FingerprintNet: Synthesized Fingerprints for Generated Image Detection

– Supplementary Material –

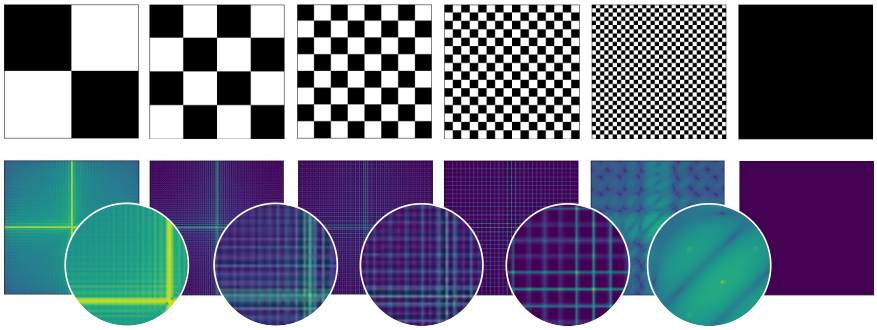

Fig. A: **The checkerboard artifacts in images and the artificial fingerprints in 2D spectra.** The greater the size of the checkerboard artifacts in images, the narrower the grids become in the artificial fingerprints in the frequency domain.

A Additional Analysis

The generator network for GAN models can be divided into two categories: the networks based on deconvolution and the networks based on interpolation. Conventionally, most previous upsampling networks for image generation gradually increase the image resolutions through the activation layer and the deconvolution layer. To improve the image quality, a number of methods [1, 2, 4, 9, 18, 19] additionally use the normalization layer, such as the batch normalization [10], adaptive instance normalization [12], and spectral normalization [15]. Also, anti-aliasing methods can be used, such as blur and low-pass filter [5, 11, 12, 16, 17]. Since the deconvolution layer for upsampling operation is known to be the cause of aliasing, the interpolation-based networks replace the deconvolution layer with interpolation, such as bicubic and linear for improved image quality [3, 5].

A number of studies have confirmed that the images generated by GAN models contain the fingerprints appearing as the unique patterns in the frequency domain and utilizing those fingerprints can be the key to robust detection of the generated images [6, 8]. As shown in Fig. A, the artificial fingerprints generated by the deconvolution layer are easily discovered in the 2D spectra. Interestingly, however, these fingerprints are not found in the generated images by the

interpolation-based methods, which can be concluded that each network creates unique fingerprints varying from each other.

We first analyze the frequency-level fingerprints generated by the deconvolution networks. To show the relation between the frequency-level fingerprints and the pixel-level checkerboard artifacts generated by the deconvolution layers, we express the pixel-level artifacts by using newly derived frequency-level fingerprints. To ease the derivation, we consider the 1-D sequence in the following derivations. When the size of the given sequence is N,

$$\mathbf{y}[n] = \mathbf{h}[n] + \mathbf{g}[n], \tag{a}$$

where n is an integer in $[0, N)$, and $\mathbf{y}$, $\mathbf{h}$, and $\mathbf{g}$ are the reconstructed sequence, the original sequence, and the pixel-level artifacts, respectively. Thus, when a_m and T represent a scale factor for m-th impulse sequence and the period between the impulse sequences, respectively, we can represent the frequency-level fingerprints by the weighted impulse train as follows:

$$\mathcal{F}\{\mathbf{g}\}[k] = \sum_{m=-M}^{m=M} a_m \delta[k - mT], \tag{b}$$

where $\mathcal{F}$ and $\delta[k]$ are the Fourier transformation function and an unit impulse sequence of the frequency component k, respectively.

Then, to acquire the pixel-level artifacts from the frequency-level fingerprints, we estimate the inverse Fourier transformation of the frequency-level fingerprints as:

$$\begin{aligned} \mathbf{g}[n] &= \frac{1}{N} \sum_{k=0}^{N-1} \sum_{m=-M}^{M} a_m \delta[k - mT] \exp\left(j \frac{2\pi k}{N} n\right) \\ &= \frac{1}{N} \sum_{m=-M}^{M} a_m \exp\left(j \frac{2\pi mT}{N} n\right), \end{aligned} \tag{c}$$

where the summation over k is canceled out since the impulse sequence has non-zero value only when $k = mT$. Because we consider the real-valued sequence, a_m satisfies the symmetry property that is $b_m = b_{-m}$ and $c_m = -c_{-m}$ where $b_m \equiv \texttt{Re}(a_m)$ and $c_m \equiv \texttt{Im}(a_m)$, *i.e.* $a_m = b_m + jc_m$. When the real and imaginary parts of $\exp\left(j\frac{2\pi mT}{N}n\right)$ are separated respectively to $p_m(n) = \cos\left(\frac{2\pi mT}{N}n\right)$ and $q_m(n) = \sin\left(\frac{2\pi mT}{N}n\right)$ by Euler's equation, the equation can be simplified with the sum formula for cosine functions as:

$$\begin{aligned} \mathbf{g}[n] &= \frac{2}{N} \sum_{m=0}^{M} b_m \cos\left(\frac{2\pi mT}{N} n\right) - c_m \sin\left(\frac{2\pi mT}{N} n\right) \\ &= \frac{2}{N} \sum_{m=0}^{M} \cos\left(\frac{2\pi mT}{N} n + \alpha\right), \end{aligned} \tag{d}$$

where $\alpha = \arctan(c_m / b_m)$. Then, by using the derivation of the summation of cosine series [14], we can acquire the final derivation of:

$$\mathbf{g}[n] = \frac{2}{N} \frac{\sin(\pi MTn/N)}{\sin(\pi Tn/N)} \cos\left(\frac{\pi(M-1)T}{N} n + \alpha\right). \tag{e}$$

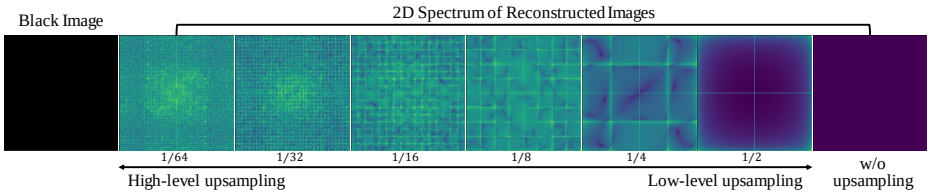

Fig. B: **The 2D spectra of the images reconstructed from zero image.** The patterns of the frequency-level fingerprints appear more frequently when more deconvolution layers are applied. Thus, by using the autoencoders with the various numbers of deconvolution layers, we can reconstruct different types of frequency-level fingerprints.

From the derivation, we can obtain three interesting characteristics of the pixel-domain artifacts. First, the fingerprints in the frequency domain are easier to discover because of their composition in the impulse train format, unlike the pixel-level artifacts based on the smooth trigonometric functions. This characteristic verifies that the impressive performance of the GAN detectors using the frequency-level fingerprints [6–8].

Second, the maximum amplitude of pixel-domain artifacts is proportional to the number of included deconvolution layers. When we estimate the upper bound of $\mathbf{g}[n]$,

$$\mathbf{g}[n] \leq \frac{2}{N} \frac{\sin(\pi MTn/N)}{\sin(\pi Tn/N)} \leq \frac{2}{N} M, \tag{f}$$

because $\sin(\pi MTn/N)/\sin(\pi Tn/N) \leq M$ of which the proof is given in Appendix. Thus, the upper bound of $\mathbf{g}[n]$ is proportional to M, which explains that the pixel-artifacts become easily distinguished with large M, as shown in Fig. 1. In Appendix A.1, we show the empirical results where the maximum value of $\mathbf{g}[n]$ increases proportionally to M.

Lastly, the greater the periods of the frequency-level fingerprints, the smaller the size of the pixel-level artifacts become. When we approximate the relation among M, T, and N, $N = (2M-1)T \approx 2MT \approx 2(M-1)T$ with large M. Thus, we can approximate the trigonometric frequencies of $\sin(\pi MTn/N)$ and $\cos(\pi(M-1)Tn/N + \alpha)$ as the constant, while the frequency of $\sin(\pi Tn/N)$ is proportional to the period T.

A.1 Proof of Equation f

We discover two interesting relations between the frequency-level fingerprints and the pixel-level artifacts of the generated images. The first relation presents that the magnitude of the pixel-level artifacts is proportional to the number of grids in the frequency-level fingerprints, which is derived by Eq. f. The second relation demonstrates that the number of grids in the frequency-level fingerprints is inversely proportional to the frequency of the checkerboards in the pixel-level artifacts.

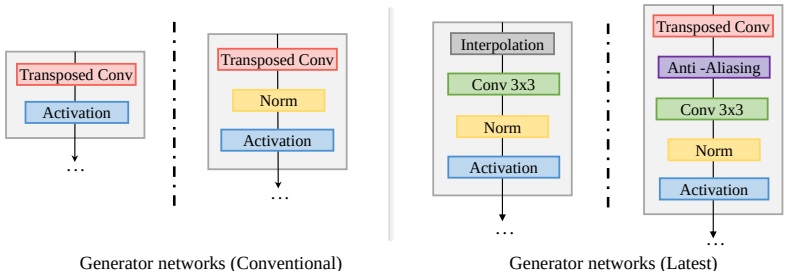

Fig. C: **The simplified upsampling process of the generator.** The solid-lined boxes indicate the essential operation, while the dashed-lined boxes indicate the selective operations.

To derive Eq. f, we need to prove the inequality of $\sin(\pi MTn/N)/\sin(\pi Tn/N) \leq M$. In this appendix, we present the proof for $|\sin(\pi MTn/N)/\sin(\pi Tn/N)| \leq M$ that is the sufficient condition of $\sin(\pi MTn/N)/\sin(\pi Tn/N) \leq M$. We first simplify $|\sin(\pi MTn/N)/\sin(\pi Tn/N)| \leq M$ as:

$$\left|\frac{\sin(M\theta)}{\sin(\theta)}\right| \leq M, \tag{g}$$

where $\theta \equiv \pi Tn/N$. Since M is larger than 0,

$$\begin{aligned} \frac{\sin^2(M\theta)}{\sin^2(\theta)} &\leq M^2 \\ 0 &\leq \sin^2(\theta) - \frac{1}{M^2}\sin^2(M\theta). \end{aligned} \tag{h}$$

By using the Euler's formula where $sin(\theta) = (e^{j\theta} - e^{-j\theta})/2j$ (j is an imaginary unit), the derivation becomes:

$$\begin{aligned} &\sin^2(\theta) - \frac{1}{M^2}\sin^2(M\theta) \\ &= -\frac{1}{4}\left(e^{j\theta} - e^{-j\theta}\right)^2 + -\frac{1}{4M^2}\left(e^{jM\theta} - e^{-jM\theta}\right)^2 \\ &= -\frac{1}{4}\left(e^{j2\theta} + e^{-j2\theta} - 2\right) + \frac{1}{4M^2}\left(e^{j2M\theta} + e^{-j2M\theta} - 2\right) \\ &= -\frac{1}{4}\left(2\cos(2\theta) - 2\right) + \frac{1}{4M^2}\left(2\cos(2M\theta) - 2\right). \end{aligned} \tag{i}$$

Then, when M is sufficiently large, $1/M^2$ goes to zero, so we can approximate the derivation as:

$$\begin{aligned} -\frac{1}{4}\left(2\cos(2\theta) - 2\right) + \frac{1}{4M^2}\left(2\cos(2M\theta) - 2\right) \\ \approx -\frac{1}{4}\left(2\cos(2\theta) - 2\right). \end{aligned} \tag{j}$$

Since

$$-1 \leq cos(2\theta) \leq 1, 0 \leq -\frac{1}{4}\left(2\cos(2\theta) - 2\right) \leq 1. \tag{k}$$

Thus,

$$|\sin(\pi MTn/N)/\sin(\pi Tn/N)| \leq M \tag{l}$$

is satisfied, which follows

$$\sin(\pi MTn/N)/\sin(\pi Tn/N) \leq M. \tag{m}$$

Then, we empirically show the relationship between the grid size of the pixel-level checkerboards and the number of lines in the frequency-level fingerprints. As shown in Fig. B, the larger the grid size of the pixel-level checkerboards, the more frequent the lines of frequency-level artifacts become. Thus, we can see that the grid size of the pixel-level checkerboards and the number of lines in the frequency-level fingerprints are positively correlated to each other.

A.2 Upsampling Process Modules

We categorize the upsampling process modules into two types: interpolation-based upsampling and deconvolution-based upsampling. Fig. C shows the difference between the two types of modules. The interpolation-based module contains the essential operations of interpolation, convolution, and activation function. In the case of a deconvolution-based module, the transposed convolution and activation function are the essential operations. In both the modules, the normalization can be selectively used to stabilize the training of the generator. To reduce the artifacts from the transposed convolution, the anti-aliasing and additional convolution layer are selectively added to the upsampling module.

B Visualization of Frequency-level Fingerprints

To find the autoencoder generating the most similar fingerprints to each GAN model, we build various autoencoders containing a different number of upsampling processes. After training the autoencoders by the images generated by the specific GAN model, we can obtain the various types of fingerprints from the numerous autoencoders. Interestingly, as shown in Fig. D, the autoencoders can generate one or more frequency-level fingerprints that are similar to those of GAN models. From the results, we can find that the multiple autoencoders can effectively reconstruct the fingerprints of various GAN models.

C Implementation Details

We resize the input images into 256×256, and the reconstructed images with the same size. Also, for detector, center crop is used for the input size of 250×250. For training, we use a single NVIDIA RTX 8000 with a batch size of 16 and

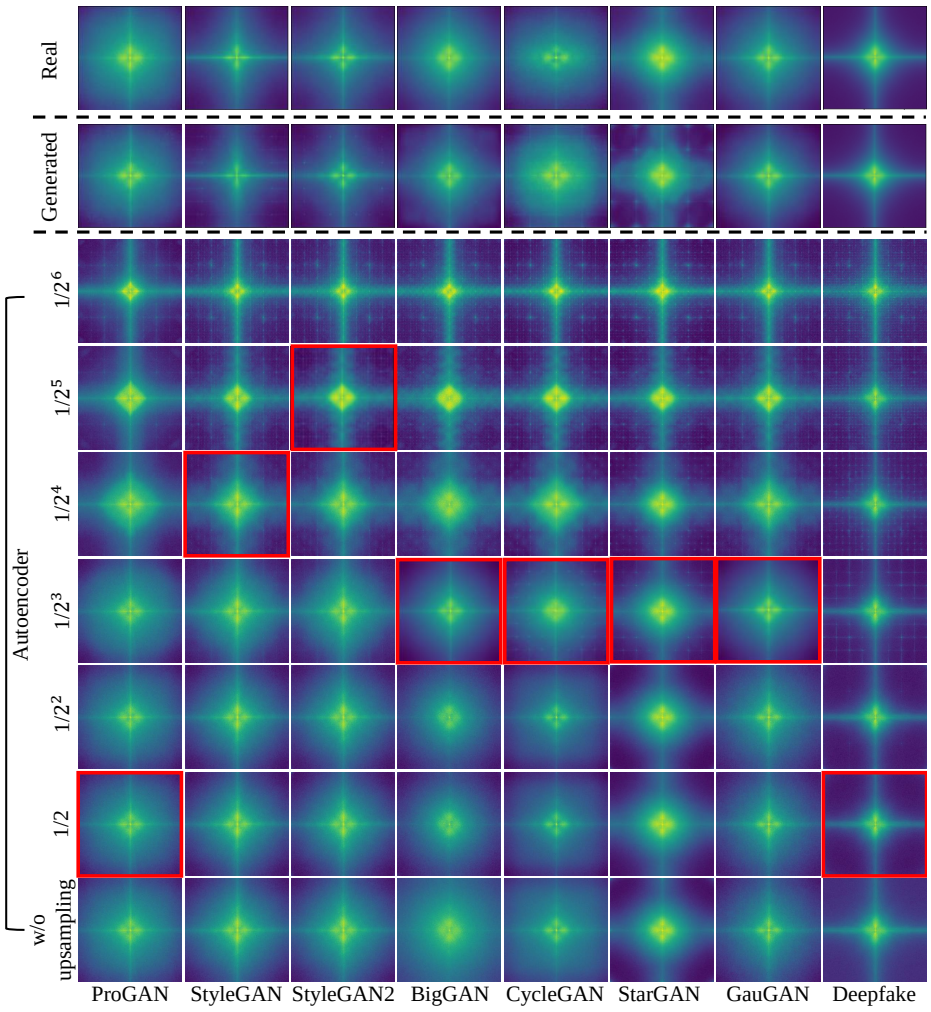

Fig. D: **Raw Results for Fingerprints of Autoencoders.** For the various GAN models, the autoencoders of different scales generate the most similar fingerprints (red boxes) to those of GAN models.

20 epochs for the generated image detector, 200 epochs for the artificial fingerprint generator. Both of the artificial fingerprint generator and generated image detector networks are trained by Adam optimizer [13] with the learning rate of 0.0001, which are the conventional hyperparameters of the previous generated image detectors.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 **30** (2017)
2. Berthelot, D., Schumm, T., Metz, L.: Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717 (2017)
3. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=B1xsqj09Fm
4. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE Conference on Computer Vision and Pattern Recognition (2018)
5. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
6. Durall, R., Keuper, M., Keuper, J.: Watch your Up-Convolution: CNN Based Generative Deep Neural Networks are Failing to Reproduce Spectral Distributions. In: IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA, United States (2020)
7. Durall, R., Keuper, M., Pfreundt, F.J., Keuper, J.: Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.00686 (2019)
8. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: International Conference on Machine Learning. pp. 3247–3258. PMLR (2020)
9. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of wasserstein gans. arXiv preprint arXiv:1704.00028 (2017)
10. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning (2015)
11. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=Hk99zCeAb
12. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
13. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (12 2014)
14. Knapp, M.P.: Sines and cosines of angles in arithmetic progression. Mathematics Magazine pp. 371–372 (2009)
15. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018)
16. Park, T., Zhu, J.Y., Wang, O., Lu, J., Shechtman, E., Efros, A.A., Zhang, R.: Swapping autoencoder for deep image manipulation. In: Advances in Neural Information Processing Systems (2020)
17. Pidhorskyi, S., Adjeroh, D.A., Doretto, G.: Adversarial latent autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14104–14113 (2020)
18. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)

19. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (2017)