Supplementary Materials
A Closer Look at Fourier Spectrum Discrepancies for CNN-generated Images Detection

Keshigeyan Chandrasegaran Ngoc-Trung Tran Ngai-Man Cheung
Singapore University of Technology and Design (SUTD)
{ keshigeyan, ngoctrung_tran, ngaiman_cheung } @sutd.edu.sg

Supplementary materials include research reproducibility details, additional results and more experiments to support our thesis statement.

1. Standard Deviation of Experiments

Standard deviation of spectral distributions were not included in the paper to allow better readability of graphs. In this section, we include the standard deviation of experiments for Baseline, Z.1.5, N.1.5 and B.1.5 experiments. The spectral distributions with standard deviations for CelebA [29], LSUN [43] and StarGAN [7] experiments are shown in Figures A, B and L respectively. We also show the Standard deviations of Spectral regularization experiments in Figure C. In all cases, we observe that the standard deviations are within acceptable range.

2. Spectral Regularization

Results of SR experiments are shown in Figure D. More specifically, SR performs generator loss scaling as there are no gradients with respect to the power spectrum difference between real and synthetic images. We were intrigued by the question on how generator loss scaling can achieve spectral consistency as claimed by [10] and noticed that the source code uses N.3.5 setup together with SR[1]. Analysing SR is out of scope for this work, but we have showed that N.3.5 setup (Main paper) is sufficient to achieve spectral consistency in identical setups.

3. Higher Resolution Experiments

In order to further investigate our thesis statement that high frequency decay discrepancies are not inherent characteristics for CNN-generated images, we extend our analysis to larger resolutions. We use image reconstruction as a representative task to investigate these effects at higher resolutions (We use 512x512).

[1] https://github.com/cc-hpc-itwm/UpConv

We select a subset of CelebA-HQ [21] dataset to train a standard autoencoder for image reconstruction. Similar to experiments in the paper, we perform experiments corresponding to Baseline, Z.1.5, N.1.5 and B.1.5 setups. We observe that Baseline and Z.1.5 setups produce high frequency Fourier discrepancies for reconstructed images, and N.1.5 and B.1.5 setups produce spectral consistent reconstructed images. The spectral distributions are shown in Figure E. This further confirms that high frequency Fourier discrepancies are not intrinsic for CNN-generated images.

4. Samples

We show more samples and spectral distributions for important CelebA [29] experiments in this section.

GAN Samples: We show extensive samples corresponding to Baseline, Z.1.5, N.1.5 and B.1.5 setups for DCGAN [34], LSGAN [31], WGAN-GP [17] in Figures F, G, H respectively.

Image-to-Image Translation: For StarGAN [7] experiments, we show an example of a reference image with corresponding translated images for Baseline, Z.1.5, N.1.5 and B.1.5 setups in Figures I. We also show the corresponding spectral distributions in Figure J.

Image Reconstruction: We show image reconstruction results for a few CelebA-HQ [21] examples in Figure K

5. FID scores

FID scores for all CelebaA [29] experiments are shown in table A. We used 50k real images and 50k GAN images to calculate each FID score. We observe that the FID scores of nearest and bilinear interpolation methods are comparable or better than the Baseline FID for all GAN setups.

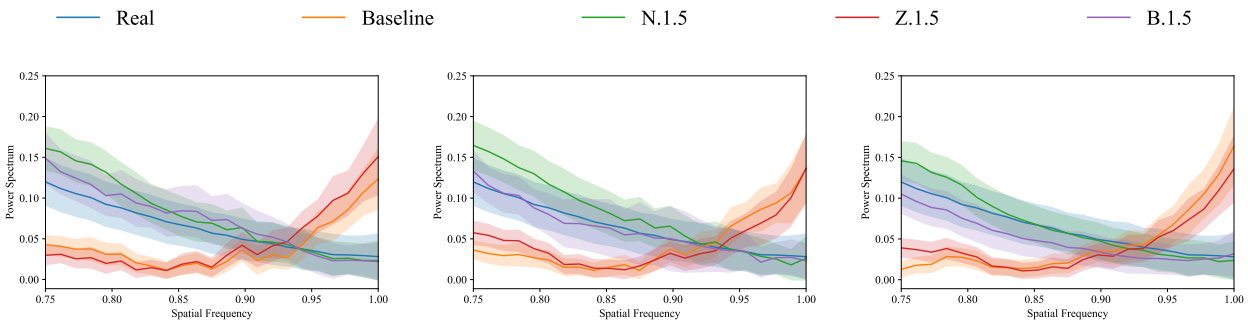

Figure A. This figure shows spectral plots from Figure 3 in the paper, with standard deviations indicated.

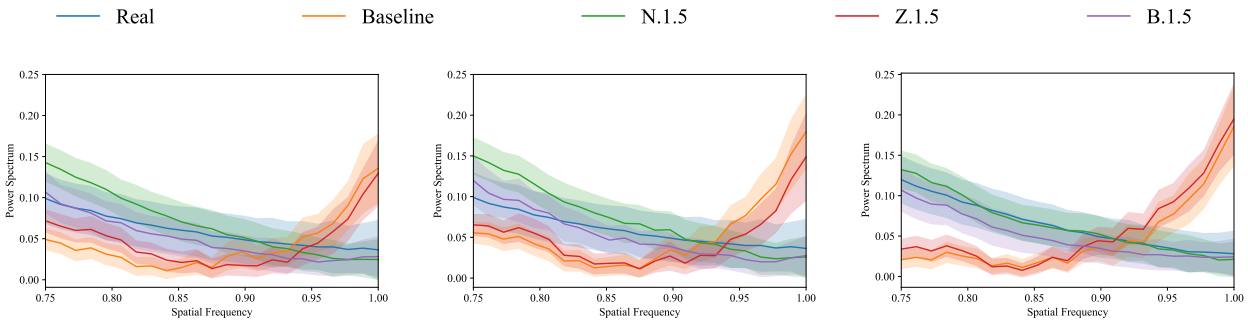

Figure B. This figure shows spectral plots from Figure 8 in the paper, with standard deviations indicated.

Setup Code	DCGAN	LSGAN	WGAN-GP
Baseline	**88.6**	**73.26**	**60.6**
N.1.5	87.52	70.69	48.69
Z.1.5	69.14	60.29	47.73
B.1.5	84.65	78.66	52.18
N.1.7	90.8	73.09	60.11
Z.1.7	71.45	59.55	43.1
B.1.7	79.92	76.33	55.28
N.1.3	93.54	74.06	58.35
Z.1.3	65.46	61.45	56.91
B.1.3	76.04	81.97	58.55
N.3.5	73.63	78.31	55.47
Z.3.5	68.41	66.27	57.59
B.3.5	80.89	72.29	54.84
SR	99.2	86.16	60.81

Table A. FID scores of GAN images trained on CelebA [29] dataset. We include the FID scores of Spectral Regularized GANs (indicated as SR) for comparison.

6. Fourier Synthetic Image Detector

We include more information and results for the synthetic image detector proposed by Dzanic *et al.* [12]. All detection rates are averaged over 10 independent runs.

6.1. Classifier Implementation Details

The exact procedure used by Dzanic *et al.* [12] to implement the classifier is shown below. For easier understanding, let us assume that we want to train a classifier to detect between real and StyleGAN images

1. Collect a repository of 1000 StyleGAN images and 1000 real images.
2. Obtain the un-normalized reduced spectrum for every real and StyleGAN image for the last 25% spatial frequencies (0.75 - 1.0).
3. Fit the reduced spectrum using power law function for every real and GAN image.
4. Extract 3 features b_1, b_2, b_3 from the fitted spectrum where b_1: start value of the fitted spectrum, b_2: decay value of the fitted spectrum, b_3: end value of the fitted spectrum. Though the authors mention only b_1, b_2 in their paper, their official implementation contained b_3 as well. This difference is acceptable since b_3 is linearly dependent on b_1, b_2 under the assumption that the power law is a good fit.
5. Train/ apply a binary KNN classifier (with k=5) using the 3 features extracted per image to predict if the GAN images are real or fake. The authors use 100 real and GAN images each to train the classifier and use the remaining 900 samples to test.

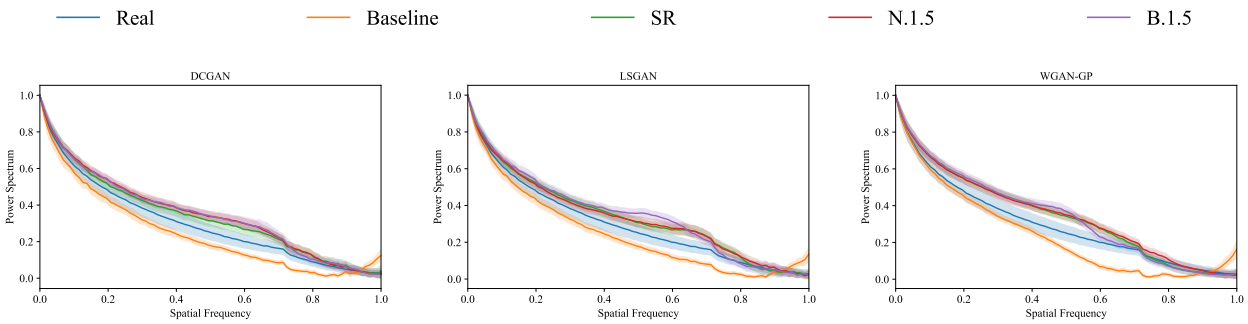

Figure C. This figure shows spectral plots from Figure 9 in the paper, with standard deviations indicated. "SR" refers to Spectral Regularization

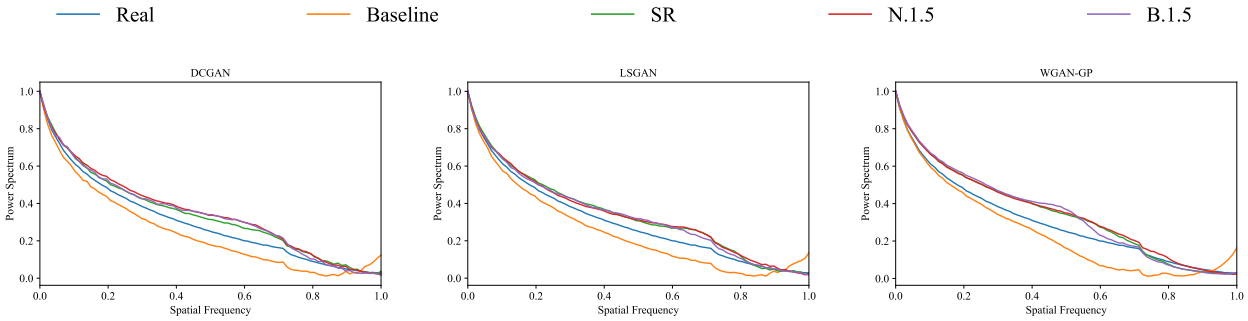

Figure D. We show the entire spectrum similar to [10]. "SR" refers to Spectral Regularization. We observe that nearest and bilinear interpolation methods produce similar spectral distributions comparing to those models trained with SR. Refer to table 1 in main paper for experiment details.

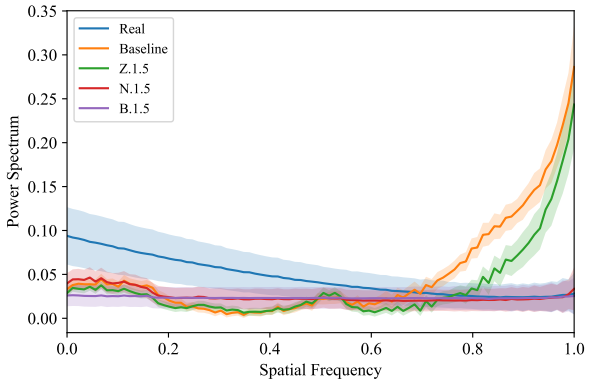

Figure E. Spectral plots for Image reconstruction at 512x512 for CelebA-HQ. The plots are averaged over 1000 samples. We show the entire spectrum similar to [10]. Similar to our other results, we observe that nearest and bilinear interpolation methods for the last upsampling step allows to obtain spectral consistent image reconstructions. We also show the standard deviation of experiments in the plot in addition to the mean.

Setup	DCGAN	LSGAN	WGAN-GP
N.1.5	$\mathbf{0.1 \pm 0\%}$	$\mathbf{0.28 \pm 0.04\%}$	$\mathbf{0.2 \pm 0\%}$
Z.1.5	$82.18 \pm 0.26\%$	$86.05 \pm 0.43\%$	$99.7 \pm 0\%$
B.1.5	$\mathbf{0 \pm 0\%}$	$\mathbf{0.1 \pm 0\%}$	$\mathbf{0.29 \pm 0.03\%}$
N.1.3	$\mathbf{0 \pm 0\%}$	$\mathbf{0 \pm 0\%}$	$\mathbf{0.3 \pm 0\%}$
N.1.7	$\mathbf{0 \pm 0\%}$	$\mathbf{0 \pm 0\%}$	$\mathbf{0.08 \pm 0.04\%}$
Z.1.3	$98.43 \pm 0.13\%$	$71.77 \pm 0.48\%$	$97.79 \pm 0.03\%$
Z.1.7	$96.55 \pm 0.07\%$	$94.59 \pm 0.09\%$	$99.9 \pm 0\%$
B.1.3	$\mathbf{0 \pm 0\%}$	$\mathbf{0.12 \pm 0.04\%}$	$\mathbf{0.1 \pm 0\%}$
B.1.7	$\mathbf{0 \pm 0\%}$	$\mathbf{0.1 \pm 0\%}$	$\mathbf{0.15 \pm 0.08\%}$
N.3.5	$\mathbf{0.2 \pm 0\%}$	$\mathbf{0 \pm 0\%}$	$\mathbf{0 \pm 0\%}$
Z.3.5	$74.07 \pm 0.68\%$	$62.17 \pm 1.05\%$	$99.97 \pm 0.05\%$
B.3.5	$\mathbf{0 \pm 0\%}$	$\mathbf{0.49 \pm 0.03\%}$	$\mathbf{0.12 \pm 0.04\%}$

Table B. Detection results for the detectors proposed by Dzanic *et al.* [12], using CelebA dataset (50% data for training). We follow exactly the procedure in [12] to train the detector for each GAN model. The table shows the successful detection rates, and we highlight the cases when the detection rates are inferior (less than 10%). The results are consistent with observations in the spectral plots.

6.2. Additional Results

Additional detection results when using 50% data to train (authors used 10% and these results are shown in the main paper) for CelebA [29], LSUN [43] and StarGAN [7] experiments are shown in table B, C, D respectively. We also conduct experiments using the reconstructed images and the detection results are shown in E.

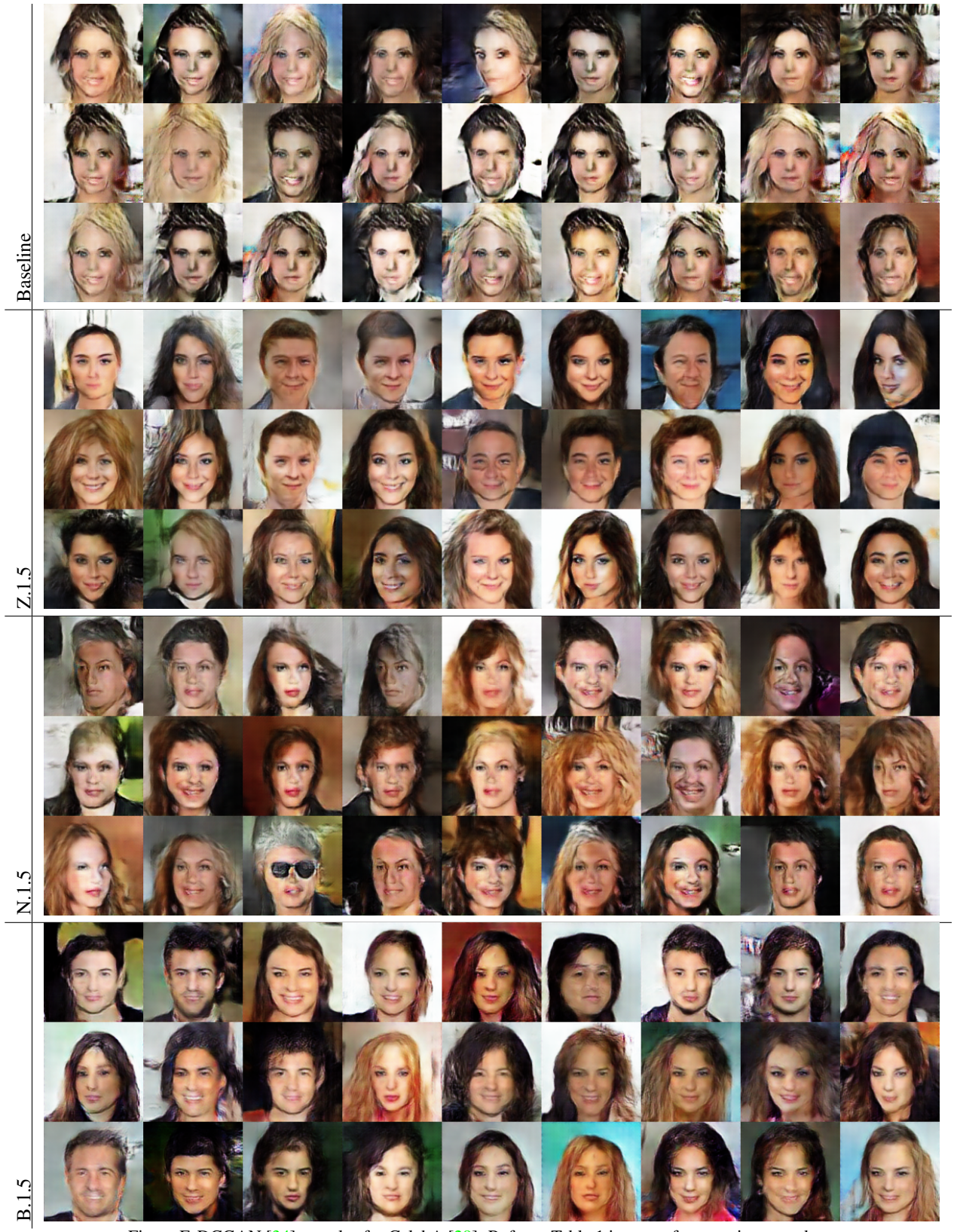

Figure F. DCGAN [34] samples for CelebA [29]. Refer to Table 1 in paper for experiment codes.

Figure G. LSGAN [31] samples for CelebA [29]. Refer to Table 1 in paper for experiment codes.

Figure H. WGAN-GP [17] samples for CelebA [29]. Refer to Table 1 in paper for experiment codes.

Figure I. Image Translation results using StarGAN. Original Image (leftmost), Baseline (column 2), Z.1.5 (column 3), N.1.5 (column 4), B.1.5 (rightmost) for attribute Blonde hair is shown. Corresponding high frequency spectral distributions are shown in Figure J. Refer to table 1 in paper for experiment codes.

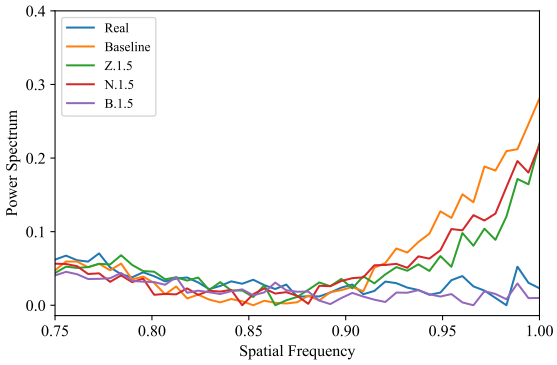

Figure J. Spectral distributions for examples shown in I. Real refers to Original image. We observe that B.1.5 setup produces spectral consistent images similar to observations recorded in the paper. Refer to table 1 in paper for experiment codes.

Setup	DCGAN	LSGAN	WGAN-GP
N.1.5	$\mathbf{0.3 \pm 0\%}$	$\mathbf{0.6 \pm 0\%}$	$\mathbf{0 \pm 0\%}$
Z.1.5	$98.41 \pm 0.15\%$	$94.84 \pm 0.07\%$	$99.93 \pm 0.05\%$
B.1.5	$\mathbf{0.1 \pm 0\%}$	$\mathbf{0 \pm 0\%}$	$\mathbf{0.06 \pm 0.05\%}$

Table C. Detection results for the detectors proposed by Dzanic *et al*. [12], using LSUN Bedrooms [43] dataset (50% data for training). The table shows the successful detection rates, and we highlight the cases when the detection rates are inferior (less than 10%).

Setup	N.1.5	Z.1.5	B.1.5
Accuracy	$53.7 \pm 0.15\%$	$64.61 \pm 0.37\%$	$\mathbf{0 \pm 0\%}$

Table D. Detection results for the forensics classifiers proposed by Dzanic [12], using CelebA [29] dataset (256x256) for StarGAN (50% data for training). We observe that B.1.5 samples easily bypasses the classifier.

7. Stronger Classifiers

7.1. SVM and MLP classifiers

We perform additional experiments using SVM and MLP classifiers (exact same setup as [12], only change in classifiers). The results are shown in Table F and Table G.

Setup	10% train data	50% train data
N.1.5	$\mathbf{0 \pm 0\%}$	$\mathbf{0 \pm 0\%}$
Z.1.5	$94.8 \pm 1.75\%$	$95.44 \pm 0.26\%$
B.1.5	$\mathbf{0 \pm 0\%}$	$\mathbf{0 \pm 0\%}$

Table E. Detection results for the forensics classifiers proposed by Dzanic [12], using reconstructed images. We observe that N.1.5 and B.1.5 samples can easily bypass the classifier.

Setup	DCGAN	LSGAN	WGAN-GP
N.1.5	$\mathbf{0.1 \pm 0\%}$	$\mathbf{0.31 \pm 0.06\%}$	$\mathbf{0.23 \pm 0.16\%}$
Z.1.5	$82.22 \pm 1.98\%$	$87.33 \pm 2.77\%$	$99.45 \pm 0.21\%$
B.1.5	$\mathbf{0 \pm 0\%}$	$\mathbf{0.11 \pm 0.09\%}$	$\mathbf{0.25 \pm 0.17\%}$
N.1.3	$\mathbf{0.01 \pm 0.03\%}$	$\mathbf{0.07 \pm 0.05\%}$	$\mathbf{0.35 \pm 0.22\%}$
N.1.7	$\mathbf{0 \pm 0\%}$	$\mathbf{0 \pm 0\%}$	$\mathbf{0.05 \pm 0.05\%}$
Z.1.3	$98.3 \pm 0.45\%$	$72.13 \pm 2.21\%$	$96.81 \pm 1.63\%$
Z.1.7	$95.81 \pm 0.93\%$	$95.55 \pm 1.23\%$	$99.24 \pm 0.43\%$
B.1.3	$\mathbf{0 \pm 0\%}$	$\mathbf{0.25 \pm 0.12\%}$	$\mathbf{0.15 \pm 0.15\%}$
B.1.7	$\mathbf{0 \pm 0\%}$	$\mathbf{0.11 \pm 0.03\%}$	$\mathbf{0.3 \pm 0.27\%}$
N.3.5	$\mathbf{0.1 \pm 0\%}$	$\mathbf{0 \pm 0\%}$	$\mathbf{0 \pm 0\%}$
Z.3.5	$74.27 \pm 3.32\%$	$65.37 \pm 6.5\%$	$93.82 \pm 0.6\%$
B.3.5	$\mathbf{0.04 \pm 0.07\%}$	$\mathbf{0.5 \pm 0.05\%}$	$\mathbf{0.21 \pm 0.14\%}$

Table F. Detection rates using SVM (RBF kernel) using same features as Dzanic *et al*. [12].

Our results are consistent: even with SVM/MLP classifiers, we can bypass them by replacing zero insertion last layer with nearest (N) or bilinear (B). For all experiments, we use 10% data for training, and the reported results are averaged over 10 runs. We also observe similar detection rates when using 50% and 80% data for training.

7.2. Using entire spectrum as features

We followed similar setup as Durall *et al*. [10] to train a SVM classifier using entire spectrum as features (88 dimensional features for 128x128 images). The finding is consistent: we observe that the features are still non-separable when using nearest (N) and bilinear (B) for the last upsampling step. The results are shown in Table H.

Figure K. Reconstruction Results. Original Image (leftmost), Baseline (column 2), Z.1.5 (column 3), N.1.5 (column 4), B.1.5 (rightmost) for CelebA-HQ [21] samples are shown. Refer to table 1 in paper for experiment codes.

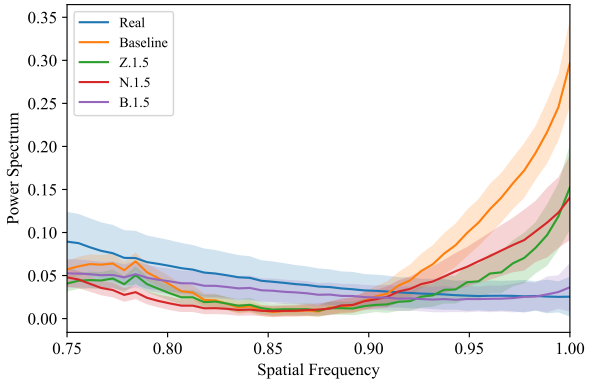

Figure L. This figure shows spectral plots from Figure 10 (StarGAN) in the paper, with standard deviations indicated

Setup	DCGAN	LSGAN	WGAN-GP
N.1.5	$\mathbf{0.1 \pm 0\%}$	$\mathbf{0.77 \pm 0.15\%}$	$\mathbf{1.53 \pm 0.32\%}$
Z.1.5	$81.14 \pm 2.9\%$	$83.88 \pm 2.59\%$	$99.77 \pm 0.09\%$
B.1.5	$\mathbf{0.04 \pm 0.1\%}$	$\mathbf{0.87 \pm 0.46\%}$	$\mathbf{3.03 \pm 0.82\%}$
N.1.3	$\mathbf{0.18 \pm 0.04\%}$	$\mathbf{0.05 \pm 0.13\%}$	$\mathbf{1.4 \pm 0.2\%}$
N.1.7	$\mathbf{0 \pm 0\%}$	$\mathbf{0.04 \pm 0.05\%}$	$\mathbf{0.67 \pm 0.18\%}$
Z.1.3	$97.54 \pm 0.41\%$	$72.65 \pm 2.64\%$	$98.11 \pm 0.44\%$
Z.1.7	$94.53 \pm 0.97\%$	$93.07 \pm 1.6\%$	$99.97 \pm 0.05\%$
B.1.3	$\mathbf{0.03 \pm 0.09\%}$	$\mathbf{1.6 \pm 0.54\%}$	$\mathbf{2.79 \pm 0.5\%}$
B.1.7	$\mathbf{0.01 \pm 0.03\%}$	$\mathbf{0.42 \pm 0.29\%}$	$\mathbf{4.63 \pm 1.01\%}$
N.3.5	$\mathbf{0.17 \pm 0.05\%}$	$\mathbf{0 \pm 0\%}$	$\mathbf{0.37 \pm 0.27\%}$
Z.3.5	$74.88 \pm 2.79\%$	$71.22 \pm 4.46\%$	$99.8 \pm 0\%$
B.3.5	$\mathbf{0.28 \pm 0.14\%}$	$\mathbf{1.89 \pm 0.45\%}$	$\mathbf{3.66 \pm 1.19\%}$

Table G. Detection rates using MLP (2 hidden layers of size 10 with sigmoid activation) using same features as Dzanic *et al.* [12].

8. Virtual KITTI

GANs belong to a larger family of computational image synthesis algorithms. In this section, we investigate the high frequency decay attributes of data created entirely using Unity Game Engine. We compare the spectral behaviour between the Official KITTI tracking benchmark [15] (Real images) and the Virtual KITTI [14] (Synthesized images). Virtual KITTI [14] recreates real-world videos from the KITTI tracking benchmark [15] inside Unity[2] game engine.

[2] https://unity.com/

Setup	DCGAN	LSGAN	WGAN-GP
N.1.5	$\mathbf{0.12 \pm 0.04\%}$	$\mathbf{1.01 \pm 0.29\%}$	$\mathbf{0.08 \pm 0.06\%}$
Z.1.5	$95.91 \pm 2.43\%$	$98.29 \pm 0.85\%$	$96.62 \pm 1.61\%$
B.1.5	$\mathbf{0.09 \pm 0.03\%}$	$\mathbf{0.24 \pm 0.11\%}$	$\mathbf{0 \pm 0\%}$
N.1.3	$\mathbf{0.26 \pm 0.07\%}$	$\mathbf{0.9 \pm 0.2\%}$	$\mathbf{0.08 \pm 0.04\%}$
N.1.7	$\mathbf{0 \pm 0\%}$	$\mathbf{0.09 \pm 0.12\%}$	$\mathbf{0.12 \pm 0.04\%}$
Z.1.3	$97.72 \pm 1.86\%$	$88.44 \pm 2.55\%$	$91.7 \pm 3.25\%$
Z.1.7	$98.43 \pm 1.43\%$	$99.29 \pm 0.5\%$	$97.09 \pm 2.8\%$
B.1.3	$\mathbf{0 \pm 0\%}$	$\mathbf{1.11 \pm 0.3\%}$	$\mathbf{0 \pm 0\%}$
B.1.7	$\mathbf{0.01 \pm 0.03\%}$	$\mathbf{0.16 \pm 0.1\%}$	$\mathbf{0.17 \pm 0.05\%}$
N.3.5	$\mathbf{0 \pm 0\%}$	$10.85 \pm 7.05\%$	$\mathbf{0 \pm 0\%}$
Z.3.5	$97.32 \pm 1.24\%$	$97.79 \pm 1.4\%$	$98.69 \pm 1.77\%$
B.3.5	$\mathbf{0.23 \pm 0.05\%}$	$\mathbf{1.09 \pm 0.17\%}$	$\mathbf{0.02 \pm 0.04\%}$

Table H. Detection rates using classifier proposed by Durall *et al.* [10]. Following [10], we use entire 1D spectrum as features.

We show some samples in Figure M. We show the entire power spectrum in Figure N and we observe that the images synthesized from the game engine do not have high frequency spectral discrepancies.

9. CRN/ IMLE

We also observe that high frequency decay discrepancies are not seen in some out-of-the-box GAN models. Specifically, we observe that CRN [6] and IMLE [28] GANs do not have such discrepancies. We show the entire power spectrum for CRN [6] and IMLE [28] GANs in Figure O and P respectively. Do note that both these models are pre-trained on GTA game data (Another instance of data synthesized from game engines). This further helps to confirm that such discrepancies are not intrinsic.

10. Dataset Details

For CelebA [29] experiments, we use the officially released train subset consisting of 162, 770 images. For LSUN [43] experiments, we select a random subset of 200, 000 images for training. For StarGAN experiments, we follow the official implementation. For autoencoder experiments, we select a random subset of 1000 images from CelebA-HQ [21].

Figure M. Samples of real images from KITTI tracking benchmark [15] dataset and the recreated images using Unity game engine obtained from Virtual KITTI [14] dataset (Right)

.

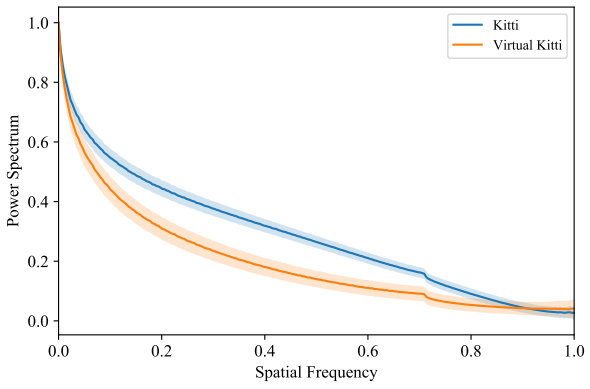

Figure N. This figure shows the spectral plots for all the frequencies for KITTI [15] and Virtual KITTI [14] datasets. All the images were center cropped to 370x370. We observe that the Virtual KITTI images do not have high frequency discrepancies.

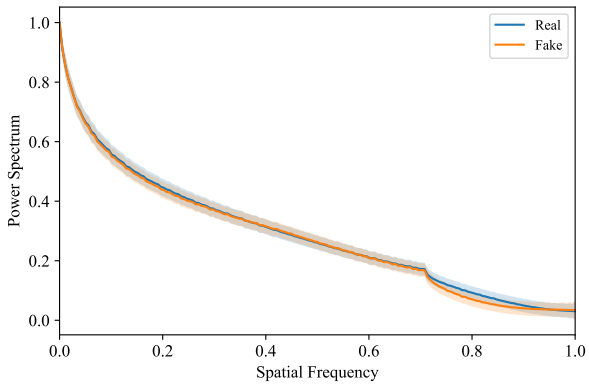

Figure O. This figure shows the spectral plots for all the frequencies for GTA (Real) images and CRN [6] synthesized images. All the images were center cropped to 256x256. We observe that the CRN generated images do not have high frequency discrepancies.

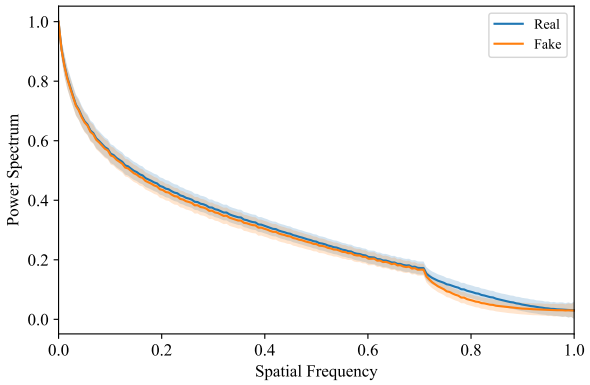

Figure P. This figure shows the spectral plots for all the frequencies for GTA (Real) images and IMLE [28] synthesized images. All the images were center cropped to 256x256. We observe that the IMLE synthesized images do not have high frequency discrepancies.

11. Implementation Details

For GAN experiments, we use Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and batch size 64. For all CelebA [29] experiments, we used an initial learning rate = 2×10^{-4}. The learning rate was reduced based on FID scores for all experiments. For DCGAN, LSGAN and WGAN-GP experiments, we use the GitHub code used by the Spectral Regularization paper[10] [3]. For StarGAN [7], we use the official implementation [4] with default hyper-parameters. For Spectral Regularization [10] experiments, we use the officially released code [5].

For all autoencoder experiments, we use Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, batch size 128, initial learning rate 2.5×10^{-3} and learning rate decay scheme that scales the learning rate by 0.9 when reconstruction error plateaus.

For Fourier synthetic detector [12] experiments, we use the officially released matlab code [6] for feature extraction and use our own script to perform KNN classification. For FID calculation, we used the open-source Pytorch FID implementation [7].

More details on hyper-parameters and research reproducibility can be found in: https://keshik6.github.io/Fourier-Discrepancies-CNN-Detection/

[3] https://github.com/LynnHo/DCGAN-LSGAN-WGAN-GP-DRAGAN-Pytorch

[4] https://github.com/yunjey/stargan

[5] https://github.com/cc-hpc-itwm/UpConv

[6] https://github.com/tarikdzanic/FourierSpectrumDiscrepancies

[7] https://github.com/mseitzer/pytorch-fid

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[2] Paul Baines. Uk election 2019: after fake keir starmer clip, how much of a problem are doctored videos?, Aug 2020.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019.

[5] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 103–120, Cham, 2020. Springer International Publishing.

[6] Qifeng Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1520–1529, 2017. 8, 9

[7] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 3, 10

[8] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[9] Danielle Citron and Robert Chesney. Deepfakes and the new disinformation war, Jun 2020.

[10] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3, 7, 8, 10

[11] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features, 2020.

[12] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. In *Thirty-fourth Annual Conference on Neural Information Processing Systems (NeurIPS)*, December 2020. 2, 3, 7, 8, 10

[13] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3247–3258. PMLR, 13–18 Jul 2020.

[14] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtualworlds as proxy for multi-object tracking analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4340–4349, 2016. 8, 9

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 8, 9

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

[17] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5767–5777. Curran Associates, Inc., 2017. 1, 6

[18] Karen Hao and Will Douglas Heaven. The year deepfakes went mainstream, Dec 2020.

[19] Ellie Harrison. Shockingly realistic tom cruise deepfakes go viral on tiktok, Feb 2021.

[20] Anil K. Jain. *Fundamentals of Digital Image Processing*. Prentice-Hall, Inc., USA, 1989.

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1, 8

[22] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12104–12114. Curran Associates, Inc., 2020.

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[25] Mahyar Khayatkhoei and Ahmed Elgammal. Spatial frequency bias in convolutional generative adversarial networks, Oct 2020.

[26] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of gans, 2017.

[27] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017.

[28] K. Li, T. Zhang, and J. Malik. Diverse image synthesis from semantic layouts via conditional imle. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4219–4228, 2019. 8, 10

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 1, 2, 3, 4, 5, 6, 7, 8, 10

[30] Sophie Maddocks. 'a deepfake porn plot intended to silence me': exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, 0(0):1–9, 2020.

[31] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, 2017. 1, 5

[32] Rachel Metz. The number of deepfake videos online is spiking. most are porn, Oct 2019.

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[34] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016. 1, 4

[35] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In H. Wallach, H. Larochelle, A. Beygelzimer,

F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 14866–14876. Curran Associates, Inc., 2019.

[36] Tom Simonite. What happened to the deepfake threat to the election?, Nov 2020.

[37] Daniel Thomas. Deepfakes: A threat to democracy or just a bit of fun?, Jan 2020.

[38] Ngoc-Trung Tran, Tuan-Anh Bui, and N. Cheung. Dist-gan: An improved gan using distance constraints. In *ECCV*, 2018.

[39] Ngoc-Trung Tran, Viet-Hung Tran, Bao-Ngoc Nguyen, Linxiao Yang, and Ngai-Man (Man) Cheung. Self-supervised gan: Analysis and improvement with multi-class minimax game. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[40] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6306–6315. Curran Associates, Inc., 2017.

[41] A. van der Schaaf and J.H. van Hateren. Modelling the power spectra of natural images: Statistics and information. *Vision Research*, 36(17):2759 – 2770, 1996.

[42] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[43] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1, 3, 7, 8

[44] X. Zhang, S. Karaman, and S. Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, 2019.

[45] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.