Robust Steganography without Embedding Based on Secure Container Synthesis and Iterative Message Recovery

Ziping Ma[1], **Yuesheng Zhu**[1,*], **Guibo Luo**[1], **Xiyao Liu**[2,*], **Gerald Schaefer**[3] and **Hui Fang**[3]
[1]Communication and Information Security Lab, Shenzhen Graduate School, Peking University
[2]School of Computer Science and Engineering, Central South University
[3]Department of Computer Science, Loughborough University
maziping.im@gmail.com, zhuys@pku.edu.cn, luoguibo@pku.edu.cn,
lxyzoewx@csu.edu.cn, gerald.schaefer@ieee.org, h.fang@lboro.ac.uk

Abstract

Synthesis-based steganography without embedding (SWE) methods transform secret messages to container images synthesised by generative networks, which eliminates distortions of container images and thus can fundamentally resist typical steganalysis tools. However, existing methods suffer from weak message recovery robustness, synthesis fidelity, and the risk of message leakage. To address these problems, we propose a novel robust steganography without embedding method in this paper. In particular, we design a secure weight modulation-based generator by introducing secure factors to hide secret messages in synthesised container images. In this manner, the synthesised results are modulated by secure factors and thus the secret messages are inaccessible when using fake factors, thus reducing the risk of message leakage. Furthermore, we design a difference predictor via the reconstruction of tampered container images together with an adversarial training strategy to iteratively update the estimation of hidden messages. This ensures robustness of recovering hidden messages, while degradation of synthesis fidelity is reduced since the generator is not included in the adversarial training. Extensive experimental results convincingly demonstrate that our proposed method is effective in avoiding message leakage and superior to other existing methods in terms of recovery robustness and synthesis fidelity.

1 Introduction

Steganography refers to the hiding of a secret message into a media carrier [Bandyopadhyay *et al.*, 2008], such as an image [Cheddad *et al.*, 2010], so as to conduct concealed communication by transmitting the carrier in a public channel [Marvel *et al.*, 1999], such as the Internet or social software. Traditional steganography methods directly embed the secret message into the carrier image by modulating some unnoticeable components, such as the least significant bits of pixel values [Van Schyndel *et al.*, 1994] or texture-rich areas [Holub and Fridrich, 2012]. To further improve hidden capacity and message recovery robustness, various deep learning-based steganography methods have been proposed [Baluja, 2017; Zhu *et al.*, 2018]. However, the embedding processing inevitably leads to abnormalities of statistical characteristics of the carrier image and, in consequence, to an inherent risk of detection by machine learning-based steganalysis tools.

To overcome this issue, steganography without embedding (SWE) methods can be designed to implement a mapping relationship between secret messages and hash sequences to select a natural image from a prepared database [Zhou *et al.*, 2015]. However, the required image quantity of such mapping-based SWE methods is exponential to their hidden capacity, which limits their efficacy. To address this problem, synthesised-based SWE methods [Hu *et al.*, 2018] implicitly represent a secret message by synthesising a container image using deep generative networks, such as GANs [Goodfellow *et al.*, 2020]. These methods can hide more than 100 message bits in one image, which is several times the hidden capacity of mapping-based methods. Due to this relatively high hidden capacity and the modification-free hiding process, which is inherently immune to typical steganalysis tools, synthesis-based SWE has gained wide interest.

However, there are still a number of drawbacks that hinder real-world applications based on existing synthesis-based SWE methods: (1) most of them are weak in terms of robustness because they fail to ensure accurate message recovery under various image attacks; (2) existing SWE methods designed to enhance robustness train their generator using adversarial training, which leads to significant degradation of image synthesis fidelity; and (3) they suffer from the risk of message leakage, since, due to the mapping between secret messages and synthesised containers, if someone else who owns the extractor they can also recover hidden messages.

To tackle the above drawbacks, in this paper, we propose **S**ecure container synthesis and **I**terative message recovery-based **S**teganography **W**ithout **E**mbedding (**SI-SWE**) as a novel robust SWE method. SI-SWE consists of a secure weight modulation-based generator, which constructs a transformation from secret messages to synthesised images, and a robust difference predictor, which predicts the difference between two unknown messages by contrasting their trans-

*Corresponding authors

formed images. Our proposed secure container synthesis mechanism introduces a secure factor to modulate the synthesis of the secure generator, while the iterative message recovery updates an estimate of the secret message using the predicted difference for a pair of container images: the original and one that is reconstructed from the estimate by the generator. Furthermore, we design a novel adversarial training strategy that uses various image attacks to only train the difference predictor. Separating the training of the generator and the difference predictor has two benefits. On one hand, using an adversarial training strategy to train the difference predictor improves the recovery robustness, since the difference predictor can accurately predict the differences from tampered images and the iterative mechanism also reduces the impact of attacks through the reconstruction. On the other hand, the generator is trained without considering attacks to reduce degradation caused by adversarial training, which guarantees high synthesis fidelity. In addition, secure factors are introduced to modulate the transformation of secret messages by the generator, which means that different secure factors will lead to different synthesised containers for hiding the same message. In this manner, using fake factors will mislead the recovery of secret messages since the reconstructed images cannot accurately represent their message differences against the container images, which in turn effectively prevents message leakage.

Our main contributions in this paper can be summarised as follows:

- **Robust message recovery:** we design a difference predictor via iterative reconstructions of container images together with an adversarial training strategy for message recovery, leading to strong robustness to accurately recover hidden secret messages even from seriously tampered container images;
- **Enhanced undetectability:** we design different strategies for the training of the weight modulation-based generator and the difference predictor, with the former trained without considering attacks and the latter using adversarial training. As a result, SI-SWE reduces degradation caused by adversarial training and guarantees high synthesis fidelity to minimise the risk of container images being suspected, while ensuring high recovery robustness;
- **Avoiding message leakage:** we introduce secure factors to modulate the transformation from secret messages to synthesised container images so that the differences between different messages can only be accurately predicted from their synthesising images modulated by the same secure factor. This ensures that hidden messages are inaccessible without knowing the correct factor, significantly reducing the risk of message leakage.

Our code is made available at https://github.com/Lemok00/SI-SWE.

2 Related Work

2.1 Steganography Based on Embedding

Traditional steganography embeds a secret message into a carrier image through (minimal) modifications. [Van Schyndel *et al.*, 1994] first proposes to replace the least significant bits (LSBs) of pixel values in images with secret message bits, while [Sharp, 2001] extends this method by adjusting the whole pixel values to avoid "pairs of values" [Westfeld and Pfitzmann, 1999]. For resistance against steganalysis, [Pevnỳ *et al.*, 2010] introduces high-dimensional models to minimise the weighted difference of feature vectors, while [Holub and Fridrich, 2012] embeds secret bits into texture-rich and noisy image regions, and [Holub *et al.*, 2014] extends this method to arbitrary domains.

Deep learning-based steganography approaches yield improvements in terms of undetectability, capacity and robustness. For improved undetectability, [Tang *et al.*, 2017; Tang *et al.*, 2020] propose the use of GANs or a reinforcement learning framework to predict probability maps for modification, which minimises pixel-level embedding costs. To enlarge the hidden capacity, [Baluja, 2017; Baluja, 2019] train deep networks to hide full-size images into another image. [Shaoping *et al.*, 2021] employs invertible neural networks to improve the recovery quality with a large capacity. For improved robustness, [Zhu *et al.*, 2018; Wengrowski and Dana, 2019] train DL-based models against perturbation pipelines or specialised datasets to ensure accurate message recovery from tampered or camera-captured images. Furthermore, [Xu *et al.*, 2022] proposes a normalising flow which models the distribution of redundant components to recover high-quality images from tampered images.

However, embedding-based steganography methods carry the inherent risk that any modification on natural images leaves traces in statistical characteristics, resulting in their lack of resistance against steganalysis tools.

2.2 Steganography Without Embedding

To overcome this issue, [Zhou *et al.*, 2015] proposes to hide messages by choosing unmodified natural images whose hash sequences are the same as message segments. To improve both hidden capacity and recovery robustness, image hashes can be generated based on the scale invariant feature transform (SIFT) [Zheng *et al.*, 2017], discrete wavelet transform coefficients [Liu *et al.*, 2020], or objects recognised by a Faster R-CNN [Luo *et al.*, 2020]. Moreover, [Zou *et al.*, 2022] uses an unsupervised clustering algorithm for the image database to improve efficiency and robustness. However, such mapping-based SWE methods require a number of database images exponential to the hidden capacity, limiting the hidden capacity.

Significantly higher hidden capacities can be achieved through image synthesis-based methods. [Hu *et al.*, 2018] first proposes to train GANs [Goodfellow *et al.*, 2020] with an extractor to achieve concealed message transmission through synthesised container images. To improve recovery reliability and synthesise diversity, [Liu *et al.*, 2022] disentangles images into structure and texture representations, and synthesise a container image by a randomly sampled texture and a structure representing a message. For application in online social networks, [You *et al.*, 2022] trains both the message hiding and recovery modules against a differential approximation of JPEG compression to synthesise container images which can defend against all possible attacks in transmission

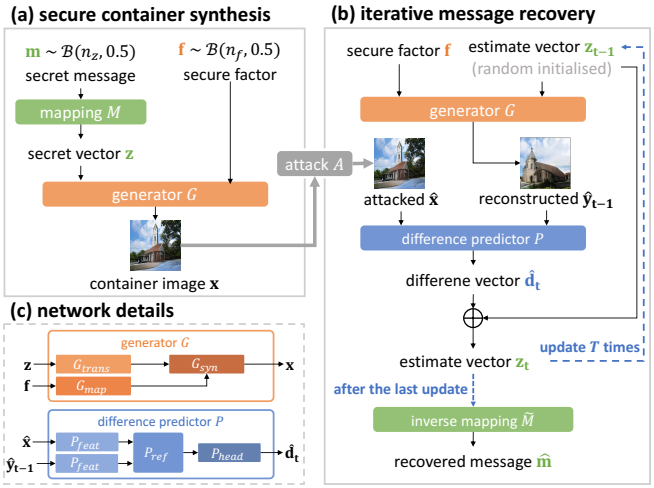

Figure 1: Overview of our proposed SI-SWE.

channels.

3 Proposed Method

3.1 Overview

An overview of our proposed SI-SWE method is given in Figure 1. Our approach consists of a secure container synthesis mechanism for message hiding, and an iterative message recovery mechanism for message recovery. For message hiding, we first map a secret message m to a secret vector z and then input z together with a secure factor f to a generator G to synthesise a container image x which can then be transmitted through noisy channels with various image attacks. For message recovery, the hidden secret message is recovered using a difference predictor P together with the same factor f and the generator G to iteratively update an estimate vector $\hat{z}_t$ and reconstruct the attacked container image $\hat{x}$. In the following, we explain the key modules of SI-SWE in detail.

3.2 Mapping Processes

Instead of directly inputting a message m to the generator, we take a vector z to achieve flexible hidden capacity and construct a mapping relationship between z and m. For this, we employ the mapping process M and its inverse process $\tilde{M}$ proposed in [Liu *et al.*, 2022], where M maps m to z through mapping its every σ-bit segment s_i to a float value z_i in range $[-1, 1]$ as

$$z_i = (h_i + 0.5)/2^{\sigma-1} - 1, \tag{1}$$

where h_i is the decimal value of s_i.

$\tilde{M}$ performs the inverse mapping

$$h_i = \lfloor (z_i + 1) \times 2^{\sigma-1} \rfloor, \tag{2}$$

and recovers the segment s_i from h_i.

3.3 Generator

In SI-SWE, the key module for message hiding is the generator G, which takes a secret vector z and a secure factor f as input to synthesise a container image x as

$$x = G(z, f) = G_{syn}(G_{trans}(z), G_{map}(f)), \tag{3}$$

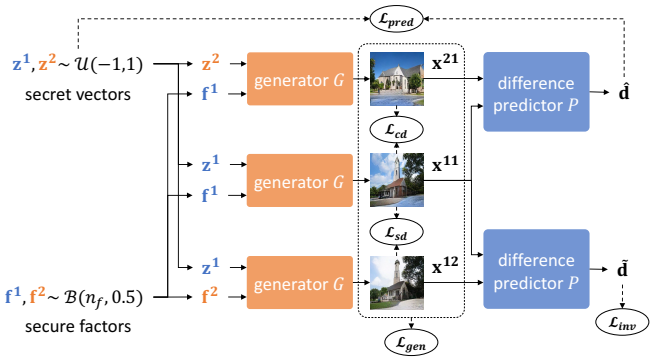

Figure 2: Generator training procedure.

where $G_{trans}(\cdot)$ transforms z to a high-dimension feature map, $G_{map}(\cdot)$ maps f to a latent vector, and $G_{syn}(\cdot)$ adopts StyleGAN2 [Karras *et al.*, 2020] as its backbone architecture to synthesise high fidelity images.

The training procedure of G is illustrated in Figure 2, where three images are firstly synthesised using different combinations of secret vectors and secure factors as

$$\begin{aligned} x^{11} &= G(z^1, f^1), \\ x^{21} &= G(z^2, f^1), \\ x^{12} &= G(z^1, f^2), \end{aligned} \tag{4}$$

where, based on to the mapping interval of M, z^1 and z^2 each consist of n_z elements sampled from a uniform distribution $\mathcal{U}(-1, 1)$, while f^1 and f^2 are sampled from a Bernoulli distribution $\mathcal{B}(n_f, 0.5)$ with n_f the length of f and 0.5 the probability.

To train the generator to synthesise containers modulated by secure factors for message hiding, we use the difference predictor P to calculate a prediction loss $\mathcal{L}_{pred}$ and an inverse loss $\mathcal{L}_{inv}$. Since x^{21} and x^{11} are synthesised using the same factor but with different vectors, $\mathcal{L}_{pred}$ is used to ensure their difference vector $d = (z^1 - z^2)$ can be accurately predicted, and we thus define the loss as

$$\mathcal{L}_{pred} = \| P(x^{11}, x^{21}), d \|^1, \tag{5}$$

where $\| \cdot \|^1$ denotes the L_1 loss. On the other hand, x^{12} is synthesised using the same vector as x^{11} but with a different factor. The inverse loss $\mathcal{L}_{inv}$ is thus designed to prevent message recovery using fake factors by maximising the predicted difference and is obtained as

$$\mathcal{L}_{inv} = \max(0, \tau_{inv} - \| P(x^{11}, x^{12}), \mathbf{0} \|^1), \tag{6}$$

where τ_{inv} refers to a margin greater than 0 and indicates that the difference vector whose norm exceeds the margin will not contribute to the loss.

To ensure realistic image synthesis, a generative loss term $\mathcal{L}_{gen}$ is introduced to make all synthesised images indiscriminative from real images and is defined as

$$\mathcal{L}_{gen} = \sum\nolimits_{x \in X} \text{softplus}(-D(x)), \tag{7}$$

where $X = \{x^{11}, x^{12}, x^{21}\}$ and D denotes a discriminator trained against the generator.

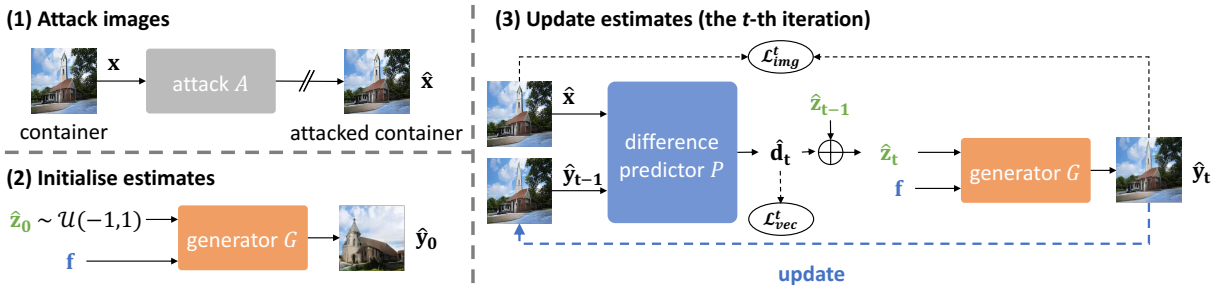

Figure 3: Difference predictor training procedure.

In addition, since using only the generative loss results in the generator synthesising images of a limited number of contents and styles [Kazemi *et al.*, 2019], we introduce content and style diversity losses, $\mathcal{L}_{cd}$ and $\mathcal{L}_{sd}$, to enhance synthesis diversity. For this, we train the generator to synthesise images with content diversity while using different secret vectors, and with style diversity while using different secure factors and define the loss terms as

$$\mathcal{L}_{cd} = \max(0, \tau_c - \mathcal{L}_{cc}(x^{11}, x^{21})), \tag{8}$$

$$\mathcal{L}_{sd} = \max(0, \tau_s - \mathcal{L}_{sc}(x^{11}, x^{12})), \tag{9}$$

where $\mathcal{L}_{cc}$ and $\mathcal{L}_{sc}$ refer to the content and style consistent losses proposed in [Kazemi *et al.*, 2019], and τ_c and τ_s are margins. We thus associate the secret vectors with the synthesised contents since the locality of contents makes the recovery of vectors more reliable, and the secure factors with the synthesised styles since the globality of styles will lead to a significant recovery error when there is a minor difference between the input factors.

The over loss of the generator is then obtained as

$$\begin{aligned}\mathcal{L}_G =& \lambda_{pred}\mathcal{L}_{pred} + \lambda_{inv}\mathcal{L}_{inv} + \lambda_{gen}\mathcal{L}_{gen} \\ &+ \lambda_{cd}\mathcal{L}_{cd} + \lambda_{sd}\mathcal{L}_{sd},\end{aligned} \tag{10}$$

where the hyper-parameters λ_{pred}, λ_{inv}, λ_{gen}, λ_{cd}, λ_{sd} control the impact of each loss term. It should be noted that, to reduce the degradation of synthesis performance, our generator is trained without considering any image attacks.

3.4 Difference Predictor

As shown in Figure 1(c), we design a difference predictor P to predict the difference between a pair of images $\hat{x}$ and $\hat{y}_{t-1}$ as

$$\hat{d}_t = P(\hat{x}, \hat{y}_{t-1}) = P_{head}(P_{ref}(P_{feat}(\hat{x}), P_{feat}(\hat{y}_{t-1}))), \tag{11}$$

where P_{feat} refers to a feature encoder, P_{ref} to a reference attention module [Yu *et al.*, 2021b], and P_{head} to a convolutional prediction head.

P_{ref} fuses encoded features ψ_x and ψ_y as

$$P_{ref}(\psi_x, \psi_y) = \psi_y + \gamma \cdot v(\psi_y)\alpha^T, \tag{12}$$

with

$$\alpha = \text{softmax}(k(\psi_x)^T q(\psi_x)), \tag{13}$$

where γ is a learnable parameter, and $k(\cdot)$, $q(\cdot)$, $v(\cdot)$ are three functions implemented by 1×1 convolutions.

attack	details/parameters
quantisation (QT)	quantise images to 8-bit RGB format
Gaussian noise (GN)	mean = 0, variance = {0.06, 0.08, 0.10}
salt & pepper noise (SPN)	noise density = {0.06, 0.08, 0.10}
speckle noise (SN)	variance = {0.06, 0.08, 0.10}
JPEG compression (JC)	quality factor = {90, 70, 50}
WebP compression (WC)	quality factor = {90, 70, 50}
Gaussian filter (GF)	variance = 1, window = {5×5, 9×9, 13×13}
averaging filter (AF)	window = {5×5, 9×9, 13×13}
median filter (MF)	window = {5×5, 9×9, 13×13}

Table 1: Attacks used in adversarial training.

The training procedure of the difference predictor is illustrate in Figure 3 and consists of three parts:

(1) Attack images. To train the difference predictor for robust message recovery, we first tamper a container image x to obtain an attacked container $\hat{x}$ as

$$\hat{x} = A(x), \tag{14}$$

where A denotes an attack randomly selected from Table 1, and the gradient back-propagation is stopped at $\hat{x}$ to avoid the effect of non-differential operations in attacks.

(2) Initialise estimates. We initialise an estimate vector $\hat{z}_0$ by randomly sampling it from the uniform distribution $\mathcal{U}(-1, 1)$ and synthesise a reconstructed container $\hat{y}_0$ through the generator G as

$$\hat{y}_0 = G(\hat{z}_0, f), \tag{15}$$

where f is the secure factor used to hide the secret vector into the container image x.

(3) Update estimates. We update the estimate vector $\hat{z}$ and the reconstructed container $\hat{y}$ iteratively T times, where T is uniformly sampled from $\{1, ..., T_{max}\}$. In particular, in the t-th iteration ($1 \leq t \leq T$), the difference predictor firstly contrasts $\hat{x}$ and $\hat{y}_{t-1}$ to predict a difference vector $\hat{d}_t$, and the estimate vector $\hat{z}_{t-1}$ is then updated to $\hat{z}_t$ as

$$\hat{z}_t = \hat{z}_{t-1} \oplus \hat{d}_t, \tag{16}$$

where $\oplus$ denotes an element-wise addition operation. Finally, the reconstructed container is updated, using $\hat{z}_t$ and f, as

$$\hat{y}_t = G(\hat{z}_t, f). \tag{17}$$

We calculate a vector recovery loss

$$\mathcal{L}^t_{vec} = \| \hat{d}_t, (z - \hat{z}_{t-1}) \|^1, \tag{18}$$

	QT	GN	SPN	SN	JC	WC	GB	AB	MB	average	capacity
DCGAN-Steg	94.83%	69.12%	60.49%	77.99%	62.48%	62.93%	64.34%	56.41%	57.48%	64.54%	100 bits
SAGAN-Steg	96.60%	63.96%	57.02%	78.28%	60.55%	63.25%	66.34%	47.70%	51.01%	61.74%	200 bits
SSteGAN	97.87%	91.33%	80.13%	95.84%	88.38%	90.45%	96.94%	83.90%	85.24%	89.21%	100 bits
WGAN-Steg	91.65%	74.03%	63.01%	82.16%	84.94%	85.73%	83.68%	72.71%	74.37%	77.87%	100 bits
GDA-Steg	72.11%	69.82%	71.74%	72.81%	73.26%	73.61%	73.90%	65.69%	66.83%	70.98%	256 bits
IDEAS	**100%**	50.51%	50.19%	57.89%	49.84%	50.01%	49.22%	50.06%	51.94%	52.20%	256 bits
CIS-Net-32	99.89%	98.50%	89.83%	99.63%	99.83%	99.24%	98.67%	82.35%	83.97%	94.12%	32 bits
CIS-Net-64	76.67%	75.12%	67.75%	76.44%	76.67%	76.59%	73.79%	70.13%	71.94%	73.62%	64 bits
SI-SWE-64	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	64 bits
SI-SWE-256	99.74%	95.47%	99.74%	98.83%	99.02%	98.93%	99.74%	99.74%	99.74%	98.92%	256 bits

Table 2: Average recovery accuracy results for all methods under different attacks on all three datasets. The best results are **bolded**, the second-best underlined.

and an image reconstruction loss

$$\mathcal{L}_{img}^{t} = \| \hat{y}_t, \hat{x} \|^1 \tag{19}$$

of the t-th iteration, where z is the hidden secret vector. The total loss of the difference predictor is then obtained as

$$\mathcal{L}_P = \sum_{t=1}^{T} (\lambda_{vec}\mathcal{L}_{vec}^{t} + \lambda_{img}\mathcal{L}_{img}^{t}), \tag{20}$$

where λ_{vec} and λ_{img} balance vector recovery and image reconstruction.

4 Experimental Results

4.1 Experimental Setup

To demonstrate the superiority of SI-SWE, we compare it with seven state-of-the-art synthesis-based SWE methods, namely DCGAN-Steg [Hu *et al.*, 2018], SAGAN-Steg [Yu *et al.*, 2021a], SSteGAN [Wang *et al.*, 2018], WGAN-Steg [Li *et al.*, 2020], GDA-Steg [Peng *et al.*, 2022], IDEAS [Liu *et al.*, 2022], and CIS-Net [You *et al.*, 2022]. We use three publicly available datasets: LSUN [Yu *et al.*, 2015] Bedrooms, LSUN Churches, and FFHQ [Karras *et al.*, 2019] to train the models. All images in the datasets are resized to a resolution of 256×256 pixels. We train SI-SWE with two different input dimensions, namely SI-SWE-64 with $n_z = n_f = 64$, and SI-SWE-256 with $n_z = n_f = 256$. The training hyperparameters are set to $\lambda_{pred} = \lambda_{vec} = 10$, $\lambda_{inv} = 5$, and $\tau_{inv} = 0.75$ to ensure both robust message recovery and preventing message leakage, while we set λ_{cd}=4 to enhance the diversity of the synthesised images' contents. We further set, empirically, $T_{max} = 10$, $\lambda_{gen} = 2$ and, $\lambda_{sd} = \lambda_{img} = \tau_c = \tau_s = 1$, while σ in the mapping process is set to 1.

4.2 Message Recovery Robustness

To compare the robustness of SI-SWE with other SWE methods, we evaluate their recovery accuracies under the attacks listed in Table 1, where each attack (except image quantisation) adopts 6 different parameters with only 3 of the 6 having been used during the adversarial training. The average accuracy results for all methods and all attacks are reported in Table 2.

From there, it is evident that our SI-SWE method achieves superior message recovery robustness, with SI-SWE-64 yielding 100% under all attacks, and SI-SWE-256 giving the second best average accuracy of 98.92%. Although CIS-Net-32 obtains slightly higher accuracies for some attacks compared to SI-SWE-256, it supports only a significantly lower hidden capacity (32 vs. 256 bits), which is also difficult to expand as demonstrated by the severely reduced accuracy results for CIS-Net-64 (73.62% vs. 94.12%).

4.3 Container Undetectability

To evaluate the undetectability of the SWE methods, we test their synthesised container images in terms of resistance against steganalysis tools and synthesis fidelity.

For resistance evaluation, we select three well-known steganalysis tools, namely StegExpose [Boehm, 2014],

	StegExpose	XuNet	YeNet
DCGAN-Steg	0.594	0.564	0.556
SAGAN-Steg	0.587	0.508	0.550
SSteGAN	0.422	0.499	0.506
WGAN-Steg	0.609	0.539	0.552
GDA-Steg	0.615	0.553	0.507
IDEAS	0.477	0.413	0.518
CIS-Net-32	0.590	0.458	0.497
CIS-Net-64	0.623	0.435	0.502
SI-SWE-64	0.523	0.425	0.503
SI-SWE-256	0.507	0.452	0.497

Table 3: AUC detection results of different steganalysis tools for all methods.

	Bedrooms	Churches	FFHQ	average
DCGAN-Steg	293.16	108.24	115.31	172.24
SAGAN-Steg	162.89	100.60	82.24	115.24
SSteGAN	177.51	220.19	173.61	190.44
WGAN-Steg	147.71	181.56	67.89	132.39
GDA-Steg	74.80	147.77	82.36	101.64
IDEAS	16.88	15.90	32.88	21.89
CIS-Net-32	54.26	31.57	42.22	42.68
CIS-Net-64	152.01	26.53	44.34	74.29
SI-SWE-64	14.61	14.65	30.53	19.93
SI-SWE-256	**13.86**	**11.68**	**28.23**	**17.92**

Table 4: FID results for all methods. The best results are **bolded**, the second-best underlined.

Figure 4: Examples of container images synthesised by all methods.

XuNet [Xu *et al.*, 2016], and YeNet [Ye *et al.*, 2017]. We obtain the receiver operating characteristic (ROC) curves of detection results and report the area under the curve (AUC) values in Table 3. As we can see from thee, all SWE methods achieve AUC values close to 0.5, which means that detection by the steganalysis tools is close to random guessing. This confirms that SWE methods are fundamentally immune to the detection by current steganalysis tools, as they synthesise realistic images to represent the secret message without modifications.

The synthesis fidelity of synthesised images determines their imperceptibility to visual inspection, and is determined by the authenticity and diversity of the generated images. We use the Fréchet inception distance (FID), a widely employed synthesis fidelity metric, to evaluate the synthesis fidelity of all SWE methods, and report the obtained results in Table 4. As is apparent from there, SI-SWE outperforms all othe SWE methods on all three datasets, with SI-SWE-64 yielding an average FID of 19.93, and SI-SWE-256 an even better average FID of 17.92 as it takes higher-dimensional secret vectors as input which leads to more diverse characteristics. Container images generated by SI-SWE are thus less likely to be suspected compared to images synthesised by the other approaches.

Examples of images synthesised by all SWE methods are shown in Figure 4 to also allow a more subjective evaluation. Looking at the container images generated by SI-SWE, it is evident that they are of high authenticity and diversity, and that it would be very hard to identify them as synthesised images.

4.4 Avoiding Message Leakage

To evaluate SI-SWE's effectiveness in avoiding message leakage, we evaluate the recovery accuracies to recover messages from quantised (QT) images using different secure factors: the real factor and other randomly sampled fake factors. Table 5 lists the obtained results and shows significantly decreased recovery accuracies when using fake factors compared to using the real secure factor for SI-SWE-64 and SI-SWE-256. This confirms that our proposed method ensures that hidden messages are inaccessible without knowing the real factor, thus significantly reducing the risk of message leakage.

	Bedrooms	Churches	FFHQ
SI-SWE-64	100% / 70.19%	100% / 69.12%	100% / 69.10%
SI-SWE-256	99.43% / 48.72%	100% / 69.04%	99.81% / 68.24%

Table 5: Recovery accuracies for quantised images using real secure factors (left number in each entry) and fake factors (right number).

		accuracy	FID
SI-SWE-64	without attacks	55.95%	15.67
	CIS-Net	98.79%	23.87
	SI-SWE	**100%**	**19.93**
SI-SWE-256	without attacks	57.81%	14.72
	CIS-Net	83.85%	21.74
	SI-SWE	**98.92%**	**17.92**

Table 6: Results for different adversarial training strategies. The results for our proposed strategy are **bolded**.

4.5 Ablation Studies

Adversarial Training Strategies

We train SI-SWE using the following adversarial training strategies and compare the resulting performance in terms of recovery and synthesis: without attacks as a baseline, using the strategy of CIS-Net, and using our proposed strategy. Table 6 shows the obtained average recovery accuracies and FID scores, and demonstrates that our proposed strategy significantly enhances the robustness of message recovery, since more attacks are used, and effectively reduces degradation of image synthesis fidelity due to the the separated training of generator and difference predictor.

τ_{inv} Settings

We perform experiments to investigate how different settings of τ_{inv} impact the performance of SI-SWE. For this, we evaluate the FID scores together with the recovery accuracies from quantised (QT) images for different factor bit error rates (f-BERs) between the real and fake factors. The factor-BER is defined as

$$\text{f-BER} = \frac{\text{number of different bits}}{\text{number of total bits}} \times 100\%, \qquad (21)$$

and we use three f-BERs in our experiments, f-BER$_1$ = 12.5%, f-BER$_2$ = 25%, and f-BER$_3$ = 50%. We evaluate for τ_{inv} = {0.25, 0.5, 0.75, 1}, and report the results in Table 7. From there, we can observe that there is a trade-off between synthesis fidelity performance of and the ability to avoid message leakage. When τ_{inv} increases, recovery accuracies decrease to prevent access to hidden messages, while FID scores increase indicating worse synthesis fidelity. Compared to $\tau_{inv} = \{0.25, 0.5\}$, setting $\tau_{inv} = 0.75$ much better avoids message leakage while FID scores increase only slightly, whereas $\tau_{inv} = 1$ significantly compromises the synthesis fidelity so that container images are more likely to arouse suspicion. We consequently recommend $\tau_{inv} = 0.75$ as the best balance to avoid message leakage while guaranteeing high synthesis fidelity to enhance undetectability.

		accuracy			
	τ_{inv}	f-BER$_1$	f-BER$_2$	f-BER$_3$	FID
SI-SWE-64	0.25	94.65%	90.20%	87.37%	17.44
	0.5	80.48%	76.45%	73.06%	18.18
	0.75	**74.62%**	**72.74%**	**71.71%**	**19.93**
	1	70.60%	67.14%	66.24%	24.67
SI-SWE-256	0.25	92.45%	90.93%	90.04%	16.81
	0.5	78.05%	76.72%	75.74%	17.11
	0.75	**71.82%**	**66.56%**	**62.35%**	**17.92**
	1	68.74%	65.24%	58.03%	20.87

Table 7: FID and recovery accuracies for quantised images for different f-BERs and different settings of τ_{inv}. The results for $\tau_{inv} = 0.75$, our recommended setting, are **bolded**.

	$\mathcal{L}_{cd}$	$\mathcal{L}_{sd}$	λ_{cd}	Bedrooms	Churches	FFHQ	average
SI-SWE-64			4	**14.61**	**14.65**	**30.53**	**19.93**
			1	27.78	16.05	56.94	33.59
	✗		-	29.98	26.62	63.46	40.02
		✗	4	16.75	14.98	38.78	23.50
	✗	✗	-	114.93	37.33	122.00	91.42
SI-SWE-256			4	**13.86**	**11.68**	**28.23**	**17.92**
			1	14.56	14.99	48.83	26.12
	✗		-	32.43	27.44	53.26	37.71
		✗	4	15.50	13.25	33.80	20.85
	✗	✗	-	94.70	58.08	36.98	63.25

Table 8: Diversity loss ablation results in terms of FID. ✗indicates removing the corresponding loss term, and the results for $\lambda_{cd} = 4$, our recommended setting, are **bolded**.

Diversity Loss Settings

To investigate the impact of the diversity loss terms $\mathcal{L}_{cd}$ and $\mathcal{L}_{sd}$, and the hyper-parameter λ_{cd}, on the synthesis fidelity of SI-SWE, we evaluate the FID scores when removing the loss items and varying the weight parameter. As shown in Table 8, removing the diversity loss term or reducing λ_{cd} leads to degraded fidelity, demonstrating the effectiveness and necessity of this component, while we recommend setting λ_{cd} to 4.

5 Conclusions

In this paper, we have proposed SI-SWE, a synthesised-based SWE method which is fundamentally immune to detection by steganalysis tools. SI-SWE not only guarantees high image synthesis fidelity, but also achieves reliable message recovery under a variety of image attacks, while the messages hidden in container images are inaccessible without knowing the right secure factor to reduce the risk of message leakage. Compared to other state-of-the-art synthesis-based SWE methods, SI-SWE is shown to effectively prevent message leakage while achieving superior performance in terms of recovery robustness and synthesis fidelity. In future work, we plan to further enlarge the hidden capacity of our approach, which, in comparison to some other techniques such as full-image hiding methods, is still relatively small.

Acknowledgements

This work was supported in part by the National Innovation 2030 Major S&T Project of China (2020AAA0104203), the National Natural Science Foundation of China (62006007, 61602527), the Natural Science Foundation of Hunan Province, China (2022GK5002, 2020JJ4746), the Special Foundation for Distinguished Young Scientists of Changsha (kq2209003), and in part by 111 Project (No. D23006).

References

[Baluja, 2017] Shumeet Baluja. Hiding images in plain sight: Deep steganography. *Advances in Neural Information Processing Systems*, 30:2069–2079, 2017.

[Baluja, 2019] Shumeet Baluja. Hiding images within images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1685–1697, 2019.

[Bandyopadhyay *et al.*, 2008] Samir K Bandyopadhyay, Debnath Bhattacharyya, Debashis Ganguly, Swarnendu Mukherjee, and Poulami Das. A tutorial review on steganography. In *International Conference on Contemporary Computing*, volume 101, pages 105–114, 2008.

[Boehm, 2014] Benedikt Boehm. StegExpose - a tool for detecting LSB steganography. *arXiv preprint arXiv:1410.6656*, 2014.

[Cheddad *et al.*, 2010] Abbas Cheddad, Joan Condell, Kevin Curran, and Paul Mc Kevitt. Digital image steganography: Survey and analysis of current methods. *Signal Processing*, 90(3):727–752, 2010.

[Goodfellow *et al.*, 2020] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[Holub and Fridrich, 2012] Vojtěch Holub and Jessica Fridrich. Designing steganographic distortion using directional filters. In *IEEE International Workshop on Information Forensics and Security*, pages 234–239, 2012.

[Holub *et al.*, 2014] Vojtěch Holub, Jessica Fridrich, and Tomáš Denemark. Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security*, 2014(1):1–13, 2014.

[Hu *et al.*, 2018] Donghui Hu, Liang Wang, Wenjie Jiang, Shuli Zheng, and Bin Li. A novel image steganography method via deep convolutional generative adversarial networks. *IEEE Access*, 6:38303–38314, 2018.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[Karras *et al.*, 2020] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[Kazemi *et al.*, 2019] Hadi Kazemi, Seyed Mehdi Iranmanesh, and Nasser Nasrabadi. Style and content disentanglement in generative adversarial networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 848–856, 2019.

[Li *et al.*, 2020] Jun Li, Ke Niu, Liwei Liao, Lijie Wang, Jia Liu, Yu Lei, and Minqing Zhang. A generative steganography method based on WGAN-GP. In *International Conference on Artificial Intelligence and Security*, pages 386–397, 2020.

[Liu *et al.*, 2020] Qiang Liu, Xuyu Xiang, Jiaohua Qin, Yun Tan, Junshan Tan, and Yuanjing Luo. Coverless steganography based on image retrieval of DenseNet features and DWT sequence mapping. *Knowledge-Based Systems*, 192:105375, 2020.

[Liu *et al.*, 2022] Xiyao Liu, Ziping Ma, Junxing Ma, Jian Zhang, Gerald Schaefer, and Hui Fang. Image disentanglement autoencoder for steganography without embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2303–2312, June 2022.

[Luo *et al.*, 2020] Yuanjing Luo, Jiaohua Qin, Xuyu Xiang, and Yun Tan. Coverless image steganography based on multi-object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[Marvel *et al.*, 1999] Lisa M Marvel, Charles G Boncelet, and Charles T Retter. Spread spectrum image steganography. *IEEE Transactions on Image Processing*, 8(8):1075–1083, 1999.

[Peng *et al.*, 2022] Fei Peng, Guanfu Chen, and Min Long. A robust coverless steganography based on generative adversarial networks and gradient descent approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[Pevnỳ *et al.*, 2010] Tomáš Pevnỳ, Tomáš Filler, and Patrick Bas. Using high-dimensional image models to perform highly undetectable steganography. In *International Workshop on Information Hiding*, pages 161–177, 2010.

[Shaoping *et al.*, 2021] Lu Shaoping, Wang Rong, Zhong Tao, and L. Rosin Paul. Large-capacity image steganography based on invertible neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10816–10825, 2021.

[Sharp, 2001] Toby Sharp. An implementation of key-based digital signal steganography. In *International Workshop on Information Hiding*, pages 13–26, 2001.

[Tang *et al.*, 2017] Weixuan Tang, Shunquan Tan, Bin Li, and Jiwu Huang. Automatic steganographic distortion learning using a generative adversarial network. *IEEE Signal Processing Letters*, 24(10):1547–1551, 2017.

[Tang *et al.*, 2020] Weixuan Tang, Bin Li, Mauro Barni, Jin Li, and Jiwu Huang. An automatic cost learning framework for image steganography using deep reinforcement learning. *IEEE Transactions on Information Forensics and Security*, 16:952–967, 2020.

[Van Schyndel *et al.*, 1994] Ron G Van Schyndel, Andrew Z Tirkel, and Charles F Osborne. A digital watermark. In *1st International Conference on Image Processing*, volume 2, pages 86–90, 1994.

[Wang *et al.*, 2018] Zihan Wang, Neng Gao, Xin Wang, Xuexin Qu, and Linghui Li. SSteGAN: Self-learning steganography based on generative adversarial networks.

In *International Conference on Neural Information Processing*, pages 253–264, 2018.

[Wengrowski and Dana, 2019] Eric Wengrowski and Kristin Dana. Light field messaging with deep photographic steganography. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1515–1524, 2019.

[Westfeld and Pfitzmann, 1999] Andreas Westfeld and Andreas Pfitzmann. Attacks on steganographic systems. In *International Workshop on Information Hiding*, pages 61–76, 1999.

[Xu *et al.*, 2016] Guanshuo Xu, Han-Zhou Wu, and Yun-Qing Shi. Structural design of convolutional neural networks for steganalysis. *IEEE Signal Processing Letters*, 23(5):708–712, 2016.

[Xu *et al.*, 2022] Youmin Xu, Chong Mou, Yujie Hu, Jingfen Xie, and Jian Zhang. Robust invertible image steganography. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7875–7884, 2022.

[Ye *et al.*, 2017] Jian Ye, Jiangqun Ni, and Yang Yi. Deep learning hierarchical representations for image steganalysis. *IEEE Transactions on Information Forensics and Security*, 12(11):2545–2557, 2017.

[You *et al.*, 2022] Zhengxin You, Qichao Ying, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Image generation network for covert transmission in online social network. In *30th ACM International Conference on Multimedia*, pages 2834–2842, 2022.

[Yu *et al.*, 2015] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[Yu *et al.*, 2021a] Cong Yu, Donghui Hu, Shuli Zheng, Wenjie Jiang, Meng Li, and Zhong-qiu Zhao. An improved steganography without embedding based on attention GAN. *Peer-to-Peer Networking and Applications*, 14(3):1446–1457, 2021.

[Yu *et al.*, 2021b] Ning Yu, Guilin Liu, Aysegul Dundar, Andrew Tao, Bryan Catanzaro, Larry S Davis, and Mario Fritz. Dual contrastive loss and attention for GANs. In *IEEE/CVF International Conference on Computer Vision*, pages 6731–6742, 2021.

[Zheng *et al.*, 2017] Shuli Zheng, Liang Wang, Baohong Ling, and Donghui Hu. Coverless information hiding based on robust image hashing. In *International Conference on Intelligent Computing*, pages 536–547, 2017.

[Zhou *et al.*, 2015] Zhili Zhou, Huiyu Sun, Rohan Harit, Xianyi Chen, and Xingming Sun. Coverless image steganography without embedding. In *International Conference on Cloud Computing and Security*, pages 123–132, 2015.

[Zhu *et al.*, 2018] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. HiDDeN: Hiding data with deep networks. In *European Conference on Computer Vision*, pages 657–672, 2018.

[Zou *et al.*, 2022] Liming Zou, Jing Li, Wenbo Wan, QM Jonathan Wu, and Jiande Sun. Robust coverless image steganography based on neglected coverless image dataset construction. *IEEE Transactions on Multimedia*, 2022.