Exploring the Asynchronous of the Frequency Spectra of GAN-Generated Facial Images

Le Minh Binh[1], Simon S. Woo[2]

[1]*Department of Software, Sungkyunkwan University, South Korea*

[2]*Department of Applied Data Science, Sungkyunkwan University, South Korea*

Abstract

The rapid progression of Generative Adversarial Networks (GANs) has raised a concern of their misuses for malicious purposes, especially in creating fake face images. Although many proposed methods succeed in detecting GAN-based synthetic images, they are still limited by the need for large quantities of the training fake image dataset, and challenges for the detector's generalizability to unknown facial images. In this paper, we propose a new approach that explores the asynchronous frequency spectra of color channels, which is simple but effective for training both unsupervised and supervised learning models to distinguish GAN-based synthetic images. We further investigate the transferability of a training model that learns from our suggested features in one source domain and validates on another target domains with prior knowledge of the features' distribution. Our experimental results show that the discrepancy of spectra in the frequency domain is a practical artifact to effectively detect various types of GAN-based generated images.

Keywords

Asynchronous of frequency, GAN-based synthetic images

1. Introduction

In recent years, there has been tremendous progress in Generative Adversarial Networks (GAN), in which two modules (generator vs. discriminator), play a minimax game to produce highly realistic data. Unfortunately, in addition to several fruitful GAN applications, attackers can exploit GANs for malicious purposes, such as spreading fake news [1] or propagating fake pornography of celebrities [2] as shown in the past. Meanwhile, several efforts have been made by researchers [3, 4, 5] to resist these nasty misuses. Wang *et al.* built a deep neural network to classify GAN-based generated images and empirically demonstrates that a classifier trained on one single dataset can generalize to different GAN datasets[4]. Especially, Dzanic and Shah [6] empirically show the systematic bias in high spatial frequencies and use this characteristic to classify real and deep network generated images. However, they did not explore the deeper statistical frequency features that we propose in our work, and they simply focused on converting the RGB to gray image, unlike ours.

Also, the checkerboard artifacts in spectrum generated by up-sampling components of GAN model were also extensively investigated by Zhang *et al.* [3] and Frank *et al.* [5]. While these techniques have shown success in terms of achieving high accuracy, they typically require a large
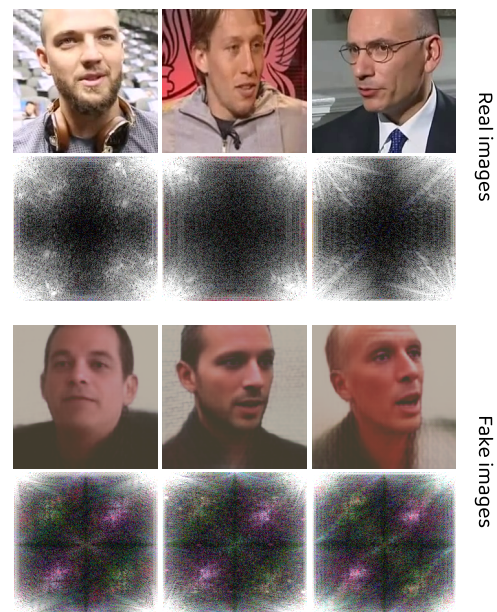
Figure 1: The frequency spectra differences of real vs. fake images. **Top**: Real images and their corresponding concurrent spectra. **Bottom**: Fake images and their corresponding (chaotic) spectra. It is difficult for human eyes to distinguish between real and fake images, but after applying DFT on each channel of images, the vital clues to distinguish real vs. fake images can be discovered.

quantities of data to train and the incurred high computational complexity, which can be prohibitively expensive
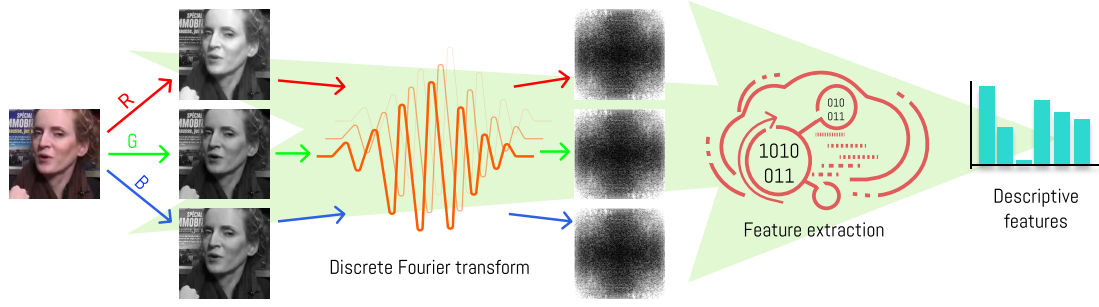
Figure 2: Illustration of our end-to-end pipeline. Unlike other research [6] that apply DFT on a grayscale image, we focus on each channel information and utilize statistical methods to obtain discriminative features.

in many practical applications. As a consequence, there is a strong need for detection methodologies that can achieve comparable levels of performance with limited training data, requiring low computational requirements and better generalization to unknown GAN-based generated images.

In this work, we first observe that the frequencies of the channels in the real images are highly correlated as shown in Fig. 1 and Fig. 3. In fact, the discriminator in the GAN model can make the generated synthesized images highly realistic, close to real images. However, to the best of our knowledge, there has not been any attempt to apply *direct correlation constraint* between channels on the output images in the frequency domain in the GAN models. From this observation, we hypothesize that the insufficiency of channel-dependent training in most of the current GAN models can produce the channel-wise asynchrony in the frequency domain. The asynchronous can be exposed in various GAN datasets, which can be effectively used to distinguish GAN-based synthetic images by both unsupervised or supervised learning methods.

In order to demonstrate the effectiveness of our proposed approach, we experiment with four types of datasets: Fake Head Talker [7], StyleGAN [8], StarGAN [9], and Adversarial Latent Auto Encoder (ALAE) [10]. Our main contributions in this work are summarized as follows:

- We firstly introduce the asynchronous in the frequency domain for detecting GAN-generated fake images.
- We propose effective unsupervised and supervised learning models using our discriminative mining features to classify real and fake images.
- We demonstrate the transferability of our proposed learning models across different GAN-based synthetic datasets using our spectra features.

2. Our Methods

2.1. Fourier Spectrum Analysis

In this section, we construct our hypothesis on the channel-wise asynchrony of GAN-based generated images. The convolutional operation at the lth layer of a generative model is formulated as follows:

$$A_i^{l+1} = Conv_i^l(A^l) = \sigma \left(\sum_{c=1}^{C_{(l)}} F_{ic}^l \circledast Up(A_c^l) \right), \quad (1)$$

where $C_{(l)}$ is the number of channels of the lth layer's output A^l, $F^l \in \mathbb{R}^{k_l \times k_l \times C_{(l+1)} \times C_{(l)}}$ is a set of $C_{(l+1)} \times C_{(l)}$ trainable $2D$ filters that have size of $k_l \times k_l$. And $Up(\cdot)$, $\circledast$ and $\sigma(\cdot)$ denote the up-sampling operator, convolutional operator and activation function, respectively. According to Khayatkhoei and Elgammal [11], we can simplify the Eq. 1 by restricting $\sigma(\cdot)$ to to rectified linear units (ReLU), which makes the $Conv_i^l(A^l)$ become locally piece-wise linear, and absorbing the up-sampling $Up(\cdot)$ into A_c^l. In this way, we transform Eq. 1 to:

$$A_i^{l+1} = Conv_i^l(A^l) = \sum_{c=1}^{C} F_{ic}^l \circledast A_c^l. \quad (2)$$

By applying the 2D discrete Fourier transform (DFT) to A_i^{l+1}, it is now viewed in the frequency domain as follows:

$$
\begin{aligned}
\tilde{A}_i^{l+1} = \mathfrak{F}(A_i^{l+1}) &= \mathfrak{F} \left(\sum_{c=1}^{C} F_{ic}^l \circledast A_c^l \right) \\
&= \sum_{c=1}^{C} \mathfrak{F} \left(F_{ic}^l \circledast A_c^l \right) \quad \textit{(linearity property of FT)} \\
&= \sum_{c=1}^{C} \mathfrak{F} \left(F_{ic}^l \right) \times \mathfrak{F} \left(A_c^l \right) \quad \textit{(conv. property of FT)} \\
&= \sum_{c=1}^{C} \tilde{F}_{ic}^l \times \tilde{A}_c^l = \langle \tilde{\boldsymbol{F}_i^l}, \tilde{\boldsymbol{A}^l} \rangle, \quad (3)
\end{aligned}
$$

where $\tilde{\boldsymbol{F}}_i^l = (\tilde{F}_{i1}^l, ..., \tilde{F}_{iC}^l)^T$, and $\tilde{\boldsymbol{A}}^l = (\tilde{A}_1^l, ..., \tilde{A}_C^l)^T$. Equation 3 indicates that in the frequency domain, every channel of in the next layer is decomposed into the combination all previous layer's channels with different sets of coefficients. If we consider A^{l+1} as the synthesized output image and fix $\tilde{\boldsymbol{A}}^l$, every vector $\tilde{\boldsymbol{F}}_{i,i=\overline{1,..C_{(l+1)}}}^l$ is trained independently to minimized the loss that applied on each $\tilde{A}_i^{l+1}$. When we consider $\tilde{\boldsymbol{A}}^l$ as a basic and $\tilde{\boldsymbol{F}}_i^l$ as the coordinate of $\tilde{A}_i^{l+1}$ with respect to $\tilde{\boldsymbol{A}}^l$, to synthesize a new image, the generative models expect that $\tilde{\boldsymbol{A}}^l$ is also good enough so that each independent coefficient vector $\tilde{\boldsymbol{F}}_i^l$ can produce corresponding single channel. In addition, these output channels should become as natural as possible in spacial domain after being stacked together in the order of three color channels: Red, Green and Blue. Small shifts of $\tilde{\boldsymbol{F}}_{i,i=\overline{1,..C_{(l+1)}}}^l$ in the frequency domain may only change fine-grained details of visualization but can produce frequency-bias when there is no direct constraint between channels.

2.2. Descriptive Features Extraction

Let $\mathcal{I}$ be a color image with three channels: Red, Green and Blue, which has width of W and height of H. To create its frequency representation, we firstly apply 2D DFT on each channel as follows:

$$\mathfrak{F}\mathcal{I}_{R/G/B}(u,v) = \sum_{x=1}^{W}\sum_{y=1}^{H} \mathcal{I}_{R/G/B}(x,y) \cdot e^{-i2\pi(\frac{ux}{W}+\frac{vy}{H})},$$
(4)

where x and y, and denote the xth and y^{th} slice in the width and height dimension of $\mathcal{I}$. For convenience, we use the notation $\mathfrak{F}\mathcal{I}_{R/G/B}$ to represent the function that is independently applied for each channel of the image. Note that $\mathfrak{F}\mathcal{I}_{R/G/B}(u,v)$ is now a complex number, *i.e.* $\mathfrak{F}\mathcal{I}_{R/G/B}(u,v) \in \mathbb{C}$, and the spectrum of each channel is obtained as follows:

$$Spec_{R/G/B}(u,v) = mod\left(\mathfrak{F}\mathcal{I}_{R/G/B}(u,v)\right), \quad (5)$$

where $mod(\cdot)$ denotes the modulus of complex number.

Although it would be challenging for human eyes to distinguish between real and GAN-generated fake images, we believe that their frequency spectra differences can be possibly exposed, when we stack the three channels' spectra of real vs. fake. Figure 1 presents our example of images' spectra from VoxCeleb2 dataset [12] and Fake Head Talker dataset [13]. In particular, in the real images, we empirically find that the spectra of three color channels are mostly concurrent when stacking together, whereas they become noisy in the fake images, as shown in Fig. 1.

Based on this important observation, we propose the following key statistical descriptive features to discriminate the GAN images in the frequency domain: $Mean$, Max, Min, $iCorr_{RG}$, $iCorr_{RB}$, and $iCorr_{GB}$, where the details are presented below:

- $Mean$. We take the average of the channel-wise spectrum differences:

$$Mean = \frac{d_{RG} + d_{RB} + d_{GB}}{3}, \qquad (6)$$

Also, we use d_{RG} that is the average spectrum differences between the spectra of the Red and Green channel in an image as follows:

$$d_{RG} = \frac{1}{WH}\sum_{u=1}^{W}\sum_{v=1}^{H} \left|Spec_R(u,v) - Spec_G(u,v)\right|,$$
(7)

and d_{RB} and d_{GB} can be similarly defined.
- Max and Min. We take the maximum and minimum values in $\{d_{RG}, d_{RB}, d_{GB}\}$.
- $i_{Corr_{RG}}$. We calculate the correlation coefficient between $Spec_R$ and $Spec_G$ and transform it to positive range value by adding 1 to its negative values as follows:

$$i_{Corr_{RG}} = -\rho(Spec_R, Spec_G) + 1, \quad (8)$$

where ρ is the Pearson correlation coefficient, and $iCorr_{RB}$ and $iCorr_{GB}$ can be similarly defined.

Our end-to-end pipeline of extracting above frequency descriptive features from a given image is visually illustrated in Fig. 2.

2.3. Binary Classifier

To demonstrate the characteristic-defining ability of the spectrum disagreement, we first employ the simple classifiers to classify real and fake images as below:

- **Gaussian Mixture Model (GMM)**. GMM is a probabilistic model that assumes the distribution of observed sampling data points is composed of a mixture of many Gaussian distributions, particularly, in our case is two distributions of real and fake class. To determine the means and variances of the two Gaussian distributions, Expectation-Maximization (EM) algorithm is used to iteratively estimate these parameters. Our descriptive features populate in the way that the higher spectrum agreement of an image will have the lower descriptive values. Therefore, we can expect that the Gaussian distribution in the mixture model, which has a smaller expectation will represent

the real images' distribution, and the other represents the fake images' distribution. By applying EM, we can classify real and fake images in an unsupervised manner in which the labels of a training dataset are not required.

- **Support Vector Machine (SVM)**. SVM is a robust supervised learning method that maximizes the margin of hyperplanes between different classes. The samples that lie along the margins are called the support vectors. In our experiment, we use SVM with the radial basis function (RBF) kernel to train with our six proposed features.

3. Experiment

3.1. Datasets

To examine the effectiveness of our proposed frequency features, we experiment four types of dataset: Fake Head Talker [7], StyleGAN [8], StarGAN [9], and Adversarial Latent Auto Encoder (ALAE) [10]. A brief description of each dataset is provided below as well as in Table 1:

- **Fake Head Talker dataset** [7]. Fake Head Talker is generated by the few-shot learning system that is pre-trained extensively on a large dataset (meta-learning). Particularly, their approach includes an embedder, a generator, and a discriminator. After training on a large corpus of talking head videos of different faces with adversarial training, their approach can transform facial landmarks from a source frame into realistically-looking personalized photographs with a few photos of a new target person, and further mimic the target.
- **StyleGAN** dataset [8]. StyleGAN is a high-level style controlling approach that governs its generator through adaptive instance normalization (AdaIN) and Gaussian noise adding in each convolutional layer. Furthermore, by proposing two novel metrics such as perceptual path length and linear separability, the generated images are less entangled and have different factors of variation.
- **StarGAN** dataset [9]. StarGAN is a unified model architecture that is able to train on multiple datasets across different domains. By proposing a simple mask vector, the StarGAN is able to flexibly utilize multiple datasets containing different label sets, and achieve competitive results in the facial attribute transfer tasks. This new approach with only a single generator and a discriminator has addressed the scalability and robustness limitations of many previous research.
- **Adversarial Latent Auto Encoder (ALAE)** dataset [10]. ALAE is an autoencoder-based generative model that is capable to learn the disentangled

representations in the latent space with adversarial settings. The ALAE model can not only synthesize high-resolution images comparing with StyleGAN, but also can further manipulate or reconstruct the new input facial images.

Details of each dataset in our experiment are summarized in Table 1. In our experiment, the number of real and GAN fake images are equal in both training and test sets. We further provide the histograms to visualize the distributions of six descriptive features of these datasets in Supp. Section A.

3.2. Experimental Results

To demonstrate the discriminative power of our proposed features, we perform three different experiments.

Binary classification. Our experimental results are shown in Table 2. We can observe that both unsupervised and supervised methods are able to produce high performance with our newly introduced frequency features. The accuracy scores of the unsupervised method on Fake Head Talker and ALAE dataset are competitive, compared to the supervised approaches. At the same time, they are still higher than 80% on StyleGAN and StarGAN. Meanwhile, the supervised method's accuracy scores are always higher than 95% on the four datasets. We can conclude that our proposed features based on the asynchronous in the frequency spectrum can effectively capture the characteristics of the GAN-generated images, and provide the foundation for distinguishing fake from real images.

Unbalanced training datasets. Furthermore, to study the feasibility of training with an unbalanced dataset using our features, we gradually reduce the number of fake images in each training dataset to $25\%, 5\%$, and 1% of the total training data size. After that, we apply the SVM as our learning model. To demonstrate our approach's effectiveness, we compare our method with FakeTalkerDetect model [13], which deployed a pre-trained AlexNet and Siamese network trained on RGB images. The results are presented in Table 3. We can observe that our method with the hand-crafted features outperforms the AlexNet and FakeTalkerDetect on both balanced and unbalanced datasets. Therefore, we can conclude that our simple yet effective features are capable to characterize the fake image much better in the unbalanced training dataset scenario, as well.

Unsupervised domain adaptation. In this task, we propose an algorithm using our proposed features that allows a pre-trained SVM model on one source dataset (*e.g.,* StyleGAN) can detect fake images in a new target dataset (*e.g.,* ALAE) with the only prior knowledge of the target feature expectations.

Table 1

Details of datasets used in our experiment.

Datasets	Resolution	Source datasets	Training size (real+ fake)	Test size (real+ fake)
Fake Head Talker [7]	224×224	VoxCeleb2 [12]	18,800	18,800
StyleGAN [8]	1024×1024	FFHQ [1]	2,000	2,000
StarGAN [9]	256×256	CelebA [14]	2,000	1,998
ALAE [10]	1024×1024	FFHQ	2,000	2,000

Table 2

Experimental results of GMM and SVM on four datasets with our discriminative features.

Datasets	GMM				SVM			
	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
Fake Head Talker	0.996	1.000	0.991	0.996	0.9972	0.994	1.000	0.997
StyleGAN	0.849	0.762	0.915	0.831	0.951	0.938	0.963	0.950
StarGAN	0.903	0.807	1.000	0.893	0.972	0.949	0.994	0.971
ALAE	0.992	0.984	1.000	0.992	0.999	0.997	1.000	0.998

Table 3

Comparison between our approach using proposed descriptive features and AlexNet and FakeTalkerDetect method on Fake Head Talker dataset. The precision, recall and F1 scores of AlexNet and FakeTalkerDetect from [13], and their values are rounded to second decimal

Methods	Accuracy	Recall	Precision	F1
AlexNet (50% fake)	0.981	0.98	0.98	0.98
FakeTalkerDetect	0.984	0.98	0.98	0.98
SVM (ours)	**0.997**	**0.994**	**1.00**	**0.997**
AlexNet (25% fake)	0.971	0.95	0.95	0.96
FakeTalkerDetect	0.986	0.98	0.98	0.98
SVM (ours)	**0.998**	**0.995**	**1.000**	**0.997**
AlexNet (5% fake)	0.964	0.98	0.80	0.87
FakeTalkerDetect	0.988	0.99	0.91	0.94
SVM (ours)	**0.997**	**0.997**	**0.997**	**0.997**
AlexNet (1% fake)	0.963	0.98	0.61	0.67
FakeTalkerDetect	0.988	0.99	0.74	0.82
SVM (ours)	**0.992**	**0.999**	**0.986**	**0.992**

Algorithm 1 Unsupervised domain adaptation with SVM using our proposed descriptive features

---

Require: Labeled source set $\{X^s, Y^s\}$, unlabeled target set X^t, where X^s and X^t includes the six proposed features $[f_1^s, .., f_6^s]$, respectively. The prior knowledge of Gaussian expectation values: $\left[m_{i,0}^s, m_{i,1}^s\right]_{i=\overline{1,...,6}}$ and $\left[m_{i,0}^t, m_{i,1}^t\right]_{i=\overline{1,...,6}}$.

1: Step 1: Scale each feature in X^s and X^t:
$$\bar{f}_i^s = \left(f_i^s - m_{i,0}^s\right) / \left(m_{i,1}^s - m_{i,0}^s\right),$$
$$\bar{f}_i^t = \left(f_i^t - m_{i,0}^t\right) / \left(m_{i,1}^t - m_{i,0}^t\right)$$
2: Step 2: Fit source set $\left\{[\bar{f}_1^s, .., \bar{f}_6^s], Y^s\right\}$ with SVM model.
3: Step 3: Use pre-trained SVM to predict target set label from $[\bar{f}_1^t, .., \bar{f}_6^t]$.

---

suggested features the pre-trained SVM shows its strong detection ability in the new target domain, where all the detection performance is above 80% of accuracy for any pair of source and target dataset. This preliminary experiment shows that our proposed features can be utilized in domain adaptation tasks with more complex learning models in the future.

4. Conclusion

Although GANs have significantly advanced in the past, we discover that there are some areas that GANs' cannot mimic the real images effectively in the frequency domain. Thus, in this work, we propose a preliminary approach that reveals the asynchronous in frequency domain of the three channels in GAN images. By mining statistical features in frequency domain, our simple yet

In particular, we first take the two Gaussian expectation values of two mixture distributions of each feature from both source and target dataset. These expectation values are kept as our prior knowledge about the target dataset. We then scale the source training set features such that their two Gaussian expectation values are normalized between 0 and 1 , to better fit the training dataset with the SVM model. In the testing phase, with our prior knowledge above, we can scale the testing features from the target dataset using the known expectation values and feed them to the pre-trained SVM model to make prediction. This adaptation learning process is summarized in the Algorithm 1.

We experiment with the four fake dataset and present the results in Table 4. We can observe that with our

Table 4
Experimental results of domain adaptation task using our proposed features

Source dataset	Target dataset	Accuracy	Recall	Precision	F1
Fake Head Talker	StyleGAN	0.800	0.908	0.745	0.819
	StarGAN	0.918	0.844	0.992	0.912
	ALAE	0.994	0.995	0.993	0.994
StyleGAN	Fake Head Talker	0.965	0.932	0.998	0.964
	StarGAN	0.906	0.814	0.998	0.896
	ALAE	0.991	0.982	1.000	0.991
StarGAN	Fake Head Talker	0.983	0.982	0.983	0.983
	StyleGAN	0.832	0.980	0.756	0.854
	ALAE	0.996	0.997	0.994	0.996
ALAE	Fake Head Talker	0.993	0.989	0.998	0.993
	StyleGAN	0.890	0.955	0.845	0.897
	StarGAN	0.929	0.871	0.985	0.925

effective unsupervised and supervised learning methods can easily discriminate the real and GAN-based synthetic facial images without utilizing deep learning methods. Our extensive experiments demonstrates that the proposed features' power in three scenarios: 1) unsupervised and supervised binary classification, 2) unbalanced training dataset, and 3) domain adaptation task. For future work, we plan to explore and exploit more on these aspects of GAN-generated images to combat against misuses from attackers, and extend our work to deepfake detection.

Acknowledgments

References

[1] T. Quandt, L. Frischlich, S. Boberg, T. Schatto-Eckrodt, Fake news, The international encyclopedia of Journalism Studies (2019) 1–6.

[2] S. Cole, We are truly fucked: Everyone is making ai-generated fake porn now, 2018. URL: https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley.

[3] X. Zhang, S. Karaman, S.-F. Chang, Detecting and simulating artifacts in gan fake images, in: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2019, pp. 1–6.

[4] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, Cnn-generated images are surprisingly easy to spot... for now, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8695–8704.

[5] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, T. Holz, Leveraging frequency analysis for deep fake image recognition, in: International Conference on Machine Learning, PMLR, 2020, pp. 3247–3258.

[6] T. Dzanic, K. Shah, F. Witherden, Fourier spectrum discrepancies in deep network generated images, arXiv preprint arXiv:1911.06465 (2019).

[7] E. Zakharov, A. Shysheya, E. Burkov, V. Lempitsky, Few-shot adversarial learning of realistic neural talking head models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9459–9468.

[8] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[9] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8789–8797.

[10] S. Pidhorskyi, D. A. Adjeroh, G. Doretto, Adver-

sarial latent autoencoders, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14104–14113.

[11] M. Khayatkhoei, A. Elgammal, Spatial frequency bias in convolutional generative adversarial networks, arXiv preprint arXiv:2010.01473 (2020).

[12] A. Nagrani, J. S. Chung, W. Xie, A. Zisserman, Voxceleb: Large-scale speaker verification in the wild, Computer Science and Language (2019).

[13] H. Jeon, Y. Bang, S. S. Woo, Faketalkerdetect: Effective and practical realistic neural talking head detection with a highly unbalanced dataset, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[14] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV), 2015.

A. Distribution of Statistical Descriptive Features

The histogram distributions of our six proposed statistical features in the frequency domains are present in the Fig. 3. We can observe that these feature distributions are highly separable between real and fake images across four datasets.
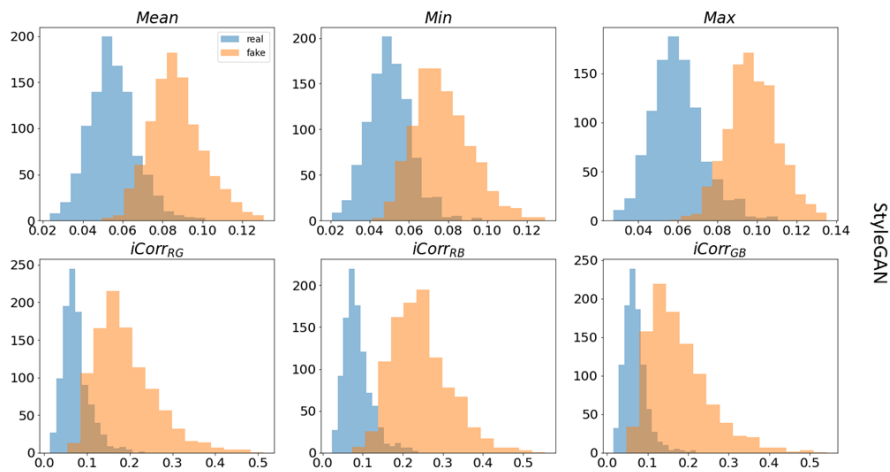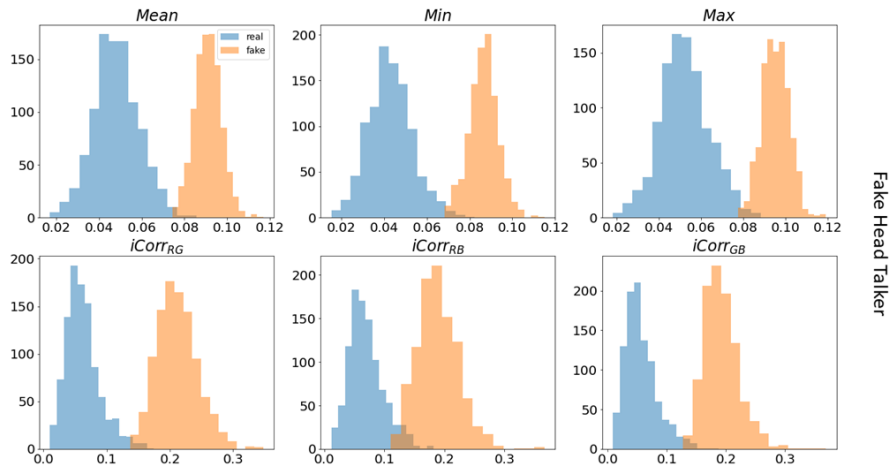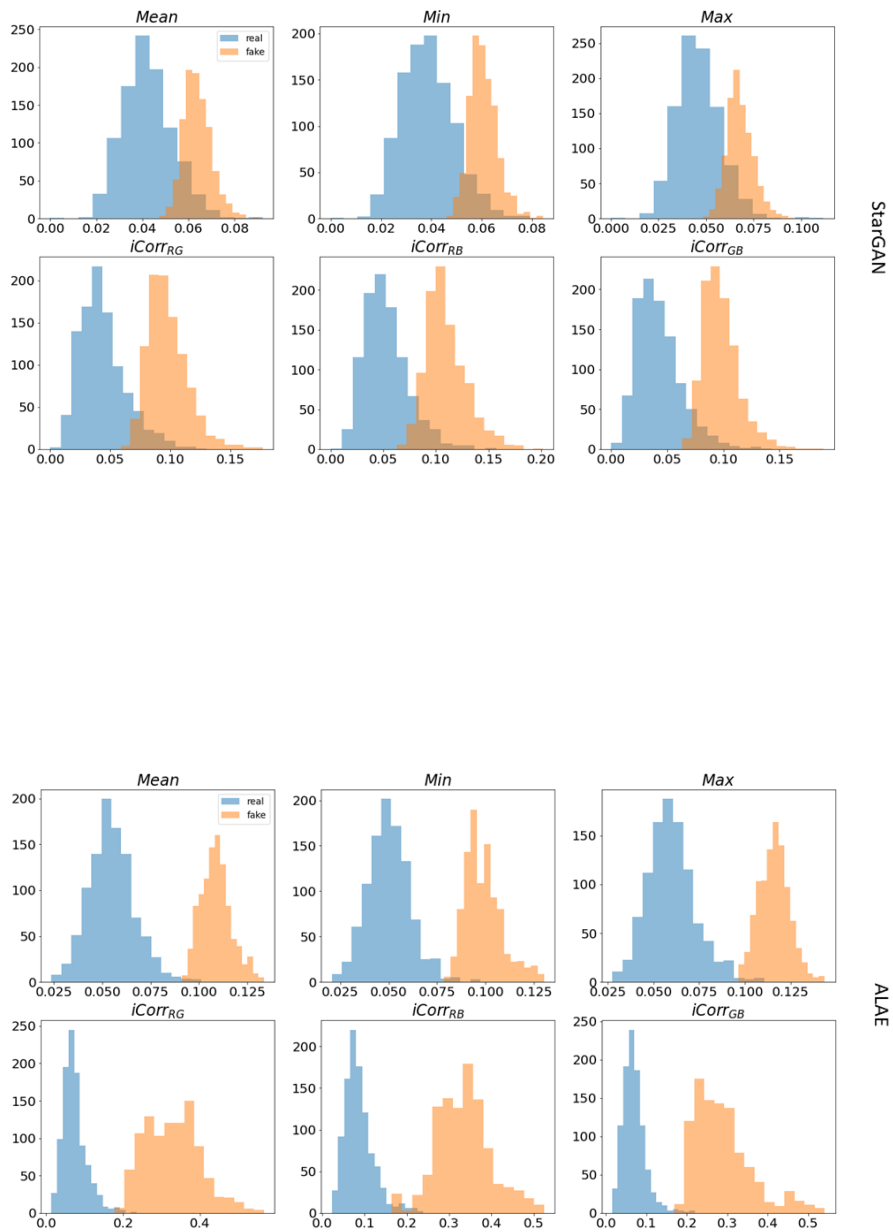
Figure 3: The histogram distributions of our six statistical descriptive features from four datasets: Fake Head Talker, StyleGAN, StarGAN and ALAE.

B. Example GAN-based Synthetic Images

We provide example images from four datasets used in our experiment.

Fake Head Talker

StyleGAN

StarGAN

ALAE