On Feasibility of Transferring Watermarks from Training Data to GAN-Generated Fingerprint Images

Venkata Srinath Mannam, Andrey Makrushin and Jana Dittmann
Department of Computer Science, Otto von Guericke University, Magdeburg, Germany

Keywords: Watermarking, Biometrics, Synthetic Fingerprints, Synthetic Data Detection, GAN, Pix2pix.

Abstract: Due to the rise of high-quality synthetic data produced by generative models and a growing mistrust in images published in social media, there is an urgent need for reliable means of synthetic image detection. Passive detection approaches cannot properly handle images created by "unknown" generative models. Embedding watermarks in synthetic images is an active detection approach which transforms the task from fake detection to watermark extraction. The focus of our study is on watermarking biometric fingerprint images produced by Generative Adversarial Networks (GAN). We propose to watermark images used for training of a GAN model and study the interplay between the watermarking algorithm, GAN architecture, and training hyperparameters to ensure the watermark transfer from training data to GAN-generated fingerprint images. A hybrid watermarking algorithm based on discrete cosine transformation, discrete wavelet transformation, and singular value decomposition is shown to produce transparent logo watermarks which are robust to pix2pix network training. The pix2pix network is applied to reconstruct realistic fingerprints from minutiae. The watermark imperceptibility and robustness to GAN training are validated by peak signal-to-noise ratio and bit error rate respectively. The influence of watermarks on reconstruction success and realism of fingerprints is measured by Verifinger matching scores and NFIQ2 scores respectively.

1 INTRODUCTION

Since the invention of Generative Adversarial Networks (GANs) (Goodfellow et al. 2014), there has been a significant rise in related research from six papers in 2015 to 762 in 2020 (Farou et al. 2020). GANs can generate realistic synthetic samples which are not tied to real persons, making such data very useful in areas with limited real data or strict restrictions on private data use such as medical research or biometrics. Synthetic images are often referred to as deepfakes because deep learning techniques are utilized for their production. Due to the security threats that may be caused by deepfakes (read synthetic images), the Chinese government banned production of deepfakes that are not watermarked (Edwards 2022). The same might happen in other countries soon.

The primary concern of our initial study was synthesis of realistic biometric fingerprint images. It has been shown in (Bahmani et al. 2021, Bouzaglo and Keller 2022, Makrushin et al. 2023) that GANs is a valid approach for this purpose. Despite all the benefits, GAN-generated biometric fingerprints can be misused for e.g. identity fraud. The study in (Bontrager et al. 2018) shows that synthetic fingerprints can mimic multiple identities without requiring a specific individual's fingerprint. Another example of malicious use of synthetic fingerprints is a fingerprint morphing attack (Makrushin et al. 2021b). Hence, our current concern is the active protection of synthetic fingerprints by watermarking them.

Indeed the passive protection approach, that is a "blind" detection of synthetic images, has its natural limits when it comes to detection of fakes produced by "unknown" generative models. Moreover, synthesis techniques and corresponding generative models constantly improve over time. Hence, embedding a watermark in all images produced by GANs is seen as a remedy to the problem of growing fake media. Watermarks enable an active protection transforming the task of fake detection to the task of watermark extraction.

In contrast to the most common goal of watermarking generative models which is intellectual property rights (IPR) protection, our motivation is linking synthetic fingerprints to a particular generative model. We currently disregard the fingerprint's integrity verification due to the technical aspects of our embedding

Mannam, V., Makrushin, A. and Dittmann, J.
On Feasibility of Transferring Watermarks from Training Data to GAN-Generated Fingerprint Images.
DOI: 10.5220/0012418100003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 4: VISAPP*, pages 435-445
ISBN: 978-989-758-679-8; ISSN: 2184-4321

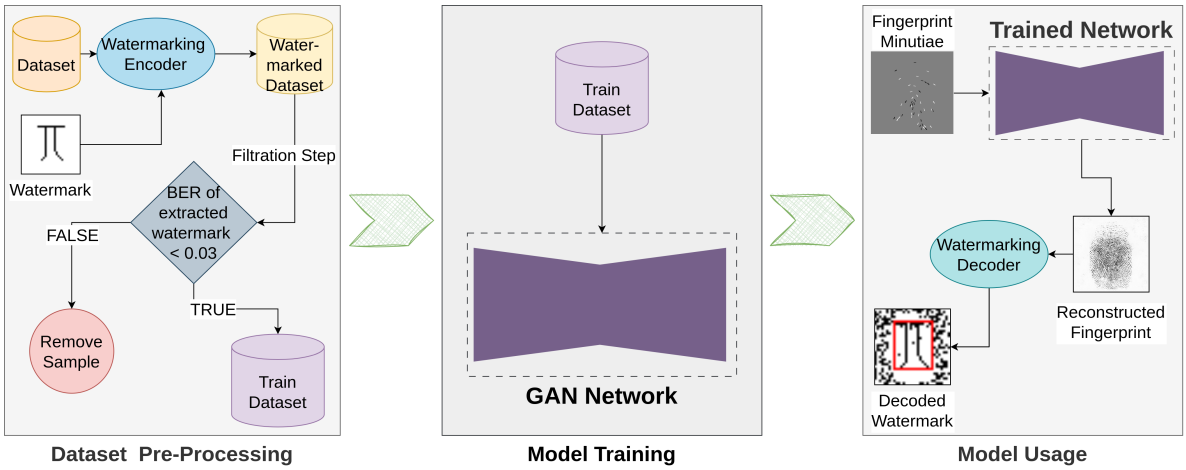

Figure 1: Overall schematic process.

algorithm. It means that our approach is applicable mainly for annotation purposes. In other words, if we have successfully extracted the watermark, we know that the fingerprint image has been produced by our model. If we cannot extract the watermark, we can say nothing about the origin of the fingerprint image. We see it also as a protection of the authors of a generative model. If the watermark has been removed by perpetrators, the sample is not authentic anymore and the model creators take no responsibility for its malicious use. The study of watermark security is subject to future research.

Note that the watermarking mechanism needs to be an integral part of the GAN model, because after sharing the model, the image generation is not controlled by model creators and there is no chance to watermark synthetic images. Hence, we propose to embed transparent and robust watermarks into training images so that after the model training the generated fingerprint images contain the same transparent watermark and are still of high utility. Figure 1 provides an overview of our experimental process.

Our evaluation addresses two objectives: assessing the retrieved watermark and evaluating the quality of the fingerprints generated by the GAN. For the former, we utilize the Peak Signal-to-Noise Ratio (PSNR) to measure imperceptibility and the Bit Error Rate (BER) to calculate the watermark's robustness. For the latter, we employ NIST Fingerprint Image Quality scores (NFIQ2) (NIST 2023) to determine the realistic appearance of the fingerprints, and Verifinger matching scores (Neurotechnology 2023) to evaluate the fingerprint reconstruction success.

Our *contributions* can be summarized as follows:

- We introduce a novel combination of a traditional watermarking algorithm based on DCT, SVD and DWT (Kang et al. 2018) and the pix2pix network for fingerprint generation (Makrushin et al. 2023) which ensures the transfer of logo watermarks from training to GAN-generated fingerprint images;
- We derive the optimal parameters of the watermarking algorithm along with optimal pix2pix hyperparameters;
- We extensively evaluate the proposed combination showing that our GAN models generate decent fingerprint images from which watermarks can be extracted.

Hereafter, the paper is structured as follows: Section 2 introduces relevant literature, followed by a detailed description of our concept and implementation in Section 3. Section 4 summarizes our experiments, Section 5 contains results and discussion and Section 6 concludes the paper offering a summary and future work.

2 RELATED WORK

2.1 Deepfake Detection

Our focus is on synthesis of fingerprint images via GANs. GANs are currently the state of the art in image generation, creating high-resolution, photorealistic images (Karras et al. 2018, Isola et al. 2017), and, most importantly, the generated images are deepfakes. These deepfakes are a threat if misused. A study reported in (Marra et al. 2019) addresses this issue by detecting GAN-generated images through the unique noise residual patterns left behind by generative models. The study in (Yu et al. 2019) introduces a neural

network classifier capable of identifying the origin of image generation. These detection approaches can be attributed as ”passive”.

2.2 Watermarks

The ease of copying, storing, or modifying data has also led to increased malicious activities. To combat such activities, watermarking techniques have been introduced, where visible or invisible information is embedded into a carrier signal. Further, watermarks are seen as additional information with which the origin is annotated. Imperceptibility, robustness, security, and recovery are the most important characteristics of watermarking. Various forms of watermarks, including text, images, audio, or video, can be embedded. Here, we focus on embedding text and images into images. Embedding a watermark into images can be done via spatial or frequency domain-based methods, each with its own advantages and drawbacks. Spatial domain techniques such as LSB, correlation-based techniques, spread spectrum techniques, and patchwork work by manipulating pixel values and bitstreams directly offer computational simplicity. Frequency domain techniques such as Discrete Wavelet Transform (DWT), Discrete Cosine Transform (DCT), Discrete Fourier Transform (DFT), Singular Value Decomposition (SVD) etc. are complex but robust against resizing or cropping (attacks). As per (Kumar et al. 2018), using a specific watermarking method may only satisfy one or two characteristics of watermarking. To address this, hybrid techniques like DCT+SVD (Tian et al. 2020) and DWT+DCT+SVD (Kang et al. 2018) have been introduced. We use a hybrid technique based on the frequency domain.

2.3 Fingerprint Synthesis via GANs

Biometrics is a field that greatly benefits from high-quality data generated by GANs. The reason is that acquiring biometric data from real persons is challenging due to high costs and privacy concerns caused by data protection regulations like the European Union (EU) General Data Protection Regulation (GDPR). Current open-source fingerprint databases are limited in quality and number of samples.

To address this, the Anguli (Ansari 2011) synthetic fingerprint generator, based on the SFinGe algorithm (Cappelli 2004), has been developed. However, the patterns generated by Anguli lack realism and therefore can be easily recognized as such. To generate more realistic synthetic fingerprints, GANs have been employed in (Bouzaglo and Keller 2022, Makrushin et al. 2023) showing their ability to create convincing synthetic fingerprints.

Fingerprint synthesis can be achieved through various approaches, including physical, statistical, or data-driven (GAN) modeling. Current statistical and physical modeling approaches tend to produce fingerprints that lack realism. They are usually visually distinguishable from real fingerprints. In contrast, GAN-based approaches usually produce realistic synthetic fingerprints. Modeling approaches can be combined by, for instance, applying CycleGAN that makes outcomes of a model-based generator appear realistic (Wyzykowski et al. 2020). Another technique in (Bahmani et al. 2021) uses StyleGAN to generate a fingerprint from a random latent vector. Also, the pix2pix network can be employed to reconstruct a fingerprint from a given minutiae template (Makrushin et al. 2023).

2.4 Watermarking Generative Models

The trend of watermarking Deep Neural Networks (DNNs) has recently gained prominence. From now on the watermarks are embedded not in media, but in functions. Given this paradigm shift and the urgent need for Intellectual Property Rights (IPR) protection in DNNs, research in (Barni et al. 2021) states the similarities, challenges, and errors to avoid in DNN watermarking in comparison to traditional watermarking. A study in (Chen et al. 2019) proposes a watermarking approach in which the watermark is directly integrated into the weights of specific layers of the network. Unlike many other DNN watermarking methods primarily focused on IPR protection, this approach also tackles the challenge of uniquely tracking users. A study in (Wu et al. 2021) employs a dual-DNN-network approach. They train a GAN model and its output is fed to another network tasked with reconstructing a predefined watermark. Their key novelty includes an objective function that calculates the watermark loss and also, a secret key that is needed to decode the watermark.

The first study addressing the transferability of artificially embedded watermarks from training images to outputs produced by GAN (Yu et al. 2021) proposes the four-step approach. The first step begins with training an encoder-decoder network. Second, the trained encoder network is utilized to embed a watermark into the training data set. Third, GAN is trained using the watermarked dataset. Fourth, the decoder is employed to extract the watermark from the GAN-created deepfakes.

Another study by (Fei et al. 2022) offers GAN

Intellectual Property Protection (IPR) using a supervised method. The methodology begins with the training of a deep learning-based image watermarking network that incorporates an imperceptible watermark into an image employing an encoder-decoder network. Following the successful training of the watermarking network, the decoder component remains fixed and is leveraged in the GAN training process to ensure the integration of the watermark within the images generated by the GAN. The novelty lies in a combined loss function comprising both the conventional GAN loss and the watermark loss. They have also introduced an image processing layer capable of performing data augmentation operations, ensuring the robustness of the embedded watermark.

In contrast to (Wu et al. 2021, Yu et al. 2021, Fei et al. 2022), our approach combines traditional watermarking with a pix2pix-based fingerprint generator, requiring no training of the watermarking part and no watermark loss function during the GAN training. The main novelty is in finding a viable combination of a GAN-based fingerprint generator and a watermarking algorithm that produces watermarks robust to GAN training.

3 OUR CONCEPT

In this research, we present a novel approach that aims to watermark the images produced by the generator of a trained GAN model using traditional digital watermarking techniques applied to GAN training images. We first embed a watermark into the training dataset with the selected digital watermarking method. Next, the data pre-processing step performs minutiae map creation from the fingerprint. Finally, training with the modified pix2pix network from (Makrushin et al. 2023) is performed. The reason for selecting pix2pix as a generative model is its ability to produce high-quality realistic fingerprint images. The main criteria for selection of the watermarking algorithm is its conformity with the pix2pix model implying that the watermark survives in a GAN training process. The repositories containing the watermarking algorithm and the GAN model code are available at https://github.com/mannam95/dct_svd_in_dwt_watermark and https://gitti.cs.uni-magdeburg.de/Andrey/gensynth-pix2pix respectively.

3.1 Watermarking Techniques

For our goal of annotating the model, we choose watermarking methods that ensure imperceptibility (making the watermarked image indistinguishable from the original) and robustness (withstanding the GAN training process).

We employ two hybrid watermarking approaches, namely DCT-SVD-in-DWT (Kang et al. 2018) and DWT-DCT-SVD (guofei 2022) to embed the watermark into the training data. The former enables the embedding of images, while the latter can embed text. During our initial studies, the text watermarking had poor results, so we focused on the DCT-SVD-in-DWT (Kang et al. 2018) method (hereafter, it is referred to as IWA). IWA involves watermark embedding and extraction. The watermark information or payload, a 2-bit 32x32 binary image, is embedded into a cover image. The binary logo size varies depending on the cover image size. For more details see (Kang et al. 2018). IWA extracts the watermark directly from the watermarked image requiring no original cover image or watermark logo/text. Note that some algorithms may require the original cover image.

In our approach, the embedding key is not given explicitly. It is rather implicit in the embedding algorithm so that the embedding key can be seen in a combination of the watermark's location, size and pattern. More precisely, our embedding key is a tuple of logo-size, logo-shape, and coordinates where the top left pixel of the logo is located. Although the watermarking algorithm is publicly known, decoding the logo without this tuple is extremely challenging. In essence, it requires an exhaustive search with all possible combinations. Randomization of the watermark location and encryption of the watermark pattern would introduce the strong security into our watermark embedding scheme. For the matter of simplicity, we currently work with a particular watermark logo at a fixed location.

At a high level, the process of watermark embedding and extraction is given by Equations 1, 2 and 3.

$$K = (coordinates, size, shape) \tag{1}$$

$$I^* = \mathrm{IWA}_{\mathrm{emb}}(I, W, K) \tag{2}$$

$$W^* = \mathrm{IWA}_{\mathrm{ext}}(I^*, K) \tag{3}$$

In Equation-1 coordinates, size, and shape are of the watermark logo. I represents the original cover image, which is a fingerprint in our case. W denotes the watermark information being embedded. The watermarked image is represented by I^*, while W^* refers to the recovered watermark. IWA_{emb}, and IWA_{ext} denote watermark embedding and extraction functions respectively. Given I and W, IWA_{emb} produces I^*. Subsequently, given a watermarked image I^*, IWA_{ext} extracts the watermark W^*.

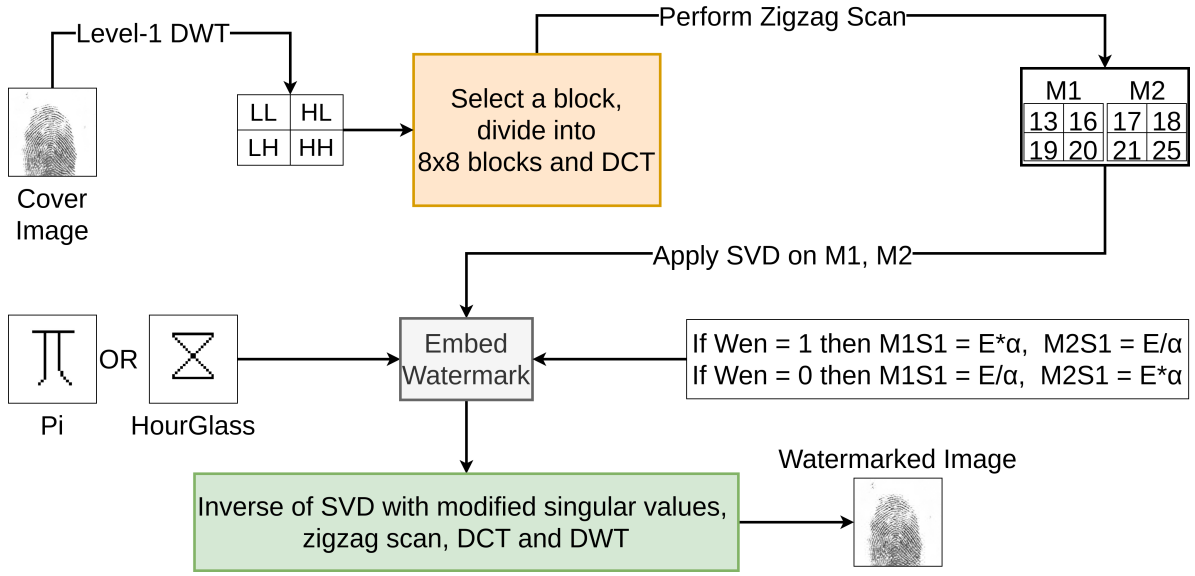

Figure 2: Watermarking process adopted from (Kang et al. 2018). "Wen" stands for the watermark image bit. "M1S1" and "M2S1" are the largest singular values of the applied SVD on the respected "M1" and "M2". "E" is the mean of the largest singular values. α is the embedding strength. Cover image is from Neurotechnology CrossMatch Dataset (Neurotechnology 2023).

The overall watermarking process is depicted in Figure 2. First, the cover image undergoes a transformation called 2D-DWT. From the resulting sub-bands (LL, HL, LH, HH), one is chosen. This selected sub-band is then divided into non-overlapping blocks of size 8x8. For each block, another transformation called 2D-DCT is applied. From the resulting coefficient matrix, 8 elements are selected based on their index using zig-zag scanning. These 8 elements are arranged in two matrices. Both matrices go through SVD, and the largest singular values are modified accordingly as shown in Figure 2. To extract the watermark, the same steps are applied. Instead of modifying the singular values, the watermark information is extracted. For more details see (Kang et al. 2018).

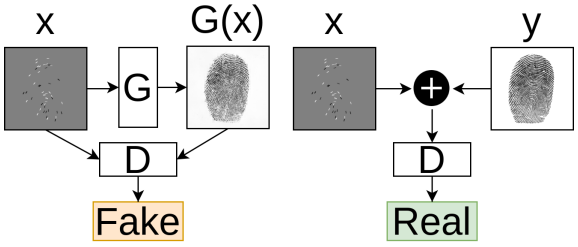

Figure 3: A high-level overview of the pix2pix architecture. G stands for generator, and D for discriminator. x is the minutiae map, y is the watermarked fingerprint, and G(x) is a fingerprint synthesized by G.

3.2 Fingerprint Generative Models

Our experimental approach utilizes the pix2pix network, specifically a conditional generative adversarial network (CGAN). The CGAN consists of two key components: a generator and a discriminator, which undergo adversarial training. On the one hand, the generator is based on U-Net architecture proposed by (Ronneberger et al. 2015) and adapted by (Isola et al. 2017), which is responsible for image-to-image translation. On the other hand, the discriminator functions as a patch-based binary classifier. The initial design of the pix2pix architecture (Isola et al. 2017) was intended for 256x256 pixel images. In our study, the fingerprint images have native resolution of 500 ppi and depicted on 515x512 pixel images. Hence, we adopt the modified version of the pix2pix network developed by (Makrushin et al. 2023). At a high level, the pix2pix network can be seen in Figure 3. For more details see (Isola et al. 2017, Makrushin et al. 2023).

The pix2pix network (Isola et al. 2017) generator requires two images: one for conditioning and the other as the true target. Here, the conditioning image is the minutiae map of the fingerprint, while the true targets are the watermarked fingerprints. To create the minutiae map, we extract minutiae using the Neurotechnology VeriFinger SDK v12.0 (Neurotechnology 2023). The extracted minutiae are then encoded into a minutiae map. The encoding methods for minutiae are directed lines and pointing minutiae as introduced in (Makrushin et al. 2023).

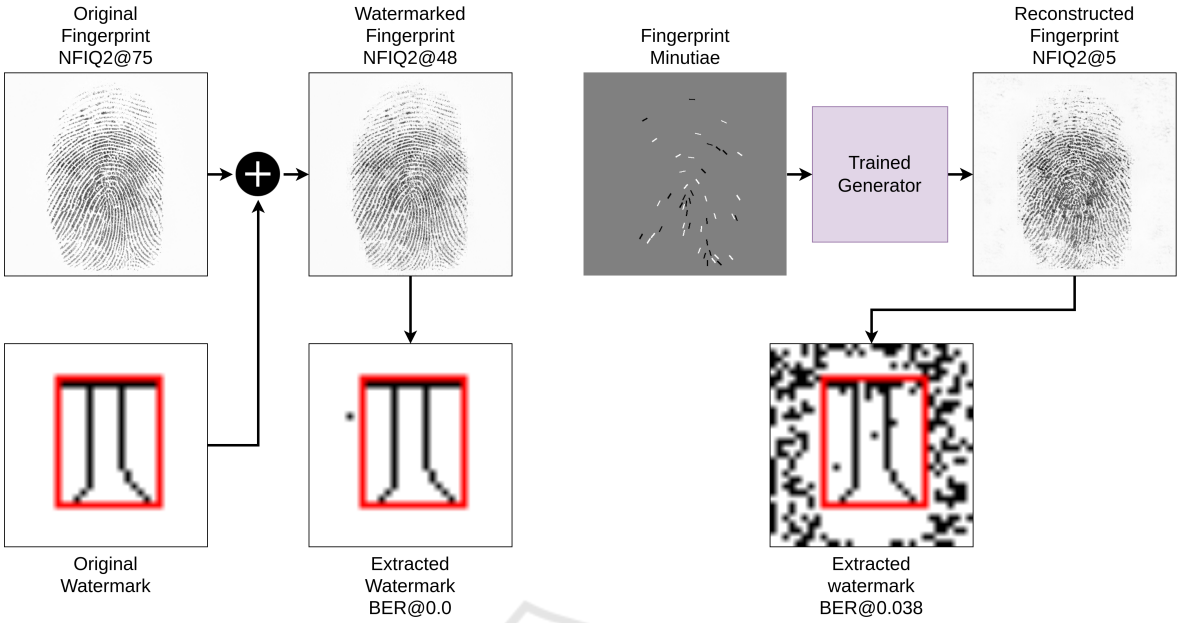

Figure 4: The process of embedding the watermark into the training dataset and extracting the watermark from the reconstructed fingerprint.

Since we focus on the transferability of watermarks and not on comparing different minutiae encodings, we simply adopt the directed lines for encoding the minutiae. The discriminator receives a concatenated tensor as input, which consists of the minutiae map supplied to the generator and the original fingerprint. The discriminator is utilized solely for training. Post-training, the generator is used alone to reconstruct fingerprints. The overall process of embedding the watermark into the training dataset and extracting the watermark from the reconstructed fingerprint is depicted in Figure 4.

4 EVALUATION

4.1 Evaluation Metrics

The watermark can be assessed via several metrics: Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM), Bit Error Rate (BER), Mean Absolute Error (MAE), and Normalized Correlation (NC). For evaluating the imperceptibility, we can use PSNR or SSIM. In contrast, the robustness of the watermarked image can be evaluated using MAE, BER, or NC. In this study, we use PSNR and BER to evaluate imperceptibility and robustness, respectively.

Following the ideas from (Makrushin et al. 2021a) we measure the realistic appearance of fingerprints by NFIQ2 scores (NIST 2023) yielding values from 0 to 100. The higher, the better utility and realism.

For fingerprint reconstruction, the True Acceptance Rate (TAR) is obtained by comparing the reconstructed and original fingerprints using the fingerprint matcher from VeriFinger SDK v12.0 (Neurotechnology 2023) which returns similarity scores from 0 to infinity. The higher, the more similar the fingerprints are. The decision thresholds are set by False Accept Rate (FAR) levels - 36, 48, and 60 for FAR levels of 0.1%, 0.01%, and 0.001%, respectively.

4.2 Training and Test Datasets

Our study utilizes a dataset of 50,000 fingerprints generated by a StyleGAN2-ada model (Karras et al. 2020) trained with 408 Neurotechnology (Neurotechnology 2023) fingerprint samples. These samples were captured using a CrossMatch Verifier 300 scanner at 500 ppi. The images were padded to 512x512 pixels prior to training. We select subsets of 2,000, 100, and 10,000 samples for training, validation, and test respectively. To ensure the diversity of dataset splits, we computed Mean Absolute Error (MAE) for all combinations, identifying a diverse range of MAE values, approximately between 30 and 230. Calculating the Verifinger scores could also help us to identify the diversity. However, we omit it in this study due to time constraints.

Further, the training dataset is embedded with the IWA watermarking algorithm and followed by a two-step filtration process. We first extract the embedded

watermark and compute the BER score between the extracted watermark logo and the original. Secondly, we pick the watermarked fingerprint samples where the BER score is less than or equal to 0.03. The reason for filtering the training set only is to include fingerprints with recoverable watermarks and ensuring their utility for our GAN training objectives. It is important to note that this filtration step is applied to the training dataset only. From this filtered dataset, 1,000 images were randomly selected for training, while the validation and test sets remained unchanged at 100 and 10,000 samples, respectively.

4.3 Experiments

We validate the proposed approach of reconstructing the watermarked fingerprints from the given minutiae maps via the following four experiments:

- Exp1: Find initial watermarking parameters.
- Exp2: Find optimal watermarking + GAN parameters.
- Exp3: Train with a different watermark logo.
- Exp4: Train with un-watermarked fingerprints.

Table 1 contains the set of parameters that are tested for the optimal performance of the watermarking algorithm IWA along with the fingerprint generative model. The optimal watermarking parameters identified in the first experiment (Exp1) include the watermark's embedding strength and the embedding level. We explore the embedding strengths of 5, 8, and 10 and assess all four wavelet subbands for embedding level: LL, HL, LH, and HH.

Table 1: Watermarking and GAN experiment configurations with training hyperparameters.

Parameter Type	Parameter Name	Selected Values
Exp1		
Watermarking	α (Alpha)	5, 8, 10
Watermarking	Embedding Level	LL, HL, LH, HH
GAN	Learning Rate	0.001
GAN	Epochs	1200
Exp2		
Watermarking	α (Alpha)	α1, α2
Watermarking	Embedding Level	EL1, EL2
GAN	Learning Rate	0.001, 0.0007
GAN	Epochs	1200, 1600, 2000
Exp3		
The best hyperparameters from **Exp2**		
Exp4		
The best hyperparameters from **Exp2**		

In our second experiment (Exp2) we select the best (in terms of highest recovery rates) two parameters for the embedding strength (α1, α2) and the embedding level (EL1, EL2). Here, we explore the GAN parameters with learning rates of 1e-3 and 7e-4 with 1200, 1600, and 2000 training epochs.

Additionally, we conduct two ablation studies in Exp3 and Exp4 to assess the impact of the watermark itself. In Exp3, we embed different watermark logo into training data. Please note that in Exp1 and Exp2, we embed "Logo-Pi" as a standard watermark, just as in Exp3 the "Logo-HourGlass". Both watermarks can be seen in Figure 2. In Exp4, the generative model is trained with raw fingerprints without any watermarks. For both Exp3 and Exp4, the best model hyperparameters are selected from the results of Exp2.

In total, we conducted 12 experiments for Exp1, trained 24 models for Exp2, and one model each for Exp3 and Exp4. All training setups utilized the Adam optimizer, batch normalization with a batch size of 64, and dropout layers are excluded. In all training runs, the learning rate linearly decays to zero after the model completes half of its training.

Notice that our GAN watermarking scheme is specified for biometric fingerprint images only. To the best of our knowledge, this is the very first study that attempts to watermark GAN-generated fingerprint images making a fair comparative study to other generic GAN watermarking approaches hardly possible.

5 RESULTS AND DISCUSSION

Table 2: Exp1 results: watermarking recovery rates. The scores (in%) represent the total number of samples out of all the test data where the "BER < 0.1". LR: learning rate, EP: epochs, α: embedding strength.

LR	EP	α	Watermarking Recovery Rates in %			
			Embedding Level			
			LL	HL	LH	HH
0.001	**1200**	5	88.99	24.78	22.12	0
		8	78.99	78.11	78.99	0.23
		10	74.60	90.25	91.86	7.47

The metric used in our **Exp1** is the Bit Error Rate (BER) between the original and recovered watermark from the pix2pix reconstructed fingerprint, specifically within the bounding box with coordinates $(x_1 = 9, y_1 = 6, x_2 = 24, y_2 = 25)$. We adopted this approach as GANs, including pix2pix, are known to generate noise around the produced images.

We consider watermark is recovered if BER is less than 0.1. This threshold is determined manually by a visual inspection. We have found that embedding the watermark in the pix2pix network training images is effective. However, recovery rates vary for each embedding level, as shown in Table-2. HH subband embedding results in near-zero recovery rates, pre-

Table 3: Exp2 results: Evaluation of watermarking and GAN parameters in 24 training configurations. EL: embedding level, LR: learning rate, EP: epochs, and "BER < 0.1" - the number of samples (in %) recovered at this threshold.

Id	**Parameters**				**TAR at FAR of**			**Avg. NFIQ2**	**Watermark Recovery Rate at BER**< 0.1	**Avg. PSNR**
	α	EL	LR	EP	0.1%	0.01%	0.001%			
1	8	HL	0.001	1200	48.64	24.89	11.27	13.70	78.11 %	31.15
2	8	HL	0.001	1600	52.31	26.58	11.46	15.47	85.69 %	31.22
3	8	HL	0.001	2000	54.64	29.57	14.79	14.75	86.44 %	31.20
4	8	HL	0.0007	1200	58.15	33.64	17.85	14.89	79.93 %	31.25
5	8	HL	0.0007	1600	58.21	34.16	17.54	15.42	86.21 %	31.22
6	8	HL	0.0007	2000	52.60	29.55	14.52	14.07	84.32 %	31.29
7	8	LH	0.001	1200	49.01	24.40	10.30	08.49	78.99 %	31.13
8	8	LH	0.001	1600	45.60	20.65	07.85	10.23	86.18 %	31.13
9	8	LH	0.001	2000	48.61	23.78	10.11	09.90	83.13 %	31.10
10	8	LH	0.0007	1200	59.45	35.58	18.62	11.61	81.51 %	31.20
11	8	LH	0.0007	1600	54.59	29.64	14.40	13.19	83.61 %	31.14
12	8	LH	0.0007	2000	59.69	35.77	19.60	11.05	82.25 %	31.20
13	10	HL	0.001	1200	52.80	27.50	12.58	12.25	90.25 %	31.23
14	10	HL	0.001	1600	50.01	26.29	14.40	12.08	**92.65 %**	31.10
15	10	HL	0.001	2000	**59.80**	**36.27**	**19.17**	10.62	89.03 %	31.20
16	10	HL	0.0007	1200	48.55	26.31	12.36	10.40	86.52 %	31.21
17	10	HL	0.0007	1600	48.22	25.64	11.65	09.63	88.60 %	31.17
18	10	HL	0.0007	2000	46.46	23.91	11.01	10.11	84.47 %	31.21
19	10	LH	0.001	1200	49.48	24.71	10.32	06.27	91.86 %	31.04
20	10	LH	0.001	1600	45.89	22.17	08.69	06.13	89.74 %	31.13
21	10	LH	0.001	2000	**63.45**	**39.29**	**20.49**	09.25	**92.92 %**	31.23
22	10	LH	0.0007	1200	51.64	26.93	12.26	07.33	87.41 %	31.12
23	10	LH	0.0007	1600	57.67	33.13	17.23	09.16	87.24 %	31.22
24	10	LH	0.0007	2000	57.21	32.68	16.45	10.87	91.21 %	31.22

sumably due to its sensitivity to filtering operations. Conversely, the LL subband shows promising recovery rates but the original image quality has degraded. The reason for that is that the LL band contains high-level image information. Changing it will directly impact the image content. HL and LH subbands have a mix of image frequencies and show high recovery rates. Thus, we select them for further experiments. Assessing the importance of embedding strength (α factor), we discovered that increasing the α improves watermark robustness, with a recovery rate exceeding 90% for both HL and LH subbands, if α is set to 10. However, the α of 8 also yields almost 80% recovery rate. Therefore, we proceed with combinations of embedding strengths 8 and 10 and embedding levels HL and LH in our further investigations.

In **Exp2**, we employ the top-performing watermarking parameters from Exp1 to refine GAN parameters across 24 training runs, as outlined in Table-3. We assess the robustness and imperceptibility of watermarked fingerprints via BER and PSNR scores, with models 14 and 21 demonstrating superior robustness with the α value of 10. The watermarks are imperceptible enough as all models exceed the acceptable PSNR threshold of 30dB. The poor visual quality and low reconstruction rates measured by NFIQ2 and Verifinger matching scores respectively reveal that our approach has a room for improvement. Fingerprints of good quality are represented by NFIQ2 scores exceeding 35, whereas scores under 6 suggest ineffective patterns. Our average NFIQ2 scores lie around 10, suggesting that the visual quality of the fingerprints is poor. Fingerprint reconstruction rates reported in (Makrushin et al. 2023) are over 80%, 70%, and 60% at FAR levels of 0.1%, 0.01%, and 0.001% respectively. Our models demonstrate significantly lower reconstruction rates, with a maximum of 63.45% and 59.80% at 0.1% FAR, using models 15 and 21 both with the α value of 10.

Table 4: Exp3 results; The "Model Id" column corresponds to the configurations ("Id") reported in Table 3.

Model Id	**Logo**	**Avg BER**
21	Pi	0.044
21	HourGlass	0.052

The results of **Exp3** are reported in Table-4. We see that watermarking capacity affects watermarking robustness. On average, the model trained with "Logo-Pi" embedded data outperforms the one trained with "Logo-HourGlass" embedded data in terms of BER scores.

The results of **Exp4** are visualized in Figure 6. The original un-watermarked data has an average NFIQ2 score of around 75. The models trained with

Figure 5: Visualization of a fingerprint image with and without a watermark. The red dots in the figure denote the locations of minutiae.

"Logo-Pi", "Logo-HourGlass", and un-watermarked data have the average NFIQ2 scores of approximately 10, 5, and 20, respectively. We see that watermarking has an impact on the visual quality of reconstructed fingerprints, but the highest degradation of fingerprints is due to the reconstruction process, as the model trained with un-watermarked data achieves an average NFIQ2 score of only 20, compared to the original raw scores of around 75. We suspect that the GAN parameters obtained may not be optimal and require further exploration.

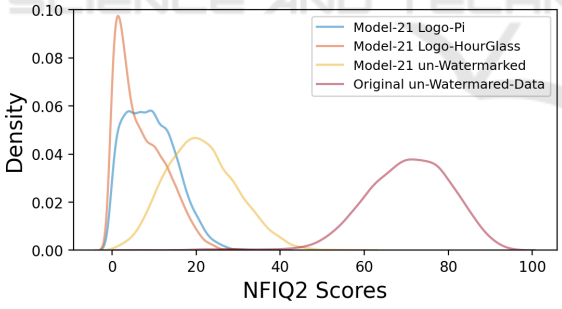

Figure 6: Distributions of NFIQ2 scores plotted for Exp3 and Exp4: Comparison of watermarks to each other and to the training with un-watermarked data.

All in all, the configurations with the embedding strength of 10, embedding levels HL or LH, the learning rate of 1e-3, and 1600 or 2000 epochs demonstrate better performance than the remaining configurations. The low fingerprint reconstruction scores could be attributed to GAN noise or suboptimal GAN parameters. A visual sample result can be seen in Figure 5.

6 CONCLUSION

Watermarking of synthetic images produced by a generative model is an important step towards protecting the creators of generative models. This paper introduces a novel approach to watermarking the training images with DCT-SVD-in-DWT (Kang et al. 2018) and training the pix2pix network with these watermarked images. Our primary goal is to create realistic synthetic fingerprints and protect the GAN authors by watermarking the model-generated images. We achieve the former one using the pix2pix model and the latter by transferring the watermark from the training dataset to the model's generated images. We experiment with various parameters of the selected watermarking technique in conjunction with training hyperparameters of GAN. For watermarking, an embedding strength of 10 results in superior outcomes, primarily when the embedding level is either HL or LH. Even though the optimal watermarking and GAN parameters enable watermark extraction from the vast majority of the reconstructed fingerprints so that the fingerprints do not loose their utility, there is a room for algorithm tuning to improve the NFIQ2 and Verifinger matching scores. The watermark recovery rate in our experiments is not high enough to consider our GAN watermarking scheme mature for application in a practical scenario. Notice that fingerprint images, due to their limited information content with black and white lines, have a very low watermark capacity. Applying our approach to colored more informative images which accommodate more watermark capacity may lead to significantly higher watermark recovery rates. An adversarial attack might find the minutiae setups that lead to vanishing watermarks in synthetic fingerprint images. Hence, improvement of the

practical effectiveness and robustness of our approach is the subject to future work. All in all, given a robust watermarking algorithm, we confirm that the watermarks can be transferred from the GAN training images to the GAN-generated images. Future work will include a thorough study of watermark transferability across various generative network architectures and extend the study to other domains like video. Furthermore, the embedding of encrypted watermarks will be studied to address the security aspect.

ACKNOWLEDGEMENTS

This research has been funded in part by the Deutsche Forschungsgemeinschaft (DFG) through the research project GENSYNTH under the number 421860227.

REFERENCES

Ansari, A. H. (2011). *Generation and storage of large synthetic fingerprint database*. M.E. Thesis, Indian Institute of Science Bangalore.

Bahmani, K., Plesh, R., Johnson, P., Schuckers, S., and Swyka, T. (2021). High fidelity fingerprint generation: Quality, uniqueness, and privacy. In *Proc. of the IEEE Int. Conf. on Image Processing (ICIP)*. IEEE.

Barni, M., Pérez-González, F., and Tondi, B. (2021). DNN watermarking: Four challenges and a funeral. In *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec '21, page 189–196, New York, NY, USA. Association for Computing Machinery.

Bontrager, P., Roy, A., Togelius, J., Memon, N., and Ross, A. (2018). DeepMasterPrints: Generating masterprints for dictionary attacks via latent variable evolution. In *Proc. BTAS*, pages 1–9.

Bouzaglo, R. and Keller, Y. (2022). Synthesis and reconstruction of fingerprints using generative adversarial networks. *CoRR*, abs/2201.06164.

Cappelli, R. (2004). SFinGe: an approach to synthetic fingerprint generation. In *Proc. of the Int. Workshop on Biometric Technologies*.

Chen, H., Rouhani, B. D., Fu, C., Zhao, J., and Koushanfar, F. (2019). DeepMarks: A secure fingerprinting framework for digital rights management of deep learning models. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, ICMR '19, page 105–113, New York, NY, USA. Association for Computing Machinery.

Edwards, B. (2022). China bans AI-generated media without watermarks. https://arstechnica.com/information-technology/2022/12/china-bans-ai-generated-media-without-watermarks/, last check 14.7.2023.

Farou, Z., Mouhoub, N., and Horváth, T. (2020). Data generation using gene expression generator. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12490:54–65.

Fei, J., Xia, Z., Tondi, B., and Barni, M. (2022). Supervised GAN watermarking for intellectual property protection. In *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6.

Goodfellow et al., I. (2014). Generative adversarial nets. In Ghahramani et al., Z., editor, *Advances in Neural Information Processing Systems (NIPS'14)*, volume 27, pages 2672–2680. Curran Associates, Inc.

guofei (2022). Blind watermark based on DWT-DCT-SVD. https://github.com/guofei9987/blind_watermark, last check 14.7.2023.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proc. CVPR*.

Kang, X., Zhao, F., Lin, G., and Chen, Y. (2018). A novel hybrid of DCT and SVD in DWT domain for robust and invisible blind image watermarking with optimal embedding strength. *Multimedia Tools and Applications*, 77:13197–13224.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. In *Proc. of the International Conference on Learning Representations (ICLR)*.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020). Training generative adversarial networks with limited data. *CoRR*, abs/2006.06676.

Kumar, C., Singh, A. K., and Kumar, P. (2018). A recent survey on image watermarking techniques and its application in e-governance. *Multimedia Tools and Applications*, 77:3597–3622.

Makrushin, A., Kauba, C., Kirchgasser, S., Seidlitz, S., Kraetzer, C., Uhl, A., and Dittmann, J. (2021a). General requirements on synthetic fingerprint images for biometric authentication and forensic investigations. In *Proc. IH&MMSec'21*, page 93–104. ACM.

Makrushin, A., Mannam, V. S., and Dittmann, J. (2023). Data-driven fingerprint reconstruction from minutiae based on real and synthetic training data. In *Proc. VISIGRAPP 2023 - Volume 4: VISAPP*, pages 229–237.

Makrushin, A., Trebeljahr, M., Seidlitz, S., and Dittmann, J. (2021b). On feasibility of GAN-based fingerprint morphing. In *Proc. of the IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6.

Marra et al., F. (2019). Do GANs leave artificial fingerprints? In *Proc. of the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511. IEEE.

Neurotechnology (2023). Neurotechnology Verifinger SDK. https://www.neurotechnology.com/verifinger.html, last check 14.7.2023.

NIST (2023). NIST Fingerprint Image Quality (NFIQ) 2. https://www.nist.gov/services-resources/software/nfiq-2, last check 14.7.2023.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Tian, C., Wen, R. H., Zou, W. P., and Gong, L. H. (2020). Robust and blind watermarking algorithm based on DCT and SVD in the contourlet domain. *Multimedia Tools and Appl.*, 79:7515–7541.

Wu, H., Liu, G., Yao, Y., and Zhang, X. (2021). Watermarking neural networks with watermarked images. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7):2591–2601.

Wyzykowski, A. B. V., Segundo, M. P., and de Paula Lemes, R. (2020). Level three synthetic fingerprint generation.

Yu, N., Davis, L., and Fritz, M. (2019). Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *Proc. of the IEEE/CVF Int. Conference on Computer Vision (ICCV)*, pages 7555–7565.

Yu, N., Skripniuk, V., Abdelnabi, S., and Fritz, M. (2021). Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proc. of the IEEE Int. Conference on Computer Vision (ICCV)*.