Steganography Applications of StyleGAN: A Short Analytical Investigation from Hiding Message in Face Images

Farhad Shadmand[1]
https://linkedin.com/in/farhadsh1992
Luiz Schirmer[2]
https://www.lschirmer.com
Nuno Gonçalves[3]
http://home.deec.uc.pt/~nunogon/

[1] University of Coimbra
Institute of Systems and Robotcs, Coimbra, Portugal
[2] University of Coimbra
Institute of Systems and Robotcs, Coimbra, Portugal
INCM Lab
Portuguese Mint and Official Printed Office Lisbon, Portugal

Abstract

In this investigation, we delve into the latent codes denoted as w, pertaining to both original and encoded images in steganography models, which are projected through StyleGAN—a generative adversarial network renowned for generating aesthetic synthesis. We present evidence of disentanglement and latent code alterations between the original and encoded images. This investigator possesses the potential to assist in the concealment of messages within images through the manipulation of latent codes within the original images, resulting in the generation of encoded images. The message into encoded renderings is facilitated by the employment of CodeFace, serving as a steganography model. CodeFace comprises an encoder and decoder architecture wherein the encoder conceals a message within an image, while the decoder retrieves the message from the encoded image. By gauging the average disparities amid the latent codes belonging to the original and encoded images, a discerning revelation of optimal channels for concealing information comes to light. Precisely orchestrated manipulation of these channels furnishes us with the means to engender novel encoded visual compositions.

1 Introduction

Image steganography constitutes a methodological approach wherein a secret message is enveloped discreetly within a host image or video, concomitantly endeavoring to increase the perceptible congruence between the encoded (I_{en}) and original (I_{or}) images, ensuring inconspicuous conflict between original and encoded images to human observers. The elucidation of steganography can be succinctly encapsulated through the subsequent mathematical expressions:

$$\Delta I = |I_{or} - I_{en}| = \alpha_{visible} + \alpha_{invisible}, \tag{1}$$

where $\alpha_{visible}$ represents visible alternative for human eyes and $\alpha_{invisible}$ denotes invisible alternative for human eyes. The prime objective intrinsic to steganography endeavor resides in the minimization prescription encapsulated by ($arg\ min\ \Delta I$), concomitantly adhering to the stipulation ($arg\ min(\alpha_{visible}) \sim 0$), thus underscoring the imperative of inconspicuity within the realm of human visual perception,

$$arg\ min\ \Delta I = \alpha_{invisible} + arg\ min(\alpha_{visible}) \tag{2}$$

The realm of focus in this project pertains to the application of image steganography, which serves the purpose of safeguarding the authenticity of documents and commodities by counteracting analogous counterfeit versions of them. Operating within this contextual framework, steganography can manifest into two categories: robust and non-robust. Steganography models characterized as robust exhibit the capacity to endure distortions arising from printing processes and digital noise, while their non-robust counterparts are meticulously tailored for operation within pristine, noise-free digital domains.

Steganography models are constrained by multiple limitations encompassing the extent of the embedded message, the coherence of encoded images, precision in decoding, and resilience against fraudulent methods. The fundamental challenges inherent to the domain of steganography, lending credence to the present research endeavor, encompass the absence of a precisely delineated frontier between the discernible parameter $\alpha_{visible}$ and the imperceptible counterpart $\alpha_{invisible}$. Concurrently, an inherent constraint lies in the endeavor to avert the complete nullification of $\alpha_{visible}$, denoted as $\alpha_{visible} \neq 0$.

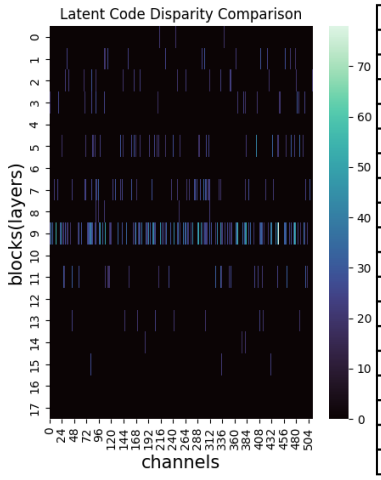
We endeavor to tackle this problem through this manuscript resides in the intricacies of reducing the differentials existing between original and encoded images, all the while ensuring optimal efficacy in the process of decoding messages embedded within printed images. Under the best optimal circumstances, imperceptible noise artifacts of hidden messages may become discernible within encoded images [8]. Our approach involves acknowledging the presence of noise artifacts that arise when hiding messages within images, while concurrently exercising authority over their configuration through a controllable image synthesis approach [3, 7]. The idea is to use a controllable image synthesis method like StyleGAN [5].

Given the substantial potential underscored within the purview of StyleGAN, a comprehensive evaluation and scrutiny of its viability in the context of steganography application is undertaken within this manuscript. The assessment delves into the alterations in latent code juxtapositions between the original and encoded images, employing the tools of CodeFace [8] and StampOne. This analytical exploration facilitates a discerning comprehension regarding the sectors of the latent space ($w+$) and specific StyleGAN blocks that bear heightened significance in preserving the covert message internally. Furthermore, insights gleaned from this investigation illuminate avenues through which style modulation can be judiciously administered and regulated during the course of the concealment process.

In Section 2, a comprehensive examination of the architecture of StyleGAN is conducted. Proceeding to Section 3, intricate elaborations are furnished concerning the procedural intricacies underlying the undertaken endeavors. Subsequently, Section 4 witnesses a comprehensive assessment of outcomes, wherein a meticulous comparison is undertaken between the latent codes derived from both the original and encoded images, through the lens of StyleGAN.

2 StyleGAN

A generative adversarial network (GAN) has two parts: generator and discriminator. The generator learns to generate plausible data. The discriminator learns to distinguish the generator's fake data from real data. In this work, we focus on StyleGAN [5] which is a cutting-edge generative model designed for producing highly realistic and high-resolution images. The key innovation of StyleGAN lies in its ability to control both high-level features, like overall structure and objects, and low-level features, such as textures and fine details, separately. This separation of control is achieved through a unique two-part generator network: a mapping network and a synthesis network. The mapping network takes a 512 dimensional random vector $z \in N(0,1)$ is mapped to an intermediate latent space $w+ \in R^{l \times 512}$, where l is the number of blocks (layers). This intermediate space is designed in such a way that different dimensions control different features of the image. For instance, one dimension might control the age of a generated face, while another might control the hairstyle. This separation of features enables fine-grained control over the generated content. The synthesis network then takes the modified latent vector from the mapping network and generates an image $I = G(w+)$. based on it. This synthesis process includes different level blocks responsible for different scales of detail. By adjusting the latent vector in the intermediate space, it's possible to manipulate specific attributes of the image, like the aforementioned age or hairstyle, while keeping other aspects constant [1]. StyleGAN represents a significant advancement in generative modeling, offering an unprecedented level of control over image generation. Its ability to generate large, diverse, and highly realistic images makes it



blocks	mean
block_9	11.1665
block_5	3.0773
block_7	2.9486
block_11	1.8379
block_3	1.4243
block_1	1.2588
block_2	1.2308
block_13	0.571
block_8	0.2821
block_15	0.2516
block_0	0.1835
block_14	0.1488
block_4	0.0368
block_10	0.0
block_12	0.0
block_6	0.0
block_16	0.0
block_17	0.0

blocks	channels	distance
9.0	445.0	78.0
9.0	400.0	57.89
9.0	366.0	55.98
9.0	271.0	55.9
9.0	477.0	55.16
9.0	13.0	54.47
9.0	418.0	52.85
9.0	383.0	52.28
9.0	290.0	51.38
9.0	109.0	48.17
9.0	463.0	48.0
9.0	99.0	47.34
9.0	93.0	46.42
9.0	216.0	46.14
9.0	494.0	45.71
9.0	296.0	45.33
9.0	25.0	45.32
9.0	349.0	44.92

Figure 1: The heatmap in the figure shows differences in latent codes between original and encoded images. The first right-side table highlights the main channels for message hiding, ranked by importance. On the left, another table lists the top eighteen channels crucial for concealing messages in images.

a valuable tool for various applications, including art, fashion, entertainment, and even data augmentation for machine learning tasks.

In here, we use StyleGAN with 18 layers (blocks), every block has 512 input channels and 1024×024 output resolution, $w+$ has 9126 dimensions (18×512).

3 Method

The experimental framework is rooted in the utilization of the IMM Face Database [6]. This dataset encompasses a substantial repository of 1830 facial images.

The experimental protocol unfolds as follows: **(1) Face Detection and Extraction:** The initial phase entails the application of a specialized face detection algorithm [4], serving to identify and subsequently extract facial components from the comprehensive source images. **(2) Encoded Image Generation:** Subsequent to the face extraction process, encoded images are engendered. These encoded counterparts serve as the carriers for concealing randomized confidential messages within the facial image. The concealment procedure is conducted across the facial images, employing CodeFace. **(3) Latent Space Projection:** A critical dimension of this investigation involves the projection of the latent space across the entire spectrum of encoded images, encompassing $w+_{CF}$, in tandem with the unaltered originals denoted as $w+_{or}$. **(4) Comparative Analysis:** The ultimate phase encompasses an intricate comparative analysis, shedding light upon the variances that distinguish the latent spaces across the aforementioned image categories. This scrutiny unveils insights regarding the disparate impact of distinct sections and individual blocks prompted by the network. By discerning the facets that experience greater influence and significance within this context, a foundation is laid for designating essential components germane to the Steganography endeavor grounded in the domain of StyleGAN networks.

3.1 Loss functions

Supplementary to the loss functions inherent to StyleGAN, an augmentation is introduced through the incorporation of pre-trained steganography decoders and the utilization of binary cross-entropy. This augmentation is strategically integrated to ensure the preservation of the concealed message throughout the latent code projection process.

3.2 Disentangled degree measurement

In the pursuit of methodological exactitude within this undertaking, it becomes imperatively requisite to interface with entities resembling Principal Component Analysis (PCA), Support Vector Machines (SVM), or attention-based classification models [2]. These avenues facilitate a thorough evaluation of an augmented dataset. Nonetheless, it is germane to accentuate that the present inquiry espouses considerably more uncomplicated instruments. Specifically, we compute the mean disparity between latent codes of original and encoded images to establish a threshold. Subsequently, channels exhibiting disparities surpassing this threshold are earmarked as pivotal channels for the concealment of messages.

4 Experiments

The initial diagram in Figure 1, situated on the left, manifests as a heatmap delineating the average discrepancy between original and encoded images. Adjacent to this, a tabular exhibition enumerates the mean disparity values corresponding to each individual block. Concluding this sequence, the table situated on the left enumerates the 18 preeminent channels identified for the covert embedding of messages. These channels, characterized by substantial alterations within latent codes between original and encoded images, emerge as focal points of significance.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: how to edit the embedded images? in 2020 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8293–8302.

[2] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International conference on learning representations*, 2018.

[3] Yuki Endo. User-controllable latent transformer for stylegan image layout editing. In *Computer Graphics Forum*, volume 41, pages 395–406. Wiley Online Library, 2022.

[4] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018.

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[6] M. M. Nordstrøm, M. Larsen, J. Sierakowski, and M. B. Stegmann. The IMM face database - an annotated dataset of 240 face images. Technical report, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, may 2004. URL `http://www2.compute.dtu.dk/pubdb/pubs/3160-full.html`.

[7] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. *arXiv preprint arXiv:2305.10973*, 2023.

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