Image Disentanglement Autoencoder for Steganography without Embedding

Xiyao Liu[1], Ziping Ma[2,1], Junxing Ma[1], Jian Zhang[1,*], Gerald Schaefer[3] and Hui Fang[3]
[1]School of Computer Science and Engineering, Central South University, Changsha, China
[2]Shenzhen Graduate School, Peking University, Shenzhen, China
[3]Department of Computer Science, Loughborough University, Loughborough, U.K.
lxyzoewx@csu.edu.cn maziping.im@gmail.com mjx2021@csu.edu.cn
jianzhang@csu.edu.cn gerald.schaefer@ieee.org H.Fang@lboro.ac.uk

Abstract

Conventional steganography approaches embed a secret message into a carrier for concealed communication but are prone to attack by recent advanced steganalysis tools. In this paper, we propose Image DisEntanglement Autoencoder for Steganography (IDEAS) as a novel steganography without embedding (SWE) technique. Instead of directly embedding the secret message into a carrier image, our approach hides it by transforming it into a synthesised image, and is thus fundamentally immune to typical steganalysis attacks. By disentangling an image into two representations for structure and texture, we exploit the stability of structure representation to improve secret message extraction while increasing synthesis diversity via randomising texture representations to enhance steganography security. In addition, we design an adaptive mapping mechanism to further enhance the diversity of synthesised images when ensuring different required extraction levels. Experimental results convincingly demonstrate IDEAS to achieve superior performance in terms of enhanced security, reliable secret message extraction and flexible adaptation for different extraction levels, compared to state-of-the-art SWE methods.

1. Introduction

Steganography allows to conceal a secret message into a carrier medium such as an image to secure its transmission without being noticed [10, 11, 25]. Traditional steganography methods embed a secret message into the least significant bits (LSBs) [29] or texture-rich regions [7] of a carrier image to aim for undetectability of the message. To further increase payload capacity, recent deep learning-based steganography methods have been proposed to achieve both acceptable imperceptibility and small extraction errors of secret messages [1, 2, 19]. However, since all these methods modify the carrier image, there is an inherent risk of the carried message being compromised through the application of machine learning-based steganalysis tools.

Steganography without embedding (SWE) is an emerging concept to hide a secret message without directly embedding it into a carrier, and thus has the unique advantage that it is immune to typical steganalysis attacks [39]. There are two types of SWE techniques. A mapping mechanism can be designed to transform the secret message into a sequence of image hashes selected from an existing image set [18, 38, 41]. The key weakness here is that the payload capacity is very small. Alternatively, deep synthesis methods, such as generative adversarial networks (GANs) [9, 30, 36], can be trained to generate a synthesised image by passing the secret message into a deep generator network.

Despite promising results achieved by synthesis-based SWE methods [16], there are three drawbacks that hinder their further use in real-world applications: (1) synthesised images from GANs are not sufficiently realistic, compromising the imperceptibility requirement; (2) the trained generator has limited synthesis diversity, thus raising potential security issues; and, more importantly, (3) it is difficult to ensure a small message extraction error when training an invertible neural network to extract the hidden message.

In this paper, we propose Image DisEntanglement Autoencoder for Steganography (IDEAS) to tackle the above issues of current synthesis-based SWE approaches. Specifically, we train an adversarial autoencoder which includes an encoder to disentangle images into structure and texture representations and a decoder to take these two representations to synthesise realistic images. Our approach can guarantee high quality image synthesis to enhance imperception of the hidden data. Furthermore, we explore the advantages that disentanglement brings, namely that the stability of structure representation improves secret message extraction accuracy while randomised texture vectors increase the style diversity of the synthesised images to enhance secu-

*Corresponding author.

rity. In addition, for scenarios where a particular acceptable extraction level is required, we design an adaptive mapping strategy to achieve enhanced hidden capacity and synthesis diversity to improve steganography performance.

Our main contribution in this paper is that we present a novel disentanglement SWE algorithm to achieve state-of-the-art steganographic performance in terms of enhanced security, reliable secret message extraction and flexible adaptation for different extraction levels. In particular, we highlight which aspects of our approach correspond to this improved performance:

- **Enhanced security:** our proposed disentanglement SWE method ensures high-quality image synthesis as well as increases style diversity to minimise the risk of compromise;
- **Reliable message extraction:** we use the structural stability of synthesised images to improve message extraction accuracy. It is worth noting that in some purposely designed settings we can achieve lossless message extraction;
- **Flexible adaptation to different extraction levels**: we design an adaptive mapping mechanism to represent flexible message-to-noise mapping. Through the use of different mapping functions, we achieve flexible hidden capacity and synthesis diversity corresponding to several secret message extraction levels which can be adapted to scenarios with specific requirements.

The code for IDEAS is made available at https://github.com/Lemok00/IDEAS.

2. Related Work

We can generally categorise image steganography methods into approaches based on embedding (SE) and without embedding (SWE), depending on whether the secret information is embedded into a carrier image directly.

2.1. Steganography based on embedding

Traditional SE methods embed secret information by modifying images based on domain knowledge. For example, [14] embeds a secret message by modifying the 4 LSBs of a carrier image. Due to the development of advanced steganalysis techniques, SE methods need to improve their undetectability against steganalysis tools. Thus, [24] proposes to use high-dimensional models covering various dependencies in natural images for highly undetectable steganography, while [7, 8] propose a wavelet-obtained weight-based (WOW) method to ensure the embedding process of only texture-rich or noisy regions and extends this to an arbitrary domain.

Deep learning (DL)-based SE methods aim for further improvements in terms of undetectability, embedding capacity or extraction robustness. For enhanced undetectability, GANs are often used to optimise pixel-level embedding costs by detecting the most suitable parts of an image [28, 34]. [27] designs an automatic cost learning framework based on deep reinforcement learning which is more appropriate for embedding cost optimisation, thus leading to improved undetectability. In order to increase the embedding capacity, [1] proposes to compromise the lossless extraction requirement to allow embedding a full-size image into another image, while [2] and [19] extend this approach to embed multiple images into a single image. Moreover, [26, 31, 40] propose to train DL-based models with perturbation pipelines or specialised dataset for extraction robustness to ensure precise message recovery from modified or camera-captured images.

2.2. Steganography without embedding

Mapping-based SWE methods establish one-to-one mappings between secret messages and image hashes. [39] generates image hash sequences by comparing the mean values of image sub-blocks and uses these images to represent the secret message. To resist subjective visual detection of transmitted images, [38] introduces a latent Dirichlet allocation (LDA)-based classification mechanism and uses DCT features for secret information mapping. [18] utilises DWT-based features as image hashes and further designs a retrieval mechanism based on DenseNet to ensure the similarity of candidate images, thus reducing the possibility of detection from manual inspection. To further enhance robustness against geometric attacks, [17, 20, 21] extract DenseNet features and object labels recognised by Faster-RCNN for secret information mapping.

Synthesis-based SWE methods utilise recent advanced computer graphic techniques to generate realistic images to represent the secret message. [32] proposes to synthesise a new arbitrary-sized texture image by mimicking texture patches sampled from an original reference image. SSteGAN [30] comprises a DL-based generator to synthesise images for steganography and a corresponding extractor to extract the secret message. [9] generates a mapping function to transfer a secret message into a random noise vector to further enhance both imperception and security. In addition, self-attention blocks [36] and Wasserstein loss with gradient penalty [15] are introduced to improve the performance of its generator. For further increased embedding capacity, [5] uses face images as steganography containers based on image selection and a StarGAN, while [4] generates images of anime characters with specific attributes such as hair colour, which represent specific binary sequences based on a codebook, for enhanced extraction stability.

3. Method

In this paper, we propose Image DisEntanglement Autoencoder for Steganography (IDEAS) as a novel image disentanglement-based SWE algorithm. Our IDEAS net-

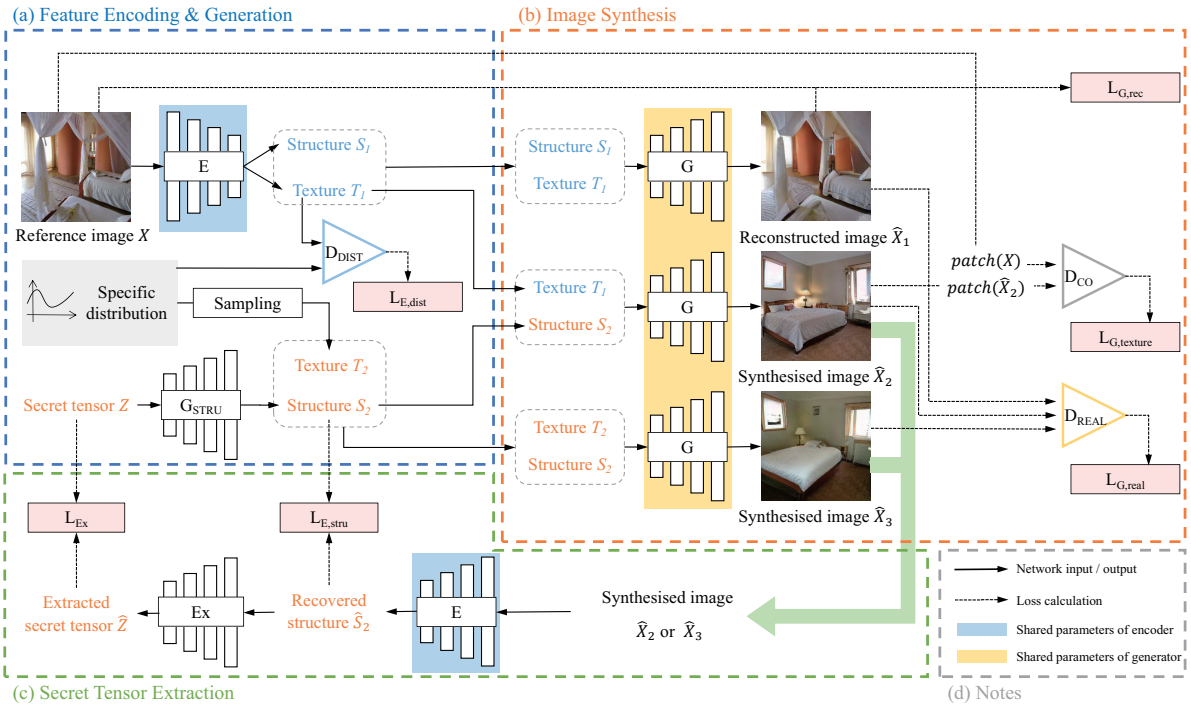

Figure 1. Training flowchart of proposed IDEAS network, which consists of three parts: (a) encoding structure and texture features, (b) synthesising images, (c) extracting input secret tensor. E=disentanglement encoder; G=generator; G_{STRU}=structure generator; Ex=tensor extractor; D_{REAL}=image-level adversarial discriminator; D_{CO}=co-occurrence discriminator; D_{DIST}=distribution discriminator.

work takes a secret message and a sampled texture vector as its input to synthesise a realistic image and extracts the hidden secret message via a disentangling encoder. In particular, we use structure representation to encode the secret message and texture representation to enrich synthesis diversity. To further enhance synthesis style diversity for steganography security, we design an adaptive mechanism to map the secret message to the structure representation based on different required levels of extraction accuracy. In the following, we explain the IDEAS network architecture, its training loss terms and its processing pipeline in detail.

3.1. IDEAS network architecture

The key components of our IDEAS network, illustrated in Figure 1, are a disentanglement encoder and a generator. The encoder allows to decompose images into structure tensor and texture vector representations and to extract the hidden secret message, while the generator takes the structure tensor generated from the secret tensor (*i.e.*, the encoded secret message) and the texture vector sampled from a uniform distribution to synthesise a realistic image for data hiding. Training is performed based on multiple loss terms to simultaneously ensure both high quality synthesis and minimal extraction error of the secret message with diversified synthesis style.

3.1.1 Feature encoding and generation

We train an encoder E to disentangle an image X into its structure feature S_1 and texture feature T_1, *i.e.*

$$(S_1, T_1) = E(X), \tag{1}$$

where X is randomly sampled from a specific dataset. Inspired by [22], a distribution discriminator D_{DIST} is used to enforce that T_1 conforms to the uniform distribution $U(-1, 1)$ by a distribution discriminative loss

$$L_{E,dist} = \text{softplus}(-D_{DIST}(T_1)), \tag{2}$$

where $\text{softplus}(x) = \log(1 + e^x)$. That is, T_1 conforms to the same distribution as a sampled T_2 from $U(-1, 1)$. Additionally, a structure generator G_{STRU} is introduced to translate a uniform sampled tensor Z to a structure distribution to obtain S_2 for hiding the secret message as

$$S_2 = G_{STRU}(Z), \tag{3}$$

where Z corresponds to the hidden secret message via a mapping mechanism (explained further below). The size of Z is $N \times \frac{H}{16} \times \frac{W}{16}$ for an RGB image X of size $3 \times H \times W$, with N a hyper-parameter.

3.1.2 Image synthesis

For disentangled image generation, three images are generated by the same generator G using different combinations

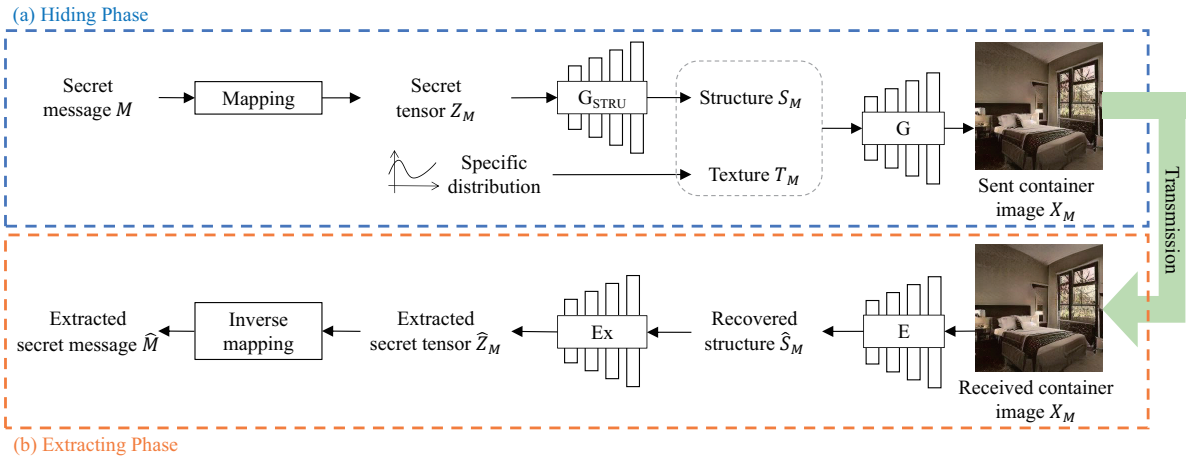

Figure 2. Flow chart of IDEAS for concealed communication, which consists of two parts: (a) hiding phase, and (b) extraction phase.

of structure and texture representations, namely

$$\begin{aligned}\hat{X}_1 &= G(S_1, T_1),\\ \hat{X}_2 &= G(S_2, T_1),\\ \hat{X}_3 &= G(S_2, T_2).\end{aligned} \tag{4}$$

We adopt StyleGAN2 [13] as the backbone architecture of our generator for high quality image synthesis. When X is reconstructed as $\hat{X}_1$, the reconstruction loss is calculated as

$$L_{G,rec} = \| X - \hat{X}_1 \|^1, \tag{5}$$

where $\| \cdot \|^1$ denotes the L1 loss between images.

$\hat{X}_2$ is generated with the same texture as X but with different structure S_2. Random cropped patches from $\hat{X}_2$ and X are passed to the co-occurrence discriminator D_{CO}, originally introduced in [23] for texture similarity comparison, to calculate the texture loss as

$$L_{G,texture} = D_{CO}(\text{patch}(\hat{X}_2), \text{patch}(X)), \tag{6}$$

where patch$(\cdot)$ denotes the random cropping function.

$\hat{X}_3$ is generated from S_2 and T_2 to synthesise a realistic image which has no relation with X. To ensure realistic synthesis, an adversarial loss term $L_{G,real}$ is introduced to make all synthesised images $\hat{X}_{1,2,3}$ indiscriminative from real images and is calculated as

$$L_{G,real} = D(\hat{X}_1) + D(\hat{X}_2) + D(\hat{X}_3), \tag{7}$$

where D denotes the discriminator with the same architecture.

3.1.3 Secret tensor extraction

To extract structure representations from the generated image and extract the input secret tensor from this structure, the encoder E and extractor Ex are designed as

$$\hat{Z} = Ex(\hat{S}_2) \tag{8}$$

with

$$\hat{S}_2 = E(\hat{X}_{2\,||\,3}). \tag{9}$$

We use two strategies at the training stage to avoid overfitting problems. First, the encoder E for extracting the structure representations from synthesised images shares parameters with the encoder to disentangle the original images. Second, for structure representation extraction from synthesised images, we train E using $\hat{X}_2$ as input for 80% of the iterations to improve recovery accuracy and $\hat{X}_3$ for the last 20% to improve recovery resilience. This split is relatively insensitive and we set the ratio empirically in our experiments.

The structure extraction loss $L_{E,stru}$ and tensor extracting loss L_{Ex} are calculated as

$$L_{E,stru} = \| \hat{S}_2 - S_2 \|^1 \tag{10}$$

and

$$L_{Ex} = \| \hat{Z} - Z \|^1, \tag{11}$$

respectively.

3.1.4 Loss function

We combine the loss terms introduced above to form the total loss function. Since $L_{G,rec}$, $L_{G,texture}$ and $L_{G,real}$ collaborate for image synthesis, they are combined to yield the generation loss

$$L_G = L_{G,rec} + L_{G,texture} + 2L_{G,real}, \tag{12}$$

where the higher weight for $L_{G,real}$ ensures generation quality.

Similarly, $L_{E,dist}$ and $L_{E,stru}$ are combined to define the encoding loss

$$L_E = L_{E,dist} + L_{E,stru}. \tag{13}$$

The total loss L_{total} to train the encoder E, generator G, structure generator G_{STRU} and extractor Ex in IDEAS algorithm is then formulated as

$$L_{total} = L_G + L_E + \lambda_{Ex} L_{Ex}, \tag{14}$$

where λ_{Ex} allows to balance between synthesis quality and extraction accuracy.

3.2. IDEAS processing pipeline

As shown in Figure 2, IDEAS comprises two stages, a secret message hiding phase, and an extraction phase.

3.2.1 Secret message hiding

This part is designed for the message sender, who first maps a secret message M to the secret tensor Z_M which is then translated to S_M by G_{STRU}. Combining S_M with a uniform sampled T_M from $U(-1, 1)$, the container image X_M is generated by G and can be transmitted to the receiver.

For the mapping function, the secret message is divided into segments of σ bits. Then, the decimal value m corresponding to each segment is mapped to a float value z by

$$z = \frac{m + 0.5}{2^{\sigma-1}} - 1 + rand(-\Delta \times r, \Delta \times r). \tag{15}$$

In this manner, flexible hidden capacities corresponding to different recovery levels can be achieved by adjusting the hyper-parameter σ.

Finally, the float values are concatenated as the secret tensor. We extend random intervals on both sides of the points to establish the adaptive mapping mechanism by adjusting the hyper-parameter Δ to achieve flexible adaptation of enhanced synthesis diversity corresponding to different recovery levels. For each point, its random intervals are of total size $2 \times \Delta \times r$, where r represents the maximum size $1/2^{\sigma-1}$ and Δ is in $[0\%, 50\%]$.

Depending on the value of Δ, our framework has three modes: (a) an extraction-prior mode with $\Delta = 0\%$ for applications requiring highly stable message extraction; (b) a balanced mode with $\Delta = 25\%$; and (c) a diversity-prior mode with $\Delta = 50\%$ for applications requiring enhanced security.

3.2.2 Secret message extraction

To extract the message hidden in X_M, the receiver uses E and Ex to extract $\hat{S}_M$ and $\hat{Z}_M$ sequentially. Then, the inverse mapping function

$$m = \text{floor}\left[(z + 1) \times 2^{\sigma-1}\right] \tag{16}$$

is applied to extract the secret message $\hat{M}$.

Compared to existing SWE methods, our method represents the secret message via the structure of a generated image to achieve higher extraction accuracy due to structural stability. Further, the container image synthesised by the StyleGAN2 generator G is of high fidelity which further improves the imperception of secret message transmission.

4. Experimental Results

4.1. Experimental setup

To demonstrate the superiority of IDEAS, we compare it with four state-of-the-art synthesis-based SWE methods, namely DCGAN-Steg [9], SAGAN-Steg [36], SSteGAN [30] and WGAN-Steg [15]. Our IDEAS and baseline models are trained on three subsets from two publicly available datasets, Bedroom and Church images from LSUN [37], and face images from FFHQ [12]. Each subset includes 70,000 randomly selected images normalised to 256×256 pixels. We train IDEAS with hyper-parameters $N = \{1, 2\}$ and $\lambda = \{2, 5, 10\}$, respectively. Note that λ is set to 10 for performance evaluation, while other settings of λ are discussed in Subsection 4.6.

We evaluate the models by comparing the area under the curve (AUC) of the receiver operating characteristic (ROC) curves and the Fréchet inception distance (FID) [6] to measure the undetectability by steganalysis tools and subjective visual perception. Meanwhile, we compare the secret message extraction accuracy of IDEAS under different hidden capacities with benchmark SWE methods. All results are obtained on an RTX 2080Ti, where IDEAS, which comprises 64.5M parameters in total, takes 12.0 ms for image synthesis and 9.4 ms for message extraction.

4.2. Security evaluation by steganalysis

To confirm immunity to steganalysis tools, we evaluate the undetectability from several well-known steganalysis tools, including StegExpose [3], XuNet [33] and YeNet [35]. In Table 1, we calculate the area under the curve (AUC) of the receiver operating characteristic (ROC) curves generated by the detections from these steganalysis tools. All AUC values of our method and the benchmark SWE methods are smaller or close to 0.5, indicating detection rates that essentially correspond to random guessing. This demonstrates that SWE methods fundamentally resist detection by current steganalysis tools since the container images are synthesised without embedding modifications.

4.3. Security evaluation by image synthesis quality

Since SWE methods are immune to typical steganalysis tools, perceptual indetectability is very important. Examples of synthesised container images of all models are given in Figure 3, while FID scores between real and synthesised images are listed in Table 2.

From Figure 3(a), we can see that images synthesised by IDEAS are of much higher fidelity with more realistic

	IDEAS $N=1$			IDEAS $N=2$			DCGAN-Steg	SAGAN-Steg	SSteGAN	WGAN-Steg
	Δ=0%	Δ=25%	Δ=50%	Δ=0%	Δ=25%	Δ=50%				
StegExpose	0.480	0.502	0.542	0.456	0.471	0.484	0.586	0.578	0.417	0.594
XuNet	0.404	0.398	0.371	0.413	0.407	0.403	0.568	0.500	0.491	0.542
YeNet	0.521	0.512	0.520	0.533	0.528	0.535	0.573	0.569	0.519	0.548

Table 1. Undetectability by steganalysis in terms of AUC of ROC.

structures and clearer textures compared to those from other algorithms. In addition, when hiding an identical secret message, as demonstrated in Figure 3(b), our method produces diversified styles due to the uniformly sampled texture representation. IDEAS is thus clearly superior to other models in terms of synthesised image quality and supporting enhanced steganography security. The reason for this is twofold: (i) the use of disentanglement structure feature ensures high image fidelity, and (ii) the use of uniformly sampled texture vectors leads to enhanced image diversity. We can also notice from Figure 3(b), that larger Δ values lead to diverse structure presentations, which demonstrates that our designed adaptive mapping mechanism can further enhance imperception of the secret message.

As shown in Table 2, IDEAS achieves the lowest FID scores on all datasets, outperforming the other techniques by a wide margin. The best results are achieved for the Bedroom dataset and the worst for the FFHQ facial images. This is because bedroom images have weaker structural constraints compared to facial images, thus making it easier for the disentanglement model to synthesise them.

	Bedrooms	Churches	FFHQ	avg. $\pm$ std.dev.
DCGAN-Steg	283.32	105.79	74.24	154.45 ± 112.71
SAGAN-Steg	159.51	99.59	82.60	113.90 ± 40.41
SSteGAN	153.48	258.80	150.37	187.55 ± 61.73
WGAN-Steg	147.45	181.20	67.95	132.20 ± 58.15
IDEAS $N=1, \sigma=1$				
$\Delta=0\%$	16.88	15.90	32.88	21.89 ± 9.53
$\Delta=25\%$	15.56	15.50	31.10	20.72 ± 8.99
$\Delta=50\%$	13.39	14.48	29.31	19.06 ± 8.89
IDEAS $N=2, \sigma=1$				
$\Delta=0\%$	14.17	17.15	29.76	20.36 ± 8.28
$\Delta=25\%$	14.01	16.32	29.02	19.79 ± 8.08
$\Delta=50\%$	13.51	16.34	28.45	19.44 ± 7.93

Table 2. FID results of synthesised images for all models.

	Bedrooms	Churches	FFHQ	capacity
DCGAN-Steg	94.01%	94.56%	96.29%	100 bits
SAGAN-Steg	96.77%	95.86%	97.12%	200 bits
SSteGAN	98.41%	97.53%	97.23%	100 bits
WGAN-Steg	92.23%	90.04%	92.85%	100 bits
IDEAS $N=1, \sigma=1$				
$\Delta=0\%$	100%	100%	100%	
$\Delta=25\%$	100%	100%	100%	256 bits
$\Delta=50\%$	99.54%	99.55%	99.49%	
IDEAS $N=2, \sigma=1$				
$\Delta=0\%$	100%	100%	100%	
$\Delta=25\%$	100%	100%	100%	512 bits
$\Delta=50\%$	99.32%	99.29%	99.42%	

Table 3. Extraction accuracy results for all models.

4.4. Extraction accuracy

Extraction accuracy results of the different SWE methods, together with their corresponding hidden capacities, are given in Table 3. It is obvious that IDEAS outperforms all other methods, providing both better secret message extraction accuracy and higher hidden capacity. In particular, when Δ is set to 0% or 25%, the obtained accuracies are 100% for all databases, a level that is not reached by any of the other models for any dataset. In addition, the hidden capacity of IDEAS is higher than that of the other approaches. When we set $N=2$, our method achieves higher hidden capacity with only an insignificant compromise on the extraction accuracy rates, outperforming the other four methods. The reason for this remarkable extraction accuracy is our use of structural stability when disentangling images for hiding secret messages.

4.5. Extraction accuracy/hidden capacity flexibility

By adjusting the values of N and σ, we can enlarge the hidden capacity of IDEAS, which is $256 \times \sigma \times N$. We thus further evaluate the extraction accuracy of secret messages of IDEAS with different hidden capacities and show

		Δ=0%	Δ=25%	Δ=50%	capacity
$N=1$ $\sigma=1$	Bedrooms	100%	100%	99.54%	256 bits
	Churches	100%	100%	99.55%	
	FFHQ	100%	100%	99.49%	
$N=2$ $\sigma=1$	Bedrooms	100%	100%	99.32%	512 bits
	Churches	100%	100%	99.29%	
	FFHQ	100%	100%	99.42%	
$N=1$ $\sigma=2$	Bedrooms	100%	100%	99.18%	512 bits
	Churches	100%	100%	99.17%	
	FFHQ	100%	100%	99.06%	
$N=2$ $\sigma=2$	Bedrooms	100%	100%	98.75%	1024 bits
	Churches	100%	100%	98.70%	
	FFHQ	100%	100%	98.99%	
$N=1$ $\sigma=3$	Bedrooms	100%	100%	98.60%	768 bits
	Churches	100%	100%	98.58%	
	FFHQ	100%	100%	98.37%	
$N=2$ $\sigma=3$	Bedrooms	100%	99.99%	97.79%	1536 bits
	Churches	100%	99.99%	97.76%	
	FFHQ	100%	100%	98.26%	

Table 4. IDEAS extraction accuracy results for different values of N and σ.

Figure 3. Examples of synthesised container images from DCGAN-Steg, SAGAN-Steg, SSteGAN, WGAN-Steg and IDEAS on LSUN Bedrooms (left), LSUN Churches (middle) and FFHQ (right). (a) comparison of fidelity of images synthesised from different secret messages. (b) comparison of diversity of images synthesised from an identical secret message.

			$\lambda_{Ex}=2$		$\lambda_{Ex}=5$		$\lambda_{Ex}=10$	
			$N=1$	$N=2$	$N=1$	$N=2$	$N=1$	$N=2$
$\sigma=1$	Bedrooms	$\Delta=0\%$	99.98% / 12.31	99.89% / 12.33	100% / 15.75	99.89% / 12.85	100% / 16.88	100% / 14.17
		$\Delta=25\%$	99.95% / 11.71	99.66% / 11.98	100% / 14.33	99.42% / 11.44	100% / 15.56	100% / 14.01
		$\Delta=50\%$	97.41% / 10.74	95.26% / 11.42	99.00% / 12.93	96.23% / 12.50	99.54% / 13.39	99.32% / 13.51
	Churches	$\Delta=0\%$	100% / 13.83	100% / 14.66	100% / 15.12	100% / 16.08	100% / 15.90	100% / 17.15
		$\Delta=25\%$	100% / 13.27	99.97% / 14.68	100% / 14.65	100% / 15.81	100% / 15.50	100% / 16.32
		$\Delta=50\%$	98.90% / 12.61	96.17% / 14.69	99.32% / 13.90	99.06% / 15.76	99.55% / 14.48	99.29% / 16.34
	FFHQ	$\Delta=0\%$	100% / 25.71	99.97% / 22.22	100% / 31.63	100% / 26.95	100% / 32.88	100% / 29.76
		$\Delta=25\%$	100% / 25.15	99.56% / 22.29	100% / 30.18	100% / 25.85	100% / 31.10	100% / 29.02
		$\Delta=50\%$	99.18% / 22.90	94.59% / 22.18	99.39% / 26.00	99.17% / 24.07	99.49% / 29.31	99.42% / 28.45
$\sigma=2$	Bedrooms	$\Delta=0\%$	99.80% / 11.14	89.49% / 11.68	99.98% / 13.32	96.99% / 11.81	100% / 14.21	100% / 13.49
		$\Delta=25\%$	99.50% / 10.84	88.21% / 11.60	99.98% / 13.11	98.42% / 12.99	100% / 13.74	100% / 13.62
		$\Delta=50\%$	95.23% / 10.39	83.80% / 11.41	98.08% / 13.10	92.94% / 12.51	99.18% / 13.37	98.75% / 13.43
	Churches	$\Delta=0\%$	99.99% / 12.59	93.30% / 14.39	100% / 14.05	100% / 15.54	100% / 14.62	100% / 16.36
		$\Delta=25\%$	99.98% / 12.69	92.40% / 14.57	100% / 14.22	100% / 15.72	100% / 14.60	100% / 16.50
		$\Delta=50\%$	97.93% / 12.52	87.86% / 14.33	98.76% / 13.91	98.22% / 16.13	99.17% / 14.64	98.70% / 16.69
	FFHQ	$\Delta=0\%$	99.99% / 24.08	94.09% / 22.67	100% / 27.45	100% / 24.54	100% / 29.56	100% / 27.26
		$\Delta=25\%$	99.99% / 23.58	92.26% / 22.33	100% / 27.16	100% / 24.36	100% / 28.08	100% / 26.70
		$\Delta=50\%$	98.37% / 22.95	87.37% / 22.25	98.89% / 25.79	98.43% / 24.05	99.06% / 27.92	98.99% / 26.61
$\sigma=3$	Bedrooms	$\Delta=0\%$	97.87% / 10.54	76.34% / 11.74	99.93% / 12.87	89.04% / 12.76	100% / 13.62	100% / 13.63
		$\Delta=25\%$	96.56% / 10.73	75.96% / 11.47	99.83% / 12.93	88.26% / 12.80	100% / 13.41	99.99% / 13.39
		$\Delta=50\%$	91.43% / 10.58	74.31% / 11.45	96.52% / 13.01	84.81% / 12.48	98.60% / 13.56	97.79% / 13.41
	Churches	$\Delta=0\%$	99.93% / 12.48	79.60% / 14.33	100% / 14.05	99.97% / 15.81	100% / 14.44	100% / 16.48
		$\Delta=25\%$	99.79% / 12.59	78.98% / 14.43	99.98% / 14.18	99.92% / 15.85	100% / 14.43	99.99% / 16.33
		$\Delta=50\%$	96.37% / 12.34	76.91% / 16.66	97.84% / 13.90	96.91% / 15.52	98.58% / 14.73	97.76% / 16.23
	FFHQ	$\Delta=0\%$	99.94% / 22.85	81.23% / 22.62	100% / 26.23	99.98% / 23.82	100% / 29.39	100% / 26.93
		$\Delta=25\%$	99.89% / 23.26	79.99% / 22.33	99.99% / 26.43	99.95% / 23.97	100% / 29.05	100% / 26.60
		$\Delta=50\%$	97.10% / 23.00	77.88% / 22.28	98.09% / 25.30	97.22% / 23.70	98.37% / 28.58	98.26% / 26.37

Table 5. IDEAS hyper-parameter analysis in terms of extraction accuracy (values on the left) and FID scores (value on the right).

the results in Table 4. From there, we can notice that the obtained extraction accuracies are excellent for all parameter settings, even with higher hidden capacities. In addition, when increasing the values of N and σ, the drop of extraction accuracy is insignificant. Combinations of different levels of extraction accuracies and hidden capacities yield flexibility for different steganography applications.

4.6. Hyper-parameter settings

Last not least, we conduct an experiment to evaluate how different hyper-parameters settings impact the performance of IDEAS. As shown in Table 5, although the obtained FID scores are fairly consistent when Δ is increased, the synthesis structure diversity is enhanced as shown in Figure 3(b) while the extraction accuracy is slightly decreased, providing a flexible adaptation for different application scenarios.

Regarding N and σ, increasing both parameters can improve the hidden capacity as shown in Table 4. From Table 5, we observe that the effect on extraction accuracy relies heavily on λ_{Ex} although an increase of both parameters leads to an accuracy drop.

When setting $\lambda_{Ex} = 10$, the accuracy drops are relatively insignificant while the synthesised image quality is somewhat more compromised. When $\lambda_{Ex} = \{2, 5\}$, the extraction accuracies are lower than 90% for the Bedrooms dataset when $\sigma = 3$. In contrast, for $\lambda_{Ex} = 10$, all settings achieve lossless extraction when $\Delta = 0\%$. In addition, all extraction accuracies with different combinations are higher than 97.76% on all datasets. Consequently, $\lambda_{Ex} = 10$ is set to maximise the synthesised image quality while ensuring high extraction accuracy.

5. Conclusions

In this paper, we have proposed IDEAS, a structure-texture disentanglement synthesis-based SWE algorithm to resist steganalysis attacks. IDEAS not only generates high-fidelity synthesised images of rich diversity but also yields high secret message extraction accuracy. Compared to other state-of-the-art synthesis-based SWE methods, IDEAS achieves better performance in terms of security and extraction accuracy of secret message.

Despite its excellent performance, the hidden capacity is relatively smaller compared to some other techniques such as full-image-to-image hiding methods. Thus, in future work, we plan to further enlarge the hidden capacity.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (61602527,U1734208), and Natural Science Foundation of Hunan Province, China (2020JJ4746).

References

[1] Shumeet Baluja. Hiding images in plain sight: Deep steganography. *Advances in Neural Information Processing Systems*, 30:2069–2079, 2017. 1, 2

[2] Shumeet Baluja. Hiding images within images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1685–1697, 2019. 1, 2

[3] Benedikt Boehm. StegExpose – a tool for detecting lsb steganography. *arXiv preprint arXiv:1410.6656*, 2014. 5

[4] Yi Cao, Zhili Zhou, QM Jonathan Wu, Chengsheng Yuan, and Xingming Sun. Coverless information hiding based on the generation of anime characters. *EURASIP Journal on Image and Video Processing*, 2020(1):1–15, 2020. 2

[5] Xianyi Chen, Zhentian Zhang, Anqi Qiu, Zhihua Xia, and Naixue Xiong. A novel coverless steganography method based on image selection and StarGAN. *IEEE Transactions on Network Science and Engineering*, 2020. 2

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 5

[7] Vojtěch Holub and Jessica Fridrich. Designing steganographic distortion using directional filters. In *IEEE International Workshop on Information Forensics and Security*, pages 234–239, 2012. 1, 2

[8] 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. 2

[9] 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. 1, 2, 5

[10] Mehdi Hussain, Ainuddin Wahid Abdul Wahab, Yamani Idna Bin Idris, Anthony TS Ho, and Ki-Hyun Jung. Image steganography in spatial domain: A survey. *Signal Processing: Image Communication*, 65:46–66, 2018. 1

[11] Inas Jawad Kadhim, Prashan Premaratne, Peter James Vial, and Brendan Halloran. Comprehensive survey of image steganography: Techniques, evaluations, and trends in future research. *Neurocomputing*, 335:299–326, 2019. 1

[12] 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. 5

[13] 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. 4

[14] Charles W Kurak Jr and John McHugh. A cautionary note on image downgrading. In *8th Annual Computer Security Application Conference*, pages 153–159, 1992. 2

[15] 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. Springer, 2020. 2, 5

[16] Qi Li, Xingyuan Wang, Bin Ma, Xiaoyu Wang, Chunpeng Wang, Zhiqiu Xia, and Yunqing Shi. Image steganography based on style transfer and quaternion exponent moments. *Applied Soft Computing*, page 107618, 2021. 1

[17] Qiang Liu, Xuyu Xiang, Jiaohua Qin, Yun Tan, and Yao Qiu. Coverless image steganography based on DenseNet feature mapping. *EURASIP Journal on Image and Video Processing*, 2020(1):1–18, 2020. 2

[18] 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. 1, 2

[19] Shao-Ping Lu, Rong Wang, Tao Zhong, and Paul L Rosin. Large-capacity image steganography based on invertible neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10816–10825, 2021. 1, 2

[20] 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. 2

[21] Yuanjing Luo, Jiaohua Qin, Xuyu Xiang, Yun Tan, Zhibin He, and Neal N Xiong. Coverless image steganography based on image segmentation. *Computers, Materials and Continua*, 64(2):1281–1295, 2020. 2

[22] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015. 3

[23] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. *Advances in Neural Information Processing Systems*, 33:7198–7211, 2020. 4

[24] 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. 2

[25] Mansi S Subhedar and Vijay H Mankar. Current status and key issues in image steganography: A survey. *Computer Science Review*, 13:95–113, 2014. 1

[26] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2117–2126, 2020. 2

[27] 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. 2

[28] 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. 2

[29] 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. 1

[30] 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. Springer, 2018. 1, 2, 5

[31] 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. 2

[32] Kuo-Chen Wu and Chung-Ming Wang. Steganography using reversible texture synthesis. *IEEE Transactions on Image Processing*, 24(1):130–139, 2014. 2

[33] 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. 5

[34] Jianhua Yang, Danyang Ruan, Jiwu Huang, Xiangui Kang, and Yun-Qing Shi. An embedding cost learning framework using gan. *IEEE Transactions on Information Forensics and Security*, 15:839–851, 2019. 2

[35] 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. 5

[36] 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. 1, 2, 5

[37] 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. 5

[38] Xiang Zhang, Fei Peng, and Min Long. Robust coverless image steganography based on dct and lda topic classification. *IEEE Transactions on Multimedia*, 20(12):3223–3238, 2018. 1, 2

[39] 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. 1, 2

[40] 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. 2

[41] Liming Zou, Jiande Sun, Min Gao, Wenbo Wan, and Brij Bhooshan Gupta. A novel coverless information hiding method based on the average pixel value of the sub-images. *Multimedia Tools and Applications*, 78(7):7965–7980, 2019. 1