Synthetic Handwritten Signature Generation Using Deep Convolutional GANs: A Security Evaluation Framework

Sameera Khan[1], Eugenio Vocaturo[2]

[1] Lincoln University College, Petaling Jaya, Malaysia ; [2] University of Calabria, Italy;

Email ID -pdf.sameerakhan@lincoln.edu.my

Abstract

Handwritten signatures remain a widely used biometric modality for authentication and forensic verification, yet the security implications of generative adversarial network (GAN)-based signature synthesis have received limited direct evaluation compared to other biometric modalities such as faces and fingerprints. This paper presents a generate-then-detect security evaluation framework applied to handwritten signatures. A Deep Convolutional GAN (DCGAN) was trained on the CEDAR genuine signature corpus under free-tier cloud GPU constraints, following an initial, ultimately infeasible attempt using StyleGAN2-ADA. The resulting synthetic signatures exhibited recognizable stroke-level structure but showed evidence of partial mode collapse, yielding a Fréchet Inception Distance (FID) of 248.95 and a mean Structural Similarity Index (SSIM) of 0.512 (± 0.077) relative to genuine samples. A ResNet-18 classifier fine-tuned to distinguish genuine from synthetic signatures achieved perfect test-set performance (Accuracy, F1-score, and AUC-ROC of 1.0000; Equal Error Rate of 0.0000), a result attributed to the limited diversity and fidelity of the generated samples rather than an inherently unbreakable detection boundary. These findings provide, to the authors' knowledge, the first generate-then-detect security evaluation conducted specifically on the handwritten signature modality, and offer a practical methodological note on generative architecture selection under resource-constrained research conditions.

Keywords: *synthetic biometrics, DCGAN, spoof detection, biometric security, signature forgery detection, ResNet-18*

1. Introduction

Handwritten signatures have been used in banking, legal and administrative system for a long time but still they have to rely on legal and institutional acceptance as one of the most widely adopted biometric modalities for authentication, contractual validation and forensic verification. The signatures are behavioral artifacts and their continued use is dependent on the assumption of their uniqueness and the lack of ease in which they are reproducible by a third party.

This is beginning to be questioned with generative adversarial networks (GANs). Recently, GAN based architectures have been able to generate biometric characteristics such as faces, fingerprints and handwritten signatures that are so realistic that they are able to fool human and automatic verification systems [3, 4]. This ability is very handy, but it can be a two-edged sword. Use of synthetic biometric data can be a solution to this problem of small datasets and demographic imbalance in biometric research, allowing to build up models without the risk of entering extra real subject biometric data. Synthetic biometric data can offer another way to solve the issue of limited datasets and demographic skews in biometric research and create models without putting further real subject biometric data at risk. A critical vulnerability of that generative ability is also a potential for spoofing where a synthetic signature created from a small reference set could have a high probability of passing through a signature-based authentication system, or even be able to pass the route through a forensic system.

To date, most of the research on synthetic biometric generation has been limited to the realism of the synthetic samples, which is the fidelity of the sample generated, and has not systematically investigated the potential security risk posed by such samples. Recently, synthetic fingerprint generation has been shown to be a viable way to augment presentation-attack-detection training [5] and parallel work in the facial domain has been performed in which the faces generated by GANs can be reliably detected by dedicated forensics fingerprint classifiers [6]. While the modality is still of great relevance in forensic and authentication applications, the 'generate then detect' paradigm has not been applied specifically to handwritten signatures, to the best of the authors' knowledge.

To address this gap, this paper proposes a model that can be used for the synthesis of handwritten signatures using a generative adversarial network (GAN) and to determine the number of synthetic samples that can be misclassified by a customized signature classifier. Unique contribution of this paper is-A GAN based pipeline- designed to produce realistic handwritten signatures for a small training set (the CEDAR genuine signature subset), the quality of the signatures generated is evaluated quantitatively by Fréchet Inception Distance (FID) and Structural Similarity Index (SSIM) along with the qualitative visual evaluation. This is assessed through a ResNet-18 binary classification system that differentiates synthetic from real signatures and not just the generativeness of the signature. The result of the detection performance obtained is discussed with respect to signature based biometric authentication systems and forensic verification systems.

The rest of this paper is organized as follows. In Section 2, a brief overview of related research in the area of synthetic generation and spoofing detection is given. A description of the dataset, the generative pipeline and the detection methodology are described in section 3. Experimental results are given in Section 4. The meaning of the results and its restrictions are discussed in Section 5 and the conclusions drawn from the paper in Section 6.

2. Related Work

Synthetic Biometric Generation

Many GANs have been proposed to generate realistic identities including faces, fingerprints and signatures due to the scarcity of data or an imbalance distribution of data in the training set. The multi-stage GAN proposed by Grosz and Jain [5] can generate fake fingerprint spoof images to significantly boost the training data for presentation-attack-detection, but does not heavily depend on the real spoof fingerprint images. Similarly, GAN-based style synthesis methods have produced quality identity images that are close enough to real images to pose an identity problem for human and automated face verification systems. While signature-specific generation is a fundamental modality in forensic and contract-authentication cases, in the above the modalities demonstrated good generative properties, security evaluation of signature-specific generation has been less studied.

Spoofing and Biometric Vulnerability – based on GAN.

There are growing number of studies that highlight how new attack surfaces are created by generative models.There are more and more studies that identify the new attack surfaces that generative models create. Othershi-Shahreza and Marcel [7] show that face recognition systems can be vulnerable to template inversion attacks, the attacks can be very successful in the white box and black box scenarios with 3D face model. The presence of deepfakes in so many modalities is not an issue regarding a specific modality, but an issue of identity verification systems in general [3, 4] and an issue of the basic

assumptions of identity such as permanence and uniqueness. This outcome strongly suggests a security point of view for signature synthesis as it is done for face and fingerprint modalities.

Synthetic Image Detection

In order to mitigate generative spoofing, steps have been taken to determine whether a biometric sample is synthetic or real. Generative spoofing threat, countermeasure to detect synthetic and real biometric samples. Say et al. [6] provide a benchmark for the performance of deepfake detection on face datasets that include a variety of synthetic images of faces generated by GANs and find that classical classifiers based on feature extraction (with RBF kernel SVM) can successfully detect synthetic faces in different sets with near-perfect accuracy. This gives a framework which indicates that detectability is not an intrinsic property of generative architecture, it's rather a property of the discriminative characteristics of the downstream classifier. Here, this detection paradigm – training a dedicated detector to tell the difference between real and synthesised samples – is applied to the handwritten signature domain where it has not been directly tested, to the authors' best knowledge.

3. Methodology

3.1 Dataset

The handwritten signature dataset CEDAR was used, containing 55 genuine and forged handwritten signatures, each of which gave their 24 genuine and 24 forged signatures (2640 images in total). For the generative modeling phase only the true signature class was kept resulting in about 1320 training images. Because of image inversion inconsistencies present in the raw images, the images were grey-scaled, and then resized to 128x128 pixels and normalized to have the signature strokes as black pixels on a uniform white background.

3.2 Synthetic Signature Generation

The synthetic signatures were generated by a Deep Convolutional Generative Adversarial Network (DCGAN) [8] composed of 4 transposed-convolutional layers in the generator network, and 4 convolutional layers in the discriminator network, both networks were trained at 64×64 resolution. After trying to build with StyleGAN2-ADA architecture [11] architecture, which was found to be too expensive for the free-tier cloud GPU resources (as CUDA kernel compilation would fail often during the training process and the data-loading process would be CPU-bound), this lightweight and fully transparent PyTorch implementation was adopted. A typical objective used in adversarial training of the DCGAN was used, and generator/discriminator trained using alternating gradients with the Adam optimizer.

The training was carried-out on a single NVIDIA T4 GPU (free tier on Kaggle), with batch size 32, for 2 thousand times of actual training, at the resolution 64x64. Model Checkpoints and sample image grids were saved every 5 epochs while the model was trained. For downstream evaluation a set of 1000 synthetic signature samples were generated using the last trained generator.

3.3 Quality Evaluation Metrics

Two complementary metrics were used to evaluate the fidelity with which generated signatures were produced.

In [9] an ideal measure of the distance between a set of real images and a set of synthetic images in a learned image space is called Fréchet Inception Distance (FID), which is a smaller distance in the image space, the closer the distributions of real and synthetic images.

Structural Similarity Index (SSIM) is a measure of perceptual similarity between two pairs of images (based on a per-pixel structural basis), calculated as:

$$SSIM(x, y) = [(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)] / [(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)]$$

where μ, σ represent local means and local standard deviations, σ_{xy} is the covariance and c_1, c_2 are stabilizing constants.

3.4 Spoof Detection Framework

A binary classifier was used to test the security properties of the synthetic signature generation system, which was trained to recognize the gaussian noise-free CEDAR signatures from synthetic signatures from DCGAN. A ResNet-18 [10] is fine-tuned for this binary classification task, which is pre-trained on ImageNet.

The total set (both authentic and synthetic) was then split into training set (70%), validation set (15%) and testing set (15%). The training time augmentation consisted of horizontal flipping, rotation by ±5°, brightness and contrast jitter. This model was then fine-tuned for 15 epochs using the Adam optimizer (learning rate = 1e-4) and step learning rate scheduler. The test set (which was never used during training) produced the following metrics: accuracy, F1 score, AUC-ROC, and Equal Error Rate (EER).

4. Results

4.1 Generative Pipeline Implementation

The initial few generations were modeled after StyleGAN2-ADA [11] but were not suitable for the free tier GPUs, and were prone to CUDA kernel compilation errors and employed unreasoned training times under CPU loaded data loading. In its place, a lighter generator–discriminator structure of four transposed-convolutional layers (generator) and four convolutional layers (discriminator) was used, and was trained at 64×64 resolution, a Deep Convolutional GAN (DCGAN) [8]. It was able to train for approximately 1.6–2.2 seconds per epoch, which was a massive improvement from the infeasible multi-hundred-hour estimate and enabled 2,000 epochs to be completed in a single free-tier GPU session.

4.2 Synthetic Signature Quality

The results of qualitative investigations of the samples indicated that there was no doubt that the samples generated during the course of training showed progressive improvement. The generated images were more of a loose texture, similar to strokes, and had low levels of structural coherence in early stages (epoch 100). For the generated samples, the pen-stroke structures were well segmented at epoch 2000, similar to hand-writing gestures in terms of structures, e.g. loop and flourish. However, it was observed that a small fraction of the training distribution was what the generator was learning and was thus a mode collapse, a phenomenon common to the training of DCGANs that is known to happen when the generator becomes too closely tuned to a small part of the input distribution. Some degree of visual resemblance was seen between a number of samples produced from a number of independent vectors of noise.

The Fréchet Inception Distance (FID) and mean Structural Similarity Index (SSIM) were 248.95 and 0.512 ± 0.077 respectively between the 1,000 samples generated and 1,320 real signature images, randomly sampled from the CEDAR database. This relatively high FID value, which is well above the common value reported for GAN-based face synthesis benchmarks, aligns the limited diversity seen qualitatively, and also demonstrates the domain mismatch between the distribution of the training images that InceptionV3 was trained on (natural images) and the low-texture, sparse nature of the handwritten signature images.

The moderate level of SSIM indicates that the samples created have structural similarity at the stroke level and that the diversity is hindered but meaningful.

4.3 Spoof Detection Performance

The ResNet-18 classifier that was optimized to classify genuine signatures from synthetic signatures performs at an accuracy of 1.0000, F1-score of 1.0000, AUC-ROC of 1.0000 and Equal Error Rate (EER) of 0.0000 on the held-out test set (n=348). The classifier was able to distinguish between genuine and fake signatures under the experimental conditions mentioned in Section 3.4.

Strikingly, but not surprisingly in the light of the limitations of generative quality discussed in Section 4.2, this result is consistent with the limitations of generative quality identified above. The partial mode collapse and high FID indicate that the output signatures of the DCGAN have a good level of local stroke structure but are not highly representative of low-level statistical artifacts and inter-sample diversity as compared to real signatures for the synthetic samples. These results are representative of architectures with low weights trained on small data sets, and have been reported in the context of face detection tasks trained using GAN-generated faces [6] as being highly discriminative for the convolutional face classifiers. This apparent separability ought to be interpreted as being due to the current limitations in generative models and not as a fundamental limit in detection: As the training time increases, the set size grows, and the model architecture grows more expressive (such as StyleGAN2), the detection performance is expected to decline.

Table 1. Summary of generative evaluation metrics.

Metric	Value
FID	248.95
Mean SSIM	0.512 ± 0.077
Training epochs	2,000
Real images	1,320
Synthetic images generated	1,000

Table 2. Spoof detection performance on the held-out test set.

Metric	Value
Accuracy	1.0000
F1-score	1.0000
AUC-ROC	1.0000
EER	0.0000
Test samples	348

5. Discussion

5.1 Signature Based Authentication and Forensic Verification implications

In fact, the main empirical finding of this work – that a finely tuned ResNet-18-based classifier can perfectly classify between synthesized and genuine signatures – is somewhat reassuring for current signature-based authentication and forensics practice, and for future work: Even with the limited

generation capabilities considered here, a visual forgery, which is hard for a naïve human observer to distinguish, leaves a detectable (and distinguishable) statistical signature. This demonstrates that automated, classifier-based screening is a potentially significant safeguard against a form of signature spoofing that relies on the easily-obtained computing power and on just a few legitimate-signature samples, which can more and more be achieved using a pre-trained generative tool.

That's not to say that the discovery is a defensive answer to the signature spoofing problem is solved, though. Here, the diversity of samples and size of the training corpus are strongly correlated with detectability, as outlined in Section 4.3, which represent generative constraints of the DCGAN architecture used. A more powerful adversary with a larger set of signatures or a more expressive generative model, such as an appropriately trained StyleGAN2 model, would likely be capable of generating synthetic signatures that are far closer to the natural ones, and that have a much wider distribution that would help to diminish the statistical gap detected by the present classifier. The security implication of being able to detect synthetic signatures is not that synthetic signatures can be detected per se but rather that as the attacker's resources grow, so does the difficulty of detection, which is a well-known phenomenon in other biometric fields such as fingerprint spoof detection [5] and face spoof detection [6].

The pipeline generates and selects generations (5.3 Generation and Selection of Pipeline Generations) and it is limited in the number of generations (5.4 Limitation of Pipeline Generations).

A secondary finding was methodologically-oriented and naturally emerged during the process of implementation. Given the limited cloud GPU compute for free, the first idea that comes to mind is to use a GAN trained with StyleGAN2-ADA which is renowned for training GANs with few data. The initial training setup was found to be impracticable because of repeated failures of the CUDA kernel, different versions of PyTorch for different environments, and excessive data-loading CPU bottlenecks. These problems were solved by the last change to a much smaller size of the DCGAN structure, which enabled the whole experimental process to be executed from generation to evaluation and detection without resorting to paid tier resources.

In general, the experience here will lead to other practical conclusions for research scientists who have to use similar compute constraints: When the toolchain or the stability of the infrastructure are not guaranteed, sophistication of architecture and cutting-edge reputation are not necessarily practical. When there is limited time for development, and/or resources, simpler, well understood architectures with less external requirements might provide a more certain route to a complete, repeatable experimental pipeline than more complex architectures with higher implementation risks.

5.3 Limitations

The following limitations must be mentioned about the present study. First, the generative model was run for a relatively small number of epochs (2000) on a small number of images of real signatures (c. 1320) compared to the numbers of images used in the typical training of GANs, so it is possible that the mode collapse and high FID scores are at least partly due to this small data set, and not just to the architecture of DCGAN. Second, the test was conducted at a relatively low resolution (64×64 for image generation and upsampled to 128×128 for evaluation), which may not perfectly represent the higher level of detail in the strokes needed in forensic-level signature analysis. Third, the spoof detection classifier was trained and evaluated exclusively on samples from one DCGAN generator which generated the training set, and not tested for generalization to other generators of synthetic signatures; this is an important direction for future evaluations as a classifier that has been trained to detect signatures from one generator might not be able to generalize to another generator. Finally the FID metric was developed and validated in the

natural photographic image domain, and may not necessarily be directly comparable with the signature domain in which FID is more commonly reported and should be interpreted primarily in relative terms within the study.

6. Conclusion

In this paper a generate-then-detect method was presented for the evaluation of security issues of the synthetic handwritten signature generation. This Deep Convolutional GAN was trained with the true corpus of genuine signatures in the CEDAR project, and trained under the free tier of the GPUs, leading to generated signatures that seemed to exhibit some element of visual structure on the stroke level with some mode collapse and higher FID than the real signature distribution. The reason for the perfect detection performance (Accuracy, F1-score and AUC-ROC at 1.0000 and EER at 0.0000) on held out test set was deemed a direct consequence of the limited diversity and fidelity of the generator, rather than a characteristic that is invulnerable to break.

The results presented here are to the authors' knowledge the first implementation of a generate-then-detect security evaluation framework for the handwritten signature modality, in addition to fingerprint and face spoof detection. The results show that, at this time, signature spoofing is still detectable in low-resource scenarios such as free-tier compute – but at the same time they suggest that the detectability of such attacks in low-resource scenarios may be lowered as the fidelity of generative models grows, and that further investigation is needed as increasingly capable generative architectures become practically available in such low-resource scenarios.

From a methodological perspective, one of the key takeaways from this work is that while the architecture of the generators becomes more complex, e.g., StyleGAN2-ADA, the use of such models in free-tier cloud resources could be challenging due to instability of data and toolchains, and simpler models, such as DCGAN, can provide a viable pathway to a stable experimental pipeline under these constraints.

Further research, including in several directions: (i) test the generalization on synthetic signatures generated by a variety of different generative architectures; (ii) evaluate on longer training, higher resolution and when possible, when compute resources allow, and to (iii) expand the present security-evaluation framework to other behavioral and physiological biometric modalities within the broader scope of synthetic biometrics research goals in which this work is embedded.

References

[1] S. Khan, M. Mishra, and V. K. Mishra, “Autogenic, prognostic, and collective signature affirmation framework based on diverse set of features,” Int. J. Biom., vol. 15, no. 5, pp. 539–559, 2023.

[2] S. Khan, M. Mishra, and V. K. Mishra, “Use of synthetic signature images for biometric authentication and forensic investigation,” Int. J. Biom., vol. 15, no. 6, pp. 685–704, 2023.

[3] S. Saha et al., “Undercover Deepfakes: Detecting Fake Segments in Videos,” in Proc. 2023 IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW), 2023, pp. 415–425.

[4] A. Firc, K. Malinka, and P. Hanáček, “Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors,” Heliyon, vol. 9, no. 4, 2023.

[5] S. A. Grosz and A. K. Jain, “SpoofGAN: Synthetic Fingerprint Spoof Images,” IEEE Trans. Inf. Forensics Secur., vol. 18, pp. 730–743, 2023.

[6] T. Say, M. Alkan, and A. Kocak, “Advancing GAN Deepfake Detection: Mixed Datasets and Comprehensive Artifact Analysis,” Appl. Sci., vol. 15, no. 2, 2025.

[7] H. O. Otroshi-Shahreza and S. Marcel, “Comprehensive Vulnerability Evaluation of Face Recognition Systems to Template Inversion Attacks via 3D Face Reconstruction,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 12, pp. 14248–14265, 2023.

[8] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” in Proc. Int. Conf. Learning Representations (ICLR), 2016.

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[11] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8110–8119.