SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

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

Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Ngai-Man Cheung
{ keshigeyan, ngoctrung_tran, ngaiman_cheung } @sutd.edu.sg

CVPR 2021 Oral

CNN-generated images have become indistinguishable from real images.

Left to right : Real, StyleGAN, StyleGAN2, PGGAN, VQ-VAE2, and ALAE generated images. [Dzanic et al. NeurIPS 2020]

With serious concerns over Deepfakes being widely used for malicious purposes, detection of deepfake multimedia content has become an important research field.

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

How to build detectors to identify CNN-generated images?

SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

How to build detectors to identify CNN-generated images?

Deep Neural Network
based detectors

Hand-crafted feature
based detectors

SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

How to build detectors to identify CNN-generated images?

Deep Neural Network based detectors

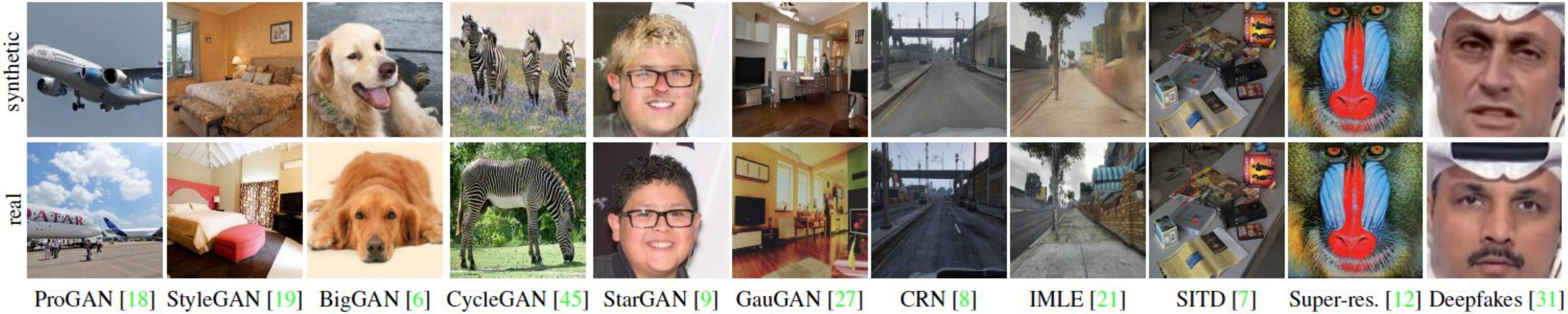
- Good detection accuracies
- Complex DNN architectures
- Computationally intensive
- Black box : learnt features used for detection are not explicit

Hand-crafted feature based detectors

- Good detection accuracies (Dzanic et al. NeurIPS 2020, Durall et al. CVPR 2020)
- Simple
- Lightweight
- Transparent

Deep Neural Network (ResNet-50) based detector

[Wang et al. CVPR 2020]



A classifier trained to detect images generated by only one CNN (ProGAN, far left) can detect those generated by many other models (remaining columns).

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

How to build detectors to identify CNN-generated images?

Deep Neural Network based detectors

- Good detection accuracies
- Complex DNN architectures
- Computationally intensive
- Black box : learnt features used for detection are not explicit

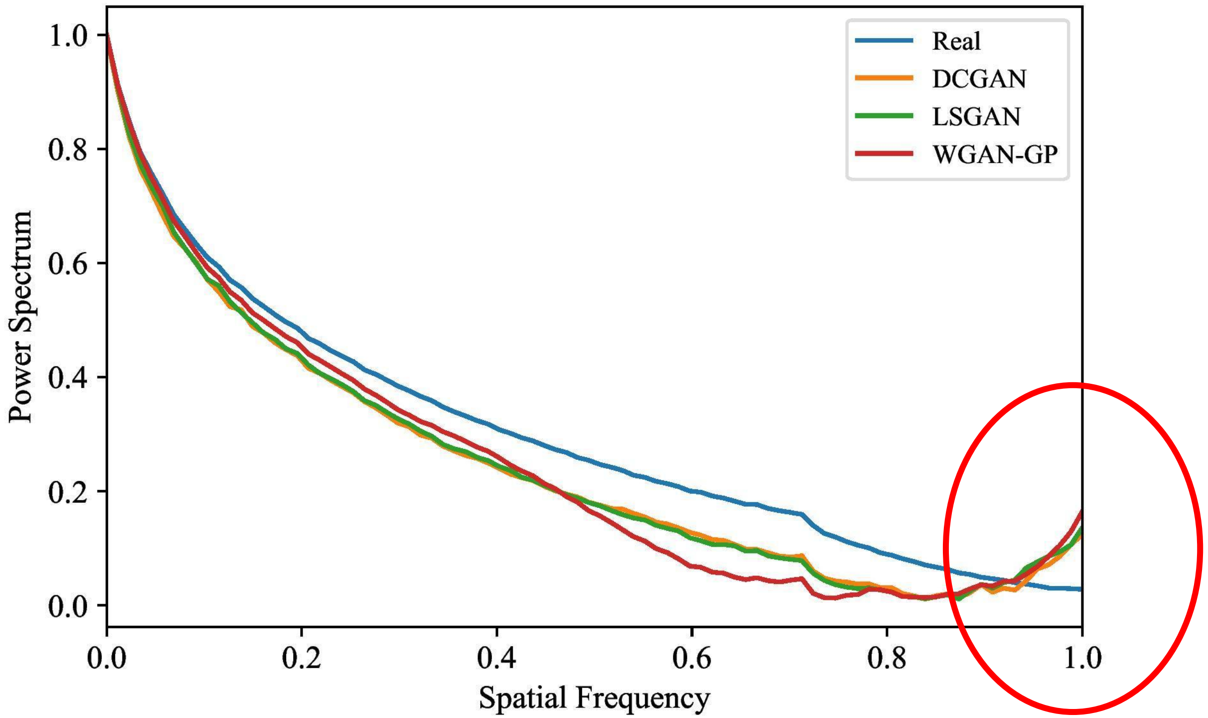
Hand-crafted feature based detectors

- Good detection accuracies (Dzanic et al. NeurIPS 2020, Durall et al. CVPR 2020)
- Simple
- Lightweight
- Transparent

SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

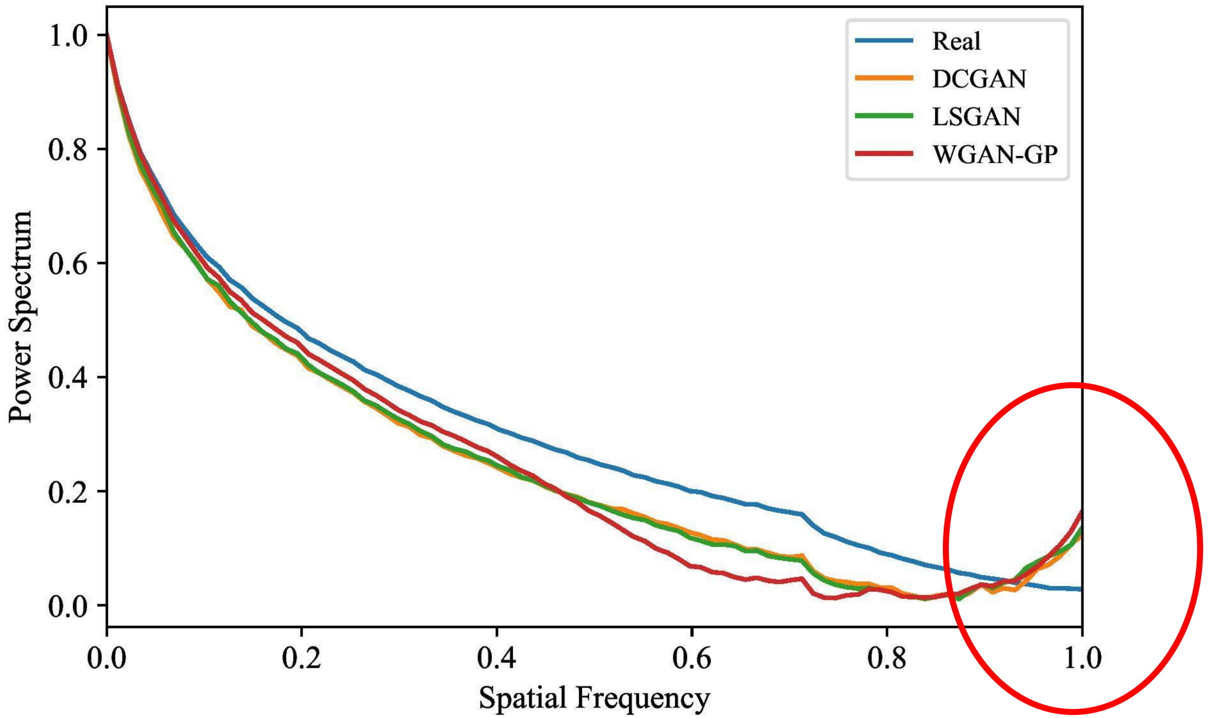
Recently some works have observed that CNN-generated images
share a systematic shortcoming in replicating
high frequency Fourier spectrum decay attributes.

Fourier Spectrum Decay Discrepancies in CNN-generated Images



CNN generated images at the highest frequencies do not decay as usually observed in real images.

Fourier Spectrum Decay Discrepancies in CNN-generated Images

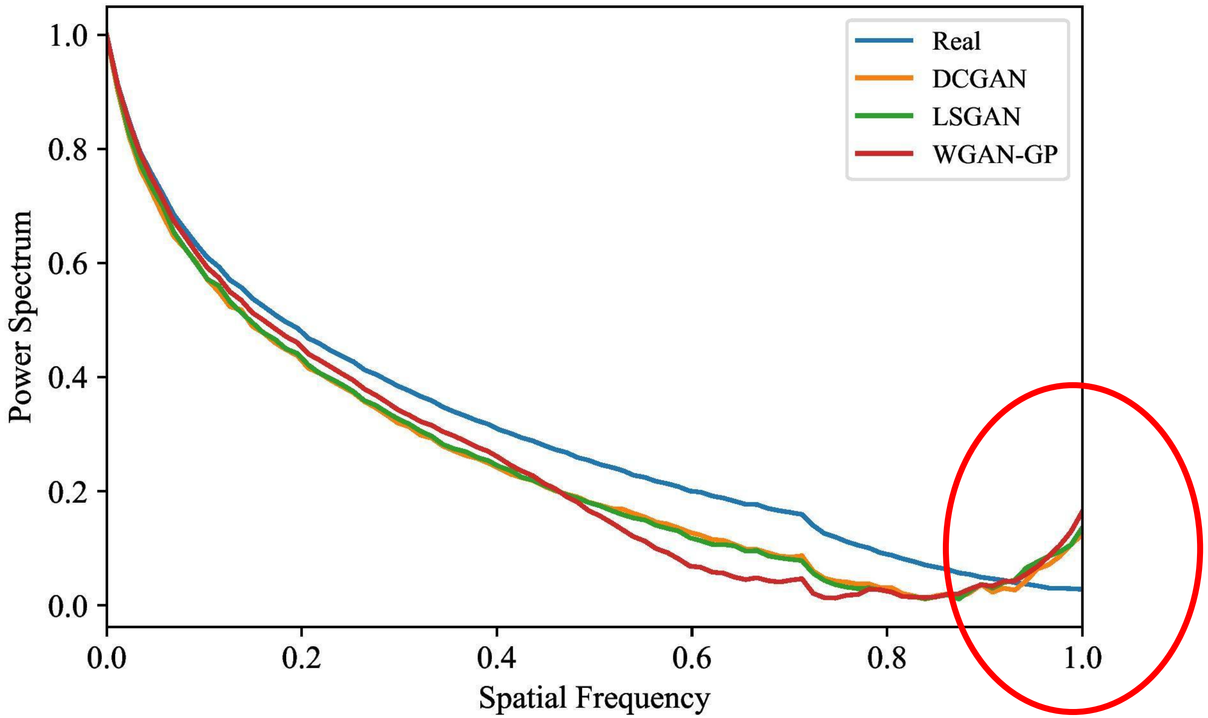


" CNN-based generative deep neural networks are failing to reproduce spectral distributions. "

" This effect is independent of the underlying architecture "

[Durall et al. CVPR 2020]

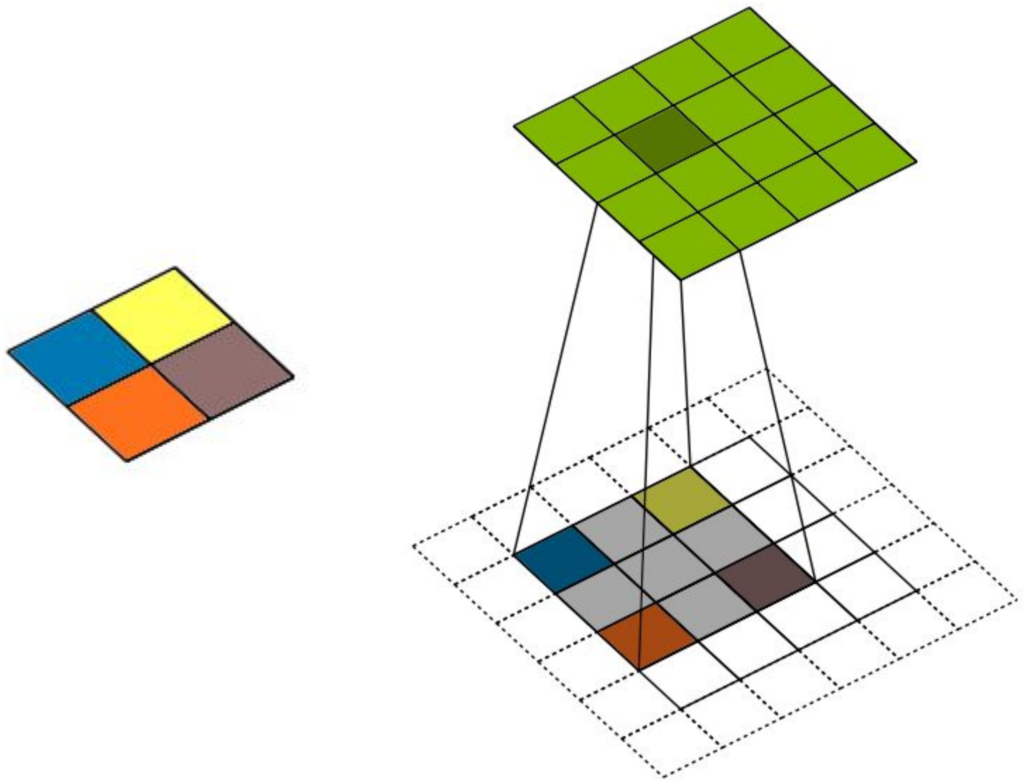
Fourier Spectrum Decay Discrepancies in CNN-generated Images



"deep network generated images share an observable, systematic shortcoming in replicating the attributes of these high-frequency modes. "

[Dzanic et al. NeurIPS 2020]

Fourier Spectrum Decay Discrepancies in CNN-generated Images



Transposed convolution: up-sampling with zero insertion + convolution with a filter kernel creates high-frequency artifacts

[Durall et al. CVPR 2020]

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Detector : High frequency decay attributes + kNN classifier

[Dzanic et al. NeurIPS 2020]

Experiment	Resolution	Compression Quality	Overall Class. Acc.	StyleGAN Class. Acc.	StyleGAN2 Class. Acc.	PGGAN Class. Acc.	VQ-VAE2 Class. Acc.	ALAE Class. Acc.
A	1024^2	100	**99.2%**	99.9%	99.5%	97.4%	99.8%	99.8%
B	1024^2	95	**94.4%**	99.2%	88.5%	88.5%	100 %	99.7%
C	1024^2	85	**83.9%**	78.9%	65.9%	78.7%	99.6%	87.4%
D	768^2	100	**98.5%**	100 %	99.1%	95.9%	99.9%	99.9%
E	768^2	95	**93.0%**	97.9%	85.4%	87.3%	100 %	99.5%
F	768^2	85	**84.6%**	77.1%	68.6%	79.3%	99.6%	85.7%
G	256^2	100	**88.8%**	85.0%	87.4%	69.0%	92.0%	90.7%
H	256^2	95	**88.1%**	81.7%	83.4%	68.2%	92.2%	87.7%
I	256^2	85	**87.4%**	67.8%	79.3%	64.8%	87.7%	80.6%

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

[Chandrasegaran, Tran, Cheung ; CVPR 2021 Oral]

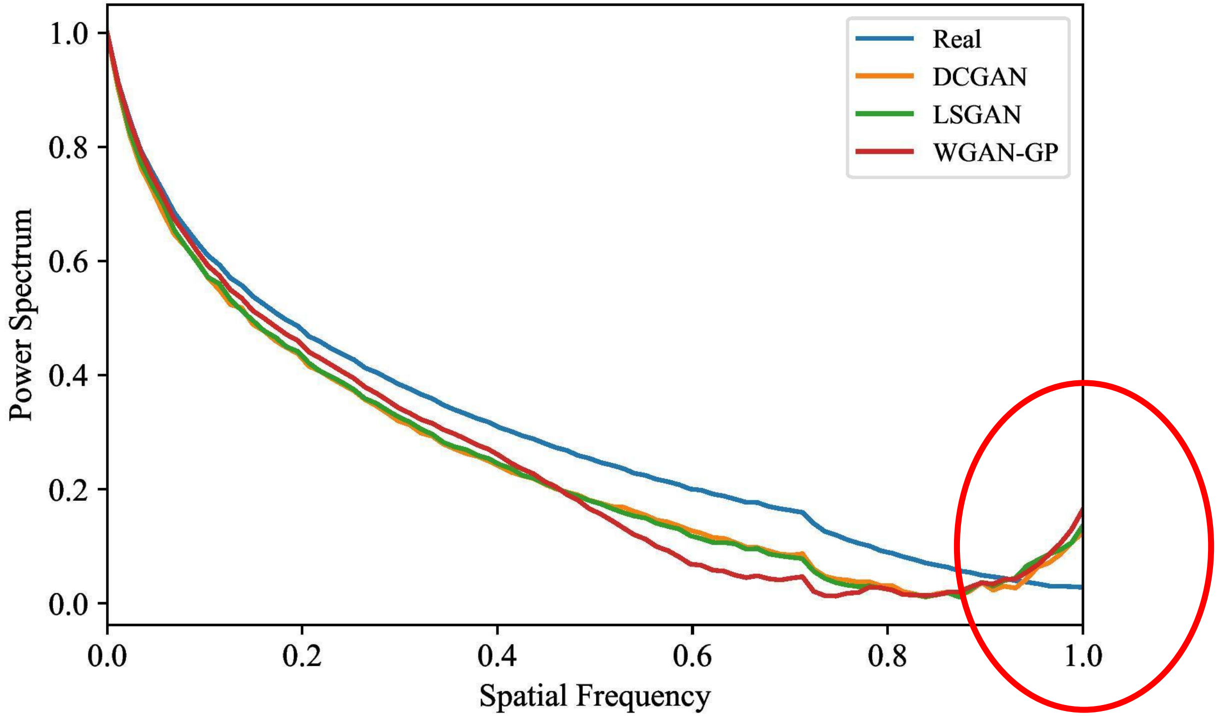
SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

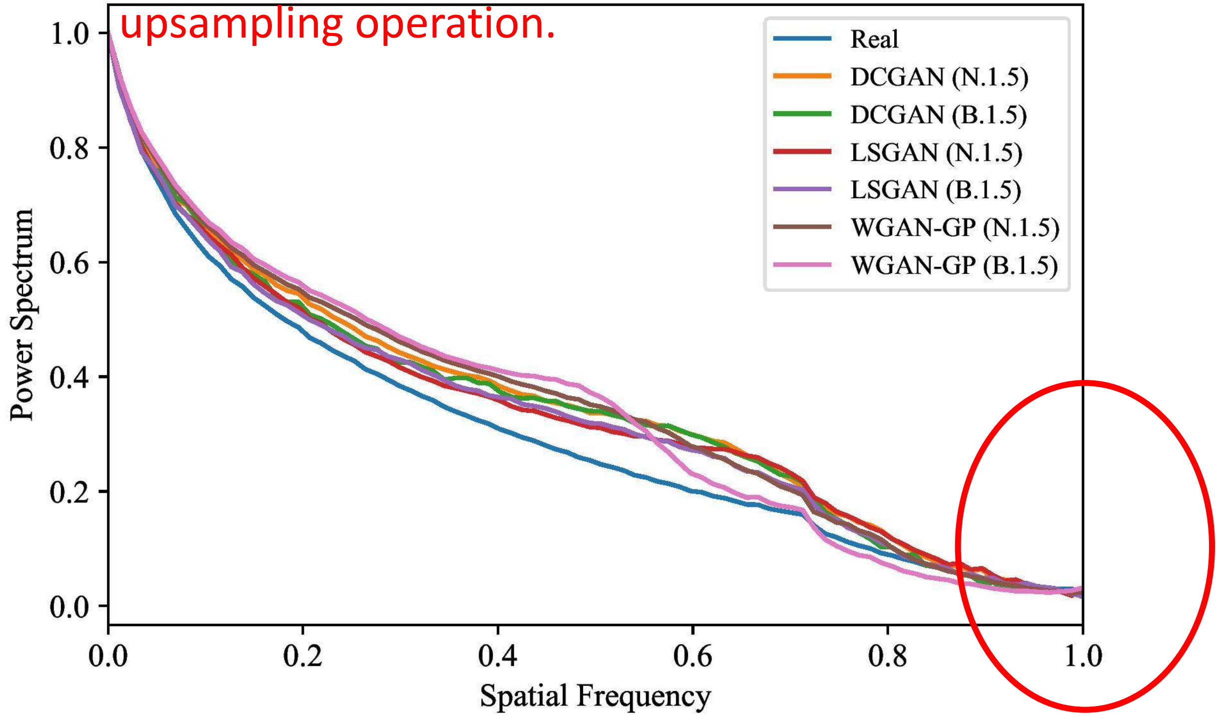
"High frequency spectral decay discrepancies are not intrinsic for CNN-generated images. Therefore, we urge re-thinking in using such features for CNN generated image detection."

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

[Chandrasegaran, Tran, Cheung ; CVPR 2021 Oral]

High frequency spectral decay discrepancy avoided by a simple architecture change in the last upsampling operation.





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

[Chandrasegaran, Tran, Cheung ; CVPR 2021 Oral]

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Setup	DCGAN	LSGAN	WGAN-GP
N.1.5	**0.09 ± 0.03%**	**0.34 ± 0.08%**	**0.14 ± 0.05%**
Z.1.5	84.82 ± 3.72%	88.16 ± 3.98%	99.75 ± 0.14%
B.1.5	**0 ± 0%**	**0.1 ± 0%**	**0.2 ± 0.12%**
N.1.3	**0 ± 0%**	**0.06 ± 0.05%**	**0.24 ± 0.13%**
N.1.7	**0 ± 0%**	**0 ± 0%**	**0.06 ± 0.05%**
Z.1.3	98.73 ± 0.56%	73.09 ± 3.5%	97.94 ± 0.87%
Z.1.7	97.23 ± 1.1%	95.66 ± 1.93%	99.94 ± 0.07%
B.1.3	**0 ± 0%**	**0.19 ± 0.1%**	**0.07 ± 0.05%**
B.1.7	**0 ± 0%**	**0.1 ± 0%**	**0.17 ± 0.13%**
N.3.5	**0.16 ± 0.05%**	**0 ± 0%**	**0 ± 0%**
Z.3.5	77.67 ± 6%	67.66 ± 11.9%	99.9 ± 0.19%
B.3.5	**0.03 ± 0.05%**	**0.48 ± 0.04%**	**0.13 ± 0.05%**

Table 2 : Detection results for the detector proposed by Dzanic et al.

We successfully by-pass the detector proposed by Dzanic et al. (NeurIPS 2020) that uses high frequency decay attributes as features.

We emphasize that we obtain these results using identical training algorithms, objective functions and network architectures (except using a different upscaling in the last upsampling operation).

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

[Chandrasegaran, Tran, Cheung ; CVPR 2021 Oral]

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

We also successfully by-pass stronger classifiers that use high frequency decay attributes as features

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

Table 10 : Detection rates using SVM (RBF kernel) using same features as Dzanic et al.

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

Table 11 : Detection rates using MLP (2 hidden layers of size 10 with sigmoid activation) using same features as Dzanic et al.

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

[Chandrasegaran, Tran, Cheung ; CVPR 2021 Oral]

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

FID scores of nearest and bilinear interpolation methods are comparable or better than the Baseline FID

Figure 4: WGAN-GP samples for Baseline (Left), N.1.5 (Middle) and B.1.5 (Right) for CelebA. We observe that the visual quality is comparable when replacing the last transpose convolutions with nearest and bilinear methods. FID scores for Baseline (Left), N.1.5 (Middle) and B.1.5 (Right) setups are 60.6, 48.69 and 52.18 respectively. (Measured using 50k real and generated samples.)

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

[Chandrasegaran, Tran, Cheung ; CVPR 2021 Oral]

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

"High frequency spectral decay discrepancies are not intrinsic for CNN-generated images. Therefore, we urge re-thinking in using such features for CNN generated image detection."

Code and pre-trained models available

Tarik Dzanic, Karan Shah, & Freddie Witherden (2020). Fourier Spectrum Discrepancies in Deep Network Generated Images. *(NeurIPS)*.

Durall, R., Keuper, M., & Keuper, J. (2020). Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. *(CVPR)*.

Wang, S.Y., Wang, O., Zhang, R., Owens, A., & Efros, A. (2020). CNN-Generated Images Are Surprisingly Easy to Spot... for Now. *(CVPR)*.