

RoSteALS: Robust Steganography using Autoencoder Latent Space

Tu Bui¹Shruti Agarwal²Ning Yu³John Collomosse^{1,2}¹University of Surrey, ²Adobe Research, ³Salesforce Research

t.v.bui@surrey.ac.uk, {shragarw, collomos}@adobe.com, ning.yu@salesforce.com

Abstract

Data hiding such as steganography and invisible watermarking has important applications in copyright protection, privacy-preserved communication and content provenance. Existing works often fall short in either preserving image quality, or robustness against perturbations or are too complex to train. We propose RoSteALS, a practical steganography technique leveraging frozen pretrained autoencoders to free the payload embedding from learning the distribution of cover images. RoSteALS has a lightweight secret encoder of just 300k parameters, is easy to train, has perfect secret recovery performance and comparable image quality on three benchmarks. Additionally, RoSteALS can be adapted for novel cover-less steganography applications in which the cover image can be sampled from noise or conditioned on text prompts via a denoising diffusion process. Our model and code are available at <https://github.com/TuBui/RoSteALS>.

1. Introduction

Deep fakes and misinformation are major societal challenges exacerbated by recent advances in generative models [31–33]. It has never been more important to understand the origins (or ‘provenance’) of digital images, to enable informed trust decisions to be made on content encountered online. Emerging standards such as the ‘Coalition on Content Provenance and Authenticity’ (C2PA) [7] embed signed provenance information within image metadata, yet this is easily stripped by attackers or as the image gets shared over the internet (*e.g.* by social media platforms). Perceptual hashing has been explored for near-duplicate search of trusted databases in order to recover stripped provenance information [2, 3, 26, 52]. However, such matches are by definition inexact, and require human-in-the-loop review when matching at scale. Image watermarking [1, 8, 44] provides a potential solution, enabling an identifier to be robustly inserted into an image, which may be used for exact search of a provenance database. Watermarking may also be used to robustly embed other indicators, such as whether an image has been produced by a generative model.

Watermarking techniques are typically steganographic –

fusing an embedded payload (or ‘secret’) within the image (‘cover’) pixels in such a way that the change in the watermarked (stego) image is *perceptually invisible* but at the same time *recoverable* via a learned secret decoder. This requirement is important for creative work where image quality should be preserved. The challenge of such processes is to learn both a representative image distribution and the secret embedding process at the same time.

To this end, we propose RoSteALS, a simple yet effective steganography technique that leverages ‘free knowledge’ from a locked (frozen) autoencoder. Our technical contributions are three-fold:

1. Latent steganographic embedding. We propose a novel method to inject the secret directly into the latent code of a locked autoencoder, enabling robust watermarking with limited training and no content specialization. Our secret encoder is small in size, easy to train, generalize well beyond training data, and can be adapted to the most advanced autoencoders to date.

2. Robust secret recovery. Our approach to secret embedding is shown to withstand severe image perturbations, critical to the use case of persisting identifiers robustly through online redistribution of content.

3. Cover-less steganography. We show that RoSteALS can be easily adapted for cover-less use (*i.e.* where a cover is synthesised on the fly for secret embedding) and for novel text-based steganographic applications. With RoSteALS, we extend the idea of cover-less steganography to an autoencoder’s latent space, where our aim is to generate a latent offset that can uniquely represent a secret. Given an autoencoder, the learned offset can be added to any random latent code (generated using an image or text using a diffusion model) of that autoencoder to produce the stego image.

2. Related work

The main goals of steganography techniques are three-fold: maximize the *length* of the secret that can be embedded, imperceptibility of the secret in the embedded image, and robustness of the secret decoding against benign and adversarial attacks. The secret can be embedded in the spatial or frequency space of the image using either hand-crafted and learning-based methods. Here we briefly describe the related techniques and place our work in context of others.

Hand-crafted methods Least significant bit (LSB) embedding was one of the first hand-crafted technique where the secret is embedded in the lowest order bits of each image pixel [45], producing images which are perceptually indistinguishable from the original cover image. Ever since there has been many techniques designed to embed the secret imperceptibly in the spatial [13, 38] and frequency [17, 18, 22, 25, 27, 28, 41] domain of the image. Even though most of these methods can embed large payloads in the image imperceptibly, they suffer with poor robustness to even minor modifications to the secret image.

Deep learning methods provide better robustness to noises while maintaining good quality of the generated image [43]. HiDDeN [57] was the first end-to-end trained watermarking network that used the encoder-decoder architecture along with a noise layer and adversarial discriminator for robustness and stego-image quality. Given a cover image and a secret of fixed length as inputs, encoder generates a stego image. The decoder takes the noisy stego image and outputs the embedded secret, while the adversarial discriminator compares the stego image and cover image for quality. Using the similar encoder-decoder approach, a variety of architectures were proposed for the improvement of stego image quality and robustness in several later works [6, 9, 19, 24, 37, 39, 46, 49]. These works mostly encode secrets and covers jointly, often with an UNet-like model and skip connections to preserve small details in the cover images [9, 39, 49]. Different from these techniques, we use a fixed pretrained auto-encoder and train a simple very small network to map the secret to the auto-encoder latent space independent of the cover image. Leveraging pretrained models for steganography has been explored recently – SSL [12] uses a ResNet50 model pretrained with DINO [5] self-supervised learning to watermark an image using back-propagation. Here, we demonstrate that the latent space of an autoencoder also provides excellent steganographic capability. Furthermore, SSL requires on-the-fly optimization and cannot operate in a blind setting as it needs a secret key to decode the secret, while our method requires just a single pass forward during inference.

Cover-less steganography aims to generate a cover image that can uniquely represent a secret, instead of changing a given cover image to embed the secret [29]. It was first introduced in [53] where the authors indexed a set of cover images to unique hash values that can be transformed to secrets. These hash values can be generated using image properties like pixel value [53], visual bag-of-words [54, 56], or histogram of oriented gradients [55]. In [10, 42], the authors learned to convert the secret message to GAN latent noise to generate appropriate container images. [23] achieves similar goals via a disentangled structure-texture autoencoder model trained on narrow domains. Other GAN-related works, [47, 48], inject the secret in form of a fingerprint to either training data [47] or model parameters [48] so that the trained models always generate images containing that fingerprint. In contrast to this, given an autoencoder, we learn

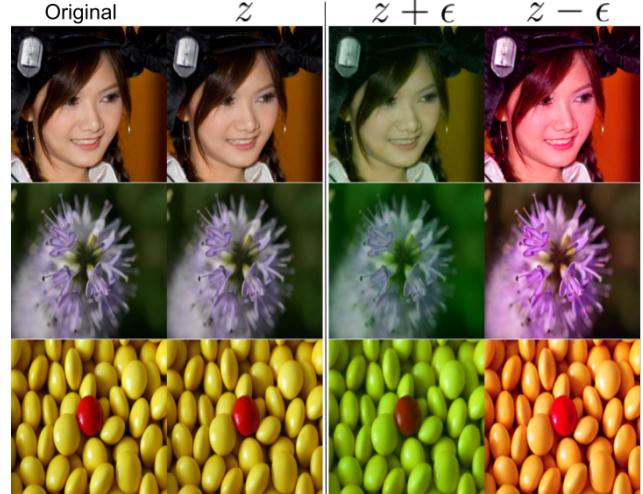


Figure 1. Correlation between changes in latent code and pixels of the reconstructed images. Here, the offset ϵ is selected as $\epsilon = k\mathcal{U}\sigma$, where \mathcal{U} is a random uniform function, σ is the std.dev. of the VQGAN latent space and noise strength $k = 0.2$.

a unique offset for every secret that can be independently applied to any latent code of that autoencoder.

3. Methodology

3.1. Leveraging pretrained AutoEncoders

To explore whether a pretrained autoencoder has capability for steganography, we experiment with the latent space of the state-of-art autoencoder, VQGAN [11]. Given an image \mathbf{x} of size $H \times W \times C$, the VQGAN encoder maps \mathbf{x} into a latent code $\mathbf{z} = E(\mathbf{x}) \in \mathbb{R}^{H' \times W' \times C}$ where H' and W' are typically $4 \times$ or $8 \times$ smaller than the original resolution. We work on the continuous space of \mathbf{z} therefore the quantization step is performed at the generator side $\bar{\mathbf{x}} = G(\mathbf{z})$ to reconstruct \mathbf{x} . We inject different kinds of noise to \mathbf{z} and observe changes on the reconstructed image. We made 2 observations: (i) certain noises cause the same perceptual change on the reconstructed images regardless of its content (Figure 1); and (ii) this embedding space of VQGAN is insensitive to small noise, $G(\mathbf{z} + \epsilon) \approx \mathbf{x}$ when $|\epsilon| < \epsilon_0$. This is probably due to the subsequent quantization process, however other models such as Kullback–Leibler regulated autoencoders do not have quantization but also exhibit similar behaviors. Regardless of the causes, (ii) inspires a possibility to inject meaningful messages (secrets) directly to the latent code without altering the image content in a noticeable way, and (i) poses an interesting question of whether the noise can be recovered given the output image.

3.2. Steganography with RoSteALS

Having prior knowledge of image distribution via the frozen autoencoder $\{E, G\}$, we aim to learn a secret encoder F to map secret $s \in \{0, 1\}^L$ to the image's latent

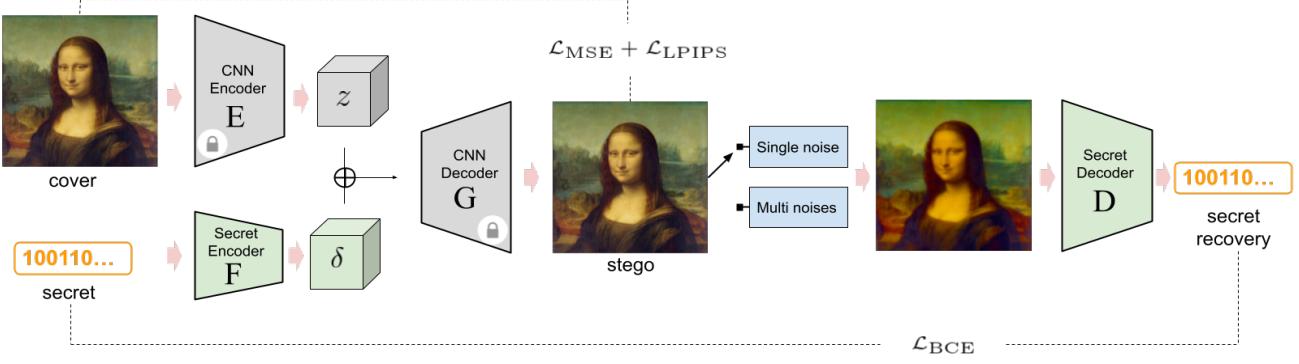


Figure 2. Architecture diagram of RoSteALS. The image encoder (E) and decoder (G) are locked during training, only updating the lightweight secret encoder (F) and decoder (D).

embedding, where L is the secret length;

$$\delta = F(\mathbf{s}) \in \mathbb{R}^{H' \times W' \times C} \quad (1)$$

F is a very small network consisting of a fully connected layer followed by SiLU [30], then reshaped and upscaled to match the dimension of \mathbf{z} , and followed by a final 1×1 convolution layer. The output δ acts as a small offset to be added to the cover embedding \mathbf{z} (Figure 2). We initialize the weight and bias of the 1×1 convolution layer to 0, following [50]. This is to ensure $\delta = 0$ in the first training iteration, so that the network has exactly the same initial behaviour as the original autoencoder. In total, our secret encoder F has just over 300K parameters for a 100-bit secret length. Interestingly, we empirically verify that conditioning F on both the cover image and secret (*e.g.* $\delta = F(\mathbf{x}, \mathbf{s})$) is not necessary (c.f. Figure 1, see also Sup.Mat.). The stego image is then constructed as $\tilde{\mathbf{x}} = G(\mathbf{z} + \delta)$ and regulated using a combination of pixel and perceptual losses;

$$\mathcal{L}_{MSE} = \|\gamma(\tilde{\mathbf{x}}) - \gamma(\mathbf{x})\|^2 \quad (2)$$

$$\mathcal{L}_{\text{quality}} = \mathcal{L}_{LPIPS}(\tilde{\mathbf{x}}, \mathbf{x}) + \alpha \mathcal{L}_{MSE} \quad (3)$$

where $\gamma(\cdot)$ is a differentiable non-parametric mapping function from RGB to the more perceptually uniform YUV space; \mathcal{L}_{LPIPS} refers to the LPIPS loss [51] commonly used as evaluator for image quality and α is a loss weight constant.

We use Resnet50 [14] as the secret decoder D , replacing the last fully connected layer to output a L -bit secret $\tilde{\mathbf{s}}$. We use BCE to compute the bit recovery loss between the predicted and groundtruth secret, $\mathcal{L}_{\text{recovery}} = \mathcal{L}_{BCE}(\mathbf{s}, \tilde{\mathbf{s}})$.

The total loss is computed as;

$$\mathcal{L} = \beta \mathcal{L}_{\text{quality}} + \mathcal{L}_{\text{recovery}} \quad (4)$$

where loss weight β controls the trade-off between stego quality and secret recovery.

Training for robustness. Being robust to noises is a desirable property of both steganography and watermarking. We insert a noise model between the image decoder

G and the secret decoder D . We use 14 noise sources from ImageNet-C [15], a rich and diverse library commonly used for evaluating model robustness. These noises can be split into 3 groups: *differentiable* including most additive and linear noises (*e.g.* brightness, saturation, contrast, ...), *approximatable with differentiable transforms* (*e.g.* jpeg compression [36]), *non-differentiable* (*e.g.* spatter). To ensure gradient can be flown back to the preceding layers, we convert any non-differentiable noise $n(\cdot)$ to additive noise via $n(\mathbf{x}) = \mathbf{x} + [n(\mathbf{x}) - \mathbf{x}]$ where $[.]$ is treated as an additive constant (gradient disconnected from computation graph). Additionally, we apply multiple common noises (*e.g.* contrast, brightness, jpeg compression) concurrently at a certain rate ($p = 0.5$), along with random individual noises. This enables the secret decoder to be trained under various data augmentations and feedback signal can be backpropagated to update the secret encoder.

4. Experiments

4.1. Datasets, training details and metrics

Datasets. We train RoSteALS using 100K images and validate on 1K images from the MIRFlickR dataset [20]. We evaluate on 3 different benchmarks - CLIC [40], MetFace [21] and Stock1K - our own collection of 1K images on Stock¹. CLIC [40] contains a mix of 530 high quality mobile and professional photographs often used to evaluate image compression and rate-distortion techniques. MetFace [21] is a narrow-domain collection of 1336 human faces extracted from artwork. Stock1K has 1000 high quality multimedia images, including photographs, vectorarts, sketches and graphic designs. We choose these 3 datasets for diversity in content, style and domain (Figure 14). To make the benchmarks more challenging, we apply a random ImageNet-C perturbation on every output stego.

Training details. Images are randomly cropped and resized to 256×256 during training. At test time, the whole images are resized to the specified resolution. We use the

¹<https://github.com/TuBui/RoSteALS>

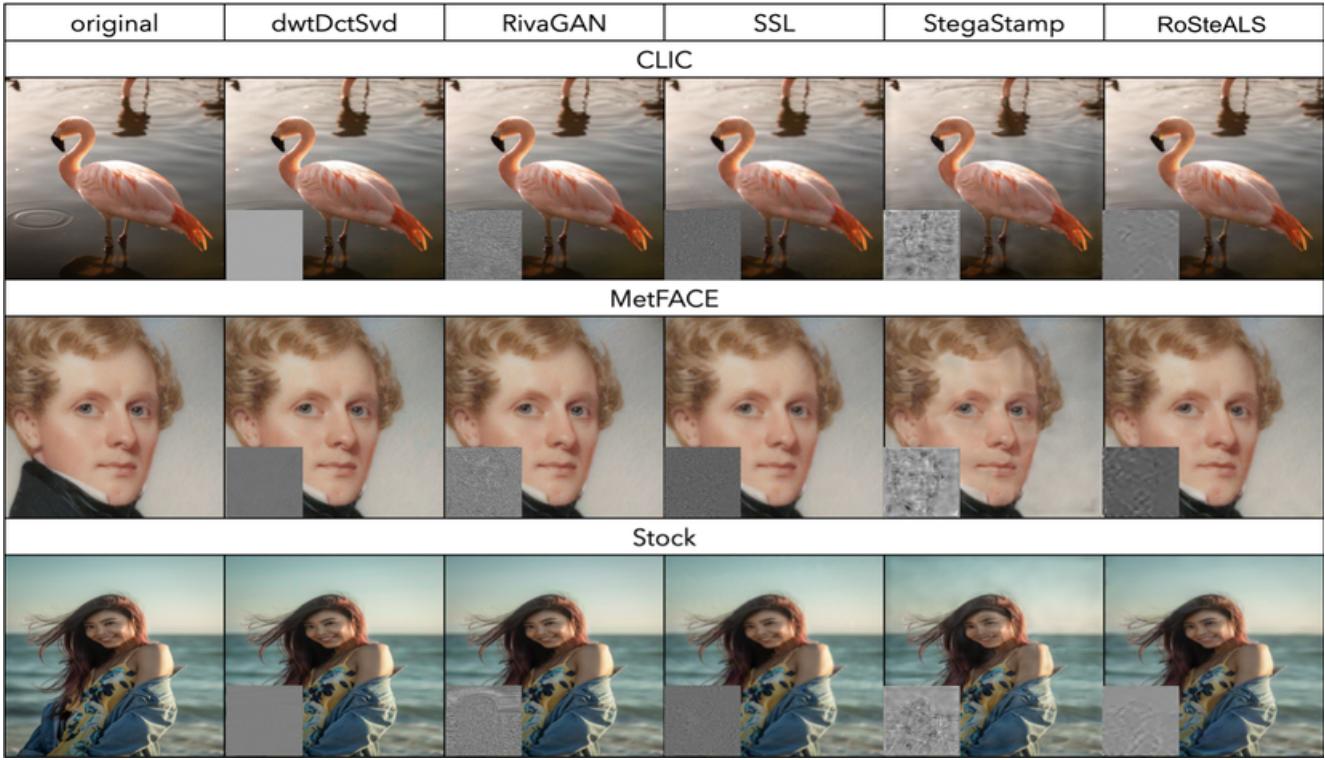


Figure 3. Qualitative examples of watermarked images from different techniques. The secret is the binary form of word string “some secrets”. In each case, except StegaStamp, the watermarked images look similar to the original images with no perceptual artifacts. Notice in the case of StegaStamp, there are visible artifacts in the entire image (second row) and on the woman’s forehead in the third row.

pretrained 256×256 VQ-f4 model of VQGAN [11] for most experiments, but also explore the 512×512 KL-f8 version (Section 4.5). Unless otherwise stated, the secret size is fixed at $L=100$ bits and uniformly sampled at random during training and testing. We use the AdamW optimizer and learning rate $8e-5$ using the Pytorch library. Training is terminated when the moving average of validation loss stops improving. We set $\alpha = 1.5$ in Equation (3) and the quality loss weight β in Equation (4) is determined dynamically. Specifically, we find that it is critical to prioritize $\mathcal{L}_{\text{recovery}}$ at the early training phase, since image quality is guaranteed in the first iteration through the locked autoencoder $\{\mathbf{E}, \mathbf{G}\}$ while the secret encoder/decoder $\{\mathbf{F}, \mathbf{D}\}$ must be learnt from scratch. We therefore start with a fixed image batch, set a low value of $\beta = 0.1$ and train our network to prioritize prediction of the random secret s . We unlock the full training image set after bit accuracy reaches a certain threshold t_1 , and linearly increase β till β_{\max} and turn on the noise model after a higher bit accuracy threshold t_2 is met. We choose $t_1 = 90\%$, $t_2 = 98\%$ and $\beta_{\max} = 10$ and do not tune it extensively. More details are in Sup.Mat.

Metrics. To evaluate stego quality versus the original cover image, we employ Peak Signal To Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), perceptual similarity score (LPIPS [51]) and Single Image Frechet Inception Distance (SIFID [35]). For secret recov-

ery evaluation, we report standard bit accuracy (*Bit acc.*) and *Bit acc.* (*ECC*) (bit accuracy with cyclic error correction code using BCH [4], following the settings in [42]). We also compute accuracy at word level, *Word acc.*, where a match is considered successful if the predicted secret differs from the corresponding groundtruth in less than 20% bits. Additionally, we report bit accuracy on clean stego for reference. Note that apart from *Bit acc.* (*clean*), three other metrics work on noised data.

4.2. Baseline comparison

We compare RoSteALS with the following baselines:

- **dwtDctSvd** [25]: a state-of-art handcrafted frequency-based method that performs secret embedding in the U frame of the YUV space after a sequence of discrete wavelet transform, discrete cosine transform and singular value decomposition. We do not report results for other popular handcrafted methods such as LSB, Jsteg [41] and OutGuess [28] because these are not designed to handle noises other than JPEG compression.
- **StegaStamp** [39]: a CNN-based method with state-of-art robustness performance.
- **RivaGAN** [49]: a GAN-based method leveraging attention mechanism and adversarial training.

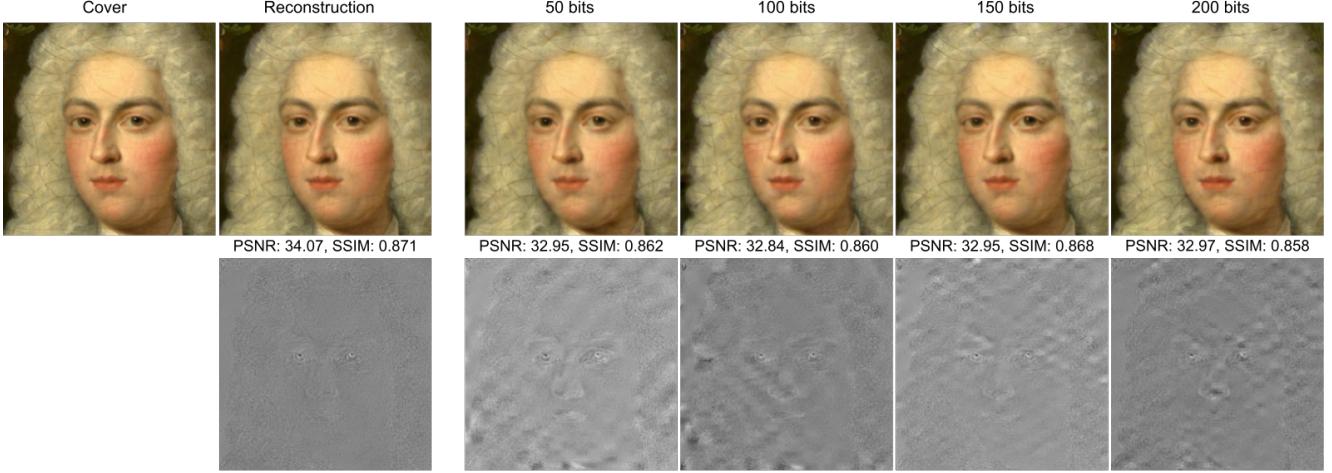


Figure 4. Change in image quality as secret length increases. Residual images are scaled to [0,255] range for visualization purpose.

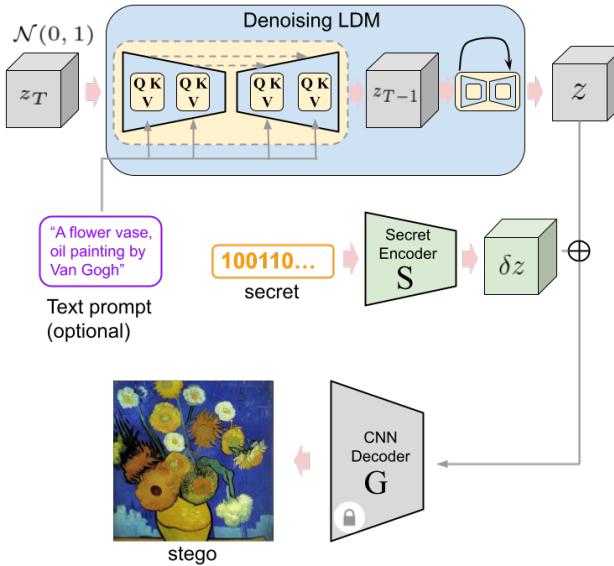


Figure 5. Text-based and cover less steganography.

- **SSL [12]**: an on-the-fly optimization method leveraging pretrained representation learning models.

For StegaStamp, we use the officially released model pre-trained on MIRFlickR for L=100 bits as we were not able to reproduce the reported results. Models for other methods are trained with the same settings as RoSteALS using their public code.

Table 1 shows image quality and secret recovery performance of all methods. We also include quality evaluation for RoSteALS’s autoencoder, VQGAN [11], for reference. There is not a clear winner in stego image quality metric. SSL [12] achieves the best PSNR and SSIM scores on all benchmarks while dwtDctSvd is most effective on LPIPS metric. On SIFID, SSL outperforms the rest on CLIC and Stock1K, but RoSteALS is the winner for MetFACE. This implies that the pretrained feature extraction model in SSL

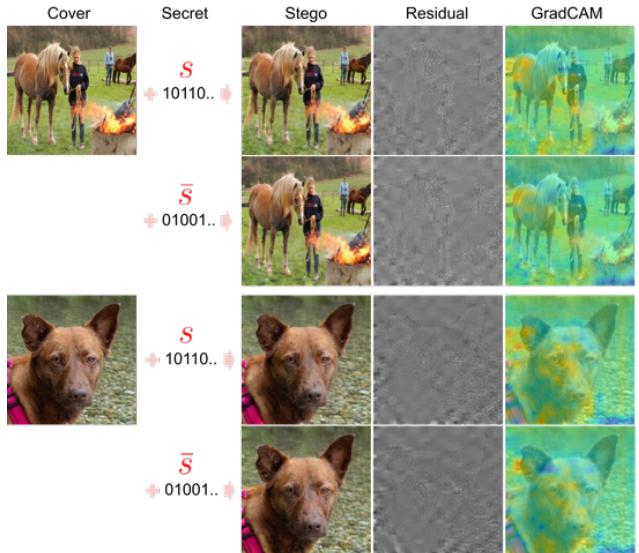


Figure 6. Studies into how the same secret and its inverted form are hidden in two different cover images. Residual images are scaled to range [0,255] for visualization purpose. Note the similarity in GradCAM heatmap between rows 1&3, and rows 2&4.

does not generalize as good as the autoencoder in RoSteALS. StegaStamp performs the worst in general, although still slightly outperforms RoSteALS on SSIM (see Section 5 for explanation). We note that RoSteALS performance is capped by its autoencoder backbone, trading just 1-2 points in each metric for secret embedding capacity and could be improved further with better backbone in future. Figure 14 shows example of stego images in each benchmark.

In term of secret recovery, all methods achieves near perfection score on clean data. RoSteALS outperforms the rest in all metrics and StegaStamp is the followup winner, while the handcrafted method and SSL have the lowest performance. The diversity in *Bit acc.* (*ECC*) shows a side effect of error correction - greatly improving performance for

Method	Image quality				Secret recovery			
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	SIFID \downarrow	Bit acc. (clean) \uparrow	Bit acc. \uparrow	Bit acc. (ECC) \uparrow	Word acc. \uparrow
CLIC								
RoSteALS	32.68 \pm 1.75	0.88 \pm 0.06	0.04 \pm 0.02	0.04 \pm 0.02	1.00	0.94 \pm 0.07	1.00	0.93
VQGAN	33.90 \pm 14.47	0.90 \pm 0.06	0.03 \pm 0.02	0.02 \pm 0.02	N/A	N/A	N/A	N/A
StegaStamp [39]	31.26 \pm 0.85	0.91 \pm 0.03	0.09 \pm 0.03	0.23 \pm 0.13	1.00	0.88 \pm 0.13	0.48 \pm 0.50	0.74
SSL [12]	41.84 \pm 0.10	0.98 \pm 0.01	0.02 \pm 0.01	0.01 \pm 0.02	0.99 \pm 0.03	0.62 \pm 0.14	0.03 \pm 0.17	0.13
RivaGAN [49]	40.32 \pm 0.15	0.98 \pm 0.01	0.02 \pm 0.02	0.07 \pm 0.06	0.98 \pm 0.03	0.77 \pm 0.16	0.22 \pm 0.41	0.45
dwtDctSvd [25]	38.96 \pm 1.41	0.97 \pm 0.02	0.01 \pm 0.01	0.02 \pm 0.02	1.00	0.61 \pm 0.20	0.16 \pm 0.34	0.21
MetFACE								
RoSteALS	34.46 \pm 1.91	0.89 \pm 0.07	0.04 \pm 0.02	0.01 \pm 0.02	1.00	0.94 \pm 0.08	1.00	0.91
VQGAN	35.98 \pm 2.45	0.90 \pm 0.07	0.02 \pm 0.02	0.01 \pm 0.02	N/A	N/A	N/A	N/A
StegaStamp [39]	32.01 \pm 0.77	0.92 \pm 0.02	0.13 \pm 0.03	0.22 \pm 0.15	1.00	0.86 \pm 0.14	0.47 \pm 0.50	0.68
SSL [12]	41.77 \pm 0.12	0.98 \pm 0.01	0.04 \pm 0.02	0.04 \pm 0.05	1.00	0.63 \pm 0.16	0.08 \pm 0.27	0.19
RivaGAN [49]	40.27 \pm 0.09	0.97 \pm 0.01	0.06 \pm 0.03	0.16 \pm 0.12	0.99 \pm 0.01	0.78 \pm 0.17	0.28 \pm 0.44	0.47
dwtDctSvd [25]	40.86 \pm 2.48	0.98 \pm 0.01	0.02 \pm 0.01	0.03 \pm 0.02	1.00	0.63 \pm 0.23	0.22 \pm 0.38	0.26
Stock1K								
RoSteALS	33.27 \pm 2.32	0.89 \pm 0.08	0.03 \pm 0.02	0.05 \pm 0.06	1.00	0.92 \pm 0.10	1.00	0.864
VQGAN	34.44 \pm 2.71	0.91 \pm 0.07	0.02 \pm 0.02	0.03 \pm 0.06	N/A	N/A	N/A	N/A
StegaStamp [39]	31.42 \pm 0.95	0.92 \pm 0.03	0.08 \pm 0.04	0.20 \pm 0.14	1.00	0.87 \pm 0.13	0.48 \pm 0.50	0.72
SSL [12]	42.07 \pm 0.50	0.99 \pm 0.01	0.02 \pm 0.02	0.01 \pm 0.02	0.95 \pm 0.09	0.59 \pm 0.12	0.02 \pm 0.13	0.09
RivaGAN [49]	40.49 \pm 0.45	0.98 \pm 0.01	0.02 \pm 0.02	0.05 \pm 0.06	0.93 \pm 0.09	0.72 \pm 0.16	0.13 \pm 0.33	0.31
dwtDctSvd [25]	39.76 \pm 2.41	0.98 \pm 0.02	0.01 \pm 0.01	0.02 \pm 0.02	0.95 \pm 0.13	0.60 \pm 0.19	0.17 \pm 0.33	0.18

Table 1. Comparison of RoSteALS and baseline methods for image quality and secret recovery performance. Values for each metric are reported as mean and standard deviation. VQGAN is the autoencoder backbone of RoSteALS. Best results are in bold.

Secret length (bits)	50	100	150	200
PSNR	32.81	32.69	32.85	32.89
SSIM	0.89	0.88	0.88	0.88
Bit acc.	0.97	0.94	0.87	0.84
Train time (epochs)	17	18	21	30

Table 2. Image quality, secret recovery performance and training speed versus secret length on the CLIC dataset. All stego images are subjected to random noises.

slightly corrupted data (as in case of RoSteALS) but also corrupting more if the data has damage beyond a certain threshold (other baselines). We attribute the performance of RoSteALS to its capability of separating the secret embedding and robustness learning tasks from image distribution learning task, resulting in perfect performance in both clean and noise data (with ECC).

4.3. Robustness

Figure 6 depicts the changes on stego image when two different secrets (inversion of each other) are embedded, as well as the effects of embedding the same secret on two different images. The residual the GradCAM heatmaps [34] indicates that the secret is embedded across the entire image. Additionally, it can be seen that the secret embedding is not dependent on image content (c.f. Figure 1). This properties coupled with the generalized VQGAN autoencoder benefit RoSteALS’s generalization on new domains unseen during training.

We analyze the secret recovery performance of RoSteALS and its closest competitor against individual noise sources in Figure 7. RoSteALS is more robust and sta-

Train volume ($\times 10^3$)	10	40	80	100
PSNR	35.07	34.51	34.36	34.46
SSIM	0.89	0.90	0.90	0.89
Bit acc.	0.91	0.91	0.93	0.94

Table 3. Image quality and secret recovery performance versus amount of training data on the MetFace dataset. All stego images are subjected to random noises.

ble than StegaStamp against all perturbations, especially on blurring (Gaussian, defocus) and heavy image enhancement effects (frost, fog). Highest performance is acquired on simple linear noises such as brightness, contrast and saturation as well as pixelate, for both methods. On the other hand, heavy jpeg compression and those that completely wipe values of certain pixels (shot, impulse and speckle noises) are the most challenging.

4.4. Dependencies and the quality/recovery tradeoff

We study the effect of secret length on RoSteALS performance and training speed. Table 2 demonstrates RoSteALS abilities to retain the same image quality as secret length increases, at the cost of decreasing secret recovery performance. Increasing secret length from 50 to 200 bits causes 13% loss in bit accuracy. Figure 4 shows an example of stego images produced by RoSteALS models trained on different secret length. Similarly, Table 3 shows that image quality is not affected by the training volume either, thanks to the frozen autoencoder. Secret recovery performance is affected due to overfitting at the secret decoder (bit accuracy increases by 3% at 10x increase in the training volume).

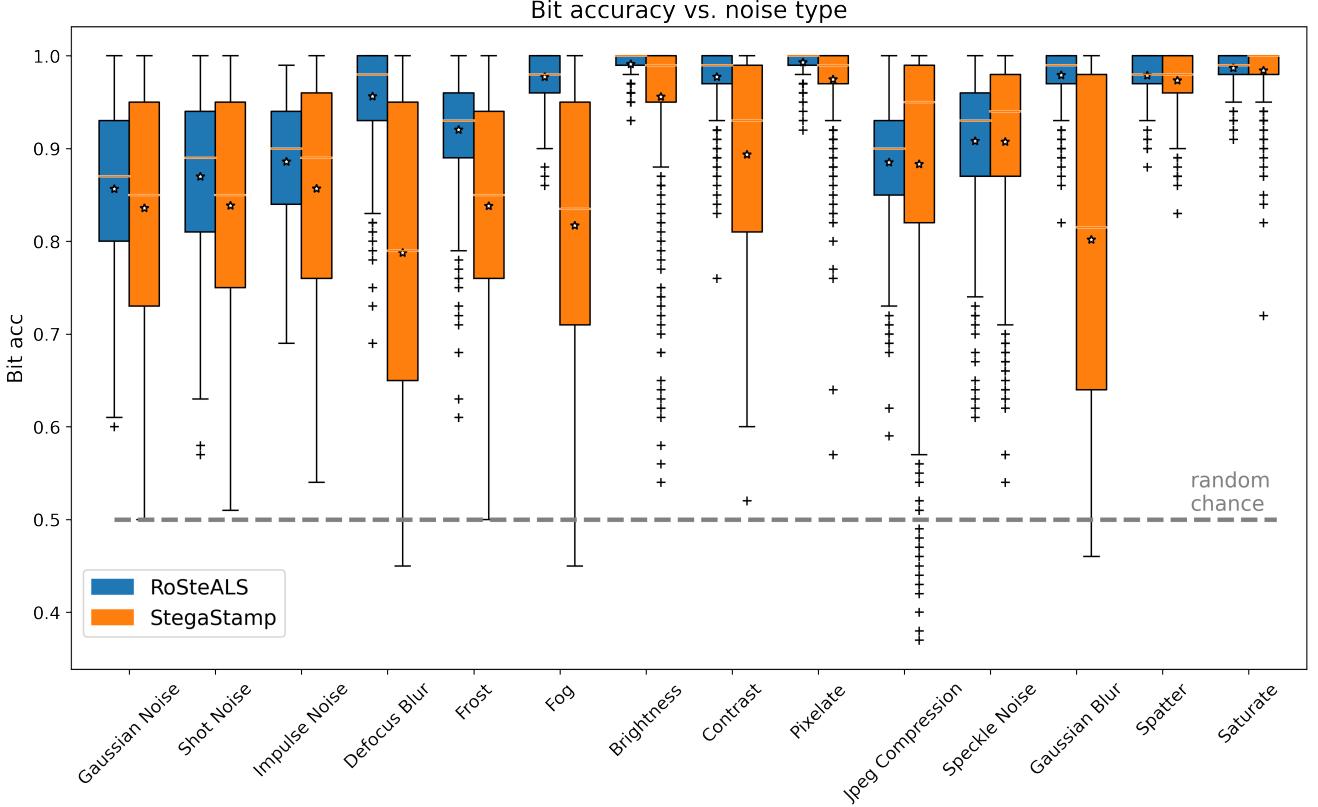


Figure 7. Effects of different perturbation sources to bit accuracy performance of RoSteALS and StegaStamp. Random chance is 50%.

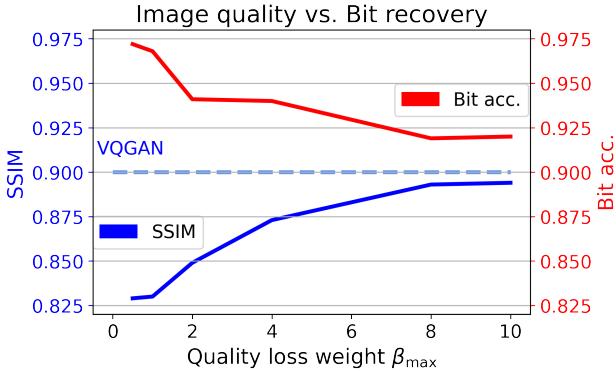


Figure 8. Stego quality and secret recovery trade-off, tested on the Stock1K benchmark. Performance of the VQGAN backbone serves as an upper limit for SSIM.

The training time also increases with longer secret length (Table 2). However, even at $L=200$ it only takes 30 epochs for the training to converge, as RoSteALS sole goal is to learn the secret encoding and decoding modules. For reference, training on SSL requires 100 passes through every images and RivaGAN requires 300 epochs to converge.

There is a trade-off between stego quality and secret recovery. Figure 8 shows that it is possible to control this trade-off in RoSteALS using the loss weight β_{\max} , which

Method	Bit acc. (clean)	Bit acc.	Word acc.
Cover-less	0.997	0.924	0.875
Text-based	0.992	0.904	0.844

Table 4. Secret recovery performance for unconditional (cover-less) and text-based conditional LDM-RoSteALS.

dictates the maximum value of β in Equation (4) (see Section 4.1). Increasing β_{\max} by 10 times boosts the SSIM score by 7points while reducing the Bit acc. score by 5%. It is worth noting that the trade-off can not go further beyond $\beta_{\max} = 10$ as the image quality is constrained by the autoencoder performance.

4.5. Text-based and Cover-less stega

RoSteALS’s design of injecting the secret signal directly to the already well-defined latent space enables novel applications in cover-less and text-based steganography. In theory, we can remove the image encoder E and inject the secret embedding to a random point in \mathbf{z} space to create a stego image without a cover. However, the distribution of latent code in VQGAN is rather complex – picking a random \mathbf{z} often lead to low quality output. Instead we can learn a mapping from a simpler distribution (*e.g.* Gaussian for cover-less stega) or from distribution of a different modal-

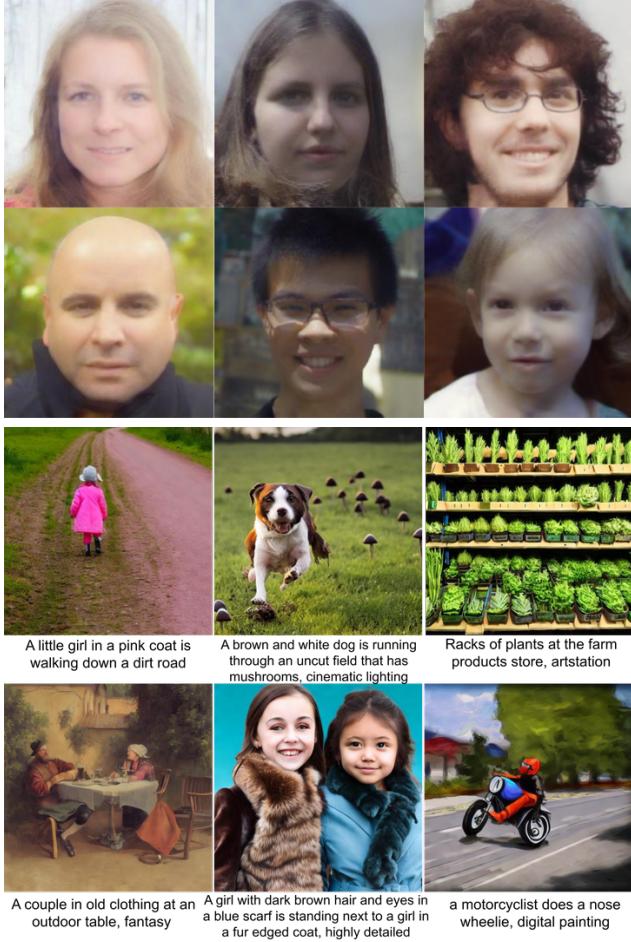


Figure 9. Random stego images are generated using our unconditional LDM-RoSteALS (top) and text conditioned LDM-RoSteALS (bottom), covering random secrets. All has bit acc. $> 99\%$. The LDM model in top row is trained on FFHQ and the LDM in bottom row is the pretrained Stable Diffusion v1.5 model.

ity (*e.g.* text). We propose to learn such mapping via the denoising process of latent diffusion (LDM) [32], as shown in Figure 5. Training the LDM is separate from RoSteALS, as long as they share the same autoencoder {E,G}. Figure 9 shows examples of cover-less and text-based stego generation. For the cover-less case, the LDM model is trained on FFHQ and we simply reuse our RoSteALS model above. For text-based stega, we use a pretrained Stable Diffusion model and train a new RoSteALS variant with the KL-f8 autoencoder backbone employed in Stable Diffusion. We use the same MIRFlickR training dataset, resize all input images to 512×512 for compatibility with the KL-f8 model and set the secret length $L=64$ bit for simplicity.

To evaluate secret recovery performance, we sample 1000 stego images for each LDM-RoSteALS models using randomly generated secrets. For text-based model, we also leverage 1000 random captions from the Flickr8K dataset [16] as the conditioning signals. To enhance the image generation quality of Stable Diffusion, we append a random

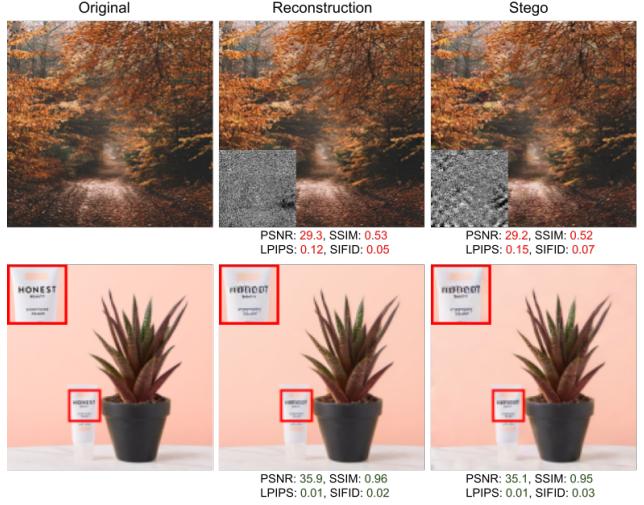


Figure 10. Failure cases are caused by VQGAN failing to reconstruct exact details of (top) clutter objects or (bottom) small text. In the top row, the change is hard to perceive but caught by quality metrics (insets are residual scaled to $[0,255]$ for visualization purpose). In the bottom row, artifacts are only local hence still yields high quality scores, but perceptually visible.

catching words (*e.g.* ‘highly detailed’, ‘cinematic lighting’, ‘artstation’, ‘sharp focus’) at the end of each caption. Table 4 shows that RoSteALS performs well in both cases, despite never ‘seeing’ a latent code generated by the LDM model during training.

5. Conclusion

We demonstrate that the latent space of a pretrained autoencoder can be leveraged to hide data, opening up a new direction for steganography and digital watermarking applications. We propose a novel secret hiding technique that offers several advantages: small model size, modular design, producing state-of-art secret recovery capability and comparable image quality versus the autoencoder. Our method, RoSteALS, can be easily adapted for novel applications such as cover-less and text-based steganography.

Limitations - RoSteALS relies on a pretrained autoencoder for image encoding and generation, thus inherits the same drawbacks of this model in preserving small details during image generation. We identify two failure cases of RoSteALS in Figure 10 involving cluttered objects that cause lots of tiny spatial shifts during image reconstruction (perceptually invisible but affecting quality measurement metrics, especially SSIM), or the challenge of reconstructing small text or face (not affecting quality metrics but could be perceptually visible). Future work could leverage more powerful autoencoders, or novel ways to carefully finetune the autoencoder that benefit steganography.

Acknowledgment This work was supported in part by DECaDE under EPSRC grant EP/T022485/1.

References

- [1] S. Baba, L. Krekor, T. Arif, and Z. Shaaban. Watermarking scheme for copyright protection of digital images. *IJCSNS*, 9(4), 2019. [1](#)
- [2] A. Bharati, D. Moreira, P.J. Flynn, A. de Rezende Rocha, K.W. Bowyer, and W.J. Scheirer. Transformation-aware embeddings for image provenance. *IEEE Trans. Info. Forensics and Sec.*, 16:2493–2507, 2021. [1](#)
- [3] Alexander Black, Tu Bui, Simon Jenni, Viswanathan Swaminathan, and John Collomosse. Vpn: Video provenance network for robust content attribution. In *Proc. CVMP*, pages 1–10, 2021. [1](#)
- [4] Raj Chandra Bose and Dwijendra K Ray-Chaudhuri. On a class of error correcting binary group codes. *Information and control*, 3(1):68–79, 1960. [4](#)
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. [2](#)
- [6] Ching-Chun Chang. Neural reversible steganography with long short-term memory. *Security and Communication Networks*, 2021:1–14, 2021. [2](#)
- [7] Coalition for Content Provenance and Authenticity. Draft technical specification 0.7. Technical report, C2PA, 2021. [1](#)
- [8] P. Devi, M. Venkatesan, and K. Duraiswamy. A fragile watermarking scheme for image authentication with tamper localization using integer wavelet transform. *J. Computer Science*, 5(11):831–837, 2019. [1](#)
- [9] Xintao Duan, Kai Jia, Baoxia Li, Daidou Guo, En Zhang, and Chuan Qin. Reversible image steganography scheme based on a u-net structure. *IEEE Access*, 7:9314–9323, 2019. [2](#)
- [10] Xintao Duan and Haoxian Song. Coverless information hiding based on generative model. *arXiv preprint arXiv:1802.03528*, 2018. [2](#)
- [11] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2020. [2, 4, 5](#)
- [12] Pierre Fernandez, Alexandre Sablayrolles, Teddy Furon, Hervé Jégou, and Matthijs Douze. Watermarking images in self-supervised latent spaces. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3054–3058. IEEE, 2022. [2, 5, 6](#)
- [13] Kazem Ghazanfari, Shahrokh Ghaemmaghami, and Saeed R Khosravi. Lsb++: An improvement to lsb+ steganography. In *TENCON 2011-2011 IEEE Region 10 Conference*, pages 364–368. IEEE, 2011. [2](#)
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. [3](#)
- [15] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. [3](#)
- [16] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. [8](#)
- [17] Vojtěch Holub and Jessica Fridrich. Designing steganographic distortion using directional filters. In *2012 IEEE International workshop on information forensics and security (WIFS)*, pages 234–239. IEEE, 2012. [2](#)
- [18] Vojtěch Holub, Jessica Fridrich, and Tomáš Denemark. Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security*, 2014:1–13, 2014. [2](#)
- [19] Chen-Hsiu Huang and Ja-Ling Wu. Image data hiding with multi-scale autoencoder network. *Electronic Imaging*, 34:1–6, 2022. [2](#)
- [20] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, 2008. [3](#)
- [21] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. [3](#)
- [22] Xiaoxia Li and Jianjun Wang. A steganographic method based upon jpeg and particle swarm optimization algorithm. *Information Sciences*, 177(15):3099–3109, 2007. [2](#)
- [23] Xiayu Liu, Ziping Ma, Junxing Ma, Jian Zhang, Gerald Schaefer, and Hui Fang. Image disentanglement autoencoder for steganography without embedding. In *Proc. CVPR*, pages 2303–2312, 2022. [2](#)
- [24] Ruohan Meng, Steven G Rice, Jin Wang, and Xingming Sun. A fusion steganographic algorithm based on faster r-cnn. *Computers, Materials & Continua*, 55(1):1–16, 2018. [2](#)
- [25] KA Navas, Mathews Cherian Ajay, M Lekshmi, Tammy S Archana, and M Sasikumar. Dwt-dct-svd based watermarking. In *2008 3rd International Conference on Communication Systems Software and Middleware and Workshops (COMSWARE'08)*, pages 271–274. IEEE, 2008. [2, 4, 6](#)
- [26] E. Nguyen, T. Bui, V. Swaminathan, and J. Collomosse. Oscar-net: Object-centric scene graph attention for image attribution. In *Proc. ICCV*, 2021. [1](#)
- [27] Tomáš Pevný, Tomáš Filler, and Patrick Bas. Using high-dimensional image models to perform highly undetectable steganography. In *Information Hiding: 12th International Conference, IH 2010, Calgary, AB, Canada, June 28-30, 2010, Revised Selected Papers 12*, pages 161–177. Springer, 2010. [2](#)
- [28] Niels Provos. Defending against statistical steganalysis. In *Usenix security symposium*, volume 10, pages 323–336, 2001. [2, 4](#)
- [29] Jiaohua Qin, Yuanjing Luo, Xuyu Xiang, Yun Tan, and Hua-jun Huang. Coverless image steganography: a survey. *IEEE Access*, 7:171372–171394, 2019. [2](#)
- [30] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Swish: a self-gated activation function. *arXiv preprint arXiv:1710.05941*, 7(1):5, 2017. [3](#)
- [31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. [1](#)
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of*

- the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 8
- [33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1
- [34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6
- [35] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019. 4
- [36] Richard Shin and Dawn Song. Jpeg-resistant adversarial images. In *NIPS 2017 Workshop on Machine Learning and Computer Security*, volume 1, page 8, 2017. 3
- [37] Nandhini Subramanian, Ismahane Cheheb, Omar Elharrouss, Somaya Al-Maadeed, and Ahmed Bouridane. End-to-end image steganography using deep convolutional autoencoders. *IEEE Access*, 9:135585–135593, 2021. 2
- [38] Mustafa Sabah Taha, Mohd Shafry Mohd Rahem, Mohammed Mahdi Hashim, and Hiyam N Khalid. High payload image steganography scheme with minimum distortion based on distinction grade value method. *Multimedia Tools and Applications*, 81(18):25913–25946, 2022. 2
- [39] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2117–2126, 2020. 2, 4, 6
- [40] G Toderici, W Shi, R Timofte, L Theis, J Ballé, E Agustsson, Nick Johnston, and F Mentzer. Workshop and challenge on learned image compression (clic2020). In *CVPR*, 2020. 3
- [41] Derek Upham. Jsteg. https://link.springer.com/chapter/10.1007/3-540-45496-9_21, 1997. 2, 4
- [42] Denis Volkonskiy, Ivan Nazarov, and Evgeny Burnaev. Steganographic generative adversarial networks. In *Twelfth international conference on machine vision (ICMV 2019)*, volume 11433, pages 991–1005. SPIE, 2020. 2, 4
- [43] Wenbo Wan, Jun Wang, Yuming Zhang, Jing Li, Hui Yu, and Jiande Sun. A comprehensive survey on robust image watermarking. *Neurocomputing*, 2022. 2
- [44] Xinyu Weng, Yongzhi Li, Lu Chi, and Yadong Mu. High-capacity convolutional video steganography with temporal residual modeling. In *Proc. ICMR*, pages 87–95, 2019. 1
- [45] Raymond B Wolfgang and Edward J Delp. A watermark for digital images. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 3, pages 219–222. IEEE, 1996. 2
- [46] Pin Wu, Yang Yang, and Xiaoqiang Li. Stegnet: Mega image steganography capacity with deep convolutional network. *Future Internet*, 10(6):54, 2018. 2
- [47] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 14448–14457, 2021. 2
- [48] Ning Yu, Vladislav Skripniuk, Dingfan Chen, Larry Davis, and Mario Fritz. Responsible disclosure of generative models using scalable fingerprinting. *International Conference on Learning Representations*, 2022. 2
- [49] Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285*, 2019. 2, 4, 6
- [50] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3
- [51] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 3, 4
- [52] X. Zhang, Z. H. Sun, S. Karaman, and S.F. Chang. Discovering image manipulation history by pairwise relation and forensics tools. *IEEE J. Selected Topics in Signal Processing.*, 14(5):1012–1023, 2020. 1
- [53] Zhili Zhou, Huiyu Sun, Rohan Harit, Xianyi Chen, and Xingming Sun. Coverless image steganography without embedding. In *Cloud Computing and Security: First International Conference, ICCCS 2015, Nanjing, China, August 13–15, 2015. Revised Selected Papers 1*, pages 123–132. Springer, 2015. 2
- [54] Zhili Zhou, QM Jonathan Wu, and Xingming Sun. Encoding multiple contextual clues for partial-duplicate image retrieval. *Pattern Recognition Letters*, 109:18–26, 2018. 2
- [55] Zhili Zhou, Q. M. Jonathan Wu, Ching-Nung Yang, Xingming Sun, and Zhaoqing Pan. Coverless image steganography using histograms of oriented gradients-based hashing algorithm. *J. Internet Technol.*, 18(5):1177–1184, Sep 2017. 2
- [56] Zhi Li Zhou, Yi Cao, and Xing Ming Sun. Coverless information hiding based on bag-of-words model of image. *J. Appl. Sci*, 34(5):527–536, 2016. 2
- [57] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 657–672, 2018. 2

A. Training details

RoSteALS is easy to train as long as it prioritizes the secret recovery loss at the early training phase. In Section 4.1 we propose a training method to overcome the complexity of the cover image domain (*e.g.* MIRFlickR is harder to train than FFHQ), the gradient flow between pretrained and learning-from-scratch modules, the challenges of large secret size, and the difficulties for the secret decoder to ‘learn’ perceptually invisible secret signals present in already high-quality images but corrupted with various perturbations. We adopt curriculum learning in our training schedule, starting from a fixed minibatch of cover images without noise corruptions, before unleashing the full training database and eventually enabling perturbations and linear loss weight ramping.

A successful training pipeline should be similar to Figure 11. We only experiment with few (t_1, t_2, β_{\max}) tuples and settle with $(t_1 = 0.90, t_2 = 0.98, \beta_{\max} = 10.0)$, therefore believe that performance could potentially be improved further with more careful parameter tuning.

B. Architecture details

RoSteALS has a very light-weight secret encoder and can be constructed using just 1 line of code using the Pytorch library. For example, for a 100-bit secret encoder:

```
secret_encoder = nn.Sequential(
    nn.Linear(100, 32*32*3), nn.SiLU(),
    Lambda(lambda x: x.view(-1, 3, 32, 32)),
    nn.Upsample((2, 2)),
    nn.Conv2d(3, 3, 3, padding=1)
)
```

We experimented with more advanced architectures and found no clear benefits over this simple module.

C. Joint cover-secret conditioning

Existing works often model secrets and covers jointly, arguing the secret embedding should depend on the cover image for optimal stego quality. We observe that is not the case for RoSteALS, as shown in Figures 1 and 6 and discussed in Section 4.3. Here, we implemented a RoSteALS alternative with the secret encoder E taking both the secret and cover as inputs. Specifically, the cover image is first blurred and downsampled to $H' \times W' \times C$, retaining only low frequency components. We then concatenate it with the upsampled secret embedding and passing to a sequence of convolution layers with SiLU activation. The weights of the last convolution layer is initialized with 0, in the same way as the proposed RoSteALS.

Table 5 shows the performance of this joint conditioning configuration, which is equal or slightly worse than the proposed approach in all metrics.

	PSNR	LPIPS	Bit acc.	Word acc.
CLIC				
Proposed	32.68 ± 1.75	0.04 ± 0.02	0.94 ± 0.07	0.93
Joint cond.	32.45±1.67	0.05±0.02	0.94±0.09	0.92
MetFace				
Proposed	34.46 ± 1.91	0.04 ± 0.02	0.94 ± 0.08	0.91
Joint cond.	33.99±1.81	0.04±0.02	0.93±0.09	0.90
Stock1K				
Proposed	33.27 ± 2.32	0.03 ± 0.02	0.92 ± 0.10	0.86
Joint cond.	33.00±2.18	0.04±0.02	0.92±0.11	0.86

Table 5. Joint cover-secret conditioning provides no benefit in RoSteALS design.

D. Perturbations

Figure 12 shows examples of 14 ImageNet-C perturbations used in our work. Note that there are 19 perturbations in ImageNet-C in total, we exclude 5 of them which are too slow to be included in training. Each perturbation has 5 levels of severity and its performance breakdown per level is shown in Figure 13. RoSteALS is most sensitive to degradation due to Gaussian, shot, impulse and speckle noises as well as jpeg compression; while being most resilient to brightness, pixelate and saturation effects.

E. More qualitative results

Figure 14 shows more qualitative examples of stego images created by RoSteALS and other baselines. The artifacts on StegaStamp generated images are perceptually visible, as if the image is covered with a transparent layer of fog. RoSteALS performance is comparable with other methods.

Figure 15 depicts several examples of our novel text-based steganography application. We note the glimpse of semantic objects visible in the residual image, however these artifacts are inevitable around the strongest edges during image generation and do not represent the secret artifacts to be picked up by the secret decoder (c.f. Figure 6 in the main paper).

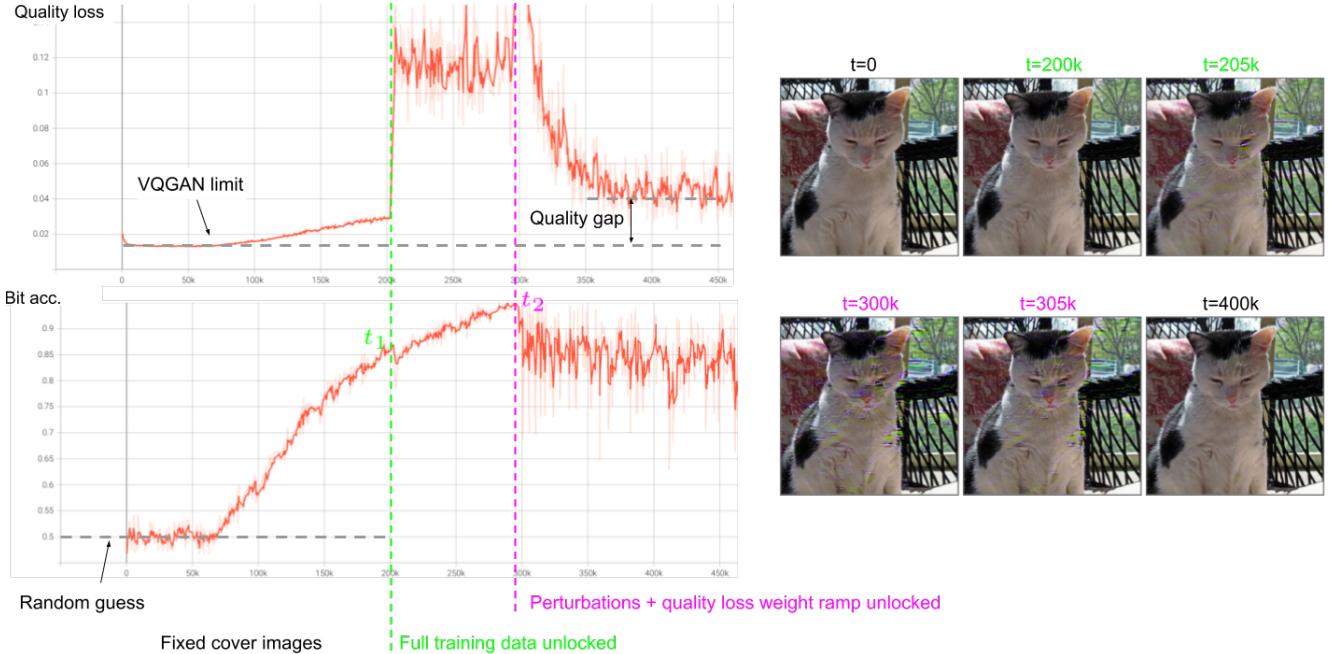


Figure 11. (Left) Quality loss and secret recovery curves for the first 18 epochs when training the 200bit-secret RoSteALS model on MIRFlickr with mini-batch size set to 4. (Right) Evolution of the stego image at different training stages.

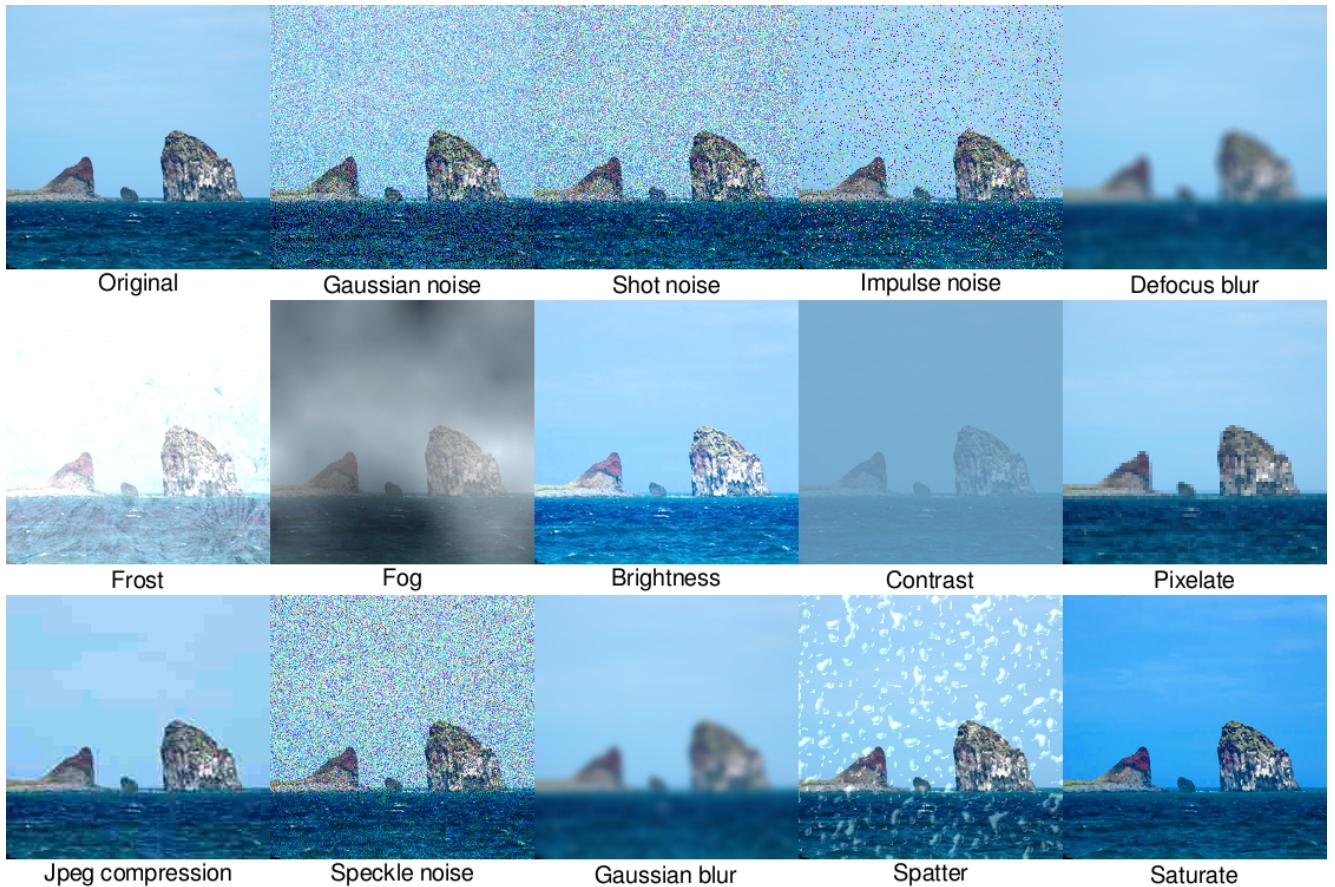


Figure 12. ImageNet-C perturbations on an example image, noise strength is set to 3 (out of 5).

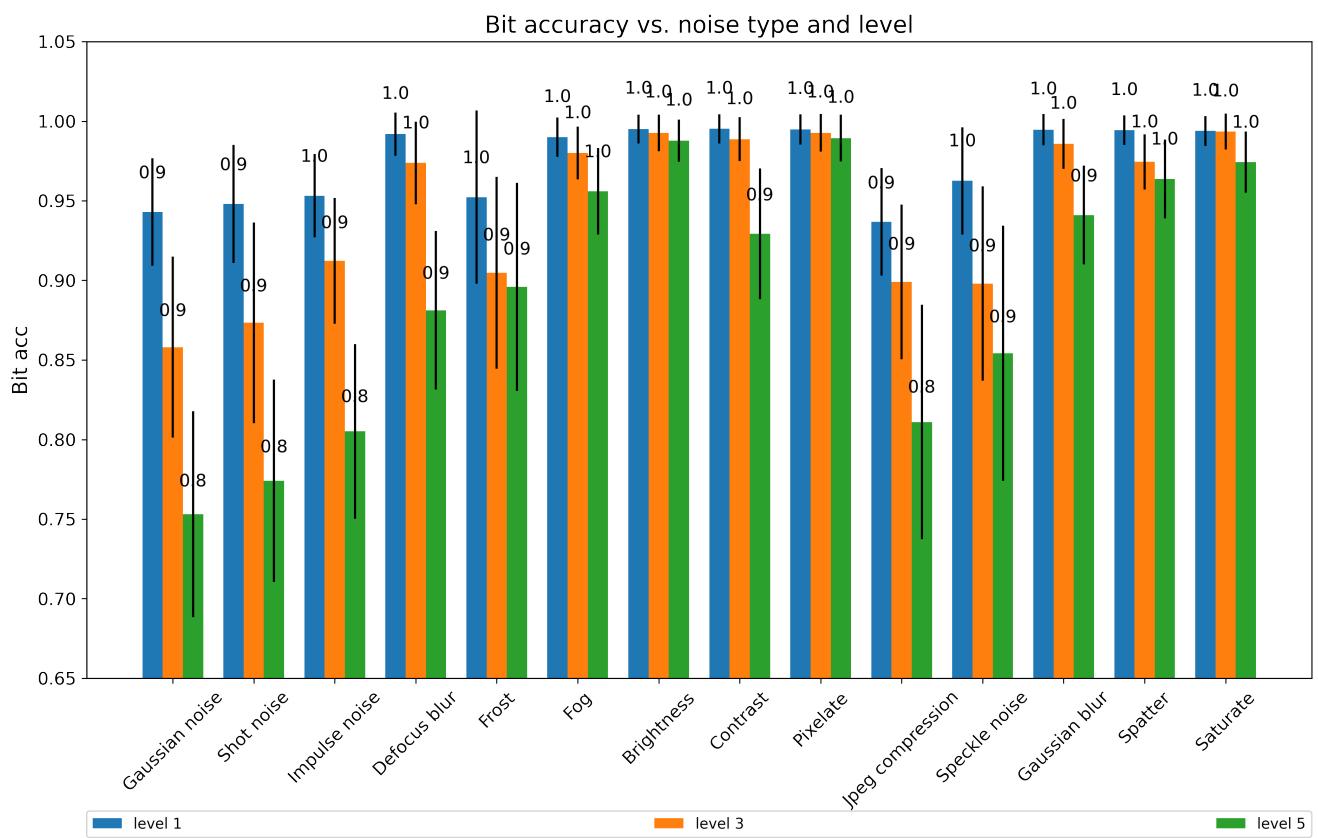
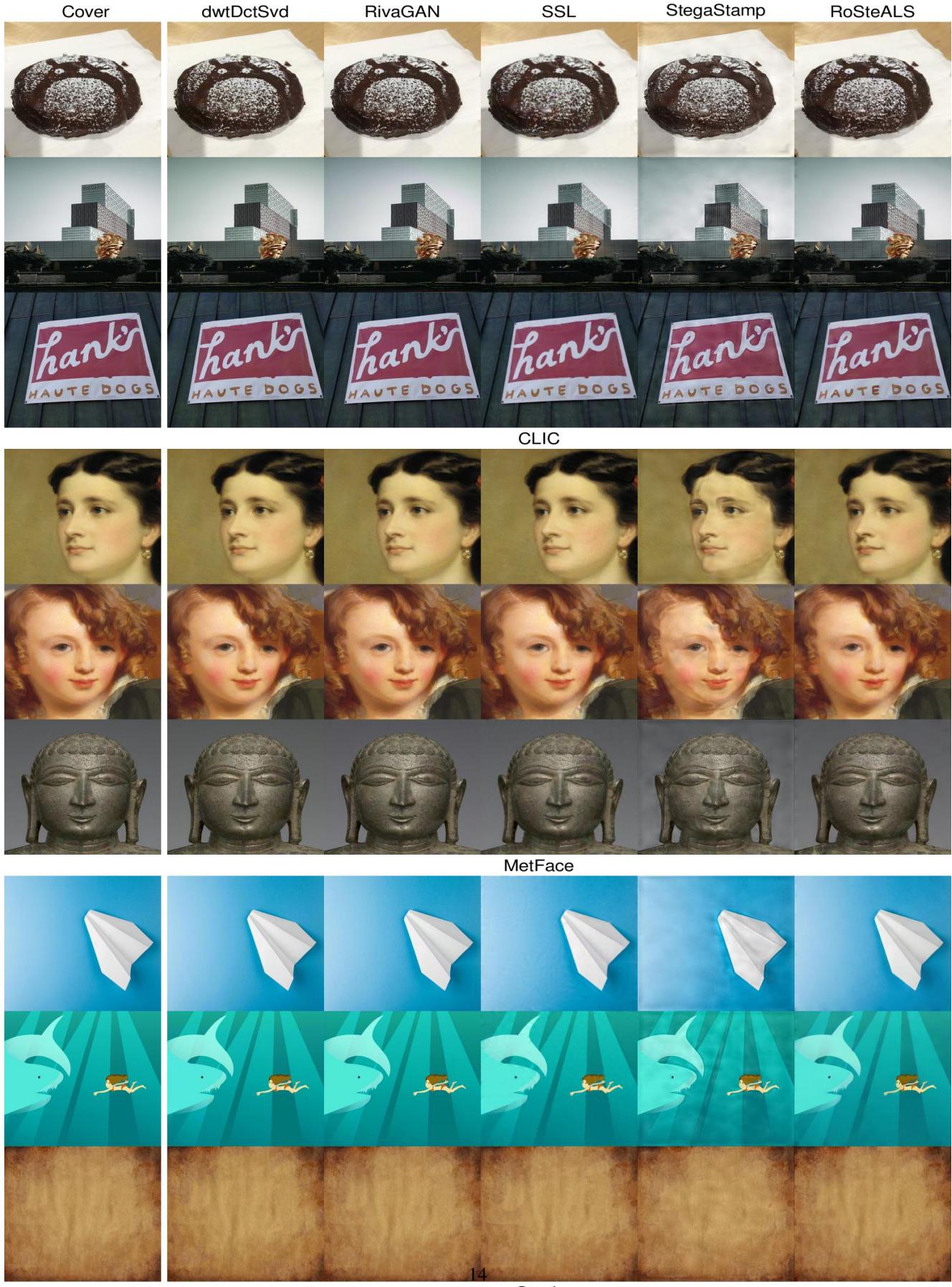


Figure 13. RoSteALS secret recovery performance breakdown for noise types and severity levels .



14
Stock

Figure 14. Stego images generated from several covers and a fixed secret. RoSteALS has better image quality than StegaStamp and perceptually comparable with other methods.

LDM



a futuristic fighter jet flying through cosmic skies. By Makoto Shinkai, Stanley Artgerm Lau, WLOP, Rossdraws, James Jean, Andrei Riabovitchev, Marc Simonetti, krenz cushart, Sakimichan, trending on ArtStation, digital art, UNSC Infinity, highly detailed, war spaceship, muted colors, dynamic lightning, F-41 Broadsword



a portrait of a cute girl with a luminous dress, eyes shut, mouth closed, long hair, wind, sky, clouds, the moon, moonlight, stars, universe, fireflies, butterflies, lights, lens flares effects, swirls, bokeh, brush effect, in style of Yoji Shinkawa, Jackson Pollock, wojtek fus, by Makoto Shinkai, concept art, celestial, amazing, astonishing, wonderful, beautiful, highly detailed, centered

LDM-RoSteALS

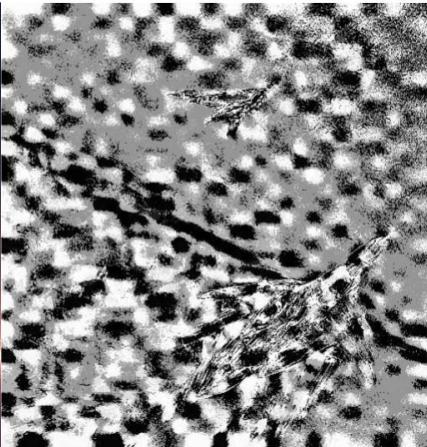


PSNR: 40.98, SSIM: 0.99
LPIPS: 0.01, SIFID: 0.00
Bit acc. (clean): 1.00



PSNR: 39.03, SSIM: 0.98
LPIPS: 0.01, SIFID: 0.00
Bit acc. (clean): 1.00

Residual



Full page concept design how to craft life poison, intricate details, infographic of alchemical diagram of how to make potions, captions, directions, ingredients, drawing, magic, wuxia



PSNR: 31.97, SSIM: 0.93
LPIPS: 0.02, SIFID: 0.03
Bit acc. (clean): 0.98

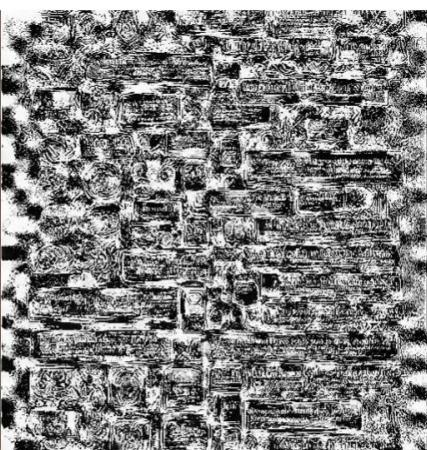


Figure 15. Text-based steganography with LDM-RoSteALS. (Left)- 512x512 images sampled from Stable Diffusion using the given prompts. (Middle) - Stegos with secret word "RoSteALS" injected. (Right) - Residual image scaled to [0,255] range.