Sanitizing Hidden Information with Diffusion Models

Preston K. Robinette[a,*], **Daniel Moyer**[a] **and Taylor T. Johnson**[a]

[a]Vanderbilt University

Abstract. Information hiding is the process of embedding data within another form of data, often to conceal its existence or prevent unauthorized access. This process is commonly used in various forms of secure communications (steganography) that can be used by bad actors to propagate malware, exfiltrate victim data, and discreetly communicate. Recent work has utilized deep neural networks to remove this hidden information in a defense mechanism known as sanitization. Previous deep learning works, however, are unable to scale efficiently beyond the MNIST dataset. In this work, we present a novel sanitization method called DM-SUDS that utilizes a diffusion model framework to sanitize/remove hidden information from image-into-image universal and dependent steganography from CIFAR-10 and ImageNet datasets. We evaluate DM-SUDS against three different baselines using MSE, PSNR, SSIM, and NCC metrics and provide further detailed analysis through an ablation study. DM-SUDS outperforms all three baselines and significantly improves image preservation MSE by 50.44%, PSNR by 12.69%, SSIM by 11.49%, and NCC by 3.26% compared to previous deep learning approaches. Additionally, we introduce a novel evaluation specification that considers the successful removal of hidden information (safety) as well as the resulting quality of the sanitized image (utility). We further demonstrate the versatility of this method with an application in an audio case study, demonstrating its broad applicability to additional domains.

1 Introduction

Steganography, or the art of hiding information in plain sight, is an area of information hiding that is used for covert communication. In steganography, a secret message is embedded in a seemingly harmless physical or digital medium, termed a cover. A container is the secret and cover together that is then relayed to an intended recipient. Due to the hidden nature of the secret message, communication can occur discreetly between parties without arousing suspicion.

This process can be applied to both physical mediums (e.g., microdots, invisible ink, embedded objects, etc.) as well as digital mediums (e.g., images, videos, audio, text, etc.). The pervasive nature of digital media makes digital steganography a more dangerous communication channel compared to its physical counterpart. Its widespread accessibility and ease of dissemination make digital platforms more susceptible to covert manipulations, potentially enabling malicious actors to spread concealed information or malware at an unprecedented scale and speed. In addition to the widespread effect of malware propagation, steganography via digital mediums can also be used to exfiltrate victim data and communicate with other bad actors. As such, advanced detection and prevention mechanisms are necessary to combat these harmful use cases. While various digital mediums can be used as covers, we focus on the use of images in this work, as these are more accessible and commonly used in the wild.

Current image steganography defenses utilize a technique known as steganalysis, or the detection of hidden messages in images [11, 1]. Steganalysis tools analyze images for known signatures and/or anomalies in pixel values, noise distributions, and other statistical measures to indicate the presence of steganographic content. Recent work has also incorporated machine learning to enhance detection accuracy [42, 39, 24]. These defense strategies, however, rely upon data curated from preexisting steganographic techniques and are referred to as *non-blind*. While adept at detecting known steganography, these defense mechanisms are useless against new forms of steganography in the wild – especially ones engineered to bypass these existing systems [15, 7, 28].

Another defense mechanism against steganography is sanitization, which eliminates the presence of any potential secret message while maintaining the integrity of the cover media. For instance, a picture of a dog containing an executable malware binary would be sanitized if the malware binary is removed and the original image is minimally changed. The authors in [26] demonstrate the success of such an approach, utilizing a variational autoencoder strategy termed SUDS. The authors show that SUDS is able to protect against least significant bit (LSB), dependent deep hiding (DDH), and universal deep hiding (UDH) steganography, each of which is an unseen method by the sanitizer prior to testing (*blind*).

While SUDS is able to successfully sanitize secrets, its ability to preserve the quality of the resulting image deteriorates as image complexity increases. This effect is a result of utilizing a variational autoencoder approach. To address the limited preservation capabilities of SUDS while maintaining sanitization performance, we propose a **d**iffusion **m**odel approach to **SUDS**, termed DM-SUDS. While diffusion models are most commonly used for their generative properties, they are trained as denoising mechanisms. By using a diffusion model to "denoise" potentially steganographic images, we believe that this approach will provide an improved alternative to SUDS, advancing the state-of-the-art in blind deep learning sanitization techniques for image-into-image steganography. All code to reproduce experiments is available at: **https://github.com/pkrobinette/dmsuds_steg**. The primary contributions of this work are the following.

1. **Implementation of a Novel Sanitization Framework:** We introduce a blind deep learning steganography sanitization method called DM-SUDS that utilizes a diffusion model framework to sanitize image-into-image universal and dependent steganography.
2. **Demonstration of Sanitizer Capabilities:** We show the benefit of our novel DM-SUDS framework on CIFAR-10 and ImageNet datasets by evaluating it on five different capabilities: comparison to three baselines, flexibility of the timestep parameter (ablation

* Corresponding Author. Email: preston.k.robinette@vanderbilt.edu

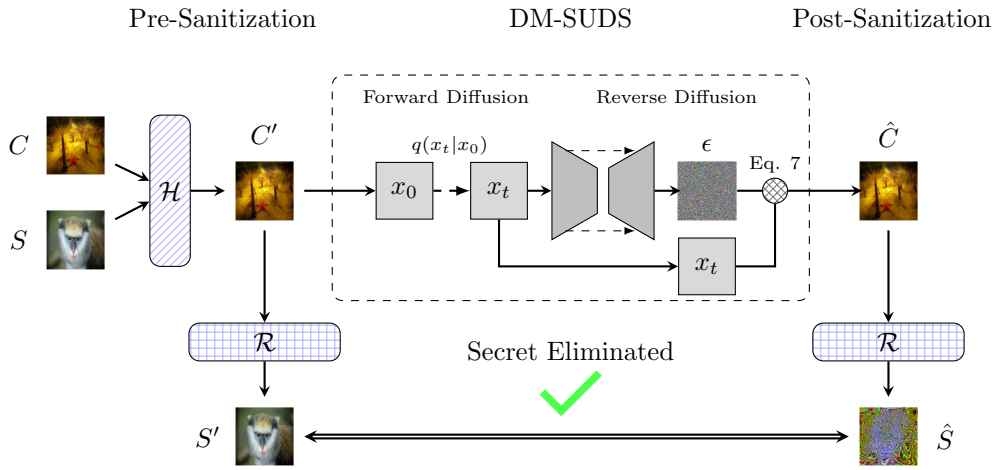


Figure 1. DM-SUDS (center) takes as input a cover or a container image $x_0 \in \{C, C'\}$ created with any type of steganographic technique. A noisy image is then sampled from this image at timestep t. In the reverse diffusion process, a denoising U-Net is used to predict the amount of noise added to the image, which is then used to recover the original image, resulting in a sanitized image $\hat{C}$. In the pre-sanitization phase, the secret is recoverable, as demonstrated in the bottom-left of the figure S'. After sanitization with DM-SUDS, however, a secret is not recoverable, as indicated by the bottom-right of the figure $\hat{S}$. The secret, therefore, is successfully eliminated.

study), ability to directly denoise (ablation study), scalability, and robustness.

3. **Development of a Novel Sanitization Specification:** We introduce a novel sanitization specification that combines secret elimination (safety) and image preservation (utility).
4. **Application to the Audio Domain**: We present a case study that applies DM-SUDS to a new domain (audio), demonstrating the versatility of this approach applied to text-into-audio steganography. This is the first approach to successfully sanitize information hidden in audio containers using deep learning.

2 Preliminaries

In this section, we introduce steganography, existing sanitization techniques, and the metrics used for evaluation. While this work is medium agnostic, **we focus on image-into-image steganography**, where an image secret is embedded into an image cover. An image is represented by a matrix (c, h, w), where c is the number of color channels, h is the height, and w is the width of the image.

For the majority of this work, we utilize the term steganography to refer to any covert or imperceptible embedding of information. This includes invisible watermarking. While steganography and watermarking utilize many of the same concepts, some watermarking methods are meant to be visible on the corresponding watermarked media, as shown in Figure 2. As such, we focus on the removal of steganographic information, which is all invisible, but acknowledge that DM-SUDS is directly applicable to the removal of invisible watermarks as well.

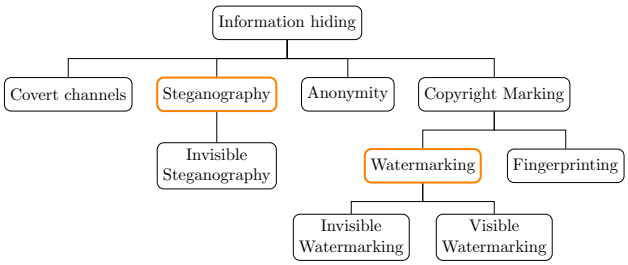


Figure 2. Information Hiding Diagram.

2.1 *Steganography*

Notation. As shown by the *Pre-Sanitization* section in Figure 1, steganography typically consists of a *cover*, *secret*, *container*, and a *revealed secret*. A *secret* S is hidden within a *cover* C using a hide function $\mathcal{H}$ to create a *container* C' such that the difference between the cover and the container is minimal, or $\mathcal{H}(C, S) = C' \mid \text{MSE}(C, C') \to 0$. The *revealed secret* S' can then be obtained from the container using a reveal function $\mathcal{R}$, which is usually the inverse of $\mathcal{H}$. The revealed secret should be minimally different from the original secret hidden in the cover, or $\mathcal{R}(C') = S' \mid \text{MSE}(S, S') \to 0$. In regard to sanitization, an image is sanitized with a sanitization function $\mathcal{P}$ to create a *sanitized image* $\hat{C}$, or $\mathcal{P}(X) = \hat{C} \mid X \in \{C, C'\}$. An attempted revealed secret from a sanitized image is denoted as $\hat{S}$, or $\mathcal{R}(\hat{C}) = \hat{S}$. In this work, $\mathcal{P} \in \{\text{SUDS}, \text{DM-SUDS}, \text{Noise}, \text{DCT-Noise}\}$, where DM-SUDS is our diffusion model approach, SUDS is an existing VAE-based approach, Noise is Gaussian Noise applied directly to the image, and DCT-Noise is Gaussian Noise applied to the Discret Cosine Transform (DCT) of the image.

Hiding Methods. There are many types of hiding techniques in steganography, including traditional and deep hiding. Traditional hiding involves either 1) modifying the pixels of the image (*spatial domain*) or 2) altering the image using a frequency distribution (*transform domain*) [4, 27, 30, 13, 9]. While information loss is inevitable, the user must strike a balance between maximizing the information hidden in the container and avoiding detection: hiding capacity vs. invisibility. With a higher hiding capacity, a secret is more likely to be detected in a container, but with a lower invisibility, less information is transferable. Traditional methods are marred by this tradeoff, in addition to ensuring robustness against container perturbations. In deep hiding, however, methods are able to incorporate a high capacity of the secret while maintaining invisibility, as these methods are capable of maintaining the pixel distributions of the cover image. Deep hiding techniques utilize deep neural networks as both the hide and reveal functions and fall into two main categories: dependent deep hiding (DDH) [32, 37, 29, 45, 2, 36, 19, 35, 10, 44, 43] and universal deep hiding (UDH) [40]. In DDH, the resulting container is cover dependent, and in UDH, the secret can be combined with any cover.

In this work, we utilize a method from each of these categories 1) traditional = least significant bit (LSB) method [14], 2) dependent deep = DDH, and 3) universal deep = UDH. The implementations for DDH and UDH are CNN-based implementations from the same code base[1], and the LSB implementation is from the SUDS codebase[2]. This hiding methods were chosen to resemble the work from [26], which is currently the only blind, deep learning method for image-into-image sanitization.

2.2 Sanitization

Traditional Sanitization. Sanitization, or active steganalysis, is the process of removing hidden information from potentially steganographic images while preserving the image quality of the cover. Traditional sanitization approaches attempt to remove hidden information by modifying the pixels of the image or the frequency distribution of the image, much like traditional steganographic approaches. Examples include flipping the bits of N LSB planes [23], adding Gaussian noise to the image or transformation of the image, and applying image filters [5]. While traditional sanitization techniques are effective on traditional hiding algorithms, they largely degrade the image quality of the restored image and are ineffective on deep hiding techniques.

Deep Learning Sanitization. Deep learning sanitization attempts to improve upon the limited capabilities of traditional sanitization, especially in regard to deep learning steganography. While most work is related to **text-into-image** steganography and watermarking, recent research has explored **image-into-image** approaches. This is significant as deep learning methods, which primarily focus on image-into-image steganography, can be adversarially trained to circumvent steganalysis techniques, including Xu-Net [34] and Ye-Net [39]. In [16], the authors present a GAN-based approach, where a U-Net generator is trained to attack watermarked images. This approach, however, is considered non-blind as it relies on known hiding techniques during the training process. Non-blind techniques are unable to defend against novel hiding strategies, as they are implemented using past, known processes. A more effective approach to sanitization is a blind methodology, where the training process does not incorporate known hiding techniques. In [12], the authors present a blind approach called PixelSteganalysis, which uses an 'analyzer' to predict pixel and edge distributions of a container, and an 'eraser' to sterilize the images by adjusting suspicious pixels indicated by the analyzer. Their approach is inefficient (e.g., a single (3,32,32) CIFAR-10 image takes an average of 3 min. to sanitize) and also requires known information about 'high-frequency areas' to train the edge distribution model. Another blind sanitization approach is implemented in [26], which utilizes a VAE-based approach called SUDS. The authors demonstrate blind sanitization capabilities against unseen traditional, dependent deep, and universal deep hiding methods on MNIST and CIFAR-10 datasets. While their approach is able to eliminate secret information from both datasets, the image preservation of the sanitized image starts to deteriorate with CIFAR-10.

Diffusion Model Applications. In this work, we utilize a diffusion model approach to sanitization. Recent work has applied diffusion models to remove adversarial noise in the spatial domain [21] and to attack a single *text-into-image* watermarking technique in [17].

[1] **DDH/UDH:** https://github.com/ChaoningZhang/Universal-Deep-Hiding
[2] **SUDS Code:** https://github.com/pkrobinette/suds-ecai-2023

2.3 Sanitization Metrics

We evaluate sanitization and reconstruction using four image quality metrics: mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and normalized correlation coefficient (NCC). MSE and PSNR quantify the absolute error between the reference and altered images. SSIM assesses perceptual quality compared to a reference, while NCC measures correlation between images, reflecting their spatial alignment. NCC ranges from -1 to 1 and is crucial in watermarking to verify watermark extraction [31, 25, 41], with 1 indicating identical images.

Unlike previous methods focusing solely on secret elimination, our approach adds image preservation to assess sanitization quality. This ensures sensitive information is removed without compromising image utility. Drawing from the watermarking domain, we use NCC to evaluate both the preservation and elimination aspects: for image preservation (IP), an NCC ≥ 0.95 indicates successful preservation using the cover image. For secret elimination (SE), an NCC ≤ 0.30 signifies successful sanitization, reflecting typical standards for data removal.

Definition 1 (Sanitization) *Combining IP and SE, an image is considered successfully sanitized in this work if: IP NCC* $\geq 0.95 \wedge$ *SE NCC* ≤ 0.30.

3 DM-SUDS Sanitization

Our goal is to improve the image preservation of sanitized images using a diffusion model approach to sanitization. A diffusion model is a generative model that consists of two main features: a forward diffusion process and a reverse diffusion process. In the forward diffusion process, an input image x_0 from a given data distribution $x_0 \sim q(x_0)$ is perturbed at timestep t with Gaussian noise with a variance $\beta_t \in (0, 1)$ to produce a sequence of latent images $x_0, x_1, ..., x_T$. This forward noising process is defined by equations 1 and 2 and is commonly reparameterized by Equation 3, which allows for direct sampling of noised latents at arbitrary steps. Here, $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{s=0}^{t} \alpha_s$, and $1 - \bar{\alpha}_t$ represents the variance of the noise for an arbitrary timestep.

$$q(x_1, \ldots, x_T|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

$$q(x_t|x_{t-1}) := \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right) \tag{2}$$

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}\right) \tag{3}$$

The model then learns to reverse this diffusion process by refining the noised sample until it resembles a sample from the target distribution. The posterior $q(x_{t-1}|x_t, x_0)$ can be calculated using Bayes theorem in terms of $\tilde{\beta}_t$ and $\tilde{\mu}_t(x_t, x_0)$, which are defined in the equations below:

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{4}$$

$$\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}\beta_t}}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t(1-\bar{\alpha}_{t-1})}}{1-\bar{\alpha}_t}x_t \tag{5}$$

$$q(x_{t-1}|x_t, x_0) = N\left(x_{t-1}; \tilde{\mu}(x_t, x_0), \tilde{\beta}_t\mathbf{I}\right) \tag{6}$$

To represent $\mu_\theta(x_t, t)$ for the reverse diffusion process, we use a U-Net model to predict the noise ϵ added to the input image. The

original image can then be predicted using Equation 7.

$$x_0 = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon \right) \tag{7}$$

While diffusion models are most commonly used for generative purposes, we instead make use of the denoising capabilities developed during training. To use the diffusion model as a sanitizer, we can apply t timesteps of noise to a potential container using Equation 3 with a cosine beta scheduler from [8], predict the noise using a neural network, and then refine the image using Equation 7, effectively preserving the image quality while maintaining sanitization performance. As this is a blind approach to sanitization, prior knowledge of steganography techniques is not required for the training process. To highlight this feature of the framework and to emphasize the accessibility of this approach, we make use of a state-of-the-art, publicly available pre-trained diffusion model from [20][3]. For the remainder of this paper, the diffusion model sanitization process is referred to as DM-SUDS.

4 Research Questions and Metrics

To evaluate the sanitization performance of DM-SUDS, we seek to answer the following four research questions.

1. **RQ1: Ability to sanitize hidden information** → *Is DM-SUDS able to successfully sanitize hidden information? How does this compare to previous sanitization methods?* To evaluate the performance of DM-SUDS, we compare DM-SUDS against three different baselines: 1) SUDS, 2) Gaussian Noise, and 3) Discrete Cosine Transform (DCT) Noise. We test each of these methods on LSB, DDH, and UDH containers created with the same 1000 image pairs from the CIFAR-10 dataset of size $\mathcal{R}^{3\times32\times32}$. CIFAR-10 is a widely used dataset for information hiding [38, 33, 22, 18], and it is also where the reconstruction capabilities of SUDS start to break down. The SUDS model used during this evaluation is from the repository provided by the authors. For the DM-SUDS model discussed in Section 3, we use a timestep variable $t = 500$ in the forward and reverse diffusion processes, which is half of the number of timesteps used during model training. We utilize the sanitization metrics discussed in Section 2.3 (MSE, PSNR, SSIM, NCC) to evaluate image preservation and secret elimination. To provide a baseline comparison to the effect of each steganographic technique on image quality, a **None** column is also calculated, which uses the metrics mentioned above to compare *covers* C with *containers* C' (**IP**) as well as *secrets* S to *revealed secrets* ***pre*** *sanitization* S' (**SE**). **RQ1** is positive if DM-SUDS successfully sanitizes according to Definition 1 while outperforming the baselines in image preservation metrics.
2. **RQ2: Effect of timestep on sanitization** → *Does the number of diffusion timesteps affect sanitization performance?* In the previous experiment, we utilize a timestep $t = 500$ in the diffusion process. This means that the noisy version of the input image is sampled from the cosine noise scheduler at $t = 500$. In the reverse diffusion process, this number is then encoded, and the neural network estimates the added noise to the image from this timestep. To determine if the number of timesteps affects sanitization performance, we evaluate DM-SUDS on LSB, DDH, and UDH containers using timestep values from $t = 25$ to $t = 1000$ in 25-step intervals, e.g., $t = [25, 50, 75, ..., 1000]$. We utilize the same container creation process and sanitization evaluation described in RQ1. **RQ2** is positive if 50% of the timesteps are successfully sanitized (Definition 1).
3. **RQ3: Ability to directly sanitize** → *Is added noise necessary for the sanitization process?* In RQ1 and RQ2, we sample a noisy image of the container at timestep t, and then denoise this image to a sanitized version of the input. Another potential way to sanitize via DM-SUDS is to treat the potentially steganographic image as the direct input to the denoiser, skipping the forward diffusion process. With this approach, however, we still have to provide an estimate t as input to the reverse diffusion neural network. We evaluate the direct sanitization capabilities of DM-SUDS using the evaluation process described in RQ2. **RQ3** is positive if the result of no noise added resembles that of RQ2 (with noise).
4. **RQ4: Scalability and robustness** → *Does DM-SUDS extend to JPEG-resistant steganography and the ImageNet dataset?* As the reconstruction ability of SUDS decreases with an increase in image complexity [26], we evaluate DM-SUDS on more complex images using the ImageNet Dataset with images of size $\mathcal{R}^{3\times128\times128}$. While DDH is the most robust of the experimental methods, it is mainly used for lossless image types, such as PNG or BMP. Recent work in dependent deep hiding has shown the promise of JPEG (lossy) resistant image steganography as well. As such, we evaluate DM-SUDS on PRIS [36], an invertible network image steganography approach that is robust under JPEG compression. This method was chosen as the authors demonstrate its improved performance against ISN [19], RIIS [35], HiNET [10], and Baluja's dependent deep learning approach [3], other competing robust image-into-image steganography techniques. As such, we test DM-SUDS on 500 DDH and PRIS containers created using ImageNet images and collect MSE, PSNR, SSIM, and NCC metrics for secret elimination and image preservation. **RQ4** is positive if DDH and PRIS containers are successfully sanitized according to Definition 1.

5 Experimental Results

5.1 Sanitization Performance of DM-SUDS

We evaluate DM-SUDS sanitization performance against three different baselines on 1000 containers made from the CIFAR-10 dataset. From the results shown in Table 1, DM-SUDS is able to successfully sanitize LSB, DDH, and UDH containers as indicated by the image preservation (IP) and secret elimination (SE) values for each metric. The low MSE, high PSNR, high SSIM, and high NCC for DM-SUDS IP demonstrate that the quality of the sanitized image is preserved, and the high MSE, low PSNR, low SSIM, and low NCC values for DM-SUDS SE validate that the secret has been successfully removed. While previous works only consider SE for successful sanitization, the *Sanitization* evaluation in the last column considers both IS and SE (IP NCC $\geq 0.95 \wedge$ SE NCC ≤ 0.30). From these results, DM-SUDS is the only method to successfully sanitize all three unseen hiding methods, outperforming each of the baselines in IP. **RQ1 is, therefore, positive.**

The results for SUDS highlight the importance of considering image preservation in addition to secret elimination for sanitization. According to SE, SUDS effectively removes the secret (low SE NCC), but it also significantly degrades the quality of the resulting image (low IP NCC). This adverse effect is further demonstrated by Figure 3. In these examples, a cover C is combined with a secret S to create a container C' using a) LSB, b) DDH, and c) UDH. This container is then sanitized by either SUDS (Figure 3a) or DM-SUDS (Figure 3b) to create a sanitized container $\hat{C}$. The sanitized secret $\hat{S}$ shows the

[3] Diffusion Model: https://github.com/openai/improved-diffusion

Table 1. Image Preservation (IP) and Secret Elimination (SE) performance metrics on LSB, DDH, and UDH containers made from 1000 CIFAR-10 image pairs. If IP NCC $\geq 0.95 \wedge$ SE NCC ≤ 0.30 (Definition 1), this results in a Successful sanitization (highlighted in green).

RQ	Hide	Method	Time (ms)	MSE (IP / SE)	PSNR (IP / SE)	SSIM (IP / SE)	NCC (IP / SE)	Sanitization (IP / SE)
RQ1	LSB	None	4.0	37.47 / 79.15	32.50 / 29.17	0.97 / 0.97	0.99 / 1.00	-
		Gaussian Noise	0.0	430.17 / 9610.91	21.80 / 8.40	0.74 / 0.01	0.92 / 0.02	Fail / Success
		DCT Noise	0.1	254.81 / 9419.64	24.19 / 8.49	0.82 / 0.01	0.95 / 0.02	Success / Success
		SUDS	0.6	359.81 / 9873.25	23.04 / 8.29	0.78 / 0.01	0.93 / -0.00	Fail / Success
		DM-SUDS*	101.4	163.18 / 9200.26	26.21 / 8.61	0.88 / 0.01	0.97 / 0.01	Success / Success
RQ1	DDH	None	15.8	114.22 / 93.43	28.59 / 29.17	0.95 / 0.95	0.98 / 0.98	-
		Gaussian Noise	0.0	516.46 / 3119.53	21.07 / 13.29	0.71 / 0.30	0.90 / 0.50	Fail / Fail
		DCT Noise	0.1	336.37 / 1363.44	23.07 / 17.33	0.80 / 0.63	0.93 / 0.83	Fail / Fail
		SUDS	0.5	391.79 / 5419.22	22.71 / 11.16	0.78 / 0.08	0.92 / 0.24	Fail / Success
		DM-SUDS*	101.8	231.27 / 5082.42	24.92 / 11.45	0.86 / 0.13	0.95 / 0.25	Success / Success
RQ1	UDH	None	15.3	23.36 / 155.69	34.69 / 26.81	0.97 / 0.92	0.99 / 0.97	-
		Gaussian Noise	0.0	415.46 / 4793.60	21.96 / 11.37	0.74 / 0.22	0.92 / 0.39	Fail / Fail
		DCT Noise	0.1	243.50 / 905.63	24.38 / 18.89	0.82 / 0.61	0.95 / 0.82	Success / Fail
		SUDS	0.6	343.71 / 14379.81	23.29 / 7.09	0.79 / 0.03	0.94 / 0.04	Fail / Success
		DM-SUDS*	101.1	148.40 / 14667.25	26.66 / 6.99	0.88 / 0.03	0.97 / 0.04	Success / Success
RQ4	DDH (ImageNet)	DM-SUDS*	2191.4	106.42 / 7607.07	28.36 / 9.63	0.84 / 0.07	0.98 / 0.04	Success / Success
RQ4	PRIS (ImageNet)	DM-SUDS*	1625.2	89.00 / 4855.80	28.91 / 11.78	0.75 / 0.24	0.96 / 0.30	Success / Success

attempted revealed secret after sanitization. While SUDS and DM-SUDS are both able to eliminate the secret as shown by the last two columns, the image quality of the sanitized container for SUDS is low. Both specifications (IP and SE) are, therefore, necessary to holistically evaluate sanitization.

5.2 *Effect of Timestep on Sanitization*

In this section, we evaluate the impact of timestep size on the sanitization process. The top two rows of Figure 4 show the sanitized imaged $\hat{C}$ and an attempted recovered secret $\hat{S}$ at various timesteps. In regard to the sanitized image $\hat{C}$, with too many timesteps, the image quality of the reconstructed image starts to deteriorate, as shown by $t = 1000$. The blurry nature of the resulting processed image indicates that the more noise that is added to a container, the more difficult it is to predict the amount of added noise, affecting the resulting refined image $\hat{C}$. With too few timesteps, however, the secret persists after sanitization. At $t = 25$, for instance, the boat of the original secret is still visible in $\hat{S}$. These results are further validated by Figure 5a, which shows IP NCC and SE NCC for DDH, UDH, and LSB sanitization. The points not in the red bar signify successful sanitization according to Definition 1. As the timestep increases, secret elimination performance improves and image preservation performance deteriorates. These results also offer an interesting insight into the robustness of each hiding method. The order of robustness as determined by the persistence of the secret with increased timesteps is DDH, UDH, and LSB. DDH, therefore, is a more robust hiding method than the other two. While the number of selected timesteps impacts performance near the edges, using a better diffusion model eliminates the need to consider the timestep for sanitization as image preservation is the limiting factor. As less than 50% of the timesteps result in successful sanitization in this experiment, **RQ2 is negative**.

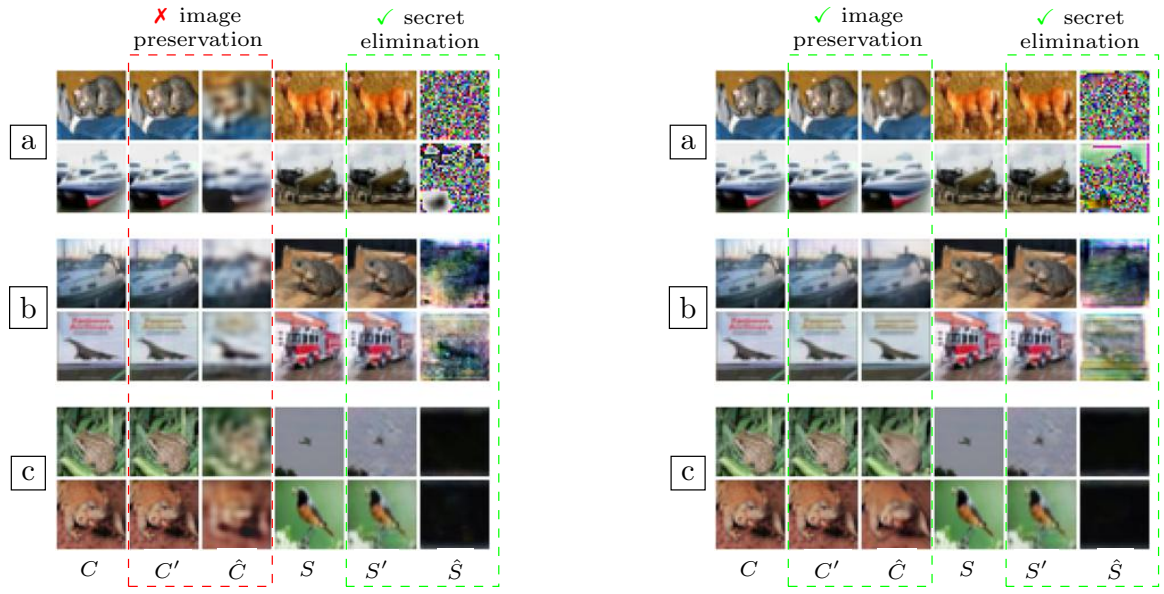
5.3 *Ability to Directly Denoise*

The plot in Figure 5b shows the IP and SE NCC metrics from sanitizing with no noise added for each hiding method across various timesteps. We see the best sanitization performance at $t = 500$. Here, the SE is lowest and the IP is above the indicated threshold. Whereas the SE from sanitizing **with** added noise continues to decrease with each timestep, the SE with no added noise starts to increase at $t = 500$, as shown by Figure 5. So much so, that the secret is not eliminated at higher timesteps. This is further demonstrated by Figure 4, which shows the image results of sanitization $\hat{C}$ and an attempted revealed secret after sanitization $\hat{S}$ for various timesteps. At $t = 1000$, the secret is still discernible from $\hat{S}$. These results indicate that the true power of the diffusion model sanitization approach lies in the added Gaussian Noise in the forward diffusion process. As these trends do not resemble those of RQ2, **RQ3 is negative.**

5.4 *Scalability to ImageNet and Robustness (JPEG)*

The results for RQ4 are detailed at the bottom of Table 1, under the labels *DDH (ImageNet)* and *PRIS (ImageNet)*. These findings affirm the efficacy of DM-SUDS even when applied to more complex datasets like ImageNet. Notably, the Image Preservation (IP) metrics—low MSE, high PSNR, high SSIM, and high NCC—highlight minimal degradation in image quality, confirming that the visual integrity of the sanitized images is maintained even with more complex images. Additionally, the robust secret elimination (SE) performance is demonstrated by high MSE, low PSNR, low SSIM, and low NCC metric values. These metrics together validate the effectiveness of DM-SUDS in sanitizing more complex datasets and secrets from more robust steganography methods. As IP NCC $\geq 0.95 \wedge$ SE NCC ≤ 0.30 is satisfied, both of these methods are successfully sanitized, and **RQ4 is positive.**

Evaluation From these experiments, RQ1 and RQ4 are positive and RQ2 and RQ3 are negative. DM-SUDS outperforms previous image-into-image deep learning methods, scales to more complex datasets, and is robust to JPEG-resistant hiding methods (PRIS). Additionally, the power of DM-SUDS lies in the added Gaussian Noise in the forward diffusion process.



(a) SUDS (b) DM-SUDS

Figure 3. A comparison between DM-SUDS and SUDS sanitization for a) LSB, b) DDH, and c) UDH steganography. Sanitization performance is determined from image preservation as well as secret elimination. While the secrets are eliminated with each method, DM-SUDS is the only method able to preserve the image.

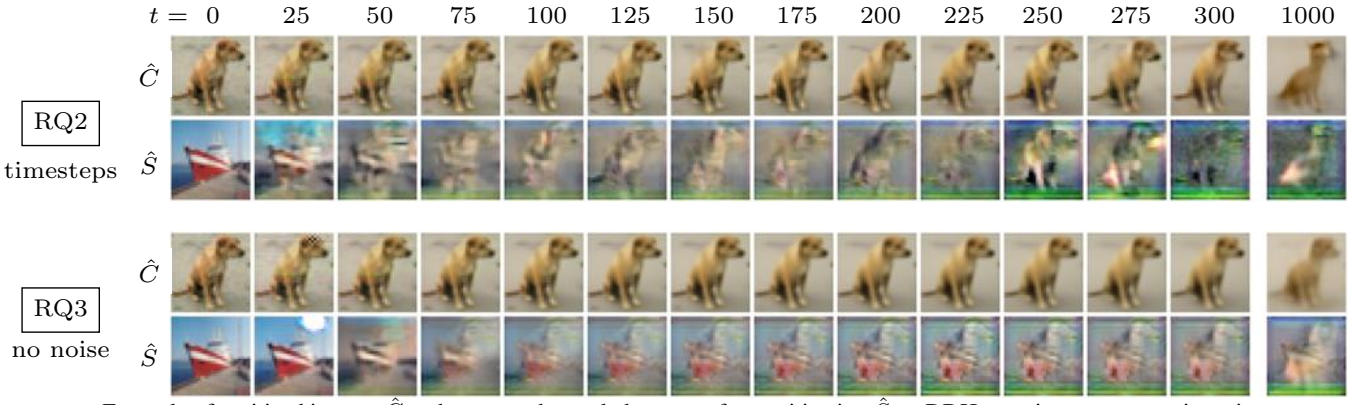
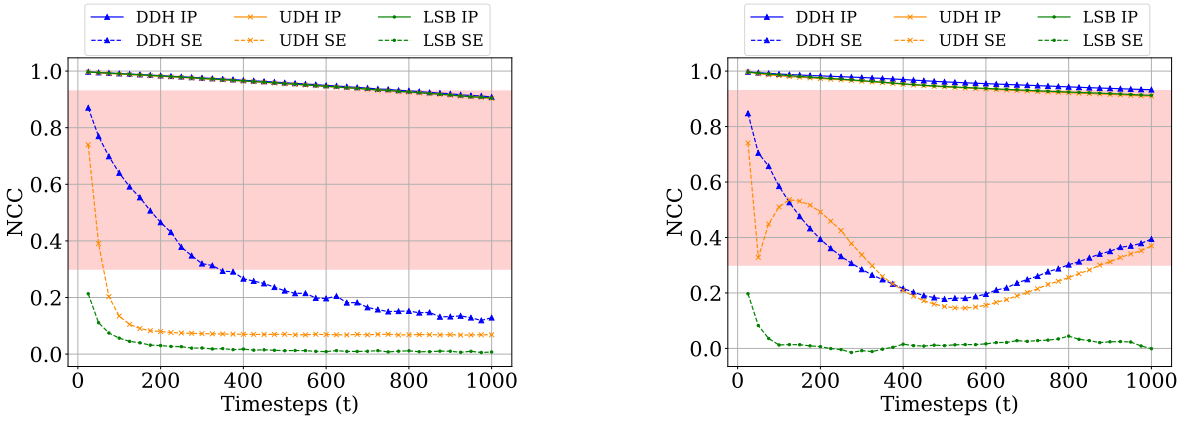
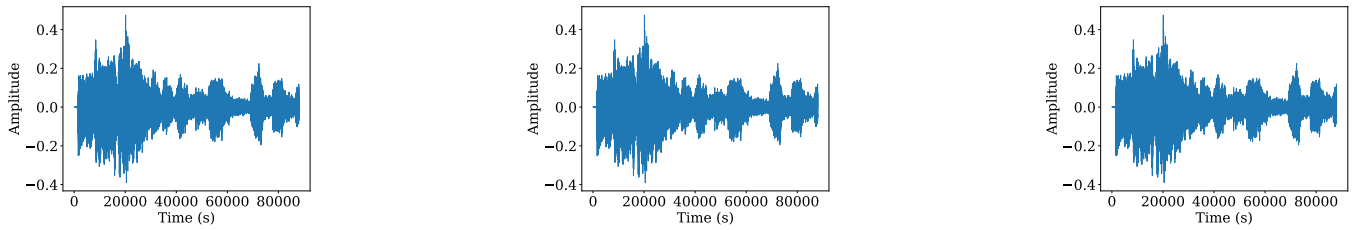


Figure 4. Example of sanitized images $\hat{C}$ and attempted revealed secrets after sanitization $\hat{S}$ on DDH containers across various timesteps.



(a) RQ2: Effect of Timesteps (b) RQ3: Effect of No Added Noise

Figure 5. Image preservation (IP) and secret elimination (SE) NCC metrics from sanitizing DDH, UDH, and LSB containers on various timesteps using added noise (a) and no added noise (b).



(a) Cover Audio (b) Container Audio (c) Sanitized Audio

Figure 6. A visual comparison between a cover audio, a container audio, and the resulting sanitized audio from DM-SUDS. The container is created by embedding the text secret S from using LSB. The container is then sanitized with DM-SUDS to create a sanitized audio.

6 Case Study: Application to Audio Steganography

To demonstrate the adaptability of this approach to other domains, we apply DM-SUDS to the sanitization of audio containers. Audio data refers to the representation of sound within a digital system. The process of hiding information in audio is very similar to hiding processes in the image domain. For this case study, we consider LSB and spread spectrum (SS) spatial methods as well as discrete wavelet transform (DWT) transform methods for hiding **text-into-audio**.

Overview We utilize 250 randomly sampled audio files from the UrbanSound8K dataset[4] to create containers with LSB, SS, and DWT. The UrbanSound8K dataset contains 8732 urban sound clips of less than 4 seconds from 10 different classes, including labels such as 'air conditioner' and 'street music'. The secret embedded in each audio file is a sentence scraped from Shakespeare's *Twelfth Night*. We use secrets of various content and sizes to stress test the sanitization process. Each of these resulting containers is sanitized using DM-SUDS. Similar to the experiments described in Section 4, we collect audio preservation and secret elimination metrics. Audio preservation is evaluated by the MSE difference between the container audio C' and the sanitized audio $\hat{C}$. As this case study utilizes text secrets, secret elimination is evaluated using bit error rate (BER) and removal rate (RR) metrics. BER is computed by comparing the bitstream extracted from the sanitized audio with the original embedded bitstream. The removal rate is calculated by Equation 8 [6]. If BER $> 0.2 \implies$ RR > 0.4, the hidden message is successfully destroyed. DM-SUDS for this case study is a publicly available diffusion model trained to generate music[5].

$$\text{RR} = 1 - ||2 \times \text{BER} - 1|| \tag{8}$$

Results From the audio preservation and secret elimination results shown in Table 2, DM-SUDS is also able to sanitize audio containers. The low MSE audio preservation values indicate that the quality of the sanitized audio is preserved during the sanitization process. Figure 6 demonstrates the audio preservation capabilities of DM-SUDS as the sanitized audio (Figure 6c) is minimally different from the container (Figure 6a). DM-SUDS is also successful at eliminating secrets. Prior to sanitization, the secret is recovered with a BER of 0.0%. This means that all secret bits are correctly recovered. After sanitization, however, the revealed secret BER is high, resulting in a high RR. As BER $> 0.2 \implies$ RR > 0.4, each of the hiding methods (LSB, SS, and DWT) is successfully sanitized with DM-SUDS.

7 Conclusion

In this work, we introduce a novel blind deep learning steganography sanitization method that utilizes a diffusion model framework to restore images called DM-SUDS. We demonstrate the success of such an approach compared to an existing VAE-based approach (SUDS), Gaussian Noise, and DCT Gaussian Noise and conduct an ablation study to determine the flexibility of the timestep parameter and the ability of DM-SUDS to directly denoise. We also demonstrate the scalability and robustness of DM-SUDS against JPEG-resistant steganography on the ImageNet dataset. Additionally, by incorporating image preservation into a novel sanitization evaluation, we effectively capture the safety and the utility of resulting sanitized images. One of the most impactful features of DM-SUDS lies in its accessibility, as any pretrained diffusion model of a target domain's data distribution can be implemented to protect against steganography. This work, therefore, introduces a highly effective and beneficial use case for a diffusion model while marking a significant advancement in the field of steganography sanitization. We also demonstrate the successful application of DM-SUDS to sanitizing containers in the audio domain. This is the first work to sanitize secrets hidden in audio containers using deep learning. In the future, we would like to address the speed of sanitization (see Table 1), and we would like to incorporate the sanitization or removal of visible watermarks in addition to the invisible information considered in this work.

Table 2. Audio sanitization results with DM-SUDS.

$\mathcal{H}$	$\mathcal{H}$ Eval	Audio Preservation	Secret Elimination	
	BER(S, S')	MSE$(C', \hat{C})$	BER$(S, \hat{S})$	RR
LSB	0.0	0.0013	0.4901	0.9273
SS	0.0	0.0013	0.4365	0.8647
DWT	0.0	0.0001	0.4851	0.9274

Acknowledgements

This paper was supported in part by a fellowship award under contract FA9550-21-F-0003 through the National Defense Science and Engineering Graduate (NDSEG) Fellowship Program, sponsored by the Air Force Research Laboratory (AFRL), the Office of Naval Research (ONR), and the Army Research Office (ARO). The material presented in this paper is based upon work supported by the National Science Foundation (NSF) through grant numbers 2220426 and 2220401, the Defense Advanced Research Projects Agency (DARPA) under contract number FA8750-23-C-0518, and the Air Force Office of Scientific Research (AFOSR) under contract number FA9550-22-1-0019 and FA9550-23-1-0135. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of AFOSR, DARPA, or NSF.

References

[1] M. Bachrach and F. Y. Shih. Image steganography and steganalysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(3):251–259, 2011.

[4] https://urbansounddataset.weebly.com/urbansound8k.html
[5] https://github.com/teticio/audio-diffusion

[2] S. Baluja. Hiding images in plain sight: Deep steganography. *Advances in neural information processing systems*, 30, 2017. https://papers.nips.cc/paper/2017/file/838e8afb1ca34354ac209f53d90c3a43-Paper.pdf.

[3] S. Baluja. Hiding images within images. *IEEE transactions on pattern analysis and machine intelligence*, 42(7):1685–1697, 2019. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8654686.

[4] A. Cheddad, J. Condell, K. Curran, and P. Mc Kevitt. Digital image steganography: Survey and analysis of current methods. *Signal processing*, 90(3):727–752, 2010. https://www.sciencedirect.com/science/article/abs/pii/S0165168409003648.

[5] S. Geetha, S. Subburam, S. Selvakumar, S. Kadry, and R. Damasevicius. Steganogram removal using multidirectional diffusion in fourier domain while preserving perceptual image quality. *Pattern Recognition Letters*, 147:197–205, 2021.

[6] L. Geng, W. Zhang, H. Chen, H. Fang, and N. Yu. Real-time attacks on robust watermarking tools in the wild by cnn. *Journal of Real-Time Image Processing*, 17:631–641, 2020.

[7] J. Hayes and G. Danezis. Generating steganographic images via adversarial training. *Advances in neural information processing systems*, 30, 2017.

[8] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

[9] V. Holub and J. Fridrich. Designing steganographic distortion using directional filters. In *2012 IEEE International workshop on information forensics and security (WIFS)*, pages 234–239. IEEE, 2012.

[10] J. Jing, X. Deng, M. Xu, J. Wang, and Z. Guan. Hinet: deep image hiding by invertible network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4733–4742, 2021.

[11] N. F. Johnson and S. Jajodia. Steganalysis of images created using current steganography software. In *International Workshop on Information Hiding*, pages 273–289. Springer, 1998.

[12] D. Jung, H. Bae, H.-S. Choi, and S. Yoon. Pixelsteganalysis: Pixel-wise hidden information removal with low visual degradation. *IEEE Transactions on Dependable and Secure Computing*, 2021.

[13] S. N. Kishor, G. K. Ramaiah, and S. Jilani. A review on steganography through multimedia. In *2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS)*, pages 1–6. IEEE, 2016. https://ieeexplore.ieee.org/document/7764373.

[14] C. Kurak and J. McHugh. A cautionary note on image downgrading. In *[1992] Proceedings Eighth Annual Computer Security Application Conference*, pages 153–159, 1992. doi: 10.1109/CSAC.1992.228224. https://ieeexplore.ieee.org/document/228224.

[15] F. Li, Y. Zeng, X. Zhang, and C. Qin. Ensemble stego selection for enhancing image steganography. *IEEE Signal Processing Letters*, 29:702–706, 2022.

[16] Q. Li, X. Wang, B. Ma, X. Wang, C. Wang, S. Gao, and Y. Shi. Concealed attack for robust watermarking based on generative model and perceptual loss. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5695–5706, 2021.

[17] X. Li. Diffwa: Diffusion models for watermark attack. In *2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS)*, pages 1–8. IEEE, 2023.

[18] Q. Liu, J. Yang, H. Jiang, J. Wu, T. Peng, T. Wang, and G. Wang. When deep learning meets steganography: Protecting inference privacy in the dark. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 590–599. IEEE, 2022.

[19] S.-P. Lu, R. Wang, T. Zhong, and P. L. Rosin. Large-capacity image steganography based on invertible neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10816–10825, 2021.

[20] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

[21] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

[22] X. Pan, S. Zhang, M. Zhang, Y. Yan, and M. Yang. House of cans: Covert transmission of internal datasets via capacity-aware neuron steganography. *Advances in Neural Information Processing Systems*, 35:24838–24850, 2022.

[23] G. Paul and I. Mukherjee. Image sterilization to prevent lsb-based steganographic transmission. *arXiv preprint arXiv:1012.5573*, 2010.

[24] Y. Qian, J. Dong, W. Wang, and T. Tan. Feature learning for steganalysis using convolutional neural networks. *Multimedia Tools and Applications*, 77:19633–19657, 2018.

[25] Y. Quan, H. Teng, Y. Chen, and H. Ji. Watermarking deep neural networks in image processing. *IEEE transactions on neural networks and learning systems*, 32(5):1852–1865, 2020.

[26] P. K. Robinette, H. D. Wang, N. Shehadeh, D. Moyer, and T. T. Johnson. Suds: Sanitizing universal and dependent steganography. *European Conference on Artificial Intelligence*, 372:1978–1985, 2023. doi: 10.3233/FAIA230489.

[27] M. S. Subhedar and V. H. Mankar. Current status and key issues in image steganography: A survey. *Computer science review*, 13:95–113, 2014. https://www.sciencedirect.com/science/article/abs/pii/S1574013714000136.

[28] W. Tang, B. Li, S. Tan, M. Barni, and J. Huang. Cnn-based adversarial embedding for image steganography. *IEEE Transactions on Information Forensics and Security*, 14(8):2074–2087, 2019.

[29] W. Tang, B. Li, M. Barni, J. Li, and J. Huang. An automatic cost learning framework for image steganography using deep reinforcement learning. *IEEE Transactions on Information Forensics and Security*, 16:952–967, 2020. https://ieeexplore.ieee.org/document/9205850.

[30] M. C. Trivedi, S. Sharma, and V. K. Yadav. Analysis of several image steganography techniques in spatial domain: A survey. In *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, pages 1–7, 2016. https://dl.acm.org/doi/10.1145/2905055.2905294.

[31] S. P. Vaidya. Fingerprint-based robust medical image watermarking in hybrid transform. *The Visual Computer*, 39(6):2245–2260, 2023.

[32] D. Volkhonskiy, I. Nazarov, and E. Burnaev. Steganographic generative adversarial networks. In *Twelfth international conference on machine vision (ICMV 2019)*, volume 11433, pages 991–1005. SPIE, 2020. https://doi.org/10.1117/12.2559429.

[33] Z. Wang, G. Wang, and Y. Yang. Nas-stegnet: Lightweight image steganography networks via neural architecture search. In *International Conference on Neural Information Processing*, pages 228–239. Springer, 2022.

[34] G. Xu, H.-Z. Wu, and Y.-Q. Shi. Structural design of convolutional neural networks for steganalysis. *IEEE Signal Processing Letters*, 23(5):708–712, 2016.

[35] Y. Xu, C. Mou, Y. Hu, J. Xie, and J. Zhang. Robust invertible image steganography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7875–7884, 2022.

[36] H. Yang, Y. Xu, X. Liu, and X. Ma. Pris: Practical robust invertible network for image steganography. *arXiv preprint arXiv:2309.13620*, 2023.

[37] J. Yang, D. Ruan, J. Huang, X. Kang, and Y.-Q. Shi. An embedding cost learning framework using gan. *IEEE Transactions on Information Forensics and Security*, 15:839–851, 2019. https://ieeexplore.ieee.org/abstract/document/8735922.

[38] Z. Yang, Z. Wang, and X. Zhang. A general steganographic framework for neural network models. *Information Sciences*, 643:119250, 2023.

[39] J. Ye, J. Ni, and Y. Yi. Deep learning hierarchical representations for image steganalysis. *IEEE Transactions on Information Forensics and Security*, 12(11):2545–2557, 2017.

[40] C. Zhang, P. Benz, A. Karjauv, G. Sun, and I. S. Kweon. Udh: Universal deep hiding for steganography, watermarking, and light field messaging. *Advances in Neural Information Processing Systems*, 33:10223–10234, 2020. https://proceedings.neurips.cc/paper/2020/file/73d02e4344f71a0b0d51a925246990e7-Paper.pdf.

[41] J. Zhang, D. Chen, J. Liao, H. Fang, W. Zhang, W. Zhou, H. Cui, and N. Yu. Model watermarking for image processing networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12805–12812, 2020.

[42] R. Zhang, F. Zhu, J. Liu, and G. Liu. Efficient feature learning and multi-size image steganalysis based on cnn. *arXiv preprint arXiv:1807.11428*, 2018.

[43] R. Zhang, S. Dong, and J. Liu. Invisible steganography via generative adversarial networks. *Multimedia tools and applications*, 78:8559–8575, 2019.

[44] L. Zhou, G. Feng, L. Shen, and X. Zhang. On security enhancement of steganography via generative adversarial image. *IEEE Signal Processing Letters*, 27:166–170, 2019.

[45] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 657–672, 2018. https://link.springer.com/chapter/10.1007/978-3-030-01267-0_40.