Towards Image Data Hiding via Facial Stego Synthesis With Generative Model

Li Dong[1,2], Jie Wang[1,2], Rangding Wang[1,2], Yuanman Li[3] and Weiwei Sun[4]

[1]Faculty of Electrical Engineering and Computer Science, Ningbo University, Zhejiang, China, 315211
[2]Southeast Digital Economic Development Institute, Zhejiang, China, 324000
[3]Shenzhen University, Guangdong, China, 518061
[4]Alibaba Group, Zhejiang, China, 310052

Abstract

Stego synthesis-based data hiding aims to directly produce a plausible natural image to convey secret message. However, most of the existing works neglected the possible communication degradations and forensic actions, which commonly occur in practice. In this paper, we devise a generative adversarial network (GAN)-based framework to synthesize facial stego images. The framework consists of four components: generator, extractor, discriminator and forensic network. Specifically, the generator is deployed to generate a realistic facial stego image from the secret message and key, while the extractor aims at extracting the secret message from the stego image with the provided secret key. To combat forensics, we explicitly integrate a forensic network into the proposed framework, which is responsible for guiding the update of generator. Three degradation layers are further incorporated, enforcing the generator to characterize the communication degradations. Experimental results demonstrate that the proposed framework could accurately extract the secret message and effectively resist the forensic detection and certain degradations, while attaining realistic facial stego images.

Keywords

data hiding, stego synthesis, generative adversarial network

1. Introduction

Data hiding aims to embed the secret message into a cover signal, without incurring awareness of an adversary. It is widely used in many applications, e.g., covert communication [1] and multimedia data protection [2, 3]. The primitive ad-hoc Least-Significant Bit (LSB) replaces the bit in least significant bit-plane of each pixel with the secret bit. While the modern data hiding methods attempt to eliminate the traces of data hiding action and improve the steganographic capacity. For example, content-adaptive steganography [1] designed sophisticated distortion function according to prior knowledge and used Syndrome-Trellis coding to embed the secret message. Recently, neural network-based data hiding is becoming one of the active research directions. Baluja [4] employed convolutional neural networks to hide an entire secret image into the cover image in an end-to-end fashion. The work SSGAN [5] attempted to exploit GAN to synthesize a cover image which is more suitable for the subsequent steganographic data embedding. ASDL-GAN [6] integrated the content-adaptive steganography and GAN, in which the generator was able to produce the modification probability maps. For the methods HayersGAN [7], HiDDeN [8] and SteganoGAN [9], they all designed an encoder-decoder alike framework based on GAN. These methods could automatically learn the suitable areas for embedding the secret bitstream message.

For the last several years, the adversarial examples to neural networks meet data hiding, and continuously drawing extensive attention from the community. Some studies, e.g., [10, 11], found that adding slight perturbations to the input data would paralyze the prediction capability of learning-based classifiers. As the opponent of data hiding, steganalysis aims to expose the data hiding on stego signal and usually involves machine-learning classifiers. Therefore, it is possible for data hiding methods to bypass steganalysis by borrowing some strategies from the adversarial examples-related works. Tang *et al.* [12] presented the Adversarial Embedding (ADV-EMB) method that adjusts the modification cost of image elements, according to the gradients that back-propagated from the target steganalytic neural network. The constructed adversarial stego could effectively fool the steganalytic network, revealing the vulnerability of the deep learning-based steganalyzer.

Note that, all aforementioned data hiding techniques are based on the cover modification. The common characteristic is that these methods can not be independent of the modification on the given cover image. As such, it inevitably leaves artifacts exposing to steganalysis. On the contrary, stego synthesis-based data hiding, e.g., [13, 14], refers to synthesizing the stego image directly from the

International Workshop on Safety & Security of Deep Learning, 21st -26th August, 2021

dongli@nbu.edu.cn (L. Dong); 1811082196@nbu.edu.cn (J. Wang); wangrangding@nbu.edu.cn (R. Wang); yuanmanli@szu.edu.cn (Y. Li); sunweiwei.sww@alibaba-inc.com (W. Sun)

secret message. It could pose more challenges for steganalysis. Under this concept, traditional methods tried to produce stego image based on some hand-crafted designations. Although the capacity was relatively higher, they were limited to synthesizing patterned images, such as textures and fingerprints. As an alternative solution, some methods [15, 16] use GAN to synthesize stego images with rich semantics, e.g., face and food. However, the accuracy of message extraction was unsatisfactory under image degradations. Moreover, the synthesized stego images can be easily identified by a well-trained forensic detector. It is thus urgent to further improve the robustness of message extraction and anti-forensic capability of stego synthesis-based data hiding methods.

In this work, we propose a Facial Stego Image Synthesis method for data hiding with GAN, which is termed as FSIS-GAN. Unlike the cover modification-based data hiding methods, FSIS-GAN is designed without providing a cover image beforehand. Compared with the existing stego synthesis-based methods, FSIS-GAN can not only synthesize realistic facial stego images, but also achieve superior performance in terms of robustness and anti-forensic capability. Experimental results conducted on the public facial dataset validate such merits of our proposed method. The main contributions of this work can be summarized as follows,

- We explicitly consider the image degradation during the covert communication, and integrate multiple degradation layers into the framework. This boost the robustness performance in terms of the message extraction.
- We incorporate a forensic network during training FSIS-GAN. By exploiting the gradients from such a forensic network, the stego image produced by the learned generator could effectively fool the forensic network.
- We explicitly adopt the secret key into the data hiding procedure of FISI-GAN, which could further improve the reliability of the secret message extraction.

The rest of this paper is organized as follows. Section II briefly reviews the related work on stego synthesis-based data hiding. Section III describes the proposed FSIS-GAN, including network architecture and loss function. Section IV presents the experimental results, and the final conclusions are drawn in Section V.

2. Stego Synthesis-based Data Hiding

The majority of data hiding method involves the modification on the given cover images. However, such cover

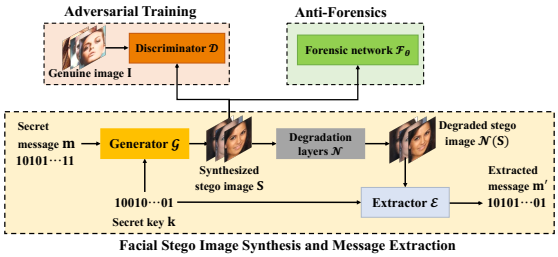


Figure 1: Overview of the proposed FSIS-GAN framework.

modification would leave embedding traces that can be detected. To resist the detection by steganalyzer, stego synthesis-based data hiding method could directly produce the stego images from the given secret message. For early attempts, Wu *et al.* citewu2014steganography proposed a texture image synthesis-based method, which selectively distributes the source patches of the original texture image onto the synthesized stego image. The message hiding and extraction depend on the choice of source patches. Motivated by the fingerprint biometrics, Li *et al.* [14] proposed to use the hologram phase constructed from the secret message to synthesize fingerprint stego image. The hologram phase consists of two phases: The first spiral phase encodes the secret message to the two-dimensional points with different polarities, and the second continuous phase is to synthesize fingerprint images. It is worth noting that conventional stego image synthesis-based methods can only synthesize patterned stego image such as textures, lacking rich semantics, which limits their practical applications.

Instead, Hu *et al.* [15] suggested using the generator of GAN to synthesize a facial stego image from the secret message. Meanwhile, the secret message can be extracted from the stego image by the corresponding extractor network. Similarly, Zhang *et al.* [16] exploited GAN to generate stego image with different semantic labels, which could improve the robustness of data extraction but significantly scarifying the steganographic capacity. The main advantage of the GAN-based works is that they could synthesize stego images with rich semantics. However, we shall note that stego images can be easily identified by some well-trained forensic networks. In addition, there is no trade-off between capacity and extraction accuracy.

3. Facial Image Data Hiding via Generative Stego Synthesis

In this section, we first give an overview of the proposed FSIS-GAN framework and then introduce each component of the framework, accompanied with thorough discussion on the loss function, network structure and train-

ing procedure.

3.1. Overview of FSIS-GAN

The proposed FSIS-GAN framework is illustrated in Figure 1. In general, it is an end-to-end framework consisting of three parts, where each part is designed to achieve a specific goal. First, the part of facial stego image synthesis and message extraction contains a generator $\mathcal{G}$, an extractor $\mathcal{E}$ and the degradation layers $\mathcal{N}$. The generator $\mathcal{G}$ is deployed to convert the secret message along with the secret key into a facial stego image. The degradation layers $\mathcal{N}$ are used to simulate possible common image degradations within the communication channel. The extractor $\mathcal{E}$ is learned to recover the secret message from the degraded stego image. Second, there is a discriminator $\mathcal{D}$ in the part of adversarial training, which aims at distinguishing the genuine data sample from the ones produced by the generator $\mathcal{G}$. Third, a well-trained existing forensic network $\mathcal{F}_\theta$ (parameterized by θ) is introduced in the part of anti-forensics, which could distinguish the genuine from the synthesized facial stego image. Note that this target forensic network is treated as a fixed adversary, and its network parameters are always frozen.

3.2. Stego Image Synthesis and Message Extraction

The part of facial stego image synthesis and message extraction achieve two functionalities. First, by using the generator $\mathcal{G}$, one can convert the given secret message into a facial stego image. Second, the extractor $\mathcal{E}$ is responsible for extracting the secret message from the input stego image. Furthermore, a secret key is introduced to ensure the communication reliability and high diversity of the generated facial stego image.

Generally, generator $\mathcal{G}$ and extractor $\mathcal{E}$ aim to learn two mappings, i.e., mapping the given secret message into a stego image, and vice versa. More formally, let $\mathbf{m} \in \{0,1\}^{l_m}$ and $\mathbf{k} \in \{0,1\}^{l_k}$ be the binary secret message and the secret key, respectively. Generator $\mathcal{G}$ is intended to learn the first mapping, transforming the message $\mathbf{m}$ along with the secret key $\mathbf{k}$ into a stego image:

$$\mathbf{S} = \mathcal{G}(\mathbf{m}, \mathbf{k}), \tag{1}$$

where $\mathbf{S}$ denotes the synthesized facial stego image of shape $C \times H \times W$. To recover the secret message, we next introduce the extractor $\mathcal{E}$. Considering that the facial stego image $\mathbf{S}$ may be degraded during transmission, the second mapping should be from the degraded stego image along with the secret key $\mathbf{k}$ to the secret message, which can be expressed by

$$\mathbf{m'} = \mathcal{E}(\mathcal{N}(\mathbf{S}), \mathbf{k}), \tag{2}$$

where $\mathcal{N}(\cdot)$ models the image degradation process, and $\mathcal{N}(\mathbf{S})$ is the degraded stego image. Here, $\mathbf{m'} \in (0,1)^{l_m}$ denotes the extracted secret message. It shall be noted that the extracted message $\mathbf{m'}$ shall be (approximately) equals the original secret message $\mathbf{m}$, and thus one can employ error correcting mechanism to fully correct the erroneous bits.

To measure the distortion between the original secret message $\mathbf{m}$ and the extracted message $\mathbf{m'}$, we use the cross-entropy loss to calculate the message *extraction loss* $\mathcal{L}_\mathcal{E}$, which is given by

$$\mathcal{L}_\mathcal{E}(\mathbf{m}, \mathbf{m'}) = -\frac{1}{l_m}\sum_{i=1}^{l_m}\big[m_i \log(m_i') + (1-m_i)\log(1-m_i')\big], \tag{3}$$

where m_i and m_i' is i-th element of $\mathbf{m}$ and $\mathbf{m'}$, respectively.

Note that, our proposed FSIS-GAN framework explicitly receiving a secret key as an input, which is designed to satisfy the Kerckhoffs' principle. It means that even the extractor $\mathcal{E}$ network is completely exposed to an attacker, the secret message $\mathbf{m}$ will be recovered only if the receiver obtain both the secret key $\mathbf{k}$ and the facial stego image $\mathbf{S}$. It is worth emphasizing that, for most of the existing GAN-based methods, e.g., [15, 16], there is no involvement of a secret key. Further notice that as the input of the extractor $\mathcal{E}$, the dimensions of secret key $\mathbf{k}$ is greatly smaller than that of the facial stego image $\mathbf{S}$. Thus, the extractor $\mathcal{E}$ tends to discard the secret key because it carries much less information. To mitigate this issue, we propose to use randomly generated incorrect secret key $\widetilde{\mathbf{k}} \in \{0,1\}^{l_k}$, where $\widetilde{\mathbf{k}} \neq \mathbf{k}$, as input during training stage. Instead of directly using the correct secret key and minimize the difference between the extracted and original message, we maximize the differences between the extracted and original message when applying incorrect secret key. Mathematically, the loss term *inverse loss* $\mathcal{L}_{\widetilde{\mathcal{E}}}$, can be expressed by the negative cross-entropy loss:

$$\mathcal{L}_{\widetilde{\mathcal{E}}}(\mathbf{m}, \widetilde{\mathbf{m}'}) = \frac{1}{l_m}\sum_{i=1}^{l_m}\big[m_i \log(\widetilde{m}_i') + (1-m_i)\log(1-\widetilde{m}_i')\big], \tag{4}$$

where $\widetilde{m}_i'$ is the i-th element of the extracted message $\widetilde{\mathbf{m}'}$ with the incorrect key $\widetilde{\mathbf{k}}$, i.e., $\widetilde{\mathbf{m}'} = \mathcal{E}(\mathcal{N}(\mathbf{S}), \widetilde{\mathbf{k}})$.

Enhancing robustness with degradation layers: In a practical communication channel, there often exists degradations on the synthesized stego image $\mathbf{S}$, when transmitting the stego to a receiver. To this end, the data hiding system requires certain robustness to ensure the accuracy of message extraction. Therefore, in this work, we take three representative degradations into account, i.e., image noise pollution, blurring, and compression. For noise pollution, we consider the one of the most widely-used noise models: Gaussian noise. For blurring, the Gaussian blurring is used. For signal compression, JPEG image compression is employed, which is extensively used for reducing the bandwidth of transmission

process. In experiments, we implement these three types of degradation as neural network layers $\mathcal{N}$ to degrade the stego image. Specifically, three network layers are used for simulating each type of degradation. Gaussian noise layer (*GNL*) is to add Gaussian noise to the facial stego image $\mathbf{S}$. Gaussian blurring layer (*GBL*) blurs $\mathbf{S}$. For JPEG compression, considering that the quantitation operation is non-differentiable, we approximate the quantization operation with a differentiable polynomial function. Such differentiating technique can also be referred to the work HiDDeN [8].

3.3. Adversarial Training Part

As aforementioned, the hand-crafted stego synthesis-based data hiding methods [13, 14] only could synthesize patterned images such as texture and fingerprint, limiting their practical applications. Synthesizing a natural image with semantics is a challenging task. However, this problem can be alleviated with the guidance of adversarial training. In this part, the purpose of the discriminator $\mathcal{D}$ is to conduct adversarial training with the generator $\mathcal{G}$ and improve the plausibility of the synthesized facial stego images.

More specifically, let $\mathbf{I}$ be the genuine facial image sample of shape $C \times H \times W$ from a publicly available genuine facial image dataset. The discriminator $\mathcal{D}$ estimates the probability that a given image sample belonging to a synthesized by the generator $\mathcal{G}$. The generator $\mathcal{G}$ attempts to fool the discriminator $\mathcal{D}$. Through such adversarial training, the generator $\mathcal{G}$ is encouraged to synthesize much more realistic facial stego images. As a variant of GAN, the network structure and loss function of BEGAN [17] provides a good reference for improving training stability. Thus, we in this work employ the adversarial training loss used in BEGAN. Mathematically, the *adversarial loss* $\mathcal{L}_{\text{adv}}$ for the generator $\mathcal{G}$ can be calculated as

$$\mathcal{L}_{\text{adv}}(\mathcal{D}(\mathbf{S}), \mathbf{S}) = \frac{1}{CHW}\Big[\big|\mathcal{D}(\mathbf{S}) - \mathbf{S}\big|\Big], \tag{5}$$

where the shape of output $\mathcal{D}(\mathbf{S})$ is same as the facial stego image. The *adversarial loss* $\mathcal{L}_{\mathcal{D}}$ for the discriminator $\mathcal{D}$ is

$$\mathcal{L}_{\mathcal{D}}(\mathbf{I}, \mathbf{S}) = \frac{1}{CHW}\Big[\big|\mathcal{D}(\mathbf{I}) - \mathbf{I}\big| - h_t \cdot \big|\mathcal{D}(\mathbf{S}) - \mathbf{S}\big|\Big], \tag{6}$$

where h_t controls the discrimination ability of $\mathcal{D}$ in the t-th training step to equilibrate the adversarial training. It can be computed as

$$h_{t+1} = h_t + \frac{\lambda}{CHW}\Big[\gamma\big|\mathcal{D}(\mathbf{I}) - \mathbf{I}\big| - \big|\mathcal{D}(\mathbf{S}) - \mathbf{S}\big|\Big]. \tag{7}$$

Here the parameter λ is the learning rate of training, and γ is a hyper-parameter to control the diversity of synthesized facial images. The quality and diversity of the facial stego images can be freely adjusted by tuning the parameter γ.

3.4. Anti-forensics Part

Remind that there is no explicit cover images involved in stego synthesis-based data hiding methods. This merit makes such type of data hiding method could effectively resist to conventional steganalysis detection. However, as pointed in [15], a well-trained forensic network could readily distinguish a synthesized stego image from the genuine one, even the synthesized stego image is of no perceptual differences to an observer.

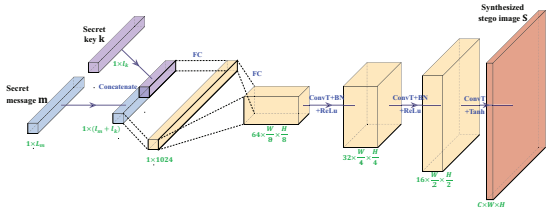
Although $\mathcal{F}_\theta$ is an expert in such a detection task, some studies [10, 11] have shown that deep neural network-based classifiers are vulnerable to adversarial examples. Inspired by this, we propose to apply strategies of obtaining adversarial examples to evade the stego detection network as a way for realizing anti-forensics. In FSIS-GAN framework, we consider a white-box scenario, i.e., assuming one has full knowledge of the target forensic network. The target forensic network $\mathcal{F}$ is trained with the genuine images from a publicly available facial dataset and the synthesized images that produced by BEGAN [17]. Then, we integrate the well-trained $\mathcal{F}_\theta$ into the FSIS-GAN framework, in which $\mathcal{F}_\theta$ receives the synthesized facial stego image $\mathbf{S}$ and output the confidence. The gradients that back-propagated by the $\mathcal{F}_\theta$ are used to update the parameters of the generator $\mathcal{G}$. To measure the loss of resisting forensic detection, we define the *anti-forensic loss* $\mathcal{L}_{\mathcal{F}_\theta}$ to computes the cross-entropy between the output of $\mathcal{F}_\theta$ and our target genuine image label:

$$\mathcal{L}_{\mathcal{F}_\theta}(\mathbf{S}) = -\log\big(1 - \mathcal{F}_\theta(\mathbf{S})\big), \tag{8}$$

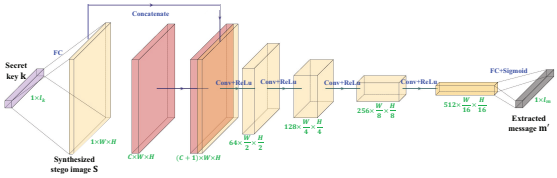
where $\mathcal{F}_\theta(\mathbf{S}) \in (0, 1)$ is the confidence output by $\mathcal{F}_\theta$. Clearly, the decrement of $\mathcal{L}_{\mathcal{F}_\theta}$ indicates the probability increment of $\mathbf{S}$ being identified as a genuine image by $\mathcal{F}_\theta$.

3.5. Network Structure and Training Strategy

The network architecture of the generator $\mathcal{G}$ and the extractor $\mathcal{E}$ are shown in Figure 2. For generator $\mathcal{G}$, the secret key vector $\mathbf{k}$ is first concatenated to the secret message vector $\mathbf{m}$ and then fed to subsequent layers. Then, $\mathcal{G}$ applies two fully-connected (FC) layers and three convtranspose (ConvT) layers to produce the facial stego image $\mathbf{S}$. In particular, after each FC layer or ConvT layer, we apply batch normalization (BN) [18] and ReLU activation function to process intermediate vectors. In experiments, we found that both $\mathbf{m}$ and $\mathbf{k}$ are composed of binary number 0 or 1, and such form is not suitable as input and the adversarial training loss would diverge. To solve this issue, additional BN layers were added, and normalization operation is carried out inside the network. Experiential results show that this trick could greatly alleviate the divergence problem.



(a) Network structure of the generator $\mathcal{G}$



(b) Network structure of the extractor $\mathcal{E}$

Figure 2: Network structure of the generator $\mathcal{G}$ and the extractor $\mathcal{E}$. "Concat", "FC", "ConvT", "BN", "Conv" denote the concatenation, fully-connected layer, convtranspose layer, batch norm, and convolution layer, respectively.

For extractor $\mathcal{E}$, we shall ensure the secret key vector $\mathbf{k}$ and the facial stego image matrix $\mathbf{S}$ in a way such that the extractor $\mathcal{E}$ would not neglect the information provided by the secret key. To this end, the extractor $\mathcal{E}$ first applies FC layer to the secret key to form the intermediate matrix, i.e., $1 \times W \times H$. Then, the facial stego image $\mathbf{S}$ and the intermediate matrix are concatenated, and then feed the fused tensor to the four convolutional (Conv) layers. Finally, the extractor $\mathcal{E}$ applies the FC layer and Sigmoid activation function to produce the message vector $\mathbf{m'}$ (or $\tilde{\mathbf{m}}$') with size of $1 \times l_m$.

For the discriminator $\mathcal{D}$, we adopt the auto-encoder alike structure from BEGAN [17]. For the target forensic network $\mathcal{F}$, we use Ye-Net [19], which is a widely-used steganalytic method.

The training process of the proposed FSIS-GAN framework is iteratively optimize the loss function of each network, except the well-trained forensic network $\mathcal{F}_\theta$. We apply the *extraction loss* $\mathcal{L}_\mathcal{E}$ and the *adversarial loss* $\mathcal{L}_\mathcal{D}$ as the loss function for the extractor $\mathcal{E}$ and the discriminator $\mathcal{D}$, respectively. In particular, The total loss $\mathcal{L}_\mathcal{G}$ for the generator $\mathcal{G}$ is a proper fusion of the four losses aforementioned as follows

$$\mathcal{L}_\mathcal{G} = \mathcal{L}_{\text{adv}} + \alpha(\mathcal{L}_\mathcal{E} + \mathcal{L}_{\widetilde{\mathcal{E}}}) + \beta\mathcal{L}_{\mathcal{F}_\theta}, \tag{9}$$

where $\mathcal{L}_{\text{adv}}$ is the adversarial loss for $\mathcal{G}$, $\mathcal{L}_{\widetilde{\mathcal{E}}}$ is the inverse loss, and $\mathcal{L}_{\mathcal{F}_\theta}$ is the anti-forensic loss. The hyper-parameters of α and β are used to control the relative importance among the four losses.

4. Experiment results

In this section, we first introduce the experimental setup. Then, to verify the robustness of our proposed FSIS-GAN, it is evaluated under image degradation and without degradation, respectively. Finally, the anti-forensic capability of FSIS-GAN is validated.

4.1. Experimental Setup

Our experiments are conducted on the CelebA dataset [20], where the region with face is identified and extracted. All images are reshaped into $3 \times 64 \times 64$. The following three metrics are used for evaluation:

- **Fréchet Inception Distance** (*FID*) [21], which is a widely-used perceptual image quality assessment metric for synthesized images. *FID* is a *de facto* metric for assessing the image quality created by generator of GANs'. Lower *FID* score indicates better consistency with human's perception on natural images.
- **Accuracy of message extraction** (*ACC*) that is computed by $ACC = \frac{L_{\text{Ext}}}{L}$, where L_{Ext} is the length of correctly extracted message and L is the length of secret message $\mathbf{m}$.
- **Probability of missed detection** (*PMD*). This metric can be calculated by $PMD = \frac{FN}{FN+TP}$, where FN (False Negative) is the ratio for case "synthesized facial image is misclassified as a genuine one", and TP (True Positive) is the ratio for case "synthesized facial image is correctly detected". Larger *PMD* indicates higher resisting ability to the forensic network.

The proposed FSIS-GAN framework is implemented with PyTorch and train on four NVIDIA GTX1080Ti GPUs with 11GB memory. The number of training epochs is set to 400 with a mini batch-size of 64. We use Adam [22] as the optimizer with a learning rate of 2×10^{-4}. For the hyper-parameters α and β in (9), with a number of trials and errors, we empirically set them as 0.1 in experiments. The parameter γ in (7) is set to 0.7, which is expected to produce reasonably diverse facial stego images. The competing method is the most related work [15]. We implement this work by ourselves because there is no publicly available code. With certain tweaking and fine-tuning, the tested results were comparable to the originally reported data from [15]. For a fair comparison, the length of the secret message l_m and the secret key l_k are all set to 100, so as to the payload is identical to that of work [15].

Figure 3: Comparison of exemplar synthesized stego images. Top: Hu *et al.* [15]; Bottom: Proposed FSIS-GAN-WD.

4.2. Performance Without Degradations

Notice that the competing method [15] does not consider the image degradations. To verify the effectiveness of the proposed method under same settings and make a fair comparison. We in this subsection to evaluate the performance without degradation layers $\mathcal{N}$. The facial stego image $\mathbf{S}$ will be transmitted to extractor $\mathcal{E}$ without any degradation. To avoid confusion, this variation of our proposed method is termed as FSIS-GAN-WD (WD is abbreviated for *Without Degradations).* We first compare the visual quality of the facial stego images with the competing method [15]. As can be seen from Figure 3, the proposed FSIS-GAN-WD could synthesize more realistic facial stego images in comparison with Hu *et al.* [15]. With more careful inspection, one can notice that the stego images produced by FSIS-GAN-WD are more vivid and with more correct semantic structures. It is difficult for a common human to aware the inauthenticity of the facial stego images synthesized by FSIS-GAN-WD. In contrast, the stego images generated by Hu *et al.* [15] are typically blurry and severely distorted, which apparently draw attentions from a forensic analyzer. For the *FID* evaluation experiment, we use $10,000$ pairs of genuine images and synthesized facial stego images to compute the *FID* score. The *FID* score of FSIS-GAN-WD is 23.20, which is much smaller than that of Hu *et al.* [15]'s 32.07.

Then, we evaluate the extraction accuracy for the case of without degradation. The results are tabulated in Table 1. To demonstrate the impact of the *inverse loss* $\mathcal{L}_{\widetilde{\mathcal{E}}}$ on the extraction accuracy, the ablation experiments are also conducted, by excluding the *inverse loss* during training. This $\mathcal{L}_{\widetilde{\mathcal{E}}}$-ablated version is denoted as FSIS-GAN-WD (ex $\mathcal{L}_{\widetilde{\mathcal{E}}}$). From the Table 1, one can draw the following conclusions. First, the extraction accuracy of FSIS-GAN-WD with the correct secret key $\mathbf{k}$ is 98.76%, which dramatically outperforms 85.23% of the competing method [15]. Second, by comparing FSIS-GAN-WD and FSIS-GAN-WD (ex $\mathcal{L}_{\widetilde{\mathcal{E}}}$), one can see that, the extraction accuracy of FSIS-GAN-WD with a correct secret key $\mathbf{k}$ slightly inferior to that of FSIS-GAN-WD (ex $\mathcal{L}_{\widetilde{\mathcal{E}}}$). This suggests that the introduced *inverse loss* would marginally harm the extraction accuracy. However, when comparing the case of incorrect key $\tilde{\mathbf{k}}$, the participation of the *inverse loss* $\mathcal{L}_{\widetilde{\mathcal{E}}}$ would significantly deduce the extraction accuracy from 97.01% to 71.50%, while FSIS-GAN-WD almost retains the same extraction accuracy. This phenomena means that the involvement of the secret key will not work if we exclude the *inverse loss.* In contrast, FSIS-GAN-WD (ex $\mathcal{L}_{\widetilde{\mathcal{E}}}$) with the incorrect key $\tilde{\mathbf{k}}$ still attains a quite high extraction accuracy of ($> 97\%$). In a short summary, without the *inverse loss* $\mathcal{L}_{\widetilde{\mathcal{E}}}$, the variant FSIS-GAN-WD (ex $\mathcal{L}_{\widetilde{\mathcal{E}}}$) will violate the Kerckhsoffs' principle.

Table 1
Comparison of message extraction accuracy (%) for the case of no communication degradations. Here, $\mathbf{k}$ and $\tilde{\mathbf{k}}$ denote the correct and incorrect secret key, respectively. FSIS-GAN-WD is a variant of the proposed method by excluding the degradation layers, and FSIS-GAN-WD (ex $\mathcal{L}_{\widetilde{\mathcal{E}}}$) represents the FSIS-GAN-WD trained without *inverse loss* $\mathcal{L}_{\widetilde{\mathcal{E}}}$.

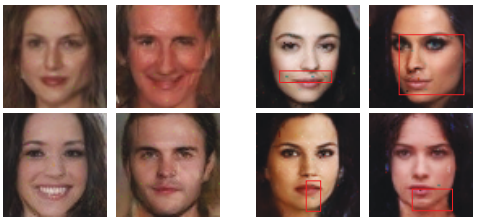
Scheme	Hu *et al.* [15]	FSIS-GAN-WD		FSIS-GAN-WD (ex $\mathcal{L}_{\widetilde{\mathcal{E}}}$)	
		with $\mathbf{k}$	with $\tilde{\mathbf{k}}$	with $\mathbf{k}$	with $\tilde{\mathbf{k}}$
Accuracy	85.23	98.76	71.50	99.41	97.01

4.3. Performance With Degradations

In this subsection, we test the robustness performance of the proposed framework under certain image degradations. The image degradation type and level are given as prior knowledge. This scenario is common in practice because one can obtain some prior knowledge on the degradation through probing the communication channel. Thus, one can fix the degradation layers $\mathcal{N}$ and its associated parameters during training stage. Specifically, in our experiments, the standard deviation σ_1 of the Gaussian noise layer (*GNL*) is set to 0.2. The kernel width d and the standard deviation σ_2 of the Gaussian blurring layer (*GBL*) are set to 3 and 1, respectively. The differentiable JPEG compression layer (*JCL*) is implemented as suggested by the work HiDDEN [8] For referring simplicity, this variation is termed as FSIS-GAN-FD (*FD* is abbreviated for *Fixed Degradation*) in the sequel.

Firstly, the stego images synthesized by FSIS-GAN-FD are provided in Figure 4. One can observe that some speckle noises emerge in the generated stego images, which can be clearly seen from the highlighted regions with red line in Figure 4 (b). Quantitatively, the *FID* score of FSIS-GAN-FD is 41.40, which is inferior to that of FSIS-GAN-WD (23.20) and Hu *et al.* [15] (32.07). Nevertheless, the stego images produced by FSIS-GAN-FD are intuitively more realistic than that of Hu *et al.* [15].

Secondly, in Table 2, we report the extraction accuracy performance under fixed degradations. Not surprisingly, one can notice that the extraction accuracy of Hu *et al.* [15] and FSIS-GAN-WD greatly degrade, which can be attributed to the overlooking on degradation-resistant message extraction issue. In contrast, FSIS-GAN-FD ex-



(a) (b)

Figure 4: The comparison of synthesized facial stego images, where four images of (a) are produced by FSIS-GAN-WD; images of (b) are stego images produced by FSIS-GAN-FD. With the introduction of degradation layers, minor speckle noises emerge (highlighted with red rectangular).

Table 2
Comparison of message extraction accuracy (%) under various degradation conditions. The bold and marked value with an asterisk (*) denote the highest extraction accuracy with correct secret key $\mathbf{k}$ and the lowest extraction accuracy with the incorrect secret key $\tilde{\mathbf{k}}$, respectively.

Scheme	Hu *et al.* [15]	FSIS-GAN-WD		FSIS-GAN-FD	
		with $\mathbf{k}$	with $\tilde{\mathbf{k}}$	with $\mathbf{k}$	with $\tilde{\mathbf{k}}$
W/o degradation	85.23	**98.76**	71.50*	98.22	72.08
Fixed *GNL*	52.72	59.78	56.23*	**95.58**	72.74
Fixed *GBL*	69.68	57.52	54.68*	**98.58**	73.78
Fixed *JCL*	65.33	61.38	58.00*	**98.46**	72.67

hibits quite promising results. Under three types of degradation layers, the extraction accuracy typically exceeds 94% (though lower than that of FSIS-GAN-WD, which is specifically designed for the non-degradation scenario). The results verify that for the case of known degradations, the proposed framework could learn to effectively resistant the fixed degradations, by employing the fixed degradation layers during the training.

Finally, to illustrate how the robustness of message extraction changes under different degradation levels, we test different degradation types with a variety of degradation levels. Due to space limit, we only report the JPEG compression degradation in Figure 5. As can be seen, with the decrement of quality factor (*QF*), the extraction accuracy generally decreases. Although the *JCL* that adopted from HiDDEN [8] could handle non-differentiable JPEG compression, it cannot perfectly reproduce the JPEG compression artifacts. Nevertheless, FSIS-GAN-FD still achieve superior robustness, when comparing with other two schemes.

4.4. Performance of Anti-forensics

Recall that, owing to that no cover images are involvement for data hiding, our method has a relatively good undetectability when exposed to a steganalyzer. However, as pointed in [15], a well-trained forensic network can effectively identify a synthesized image. To solve this issue, we explicitly considered the anti-forensics scenario and introduce the *anti-forensic loss* $\mathcal{L}_{\mathcal{F}_\theta}$.

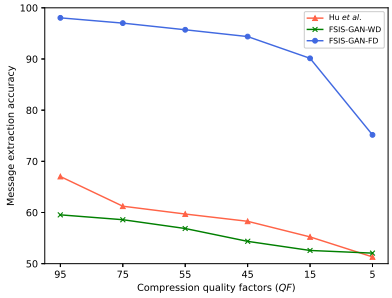


Figure 5: Comparison of the message extraction accuracy (%) under various levels of JPEG compression degradation.

To demonstrate the influence of *anti-forensic loss* $\mathcal{L}_{\mathcal{F}_\theta}$, we conduct the ablation experiment by excluding the loss term $\mathcal{L}_{\mathcal{F}_\theta}$, and thus this variant is termed as FSIS-GAN (ex $\mathcal{L}_{\mathcal{F}_\theta}$). For a concrete example, we employ the well-trained forensic network Ye-Net [19] $\mathcal{F}_\theta$ to detect 3000 facial stego images produced by different methods, and record the probability of missed detection (*PMD*). The *PMD*'s of Hu *et al.* [15], FSIS-GAN (ex $\mathcal{L}_{\mathcal{F}_\theta}$), and FSIS-GAN are 3.23%, 8.84%, and 89.91, respectively. As clearly shown, for FSIS-GAN (ex $\mathcal{L}_{\mathcal{F}_\theta}$), despite the facial stego images look natural for human, they are easily exposed to the forensic network, where the *PMD* value is lower than 10%. In contrast, by introducing the *anti-forensic loss* term, the value of *PMD* of FSIS-GAN could reach 89.91%. This means the proposed method FSIS-GAN could effectively bypass the existing forensic network, retaining an nice anti-forensic capability.

5. Conclusion

In this work, we proposed a stego-synthesis based data hiding method using generative neural network, by explicitly considering the image degradation and anti-forensic need. Specifically, the generator is to synthesize a facial stego image from the given secret message and secret key. The extractor aims to recover the secret message with the secret key. Through the adversarial training with the discriminator, the generator could produce realistic facial stego images. The degradation layers are introduced during the training, which significantly enhance the robustness of message extraction. A forensic network is incorporated during training, in response to the possible adversarial forensic analysis in communication channel. Experimental results verified that, our approach could generate more natural facial stego images, while retaining higher message extraction accuracy and nice anti-forensic ability.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61901237, in part by the Open Project Program of the State Key Laboratory of CADCG, Zhejiang University under Grant A2006, and in part by the Ningbo Natural Science Foundation under Grant 2019A610103. Thanks to Southeast Digital Economic Development Institute for supporting the computing facility.

References

[1] V. Sedighi, R. Cogranne, J. Fridrich, Content-adaptive steganography by minimizing statistical detectability, IEEE Trans. Inf. Forensics Security 11 (2015) 221–234.

[2] J. Zhou, W. Sun, L. Dong, X. Liu, O. C. Au, Y. Y. Tang, Secure reversible image data hiding over encrypted domain via key modulation, IEEE Trans. Circuits Syst. Video Technol. 26 (2015) 441–452.

[3] L. Dong, J. Zhou, W. Sun, D. Yan, R. Wang, First steps toward concealing the traces left by reversible image data hiding, IEEE Trans. Circuits Syst. II, Exp. Briefs 67 (2020) 951–955.

[4] S. Baluja, Hiding images within images, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 1685–1697.

[5] H. Shi, J. Dong, W. Wang, Y. Qian, X. Zhang, SSGAN: secure steganography based on generative adversarial networks, in: Pacific Rim Conference on Multimedia, 2017, pp. 534–544.

[6] W. Tang, S. Tan, B. Li, J. Huang, Automatic steganographic distortion learning using a generative adversarial network, IEEE Signal Process. Lett. 24 (2017) 1547–1551.

[7] J. Hayes, G. Danezis, Generating steganographic images via adversarial training, in: Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 1954–1963.

[8] J. Zhu, R. Kaplan, J. Johnson, F. Li, HiDDen: Hiding data with deep networks, in: Proc. Eur. Conf. Comput. Vis., 2018, pp. 657–672.

[9] K. A. Zhang, A. Cuesta-Infante, L. Xu, K. Veeramachaneni, SteganoGAN: High capacity image steganography with GANs, arXiv preprint arXiv:1901.03892 (2019).

[10] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199 (2013).

[11] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, arXiv preprint arXiv:1412.6572 (2014).

[12] W. Tang, B. Li, S. Tan, M. Barni, J. Huang, CNN-based adversarial embedding for image steganography, IEEE Trans. Inf. Forensics Security 14 (2019) 2074–2087.

[13] K. Wu, C. Wang, Steganography using reversible texture synthesis, IEEE Trans. Image Process. 24 (2014) 130–139.

[14] S. Li, X. Zhang, Toward construction-based data hiding: From secrets to fingerprint images, IEEE Trans. Image Process. 28 (2018) 1482–1497.

[15] D. Hu, L. Wang, W. Jiang, S. Zheng, B. Li, A novel image steganography method via deep convolutional generative adversarial networks, IEEE Access 6 (2018) 38303–38314.

[16] Z. Zhang, G. Fu, R. Ni, J. Liu, X. Yang, A generative method for steganography by cover synthesis with auxiliary semantics, Tsinghua Science and Technology 25 (2020) 516–527.

[17] D. Berthelot, T. Schumm, L. Metz, BEGAN: Boundary equilibrium generative adversarial networks, arXiv preprint arXiv:1703.10717 (2017).

[18] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167 (2015).

[19] J. Ye, J. Ni, Y. Yi, Deep learning hierarchical representations for image steganalysis, IEEE Trans. Inf. Forensics Security 12 (2017) 2545–2557.

[20] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proc. IEEE Int. Conf. Comput. Vis., 2015.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, in: Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 6629–6640.

[22] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).