Attributing and Detecting Fake Images Generated by Known GANs

Matthew Joslin and Shuang Hao

University of Texas at Dallas
{matthew.joslin, shao}@utdallas.edu

***Abstract*—The quality of GAN-generated fake images has improved significantly, and recent GAN approaches, such as StyleGAN, achieve near indistinguishability from real images for the naked eye. As a result, adversaries are attracted to using GAN-generated fake images for disinformation campaigns and fraud on social networks. However, training an image generation network to produce realistic-looking samples remains a time-consuming and difficult problem, so adversaries are more likely to use published GAN models to generate fake images.**

In this paper, we analyze the frequency domain to attribute and detect fake images generated by a known GAN model. We derive a similarity metric on the frequency domain and develop a new approach for GAN image attribution. We conduct experiments on four trained GAN models and two real image datasets. Our results show high attribution accuracy against real images and those from other GAN models. We further analyze our method under evasion attempts and find the frequency-based approach is comparatively robust.

I. Introduction

Since the introduction of Generative Adversarial Networks (GANs) which can be used to generate superficially realistic images, the quality of these synthetic images has progressively improved. Although early attempts at fake image generation left obvious visual cues [1, 2], recent GAN approaches, such as StyleGAN [3], achieve near indistinguishability from real images for the naked eye. Consequently cybercriminals have begun utilizing GAN-generated fakes to spread disinformation and masquerade on social networks, such as establishing fake accounts [4, 5]. However, training an image generation network, such as StyleGAN, remains a time-consuming and difficult task. Therefore, cybercriminals may prefer to use pre-trained GAN models available online to generate fake pictures. For example, training StyleGAN for high-quality samples requires almost a week with eight state-of-the-art GPUs on a $100k server. Besides the cost required, training an effective GAN model also demands a high-level of expertise, e.g., dealing with mode collapse, stability between the discriminator and the generator, and hyperparameter settings.

We consider the problem of how to automatically estimate whether arbitrary images were created by a known GAN model, given the GAN model itself or a small set of output samples. The goal is to distinguish from real images or images produced by other GANs (whose models we do not need to know). Though attributing and detecting GAN images have recently drawn increasing research attention, previous studies mostly rely on legacy artifacts of early GANs [6–8] or training black-box classifiers [9–11]. However, state-of-the-art GANs do not contain obvious legacy visual artifacts, and typically the training process demands a large set of GAN image samples (e.g., 100,000) [9, 11], which is not feasible to use in practice.

In this paper, we design a novel approach to efficiently detect and attribute fake images generated by a known GAN model. We analyze the frequency spectrum of images, where each frequency point is described by a magnitude and a phase. The frequency domain provides a new perspective to analyze GAN images, compared to the hardly distinguishable pixel domain. We then derive a similarity metric to measure the correlation between the frequency domains of images. By averaging the frequency samples from a known GAN model, we obtain the fingerprint introduced by the GAN model. Our approach requires a comparatively small amount of samples (1,000 or less). We performed experiments on four trained GAN models, based on three different GAN architectures (including ProGAN [12], StyleGAN [3], and StarGAN [13]) and two real facial datasets (including CelebA [14] and FFHQ [3]). Our results demonstrate high accuracy for distinguishing images generated by a known GAN model from real images or images produced by other GAN models. We examined evasion effects of common image manipulations, including noise addition, blurring, and JPEG compression. Our approach remain robust against such evasion attempts.

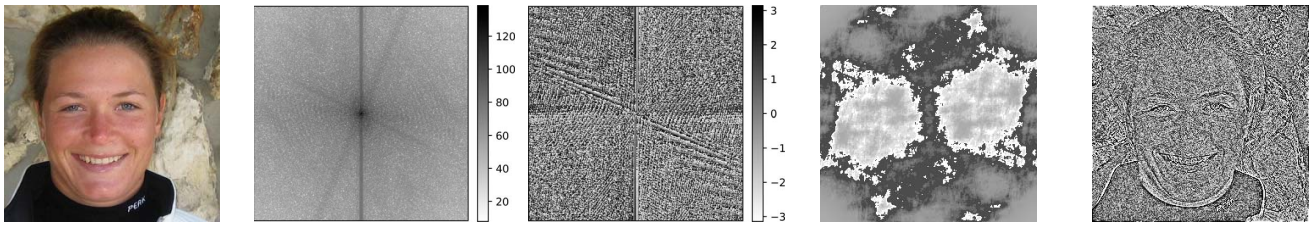
To summarize, our work makes the following contributions.

- We analyze the frequency spectrum of images to provide a new perspective towards attributing synthesized images to a known GAN model. We also develop a similarity metric to measure the correlation on the image frequency domain.
- We design a novel approach based on the frequency spectrum to estimate the likelihood that images were generated by a known GAN model. Our approach needs a relatively small amount of samples to build the fingerprint of GAN models.
- We perform experiments on state-of-the-art GAN models. The results show that our approach achieves high accuracy on image attribution (against real images and images generated by other GAN models) and is robust against evasion manipulations.

II. Methodology

Our heuristic is that GAN models embed unique fingerprints to the generated images, which we can leverage to distinguish

DOI 10.1109/SPW50608.2020.00019



(a) CelebA original image (b) Magnitude (dB) (c) Phase (radians) (d) Reconstructed image only from magnitude (e) Reconstructed image only from phase

Figure 1: Example real image (from CelebA), magnitude and phase, and reconstruction from only the magnitude and phase. The subfigure (e) is the reconstruction based only on the phase, which preserves the main information of the original image.

from both real images and the images produced by other GAN models. We analyze the frequency domain of images, which allows for more explicit interpretation of the signals and increase the robustness against evasion attempts. We then derive a metric to measure the similarity in the frequency domain, and develop an efficient approach to attribute synthetic images to a known GAN model.

Frequency domain interpretation. The frequency domain of an image describes the image in the spatial domain as the sum of sine functions. Mathematically, the frequency domain can be represented with complex numbers for arithmetic operations. The most common method for converting an image into the frequency domain is the discrete Fourier transform (DFT). Although the naive DFT algorithm suggests a complexity of $O(n^2)$ (where n is the number of pixels), the fast Fourier transform (FFT) algorithm requires only $O(n \log(n))$ operations. As a result, performing analysis in the frequency domain can be efficient.

Prior work has suggested that GANs may unintentionally add fingerprints to the output [15, 16]. However, these studies relied on legacy visual artifacts, which are greatly diminished in the state-of-the-art GANs. We analyze the frequency domain in our work, which provides a different perspective for interpretation. For example, low frequency components define the image's overall structure (such as shape and location of a face), while the high frequency portion describes the details in an image (such as hair, eyes, and teeth). In addition, each frequency component can be described by a magnitude and a phase. The magnitude determines how large of an impact the component has on the final output, and the phase describes where in the image the component will have its maximum value. Prior work has shown that phase information is more important to construct an image compared to the magnitude information [17, 18]. To demonstrate the effect of the magnitude and phase, Figure 1 shows an example image in the pixel domain, the magnitude and phase in the frequency domain, and the reconstructed images based on the magnitude and phase respectively. The reconstruction of Figure 1(d) is the inverse FFT on the magnitude of the original image (Figure 1(b)) but fixed phase of 90°. Figure 1(e) is the inverse FFT on the phase of the original image (Figure 1(c)) but fixed magnitude of 1. To increase readability, we apply contrast enhancement on the final output for both cases. Noticeably, the phase reconstruction of Figure 1(e) resembles the original image more (Figure 1(a)). This observation leads us to focus our analysis on the phase information.

Derivation of similarity metrics. We then derive a metric to assess the similarity in the two-dimensional frequency domain. Traditional Euclidean distance-based metrics do not provide a method for normalizing the similarity. We adapt the cosine similarity as our basic metric. For two frequency components A and B (with the same frequency, but possibly different magnitudes and phases), we represent them in the complex number formats, calculate the product between the two complex numbers (with B converted to the complex conjugate), and take the real part normalized by the magnitude of both components, as shown in Equation 1. The value range is between $[-1, 1]$, and the cosine similarity is commutative.

$$\cos(A, B) = \frac{\text{Re}(A \cdot \overline{B})}{\|A\| \, \|B\|} \tag{1}$$

The geometric interpretation of the similarity is to measure how the two vectors in the complex number space align, which corresponds to the phase difference of the two signals. As discussed above, phase information is more important to construct an image compared to the magnitude information. To measure the aggregate similarity between two images, we calculate the average of the cosine similarities across different points (corresponding to different frequency components) in the frequency domain. Suppose the frequency domains of two images are F_A and F_B. The similarity metric is calculated as in Equation 2, where N is the number of points in the frequency domain, and F_A^i and F_B^i are the frequency components (which can be represented as complex numbers) at the point i.

$$\text{Sim}(F_A, F_B) = \frac{\sum_{i=1}^{N} \cos(F_A^i, F_B^i)}{N} \tag{2}$$

Attributing to a GAN model. Given an established GAN model, we develop an attribution approach to estimate the likelihood that images were generated by this particular model. In our scenario, the required information is a target GAN model, and we can acquire synthetic image samples, e.g., by executing the generator or being provided with generated image samples. For the new images produced by this target

GAN model, we expect our system will give high similarity scores. On the other hand, for real images or images generated by other GAN models, our system will provide low similarity scores. We use the average of the synthesized image samples as the model signature and the metric derived above to measure the similarity on the frequency domain. Figure 2 shows the workflow of our approach. Note that a color image contains three channels (RGB). The detailed steps to process color images are described below.

1) We obtain a set of sample images synthesized by the GAN model (e.g., 1,000 images). We can execute the GAN model to generate random samples, or the samples can be provided by a third party (if the GAN model is not directly accessible). We perform FFT on each RGB channel of these synthetic images to get the two-dimensional frequency domain outputs.
2) We take the point-wise average of the frequency domain for each RGB channel across the sampled GAN-synthesized images. The per-channel frequency average is used as the signature of the GAN model.
3) When testing an image (to estimate whether the image is generated by this GAN model), we apply FFT to convert each channel of the image into the two-dimensional frequency domain.
4) We calculate the similarity between the frequency domain of the test image and the signature of the GAN model, by using Equation 2 on each RGB channel. We further take the average across the three channels to get the final similarity score. A higher similarity score indicates the image is more likely to be generated by this GAN model.

Our approach has high computational efficiency. The model signature can be extracted based on a comparatively small amount of image samples. In our experiments in Section III, 1,024 GAN-model images are sufficient to form the model signature. On the other hand, prior approaches that relied on training a CNN used a large set of samples (typically 100,000) [9, 11]. We use the FFT algorithm to convert images to the frequency domain, which make the image processing efficient as mentioned above.

III. Experiments and Results

Next we describe the GAN models and datasets in our experiments, present the attribution performance to a known GAN model, explore the effect of possible evasions, and investigate the effect of different frequency components.

Experiment setup. We experiment on three GAN architectures, ProGAN [12], StyleGAN [3], and StarGAN [13]. Except for StarGAN which generates 256×256 resolution images, the resolutions of the images produced by other GAN models are 1024×1024. These GANs represent a challenging problem for attribution as they remain near the state of the art. In fact, to the best of our knowledge, no prior work has attempted to detect StyleGAN generated fakes and even ProGAN has received relatively little attention. In addition, the selected GANs generate images of human portraits, making them appealing to cybercriminals to abuse in social engineering attacks.

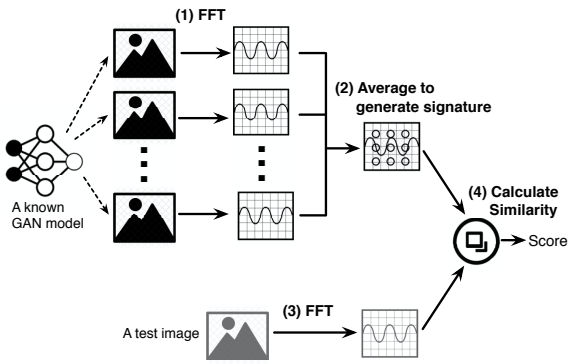


Figure 2: Our approach for determining whether an image is generated by a known GAN model.

We use the popular CelebA dataset and the relatively new FFHQ dataset as real images. The CelebA dataset provides a common point of comparison as all the GANs that we test have models trained on CelebA. To improve image generation performance, the FFHQ dataset was collected from the Flickr website and released recently. We have an additional StyleGAN model trained on the FFHQ dataset. Presumably the CelebA and FFHQ images should look similar to the GAN images, since the data was used to train the GAN models, which makes the image attribution more challenging. Our evaluation targets the rigorous cases, and all of the images are faces. For the model naming convention, we use the combination of the GAN architecture and the training dataset to refer to the GAN model, e.g., "StyleGAN CelebA" indicates the model using the StyleGAN architecture and trained on CelebA. Figure 3 shows samples from the two real datasets and all four GAN models. Note that the GAN generated samples appear impressively realistic.

We generate 1,024 samples for each GAN to calculate the model signature as described in Section II. To evaluate the performance of our method, we randomly select 1,000 samples from each GAN (distinct from those used to generate the signature) and 1,000 real images from CelebA and FFHQ respectively. To run our experiments, we run all the GANs on a NVIDIA Titan RTX GPU with 24 GB RAM running CUDA version 10.2, and use a server with 32 CPU cores and 196 GB of RAM to evaluate our method. To compare images with differing resolutions (for example ProGAN to StarGAN), we resize the image to the resolution of the target GAN.

Attribution results. We first use the scenario where the known GAN model is "StyleGAN CelebA" to demonstrate the potential of our method for GAN image attribution. The test datasets are the two real image datasets (CelebA and FFHQ) and the fake images generated by different GAN models. Figure 4 shows the distributions of the correlation with the "StyleGAN CelebA" signature. Each curve indicates a given test dataset with the "StyleGAN CelebA" signature. The x-axis

(a) Real CelebA (b) Real FFHQ (c) StyleGAN CelebA (d) StyleGAN FFHQ (e) ProGAN CelebA (f) StarGAN CelebA

Figure 3: Example images from real datasets and GAN datasets.

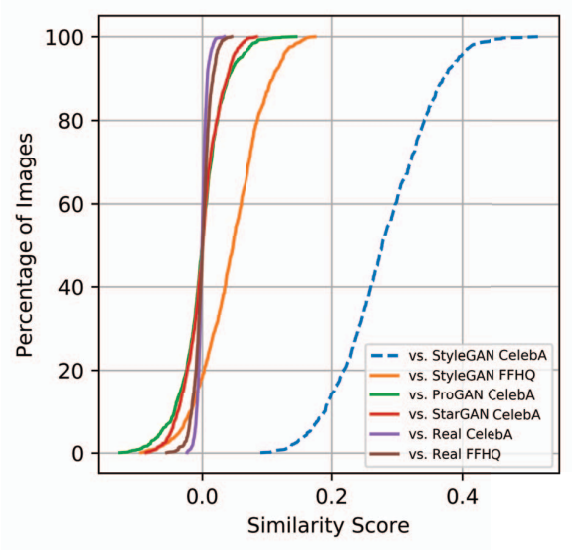


Figure 4: CDF of the similarity scores between the "StyleGAN CelebA" signature and all datasets. The x-axis is the similarity score and the y-axis denotes the CDF for each dataset. The dashed curve indicates the similarity between the "StyleGAN CelebA" testing samples and the "StyleGAN CelebA" signature which is clearly separated from the other curves. The observation means that the "StyleGAN CelebA" test set can be correctly attributed to the model.

shows the similarity score and the y-axis means the percentage of images in each dataset. The CDFs show a clean separation between "StyleGAN CelebA" samples (the dashed curve) and real images and only slight overlap with samples generated by different GAN models. These results provide multiple observations: (1) Our method can successfully distinguish from not only unrelated real, face images (FFHQ), but also real images (CelebA) which are related to the data for training "StyleGAN CelebA" model. (2) Samples from other GAN models ("StarGAN CelebA" and "ProGAN CelebA") are well separated from the known "StyleGAN CelebA" model, even though they share the same training dataset. (3) For the GAN model with the same architecture e.g., "StyleGAN FFHQ", we still observe clear separation between "StyleGAN CelebA" and "StyleGAN FFHQ". The findings indicate that our method achieves strong capability to attribute images to the correct GAN model. We observe similar trends when attributing to other GAN models, across "StyleGAN FFHQ", "ProGAN CelebA", and "StarGAN CelebA".

Next we evaluate the overall performance with the area under the curve (AUC) statistic which is the area under the receiver operating characteristic curve (ROC). The AUC corresponds to the likelihood that our method will produce a higher correlation for a positive sample (from the same GAN as the signature) than a negative sample (a real image or a sample from another GAN). For example, an AUC of 1.0 gives perfect attribution, while an AUC of 0.5 is equivalent to random guessing. Table I shows all the results for each dataset and signature combination. The row heading gives the signature of the known GAN model used to compute the correlation score and the column heading shows the source dataset being distinguished from the samples generated by the known GAN model. For example, row "StyleGAN CelebA" and column "ProGAN CelebA" means that the "StyleGAN CelebA" signature is used to attribute between "StyleGAN CelebA" images and "ProGAN CelebA" images. The symbol "—" indicates pairs where the samples from the known GAN model would be compared with themselves.

Overall, we find that our method achieves high performance in attributing images to the GAN which generated them. Interestingly, "StyleGAN CelebA" and "StyleGAN FFHQ" exhibit high detection rates for both real images and images generated by other GANs. Although StyleGAN produces the most photo-realistic samples in the pixel domain, which are difficult to identify by human eyes, the samples show obvious patterns in the frequency domain. The results for StyleGAN also indicate that for networks with the same architecture but different training sets can still yield correct attribution (between "StyleGAN CelebA" and "StyleGAN FFHQ"). We also find that our method can distinguish between disparate architectures. For example, "StarGAN CelebA" uses a considerably different type of architecture (image-to-image translation changing a given attribute), taking in real images as input instead of random initial vectors as other common architectures use, e.g., StyleGAN and ProGAN. We use all types of attribute changes in StarGAN for a more diverse dataset, although prior work focused on detecting a single type of attribute change (such as gender) [16]. Even though StarGAN uses real images as seeds, our method is still capable of distinguishing StarGAN samples from other types of images.

Evasion analysis. To evaluate the robustness of our detection approach against evasion, we apply three common evasion

Known Model \ *Comparison Dataset*	StarGAN CelebA	ProGAN CelebA	StyleGAN CelebA	StyleGAN FFHQ	Real CelebA	Real FFHQ
StarGAN CelebA	—	0.937	0.934	0.950	0.956	0.947
ProGAN CelebA	0.919	—	0.973	0.950	0.953	0.947
StyleGAN CelebA	1.000	1.000	—	0.999	1.000	1.000
StyleGAN FFHQ	1.000	1.000	0.997	—	1.000	1.000

Table I: Attribution results with the area under the curve (AUC) for distinguishing between a set of GAN images and another dataset. The row title indicates the GAN signature used for attribution and the column title denotes the source dataset being distinguished between the GAN generated samples. For example, row "StyleGAN CelebA" and column "ProGAN CelebA" means that the "StyleGAN CelebA" signature is used to attribute between "StyleGAN CelebA" images and "ProGAN CelebA" images. The missing values "—" indicate pairs where the GAN samples would be compared with themselves.

techniques, adding noise, blurring, and JPEG compression. For the noise attack, we add Gaussian noise to each pixel in an image for an average change of 10% per pixel. When evaluating the blur attack, we use a uniform filter with size 4 x 4. Finally, for JPEG compression we set the quality level to 10% (obvious visual effects). To illustrate how much the evasion attacks affect the perceptual quality of the GAN generated images, we include an example of each evasion and the original GAN image in Figure 5. For each evasion, the image shows strong visual changes.

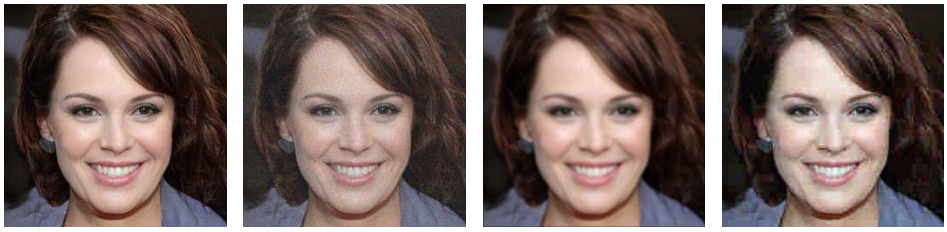
We evaluate each evasion with 1,000 real and fake images per dataset. Table II shows the AUC for distinguishing between a GAN image after evasion and the corresponding real images, e.g., between "StyleGAN FFHQ" and real FFHQ images or between "StyleGAN CelebA" and real CelebA images. The rightmost column corresponds to the case with no evasion (from Table I). While we do observe degraded performance for most of the evasions, our method proves fairly robust against these evasions, as compared to prior work which experienced dramatically decreased accuracy without proper retraining [10, 19]. In particular, we observe that JPEG compression affects attribution the least for StyleGAN. In fact, "StyleGAN FFHQ" experiences almost no degradation in performance. The robustness of our method to JPEG compression suggests that JPEG compression preserves the phase for the most important frequency domain components. However "ProGAN CelebA" and "StarGAN CelebA" both exhibit reduced performance indicating that the GAN signatures for these GANs may be more concentrated in high frequency components. Both blur and noise affect performance more significantly for all of the GANs under test. Noise and blurring effectively modify the high frequency components of the spectrum more strongly than lower frequencies. We hypothesize that the lowered performance may stem from the fact that part of the unique signature for a given GAN is in the high frequency.

Frequency component analysis. We conduct further experiments to understand the effect of different frequency components on our method's performance. First, we separate the high and low frequency regions in the spectrum. We use the highest and lowest 25% of the frequency domain respectively to calculate the similarity. We observe mixed effects when using partial spectrums for attribution. For example, if only the high frequency components are used, the "StarGAN CelebA" signature sees improved performance of 0.949 when distinguishing between "StarGAN CelebA" samples and real FFHQ images, but lowered performance of 0.916 with only the low frequency spectrum. Meanwhile, the "ProGAN CelebA" signature shows degraded performance of 0.824 for high frequency components and enhanced performance of 0.965 for low frequencies when comparing "ProGAN CelebA" images and real FFHQ images. These results indicate that different signatures are concentrated in different parts of the spectrum.

Known Model	Noise	Blur	JPEG	No Evasion
StarGAN CelebA	0.814	0.837	0.804	0.956
ProGAN CelebA	0.819	0.861	0.853	0.953
StyleGAN CelebA	0.835	0.867	0.991	1.000
StyleGAN FFHQ	0.825	0.911	0.998	1.000

Table II: Attribution results with AUC after evasions of adding noise, blurring, and JPEG compression. Each signature from the known GAN model is compared with the corresponding samples from the known GAN and the real images which provided the training data for the known GAN model. For example, the "StyleGAN CelebA" signature is compared with "StyleGAN CelebA" samples and real CelebA images. The last column is the original results with no evasion taken from Table I.

Second, to estimate which specific frequency components most affect attribution, we rank and select the top frequency components with the highest correlation to perform attribution. To rank the frequency components, we use an additional dataset of 1,000 samples from each GAN and compute the average correlation (with the corresponding GAN signature) for each frequency component. We then select and use 25% of the frequency components with the highest average correlation to perform attribution. The results generally show an improvement when compared to using all of the frequency components for attribution. On the other hand, we also test using the top 25% of components under evasion, and we observe a slight degradation in performance. For example, the "StarGAN CelebA" signature increases the AUC from 0.956 to 0.975 when comparing with real CelebA images, but the AUC falls from 0.837 to 0.822 when the "StarGAN CelebA" samples are blurred. The reduced performance indicates that using the entire spectrum helps under evasion since it uses the maximum amount of available information. In practice, using only part of



(a) Original from "StyleGAN CelebA" (b) Noise (c) Blur (d) JPEG compression

Figure 5: "StyleGAN CelebA" image before and after the noise, blurring, and JPEG compression evasion attacks. For the noise attack we add Gaussian noise to each pixel for an average change of 10% per pixel, the blurring attack uses a uniform filter with size 4x4, and the JPEG compression attack degrades the quality to 10% of the original image. Note that all of the evasion settings we use produce visual degradation compared to the original image.

the frequency domain seems to offer tradeoffs in performance and evasion resistance. Operators can choose the top frequency components as the signature to reduce computational costs.

IV. Related Work

GANs [2] have steadily improved in their ability to generate realistic fake images either from random seeds [1, 20–22] or from real images [23–25]. In this paper, we examine both unconditional GANs (ProGAN and StyleGAN) and image-to-image translation GANs (StarGAN). Both ProGAN [12] and StyleGAN [3] follow a similar structure, and generate an image from a high-dimensional random vector through repeated convolutions, upsampling, and non-linearity. The key innovation of ProGAN is to progressively increase the resolution by training additional layers. StyleGAN builds upon ProGAN's success by modulating the intermediate output with a style vector and introducing noise at each layer. In contrast to ProGAN and StyleGAN which generate fake images from scratch, StarGAN [13] takes in a real image and transforms it to fake image with a given attribute changed. In particular, StyleGAN produces highly realistic images which often cannot be distinguished from real images.

Detecting and attributing GAN images has begun to receive growing attention in recent years. Previous work has suggested that GANs include patterns in the output, such as upsampling patterns [15] or inconsistent colors [7, 8]. Marra et al. [16] adapted the photo-response non-uniformity (PRNU) method from image forensics, but they used the pixel domain and captured only the noise patterns in the image. Yang et al. [6] detected GAN-generated faces by using inconsistencies in the facial features. Their approach does not achieve high performance and cannot detect GAN generated faces with consistent facial features such as those produced by StyleGAN. Albright and McCloskey [26] considered the problem of attributing an image to a known model by inverting the output through the network. However, inverting the output requires the entire network rather than simply samples from the network. Another category of approaches focuses on training black-box classifiers. Yu et al. [19] trained a deep convolutional network to analyze the visual fingerprints. The approach requires a large number of samples and experiences poor performance under evasion attacks (without retraining under the correct evasion settings). Though Zhang et al. [10] used the frequency domain for detecting GAN images, they train a classifier based on the legacy upsampling artifacts in the magnitudes and only support images generated by image-to-image GANs (e.g., StarGAN, but not ProGAN and StyleGAN). Hsu et al. [11] used the contrastive loss between a set of real and fake images to train a classifier. Nataraj et al. [27] trained a classifier on the co-occurrence matrix to detect samples from image-to-image translation networks. In contrast, our method analyzes the frequency domain which more effectively captures the GAN signature, requires a relatively small number of samples, and does not use black-box classification.

V. Conclusion

In this paper, we develop a frequency-based approach to attribute and detect images generated by a known GAN model. We provide a similarity metric on the frequency domain to compare two spectrums, design an approach to compute a GAN's signature, and use them to distinguish between GAN images from other GANs and real images. Our method achieves high attribution performance even for highly realistic GAN images generated by state-of-the-art GANs. We also test three common types of evasions, adding noise, blurring, and JPEG compression, on GAN generated images and find that our technique remains relatively robust to these attacks despite the obvious visual degradation to the original image. Our proposed approach provides a promising direction to efficiently attribute and detect GAN images.

References

[1] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *arXiv preprint arXiv:1511.06434* (2015).

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *Neural Information Processing Systems (NIPS)*. 2014.

[3] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[4] Raphael Satter. *Experts: Spy Used AI-generated Face to Connect with Targets*. June 2019. URL: https : / / apnews . com / bc2f19097a4c4fffaa00de6770b8a60d.

[5] Paris Martineau. *Facebook Removes Accounts with AI-Generated Profile Photos*. Dec. 2019. URL: https://www.wired.com/story/facebook-removes-accounts-ai-generated-photos/.

[6] Xin Yang, Yuezun Li, Honggang Qi, and Siwei Lyu. "Exposing GAN-synthesized Faces Using Landmark Locations". In: *ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*. 2019.

[7] Scott McCloskey and Michael Albright. "Detecting GAN-generated Imagery using Color Cues". In: *arXiv preprint arXiv:1812.08247* (2018).

[8] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang. "Detection of Deep Network Generated Images Using Disparities in Color Components". In: *arXiv preprint arXiv:1808.07276* (2018).

[9] Ning Yu, Larry Davis, and Mario Fritz. "Attributing Fake Images to GANs: Analyzing Fingerprints in Generated Images". In: *arXiv preprint arXiv:1811.08180* (2018).

[10] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. "Detecting and Simulating Artifacts in GAN Fake Images". In: *arXiv preprint arXiv:1907.06515* (2019).

[11] Chih-Chung Hsu, Chia-Yen Lee, and Yi-Xiu Zhuang. "Learning to Detect Fake Face Images in the Wild". In: *International Symposium on Computer, Consumer and Control (IS3C)*. 2018.

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. "Progressive Growing of GANs for Improved Quality, Stability, and Variation". In: *CoRR* abs/1710.10196 (2017).

[13] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[14] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. "Deep Learning Face Attributes in the Wild". In: *International Conference on Computer Vision (ICCV)*. 2015.

[15] Augustus Odena, Vincent Dumoulin, and Chris Olah. "Deconvolution and Checkerboard Artifacts". In: *Distill* (2016).

[16] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. "Do GANs Leave Artificial Fingerprints?" In: *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. 2019.

[17] Alan. V. Oppenheim and Jae. S. Lim. "The Importance of Phase in Signals". In: *Proceedings of the IEEE*. Vol. 69. 5. 1981.

[18] MJ Morgan, J Ross, and A Hayes. "The Relative Importance of Local Phase and Local Amplitude in Patchwise Image Reconstruction". In: *Biological cybernetics* 65.2 (1991).

[19] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. "Generative Image Inpainting with Contextual Attention". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018.

[20] Andrew Brock, Jeff Donahue, and Karen Simonyan. *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. 2018.

[21] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabs Pczos. "MMD GAN: Towards Deeper Understanding of Moment Matching Network". In: *Neural Information Processing Systems (NIPS)*. 2017.

[22] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rmi Munos. "The Cramer Distance as a Solution to Biased Wasserstein Gradients". In: *arXiv preprint arXiv:1705.10743* (2017).

[23] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In: *International Conference on Computer Vision (ICCV)*. 2017.

[24] Mehdi Mirza and Simon Osindero. "Conditional Generative Adversarial Nets". In: *arXiv preprint arXiv:1411.1784* (2014).

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. "Image-to-Image Translation with Conditional Adversarial Networks". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[26] Michael Albright and Scott McCloskey. "Source Generator Attribution via Inversion". In: *CVPR Workshop on Media Forensics*. 2019.

[27] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, BS Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, and Amit K Roy-Chowdhury. "Detecting GAN Generated Fake Images Using Co-occurrence Matrices". In: *Electronic Imaging* (2019).