nabla_G L = \frac{\partial}{\partial \theta_G} E_{z\sim Pdata(z)}[\log(1-D(G(z,\theta_G),\theta_D))] \tag{3}$$

Wasserstein distance is also called the EM (Earth-Mover) distance

$$W(P_r,P_g) = \inf_{\gamma\sim\prod(P_r,P_g)} E_{(x,y)\sim\gamma}[\|x-y\|] \tag{4}$$

Where $\prod(P_r,P_g)$ denotes the set of all joint distributions $\gamma(x,y)$ whose marginal are respectively P_r and P_g. Intuitively, $\gamma(x,y)$ indicates how much "mass" must be transported from x to y in order to transform the distributions P_r into the distribution P_g. The EM distance then is the "cost" of the optimal transport plan.

4 Experimental results and analysis

In this paper, 5,000 images are randomly selected from the CelebA dataset to experiment, and the results show that the coverless image information steganography based on generative model method can be implemented well. The sender and receiver share the same dataset and the same parameters. As shown in Fig. 4 and Fig. 5, we feed the secret image *img* into the generative model, generating the meaning-normal and independent *IMG'* which is not related to the secret image we want to transmit.

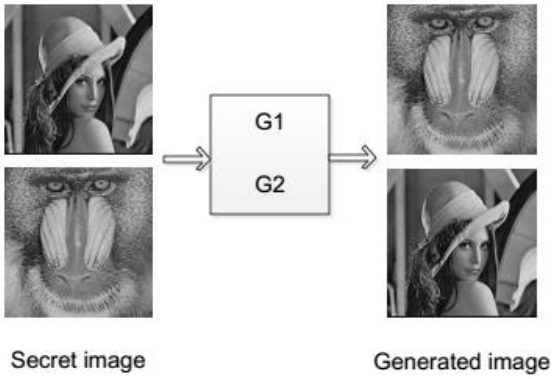

Figure 4: Training generative model G1, G2

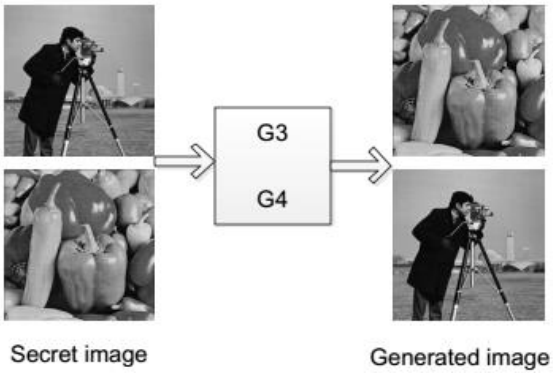

Figure 5: Training generative model G3, G4

As shown above, we choose *Lena* as the secret image *img*, it can generate the *IMG'* visually the same as *Baboon* we want to transmit. In the meantime, we also trained *Baboon* to generate the *IMG'* visually the same as *Lena* through the WGAN. We save the corresponding generative model G1 and G2 of generating visually the same as *Baboon* and *Lena* respectively. Using the same method, we take the *Cameraman* and *Peppers* as the secret image to experiment respectively, and they can generate corresponding *Peppers* and *Cameraman*. We also save the corresponding generative model G3 and G4 of generating visually the same as *Peppers* and *Cameraman* respectively, and apply them to the next experiment, instead of the WGAN. We put the generative model G1, G2, G3 and G4 of generating visually the same as *Baboon*, *Lena*, *Peppers* and *Cameraman* in a database respectively, so that the generative model database is built. Since both the sender and the receiver train well the generative model database, we perform experiments as shown in Fig. 6 and Fig. 7.

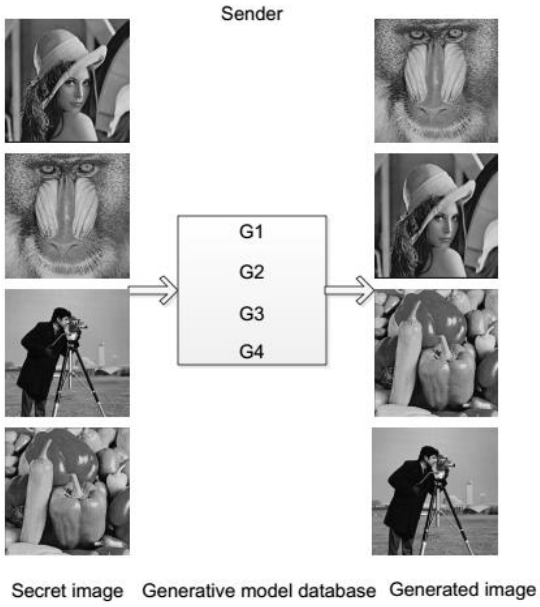

Figure 6: Training generative model database for sender

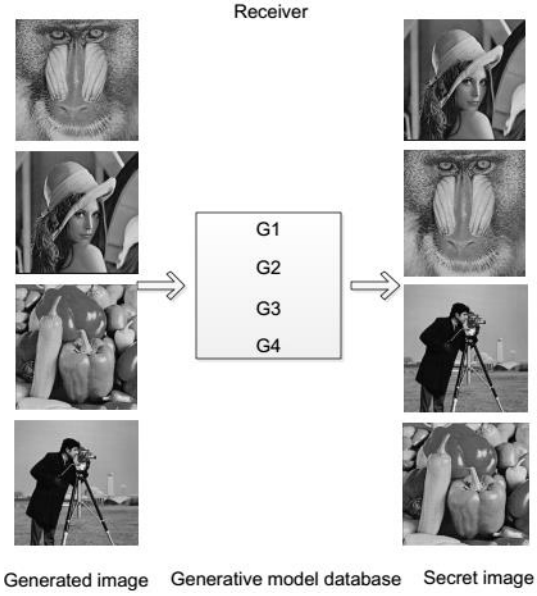

Figure 7: Training generative model database for receiver

As shown above, when the sender wants to transmit the secret image *Lena*, the generated image *Baboon* can be transmitted to the receiver to generate a generated image visually the same as the secret image *Lena*, similarly, if you want to transmit *Baboon*, you can transmit the generated image *Lena*, if you want to transmit *Cameraman*, you can transmit the generated image *Peppers*, if you want to transmit *Peppers*, you can transmit the generated image *Cameraman*. In this experiment, we have successfully achieved the

effect of coverless image steganographic scheme based on a generative model by feeding a secret image to generate a meaning-normal and independent image which is not related to the secret image we want to transmit, and when the secret image is given, the transmitted image is unique and specific. Consequently, the image steganographic scheme proposed in this paper is feasible. In practical application, we are more concerned with the content of the image rather than the pixels in addition to professional image workers, this scheme can produce a meaning-normal and independent image which is not related to the secret image we want to transmit, which can satisfy most requirements, thereby, we suppose that if you want to send a secret image, you only need to transmit a meaning-normal and independent image to the receiver, the receiver only need to feed transmitted image to the generative model database, generate an image visually the same as the secret one, no needing direct transmission of the secret image. Besides, the transmitted image does not embed any information of the secret image, so it does not give visual cues to attackers, and the image steganography analysis does not work. This scheme can resist detection of all the existing steganalysis tools, and improve the security of the image.

The experimental results show that the image is completely different from the secret image based on the method of generative model. The attacker cannot know what the secret image to be transmitted is, and the generated image is visually the same as the secret image, which meet the practical application standard, In addition to the visually qualitative analysis, the histogram of Fig. 8 can also obtain the same quantitative analysis results.

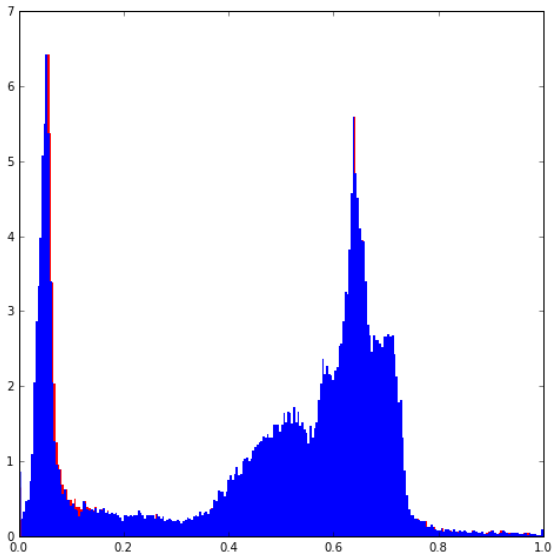

Figure 8: The histogram of the generated image and secret image distribution

As shown in Fig. 8, the red portion represents the secret image, and the blue portion represents the generated image. It can be seen from this histogram that the distribution of the generated images and the secret images are almost identical, and the small differences are almost negligible.

5 Conclusion

To sum up, the paper proposed the coverless image steganographic scheme based on a generative model. An image visually the same as the secret image is generated by transmitting a normal-meaningful image to the receiver. A fed image corresponds uniquely to a secret image. This method is practical. Therefore, it can be applied to image steganography and image protection.

Acknowledgement: This paper was supported by the National Natural Science Foundation of China (No. U1204606), the Key Programs for Science and Technology Development of Henan Province (No. 172102210335), Key Scientific Research Projects in Henan Universities (No. 16A520058).

References

Arjovsky, M.; Bottou, L. (2017): Towards principled methods for training generative adversarial networks. *International Conference on Learning Representations.*

Arjovsky, M.; Chintala, S.; Bottou, L. (2017): Wasserstein generative adversarial networks. *Proceedings of the 34nd International Conference on Machine Learning*, vol. 70, pp. 214-223.

Bengio, Y.; Laufer, E.; Alain, G.; Yosinski, J. (2014): Deep generative stochastic networks trainable by backprop. *International Conference on Machine Learning*, pp. 226-234.

Bengio, Y.; Mesnil, G.; Dauphin, Y. N.; Rifai, S. (2013): Better mixing via deep representations. *International Conference on Machine Learning*, pp. 552-560.

Bengio, Y.; Yao, L.; Alain, G.; Vincent, P. (2013): Generalized denoising auto-encoders as generative models. *Advances in Neural Information Processing Systems*, pp. 899-907.

Chen, X.; Chen, S.; Wu, Y. (2017): Coverless information hiding method based on the Chinese character encoding. *Journal of Internet Technology*, vol. 18, no. 2, pp. 313-320.

Cox, I. J.; Kilian, J.; Leighton, F. T.; Shamoon, T. G. (1997): Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673-1687.

Glorot, X.; Bordes, A.; Bengio, Y.; Glorot, X.; Bordes, A. et al. (2012): Deep sparse rectifier neural networks. *International Conference on Artificial Intelligence and Statistics*, pp. 315-323.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D. et al. (2014): Generative adversarial nets. *International Conference on Neural Information Processing Systems*, pp. 2672-2680.

Hinton, G. E.; Osindero, S.; Teh, Y. W. (2006): A fast learning algorithm for deep belief nets. *Neural Computation*, vol. 18, no. 7, pp. 1527-1554.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. R. (2012): Improving neural networks by preventing co-adaptation of feature detectors.

Computer Science, vol. 3, no. 4, pp. 212-223.

Holub, V.; Fridrich, J. (2012): Designing steganographic distortion using directional filters. *IEEE International Workshop on Information Forensics and Security*, pp. 234-239.

Holub, V.; Fridrich, J.; Denemark, T. (2014): Universal distortion function for steganography in an arbitrary domain. *Eurasip Journal on Information Security*, vol. 2014, no. 1, pp. 1-13.

Lin, W. H.; Horng, S. J.; Kao, T. W.; Fan, P.; Lee, C. L. et al. (2008): An efficient watermarking method based on significant difference of wavelet coefficient quantization. *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 746-757.

Ma, Y.; Luo, X.; Li, X.; Bao, Z.; Zhang, Y. (2018): Selection of rich model steganalysis features based on decision rough set α-positive region reduction. *IEEE Transactions on Circuits & Systems for Video Technology*, pp. 1.

Pevny, T.; Filler, T.; Bas, P. (2010): Using high-dimensional image models to perform highly undetectable steganography. In Böhme, R.; Fong, P. W. L.; Safavi-Naini, R. (Eds.): *Information Hiding*, pp. 161-177. Springer Berlin Heidelberg.

Qin, C.; Chang, C.; Chiu, Y. (2014): A novel joint data-hiding and compression scheme based on smvq and image inpainting. *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 969-978.

Qin, C.; Chang, C. C.; Hsu, T. J. (2015): Reversible data hiding scheme based on exploiting modification direction with two steganographic images. *Multimedia Tools & Applications*, vol. 74, no. 15, pp. 5861-5872.

Qin, C.; Chen, X.; Luo, X.; Zhang, X.; Sun, X. (2018): Perceptual image hashing via dual-cross pattern encoding and salient structure detection. *Information Sciences*, vol. 423, pp. 284-302.

Qin, C.; Ji, P.; Chang, C. C.; Dong, J.; Sun, X. (2018): Non-uniform watermark sharing based on optimal iterative BTC for image tampering recovery. *IEEE Multimedia*.

Qin, C.; Ji, P.; Zhang, X.; Dong, J.; Wang, J. (2017): Fragile image watermarking with pixel-wise recovery based on overlapping embedding strategy. *Signal-Processing*, vol. 138, pp. 280-293.

Qin, C.; Zhang, X. (2015): Effective reversible data hiding in encrypted image with privacy protection for image content. *Journal of Visual Communication & Image Representation*, vol. 31, pp. 154-164.

Ruanaidh, J. J. K. O.; Dowling, W. J.; Boland, F. M. (1996): Phase watermarking of digital images. *Proceedings of 3rd IEEE International Conference on Image Processing*, vol. 3, pp. 239-242.

Smolensky, P. (1986): Information processing in dynamical systems: Foundations of harmony theory. *In Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pp. 194-281. MIT Press, Cambridge, MA.

Srivastava, N.; Salakhutdinov, R. (2012): Multimodal learning with deep boltzmann machines. *International Conference on Neural Information Processing Systems*, pp. 2222-2230.

Wu, Y.; Chen, X.; Sun, X. (2018): Coverless steganography based on english texts using binary tags protocol. *Journal of Internet Technology*, vol. 19, no. 2, pp. 599-606.

Xia, Z.; Li, X.; Wu, Y. (2017): Coverless information hiding method based on LSB of the character's unicode. *Journal of Internet Technology*, vol. 18, no. 6, pp. 1353-1360.

Yang, C.; Weng, C.; Wang, S.; Sun, H. (2008): Adaptive data hiding in edge areas of images with spatial lsb domain systems. *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 3, pp. 488-497.

Yuan, C.; Xia, Z.; Sun, X. (2017): Coverless image steganography based on SIFT and BOF. *Journal of Internet Technology*, vol. 18, no. 2, pp. 435-442.

Zhang, Y.; Qin, C.; Zhang, W.; Liu, F.; Luo, X. (2018): On the fault-tolerant performance for a class of robust image steganography. *Signal Processing*, vol. 146, pp. 99-111.

Zhou, Z.; Sun, H.; Harit, R.; Chen, X. Y.; Sun, X. M. (2015): Coverless image steganography without embedding. *International Conference on Cloud Computing and Security*, pp. 123-132.

Zhou, Z.; Wu, Q. M. J.; Yang, C. N.; Sun, X. M.; Pan, Z. Q. (2017): Coverless image steganography using histograms of oriented gradients-based hashing algorithm. *Journal of Internet Technology*, vol. 18, no. 5, pp. 1177-1184.