A Journey Through Steganography Security Marks: Tracing Innovations from StegaStamp to StampOne

Farhad Shadmand
farhad.shadmand@isr.uc.pt
Institute of Systems and Robotics,
University of Coimbra,
Portugal

Luiz Schirmer
luizschirmer@unisinos.br
University of the Sinos
River Valley Rio de Janeiro,
Brazil

Nuno Gonçalves
nunogon@deec.uc.pt
Institute of Systems and Robotics,
University of Coimbra, Portugal
INCM, Lisbon, Portugal

INTRODUCTION

Modern machine-readable travel documents (MRTDs) and industrial authentication systems increasingly integrate a combination of biometric identifiers and security markings to prevent unauthorized replication or tampering. Beyond MRTDs, security pattern technologies are widely employed in other domains, particularly in brand protection and tax validation. Their use is prevalent in industries like luxury goods, wine and spirits, pharmaceuticals, and medical packaging.

This abstract introduces a set of advanced techniques for visual information embedding, drawing on the principles of steganography and digital watermarking. These methods are built upon modern deep learning frameworks and are designed to meet stringent requirements for security, robustness, and imperceptibility. The proposed solutions enable secure data integration within visual media and are applicable to high-stakes scenarios such as identity verification, counterfeit prevention, and brand authentication.

Steganography is the practice of embedding information within another medium in a manner that conceals the very existence of the hidden data. In digital image-based steganography, this typically involves two independent image-to-image neural networks: an encoder and a decoder. The encoder learns to embed a secret payload—such as text, another image, or binary code—into a cover image, producing an encoded image that remains perceptually similar to the original. The decoder, operating independently, is trained to extract and reconstruct the hidden message from the encoded image, even under conditions of distortion or noise.

The first highly robust steganography model in this field was StegaStamp [8], which introduced a pioneering approach by simulating a wide range of digital and physical distortions—including printer and scanner noise—during the training process. Its architecture employed a specialized U-Net encoder with a bottleneck layer for embedding the payload. However, while StegaStamp demonstrated resilience to various distortions, it struggled to preserve the structural integrity of input images. This limitation became particularly evident on semantically sensitive datasets, such as frontal face images, where it exhibited poor perceptual performance.

To overcome these challenges, we propose CodeFace [5] as a next-generation steganography framework that significantly improves the perceptual quality of encoded face images. Designed specifically for frontal facial data, CodeFace integrates a face-aware pipeline combining face detection and deep feature extraction to guide the encoding process. This enables the model to minimize perceptual discrepancies between the original and encoded images. We further deploy CodeFace as a security-enhancing layer for face images in Machine-Readable Travel Documents (MRTDs), offering both robustness and high visual fidelity in identity-sensitive applications.

Subsequently, RiemStega [3] was introduced to further enhance the performance of both the encoder and decoder components. This model incorporates a novel covariance-based loss function that operates in a Riemannian geometry space, encouraging the preservation of statistical consistency between original and encoded image features. In addition, RiemStega replaces the bottleneck structure used in StegaStamp's U-Net [4] with a self-attention mechanism, enabling more effective global feature interactions and improving the model's ability to embed and recover information with higher fidelity.

RoSteALS [1] is a lightweight and highly robust steganography framework based on generative adversarial networks (GANs), comprising only $300k$ parameters. Despite its compact architecture, the model exhibits strong resilience against various digital noise simulations, making it well-suited for purely digital communication scenarios. However, a key limitation of RoSteALS lies in its decoder's inability to accurately recover hidden messages from printed and re-scanned images, thereby restricting its applicability in print-based or physical media steganography.

Finally, we introduced StampOne [7], a steganography framework that bridges the gap between robust and non-robust models by placing greater emphasis on enhancing print-ability and resilience to real-world distortions. StampOne proposes a novel Reinforcement High-Frequency Strategy, designed to improve the robustness of embedded messages against transformations introduced by printing and scanning processes. The model incorporates a dedicated analysis-and-conversion module that preprocessed input data before encoding and decoding. This module aims to optimize the spectral distribution of features—specifically by enhancing high-frequency components and ensuring balanced frequency representations, thereby improving both visual fidelity and message recoverability in the final encoded images.

KEY CHALLENGES AND DESIGN TRADE-OFFS IN STEGANOGRAPHY STAMPS

The methods presented in this dissertation are grounded in the intersection of steganography, digital watermarking, and deep learning. By harnessing the advanced feature extraction and representational power of neural networks, we propose techniques that strive to optimize the trade-offs among several key performance criteria:

- **Perceptual Quality:** Maintaining a high degree of visual similarity between the encoded and original images, thereby concealing the presence of embedded information and preserving the natural appearance of the cover image.
- **Robustness:** Ensuring reliable extraction of the hidden payload under a wide range of digital and physical perturbations, including compression artifacts, additive noise, geometric distortions, and surface damage such as scratches or folds.
- **Capacity:** Maximizing the volume of information that can be embedded without degrading perceptual quality, while maintaining decoder reliability.
- **Security:** Providing strong protection against unauthorized decoding or tampering by designing models that resist reverse engineering, brute-force extraction, and adversarial attacks.

These objectives guide the design of the proposed frameworks, enabling the development of practical, scalable, and secure steganography systems suitable for real-world deployment.

PERFORMANCE COMPARING BETWEEN MODELS

Table 1 (A) presents the perceptual quality evaluation of encoded images. Among the compared models—CodeFace, StegaStamp, and Stam-

Table 1: (A) Quantitative evaluation of encoded image quality using perceptual similarity metrics. (B) Decoding performance on 40 printed encoded images, captured using a Samsung S22 Ultra smartphone. Models M1 and M2 correspond to the StampOne architecture employing Attention-VNet and UNetPlus backbones, respectively. Model M3 represents a non-robust baseline constructed with two independent Attention-VNet networks. The first four rows present results from high-robustness models, while the final two rows provide non-robust references, serving as a benchmark for evaluating decoder reliability under real-world print-capture conditions.

	(A) Encoded images quality			(B) **Bit acc (%)** - VGGFace2 [2]				
Methods	SSIM (⇑)	LPIPS (⇓)	ColorHisto (⇓)	6×6 cm	5×5 cm	4×4 cm	3×3 cm	2×2cm
StegaStamp [8]	0.93 ± 0.001	4.92 ± 1.6	6.11 ± 10.5	78	72	70	65	48
CodeFace [6]	0.95 ± 0.0002	3.06 ± 0.9	7.32 ± 6.1	55	55	50	38	15
StampOne (M1)	$\mathbf{0.98 \pm 0.00002}$	$\mathbf{1.25 \pm 0.4}$	$\mathbf{5.38 \pm 4.9}$	**100**	**100**	**100**	**95**	**62**
StampOne (M2)	0.96 ± 0.00007	2.74 ± 2.38	6.30 ± 4.07	88	85	72	63	43
Non-robust (M3)	0.92 ± 0.001623	1.04 ± 1.69	2.80 ± 60.8	0	0	0	0	0
RoSteALS [1]	0.95 ± 0.0006	$\mathbf{0.04 \pm 0.0003}$	$\mathbf{0.09 \pm 0.003}$	0	0	0	0	0

Table 2: Impact of three types of image under different noise types. 1000 images from COCO test dataset are used for the decoder performance evaluation. Bit accuracy (%) during decoding from encoded images is evaluated under various types and levels of noise. M1 and M2 represent StampOne models utilizing the Attention−VNet and UNetPlus architectures, respectively. On the other hand, M3 refers to a non-robust model constructed through the utilization of two instances of Attention−VNet.

	JPEG (%)			Gaussian (Std 0 to 1)			Resolution (Pixel)		
Methods	70	60	50	0.08	0.06	0.04	(60 × 60)	(80 × 80)	(100 × 100)
StegaStamp [8]	**100**	**100**	**100**	**100**	**100**	**100**	55	80	91
CodeFace [6]	80	88	88	55	75	86	2	11	36
RoSteALS [1]	87	90	94	23	35	53	**96**	97	98
StampOne (M1)	**100**	**100**	**100**	98	**100**	**100**	74	**98**	**100**
StampOne (M2)	97	99	**100**	88	96	99	72	94	99
Non-robust (M3)	0	0	0	13	46	84	0	0	22

pOne—StampOne demonstrates superior overall performance, particularly when using the Attention-VNet and UNetPlus backbones, as indicated in the table. Further evaluations of StampOne with alternative architectures are provided in the supplementary material.

In terms of SSIM, StampOne consistently achieves the highest scores among all robust models, indicating strong structural preservation. Although RoSteALS achieves slightly better results in the Color Histogram and LPIPS metrics, it fails to recover any messages from printed encoded images, limiting its practical applicability. In contrast, StampOne maintains both high perceptual quality and robustness to real-world printing conditions.

For print-based evaluation, a set of forty frontal face images from the VGGFace2 dataset was encoded and printed at various physical sizes, ranging from 2×2 cm to 6×6 cm (width $\times$ height), using a standard consumer-grade Brother L3270CDW color printer. To simulate real-world deployment conditions, decoding was conducted under uncontrolled lighting environments, with video recordings captured using a Samsung S22 Ultra smartphone.

The decoding performance of our proposed models—employing AttentionVNet and UNetPlus architectures—was benchmarked against established methods, including StegaStamp and CodeFace. As presented in Table 1(B), the Attention-VNet–based model consistently achieved the highest recovery accuracy from printed images, demonstrating superior robustness and confirming its effectiveness for printer-resilient steganographic applications. Additional cross-device results obtained using different smartphones are included in the supplementary material.

To assess decoder performance under real-world distortions, we conducted a series of experiments involving various noise conditions, including JPEG compression, Gaussian noise, resolution reduction, and contrast and brightness variations. Decoder effectiveness was quantified by the percentage of successfully recovered messages from the encoded images.

The results, summarized in Table 2, indicate that StampOne consistently outperforms competing models across most distortion scenarios. Notably, StegaStamp exhibits comparable robustness to StampOne under specific conditions, particularly JPEG compression and Gaussian noise, highlighting its resilience in digitally degraded environments.

REFERENCES

[1] Tu Bui, Shruti Agarwal, Ning Yu, and John Collomosse. Rosteals: Robust steganography using autoencoder latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 933–942, 2023.

[2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[3] Aniana Cruz, Guilherme Schardong, Luiz Schirmer, João Marcos, Farhad Shadmand, and Nuno Gonçalves. Riemstega: Covariance-based loss for print-proof transmission of data in images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Tucson, USA, 2025.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015.

[5] Farhad Sadmand, Iurii Medvedev, and Nuno Gonçalves. Codeface: a deep learning printer-proof steganography for face portraits. *IEEE Access*, pages 1–1, 2021. doi: 10.1109/ACCESS.2021.3132581.

[6] Farhad Shadmand, Iurii Medvedev, and Nuno Gonçalves. Code face: A deep learning printer-proof steganography for face portraits. *IEEE Access*, 9:167282–167291, 2021.

[7] Farhad Shadmand, Ivan Medvedev, Lucas Schirmer, Joao Marcos, and Nuno Gonçalves. Stampone: Addressing frequency balance in printer-proof steganography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4367–4376, 2024.

[8] 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.