RoSteALS: Robust Steganography using Autoencoder Latent Space

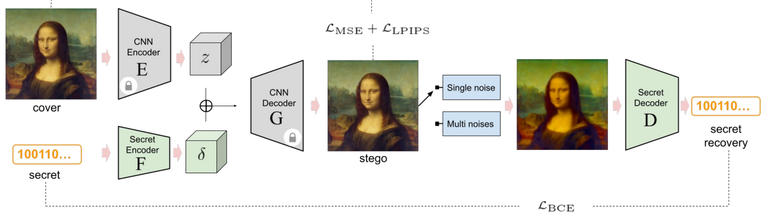
Tu Bui[1] Shruti Agarwal[2] Ning Yu[3] John Collomosse[1,2]
[1]University of Surrey, [2]Adobe Research, [3]Salesforce Research
t.v.bui@surrey.ac.uk, {shragarw,collomos}@adobe.com, ning.yu@salesforce.com

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

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 light-weight 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.

Method, Dataset, and Results



Shown above is the architecture diagram of our watermarking method -- RoSteALS. The image encoder (E) and decoder (G) are locked during training, only updating the lightweight secret encoder (F) and decoder (D). F is a very small network consisting of a fully-connected layer followed by SiLU[1]. The output δ acts as a small offset to be added to the cover embedding z. The stego image is then constructed as x = G(z + δ) and regulated using a combination of pixel and perceptual losses. We use Resnet50 [2] as the secret decoder D, replacing the last fully connected layer to output a L-bit secret. We use Binary Cross-Entropy (BCE) to compute the bit recovery loss between the predicted and the ground truth secret.

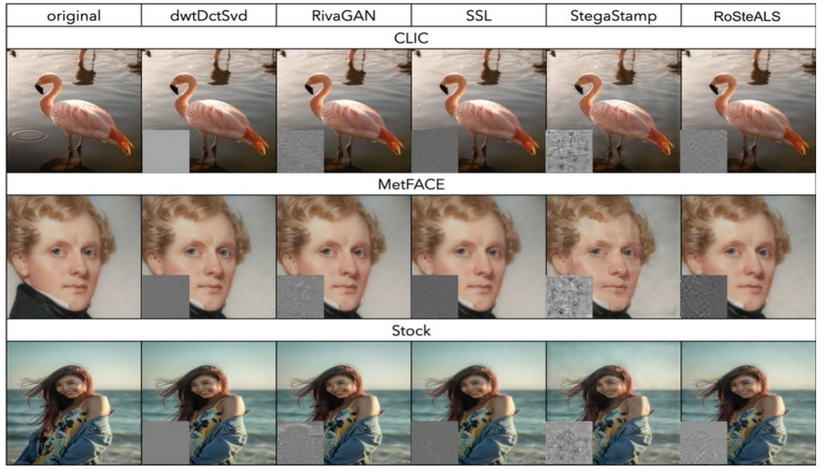
We train RoSteALS using 100K images and validate on 1K images from the MIRFlickR dataset [3]. We evaluate on 3 different benchmarks - CLIC [4], Met-Face [5] and Stock1K - our own collection of 1K images from Adobe Stock.

Source Code

Model https://github.com/TuBui/RoSteALS

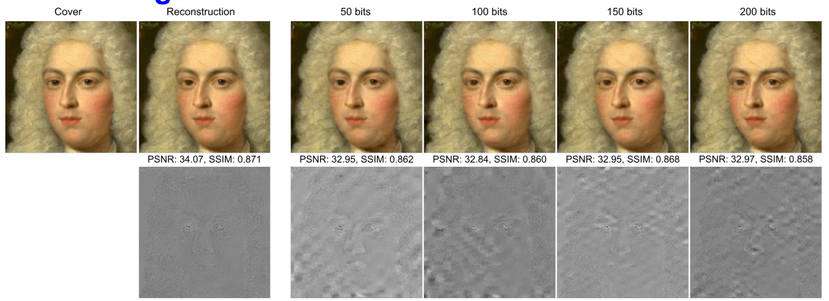
	Method\ Measure	LPIPS ↓	SSIM ↑	Accuracy Clean ↑	Accuracy Noise ↑	Accuracy Noise (ECC) ↑
handcrafted	Jsteg[6]	**0.01**	**0.99**	**1.00**	0.55	0.13
	OutGuess[7]	**0.01**	**0.99**	0.82	0.47	0.10
	dctDWTSVD[8]	0.02	0.98	0.99	0.63	0.17
data-driven	rivaGAN[9]	0.03	0.97	0.99	0.78	0.13
	Stegastamp[10]	0.29	0.64	0.99	0.87	0.48
	SSL[11]	0.04	0.98	**1.00**	0.63	0.02
	RoSteALS	0.04	0.91	**1.00**	**0.92**	**1.00**

Shown is the performance of our method in comparison to the state-of-the-art watermarking methods. For stego quality, we report SSIM and LPIPS. For secret recovery, we report standard bit accuracy using clean stego, noisy stego, and noisy stego after ECC (cyclic error correction using BCH [12]). Even though there is not a clear winner in stego image quality, RoSteALS achieves the best secret retrieval performance both before and after stego image corruption, resulting in perfect performance in both clean and noise data (with ECC).



Qualitative examples of watermarked images from different techniques. In each case, except StegaStamp, the watermarked images look like the original images with no perceptual artifacts. Notice in the case of StegaStamp, there are visible artifacts in the entire image.

Secret-Length



Change in image quality as secret length increases. Residual images are scaled to [0,255] range for visualization purpose. The average image quality (SSIM) for 50-, 100-, 150-, 200-bits secret length is 0.89, 0.88, 0.88, 0.88, while the bit-accuracy in noisy stego is 0.97, 0.94, 0.87, and 0.84.

Strengths

- better robustness and image quality
- faster training and inference time
- adabtable to secrets of different lengths 50 to 200 without loss of image quality
- generalizability to unseen image domain
- application to coverless steganography

Limitations

- image quality limited by the performance of pretrained autoencoder
- struggles in reconstructing small text or face
- struggles in reconstructing images with cluttered objects

References

1. P. Ramachandran, et al. Swish: a self-gated activation function. arXiv. 2017.
2. K. He, et al. Deep residual learning for image recognition. CVPR. 2016.
3. M. J. Huiskes, et al. The mir flickr retrieval evaluation. *ACM International Conference on Multimedia Information Retrieval*. 2008.
4. G Toderici, et al. Workshop and challenge on learned image compression (CLIC2020). CVPR, 2020.
5. T. Karras, et al. Training generative adversarial networks with limited data. *NeurIPS*. 2020.
6. X. Li, et al. A steganographic method based upon jpeg and particle swarm optimization algorithm. *Information Sciences*. 2007.
7. N. Provos. Defending Against Statistical Steganalysis. *Usenix Security Symposium*. 2001.
8. K.A .Navas, et al. DWT-DCT-SVD based watermarking. *COMSWARE*. 2008.
9. K.A. Zhang, et al. Robust invisible video watermarking with attention. *arXiv*. 2019.
10. M. Tancik, et.al. Stegastamp: Invisible hyperlinks in physical photographs. *CVPR*. 2020.
11. P. Fernandez, et al. Watermarking images in self-supervised latent spaces. *ICASSP*. 2022.
12. R. C. Bose, et al. On a class of error correcting binary group codes. *Information and control*. 1960.

Acknowledgements

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