FakeSafe: Human Level Steganography Techniques by Disinformation Mapping using Cycle-consistent Adversarial Network

HE ZHU¹, DIANBO LIU.²
¹Mila - Quebec AI Institute (e-mail: he.zhu@mila.quebec)
²Mila - Quebec AI Institute (e-mail: dianbo.liu@mila.quebec)

ABSTRACT Steganography is the task of concealing a message within an overt medium such that the presence of the hidden message is barely detectable. Recently, myriads of works have introduced the inchoate techniques of deep learning to the field of steganography. Nevertheless, existing issues like small payload capacity and image distortion have exceedingly suffocated the steganographic research. In this paper, we propose FakeSafe, a novel cycle-consistent adversarial network proffering human-level steganography. Mapping the confidential information into fake messages, FakeSafe efficaciously precludes the detection of steganalysis algorithms and human eyes. There are three contributions in our work: (i) we construct a multi-step FakeSafe mapping, which significantly impedes the steganalysis models to identify and recover the hidden message; (ii) our steganographic models are robust enough since they are applicable to multifarious data domains, including image and text information; (iii) we introduce a coverless solution to embed the clandestine message within a medium of a specific type in lieu of a dedicated cover. We have conducted experiments using both benchmark and real-world data sets to demonstrate potential applications of FakeSafe, whose open source library is available online at: https://github.com/mikemikezhu/fake-safe

INDEX TERMS Generative Adversarial Networks, Machine Learning, Privacy

I. INTRODUCTION

STEGANOGRAPHY is the art of covered writing. The term dates back to at least 440 B.C. when the Early Greek sovereigns would shave off the hair of slaves, tattoo the furtive messages onto their scalps, and wait for their hair to regrow to shroud the messages to be delivered [1]. Although the avant-garde steganographic techniques used nowadays are drastically intricate compared with those in the ancient time, the underlying principles are still identical: cloaking the information in an overt medium such that its presence is virtually undetectable.

The modern steganography is pervasively applicable in various fields. In medicine, steganography could efficaciously conceal patients’ private information in images such as X-rays and MRIs [2]. Moreover, steganographic techniques are instrumental in transmitting secret information, watermark, and copyright certification in the media sphere [3].

The steganography algorithms are predominantly measurable by their capacity, distortion, and secrecy. Specifically, the capacity refers to the amount of data that can be concealed inside a medium, whose evaluation metric, bits-per-pixel (bpp), essentially calculates the average number of bits embedded into each cover image’s pixel. The distortion, on the other hand, gauges the similarity between the steganography image and its corresponding cover image. Furthermore, the secrecy describes the capability to forestall detection by naked eyes or steganalysis tools.

To ameliorate the steganography algorithms, a slew of researchers have made tremendous progress in the steganographic research. Traditional approaches to image steganography conceal the secret messages in either spatial or transform domain [4]. Marvelous invisibility and security though they have achieved, these methods are primarily shackled by the small payload capacity of approximately 0.4 bpp [5]. Fortunately, the explosive development of deep learning has inspired myriads of researchers to use nascent technology of generative adversarial networks (GANs) in steganography.
For example, Volkhonskiy first proposed a steganography enhancement algorithm SGAN based on GANs [8]. Similarly, Shi produced SSGAN based on WGAN and achieved a preferable outcome [9]. Nonetheless, the images generated by both SGAN and SSGAN are warping in semantics, thus potentially sabotaging the model’s security.

Meanwhile, Tang put forward another steganography method ASDL-GAN, where the generator could train itself to discover the most apposite pixels to embed secret data [10]. Though progressively enhanced, the model still has low capacity, and its security has not even transcended the conventional steganography algorithms such as S-UNIWARD [4]. Baluja proposed a convolutional neural network with a peculiar encoder-decoder architecture [11]. Although the model has accomplished large capacity and strong invisibility, the generated images are distorted in colour, thus conspicuously detectable by either human eyes or steganalysis tools. Atique also introduced a comparable encoder-decoder model [12], but their steganography images are yellowing and effortlessly recognisable, which remarkably stifes the steganographic performance.

Inspired by previous steganographic research, we combined GANs and consistency loss to develop a novel steganography method named FakeSafe, which maps the private information onto a fake message visually indistinguishable from the real messages. Besides, FakeSafe can be employed during data transfer, data storage, data usage or other scenarios to consolidate traditional encryption and security technologies.

Our contributions through this paper are:

- We construct a multi-step FakeSafe mapping with a cascade of stenographic functions, which significantly ensures the safety of sensitive data. Even if the attackers know the message is fake, they may not recognise how many steps the messages were mapped.
- We design a steganography method applicable to various data domains, including image and text information. The fake message can be either from the same domain as the original private information or from a completely different domain, which drastically enhances the framework’s robustness.
- We introduce a coverless solution to conduct steganography. Unlike the conventional steganography methods, which require a dedicated cover for secret information embedding, our model enshrouds the hidden messages in the medium of a particular category. This approach greatly satiates the demands of those who wish to simplify the steganographic procedure without a premeditated container.

The rest of the paper is organised as follows. Section 2 introduces the motivations and architecture details of FakeSafe. Section 3 conducts proof-of-concept experiments to explore how well FakeSafe can help protect private information at the human level and the quality of reconstructed data from the fake domain. Section 4 demonstrates multi-step FakeSafe mapping. Finally, Section 5 concludes the paper with relevant discussion, followed by the appendix in Section 6.

II. FAKESAFE

A. MOTIVATION AND FORMULATION

The FakeSafe method aims to map the original private information onto a fake but realistically looking message. The method consists of two parts: 1) a function $F$ that maps a private message $X$ into a fake message $X^{\text{fake}}$, i.e., $X^{\text{fake}} = F(X)$. $X^{\text{fake}}$ can be from the same domain as $X$, such as a human face image, or a completely different domain. 2) a reconstruction function $R$ that maps the fake message back to the original message. $F$ and $R$ are specific to each data set.

B. SYSTEM

We assume a sender of private information and a targeted receiver of information in our steganographic system. The data transfer and storage infrastructure is not 100% safe, and malicious attackers might surreptitiously purloin private information from it. Only the data sender can access the function $F$ to map the private data to the fake domain using steganography algorithms. Meanwhile, only the targeted receiver is accessible to function $R$ to recover the original data. Even if the eavesdroppers fortuitously procure fake data $X^{\text{fake}}$, without additional information, it is exceedingly gruelling for them to recognise that the data is fake, because $X^{\text{fake}}$ is visually realistic. Even if both $X$ and $X^{\text{fake}}$ are from the same data domain (e.g. mapping a set of human faces into another set of human faces), it is still taxing for the stealthy snoops to puzzle out the original data without an appropriate mapping function.

C. GENERATIVE ADVERSARIAL NETWORKS(GAN) WITH CYCLE CONSISTENCY LOSS

GAN was adopted as the steganography function $F$ to map private information to fake message in this study based on its marvellous performance in generating fake data sets visually realistic to humans. Generative model $F(X)$ is trained against discriminator $D$ to make the outputs of $X^{\text{fake}} = F(X)$ visually indistinguishable from the samples used to train $D$. We name this $X^{\text{fake}}$ messages as FakeSafe messages. $D$ and $F$ were trained in an alternating manner. The objective loss function for training generator and discriminator is:

$$L(x, x^{\text{fake}}, F, D) = E[\log(D(X^{\text{fake}}))] + E[\log(1 - D(F(x)))]$$

$F$ generates data points that indistinguishably resemble real data in the fake message domain. Least loss was used to train the GAN on account of its reported stability [7]. Therefore, when training GAN, we train $F$ to minimize $E[1 - D(F(x))]^2$ and train $D$ to minimize $E[D(F(x))]^2$. After the other party receives the FakeSafe message, it will be recovered using a trained model $R$ such that $R(F(X)) \approx X$. To train $R$ to retrieve the original message from $X^{\text{fake}}$, we define the reconstruction objective as:

$$L(x^{\text{fake}}, x, R, F) = E[\log(R(x^{\text{fake}}))]$$

In practice, we train $F$ and $D$ alternately, then $R$ is fine-tuned using $X^{\text{fake}}$ and $X$. This cycle continues until the reconstruction error is negligible.
FIGURE 1. Human level data protection by mapping original information into fake data domain (disinformation). The original information can be recovered from fake messages using a trained reconstruction function.

cycle-consistency was used to make reconstructed data $\hat{R}(X_{fake})$ matching the original data $X$. The loss function is

$$L(F, R, X) = E[|R(F(X)) - X|]$$

For reconstruction errors, we used the absolute loss. A fully connected neural network with leaky ReLU was employed in both generator and discriminator models for simplicity.

D. MODEL IMPLEMENTATION

As this is a proof-of-concept study, 1) for the image-image generator model, we adopted a plain three-layer fully connected neural network with 256, 512 and 1024 units. 2) for the text-image generator model, we used a four-layer fully connected neural network with 64, 256, 512 and 1024 units. Leaky ReLU was employed as the activation functions for hidden layers, and batch normalization was applied in both image-image and text-image generator model. 3) for the image-text generator model, we used a four-layer fully connected neural network with 128, 256, 512 and 1024 units. Leaky ReLU was also used as the activation function, and a dropout with rate 0.2 was introduced to preclude over-fitting. The adam optimization with a learning rate of 0.0002 was used in the above 3 cases.

III. EXPERIMENTS AND RESULTS

To unravel whether our FakeSafe method is potent enough in steganography, we conducted three types of proof-of-concept experiments:

- We encoded information into fake messages from the same data domain, using MNIST and MNIST fashion as an example.
- We encoded information into fake messages from a different domain, such as MNIST digits to MNIST fashion images.
- We explored the feasibility of multi-step FakeSafe encoding of information.

The reconstructing accuracy decreased as we increased the number of steps to encode information. We evaluate its potential values in a real-world application using a face video frame from an open-source clinical data set.

A. DATE SET

To conduct proof-of-concept experiments, four data sets were used in this study:

- MNIST hand written digits data set
- MNIST fashion data set
- Tatoeba English text data set
- The UNBC-McMaster Shoulder Pain Expression Archive Data set [13]

UNBC-McMaster Shoulder Pain is a real-world data set from the clinical setting, consisting of human face video frames from various individuals with shoulder pain. We utilise this data set as an example of the real use case of FakeSafe in the medical setting.

B. FAKESAFE MAPPING ONTO THE SAME DATA DOMAIN

Our potential application of FakeSafe is to map private information on other same data domain but different data points.

We conducted the following four experiments:

- MNIST→F→MNIST→R→MNIST
- Fashion→F→Fashion→R→Fashion
- Face→F→Face→R→Face
- Word→F→Word→R→Word

When conducting experiments on MNIST, Models $F, D, R$ were all trained using the training sets with images of 10 hand-written digits. Therefore, $X_{fake} = F(X)$ can be any possible number from the training set and might not have to be the same digits as $X$. As shown in figure III , the recovered images $R(F(X))$ have the same labels as the original images $X$, while the FakeSafe images $F(X)$ are different.
Similarly, when conducting experiments on MNIST fashion data set which contain objects from 10 different categories, such as "shoe" or "dress", \( R(F(X)) \) have the same labels as the original message \( X \) and could differ from labels of \( X^{Fake} = F(X) \).

When conducting experiments on human face images, original data \( X \) is a human face image which was mapped to another human image \( X^{Fake} = F(X) \) that could be from the same person or a different person.

When conducting experiments on English words, original data \( X \) is a 50-dimension word embeddings which was mapped to another 300-dimension word embeddings \( X^{Fake} = F(X) \) that could be from the same word or a different word.

Three evaluation metrics were used to gauge the quality of the reconstructed message \( R(F(X)) \).

- Reconstruction errors between \( R(F(X)) \) and \( X \) were calculated as mean squared errors.
- In order to examine whether the reconstructed messages \( R(F(X)) \) still look like from the same class or individual as \( X \) to human, we trained a classifier \( C \) on \( X \) in training set to classify their labels, i.e. the digits, fashion category or individual ID, and apply \( C \) onto reconstructed data \( R(F(X)) \). The accuracy, F1 score, precision and recall of \( C(R(F(X))) \) were compared with the original labels of \( X \).
- Concerning the image domains, we also introduced the Structure Similarity Index (SSIM) [14] and Peak Signal to Noise Ratio (PSNR) [15] to evaluate the similarity...
The MNIST→F→MNIST→R→MNIST FakeSafe experiment achieved a reconstruction error of 1.62, classifier precision of 0.90, recall of 0.80, F1 score of 0.81, SSIM of 0.74, and PSNR of 15.65. In the Fashion→F→Fashion→R→Fashion experiment (Table 1), FakeSafe achieved a reconstruction error of 1.2, classifier precision of 0.71, recall of 0.72, F1 score of 0.70, SSIM of 0.63, and PSNR of 17.02. The real-world human face image data set, FakeSafe method achieved a reconstruction error of 1.06, precision of 0.95, recall of 0.92, F1 score of 0.92, SSIM of 0.83, and PSNR of 25.74.

C. FAKESAFE MAPPING ONTO A DIFFERENT DATA DOMAIN
Concealing confidential data in fake messages of the same type is conducive to hidden information preservation by misleading the malicious eavesdroppers. Nevertheless, sometimes it is better not to expose the original information domain at all. Therefore, we conducted experiments to FakeSafe map information into the message in a different domain.

We conducted the following four experiments:
- MNIST→R→Fashion→F→MNIST
- Fashion→R→MNIST→F→Fashion
- Face→R→MNIST→F→Face
- Word→R→Fashion→F→Word

The performances are comparable to FakeSafe mapping onto the same data domain (Table 1).

Specifically, for the experiment Word→F→Fashion→R→Word experiment, we have tried two different approaches to map the original messages.
- In the first approach, we will tokenize the original messages, which are the English words, and then map the tokens to MNIST fashion images using FakeSafe. We will map the MNIST fashion images back to tokens during the decoding process, which will be eventually converted back to English words.
- In the second approach, we will first convert the words to word embeddings with 50 dimensions, using GloVe Word Embeddings, and then map the word embeddings to MNIST fashion images. We will use FakeSafe to map the MNIST fashion images back to word embeddings during the decoding process and then decode back to the original words by finding the word with the smallest cosine similarity with the decoded word embeddings.

It is noteworthy that, the second approach, which uses word embeddings as the original messages, is proved to achieve better performance than the first approach, which only uses word tokens as the original messages.

IV. DEEPER FAKESAFE MAPPING
To guarantee the safety of sensitive data, one may ask why not map the original private multiple times using a cascade of different \( F \) functions, so that even if the attacker knows the message is fake, he or she will not know how many steps the messages were mapped. In order to explore the feasibility of deeper FakeSafe mapping, we conducted a series of experiments of 2-step and 3-step FakeSafe using MNIST, fashion and face images (Figure III-B and Table 2). Our results suggest that even it is possible to conduct multi-step FakeSafe mapping, the reconstruction error increased, and classification accuracy decreased dramatically.

V. CONCLUSION
In this article, we propose a method, named FakeSafe, to proffer human-level steganography by mapping each data point into a fake message visually realistic to human. We utilised GANs with cycle-consistency to build a function to map the original data to fake message and another function to map the fake message back to the original data. Both functions are data set specific and are easily tractable for the other data sets. FakeSafe method furnishes users the flexibility to map private data onto various data domains depending on use cases. FakeSafe can be efficiently utilised in combination with conventional data protection technologies but focuses on human-level steganography which considers human factors in data security and privacy protection.

APPENDIX
Furthermore, we have also conducted a supplementary FakeSafe experiment case Sentence→F→Fashion→R→Sentence, the performance of which is demonstrated in Table 3.

We have tried two different approaches to conduct the
- In the first approach, we have trained a Seq2Seq model using the GRU layer, which will encode the sentence sequence to the internal hidden states, and then decode back to the original sentence sequence. Then we will map the internal states, which are generated by the Seq2Seq encoder, to MNIST fashion images. During the decoding process, we will use FakeSafe to map the MNIST fashion images back to the internal states used by the Seq2Seq model, and further decode back to the sentence.
- In the second approach, considering that Word→F→Fashion→R→Word model achieves good performance on 50-dimension words embeddings, we have attempted to train a Word→F→Fashion→R→Word model first. Then we have split the sentence into a list of words, and each of the words will be encoded into a MNIST fashion image. Eventually, the MNIST fashion images will be decoded back to the list of words, which will be further converted to the sentences.

REFERENCES
[1] A. Siper, R. Farley, and C. Lombardo, “The rise of steganography,” Proceedings of Student/Faculty Research Day, 2005.
[2] Y. Srinivasan, B. Nutter, S. Mitra, B. Phillips, and D. Ferris, “Secure transmission of medical records using high capacity steganography,” Proceedings. 17th IEEE Symposium on Computer-Based Medical Systems, 2004.
FIGURE 5. Multi-step FakeSafe mapping. The reconstruction errors increase with depth.

TABLE 1. Performance of FakeSafe mapping onto the same domain or different domain

| Experiment | Original Message | FakeSafe Message | Reconstruction Error | Precision | Recall | F1 Score | SSIM | PSNR | BPP | Typ | Remark |
|------------|------------------|------------------|----------------------|-----------|--------|---------|-------|------|-----|------|-------|
| Face->Face  | Face image       | Face image       | 1.06                 | 0.99      | 0.48   | 0.68    | 0.81  | 18.84 |     |      | Same domain. |
| Fashion->R | Fashion image    | Fashion image    | 1.2                  | 0.11      | 0.72   | 0.85    | 17.02 | 8     |     |     | Same domain. |
| World2-Fashion->R | English words (50d embeddings) | English words (50d embeddings) | NA | 0.85 | 0.66 | 0.76 | 17.02 | Same domain. |
| World2-MNIST->MNIST | English words (50d embeddings) | English words (50d embeddings) | NA | 0.85 | 0.66 | 0.76 | 17.02 | Same domain. |
| World2-Fashion->R | English words (50d embeddings) | Fashion image | NA | 0.86 | 0.66 | 0.76 | 17.02 | Same domain. |
| World2-Fashion->Fashion | English words (50d embeddings) | Fashion image | NA | 0.86 | 0.66 | 0.76 | 17.02 | Same domain. |

TABLE 2. Performance of multi-step FakeSafe mapping

| Experiment | Reconstruction Error | Precision | Recall | F1 Score | SSIM | PSNR | BPP | Typ | Remark |
|------------|----------------------|-----------|--------|---------|------|------|-----|------|-------|
| mnist->mnist->mnist->mnist->mnist | 37.1 | 0.66 | 0.66 | 0.73 | 0.73 | 0.47 | 13.36 | 8 two-step |
| fashion->mnist->mnist->mnist->fashion | 17.3 | 0.48 | 0.48 | 0.39 | 0.42 | 11.51 | 8 three-step |

TABLE 3. Performance of Supplementary Experiments

| Experiment | Original message | FakeSafe message | BLU score | Type | Data set | Remark |
|------------|------------------|------------------|-----------|------|----------|-------|
| Sentence->R->Sentence | Sentence states | Fashion image | 0.04 | Cross domain | Tatoeba (7061 words) | Use seq2seq model |
| Sentence->R->Sentence | Sentence states | Fashion image | 0.03 | Cross domain | small_vocabulary (202 words) | Use seq2seq model |
| Sentence->R->Sentence | English words (50d embeddings) | Fashion image | 0.48 | Cross domain | small_vocabulary_en (202 words) | Use word embeddings |

[3] S. U. Maheswari and D. J. Hemanth, “Frequency domain QR code based image steganography using Fresnelet transform,” *AEU - International Journal of Electronics and Communications*, 2015.

[4] V. Holub, J. Fridrich, and T. Denemark, “Universal distortion function for steganography in an arbitrary domain,” *EURASIP Journal on Information Security*, 2014.

[5] P. Povy, T. Filler, and P. Bas, “Using high-dimensional image models to perform highly undetectable steganography,” *Information Hiding*, 2010.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[8] D. Volkonskiy, I. Nazarov, and E. Burnaev, “Steganographic generative adversarial networks,” *ICLR 2016*, 2016.

[9] H. Shi, J. Dong, W. Wang, Y. Qian, and X. Zhang, “SSGAN: Secure steganography based on generative adversarial networks,” in *Advances in Multimedia Information Processing*, *PCM 2017*, 2017.

[10] W. Tang, S. Tan, B. Li, and J. Huang, “Automatic steganographic distortion learning using a generative adversarial network,” *IEEE Signal Processing Letters*, 2017.

[11] S. Baluja, “Hiding images in plain sight: Deep steganography,” in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

[12] A. Rehman, R. Rahim, S. Nadeem, and S. Hussain, “End-to-end trained CNN encode-decoder networks for image steganography,” *CoRR*, 2017.

[13] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, “Painful data: The uncle-memahor shadow pain expression archive database,” in *Face and Gesture 2011*, IEEE, 2011, pp. 57–64.

[14] Z. Wang and H. Sheikh, “Image quality assessment: From error visibility to structural similarity,” *IEEE Transactions on Image Processing*, 2004.

[15] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “A comprehensive evaluation of full reference image quality assessment algorithms,” in *Proceedings of the 19th IEEE International Conference on Image Processing*, 2012.
HE ZHU received the BSSc degree (First Class Honours) in digital graphic technology from Hong Kong Baptist University, Hong Kong, China, in 2015, and the MSc degree in multimedia technology from the Hong Kong Polytechnic University, Hong Kong, China, in 2016.

From 2016 to 2020, he worked as a Software Engineer in Hong Kong. Since 2021, he has been a Researcher at Mila - Quebec AI Institute. His current research interests include domain generalization, federated learning, and health informatics.

DIANBO LIU Dianbo Liu was born in the west part of China in 1988. After high school, he was educated at University of Edinburgh, University of Dundee, MIT and Harvard university for computational modeling, machine learning and biomedical informatics.

Dianbo works in Bengio group, Mila, Quebec, Canada as a postdoc fellow. Before that, he worked at Harvard university medical school, Boston children’s hospital, the Broad institute of MIT and Harvard. He was also co-founder and previous CTO of secure AI labs, Cambridge, MA, USA.