- Supplemental Material -
Watch your Up-Convolution: CNN Based Generative Deep Neural Networks are Failing to Reproduce Spectral Distributions

The supplementary material of our paper contains additional details on the presented experiments, as well as some support experiments that might help to get a better understanding of the spectral properties of up-convolution units.

1. Using Spectral Distortions to Detect Deepfakes

In this section, we provide more detailed results of the experiments presented in section 4.1 of the paper.

1.1. More Details on the used Datasets

1.1.1 Faces-HQ

To the best of our knowledge, currently no public dataset is providing high resolution images with annotated fake and real faces. Therefore, we have created our own data set from established sources, called *Faces-HQ*[1]. In order to have a sufficient variety of faces, we have chosen to download and label the images available from the *CelebA-HQ* data set [3], *Flickr-Faces-HQ* data set [4], 100K Faces project [1] and *www.thispersondoesnotexist.com*. In total, we have collected 40K high quality images, half of them real and the other half fake faces. Table 1 contains a summary.

Training Setting: we divide the transformed data into training and testing sets, with 20% for the testing stage and use the remaining 80% as the training set. Then, we train a classifier with the training data and finally evaluate the accuracy on the testing set.

	# of samples	category	label
CelebA-HQ data set [3]	10000	Real	0
Flickr-Faces-HQ data set [4]	10000	Real	0
100K Faces project [1]	10000	Fake	1
www.thispersondoesnotexist.com	10000	Fake	1

Table 1: *Faces-HQ* data set structure.

[1]Faces-HQ data has a size of 19GB. Download: https://cutt.ly/6enDLYG

1.1.2 CelebA

The CelebFaces Attributes (*CelebA*) dataset [6] consists of 202,599 celebrity face images with 40 variations in facial attributes. The dimensions of the face images are 178x218x3, which can be considered to be a medium-resolution in our context.

Training Setting: While we can use the real images from the *CelebA* dataset directly, we need to generate the fake examples on our own.Therefore we use the real dataset to train one DCGAN [8], one DRAGAN [5], one LSGAN [7] and one WGAN-GP [2] to generate realistic fake images. We split the dataset into 162,770 images for training and 39,829 for testing, and we crop and resize the initial 178x218x3 size images to 128x128x3. Once the model is trained, we can conduct the classification experiments on medium-resolution scale.

1.1.3 FaceForensics++

FaceForensics++ [9] is a collection of image forensic datasets, containing video sequences that have been modified with different automated face manipulation methods. One subset is the DeepFakeDetection Dataset, which contains 363 original sequences from 28 paid actors in 16 different scenes as well as over 3000 manipulated videos using DeepFakes and their corresponding binary masks. All videos contain a trackable, mostly frontal face without occlusions which enables automated tampering methods to generate realistic forgeries.

Training Setting: the employed pipeline for this dataset is the same as for *Faces-HQ* dataset and *CelebA*, but with an additional block. Since the DeepFakeDetection dataset contains videos, we first need to extract the frame and then crop the inner faces from them. Due to the different content of the scenes of the videos, these cropped faces have different sizes. Therefore, we interpolate the 1D Power Spectrum to a fix size (300) and normalizes it dividing it by the 0^{th} frequency component.

1.2. Experimental Results

1.2.1 Spectral Distributions

The following figures 1, 2 and 3 show the spectral (AI) distributions of all datasets. In all three cases, it is evident that a classifier should be able to separate real and fake samples. Also, based on our theoretical analysis (see section 2.3 in the paper), one can assume that the generators in used FaceHQ and FaceForensics++ datasets used *up+conv* based upconvolutions or successively blurred the generated images (due to the drop in high frequencies). CelebA based fakes used *transconv*.

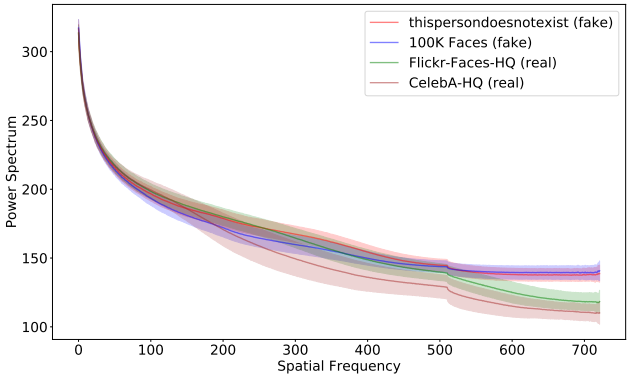

Figure 1: Statistics (mean and variance) of the Faces-HQ dataset.

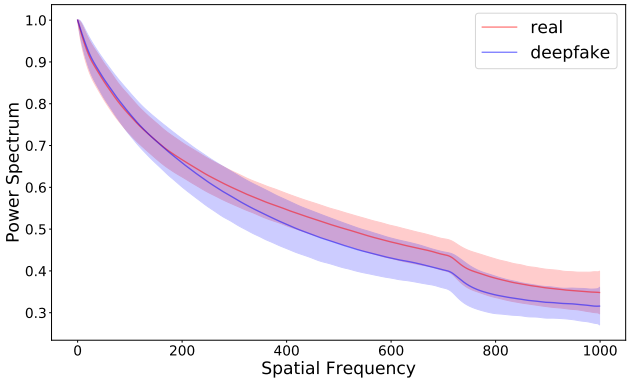

Figure 2: Statistics (mean and variance) of the FaceForensics++, DeepFakeDetection dataset.

Figure 4 gives some additional data examples and their according spectral properties for the FaceForensics++ data.

1.2.2 T-SNE Evaluation

Figure 5 shows the clustering properties of our AI features. It is quite obvious that a classifier should not have problems to separate both classes (real and fake).

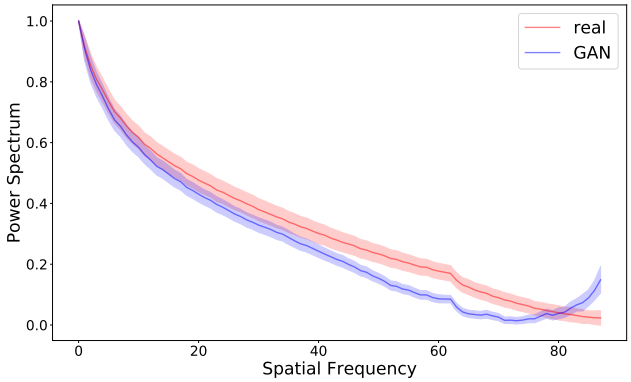

Figure 3: Statistics (mean and variance) of the CelebA dataset: average of images generated by the different GAN schemes (DCGAN, DRAGAN, LSGAN and WGAN-GP).

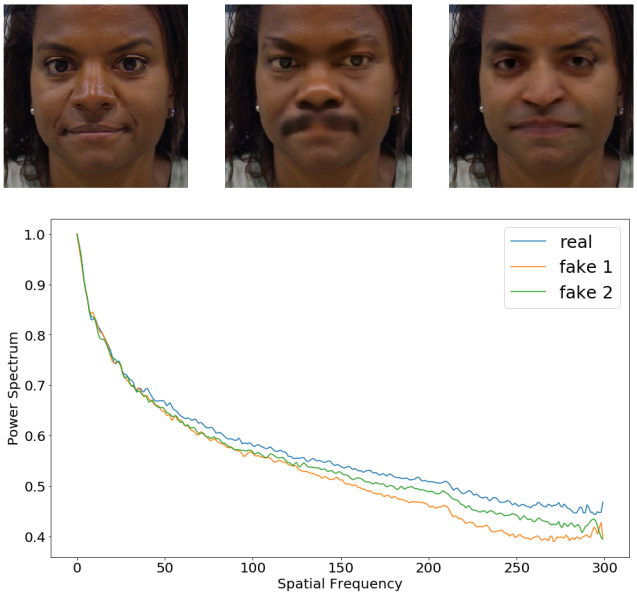

Figure 4: FaceForensics++ data. **Top:** example of one real face (left) and two deepfake faces, fake 1 (center) and fake 2 (right). Notice that the modifications only affect the inner face. **Bottom:** normalized and interpolated 1D Power Spectrum from the previous images.

1.2.3 Detection Results Depending on the Number of Available Samples

In this section, we show some additional results on the DeepFake detection task (table 1 in the paper). In tables 2, 3 and 4, we focus on the effect of the available number of data samples during training. As shown in the paper, our approach works quite well in an unsupervised setting and needs as little as 16 annotated training samples to achieve

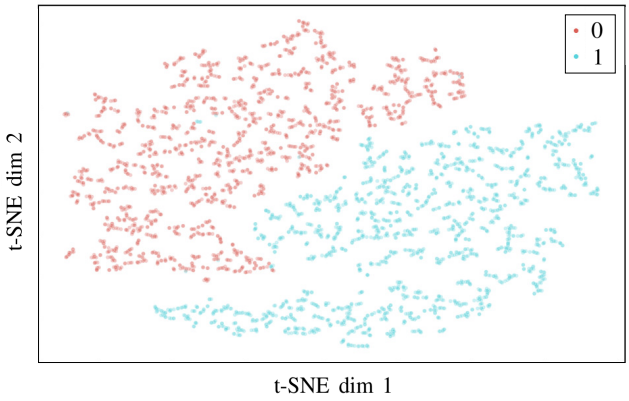

Figure 5: T-SNE visualization of 1D Power Spectrum on a random subset from *Faces-HQ* data set. We used a perplexity of 4 and 4000 iterations to produce the plot.

100% classification accuracy in a supervised setting.

	80% (train) - 20% (test)		
# samples	SVM	Logistic Reg.	K-Means
4000	100%	100%	82%
1000	100%	100%	82%
100	100%	100%	81%
20	100%	100%	75%

Table 2: Faces-HQ: Test accuracy using SVM, logistic regression and k-means under different data settings.

	80% (train) - 20% (test)		
# samples	SVM	Logistic Reg.	K-Means
2000	100%	100%	96%
100	100%	95%	100%
20	100%	85%	100%

Table 3: CelebA: Test accuracy using SVM, logistic regression and k-means.

	80% (train) - 20% (test)	
# samples	SVM	Logistic Reg.
2000	85%	78%
1000	82%	76%
200	77%	73%
20	66%	76%

Table 4: FaceForensics++: Test accuracy using SVM classifier and logistic regression classifier under different data settings. Evaluated on single frames.

2. Spectral Regularization on Auto-Encoder

In this second section, we show some additional results from our AE experiments (see figure 4 of the paper).

2.1. Loss during Training

Figure 6 shows the evaluation of the loss (see equations 10 and 11 in the paper) with and without spectral regularization for a decoder with 3 convolutional layers and 3 filters of kernel size 5×5 each.

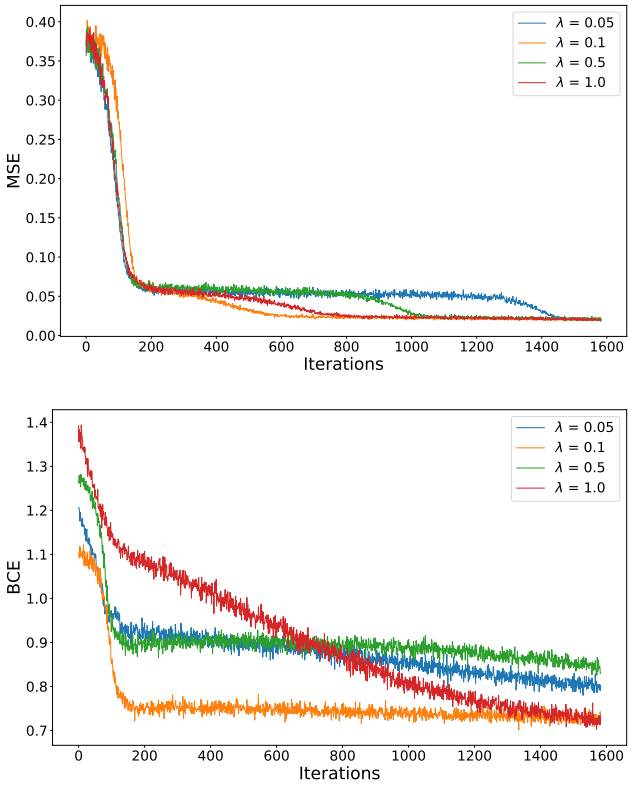

Figure 6: Evolution of the different losses that define the $\mathcal{L}_{\text{final}}$ from our AE. **Top:** Mean Square Error (MSE) during the training ($\mathcal{L}_{\text{Reconstruction}}$). **Bottom:** Binary Cross-Entropy loss (BCE) during the training ($\mathcal{L}_{\text{Spectral}}$).

These results show that the spectral regularization also has a positive effect on the convergence of the AE and the quality of the generated output images (in terms of MSE).

2.2. Effect of the Spectral Regularization

Figure 7 shows the impact of the spectral regularization on the AE problem. We can notice how both *transconv* and *up+conv* suffer from different behaviour on the frequency spectrum domain, specially in high frequency components. Nevertheless, after applying our spectral regularization technique, the results get much closer to the real 1D Power Spectrum distribution, generating images closer to the real distribution.

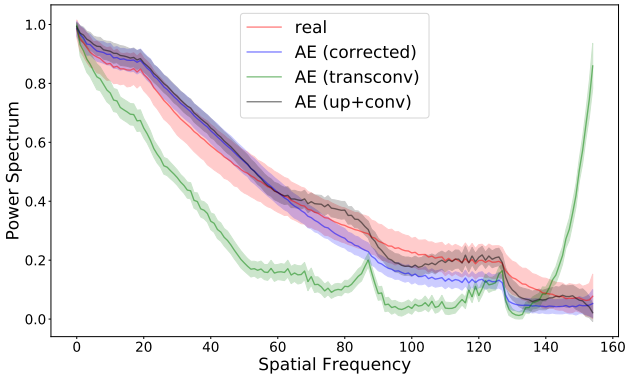

Figure 7: AE results for the baselines (*transconv* and *up+conv*) and for the proposal with spectral loss (*corrected*). The corrected AE has 3 additional convolutional layers after the last *transconv* layer. Each layer has 32 filters of size 5x5 and $\lambda = 0.5$

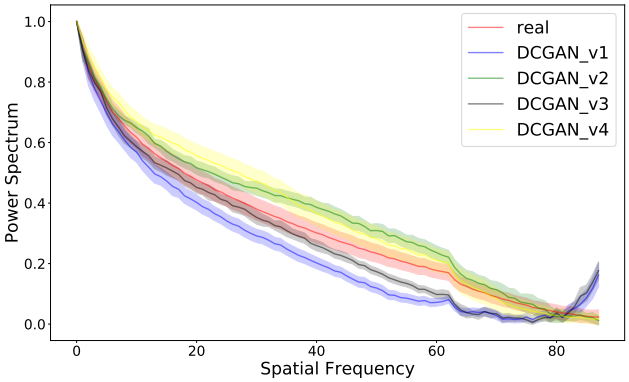

Figure 8: AE results for different topologies applied to DCGAN. Each version incorporates different amounts of convolutional layers to its DCGAN structure.

2.3. Effect of different Topologies

In this experiment, we evaluate the impact of different topology design choices. Figure 8 shows statistics of the spectral distributions for some topologies:

- Real: original face images from CelebA
- DCGAN_v1: a DCGAN topology with spectral regularization and one convolution layer (32 5x5 filters) after the last two up-convolutions.
- DCGAN_v2: a DCGAN topology with spectral regularization and two convolution layers (32 5x5 filters) after the last up-convolution.
- DCGAN_v3: a DCGAN topology with spectral regularization and one convolution layer (32 5x5 filters) after the every up-convolution.
- DCGAN_v4: a DCGAN topology with spectral regularization and three convolution layers (32 5x5 filters) after the last up-convolution.

Following the theoretical analysis and after a rough topology search for verification, we conclude that it is sufficient to add 3 5x5 convolutional layers after the last up-convolution in order to utilize the spectral regularization.

References

[1] 100,000 faces generated. `https://generated.photos/`.

[2] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[3] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[4] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[5] N. Kodali, J. Abernethy, J. Hays, and Z. Kira. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017.

[6] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[7] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

[8] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[9] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.